MWE 2026

The 22nd Workshop on Multiword Expressions (MWE 2026)

Proceedings of the Workshop

March 28, 2026

The MWE organizers gratefully acknowledge the support from the following organizations.

Gold

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-363-0

Introduction

The 22nd Workshop on Multiword Expressions (MWE 2026) took place on 28th March 2026, in Rabat, Marocco, collocated with the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026).
MWE 2026 is supported by the Special Interest Group on the Lexicon (SIGLEX) of the Association for Computational Linguistics (ACL), SIGLEX's Multiword Expressions Section (SIGLEX-MWE), and UniDive COST Action CA21167.
The notion of multiword expressions (MWEs) encompasses a range of closely related phenomena: idioms, compounds, light-verb constructions, phrasal verbs, rhetorical figures, collocations, institutionalized phrases, etc. They exhibit lexical, syntactic, semantic, pragmatic, and/or statistical idiosyncrasies. Given their irregular nature, MWEs often pose complex problems in various natural language processing (NLP) tasks such as language understanding and machine translation, hence still represent an open issue for computational linguistics.
For the 22nd edition of the workshop, our call for papers focused particularly on the following topics:

- Annotation (expert, crowdsourcing, automatic) and representation in resources such as corpora, treebanks, e-lexicons, WordNets, constructions (also for low-resource languages);
- Processing in syntactic and semantic frameworks (e.g. CCG, CxG, HPSG, LFG, TAG, UD, etc.);
- Evaluation of annotation and processing techniques;
- Discovery and identification methods, including for specialized languages and domains such as clinical or biomedical NLP;
- Interpretation of MWEs and understanding of text containing them;
- Language acquisition, language learning, and non-standard language (e.g. tweets, speech);
- Computationally-applicable theoretical work in psycholinguistics and corpus linguistics;
- Processing for end-user applications (e.g. MT, NLU, summarisation, language learning, etc.);
- Implicit and explicit representation in pre-trained language models and end-user applications;
- Evaluation and probing of pre-trained language models;
- Resources and tools (e.g. lexicons, identifiers) and their integration into end-user applications;
- Adaptation and transfer of annotations and related resources to new languages and domains including low-resource ones.

For this edition, 75% of the submitted papers were accepted.
The dominating trend in MWE research presented at the workshop focuses on large language models, particularly examining their capabilities and limitations across diverse tasks. Multiple studies compare performance of task-specific fine-tuned models to general-purpose generative LLMs on MWE identification tasks. Domain-specific applications (in the technical domain, bioinformatics, etc.) reveal how MWE extraction methods can be tailored to specialised contexts. Several papers identify systematic features that current LLMs fail to capture with respect to MWEs: cognitive studies reveal distinct processing signatures for different MWE types that computational models miss, while semantic interpretation tasks expose fundamental limitations in achieving fine-grained understanding of compound meaning and implicit relations between constituents.
The workshop demonstrates substantial multilingual coverage, addressing languages generally underrepresented in MWE research and exposing cross-linguistic variation in MWE behavior. Papers present

research on Swedish, Ukrainian, Romanian, Galician, Marathi, Turkish, Korean, Sinhala, and Chinese. Several studies explicitly employ the PARSEME annotation framework to enable cross-linguistic comparison, discussing the methodological challenges in ensuring annotation consistency and inter-annotator agreement across languages.
Research focuses on language-specific phenomena that challenge existing annotation schemes and identification methodologies, particularly in morphologically complex and typologically distinct languages. Studies on agglutinative and postpositional languages identify constructions that are difficult to categorise within existing frameworks. Low-resource language studies discuss limitations in handling culturally-specific phenomena, rare expressions and MWEs absent in training corpora, a challenge that persists despite advances in LLMs.
The MWE 2026 Workshop hosted two shared tasks: PARSEME 2.0, whose objective is to identify and paraphrase MWEs in written text, and AdMIRe 2 (Advancing Multimodal Idiomaticity Representation), which explores the comprehension ability of multimodal models for MWEs in a variety of languages.

Verginica Barbu Mititelu, Mathieu Constant, A. Seza Doğruöz, Atul Kr. Ojha, Alexandre Rademaker, Ivelina Stoyanova (MWE-2026 Organizers and Co-Chairs)

Organizing Committee

Workshop Chairs

Mathieu Constant, Université de Lorraine, CNRS, ATILF
A. Seza Doğruöz, Ghent University
Ivelina Stoyanova, Institute for Bulgarian Language, Bulgarian Academy of Sciences
Verginica Barbu Mititelu, Romanian Academy Research Institute for Artificial Intelligence
Atul Kr. Ojha, Insight Research Ireland Centre for Data Analytics, DSI, University of Galway, Ireland and Panlingua Languague Processing LLP, India
Alexandre Rademaker, School of Applied Mathematics of Getulio Vargas Foundation

Program Committee

Abigail Walsh, Dublin City University
Agata Savary, Université Paris-Saclay
Ahmet Erdem, Istanbul Technical University
Alberto Barrón-Cedeño, Università di Bologna
Ali Azmoudeh, Istanbul Technical University
Andrea Horbach, Leibniz Institute for Science and Mathematics Education
Andrei Tiberiu Carp, Tomorrow University of Applied Sciences
Anna Hülsing, Christian-Albrechts-Universität Kiel
Atakan Site, Istanbul Technical University
Barış Bilen, Istanbul Technical University
Beata Trawinski, Leibniz Institute for the German Language
Bora Şenceylan, Istanbul Technical University
Carlos Ramisch, LIS - Laboratoire d'Informatique et Systèmes
Chikara Hashimoto, Rakuten Institute of Technology
Cristea Alexandru-Marian, University of Bucharest
Cvetana Krstev, University of Belgrade, Faculty of Philology
David Cotigă, University of Bucharest
Debora Ciminari, University of Bologna
Doğukan Arslan, Istanbul Technical University
Elif Bayraktar, Istanbul Technical University
Emmanuele Chersoni, The Hong Kong Polytechnic University
Eric G C Laporte, Université Gustave Eiffel
Gaël Dias, University of Caen Normandy
Gražina Korvel, Vilnius University
Gülşen Eryiğit, Istanbul Technical University
Irina Lobzhanidze, Ilia Chavchavadze State University
Irina Moise, University of Bucharest
Ismail El Maarouf, Imprevicible
Ivelina Stoyanova, Deaf Studies Institute
Jan Odijk, Utrecht University
John Philip McCrae, University of Galway
Kenneth Church, Northeastern University
Kubilay Kağan Kömürcü, Istanbul Technical University
Laura A. Michaelis, University of Colorado at Boulder
Le Qiu, The Hong Kong Polytechnic University
Manfred Sailer, Johann Wolfgang Goethe Universität Frankfurt am Main

Manon Scholivet, Université Paris-Saclay
Maria Mitrofan, Research Institute for Artificial Intelligence
Mathieu Constant, Université de Lorraine, CNRS, ATILF
Matthew Shardlow, The Manchester Metropolitan University
Meghdad Farahmand, University of Genoa
Mehmet Utku Colak, Istanbul Technical University
Miriam Butt, Universität Konstanz
Monika Czerepowicka, University of Wamia and Masuria
Muhammed Abdullah Gümüş, International Technological University
Nina Hosseini-Kivanani, RTL
Oğuz Ali Arslan, Istanbul Technical University
Özge Umut, Istanbul Technical University
Oguzhan Karaarslan, Istanbul Technical University
Paul Cook, University of New Brunswick
Petya Osenova, Sofia University St. Kliment Ohridski
Ranka Stanković Stanković, University of Belgrade
Rares-Alexandru Roscan, University of Bucharest
Sabine Schulte im Walde, University of Stuttgart
Sergiu Nisioi, University of Bucharest
Shiva Taslimipoor, University of Cambridge
Stan Szpakowicz, University of Ottawa
Stella Markantonatou, ATHENA RIC
Tugce Temel, Istanbul Technical University
Tiberiu Boros, Adobe Systems
Tunga Gungor, Bogazici University
Veronika Vincze, University of Szeged
Yu-Yin Hsu, The Hong Kong Polytechnic University
Yunus Karatepe, International Technological University

Table of Contents

Program

Saturday, March 28, 2026

09:00 - 09:15 *Welcome and Introduction to 22nd MWE Workshop*

09:15 - 09:45 *Findings of the MWE 2026 Shared Tasks*

Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions
Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Mitrofan and Vasile Pais

MWE-2026 Shared Task: AdMIRe 2 Advancing Multimodal Idiomaticity Representation
Doğukan Arslan, Rodrigo Wilkens, Wei He, Dilara Torunoglu Selamet, Thomas Pickard, Aline Villavicencio, Adriana Silvina Pagano and Gülşen Eryiğit

09:45 - 10:30 *Oral Session*

Swedish Multiword Expression Corpora in PARSEME
Sara Stymne, Astrid Berntsson Ingelstam and Eva Pettersson

Cognitive Signatures of Multi-Word Expressions: Reading-Time and Surprisal
Diego Alves, Sergei Bagdasarov and Elke Teich

Diversity patterns run deep: Impact of diversity intake on multiword expression identification
Mathilde Deletombe, Manon Scholivet, Louis Estève, Thomas Lavergne and Agata Savary

10:30 - 11:00 *Coffee Break*

11:00 - 12:00 *Poster Session*

Large Language Models Put to the Test on Chinese Noun Compounds: Experiments on Natural Language Inference and Compound Semantics
Le Qiu, Emmanuele Chersoni, He Zhou and Yu-Yin Hsu

SinFoS: A Parallel Dataset for Translating Sinhala Figures of Speech
Johan Nevin Sofalas, Dilushri Pavithra, Nevidu Jayatilleke and Ruvan Weerasinghe

Ukrainian Multiword Expressions Corpus: Creation, Annotation, and Linguistic Analysis
Hanna Sytar, Maria Shvedova and Olha Kanishcheva

Saturday, March 28, 2026 (continued)

DCSN-NLP at MWE-2026 AdMIRe 2: Bridging Literal and Figurative Meaning Through Hierarchical Multimodal Reasoning
David Cotigă and Sergiu Nisioi

ITUNLP at MWE-2026 AdMIRe 2: A Zero-Shot LLM Pipeline for Multimodal Idiom Understanding and Ranking
Atakan Site, Oğuz Ali Arslan and Gülşen Eryiğit

Archaeology at WE-2026 PARSEME 2.0 Subtask 1 and 2: Parsing is for Encoders, Paraphrasing is for LLMs
Rares-Alexandru Roscan and Sergiu Nisioi

ITUNLP2 at MWE-2026 AdMIRe 2: Modular Zero-Shot Pipelines for Multimodal Idiom Grounding and Ranking
Özge Umut and Bora Şenceylan

12:00 - 12:20 *Community discussion*

12:20 - 12:30 *Concluding Remarks*

Large Language Models Put to the Test on Chinese Noun Compounds: Experiments on Natural Language Inference and Compound Semantics

Le QIU and Emmanuele Chersoni and He ZHOU and Yu-yin HSU
Department of Language Science and Technology, The Hong Kong Polytechnic University
11 Yuk Choi Road, Hung Hom, Kowloon, Hong Kong, China
lani.qiu@connect.polyu.hk,
{emmanuele.chersoni,he.zhou,yu-yin.hsu}@polyu.edu.hk

Abstract

Noun compounds are generally considered an open challenge for NLP systems, given to the difficulty of interpreting the implicit semantic relation between modifier and head, although the advent of Large Language Models (LLMs) recently led to remarkable performance leaps. However, most evaluations have been carried out on English benchmarks.

In our work, we test LLMs on compound semantics understanding in Chinese, adopting two different evaluation scenarios: an extrinsic evaluation in a Natural Language Inference task, and an intrinsic evaluation in which models are directly asked to predict the semantic relation linking the two constituents.

Our results show that the bigger and more recent LLMs are able to surpass supervised baselines in the inference task, especially when tested under the few-shot setting. In the more challenging task of selecting the correct interpretation of the compounds out of a fine-grained typology of semantic relations between head and modifier, the best Chinese LLM (Qwen-plus) manages to select the correct option in about one third of the cases.

1 Introduction

Noun-noun compounds are ubiquitous in natural languages, and they notoriously represent a challenge for NLP applications due to the ambiguity of the implicit semantic relation linking the two nouns, the modifier and the head (Nakov, 2008b; Libben, 2014). The correct interpretation of a compound may be essential for the correct understanding of the semantics of a sentence, and for the appropriateness of an automatic translation: when an English speaker hears about a *carrot cake*, s/he should understand that the cake *is made* of carrot; when a Chinese speaker hears about a 爱情故事 (*love story*), s/he should understand that the story is *about* love. Significantly, native speakers are able to identify similar relationships even in compounds that have never been met before (Van Jaarsveld and Rattink, 1988), with entities having similar semantic features, which explains why compounding is a very productive mechanism for creating novel words. NLP evaluation generally focused on eliciting a plausible paraphrasing of a noun compound from the models, typically in the form of a verb phrase (e.g. *flu virus* → *virus that causes flu*) (Nakov, 2008a; Butnariu et al., 2009; Hendrickx et al., 2013; Shwartz and Waterson, 2018; Shwartz, 2019; Coil and Shwartz, 2023; Rambelli et al., 2024), and mainly using English as the language of study.

In our work, we test the understanding of Chinese noun compound semantics in current LLMs. Our evaluation is first carried out in an *extrinsic* task, where models are required to grasp the meaning of the compound to perform natural language inference (NLI) (Bowman et al., 2015); and then in an *intrinsic* task, where they are asked to select a semantic relation from a limited inventory, representing the link between modifier and head. We observed that, while the best Chinese LLMs and GPT-4 perform similarly for the NLI task (even beating supervised models when prompted with few-shots), the former are more accurate in selecting specific, human-like semantic interpretations, with Qwen-plus achieving the top performance. [1]

2 Related Work

Some studies in the Chinese NLP literature tackled the challenge of interpreting noun compounds (Wang et al., 2010; Gu et al., 2016; Wang et al., 2016), but they share the main limitation that their evaluation datasets were not made available. The study of Liu et al. (2022) adopted a hybrid approach to the interpretation problem, first employing a classifier to identify the relations of compound nouns,

[1] Code and data are available at https://github.com/Laniqiu/zh.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 1–7

and later utilized a paraphrasing model to interpret those that were labeled with an arbitrary undefined relation. Liu and colleagues did release their benchmark, a dataset in the life service domain containing 1,478 compounds with annotated relation labels. However, the number includes different types of compounds, such as for example adjective-noun compounds, so that the total number of actual noun-noun compounds available for evaluation is relatively small. Moreover, the labels in the dataset refer to the type of meaning carried by the modifier, rather than to the relationship between the modifier and the head: for example, in 生日礼服 (*birthday dress*), the labeled relation is `time`, as the modifier indicates the occasion on which the dress is worn.

Using the noun compounds in Liu and colleagues' data, Zhou et al. (2024) adopted a template-based approach to generate a NLI dataset, where the premise always contains a noun compound and the hypothesis label (`entailment`, `neutral` or `contradiction`) depends on the correct understanding of the compound meaning. Using a total of 66 templates on 625 of the compounds from Liu et al. (2022), they obtained an evaluation dataset of 3,740 premise-hypothesis pairs. Some examples of the dataset items are shown below:

(1) 前提: 运动员有一个不锈钢饭盒。
假设: 不锈钢是饭盒的制作材料。
类别: 蕴含
Premise: The athlete has a stainless steel lunch box.
Hypothesis: Stainless steel is the material that the lunch box is made of.
Category: Entailment

(2) 前提: 清洁工昨天吃了巧克力蛋糕。
假设: 清洁工吃的蛋糕里没有巧克力。
类别: 矛盾
Premise: The janitor ate chocolate cake yesterday.
Hypothesis: There was no chocolate in the cake that the janitor ate.
Category: Contradiction

Although their study only tested relatively small models (i.e. Qwen and Chinese Alpaca in their 7B parameter versions), they found that such models already perform competitively with fine-tuned encoders (i.e. BERT and RoBERTa).

3 Experimental Settings

3.1 Evaluation Datasets

We ran our LLM evaluation on two datasets. The first one is **NCNLI**, a NLI dataset introduced by Zhou et al. (2024): it includes 3,740 premise-hypothesis pairs, 1,564 labeled as 'entailment', 1,092 as 'contradiction' and 1,084 as 'neutral'.

The second one is a newly-constructed dataset for noun compound interpretation in Chinese. The data are noun compounds extracted from the *New Era People's Daily Corpus* (Huang and Wang, 2019), after applying POS Tagging with Jieba. By definition, a noun-noun compound consists of two nouns standing next to each other. A preliminary list of such compounds was automatically extracted, and then filtered by one of the authors (a native speaker of Mandarin Chinese with a PhD in Computational Linguistics) to exclude cases of POS ambiguity and tagger error. This left us with 2,083 compounds in total. Henceforth we refer to this dataset as **NEPD**, to indicate the original source of the data.

To determine the compounding relation of each word, we recruited three graduate students in Chinese linguistics for the annotation. Specifically, we predefined 11 semantic categories of compounding relations: *CAUSE*, *MAKE*, *HAVE*, *USE*, *BE*, *IN*, *FOR*, *FROM*, *ABOUT*, *AND*, *OR*[2], using the hierarchy constructed by Liu and Liu (2019). Prior to annotation, annotators received training on the guidelines and examples. Each annotator was asked to assign one or more semantic relations to each compound. If none of the predefined categories were deemed as appropriate, the annotators were instructed to select the *OTHER* label.

Each compound was annotated by three annotators, and their input was reviewed by a more experienced linguist and annotator (one of the authors of this study) for additional quality checking. We assigned each compound the majority relation, that is, the relation on which at least two of the annotators agreed. To assess consistency between annotations, we used the Jaccard similarity coefficient to measure the overlap between pairs of annotators: this metric calculates the percentage of labels selected by both annotators out of all labels selected by either one of them. On average, we obtained a coefficient value of 0.412, indicating a moderate level of agreement in the task.

[2]Definition for each category will be given in the prompt.

Compounds for which no dominant relation could be identified (i.e. those for which the three annotators chose three different relations) were discarded. As a result, the final dataset comprised 1,514 compounds. Some examples are in Table 1, while relation frequencies can be seen in Table 2.

Compounds	Relation (s)
岛国 (*island country*), 水草 (*water plant*)	IN
风雨 (*wind and rain*), 书画 (*painting and calligraphy*)	AND
中国画 (*Chinese painting*), 民间舞 (*folk dance*)	FROM, ABOUT

Table 1: Example compounds with full agreement (first 2 rows) and partial agreement (the last row). Agreement statistics can be found in Table 6 of the Appendix.

Relation	Frequency	Majority
CAUSE	133	30
MAKE	454	116
HAVE	774	162
USE	137	27
BE	476	79
IN	679	178
FOR	1501	430
FROM	367	71
ABOUT	1327	356
AND	193	56
OR	76	9
OTHER	212	24

Table 2: Frequency of semantic relations in the NEPD data (note that compounds can be annotated with multiple relations) and their frequency as majority relation.

3.2 Models and Settings

We tested a pool of smaller (i.e. around a 7 billion parameter size) and larger Chinese LLMs on both task: **Qwen-7B** (Bai et al., 2023), **Chinese Alpaca 7B** (Cui et al., 2023), **DeepSeek-7B** (Bi et al., 2024) and **Qwen2.5-7B** (Yang et al., 2024) (all of them in their instruction-tuned versions) were tested on our server, while **Qwen-plus** and **DeepSeek-chat** were queried via the online interfaces. Additionally, **GPT-4o-mini**[3] for the sake of comparison with one of the most capable and popular Western models. The prompts we crafted can be found in the Appendix.

For comparison with pretrained supervised models on the NLI task, we reimplement the Chinese BERT- (Devlin et al., 2019; Cui et al., 2019, 2020) and RoBERTa-based (Liu et al., 2019; Cui et al., 2019, 2020) baselines from Zhou et al. (2024).

[3] https://platform.openai.com/docs/models/gpt-4o-mini.

3.3 Metrics

For the NLI task, we evaluate models in terms of standard **Accuracy** and **F1-Macro** score. For compound interpretation, we use both **Accuracy** (the number of times the model output exactly the majority relation for the target compound, divided by the total number of samples) and **R- Rank** (Camacho-Collados et al., 2018), defined as:

$$R - rank = \frac{1}{n}\sum_{i=1}^{n} rank_i \quad (1)$$

where n is the total number of samples, while $rank_i$ is the rank of the majority relation for the i-th compound sample. Since the correct relation may not always appear in the prediction list, we add **Hit Ratio** (Alsini et al., 2020) as a supplementary metric. Hit@k (or hit ratio @k) is the proportion of test cases in which the correct item appears within the top k positions in the model ranking.

4 Results

A general summary of the results can be seen in Table 3, including both the scores for the LLMs on the two datasets (3a) and the performance of the fine-tuned baselines on the NCNLI data (3b). On NCNLI, the best LLMs in a zero-shot setting are close to the fine-tuned RoBERTa baseline: only the more recent models of the Qwen family are able to consistently surpass it. Perhaps surprisingly, the smaller Qwen2.5-7B model is the one getting the highest accuracy and F1-score in this setting. A different trend becomes visible under the a few-shot setting: while smaller models seem to be inconsistent, bigger LLMs show clear gains from exposure to task examples. GPT-4o-mini, Qwen-plus and DeepSeek-chat all see noticeable boosts, and particularly GPT-4o-mini, which achieves the best score overall. Among the small models, Qwen2.5-7B keeps being competitive, but does not have any gain from few shots. In other words, only bigger models seem to be able to consistently perform in-context learning from the additional examples.

The interpretation task, as expected, is more challenging, given the high number of semantic relations to choose from and the subtle nature of the compound intepretation. A noticeable figure is the

	NCNLI, 0-shot		NLI, 3-shot		NEPD		
	Acc	F1	Acc	F1	Acc	R-rank	Hit@5
Alpaca 7B	71.35	64.54	52.14	42.00	2.01	8.29	18.54
Qwen 7B	64.15	51.44	64.74	56.05	6.74	11.08	2.03
DeepSeek-7B	55.98	45.73	63.84	50.57	9.91	7.97	38.78
Qwen2.5-7B	77.90	78.26	76.52	77.17	13.28	10.25	15.17
GPT-4o-mini	68.29	69.35	**80.41**	**80.24**	8.23	5.08	69.22
DeepSeek-chat	71.00	70.58	74.64	74.59	23.38	5.83	60.74
Qwen-plus	76.67	76.91	79.03	79.67	**36.13**	**4.42**	**73.34**

(a) Results with LLM prompting.

	Acc	F1
BERT	57.70	52.59
RoBERTa	72.94	71.49

(b) NCNLI results with baselines from Zhou et al. (2024), fine-tuned on 50k examples from the OCNLI dataset (Hu et al., 2020).

Table 3: Evaluation results. Only valid outputs counted for the metrics.[4] All scores are reported as the average of 3 runs and reported in %, except for R-rank. Best scores on each dataset are in **bold**.

value of the hit ratio@5: it is clear that small models have a hard time in this task, as they all fall short of ranking the correct semantic relation in the top 5 in most cases, and have accuracy scores mostly in single digits (notice that a random baseline with uniform probability distribution would get around 8.3% of correct answers). A remarkable improvement in the hit ratio can be seen with bigger models, but with an important distinction: while GPT-4o-mini manages to include the right answers at the top of the rank in most cases (over 69%), its accuracy and R-rank are not significantly better than those of the smaller models. This suggests that, while GPT-4o-mini does a better job in selecting plausible semantic relations for the intepretation of a compound, it still struggles in identifying the correct ones within a pool of plausible options.

On the other hand, the two Chinese competitors achieve better scores for those metrics. Qwen-plus is particularly impressive, managing to align with humans on the most plausible relation in about one third of the cases and to achieve a very low R-rank value, suggesting that the correct option is almost always close to the top of the rank. If we consider smaller and bigger models separately, it is interesting to notice that in both "categories" a model of the Qwen family emerges as the best performing one across the two tasks. As it can be seen from Table 4, bigger models tend to predict more frequent semantic relations more accurately, with weak-to-moderate positive correlations between accuracy and relation frequency; two relatively rare relations such as OR and FROM are more challenging for most LLMs (average accuracy < 3%, cf. Table 5).

[4]For the NLI task, Alpaca 7B rejected 11.41% of samples in both settings, Qwen 7B rejected 0.09% (zero-shot) and 0.02% (few-shot); others showed no rejections. For relation interpretation, rejection rates were 1.43% for Alpaca 7B, 0.48% for Qwen 7B, and 0.07% for Qwen-plus, and none for others.

Model	Correlation
Alpaca 7B	-0.30
Qwen 7B	0.12
DeepSeek-7B	-0.03
Qwen2.5-7B	0.20
GPT-4o-mini	0.19
DeepSeek-chat	0.39
Qwen-plus	0.41

Table 4: Spearman correlation index between relation frequency (as majority relation) and accuracy.

Relation	Average Performance		
	Acc	R-rank	Hit@5
CAUSE, MAKE	> 40%	< 2	> 60%
OR, FROM	< 3%	≈ 4	< 15%

Table 5: Hardest (top) and easiest (bottom) relation categories on average.

5 Conclusions

In our work, we evaluated LLMs on noun compounds semantics using two different tasks: a NLI task where the understanding of the inference depends on the correct interpretation of the compound, and a task focusing on identifying the specific semantic relation existing between modifier and head, for which we collected a new dataset.

As for NLI, we found LLMs to be already improving over the performance of pretrained models, with larger LLMs taking the most advantage from task examples in the few-shot setting. Selecting the correct relation in the interpretation task is more challenging, given the ambiguity of noun compounds and the greater number of classes to choose from. As in the NLI task, we observed the most consistent performance from Qwen models, with Qwen-plus being the most aligned with human intuitions of compound semantics.

Limitations

An important limitation of the study lies in our dataset for compound interpretation, since we aimed at recruiting annotators with a high level of expertise (PhD students in linguistics) and, as a consequence, we had a relatively low number of annotators (3) for each dataset instance. Although the annotations were quality checked by one of the authors, who has expert level knowledge of the subject, our choice might have favored some idiosyncratic interpretation of the compounds.

As a term of comparison for Western LLMs we used GPT-4o-mini, which proved to be a cost-efficient and performance-effective option. However, we did not have the time to test more recently-released models, such as GPT-5.

Acknowledgements

EC was supported by a GRF grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 15612222).

References

Areej Alsini, Du Q Huynh, and Amitava Datta. 2020. Hit Ratio: An Evaluation Metric for Hashtag Recommendation. *arXiv preprint arXiv:2010.01258*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609*.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, and 1 others. 2024. Deepseek LLM: Scaling Open-source Language Models with Longtermism. *arXiv preprint arXiv:2401.02954*.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of EMNLP*.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid O Séaghdha, Stan Szpakowicz, and Tony Veale. 2009. SemEval-2010 Task 9: The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*.

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. SemEval-2018 Task 9: Hypernym Discovery. In *Proceedings of SemEval*.

Jordan Coil and Vered Shwartz. 2023. From Chocolate Bunny to Chocolate Crocodile: Do Language Models Understand Noun Compounds? In *Findings of ACL*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *Findings of EMNLP*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-Training with Whole Word Masking for Chinese BERT. *arXiv preprint arXiv:1906.08101*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and Effective Text Encoding for Chinese Llama and Alpaca. *arXiv preprint arXiv:2304.08177*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Min Gu, Yanhui Gu, Fang Xu, Bin Li, Bin Zhao, and Weiguang Qu. 2016. Research on Chinese Noun+Noun Compounds Semantic Classification and Automatic Interpretation. *ICIC Express Letters. Part B, Applications: An International Journal of Research and Surveys*, 7(1):173–179.

Iris Hendrickx, Preslav Nakov, Stan Szpakowicz, Zornitsa Kozareva, Diarmuid O Séaghdha, and Tony Veale. 2013. SemEval-2013 Task 4: Free Paraphrases of Noun Compounds. *Proceedings of SemEval*.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kuebler, and Larry Moss. 2020. OCNLI: Original Chinese Natural Language Inference. In *Findings of EMNLP*.

Shuiqing Huang and Dongbo Wang. 2019. Construction, Performance and Application of New Era People's Daily Segmented Corpus (I) – Construction and Evaluation Corpus. *Library and Information Service*, 63(22):5–12.

Gary Libben. 2014. The Nature of Compounds: A Psychocentric Perspective. *Cognitive Neuropsychology*, 31(1-2):8–25.

Jingping Liu, Juntao Liu, Lihan Chen, Jiaqing Liang, Yanghua Xiao, Huimin Xu, Fubao Zhang, Zongyu Wang, and Rui Xie. 2022. Noun Compound Interpretation with Relation Classification and Paraphrasing. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8757–8769.

Pengyuan Liu and Yujie Liu. 2019. Semantic Relations Hierarchy and Knowledge Base Construction of Chinese Basic Noun Compounds. *Journal of Chinese Information Processing*, 33(4):20–28.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Preslav Nakov. 2008a. Noun Compound Interpretation Using Paraphrasing Verbs: Feasibility Study. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 103–117. Springer.

Preslav Nakov. 2008b. Paraphrasing Verbs for Noun Compound Interpretation. In *Proceedings of the LREC Workshop on Multiword Expressions*.

Giulia Rambelli, Emmanuele Chersoni, Claudia Collacciani, and Marianna Bolognesi. 2024. Can Large Language Models Interpret Noun-Noun Compounds? A Linguistically-Motivated Study on Lexicalized and Novel Compounds. In *Proceedings of ACL*.

Vered Shwartz. 2019. A Systematic Comparison of English Noun Compound Representations. In *Proceedings of the ACL Workshop on Multiword Expressions*.

Vered Shwartz and Chris Waterson. 2018. Olive Oil Is Made of Olives, Baby Oil Is Made for Babies: Interpreting Noun Compounds Using Paraphrases in a Neural Model. In *Proceedings of NAACL*.

Henk J Van Jaarsveld and Gilbert E Rattink. 1988. Frequency Effects in the Processing of Lexicalized and Novel Nominal Compounds. *Journal of Psycholinguistic Research*, 17:447–473.

Meng Wang, CR Huang, Shiwen Yu, Bin Li, and 1 others. 2010. Chinese Noun Compound Interpretation based on Paraphrasing Verbs. *(Journal of Chinese Information Processing)*, 24(6):3–9.

Meng Wang, Lulu Wang, Na Tian, and Bin Li. 2016. Automatic Interpretation of Chinese Noun Compounds Based on Word Similarity. *ICIC Express Letters, Part B: Applications*, 7(6):1215–1221.

An Yang, Baosong Yang, B Zhang, B Hui, B Zheng, B Yu, Chengpeng Li, D Liu, F Huang, H Wei, and 1 others. 2024. Qwen2.5 Technical Teport. *arXiv preprint arXiv:2412.15115*.

He Zhou, Yu Yin Hsu, and Emmanuele Chersoni. 2024. Evaluating Chinese Noun Compound Interpretation in Natural Language Inference. In *Proceedings of the Chinese Lexical Semantics Workshop (CLSW 2024)*.

Appendix

Prompts

5.0.1 NCNLI

For the NLI task, the prompt follows a fixed template for both settings:

> 请将下面前提（P）和假设（H）之间存在的逻辑推理关系分为以下类别之一：蕴含,矛盾或中立。只需回答类别。
> P: xxx H: xxx
>
> Please identify the semantic relation between the premise (P) and the hypothesis (H) and respond with one of the following semantic relations: Entailment, Contradiction or Neutral. Return their relation only.
> P: xxx H: xxx

In the few-shot setting, we concatenate a representative example for each category to the template above, while ensuring that none of them overlaps with the evaluation set.

> 前提: 运动员有一个不锈钢饭盒。
> 假设: 不锈钢是饭盒的制作材料。
> 输出: 蕴含
>
> **P**: The athlete has a stainless steel lunch box.
> **H**: Stainless steel is the material that the lunch box is made of.
> **Output**: Entailment
>
> 前提: 清洁工昨天吃了巧克力蛋糕。
> 假设: 清洁工吃的蛋糕里没有巧克力。
> 输出: 矛盾
>
> **P**: The janitor ate chocolate cake yesterday.
> **H**: There was no chocolate in the cake that the janitor ate.
> **Output**: Contradiction
>
> 前提: 科学家最喜欢的是椒盐牛蛙。
> 假设: 科学家只吃过椒盐口味的牛蛙。
> 输出: 中立
>
> **P**: The scientist's favorite is salt-and-pepper bullfrogs.
> **H**: The scientist has only eaten bullfrogs with a salt-and-pepper flavor.
> **Output**: Neutral

5.0.2 NEPD

For the NEPD task, the prompt is derived from a concise summarization of the guidelines and instructions we provided to human annotators.

> 给定一个中文复合词语，该词语由两个名词复合构成，请对其名词成分之间的语义关系进行分类，共11个预定义类别，分别为：CAUSE（表示因果关系）、MAKE（表示组成）、HAVE（表示拥有、具备）、USE（表示

使用、利用工具或手段）、BE（表示说明和补充）、IN（表示空间上的包含关系）、FOR（表示目的、用途）、FROM（表示来源）、ABOUT（表示主题或相关内容）、AND（表示并列、组合关系）、OR（表示选择或替代关系）。请仅返回最可能的类别名称，并按可能性从高到低排序。若该词语不属于上述任何类别，请返回OTHER。
复合词: xx

Given a Chinese compound word formed by two nouns, classify the semantic relationship between its noun components into one of 11 predefined categories: CAUSE (indicating causal relation), MAKE (indicating composition), HAVE (indicating possession or having), USE (indicating usage or utilization of tools or means), BE (indicating explanation or description), IN (indicating spatial inclusion), FOR (indicating purpose or function), FROM (indicating source, origin, or starting point), ABOUT (indicating topic or related content), AND (indicating coordination or combination), OR (indicating choice or alternative). Please return only the most likely category names, ordered from highest to lowest likelihood. If the compound word does not belong to any of the above categories, return OTHER.
Compound: xx

Dataset Statistics

	Count
Fully agreed	458
Partially agreed	1,056
Tied	569
Total	2,083

Table 6: Annotation agreement statistics.

We categorized the annotation results into three groups based on the level of inter-annotator agreement among three annotators for each compound. 'Fully agreed' indicates consensus among all three annotators; 'partially agreed' indicates consensus between two annotators with the third disagreeing; and 'tied' indicates three distinct judgement with no agreement between any pair. In addition, we calculated that each annotator provided an average of 2.10 relations, indicating that the annotation task involves a considerable level of difficulty.

SINFOS: A Parallel Dataset for Translating Sinhala Figures of Speech

Johan Sofalas[a], **Dilushri Pavithra**[a], **Nevidu Jayatilleke**[b] **and Ruvan Weerasinghe**[a]
[a]Research Department, Informatics Institute of Technology, Sri Lanka
{johan.s, pavithra.r, ruvan.w}@iit.ac.lk,
[b]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
nevidu.25@cse.mrt.ac.lk

Abstract

Figures of Speech (FoS) consist of multi-word phrases that are deeply intertwined with culture. While *Neural Machine Translation* (NMT) performs relatively well with the figurative expressions of high-resource languages, it often faces challenges when dealing with low-resource languages like Sinhala due to limited available data. To address this limitation, we introduce SINFOS, a dataset of 2,344 Sinhala figures of speech with cultural and cross-lingual annotations. We examine this dataset to classify the cultural origins of the FoS and to identify their cross-lingual equivalents. Additionally, we have developed a binary classifier to differentiate between two types of FoS in the dataset, achieving an accuracy rate of approximately 92%. We also evaluate the performance of existing LLMs on this dataset. Our findings reveal significant shortcomings in the current capabilities of LLMs, as these models often struggle to accurately convey idiomatic meanings. By making this dataset publicly available, we offer a crucial benchmark for future research in low-resource NLP and culturally aware machine translation.

1 Introduction

Language and culture are deeply interrelated and significant mutual influence in multiple ways (Hamidi, 2023). FoS are the tools that make language expression more vivid, attractive, and effective (Regmi, 2015). They are built through a small set of meaning-construction mechanisms where speakers reuse familiar knowledge structures in new contexts (Dancygier and Sweetser, 2014). Speakers utilise various figurative forms, such as exaggeration and idioms, as they often achieve discourse goals more effectively than literal words (Roberts and Kreuz, 1994). While idioms are universal, each language features unique expressions with specific meanings, complicating the translation process and creating a sophisticated challenge (Medagama, 2021).

The Sinhala language is part of the Indo-Aryan branch of the Indo-European language family with a rich and diverse literary heritage that has evolved over several millennia. It uses a unique script that is derived from the ancient Indian Brahmi script (Jayatilleke and de Silva, 2025b). The origins of Sinhala can be traced back to between the 3rd and 2nd centuries BCE. Sinhala is the primary language of the Sinhalese people, who make up the largest ethnic group in Sri Lanka, and it is recognised as the first language (L1) for approximately 16 million individuals (De Silva, 2025; Jayatilleke and de Silva, 2025a). According to the criteria established by Ranathunga and de Silva (2022), Sinhala is classified as a lower-resourced language (Category 2).

Sinhala has a long and well-documented tradition of FoS (භාෂා අලංකාර \ bʰɑːʃɑː ʌlʌnkʌrʌ) that appears in both literary and everyday communication (Senaveratna, 2005). They emerged gradually as Sinhala speakers and writers needed brief ways to support religious, educational, and courtly objectives, communicate indirectly and memorably in everyday conversation, and enhance the aesthetic quality of their poetry (Nawaz et al., 2025; Mieder, 1997). Currently, Sinhala FoS are mainly preserved in collections such as books and dictionaries, with many manuscripts held by national institutions and temples (Mieder, 1997). In this study, we present SINFOS [1], the first Sinhala dataset of its kind with essential data to support the task of machine translation (target language: English).

2 Related Works

A substantial body of research has examined FoS, including idioms (Sporleder et al., 2010), metaphors (Dodge et al., 2015), proverbs (Bonin et al., 2017), and other forms of figurative language (Kabra et al., 2023).

[1] https://huggingface.co/datasets/SloppyCalculator/SinFoS

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 8–26

2.1 Existing FoS Corpora

Resources are predominantly English-focused, whereas a smaller subset provides broader multilingual coverage, including European Portuguese, Danish, Chinese, and multi-language compilations such as MABL and ID10M. (Kabra et al., 2023; Tedeschi et al., 2022). The datasets ranged in size from moderate idiom/proverb collections, small lexicons (hundreds to 1,000 items) (Zhou et al., 2021; Moussallem et al., 2018), to (1̃,000–10,000) (Stowe et al., 2022; Reddy et al., 2011), with a few large-scale corpora (tens of thousands of instances/pairs or even larger textual corpora) (Zheng et al., 2019; Krennmayr and Steen, 2017). Moreover, a limited number of datasets, such as Adewumi et al. (2022), have a multi-phenomenon architecture that covers a greater variety of figurative categories, whereas many datasets are single-phenomenon resources that primarily target idioms or metaphors (Sporleder et al., 2010; Dodge et al., 2015; Prochnow et al., 2024; Shaikh et al., 2024).

Shaikh et al. (2024) introduce KonIdioms[2], an annotated Konkani idiom corpus (4,332 sentences and 817 potentially idiomatic expressions) designed to support automatic idiom identification and evaluation for this low-resource language. Furthermore, the PARSEME [3] dataset release 1.3 provides multilingual annotations of *Verbal Multiword Expressions* (VMWEs) across Arabic, Bulgarian, Chinese, Croatian, Greek, Hebrew, Hindi, Irish, Latvian, Lithuanian, Maltese, Slovene, and Turkish languages, including a dedicated category for verbal idioms alongside other VMWEs types (Savary et al., 2023). In contrast, the SemEval-2022 Task 2 dataset[4] by Tayyar Madabushi et al. (2022) focuses on idiomaticity-related modelling through sentence-level evaluation in English, Portuguese, and Galician, supporting tasks such as idiom detection and representation learning. Additionally, IMIL[5]introduces an Idiom Mapping for Indian Languages resource that links idioms across Bengali, Gujarati, Hindi, Marathi, Punjabi, Tamil, and Urdu (with English mappings), enabling cross-lingual comparison and transfer for idiom processing (Agrawal et al., 2018).

It is clear that datasets related to FoS are a significant area of focus for researchers in the field, including languages like Konkani (Shaikh et al., 2024), which falls under the same language resource category (Category 2) as Sinhala (Ranathunga and de Silva, 2022). We have also discussed various existing FoS datasets for different languages in detail in Appendix A.

2.2 Classification of FoS

Many studies have classified FoS into multiple categories, each supported by explicit definitions (Banou et al., 2025). Jang et al. (2023) categorised the FLUTE [6] dataset into four categories, such as sarcasm, similes, idioms, and metaphors. Early work, such as the *SemEval-2015 Task 11* by Ghosh et al. (2015) and the discourse-oriented analysis by Musolff (2017), primarily focused on the interplay between sentiment and specific tropes, particularly irony, sarcasm, and metaphor, in social media and public discourse. Moreover, Chakrabarty et al. (2021a) redefined figurative language data as instances of *Recognising Textual Entailment* (RTE), structuring sentence pairs that comprise a premise and a hypothesis with an associated entailment label, by drawing on five pre-existing datasets (Figurative-NLI[7] (Chakrabarty et al., 2020), datasets on *irony* compiled by Van Hee et al. (2018)[8] and Ghosh et al. (2020)[9], Sarcasm SIGN[10] (Peled and Reichart, 2017), a metaphor dataset[11] by Chakrabarty et al. (2021b)) annotated for simile, metaphor, and irony, thereby constructing a corpus of more than 12,500 RTE examples. Hayani (2016) has classified the figurative texts into 12 categories, such as metaphor, personification, hyperbole, simile, metonymy, synecdoche, irony, antithesis, symbolism, and paradox.

2.3 LLMs based Machine Translations

As mentioned by Pramodya (2023), NMT systems for low-resource, morphologically rich languages such as Sinhala increasingly adopts transfer learning and fine-tuning of multilingual sequence-to-sequence LLMs rather than SMT. As mentioned by Thillainathan et al. (2025), systematic pretraining on monolingual data followed by intermediate-task transfer provides better results than conventional single-stage fine-tuning of multilingual LLM-based MT systems in Sinhala-to-English translation. Despite these advancements, translating figurative lan-

[2] https://bit.ly/3Y4LGd3
[3] http://hdl.handle.net/11372/LRT-5124
[4] https://bit.ly/4s8k4Bt
[5] https://bit.ly/4p4SUsC
[6] https://huggingface.co/datasets/ColumbiaNLP/FLUTE
[7] https://github.com/tuhinjubcse/Figurative-NLI
[8] https://competitions.codalab.org/competitions/17468
[9] https://bit.ly/44D3O1q
[10] https://github.com/lotemp/SarcasmSIGN
[11] https://bit.ly/4rfWrGc

guage remains a challenging task. While retrieval-augmented prompting can improve the translation of idioms by offering helpful definitions or context (Donthi et al., 2025), comparative analyses show that, compared to human translations, outputs from LLMs often lack cultural nuance and tend to simplify creative metaphors (Sahari et al., 2024; Karakanta et al., 2025).

Based on existing studies, it is evident that Sinhala figurative language is underexplored in the field of computational linguistics. Incorporating this resource by identifying the dominant semantic and cultural domains reflected in Sinhala figurative language, along with translating these data from Sinhala to English, will be significant for future research. Therefore, the purpose of this work is to present a dataset of Sinhala figurative language, capture its cultural nuances, and provide an essential resource for the task of machine translation from Sinhala to English.

3 Data Collection and Annotation

The SINFOS dataset consists of 2,344 unique FoS and was compiled from a carefully curated selection of authoritative resources, including various Sinhala literary works and selected Wikipedia entries. This section provides a detailed overview of the processes involved in assembling, annotating, and preprocessing the data. An example of a record from the dataset that underwent these steps is illustrated in Figure 4 in Appendix D.

3.1 Data Assembly

A significant portion of the data, approximately 65%, was sourced from the prominent Sinhala books in this field. වාග්සම්ප්‍රදා \ vagsʌmprʌðʌ - Idioms (Department of Official Languages), අතීත වාක්‍ය දීපනිය \ ʌθi:θʌ va:kyʌ ði:pʌnɪjʌ - Atheetha Wakya Deepanya (Senanayaka, 1880), and the Dictionary of Proverbs of the Sinhalese (Senaveratna, 2005), while the remaining 35% was extracted from Wikipedia [12].To ensure high fidelity to the source material, the core Sinhala expression was collected as the primary data entry. This is a foundational practice validated by benchmarks like the IDIX (Sporleder et al., 2010) and the ChID (Zheng et al., 2019) corpora, which rely on the collection of specific linguistic expressions as the base unit for identification.

[12] https://bit.ly/4qdZy08

3.2 Annotation Process

To ensure the accuracy of the sources, the annotation process closely followed the resources outlined in subsection 3.1 and was carried out by native Sinhala speakers. Importantly, when primary sources lacked the expected information related to translations (although the attributes *Literal / Visual Image* and *Type of FoS* involved some human annotation as detailed in subsections 3.2.1 and 3.2.2), the annotators refrained from using personal knowledge to avoid potential subjective interpretations. Instead, they strictly drew from previously verified resources. For example, *What it really implies* was derived directly from the *Corresponding FoS in English* found in the source books, utilising standard references such as Merriam-Webster (Dictionary, 2002) and the Cambridge Dictionary (Brown et al., 2013) for validation. Similarly, missing *Literal Image* entries were translated strictly from the FoS text, while *Type of FoS* categories were assigned based solely on the logical frameworks outlined in subsection 4.1 and Appendices B, C. A final comprehensive review confirmed that all entries were grounded in these external standards, ensuring high data integrity. As a result of the procedures followed, certain records did not include some attributes, as shown in Table 1.

Attribute	Count
Sinhala (සිංහල \ sɪŋhʌlʌ)	2344
Type of FoS	2344
Literal / visual image	2344
Corresponding FoS in English	1571
What it really implies	2059
Additional Context	125

Table 1: Distribution of annotated fields in the dataset.

3.2.1 *Type of FoS*

To clarify the figurative language associated with each record, the dataset includes a "*Type of FoS*" attribute. This granularity was essential for determining the distinct processing strategies required for different figurative types, a necessity highlighted by the PIE corpus (Adewumi et al., 2022), which classifies data into specific types like metaphors and similes, and the IMPLI study (Stowe et al., 2022), which demonstrates that models process idioms and metaphors differently.

The entries are organised into five main categories, as detailed in Table 2. Most of the idioms were obtained from (Department of Official Lan-

Type of FoS	Number of Entries
Proverbs (ප්‍රස්තාපිරුළු \ prʌsθapɪrulu)	988
Idioms (වාග්සම්ප්‍රදා \ vagsʌmprʌðʌ)	1319
Adages (ආප්තෝපදේශ \ ɑ:pθɔ:pʌðe:ʒʌ)	15
Idiosyncratic (පුද්ගලික \ puðgʌlɪkʌ:ʒʌ)	11
Sayings (කියමන් \ kɪjʌmʌn)	11

Table 2: Distribution of Entries by Figure of Speech Type.

guages), while the majority of the proverbs were gathered from (Senaveratna, 2005). For certain FoS, specific types of FoS annotations were readily available, allowing us to directly categorise them within our classification strategy and document them accordingly. The remaining FoS were annotated based on the criteria outlined in subsection 4.1. The guidelines provided in Appendix C were used to distinguish between proverbs and idioms. Additionally, proverbs were categorised into three subcategories based on their intent, origin, and conclusion. These annotations were performed according to the criteria in Appendix B. Proverbs were assigned tags corresponding to the three categories mentioned earlier, while the other types of figurative speech were labelled directly, using their Sinhala names.

3.2.2 *Literal / Visual Image*

SINFOS uses a "*Literal / Visual Image*" annotation for each entry to provide a visual reference for non-native speakers by eliminating all abstract concepts, emotions, and symbolism. Documenting the literal imagery aligned with psycholinguistic research on imageability and methodologies for testing compositionality. Since the majority of these expressions are figurative, capturing the mental image was highly necessary. Furthermore, the inclusion of the implied meaning provided the ground truth required to test a model's ability to transcend surface definitions, mirroring the "real vs. false definition" methodology of the *Danish Idiom Dataset* (Sørensen et al., 2025).

Majority of the annotation was done using the above given sources as the relevant visual details were provided by them, whilst the others were annotated by translating the Sinhala FoS, word by word (e.g.,එක හඬින් \ ɛkʌ hʌndɪn as "With one voice"). The annotation process adhered to precise guidelines for aligning words, ensuring direct correspondence between the nouns and verbs in the original Sinhala text and their English descriptions. To maintain a "Semantic Ground Truth" and avoid introducing an outside context, only tangible objects and specific actions were documented. Furthermore, non-translatable "cultural objects" were preserved in their original form. For example වැඩි පද ගහන්නේ තොවිලෙ කැතවෙන්නයි \vædɪ pʌðʌ gʌhʌnne: θɒvɪle: kæθʌvɛnnʌj was annotated as "Too much tom-toming means that the tovila is going to be spoilt", retaining the word "තොවිලෙ \ θɒvɪlɛ - Tovila (devil-ceremony, exorcism)". This method helps prevent "translation loss" and ensures that the dataset's literal accuracy is preserved, avoiding misleading interpretations that could arise from forced or inaccurate translations of culturally specific items.

3.2.3 *Corresponding FoS in English*

The attribute "*Corresponding FoS in English*" refers to the equivalent English figurative expression (FoS) for its Sinhala counterpart. One of the techniques explored by translators is direct substitution, which effectively facilitates the understanding of figurative language across different languages, even without explicit meanings (Adelnia and Dastjerdi, 2011). This process further enabled the identification of cross-lingual equivalence and cultural parallels, a parallel alignment approach that was validated through the cross-linguistic mapping of proverbs in *PROMETHEUS* (Özbal et al., 2016) and the alignment protocols of *ParaDiom* (Donaj and Antloga, 2023).

The FoS obtained from Department of Official Languages included corresponding English FoS for all entries, whereas Senaveratna (2005) provided corresponding English FoS for only some entries, which were used for annotation. Additionally, the process of annotating this data also aided in determining the "*What it really implies*" aspect for certain FoS.

3.2.4 *What it really implies*

The "*What it really implies*" column was established to clearly explain Sinhala figurative phrases in English, capturing their deeper meaning. It translates each Sinhala figurative expression into a shared human experience. Given that recognition of FoS is highly context-dependent, additional context is included to assist in disambiguation and cultural grounding. This field captures terms specific to Sinhalese culture, regional variations, and the folklore or stories behind specific figures of speech, ensuring the dataset serves as a comprehensive resource for understanding the "naked truth" behind

the language. This is supported by the context-dependent annotation standards of *EPIE* (Saxena and Paul, 2020) and the cultural analysis frameworks of *PROMETHEUS* (Özbal et al., 2016).

To maintain clarity in the data and prevent lengthy explanations, the annotation process prioritised brevity over excessive detail. Only essential translations were included, omitting additional context or details that could complicate data analysis. Most implications in the expressions were derived directly from primary reference sources mentioned in the subsection 3.1. However, when a corresponding English equivalent was identified, the meaning was modified to align with the common interpretation of that English idiom. To guarantee reliable data, entries lacking a source-based explanation or an English equivalent were excluded. This mitigates the risk of inaccuracies or subjective misinterpretations. The annotations adhere to a specific format to aid in computational modelling. Behavioural advice and actions are expressed in the infinitive form. Character types or scenarios are described in formal terms. By eliminating secondary imagery and metaphorical elements, this approach clarifies the meaning for non-native speakers. It offers a clear "ground truth " for comparing the literal interpretation of a phrase with its actual significance.

3.3 Data Pre-processing

During the pre-processing stage, meticulous attention was devoted to punctuation, particularly in the context of FoS. The retention of punctuation marks in these instances is crucial, as they play a significant role in determining both prosody and syntactic structure, which are essential for achieving accurate processing. To ensure this dataset does not leak important information about figurative language, no further word-level or sentence-level filtration was conducted on any records, including those containing stereotypes, to facilitate authentic cultural analysis and the study of historical societal norms.

4 Analysis of SINFOS

The SINFOS dataset comprises 2,344 FoS, totalling 8,903 words. The literal image section includes 14,383 words, while the "What it really implies" section has 19,386 words. On average, each Sinhala FoS consists of 3.798 words. A brief overview of the dataset statistics is shown in Table 3.

Category	N	Mean	Median	Max	Total
Sinhala FoS	**2344**	**3.80**	**3**	**24**	**8903**
Literal / visual image	2344	6.14	5	38	14383
What it really implies	2059	9.41	8	56	19366
Corresponding FoS in English	1571	3.44	3	21	5401

Table 3: Summary statistics of word counts across different categories.

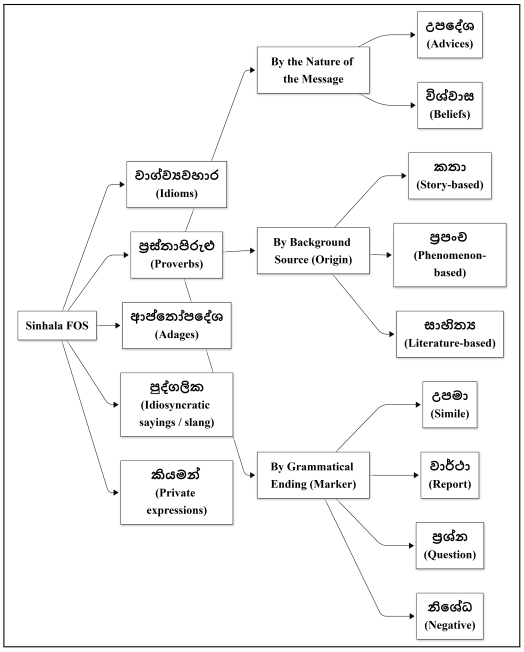


Figure 1: Summary of Sinhala FoS Dataset Classification

4.1 Classification of FoS

The classification of Sinhala FoS (භාෂා අලංකාර \ bʰɑ:ʃɑ: ʌlʌnkʌrʌ) is complex due to the fluidity of the language and its deep rooting in oral tradition. As mentioned in the subsection 3.2.1, this study classified Sinhala FoS into five main categories. The etymological roots of these terms provide a necessary framework for understanding their usage.

වාග්සම්ප්‍රදා \ vagsʌmprʌðʌ (Sinhala idioms): Derived from the Sanskrit roots " වාග්/වාක් \ vɑ:g/vɑ:k " (speech/word) and " සම්ප්‍රදාය \ sʌmprʌðɑ:jʌ " (tradition/heritage), this term refers to speech patterns established by long-standing usage. Unlike proverbs, which are often wisdom-based, these are usage-based constructs where the meaning transcends the literal definitions of the individual words. These are typically incomplete phrases or fragments, often ending in a verb. For example ආවාට ගියාට \ ɑ:vɑ:tʌ gɪjɑ:tʌ literally translates to "For coming and going" while it actually means "not friendly, and showing

little interest in other people in a way that seems slightly rude".

ප්‍රස්තාපිරුළු \ prʌsθapɪrulu (Sinhala proverbs): This is a compound of " ප්‍රස්තා \ prʌsθa: " denoting a specific occasion, moment, or opportunity, and " පිරුළු \ pɪrulu " referring to a simile, reply, or adage. Consequently, this functions as a "situational simile", a pre-packaged linguistic unit invoked to address a specific incident by comparing it to a known truth. In contrast to Sinhala idioms, Sinhala proverbs are syntactically complete sentences or clauses that can stand alone. For example ඉඟුරු දීලා මිරිස් ගන්නා වගෙයි \ ɪnguru ði:lɑ: mɪrɪs gʌθθɑ: vʌgɛj (Like exchanging ginger for chili). To provide a granular analysis, Sinhala proverbs were further classified based on the nature of the message, the source of the background, and the grammatical ending as mentioned in Appendix B.

ආප්තෝපදේශ \ ɑ:pθɔ:pʌðe:ʒʌ (Sinhala adages): Unlike figurative proverbs, these are literal directives. They represent the prescriptive aspect of the language (what one should do), distinct from the descriptive nature of idioms. An example of adages in SINFOS is ඉගෙනීම නොනැසෙන ධනයකි \ɪgʌni:mʌ nɒnæsɛnʌ ðʰʌnʌjʌkɪ (Education is an indestructible form of wealth).

පුද්ගලික \ puðgʌlɪkʌ:ʒʌ (Idiosyncratic): These are hyper-local sayings used by individuals or small groups. While not yet FoS in the public domain (Crocker, 1977), they represent the genesis point of language evolution, where personal metaphors potentially graduate into public idioms over time. Slang also falls under this category. For example the phrase අභිධර්ම මුදලාලිගෙ හෝටලේ වගේ \ ʌbʰɪðʰʌrmʌ mʊðʌlɑ:lɪge: hɔ:tʌle: vʌge: (Like Abhidharma mudalali's hotel) would be well understood by the people living in the surroundings but not by everyone.

කියමන් \ kɪjʌmʌn (Sinhala sayings): Concise verbal phrases are commonly used in daily conversation to express a thought, comment, or observation. In contrast to proverbs or idioms, these do not inherently possess a moral lesson, universal truth, or established figurative interpretation recognised by a large group. As an example මරුවා ආ දාට බාදා නැහිලු \ mʌru:vɑ: ɑ: ðɑ:tʌ ba:ðɑ: næθɪlu: (When death comes, there is no let or hindrance).

The dataset primarily consists of Sinhalese proverbs and idioms, leading to the creation of a binary classification model aimed at distinguishing between proverbs and idioms. A Voting Ensemble model, incorporating Support Vector Machines (SVM), Random Forest, and XGBoost with TF-IDF Character 3-Gram vectorisation, achieved an impressive accuracy of 90.56%. This approach, based on character-level processing, effectively tackled the intricacies of Sinhala morphology (Priyanga et al., 2017) by detecting subword elements rather than relying solely on exact phrases. The implementation of Word2Vec embeddings significantly improved performance compared to experiments based on TF-IDF (sparse vector representation). This includes the accuracy of the TF-IDF Character 3-Gram in both the Gaussian Naive Bayes and Linear SVC models, achieving an accuracy of 90.34% in each case. The analysis indicated that specific verb endings served as strong indicators of idiomatic expressions, while comparative particles and rhythmic consonant clusters were associated with proverbs. Incorporating 3-gram TF-IDF was used to leverage the identified patterns, resulting in models with these embeddings performing better than their word-level counterparts. The semantic understanding provided by dense embeddings, such as Word2Vec, also proved effective in recognising these patterns. Ultimately, utilising a Deep Feed Forward Neural Network (Deep NN), which offers superior semantic understanding, achieved the highest overall accuracy of 92.7% and the best recall for proverbs at 94%. The embeddings for the LSTM and Deep NN models are not specified in Table 8, as they relied on the standard TensorFlow Keras embeddings that learned directly from the training data.

Model	Vectorization	Accuracy	P-Rec.	I-Rec.
Gaussian Naive-Bayes	Word2Vec	90.34%	92%	89%
LinearSVC	Word2Vec	90.34%	90%	91%
Random Forets	TF-IDF (Char 3)	89.27%	83%	94%
XG-Boost	TF-IDF (Char 3)	90.13%	86%	93%
Ensemble (SVC+RFM+XGB)	TF-IDF (Char 3)	90.56%	85%	95%
Bi-LSTM	-	91.63%	86%	96%
Deep NN	-	92.7%	94%	92%

Table 4: Model Performance Comparison. Further details in Appendix F. *Note that P-Rec = Recall for Proverbs and I-Rec = Recall for Idioms.

4.2 Cultural Analysis

This research employed a hybrid methodological approach that combined both inductive and deductive thematic analysis to explore the relationship between physical imagery and cultural significance

in Sinhala FoS. This computational analysis was conducted on English translations of the dataset. The analysis identified two main aspects of the FoS: "*Literal / Visual Image*" (Source Domain), which consists of the tangible visual components that make up the figure of speech, and "*What it Really Implies*" (Target Theme), which signifies the deeper abstract or cultural meanings conveyed by the text. To minimise researcher bias and ensure that the coding frameworks were derived from raw data rather than from preconceived notions, we emphasised a bottom-up discovery phase. This inductive stage employed unsupervised machine learning methods to uncover naturally occurring patterns. Specifically, we applied TF-IDF vectorisation (using unigrams and a maximum of 2,000 features) along with K-Means clustering (k=5) to analyse the "What it Really Implies" dimension and uncover hidden linguistic clusters.

Additionally, we conducted a frequency analysis using a *Bag-of-Words* (BoW) model for both the "*Literal / Visual Image*" and "*What it Really Implies*" dimensions. This analysis allowed us to identify the most frequent and significant terms in each cluster, categorising specific words under different themes and establishing a data-driven basis for the theoretical coding frameworks. After completion of the exploratory phase, the recognised patterns were compiled into an organised dictionary for the deductive phase. We employed a rule-based classification system, using the specific keywords identified in the earlier phase as indicators of broader cultural categories. The algorithm compared the text against this predefined dictionary; if a keyword associated with a certain category was found, that category was assigned to the entry. This approach enabled multi-label classification, assuming that the subject matter remained consistent across the figurative language, thereby confirming that the detected keywords were suitable representations of the main concepts.

Lastly, a bivariate cross-tabulation was performed to quantitatively evaluate the connections and dependencies between the identified Source Domains and Target Themes. The findings reveal that Somatic (Body) and Agrarian (Nature) imagery are the most prevalent source domains, with notable mentions of the hand (n=56), water (n=46), and trees (n=43). The most frequently encountered themes are Ethics & Moral Character (n=162) and Karma & Consequence (n=127). This suggests a distinct metaphorical framework in which nature-related metaphors primarily promote moral conduct (n=20), while physical imagery specifically illustrates the tangible repercussions of karmic consequences (n=14). The distribution of literal source domains and abstract cultural themes observed in SINFOS is summarised in Table 7 in Appendix E. This implies that these FOS primarily serve as mechanisms for reinforcing social norms rather than simply providing descriptive observations.

4.3 Cross-Lingual Equivalence Analysis

This study investigates a collection of 1,571 Sinhala phrases that have English "Literal/ Visual Image" translations. This sample is derived from the initial dataset of 2,344 phrases, as the remaining 773 lack direct English equivalents. The findings indicate a notable cultural divergence, demonstrated by a symbolic overlap score of merely 0.05 using the Jaccard Index and a lexical similarity score of 0.32. The lexical similarity was calculated using the sequence matcher in the `difflib` [13] library, which employs the *Ratcliff/Obershelp Algorithm* (Ratcliff and Metzener, 1988). This implies that although the functional meanings align, the underlying metaphors originate from distinct contexts.

For example, Sinhala employs the expression "exchanging ginger for chilli," while English phrases refer to "jumping out of the frying pan into the fire." In terms of structure, 93.3% of the phrases retain their original form, while 4.9% transition from Sinhala similes into English metaphors. An illustration is "Like the eye," which transforms into "Apple of one's eye."

Furthermore, expressions in Sinhala are, on average, 32% longer than their English counterparts, yielding a ratio of 1.32. This distinction is effectively showcased by the English phrase "red herring," which in Sinhala translates to an elaborate depiction where "the fox conceals the fowl in the forest and scurries about, swinging a coconut husk from its mouth."

5 Benchmarking on LLMs

In this section, we use SINFOS as a benchmark to evaluate the performance of selected LLMs and *Small Language Models* (SLMs) in translating these complex expressions. A subset of 499 FoS was curated based on specific criteria: they represent diverse categories and possess intricate meanings that are particularly challenging for mod-

[13] https://bit.ly/4p48y7o

System/Context:

You are an expert linguist specialising in Sinhala (Sri Lanka) language and folklore.

Task:

I will provide a list of Sinhala Figures of Speech. For each item, provide only the English Figurative Meaning (what it really implies in a specific context). Do not translate literally. Do not explain the literal words.

Figure 2: Prompt used to generate responses from LMs.

els to interpret (Tayyar Madabushi et al., 2022). To ensure a comprehensive evaluation, we employed stratified sampling, purposefully oversampling rare categories, such as adages (11), "private" expressions (10), and sayings (3), which are often overshadowed by dominant idioms (190) and proverbs (285). This approach allows for a robust assessment of model capabilities across the full spectrum of figurative language, prioritising interpretative difficulty to test the distinction between literal cues and cultural nuances (Tayyar Madabushi et al., 2022). Furthermore, proverbs were broken down into their core elements (story, nature, and literature) to better analyse the depth of cultural understanding.

We used the same prompt for all models to establish a consistent evaluation baseline. Figure 2 shows the prompt provided to the *Language Models* (LMs) to elicit the meanings of the FoS. This method helps avoid prompt-induced bias, as small variations in wording could unintentionally favour one LM over another, ensuring that the responses are directly comparable.

Model	Cosine Similarity	Fidelity Scores
Gemini 3 Pro	0.6678	0.3117
Llama 4 Maverick	0.6400	0.2351
Grok 4.1	0.6354	0.2361
GPT 5.2	0.6221	0.2090
DeepSeek-V3.2	0.6126	0.2052
Claude Sonnet 4	0.5972	0.1628
Gemma	0.6024	0.1914
GPT 4.1 mini	0.5816	0.1300
Qwen 3	0.5596	0.1247

Table 5: Performance of language models on Sinhala FoS.

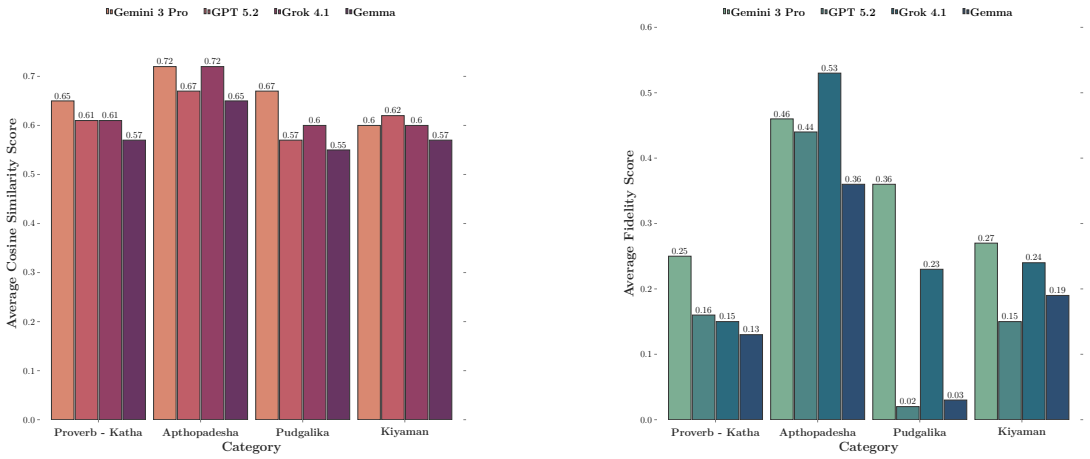
To evaluate how effectively LMs grasp FoS in Sinhala, this research employs a dual framework that examines both context retrieval and logical comprehension. This method reflects the two-step process of theme identification and truth condition mapping by Reimers and Gurevych (2019). The initial phase utilises a bi-encoder architecture with `FlagEmbedding` (specifically the *BAAI/bge-large-en-v1.5* [14] model) to calculate Cosine Similarity between the outputs of the model and the meanings annotated in the dataset. This model was selected for its state-of-the-art performance on the *Massive Text Embedding Benchmark* (MTEB), ensuring precise high-dimensional mapping that outperforms standard baselines in capturing "Semantic Relatedness" (Chen et al., 2024; Tayyar Madabushi et al., 2022).

Although this segment efficiently penalises thematic discrepancies, such as mixing "betrayal" with "love," it may be influenced by the "Keyword Bag" problem, in which comparable terms obscure gaps in logical coherence. For example, the idiom කොහොඹ ගහට කරවිල වැල ගියා වගෙයි \ kɒhɒmbʌ gʌhʌtʌ kʌrʌvilʌ vælʌ gɪjɑ: vʌge: which implies the compatible union of two negative forces (literal image: 'like the karawila creeper twining round the kohomba tree') received a high similarity score of 0.805 for the DeepSeek V3 translation, 'a mismatched or absurd pairing', despite the model's output conveying the exact opposite meaning.

To tackle this issue, the second segment measures the Fidelity Score, which implements a Cross-encoder (*stsb-roberta-large*) to evaluate intricate dependencies by analysing sentences concurrently (Reimers and Gurevych, 2019). In this context, Fidelity represents the semantic faithfulness of the model's output to the ground truth. This functions as a replacement for "Semantic Entailment," aiding in the differentiation between sentences that share similar phrasing but convey distinct meanings, such as "the dog bit the man" versus "the man bit the dog" (Li et al., 2024). By utilising the full self-attention mechanism of the Cross-encoder, the framework captures the syntactic nuances often missed by Bi-encoder models. Integrating this Fidelity Score with the first segment provides robust safeguards against "Low-Resource Hallucination," enabling a comprehensive assessment of Language Models in the Sinhala language (Benkirane et al., 2024).

At the same time, the Fidelity scores struggle with something known as the "Hyper-Literal" problem, where creative paraphrasing could be penalised. For example, the phrase බුරන බල්ලෝ හපාකන්නේ නැහැ \ bʊrʌnʌ bʌllɔ: hʌpɑ: kʌnne: næhæ is directly translated as "Barking dogs don't bite" by DeepSeek V3. In the case of translating FoS, substitution with a valid FoS

[14] https://huggingface.co/BAAI/bge-large-en-v1.5



(a) Cosine Similarity Score Comparison for Selected Categories

(b) Fidelity Score Comparison for Selected Categories

Figure 3: Benchmarking LLM Performance: (a) Cosine Similarity and (b) Fidelity Score. Information on all LLM performances could be found in Appendix G.

is considered to be a valid form of translation (Adelnia and Dastjerdi, 2011), but Fidelity gives it a modest score of 0.0089, as both phrases do not have lexical overlap. Relying only on one of these metrics can cause blind spots and skew evaluation results.

Therefore, by including both metrics, we can better assess the model's performance. This method identifies "hallucinated relevance," where high Cosine scores suggest understanding, but low Fidelity scores indicate a lack of grasp on underlying intent. This helps benchmark true understanding over mere statistical matching. Table 5 displays the average Cosine Similarity Scores and average Fidelity Scores obtained by each of these models across all the FoS available on the stratified sample obtained on the dataset based on the types of FoS, difficulty and figurativeness.

The assessment of nine advanced models reveals that `Gemini` stands out in its ability to analyse Sinhala FoS, achieving the top scores in Cosine Similarity and Fidelity. The success of the smaller `Gemma` model indicates that cultural relevance takes precedence over the model size. Nonetheless, there is an issue known as the "illusion of competence." Some models can effectively retrieve context but falter in logical comprehension. As a result, they may identify the correct domain but often misinterpret the meanings. Conventional metrics, such as BLEU, do not adequately address this challenge. Furthermore, models such as `GPT-4o mini` and `Qwen` exhibit "broken figurative triggers," offering literal interpretations instead of figurative ones for specific expressions. While most models perform well with sayings that align with Western proverbs, they tend to struggle with distinct and folklore-inspired proverbs. This stems from their literal approach to translation, which neglects the cultural context needed to understand nuances.

6 Conclusion

This study introduces SINFOS, a dataset containing 2,344 Sinhala FoS accompanied by expert-verified explanations. The annotation process is comprehensively explained in the paper. The available details were entered into the dataset, and the missing details were handled in a manner consistent with the structure of the entered details to ensure the dataset's accuracy and validity.

The analysis of the dataset emphasises a significant disparity in meaning between Sinhala idioms and their English equivalents. The cross-linguistic examination revealed the disparities among the languages, while the cultural analysis showcased the distinct culture reflected in the FoS, emphasising the challenges of translation. While LLMs can effectively handle FoS with direct English translations, they often struggle with culturally specific terminology. This can result in inaccuracies or literal conversions. Future research should focus on improving the verification of these results by implementing ablation studies and presenting statistical significance. Consequently, SINFOS serves as a vital resource for developing novel approaches in Machine Translation and modelling frameworks that seek to integrate cultural insights into languages with fewer resources.

Limitations

Sinhala meaning unavailability: A key limitation of this study is the incomplete availability of English meanings for some Sinhala FoS. In several cases, authoritative definitions or consensus interpretations were not available in accessible reference sources, which constrained some of the analysis, such as where cross-lingual analysis could not be performed across all the FoS, and the domains spoken by these FoS could not be analysed in the cultural analysis.

Meaning loss in English rendering: Some Sinhala FoS are highly culture-bound, context-dependent, or rely on implicit background knowledge, making direct English rendering difficult and increasing the risk of ambiguity or meaning loss. As a result, a portion of the dataset may contain paraphrased or approximate meanings rather than fully equivalent English interpretations, which can affect translation quality and downstream classification performance.

Class imbalance in පුද්ගලික \ puðgʌlikʌ:ʒʌ and කියමන් \ kɪjʌmʌn categories: The dataset exhibits class imbalance, particularly within the පුද්ගලික \ puðgʌlikʌ:ʒʌ and කියමන් \ kɪjʌmʌn categories, where only 11 instances were available for both categories. Therefore, the analysis done was heavily influenced by the dominant idioms and proverbs. A classification model could not be trained to classify all FoS due to the class imbalance.

References

Amineh Adelnia and Hossein Vahid Dastjerdi. 2011. Translation of idioms: A hard task for the translator. *Theory and Practice in Language Studies*, 1(7):879–883.

Tosin Adewumi, Roshanak Vadoodi, Aparajita Tripathy, Konstantina Nikolaido, Foteini Liwicki, and Marcus Liwicki. 2022. Potential idiomatic expression (PIE)-English: Corpus for classes of idioms. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 689–696, Marseille, France. European Language Resources Association.

Ruchit Agrawal, Vighnesh Chenthil Kumar, Vigneshwaran Muralidharan, and Dipti Sharma. 2018. No more beating about the bush : A step towards idiom handling for Indian language NLP. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Katsiaryna Aharodnik, Anna Feldman, and Jing Peng. 2018. Designing a Russian idiom-annotated corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Israa Alsiyat, Scott Piao, and Mansour Almansour. 2023. Arabic metaphor corpus (amc) with semantic and sentiment annotation. page 1. The twelfth International Corpus Linguistics Conference, CL2023 ; Conference date: 03-07-2023 Through 06-07-2023.

David Antunes, Jorge Baptista, and Nuno J. Mamede. 2025. A European Portuguese corpus annotated for verbal idioms. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 58–66, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Zouheir Banou, Sanaa El Filali, El Habib Benlahmar, Fatima-Zahra Alaoui, Laila El Jiani, and Hasnae Sakhi. 2025. A systematic review of figurative language detection: Methods, challenges, and multilingual perspectives. *Natural Language Processing Journal*, 13:100192.

Kenza Benkirane, Laura Gongas, Shahar Pelles, Naomi Fuchs, Joshua Darmon, Pontus Stenetorp, David Ifeoluwa Adelani, and Eduardo Sánchez. 2024. Machine translation hallucination detection for low and high resource languages using large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9647–9665, Miami, Florida, USA. Association for Computational Linguistics.

Patrick Bonin, Alain Méot, Jean-Michel Boucheix, and Aurélia Bugaiska. 2017. Psycholinguistic norms for 320 fixed expressions (idioms and proverbs) in french. *The Quarterly Journal of Experimental Psychology*, 71:1–37.

Edward Keith Brown, James Edward Miller, and James Edward Miller. 2013. *The Cambridge dictionary of linguistics*. Cambridge University Press.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021a. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020. Generating similes effortlessly like a pro: A style transfer approach for simile generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6455–6469, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021b. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4250–4261, Online. Association for Computational Linguistics.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

J Christopher Crocker. 1977. The social functions of rhetorical forms. *The social use of metaphor: Essays on the anthropology of rhetoric*, 2:33–66.

Barbara Dancygier and Eve Sweetser. 2014. *Figurative Language*. Cambridge Textbooks in Linguistics. Cambridge University Press, New York, NY, USA. Also available as paperback ISBN 978-0-521-18473-1 and PDF ISBN 978-1-107-77687-6.

Nisansa De Silva. 2025. Survey on publicly available sinhala natural language processing tools and research. *arXiv preprint arXiv:1906.02358*.

Department of Official Languages. *Idioms*. Department of Official Languages, Sri Lanka, Colombo, Sri Lanka.

Merriam-Webster Dictionary. 2002. Merriam-webster. *On-line at http://www. mw. com/home. htm*, 8(2):23.

Ellen Dodge, Jisup Hong, and Elise Stickles. 2015. MetaNet: Deep semantic automatic metaphor analysis. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 40–49, Denver, Colorado. Association for Computational Linguistics.

Gregor Donaj and Špela Antloga. 2023. ParaDiom: A parallel corpus of idiomatic texts. In *Text, Speech, and Dialogue*, page 147–158.

Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Doh, and Eid Rodan. 2025. Improving llm abilities in idiomatic translation. In *Proceedings of the 1st Workshop on Language-Oriented Research in SLMs (LoResLM)*. Also available as arXiv:2407.16470.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Eirini Florou, Konstantinos Perifanos, and Dionysis Goutsos. 2018. Neural embeddings for metaphor detection in a corpus of Greek texts. In *2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–4.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–478, Denver, Colorado. Association for Computational Linguistics.

Debanjan Ghosh, Elena Musi, and Smaranda Muresan. 2020. Interpreting verbal irony: Linguistic strategies and the connection to theType of semantic incongruity. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 82–93, New York, New York. Association for Computational Linguistics.

Sayan Ghosh and Shashank Srivastava. 2022. ePiC: Employing proverbs in context as a benchmark for abstract language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3989–4004, Dublin, Ireland. Association for Computational Linguistics.

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.

Souad Hamidi. 2023. The relationship between language, culture, and identity and their influence on one another. 3.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264, Dubrovnik, Croatia. Association for Computational Linguistics.

Risma Hayani. 2016. Figurative language on Maya Angelou selected poetries. *Script Journal: Journal of Linguistic and English Teaching*, 1:131.

Yusuke Ide, Joshua Tanner, Adam Nohejl, Jacob Hoffman, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. 2025. CoAM: Corpus of all-type multiword expressions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27004–27021, Vienna, Austria. Association for Computational Linguistics.

Hyewon Jang, Qi Yu, and Diego Frassinelli. 2023. Figurative language processing: A linguistically informed feature analysis of the behavior of language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9816–9832, Toronto, Canada. Association for Computational Linguistics.

Nevidu Jayatilleke and Nisansa de Silva. 2025a. Sidiac: Sinhala diachronic corpus. *arXiv preprint arXiv:2509.17912*.

Nevidu Jayatilleke and Nisansa de Silva. 2025b. Zero-shot OCR accuracy of low-resourced languages: A comparative analysis on Sinhala and Tamil. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 471–480, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Zhiying Jiang, Boliang Zhang, Lifu Huang, and Heng Ji. 2018. Chengyu cloze test. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 154–158, New Orleans, Louisiana. Association for Computational Linguistics.

Charles Jochim, Francesca Bonin, Roy Bar-Haim, and Noam Slonim. 2018. Slide - a sentiment lexicon of common idioms. In *International Conference on Language Resources and Evaluation*.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Aji, Genta Winata, Samuel Cahyawijaya, A. Aremu, Perez Ogayo, and Graham Neubig. 2023. Multilingual and multi-cultural figurative language understanding. pages 8269–8284.

Alina Karakanta, Mayra Nas, and Aletta G. Dorst. 2025. Metaphors in literary machine translation: Close but no cigar? In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 276–286, Geneva, Switzerland. European Association for Machine Translation.

Muhammad Farmal Khan and Mousumi Akter. 2025. Evaluating large language models on Urdu idiom translation. *Preprint*, arXiv:2510.17460.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. SemEval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47, Atlanta, Georgia, USA. Association for Computational Linguistics.

Tina Krennmayr and Gerard Steen. 2017. *VU Amsterdam Metaphor Corpus*, pages 1053–1071. Springer Netherlands, Dordrecht.

Murathan Kurfalı, Robert Östling, Johan Sjons, and Mats Wirén. 2020. A multi-word expression dataset for Swedish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4402–4409, Marseille, France. European Language Resources Association.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18601–18609.

Chaya Liebeskind and Yaakov HaCohen-Kerner. 2016. A lexical resource of Hebrew verb-noun multi-word expressions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 522–527, Portorož, Slovenia. European Language Resources Association (ELRA).

Changsheng Liu and Rebecca Hwa. 2017. Representations of context in recognizing the figurative and literal usages of idioms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31.

Thisiri Medagama. 2021. Idiomatic language complexities in translation with special reference to sinhalese and english. *Journal of Research in Humanities and Social Science*.

Wolfgang Mieder. 1997. Modern paremiology in retrospect and prospect. *Paremia*, 6:399–416.

Diego Moussallem, Mohamed Ahmed Sherif, Diego Esteves, Marcos Zampieri, and Axel-Cyrille Ngonga Ngomo. 2018. Lidioms: A multilingual linked idioms data set. *Preprint*, arXiv:1802.08148.

Andreas Musolff. 2017. Metaphor, irony and sarcasm in public discourse. *Journal of Pragmatics*, 109:95–104.

Farzana Nawaz, Tahira Jabeen, and Sadia Rather. 2025. The power of language and religious thoughts: A pragma-rhetorical analysis of israr ahmed's speech. *AGATHOS*, 16(2):167–182. Issue 31.

Gözde Özbal, Carlo Strapparava, and Serra Sinem Tekiroğlu. 2016. PROMETHEUS: A corpus of proverbs annotated with metaphors. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3787–3793, Portorož, Slovenia. European Language Resources Association (ELRA).

John Pavlopoulos, Panos Louridas, and Panagiotis Filos. 2024. Towards a Greek proverb atlas: Computational spatial exploration and attribution of Greek proverbs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11842–11854, Miami, Florida, USA. Association for Computational Linguistics.

Lotem Peled and Roi Reichart. 2017. Sarcasm SIGN: Interpreting sarcasm with sentiment based monolingual machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1690–1700, Vancouver, Canada. Association for Computational Linguistics.

Maria Pershina, Yifan He, and Ralph Grishman. 2015. Idiom paraphrases: Seventh heaven vs cloud nine. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 76–82, Lisbon, Portugal. Association for Computational Linguistics.

Ashmari Pramodya. 2023. Exploring low-resource neural machine translation for Sinhala-Tamil language pair. In *Proceedings of the 8th Student Research Workshop associated with the International Conference Recent Advances in Natural Language Processing*, pages 87–97, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

R. Priyanga, Surangika Ranathunga, and G. Dias. 2017. Sinhala word joiner. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 265–274, Kolkata, India. NLP Association of India.

Alexander Prochnow, Johannes E. Bendler, Caroline Lange, Foivos Ioannis Tzavellos, Bas Marco Göritzer, Marijn ten Thij, and Riza Batista-Navarro. 2024. IDEM: The IDioms with EMotions dataset for emotion recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8569–8579, Torino, Italia. ELRA and ICCL.

Carlos Ramisch, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, and Aline Villavicencio. 2016. How naked is the naked truth? a multilingual lexicon of nominal compound compositionality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–161, Berlin, Germany. Association for Computational Linguistics.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848, Online only. Association for Computational Linguistics.

John W. Ratcliff and David E. Metzener. 1988. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, 13(7):46–51.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Lok Regmi. 2015. Analysis and use of figures of speech. *Journal of NELTA Surkhet*, 4.

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. Association for Computational Linguistics.

Richard M. Roberts and Roger J. Kreuz. 1994. Why do people use figurative language? *Psychological Science*, 5(3):159–163.

Yousef Sahari, Fawaz Qasem, Eisa Asiri, Ibrahim Alasmri, Ahmad Assiri, Shafi Alqahtani, and Hassan Mahdi. 2024. Evaluating the translation of figurative language: A comparative study of chatgpt and human translators. *CALR Linguistics Journal - Issue 15*.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, and 9 others. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Prateek Saxena and Soma Paul. 2020. EPIE dataset: A corpus for possible idiomatic expressions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4529–4536, Marseille, France. European Language Resources Association.

A.M. Senanayaka. 1880. *Athetha Wakya Deepanya*. Catholic Press.

John M. Senaveratna. 2005. *DictonarY of Proverbs of the Sinhalese*. Asian Educational Services, New Delhi, India.

Naziya Mahamdul Shaikh, Jyoti D. Pawar, and Mubarak Banu Sayed. 2024. Konidioms corpus: A dataset of idioms in Konkani language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9932–9940, Torino, Italia. ELRA and ICCL.

Dhirendra Singh, Sudha Bhingardive, and Pushpak Bhattacharyya. 2016. Multiword expressions dataset for Indian languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2331–2335, Portorož, Slovenia. European Language Resources Association (ELRA).

Nathalie Hau Sørensen, Sanni Nimb, Agnes Aggergaard Mikkelsen, and Jonas Jensen. 2025. The Danish idiom dataset: A collection of 1000 Danish idioms and

fixed expressions. In *Proceedings of the 1st Workshop on Nordic-Baltic Responsible Evaluation and Alignment of Language Models (NB-REAL 2025)*, pages 55–63, Tallinn, Estonia. The University of Tartu Library.

Caroline Sporleder, Linlin Li, Philip Gorinski, and Xaver Koch. 2010. Idioms in context: The IDIX corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Kenan Tang. 2022. Petci: A parallel English translation dataset of Chinese idioms. *Preprint*, arXiv:2202.09509.

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.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Sarubi Thillainathan, Songchen Yuan, En-Shiun Annie Lee, Sanath Jayasena, and Surangika Ranathunga. 2025. Beyond vanilla fine-tuning: Leveraging multistage, multilingual, and domain-specific methods for low-resource machine translation. *Preprint*, arXiv:2503.22582.

Michael Toker, Oren Mishali, Ophir Münz-Manor, Benny Kimelfeld, and Yonatan Belinkov. 2024. A dataset for metaphor detection in early medieval Hebrew poetry. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 443–453, St. Julian's, Malta. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A large-scale Chinese IDiom dataset for cloze test. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 777–787, Florence, Italy. Association for Computational Linguistics.

Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48, Online. Association for Computational Linguistics.

A Existing Datasets Utilised

A.1 Germanic-Language Corpora

Sporleder et al. (2010) have introduced `IDIX` dataset which contains English idioms. In there, they have mentioned idioms as a contextual disambiguation problem. Rather than focusing on token-level labels, Haagsma et al. (2020) emphasises an inventory of potentially idiomatic expression types in English, that may be idiomatic depending on usage. The `PIE` dataset presented by Zhou et al. (2021) has been constructed to aid in the analysis of idiom paraphrasing by connecting idiomatic statements to alternatives that preserve meaning. `PIE` dataset by Adewumi et al. (2022), constructed from `BNC` and `UKWaC`, provides an additional comprehensive English-only structure where instances are labelled across different FoS, such as metaphor, simile, euphemism, and irony, alongside literal examples. This extends beyond a binary idiom/literal structure to facilitate fine-grained multi-class categorisation of figurative language.

The `VU Amsterdam Metaphor Corpus` by Krennmayr and Steen (2017) provides extensive manually annotated text that allows metaphor recognition in natural language for metaphor processing in English. It is frequently used to assess and train systems that need to recognise metaphorical usage on a large scale. Moreover, Saxena and Paul (2020) presented a more condensed English idiom-oriented dataset with an emphasis on modelling idiomatic phrases as evaluative targets. It is typically employed to determine whether representations capture the conventionalised meanings underlying idioms or address them compositionally. The

benchmark in Stowe et al. (2022) utilises paired instances to recast figurative understanding into a controlled evaluation format through the combination of a large semi-automatic section with a smaller manually selected gold set. Instead of focusing on the use of surface-level clues, it is meant to assess how effectively models handle figurative meaning, such as idioms and metaphors.

The dataset by Reddy et al. (2011) provides human judgments on the transparency of a compound's meaning and the strength of its components' contributions. This dataset serves as a common baseline for forecasting noun compounds' levels of (non-)compositionality. Liu and Hwa (2017) presents evaluation material for phrase-level robustness and rewriting where systems have to maintain meaning despite phrase replacements. This is helpful for evaluating phrase semantics and idiom-aware paraphrasing. In addition, `CoAM` by Ide et al. (2025) focuses on the behaviour of multiword expressions in English and supports identification studies in which *Multi-word Expressions* (MWEs) need to be regarded as single lexicalised components instead of distinct words. Furthermore, a number of English-only idiom benchmarks focus specifically on evaluating idiom competence rather than building linked lexical resources. Notable examples include `IDEM` by Prochnow et al. (2024), `IDIOMEM` by Haviv et al. (2023), and `SLIDE` by Jochim et al. (2018). With the objective to facilitate benchmarking and descriptive linguistic analysis in Danish, `The Danish Idiom Dataset` provides a selective collection of idioms and fixed expressions (Sørensen et al., 2025). Swedish resources enhance this idiom-specific focus by extending coverage to MWEs more broadly. This allows for wider-coverage modelling of formulaic language and provides annotated material for recognising lexicalised MWEs beyond idioms (Kurfalı et al., 2020). Furthermore, Germanic-language research frequently interacts with translation evaluation, especially in English-German contexts where specific idiom translation data allows for the methodical evaluation of MT/LLM errors such as literalization, semantic drift, and attenuation of figurative meaning during translation (Fadaee et al., 2018).

A.2 Indic-Language Corpora

`The Idiom Handling Dataset for Indian Languages` by Agrawal et al. (2018) provides idiom processing across several Indic languages such as Hindi, Urdu, Bengali, Tamil, Gujarati, Malayalam, Telugu, and typically includes mappings that enable cross-lingual handling, extending the coverage in Indic languages. In low-resource contexts, multilingual assessment and comparative analysis are enabled by (Agrawal et al., 2018).

In addition, the dataset presented by (Singh et al., 2016) focuses on Hindi and Marathi idioms/MWEs within Indic languages, offering annotated content for MWE/idiom recognition and mode ling in these languages. `Konidiom` by(Shaikh et al., 2024) provides idiom data for Konkani, a smaller, language-specific idiom resource that supports idiom research and resource development in a low-resource environment.

Khan and Akter (2025)'s dataset for Urdu focuses on translating idioms from Urdu and Roman Urdu, utilising idiom-focused test material to assess whether modern structures can preserve idiomatic meaning across script and language diversity. This is primarily an evaluation resource for translation behaviour under idiomaticity.

A.3 Romance-Language Corpora

Romance-language resources support a coherent discussion of how figurative meaning is represented within closely related languages and how well models transfer across them. By providing naturally grounded instances that allow idiom detection and interpretation in practical circumstances, `VIDiom-PT` supports this viewpoint for European Portuguese (Antunes et al., 2025). In contrast, `Prometheus` emphasises meaning recovery at the discourse level and is proverb-oriented, making it simpler to comprehend multilingual proverbs through English–Italian data. By allowing systematic comparison between related Romance languages, standardised multilingual assessment strengthens these language-specific techniques. `SemEval-2022 Task 2` provides a common benchmark for English, Portuguese, and Galician in similar language circumstances, allowing for controlled assessment of cross-lingual generalisation and transfer (Tayyar Madabushi et al., 2022).

A.4 Cross-Lingual Figurative Language Corpora

The large-scale cloze benchmark `ChID` by Zheng et al. (2019) is employed to evaluate idiom comprehension in Chinese resources. It requires models to select a suitable idiom to fill in a passage's blank. In addition to testing contextual idiom understanding through blank-filling. In addition to assessing

contextual idiom comprehending by blank-filling, the `Chengyu Cloze Test Dataset` by Jiang et al. (2018) emphasises semantic fit and discourse compatibility and delivers an invaluable, nearly equivalent evaluation environment.

Moreover, `PETCI` by (Tang, 2022) provides Chinese idioms related to English translations, facilitating the assessment of whether MT/LLM systems retain idiomatic meaning instead of generating literal, word-by-word renditions. Given this, it is extremely beneficial for controlled idiom translation testing. By enabling idiom identification as well as analysis in morphosyntactically rich contexts, where inflexion and flexible surface forms can complicate detection and interpretation, Slavic-language corpora expand figurative language study beyond English (Aharodnik et al., 2018; Donaj and Antloga, 2023). In order to allow both proverb retrieval/analysis and computational metaphor identification in a non-English setting, Greek corpora usually integrate structured proverb repositories with metaphor-annotated datasets (Pavlopoulos et al., 2024; Garcia et al., 2021). Through Hebrew and Arabic resources which facilitate MWE identification and metaphor detection in domain-specific contexts, including historically and stylistically unique texts that present additional model transfer challenges, Semitic corpora broaden coverage (Liebeskind and HaCohen-Kerner, 2016; Toker et al., 2024; Alsiyat et al., 2023).

As a way to improve cross-lingual mapping and interoperability, multilingual linked idiom resources represent idioms as interconnected lexical entities across languages and link them to external lexical-semantic inventories (Moussallem et al., 2018). Furthermore, multilingual shared benchmarks support systematic analysis of cross-lingual generalisation and provide consistent comparison of systems on MWEs, idiomaticity, and phrase-level semantics through providing standardised annotation guidelines and evaluation protocols across various languages (Savary et al., 2023; Korkontzelos et al., 2013; Tayyar Madabushi et al., 2022; Tedeschi et al., 2022). A summary of existing corpora, indicating the languages covered and the FoS addressed in the above studies, is shown in Table 6.

B Classification of Sinhalese Proverbs

Here we discuss the classification of Sinhalese proverbs based on different criteria as given below.

B.1 By the Nature of the Message (The Shape of the Message)

උපදේශ \ʊpʌðe:ʒʌ : Proverbs that contain a moral lesson or advice. While not all proverbs are adages, some are interchangeably used to provide direct guidance, such as "Don't burn your hand while the tongs are there".

විශ්වාස \vɪʒvɑ:sʌ: Proverbs that express a commonly accepted social truth or collective belief rather than a direct instruction. These are sometimes referred to as "Truth-principle proverbs" (Sathyadharma Pirulu). Examples include "A barking dog does not bite" or "Like eating the ear while sitting on the horn".

B.2 By Background Source (The Origin)

උපමා\ʊpʌmʌ : Ends in comparative markers. (වගේ*vʌge:*, සේ*se:*, මෙන්*mɛn* , වැනි*vænɪ*).

වාර්තා\vɑ:rθɑ:: Ends in hearsay markers (ලූ*lu:*).

ප්‍රශ්න\prʌʒnʌ : Ends in interrogative markers (ද*ðʌ*), often acting as rhetorical devices to prompt self-reflection (e.g.," සිත ඇත්නම් පත කුඩා ද? \ sɪθʌ æθnʌm pʌθʌ kʊdɑ: ðʌ?"

නිශේධ\nɪʒe:ðʰʌ (Negative): Ends in negation. (නැහැ*næhæ*, බෑ*bae:*, නෑ*nae:*).

B.3 By Grammatical Ending (The Marker)

කතා \ kʌθɑ: (Story-based): These proverbs rely on shared cultural memory. They are often unintelligible without knowledge of the specific folktale or historical event (e.g., "අන්දරේ සීනි කෑවා\ *ʌnðʌre: si:nɪ kæ:vɑ: vʌgɛj* " - Like Andare eating sugar).

ප්‍රපංච \ prpʌŋʧʌ (Phenomenon-based): These are derived from empirical observations of the agrarian environment, nature, or daily life (e.g., "පිණි දිය දැක නොතලන් නෙලා පලා*pɪnɪ ðɪjʌ ðækʌ nʊθʌlʌn nɛlɑ: pʌlɑ:* " - Do not crush the greens, seeing the dew).

සාහිත්‍ය \ sɑ:hɪθjʌ (Literature-based): These originate from classical texts such as the පන්සිය පනස් ජාතක\ *pʌnsɪjʌ pʌnʌs dʒɑ:θʌkʌ* or සුභාශිතය*subʰɑ:ʒɪθʌjʌ*, reflecting the influence of Buddhism and literacy on folk speech.

Among these, උපමා \ ʊpʌmʌ (Simile) sub-category is the most prevalent. This indicates that analogi-

Dataset	Languages	FOS Explored
IDIX (Sporleder et al., 2010)	English	Idioms
MAGPIE(Haagsma et al., 2020)	English	Potentially Idiomatic Expressions
PIE (Zhou et al., 2021)	English	Idiomatic Expressions (IE)
PIE(BNC and UKWaC) (Adewumi et al., 2022)	English	Metaphor, simile, euphemism, parallelism, personification, oxymoron, paradox, hyperbole, irony, and literal
MABL (Kabra et al., 2023)	Hindi, Indonesian, Javanese, Kannada, Sundanese, Swahili and Yoruba	Figurative language
VIDiom-PT (Antunes et al., 2025)	European Portuguese	Verbal Idioms
The Danish Idiom Dataset (Sørensen et al., 2025)	Danish	Idiomatic expressions and fixed expressions
LIDIOMS, DBnary,BabelNet (Moussallem et al., 2018)	English, German, Italian, Portuguese, and Russian	Idioms
Prometheus (Özbal et al., 2016)	English, Italian	Proverbs
VU Amsterdam Metaphor Corpus (Krennmayr and Steen, 2017)	English	Metaphors
MetaNet (Dodge et al., 2015)	English, Russian, Mexican Spanish, Iranian Farsi	Metaphors
EPIE (Saxena and Paul, 2020)	English	Idiomatic Expressions
IMPLI (Stowe et al., 2022)	English	Idiom, Metaphor
ePiC (Ghosh and Srivastava, 2022)	English	Proverbs
ChID (Zheng et al., 2019)	Chinese	Metaphor, Near-synonymy
UPD*(Reddy et al., 2011)	English	Compound Nouns
SemEval-2013 Task 5 Dataset (Korkontzelos et al., 2013)	English, German, Italian	Phrases
IdiomKB (Li et al., 2024)	English, Chinese, Japanese	Idioms
IDEM (Prochnow et al., 2024)	English	Idioms
IDIOMEM. (Haviv et al., 2023)	English	Idioms
ID10M (Tedeschi et al., 2022)	English, Chinese, Spanish, Dutch, French, German, Italian, Japanese, Polish, Portuguese	Idioms
PETCI (Tang, 2022)	Chinese, English	Idioms
AStitchInLanguageModels Dataset (Tayyar Madabushi et al., 2021)	English, Portuguese	Idioms
UPD* (Garcia et al., 2021)	English	Idioms
UPD* (Cordeiro et al., 2019)	English	Nominal Compounds
SLIDE (Jochim et al., 2018)	English	Idioms
Russian Idiom-Annotated Corpus (Aharodnik et al., 2018)	Russian	Idiom
UPD*(Fadaee et al., 2018)	English, German	Idioms,Idiom Translation Dataset
Idiom Handling Dataset for Indian Languages (Agrawal et al., 2018)	English, Hindi, Urdu, Bengali, Tamil, Gujarati, Malayalam, Telugu	Idioms
Chengyu Cloze Test Dataset (Jiang et al., 2018)	Chinese	Idioms
Multilingual Lexicon of Nominal Compound Compositionality (Ramisch et al., 2016)	English, French, Portuguese	Nominal Compounds
UPD* (Pershina et al., 2015)	English,Idioms	Idiom Paraphrase Dataset
Phrasal Substitution Dataset (Liu and Hwa, 2017)	English	Idiomatic Expressions
CoAM (Ide et al., 2025)	English	MWEs
ParaDiom (Donaj and Antloga, 2023)	Slovene, English	Idiomatic Texts
Konidioms Corpus (Shaikh et al., 2024)	Konkani	Idioms
Multi-word Expression Dataset for Swedish (Kurfalı et al., 2020)	Swedish	Multi-word Expression
PARSEME Corpus Release 1.3 (VMWEs) (Savary et al., 2023)	Arabic, Bulgarian, Chinese, Croatian, Greek, Hebrew, Hindi, Irish, Latvian, Lithuanian, Maltese, Slovene, Turkish	Idioms, multiword expressions (verbal MWEs)
SemEval-2022 Task 2 Dataset (Tayyar Madabushi et al., 2022)	English, Portuguese, Galician	Idioms
UPD*(Singh et al., 2016)	Hindi, Marathi	Idioms, MWEs
UPD* (Liebeskind and HaCohen-Kerner, 2016)	Hebrew	MWEs (incl. idiom-like fixed expressions)
Greek Proverb Atlas(Pavlopoulos et al., 2024)	Greek	Proverbs
UPD* (Florou et al., 2018)	Greek	Metaphor
UPD* (Toker et al., 2024)	Hebrew	Metaphor
UPD* (Khan and Akter, 2025)	Urdu, Roman Urdu	Idioms
AMC (Alsiyat et al., 2023)	Arabic	Metaphor

Table 6: Existing Datasets Summary. *Corpora named 'UPD' represent the *Unnamed Primary Dataset(s)*, which includes papers that have released/utilised datasets without specific names.

cal reasoning, understanding one concept in terms of another, is the primary cognitive tool used in Sinhala folk wisdom. වාර්තා \vaːrθaː (Report) category is the second most common proverb structure. The prevalence of the particle ලු \ *luː* (it is said) underscores the importance of oral tradition and collective knowledge in Sri Lankan culture, wisdom is validated not by the speaker's authority, but by the fact that "it has been said" by ancestors.

C Sinhala Proverbs vs Sinhala Idioms

The Dichotomy of Sinhala Proverbs and Sinhala Idioms: While both categories function as figurative devices, they are distinguishable through three primary dimensions: Syntactic Structure, Semantic Deductibility, and Pragmatic Function.

Sinhala (සිංහල \ sɪŋhʌlʌ) : ඉත්තෑවගේ ගුලේ කබල්ලෑව වැදී "මුත්තාපා කීවත් යන්නේ නෑ" කිව්වාලු \ ɪθθæːvʌgeː guleː kʌbʌlːlæːvʌ vʌðiː muθθɑːpɑː kiːvʌθ jʌnneː næː kɪvvɑːlʊ

Type of FOS : [විශ්වාස][කතා][උපමා] \ [vɪʒvɑːsʌ][kʌθɑː][ʊpʌmɑː]

Literal / visual image : The ant-eater, who forcibly occupied the porcupine's hole, swore by his forbears that he would never leave it.

Corresponding FOS in English : Possession is nine-tenths of the law.

What it really implies : Taking possession of other people's property through deceit.

Additional Context : In a certain forest, a porcupine lived inside a cave. One day, an anteater, who had lost his way while traveling, came across this cave and asked the porcupine if he could stay there for shelter. The porcupine kindly agreed and allowed the anteater to stay inside. However, the next morning, the anteater showed no sign of leaving. Since the cave was too small for both of them to live together, the porcupine politely requested the anteater to leave. "If you don't like it, then you can go and find another place. I'm quite comfortable here," the anteater replied. Angered, the porcupine raised his sharp quills and attacked the anteater. But the anteater's body was covered in thick, coarse hide, so the porcupine's blows had no effect. The anteater remained in the cave, while the porcupine was forced to leave and find another shelter.

Figure 4: An example of a record on the dataset.

Semantic Deductibility (Opacity vs. Transparency): Idioms in Sinhala often exhibit high semantic opacity; a learner cannot easily deduce that " කහ \ kʌhʌ" in " ඌරන්ට කහ \ uːrʌntʌ kʌhʌ" implies "wasting resources." However, Proverbs are often semantically translucent. Even a first-time listener can deduce the meaning of " ගහ දන්න අයට කොළ කඩා පානවා \ gʌhʌ ðʌnnʌ ʌjʌtʌ kɒlʌ kʌdɑː pɑːnʌvɑː" (Showing leaves to those who know the tree) based on the imagery of deception and expertise.

Pragmatic Function: ප්‍රස්තාපිරුළු \ prʌsθapɪrulu are didactic; they convey general truths, social beliefs, or moral advice (උපදේශ \ ʊpʌðeːʒʌ). වාග්සම්ප්‍රදා \ vɑːg sʌmprʌðɑː are descriptive; they categorise a state of being or an action without necessarily offering a moral judgment.

Dominance of Idioms: වාග්සම්ප්‍රදා \ vɑːg sʌmprʌðɑː constitute the overwhelming majority of the dataset. This quantitative dominance suggests that Sinhala speakers prioritise "descriptive efficiency" in daily language, using short, culturally loaded phrases to quickly describe complex situations, over the more formal, structured wisdom of proverbs.

D Dataset Annotation

The dataset was annotated by filling in the fields. Not all fields were filled in for all records, as shown in Table 1. Figure 4 contains an example of a record in the dataset.

E Cultural Analysis

The Source Domain explores the abstract imagery and objects used in the FoS to deliver the message, whilst the target theme is used to identify the messages delivered by the various FoS.

E.1 Specific Cultural Codes

Certain symbols carry specific, unchangeable meanings in the Sinhala cultural lexicon. The following are some of the examples utilised in Sinhala FoS.

The Elephant (Power & Scale): The elephant is the cultural yardstick for greatness. It is used to contrast "the great" with "the small." It represents forces that are often too big to manage or criticise.

The Dog (Low Status): In contrast to the elephant, the dog is consistently used to represent unworthiness or low social status. It serves as a warning of what happens when one lacks dignity.

The Tree (Character): Trees are almost always metaphors for moral character. A person is judged like a tree, by their "fruit" (utility to society) or their "wood" (strength/weakness).

E.2 Emotional Landscape

The sentiment analysis shows that the vast majority of FoS (83% of the data) are *Neutral*. They are not optimistic or pessimistic; they are descriptive. The culture does not say "Life is good" or "Life is bad"; it says, "If you take this action, the corresponding outcome will occur inevitably." It values truth over comfort.

Table 7 provides a comprehensive overview of the cultural analysis, summarising the frequency of literal imagery and the specific thematic domains explored within the SINFOS dataset.

Category	No. of Occurrences
Source Domain (Literal Imagery)	
Body & Senses (Somatic)	242
Nature & Agriculture	194
Animals (Fauna)	148
Household & Daily Life	144
Target Theme (Cultural Meaning)	
Ethics & Moral Character	162
Karma & Consequence	127
Impermanence & Uncertainty	121
Social Status & Hierarchy	73
Human Relations & Conflict	64

Table 7: Distribution of literal source domains and abstract cultural themes observed in the SINFOS dataset via hybrid thematic analysis.

F Model Classification

The results of classifying proverbs and idioms are summarised in Table 8. Word2Vec showed the best performance for Naive-Bayes and Linear SVC in terms of recall and accuracy. In contrast, TF-IDF 3-gram vectorisation excelled with Random Forest, XGBoost, and the ensemble model combining these with Linear SVC.

Experiments	Acc.	P-Rec.	I-Rec.
TF-IDF (Unigram) Multimodal Naive-Bayes	0.7167	0.59	0.81
TF-IDF (Char 3-gram) Multimodal Naive-Bayes	0.8348	0.79	0.87
Gaussian Naive-Bayes Word2Vec	0.9034	0.92	0.89
TF-IDF (Unigram) Linear SVC	0.7639	0.64	0.86
TF-IDF (Char 3-gram) Linear SVC	0.8712	0.81	0.92
Linear SVC Word2Vec	0.9034	0.90	0.91
Random Forest (tuning) TF-IDF (Unigram)	0.8026	0.66	0.91
Random Forest (tuning) TF-IDF (Char 3-gram)	0.8927	0.83	0.94
Random Forest (tuning) Word2Vec	0.8755	0.81	0.93
XGBoost (tuning) TF-IDF (Unigram)	0.8090	0.65	0.93
XGBoost (tuning) TF-IDF (Char 3-gram)	0.9013	0.86	0.93
XGBoost (tuning) Word2Vec	0.8690	0.83	0.90
TF-IDF (Char 3-gram) Voting Ensemble	0.9056	0.85	0.95
Voting Ensemble Word2Vec	0.8884	0.85	0.92
Bi-LSTM	0.9163	0.86	0.96
Deep NN	0.9270	0.94	0.92

Table 8: Model Performance: Accuracy, Proverbs Recall (P-Rec.), and Idioms Recall (I-Rec.).

G Performance of all LLMs

A brief overview of each metric's blind spots and how each metric mitigates the other's is provided in Table 9.

Metric	Blind Spot	Mitigation Strategy
Cosine Similarity	The "Keyword Bag" Problem: the model may achieve high scores by guessing relevant keywords even if the grammatical structure is flawed.	Fidelity acts as a "Logic Gate," requiring semantic validity rather than just keyword overlap.
Fidelity Score	The "Hyper-Literal" Problem: creative paraphrases with different structures might be penalised.	Cosine Similarity permits creative phrasing; high similarity with low fidelity suggests a valid non-standard translation.

Table 9: Evaluation Metrics and Mitigation of their Blind Spots

The Figures 5 and 6 represent the Cosine Similarity scores and Fidelity Scores of all the models across seven different categories. Along with ආප්තෝපදේශ \ ɑːpθɔːpʌðeːʒʌ, the models seem to have decent performances for proverbs associated with nature as they seem to be able to decipher the meaning using the phenomenon. In the case of කියමන් \ kɪjʌmʌn though, as in proverbs based on folklore, the language models seem to struggle. This is tied with the fact that unlike ආප්තෝපදේශ \ ɑːpθɔːpʌðeːʒʌ, කියමන් \ kɪjʌmʌn are more specific to the language.

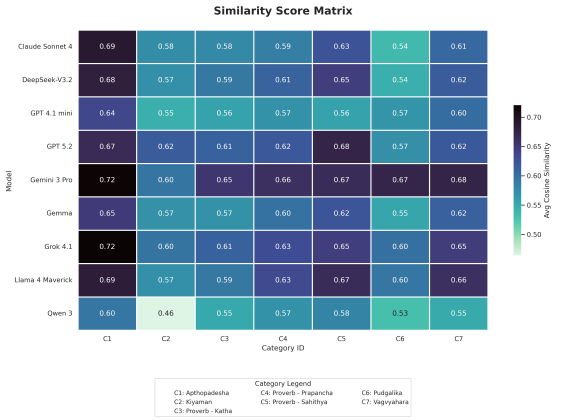


Figure 5: Benchmarking LLM Performance: Cosine Similarity.

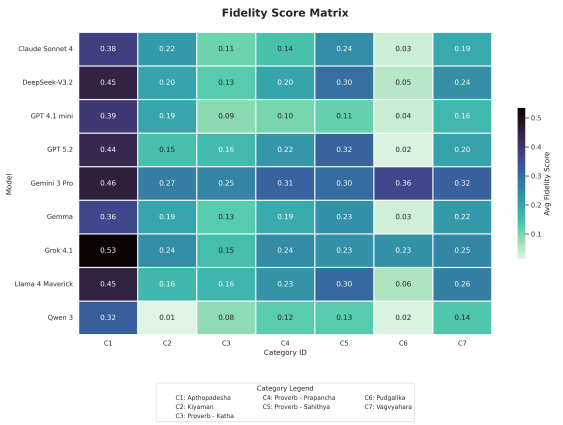


Figure 6: Benchmarking LLM Performance: Fidelity Scores.

Swedish Multiword Expression Corpora in PARSEME

Sara Stymne, Astrid Berntsson Ingelstam, and Eva Pettersson
Department of Linguistics and Philology
Uppsala University, Sweden
[first_name.last_name]@lingfil.uu.se

Abstract

We present the annotation of Swedish multiword expressions under the PARSEME annotation scheme, including a new release and a historical overview of previous releases. We provide an overview of the evolution of the Swedish datasets and of inter-annotator agreement. We discuss general guidelines and the development of Swedish-specific guidelines for particle verbs and multiword tokens, as well as additional challenges for the Swedish annotation. We also conduct an initial comparison of Swedish and other Germanic languages, identifying aspects where the PARSEME guidelines require revision to ensure better consistency across languages.

1 Introduction

Multi-word expressions (MWEs) are non-compositional expressions like *hålla ett öga på* ('keep an eye on') or *ge upp* ('give up'), where the meaning of the full expression cannot be directly inferred from the meaning of its parts. MWEs are challenging for many NLP applications, due to their irregularities, and have been called "a pain in the neck for NLP" (Sag et al., 2002; Shwartz and Dagan, 2019). The processing of MWEs is challenging for classical NLP, both for MWE-centered tasks, such as MWE discovery and identification, and for handling MWEs in other tasks, such as parsing and machine translation Constant et al. (2017). MWEs remain challenging for LLMs; for instance, Miletić and Schulte im Walde (2024) find that transformer models struggle with MWEs, largely relying on surface patterns and memorized information.

PARSEME[1] is an initiative that provides universal guidelines for consistent annotation of MWEs across languages. Their first iteration, version 1, only covered verbal MWEs (VMWEs), but since version 2.0, the coverage has been extended to all types of MWEs. PARSEME is based on general MWE categories, present across languages, with an annotation framework based on decision diagrams for annotation. In some cases, language-specific rules are added as a complement.

In this paper, we focus on Swedish and describe the evolution of Swedish MWE resources in PARSEME. We give an overview of the Swedish resources in the existing PARSEME releases and describe the work on a coming release, which we refer to as release 2.1. We discuss the annotation process, inter-annotator agreement, and MWE distribution, as well as Swedish-specific guidelines and challenges, related to phenomena such as multiword tokens and particle verbs. Finally, we compare the MWE distribution for the Germanic languages available in PARSEME.

2 Related work

In this section, we describe the PARSEME framework for MWE annotation, followed by a discussion of other work on MWE resources for Swedish and other languages.

2.1 PARSEME

The PARSEME scientific network has produced guidelines and corpora for the annotation of MWEs since 2017 (Savary et al., 2017; Ramisch et al., 2018, 2020; Savary et al., 2023a, 2026). In version 1, the resources covered only verbal MWEs, but since version 2, all types of MWEs are covered.[2] The VMWE guidelines are thus tested in several iterations, whereas the extended guidelines for other MWE types are currently in their first iteration. Most PARSEME releases have been tied to a shared task on MWE identification (e.g. Scholivet et al., 2026), but version 1.3 was released inde-

[1] https://gitlab.com/parseme/corpora/

[2] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 27–37

VMWE	VID	verbal idiom	*att **träda i kraft*** (lit. 'to step in force') 'to come into effect'
	LVC.full	light-verb construction; bleached verb	*att **hålla tal*** (lit. 'to hold speech') 'to make a speech'
	LVC.cause	light-verb construction; causal verb	*att **väcka hopp*** (lit. 'to wake hope') 'to inspire hope'
	IRV	inherently reflexive verb	*att **gifta sig*** (lit. 'to marry oneself') 'to get married'
	IVPC.full	idiomatic verb-particle construction	*att **höra till*** (lit. 'to hear to') 'to belong'
	IVPC.semi	semi-idiomatic verb-particle constr.	*att **fråga ut*** (lit. 'to ask out') 'to interrogate / to invite out'
	MVC	multi-verb construction	*No instances found in the Swedish data*
	IAV	inherently adpositional verb	*att **bero på*** 'to depend on' (*Optional and experimental MWE category not annotated in Swedish)*
NMWE	NID	nominal idiom	***flod|häst*** (lit. 'river horse') 'hippopotamus'
	PronID	pronominal idiom	***en del*** (lit. 'a part') 'some'
	NV	deverbal nominal MWE	***ut|gåva*** (lit. 'out-gift') 'edition' (cf. *att **ge ut*** (lit. 'give out') 'publish')
AMWE	AdjID	adjectival idiom	***så kallad*** 'so-called'
	AdvID	adverbial idiom	***över huvud taget*** (lit. 'over head taken') 'at all, even'
	AV	deverbal adjectival/adverbial MWE	***igen|känd*** (lit. 'again known') 'recognized' (cf. *att **känna igen*** (lit. 'know again') 'recognize')
FuncMWE	DetID	determiner idiom	***ett par*** (lit. 'a pair') 'a couple of'
	AdpID	adpositional idiom	***i och med*** (lit. 'in and with') 'due to, because of'
	ConjID	conjunction idiom	***såväl som*** 'as well as'
	IntjID	interjection idiom	***oj då*** (lit. 'oh then') 'oh dear / ouch' (*No instances found in the Swedish data*)

Table 1: PARSEME typology of MWEs with Swedish examples.

pendently. Overall, the PARSEME corpora cover 33 languages, but the language coverage varies between releases. Release 1.3 includes 26 languages, and release 2.0 includes 17 languages.

The main MWE categories are verbal, nominal, adjectival/adverbial, and functional MWEs, each of which has several subtypes. Table 1 provides an overview of all MWE categories, accompanied by Swedish examples. The verbal category has six main categories: verbal idioms, VID, light verb constructions, LVC, inherently reflexive verbs, IRV, idiomatic verb-particle constructions, IVPC[3], multi-verb constructions, MVC, and inherently adpositional verbs, IAV. IAV is an optional and experimental category that has not been included in any of the Swedish releases. Nominal MWEs are split into nominal, NID, and pronominal, PronID idioms; adjectival/adverbial MWEs into adjectival, AdjID, and adverbial, AdvID, idioms; and functional MWEs into determiner, DetID, adpositional, AdpID, conjunctional, ConjID, and interjection, IntjID, idioms. Nominal and adjectival/adverbial MWEs further have a deverbal category, for expressions that can be rephrased into a verbal expression that passes the verbal MWE tests. These are further subcategorized into the verbal subtypes, but in this paper, we group them into the main classes: deverbal nominal and deverbal adjectival/adverbial. Some meaning-preserving variants that were annotated as VMWEs prior to release 2.0, are now instead considered deverbal.

The PARSEME guidelines are organized as a decision tree, where the first question concerns the distribution of the candidate expression, specifically whether it is verbal, nominal, adjectival/adverbial, or functional. This is followed by a specific decision tree for each MWE subtype, which includes several tests to determine whether a candidate meets the MWE criteria. The goal is to annotate expressions that are non-compositional. Since it is hard to directly judge the semantic idiomacity or non-compositionality of expressions, the tests are based on the fact that non-compositionality correlates with syntactic and morphological inflexibility (Sag et al., 2002). The rules thus test aspects such as whether a regular syntactic or morphological change leads to unexpected meaning change, whether a cranberry word is present, or whether the internal syntactic structure of the expression is irregular. MWE candidates must consist of multiple tokens, but they can also be multiword tokens (MWTs), i.e., cases where one token contains several words, such as compounds (e.g. *sky|skrapa* 'skyscraper'),[4] particle verbs (e.g. *in|gripa* (lit. 'in seize') 'intervene'), and contractions (e.g. 'don't'). MWTs are common in Germanic languages, which are compounding languages, except for English.

MWEs are annotated using the FLAT annotation tool.[5] Annotators first identify MWE candidate expressions, and then label them as MWEs if they

[3]IVPC was called VPC in releases prior to 2.0. We will consistently use the term IVPC even for earlier releases.

[4]Vertical bars are used to indicate token boundaries in MWTs; they are not part of the Swedish orthography.

[5]https://github.com/proycon/flat

pass the tests in the PARSEME decision trees. In most cases, a file is annotated by a single annotator due to annotator availability; however, inter-annotator checks are regularly performed (Ramisch et al., 2018, 2020). To further improve the quality, the PARSEME protocol of annotation includes a consistency-check step, where all annotations are reconsidered across files (Savary et al., 2017). Inconsistencies can be identified by the use of a custom tool that highlights examples of potential MWEs across all texts, based on lemmas, by searching for all MWEs that have been annotated at least once. A single annotator can then change inconsistently annotated examples, or add/remove examples that have been missed or erroneously annotated. This process resolves many inconsistencies between annotators.

PARSEME annotations use a format called CUPT (Ramisch et al., 2018), which is an extension of the ConllU format from the Universal Dependencies (UD) initiative (Nivre et al., 2020; de Marneffe et al., 2021), adding a column with MWE annotations. UD is another cross-linguistic initiative for morphosyntactic annotation, focusing on lemmas, morphology, part-of-speech tags, and dependency annotations. PARSEME corpora, from release 1.1, include UD annotations, either by annotating MWEs on top of existing UD treebanks, as for Swedish, or by parsing other corpora into the UD format. Discussions on how to further unify PARSEME and UD are ongoing (Savary et al., 2023b).

2.2 PARSEME MWE Annotations in Other Languages

The annotation of VMWEs for the Turkish PARSEME corpus posed problems due to the agglutinative nature of the language (Ozturk et al., 2022). Specifically, the automatic lemmatization of the Turkish corpus was often incorrect, with suffixes being incorrectly or insufficiently stripped, causing inconsistencies in the finalized corpus, as a single MWE type is represented by a sequence of included lemmas.

Hadj Mohamed et al. (2025) note that since Arabic is an agglutinative language, with many MWTs, only parts of an MWT may belong to an MWE. They also note that Arabic has a high rate of discontinuous MWEs. Walsh et al. (2020) note several issues, including the difficulty of distinguishing between particle verbs and inherently adpositional verbs, where particles often are homonymous with prepositions. They note that language-specific tests for this issue are needed for Irish.

2.3 MWE Annotations Beyond PARSEME

There are many datasets available for MWEs for a single language, sometimes focusing on specific types of MWEs only, such as Estonian particle verbs (Kaalep and Muischnek, 2006) or French adverbial MWEs (Laporte et al., 2008). Datasets covering many languages are not as common. ID10M contains annotations of idioms for 10 languages, created automatically based on Wiktionary entries marked as idiomatic or literal, with matching occurrences from Wikipedia text, and a manual curation for four languages (Tedeschi et al., 2022). Other multilingual resources of MWEs are often created for specific tasks in a few languages, such as multimodal MWE comprehension in English and Turkish (Pickard et al., 2025) or idiomaticity and semantic text similarity for English, Portuguese, and Galician (Tayyar Madabushi et al., 2022).

2.3.1 Work on Swedish MWEs

The Swedish lexicon resource SALDO (Svenskt associationslexikon 2, Borin et al., 2013) contains some lexicalized multiword and compound word entries. When describing SALDO, Borin et al. (2013) discuss the frequent compounding in Swedish, and state that a single token often corresponds to an MWE in other languages. Such lexicalized compound words often have a compositional sense that differs from the literal reading; a trait that is emphasized in the PARSEME guidelines. For instance, the noun compound *husbil* (*hus*+*bil*) means ‘camper’ or ‘trailer’, but the literal meaning reading is ‘house car’.

Kurfalı et al. (2020) form a dataset of 96 Swedish verbal, nominal, and prepositional MWEs from SALDO paired with human-judged scores for degree of compositionality. They compare the human judgment with scores from computational vector models, and find that these agree poorly. Furthermore, Tiedemann et al. (2022, 2024) attempt to relate MWEs extracted from SALDO to the language proficiency levels of the Common European Framework of Reference (CEFR). They compare difficulty rankings of MWEs by L2 learners and teachers of Swedish to frequencies in course books and learner-produced texts at different proficiency levels. Tiedemann et al. (2024) also suggest that more transparent MWEs with a lower compositionality are easier for L2 speakers to learn, whereas

less transparent MWEs are more difficult.

3 Corpus Description

In this section, we give an overview of the Swedish PARSEME releases, including our latest annotations, and discuss inter-annotator agreement.

3.1 Swedish PARSEME Releases

Swedish was present in the first PARSEME release, version 1.0 (Savary et al., 2017), with a very small dataset covering 200 sentences, sampled from the Swedish Newspaper *Göteborgsposten* (GP). Three annotators were involved in this release. Swedish was not part of release 1.1 (Ramisch et al., 2018).

From release 1.2 (Ramisch et al., 2020), Swedish was added again, now with considerably larger corpora. From this release, the data to annotate was taken from the Swedish Universal Dependencies treebanks Talbanken[6] and PUD[7], containing human-annotated morphosyntactic information and lemmas. This data was annotated for MWEs by a group of six annotators, covering the training section of Talbanken. Release 1.3 (Savary et al., 2023a) was an extension of 1.2, annotated by two annotators from release 1.2, covering all of Talbanken. Up to release 1.3, only VMWEs were annotated. The annotation group from releases 1.2 and 1.3 received initial training, followed by regular email discussions about borderline cases.

Release 2.0 covered all MWE types, and was annotated by a team of five annotators, partially overlapping with the team from version 1.2/1.3. Most of the previously annotated data from Talbanken were reannotated with the new guidelines. In addition, all data in the PUD treebank was annotated from scratch. The annotator team received initial training, followed by regular meetings to discuss annotation guidelines and tricky cases.

In this paper, we also describe a new annotation effort, which we will tentatively refer to as release 2.1, where the annotations from release 2.0 are revisited, and the remaining part of Talbanken is added. In this annotation phase, two annotators from release 2.0 participated. These annotators held discussions to resolve disagreements over the interpretation of the guidelines from version 2.0.

[6] https://github.com/UniversalDependencies/UD_Swedish-Talbanken/blob/master

[7] https://github.com/UniversalDependencies/UD_Swedish-PUD/tree/master

3.2 Corpus Statistics

Table 2 gives an overview of all Swedish PARSEME releases, and the distribution of VMWEs. Table 3 gives further details on the MWE distribution for the two releases that contain all MWE types. Percent distributions of the MWE categories are shown in Tables 6 and 7. Except for release 1.0, all releases are reasonably sized, covering at least 4,000 sentences, and range between 1,991 and 4,904 MWEs. Among the verbal categories, it is clear that IVPCs are the most common type, followed by LVCs and VIDs. The only rare category is LVC.cause; all other categories have close to or over 100 instances in our latest release.

Among the new MWE types added in release 2.0, AdvID is by far the most common type, followed by NID and NV. The AdvID category largely consists of prepositions followed by nouns for which a morphological inflection would lead to ungrammaticality or an unexpected change in meaning. In many cases, the noun can only take the indefinite singular form, as in (1). In other cases, the noun is always in the definite form instead, as in (2).

(1) **till exempel**
to example
'for example'

(2) **i onödan**
in non-necessity.DEF
'unnecessarily'

We note that in all releases, except 1.0, that do not cover them, multiword tokens are common. One clear distinction between the releases is the proportion of multiword tokens (%MWT). The guidelines for when to consider a token as an MWT changed between release 1.3 and 2.0, as described in Section 4.2, resulting in a decrease in the proportion of verbal MWTs from approximately 50% in releases 1.2 and 1.3 to under 30% in releases 2.0 and 2.1. This change especially affected IVPCs, with a much lower proportion in releases 2.0/2.1. This change is also reflected in the number of MWEs per sentence. The MWT rate is overall high in releases 2.0 and 2.1, though, to a high extent due to new categories with a very high proportion of MWEs, such as NID (97%), NV (100%), and AV (100%). Another notable change is the increase of ConjID from release 2.0 to 2.1, which is mainly caused by a changed decision for common correlative conjunctions like ***varken*** … ***eller*** 'neither … nor' and ***både*** … ***och*** 'both … and', which are considered as MWEs in release 2.1, but not in 2.0.

Release	Text source	Sent.	Tokens	VMWE			VID	IRV	LVC	LVC	IVPC	IVPC
				Total	%MWT	Per sent.			full	cause	full	semi
SV 1.0	News (GP)	200	3,376	56	0	0.28	9	3	13		31	
SV 1.2	Talbanken	4,304	65,482	1,991	48.6	0.46	291	115	279	11	871	424
SV 1.3	Talbanken	6,026	96,820	3,155	51.2	0.52	441	237	417	10	1461	589
SV 2.0	Talbanken, PUD	5,553	90,392	1,779	27.9	0.32	396	186	389	19	494	294
SV 2.1	Talbanken, PUD	7,026	115,944	2,275	27.7	0.32	480	295	500	29	580	391

Table 2: Corpus overview and distribution of verbal MWEs for the Swedish PARSEME releases. Note that no distinction was made between subtypes of LVC and VPC in release 1.0.

Release	MWE			Verbal	Nominal			Adjectival/adverbial			Functional		
	Total	%MWT	Per sent.	Total	NID	PronID	NV	AdjID	AdvID	AV	AdpID	ConjID	DetID
SV 2.0	3,678	58.0	0.66	1,779	331	91	232	76	775	84	127	91	94
SV 2.1	4,904	53.7	0.70	2,275	492	111	318	98	929	116	187	222	156

Table 3: Overview and distribution of MWEs in editions that annotate all MWE types. For verbal distribution, see Table 2.

		F1-score	
	Sent.	VMWE	All MWE
SV 1.2	700	73.4	–
SV 2.0	500	46.3	47.0
SV 2.1	300	88.4	54.0

Table 4: Inter-annotator agreement as F1-score across two annotators for three versions of the Swedish PARSEME corpus.

3.3 Corpus Quality

For three of the PARSEME releases, we performed double annotation to assess inter-annotator agreement. For version 1.2, two annotators annotated the same 700 sentences for VMWEs. In version 2.0, the first version to contain MWEs beyond VMWEs, 500 sentences were annotated by a different set of two annotators from scratch. For version 2.1, we had no completely fresh data, so two annotators separately annotated on top of the VMWE annotations from version 1.3. Note, however, that the guidelines for MWTs, as well as some guideline interpretations, had changed between these two iterations, which affected a relatively large number of VMWE instances.

To calculate agreement, we followed Savary et al. (2026), and used the F1-score between the two annotations, since calculating chance agreement, which is needed for most agreement measures, is challenging in cases like these, where relatively few words in a text are annotated. Table 4 provides an overview of the results, reported both for the full annotations and for the subset of version 2 annotations that are VMWEs. We note that the agreement decreased from version 1.2 to 2.0, possibly due to the much more diverse sets of MWEs being annotated, and new MWT guidelines being developed during the course of the annotation (see Section 4.2). Between releases 2.0 and 2.1, the two annotators had a discussion about disagreements, borderline cases, and the interpretation of the guidelines in release 2.0, which led to a higher level of agreement in version 2.1. We especially note that, although the annotations for verbal MWEs were performed on version 1.3, the agreement scores are higher for VMWEs than for version 1.2.

While the agreement scores for non-verbal categories are low, the consistency of the final dataset is improved by the use of consistency checks across files (see Section 2.1), where one annotator goes through all annotations across files before each release. To give an indication of the proportion of changed decisions, Table 5 shows the F1-score of comparing the annotations before and after consistency-checking for the top categories of MWEs. For categories with subcategories, we report macro F1-scores. Overall, VMWEs show relatively few changes, with scores over 90 for all types, further supporting that the annotation of these categories is more consistent than for other categories. Nominals, conjunctions, and deverbals have considerably lower scores, between 62 and 84. For conjunctions, this can mostly be attributed to the decision to include correlative conjunctions as MWEs (see Section 3.2).

As a further point of comparison, across languages for version 2.0, the average inter-annotator agreement F1-score was 60.6, with scores ranging from 20 to 99 (Savary et al., 2026). For 7 of the 14 languages for which inter-annotator agreement was calculated, the F-score was below 65. This warrants a detailed cross-lingual investigation into agreement across PARSEME languages to deter-

VID	IRV	LVC	IVPC	AV	AdjID	AdpID	AdviD	ConjID	DetID	NID	NV
92	97	97	94	78	90	89	90	62	76	84	81

Table 5: F-scores per category when comparing corpus release 2.1 before and after consistency checks. For IVPC, LVC, and deverbal categories, results are macro-averaged across subclasses.

mine whether the widespread low agreement scores are primarily due to the new MWE categories in other languages than Swedish as well. In addition, a more detailed investigation of which categories pose problems is needed. Such an investigation could feed into a new, improved version of the PARSEME guidelines.

3.4 Releases

Versions 1.0 to 2.0 described in this paper are already released, as part of multilingual PARSEME releases. For the latest annotations, we plan to release them in two ways: as part of the next PARSEME release, 2.1 (which is currently not yet scheduled), and as part of the next UD release, by adding the PARSEME annotations to Talbanken and PUD. The release of UD version 2.18 is scheduled for May 2026. Both releases will be under permissive licenses.

4 Swedish-Specific Considerations

In this section, we discuss considerations for the Swedish PARSEME annotation, first by describing Swedish language-specific guidelines, and then by discussing issues encountered during annotation. While our discussion is based on a Swedish perspective, several issues are also relevant to other languages, and we include some comparisons to other Germanic languages.

4.1 Swedish-specific PARSEME Guidelines

The PARSEME guidelines include language-specific tests that may relate to language-specific MWE categories, more specific tests for some MWE types, or elementary language features. Swedish has language-specific tests for two issues of the latter category, particles and MWTs.

4.1.1 Particles versus Prepositions

Particle verbs or phrasal verbs are pervasive in the Germanic languages, and idiomatic uses of them are covered by the IVPC category in PARSEME. In many cases, particles are homonymous with prepositions in prepositional complements or verb prefixes, and there is thus a need to be able to distinguish these cases. Language-specific rules currently exist for Swedish, English, and German,[8], but the need is also noted for languages from other families, like Irish (Walsh et al., 2020). Example (3) shows an ambiguous sentence in Swedish, where a particle verb reading corresponds to the English translation 'visits', whereas a prepositional reading translates into 'greets'. The main way to distinguish particles and prepositions in Swedish is through stress patterns, where particle verbs have the primary stress on the particle, whereas prepositional verbs have the main stress on the verb, with an unstressed preposition (Svenonius, 2003). Thus, the stress pattern is the basis of the decision rule for Swedish in PARSEME.

(3) Hon **hälsar på** oss.
She greets on us
'She visits/greets us.'

This rule contrasts with the English and German rules, which are based on tests of movement and insertion. In both English and German, the particle can be placed at the end of the sentence, as in 'she takes her clients in', whereas in Swedish it cannot. Additionally, English has a test based on adjunct insertion, and German has a test for separable verb prefixes, which are not applicable to Swedish; in contrast to German, where separated and compounded forms of particle verbs are governed syntactically, Swedish particle verbs occur either in a separated or compounded form based on the expression, where some expressions can occur in both forms, but sometimes with shifted semantics between the forms (Norén, 1995). There could be a possibility of a syntactic test for Swedish, based on topicalizing the prepositional object, as in (4), which is not possible with a particle reading. However, this test can be challenging to apply in certain cases (Svenonius, 2003), so even if it were added as an additional test for particles in Swedish, we still believe the test for stress patterns is necessary.

[8] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/?page=060_Language-specific_tests/010_Particles,_prepositions,_prefixes

(4) **På** oss **hälsar** hon.
On us greets she
'Us, she greets. *Us, she visits'

4.2 Multiword Tokens

To unambiguously identify multiword tokens (MWTs) is crucial for identifying MWE candidates. However, so far, Swedish is the only PARSEME language with language-specific rules for MWTs. We believe such rules should be added for other languages as well, including German and Dutch, which have a similar structure to Swedish regarding MWTs.

For Swedish, there have been two iterations of rules for MWTs, one for releases 1.2–1.3, and one from release 2.0 forward. The first set of rules was part of releases that covered only VMWEs, meaning the rules for MWTs also covered only verbal expressions. Thus, there was a first rule for testing whether the expression was verbal. The second rule, *splittable*, states that the expression is an MWT if it can be used in its split form, with the same or slightly shifted semantics. The third rule is very permissive, and states that an expression is an MWT if all of its components can be used as standalone words with the same part-of-speech as in the full token.

As discussed in Section 3.2, these rules have a considerable impact on the number of MWTs and thus also on the number of MWEs, especially for IVPCs, where the first set of MWT rules is highly permissive. They lead to many instances of MWTs that passed the MWE guidelines but did not align with the annotators' intuition. Additionally, the MWT rules needed to be extended to cover non-verbal expressions for release 2.0. Keeping these rules would have led to inconsistencies, such as: not considering the verb *för|lora* (lit. 'for LORA') 'lose' an MWT, but considering the related noun *för|lust* (lit. 'for lust') 'loss' an MWT and also an MWE, since 'lust' but not 'lora' happens to be a stand-alone word. These issues caused discussions among the annotators of the 2.0 corpus version, leading us to revise the rules before finalizing the annotations for the release.

The rules for release 2.0/2.1 first test whether the expression is a noun-noun compound (NNC). Noun-noun compounding is very frequent in Swedish, and it is ungrammatical to directly split an NNC into separate graphical tokens, while some NNCs can be split using prepositional complements. We still wanted to consider all NNCs as MWTs to achieve a better cross-lingual comparison with the many languages where NNCs are written as graphically distinct words, such as in English, and thus are considered MWE candidates in those languages. This means that an NNC like ***bergs|kedja*** (lit. 'mountain chain') 'mountain range' would be considered an MWT and thus it can also be an MWE as can its English counterpart.

For any remaining MWT candidates, we keep the *splittable* rule from before. We also note that many nominal, adjectival, or adverbial expressions that are deverbal could be split in their verbal form, but not in their original form. We thus add a third rule: if a non-verbal expression has a corresponding verbal form that can be split with the same or slightly changed semantics, it should be considered an MWT, but only the tests for the deverbal classes NV or AV should be considered. For examples of deverbals, see NV and AV in Table 1.

As a result of the updated MWT rules, candidates like *genom|snitt* (lit. 'through incision') 'average', *där|för* (lit. 'there fore') 'because', and *för|lust* (lit. 'for lust') 'loss' are no longer considered as MWTs, whereas all NNCs are considered as MWTs, so that those that fill the criteria for NID are annotated as such. In addition, we now cover the quite common class of deverbal MWTs.

The change in MWT guidelines affects the annotation of phrasal verbs (IVPC.full/semi), including deverbal IVPCs. Swedish phrasal verbs that have a split form, such as *att* ***hälsa på*** (lit. 'to greet on') 'to visit' and *att* ***fråga ut*** (lit. 'to ask out') 'to interrogate', can often be compounded into particle+verb. Deverbal forms of phrasal verbs are commonly compounded in such a way, for example ***på|hälsad*** (lit. 'on greeted') 'visited' (AV.IVPC.full) and ***ut|frågning*** (lit. 'out asking') 'interrogation' (NV.IVPC.semi). There are many cases of deverbal particle+verb compounds, such as ***ut|bildad*** (lit. 'out educated') 'educated' (from ***ut|bilda*** (lit. 'out educate') 'educate'), where the split verbal version is either not possible (**bilda ut*) or has a completely different sense from the compounded verb. According to the previous MWT guidelines, such cases would be annotated as deverbal IVPCs as long as the particle has primary stress. With the new guidelines, however, such words are not considered MWTs and are therefore not annotated as MWEs.

4.3 Issues for Swedish MWE Annotation

4.3.1 MWEs as Part of a Token

Due to frequent compounding in Swedish, the corpus contains instances in which only part of a graphical token belongs to an MWE, while other parts do not. One example is the LVC.full in (5), where only part of the compound word is a part of the LVC. In other cases, an MWE can be embedded within an MWT, such as the NV.IVPC.semi *till|syn* (lit. 'to see') 'supervision' in (6) which has the verbal correspondent ***se till*** (lit. 'see to') 'look after'. With the current PARSEME annotation framework, it is not possible to annotate subtokens, which would be preferable in this case, and thus, the full token is annotated as an MWE. This issue has previously been discussed with a proposed solution (Savary et al., 2023b), but it has not yet been implemented, as it requires changes to the entire annotation process, including the ability to mark subtokens in the annotation tool, process them during consistency checking, and update the CUPT format.

(5) att **ha** vårdnads|**rätt**
to have custody-right
'to have right to custody'

(6) barn**tillsyn**sproblem
barn **till|syn**s problem
child supervision problem
'child supervision issue'

4.3.2 Reflexive Particle Verbs

PARSEME recognizes several subtypes of VMWEs, including idiomatic particle verbs (IVPCs) and inherently reflective verbs (IRVs). In Swedish, it is common for verbs to take both a particle and a reflexive, as in (7) and (8). However, there is no specific category for this combined category; they thus need to be annotated as verbal idioms (VIDs), which is the only possible category for verbs with more than one argument. We would advocate for the inclusion of an IVPC-IRV category for these cases in PARSEME. This issue would also need to be discussed in connection with inherently adpositional verbs (IAVs), currently experimental and not annotated for Swedish, since they can also be combined with reflexives and particles, as in ***slå sig ihop med*** (lit. 'hit oneself together with') 'gang up with'.

(7) att **bry sig om**
to care oneself about
'to care about'

(8) att **ta med sig**
to take with oneself
'to bring'

4.3.3 Splittability of Multiword Tokens

The updated guidelines for determining whether a compound word is an MWE candidate (see Section 4.2) build on splittability, but it is not always clear if a certain compound token is splittable or not. Many compounds, especially adverbial compounds, originate from a co-occurrence of two separate tokens that develop a new, compositional sense, and are accordingly written together as one token more frequently. For instance, *över|allt* (lit. 'over everything') 'everywhere' originates from the two words 'över allt', but it cannot be used in a split form anymore, and it is thus not an MWT according to release 2.0 guidelines. In other cases, as with ***i|dag*** (lit. 'in day') 'today' the compositional sense 'today' can correctly be written also as two separate tokens, 'i dag', which means that the form 'idag' is considered an MWT, according to the *splittable* rule.

5 Comparison of Germanic Language MWEs

To further contextualize the Swedish MWE annotations, we compare the MWE distribution across MWE types with that of other Germanic languages. Since not all languages are available in each release, we select the latest release for each Germanic language present in PARSEME: English 1.2, German 1.3, and Dutch 2.0. We compare this with the two largest Swedish releases, 1.3 and 2.1. Table 6 contains an overview of this comparison for verbal MWEs, and Table 7 contains an overview for Swedish 2.1 and Dutch 2.0, which also contain other MWE types. Some languages annotated the experimental IAV category (Dutch, 80 instances; English, 71 instances), which are not present in the other treebanks. For better comparability, these are excluded from the tables and analysis.

For all languages except English, as expected, the proportion of MWTs is quite high, around 30%. We note that our Swedish guidelines from release 2.1 give an MWT proportion that is more similar to German and Dutch than the permissive rules from release 1.3. We also note that the MWT proportion for types other than VMWEs is considerably higher for Swedish than for Dutch, with a large difference for many categories , such as NIDs (97% versus 13%) and AVs (100% versus 11%). This

	Total	%MWT	IRV	IVPC.full	IVPC.semi	LVC.full	LVC.cause	MVC	VID
Swedish 1.3	3155	51.2	7.5	46.3	18.7	13.2	0.3	0.0	14.0
Swedish 2.1	2275	27.7	13.0	25.5	17.2	22.0	1.3	0.0	21.1
Dutch 2.0	251	26.9	6.0	36.7	3.6	12.7	1.2	0.8	39.0
German 1.2	4041	30.7	8.0	43.2	4.8	7.7	0.8	0.0	35.6
English 1.3	1043	0.35	0.0	35.3	5.1	31.9	4.9	4.9	17.9

Table 6: Distribution of VMWEs for Germanic language releases in PARSEME and proportion of MWTs.

	Total	%MWT	AV	AdjID	AdpID	AdviD	ConjID	DetID	IntjID	NID	NV	Verbal
Swedish 2.1	4793	53.7	2.4	2.0	3.9	19.4	4.6	3.3	0.0	10.3	6.6	47.5
Dutch 2.0	527	25.6	2.8	1.3	3.1	20.1	1.5	4.4	0.4	6.8	4.6	54.5

Table 7: Distribution of MWEs for Germanic language releases in PARSEME and proportion of MWTs.

discrepancy calls for the synchronization of MWT guidelines across the Germanic languages.

Among the non-verbal categories, there are no major differences in distribution between Dutch and Swedish, except that ConjID and NID are more common in Swedish, which may interact with the MWT decisions. For the distribution of VMWEs, English, as expected, stands out from the other Germanic languages, with no IRVs and a high number of LVCs. Compared to Dutch and German, Swedish release 2.1 has a higher proportion of IVPC.semi and IRV, and a lower proportion of IVPC.full. We believe that this is partially due to guideline interpretations, which warrant a more in-depth comparison and discussion for the coming releases. A notable difference between the languages is that Dutch and English have annotated MVCs, whereas Swedish and German do not. This is mainly due to the annotation of expressions with (en) 'let'/ (nl) 'laten', as in *to **let** someone **know***. While this construction exists in Swedish ('låta') and German ('lassen'), it has not been considered an MVC for those languages. The PARSEME guidelines currently have detailed language-specific MVC rules for Hindi and Chinese, with only a single rule on lexical inflexibility applicable to other languages. Our interpretation is that the *let* construction does not pass this rule and should be treated as a regular syntactic construction; however, this needs to be revisited in future guideline discussions. English also includes 'get rid' and 'cross examine' as MVCs.

6 Recommendations

Here, we summarize issues that we recommend the PARSEME community to address in order to strengthen the annotation framework.

Agreement Perform an in-depth analysis of the inter-annotator agreement across languages in release 2.0, to see if there are general difficulties across languages, and update the guidelines to address these issues.

Subtoken annotation Update the PARSEME framework so that subtoken annotation can be used for MWEs that include only a part of a graphical word (See also Savary et al., 2023b).

IVPC-IRV Create joint categories of VMWE types that can co-occur, such as reflexives and particles, and investigate how these classes interact with the IAV class.

MWTs and particles To determine whether a token is an MWT or not, and to distinguish particles from prepositions, create language-specific rules where needed, and synchronize these rules across languages.

Synchronize guidelines Synchronize the guidelines for Germanic languages, especially for IVPCs and IRVs.

MVCs Improve the language-independent guidelines for MVCs and discuss the status of the *let* construction.

7 Conclusion

We discuss the Swedish PARSEME corpus version 2.1, and give a historical overview of previous Swedish PARSEME releases. We present an overview of the Swedish annotations and discuss Swedish-specific considerations, such as the handling of particle verbs and multiword tokens. We also provide an initial comparison of the annotation of Swedish and other Germanic languages, revealing some inconsistencies. The inter-annotator agreement for Swedish is good for verbal MWEs, but quite low for other types, which were recently added to PARSEME. There is thus a need for a renewed overview of the PARSEME guidelines for Swedish as well as for other Germanic languages.

Acknowledgments

This work is supported by the Swedish national research infrastructure Språkbanken, jointly financially supported by the Swedish Research Council (2018–2028; grants 2017-00626 and 2023-00161) and the 10 participating partner institutions. We also received support from the CA21167 COST action UniDive and the IC1207 COST action PARSEME, both funded by the European Union via COST (European Cooperation in Science and Technology).

We thank members of the PARSEME and UniDive communities, especially Agata Savary and Carlos Ramisch, for their support, assistance, and discussions throughout the work on all Swedish releases. We also thank the anonymous reviewers for their feedback and suggestions.

We thank all Swedish annotators and language leaders (LL) both for their annotation work, and for valuable discussions about annotation: Fabienne Cap (LL v1.0), Elsa Erenmalm (v1.2/1.3), Gustav Finnveden (v1.2), Bernadeta Griciūtė (v1.2), Ellinor Lindqvist (v1.2), Stella Lundqvist (v2.0), Ida Nilsson (v2.0), and Joakim Nivre (v1.0), who annotated in addition to the three authors: Astrid Berntsson Ingelstam (v2.0/2.1), Eva Pettersson (v1.0/1.2/2.0), and Sara Stymne (LL v1.2/1.3/2.0/2.1, annotator v1.0/1.2/1.3/2.0/2.1).

References

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47:1191–1211.

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.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Najet Hadj Mohamed, Cherifa Ben Khelil, Agata Savary, Iskander Keskes, Jean Yves Antoine, and Lamia Belguith Hadrich. 2025. PARSEME-AR: Arabic reference corpus for multiword expressions using PARSEME annotation guidelines. *Language Resources and Evaluation*, 59:1331–1361.

Heiki-Jaan Kaalep and Kadri Muischnek. 2006. Multiword verbs in a flective language: the case of Estonian. In *Proceedings of the Workshop on Multi-word-expressions in a multilingual context*.

Murathan Kurfalı, Robert Östling, Johan Sjons, and Mats Wirén. 2020. A multi-word expression dataset for Swedish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4402–4409, Marseille, France. European Language Resources Association.

Eric Laporte, Takuya Nakamura, and Stavroula Voyatzi. 2008. A French corpus annotated for multiword expressions with adverbial function. In *Proceedings of the Language Resources and Evaluation Conference (LREC). Linguistic Annotation Workshop*, pages 48–51, Marrakech, Morocco.

Filip Miletić and Sabine Schulte im Walde. 2024. Semantics of multiword expressions in transformer-based models: A survey. *Transactions of the Association for Computational Linguistics*, 12:593–612.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Kerstin Norén. 1995. Partikelverb och lexikon: kriterier för att finna lexikaliserade partikelverb. In *Nordiske Studier i Leksikografi 3*, pages 321–330. Rapport fra Konference om leksikografi i Norden, Reykjavik, 7–10 juni 1995.

Yagmur Ozturk, Najet Hadj Mohamed, Adam Lion-Bouton, and Agata Savary. 2022. Enhancing the PARSEME Turkish Corpus of Verbal Multiword Expressions. In *18th Workshop on Multiword Expressions (MWE 2022) @LREC2022*, Marseille, France.

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.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, and 6 others. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang,

Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

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, Mexico City, Mexico.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, and 9 others. 2023a. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Agata Savary, Manon Scholivet, Carlos Ramisch, Takuya Nakamura, Eric Bilinski, Sara Stymne, Voula Giouli, Stella Markantonatou, Vasile Păiș, Maria Mitrofan, Louis Estève, Bruno Guillaume, Verginica Barbu Mititelu, Jaka Čibej, Roberto A. Díaz Hernández, Victoria Fendel, Polona Gantar, Olha Kanishcheva, Cvetana Krstev, and 9 others. 2026. PARSEME 2.0 multilingual corpus of multiword expressions. Submitted manuscript `https://gitlab.com/parseme/corpora/-/blob/master/pre-prints/PARSEME_corpus_2.0_pre-print.pdf`.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023b. PARSEME meets Universal Dependencies: Getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*, 9.

Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Carp, and Vasile Păiș. 2026. Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Peter Svenonius. 2003. Swedish particles and directional prepositions. In Lars-Olof Delsing, Cecilia Falk, Gunlög Josefsson, and Halldór Á. Sigurðsson, editors, *Grammar in Focus: Festschrift for Christer Platzack 18 November 2003*, volume II, pages 343–351. Department of Scandinavian Languages, Lund University, Lund.

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.

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.

Therese Lindström Tiedemann, David Alfter, Yousuf Ali, Daniela Piipponen Mohammed, Beatrice Silén, and Elena Volodina. 2024. Multiword expressions in Swedish as a second language: Taxonomy, annotation, and initial results. *Multiword expressions in lexical resources*, page 309.

Therese Lindström Tiedemann, David Alfter, and Elena Volodina. 2022. CEFR-nivåer och svenska flerordsuttryck. In *Svenskan i Finland 19: Föredrag vid den nittonde sammankomsten för beskrivningen av svenskan i Finland, Vasa den 6–7 maj 2021*, pages 218–233. Svensk-österbottniska samfundet.

Abigail Walsh, Teresa Lynn, and Jennifer Foster. 2020. Annotating verbal MWEs in Irish for the PARSEME shared task 1.2. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 58–65, online. Association for Computational Linguistics.

Ukrainian Multiword Expressions Corpus: Creation, Annotation, and Linguistic Analysis

Hanna Sytar
Institute of Slavonic Studies of the Czech Academy of Sciences (ISS CAS), Valentinska 91/1, 11000 Praha
hanna.sytar@slu.cas.cz

Maria Shvedova
National Technical University "Kharkiv Polytechnical Institute"
Kyrpychova str. 2, 61002, Kharkiv
mariia.shvedova@khpi.edu.ua

Olha Kanishcheva
Heidelberg University
Grabengasse 1, 69117 Heidelberg
SET University
Mykoly Shpaka St. 3, 03113, Kyiv
kanichshevaolga@gmail.com

Abstract

This paper presents the development of a corpus of annotated multiword expressions (MWEs) for Ukrainian. The resource covers four major categories of MWEs: verbal, nominal, adjectival/adverbial, and functional. We describe the methodology used for data selection, the annotation scheme, and the procedures employed during annotation. In addition, the paper discusses some specific types of MWE constructions, illustrating their usage with numerous examples and addressing complex and borderline cases. The resulting corpus is an important resource for linguistic studies and NLP tasks involving MWEs, and is publicly accessible here.

1 Introduction

Multiword expressions (MWEs) – such as idioms, light verb constructions, and collocations – play a crucial role in both linguistic theory and natural language processing (NLP) (Constant et al., 2017; Giouli and Barbu Mititelu, 2024). They represent combinations of words whose meaning cannot always be reasoned from their components (non-compositional), and they are essential for accurate parsing, translation, and lexical semantics. The identification and correct processing of MWEs have been shown to improve performance across a wide range of NLP tasks, including machine translation, information extraction, and language modeling. Therefore, the availability of high-quality MWE-annotated corpora is very important for the development of language technologies that can handle idiomatic and non-compositional constructions effectively (Savary et al., 2017).

For the Ukrainian language, MWE research continues to be fairly underexplored. However, despite recent progress in Ukrainian NLP, there is still a lack of systematically annotated data that represent the diversity and complexity of MWEs. This gap is partly due to the linguistic characteristics of Ukrainian – a morphologically rich and syntactically flexible language, where free word order, inflectional variation, and the presence of MWE variants make automatic identification of MWEs particularly challenging. Moreover, the lack of existing linguistic resources limits the ability to train and evaluate computational models for Ukrainian MWE detection.

In this article, we describe our experience in creating, annotating, and analyzing a new corpus of Ukrainian multi-word expressions as part of the multilingual PARSEME shared task[1]. The corpus includes manually annotated MWEs and is designed to support both linguistic research and computational modeling.

This article is structured as follows. Section 2 discusses how phenomena corresponding to multiword expressions in the PARSEME framework are treated within different areas of traditional Ukrainian linguistics, including phraseology, and different branches of grammar. Section 3 describes the corpus structure and data sources. Section 4 outlines the annotation scheme and process, as well as the types of MWEs, with examples. Section 5 addresses complex and borderline cases, including specific constructions not fully covered by the PARSEME scheme, such as multiword particles, challenging instances of inherently adpositional verbs, multiword adpositions, and the variation observed within MWEs, based on the results of the corpus analysis. Section 6 concludes the article and outlines plans.

2 Multiword Expressions in Ukrainian Linguistics

In Ukrainian linguistics, the term MWE is still not widely used. Different types of MWEs are studied within both phraseology and grammar.

Firstly, Ukrainian phraseology traditionally ap-

[1] https://gitlab.com/parseme/corpora/-/wikis/home

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 38–47

plies a narrow approach to fixed expressions, which includes only phraseologisms proper (idioms). Proverbs and sayings, language clichés, etiquette formulas, phraseme-like constructions, and other fixed expressions are not included in this analysis (Alefirenko, 1988; Bilonozhenko et al., 1993; Uzhchenko and Uzhchenko, 2005). Accordingly, well-known Ukrainian phraseological dictionaries contain only idioms (Bilonozhenko et al., 2003, 1993), while proverbs and sayings are compiled separately (Nomys, 1993). Under this traditional approach, most types of MWEs are not represented in lexicographic resources and are not considered in the development of computational tools.

Secondly, various types of MWEs are partially described across separate branches of Ukrainian grammar under different terminological labels. For example, light verb constructions (LVC) correspond to *periphrastic verb-noun constructions* or *periphrastic predicates*, studied within functional-communicative or semantic syntax (Zahnitko, 2001; Sytar, 2010). The term adposition idiom (AdpID) corresponds to what is called a *secondary compound preposition* in Ukrainian grammar (Vykhovanets', 1980; Vykhovanets' et al., 2017), or alternatively *prepositional equivalent* or *prepositional analogue* (Luchyk, 2006; Zahnitko et al., 2007; Kushch, 2008; Zahnitko et al., 2009). Conjunction idioms (ConjID), known as *secondary compound conjunctions*, are recognized as a distinct structural type of conjunctions (Vykhovanets' et al., 2017) and are described lexicographically in (Horodens'ka, 2007; Luchyk, 2006).

The definition of inherently adpositional verbs (IAV) is closely related to the well-developed concept of verbal valency in Ukrainian grammar, including valency-determined obligatory argument positions of the verbal predicate (Vykhovanets', 1988; Zahnitko, 1996; Masyts'ka, 1998) and the concept of verbal government, which has also been lexicographically documented (Kolibaba and Fursa, 2025).

3 Corpus Design and Data Sources

Research on multiword expressions has attracted increasing attention in recent decades, with multilingual NLP initiatives – such as the PARSEME shared tasks (Savary et al., 2017; Ramisch et al., 2020) – establishing common typologies and annotation standards for over 30 languages. These efforts have produced multilingual corpora that now serve as essential benchmarks for automatic MWE processing.

For Slavic languages, existing resources (e.g., for Polish, Czech, and Bulgarian) demonstrate that rich morphology and flexible syntax consistently complicate both annotation and automatic detection (Savary and Waszczuk, 2020; Stoyanova et al., 2016; Pala et al., 2008). Until now, Ukrainian has lacked a fully comprehensive systematically annotated MWE corpus. UD_Ukrainian-ParlaMint (Shvedova et al., 2025) contains MWE information partially through *fixed* dependency relations, with heads annotated using *ExtPos* tags (external POS feature indicating the effective part of speech of an expression); however, this annotation covers only two PARSEME categories (adjectival/adverbial MWEs and functional MWEs)[2].

3.1 Data Sources

All annotated data originate from the General Regionally Annotated Corpus of Ukrainian (GRAC)[3] (Shvedova, 2020). The selected texts come from the *Ukrainian Week* newspaper (2013-2016), and the data type is interview. Initially, the corpus was automatically annotated using the UDPipe 2 model for Ukrainian (ukrainian-iu-ud-2.15-241121)[4](Straka, 2018), providing lemmas, UPOS and XPOS tags, and morphological features in accordance with Universal Dependencies conventions. Multiword expressions were then manually annotated with the FoLiA Linguistic Annotation Tool (FLAT)[5], following the PARSEME MWE 2.0 guidelines.

3.2 Corpus Statistics

The current version of the corpus contains 12,078 sentences with a total of 198,555 tokens, including 5,993 annotated multiword expressions. Each document is enriched with metadata detailing its source, genre, and publication year. Annotations are provided in CoNLL-U format, ensuring compatibility with Universal Dependencies resources and other corpus analysis tools.

The annotated MWEs are categorized as follows: verbal – 2,804, nominal – 818, adjectival and adverbial – 1,017, functional – 1,354, deverbal nouns – 345, and idioms – 1,134.

[2]https://universaldependencies.org/uk/feat/ExtPos.html
[3]https://uacorpus.org/
[4]https://ufal.mff.cuni.cz/udpipe/2/models
[5]https://flat.readthedocs.io/

It is important that our data consists of contemporary journalistic texts containing newly emerging multiword expressions that have not been documented in phraseological dictionaries. Ukrainian phraseological dictionaries were compiled in the late 20th and early 21st centuries, with their primary sources being folklore and works of Ukrainian literature from the 19th to 20th centuries. Examples of such new expressions include *vnutrišn'o peremiščena osoba* 'internally displaced person', *tymčasovo okupovana terytorija* 'temporarily occupied territory', *zeleni čolovičky* 'little green men' (unmarked soldiers), *hlyboka sturbovanist'* 'deep concern' (diplomatic formality masking inaction).

The corpus also contains colloquial variants of fixed expressions, e.g., *vymušenyj pereselenec'* 'forced migrant', *povna majačnja* 'complete nonsense', *vse po fen-šuju* 'everything as it should be' (lit. 'everything according to Feng Shui'), *vidpravyty na try litery* 'tell someone to go to hell' (lit. 'send to three letters').

At the same time, the annotated MWE set includes prepositional units that have not previously been described in Ukrainian grammars or dictionaries, such as *komitet* ***u spravax*** *nacional'nostej* 'committee **on** nationalities'; ***na moment*** *svoho vidkryttja* '**at the time of** its opening'; *Riven' ïï pidtrymky kolyvajet'sja* ***v korydori*** *60-70%* 'its support level fluctuates **between** 60 and 70%'.

Additionally, cases have been observed where the meaning of well-known idioms has shifted; e.g., *imperija zla* 'evil empire', a phrase used by Ronald Reagan in a 1983 speech to refer to the USSR, is used in Ukrainian texts of recent years to denote Russia as a country that continues the totalitarian and imperial policies of the Soviet Union: *– Why do so few Russians sympathize with the Maidan? – Because Russia is an empire.* ***An evil empire.*** *A fragment of the Soviet Union, not yet ready for something different. They want to rule over others; the empire is still coursing through their blood. (Ukrainian Week, 2014; our transl. from Ukr.)*

Therefore, the created corpus can partly compensate for the incompleteness of existing phraseological and grammatical dictionaries of Ukrainian and serve as a valuable resource for addressing various NLP tasks.

4 Annotation Scheme

The annotation scheme for the Ukrainian MWE corpus follows the general principles of the PARSEME Shared Task 2.0 guidelines[6], with adaptations that reflect the specific grammatical and lexical properties of Ukrainian. The goal of the scheme is to maintain cross-linguistic compatibility while accurately capturing constructions characteristic of Ukrainian. Figure 1 presents an example of an output file from our corpus and illustrates the corresponding format.

The Ukrainian MWE corpus is distributed in the standard .cupt format. The linguistic annotation follows the cupt column structure. Lemmas (column 3), UPOS tags (column 4), XPOS tags (column 5), morphological features (column 6), as well as syntactic heads and dependency relations (columns 7-8), are automatically generated using UDPipe 2. The UPOS and FEATS columns follow the Universal Dependencies tagsets, while XPOS is likely based on the AnCora tagset. Additional metadata in the MISC column (column 10) is also automatically provided. The PARSEME:MWE column (column 11) contains manually assigned labels for multiword expression categories, including VID, LVC.full, LVC.cause, IRV, and the experimentally annotated IAV category. All automatic annotations were produced using the UDPipe 2 model[7].

In the following, we describe the main MWE categories and subcategories, together with representative examples of multiword expressions in Ukrainian.

4.1 MWE Types

The top-level categories cover all syntactic types of MWEs and include verbal MWEs (VMWEs), nominal MWEs (NMWEs), adjectival and adverbial MWEs (AMWEs), and functional MWEs (FuncMWEs). This comprehensive classification is introduced in version 2.0 of the annotation guidelines, extending earlier versions of PARSEME that covered verbal MWEs only.

During manual annotation, candidate multiword expressions are classified using category-specific decision diagrams. The annotation scheme distinguishes four major MWE classes: verbal, nominal, adjectival/adverbial, and functional.

Verbal MWEs (VMWEs) are subdivided into universal, quasi-universal, language-specific, and

[6]https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/
[7]https://ufal.mff.cuni.cz/udpipe/2/models

```
# source_sent_id = . . news-69-27
# text = У той самий час наша демократія розпадається на друзки .
1   У    у    ADP Spsa    Case=Acc    4   case    _   _   *
2   той той DET Pd--mnsaa   Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing|PronType=Dem  4   det _   _   1:DetID
3   самий   самий   DET Px--mnsaa   Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing|PronType=Prs|Reflex=Yes   4   det _   _   1
4   час час NOUN    Ncmsan  Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing   7   obl _   _   *
5   наша    наш DET Ppslf-sna   Case=Nom|Gender=Fem|Number=Sing|Person=1|Poss=Yes|PronType=Prs  6   det _   _   *
6   демократія  демократія  NOUN    Ncfsnn  Animacy=Inan|Case=Nom|Gender=Fem|Number=Sing    7   nsubj   _   _   *
7   розпадається    розпадатися VERB    Vmpip3s Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin    0   root
8   на  на  ADP Spsa    Case=Acc    9   case    _   _   2
9   друзки  друзка  NOUN    Ncmpan  Animacy=Inan|Case=Acc|Gender=Masc|Number=Plur   7   obl _   SpaceAfter=No   2
10  .   .   PUNCT   U   _   7   punct   _   _   *
```

Figure 1: The output format of the Ukrainian MWE corpus is illustrated below using a sample sentence: *U toj samyj čas naša demokratija rozpadajet'sja na druzky.* 'At the same time, our democracy is falling apart.' In this example, the multi-word expressions *toj samyj* 'the same' (DetID) and *rozpadajet'sja na druzky* 'falling apart' (VID) are annotated (MWEs are in red blocks).

an optional experimental category. Universal VMWEs include light verb constructions and verbal idioms, while quasi-universal categories cover inherently reflexive verbs, idiomatic verb-particle constructions, and multi-verb constructions. Language-specific categories are defined separately for each language, and an experimental category is introduced for inherently adpositional verbs.

Nominal MWEs (NMWEs) comprise nominal idioms, pronominal idioms, and deverbal nominal MWEs derived from verbal MWEs, inheriting their subcategorization.

Adjectival and adverbial MWEs (AMWEs) include adjectival idioms, adverbial idioms, and deverbal MWEs derived from verbal constructions.

Finally, **functional MWEs (FuncMWEs)** form a universal class and include determiner, adposition, conjunction, and interjection idioms.

More details about the MWE subtypes and examples can be seen in Table 1.

This typology ensures comprehensive coverage of syntactic and functional MWE types, providing a consistent framework for manual annotation and supporting subsequent computational processing.

Fig. 2 shows the distribution of Ukrainian multi-word expression types by frequency. The vertical axis lists the MWE types, while the horizontal axis represents the number of occurrences of each type in the corpus.

The distribution is uneven and highly skewed. The most frequent type is IAV, which clearly dominates all other categories with more than 1,200 instances. Other high-frequency types include LVC.full, AdpID, and AdvID, each represented by several hundred occurrences.

A noticeable but lower frequency is observed for VID and NID, which form a medium-frequency group. In contrast, many MWE types (such as AV.LVC.cause, IVPC.full, MVC, and NV.VID) are represented by only a few instances.

Overall, the diagram demonstrates a long-tail distribution typical of linguistic data: a small number of MWE types account for the majority of occurrences, while most types occur rarely.

4.2 Annotation Process

Annotation was performed manually by two linguists using the FLAT annotation platform[8]. The annotators followed detailed written guidelines derived from the PARSEME framework[9]. Ambiguous cases and borderline expressions were discussed collaboratively to ensure consistency.

4.3 Quality Control and Agreement

To assess annotation reliability, a subset of 20 files was independently annotated by two researchers. Inter-annotator agreement (IAA) was calculated using the MWE-based F-measure (Savary et al., 2017), resulting in an MWE-based F-score of 54. Disagreements were resolved through discussion and guideline analysis.

The evaluation of MWE annotation shows significant variation across categories, with functional expressions showing the highest reliability. AdpID and AdvID achieved the most robust results, with F1-scores of 0.845 and 0.754 respectively, suggesting that adpositional and adverbial idioms are more easily identifiable in Ukrainian. Conversely, verbal constructions such as IAV and LVC.cause suffer from a severe *recall gap* (0.158 and 0.156), where high precision indicates that while annotations are accurate, a vast majority of instances remain undetected. More detailed information about the evaluation scores for each MWE class can be seen in Figure 3.

[8] https://flat.readthedocs.io/en

[9] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/

Type / Subtype	Examples / Description
LVC.full	Semantically bleached verb (e.g., *vyslovyty zaperečennja* 'to raise an objection')
LVC.cause	Verb adds causative meaning (e.g., *spryčynyty rujnuvannja* 'to cause destruction')
VID	Verbal idioms (e.g., *skakaty v hrečku* 'to commit adultery', lit. 'to jump into buckwheat')
IRV	Inherently reflexive verbs (e.g., *dozvolyty sobi* 'to afford', lit. 'to allow oneself')
IVPC.full / IVPC.semi MVC	Multi-verb constructions (e.g., *daty (komus') zrozumity* 'let someone know / make clear')
IAV (experimental)	Inherently adpositional verbs (e.g., *vplyvaty na* 'influence smth.')
NID	Nominal idioms (e.g., *prymxa doli* 'whim of fate')
PronID	Pronominal idioms (e.g., *odyn odnoho* 'one another')
NV	Deverbal nominal MWEs derived from VMWEs (e.g., *znjattja sankcij* 'lifting of sanctions')
AdjID	Adjectival idioms (e.g., *tak zvanyj* 'so-called')
AdvID	Adverbial idioms (e.g., *ostannim časom* 'recently', lit. 'in recent times')
AV	Deverbal AMWEs derived from VMWEs (e.g., *ozbrojenyj do zubiv* 'heavily armed', lit. 'armed to the teeth')
DetID	Determiner idioms (e.g., *toj čy inšyj* 'a particular', lit. 'one or another')
AdpID	Adposition idioms (e.g., *pid čas* 'during', lit. 'under time')
ConjID	Conjunction idioms (e.g., *dlja toho, ščob* 'in order to' lit. 'for that to')
IntjID	Interjection idioms (e.g., *Slava Bohu!* 'Thank God!')

Table 1: Classification of multiword expression types with their main categories and examples.

5 Discussion

In this section, we discuss several notable features and challenges encountered during the annotation of Ukrainian MWEs, highlighting patterns that may be relevant for other Slavic languages and suggesting potential extensions to the existing classification framework.

5.1 Particle Idioms

Our annotation experience with Ukrainian MWEs suggests that the current classification employed in the project would benefit from the inclusion of an additional type, **Particle Idioms (PartID)**. The news corpus contains a considerable number of multiword particles that, during annotation, were assigned to the *Other* category: *vse ž taky* 'after all / still / nevertheless', *navrjad čy* 'hardly / unlikely', *xiba ščo* 'unless / except perhaps', *xoč by* 'at least', *xoča b* 'at least', etc.

We assume that multicomponent (compound) particles are not specific to Ukrainian alone (Zahnitko and Karataieva, 2012), but are also characteristic of other Slavic languages; cf. Czech *kéž by* 'if only / I wish', *ještě aby* 'as if (… were to)': Czech.: ***Kéž by** se mu to povedlo!* 'May he succeed in this!' ***Ještě aby** si stěžoval!* Lit. 'As if he were to complain!'; idiomatic meaning: 'He has no right to complain.' Polish: *Trzeba próbować, **a nuż** się uda?* 'You should try, **what if** it works?' *Płowa zwierzyna to **bądź co bądź** zwierzyna szlachetna.* 'An ungulate is, **after all**, a noble animal', lit. *bądź co bądź* 'be what be'.

During annotation, cases were found where combinations such as *particle+preposition*, *pronoun+preposition*, etc. are used as compound particles, i.e., in contemporary Ukrainian, they function as **Particle Idiom (PartID)**: *Jakščo ljudyna xoče provezty vodu, ščodennyky, olivci, to **do čoho** tut Služba bezpeky?* 'If a person wants to transport water, diaries, pencils, **what** does the Security Service **have to do with** it?' ***Ščo za** dyvyna taka xovajet'sja za cym terminom, my šče pohovorymo nyžče.* '**What kind of** wonder lies behind this term, we will discuss below.' *Novyny ne dyvljusja – ščos' meni **ne do** nyx.* 'I don't watch the news – I am **not in the mood for** it'.

These cases are a zone of intersection between MWEs and phraseme constructions. In the COST Action CA22115[10] Memorandum is indicated that phraseme construction (PhraCons) is a construction that "consist of one or more lexically fixed

[10]https://www.phraconrep.com/

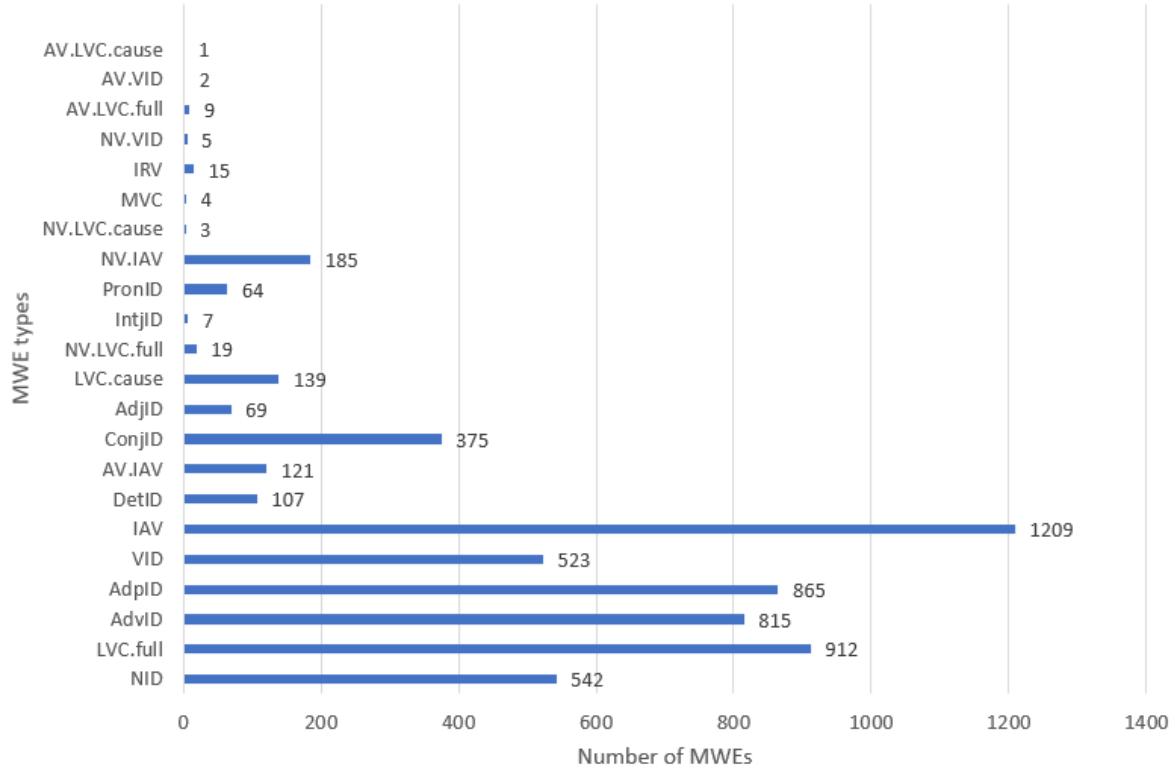


Figure 2: Distribution of Ukrainian MWE types by frequency.

element/s (=anchor/s) and one or more open slot/s (Dobrovol'skij, 2011). The slots must be filled by lexical elements (=fillers) according to lexical, grammatical, communicative, stylistic and intonational rules. Although PhraCons are partially schematic, they have an abstract overall meaning that is usually idiomatic, which means it cannot be attained simply by adding up the meanings of its constituents". This specific type of construction is the focus of *COST Action CA22115* - A Multilingual Repository of Phraseme Constructions in Central and Eastern European Languages (PhraConRep) (Braxatorisová, 2024), where Ukrainian is among the 15 languages under study. The structural, semantic, and pragmatic properties of phraseme constructions in Ukrainian from the perspective of construction grammar are described in (Sytar, 2017). It should be noted that in such contexts the phraseme construction is broader than the MWE: the phraseme construction corresponds to the pattern N_{dat} *ne do* N_{gen} (*Meni* ***ne do*** *novyn.* 'I am **not in the mood for** news.'), whereas the MWE is limited to *ne do*.

5.2 Inherently Adpositional Verb (IAV)

As shown in Figure 2, our text corpus revealed 1,209 contexts of inherently adpositional verbs (IAVs), which constitutes the absolute majority, exceeding the predictably frequent nominal, adverbial, and verbal idioms (912, 815, and 523, respectively). Cases of special optional and experimental inherently adpositional verbs (or prepositional verbs) caused the greatest difficulties and required discussion and agreement among annotators.

According to the project documentation, this MWE type encompasses two groups of cases: "It consists of a verb or VMWE and an idiomatic selected preposition or postposition that is either always required or, if absent, changes the meaning of the verb or VMWE significantly."[11] Both subtypes are present in Ukrainian:

a) verbs with mandatory postverbal prepositional complement: *asocijuvatysja z*+Ins 'to be associated with smth.', *vplyvaty na* +Acc 'to influence smth.', *gruntuvatysja na*+Dat 'to be based on smth.', *narażatysja na*+Acc 'to face smth.'.

b) polysemous verbs, which have different meanings with and without a prepositional complement:

[11] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/index.php?page=iav#iav

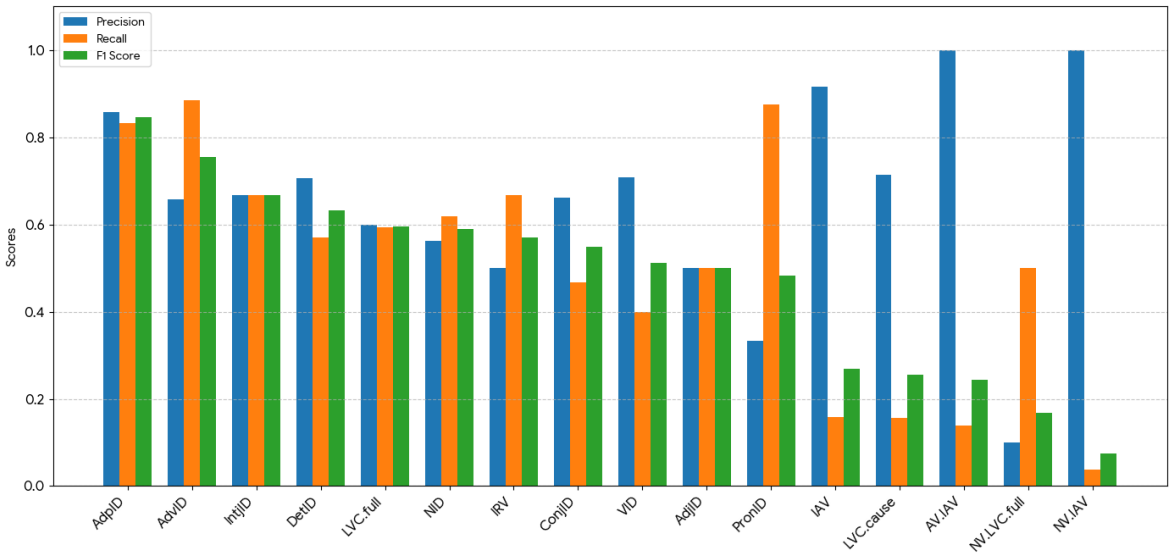


Figure 3: Precision, Recall, and F1-score across Ukrainian MWE classes.

rozraxovuvaty+Acc ‘to calculate smth.’ vs. *rozraxovuvaty na*+Acc ‘to count on smth.’, *zvil’nytysja* ‘to become vacant’ vs. *zvil’nytysja vid*+Gen ‘to get rid of smth.’.

Moreover, in the Ukrainian material we can identify polysemous verbs that occur both without a prepositional complement and with different prepositional complements in different meanings, cf.: *tjažity* ‘to gravitate’, *tjažity nad*+Ins ‘to weigh upon smth.’, *tjažity do*+Gen ‘to gravitate towards smth.’

Finally, the theoretically well-developed concepts of valency-predicted obligatory position and valency-predicted optional position (Vykhovanets’, 1993; Zahnitko, 2001) proved difficult to differentiate in practice. Firstly, this is due to the possibility of ellipsis – the omission of certain structural components in a sentence, including prepositions, which can be easily recovered from context, cf.: *Firmy* ***konkurujut’****, ščob prodaty teplo v merežu.* ‘Companies **are competing** for the right to sell heat to the network.’ (i.e. *Firmy* ***konkurujut’ odna z odnoju*** ‘Companies **compete with one another**’ = *Firmy* ***konkurujut’ miž soboju*** ‘Companies **compete among themselves**’). Secondly, in colloquial speech and social media posts, verbs may be used without their prepositional complements in non-normative way, cf.: *To* ***zaležyt’ vid*** *bahat’ox čynnykiv.* ‘It **depends on** many factors.’ (normative) vs. *To* ***zaležyt’****.* ‘It **depends**’ (colloquial). *To* ***zaležyt’ vid*** *toho, jak pytaty.* ‘It **depends on** how you ask.’ (normative) vs. *To* ***zaležyt’****, jak pytaty.* (Twitter 2017; colloquial, prepositional complement omitted).

5.3 Adposition idiom (AdpID)

A distinctive feature of Ukrainian MWEs is the significant number of secondary multi-component (compound) prepositions (Adposition idiom (AdpID)): ***u mežax*** *spravy* ’**within the scope of** proceedings’, ***za pidsumkamy*** *vizytu* ‘**based on the results of** the visit’. This is the third most common class of MWEs, comprising 865 units (see Figure 2). Such units reflect the processes of grammaticalization and phraseologization that are ongoing in the current stage of Ukrainian language development. For more information on the expansion of the group of secondary prepositions in Ukrainian, see (Zahnitko et al., 2007; Sytar and Zahnitko, 2025).

Similar processes within the prepositional and conjunctional subsystems of Ukrainian are noteworthy, and one can conclude that these units compete, cf.: ***nezvažajučy na*** *partiju* ‘**regardless of** party’ (preposition) vs. ***nezvažajučy na*** *te, xto do jakoï partiï naležyt’* ‘**regardless of** who belongs to which party’ (conjunction); ***nezaležno vid*** *rivnja osvity* ‘**regardless of** education level’ (preposition) vs. ***nezaležno vid*** *toho, jakyj riven’ osvity vin maje* ‘**regardless of** what education level he has’ (conjunction); ***vidpovidno do*** *real’nyx doxodiv* ‘**according to** actual income’ (preposition) vs. ***vidpovidno do*** *toho, jaki ïx real’ni doxody* ‘**according to** what their actual income is’ (conjunction), etc.

5.4 Variants of Idioms

Despite the fact that one of the characteristics of MWEs is their fixedness or limited flexibility, variants of idioms that are easily identified by native speakers and do not alter the holistic meaning of MWEs have proven to be characteristic of contemporary Ukrainian. However, this very variability of MWEs can complicate their automatic identification in text and the performance of other NLP tasks. According to our observations, idiom variants arise through the following transformations:

a) constructions with zero copula, typical of Ukrainian and other East Slavic languages: *Istyna zavždy poseredyni.* 'The truth always **lies** in the middle.' *U straxu velyki oči.* 'Fear **has** big eyes.' We classified such cases as verbal MWEs despite the formal absence of the verb.

b) introduction of additional components into MWEs: verbal phrase *prolyty svitlo* 'to shed light' modified by the adverbial particle *troxy* 'a little': *U rozmovi vin **prolyv troxy svitla** na perspektyvy ukraïns'kyx bankiv* 'In conversation, he **shed some light** on the prospects for Ukrainian banks.' The verbal phrase *povernutysja na Olimp* 'return to Olympus' is modified by the introduction of the possessive adjective *kyïvs'kyj*: *povernutysja na kyïvs'kyj Olimp* 'return to the Kyiv Olympus'. Common nominal idioms *krok upered* 'step forward' and *krok nazad* 'step back': *Te, ščo my robymo, - **krok upered, try vbik, potim odyn nazad**.* 'What we are doing is one **step forward, three steps to the side, then one step back**.'

c) replacement of components: the biblical expression *prodaty za mysku sočevyčnoï jušky* 'to sell (something) for a bowl of lentil stew' is transformed into *prodaty za tarilku boršču* 'to sell (something) for a plate of borshch' (a traditional Ukrainian dish): *Such propaganda exploits the servile mentality of the "nostalgic Soviet type", who is willing to sell freedom for a plate of borshch (Ukrainian Week, 2014; our transl. from Ukr.).*

d) omission of components: In the verbal phrase *zaxyščaty čest' [svoho] mundyra* 'to defend the honor of [one's] uniform', the noun for 'honor' is omitted in the interview text: *Vony* [pracivnyky sylovyx struktur] *duže **zaxyščajut' svij mundyr**.* 'They [law enforcement officers] strongly **defend their uniform**'.

Special attention was required during annotation for cases in which multiple types of MWEs were combined: *Amerykans'ka delehacija **vxodyt' do skladu** Parlaments'koï asambleï* 'The American delegation **forms part of** the Parliamentary Assembly.': *vxodyt' do*+Gen 'to be part of smth.' is an inherently adpositional verb, and *do skladu*+Gen is an adposition idiom, 'in smth.', lit. 'into the composition of smth.'. *Centr protydiï teroryzmu ta hibrydnym zahrozam **sprjamovuje** svoï **zusyllja na** vidbyttja kiberatak* 'The Center for Countering Terrorism and Hybrid Threats **directs its efforts toward** repelling cyberattacks.': *sprjamovuje zusyllja* 'directs efforts' is a light verb constructions, and *sprjamovuje na* 'directs toward' is an inherently adpositional verb.

These observations highlight the complexity of MWE phenomena in Ukrainian and point to areas requiring further investigation and annotation improvement. Currently, the annotation scheme does not provide a mechanism to link the variants as alternative forms of the same multiword expression. In future work, it would be beneficial to incorporate this functionality into the annotation framework.

6 Conclusions and Future Plans

The obtained results show an imbalance in the distribution of MWE types in Ukrainian interview texts. On the one hand, this imbalance highlights specific features of Ukrainian phraseology and grammar. On the other hand, these findings require further validation on a larger corpus and through the inclusion of data from other text styles.

The analysis of Ukrainian data also indicates the need to refine the existing MWE classification. In particular, we propose introducing a separate category within functional MWEs, namely Particle Idioms.

As a direction for future research, we plan to identify and analyze cases involving overlaps between different MWE types, which will contribute to a more precise and comprehensive description of multiword expressions in Ukrainian.

These findings and future research directions are enabled by the creation of a dedicated Ukrainian MWE resource. With the expansion of PARSEME in 2025 to include additional MWE types and languages, Ukrainian became part of the shared task for the first time. Supported by the UniDive project[12], we successfully integrated Ukrainian into this international initiative. Based on established annotation guidelines and previous linguistic research, the resulting resource represents the first

[12] https://unidive.lisn.upsaclay.fr/

comprehensive corpus of Ukrainian multiword expressions and supports both theoretical research and computational modeling.

Limitations

The presented resource has several limitations that should be mentioned. First, the size of the corpus is still relatively small compared to MWE datasets for Romanian, Hebrew, and Polish languages (more than 13,000 MWEs). Although it is useful for initial research, some rare constructions may not be well represented. In the future, the corpus should be expanded to include more text types and topics.

Second, even though the annotation scheme follows the PARSEME Shared Task 2.0 guidelines, some Ukrainian-specific constructions do not fit perfectly into the existing categories. In such cases, annotators had to rely on internal decisions, which may lead to small inconsistencies or unclear borderline cases.

Third, the automatic linguistic annotation produced by UDPipe 2 (such as tokenisation, lemmas, POS tags, or dependencies) may contain errors. While MWEs were annotated manually and independently of these layers, such automatic mistakes can still influence how some expressions are interpreted.

Finally, the annotation was carried out by a small group of annotators. Although they worked together and discussed difficult cases to ensure consistent decisions, involving more annotators and calculating formal inter-annotator agreement in future work would further increase the reliability of the resource.

These limitations point to several directions for improvement, such as extending the corpus, refining annotation rules, and adding more quality-control procedures.

Acknowledgments

We would like to thank the reviewers for their time and effort in reviewing this manuscript. We sincerely appreciate their valuable comments and suggestions, which greatly helped us improve the quality of the work. This research was partially funded by the Alexander von Humboldt Foundation, and this work received support from the COST Action CA21167 'UniDive'[13] (European Cooperation in Science and Technology). The authors are also grateful to Friedrich Schiller University Jena for providing the research facilities and support that made this work possible.

[13]https://unidive.lisn.upsaclay.fr/

References

Mykola Alefirenko. 1988. *Teoretychni pytannia frazeolohii*. Vyshcha shkola, Kharkiv.

Vira Bilonozhenko and 1 others. 1993. *Frazeolohichnyi slovnyk ukrainskoi movy*. Naukova dumka, Kyiv.

Vira Bilonozhenko and 1 others. 2003. *Slovnyk frazeolohizmiv ukrainskoi movy*. Naukova dumka, Kyiv.

Anita Braxatorisová, editor. 2024. *Synsemantika in Phrasem-Konstruktionen im Deutschen und anderen Sprachen*. Logos Verlag Berlin GmbH, Berlin.

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

Dmitrij Dobrovol'skij. 2011. Phraseologie und konstruktionsgrammatik. In Alexander Lasch and Alexander Ziem, editors, *Konstruktionsgrammatik III. Aktuelle Fragen und Lösungsansätze*, pages 110–130. Tübingen.

Voula Giouli and Verginica Barbu Mititelu, editors. 2024. *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*. Number 6 in Phraseology and Multiword Expressions. Language Science Press, Berlin. Published under CC BY 4.0.

Kateryna Horodens'ka. 2007. *Hramatychnyi slovnyk ukrainskoi movy: spoluchnyky*. Vydavnytstvo KhDU, Kherson.

Larysa Kolibaba and Valentyna Fursa. 2025. *Slovnyk diieslivnoho keruvannia: u 2 t.* Pidruchnyky i posibnyky, Kyiv.

Natalija Kushch. 2008. *Pryimennykova ekvivalentnist v ukrainskii hramatytsi: struktura, semantyka, funktsii*. Ph.D. thesis, Donetsk.

Alla Luchyk. 2006. *Slovnyk ekvivalentiv slova ukrainskoi movy*. Wydawnictwo Uniwersytetu Śląskiego, Katowice.

Tetiana Masyts'ka. 1998. *Hramatychna struktura dieslivnoi valentnosti*. RVV "Vezha" VDU im. Lesi Ukrainky, Lutsk.

Matvij Nomys. 1993. *Ukrainski prykazky, pryslivia i take inshe*. Lybid, Kyiv.

Karel Pala, Lukáš Svoboda, and Pavel Šmerk. 2008. Czech MWE database. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Agata Savary and Jakub Waszczuk. 2020. Polish corpus of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 32–43, online. Association for Computational Linguistics.

Maria Shvedova. 2020. The general regionally annotated corpus of ukrainian (grac, uacorpus.org): Architecture and functionality. In *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020), Volume I: Main Conference*, CEUR Workshop Proceedings, pages 489–506, Lviv, Ukraine.

Maria Shvedova, Arsenii Lukashevskyi, and Andriy Rysin. 2025. Developing a Universal Dependencies treebank for Ukrainian parliamentary speech. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 55–63, Vienna, Austria (online). Association for Computational Linguistics.

Ivelina Stoyanova, Svetlozara Leseva, and Maria Todorova. 2016. Towards the automatic identification of light verb constructions in bulgarian. In *Proceedings of CLIB 2016*, pages 28–37, Sofia, Bulgaria.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Hanna Sytar. 2010. Opysovi predykaty z modalnym komponentom v ukrainskii movi: struktura y semantyka. *Linhvistychni studii*, (20):145–151.

Hanna Sytar. 2017. *Syntaksychni frazeolohizmy v rozrizi konstruktsiinoi hramatyky*. TOV "Nilan-LTD", Vinnytsia. Monograph.

Hanna Sytar and Anatolii Zahnitko. 2025. Sekundární předložky s variantní valencí vyjadřující determinací akce v ukrajinštině. *Prace Filologiczne*, 80:325–350.

Viktor Uzhchenko and Dmytro Uzhchenko. 2005. *Frazeolohiia suchasnoi ukrainskoi movy*. Almamater, Luhansk.

Ivan Vykhovanets'. 1980. *Pryimennykova systema ukrainskoi movy*. Naukova dumka, Kyiv.

Ivan Vykhovanets'. 1988. *Chastyny movy v semantykohramatychnomu aspekti*. Naukova dumka, Kyiv.

Ivan Vykhovanets'. 1993. *Hramatyka ukrainskoi movy. Syntaksys*. Lybid, Kyiv.

Ivan Vykhovanets', Kateryna Horodens'ka, Anatolii Zahnitko, and Svitlana Sokolova. 2017. *Hramatyka suchasnoi ukrainskoi literaturnoi movy. Morfolohiia*. Vydavnychyi dim Dmytra Buraho, Kyiv.

Anatolii Zahnitko. 1996. *Teoretychna hramatyka ukrainskoi movy. Morfolohiia*. DonDU, Donetsk.

Anatolii Zahnitko. 2001. *Teoretychna hramatyka ukrainskoi movy. Syntaksys*. DonNU, Donetsk.

Anatolii Zahnitko, Illya Danyliuk, Hanna Sytar, and Inna Shchukina. 2007. *Slovnyk ukrainskykh pryimennykiv*. TOV VKF "BAO", Donetsk.

Anatolii Zahnitko and Anna Karataieva. 2012. *Slovnyk chastok: materialy i statti*. DonNU, Donetsk.

Anatolii Zahnitko, Kateryna Vynohradova, Illya Danyliuk, Nadija Zahnitko, Natalija Kushch, Maryna Orans'ka, Tetjana Kitaieva, Hanna Sytar, Valerija Chekalina, and Inna Shchukina. 2009. *Funktsionalno-komunikatyvna i tekstova paradyhma ukrainskykh pryimennykiv ta yikhnikh ekvivalentiv*. Weber (Donetska filiia), Donetsk.

Cognitive Signatures of Multi-Word Expressions: Reading-Time and Surprisal

Diego Alves and Sergei Bagdasarov and Elke Teich
Saarland University
Saarbrücken, Germany
diego.alves@uni-saarland.de, sergeiba@lst.uni-saarland.de,
elke.teich@uni-saarland.de

Abstract

This study investigates whether eye-tracking measures predict if a word is the final token of a multi-word expression (MWE), focusing on two understudied MWE types: fixed expressions (e.g., *due to*) and phrasal verbs (e.g., *turn out*). Using mixed-effects logistic regression, we compared tokens in MWE contexts with the same tokens in non-MWE contexts. Results reveal a clear difference in processing. For fixed expressions, reading-time measures significantly predict MWEhood. In contrast, phrasal verbs show no consistent predictive effects. Additionally, we compared the reading-time models to models that included GPT-2 surprisal as a predictor. While surprisal does predict MWEhood, it fails to capture the distinction between types. These findings highlight the need to consider MWE typology in models of formulaic language processing.

1 Introduction

Across languages, certain word combinations, known as multi-word expressions (MWEs), are conventional patterns associated with specific meanings or connotations. MWEs take diverse forms, ranging from structurally fixed idioms with figurative meanings (e.g., *break the ice*), to compounds (e.g., *sea water*), which vary in compositionality, and phrasal verbs (e.g., *carry out*), which can be either compositional or idiomatic and are often lexically productive (Avgustinova and Iomdin, 2019).

MWEs are ubiquitous because they enhance language efficiency through predictable transitions between words. Highly conventionalised MWEs can be retrieved holistically from the lexicon rather than incrementally processed, providing a processing advantage over novel sequences (Siyanova-Chanturia et al., 2017). From a communicative perspective, MWEs reduce cognitive load for language users, serving as devices that streamline processing and facilitate comprehension (Conklin and Schmitt, 2012). Many studies have demonstrated the processing advantages of MWEs using eye-tracking and event-related potentials (ERP). These studies show that MWEs are generally read and processed more efficiently than novel sequences, with facilitation influenced by factors such as frequency, predictability, familiarity, and type-specific properties (e.g., Siyanova (2010); Carrol and Conklin (2020); Kessler et al. (2021)).

As shown by Carrol and Conklin (2020), different types of MWEs exhibit different cognitive processing patterns. In the present study, we focus on two types: fixed expressions (e.g., *due to*, *out of*) and phrasal verbs (e.g., *turn out*, *rush in*). These small lexical units have been understudied in research on MWE processing. Our analysis focuses on the final token of each sequence because MWEs are characterized by highly predictable transitions between constituent tokens. This predictability advantage is expected to manifest most clearly at the final token, which is processed more rapidly when it completes an MWE than when the same token appears in a non-MWE context. To do this, we compare tokens appearing in MWEs with the same tokens when they occur in non-MWE contexts. Additionally, we compare the results obtained using reading-time predictors with models based on surprisal estimates from a large language model, to examine whether the surprisal behaviour of the final token also differs according to MWE type.

2 Related Work

Eye-tracking studies have long shown that gaze patterns are sensitive to linguistic and contextual factors, including lexical frequency, verb complexity, and ambiguity (Rayner (1975); Rayner and Duffy (1986); Rayner et al. (2012)), providing a foundation for understanding real-time processing of multi-word expressions (MWEs) and formulaic language.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 48–53

Frequency strongly influences MWE processing. Siyanova (2010) found that high-frequency MWEs are processed more efficiently by native speakers, whereas non-native speakers benefit mainly from very high-frequency items. Conklin and Schmitt (2012) review evidence that MWEs are generally read faster than novel sequences, with speed modulated by frequency, predictability, and transparency. Pellicer-Sánchez and Perez (2024) similarly highlight frequency, familiarity, predictability, and decomposability as robust predictors of processing ease, especially for L1 readers.

Different MWE types exhibit distinct patterns. Carrol and Conklin (2020) reported a general processing advantage for idioms, binomials, and collocations, with type-specific effects: idioms were sensitive to frequency, familiarity, and decomposability; binomials to predictability and semantic association; collocations to mutual information. Kessler et al. (2021) extended this to spoken idioms, showing listeners fixate predicted completions and early semantic associates, with ERP data indicating facilitated processing for correct completions.

Late gaze measures, including regressions and re-reading, reliably distinguish MWEs from novel sequences. Rohanian et al. (2017) showed that combining gaze features with part-of-speech and frequency enables computational models to predict MWEs, consistent with findings that early gaze measures are less informative (Siyanova-Chanturia, 2013).

The predictability of a final MWE element can also be formalized with surprisal, the negative log probability of an event (Shannon (1948)), with higher surprisal leading to longer fixations. Onnis and Huettig (2021) applied this to MWEs, showing that frequent and predictable sequences are easier to integrate, whether stored as chunks or composed. Moreover, Alves et al. (2025) show that the negative surprisal slope over token sequences is a strong predictor of MWEhood.

In this study, we focus on two under-studied MWE types, using regression models to examine whether reading-time measures predict MWEhood. We also compare these effects with surprisal estimates from a large language model, which have been shown to predict reading times (Wilcox et al., 2023).

3 Methodology

3.1 Data

We used two eye-tracking corpora: UCL (Frank et al., 2013) and Provo (Luke and Christianson, 2018).

The UCL dataset includes self-paced reading times and eye-tracking data from 361 English sentences drawn from three novels. The participants were native speakers and first-year psychology students (104 self-paced readers and 42 eye-tracking participants; mostly native speakers). Reading-time measures include word-by-word response times, first-fixation, first-pass gaze duration, and total fixation.

The Provo Corpus contains eye-tracking data from 84 native English-speaking adults reading 55 short passages (134 sentences, 2,745 words) from news, fiction, and popular-science texts. Measures include fixation durations, number of fixations, skipping, regressions, and cloze-based predictability norms. Unlike isolated sentence corpora, Provo captures more naturalistic, continuous reading, making it particularly suitable for studies of predictive processing.

Sentences from both corpora were automatically annotated using the Universal Dependencies framework with Stanza (Qi et al., 2020) and the combined English model. Fixed expressions and phrasal verbs were identified from tokens labeled as `fixed` and `compound:prt`, respectively, and assigned a value of 1 (MWE), while tokens with the same surface form but different labels were assigned 0 (non-MWE).

3.2 Reading-time Measures and Surprisal

In this study, we focus on three widely used reading-time measures (Rayner, 1998). First fixation duration refers to the duration of the initial fixation on a word during first-pass reading. Gaze duration is the sum of all first-pass fixations on a word, while total fixation duration represents the total time spent fixating on a word, including regressions.

First fixation duration reflects early lexical access, gaze duration captures lexical and syntactic processing during initial reading, and total fixation duration indexes later comprehension stages such as reanalysis and integration difficulties (Rayner, 1998).

For the comparison of reading-time measures with surprisal, we estimated the surprisal of each

word using the smallest GPT-2 model[1] (Radford et al., 2019). We use GPT-2 because prior work has shown that surprisal estimates from larger transformer-based language models often provide a poorer fit to human reading times than smaller models, likely because increased capacity leads to representations that diverge from human incremental processing (Oh and Schuler, 2023). Surprisal values were extracted using the `surprisal`[2] Python library. Word-level surprisal was computed by summing the surprisal values of the constituent subword tokens.

3.3 Regression Models

We performed logistic mixed-effects regression analyses in R, using lme4 library (Bates et al., 2015), to examine whether reading-time (RT) measures, tested one at a time, predict the likelihood that a word is part of a multi-word expression (MWEhood). Our analysis focusses specifically on the final tokens that occur either in fixed expressions or in phrasal verbs, comparing their behaviour when they appear in MWE contexts versus non-MWE contexts. For each token, we fitted a logistic mixed-effects model with predictors including the current word's RT, word length, their interaction, spillover RTs and word lengths of the two preceding words, and random intercepts for participants (Equation 1).

$$\begin{aligned} \text{MWEhood}_{ij} \sim\ & \text{RT}_{ij} \times \text{WordLength}_{ij} \\ + \text{RT}_{i,j-1} + & \text{RT}_{i,j-2} + \text{WordLength}_{i,j-1} \\ + \text{WordLength}_{i,j-2} + & (1 \mid \text{Subject}_i) \end{aligned} \quad (1)$$

The same type of regression was conducted in a second step, replacing the reading-time measure with surprisal estimates derived from a GPT-2 language model for the occurrences of the MWEs in the Brown corpus (Francis, 1965).

Finally, to complement our analysis, we calculated the pointwise mutual information (PMI) for each fixed expression and phrasal verb identified in the corpora[3]. The idea is to test whether PMI values can account for the differences observed between the reading-time models and the surprisal models.

[1] https://huggingface.co/openai-community/gpt2

[2] https://pypi.org/project/surprisal/

[3] All MWEs extracted from the corpora for this study were bigrams.

4 Results

4.1 MWE Identification

From the parsed sentences of both corpora, we extracted several MWEs. In the case of fixed expressions, six were identified in the UCL corpus; however, for four of these, the final token did not appear in a non-MWE context (e.g., *at least*, *in order*). Consequently, only *instead of* and *out of* were considered, with the preposition *of* as the analyzed token. In the Provo corpus, nine fixed expressions were identified, but only *due to* and *out of* had final tokens that also occurred in non-MWE contexts.

Regarding phrasal verbs, seventy were extracted from the UCL corpus. The ones retained for our analysis included three with the particle *on* (*knock on*, *go on*, and *caught on*), six with *in* (e.g., *fill in*, *step in*), and seventeen with *out* (e.g., *let out*, *knock out*). In the Provo corpus, fifteen phrasal verbs were extracted, of which one had the particle *in* (*rush in*) and five had *out* (*turned out*, *help out*, *dig out*, *built out*, *looked out*).

4.2 Reading Time as MWEhood Predictor

Table 1 shows the significance and AIC values of first fixation, gaze, and total fixation for the final tokens of fixed expressions and phrasal verbs in the UCL and Provo corpora, in predicting whether a token is part of an MWE. Stars indicate statistical significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, ns = not significant), and numbers in the adjacent column represent the corresponding AIC values of the regression models.

It is can be observed that reading-time measures are statistically significant predictors of MWEhood for fixed expressions in both corpora. The coefficients indicate, consistent with previous work, that tokens are read faster when they form part of an MWE. However, in contrast to earlier findings (cf. Siyanova-Chanturia (2013)), we also find significant and consistent effects for first-fixation duration, suggesting that MWE processing advantages can emerge at earlier stages of lexical access than previously reported.

On the other hand, for phrasal verbs, we observed no significant effects (with the exception of the particle *out* in the Provo corpus). Although Kissane et al. (2024) reported that phrasal-verb particles tend to be read more rapidly than verb–preposition bundles, our results align with the findings of Yaneva et al. (2017), who showed that

Corpus	Token	MWE Type	First Fix.	AIC	Gaze	AIC	Total Fix.	AIC
Provo	to	Fixed Expression	**	294	*	294	***	1959
	of	Fixed Expression	*	107	ns	113	ns	491
UCL	of	Fixed Expression	**	522	**	525	ns	531
Provo	in	Phrasal Verb	ns	57	ns	57	ns	508
	out	Phrasal Verb	ns	230	*	224	ns	1045
UCL	on	Phrasal Verb	ns	513	ns	493	ns	504
	in	Phrasal Verb	ns	655	ns	658	ns	661
	out	Phrasal Verb	ns	806	ns	805	ns	805

Table 1: Significance and AIC values of reading-time measures for fixed expressions and phrasal verbs across corpora. Stars indicate statistical significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, ns = not significant).

the final word in verb–particle combinations does not differ in processing between MWEs and control phrases for either native or non-native speakers. This is likely because readers often extract sufficient information about particles before directly fixating on them, resulting in high skipping rates.

In terms of AIC, models trained on the Provo data show better predictive accuracy compared to UCL models for first fixation and gaze duration. However, when total fixation duration is used as a predictor, the Provo-based models are less predictive of MWEhood, suggesting that the status of the token as part of an MWE has a stronger influence during earlier stages of cognitive processing.

When replacing the reading-time measures in equation 1 with surprisal estimates from GPT-2 (also for the previous tokens), we observe a significant effect (*** $p < 0.001$) for all tokens except *in*. This suggests that surprisal does not distinguish between MWE types in a way that reflects the cognitive patterns observed in the eye-tracking data. Additionally, the AIC values for models using surprisal as a predictor are relatively high, over 3,000 for fixed expressions and over 1,200 for phrasal verbs, indicating lower predictive accuracy compared to models using reading-time measures as predictors.

The differences observed in the reading-time models may be due to structural differences between fixed expressions and phrasal verbs. While the former function as grammatical units, the latter behave as lexical items, which entails differences in their overall cognitive processing. Moreover, reading time reflects multiple stages of processing such as lexical access, syntactic integration, and comprehension, whereas surprisal is more limited, capturing only how predictable a token is given the preceding context. Table 2 presents the mean PMI values for the fixed expressions and phrasal-verb particles included in our analysis.

Token	MWE Type	Mean PMI
to	fixed	2.53
of	fixed	3.81
on	PV	3.42
in	PV	1.65
out	PV	3.94

Table 2: Mean PMI values for fixed expressions and phrasal verbs.

Analysing the PMI values of the fixed expressions and phrasal verbs shows that fixed expressions and verb–particle combinations with *out* and *on* generally show the highest PMI scores, although some variability is evident (e.g., the low PMI of *not to*, 0.81). In contrast, phrasal verbs with *in* show the lowest PMI values, which may help explain the lack of significant effects when using surprisal as a predictor. Overall, these results indicate that PMI alone cannot account for the differences observed in the cognitive processing of fixed expressions and phrasal verbs.

5 Conclusion and Future Work

This study examined whether eye-tracking measures predict whether a word is the final token of a multi-word expression (MWE), focusing on fixed expressions (e.g., *due to*) and phrasal verbs (e.g., *go out*). Logistic mixed-effects regression analyses were used to compare reading-time measures for tokens appearing in MWEs versus the same tokens in non-MWE contexts.

The results reveal a clear processing distinction between these MWE types. For fixed expressions, reading times, including early measures such as first-fixation duration, significantly predicted

MWEhood. In contrast, phrasal verbs showed no consistent reading-time differences. Additionally, while surprisal estimates from GPT-2 generally predicted MWEhood, they did not capture this type-specific distinction, and PMI values also failed to account for the observed processing differences.

These findings highlight that MWE type matters: fixed expressions, which function as grammatical units, and phrasal verbs, which behave as lexical items, engage distinct cognitive mechanisms despite both being formulaic.

Although the present study focuses specific classes of English MWEs, the proposed approach is not inherently language-specific. It could be extended to other languages by leveraging tokens labelled as fixed in the Universal Dependencies (UD) framework, which capture a wide range of multiword expressions cross-linguistically. Moreover, while phrasal verbs are characteristic of English, the same methodology could be applied to other MWE types, such as light verb constructions, which are prominent in many languages.

Future work should extend this investigation to additional eye-tracking corpora and other types of MWEs not included in the present study.

Limitations

The findings of this study should be considered in light of its limitations. First, the analysis relies on data from only two eye-tracking corpora (UCL and Provo), which constrains the number and variety of multi-word expressions (MWEs) available for examination. Consequently, many fixed expressions and phrasal verbs were excluded because their final tokens did not appear in comparable non-MWE contexts, reducing statistical power and generalisability. Second, the findings are specific to two MWE types (fixed expressions and phrasal verbs); other important categories were not tested with the same regression approach. Consequently, idiomaticity, transparency, and semantic compositionality are not examined in this paper. Ideally, future eye-tracking experiments would include compounds in both compositional and non-compositional contexts, enabling direct comparison of reading-time measures.

Acknowledgements

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

Diego Alves, Sergei Bagdasarov, and Elke Teich. 2025. Surprisal dynamics for the detection of multi-word expressions in english. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1185–1194.

Tania Avgustinova and Leonid Iomdin. 2019. Towards a typology of microsyntactic constructions. In *International Conference on Computational and Corpus-Based Phraseology*, pages 15–30. Springer.

Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Gabor Grothendieck, Peter Green, and Maintainer Ben Bolker. 2015. Package 'lme4'. *convergence*, 12(1):2.

Gareth Carrol and Kathy Conklin. 2020. Is all formulaic language created equal? unpacking the processing advantage for different types of formulaic sequences. *Language and speech*, 63(1):95–122.

Kathy Conklin and Norbert Schmitt. 2012. The processing of formulaic language. *Annual review of applied linguistics*, 32:45–61.

W Nelson Francis. 1965. A standard corpus of edited present-day american english. *College English*, 26(4):267–273.

Stefan L Frank, Irene Fernandez Monsalve, Robin L Thompson, and Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of english sentence processing. *Behavior research methods*, 45(4):1182–1190.

Ruth Kessler, Andrea Weber, and Claudia K Friedrich. 2021. Activation of literal word meanings in idioms: Evidence from eye-tracking and erp experiments. *Language and Speech*, 64(3):594–624.

Hassane Kissane, Konstantin Tziridis, Achim Schilling, Patrick Krauss, and Thomas Herbst. 2024. Cognitive dynamics of verb-particle constructions: An eye-tracking study. *bioRxiv*, pages 2024–12.

Steven G Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50(2):826–833.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Luca Onnis and Falk Huettig. 2021. Can prediction and retrodiction explain whether frequent multi-word phrases are accessed 'precompiled' from memory or compositionally constructed on the fly? *Brain Research*, 1772:147674.

Ana Pellicer-Sánchez and Maribel Montero Perez. 2024. Eye-tracking in vocabulary research: Introduction to the special issue. *Research methods in applied linguistics*, 3(1):100095.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Keith Rayner. 1975. The perceptual span and peripheral cues in reading. *Cognitive psychology*, 7(1):65–81.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Keith Rayner and Susan A Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & cognition*, 14(3):191–201.

Keith Rayner, Alexander Pollatsek, Jane Ashby, and Charles Clifton Jr. 2012. *Psychology of reading*. Psychology Press.

Omid Rohanian, Shiva Taslimipoor, Victoria Yaneva, and Le An Ha. 2017. Using gaze data to predict multiword expressions.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Anna Siyanova. 2010. On-line processing of multi-word sequences in a first and second language: Evidence from eye-tracking and erp. Technical report, University of Nottingham.

Anna Siyanova-Chanturia. 2013. Eye-tracking and erps in multi-word expression research: A state-of-the-art review of the method and findings. *The Mental Lexicon*, 8(2):245–268.

Anna Siyanova-Chanturia, Kathy Conklin, Sendy Caffarra, Edith Kaan, and Walter JB van Heuven. 2017. Representation and processing of multi-word expressions in the brain. *Brain and language*, 175:111–122.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Victoria Yaneva, Shiva Taslimipoor, Omid Rohanian, and Le An Ha. 2017. Cognitive processing of multiword expressions in native and non-native speakers of english: Evidence from gaze data. In *International conference on computational and corpus-based phraseology*, pages 363–379. Springer.

Cheese it up: CamemBERT Outperforms Large Language Models for Identification of French Multi-word Expressions

Sergei Bagdasarov and Diego Alves and Elke Teich
Saarland University, Germany
sergeiba@lst.uni-saarland.de, diego.alves@uni-saarland.de, e.teich@mx.uni-saarland.de

Abstract

In recent years, language models, both encoder-only and generative, have been applied to a variety of downstream NLP tasks, including sequence labeling tasks like automatic multi-word expression identification (MWEI). Multiple studies show that, in general, fine-tuned encoder-only models like BERT tend to outperform pretrained generative LLMs on downstream tasks (Arzideh et al., 2025; Ochoa et al., 2025; Bucher and Martini, 2024; Sebők et al., 2025). However, such comparisons are sparse for MWEI, in particular for French, in part due to the lack of comprehensive gold-standard datasets. In this study, we address this research gap by comparing CamemBERT with gpt-oss and Qwen3 for MWEI, using the French subcorpus of the newly released PARSEME dataset. CamemBERT outperforms both LLMs by large margins in precision, recall, and F1. We complement this numerical evaluation with a qualitative analysis of prediction errors.

1 Introduction

Multi-word expressions (MWEs) are prefabricated sequences of words that tend to be stored and processed as whole units in memory, rather than composed online from their individual components (e.g. *briser la glace* (break the ice), *lors de* (at the time of), *poser [une] question* (ask [a] question) (Siyanova-Chanturia, 2013; Siyanova-Chanturia et al., 2017). MWEs are ubiquitous in language and are widely attested in different text types, making automatic MWE identification (MWEI) a crucial task in many natural language processing (NLP) applications. At the same time, MWEI remains a challenging problem due to the structural diversity, variability, as well as semantic and syntactic idiosyncrasies of MWEs.

In recent years, transformer-based encoder-only language models have achieved strong performance across a wide range of downstream NLP tasks, including sequence labeling tasks like MWEI (Bello et al., 2023; Garrido-Merchan et al., 2023; Bui and Savary, 2024; Labusch et al., 2019). In parallel, generative large language models (LLMs) have gained increasing attention since 2022 and are now widely used by scholars for diverse NLP tasks, often in zero-shot or few-shot settings (Gilardi et al., 2023; Törnberg, 2023).

Despite this progress, systematic comparisons between fine-tuned pretrained models and generative LLMs for MWEI remain limited, particularly for languages other than English. This paper addresses this gap by comparing a fine-tuned `CamemBERT` model (Martin et al., 2019) with two open-source LLMs, `gpt-oss-20b` (OpenAI, 2025) and `Qwen3-32B-AWQ` (QwenTeam, 2025), providing a focused evaluation in the context of automatic identification of French MWEs.

2 Related Work

Since the introduction of transformers technology in 2017, pretrained encoder-only models like BERT have proven to be powerful and versatile tools that found their applications for MWEI as well. Fine-tuned on gold-standard datasets, they show excellent performance, achieving state-of-the art results (Gombert and Bartsch, 2020; Premasiri and Ranasinghe, 2022). More recently, autoregressive LLMs like GPT or Qwen also started to be used for non-generative tasks, and researchers explored their capability to detect MWEs (Hashiloni et al., 2025; Ide et al., 2025).

However, many endeavors in this field are only tailored to detect MWEs in English (e.g., Schneider et al. (2016)), which has much more training data available in comparison to other languages. Approaches developed as part of PARSEME shared tasks (Savary et al., 2017; Ramisch et al., 2020) or using PARSEME data (Savary et al., 2023) do foster multi-linguality and often consider French MWEs (Bui and Savary, 2024). But PARSEME

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 54–60

data released before 2025 only focused on verbal MWEs, limiting considerably the scope of extraction.

For this reason, the recent publication of the new PARSEME 2.0 corpora,[1] which extend MWE annotation to all grammatical categories, is an important milestone in the MWEI task. Using the French subcorpus of this dataset, we aim to address the existing research gap in the analysis of performance of masked models and LLMs in identifying French MWEs of all types.

3 Data and Models

3.1 Data

We use the updated French subset of the newly released data for PARSEME 2.0 shared task. In comparison to the previous releases, which covered only verbal MWEs, this dataset includes MWEs of all structural types. Since annotated test data was not available at the time this study was carried out, we use the `train` split for fine-tuning and extraction of few-shot examples and subsequently test the models on the `dev` split. Table 1 below summarizes some statistics about the data, and Table 6 provided in Appendix B gives an overview of MWE classes represented in the dataset.

Split	Sentences	Tokens	MWEs	MWE Classes
train	3,357	80,559	4,604	16
dev	373	9,353	531	13

Table 1: Data overview.

3.2 Models

We use the large version of `CamemBERT`[2] (Martin et al., 2019) as our main model for MWEI task. `CamemBERT` is based on `RoBERTa` and has been specifically trained on French data, making it ideal for working with French MWEs. We further fine-tune it for token classification using the `train` split of the French PARSEME dataset. PARSEME annotations are converted to BIO format, preserving the original tokenization, including split French contractions. Table 3 illustrates how BIO annotation was implemented. Fine-tuning is performed for three epochs using the `transformers`[3] Python library, with the best hyperparameters selected via grid search: a learning rate of **5e-05** and a batch size of **16** (see Table 2).

LR	BS	Precision	Recall	F1
2e-05	16	.76	.79	.78
2e-05	32	.69	.73	.71
3e-05	16	.78	.83	.80
3e-05	32	.74	.78	.76
5e-05	16	.79	.82	.81
5e-05	32	.78	.82	.80

Table 2: Grid search results.

Additionally, we use two open-source LLMs for prompt-based MWEI: `gpt-oss-20b`[4] (OpenAI, 2025) and `Qwen3-32B-AWQ` (QwenTeam, 2025),[5] which we will refer to as `gpt-oss` and `Qwen` for brevity. Both models have strong multilingual capabilities, including in French. For instance, `Qwen` shows similar performance on MMLU (Hendrycks et al., 2021) and MMMLU[6] benchmarks, which test models' multitask language understanding in English (MMLU) and other languages (MMMLU). On MMMLU, `Qwen` outperforms other models of similar size (QwenTeam, 2025). While `gpt-oss` performs slightly worse on MMMLU in comparison to MMLU, it achieves one of the highest scores in French among languages other than English (OpenAI, 2025).

We prompt both models via the `transformers` library using the text generation pipeline and setting the temperature to 0.2 for more deterministic responses. `Qwen` was prompted both with and without reasoning, while for `gpt-oss` the reasoning was set to medium. The prompt consisted of a system message and a user message. The system message described the models' role, defined the MWEI task, and provided examples (at least three per MWE type). The user message included the instruction to identify MWEs in the provided target sentence. We tested two versions of the prompt: in the first one, the models were required to return MWEs as pipe-separated strings, while in the second one the models were prompted to annotate each sentence token following the BIO scheme. For the final experiment, we chose the former because the latter resulted in an extremely high proportion of invalid responses. The full final prompt is provided in Appendix A.

[1] https://gitlab.com/parseme/sharedtask-data/-/tree/master/2.0/subtask1?ref_type=heads

[2] https://huggingface.co/almanach/camembert-large

[3] https://pypi.org/project/transformers/

[4] https://huggingface.co/openai/gpt-oss-20b

[5] https://huggingface.co/Qwen/Qwen3-32B-AWQ

[6] https://huggingface.co/datasets/openai/MMMLU/viewer/default

Tokens	À	le	delà	de	cette	querelle	,	l'	affaire	est	triste	.
PARSEME	1:AdvID	1	1	*	*	*	*	*	*	*	*	*
BIO	B	I	I	O	O	O	O	O	O	O	O	O

Table 3: Example of BIO labels compared to PARSEME labels. Original tokenization is preserved. English translation of the example: *Beyond this quarrel, the matter is sad.*

4 Results

4.1 MWE Identification

We evaluate the models on MWE level with a conservative procedure (i.e. full match with true MWEs required), considering both macro and category-specific performance. As shown in Table 4, `CamemBERT` outperforms by large both LLMs, achieving an F1-score of 0.74. Open AI's `gpt-oss` ranks higher than `Qwen` in terms of precision, however achieves considerably lower recall. This may be due to a high proportion of invalid responses delivered by this model. In 128 cases out of 373, it either produced endless repetitions or generated irrelevant text output. For comparison, the proportion of invalid responses by `Qwen` is notably lower: nine in the run without reasoning and 36 in the run with reasoning. As expected, enabling the reasoning for `Qwen` improves the performance, albeit only in terms of precision.

Model	**Precision**	**Recall**	**F1**
CamemBERT	.73	.75	.74
gpt-oss	.29	.21	.24
Qwen (no reasoning)	.13	.36	.19
Qwen (with reasoning)	.21	.36	.27

Table 4: Model performance on MWE level (macro).

For the more fine-grained analysis across MWE classes, we grouped subcategories of the same class into a single category. For example, light-verb constructions with semantically bleached verbs (**LVC.full**) and light-verb constructions with causative meaning (**LVC.cause**) were both aggregated under the category of light-verb constructions. Since the task focused on identifying MWEs without assigning specific class labels, we report only recall, computed based on the labels provided in the dataset.

As shown in Figure 1, `CamemBERT` achieves its highest performance (over 0.8) for adpositional MWEs (**AdpID**), complex conjunctions (**ConjID**), deverbal nominal MWEs (**NV**), and pronominal MWEs (**PronID**). Performance decreases notably for light-verb constructions (**LVC**) and verbal idioms (**VID**), dropping to 0.62 and 0.56, respectively.

`Qwen` with reasoning enabled performs comparably to `CamemBERT` in identifying **ConjIDs** and **NVs**. Interestingly, the version with reasoning disabled outperforms the reasoning-enabled variant for all verbal MWEs, as well as for **AdpIDs** and **PronIDs**. In general, `Qwen` achieves better results than `gpt-oss` in all but one MWE category.

4.2 Qualitative Error Analysis

In this section, we conduct an error analysis, focusing on false positives, i.e. cases where word sequences identified as MWEs by the models are not considered as such in the dataset. For `Qwen`, we consider only the results produced with reasoning enabled.

When considering the false positives common to `gpt-oss` and `Qwen`, a total of 143 cases is identified. Of these, 70 token sequences are partial matches, usually containing more tokens than those specified in the PARSEME annotations (e.g., *faire confiance à* (to trust + `adposition`) instead of *faire confiance* (to trust)). In a few cases, both models identified single tokens as MWEs (e.g., *Ajaccio* (Ajaccio), *contenu* (content)), as well as single tokens followed by punctuation (e.g., *comment?* (how?)).

By contrast, `CamemBERT` false positives, when partially matching PARSEME MWEs, tend to contain fewer tokens than required by the annotation (e.g., *sur la table* (on the table) instead of *être sur la table* (to be on the table)).

The FPs from `gpt-oss` and `Qwen` which are not partial matches mainly fall into a small number of recurrent categories (see Table 5). Most of them are noun phrases (NPs), including fixed expressions, domain-specific terminology, institutional names, and proper nouns that do not correspond to PARSEME MWEs. Verb phrases (VPs) and prepositional phrases (PPs) form another important group, typically involving frequent and formulaic constructions. Less frequent false positives include adverbial phrases (AdvPs), conjunctive phrases (ConjPs), and discourse expressions, as well as a small set of miscellaneous cases related to tok-

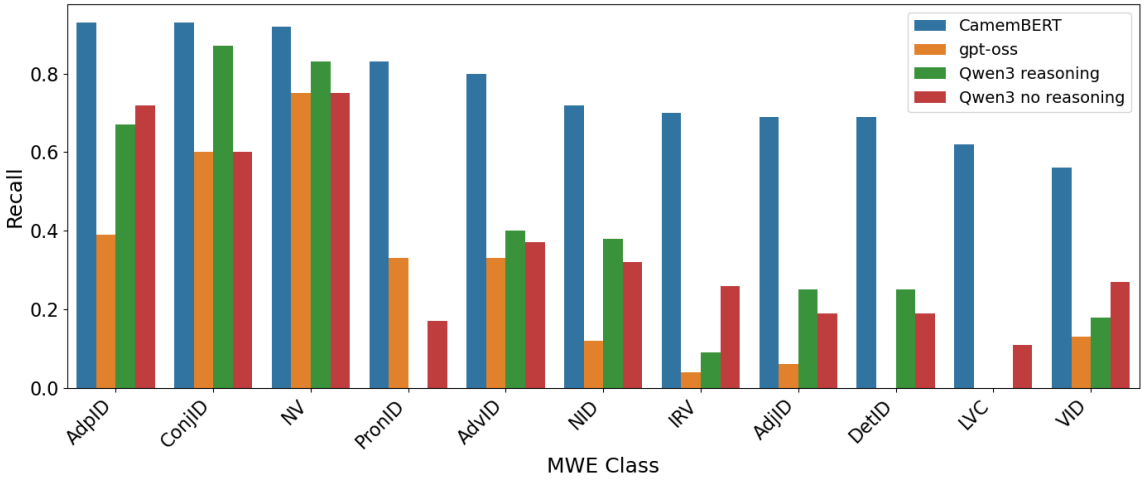


Figure 1: Model recall performance across MWE classes.

FP Class	Examples
NP	*prison ferme*, *durée de conservation*, *Royaume-Uni*, *Conseil supérieur de la magistrature*
VP	*faire aboutir*, *se focaliser sur*, *nous contenter de*, *se concentrer sur*
PP	*sur le fondement de*, *sur le thème de*, *non loin de*, *en faveur de*
AdvP	*une nouvelle fois*, *dans son ensemble*, *fort heureusement*, *mercredi soir*
ConjP	*c'est pourquoi*, *ni l'un ni l'autre*
DiscE	*voir aussi*, *chers collègues*, *monsieur le président*
Other	*de le*, *eu/1/04/289/001*, *comment ?*

Table 5: Main categories of false positives produced by GPT and Qwen, with representative examples for each class.

enization or metadata.

When restricting the analysis to `CamemBERT` false positives that are not partial matches, most errors correspond to isolated lexical units or complete but non-idiomatic phrases. A large share consists of noun phrases, including abstract nouns, technical terms, and institutional or domain-specific expressions (e.g., *secteur économique* (economic sector), *aide sociale* (social assistance), *gouvernement fantôme* (shadow government)). Another common type consists of prepositional or adverbial phrases that are syntactically well-formed collocations but are not considered MWEs under the PARSEME framework (e.g., *en fuite* (on the run), *dans les délais* (on time / within the deadline), *non loin de* (not far from)). `CamemBERT` also predicts single verbs or function words and truncated elements that do not constitute MWEs (e.g., *laisser* (to let), *parlez* (speak / talk [imperative or plural]) , *il* (he/it), *de* (of/from), *s'* (oneself / reflexive clitic)).

5 Discussion

Our findings confirm the initial expectation that fine-tuned encoder-only language models outperform pretrained LLMs on the MWEI task. Notably, even fine-tuning LLMs or employing complex prompting setups, such as chain-of-thought reasoning, are unlikely to achieve performance comparable to that of encoder-only models (see, for example, the scores reported by Hashiloni et al. (2025) on the CoAM dataset, which is similar to PARSEME corpus). Besides, it is worth noting that our fine-tuning approach was rather straightforward; more sophisticated strategies could potentially result in even higher performance (cf. Bui and Savary (2024)).

As shown in the previous section, all models are better in identifying **AdpID**, **ConjID**, and **NVs**. The first two of these classes are arguably the most highly grammaticalized MWEs, which likely facilitates recognition by models. In contrast, the lowest performance is observed for verbal MWEs (**VID** and **LVC**), which show greater variability and are frequently realized in discontinuous forms, posing a substantial challenge for MWEI systems (cf. Bui and Savary (2024); Constant et al. (2017)).

The error analysis of false positives reveals that, due to the large number of partial matches, one of the main challenges lies in correctly defining MWE boundaries. Moreover, the presence of discontinu-

ous forms may also explain the substantial number of false positives consisting of single tokens.

6 Conclusion and Future Work

In this study, we compared the performance of fine-tuned `CamemBERT` with two LLMs (`gpt-oss-20b` and `Qwen3-32B-AWQ`) in the task of automatic identification of French MWEs. Our results show that the fine-tuned model outperforms both LLMs by a large margin, although `Qwen` shows comparable performance in the identification of some MWE types. Our error analysis reveals that `CamemBERT` tends to underpredict MWE tokens, while both LLMs generally return longer MWE sequences, often including function words in MWEs. Future work will incorporate multi-lingual masked models and a wider selection of LLMs, including LLMs predominantly trained on French data. Additionally, we will explore if using more fine-grained label schemes can improve model performance.

Limitations

This study is subject to some limitations that should be taken into account. First, our evaluations were based on the `dev` data split due to the unavailability of annotated test data. Second, this study used a conservative evaluation strategy requiring a full match between a true MWE and model predictions, which penalises model performance. Nevertheless, while a token-based evaluation will yield higher scores, the margin is likely to remain in favour of `CamemBERT`. Third, LLMs were prompted only in a few-shot setting. Testing other prompting strategies would give a more complete understanding of LLM capabilities for the identification of French MWEs.

Acknowledgments

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

Kamyar Arzideh, Henning Schäfer, Héctor Allende-Cid, Giulia Baldini, Thomas Hilser, Ahmad Idrissi-Yaghir, Katharina Laue, Nilesh Chakraborty, Niclas Doll, Dario Antweiler, Katrin Klug, and Niklas Beck et al. 2025. From BERT to Generative AI - Comparing Encoder-only vs. Large Language Models in a Cohort of Lung Cancer Patients for Named Entity Recognition in Unstructured Medical Reports. *Computers in Biology and Medicine*, 195:110665.

Abayomi Bello, Sin-Chun Ng, and Man-Fai Leung. 2023. A BERT Framework to Sentiment Analysis of Tweets. *Sensors*, 23(1).

Martin Bucher and Marco Martini. 2024. Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification. *ArXiv*, abs/2406.08660.

Van-Tuan Bui and Agata Savary. 2024. Cross-type French Multiword Expression Identification with Pre-trained Masked Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4198–4204, Torino, Italia. ELRA and ICCL.

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

Eduardo C. Garrido-Merchan, Roberto Gozalo-Brizuela, and Santiago Gonzalez-Carvajal. 2023. Comparing BERT Against Traditional Machine Learning Models in Text Classification. *Journal of Computational and Cognitive Engineering*, 2(4):352–356.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd Workers for Text-annotation Tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Sebastian Gombert and Sabine Bartsch. 2020. MultiVitaminBooster at PARSEME Shared Task 2020: Combining window- and dependency-based features with multilingual contextualised word embeddings for VMWE detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 149–155, online. Association for Computational Linguistics.

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 23782–23801, Suzhou, China. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Yusuke Ide, Joshua Tanner, Adam Nohejl, Jacob Hoffman, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. 2025. CoAM: Corpus of All-Type Multiword Expressions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27004–27021, Vienna, Austria. Association for Computational Linguistics.

Kai Labusch, Preußischer Kulturbesitz, Clemens Neudecker, and David Zellhöfer. 2019. BERT for

Named Entity Recognition in Contemporary and Historical German. In *Proceedings of the 15th conference on natural language processing*, pages 9–11.

Louis Martin, Benjamin Muller, Pedro Ortiz Suarez, Yoann Dupont, Laurent Romary, Eric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. CamemBERT: a Tasty French Language Model. In *Annual Meeting of the Association for Computational Linguistics*.

Juan G. Diaz Ochoa, Natalie Layer, Jonas Mahr, Faizan E. Mustafa, Christian U. Menzel, Martina Müller, Tobias Schilling, Gerald Illerhaus, Markus Knott, and Alexander Krohn. 2025. Optimized BERT-based NLP Outperforms Zero-shot Methods for Automated Symptom Detection in Clinical Practice. *Frontiers in Digital Health*, 7:1623922.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *Preprint*, arXiv:2508.10925.

Damith Premasiri and Tharindu Ranasinghe. 2022. BERT(s) to Detect Multiword Expressions. *Preprint*, arXiv:2208.07832.

QwenTeam. 2025. Qwen3 Technical Report. *Preprint*, arXiv:2505.09388.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, and Tunga et al. Güngör. 2020. Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, and Polona et al. Gantar. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 Task 10: Detecting Minimal Semantic Units and their Meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559, San Diego, California. Association for Computational Linguistics.

Miklós Sebők, Viktor Kovács, Martin Bánóczy, Daniel Møller Eriksen, Nathalie Neptune, and Philippe Roussille. 2025. Beyond token limits: Assessing language model performance on long text classification. *Preprint*, arXiv:2509.10199.

Anna Siyanova-Chanturia. 2013. Eye-tracking and ERPs in multi-word expression research: A state-of-the-art review of the method and findings. *The Mental Lexicon*, 8(2):245–268.

Anna Siyanova-Chanturia, Kathy Conklin, Sendy Caffarra, Edith Kaan, and Walter JB van Heuven. 2017. Representation and processing of multi-word expressions in the brain. *Brain and language*, 175:111–122.

Petter Törnberg. 2023. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. *Preprint*, arXiv:2304.06588.

A LLM Prompt

System message:

#Role

You are a professional annotator with strong linguistic background. Your area of specialization is multi-word expressions (MWEs).

#Task

MWEs are sequencies of at least two words that show some degree of orthographic, morphological, syntactic and/or semantic idiosyncrasy with respect to what is considered general grammar rules of a language. Your task will be to identify MWEs in French texts.

#Input

You will be provided with a sentence in French that may or may not contain MWEs.

#Instructions

Extract MWEs detected in the sentence if any. If more than one MWE is detected, separate the MWEs with a pipe (|). MWEs can be discontinous, i.e. contain optional tokens that are not part of an MWE. Do not extract such tokens.

#Output

Return the extracted MWEs. If no MWEs detected, return the original sentence. Do not include any additional text or comments.

#Examples

FEW-SHOT EXAMPLES

User message:

Identify MWEs in the following sentence: TARGET SENTENCE

B MWE Classes

MWE Class	Description	Examples
LVC.cause	Causative light-verb constructions	*susciter intérêt, entraîner réduction, apporter cohérence*
LVC.full	Light-verb constructions with bleached verbs	*avoir droit, faire appel, donner instructions*
VID	Verbal idioms	*venir en aide, donner un blanc-seing, avoir lieu*
IRV	Inherently reflexive verbs	*s'apparenter, se trouver, se rendre*
MVC	Multi-verb constructions	*faire valoir, laisser faire, entendre parler*
NID	Nominal idioms	*homme d'affaires, chaîne de radio, droit de vote*
NV.VID	Deverbal nominal MWEs stemming from a VID	*mise au point, retour à la normale, entrée en vigueur*
NV.MVC	Deverbal nominal MWEs stemming from an MVC	*savoir-faire*
NV.LVC.full	Deverbal nominal MWEs stemming from an LVC.full	*ayant droit, prises de position*
AdjID	Adjectival idioms	*à part entière, en désaccord, de premier plan*
AdvID	Adverbial idioms	*en même temps, en principe, c'est-à-dire*
AdpID	Adposition idioms	*grâce à, lors de, aux termes de*
ConjID	Conjunction idioms	*parce que, à moins que, ou bien*
DetID	Determiner idioms	*plein de, majorité de, un peu de*
IntjID	Interjection idioms	*eh bien, bonne chance*
PronID	Pronominal idioms	*elle-même, tout le monde, quelque chose*

Table 6: MWE classes in French PARSEME data.

Extracting Multi-Word Expressions Representing Technical Terms and Proper Nouns in Log Messages

Kilian Dangendorf, Sven-Ove Hänsel, Jannik Rosendahl,
Felix Heine, Carsten Kleiner, Christian Wartena
Institute for Applied Data Science Hannover (Data|H)
Hochschule Hannover
{kilian.dangendorf, sven-ove.haensel, jannik.rosendahl,
felix.heine, carsten.kleiner, christian.wartena}@hs-hannover.de

Abstract

IT-systems generate log messages containing important information about the system's health. To gather information about system entities, we extract multi-word expressions (MWEs) representing technical terms and proper nouns from a wide range of log messages from 16 different real systems. We apply Gries' information-theoretic approach which iteratively calculates the best MWE candidates using an eight-dimensional ranking method. These candidates are evaluated in an annotation study, achieving a precision of 66 %. This value is significantly higher than evaluations on general-purpose texts, demonstrating the higher occurrence of compound technical terms and proper nouns in log messages. The MWEs found can be used to reduce the number of nodes in a system behavior graph while increasing the information density of the nodes.

1 Introduction

Every computer system generates log messages. Analyzing and interpreting these messages allows conclusions to be drawn about the health of the system and possible cyberattacks. Log messages are intended to be read by developers who are familiar with the computer domain and therefore contain names of system parts, numbers, dates, memory addresses, etc. Logs are often short incomplete sentences (e.g. in most cases the subject is missing) with few or no verbs but many technical terms (see Listing 1 for examples). Thus, multi-word expressions (MWEs) make up a larger proportion than in general English texts.

This work forms the basis for our subsequent research, in which we construct a behavior graph from log messages. In the first modeling step, each token represents a node in the graph, resulting in large graphs. The more nodes we can eliminate from this point onwards while retaining all relevant information, the more efficiently the downstream graph neural network (GNN) can work, since there will be fewer nodes and higher information density.

Established methods for finding MWEs can help, as several tokens can be merged into one that is more descriptive. Technical terms and proper nouns in log messages are valuable because they describe or reference an entity in the system. This allows us to reduce the number of tokens as well as specify the systems entities more precisely.

The concepts of log parsing and current work in the applications of NLP on log messages as well as MWE extraction methods are described in Section 2. Section 3 shows how we process log messages and extract MWEs as technical terms and proper nouns. The results of our annotation study are presented in Section 4. Finally, we conclude our findings in Section 5.

2 Related Work

Listing 1: Example log lines from the Linux dataset containing month, date, time, log level, logging component and the log message (bold).

```
Oct 23 12:40:02 combo cups-lpd[22514]:
 ↪ Unable to get command line from client!
Oct 25 10:08:47 combo kernel:
 ↪ Inode-cache cache hash table entries: 16384 (order: 4,
 ↪ 65536 bytes)
Nov 22 14:31:58 combo kernel:
 ↪ httpd: page allocation failure. order:0, mode:0x1d2
Nov 22 14:32:24 combo kernel:
 ↪ sendmail: page allocation failure. order:0, mode:0x1d2
Dec  6 12:22:57 combo kernel:
 ↪ PID hash table entries: 512 (order: 9, 4096 bytes)
```

2.1 Log Parsing

System logs are notoriously unstructured and often include metadata such as timestamps, log levels, or logging components. Parsing these raw logs into their building blocks, metadata, and log templates by identifying static and variable parts is a common method. Zhu et al. (2019) provide an open-source toolkit to compare the accuracy of 13 different log parsers on a benchmark dataset, which was later published independently as *Loghub* (Zhu

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 61–65

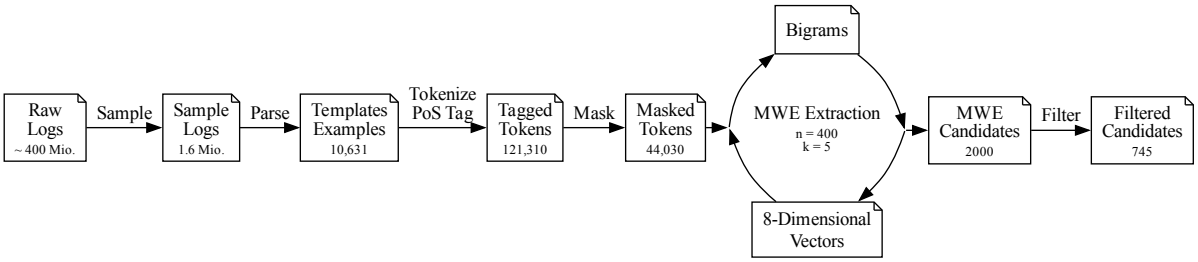


Figure 1: Our automated data processing pipeline applying Gries' MWE extraction method.

et al., 2023). The aim is to correctly assign each log message to the corresponding template.

Listing 1 shows some example lines of the Linux dataset with metadata, the message itself is highlighted. Splitting metadata and messages from Listing 1 and parsing them leads to the templates shown in Listing 2, where each variable part was replaced by <*>. Large amounts of log messages can be reduced to a small number of templates. E.g., the Linux dataset consists of 25,000 log messages, which result in 488 templates (Zhu et al., 2019).

Listing 2: Templates resulting from Listing 1. Underlined parts indicate the desired MWEs.

```
Unable to get command line from client!
<*> hash table entries: <*> (order <*>: <*> bytes)
<*>: page allocation failure. order:<*>, mode:<*>
```

2.2 NLP on Log Messages

NLP techniques are increasingly being used in log parsing. *PosParser* (Jiang et al., 2024) uses PoS tags to identify variable and static parts of log messages and thus infer the corresponding template. The sequence of all verbs and the first noun play a key role here. A similar approach is that of *NLog* (Pi et al., 2019) in which all nouns are selected as key candidates and filtered by frequency in the dataset. The tool *NERLogParser* (Studiawan et al., 2018) uses deep learning techniques to identify and extract named entities from log messages. Shikha and Timalsina (2022) use a similar deep learning based named entity recognition (NER) method to find templates in log messages.

Lv et al. (2021) use PoS tags to remove *unnecessary* words, and apply a multi-layer LSTM to the embeddings of the remaining words to detect anomalies. Li et al. (2022) use PoS tagging and NER to generate a template vector again using Word2vec that is used as input to a deep neural network to find anomalies.

2.3 MWE Extraction

EXTERLOG is a tool for terminology extraction from log files (Saneifar et al., 2009a,b). Co-occurrences are extracted using PoS tags and the resulting syntactic patterns. Pointwise mutual information (PMI) and the Dice coefficient are used as ranking metrics. Gries (2022) uses 8 dimensions, representing different, most information theoretic properties of MWEs, in an iterative apporach.

Ben Youssef (2024) evaluated resulting MWE candidates and reported a precision of 48% on English texts. Bagdasarov and Teich (2025) demonstrate that Gries' method can be transferred to German and report a precision of 61 %.

3 From Logs to MWE Candidates

We use the 16 datasets from Loghub. They represent a wide range of software systems, such as distributed systems, supercomputers, operating systems, mobile systems, server applications, and standalone software. In total, these datasets contain over 400 million log lines.

We sampled 100,000 random lines from each of the 16 datasets resulting in a total of 1.6 million log messages. To achieve a sample that is diverse and does not have many repetitions of almost identical log lines, we used the established log parser Drain (He et al., 2017) to extract log templates. For each unique template we select one random example log message. This prevents frequent templates from being given excessive importance. After sampling we have 10,631 log messages consisting of 121,310 tokens (see Figure 1).

To target technical terms, we pre-filter relevant word types. Saneifar et al. (2009b) achieved better results when they pre-filtered according to syntactic patterns of PoS. With POSLOG (Dangendorf et al., 2025), we have a tool that was developed specifically for tagging log messages[1]. POSLOG tags in Universal PoS-tags with 17 tags (Petrov et al., 2012; Nivre et al., 2016).

For technical terms and proper nouns, nouns, proper nouns, adjectives and adverbs are of special

[1] POSLOG also comes with a tokenizer that specializes in log messages, which we have also used here.

interest[2]. We mask tokens of other PoS so as not to destroy the sentence structure. When forming bigrams, masked tokens are skipped. For the MWE extraction, 44,030 tokens remain unchanged.

3.1 Identifying MWE Candidates

To extract the MWEs, we use the method of Gries (2022). This involves ranking all bigrams in several iterations, with the best candidates being merged into one token before the next iteration starts. This way, even long MWEs are found if they achieve a high ranking. For the ranking, an 8-dimensional vector is generated for each bigram ab:

Frequency (1) is the overall occurrence of the bigram, logarithmically dampened.

Dispersion (2) states how much the distribution of the bigram across the 16 datasets deviates from the distribution of the proportions of the respective datasets using Kullback-Leibler divergence (KLD).

Type Frequency (3 & 4) counts how many different types follow a and precede b, dampened logarithmically.

Entropy (5 & 6) calculates the normalized entropy after a and before b. Similar to dimensions 3 and 4, but here the frequency distribution of the following/preceding types is taken into account.

Association (7 & 8) gives the degree to which a attracts b and b attracts a (calculated by KLD).

If not already normalized, these dimension values are min/max normalized to the interval [0;1] for each iteration n. The dimensions are designed in such a way that good MWE candidates generate high values. In each iteration, the Euclidean distance to the origin of the hypercube is calculated. The k MWE candidates with the highest distance are selected as MWEs from the iteration and are merged. For our experiment, we choose n=400 iterations and k=5 and we get 2,000 MWE candidates.

Referring to the initial example in Listing 2, *hash table* was found in iteration 22 and *hash table entries* in iteration 47, *page allocation* in iteration 64 and *page allocation failure* in iteration 77, and *command line* in iteration 159.

3.2 Post-Filter

We dropped MWE candidates based on three rules: First, we filter out candidates whose tokens contain punctuation characters like URLs, paths and other tokens that are atypical for words (e.g., `ccfile::copyfile`, `krbtgt/#24#@#24#`). Second, we filter out words with more than 15 characters[3] (e.g., `ksserverupdaterequestdelegate`). Third, we exclude candidates whose last token does not have the PoS tag `NOUN` or `PROPN` (e.g., *too many*, *so far*). After filtering, we have 745 candidates.

[2]A technical term built from adjective and noun would be *public key* for example.

Table 1: Proportions of successful and unsuccessful MWE candidates with majority vote.

MWE Candidates	Count	Rel.
Unsuccessful	253	34 %
Successful	492	66 %
Total	745	100 %

4 Results

Three computer scientists annotated these 745 candidates. They annotated the extracted candidates as being perceived as a technical term in the domain of the dataset or not being a typical term in that domain. The annotators agreed on 486 candidates: 346 accepted and 140 rejected. This results in a Fleiss' kappa for all annotators of 0.49, which is in the middle range of moderate agreement. The deviation in the annotation can be explained by the fact that log messages are application-specific, and each annotator has a different background of experience. Additionally some phrases can be interpreted in multiple ways. The most common reason for agreement deviation were noun phrases consisting of adjective and noun. For example, *public key*, *floating point*, and *local host* are established terms in computer science and definitely belong together. Uncertain candidates include, for example *real time*, *excessive wakeups*, *read-only filesystem*, and *remote host*, where the annotators disagreed.

In the following we use the majority vote as gold standard. For 66 % of the extracted MWEs (see Table 1) the majority perceived the candidate as a real technical term or proper noun, which proves the effectiveness of the extraction method.

4.1 Classification

The annotators' second task was to classify the successful candidates, distinguishing between technical terms and proper nouns. The results can be found in Table 2. A total of 424 technical terms and

[3]An analysis of over one million English words found 136 words longer or equal to 15 characters. We therefore choose a limit of 15 characters to exclude long tokens that are unlikely to be words. Archived link: https://web.archive.org/web/20090427054251/http://www.maltron.com/words/words-longest-modern.html

30 proper nouns were identified. The proportion of proper nouns may seem small at first glance. However, we are dealing with MWEs, so single-token proper nouns such as *Linux* or *Google* are not included. The extracted proper nouns can be divided into three subclasses: We found *red hat linux* in the Thunderbird dataset as an example for a *company*, *dave jones* in the Linux dataset as an example for a *person*, and *internet systems consortium* in the Thunderbird dataset as an example for a *group*.

Table 2: Classification results on successful MWE candidates with majority vote. Parity arises when two annotators voted for successful, but different classes.

Class	Count	Rel.
Technical Terms	424	86 %
Proper Nouns	30	6 %
Majority Vote	454	93 %
Parity	38	8 %
Total	492	100 %

4.2 Most Widely Spread Examples

The term *file descriptor* appears scattered across most datasets: five times in four datasets in total. Spread across three datasets each there are the following exemplary terms: *configuration file* appears seven times, *network connection* six times, *exit status* also six times, and *lock file* three times.

The low number of the same technical terms across multiple datasets indicates that the language used in the respective systems is largely independent domain-specific terminology.

4.3 Occurrence in the Datasets

We found at least one MWE in every dataset sample. In total, we found 454 technical terms and proper nouns in 1,513 of the 10,631 (14.2 %) messages.

Combining these MWE tokens into one token saves 1,859 of the 121,310 total tokens. This corresponds to a savings rate of 1.5 %. The distribution across the datasets is shown in Table 3. While the Thunderbird dataset provides about a quarter of the MWEs in this work, only one MWE was found in the Apache dataset.

4.4 Longest Terms

The distribution of the successful MWEs candidates by their count of tokens can be found in Table 4. Almost 80 % of MWEs consist of two tokens, 15.6 % of three tokens, and 2.9 % of four. MWEs consisting of five or more tokens make up the remaining 1.3 %. Examples are *pci hot plug pci core* from Linux and *google software update installer* from Mac. The longest technical terms we found consist of six tokens coming from Thunderbird: *usb universal host controller interface driver*.

Table 3: Token replacement counts per dataset. For example, with the help of the successful MWE candidates, 490 tokens could be reduced in the Thunderbird dataset.

Dataset	Count	Rel.	Dataset	Count	Rel.
Thunderbird	490	26.36 %	OpenSSH	16	0.86 %
Android	361	19.42 %	HPC	13	0.70 %
BGL	342	18.40 %	Zookeeper	10	0.54 %
Mac	305	16.41 %	Proxifier	5	0.27 %
Linux	168	9.04 %	OpenStack	4	0.22 %
Hadoop	70	3.77 %	HDFS	3	0.16 %
Windows	52	2.80 %	HealthApp	2	0.11 %
Spark	17	0.91 %	Apache	1	0.05 %
		Total 1,859	100 %		

Table 4: Distribution of successful MWE candidates by their length in tokens.

Tokens	Count	Rel.
2	362	79.7 %
3	71	15.6 %
4	13	2.9 %
5	5	1.1 %
6	1	0.2 %
Total	454	100.0 %

5 Conclusion

Applying Gries' method on log messages results in a precision of 66 % for automatically finding technical terms and proper nouns, exceeding related work and indicating more MWEs in log messages.

Limiting the results to these two classes reduces the overall quality of finding MWEs in this evaluation. Without pre- and post-filtering this method also finds useful MWEs such as *no such file or directory*, *too many open files*, or *file not found*.

As only very few MWEs appear in more than three datasets, we conclude that logs use highly domain-specific terminology. In other words, it is to be expected that only a few of the MWEs extracted here will appear in a new dataset.

Through multiple iterations, MWEs consisting of up to six tokens were found. By combining all successful MWEs candidates in the entire data selection, we can reduce the number of tokens by 1.5 %. It should be noted that we have sampled the data by templates, so this value may vary when counted across all messages. However, this takes us a step closer to our goal of reducing the number of tokens and specifying the system entities more precisely.

References

Sergei Bagdasarov and Elke Teich. 2025. Applying an information-theoretic approach for automatic identification of German multi-word expressions. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, pages 295–305, Hannover, Germany. HsH Applied Academics.

Chadi Ben Youssef. 2024. *mMERGE: a corpus driven Multiword Expressions discovery algorithm*. Ph.D. thesis, UC Santa Barbara. ProQuest ID: BenYoussef_ucsb_0035D_16720. Merritt ID: ark:/13030/m5457324. Retrieved from https://escholarship.org/uc/item/4ph4517b.

Kilian Dangendorf, Sven-Ove Hänsel, Jannik Rosendahl, Felix Heine, Carsten Kleiner, and Christian Wartena. 2025. PosLog: Creating a Part of Speech Tagger for Log Messages. In *2025 IEEE 13th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, pages 1444–1449.

Stefan Th Gries. 2022. Multi-word units (and tokenization more generally): a multi-dimensional and largely information-theoretic approach. *Lexis*, (19).

Pinjia He, Jieming Zhu, Zibin Zheng, and Michael R. Lyu. 2017. Drain: An Online Log Parsing Approach with Fixed Depth Tree. In *2017 IEEE International Conference on Web Services (ICWS)*, pages 33–40.

Jinzhao Jiang, Yuanyuan Fu, and Jian Xu. 2024. Posparser: A heuristic online log parsing method based on part-of-speech tagging. *IEEE Transactions on Big Data*, pages 1–12.

Zezhou Li, Jing Zhang, Xianbo Zhang, Feng Lin, Chao Wang, and Xingye Cai. 2022. Natural language processing-based model for log anomaly detection. In *2022 IEEE 2nd International Conference on Software Engineering and Artificial Intelligence (SEAI)*, pages 129–134.

Dan Lv, Nurbol Luktarhan, and Yiyong Chen. 2021. Conanomaly: Content-based anomaly detection for system logs. *Sensors*, 21(18).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Aidi Pi, Wei Chen, Will Zeller, and Xiaobo Zhou. 2019. It can understand the logs, literally. In *2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 446–451.

Hassan Saneifar, Stéphane Bonniol, Anne Laurent, Pascal Poncelet, and Mathieu Roche. 2009a. Mining for relevant terms from log files. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, pages 77–84. SciTePress - Science and and Technology Publications.

Hassan Saneifar, Stéphane Bonniol, Anne Laurent, Pascal Poncelet, and Mathieu Roche. 2009b. Terminology extraction from log files. In *Database and Expert Systems Applications*, pages 769–776, Berlin, Heidelberg. Springer Berlin Heidelberg.

Saloni Shikha and Arun Kumar Timalsina. 2022. Automated log parsing through named entity recognition. In *Proc. of the 12th IOE Graduate Conference*, volume 12, pages 1747–1753.

Hudan Studiawan, Ferdous Sohel, and Christian Payne. 2018. Automatic log parser to support forensic analysis.

Jieming Zhu, Shilin He, Pinjia He, Jinyang Liu, and Michael R. Lyu. 2023. Loghub: A large collection of system log datasets for ai-driven log analytics. In *IEEE International Symposium on Software Reliability Engineering (ISSRE)*.

Jieming Zhu, Shilin He, Jinyang Liu, Pinjia He, Qi Xie, Zibin Zheng, and Michael R. Lyu. 2019. Tools and benchmarks for automated log parsing. In *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice*, ICSE-SEIP '19, page 121–130. IEEE Press.

Two Birds with One Stone: Annotating Romanian Multiword Expressions with an Eye to the PARSEME 2.0 Guidelines Applicability

Verginica Barbu Mititelu[1], Mihaela Cristescu[2], Elena Irimia[1], Carmen Mîrzea Vasile[2]

[1]Romanian Academy Research Institute for Artificial Intelligence, [2]University of Bucharest
vergi@racai.ro,mihaela.cristescu@litere.unibuc.ro,elena@racai.ro,carmen_marzea@yahoo.fr

Abstract

This paper presents an enhanced version of the Romanian corpus previously annotated only for verbal multiword expressions. The new release extends the annotation to multiword expressions of other parts of speech, following version 2.0 of the PARSEME guidelines. The corpus has been expanded, its new part was automatically morpho-syntactically annotated based on the Universal Dependencies framework, followed by extensive semi-automatic annotation of multiword expressions across all morphological categories. The paper also reports quantitative data on the updated corpus and discusses the distribution and characteristics of Romanian multiword expressions. We also highlight language-specific annotation challenges and issues arising from the PARSEME 2.0 guidelines.

1 Introduction

Multiword expressions (MWEs) are everywhere, yet notoriously slippery to define and analyze. From idioms like *go bananas* to collocations such as *rancid butter* and phrasal verbs like *put up with*, these fixed or semi-fixed combinations of words play a key role in how meaning is packaged and conveyed. Understanding how MWEs function is easier said than done, as their meaning often goes beyond the sum of their parts. However, besides the semantic non-compositionality, MWEs also exhibit idiosyncrasies at other linguistic levels (Baldwin and Kim, 2010): lexical, syntactic, pragmatic and even statistical.

We present below the process of quantitatively and qualitatively enriching the Romanian component of the PARSEME corpus (Savary et al., 2023). Its initial version (Barbu Mititelu et al., 2019) contained annotations of only verbal MWEs, while now MWE of all parts of speech have been annotated. Another contribution of this paper is that of offering feedback regarding the PARSEME 2.0 guidelines[1] that were observed during the annotation.

The paper is structured as follows: Section 2 presents the current work concerning MWEs both within Romanian linguistics and in an international context. In Section 3 statistics of the enriched corpus is given. The types of MWEs annotated in the data are inventoried in Section 4, alongside the challenges their identification raises. We describe our work methodology in Section 5 and make some remarks on the frequency and variety of MWEs in this corpus in Section 6. The issues we had in the application of the decision trees of the PARSEME guidelines 2.0 are presented in Section 7, before concluding the paper.

2 Related Work

Related work concerns, on the one hand, the current situation of research on MWE in the Romanian language and, on the other hand, the larger international background against which our work has been carried out.

The systematic study of MWEs in Romanian linguistics dates back to the 1950s (Ioaniţescu, 1956), with early focus on verbal MWEs (Dimitrescu, 1958). Over time, scholars have used varied terminology and offered differing views on their definition, classification, and structure (Căpățână, 2007). A recent contribution to this field (Pană Dindelegan et al., 2025) provides a detailed overview of Romanian MWEs, covering theoretical issues (concepts, delimitation criteria, graduality, and terminology) and practical aspects (types, analysis methods, and exercises). Aimed primarily at educational and applied purposes, it offers representative inventories and solutions to common difficulties, making it a useful resource for students, teachers, and researchers of Romanian grammar.

[1] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 66–74

In recent years, there has been growing interest in the computational analysis of Romanian MWEs (for a larger context, see (Barbu Mititelu et al., 2025)). A corpus annotated with verbal MWEs is already mentioned above and was developed within PARSEME COST Action[2]. The lexicon of Romanian verbal MWEs (Leseva et al., 2024) provides uniform descriptions of such MWEs at several linguistic levels (lexical, morphologic, syntactic, semantic, stylistic) (and in comparison with a language historically in contact with Romanian, i.e. Bulgarian, and with English). Additionally, the Romanian Reference Treebank (Barbu Mititelu, 2013) has been annotated with multiword conjunctions (Barbu Mititelu and Voicu, 2024) which were assigned Penn Discourse Treebank relations (Webber et al., 2019), facilitating deeper insights into their syntactic and semantic roles.

Besides developing language resources (a corpus and a lexicon), researchers have also been interested in developing systems dedicated to the task of identifying MWEs in Romanian corpora (Boros et al., 2017; Avram et al., 2023).

PARSEME has conducted a series of multilingual annotation campaigns and shared tasks dedicated to MWEs. Corpora for 20+ languages were annotated with verbal MWEs and further used as training, tuning and testing data for systems in three shared tasks for automatic MWE identification in corpora. One factor that made this effort even more valuable was the annotation of MWEs observing common guidelines developed with an eye to universality: a common typology of MWEs was created for and tested on all languages involved, at the same time making space for any particular language specificity, i.e. language specific MWE types were accepted and described.

The PARSEME annotation framework formalized a consistent procedure for identifying MWEs across languages, combining a decision-tree approach with cross-lingual validation and language-specific clarifications (Savary et al., 2018). This approach has been crucial for ensuring both comparability and adaptability across typologically diverse languages. With PARSEME 1.3 (Savary et al., 2023), the multilingual corpus expanded to 26 languages, was aligned with Universal Dependencies v.2, and further enhanced its linguistic coverage, consolidating its position as a major resource for multilingual MWE research.

[2] https://typo.uni-konstanz.de/parseme/

3 The Corpus

The PARSEME-Ro corpus consists exclusively of journalistic texts published between 2003 and 2017, and includes 56,703 sentences totaling 1,015,623 tokens (i.e. syntactic words and punctuation). In this annotation campaign, we expanded the corpus with 14,517 sentences (407,801 tokens), also drawn from journalistic sources, in order to maintain genre homogeneity throughout the corpus. Table 1 shows that the newly added data contains longer sentences and represents about a third of the final corpus.

The number and proportion of MWEs of different parts of speech annotated in the corpus are rendered in Table 2. We can see that functional MWEs are the most frequent in the corpus. It is a closed morphological class but indispensable in rendering logical connection between syntactic units in sentences. The ratio token/MWE is 22 (i.e., there is an average frequency of one MWE per 22 tokens), which, correlated with the average sentence length (20 tokens, see Table 1), means that each sentence contains an average of 1.1 MWEs.

4 MWEs in Romanian

4.1 Types of MWEs occurring in Romanian

The types of MWEs currently represented in the Romanian corpus follow the PARSEME guidelines 2.0 and are as follows:

- Verbal
  - Verbal Idioms (VID): *avea de gând* (‘have of thought’ “intend”), *da viață* (‘bring life’ “bring to life”)
  - Light Verb Constructions (LVC)
    - full: *avea grijă* (‘have care’ “take care”), *da citire* (‘give reading’ “read”)
    - cause: *da asigurare* (‘give assurance’ “assure”), *pune la dispoziție* (‘put at disposal’ “provide”)
  - Reflexive Verbs (IRV): *se gândi* (“think”), *se abține* (“refrain”)
  - Inherently Adpositional Verbs (IAV): *conta pe* (“count on”), *depinde de* (“depend on”)
- Nominal
  - Nominal Idioms (NID): *bani gheață* (‘money ice’ “cash”), *bătaie de cap* (‘beating of head’ “trouble, nuisance”)

	Older data	New data	TOTAL data
sentences	56703	14517	71220
tokens	1015623	407801	1423424
tokens/sentence	18	28	20

Table 1: Statistics of the PARSEME-Ro 2.0.

MWE PoS	#	%
verbal	18084	28
nominal	8752	13
adjective/adverb	14672	23
functional	23093	36
TOTAL	64601	

Table 2: Statistics on MWEs in PARSEME-Ro 2.0.

  - Pronominal Idioms (PronID): *câte ceva* (“something”), *Excelența Sa* (“His Excellency”)
  - Deverbal Nominal MWEs (NV): *aducere aminte* (‘bringing to memory’ “memory, remembrance”), *băgare de seamă* (‘putting of notice’ “attention, observation”)

- Modifier
  - Adjectival Idioms (AdjID): *cât un purice* (‘as big as a flea’ “very small”), *de vină* (‘of guilt’ “guilty”)
  - Adverbial Idioms (AdvID): *de asemenea* (‘of the same’ “also, likewise”), *și așa mai departe* (‘and thus further’ “and so on”)
  - Deverbal adjectival / adverbial MWEs (AV): *cu luare aminte* (‘with taking notice’ “attentively”), *avut în vedere* (‘had in view’ “considered, took into account”)

- Functional
  - Determiner Idioms (DetID): *tot felul de* (“all kinds of”), *ca atare* (“as such”)
  - Adposition Idioms (AdpID): *în legătură cu* (‘in connection with’ “regarding”), *cu excepția* (‘with the exception’ “except for”)
  - Conjunction Idioms (ConjID): *astfel încât* (“so that”), *pentru că* (‘for that’ “because”)
  - Interjection Idioms (IntjID): *așa să fie* (“so be it, amen”), *nici vorbă* (‘no word’ “no way, not a chance”)

4.2 Challenges in MWE Identification in Romanian

As mentioned above, the MWE status is decided based on the battery of tests organized as a decision tree in the PARSEME annotation guidelines. Inherent difficulties are detailed in the annotation guidelines (see Section 3 therein, e.g., the solutions provided for problematic reflexive expressions). Beyond these general challenges, MWE identification in Romanian faces additional difficulties stemming from the language’s specific lexico-grammatical features.

Romanian is a language with rich morphology (Pană Dindelegan, 2013), where some grammatical categories have analytical expression. There are specific word strings that warrant examination at the morphology-lexicon interface. This concerns primarily grammatical categories (e.g., the comparative of superiority) and morpho-lexical (sub)types (e.g., the supine, ordinal numerals, distributive numerals, etc.). Some of these analytic forms may acquire a specific, idiosyncratic meaning distinct from the compositional meaning specific to their paradigm. These sequences with a fixed idiosyncratic meaning have been treated as MWEs (not as converted analytical forms): e.g., the expression of comparison (Mîrzea Vasile, 2012, 32-33) and intensity (*mai ales* “especially”, *mai curând* “rather”, *mai mult* “more”, *mai puțin* “except for”, *cel mai tare* “especially”, etc.).

The supine is a non-finite verb form containing a grammaticalized functional preposition (*de* “of”) and a deverbal abstract noun (Pană Dindelegan, 2013, 233-243). Supine forms with an idiosyncratic meaning were considered MWEs in our annotation (AdvIDs or AdjIDs, e.g., *de împrumut* “borrowed, unfitting”, *de neuitat* “unforgettable”, *de neînchipuit* “unimaginable”), while those in free syntactic configurations were omitted (e.g., *termină de scris* “finishes writing”, *instrument de scris* “writing instrument”, *apă bună de băut* “water good for drinking”, *De băut, am băut.* “As for drinking, I drank.”).

Another characteristic of Romanian is the con-

version (or zero-derivation) of adjectives into adverbs. The few dozen adverbs suffixed with *-ește* in contemporary Romanian are most frequently used in fixed expressions: *a împărți frățește* "to share fraternally", *a fi răsplătit regește* "to be rewarded royally", etc. (Mîrzea Vasile, 2012, 91-128).[3] There are also many compositional adverbials with quite regular structures, which were not considered MWEs; e.g., *în mod* 'in manner' + adjective: *în mod special* "in a special manner", *în mod necinstit* "in a dishonest manner"; *cu/fără* "with/without" + abstract quality noun: *cu prietenie* "with friendship", *cu dragoste* "with love", *fără plăcere* "without pleasure", *fără frică* "without fear". The expressions that we retained are those which passed the PARSEME tests: *Cu plăcere!* "My pleasure!" (adverbial used as IntjID), *fără seamăn* "peerless, incomparably" (AdjID/AdvID containing the cranberry old noun *seamăn* "resemblance", cf. current equivalent *asemănare* "resemblance"), *fără stare* "restless, agitated" (AdjID, in which the noun *stare* has a special meaning, "calm, tranquility"), etc.

In Romanian, there are the variable elements *al* and *cel*, with semi-functional or functional status depending on the context: e.g., *al* can be an obligatory unbound possessive morpheme in contexts of non-adjacency with the definite enclitic article, and can function as a pronoun that cannot appear independently; *cel* is a morpheme of the relative superlative degree, but can also have a status similar to that of a pronoun, etc. (Pană Dindelegan, 2013, 265-267, 309-318). Compositional constructions with these elements were omitted, and those which have developed a special meaning were annotated as MWEs; e.g., *ai mei* "my folks" (but not *ai mei* "mine" from: *Pantofii tăi sunt curați, ai mei nu sunt.* "Your shoes are clean, mine are not."), *Cel de Sus* (please notice the capitalized words) "God" (but not *cel de sus* "the above one": *El culege mărul de jos, nu pe cel de sus.* "He picks the apple from below, not the one from above").

5 Work Methodology

Our objective in this annotation campaign was to automate part of the workflow to optimize the overall process. The main motivations for introducing automation were the significantly larger number of non-verbal MWEs targeted for annotation and the inherent redundancy of functional MWEs, which makes them particularly amenable to automatic processing. All automatically generated annotations were manually validated and, when necessary, corrected, given that, as previously noted, the accurate identification and labeling of MWEs remains a challenging task. The following subsections outline the steps undertaken in the annotation process of the PARSEME-Ro corpus in its 2.0 version.

5.1 Automatic Retrieval of MWEs from Dictionaries

The automatic annotation approach was a resource-based one, involving Romanian idioms and expressions dictionaries. In selecting these linguistic resources, we restricted our focus to dictionaries that were already digitised, available online or in standard digital formats (e.g., .DOCX, .XLS, .PDF), enabling automatic processing and eliminating the need for manual scanning and subsequent OCR processing.

Using dedicated scripts, information was extracted from PDF files in the case of three dictionaries: DELS (Mărănduc, 2010), *Dicționar de expresii românești în contexte*, Vol. 1-4 (Dictionary of Romanian Expressions in Context, (Ilincan, 2015)) and *Dicționar frazeologic al limbii române* (Phraseological Dictionary of the Romanian Language, (Tomici, 2009)). Through a combination of automatic and manual processing, the resulting TXT files were parsed and curated to: (i) exclude verbal expressions (as already annotated in the previous versions of the corpus), (ii) remove definitions, examples, usage notes, variants, lexicographic cross-references, etc., (iii) expand expressions in case of variants rendered as alternations (e.g., the unique entry *Majestatea Ta/Sa/Voastră* was split into three different PronIDs, namely *Majestatea Ta* "Your Majesty" (2nd person singular), *Majestatea Sa* "His/Her Majesty" (3rd person singular) and *Majestatea Voastră* "Your Majesty" (2nd person plural)), (iv) correct errors arising from the automatic content extraction from PDFs (such as end-of-line word segmentation, diacritics misencoding), and (v) format candidate MWEs as lists with one entry per line.

The online dictionary *Dicționarul ortografic, ortoepic și morfologic al limbii române*[4] (DOOM, The Orthographic, Orthoepic, and Morphological Dictionary of the Romanian Language) offers the possibility to retrieve and download lists of idiomatic expressions through queries targeting a

[3] Such examples do not occur in the corpus.

[4] https://doom.lingv.ro/

specific part of speech. The online version of *Dicționarul explicativ al limbii române*[5] (DEX, The Explanatory Dictionary of the Romanian Language) was not downloaded, but it is a comprehensive resource that was manually consulted at all stages of human validation and annotation.

The MWEs extracted from the aforementioned dictionaries were consolidated into a single inventory after duplicate entries were removed. This inventory was then automatically matched to the corpus at the word-form level, and the resulting list of matched MWEs (2,034 unique occurrences) was carried forward to the next annotation stage.

5.2 Curation of the List of MWEs Extracted from Dictionaries

The automatically matched MWEs list was manually validated and labeled with PARSEME MWE categories by a team of six linguists, following a two-step procedure to allow cross-validation. Each expression was assigned one or more labels from the label set in the PARSEME guidelines version 2.0, applying the test battery therein. Some of these expressions were clearly erroneous, arising from errors in the automatic extraction process, while others were plausible MWEs but failed the PARSEME tests corresponding to their specific part of speech. All such expressions were subsequently labeled as NOT MWE.

Entries that could be assigned two different parts of speech given their possible distributions, and consequently two distinct MWE labels, were expanded into separate corresponding entries. For example, the entry *de mână* ('by hand') was split into *de mână* (AdjID), as in *scris de mână* ('writing of hand' "handwriting") (i.e., when having a noun as its syntactic head), and *de mână* (AdvID), as in *scrie de mână* ('writes by hand') (i.e., when having a verb as its syntactic head). This duplication procedure was not applied to MWEs exhibiting polysemy and, thus, occurring with the same part of speech (e.g., *în parte*, labeled as AdvID, has the meanings "partially" (*Ai dreptate, în parte.* "You are right, partially."), but also "separately, one by one" (*El răspunde la fiecare întrebare în parte.* "He answers each question separately.").

Cross-validation was conducted, with each entry being independently validated by two annotators. The overall inter-annotator agreement rate was 57.9% (1,178 out of 2,034 total entries). A third round of validation, carried out by a linguist, was performed on the consensus dataset, while the disagreement dataset was analysed in expert team meetings until a consensus was reached. In certain cases, conflicting annotations were both retained, particularly when one of the annotators had expanded an entry to account for two possible part of speech labels. This is the case of the expression *de mână* discussed above. In most cases, only one of the competing labels was selected as correct, the other(s) being annotation errors. In other cases, the expression was reclassified as NOT MWE for failing to pass the PARSEME tests battery, in spite of being considered MWEs by the authors of the dictionary from which they were automatically extracted, which once again shows that there is no universally accepted definition of MWEs. For example, *fără risipă* ("without waste"), initially labeled as AdjID, was ultimately retained as NOT MWE, while *cu judecată* ("with judgment, rationally"), initially tagged as AdvID and AdjID, was similarly reclassified as NOT MWE in the third validation round (see Subsection 4.2 above).

Overall, 46.9% of the agreed-upon and 30.8% of the disagreed-upon expressions (816 in total) were classified as NOT MWE for failing the PARSEME tests and were, consequently, excluded. The final dataset, after expanding homonymous entries, comprised 2,010 MWEs.

5.3 Automatic MWE Annotation

The manually validated resulting list was used to automatically annotate the PARSEME-Ro corpus, by performing word-form level matching. When identifying an expression for annotation, a window of up to two intervening tokens was permitted between any of its components, accommodating insertions typical of some MWEs. Although this approach does not ensure full recall, it offers a practical trade-off with precision, since the likelihood of false positives increases with the number of allowed intervening words.

5.4 Manual Correction of the Automatic Annotation

A dedicated FLAT platform instance, configured for the PARSEME project and providing individual accounts for each annotator, was used for manual validation of the automatically annotated corpus. The same team of six linguists participated in this stage, which involved: (i) removing one of the two annotations in cases of homonymous expressions

[5] https://dexonline.ro/

bearing two possible labels that could only be disambiguated in context: see the expression *de mână* explained in subsection 5.2; (ii) correct the MWE type label when a wrong one was automatically assigned; (iii) adding or removing components of an expression and (iv) deleting an annotation in case of false positives resulting either from the allowance of intervening tokens or from instances with a compositional meaning. For example, the construction *de preț* 'of price' "precious" is considered AdjID only in contexts such as *Am amintiri de preț din acel concediu.* ('Have-I memories of price from that vacation' "I have precious memories from that vacation."). However, the automatic annotation marked *de preț* as AdjID even in contexts such as *Marcatorul de preț este un aparat care...* ('Marker of price is a device that...' "A labelling system is a device that..."), in which case the annotator removed the label.

At the same time, expressions that were not automatically detected required manual annotation. This occurred primarily for three reasons:

- the list of manually validated MWEs was not exhaustive and therefore did not include some expressions that occur in our corpus. Such examples include expressions referring to meteorological warnings and alerts (*cod portocaliu* "orange code", *cod roșu* "red code"), which were classified as NIDs;
- the number of intervening words between the components of an expression exceed the predefined limit of two, e.g. *Ei **pun**, fără nicio îndoială, **bazele** statului modern.* ('They put, without any doubt, the foundations of the modern state.' "They are undoubtedly laying the foundations for the modern state."); a few functional MWEs also allow for this: e.g., *Au acționat **fără ca** măcăr atunci, în ultimul ceas, **să** le pese de ce simt ceilalți.* "They have acted without at least then, in the last hour, to care about what others feel.";
- the notion of MWE sometimes cover more than what is traditionally called expression (see the case of terms that observe the PARSEME definition of a NID), e.g., *date personale* "personal data".

As shown in Table 3, the automation of the procedure significantly reduced the number of required manual operations (see the great number of automatically annotated MWEs left unchanged after manual correction: 18,953). When counting unchanged MWEs, verbal expressions annotated in previous stages of the project were not taken into account, but any insertion, deletion or modification of verbal MWE was counted.

Operations	**#**
Insertions	4579
Deletions	5177
Modifications	1474
Unchanged	24844
MWEs after manual correction phase	36074
Existing verbal MWEs	5891
Unchanged minus verbal MWEs	18953

Table 3: Operations done in the manual correction of the automatic annotation step.

Unfortunately, due to time constraints and reduced number of staff, the files of the corpus were manually checked only by one linguist. However, in order to understand the extent to which the team of annotators agree in their evaluation, a part of the corpus was doubly annotated: 2000 sentences were randomly selected from the automatically annotated ones and four pairs of annotators manually checked them. We calculated the inter-annotator agreement score using the scripts made available by the PARSEME team (see (Savary et al., 2017)). Its value is 0.78, which shows high consistency among the annotators in our team.

5.5 Ensuring Annotation Consistency

A methodology to ensure consistency in the PARSEME annotation, implemented at the project level through a suite of Python libraries, reflects the initiative's commitment to producing high-quality datasets. All annotations in the corpus are automatically extracted and grouped according to the unique MWE they pertain to, alongside occurrences of the same sequences that were skipped during the automatic or manual annotation stages. This setup allows human validators to examine all contexts of a given word sequence and assess the MWE status and assign a label for each occurrence. By presenting the user with MWEs in contexts, both similar and divergent ones together, the process facilitates more accurate and consistent annotation of MWEs.

When the manual consistency check was over, the F-measure between the manual annotations and the outcomes of the consistency check was calculated and its value is 86, which is indicative of a fairly consistent corpus.

6 Remarks on the Frequency and Variety of MWEs in the PARSEME-Ro Corpus

As shown in Section 3, Table 2, functional MWEs constitute the most frequent category (36% of total MWEs), followed by verbal MWEs (28%), adjectival and adverbial MWEs (23%), and nominal MWEs (13%). Within these main categories, the distribution of subtypes exhibits varying degrees of asymmetry: major imbalances (e.g., PronIDs are significantly less frequent than fully lexical nominal MWEs; DetIDs are extremely underrepresented compared to other functional MWEs), moderate imbalances (AdpIDs show the highest frequency, yet ConjIDs still register a substantial number of occurrences), or balanced distributions (AdjIDs and AdvIDs).

As expected, the content categories (nominal, verbal, adjectival, and adverbial expressions) exhibit greater variety in the corpus compared to functional ones (prepositional and conjunctional): e.g., the expressions (NIDs) *amor propriu* "self-esteem", *cotă de piață* "market share", (AdjIDs) *în putere* "in force, powerful", *la cheie* "turnkey", (AdvIDs) *an de an* "year after year", *cu zâmbetul pe buze* "with a smile on one's face", (VIDs) *a sări în ochi* "to catch the eye", *a aduce pe lume* "to bring into the world, to give birth" have fewer than 5 occurrences each, whereas functional ones (ConjIDs) *pentru că* "because", *după ce* "after", *după cum* "as", (AdpIDs) *în cazul* "in the case of", *față de* "compared to" have between 200 and 500 occurrences each.

Also as expected, the list of MWEs from dictionaries is only partially found in the corpus (see Section 5.1); conversely, new MWEs not recorded in dictionaries were identified. Thus, of the 94 forms of polite pronominal expressions listed in DOOM (grouped into 70 separate dictionary entries), the corpus contains only scattered occurrences of fewer than 10 distinct PronIDs (*Majestatea Sa* "His/Her Majesty", *Sanctitatea Sa* "His/Her Holiness", *Preasfinția Sa* "His/Her Grace", etc.). Therefore, the vast majority of polite pronominal expressions included in the contemporary prescriptive dictionary DOOM (*Luminarea Voastră* "Your Reverence", *Panevghenia Ta* "Your Eminence", *Preacucernicia Sa* "Your Piety", etc.) seem not to be in current use. Manually identified MWEs that are not included in consulted dictionaries (i.e. 53% of the all unique MWEs annotated in the corpus) include AdjIDs (e.g., *de tristă amintire* "of bitter memory" (from the communist period in Romania), NIDs (e.g., *cod galben* "yellow code", *chestiune arzătoare* "a burning issue"), AdpIDs (e.g., *funcție de* "depending on"), etc.

As in other languages, some MWEs have been borrowed or partially or totally loan-translated. During the decision tree application process, we observed that certain diagnostic features of these expressions are due to the source language. For example, the plural form *forțe* "forces" in *forțele armate* "armed forces" (NID) is borrowed through translation from Fr. *forces armées*; the definite articulated form of the noun in *cu forța* "by force" can be similarly explained, cf. Fr. *avec la force*; in contrast, the noun in *cu putere* "forcefully" lacks an article. Some MWE elements have meanings difficult to grasp outside these expressions; e.g., the terminological senses of expressions containing the noun *fond* (e.g., *instanță de fond* "court of first instance" (cf. Fr. *juridiction de fond*), *zgomot de fond* "background noise" (cf. Fr. *bruit de fond*), *a judeca pe fond* "to judge on the merits" (cf. Fr. *juger sur le fond*)); in the NID *fond de rulment* "working capital" (cf. Fr. *fonds de roulement*), *rulment* is a cranberry word — it does not occur independently in Romanian nor in other expressions (it may be mistaken for its homonym denoting a ball bearing).

Some MWEs, particularly those fixed in specialized registers, exhibit pleonastic features; e.g., *funcționar public* (NID) "civil servant" (by definition, see DEX, such employees work exclusively in the public sector), *drept pentru care* (ConjID) "wherefore, because" (both prepositions *drept* and *pentru* express causality, the former being archaic).

7 Issues in Tests Application

In assigning the MWE status, as defined in the project, the application of the proposed tests may encounter several issues, a situation potentially applicable to other languages, not only Romanian. Among these issues, three are particularly noteworthy:

- The cranberry status of a component within an expression, a strong diagnostic test for all MWE types, is sometimes difficult to determine, as it involves relatively subjective evaluations of (i) whether an element is archaic or obsolete (no strict delimitation exists: some outdated lexemes remain familiar to certain speakers through literature, historical texts, etc.), and (ii) the size of the closed set of contexts in which it still occurs. For example, the

conjunction *necum* 'not-how' is perceived as archaic both when used independently (*Ea nu a văzut arma nici măcar în poze, necum în realitate.* "She hasn't seen the weapon even in pictures, let alone in reality.") and when it appears in two other MWEs: *necum să* "so much the less" (ConjID) and *cum-necum* "one way or another" (AdvID); the stand-alone noun *preajmă* is out of use and appears in two AdpIDs: *în preajma* "in the vicinity of" and *din preajma* "from the vicinity of".

- The morphological inflexibility test is inoperative for expressions containing generic nouns or other types of defective forms, as morphological non-prototypicality in such cases has internal semantic constraints. For example, abstract nouns such as *plac* "liking", *libertate* "freedom", *dispoziție* "mood", *pace* "peace", *siguranță* "safety" are inherently morphologically restricted as singularia tantum, and this restriction carries over into expressions: *pe plac* "to one's liking" (AdvID), *în libertate* "free(ly)" (AdvID/AdjID), *și pace!* "and that's it!" (IntjID), etc. Furthermore, in Romanian, expressive or intensifying shifts between number values sometimes occur, as in *la început* (sg.) → *la începuturi* (pl.) "at the beginning" or *fără margini* "boundless" (pl.) → *fără margine* (sg.) "without limit(s)".

- In Romanian, the prohibited modification test (available for NIDs, AdvIDs, AdjIDs, DetIDs), like the morpho-syntactic inflexibility test, fails to determine the MWE status for specialized terms when relative adjectives are present (e.g., *persoană fizică* "natural person", *placă turnantă* "turntable", *politică externă* "foreign policy", *relații publice* "public relations"), as this class systematically blocks gradation (**persoană foarte fizică* 'person very physical') and anteposition (**fizică persoană* 'physical person', (Pană Dindelegan, 2013, 418-419)).

The annotation team discussed these issues and established consistent guidelines.

8 Conclusions and Future Work

We presented here the most recent work for enlarging and enhancing the Romanian corpus annotated with MWEs, namely PARSEME-Ro. It is a resource that will be coupled with an electronic lexicon of MWEs for Romanian (Leseva et al., 2024), so that the linguistic description of the MWEs gets more detailed.

Besides its importance for shared tasks and evaluation campaigns (the annotated corpus was used in the PARSEME 2.0 shared task[6]), the corpus can also serve as a resource for language learning, especially second language learning, as foreigners can find here various contexts for a large number of expressions.

The corpus is also relevant for the language specialists, who can find here a current snapshot of the number, frequency, and inventory of MWEs in the journalistic genre. When compared with existing dictionaries, this can show the tendency of some MWEs to become obsolete, as well as the appearance of new ones. Of course, they need to take this with a grain of salt, given that our corpus is not representative of the whole journalistic writing, let alone of the language in general. We are also interested in extending the analysis of MWEs by adding new text genres to the corpus and then notice similarities and differences between them and the journalistic one.

9 Acknowledgment

Part of the work presented here was was carried out within the "Large Language Models for the European Union (LLMs4EU)", project no. 101198470, call DIGITAL-2024-AI-B-06-LANGUAGE, funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them. Another part of this work was supported by a grant of the Ministry of Research, Innovation and Digitalization - UEFISCDI, project number PN-IV-P8-8.2-EUD-2025-0061, within PNCDI IV. Another part of the work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

References

Andrei-Marius Avram, Verginica Barbu Mititelu, and Dumitru-Clementin Cercel. 2023. Romanian multiword expression detection using multilingual adver-

[6] https://unidive.lisn.upsaclay.fr/doku.php?id=other-events:parseme-st

sarial training and lateral inhibition. *arXiv preprint arXiv:2304.11350*, no volume:no pages.

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, USA.

Verginica Barbu Mititelu. 2013. istemul all-inclusive în reprezentarea cunoştinţelor lexicale. In Ofelia Ichim, editor, *Tradiţie/inovaţie - identitate/ alteritate: paradigme în evoluţia limbii şi culturii române*, pages 9–18. Editura Universităţii „Alexandru Ioan Cuza", Iași.

Verginica Barbu Mititelu, Mihaela Cristescu, and Mihaela Onofrei. 2019. The Romanian corpus annotated with verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 13–21, Florence, Italy. Association for Computational Linguistics.

Verginica Barbu Mititelu, Voula Giouli, Gražina Korvel, Chaya Liebeskind, Irina Lobzhanidze, Rusudan Makhachashvili, Stella Markantonatou, Aleksandra Markovic, and Ivelina Stoyanova. 2025. Survey on lexical resources focused on multiword expressions for the purposes of NLP. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 41–57, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Verginica Barbu Mititelu and Tudor Voicu. 2024. Function multiword expressions annotated with discourse relations in the Romanian reference treebank. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*, pages 90–97, Sofia, Bulgaria. Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

Tiberiu Boros, Sonia Pipa, Verginica Barbu Mititelu, and Dan Tufiș. 2017. A data-driven approach to verbal multiword expression detection. PARSEME shared task system description paper. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 121–126, Valencia, Spain. Association for Computational Linguistics.

Cecilia Căpățână. 2007. *Elemente de frazeologie*. Editura Universitaria, Craiova.

Florica Dimitrescu. 1958. *Locuțiunile verbale în limba română*. Editura Academiei, Bucharest.

Vasile Ilincan. 2015. *Dicționar de expresii românești în contexte [DERC]*. Presa Universitară Clujeană, Cluj-Napoca.

Eugen Ioaniţescu. 1956. Locuţiunile. *Limba română*, 6:48–54.

Svetlozara Leseva, Verginica Barbu Mititelu, Ivelina Stoyanova, and Mihaela Cristescu. 2024. A uniform multilingual approach to the description of multiword expressions. In Voula Giouli and Verginica Barbu Mititelu, editors, *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, pages 73–116. Language Science Press, Berlin.

Carmen Mîrzea Vasile. 2012. *Eterogenitatea adverbului românesc. Tipologie și descriere*. Editura Universității din București, Bucharest.

Cătălina Mărănduc. 2010. *Dicționar de expresii, sintagme și locuțiuni ale limbii române, DELS*. Corint, Bucharest.

Gabriela Pană Dindelegan, editor. 2013. *The Grammar of Romanian*. Oxford University Press, Oxford.

Gabriela Pană Dindelegan, Raluca Brăescu, and Cristiana Aranghelovici. 2025. *Locuțiunile limbii române*. Univers Enciclopedic, Bucharest.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, and 9 others. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomir Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, and 3 others. 2018. *PARSEME multilingual corpus of verbal multiword expressions*. Number 2 in Phraseology and Multiword Expressions. Language Science Press, Berlin.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Mile Tomici. 2009. *Dicţionar frazeologic al limbii române*. Editura Saeculum Vizua, Bucharest.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. Technical report, University of Pennsylvania.

Incorporating Multiword Expressions in Galician Neural Machine Translation: Compositionality, Efficiency, and Performance

Daniel Solla, Paula Pinto-Ferro, Laura Castro
Pablo Gamallo and **Marcos Garcia**
CiTIUS - Centro Singular de Investigación en Tecnoloxías da Información
Universidade de Santiago de Compostela
{pablo.gamallo,marcos.garcia.gonzalez}@usc.gal

Abstract

This paper explores the behavior of neural machine translation models on two newly introduced datasets containing noun-adjective MWEs with different degrees of semantic ambiguity and compositionality. We compare general-domain machine translation systems with fine-tuned models exposed to small subsets of the target MWEs. By assessing the effects of the learning steps and corpus size, we found that carefully designed fine-tuned may improve MWE handling while mitigating catastrophic forgetting. However, our error analysis reveals that models still struggle in several scenarios, particularly when translating MWEs with idiomatic meanings. Both the datasets and the experiments focus on translation involving Galician, English, and Spanish.

1 Introduction

In the last decade, Neural Machine Translation (NMT) evolved from recurrent networks (Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2015) to Transformer-based sequence-to-sequence (*seq2seq*) architectures (Vaswani et al., 2017). More recently, the use of Large Language Models (LLMs) as translation engines has gained popularity, especially due to their impressive performance in high-resource languages (Zhu et al., 2024). However, the performance of *seq2seq* models in low-resource languages remains competitive with current LLMs (Robinson et al., 2023; Gibert et al., 2025), while also providing other advantages such as faster inference and lower computational energy costs.

In either approach, incorporating new lexicon, contexts, or textual domains in a Machine Translation (MT) system can be challenging. On the one hand, training a model from scratch and exploring different hyperparameter configurations incurs substantial computational costs. On the other hand, fine-tuning (FT) an existing model with new datasets risks catastrophic forgetting, where gains on the new data can lead to reduced performance on previously learned contexts or domains (McCloskey and Cohen, 1989; Gu and Feng, 2020).

One particular challenge that NMT systems often struggle with is the accurate translation of Multiword Expressions (MWEs). In addition to being pervasive across all languages (Ramisch, 2023), MWEs present ambiguities at multiple levels. First, from the perspective of semantic compositionality, MWEs exhibit varying degrees of idiomaticity (e.g., the compositional *apple tree*, or *red herring* meaning a 'misleading clue' idiomatically). Second, at the level of individual components, some words within MWEs have high degrees of polysemy (e.g., *green bank* where 'bank' may refer to a financial institution or to a river bank, while 'green' may also have multiple senses). Finally, like single words, MWEs can convey different meanings depending on the context (e.g., *glass ceiling* as an 'invisible barrier' or as a physical structure). Among other factors, these make MWEs particularly challenging to model, not only for NMT but also for neural causal language modeling (Dankers et al., 2022; Liu et al., 2025; He et al., 2025).

This paper investigates continual learning approaches for integrating MWEs into NMT models, and evaluates their performance on such expressions. We focus on the translation from Galician (GL) to two languages: English (EN) —as an example of a distant language— and Spanish (ES) —a related language that exerts influence on Galician. Specifically, we present *i)* two new parallel datasets (Galician-English and Galician-Spanish) composed of sentences with MWEs with different types of ambiguity;[1] *ii)* an extensive set of experiments assessing the influence of learning steps and

[1]The dataset is available at https://github.com/marcospln/parallel_noun-adj_gl-en-es/. A subset of manually created sentences is kept private to prevent data contamination in future research.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 75–85

size of the training corpus; *iii)* quantitative analyses comparing base and fine-tuned models; and *iv)* a qualitative error analysis showing the effects of the compositionality and frequency levels of the MWEs on their translations.

Our findings show that *i)* the degree of overlap in MWE composition seems to be a function of linguistic proximity (i.e., Galician-Spanish MWEs are more similar than Galician-English equivalents), and this crucially affects translation performance; *ii)* targeted fine-tuning can be an effective and efficient strategy for enhancing MWE translation, although only in some cases; *iii)* NMT still struggles to translate MWEs, primarily due to idiomaticity, with factors such as frequency and linguistic distance also contributing to the difficulty.

2 Related work

Recent studies in low-resource NMT demonstrate that combining multilingual pre-trained models with synthetic corpora can achieve strong translation performance even when high-quality parallel data is scarce (Sant et al., 2024). Fine-tuning self-supervised multilingual models such as mBART on small parallel corpora has proven effective for adapting to new language pairs (Thillainathan et al., 2021). A key challenge in continual training is catastrophic forgetting (McCloskey and Cohen, 1989). Several works analyze forgetting at both module and parameter levels, showing that excessive parameter drift during domain adaptation can degrade general-domain performance (Gu and Feng, 2020). Complementary research highlights that the extent of forgetting is related to the properties of the adaptation data, such as the introduction of new target vocabulary (Saunders and DeNeefe, 2024). These findings suggest that careful, targeted fine-tuning is less likely to harm overall translation quality than broad domain shifts or large, uncontrolled adaptation datasets.

Instruction-based fine-tuning, originally developed for LLMs, has also been adapted to traditional encoder-decoder NMT systems. Such methods allow models to learn multiple translation customization tasks jointly through compact fine-tuning stages, demonstrating that specialized behaviors can be efficiently acquired without full retraining (Raunak et al., 2024). According to recent reviews, multilingual pre-training on generic language tasks allows models to internalize shared structures across languages. This shared knowledge helps compensate for the lack of parallel data during fine-tuning for translation (Ataman et al., 2025).

In the context of MWEs and idiomaticity, early work demonstrated the benefits of detecting and specially handling phrasal verbs, which significantly improved translation consistency and fluency (Kordoni and Simova, 2014). In the same line, MWE-aware NMT approaches using annotation and data augmentation with external linguistic resources have shown substantial improvements (Zaninello and Birch, 2020). More recent research on Transformer-based NMT demonstrates that these systems exhibit an excessive bias toward compositionality, leading to systematic difficulties in modeling non-compositional expressions (Dankers et al., 2022; Liu et al., 2025)

Several datasets have been released aimed at evaluating the performance of NMT and other models on MWEs with different degrees of idiomaticity, such as the English-based MAGPIE (Haagsma et al., 2020), MultiMWE, including Chinese, English, and German (Han et al., 2020b), or the also multilingual AlphaMWE —with the same languages and Polish— (Han et al., 2020a), recently expanded to other varieties including Italian and Arabic (Han et al., 2025).

Regarding NMT for Galician, the *Proxecto Nós* (*Nós Project*) recently released state-of-the-art models[2], and its research reported benefits from architectural adaptations such as smaller BPE vocabularies, which consistently improve performance across data scales (Outeirinho et al., 2024). New resources, including the CorpusNÓS (de Dios-Flores et al., 2024) and parallel datasets[3], provide large-scale training material for both NMT and LLM-based translation models enabling improved translation quality in low-resource settings.

To the best of our knowledge, there is currently no parallel MWE dataset for Galician that provides manually curated translations together with annotations for idiomaticity class, sense, and frequency. Moreover, no prior work has specifically investigated continual learning strategies for NMT with a focus on MWEs including Galician. Our work directly addresses these gaps.

[2] https://nos.gal/gl/proxecto-nos

[3] https://github.com/proxectonos/corpora#traducion-automatica

3 New parallel corpus of noun-adjective MWEs

This section introduces two new parallel corpora (Galician-Spanish and Galician-English) composed of sentences containing fine-grained annotation of noun-adjective MWEs.

Source data: The source data is the dataset presented by Castro et al. (2025), which comprises 240 noun-adjective MWEs in Galician conveying 322 different senses. Each of these senses is contextualized in up to 6 sentences, totaling 1,858 examples (average of 5.77 per sense). The initial identification of the 240 MWEs was performed using *i)* a dependency-based approach, i.e., extracting contiguous and non-contiguous noun adjective pairs linked by a dependency relation, and *ii)* a frequency-based criterion, selecting expressions from two ranges: high and low frequency.

MWEs' properties: At the token-level, each contextualized MWE is classified according to its compositionality class in the given context as *idiomatic* (e.g. *obra morta*, 'freeboard', literally 'death job'), *compositional* (e.g. *centro comercial*, 'shopping center'), or *partial* (e.g. *campo magnético*, 'magnetic field'). At the type-level, MWEs that may have different compositionality scales depending on the context are marked as *potentially idiomatic expressions* (e.g. *montaña rusa*, literally 'Russian mountain' which can refer idiomatically to a roller coaster or to an actual mountain in Russia).

Translations: Each of the 1,858 sentences was manually translated into Spanish and English by a professional translator, and then reviewed by a second one, ensuring the quality of the parallel resource. The translators were asked to perform an adequate translation taking into account both the meaning of the MWE and of the whole sentence, so that the original MWEs (in Galician) were not always translated by a MWE in the target languages (e.g. *fondo mariño*, literally 'sea bottom', translated as 'seabed'). During this process, the translators also identified those elements in the new sentences conveying the meaning of the original MWE, allowing for further qualitative analyses. The final resources are two parallel corpus composed of bilingual pairs of 1,858 sentences. Table 1 includes some examples of the original sentences and their English translations.

Semantic phenomena: The dataset contains three main types of linguistic phenomena that may challenge language modeling in general and NMT in particular: *i)* compositionality class, including idiomatic, partially idiomatic, and compositional expressions (see examples above); *ii)* ambiguity of the components: a component of the MWE, namely the head, may have different meanings: e.g., *número enteiro*, where *número* ('number') may have the following three senses and corresponding translations: mathematical ('integer'), graphical (e.g., 'whole number'), journalistic ('full issue', as in the complete volumes of a journal); *iii)* ambiguity of the whole MWE, e.g., *auga doce* (literally 'sweet water') which may refer to 'fresh water' (vs. 'salty water'), or to 'sweetened water' (i.e., with sugar). It is worth mentioning that the former ambiguities (at the word and at the MWE level) may occur in the same compositionality class or across different scales.

Note that the degree of semantic divergence is considerably smaller between Galician and Spanish than between Galician and English. For many MWEs, a literal, word-for-word translation from Galician into Spanish is often accurate and preserves both meaning and structure, whereas this is not the case for Galician-English. This contrast further motivates the use of GL-EN MWEs as a challenging evaluation setting, since the model must resolve non-literal correspondences that are not predictable from morphology or word-level semantics alone.

4 Materials and methods

Models: All experiments are based on the NMT models developed within the *Proxecto Nós*, publicly available through HuggingFace.[4] These models are implemented using the OpenNMT framework (Klein et al., 2017) and follow a Transformer-based encoder-decoder architecture. They represent state-of-the-art systems for Galician-centric MT (Buján et al., 2025).

Fine-tuning data: We use the new MWE datasets (§ 3) to both fine-tune and evaluate the performance of the NMT models. As mentioned, each parallel corpus is composed of up to 6 sentences for each of the 322 MWE senses. We first select two sets of 322 sentences, using one for fine-tuning and the other for evaluation. We then increase the

[4] https://huggingface.co/collections/proxectonos/mt

MWE	Sentence (GL)	Sentence (EN)
fonte principal	...a principal fonte da cidade...	...the main fountain of the city...
fonte principal	...eran a fonte principal de recrutamento das súas tropas.	...and they were also the main source of recruitment for their troops.
zona húmida	As zonas máis húmidas son as rexións occidental e central...	The western and central regions are the most humid areas...
zona húmida	...das zonas húmidas europeas foron totalmente destruídos.	...of European wetlands were totally destroyed.
centro comercial	É o principal centro comercial e industrial do estado...	It is the main commercial and industrial hub in the state...
centro comercial	O centro comercial tiña no seu proxecto inicial...	The initial design for the shopping mall included...

Table 1: Examples of the original MWEs and sentences in Galician and their English translations.

training data with more sets to analyze the impact of the amount of learning data.[5]

Evaluation sets: We assess the impact of fine-tuning in both generic parallel corpus and in a subset of the MWE dataset for each pair of languages. As standard datasets, we use parallel sentences from Flores (Goyal et al., 2022), Tatoeba (Tiedemann, 2020), and from the *Nós Project*, which contain generic datasets and a detailed test-suite focused on particular linguistic phenomena.[6] Table 2 shows the size of each of the evaluation sets. These test sets allow us to measure both general MT performance and improvements on targeted MWEs, providing a balanced evaluation of fine-tuning effects and potential catastrophic forgetting.

Dataset	GL–EN	GL–ES
Flores	1,012	1,012
Tatoeba	1,018	3,131
Nós_MT_Gold_1	1,777	1,998
Nós_MT_Gold_2	1,777	1,998
Nós_MT_Test-suite	364	334
MWE dataset	322	322

Table 2: Size (in number of sentences) of the test sets.

Evaluation metrics: For general translation quality, we use the standard BLEU and TER metrics.[7] The results on the general MT corpora are reported using the micro-average across the five datasets (Table 2). Besides, during experimentation, model selection is guided by *Score_mwe*, a metric proposed by Zaninello and Birch (2020). This metric is specifically designed to evaluate how accurately MT systems render MWEs, without relying on explicit phrase alignment. It operates at the sentence level by comparing each word of the reference translation of a source-side MWE with the closest matching word in the system hypothesis, using a character-based Levenshtein distance.

5 Experiments

All experiments were conducted using the OpenNMT framework.[8] Fine-tuning was performed on a single NVIDIA Tesla V100S GPU (32GB) on an HPC cluster. All models share the same architecture and base training configuration (detailed in Appendix A). Fine-tuning experiments use the hyperparameter settings of the original models, and differ only in the number of FT steps and the size of the MWE corpus.

As the fine-tuning datasets were designed using Galician MWEs as source, our analyses focus on NMT systems from Galician, although we also refer to the other translation directions when necessary.

Experiment 1. Learning steps: We conducted a series of experiments varying the number of fine-tuning steps, using values from 100 to 2,000 in increments of 100. For each translation direction we used one set of 322 sentences for fine-tuning, and the other for evaluating.[9]

[5] Some of the final sets contain slightly fewer than 322 sentences, particularly for low-frequency senses.

[6] https://github.com/proxectonos/corpora

[7] We also computed ChrF, BERTScore, and METEOR, but we mostly rely on BLEU to discuss the results as in general all of these metrics correlate in our experiments.

[8] https://opennmt.net

[9] To ensure that the results do not depend on the sets used, we also reverse their role for training and testing, with almost

Experiment 2. Corpus size: To assess the impact of the number of examples used for continual learning, we incrementally fine-tuned the model with additional sets of ≈322 sentence pairs (from 1 to 5 sets), until reaching 1,536 instances, i.e., the full dataset except for the held-out set. In each case, we fine-tuned the models during 600 and 1,200 steps.

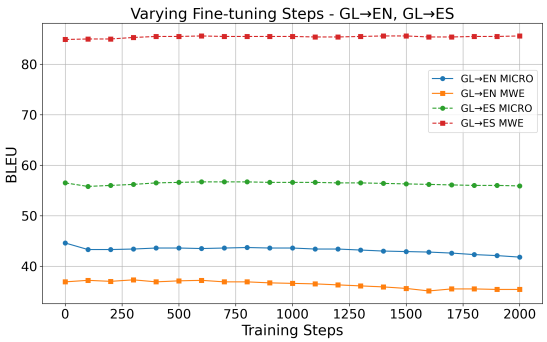
6 Results and discussion

Experiment 1. Number of Fine-Tuning Steps: Figure 1 displays the BLEU learning curves of GL-EN and GL-ES from 0 (original model) to 2,000 fine-tuning steps. Figure 1a shows the micro-average in the general corpora and in the MWE test sets, while Figure 1b plots the results of the *Score_mwe* metric. In GL→EN, as fine-tuning progresses, average scores steadily decline in BLEU, indicating gradual drops in overall translation quality. Meanwhile, MWE evaluations show small gains in BLEU and *Score_mwe* early on, but the improvements do not hold and drop rapidly, showing that targeted MWE updates offer limited benefits and cannot compensate for the declining general performance. In GL→ES, average scores remain stable throughout fine-tuning, with BLEU around 56 and consistent general translation quality. MWE evaluations show strong and sustained improvements, reaching *Score_mwe* above 0.98. In this respect, it is worth recalling that, due to the linguistic similarity between the two languages, MWEs in Galician can be translated literally into Spanish in most cases (see § 3).

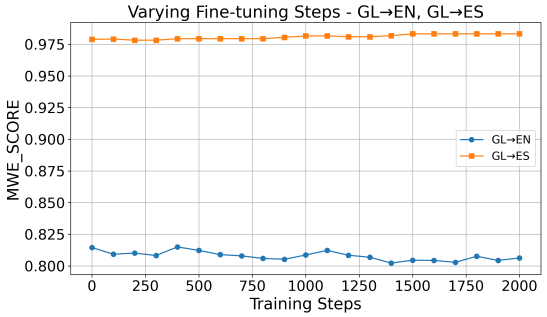
In general, the number of fine-tuning steps influences both overall translation quality and MWE learning. Early steps primarily stabilize general performance, while moderate fine-tuning (600–1000 steps, depending on the translation direction) tends to maximize MWE gains without significantly destabilizing the model. However, excessive fine-tuning steps can lead to overfitting or gradual drops in general BLEU scores, showing the need for careful step selection to balance MWE learning with overall translation quality.

In sum, the results of this set of experiments suggest that continual learning of MWEs, especially for a distant linguistic variety and with limited data (just 322 sentence pairs) is a hard task that requires careful analysis of the trade-off between learning new expressions and overall performance of the

identical results.



(a) BLEU scores GL→EN, GL→ES



(b) *Score_mwe* GL→EN, GL→ES

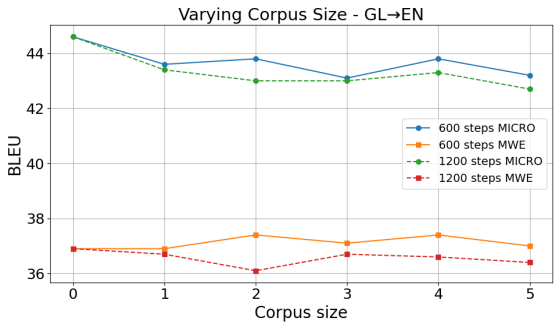
Figure 1: Effect of varying fine-tuning steps on BLEU and *Score_mwe* across different translation directions.

NMT systems. In the next experiment, we explore whether more training data enables improvements in the performance of the models.

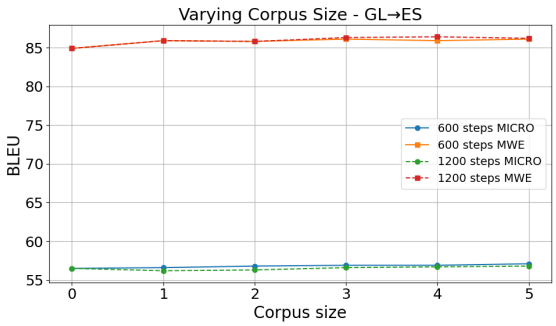
Experiment 2. Corpus Size: In GL-EN, increasing the MWE learning data involves a gradual loss in overall translation quality, although with modest gains in the performance on the MWE dataset (left column of Figure 2). In GL-ES, more training data also generally improves the quality of MWE translations while maintaining competitive results in the general domain, although the BLEU in the MWE dataset was high in the original model (Figure 2, right column).

Overall, the impact of increasing the MWE fine-tuning corpus size is highly direction-dependent. While larger data can substantially improve MWE translation, these gains are not always aligned with improvements in general translation quality. Taken together, these results show that scaling MWE fine-tuning data is more effective when combined with moderate fine-tuning steps and when the translation direction exhibits inherent robustness to domain-specific updates.

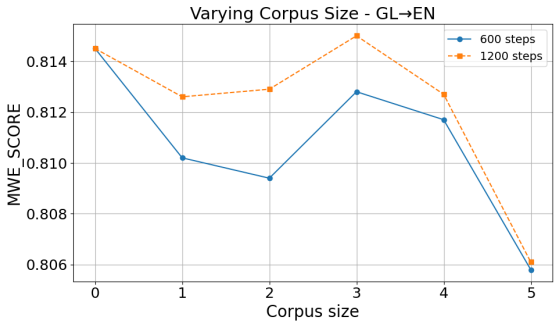
Best model selection: The best-performing models for each translation direction were selected



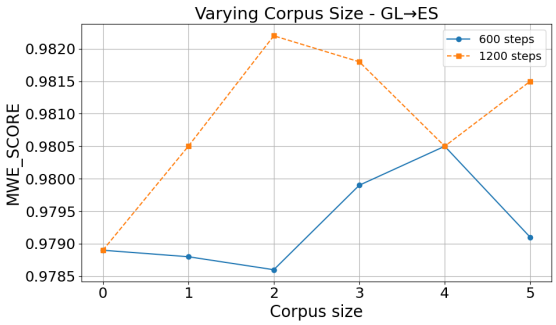
(a) BLEU scores



(c) BLEU scores



(b) *Score_mwe*



(d) *Score_mwe*

(A) GL→EN translation vs. MWE corpus size.

(B) GL→ES translation vs. MWE corpus size.

Figure 2: Translation performance as a function of MWE corpus size for GL→EN and GL→ES.

based on a balanced evaluation of general-domain translation quality and MWE-specific performance. In particular, model selection aimed to maximize improvements in MWE translation accuracy while preserving performance on general test sets, thereby minimizing the risk of catastrophic forgetting. This selection criterion ensures that improvements in MWE handling are not obtained at the expense of overall translation quality.

For each of the four translation directions, we compare the original baseline model with its fine-tuned counterpart. General translation quality is assessed using standard MT metrics (BLEU and TER) computed over multiple general-domain test sets and aggregated using micro-averaging. In parallel, MWE-specific performance is evaluated on a dedicated MWE test set using the same metrics, together with the value of the *Score_mwe*. Table 3 summarizes the results for the models (original and fine-tuned versions), reporting micro-average BLEU and TER scores on the general-domain test sets alongside the corresponding scores on the MWE test set. Although they are not the focus of this analysis, we include the results of the models from English and Spanish to Galician.

Model	**General**		**MWE**	
	BLEU	**TER**	**BLEU**	**TER**
GL→EN B	44.6	40.4	36.9	46.5
GL→EN FT	43.6	40.6	36.9	**46.3**
EN→GL B	38.6	45.9	37.9	47.7
EN→GL FT	**38.9**	**45.1**	**40.2**	**45.5**
GL→ES B	56.5	33.2	84.9	8.8
GL→ES FT	**56.6**	**32.7**	**86.3**	**7.9**
ES→GL B	52.3	36.3	83.7	9.2
ES→GL FT	51.6	**36.1**	83.4	**9.0**

Table 3: Summary of best models (Base and Fined-Tuned) for all translation directions. Results are general-domain and MWE-specific micro-averaged BLEU and TER. Numbers in bold are FT models with better results.

7 Qualitative Analysis

To complement the quantitative results, we use the best-performing models to carry out a series of qualitative analyses of the MWE translations from Galician to EN and ES. This analysis also allows us to gain a finer-grained understanding of how fine-tuning affects the translation of noun–adjective MWEs with different semantic properties.

MWE (GL)	Base model output	Fine-tuned model output
zona húmida	wet zone	wetland
liña férrea	iron line	railway line
montaña rusa	mountain of Russian	rollercoaster
réxime franquista	Franco regime	Francoist regime
banda estadounidense	band	American band

Table 4: Examples of MWE translations improved after fine-tuning (GL→EN)

To do so, a language expert manually classified each MWE translation in the test sets as *Correct*, *Variant*, and *Incorrect*.[10] An instance was labeled as *Correct* when the MWE was translated with the same expression as the reference. *Variant* labels were assigned to meaning-preserving alternatives that differed lexically or stylistically from the reference. *Incorrect* cases were subsequently classified as *i)* inadequate literal translations, *ii)* wrong sense disambiguation, *iii)* partial or full omission of the MWE, and *iv)* other errors (e.g., spelling errors, untranslated source-language words, or otherwise unintelligible outputs).

MWE translation accuracy: The first analysis examined the influence of the reference sentences on the quantitative results by computing accuracy under two evaluation criteria: (i) only *Correct* instances are counted as correct, and (ii) both *Correct* and *Variant* instances are treated as adequate translations: For GL-ES, the results were very similar (accuracies around 0.97 in every case), while for GL-EN the results increased from 0.62 (original and fine-tuned models in the first scenario) to 0.77 (both models in the second evaluation). These results reinforce the need for qualitative evaluation of MT systems.

The manual review of all translation allowed us to observe MWE translation differences between the original and the fine-tuned models. In this regard, in some GL-EN cases, FT enabled the model to correctly translate several MWEs that were previously mistranslated by the base model. These include cases involving non-compositional meaning or strong sense shifts, such as *zona húmida* (*wet zone→wetland*) or the idiomatic *montaña rusa* (*Russian mountain→roller coaster*). Additional improvements were observed in cases where the fine-tuned model selected a more idiomatic or semantically precise variant (e.g. *Franco regime→Francoist regime*; *harsh blow→hard blow*). Table 4 highlights cases where fine-tuning successfully corrected MWE translations that were previously incorrect in the base model. At the same time, a small number of MWEs that were correctly translated by the base model became incorrect after fine-tuning. These regressions typically involved increased literalness or minor lexical degradation, such as *main fountain* being rendered as *main source*, or reduced lexical realization (e.g. *Olympic gold medals → Olympic golds*). Although these cases are relatively few, they illustrate the delicate balance between specialization and generalization in continual learning. Table 5 presents representative examples of observed error types from the GL→EN experiments.

Despite the high performance of the original GL-ES model, fine-tuning still leads to targeted improvements. Several MWEs previously mistranslated by the base model are correctly handled after adaptation (e.g., *yacimiento→yacimiento arqueológico*, *piedra filosofalda→piedra filosofal*, *pescado azul→pez azul*). Improvements also occur in stylistic variants (e.g., *pariente cercano→pariente próximo*). Some representative GL-ES error examples can be seen in Table 6.

Error types: Table 7 shows the distribution of the translation error types of the MWEs. In every case, wrong sense disambiguation and wrong literal translations account for the majority of errors, confirming the difficulty of resolving meaning for non-compositional MWEs. While in GL-ES there are no significant differences between the base and fine-tuned models, in GL-EN, the fine-tuned models seem to perform better semantic disambiguation, but also produce more (inadequate) literal translations.

Frequency effects: We take advantage of the MWE frequency classification included in the original dataset to observe if it has any effect in the quality of the translation. The results (Table 8) indicate that in most cases high-frequency MWEs

[10] In a first step, we included the category *Doubt* reserved for borderline cases, which were solved before the final analysis.

Error type	MWE (GL)	System output (EN)	Reference (EN)
Literal translation	xénero musical	musical gender	music genre
Wrong sense disamb.	montaña rusa	Russian mountain	roller coaster
Omission	terra firme	land	dry land
Others	fosa común	mass mass	mass grave

Table 5: Representative error types in GL→EN MWE translation

Error type	MWE (GL)	System output (ES)	Reference (ES)
Literal translation	letra grosa	letra gruesa	letra negrita
Wrong sense disambiguation	peixe azul	pescado azul	pez azul
Spelling error	cambio climático	cambio climatico	cambio climático
Untranslated words	xogador novo	jugadoras novas	jugadoras jóvenes
Others	pedra filosofal	piedra filosofalda	piedra filosofal

Table 6: Representative error types in GL→ES MWE translation

Error type	GL-EN		GL-ES	
	Base	FT	Base	FT
Literal trans.	0.41	0.44	0.38	0.38
Wrong sense dis.	0.53	0.48	0.25	0.24
Omission	0.05	0.07	0.12	0.00
Others	0.01	0.01	0.25	0.38

Table 7: Percentage of MWE translation error types.

are more easily translated than low frequency ones. While in GL–ES the translation accuracy of frequent MWEs is, on average, less than 1% higher than that of low-frequency MWEs, in GN–EN the gap is much more pronounced, reaching almost 10% on average.

Pair	Frequency	Base	FT
GL-EN	High	81.53%	82.16%
	Low	72.73%	71.51%
	Overall	77.02%	76.71%
GL-ES	High	98.09%	97.45%
	Low	96.97%	97.57%
	Overall	97.51%	97.51%

Table 8: Percentage accuracy of MWE translation by frequency.

Compositionality effects: However, because MWEs can display different contextual senses independently of their frequency, we additionally examine translation performance across token-level compositionality classes. The results in Table 9 show that, namely in GL-EN (as in the previous analyses), idiomatic expressions remain challenging for NMT, both for generic systems and for FT models trained with examples of the target MWEs, yielding the lowest average accuracy (54%). By contrast, performance is substantially higher for partial MWEs (77%) and fully compositional MWEs (82.5%), confirming a clear correlation between degree of compositionality and translation accuracy.

Pair	Comp. class	Base	FT
GL-EN	Idiomatic	56.25%	52.08%
	Partial	76.47%	77.65%
	Compositional	82.54%	82.54%
GL-ES	Idiomatic	93.75%	95.83%
	Partial	96.47%	95.29%
	Compositional	98.94%	98.94%

Table 9: Percentage accuracy of MWE translation by compositionality class.

8 Conclusions and further work

This study explored the effects of targeted fine-tuning on MWE translation in Galician-English and Galician-Spanish, systematically analyzing the impact of both the number of learning steps and the size of MWE corpora. We release two new manually created datasets composed of pairs of 1,858 sentences with detailed annotation of the MWEs.

The results of systematic evaluations suggest that moderate FT (around 600–1000 steps) generally provides the best balance between general translation quality and MWE-specific improvements, but this depends on the language pair under evaluation. In this regard, for similar varieties where the semantics of MWEs may be less divergent (Galician and Spanish vs. Galician and English), the performance

of the original models is competitive.

Regarding the datasets, increasing the size of the MWE fine-tuning corpus does not always guarantee improvements. GL→EN models show limited sensitivity, and ES→GL only exhibits modest gains, indicating again that language-specific characteristics influence how effectively additional MWE data can be leveraged.

A qualitative analysis allowed us to observe that high-frequency MWEs are generally easier to translate, and that idiomatic ones are harder to translate than compositional MWEs, indicating that non-compositional meaning remains difficult to capture.

Building on the insights from this study, several directions for future research are suggested: First, to explore adaptive FT methods that adjust the number of steps or learning rate based on model performance on both general and MWE-specific validation sets. Second, to experiment with synthetic or automatically extracted MWE corpora to increase coverage of rare and idiomatic expressions, and assess their impact on model robustness and generalization. Finally, we plan to extend the proposed FT approach to types of MWEs (e.g., verb–object constructions) and semantic phenomena (e.g., various types of lexical ambiguity).

Limitations

This study presents a controlled analysis of targeted fine-tuning for MWE translation. However, several limitations should be acknowledged.

First, the size of the manually curated MWE datasets is relatively small. Although the corpora were carefully constructed and translated to ensure high linguistic quality, the limited number of sentences per MWE sense restricts the statistical power of the results and may limit their generalizability to broader domains or unseen MWE types.

Second, the experiments focus on a restricted set of MWE categories, noun-adjective pairs. Other types of MWEs are not covered, and as a result, the conclusions may not directly transfer to all forms of non-compositional language.

Third, the analysis is limited to four translation directions involving Galician, English, and Spanish, with a primary focus on translations from Galician. While these directions offer valuable insights into low-resource and asymmetric translation settings, the findings may not generalize to languages with different typological properties. Furthermore, the original MWEs were compiled in only one of the languages (Galician).

Finally, fine-tuning was performed under controlled experimental conditions using fixed architectures and hyperparameters. Alternative model architectures, parameter-efficient fine-tuning methods, or dynamic training strategies were not explored and could lead to different trade-offs between MWE performance and general translation quality.

Acknowledgments

This paper was funded by MCIU/AEI/10.13039/501100011033 (grants with references PID 2021-128811OA-I00, PID2024-161928OB-I00, CNS2024-154902, and AIA2025-163322-C62), by the Galician Government (ED431G 2023/04 and ED431B 2025/16), and by the *Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia* - Funded by EU — NextGenerationEU within the framework of the project *Desarrollo Modelos ALIA*.

References

Duygu Ataman, Alexandra Birch, Nizar Habash, Marcello Federico, Philipp Koehn, and Kyunghyun Cho. 2025. Machine Translation in the Era of Large Language Models:A Survey of Historical and Emerging Problems. *Information*, 16(9):723. Publisher: Multidisciplinary Digital Publishing Institute.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Saúl Buján, Daniel Bardanca Outeiriño, Pablo Gamallo, Iria de Dios Flores, and José Ramón Pichel Campos. 2025. Machine translation for low-resource languages: Performance trade-offs between seq2seq and generative approaches. *Procesamiento del Lenguaje Natural*, 75:297–315.

Laura Castro, Anna Temerko, and Marcos Garcia. 2025. Compositionality and Ambiguity in Multiword Expressions: A Dataset for the Evaluation of Language Models in Galician. In *Progress in Artificial Intelligence*, pages 228–240, Cham. Springer Nature Switzerland.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Iria de Dios-Flores, Silvia Paniagua Suárez, Cristina Carbajal Pérez, Daniel Bardanca Outeiriño, Marcos Garcia, and Pablo Gamallo. 2024. CorpusNÓS: A massive Galician corpus for training large language models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 593–599, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Ona de Gibert, Dayyán O'Brien, Dušan Variš, and Jörg Tiedemann. 2025. Mind the gap: Diverse NMT models for resource-constrained environments. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 209–216, Tallinn, Estonia. University of Tartu Library.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Shuhao Gu and Yang Feng. 2020. Investigating catastrophic forgetting during continual training for neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4315–4326, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020a. AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020b. MultiMWE: Building a multi-lingual multi-word expression (MWE) parallel corpora. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2970–2979, Marseille, France. European Language Resources Association.

Lifeng Han, Najet Hadj Mohamed, Malak Rassem, Gareth Jones, Alan Smeaton, and Goran Nenadic. 2025. Towards a resource for multilingual lexicons: an mt assisted and human-in-the-loop multilingual parallel corpus with multi-word expression annotation. *Language Resources and Evaluation*. Forthcoming.

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

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Valia Kordoni and Iliana Simova. 2014. Multiword Expressions in Machine Translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1208–1211, Reykjavik, Iceland. European Language Resources Association (ELRA).

Linfeng Liu, Saptarshi Ghosh, and Tianyu Jiang. 2025. Evaluating the impact of verbal multiword expressions on machine translation. *Preprint*, arXiv:2508.17458.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation - Advances in Research and Theory*, 24(C):109–165.

Daniel Bardanca Outeirinho, Pablo Gamallo Otero, Iria de Dios-Flores, and José Ramom Pichel Campos. 2024. Exploring the effects of vocabulary size in neural machine translation: Galician as a target language. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 600–604, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Carlos Ramisch. 2023. Multiword expressions in computational linguistics. Habilitation à diriger des recherches. Aix Marseille Université (AMU).

Vikas Raunak, Roman Grundkiewicz, and Marcin Junczys-Dowmunt. 2024. On instruction-finetuning neural machine translation models. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1155–1166, Miami, Florida, USA. Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Aleix Sant, Daniel Bardanca, José Ramom Pichel Campos, Francesca De Luca Fornaciari, Carlos Escolano, Javier Garcia Gilabert, Pablo Gamallo, Audrey Mash, Xixian Liao, and Maite Melero. 2024. Training and

Fine-Tuning NMT Models for Low-Resource Languages Using Apertium-Based Synthetic Corpora. In *Proceedings of the Ninth Conference on Machine Translation*, pages 925–933, Miami, Florida, USA. Association for Computational Linguistics.

Danielle Saunders and Steve DeNeefe. 2024. Domain adapted machine translation: What does catastrophic forgetting forget and why? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12660–12671, Miami, Florida, USA. Association for Computational Linguistics.

Sarubi Thillainathan, Surangika Ranathunga, and Sanath Jayasena. 2021. Fine-Tuning Self-Supervised Multilingual Sequence-To-Sequence Models for Extremely Low-Resource NMT. In *2021 Moratuwa Engineering Research Conference (MERCon)*, pages 432–437. ISSN: 2691-364X.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Andrea Zaninello and Alexandra Birch. 2020. Multiword Expression aware Neural Machine Translation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3816–3825, Marseille, France. European Language Resources Association.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Appendix

A Training and Fine-Tuning Hyperparameters

Table 10 summarizes the architecture and training hyperparameters used for both base training and continual fine-tuning of the Seq2Seq Transformer models. All models were trained using the OpenNMT framework, using the same configuration across all translation directions.

Parameter	**Value**
Model architecture	
Encoder layers	12
Decoder layers	12
Attention heads	16
Hidden size	512
Feed-forward size	2048
Dropout	0.1
Label smoothing	0.1
Position encoding	Enabled
Training and fine-tuning	
Optimizer	Adam
β_2	0.998
Learning rate	2
Warmup steps	8,000
Batch size	4,096 tokens
Gradient accumulation	4 steps
Maximum sequence length	150 tokens
Training precision	FP32

Table 10: Model architecture and training hyperparameters

Beyond Single Words: MWE Identification in Bioinformatics Research Articles and Dispersion Profiling Across IMRaD

Jurgi Giraud and Andrew Gargett
School of Languages and Applied Linguistics
The Open University
Milton Keynes, UK
jurgi.giraud@open.ac.uk andrew.gargett@open.ac.uk

Abstract

Multiword Expressions (MWEs) are pervasive in scientific writing, and in specialized domains they include both multiword terminology (e.g., noun compounds) and recurrent academic phrasing. This study profiles MWEs in a large corpus of bioinformatics research articles segmented by IMRaD sections. Building on recent multi-method approaches to scientific MWE identification, we extract MWEs using complementary automated strategies (semantic matching, dependency parsing, controlled vocabularies, and academic formula lists) and compare the resulting inventories by size, form, and IMRaD section distribution. We further quantify cross-document dispersion using document frequency and Gries' DP to distinguish widely reused expressions from items concentrated in a small subset of articles. Results show that bioinformatics MWEs are predominantly short and nominal, but that extraction methods differ in the extent to which they recover discourse and reporting phraseology. Dispersion is strongly long-tailed across sections with most MWEs being document-specific, while a smaller recurrent core aligns with section function and is enriched for conventional templates and standardized multiword terms. Overall, the findings argue for combining complementary identification methods with dispersion profiling to characterize domain "multiwordness" in a principled and section-sensitive way.

1 Introduction and Background

Multiword Expressions (MWEs) refer to a broad class of linguistic forms that span conventional word boundaries, consisting of two or more words that function as a single unit with semantic, syntactic, and/or lexical properties (Constant et al., 2017; Sag et al., 2002). They encompass a heterogeneous set of items including idioms, collocations, phrasal verbs, fixed or semi-fixed phrases, lexicalized compounds, and institutionalized expressions, which appear across languages with varying degrees of compositionality and predictability (Villavicencio et al., 2005; Constant et al., 2017; Masini, 2019). MWEs are not just a feature of general language: they are central to specialized discourse and scientific writing, with prior work applying MWE extraction in scientific corpora (Kim et al., 2018; Premasiri et al., 2023; Alves et al., 2024; Bagdasarov and Teich, 2024; Alves et al., 2025; Florescu and Ohniwa, 2025). Scientific research articles, in particular, make extensive use of dense nominal style and increasingly complex noun phrases to pack information efficiently (Biber and Gray, 2016; Degaetano-Ortlieb and Teich, 2018; Bagdasarov and Teich, 2024). MWEs often correspond to key domain concepts (e.g., *gene expression profile*) or conventional academic phrases (e.g., *in this study*). This tendency is especially salient in bioinformatics, an interdisciplinary field that sits at the crossroads of biology, biomedical science, and computer science (Nakaya, 2021) with a rapidly evolving terminology.

Research articles also exhibit systematic variation across sections. The IMRaD convention (Introduction–Methods–Results–Discussion), now widely adopted in the biomedical sciences, reflects distinct communicative purposes and is associated with measurable differences in rhetorical and lexico-grammatical choices (Sollaci and Pereira, 2004; Wu, 2011). A section-aware perspective is therefore valuable for locating where domain-specific terminology concentrates (e.g., procedural labels in Methods) and where formulaic discourse markers cluster (e.g., result-reporting patterns) (Sollaci and Pereira, 2004; Wu, 2011; Hyland, 2012).

Beyond identifying MWEs, it is also important to determine whether they constitute broadly shared phraseological resources or remain localized to specific papers. Dispersion profiling provides this functional perspective by quantifying how evenly an expression is distributed across doc-

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 86–95

uments within a corpus or subcorpus (Gries, 2021). In corpus linguistics, dispersion measures complement frequency by distinguishing items that are frequent because they recur widely from items that are frequent but concentrated in a small subset of texts (Gries, 2021). This distinction is particularly relevant for scientific MWEs as some are productive and topic- or dataset-contingent (e.g., novel noun compounds), while others behave like reusable templates (e.g., reporting or framing formulas), and these differences are expected to vary by IMRaD section.

To support section-aware analyses of bioinformatics phraseology, we compiled BIOMONO_EN, a large English corpus of open-access bioinformatics research articles. Using complementary MWE identification strategies (lexicon-based semantic tagging, dependency-based extraction, ontology matching, and academic formula lists), we characterize bioinformatics MWEs across IMRaD sections and profile their dispersion using document frequency and Gries' DP (Gries, 2021). We show that (i) nominal multiword terminology dominates, with noun compounds accounting for the majority of dependency-derived MWEs; (ii) formulaic academic expressions are widely attested and contribute to the most evenly dispersed MWEs; and (iii) dispersion is strongly long-tailed, with most MWEs occurring in single documents while a smaller recurrent core is shaped by section function and enriched for multi-source overlaps and standardized terminology.

2 Methods and Materials

2.1 Corpus Compilation and IMRaD Subcorpora

To study bioinformatics MWEs in natural text, we compiled a large in-domain corpus of English research articles, named BIOMONO_EN. We leveraged the ALLOFPLOS[1] collection, a repository of ∼200k open-access articles from PLOS journals (Seiver et al., 2018). We filtered this collection by subject area metadata to retrieve articles classified under "bioinformatics". This yielded 4,707 full-text articles from journals such as *PLOS One* and *PLOS Computational Biology*, totaling approximately 24,234,000 words.

Each article was partitioned into its main sections according to the IMRaD structure (following each article's XML section tags). We extracted six section-based subcorpora: Abstracts, Introductions, Methods (including Materials), Results, Discussions, and Conclusions. This stratification enables analysis of MWEs in different communicative contexts.

Table 1 summarizes the size of each subcorpus in number of words.

Section	Word count
Abstracts	1,047,099
Introductions	3,289,196
Methods	5,770,217
Results	7,685,586
Discussions	5,941,263
Conclusions	500,764

Table 1: BIOMONO_EN corpus statistics: total words per IMRaD section.

2.2 MWE Identification Techniques

Following multi-method approaches to scientific MWE extraction (Alves et al., 2024), we used complementary automated strategies designed to capture different facets of multiwordness: (i) lexicon-based semantic matching (USAS), (ii) dependency-linked constructions (UD), (iii) controlled-vocabulary terminology (MeSH), and (iv) list-based academic formulas (AFL/ARTES). We treat these outputs as partially overlapping views rather than interchangeable inventories.

Lexicon-based semantic matching (USAS). We used the UCREL Semantic Analysis System (USAS), which supports multiword matching via lexical resources and disambiguation (Piao et al., 2003; Rayson et al., 2004). Tagging and extraction were carried out using `PyMUSAS`[2]

Dependency-based extraction (UD). We applied UD dependency parsing and extracted MWEs as sequences connected by relations commonly associated with multiword constructions, following Alves et al. (2024): `compound`, `compound:prt`, `fixed`, `flat`, and `flat:foreign`. Parsing was performed with `Stanza` (de Marneffe et al., 2021; Qi et al., 2020).

Ontology term matching (MeSH). To isolate standardized domain terminology, we performed string matching against Medical Subject Headings (MeSH)[3]. This provides high-precision matches to

[1] https://plos.org/text-and-data-mining/

[2] https://ucrel.github.io/pymusas/

[3] https://www.nlm.nih.gov/mesh/meshhome.html

controlled-vocabulary terms but does not capture novel terms absent from the ontology.

Academic formula lists (AFL/ARTES). To capture conventional academic and scientific phraseology, we matched expressions from the Academic Formulas List (AFL) (Simpson-Vlach and Ellis, 2010) and the ARTES scientific phraseology database (Kübler and Pecman, 2011). List matching estimates coverage of known formulas and complements open-ended extraction approaches. From AFL, we took the "core" list of 207 expressions (frequent in both spoken and written academia) and the "written" list of 200 expressions (specific to academic writing). From ARTES, we extracted 830 English expressions from the scientific dictionary and 420 from the cross-disciplinary dictionary.

2.3 Dispersion Analysis

To complement frequency-based profiling, we analyzed how evenly MWEs are distributed across documents within each IMRaD subcorpus. For each section, we treated each article section instance (e.g., one abstract, one introduction) as a separate document and computed two dispersion indicators for every attested MWE type. First, we calculated document frequency (DF), i.e., the number of documents in which an MWE occurs at least once, reported both as a count and as a percentage of documents (DF%). Second, we computed Gries' DP (Gries, 2021), a dispersion coefficient that quantifies the deviation of an item's observed distribution across corpus parts from an equal-share baseline. DP approaches 0 when an MWE is distributed relatively evenly across documents and approaches 1 when it is concentrated in a small subset of documents. We report DP alongside occurrences and DF/DF% to distinguish MWEs that are frequent because they recur broadly from those that are frequent but locally concentrated. Dispersion statistics were computed separately per section and stratified by MWE source (UD, USAS, MeSH, formula lists, and their overlaps) to characterize how extraction strategies differ in the degree to which they capture section-general phraseological templates versus document-specific constructions.

3 Results

3.1 MWE Extraction Results

Table 2 summarizes total and unique MWEs/entities extracted by USAS and UD, along with MeSH matches, across sections. Because sections differ in size (Table 1), raw totals partially reflect section length. To control for this, Table 3 reports rates per million words, revealing differences in extraction density that are not visible from raw counts alone. Notably, sections that are largest in raw totals (e.g., Results and Discussions) are not necessarily the highest in per-word MWE yield, underscoring the importance of normalization for section-wise comparison.

USAS method

The USAS method identified substantial inventories of MWEs across sections (Table 2). In raw terms, the largest sections contain the most MWEs. However, the contrast between raw and normalized counts is particularly informative. Although Results has the largest raw USAS total (Table 2), Methods has the highest USAS density once normalized (75,767 total USAS MWEs per million words in Methods vs. 57,119 in Results; Table 3). Similarly, Abstracts show the highest rate of unique USAS MWEs per million words (41,766), indicating comparatively high type diversity per unit of text despite being much smaller in raw size.

Figure 1a displays MWE lengths across BIOMONO_EN sections. The majority of USAS-extracted MWEs are two-word MWEs, with 59.94% for abstracts, 54.34% for introductions, 60.56% for methods, 61.15% for results, 56.31% for discussions, and 58.04% for conclusions. Three-word MWEs also represent a large proportion of extracted MWEs above 30% for each section, notably 37.77% for introductions, 35.66% for discussions, and 33.07% for conclusions.

We also quantified the prevalence of nominal MWEs in the USAS-derived inventories (Table 4). Nominal MWEs constitute a majority of unique USAS MWEs in all sections, but the proportion varies with Abstracts (86.39%), Methods (84.01%), and Discussions (84.08%) showing high nominal shares, whereas Results is notably lower (58.14%). This variability suggests that lexicon-recognized MWEs in Results include a larger proportion of non-nominal phraseology.

UD method

The dependency-based UD method produced a larger inventory of MWEs than the USAS method in all sections (Table 2). This is consistent with the productivity of compound formation and the broad coverage of dependency relations used to encode multiword constructions. Figure 1b shows that, as

Section	USAS MWEs		UD MWEs		MeSH entities	
	Total	Unique	Total	Unique	Total	Unique
Abstracts	67,448	43,733	134,817	77,006	86,669	5,226
Introductions	197,897	107,321	370,386	172,427	264,580	8,310
Methods	437,194	185,414	879,530	350,688	346,715	6,655
Results	438,990	180,405	1,310,774	372,933	421,019	7,239
Discussions	397,563	191,049	1,353,212	285,419	374,720	8,239
Conclusions	29,810	18,797	53,760	31,450	31,433	2,617

Table 2: Raw counts of total (instances) and unique MWEs/entities by section.

Section	USAS (per million words)		UD (per million words)		MeSH (per million words)	
	Total	Unique	Total	Unique	Total	Unique
Abstracts	64,414	41,766	128,753	73,542	82,771	4,991
Introductions	60,166	32,628	112,607	52,422	80,439	2,526
Methods	75,767	32,133	152,426	60,776	60,087	1,153
Results	57,119	23,473	170,550	48,524	54,780	942
Discussions	66,916	32,156	227,765	48,040	63,071	1,387
Conclusions	59,529	37,537	107,356	62,804	62,770	5,226

Table 3: Length-normalized counts (per million words) of total (instances) and unique MWEs/entities by section, computed from Tables 1 and 2.

Section		Nominal USAS MWEs
Abstracts	n=	37,782
	%	86.39%
Introductions	n=	72,494
	%	67.55%
Methods	n=	155,768
	%	84.01%
Results	n=	104,895
	%	58.14%
Discussions	n=	160,630
	%	84.08%
Conclusions	n=	14,859
	%	79.05%

Table 4: Number and percentage of unique nominal MWEs as extracted by the USAS method.

with USAS, two-word sequences dominate the UD-derived inventories, reflecting the prominence of binary compounds and short fixed constructions.

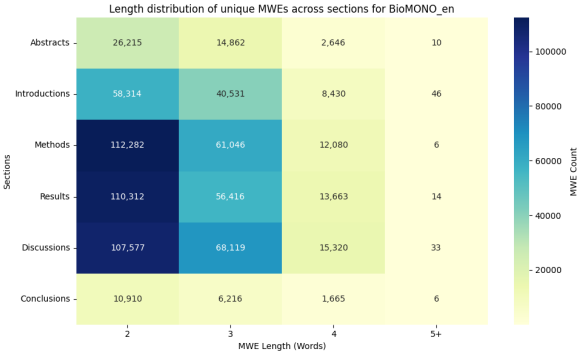
Normalization also changes how section differences are interpreted for UD MWEs. In raw terms, Discussions and Results dominate because they are long sections (Table 2). However, per-million rates show that Discussions is the densest site of UD MWEs (227,765 per million words), exceeding Results (170,550) and Methods (152,426) (Table 3).

To examine what kinds of dependency-linked constructions dominate, Table 5 reports the distribution of unique UD MWEs by relation category. More than 90% of unique MWEs extracted across BIOMONO_EN sections belong to the `compound` category. `flat` is also the second most prominent category, with for instance 8.07% of extracted MWEs from the Methods section belonging to that category, and 8.04% for the Results section. Methods and Results are the only two sections containing `flat:foreign` MWEs, although in very small number.

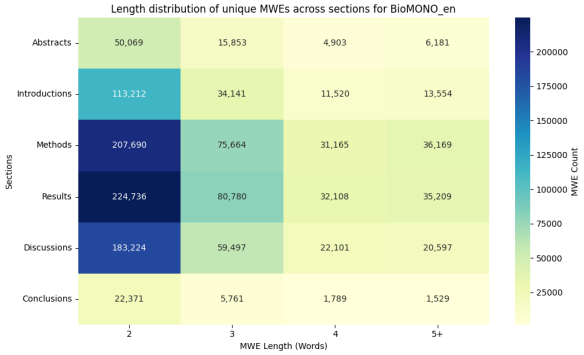
MeSH method

MeSH matching yields fewer unique items than the open-ended UD and USAS inventories (Table 2), as expected for a controlled vocabulary. However, the matches are domain-relevant by construction and provide a high-precision view of standardized terminology. For MeSH, normalized rates highlight a different profile from raw totals. Abstracts and Introductions show the highest MeSH token density (82,771 and 80,439 per million words, respectively), even though they do not contain the most raw MeSH matches (Table 3). Conversely, Methods and Results exhibit much lower unique MeSH rates per million words (1,153 and 942), suggesting heavier repetition of a narrower standardized vocabulary within those sections, whereas Conclusions show a comparatively high unique MeSH density given their short length (Table 3).

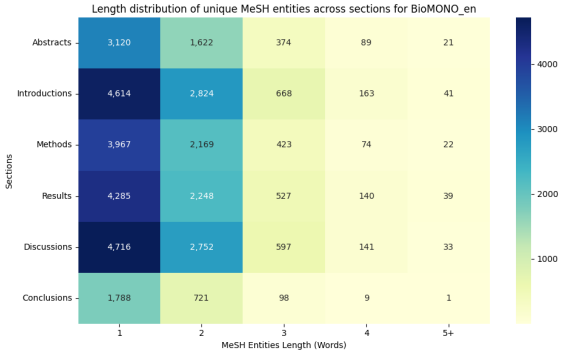
MeSH entities also vary in length from single-word entities to MWEs up to five-word long and more, as seen in Figure 1c. Single-word entities are the majority in every section: 59.70% of unique MeSH entities in Abstracts are single-word terms, 55.52% in Introductions, 59.61% in Methods, 59.19% in Results, 57.24% in Discussions, and 68.32% in Conclusions. The remaining entities are MWEs, among which two-word terms are most prevalent, and longer terms occur up to five words and beyond. This pattern indicates that controlled-



(a) Distributions of unique USAS MWEs found across BIOMONO_EN sections by length (in words).



(b) Distributions of unique UD MWEs found across BIOMONO_EN sections by length (in words).



(c) Distributions of unique MeSH entities found across BIOMONO_EN sections by length (in words).

Figure 1: Length distribution of MWEs across extraction techniques.

vocabulary terminology in bioinformatics is partly multiword and that a substantial share of standardized concepts may be missed by analyses limited to single-word types.

Academic formulaic expressions method

Table 6 presents the number and percentage of unique MWEs found in each section of the BIOMONO_EN corpus across MWE lists. The AFL lists (core and written) show near-complete coverage across all sections, with high percentages overall (82.5–100%), reaching 100% in several sections. In contrast, the ARTES lists (scientific and cross) show lower coverage, with percentages ranging between approximately 29% and 52%. Notably, the Results and Discussions sections consistently contain the highest proportion of MWEs across all lists, particularly for the ARTES lists.

3.2 Dispersion results

Dispersion is dominated by a pronounced long tail, as seen in Figure 2. The median MWE occurs once and appears in exactly one document in every section (median DF% $\approx$ 0.02–0.06%, depending on section size). Consequently, most MWEs are maximally clustered, with median DP values close to 1 throughout ($\approx$ 0.9990–0.9998). The proportion of MWEs attested in a single document is very high across the board (83.7% in Abstracts, 80.4% in Introductions, 75.4% in Methods, 81.6% in Results, 82.9% in Discussions, and 87.0% in Conclusions), rising to $\geq$89.5% in all sections when considering MWEs occurring in at most two documents. This pattern indicates that the MWE inventory is overwhelmingly driven by low-frequency, document-specific units, with only a small minority recurring across texts.

Type inventories are dominated by UD-only (53.1–62.3%) and USAS-only (25.9–32.8%) MWEs, with a stable UD+USAS overlap (10.9–13.0%). MeSH is rare (0.61–1.99%) and Formulas rarer (0.17–0.63%; Conclusions: 1.50%). Source behavior separates the long tail from the core: UD-only/USAS-only MWEs are the most document-specific (singletons: 81.2–88.5% / 77.1–90.3%), whereas overlap MWEs recur more broadly (singletons: 60.5% in Methods; 71.2% in Results). MeSH units are fewer but less singleton-heavy (53.3–67.2%), and MeSH+UD+USAS shows the

Section		compound	compound:prt	fixed	flat	flat:foreign
Abstracts	n=	73,807	132	102	2,965	-
	%	(95.85%)	(0.17%)	(0.13%)	(3.85%)	
Introductions	n=	162,188	441	327	9,471	-
	%	(94.06%)	(0.26%)	(0.19%)	(5.49%)	
Methods	n=	321,370	650	367	28,291	10
	%	(91.64%)	(0.19%)	(0.10%)	(8.07%)	(0.003%)
Results	n=	341,773	679	397	29,983	1
	%	(91.67%)	(0.18%)	(0.11%)	(8.04%)	(<0.001%)
Discussions	n=	265,856	605	434	18,524	-
	%	(93.15%)	(0.21%)	(0.15%)	(6.49%)	
Conclusions	n=	30,014	137	87	1,212	-
	%	(95.43%)	(0.44%)	(0.28%)	(3.85%)	

Table 5: Frequency and percentage of UD relation categories across the different sections of BIOMONO_EN based on unique UD MWEs.

		Abstracts	Introductions	Methods	Results	Discussions	Conclusions
ARTES	n=	250	343	277	342	397	240
scientific	%	30.86%	42.35%	34.20%	42.22%	49.01%	29.63%
ARTES	n=	123	203	159	183	216	125
cross	%	29.71%	49.03%	38.41%	44.20%	52.17%	30.20%
AFL	n=	189	205	200	204	205	189
core	%	91.30%	99.03%	96.62%	98.56%	99.03%	91.30%
AFL	n=	165	199	195	200	200	189
written	%	82.50%	99.50%	97.50%	100%	100%	94.50%

Table 6: Number of unique MWEs found in BIOMONO_EN sections across MWE lists. Percentages are calculated against the total number of unique MWEs in each list.

strongest terminological stability (singletons ∼20–33% in Abstracts–Discussions; 47% in Conclusions). Finally, Formulas behave most "core-like" (singletons: 8.9–17.8%) and are over-represented among the most evenly dispersed MWEs (share in the 500 lowest-DP MWEs: 28.0% Abstracts, 40.0% Introductions, 13.2% Methods, 24.4% Results, 37.2% Discussions, 46.4% Conclusions).

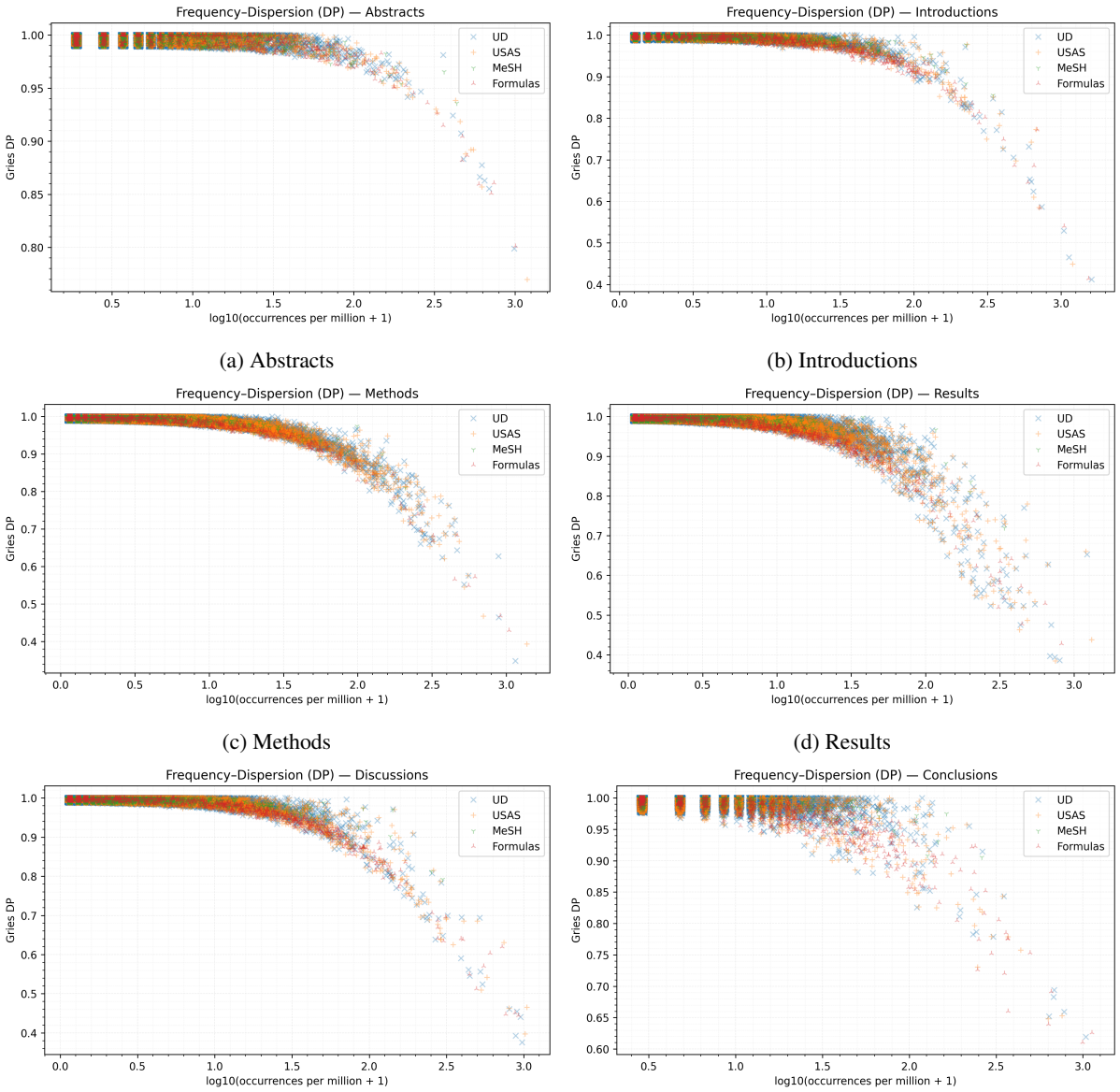
Overall, dispersion reflects a two-layered structure with a large extractor-driven long tail (UD/USAS-only) and a smaller, section-shaped recurrent core enriched for Formulas, overlap MWEs, and MeSH-overlap terminology.

4 Discussion

Across methods, our results highlight the centrality of multiwordness in bioinformatics discourse, but they also show that "MWE dominance" depends on the extraction method. The UD-based inventories are overwhelmingly compound-dominated (Table 5), confirming that noun-compound formation is a primary structural resource for expressing domain concepts succinctly. This is consistent with long-standing accounts of scientific prose as a "compressed code" that favors dense noun phrase packaging over more clausal, elaborated alternatives (Biber and Gray, 2016). This pattern also closely aligns with recent computational evidence that (i) compounds constitute the dominant UD MWE class in scientific writing overall and exhibit a clear increase over time (cf. compounds at 80.2% of UD MWEs; Alves et al., 2024), and (ii) biomedical abstracts are likewise strongly compound-heavy in UD-based inventories, with Bagdasarov and Teich (2024) reporting compound shares above 90%. For Natural Language Processing (NLP), this underscores that a large portion of domain "terminology" is not a closed list but a productive constructional space.

The lexicon-based USAS inventories also contain a large proportion of nominal MWEs in most sections (Table 4), but the proportion is notably lower in Results. This divergence is informative because lexicon-based approaches recover a broader mix of discourse and reporting phraseology, and Results is precisely the section where comparison and evidential framing are most prominent. Practically, this indicates that MWE resources for domain NLP should be section-aware, as the phraseological targets relevant to information extraction or summarization are not uniform across IMRaD.

MeSH matching provides a complementary view of standardized terminology. While most unique MeSH entities are single-word terms, a substantial fraction are multiword (Figure 1c), demonstrat-



(a) Abstracts

(b) Introductions

(c) Methods

(d) Results

(e) Discussions

(f) Conclusions

Figure 2: Frequency–dispersion profiles (occurrences vs. Gries' DP) of MWEs in BIOMONO_EN by IMRaD section. Each point corresponds to one MWE type; DP values closer to 0 indicate more even dispersion across documents, while values closer to 1 indicate stronger clustering. Points are colored by MWE source (UD, USAS, MeSH, and formulas).

ing that controlled-vocabulary terminology is not reducible to single-word naming. The elevated unique MeSH rate in Abstracts and Conclusions (Table 3) further suggests that summary sections foreground canonical entity naming, even when the running text is shorter. For downstream applications (e.g., normalization, retrieval, and MT terminology control), these multiword entities constitute high-value targets that are easy to miss under unigram-centric preprocessing.

Finally, list-based formula coverage indicates that bioinformatics writing draws broadly on general academic formulas (AFL) and on a sizeable subset of scientific phraseology documented in ARTES (Table 6). This signals substantial transferability of general scientific phraseology resources to bioinformatics, while dispersion results clarify where these templates function as shared scaffolding versus local phrasing.

Dispersion profiling adds a functional perspective on these inventories. Across IMRaD, MWE types exhibit a strongly long-tailed distribution: most occur in a single document, while a much smaller set forms a recurrent "core" whose composition shifts by section. This pattern is expected when dispersion is assessed over many documents and is captured by Gries' DP, which explicitly distinguishes frequency from distributional evenness (Gries, 2021). Importantly, the long tail should not be interpreted as a lack of phraseological structure: many UD MWEs are *productively constructed* (e.g., novel or dataset-contingent compounds) rather than retrieved as fixed strings, increasing type counts while limiting cross-document recurrence (Biber and Gray, 2016). For MWE research, this reinforces the need to separate productive constructions from reusable templates. For NLP, it suggests that robust domain handling requires both (i) mechanisms for generalizing over productive compounds and (ii) explicit modeling of recurrent templates that shape section-level discourse.

Source-specific dispersion further clarifies what constitutes the recurrent backbone. UD-only and USAS-only MWEs contribute most of the document-specific tail, whereas overlap MWEs (UD+USAS) recur more broadly, suggesting that multi-method confirmation captures sequences that are simultaneously structurally cohesive and functionally salient. The strongest "core-like" behavior is observed for list-derived formulas, which are rare as types yet disproportionately represented among the most evenly dispersed MWEs. This aligns with corpus work showing that recurrent MWEs and formulaic sequences function as register-specific building blocks in academic discourse (Biber et al., 2004; Hyland, 2008; Wray, 2002). MeSH matching complements this picture by isolating a compact set of standardized multiword terms that recur across documents when they are also recoverable by general extraction, consistent with the stabilizing role of controlled vocabularies in scientific naming.

5 Conclusion

Using complementary MWE identification strategies and dispersion profiling, this study maps bioinformatics "multiwordness" across IMRaD sections. MWEs are predominantly short and nominal, reflecting compound-heavy phrasal compression in scientific prose (Biber and Gray, 2016; Alves et al., 2024), while other extraction methods recover additional reporting and procedural templates. Dispersion is strongly long-tailed: most MWEs are document-specific, but a smaller recurrent core aligns with section function and is enriched for conventional templates and standardized multiword terminology. For NLP and MWE research, the main implication is that domain phraseology is best operationalized as a two-layer system (productive constructions plus reusable templates) and that combining multi-method identification with dispersion analysis provides a principled way to prioritize MWEs for domain-adapted preprocessing and downstream applications.

6 Limitations and Future Work

This study relies on automated MWE identification and therefore inherits method-specific biases, such as parsing sensitivity, tokenization, disambiguation, and vocabulary coverage.

A priority next step is to build a small, section-stratified manually verified subset to quantify boundary errors, false positives/negatives, and overlap reliability, enabling precision-oriented reporting in addition to coverage. Beyond validation, two extensions are particularly relevant: (i) multilingual replication to test whether the same section-conditioned multiword patterns hold under different morphosyntactic systems, and (ii) downstream evaluation to assess whether MWE-aware resources improve domain tasks such as terminology normalization, information extraction, retrieval, or domain-specific Machine Translation.

Acknowledgments

This work was supported by the Open-Oxford-Cambridge Arts and Humanities Research Council (AHRC) Doctoral Training Partnership (OOC-DTP), project reference 2739531.

Data and Code Availability

Data and code are available at https://github.com/jurgigi/BioMONO

References

Diego Alves, Stefan Fischer, Stefania Degaetano-Ortlieb, and Elke Teich. 2024. Multi-word Expressions in English Scientific Writing. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 67–76, St. Julians, Malta. Association for Computational Linguistics.

Diego Alves, Stefan Fischer, and Elke Teich. 2025. Syntagmatic Productivity of MWEs in Scientific English. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 1–6, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Sergei Bagdasarov and Elke Teich. 2024. Multi-word expressions in biomedical abstracts and their plain English adaptations. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 483–488, Miami, USA. Association for Computational Linguistics.

Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. If you look at . . . : Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics*, 25(3):371–405.

Douglas Biber and Bethany Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Studies in English Language. Cambridge University Press, Cambridge.

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

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308. Place: Cambridge, MA Publisher: MIT Press.

Stefania Degaetano-Ortlieb and Elke Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 22–33, Santa Fe, New Mexico. Association for Computational Linguistics.

Cosmin Mihail Florescu and Ryosuke L. Ohniwa. 2025. On the creation of a corpus-derived medical multiword term list. *Information*, 16(2):118.

Stefan Th. Gries. 2021. Analyzing dispersion. In Magali Paquot and Stefan Th. Gries, editors, *A Practical Handbook of Corpus Linguistics*, pages 99–118. Springer, Cham.

Ken Hyland. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1):4–21.

Ken Hyland. 2012. Bundles in Academic Discourse. *Annual Review of Applied Linguistics*, 32:150–169.

Sun Kim, Lana Yeganova, Donald C. Comeau, W. John Wilbur, and Zhiyong Lu. 2018. Pubmed phrases, an open set of coherent phrases for searching biomedical literature. *Scientific Data*, 5:180104.

Natalie Kübler and Mojca Pecman. 2011. ARTES: an online lexical database for research and teaching in specialized translation and communication. In *ESSLLI 2011, International Workshop on Lexical Resources (WoLeR)*, Ljubljana, Slovenia.

Francesca Masini. 2019. Multi-Word Expressions and Morphology. In *Oxford Research Encyclopedia of Linguistics*.

Helder I. Nakaya. 2021. *Bioinformatics*. Exon Publications, Australia.

Scott S. L. Piao, Paul Rayson, Dawn Archer, Andrew Wilson, and Tony McEnery. 2003. Extracting Multiword Expressions with A Semantic Tagger. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 49–56, Sapporo, Japan. Association for Computational Linguistics.

Damith Premasiri, Amal Haddad Haddad, Tharindu Ranasinghe, and Ruslan Mitkov. 2023. Deep learning methods for identification of multiword flower and plant names. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 879–887, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Paul Rayson, Dawn Archer, and Scott Piao. 2004. The UCREL semantic analysis system. In *Proceedings of the Beyond Named Entity Recognition Workshop*, Lisbon, Portugal.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Elizabeth Seiver, M Pacer, and Sebastian Bassi. 2018. Text and data mining scientific articles with allofplos. In *Proceedings of the 17th Python in Science Conference*, pages 61 – 64.

Rita Simpson-Vlach and Nick C. Ellis. 2010. An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, 31(4):487–512.

Luciana B. Sollaci and Mauricio G. Pereira. 2004. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *Journal of the Medical Library Association*, 92(3):364–371.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. Editorial: Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Comput. Speech Lang.*, 19(4):365–377.

Alison Wray. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press, Cambridge.

Jianguo Wu. 2011. Improving the writing of research papers: IMRAD and beyond. *Landscape Ecology*, 26(10):1345–1349.

The Lock, Stock, and Barrel of Marathi Multiwords

Aakanksha Padhye, Ashwini Vaidya
Indian Institute of Technology Delhi
{aakanksha.p, avaidya} @hss.iitd.ac.in

Abstract

Multiword expressions are an important area of study in linguistics and natural language processing as they represent combination of words that function as a single unit, and display properties that cannot be predicated fully from their individual components. This paper describes annotated corpora of about 3000 multiword expressions across syntactic categories in Marathi. This is the first exhaustive resource for Marathi which includes both verbal and non-verbal multiwords. In order to develop the guidelines for annotation, we have used the existing literature on the identification and classification of these expressions. Following the PARSEME 2.0 guidelines, we discuss the categories of multiwords and their behaviour in the corpus. Throughout the annotation process, we encounter variability in compositionality and syntactic realization and discuss our design decisions during annotation. Such a dataset will further our understanding of how grammatical structure can be integrated with lexically stored multiword units in Marathi.

1 Introduction

Multiword expressions (MWEs) are a pervasive and heterogeneous class of linguistic units that are central to research in linguistics and natural language processing. Linguistically, these constructions challenge the notions of compositionality, argument structure, and the division of labor between syntax and lexicon. Efficient handling of MWEs would prove beneficial for natural language processing tasks like machine translation (Constant et al., 2017), semantic processing (Korkontzelos, 2010), information extraction, word sense disambiguation (Singh et al., 2016); and psycholinguistic studies like MWE representation and processing (Wittenberg and Piñango, 2011; Nenonen et al., 2002), etc.

Previous attempts to annotate Marathi MWEs restrict themselves to compound nouns and light verb constructions alone (Singh et al., 2016). In this paper, we describe our effort at creation of a more comprehensive database of MWEs in Marathi under the PARSEME project (Savary et al., submitted).[1] The annotation tags, annotation platform, and annotation schema strictly adhere to the guidelines of the PARSEME project (Savary et al., submitted). As a result, we do not revisit these details here. Instead, we primarily focus on reporting the methodological decisions adopted during the process of annotation, and the linguistic and empirical challenges encountered during the process.

2 Corpora and Annotation

The corpora needs to be representative and balanced (Pustejovsky and Stubbs, 2012) in order to capture the entire range of MWEs in Marathi. This is achieved by carefully determining the genre of the data. Ozarkar (2014) observes that Marathi light verb constructions may have originated in informal contexts. Keeping this in mind, we have chosen Marathi UD Treebank (Ravishankar, 2017), and Anuvaad (Tiedemann, 2012) corpora, primarily comprising stories from Wikisource and the lifestyle genre, respectively. Additionally, we have web-crawled children's stories that are randomly sampled from different sources. Table 1 reports the number of tokens in each type of corpus in this dataset.

Marathi UD Treebank (Ravishankar, 2017) is already annotated with gold standard POS tags, syntactic structures, and semantic relations in ConLL-U format. The remaining two corpora are raw. For the Anuvaad corpus and the children's stories, we used UDPipe for parsing and tagging the

[1]The data will be released under the PARSEME 2.0 (Savary et al., submitted) initiative.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 96–102

Corpora	Tokens
Marathi UD Treebank	3849
Anuvaad Corpus	27956
Children's Stories	4287

Table 1: Tokens in corpora chosen

raw Marathi text (Straka and Straková, 2017). We found that UDPipe (Straka and Straková, 2017) for Marathi is not very accurate, and we find errors for POS tagging and sentence segmentation. The Anuvaad (Tiedemann, 2012) corpus, and children's stories contain only 'silver standard' tags and sentence segments. For this present work, we have prioritized the annotation of MWEs by not letting the errors influence the annotations.

The annotation task is carried out by a single annotator. Hence, we are unable to calculate inter-annotator agreement.

The upcoming sections discuss all possible MWEs in Marathi. Based on (Savary et al., submitted) guidelines, we broadly classify them into two categories: verbal MWEs and non-verbal MWEs. The latter is a broader class comprising nominal, adjectival, and adverbial MWEs.

3 Verbal MWEs

Structurally, verbal MWEs in Marathi are broadly classified into verb-verb and preverb-verb constructions in the literature. PARSEME 2.0 (Savary et al., submitted) refers to these constructions as multi-verb constructions, and light verb constructions respectively. This section presents the categories incorporated by these constructions, their identification along with the semantics they render.

3.1 Multi-Verb Constructions

PARSEME 2.0 (Savary et al., submitted) identifies multi-verb constructions (MVCs) as a sequence of two verbs functioning as a single predicate, having the same subject. referring to a single event, and denoting a single tense, aspect and polarity value. These characteristics are identified using Ozarkar (2014)'s classification of multi-verbs. The constructions below are annotated as MVCs in the corpus following her classification:

1. **Complex predicates (CPs)**: monoclausal and monoeventual sequences like basun rahɳe 'sit stay'

2. **Factor verbs**: expressions stored in the mental lexicon as a single unit or a set formula. Example: nigʰun ʤaɳe | lit. 'emerge go' ('depart')

3. **Manner-adverbial CPs**: sequences like ɖ̪ʰawət̪ jeɳe 'run come'. The author notes that Marathi, unlike Hindi, does not give a serial reading for such constructions.

The MVCs usually occur in two forms in Marathi. Firstly, we have verb-verb sequences conjoined by the conjunctive particle (-un). Secondly, there are verb-verb sequences conjoined by an imperfective marker (-ət̪).

Ozarkar (2014) identifies a list of verbs conjoined by the conjunctive particle (-un). It includes ʤaɳe 'go', jeɳe 'come', ɖ̪eɳe 'give', gʰeɳe 'take', ʈakɳe 'throw', ʈʰewɳe 'keep', bəsɳe 'sit', kaɖʰɳe 'draw out', and rahɳe 'stay' as light verbs. Pardeshi et al. (2006) add gʰalɳe 'put', pəɖɳe 'fall', and aɳɳe 'bring' to the list.

On the other hand, for the imperfective marker (-ət̪), Ozarkar (2014) observes that light verbs like bəsɳe 'sit', suʈɳe 'be released', tsalɳe 'walk', ʤaɳe 'go', jeɳe 'come', and rahɳe 'stay' can be found. Kume (2011) mentions that certain perception verbs like pahɳe 'see' also function as light verbs.

We use the verb list for MVCs mentioned in the literature, and empirically investigate the occurrences of these verbs in the corpora. We have identified thirteen verbs functioning as light in the verb-verb sequence. Figure 1 shows the verbs with the highest frequencies in the corpora. These verbs are followed by ɖ̪eɳe 'give', ʈakɳe 'throw', kaɖʰɳe 'draw out', bəsɳe 'sit', tsalɳe 'walk' and pahɳe 'see'. Light verbs with the lowest frequencies are suʈɳe 'be released' and aɳɳe 'bring'.

Verbal reduplication is also attested in the corpus in a few rare examples. Instances like tsaləʈ̪ tsaləʈ̪ 'walk walk' are considered as MVCs as the entire verb is reduplicated to form a verb-verb sequence.

It should be noted that not all verbs mentioned above function as light verbs in all contexts when they appear as a second verb in the verb-verb sequence. Verbs like ʤaɳe 'go' and jeɳe 'come' can also function as passive markers. The passive constructions are not MVCs. Similarly, modals and auxiliaries do not constitute MVCs. Accordingly, passives, auxiliaries, modals, permissives,

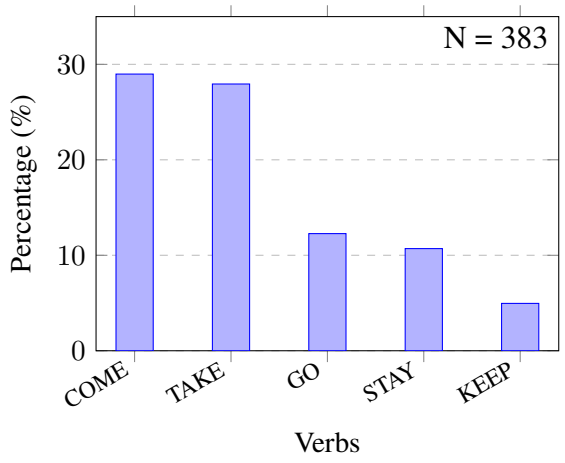


Figure 1: Distribution of verbs functioning as light in the MVC class. The figure illustrates top five light verbs with the highest frequencies amongst the thirteen light verbs identified.

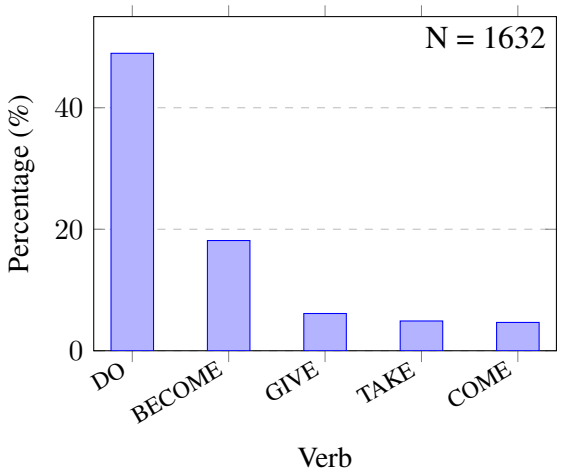


Figure 2: Distribution of verbs functioning as light in the LVC.full class. The figure illustrates top five light verbs with the highest frequencies amongst the twenty-three light verbs identified.

other such seemingly similar structures are carefully separated from the actual MVCs during the annotation process.

3.2 Light Verb Constructions

PARSEME 2.0 (Savary et al., submitted) defines light verb constructions (LVCs) as expressions formed by a verb, and a wide range of preverbs like nouns, adjectives, prepositions, etc (Family, 2014). Similar to Hindi, Marathi has nouns and adjectives as preverbs. There is also evidence of adverbs as preverbs in the database. In this subsection, we talk about the characteristics of these preverbs, and also discuss their identification strategies.

Literature on LVCs in Marathi is rather sparse, though Kulkarni (2019) and Hook and Pardeshi (2009) touch upon marɳe 'hit' and kʰaɳe 'eat' expressions briefly. Family (2014)'s identification of Persian light verbs under this category can be extended to Marathi for the purposes of annotation. These light verbs include kərɳe 'do', pəɖɳe 'fall', hoɳe or bənɳe 'become', miɭɳe 'get', aɳɳe 'bring', ʤaɳe 'go', and jeɳe 'come'. Certain perception verbs like pahɳe 'see' also function as light verbs (Kume, 2011). Accordingly, we have considered synonyms of 'see' like d̪isɳe 'see' as light verbs. Additionally, we have annotated verbs like lagɳe 'be attached', waʈɳe 'seem', along with some of the verbs recognized as light verbs in the MVC category like kaɖʰɳe 'draw out'.

We refer to this list to annotate the verbs, and to examine their empirical distribution. We have identified twenty-three such verbs. Figure 2 shows the verbs with the highest frequencies in the data. Verbs like kʰaɳe 'eat' and suʈɳe 'be released' have the lowest frequencies. We encounter certain verbs like suʈɳe 'be released' and soɖɳe 'be release-cause', bənɳe 'make' and bənəwɳe 'make-cause', wherein the second verb is in the causative form of the first verb. The causative light verb has been annotated as LVC.cause, following PARSEME 2.0 (Savary et al., submitted) guidelines and the non-causative form is LVC.full. LVC.cause are very few in number, and only attested with four verbs like soɖɳe 'be release-cause', bənəwɳe 'make-cause', miɭəwɳe 'get-cause', and d̪akʰəwɳe 'see-cause'. Verbs like bənəwɳe 'make-cause' and miɭəwɳe 'get-cause' have the highest frequencies while soɖɳe 'be release-cause' and d̪akʰəwɳe 'see-cause', the lowest.

Bonial (2021) notes that the event semantics of LVCs stems from the nouns (and other preverbs), rather than the verbs alone. These nouns (and other preverbs) also distinguish these verbs from their full verb and light verb usages. Verbs in their full verb usages denote their literal, canonical sense, while the light verbs constitute the non-literal senses. When the nominal preverbs are abstract denoting events or states the verb is light, else full. The corpora show that while this holds true for most of the light verbs there are certain light verbs that have no such selectional restrictions. Light verbs like hoɳe 'become', ʤaɳe 'go', bənɳe 'make', kərɳe 'do', also as noted by Fam-

ily (2014), select nominal preverbs that are not abstract.

Based on this understanding, we come up with certain heuristics to distinguish the verbs into their light and full versions. The rephrasing test states that such constructions can be rephrased with one-word predicate. The Marathi corpora have examples like uttər ɖeɳe 'answer give' which can be paraphrased into corresponding single verb - uttərɳe 'answer'. But this test is not valid for most of the expressions as they cannot be mapped to their corresponding single verb forms. Thus, we come up with the following diagnostics beyond the rephrasing test to identify these monoclausal constructions:

1. **Omission of the preverb**: In the preverb-verb sequence, the preverb cannot be omitted. Example: ramne kʰoli swəʧʧʰə ʈʰewli | lit. 'Ram room clean keep' ('Ram kept the room clean') cannot be rewritten as *ramne kʰoli ʈʰewli 'Ram room kept'

2. **Co-ordination**: Event nouns as preverbs cannot be co-ordinated. Example: *ʈjane bʰeʈ aɳi məɖət ɖili 'he visit and help give'

3. **Limited compatibility with light verbs**: Certain nouns functioning as preverbs like bʰaʃəɳ 'speech' allow certain light verbs like kərɳe 'do' or ɖeɳe 'give'.

The LVC category is the most productive as compared to all other MWEs in Marathi across all the corpora that were examined (See Table 2).

3.3 Verbal Idioms

Verbal Idioms (VIDs) are a sequence whose meaning does not arise from the meaning of either of the component verbs. Example: aɖʰeweɖʰe gʰeɳe | lit. 'roundabout take' ('to make excuses'). These are relatively fewer in number as compared to MVCs and LVCs in the corpora.

4 Non-verbal MWEs

Non-verbal MWEs consist of a broad category of MWEs based on their syntactic role - nominal (NID), adjectival (AdjID), adverbial (AdvID), and other MWEs with other functional categories. Constant et al. (2017) assert that non-verbal MWEs can be grouped into the following schemes that are non-exhaustive and often overlapping. We follow the grouping to categorize the non-verbal MWEs identified in the Marathi data.

1. **Compounds**: can be further divided into two types: closed and open compounds. Closed compounds like kagaɖpəttrə 'document' are formed by two or more words functioning as a single token, and open compounds like mittrə-məitriɳi 'friends' are formed from lexemes separated by spaces or hyphens.

2. **Mutiword term**: a multiword designation of a general concept in a specific subject field. Example: uʧʧə rəktɖab 'high blood pressure'

3. **Complex function word**: functional word formed by one or more lexeme. Example: ʤəwəlpas 'nearby, almost'

4. **Idioms**: a group of lexemes whose meaning is established by convention. Example: ʤiw ki praɳ | lit. 'heart or spirit' ('immense love')

Constant et al. (2017) state that NIDs can also be classified into multiword named entities designating real-world entities like persons, organizations, locations, etc. The PARSEME 2.0 (Savary et al., submitted) guidelines do not identify these expressions as MWEs. Therefore, they are not annotated.

Reduplication is a morphophonological phenomenon found in several Indic languages. In this dataset, we look at them from the point of view of multi-word expressions. There are varieties of reduplication in the data. Total reduplication appears in examples like gərəm gərəm 'hot hot', and onomatopoeic expressions like fəɳ-fəɳ 'a kind of sound' while partial reduplication can be seen in expressions like awəɖi niwəɖi 'likes dislikes'. Moreover, semantic reduplication like ʈʰəɳɖəgar | lit. 'cold cold' ('very cold') is found in the datasets. Pandharipande (1998) refers to such expressions as emphatic compounds, as this process intensifies the meaning of the first noun by the use of a synonym.

The semantic properties of the compound expressions are also taken into account during annotation. Pandharipande (1998) refers to expressions like haʈ-paj 'hand-feet' as a superordinate compound, as the two nouns belong to the same semantic class and there is no hierarchical head-embedding between the two. The expression overlaps with the class of a copulative compound. We have annotated these expressions, depending upon the class of the individual components. Accord-

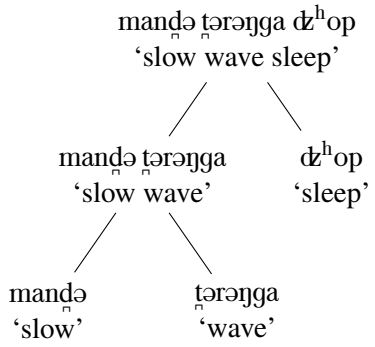


Figure 3: Stacked annotation observed in NID. The tree depicts one of the possible annotations of the NID appearing as a closed compound in the dataset. It suggests that NID has its own internal structure, and needs to be combined in the specific order to render desired meaning.

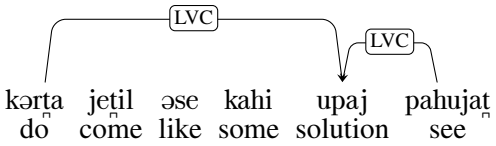


Figure 4: The sentence can be roughly glossed as - Let us look at some solutions that can be done. The figure indicates that noun ‘solution’ is shared by the two verbs - ‘do’ and ‘see’.

ingly, the example mentioned above is labeled as NID.

The author states that in adjective-noun compounds like uʧʧə rəkṭəḍab ‘high blood pressure’, noun is the semantic head, and adjective modifies the noun. The resulting expression is a noun. We have classified such examples as NID. However, noun-adjective expressions like praɳəgʰaṭək ‘life-threatening’ are annotated as AdjIDs because the resulting expression is an adjective. Expressions like lakuɖtoɖ| lit. ‘wood break’ (‘the act of breaking a log of wood’) are noun-verb compounds wherein the derived compound functions as a noun. They are rarely found in the data, and following Pandharipande (1998), they are tagged as NID.

The corpora have certain expressions that have an internal structure, and span over multiple tokens. They result in nested annotations as seen in Figure 3. The current guidelines for annotation do not permit nested annotations for closed compounds. Therefore, we have annotated them as a flat structure.

5 Properties of Marathi MWEs

Marathi MWEs, as observed in the data, possess certain properties that are challenging for their annotation and representation. In this section, we briefly present an overview of their characteristics that provide the rationale underlying annotation decisions.

- **Heterogeneity**: Section 3 and Section 4 in the paper show that the annotated MWEs are not confined to any specific syntactic construction. They are linguistically diverse, and cannot be restricted to only compound nouns and light verbs.

- **Non-compositionality**: Within the entire class of Marathi MWEs, idioms are highly non-compositional. The rest of the categories fall on a continuum between compositionality and non-compositionality.

- **Overlap** (Schneider, 2014): There are some overlapping MWE instances. Figure 4 shows that the noun *solution* overlaps with two distinct verbs, acting as a preverb for both light verbs.

- **Gappy grouping** There are intervening elements between the components of MWEs, making them discontinuous (Constant et al., 2017). Schneider (2014) classifies the ‘gap’ as the argument gap formed by an argument of the predicate, and the modifier gap created due to the intervening adjective, adverb, or determiner.

  (1) ha prajog < aʈʰəwɖjaṭun ek weɭa > kəra
  this experiment < in a week one time > do
  Perform this experiment once a week.

  In (1), the LVC.full in blue is discontinuous, separated by an adverbial modifier.

  Most of the constructions exhibiting this property belong to the LVC class. Marathi corpora reinforce the fact that MVCs are tightly integrated verbal units with restricted internal syntax, while LVCs permit intervening linguistic material (Butt, 1995).

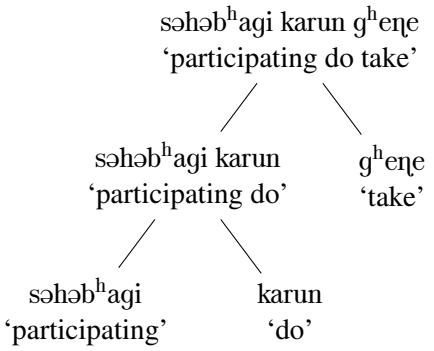


Figure 5: Stacked annotation observed in a verbal MWE. The root of the tree is an MVC. The terminal nodes together form an LVC. This construction is nested within a larger MVC with the light verb 'take'.

- **Stacked annotation**: When an MWE contains another MWE it leads to a hierarchical structure. While this phenomenon is observed in Hindi (Jain and Vaidya, 2024), it is also found in both the verbal and non-verbal MWEs of Marathi. Figure 5 illustrates a verbal MWE within a verbal MWE. Whenever it is possible to preserve the embedded structure, that representation is preferred. However, closed compounds as discussed in Figure 3 are annotated as a flat structure.

6 Summary and Conclusion

We develop an exhaustive knowledge base of Marathi MWEs of all syntactic types - verbal, nominal, adjectival, adverbial, and other functional types. The consistency checks have been performed as per PARSEME 2.0 (Savary et al., submitted) guidelines. The Table 2 mentions the distributional patterns of MWEs in Marathi, revealing both frequent patterns and exceptional cases in the language.

Wherever the precise MWE identificational criteria are not studied, we attempt to propose them based on the empirical evidence from the corpora. However, determining the MWE-hood status of these expressions remains challenging, owing to their structural and semantic properties.

7 Limitations

There are certain limitations affecting the applicability of the resource. First, the annotations are performed by a single annotator. Therefore, inter-annotator agreement cannot be reported. Secondly, the resource depends on the automatic preprocessing done using UDPipe (Straka and Straková, 2017). This has led to errors in tokenization, POS tagging. Though we have prioritized the MWE annotations, we plan to manually review and correct these automatically generated annotations for future release.

References

Claire Bonial. 2021. Précis of take a look at this! form, function, and productivity of english light verb constructions. *Colorado Research in Linguistics*, 25.

Miriam Butt. 1995. *The structure of complex predicates in Urdu*. Center for the Study of Language (CSLI).

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

Neiloufar Family. 2014. *Semantic spaces of Persian light verbs: A constructionist account*, volume 6. Brill.

Peter Hook and Prashant Pardeshi. 2009. A taxonomy of eat expressions in marathi. *Annual Review of South Asian Languages and Linguistics*, pages 41–63.

Kanishka Jain and Ashwini Vaidya. 2024. Revisiting VMWEs in Hindi: Annotating layers of predication. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 98–105, Torino, Italia. ELRA and ICCL.

Ioannis Korkontzelos. 2010. Unsupervised learning of multiword expressions. Unpublished.

Aaditya Kulkarni. 2019. Semantics of hit expressions in marathi.

Yusuke Kume. 2011. On the complement structures and grammaticalization of see as a light verb. *The Electronic Library*, 28:206–221.

Marja Nenonen, Jussi Niemi, and Matti Laine. 2002. Representation and processing of idioms: Evidence from aphasia. *Journal of Neurolinguistics*, 15(1):43–58.

Renuka Ozarkar. 2014. *Structures of Marathi verbs*. Ph.D. thesis, Doctoral dissertation, University of Mumbai.

Rajeshwari V Pandharipande. 1998. *Marathi*. Routledge.

Corpora	MVC	LVC.full	LVC.cause	VID	NID	AdjID	AdvID
UD Treebank (%)	46 (1.19)	120 (3.11)	1 (0.02)	17 (0.44)	107 (2.77)	12 (0.31)	15 (0.38)
Anuvaad Corpus (%)	243 (0.86)	1279 (4.57)	2 (0.007)	2 (0.007)	868 (3.10)	55 (0.19)	62 (0.22)
Children's Stories (%)	94 (2.19)	233 (5.43)	5 (0.11)	3 (0.02)	24 (0.55)	11 (0.25)	13 (0.30)
Total (%)	**383**(1.06)	**1632** (4.52)	**8** (0.02)	**22** (0.06)	**999** (2.76)	**78** (0.21)	**90** (0.24)

Table 2: Distribution of MWEs in Marathi. The table presents the entire landscape of MWEs identified and annotated. PARSEME 2.0 (Savary et al., submitted) identifies other constructions like inherently reflexive verbs (IRV), inherently adpositional verbs (IAV), etc. They are not attested in Marathi data.

Prashant Pardeshi, Peter E. Hook, and Sung-Yeo Chung. 2006. In search of the origins of compound verbs in marathi. Handout of the presentation made at SALA 26, CIIL, Mysore.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.".

Vinit Ravishankar. 2017. A universal dependencies treebank for marathi. In *Proceedings of the 16th international workshop on treebanks and linguistic theories*, pages 190–200.

Agata Savary, Manon Scholivet, Carlos Ramisch, Takuya Nakamura, Eric Bilinski, Sara Stymne, Voula Giouli, Stella Markantonatou, Vasile Păiş, Maria Mitrofan, Louis Estève, Bruno Guillaume, Verginica Barbu Mititelu, Jaka Čibej, Roberto A. Díaz Hernández, Victoria Fendel, Polona Gantar, Olha Kanishcheva, Cvetana Krstev, and 9 others. submitted. PARSEME 2.0: Multilingual corpus of multiword expressions. Submitted to LREC 2026.

Nathan Schneider. 2014. *Lexical Semantic Analysis in Natural Language Text*. Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

Dhirendra Singh, Sudha Bhingardive, and Pushpak Bhattacharyya. 2016. Multiword expressions dataset for indian languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2331–2335.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, pages 88–99.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Eva Wittenberg and Maria Piñango. 2011. Processing light verb constructions. *The Mental Lexicon*, 6.

An Idiom Benchmark for Turkish

Ebru Çavuşoğlu
Translation and Intercultural Studies
Samsun University
ebru.cavusoglu@samsun.edu.tr

Çağrı Çöltekin
Department of Linguistics
University of Tübingen
cagri.coeltekin@uni-tuebingen.de

Abstract

Despite recent significant advances, idioms, like other forms of figurative language, present a challenge to natural language processing (NLP). Benchmark corpora are essential for improving the current models on understanding idioms. However, such corpora are only available for a limited set of languages. In this paper, we introduce our ongoing work on a benchmark corpus of Turkish idioms. Our corpus is structured for testing both idiom recognition and idiom understanding. The corpus currently consists of 200 instances with sentences including idiomatic use, their literal paraphrases, similar sentences with no entailment, and non-idiomatic use of the idiomatic expressions when possible. We describe the methodology used to create the corpus, as well as initial experiments with a selection of LLMs.

1 Introduction

Idioms are multi-word expressions (MWEs) with a conventionalized interpretation. The meanings of idioms cannot be inferred from compositionality from the individual words. The correct interpretation of idioms requires familiarity with the idiom as a conventionalized unit of meaning within the particular language. Furthermore, many idiomatic expressions can also be used literally (Savary et al., 2019), leading to a possible ambiguity that has to be resolved based on the context. Similar to the other forms of figurative expression, like metaphors, proverbs, and irony, idiom understanding necessitates cultural awareness and pragmatic reasoning beyond compositional semantics because of their strong dependence on broader linguistic and non-linguistic context. As a result, idioms present challenges for non-proficient speakers, as well as the natural language processing (NLP) systems (Baldwin and Kim, 2010).

Recent developments in pretrained language models have significantly improved their performance in various tasks related to natural language generation and comprehension. However, figurative language language understanding remains to be one of the key challenges even for state-of-the-art language models (Tayyar Madabushi et al., 2021; Mi et al., 2025). Measuring and improving NLP systems beyond current state-of-the-art on figurative language processing requires high-quality and diverse benchmark datasets. However, the majority of current figurative language benchmark datasets concentrate on English or a limited number of high-resource languages. Although some multilingual idiom datasets exist (e.g., Tedeschi et al., 2022; Moussallem et al., 2018), the datasets for other languages are rather scarce.

In this paper, we present a benchmark corpus of Turkish idiomatic expressions that can be used to test idiom recognition, idiom understanding, paraphrasing, and contextual disambiguation. Each idiom in the corpus includes (1) the general form of the idiomatic expression (IE), (2) the description possibly with examples from a dictionary definition, (3) an example sentence with idiomatic use of the IE, (4) an example sentence with non-idiomatic, literal use of the IE, (5) a literal paraphrase of the idiomatic sentence (entailing (3)), and (6) a sentence with semantic/surface similarity to (3) without entailment. An example from the corpus is presented in Table 1. The fields (1) and (2) were obtained from online dictionaries, while fields (3)-(6) were created in this study. The primary objective is to provide a reliable and reusable benchmark that accurately captures linguistic variation and authentic usage of idioms in Turkish. Although multiple corpora of idiomatic expressions exist for Turkish (e.g., Berk et al., 2018; Eryiğit et al., 2023), these corpora focus on idiom detection tasks. To the best of our knowledge, a manually constructed corpus similar to our corpus does not exist for Turkish. Besides as a benchmark for

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 103–109

assessing idiom understanding of language models, the present dataset is also a useful resource for linguistic analysis of multi-word expressions and figurative language use, and for educational applications.

In the remainder of this paper, we briefly summarize some of the earlier work in the field (Section 2), describe the methodology used during corpus creation and provide some statistics on the corpus in Section 3. In Section 4 we present results on a selection of large language models (LLMs) for idiom detection and idiom understanding tasks evaluated on the present benchmark data, before concluding in Section 5.

2 Related work

Computational study of idioms typically overlap with studies of multi-word expressions (MWEs), as well as studies that focus on figurative language. While computational models of idiom understanding have a long history, the number of studies and the number of corpora annotated for idiomatic expressions has recently grown more rapidly (see Flor et al., 2025, for a recent survey of datasets).

As in other areas of natural language processing, many influential datasets are for English (Cook et al., 2008; Liu and Hwa, 2016; Stowe et al., 2022; Chakrabarty et al., 2022; Haviv et al., 2023, e.g.,). Recently, idiom datasets for other languages, such as Korean (Wang et al., 2025) and Danish (Sørensen et al., 2025), and even for truly low-resource languages, like Nepali (Pokharel and Agrawal, 2025) and Konkani (Shaikh et al., 2024) have also been published. Another relatively recent direction is multilingual datasets like AStitchInLanguageModels (Tayyar Madabushi et al., 2021) (English and Portuguese), ID10M (Tedeschi et al., 2022) which includes 10 languages, LIdioms (Moussallem et al., 2018) which also links idiomatic expressions in the languages covered. Khoshtab et al. (2025) also unifies a number of earlier idiomatic expression datasets, as well as introducing a new one in Persian. None of these multilingual datasets include Turkish. A recent study creates a Turkish idiomatic expressions dataset (Kim et al., 2025). However, the data is not released due to copyright concerns.

There has also been a number of shared tasks with idiom-related tasks, including FigLang (Saakyan et al., 2022), and PARSEME (Ramisch et al., 2018, 2020; Savary et al., 2023) shared task. PARSEME shared task also features a Turkish MWE dataset (including idioms) which was created and improved along with the shared task (Berk et al., 2018; Ozturk et al., 2022). Besides the PARSEME data, Eryiğit et al. (2023) is another manually created idiom dataset for Turkish. Like most idiom datasets for other languages, Turkish idiom datasets so far target the idiom (span) detection task. Our work differs from these corpora as it can be used probing understanding of idiomatic expressions through entailment, paraphrasing idioms, or even for idiom generation. Furthermore, current Turkish idiomatic expression datasets typically cover a small number of potentially idiomatic expressions (with a large number of figurative/literal example sentences), while our aim is to include a large number of diverse potential idiomatic expressions.

3 Corpus Creation and Corpus statistics

We selected a large set of idioms from a number of online idiom and proverb dictionaries.[1] We removed the proverbs, based on the indication in each dictionary, and eliminated exact duplicates. This resulted in 10 970 idioms and their descriptions. Some of the descriptions also include example uses of the idiom from literature. Turkish is an agglutinative language with a wide range of inflectional and derivational morphology, as well as a flexible word order. As a result, Turkish idioms often undergo morphological changes, such as shifts in tense, person, or voice, while retaining their metaphorical meaning. For instance, the idiom *burnu sürtülmek* shows up as *burnu sürtüldü* and *burnu sürtülsün* in different examples in Table 1. The potential variation is much wider, (e.g., *sürtülmüş büyük burunları* 'their big noses are (evidentially) scraped (lit.)' can also be perfectly fine in the appropriate context).

Another variation related to the corpus creation is the potential literal use of the idiomatic expressions. Some expressions are very likely to be used in their literal meaning (e.g., *baskın yapmak* 'to raid (lit.) / to visit someone unexpectedly (fig.)'), while others are very unlikely to be used literally (e.g., *burnu havada olmak* 'to have one's nose on

[1]The dictionary of Turkish Language association (https://sozluk.gov.tr/, Wiktionary (https://en.wiktionary.org/wiki/Category:Turkish_idioms), and a Learner's dictionary of Proverbs and Idioms (https://www.turkcedersi.net/deyimler-ve-deyimlerin-anlamlari/).

Field	Example
Form	*burnu sürtülmek* ‘to have (ones) nose scraped (lit.)’
Description	*Sıkıntı çektikten sonra daha önce beğenmediği bir durumu kabul etmek, gururundan vazgeçmek.* ‘To learn a lesson, accept an (unfavorable) condition after an unpleasant experience.’
Figurative	*Sözümüzü dinlemediği için burnu sürtülsün diye bıraktık.* ‘Since he/she did not listen, we left him/her there to teach him/her a lesson.’
Lit. paraphrase	*Sözümüzü dinlemediği için sıkıntı çeksin diye bıraktık.* ‘Since he/she did not listen, we left him/her there for him/her to suffer (and learn).’
Similar	*Sözümü dinlemedi ve burnu büyük diye ameliyat oldu ama sonrasından sıkıntı çekti.* ‘She/he did not listen to me and had a nose operation, but suffered a lot afterwards.’
Literal	*Kapıyı yüzüne birden kapatınca burnu sürtüldü.* ‘When the door was shut on her/his face, his nose scraped/scratched.’

Table 1: An example from the corpus.

the air (lit.) / to be arrogant (fig.)’). All idiomatic expressions in our corpus are MWEs. Most idiomatic expressions in the corpus are verbal constructions (including nominal object/oblique modifiers) similar to ones exemplified so far (89 %). However, there are also a number of conventionalized metaphors like *boncuk gibi* ‘like a bead’, or other expressions like *boğazına kadar* ‘up to his/her neck (lit.)’ and *babasının çiftliği* ‘one’s fathers farm (lit.)’. Currently we do not classify the idiomatic expressions based on any of these variations.

Ideally, to have a varied benchmark, all the above-mentioned variation should be considered while selecting idioms. Unfortunately, many of these are not quantifiable. As a result, we tried to balance the frequency of the potential idiomatic expressions based on their frequency in the Leipzig web corpus (Goldhahn et al., 2012), and selecting the first 200 instances we annotate from different frequency ranges. About 30 % of the 200-idiom corpus is not observed in the corpus, while the most frequent idiomatic expression occurs 7900 times per million sentences. All 200 idiomatic forms in the current corpus occur 36 000 times per million sentences.

After selecting the 200 instances, a researcher with background in translation studies (the first author) generated sample sentences following the guidelines listed below.

- Idiomatic use of the MWE, where we aimed at natural use of the idiom in typical (informal) communication settings, where the text alone is clear enough to signal idiomatic use. We avoided the use of other idioms in the generated sentence.
- Literal paraphrase of the sentence, where the sentence with idiomatic use would entail the sentence with the literal use. We avoided paraphrasing an idiom with another idiom.
- A sentence that is similar to the sentence with the idiomatic expression, but without an entailment relation – either contradictory with the idiomatic use or irrelevant.
- Non-idiomatic use of the same MWE. Again, we avoided the use of other potentially idiomatic expressions for this sentence as well. In a few cases (3 out of 200), a non-idiomatic use did not lead to a plausible sentence (e.g., *ayağının pabucu olmak* ‘to be shoe of one’s feet (lit.) / to be worthless in comparison to someone (fig.)’.

The resulting corpus contains 200 idioms (797 example sentences, and dictionary descriptions). The length of the sample sentences are approximately 9 tokens on average.

4 Computational Experiments

In this section we present results of idiomaticity detection and textual entailment recognition tasks on a sample of large language models, namely Google Gemini (Gemini Team et al., 2025), OpenAI GPT 4 (OpenAI et al., 2024), and a number of smaller open models from the Llama family (Meta AI, 2024). The models are asked to perform binary classification tasks. The first task asks

Model	Detection	Entailment
Gemini 2.5-flash	0.609	0.532
GPT-4o	0.594	0.520
Llama-3 70B-Instruct	0.614	0.545
Llama-3 8B-Instruct	0.544	0.517
Llama-3 3B-Instruct	0.521	0.495
Llama-3 1B-Instruct	0.496	0.475

Table 2: Accuracy of idiom detection and entailment of a selection of LLMs on the current Turkish idiom dataset.

whether there is an idiomatic expression used figuratively in the given sentence or not. In the second task, the language model is given the idiomatic sample sentence as the premise, and either literal rephrase or semantically similar non-entailing sentence, and asked whether there is entailment or contradiction. We prompted each language model with the simple zero-shot prompts (provided in Appendix A). Prompts are given to all models in English. We experimented changing the prompting language to Turkish, and also including expressions like 'you are an expert linguist' as part of the system prompt. However, the basic prompts presented in Appendix A worked best for most cases, with some variation without clear trends. We did not experiment with few-shot or CoT prompting as our aim is not to obtain best scores, but assess the 'understanding' of the idioms by the language models without further aid, similar to what would be expected in normal language use.

Even though they were asked for a restricted set of labels, the models, especially the larger ones, occasionally offered their unsolicited reasoning. In such cases if the first or the last word still was a valid label, we used it. For a few without an identifiable label, we read the text and determined the label manually. We report accuracy as the class distribution is balanced in both tasks. Table 2 presents the accuracies of all models we experimented with in this study.

Larger models perform around 60 % of accuracy in idiomaticity detection, while smaller models perform by chance or close to chance level. There is no noticeable difference between two commercial large language models and 80B parameter Llama 3. Textual entailment scores are generally worse, again, smaller models perform at chance level. Larger models perform better than chance, but also not much better than a random baseline. Looking closely at the labels, all models seem to prefer one of the labels heavily. Larger models typically prefer the positive answer ('entailment' or 'yes' to idiomaticity), but smaller models' label preference may also vary across different runs. For the idiom instances that were not found in the Leipzig corpus, performances of large models also drop to the level of a random baseline.

5 Conclusions and Future Directions

We presented a fully manually created corpus of Turkish idioms. The corpus is built on a selection of potentially idiomatic expressions based on their frequency, and includes newly-created sample sentences including idiomatic and non-idiomatic uses of the potentially idiomatic expressions, as well as literal sentences that is in entailment or contradiction relation with the idiomatic sentence. The information in the corpus can be useful for testing idiom understanding of NLP systems through textual entailment task, paraphrasing idiomatic expressions as literal expressions, idiom generation, as well as idiom identification.

The preliminary computational experiments show that the current dataset is challenging for large language models. Even state-of-the-art commercial LLMs seem to do barely above chance level on the entailment task. We also show that the performance of the models further decreases for the low-frequency idioms. This finding is in line with earlier observations that current idiomatic expression datasets lack variety and are not challenging enough (e.g., Haagsma et al., 2019; De Luca Fornaciari et al., 2024).

The initial corpus we presented is part of an ongoing work. We plan to extend the coverage of the corpus both with respect to the number of idiomatic expressions, and with respect to sample sentences for each idiomatic expression. We also plan to classify the idiomatic expressions further, particularly the classes of expressions identified in linguistics, psychology and translation studies. These could particularly be interesting in comparing human and LM idiom usage or difficulties (e.g., attributes of idioms like 'concreteness' or 'imageability' are likely to have different difficulties for humans and LMs). The current version of the dataset is available at `https://github.com/coltekin/turkish-idioms`.

Limitations

The small size is currently the major limitation of the dataset, which also affects the reliability of the results obtained in computational experiments.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. pages 267–292.

Gözde Berk, Berna Erden, and Tunga Güngör. 2018. Turkish verbal multiword expressions corpus. In *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

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.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22.

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.

GülŞen Eryiğit, Ali Şentaş, and Johanna Monti. 2023. Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering*, 29(4):909–941.

Michael Flor, Xinyi Liu, and Anna Feldman. 2025. A survey of idiom datasets for psycholinguistic and computational research. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, pages 90–100, Hannover, Germany. HsH Applied Academics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 60 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Hessel Haagsma, Malvina Nissim, and Johan Bos. 2019. Casting a wide net: Robust extraction of potentially idiomatic expressions. *arXiv preprint arXiv:1911.08829*.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264, Dubrovnik, Croatia. Association for Computational Linguistics.

Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2025. Comparative study of multilingual idioms and similes in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8680–8698, Abu Dhabi, UAE. Association for Computational Linguistics.

Jisu Kim, Youngwoo Shin, Uiji Hwang, Jihun Choi, Richeng Xuan, and Taeuk Kim. 2025. Memorization or reasoning? exploring the idiom understanding of LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21689–21710, Suzhou, China. Association for Computational Linguistics.

Changsheng Liu and Rebecca Hwa. 2016. Phrasal substitution of idiomatic expressions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 363–373, San Diego, California. Association for Computational Linguistics.

Meta AI. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Diego Moussallem, Mohamed Ahmed Sherif, Diego Esteves, Marcos Zampieri, and Axel-Cyrille Ngonga Ngomo. 2018. LIdioms: A multilingual linked idioms data set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin,

Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 103 others. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Yagmur Ozturk, Najet Hadj Mohamed, Adam Lion-Bouton, and Agata Savary. 2022. Enhancing the PARSEME Turkish corpus of verbal multiword expressions. In *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, pages 100–104, Marseille, France. European Language Resources Association.

Rhitabrat Pokharel and Ameeta Agrawal. 2025. neDIOM: Dataset and analysis of Nepali idioms. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 160–171, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, and 6 others. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakrabarty, Debanjan Ghosh, and Smaranda Muresan. 2022. A report on the FigLang 2022 shared task on understanding figurative language. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 178–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, and 9 others. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Silvio Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta, and Voula Giouli. 2019. Literal occurrences of multiword expressions: Rare birds that cause a stir. *The Prague Bulletin of Mathematical Linguistics*.

Naziya Mahamdul Shaikh, Jyoti D. Pawar, and Mubarak Banu Sayed. 2024. Konidioms corpus: A dataset of idioms in Konkani language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9932–9940, Torino, Italia. ELRA and ICCL.

Nathalie Hau Sørensen, Sanni Nimb, Agnes Aggergaard Mikkelsen, and Jonas Jensen. 2025. The Danish idiom dataset: A collection of 1000 Danish idioms and fixed expressions. In *Proceedings of the 1st Workshop on Nordic-Baltic Responsible Evaluation and Alignment of Language Models (NB-REAL 2025)*, pages 55–63, Tallinn, Estonia. The University of Tartu Library.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Xiaonan Wang, Seoyoon Park, and Hansaem Kim. 2025. Benchmarking Korean idiom understanding: A comparative analysis of local and global models. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 1341–1351, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

A Prompts

The following are the prompts used for the experiments reported in the paper.

```
I will provide you with a
pair of sentences in Turkish
consisting of a premise and
a hypothesis. Is there a
contradiction or entailment
between the premise and
hypothesis? Answer only with
"contradiction" or "entailment".
Premise: [P]
Hypothesis: [H]
Label:
```

```
Does the following Turkish
sentence contain an idiom which
is used figuratively? Answer
only with "yes" or "no".
Sentence: [S]
Answer:
```

Diversity patterns run deep: Impact of diversity intake on multiword expression identification

Mathilde Deletombe, Manon Scholivet, Louis Estève, Thomas Lavergne, Agata Savary
Université Paris-Saclay, CNRS, LISN
first.last@lisn.fr

Abstract

Multiword expressions (MWEs) are good examples of a phenomenon where identification systems struggle with generalisation: MWE present in the test set but absent in the training set are rarely identified. This raises the question of the diversity of the test set, relative to that of the train set, and how this impacts performance. We set out to measure how much diversity of a train corpus increases when adding individual MWEs from the test corpus, and how this increase impacts MWE identification performance. We measure diversity across a three-dimension framework and find mostly consistent negative correlations with performance in 14 languages and 8 systems.

1 Introduction

Multiword expression (MWEs), such as *to **pay** a **visit***, *to **take off*** or *to **call it a day***, have been an object of interest and a major challenge in Natural Language Processing (NLP) for decades (Sag et al., 2002; Shwartz and Dagan, 2019), notably due to their prevalence in texts (Candito et al., 2021) and their semantic non-compositionality (Nandakumar et al., 2018; Cordeiro et al., 2019; Miletić and Schulte im Walde, 2025). MWE-related tasks defined by the NLP community include MWE identification in running text (Constant et al., 2017).

This task has received attention in the past decade, notably due to shared tasks such as DiMSUM for English (Schneider et al., 2016), and the PARSEME shared task on automatic identification of verbal MWEs (VMWEs) in up to 20 languages, with its three editions: 1.0 (Savary et al., 2017), 1.1 (Ramisch et al., 2018), and 1.2 (Ramisch et al., 2020). Edition 1.2, building on the findings from edition 1.1, introduced a focus on *unseen* VMWEs. A VMWE from the test corpus is considered seen if another VMWE with the same multiset of lemmas is annotated at least once in the train or the development corpus. Otherwise it is considered unseen. Ramisch et al. (2020) showed that the performances of VMWE identification systems more strongly (inversely) correlate with the number of unseen VMWEs than with the size of the train corpus. Savary et al. (2019) argued that this is due to the very nature of the MWE phenomenon and its distributional properties.

The number of MWEs seen in the test but not in the train can be interpreted as the lack of MWE diversity in the train, relative to the test. But diversity has many facets (Stirling, 1994, 2007; Ramaciotti Morales et al., 2021; Estève et al., 2025) and can refer not only to the number of categories (*variety*) but also to the evenness of their distribution (*balance*) and to their relative differences (*disparity*). Given that unseen data had such a predominant impact on system performance in PARSEME 1.2 shared task, we wish to examine how far these observations can be generalised to more widely understood diversity aspects. We introduce the notion of *train/test diversity intake*, or *diversity intake* for short, to denote the diversity that the test corpus adds to the train corpus.[1] In other words, we are interested in relative rather than absolute diversity quantification, as defined by Estève et al. (2025).

We address two research questions:

RQ1 How to estimate the train/test diversity intake?
RQ2 Does this intake correlate with performance in the MWE identification task?

To address RQ1, we take inspiration from interdisciplinary work on diversity, where this notion has been thoroughly conceptualised. We select three diversity indicators: richness delta, negated Zipfian curvature delta, and minimum tree edit distance. To tackle RQ2, we use the PARSEME 1.2 shared task corpora and system predictions. For the MWEs from a test corpus, we measure their

[1] For the sake of brevity, we consider that the development corpus (if any) is part of the train corpus.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 110–116
March 28, 2026

individual share in the diversity intake. We then calculate the correlation between this share and the fact of being correctly or wrongly predicted by a system. Our hypothesis is that diversity intake and performance are inversely correlated.

Data, codes and results of our experiments are openly available.[2]

2 PARSEME data

To examine how diversity intake correlates with performance, we use the open source corpora and system predictions from the PARSEME shared task 1.2.[3] The corpora cover 14 languages: Basque (EU), Chinese (ZH), French (FR), German (DE), Greek (EL), Hebrew (HE), Hindi (HI), Irish (GA), Italian (IT), Polish (PL), Brazilian Portuguese (PT), Romanian (RO), Swedish (SV), and Turkish (TR). Their sizes range from 35 thousand to over 1 million tokens per language, with 1 thousand to 9 thousand manually-annotated VMWEs.[4] They also include UD-style[5] morphosyntactic annotations.

We use the predictions of 8 out of 9 systems participating in the shared task.[6] MTLB-STRUCT, TRAVIS-multi and TRAVIS-mono were based on BERT-finetuning; ERMI and MultiVitamin used simpler neural networks; HMSid and Seen2Unseen applied association measures; Seen2Seen and FipsCo were rule-based.

3 Diversity measures

For diversity quantification, we use the conceptual framework defined by Stirling (1994, 2007) to unify previous work in several scientific fields, most prominently ecology (Ricotta and Szeidl, 2006; Leinster and Cobbold, 2012; Scheiner, 2012; Chao et al., 2014; Chao and Ricotta, 2019). This framework has been recently applied in NLP (Estève et al., 2025), and to MWEs in particular (Lion-Bouton et al., 2022). It assumes that diversity is a property of sets whose *elements* can be aportioned into *categories*. Like Lion-Bouton et al. (2022), we define categories as VMWE canonical forms represented by multisets of lemmas of their components. For instance, for the MWE *to* ***call a spade a spade***, the multiset of lemmas is {*a*, *a*, *call*, *spade*, *spade*}. Elements are occurrences of these MWE canonical forms.[7]

Given the category/element dychotomy, diversity can be characterized along three dimensions (Stirling, 2007): *variety*, *balance*, and *disparity*. All other things being equal, the higher the variety the higher the diversity of a set. The same holds for balance and disparity.

Variety relates to the number of categories. A simple and widely used variety measure is *richness*, i.e. simply the number of categories, and we will use it for estimating variety intake.

Balance relates to the evenness of the distribution of the elements in categories. Balance reaches its optimum when the distribution is perfectly uniform. In fields like ecology, the distribution of categories (e.g. species) is often hard to estimate reliably, and then so-called non-parametric diversity measures, like Shannon evenness (Smith and Wilson, 1996), are used. But if a particular distribution can be assumed, so-called parametric measures apply (Magurran, 2004). In NLP, Zipfian distributions are frequently encountered and apply to MWEs (Ryland Williams et al., 2015). A Zipfian distribution is characterised by the probability mass function $Z_{s,n}(i) = i^{-s}\left(\sum_{j=1}^{n} j^{-s}\right)^{-1}$, where, in our case, n is the number of VMWE categories, i is the rank of the i's most frequent VMWE category, and s is the exponent characterizing the curvature of the distribution. When $s = 0$, the distribution is uniform, and the higher s, the more curved (more unbalanced) the distribution is. Therefore, the opposite of curvature, i.e. $-s$, can be considered a measure of balance (Zhang et al., 2023).

Disparity reflects the extent to which categories are different from each other, which calls for an appropriate distance measure between categories. Recently, various disparity measures using semantic vector spaces have been used in NLP (Yang et al., 2024; Yu et al., 2022; Puranik et al., 2023; E et al., 2023; Kim et al., 2023; Cao and Wan, 2020) but it was also shown that such measures strongly correlate with variety, in particular when MWE are concerned (Estève et al., 2024). Therefore, in-

[2] https://gitlab.lisn.upsaclay.fr/deletombe/repo

[3] https://gitlab.com/parseme/sharedtask-data

[4] The annotation follows unified guidelines with a VMWE taxonomy including verbal idioms (***go bananas***), light verb constructions (***pay a visit***, ***grants rights***), inherently reflexive verbs (***help oneself***), verb-particle construction (***do in***), multi-verb constructions (***let go***) and inherently adpositional verbs (***rely on***).

[5] https://universaldependencies.org/format.html

[6] The 9th system, MultiVitaminBooster, had per-language scores below 1% F-measure.

[7] Note that a non-idiomatic co-occurrence of a multiset of lemmas does not count as an element, e.g. in *she called this thing a spade but a spade is something else*.

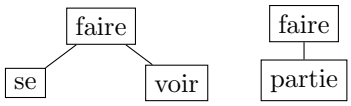


Figure 1: Syntactic trees of two French VMWEs: ***se faire voir*** (lit. 'make oneself see') 'show oneself in a fancy place' and ***faire partie*** (lit. 'make part') 'belong'.

spired by Guo et al. (2024), we endorse interest in syntactic (and partly semantic) diversity.

As the distance underlying our disparity measure, we use the *tree edit distance* by Zhang and Shasha (1989) between two VMWEs seen as simplified syntactic dependency subtrees, where only the lemmas and the head-dependent relations are represented, as in Figure 1. For such trees, we consider three elementary edit operations, allowed both for leaves and internal nodes, in any order: (i) deletion of a node, (ii) insertion of a node, (iii) replacement of the lemma of a node by another lemma. Operations (i) and (ii) each cost 1. The cost of (iii) is half of the cosine distance between the vectors representing the original lemma and the replacement lemma (i.e. its range is $[0, 1]$). We use the Word2Vec-style (Mikolov et al., 2013) vector spaces by Estève et al. (2024),[8] trained on the PARSEME corpus edition 1.3.[9] The tree edit distance is the cost of a minimal sequence of elementary edit operations transforming one tree into another. For instance, the edit distance between the two trees in Figure 1 is 1.282: 1 for the deletion of *se* 'oneself' and $0.564/2$ for the replacement of *voir* 'see' by *partie* 'part'.

4 Diversity intake

To address RQ1, we estimate diversity intake (DI), for each VMWE individually, along the three dimensions of diversity. For each diversity dimension, the DI of the test corpus is represented by a vector, comprising all individual DIs. For a given language L, let TRAIN and TEST be its train and test corpora. Let $E_{\text{TEST}} = (e_1, \ldots, e_n)$ be the list of the VMWE categories from TEST. Consider the toy example of TRAIN (1) and TEST (2) in French. Here, E_{TEST} =({*faire* 'do', *partie* 'part'}, {*faire* 'do', *se* 'oneself', *voir* 'see'}, {*faire* 'do', *sembler* 'seem'}).

(1) Il **s'agissait** de **faire partie** du show, donc il en **faisait partie**.
It was about taking part in the show, so he took part in it.

(2) Même s'il n'en **faisait** pas **partie**, il s'y **ferait voir** et **ferait semblant**.
Event if he didn't take part in it, we would show up and pretend.

We define the *variety intake* $DI_v(e_i)$ to be 1 if e_i is absent from TRAIN (i.e. it adds to TRAIN's variety), and 0 otherwise. Then the DI for the whole TEST is $DI_v = (DI_v(e_1), \ldots, DI_v(e_n))$. In (2) we have $DI_v = (0, 1, 1)$.

To calculate the *balance intake* $DI_b(e_i)$, we add e_i to the set of VMWEs from TRAIN and recalculate its Zipfian curvature s. The difference between $-s$ in TRAIN with and without e_i is the value of $DI_b(e_i)$. In (2), adding e_2 or e_3 to TRAIN flattens the curvature but adding e_1 increases the frequency of the most frequent category. Therefore, the DI_b vector has positive values at positions 2 and 3 and a negative one at position 1.

The *disparity intake* follows a slightly different logic than variety and balance intake. The idea is that a system might correctly identify e_i on the basis of a VMWE from TRAIN which is similar, even if not identical, to e_i. Therefore, $DI_d(e_i)$ is defined as the minimum edit distance between e_i and any VMWE in TRAIN. In (2), we have $DI_d \approx (0.00, 1.11, 0.16)$ respectively for **faire partie** / **faire partie** (identical means 0 distance), **s'agit** / **se faire voir** (0 distance between **se** and **se**, ≈ 0.11 between **agir** and **faire**, 1 to add **voir**), and **faire partie** / **faire semblant** (0 distance between **faire** and **faire**, ≈ 0.16 between **partie** and **semblant**).

To account for performance of system S on expression e_i, we define $Perf_S(e_i)$ to be 1 if S has correctly identified e_i and 0 otherwise. Then $Perf_S = (Perf_S(e_1), \ldots, Perf_S(e_n))$. In (2), if only the first two expressions are true positives, then $Perf_S = (1, 1, 0)$.

To address RQ2, in each language we calculate the diversity intake vectors, DI_v, DI_b and DI_d. We then measure the Pearson correlation between each of them and the performance vector $Perf_S$, for each system S. The results are described in the following section.

[8] This semantic space represents both single words and VMWEs. Here, we only use vectors for the former.

[9] Edition 1.3 contains consolidated versions of the VMWE-annotated corpora from 3 shared task editions in 26 languages.

Table 1: Pearson correlation measurement between variety/balance/disparity intake and performance

		DE	EL	EU	FR	GA	HE	HI	IT	PL	PT	RO	SV	TR	ZH
Variety	ERMI	−0.46	−0.47	−0.47	−0.43	−0.50	−0.61	−0.48	−0.46	−0.51	−0.38	−0.60	−0.53	−0.43	−0.44
	FipsCo	−0.18	−0.28		−0.33										
	HMSid				−0.26										
	MTLB-STRUCT	−0.53	−0.52	−0.56	−0.54	−0.48	−0.74	−0.42	−0.61	−0.58	−0.47	−0.63	−0.52	−0.44	−0.39
	Seen2Seen	**−0.89**	**−0.87**	**−0.86**	**−0.90**	**−0.74**	**−0.87**	**−0.75**	**−0.88**	**−0.94**	**−0.89**	**−0.65**	**−0.85**	**−0.88**	**−0.89**
	Seen2Unseen	−0.86	−0.82	−0.80	−0.78	−0.58	**−0.87**	−0.39	−0.84	−0.89	−0.81	−0.63	−0.81	−0.82	**−0.89**
	TRAVIS-mono	−0.41	−0.14		−0.49			−0.30	−0.48	−0.54		−0.52	−0.45	−0.38	−0.31
	TRAVIS-multi	−0.46	−0.54	−0.44	−0.55	−0.22	−0.67	−0.39	−0.53	−0.54		−0.47	−0.49	−0.44	−0.40
Balance	ERMI	**−0.31**	−0.31	**−0.35**	**−0.34**	−0.49	**−0.56**	−0.32	**−0.48**	**−0.39**	−0.35	−0.19	**−0.48**	**−0.36**	−0.32
	FipsCo	0.16	−0.07		−0.31										
	HMSid				−0.21										
	MTLB-STRUCT	−0.21	−0.23	−0.28	−0.25	−0.42	−0.43	−0.26	−0.35	−0.31	−0.29	−0.15	−0.40	−0.31	−0.23
	Seen2Seen	−0.27	−0.31	**−0.35**	−0.31	**−0.65**	−0.51	**−0.53**	−0.34	−0.36	**−0.36**	**−0.22**	**−0.48**	**−0.36**	**−0.33**
	Seen2Unseen	−0.27	−0.29	−0.32	−0.28	−0.54	−0.50	−0.36	−0.33	−0.34	−0.34	**−0.22**	−0.47	−0.34	**−0.33**
	TRAVIS-mono	−0.21	**−0.44**		−0.24			−0.50	−0.34	−0.28		−0.16	−0.40	−0.29	−0.21
	TRAVIS-multi	−0.23	−0.25	−0.31	−0.28	−0.34	−0.47	−0.33	−0.36	−0.32		**−0.22**	−0.42	−0.29	−0.25
Disparity	ERMI	−0.33	**−0.26**	**−0.33**	−0.26	−0.19	−0.23	−0.38	**−0.39**	−0.41	**−0.35**	−0.26	−0.27	**−0.27**	−0.02
	FipsCo	−0.34	−0.13		**−0.32**										
	HMSid				−0.29										
	MTLB-STRUCT	**−0.36**	−0.23	−0.32	−0.22	**−0.24**	−0.27	−0.44	**−0.39**	−0.41	−0.34	−0.29	−0.28	−0.24	**−0.09**
	Seen2Seen	**−0.36**	−0.21	−0.21	−0.21	−0.11	−0.23	−0.30	−0.35	−0.35	−0.27	−0.12	−0.25	−0.19	**−0.09**
	Seen2Unseen	**−0.36**	−0.22	−0.23	−0.24	−0.15	−0.24	−0.36	−0.36	−0.36	−0.29	−0.11	−0.25	−0.23	−0.08
	TRAVIS-mono	−0.32	−0.02		−0.21			−0.15	−0.37	−0.35		**−0.33**	−0.25	−0.22	−0.05
	TRAVIS-multi	−0.35	−0.21	**−0.33**	−0.23	−0.15	**−0.30**	**−0.46**	−0.36	**−0.42**		−0.28	**−0.32**	−0.26	−0.06

5 Results

The results are shown in Table 1, with the strongest correlation for each language highlighted in bold. All results, except for TRAVIS-multi in ZH, ERMI for the same langage and TRAVIS-mono for EL, are statistically significant with threshold 0.05. We observe mostly negative correlation (with only three exceptions), which suggests that higher train/test diversity intake is associated with weaker system performance, and vice versa, which corroborates our hypothesis. Negative correlation is considered (i) strong, (ii) moderate and (iii) weak, if the scores fall (i) below -0.7 or above 0.7, (ii) from -0.7 to -0.3 or from 0.3 to 0.7 and (iii) from -0.3 to 0.3.

Regarding variety, a strong negative correlation with system performance is noticeable. Out of the 83 scores, the majority consists of moderate (53) and strong (25) correlations. The Seen2Seen and Seen2Unseen systems display the strongest negative correlation, reaching -0.94 for Seen2Seen in PL and -0.89 for Seen2Unseen in ZH and PL. This is expected, given that these systems focus on the MWEs seen in train. Conversely, FipsCo relies of external MWE lexicons, which is consistent with its relatively weak correlation with variety intake.

For balance intake we observe a weaker but still non-negligible negative correlation, with 52 moderate and 31 weak scores. The highest scores are shown for Seen2Seen (reaching -0.65 in GA) and ERMI (reaching -0.56 in HE). This might be partly due to the fact that unseen VMWEs, when added to TRAIN systematically increase both its variety and balance. This might strongly influence the systems like Seen2Seen and ERMI which use no external data (e.g. lexicons, pre-trained models).

As to disparity intake, a majority of correlations (50) are weak, which indicates that disparity intake has little or no impact on performance. This might mean that systems hardly capture syntactic and semantic similarities between VMWEs. There are only 33 moderate correlations, notably for TRAVIS-multi in HI and PL (-0.46 and -0.42), ERMI in PL (-0.41), and MTLB-STRUCT in HI and PL (-0.44 and -0.41).

6 Conclusions and future work

In this paper, we investigated the correlation between train/test diversity intake and the performance of VMWE identification systems. We confirmed prior findings from the PARSEME shared tasks showing a strong relationship between system performance and the rate of unseen VMWEs, which we re-interpreted as variety intake.

We extended previous findings to the two other dimensions of diversity — balance and disparity — by studying whether similar correlations could be observed. Balance intake, quantified through the curvature of the Zipfian distribution, was found to exhibit a moderate correlation with system performance. In contrast, disparity intake, modelled using a novel approach based on tree edit distance,

showed only weak correlation with system performance in our experiments.

In the future, we intend to apply the methods presented here to the corpora and system predictions from the latest edition 2.0 of the PARSEME shared task (Scholivet et al., 2026). We also wish to study alternative ways to measure diversity that may better correlate with performance. Such measures may help evaluate the systems but also improve their training and contribute to the creation of more balanced datasets.

7 Acknowledgements

This work received support from: (i) COST (European Cooperation in Science and Technology) through the CA21167 COST action UniDive, (ii) the French Agence Nationale pour la Recherche, through the SELEXINI project (ANR-21-CE23-0033-01), (iii) the "*Plan Blanc*" (White Plan) doctoral funding from Université Paris-Saclay (France).

8 Limitations

This work uses vector spaces to quantify dissimilarity between words, which, as any vector space trained on real-world data, cannot account for all possible structures and variations. The quality of the vector spaces, and subsequent experiments relying on them, while reasonable, cannot be a complete and perfect representation of these phenomena at work.

Diversity quantification has a very rich bibliography and many other diversity measures exist which could be applied in our context. More thorough criteria for selecting the most accurate measures are needed.

References

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier, and Silvio Cordeiro. 2021. A french corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*, 8(2).

Yue Cao and Xiaojun Wan. 2020. DivGAN: Towards diverse paraphrase generation via diversified generative adversarial network. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2411–2421, Online. Association for Computational Linguistics.

Anne Chao, Chun-Huo Chiu, and Lou Jost. 2014. Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution, and Systematics*, 45:297–324. Publisher: Annual Reviews.

Anne Chao and Carlo Ricotta. 2019. Quantifying evenness and linking it to diversity, beta diversity, and similarity. *Ecology*, 100(12):e02852. Number: 12.

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

Silvio Ricardo Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57. Impact Factor: 1.319. http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00341.

Venkatesh E, Kaushal Maurya, Deepak Kumar, and Maunendra Sankar Desarkar. 2023. DivHSK: Diverse headline generation using self-attention based keyword selection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1879–1891, Toronto, Canada. Association for Computational Linguistics.

Louis Estève, Agata Savary, and Thomas Lavergne. 2024. Vector spaces for quantifying disparity of multiword expressions in annotated text. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 110–130, Bangkok, Thailand. Association for Computational Linguistics.

Louis Estève, Marie-Catherine de Marneffe, Nurit Melnik, Agata Savary, and Olha Kanishcheva. 2025. A survey of diversity quantification in natural language processing: The why, what, where and how. *Preprint*, arXiv:2507.20858.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.

Donghyun Kim, Youbin Ahn, Wongyu Kim, Chanhee Lee, Kyungchan Lee, Kyong-Ho Lee, Jeonguk Kim, Donghoon Shin, and Yeonsoo Lee. 2023. Persona expansion with commonsense knowledge for diverse and consistent response generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1139–1149, Dubrovnik, Croatia. Association for Computational Linguistics.

Tom Leinster and Christina A. Cobbold. 2012. Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489. Number: 3.

Adam Lion-Bouton, Yagmur Ozturk, Agata Savary, and Jean-Yves Antoine. 2022. Evaluating diversity of multiword expressions in annotated text. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3285–3295, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Anne E. Magurran. 2004. *Measuring biological diversity*. Oxford: Blackwell Publishing Company , 2004, Oxford.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Filip Miletić and Sabine Schulte im Walde. 2025. Modeling the evolution of English noun compounds with feature-rich diachronic compositionality prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20071–20092, Vienna, Austria. Association for Computational Linguistics.

Navnita Nandakumar, Bahar Salehi, and Timothy Baldwin. 2018. A comparative study of embedding models in predicting the compositionality of multiword expressions. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 71–76, Dunedin, New Zealand.

Vinayak Puranik, Anirban Majumder, and Vineet Chaoji. 2023. PROTEGE: Prompt-based diverse question generation from web articles. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5449–5463, Singapore. Association for Computational Linguistics.

Pedro Ramaciotti Morales, Robin Lamarche-Perrin, Raphaël Fournier-S'Niehotta, Rémy Poulain, Lionel Tabourier, and Fabien Tarissan. 2021. Measuring diversity in heterogeneous information networks. *Theoretical Computer Science*, 859:80–115. Publisher: Elsevier.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, and 6 others. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Carlo Ricotta and Laszlo Szeidl. 2006. Towards a unifying approach to diversity measures: bridging the gap between the Shannon entropy and Rao's quadratic index. *Theoretical Population Biology*, 70(3):237–243. Number: 3.

Jake Ryland Williams, Paul R. Lessard, Suma Desu, Eric M. Clark, James P. Bagrow, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Zipf's law holds for phrases, not words. *Scientific Reports*, 5.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expresions: A Pain in the Neck for NLP. In *Proceedings of CICLING'02*. Springer.

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91, Florence, Italy. Association for Computational Linguistics.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Samuel M. Scheiner. 2012. A metric of biodiversity that integrates abundance, phylogeny, and function. *Oikos*, 121(8):1191–1202. Number: 8.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559, San Diego, California. Association for Computational Linguistics.

Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Carp, and Vasile Păiș. 2026. Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Benjamin Smith and J. Bastow Wilson. 1996. A Consumer's Guide to Evenness Indices. *Oikos*, 76(1):70–82. Number: 1 Publisher: [Nordic Society Oikos, Wiley].

Andrew Stirling. 1994. Diversity and ignorance in electricity supply investment. *Energy Policy*, 22(3):195–216.

Andy Stirling. 2007. A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4(15):707–719. Number: 15 Publisher: Royal Society.

Yuting Yang, Pei Huang, Feifei Ma, Juan Cao, and Jintao Li. 2024. PAD: A robustness enhancement ensemble method via promoting attention diversity. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12574–12584, Torino, Italia. ELRA and ICCL.

Yu Yu, Shahram Khadivi, and Jia Xu. 2022. Can data diversity enhance learning generalization? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4933–4945, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kaizhong Zhang and Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal on Computing*, 18(6):1245–1262.

Xinran Zhang, Maosong Sun, Jiafeng Liu, and Xiaobing Li. 2023. Lingxi: A diversity-aware Chinese modern poetry generation system. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 63–75, Toronto, Canada. Association for Computational Linguistics.

A Curious Class of Adpositional Multiword Expressions in Korean

Junghyun Min[1] **Na-Rae Han**[2] **Jena D. Hwang**[3] **Nathan Schneider**[1]
[1]Georgetown University [2]University of Pittsburgh [3]Allen Institute for AI
{jm3743, nathan.schneider}@georgetown.edu
naraehan@pitt.edu jenah@allenai.org

Abstract

Multiword expressions (MWEs) have been widely studied in cross-lingual annotation frameworks such as PARSEME. However, Korean MWEs remain underrepresented in these efforts. In particular, Korean multiword adpositions lack systematic analysis, annotated resources, and integration into existing multilingual frameworks. In this paper, we study a class of Korean functional multiword expressions: postpositional verb-based constructions (PVCs). Using data from Korean Wikipedia, we survey and analyze several PVC expressions and contrast them with non-MWEs and light verb constructions (LVCs) with similar structure. Building on this analysis, we propose annotation guidelines designed to support future work in Korean multiword adpositions and facilitate alignment with cross-lingual frameworks.

1 Introduction

PARSing and Multiword Expressions (PARSEME; Savary et al., 2015) is a robust multilingual framework for annotating idiomatic word combinations, known as multiword expressions (MWEs). They are distinguished from literal and fully productive combinations via a suite of linguistic tests, and are classified into subcategories based on their grammatical structure. Similarly to words, MWEs are categorized according to their morphosyntactic and functional criteria (Baldwin and Kim, 2010; Savary et al., 2023).

One category of MWEs is expressions that act like prepositions or postpositions (collectively 'adpositions'). The PARSEME 2.0 guidelines encompass a wide range of MWE types, including adpositional MWEs as a subtype of functional MWEs.[1] An English example (sometimes termed a *complex preposition*) is **in front of**. With respect to its syntactic distribution, **in front of** is similar to single-word adpositions like **behind** and **near**. It also carries a relational, and in particular spatial, meaning typical of adpositions. Finally, it is an MWE by PARSEME guidelines since it exhibits grammatical fixedness (fossilization): it has been lexicalized to not allow morphological changes (**in fronts of*) or modifiers (**in far front of*).

In this paper, we consider adpositional MWEs in Korean against the backdrop of the PARSEME framework. We focus on a pattern of grammaticalized adpositional MWEs that we term **postpositional verb-based constructions (PVCs)**. An example appears in (1):

(1)

*게*에	관한	*책*
key-**ey**	**kwanha-n**	chayk
crab-OBL	relate-ADN	book

'a book about crabs'

In (1), the semantically bleached postposition **-**에 **ey** combines with the verb **관한** ***kwanhan***—fossilized in its adnominal (or attributive[2]) form—to constitute the adpositional MWE **-**에 관한 **ey kwanhan** 'about'.[3] As a unit, the MWE marks the topic argument of the head nominal 'book' to mean 'book **about** crabs.'

To the best of our knowledge, this is the first study of PVCs. Prior studies of Korean have examined a range of classes of MWEs and idiomatic expressions, including a range of adpositional expressions, and the polysemy of grammatical markers that come into play in PVCs (§5). Moreover, adpositions and adpositional MWEs have been semantically annotated in the multilingual SNACS framework (Schneider et al., 2018, 2022; Arora

[1] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/

[2] 'Attributive' contrasts with predicative; 'adnominal' focuses on its function as a modifier of a noun. Both refer to the same set verb endings in Korean.

[3] Throughout this paper, we format **adpositions** (including PVCs), ***verbs***, and *other categories, including suffixes*.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 117–126

et al., 2021, *inter alia*), but the Korean implementation of SNACS (Hwang et al., 2020) has not annotated adpositional MWEs.[4]

In this paper, we develop an account of this construction and a proposal for how its instances should be annotated in the PARSEME framework. We contribute:

- an initial list of such Korean PVCs, drawn from Korean Wikipedia;
- a proposal for how to annotate them under current PARSEME guidelines; and
- a proposed linguistic analysis of PVCs, including a discussion on distinguishing them from adnominal verbs.

2 Defining PVCs

Single-word Korean adpositions are postpositions called 조사 *cosa*[5] that are orthographically and phonologically bound to a noun phrase (Martin, 1992; Sohn, 2001; Yeon and Brown, 2019). They function as case markers, conjunctions, and markers of discourse-pragmatic meaning like topic and focus (Hwang et al., 2020).

We define Korean PVCs as adpositional multiword units composed of two parts: a *postposition* that attaches to a noun phrase and a verb that undergoes limited inflection.[6]

This paper focuses on *하다 -hata* PVCs, whose verb portion is composed of a *bound stem* and a *verbalization suffix 하다 -hata* that attaches to the bound stem.[7] We thus describe them as having three components: a postposition, a bound stem, and *하다 -hata* suffix. Together, they mark the relationship between the head and the object noun. As illustrated in (1), the adposition **-에 ey** and the bound stem **대 *tay*** with the verbalization suffix in the adnominal form (*-한 han*) comprise a PVC that corresponds to the English preposition **about**.

A key characteristic of Korean postpositional MWEs is that they are fossilized in meaning. Although the adnominal (e.g., **관한 *kwanhan***) in example (1) is in the attributive form of a verb (e.g., **관하다 *kwanhata*** 'to relate to'), it cannot be productively used as a matrix verb of a sentence and does not participate in regular verbal inflection as exemplified in (2). Additionally, it cannot freely undergo regular morphological change as shown in (3), appearing almost exclusively within a limited set of constructions allowed for the MWE, further discussed in §4.1.

(2)

* *책이*	*게에*	관했다
chayk-i	key-**ey**	**kwanha-yss-ta**
book-NOM	crab-OBL	relate-PAST-DECL

'A book was about crabs.'

(3)

게에	관한 / *했던	*토론*
key-**ey**	**kwanha-n** / ***-yss-ten**	tholon
crab-OBL	relate-ADN / -PAST-ADN	debate

'a debate that used to be about crabs'

This is in contrast to the full predicate verbs with suffix ***-hata***. Example (4) shows the predicate **구하다 *kuhata*** 'to rescue' appearing in the same adnominal construction as PVCs. But the key distinguishing factor is that this can also be freely used as a matrix verb with a full range of inflectional endings while retaining its core meaning as seen in example (4).

(4)

친구를	구한 / 했던	*강아지*
chinku-**lul**	**kuha-n** / **-yss-ten**	kangaci
friend-ACC	rescue-ADN / -PAST-ADN	puppy

'a puppy that rescued a friend'

These characteristics notably pass two of the tests for fixedness under the PARSEME guidelines (Savary et al., 2015).

3 Extracting a list of PVCs

A range of expressions fall in the PVC class. To compile an initial list of these, we develop a pipeline to extract and filter candidates from corpora. The main challenge is disambiguating them from false positives (verbs and light verb constructions appearing in the same adnominal construction as PVCs, as in (4)).

Candidate extraction First, we take the May 2024 dump of Korean Wikipedia totaling 515k articles (Chang, 2024; Wikimedia Foundation, 2024). Then, we analyze the main text of each article with `konlpy`'s (Park and Cho, 2014) `Mecab` morphological analyzer (Kudo, 2013). Finally, using a regular expression, we look for sequences of an adposition

[4] And likewise for Japanese, which is structurally similar (Aoyama et al., 2024).

[5] Pronounced 'josa'.

[6] While we do not discuss Korean adpositional MWEs beyond PVCs, we are aware of several postpositional units that may be considered multiword expressions, like stacked postpositions (Hwang et al., 2020) and other postposition equivalents (Moon, 2015).

[7] PVCs also extend to constructions that do not contain the *-hata* suffix such as **-에 따르면 -ey ttalumyen** 'according to' containing a bound stem lexicalized from the verb **따르다 *ttaluta*** 'to follow'. As with *하다 -hata* PVCs, the verb is limited in inflection.

Adposition	Bound stem	Meaning	Suffix forms with 하다 hata inflections
-에 *ey*; oblique	대 *tay*	about	-한 *han*, -해 *hay*, -해서 *hayse*, -하여 *haye*
-에 *ey*	의 *ui*	by	-한 *han*, -해 *hay*, -해서 *hayse*, -하여 *haye*
-를 *lul*; accusative	통 *thong*	via, through	-한 *han*, -해 *hay*, -해서 *hayse*, -하여 *haye*
-를 *lul*	위 *wi*	for	-한 *han*, -해 *hay*, -해서 *hayse*, -하여 *haye*
-로 *lo; dative*	인 *in*	due to	-한 *han*, -해 *hay*, -해서 *hayse*, -하여 *haye*
-에 *ey*	관 *kwan*	about	-한 *han*, -해 *hay*, -해서 *hayse*, -하여 *haye*
-에 *ey*	속 *sok* †	in, belong to	-한 *han*, -해 *hay*
-로 *lo*, 를 *lul*	향 *hyang* †	towards	-한 *han*, -해 *hay*
-에 *ey*	비 *pi*	than, compared to	-한 *han*, -해 *hay*, -해서 *hayse*, -하여 *haye*
-에도 *eyto*; oblique + additive	불구 *pwulkwu*	although	-하고 *-hako*
-를 *lul*	비롯 *piros*	such as	-한 *han*, -해 *hay*, -해서 *hayse*, -하여 *haye*
-를 *lul*	기 *ki* †	since	-한 *han*, -해 *hay*
-에 *ey*	반 *pan*	against, unlike	-한 *han*, -해 *hay*, -해서 *hayse*, -하여 *haye*
-를 *lul*	위시 *wisi*	such as	-한 *han*, -해 *hay*, -해서 *hayse*, -하여 *haye*

Table 1: Non-exhaustive list of distributional properties of 14 Korean PVCs, in order of frequency in Korean Wikipedia. Additional adpositions and suffix forms are possible for select arguments as described in §4. *-해 -hay*, *-해서 hayse*, *-하여 haye* are resultative connective suffixes; *-하고 hako* is an conjunctive connective suffix; *-한 han* is an adnominal suffix. †Unlike others, these bound stems can serve as the main predicate in a sentence, as discussed in §4.1.

(with a J* tag), a bound stem (XR but commonly erroneously parsed as a noun NN*), and some inflection of suffix *-hata*.[8]

For each stem, we retrieve a list of adpositions, suffixes, and adposition-stem-suffix sequences that occur with it. As PVCs are more lexicalized than the usual verb-argument construction, bound stems that form PVCs co-occur with smaller numbers of adpositions, suffixes, and sequences, which may be used to select bound stems that can form PVCs. However, to ensure accurate retrieval, we take the 300 most frequently occurring stems in such sequences out of 3.5k candidates and manually verify whether the stem forms a PVC.

Annotating candidates To verify the PVC-hood of a candidate, we exploit the properties of PVCs described in §4 and connected to those of functional MWEs (Savary et al., 2015). PVC verbs:

- exhibit limited capacity for morphological inflection (see (3)),
- cannot be modified (5), and
- do not serve as a main predicate of a sentence or a clause (discussed further in §4.1). Exceptions exist, which we do not treat as instances of PVCs; see Appendix B.

Among the 300 stems, 12 are bound stems that form PVCs, and the rest are verbs (mostly light verb constructions) with cased arguments. Of the 12, 3 stems can also serve as a main predicate of a sentence or a clause; they only form PVCs when they do not (§4.1). Our additional manual analysis uncovers 2 additional, less frequent bound stems that form PVCs.

[8]We release our pipeline at https://github.com/aatlantise/korean-multiword-postpositions.

4 Properties of PVCs

The 14 identified PVCs, their components, and their meanings appear in Table 1, sorted from most to least frequent in Korean Wikipedia. The list of applicable postpositions and suffixes is not exhaustive. For example, **-**로 향한 **lo hyanghan** 'towards' can surface with **-**에게 **eykey** when it attaches to animate arguments (**-**에게 향한 **eykey hyanghan**).

We verify that PVCs are indeed MWEs and adpositions, in accordance with the PARSEME guidelines (Savary et al., 2015). Distributionally, they can be replaced with single-word adpositions: for example, **-**를 위한 **lul wihan** 'for' can be replaced by **-**의 **ui** 'of' or **-**같은 **kath'un** 'like' while retaining grammaticality and acceptability.

As expected, Korean PVCs exhibit morphological or lexical inflexibility that characterize MWEs. Examples (2), (3), and (5) show that modification of a component (the verb) and regular morphological change into the past tense results in questionable acceptability, where both examples are better analyzed as a predicate in a relative clause. We note that despite the lexicalization, PVCs exhibit neither irregular syntactic structure nor semantic idiomaticity; we thus describe them as weak MWEs (Schneider et al., 2014).

(5) * *게에* *약간* 관한 *책*
key-**ey** yakkan **kwanha-n** chayk
crab-OBL somewhat relate-ADN book

'a book somewhat about crabs'

4.1 Adnominal and connective forms only

A distinctive property of Korean PVCs, resulting from their lexicalization, is that the verb suffix *-하다 hata* present in the PVC can only be inflected in limited cases discussed below.

Unlike regular verbs, many that form PVCs entirely resist appearing as a tensed matrix predicate: progressive ***관한다 *kwanhanta*** and past ***관했다 *kwanhayssta*** would be ungrammatical for (1). For others, doing so forces a change in meaning: the PVC **-에 대한 ey tayhan** means 'about', but **-을 대했다 *ul tayhayssta*** in the past-tense matrix form require a different postposition *and* changes the meaning to 'treated'.

As a result, *-hata* suffixes in PVCs have a constrained distribution tied to specific inflectional endings: they only appear in a pre-nominal position with the adnominal ending *-한 han* (shown in (1)) or in a pre-verbal position with connective endings (*-해 hay*, *-해서 hayse*, *-하여 haye*).[9]

A small set of PVCs exhibit exceptions: **속하다 *sokhata*** 'be a part of', **향하다 *hyanghata*** 'face, head towards', and **기하다 *kihata*** 'set as time for something to begin'. These permit predicative use without a major semantic shift. However, we treat these as PVCs, as the predicative forms are generally dispreferred in favor of their respective PVC forms or alternate constructions. See Appendix B for more discussion.

4.2 PVC verbs are not light verb constructions

In addition to the predicate verb and verbal suffix, **하다 *hata*** 'to do' can function as a light verb that combines with a noun complement to become lexicalized and form light verb constructions (example (6); Chae, 1996; Han, 2000). In fact, just like the full predicate, these LVCs can also participate in verbal inflections. As a result, verbs in PVCs like 관한 **kwanhan** (example (1)) can resemble LVCs, especially when the LVC appears as an adnominal.

(6)
공원을	*산책하다*
kongwuen-lul	sanchayk-ha-ta
park-ACC	stroll-LV-DECL

'take a stroll at the park'

(7)
공원을	*산책한*	*사람*
kongwuen-lul	sanchayk-ha-n	salam
park-ACC	stroll-LV-ADN	person

'person that took a stroll at the park'

In an LVC, most of the semantic content is contributed by the noun (Jespersen, 1965; Cattell, 1984; Baldwin and Kim, 2010). The noun in an LVC like *산책 sanchayk* 'stroll' is a fully unbound lexical item (Chae, 1996), such that can be reduced to a nominal form, as in (8).

(8)
공원	*산책*
kongwuen	sanchayk
park	stroll

'a stroll at the park'

(9)
**게*	*관*
key	kwan
crab	relate

In a PVC, while it could be argued that the semantic content of **관하다 *kwanhata*** 'relate' stems from *관 kwan*,[10] PVC verb stems are bound and cannot appear on their own—as shown in example (9), which is structurally analogous to (8). They are comparable to English bound morphemes like *fer* in ***refer*** or ***confer***. Thus, they are unable to be reduced to nominal forms, as required by PARSEME guidelines (Savary et al., 2015).

4.3 Other constructions with same structure

In this section, we discuss how PVCs with predicative and non-predicative verbs ((2), (3), and (5)) compare to LVCs (6) and non-MWEs (4) with cased arguments, as they can form the same surface structure consisting of a noun with adposition, a stem (or noun), and a *-hata* verbal suffix. We consider 3 constructions: PVCs with non-predicative verbs (**PVC-n**), PVCs with possibly predicative verbs (**PVC-p**; denoted with † in Table 1), and verbs with cased arguments (**verb + arg**), in the order from most to least lexicalized. We compare them in 4 forms: adnominal (attributive) verb (1), main predicative verb (e.g. **향하다 *hyanghata*** 'face, head towards' in §4.1), verb modification (5), and serial verb modification.

Table 2 offers grammaticality judgments. Despite the same surface structure across the 4 forms, the grammaticality of these forms varies across

[9] The connective suffix forms are generally analyzed in the literature to form serial verb constructions that together denote a single event (Kim, 2010; Im and Lee, 2001).

[10] *관 kwan* is a Sino-Korean stem from Middle Chinese *關 kwaen* meaning 'relation' or 'barrier' (Baxter, 2010).

Forms	PVC-n	PVC-p	verb + arg
Adnominal verb	✓	✓	✓
Main predicate		✓ †	✓
Adn. verb mod.		✓ †	✓
Conn. verb mod.			✓

Table 2: Grammaticality of 3 categories of constructions with a PVC-like surface form. PVCs with non-predicative verbs are the most lexicalized, only able to appear in adnominal (attributive) forms. †Predicative verbs that form PVCs may undergo modification or serve as the main predicate in a sentence, but with such change they no longer carry adpositional meaning and instead behave like a regular verb-argument construction.

the 3 constructions. PVCs with non-predicative verbs, comprising the majority of PVCs, are the most lexicalized, only able to appear in adnominal and connective forms. While predicative verbs that form PVCs are more flexible, they are only PVCs when they do not appear as a main predicative verb, as they can be modified and undergo inflectional morphology as main predicates. They are PVCs when they appear as the nonfinal component of a serialized verb. We provide explicit examples and further explanations on grammaticality judgments of these constructions in Appendix C.

5 Related Work

We review studies of Korean MWEs and idiomaticity.

Traditional accounts. Traditional Korean grammar identifies several categories of idiomatic expressions. Examples include 관용어 *kwanyong-e* "habitual word"; 숙어 *swuk-e* "familiar word"; and 연어 *yen-e* "connected words, collocations". The boundaries of these categories tend to be fuzzy, with dictionaries disagreeing with each other. See Appendix A for further details.

Contemporary linguistic accounts. Linguistic accounts have offered taxonomical and theoretical analyses of Korean MWEs. Prior work on Korean MWEs and LVCs describes their lexical and morphological inflexibility, analyzes LVC subtypes (e.g., common noun vs. serial verb constructions, often involving Sino-Korean nouns), and provides generative accounts of nominal–light verb combinations, noting the diverse realizations of light verbs and adjectives (Han and Rambow, 2000; Lee, 2011; Im and Lee, 2001; Chae, 1996; Bak, 2012). The form 하다 ***hata*** 'to do' has been a particular subject of study in its range of grammatical and semantic realizations, as a light verb, as a light adjective, or as a suffix (Han, 2000). It is important to note that while verbs that appear in PVCs contain *-하다 hata*, they are suffixes that attach to bound stems rather than light or support verbs that attach to nouns (Chae, 1996, 2013).

Postpositional MWEs. Early work on modern Korean discusses postpositional phrases and compound postpositions (Underwood, 1890; Roth, 1937). More recent work following Kim (2002) has used the term *postposition equivalents* for units that function like postpositions. While many postposition equivalents are cased nouns denoting spatial relations (e.g. 앞에 **ap-ey** front-LOC '**at** the front, **in front of**'; Suh, 2004), some Korean postposition equivalents have recently been described as adpositional multiword expressions that are metaphoric (Han et al., 2024), although the authors focus on (experiential) metaphoricity and idiomaticity of these MWEs rather than their MWE-hood, composition, or structure. Other postposition equivalents have been analyzed as collocations or ongoing processes of grammaticalization (Kim, 2002; Moon, 2015). No prior work has focused on PVCs.

Korean MWE processing for machine translation. Outside of linguistics, work in machine translation has discussed Korean MWEs, including the PVCs we discuss in this paper, despite inconsistent terminology. In their study of the challenges of English *suk-e* in English-Korean machine translation, Lee and Kim (1993) discuss English MWEs and attempt to create a dictionary mapping each of them to Korean counterparts. Finally, in a study of Korean-to-Japanese multiword "translation units," Moon and Lee (2000) include a discussion on "semi-words" **-**를 위한 **lul wihan** 'for' and **-**のための **no tameno** ('for'), the former of which we describe as a PVC in this paper.

6 Conclusion

In this paper, we describe a class of Korean multiword adpositions: postpositional verb-based constructions (PVCs). We offer a list of 14 PVCs extracted from a corpus (Table 1), analyze their verbs and their distributions (§4.1, §4.2), and compare them to similar constructions with the same surface form (§4.3). This work is the first of its kind to analyze Korean multiword adpositions and draw a connection to cross-lingual annotation frameworks like PARSEME.

Limitations and future work

We have highlighted postpositional verb-based constructions (PVCs) in Korean and examined their linguistic properties towards annotating them in the PARSEME framework. Future work should pursue comprehensive annotation of Korean adpositional MWEs—including PVCs—in corpus data. This will shed light on properties such as ambiguity and fossilization.

Our analysis is not exhaustive: we note distributional variation across domains. For example, the gerundial -*함 ham* suffixes that denote a state are frequently attested in legal text. Our analysis will benefit from future work on such variation.

Our corpus analysis relies on the `Mecab` morphological analyzer (Kudo, 2013). While efficient and effective, the analyzer is not error-free; we thus do not report various metrics we measure in this paper, including the number of adpositions, suffixes, and sequences with which each stem occurs. The metrics are still available on our codebase.

In addition, comparative analysis with similar Japanese constructions formed with a postposition, Sino-Japanese stem, and an inflection of Japanese suffix -*する suru* (e.g., -に対して **ni taishite** 'about') may yield fruitful insights. Note Sino-Japanese character 対 corresponds to *대 tay* in -에 대한 **ey tayhan** 'about' from Sino-Korean 對.

Ethics statement

In our research, we made use of publicly available datasets published on the web. We acknowledge that the data obtained from the web may contain potential biases. We ensured that all datasets employed in our study were accessed and used in a manner that respects their intended use and complies with any associated licenses or terms of service.

We disclose our use of Gemini[11] as a coding assistant. We acknowledge the environmental and ethical considerations associated with the use of such AI technology, and have thoroughly reviewed the content to ensure that it does not include any unethical material.

Acknowledgements

This paper is based on a project for the Georgetown University course "All About Prepositions," taught by Nathan Schneider. We thank Youngjin Kim and Youngbin Noh for their discussions on native speaker acceptability judgments; Wesley Scivetti, Yujin Seo, and students in All About Prepositions for their helpful comments on the project; and Amir Zeldes and Kohei Kajikawa for their help with Japanese analogs. This research was supported in part by NSF award IIS-2144881.

[11] https://gemini.google.com/

References

Tatsuya Aoyama, Chihiro Taguchi, and Nathan Schneider. 2024. J-SNACS: Adposition and case supersenses for Japanese joshi. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9604–9614, Torino, Italia. ELRA and ICCL.

Aryaman Arora, Nitin Venkateswaran, and Nathan Schneider. 2021. SNACS annotation of case markers and adpositions in Hindi. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 454–458, Online. Association for Computational Linguistics.

Jaehee Bak. 2012. *The light verb construction in Korean*. Ph.D. thesis, University of Toronto.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of natural language processing*, 2:267–292.

W.H. Baxter. 2010. *A Handbook of Old Chinese Phonology*. Trends in Linguistics. Studies and Monographs [TiLSM]. De Gruyter.

N.R. Cattell. 1984. *Composite Predicates in English*. Syntax and semantics. Academic Press.

Hee-Rahk Chae. 1996. Light verb constructions and structural ambiguity. In *Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation*, pages 99–107, Seoul, Korea. Kyung Hee University.

Hee-Rahk Chae. 2013. Myths in Korean morphology and their computational implications. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, pages 505–511, Taipei, Taiwan. Department of English, National Chengchi University.

W D Chang. 2024. Korean Wikipedia 2024-05-01.

Haehyeong Choi. 2016. 베트남어와 한국어의 성어(成語) 개념 비교 연구 (A Comparative Study of the Concept of Idioms in Vietnamese and Korean). *인문학연구 (Humanities Research)*, 52:305–332.

Doosan Corporation. n.d.a. Kwanyonggu [idiom] (관용구). In Doosan Encyclopedia.

Doosan Corporation. n.d.b. Sugeo [idiom] (숙어). In Doosan Encyclopedia.

Doosan Donga. 2003. *New Korean Dictionary (새국어 사전)*, 4th edition. Doosan Donga, Seoul.

H.A. Giles. 1873. *A Dictionary of Colloquial Idioms in the Mandarin Dialect*. A.H. De Carvalho.

Chung-hye Han and Owen Rambow. 2000. The Sino-Korean light verb construction and lexical argument structure. In *Proceedings of the Fifth International Workshop on Tree Adjoining Grammar and Related Frameworks (TAG+5)*, pages 93–100, Université Paris 7.

Jeong Han Han, Myunghee Cha, and Hye Gyeong Yoon. 2024. 국어의 조사를 대체하는 다단어 표현 연구: 체계기능언어학의 경험적 은유의 관점 a study on multiword expressions that substitute for Korean postpositional markers: From the perspective of experiential metaphors in Systemic Functional Linguistics. *The Journal of Learner-Centered Curriculum and Instruction*, 24:781–801.

Sunhae Han. 2000. *Les prédicats nominaux en coréen: Constructions à verbe support hata*. PhD thesis, Université Paris Diderot (Paris 7), Paris, France. Thèse de doctorat dirigée par Maurice Gross.

Phuong Nguyen Hoang, Jung In Woong, An Mi Ji, Park Yoon Seo, Jun Ji Hyeon, and Kim Ju Hyun. 2022. Vietnamese and South Korean proverbs and idioms of social relations in comparisons. *South Asian Research Journal of Humanities and Social Sciences*.

Jena D. Hwang, Hanwool Choe, Na-Rae Han, and Nathan Schneider. 2020. K-SNACS: Annotating Korean adposition semantics. In *Proceedings of the Second International Workshop on Designing Meaning Representations*, pages 53–66, Barcelona Spain (online). Association for Computational Linguistics.

Seohyun Im and Chungmin Lee. 2001. Type construction of nouns with the verb ha- 'do'. In *Proceedings of the 16th Pacific Asia Conference on Language, Information and Computation*, pages 103–112, Jeju, Korea. The Korean Society for Language and Information.

Otto Jespersen. 1965. *A Modern English Grammar on Historical Principles: Morphology*. Allen & Unwin.

Jong-Bok Kim. 2010. Argument composition in Korean serial verb constructions. *Studies in Modern Grammar*, (61):1–24.

Seonhyo Kim. 2002. *현대 국어의 관형어 연구: A Study on Noun Modifiers in Modern Korean*. Ph.D. thesis, Seoul National University.

Taku Kudo. 2013. MeCab: Yet another part-of-speech and morphological analyzer.

Hoseok Lee and Youngtaek Kim. 1993. 영어-한국어 기계번역을 위한 연어와 숙어 트랜스퍼 사전 (A Collocation and Idiom Transfer Dictionary for English–Korean Machine Translation). *(구) 정보과학회논문지 (Journal of the Information Science Society)*, 20(7):976–987.

Juwon Lee. 2011. Two types of Korean light verb constructions in a typed feature structure grammar. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 40–48, Portland, Oregon, USA. Association for Computational Linguistics.

Geunseok Lim. 2011. 한국어 연어 연구의 전개와 쟁점에 대하여 (On the Development and Issues in Korean Collocation Research). *국어학 (國語學) (Korean Linguistics)*, 61:359–387.

S.E. Martin. 1992. *A Reference Grammar of Korean: A Complete Guide to the Grammar and History of the Korean Language*. Tuttle language library. C.E. Tuttle.

Byoungyul Moon. 2015. *한국어 조사 상당 구성에 대한 연구: Study on the constructions corresponding to particles in the Korean language*. Ph.D. thesis, Seoul National University.

Kyonghi Moon and Jong-Hyeok Lee. 2000. Representation and recognition method for multi-word translation units in Korean-to-Japanese MT system. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

National Institute of Korean Language. 1999. *Standard Korean Language Dictionary (표준국어대사전)*. Doosan Donga, Seoul. Government-authorized edition.

Eunjeong L. Park and Sungzoon Cho. 2014. KoNLPy: Korean natural language processing in Python. In *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology*, Chuncheon, Korea.

L. Roth. 1937. *Han-moun: Hilfsbuch zur Grammatik der koreanischen Sprache*. Abtei St. Benedikt.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002 Mexico City, Mexico, February 17–23, 2002 Proceedings 3*, pages 1–15. Springer.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika

Vincze, and Abigail Walsh. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. 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, et al. 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)*.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Archna Bhatia, Na-Rae Han, Tim O'Gorman, Sarah R. Moeller, Omri Abend, Adi Shalev, Austin Blodgett, and Jakob Prange. 2022. Adposition and case supersenses v2.6: Guidelines for English. *Preprint*, arXiv:1704.02134.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend. 2018. Comprehensive supersense disambiguation of English prepositions and possessives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 185–196, Melbourne, Australia. Association for Computational Linguistics.

Jiyeon Shim. 2009. 국어 관용어의 인지의미론적 연구 (A Cognitive Semantic Study of Korean Idiomatic Expressions). *고려대학교 대학원 박사학위 논문 (PhD Dissertation, Graduate School of Korea University)*.

H.M. Sohn. 2001. *The Korean Language*. Cambridge Language Surveys. Cambridge University Press.

Kyoungsook Suh. 2004. 현대 국어의 조사 상당어에 대한 연구: A study on constructions corresponding to particles in Modern Korean. Master's thesis, Seoul National University.

H.G. Underwood. 1890. *한영문법: An Introduction to the Korean Spoken Language*. Kelly & Walsh.

Wikimedia Foundation. 2024. Wikimedia downloads.

J. Yeon and L. Brown. 2019. *Korean: A Comprehensive Grammar*. Routledge Comprehensive Grammars. Taylor & Francis.

A Traditional approaches to Korean idiomatic expressions

Traditional frameworks in Korean grammar comment on the use and history of idiomatic expressions. Following Sag et al.'s (2002) definition of the multiword expression as "lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical syntactic, semantic, pragmatic, and/or statistical idiomaticity," we discuss the following terms: 관용어 *kwanyong-e* "habitual word"; 숙어 *swuk-e* "familiar word"; 성어 *seng-e* "word from the old times"; 연어 *yen-e* "connected words," 'collocations'; 속담 *sokdam* "earthly talk," 'proverbs'.

Korean *yen-e* and *sokdam* exhibit categorical differences, as they can respectively be compared to collocations and proverbs (Lim, 2011; Hoang et al., 2022), although Shim (2009) describes "syntactic *kwanyong-e*" to be "words that must go with each other": collocations.

Doosan Donga (2003) defines *kwanyong-e* as "expressions used habitually by the ordinary person", *swuk-e* as "expressions that comprise of two or more words frequently used to function as a single word," and *seng-e* as "frequently quoted expressions since the old times." Despite distinct definitions offered, there is no clear categorical difference between the terms. Some dictionaries equate the terms *kwanyong-e*, *suk-e*, and *seng-e* (National Institute of Korean Language, 1999; Doosan Donga, 2003); others explicitly differentiate them. Doosan Encyclopedia provides English examples in its entries of the terms. Example of a *kwanyong-e* is "It rains cats and dogs" and "He kicked the bucket" (Doosan Corporation, n.d.a) while an example of *suk-e* is "get in" (Doosan Corporation, n.d.b), whose function as a single word is emphasized. From this, we infer the encyclopedia categorically distinguishes the highly idiomatic "cats and dogs" and "kick the bucket" from other shorter, less idiomatic light verb constructions, prepositional verbs, and verb-particle constructions. On the other hand, a *seng-e* is explained to usually have a Chinese origin, possibly due to influence from its subcategory *saca-seng-e*, old 4-letter Chinese proverbs (Giles, 1873; Choi, 2016). Overall, we find that while there are several terms that can refer to multi-word expressions, their boundaries are often unclear and fuzzy to the point that dictionaries disagree with each other.

B Exceptions to PVC syntactic limitations

There are three notable exceptions to the limitations that verbs that form PVCs take: **속하다 *sokhata*** 'be a part of', **향하다 *hyanghata*** 'face, head towards', and **기하다 *kihata*** 'set as time for something to begin' as outlined in Table 1. These verbs can partici-

pate in the predicative from without deep change in meaning (10). We treat these as PVCs due to their restricted distribution and the general preference for corresponding PVC forms or alternative constructions (11).

(10) *하늘로* *갑자기* *향했다*
hanul-lo kapcaki hyangha-yss-ta
sky-DAT suddenly head-PAST-DECL

'suddenly headed towards the sky'

(11) *하늘로* 향해 *날아가다*
hanul-**lo** **hyangha-y** nalaka-ta
sky-DAT head-CONN fly-DECL

'fly towards the sky'

C Grammaticality judgments of PVCs and related constructions in various forms

In this section, we provide glossed examples of various constructions and their grammaticality judgments discussed in Table 2 and §4.3. For each of the 3 constructions, we assign an event to be consistent in this section. The non-predicative PVC example will continue featuring the book about crabs (1 et al.). The predicative PVC example will feature a ball headed towards the sky (10, 11). The regular verb-argument construction examples will feature a puppy that rescued a friend. All three verb lemmas are similar in appearance with a single-character bound stem and a *-hata* suffix: **관하다 *kwanhata*** 'relate,' a non-predicative PVC verb (**PVC-n**), **향하다 *hyanghata*** 'go towards,' a possibly predicative PVC verb (**PVC-p**), and **구하다 *kuhata*** 'rescue,' a regular verb with a cased argument.

Adnominal forms. All three verbs are able to take the adnominal form (12–14).

(12) *게에* 관한 *책*
key-**ey** **kwanha-n** chayk
crab-OBL relate-ADN book

'a book about crabs' [PVC-n]

(13) *하늘을* 향한 *공*
hanul-**ul** **hyangha-n** kong
sky-ACC face-ADN ball

'a ball (headed) towards the sky' [PVC-p]

(14) *친구를* 구한 *강아지*
chinku-**lul** kuha-n kangaci
friend-ACC rescue-ADN puppy

'a puppy that rescued a friend' [verb+arg]

Main predicate forms. The non-predicative PVC's verb is unable to serve as a main predicate of a sentence (15). While the predicative PVC's can (16), it no longer forms a multiword expression, and behaves like the verb-argument construction.

(15) * *책이* *게에* 관했다
chayk-i key-**ey** **kwanha-yss-ta**
book-NOM crab-OBL relate-PAST-DECL

'A book was about crabs.' [PVC-n]

Example (15) is generally grammatically unacceptable. However, in legal texts, the same construction is acceptable and even frequent when the verb takes the gerundial *-함 ham* suffix (e.g. to signal a state in which a book is about crabs).

(16) *공이* *하늘을* *향했다*
kong-**i** hanul-**ul** hyangha-yss-ta
ball-NOM sky-ACC face-PAST-DECL

'A ball headed towards the sky.' [PVC-p]

While grammatically acceptable, the verb in (16) carries predicative meaning instead of the adpositional meaning carried by (13). In this usage, then, **향하다 *hyanghata*** 'to head' is treated as a full lexical verb akin to the non-PVC verb **구하다 *kuhata*** 'to rescue' (below in 17), which is subject to no restriction as a matrix verb.

(17) *강아지가* *친구를* *구했다*
kangaci-**ka** chinku-**lul** kuha-yss-ta
puppy-NOM friend-ACC rescue-PAST-DECL

'A puppy rescued a friend.' [verb+arg]

Adnominal verb modification. Attempting to modify the verb shows that PVC verbs are not able to undergo modification without ungrammaticality (18) or shift in meaning (PVC to regular verb with argument; 19). The PVC as a whole may be freely modified (20, 21). The regular verb may be modified with or without the argument freely (22).

(18) * *게에* *명백히* 관한 *책*
key-**ey** myengpaykhi **kwanha-n** chayk
crab-OBL clearly relate-ADN book

'a book clearly about crabs' [PVC-n]

(19) *하늘을* *확실히* 향한 *공*
hanul-**ul** hwaksilhi **hyangha-n** kong
sky-ACC surely face-ADN ball

'a ball that is surely headed towards the sky' [PVC-p]

(20) *명백히* *게*에 관한 *책*
myengpaykhi key-**ey** **kwanha-n** chayk
clearly crab-OBL relate-ADN book

'a book clearly about crabs' [PVC-n]

In (19), insertion of an adverb inside what would have been a PVC/MWE still results in a grammatical sentence. This strongly points to an alternative analysis for this sentence, one in which *하늘*을 **향한** *공* *hanul*-**ul** ***hyanghan*** *kong* 'ball towards the sky' forms a relative clause, as reflected in the translation. This would, conversely, mean that treatment of (21) could go either way, i.e., one that involves a PVC/MWE *or* one with a relative clause.

(21) *확실히* *하늘*을 향한 *공*
hwaksilhi hanul-**ul** **hyangha-n** kong
surely sky-ACC face-ADN ball

'a ball surely (headed) towards the sky' [PVC-p]

(22) *친구*를 *용감하게* 구한 *강아지*
chinku-**lul** yongkamhakey kuha-n kangaci
friend-ACC courageously rescue-ADN puppy

'a puppy who courageously rescued a friend' [verb+arg]

Connective verb modification. In a verb serialization construction, the final verb is the main predicate (Kim, 2010). In this case, PVC verbs take the connective suffix to form PVCs, and cannot be modified (23, 24). Even non-predicative PVC verbs cannot take the predicative form due to their non-final verb position and remain lexicalized. Similarly to (20, 21), PVCs as a whole can be modified. Non-PVC verbs can be freely modified (25).

(23) * *게*에 *명백히* 관해
key-**ey** myengpaykhi **kwanha-y**
crab-OBL clearly relate-CONN
서술한 *책*
seswulhan chayk
describe-ADN book

'book written clearly about crabs' [PVC-n]

(24) * *하늘*을 *확실히* 향해 *날아간* *공*
hanul-**ul** hwaksilhi **hyangha-y** nalakan kong
sky-ACC surely face-CONN fly-ADN ball

'ball flying while surely heading towards the sky' [PVC-p]

(25) *친구*를 *용감하게* *구해*
chinku-**lul** yongkamhakey kuha-y
friend-ACC courageously rescue-CONN
살린 *강아지*
salli-n kangaci
save-ADN puppy

'puppy who courageously rescued and saved the friend' [verb+arg]

PolyFrame at MWE-2026 AdMIRe 2: When Words Are Not Enough: Multimodal Idiom Disambiguation

Nina Hosseini-Kivanani
University of Luxembourg & RTL
Esch-sur-Alzette, Luxembourg
nina.hosseinikivanani@uni.lu

Abstract

Multimodal models struggle with idiomatic expressions due to their non-compositional meanings, a challenge amplified in multilingual settings. We introduced PolyFrame, our system for the MWE-2026 AdMIRe 2 shared task on multimodal idiom disambiguation, featuring a unified pipeline for both image+text ranking (Subtask A) and text-only caption ranking (Subtask B). All model variants retain frozen CLIP-style vision–language encoders and the multilingual BGE M3 encoder, training only lightweight modules: a logistic regression and LLM-based sentence-type predictor, idiom synonym substitution, distractor-aware scoring, and Borda rank fusion. Starting from a CLIP baseline (26.7% Top-1 on English dev, 6.7% on English test), adding idiom-aware paraphrasing and explicit sentence-type classification increased performance to 60.0% Top-1 on English, and 60.0% Top-1 (0.822 NDCG@5) in zero-shot transfer to Portuguese. On the multilingual blind test, our systems achieved average Top-1/NDCG scores of 0.35/0.73 for Subtask A and 0.32/0.71 for Subtask B across 15 languages. Ablation results highlight idiom-aware rewriting as the main contributor to performance, while sentence-type prediction and multimodal fusion enhance robustness. These findings suggest that effective idiom disambiguation is feasible without fine-tuning large multimodal encoders.

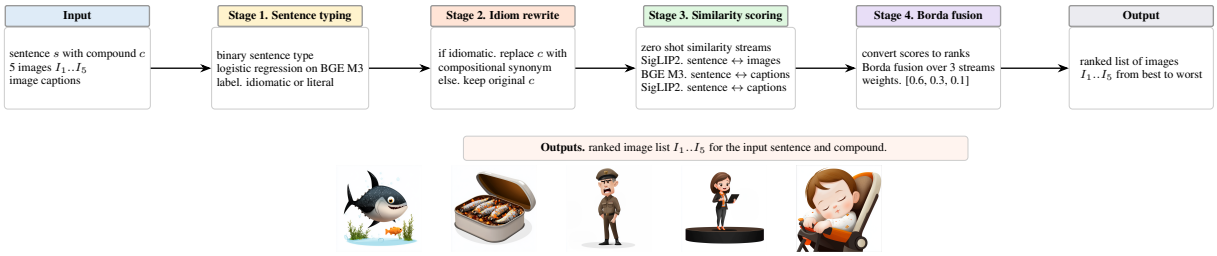
1 Introduction

Idiomatic expressions are a long-standing problem for computational semantics because their figurative meanings are not compositionally predictable from surface form (Flor et al., 2025; Zeng and Bhat, 2021). Models trained primarily on literal data often interpret idioms word by word, which leads to semantic drift in downstream tasks such as translation, retrieval, and captioning (Pickard et al., 2025a). This effect is amplified in multilingual settings, where idioms differ in lexicalisation and transparency across languages (Domhan et al., 2022; Cap et al., 2015). Recent work treats idiomaticity as a structured prediction problem, for example, by classifying usages as literal or idiomatic or by ranking contextually appropriate interpretations (Endalie et al., 2023).

AdMIRe 2.0 provides sentences containing potentially idiomatic nominal compounds together with five candidate images and parallel caption-based variants (Arslan et al., 2026). Subtask A asks systems to rank images according to how well they reflect the intended meaning of the compound in context. Subtask B replaces images with captions and prompts to probe text-only reasoning under the same schema. Both subtasks implicitly require robust sentence-type judgements (idiomatic vs. literal) and fine-grained alignment of the chosen sense with visual or textual cues.

This paper presents our system description for AdMIRe 2.0 (Arslan et al., 2026). We build a shared data and evaluation pipeline and instantiate three variants within a single architecture: (i) a CLIP-based multimodal baseline for Subtask A; (ii) an improved multimodal ranker with supervised sentence-type classification (logistic regression plus a literal-first LLM classifier), idiom synonym replacement, distractor-aware scoring, and fusion of visual and caption-based scores; and (iii) a text-only counterpart of the improved ranker for Subtask B (Figure 1). All systems use frozen CLIP-style vision–language encoders and the multilingual BGE M3 sentence encoder (Chen et al., 2024), with learning restricted to lightweight idiomaticity classifiers, scalar fusion weights, and rank-level Borda ensembles over multiple CLIP and LLM backends. Our experiments examine how sentence-type prediction, idiom-aware rewriting, and multimodal fusion jointly improve idiom-sensitive image and caption ranking across languages.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 127–133



Example candidate images for the idiom “big fish”.

Figure 1: Overview of the final POLYFRAME pipeline. Sentence typing via logistic regression, idiom replacement for idiomatic cases, three zero shot similarity streams with SigLIP2 and BGE M3, and Borda fusion

2 Related Work

Idiom processing and idiomaticity detection. Text-based idiom processing has been studied extensively through shared tasks and benchmarks that ask models to decide whether a target expression is used literally or idiomatically in context (e.g., (Jakhotiya et al., 2022; Boisson et al., 2022)). SemEval 2022 Task 2 introduced multilingual idiomaticity detection for verbal multiword expressions and showed that glosses, translations, and lexical knowledge bases can improve cross-lingual disambiguation (Phelps et al., 2022). Subsequent work has evaluated large language models on idiom detection and translation, reporting that they still over-predict idiomatic readings in literal contexts and require careful prompting to approach the performance of supervised baselines (Phelps et al., 2024). Other studies have proposed targeted metrics for evaluating idiom translation in neural machine translation and documented a strong tendency toward literal renderings (Khan and Akter, 2025). The AdMIRe shared tasks extend this line of work by framing idiomaticity as a multimodal ranking problem and by providing graded relevance annotations for sentence type and image suitability (Torunoğlu-Selamet et al., 2026; Pickard et al., 2025b).

Multimodal idiom understanding and vision–language models. Multimodal encoders such as CLIP learn aligned image–text representations via large-scale contrastive learning and have become standard backbones for zero-shot classification and retrieval. However, their pretraining data and objectives are not tailored to figurative language, and idiomatic usages are often under-represented. Recent approaches adapt CLIP to idiom interpretation by enriching prompts with natural-language definitions, few-shot examples, or idiomatic paraphrases, and by combining visual similarity with textual sentence embeddings (Markchom et al., 2025; Wang et al., 2025). AdMIRe 2.0 offers a test bed for such strategies by providing both image-based and caption-based variants of the same task. Our systems follow this direction: we keep OpenCLIP-style encoders frozen, pair them with BGE M3 for multilingual sentence embeddings, and focus learning capacity on lightweight idiomaticity classifiers and fusion mechanisms that combine visual, caption-based, and definition-based scores within a single architecture.

3 Methodology[1]

We describe our systems for the two AdMIRe 2.0 subtasks, which jointly probe sentence-level idiomaticity and multimodal disambiguation (see Figure 1 for an overview). Subtask A presents a context sentence containing a potentially idiomatic nominal compound and five candidate images; the system must rank the images by appropriateness. Subtask B reuses the same TSV schema but replaces images with captions and prompts, yielding a text-only ranking task. Sentence-type prediction is evaluated by accuracy, and ranking by Top-1 accuracy and NDCG@5.

3.1 Task setup and system variants

For Subtask A, the organisers provide supervised splits for English (EN) and Portuguese (PT) with train, development, test, and cross-evaluation partitions. Each instance consists of a sentence, the target compound, five image file names organised under idiom-specific directories, and optionally image captions and prompts. For evaluation, we run our system on the official blind-test bundles covering 15 languages, resolving image paths via a language-aware global root and tracking the number of processed instances per language (approximately 50–360) to ensure full coverage.

[1]https://github.com/NinaKivanani/PolyFrame

For Subtask B, we reuse the TSV structure but operate in text-only mode: image identifiers are retained for compatibility, while ranking uses captions and prompts only. Within a shared architecture, we instantiate three system variants to study the impact of modality, sentence-type modelling, and idiom replacement: (i) a CLIP-based multimodal baseline for Subtask A, (ii) an improved multimodal ranker with explicit sentence-type classification, idiom replacement, and distractor-aware fusion, and (iii) a text-only variant of the improved ranker for Subtask B.

3.2 Representations and encoders

Each TSV row is parsed into a context sentence x, a target compound c, five image identifiers $\{i_1, \ldots, i_5\}$, and, when available, image captions $\{c_k\}_{k=1}^{5}$. Supervised splits additionally provide a gold sentence type $y_{\text{sent}} \in \{\text{idiomatic}, \text{literal}\}$ and a gold ranking $\pi^* = [\pi_1^*, \ldots, \pi_5^*]$.

All systems use vision-language encoders for multimodal understanding. The baseline uses a multilingual CLIP variant (xlm-roberta-large-ViT-H-14) with frozen LAION-5B weights. The improved systems employ a dual-encoder architecture: SigLIP2 (ViT-SO400M-14-SigLIP2-378) (Tschannen et al., 2025) as the primary vision-language encoder for image-text similarity, combined with BGE-M3 as a separate multilingual text encoder for text-only ranking via caption matching. Rankings from both encoders are fused using Borda-count aggregation. Additional CLIP or SigLIP models can be added in ensemble mode with weighted fusion.

For sentence and caption representations we additionally use BGE M3, a multilingual dense retrieval encoder. Given a sentence–compound pair (x, c), we concatenate them and obtain an embedding $g(x, c) \in \mathbb{R}^d$, which provides text-only similarity scores for caption-based ranking and serves as the feature space for a lightweight binary sentence-type classifier.

3.3 Sentence-type prediction and ranking

Sentence-type prediction. Sentence-type prediction combines a supervised classifier and an LLM-based component. We train a two-way logistic regression model on EN data using BGE M3 features $g(x, c)$ to predict $y_{\text{sent}} \in \{\text{literal}, \text{idiomatic}\}$. This serves as the primary sentence-type signal when available.

As a complementary mechanism and cross-lingual fallback, we use literal-first prompting with external LLMs (GPT-4o, Qwen3-32B, Llama 3.1-70B, DeepSeek-v3, Mistral). For each compound–language pair, the system first asks the LLM to generate a small set of clearly literal example sentences, which are cached. The final prompt presents these examples together with the target sentence and instructions to decide between LITERAL and IDIOMATIC. We allow brief reasoning but constrain the output to a single label.

Baseline multimodal ranker. The baseline first predicts sentence type heuristically, then constructs prompts, and finally performs visual and textual ranking. Given the predicted type, we build a small set of text queries that combine the sentence, the compound, short natural-language definitions from a hand-curated idiom lexicon, and few-shot examples. These queries are encoded with the CLIP text tower and averaged to obtain a single query embedding q.

For visual ranking, each image i_k is mapped to an embedding $v_k = f_{\text{img}}(i_k)$, and similarity scores are computed via cosine similarity between q and v_k, optionally scaled by a temperature τ to sharpen the distribution over the five candidates; images are then ranked by descending similarity. Caption-based ranking is analogous, using captions instead of images. Visual and caption-based rankings are fused via a Borda-style scheme that converts rank positions into scores and interpolates them, typically with a higher weight on visual information.

Improved multimodal ranker. The improved ranker extends the baseline with supervised sentence-type classification, idiom synonym replacement, and distractor-aware fusion. For sentences classified as idiomatic, we rewrite the context by replacing occurrences of c with a compositional paraphrase that makes the figurative meaning explicit; the modified sentence is then used to construct CLIP prompts, while literal sentences are left unchanged.

The final ranking fuses three score streams: (1) SigLIP2 vision-language similarity (sentence $\leftrightarrow$ images), (2) BGE-M3 text similarity (sentence $\leftrightarrow$ captions), and (3) SigLIP2 text similarity (sentence $\leftrightarrow$ captions via SigLIP2's text encoder). In our submitted configuration we use weighted Borda-count aggregation with fixed weights of [0.6, 0.3, 0.1] for vision-language similarity, BGE-M3 text similarity, and SigLIP2 caption similarity, respectively, in image+text mode. For text-only mode, weights are adjusted to [0.0, 0.7, 0.3] to focus on text components while ignoring vision. A confidence measure

derived from the gap between the top two scores in each stream can adjust these weights. For non-English languages, an optional cross-lingual mode combines scores computed on the original sentence with scores computed on an English translation obtained from the same LLM used for literal-first classification.

Text-only ranker. The text-only system reuses the same architecture in caption-only mode. Image embeddings are set to zero and receive zero weight in the fusion stage; sentence-type prediction, idiom synonym replacement, and caption-based scoring with BGE M3 and SigLIP2 remain unchanged. This design ensures that improvements over random ranking are attributable purely to text modelling of captions and that ablations of the classifier, literal-first prompting, or idiom replacement are directly comparable across subtasks.

3.4 Ensembles, transfer strategies

Model ensembles and Borda fusion. To improve robustness and support zero-shot transfer, we employ ensembles at both encoder and classifier level with a unified rank-level fusion scheme. On the vision side, we optionally use an ensemble of CLIP variants. Our best-performing configuration uses a single SigLIP2 model (ViT-SO400M-14-SigLIP2-378) rather than an ensemble, as this provided optimal performance while maintaining efficiency. The system supports ensemble configurations where multiple CLIP or SigLIP variants can be combined using weighted Borda fusion, but the submitted system operated in single-model mode.

For sentence-type prediction, we evaluate several LLMs under the same literal-first prompting strategy; in the final configuration, we combine Qwen3-32B and GPT-4o with weights $[0.6, 0.4]$, selected based on development performance, while other LLMs act as fallbacks. In all cases, we apply weighted Borda fusion: each candidate's rank in each list is converted into a score (higher for better ranks), multiplied by the corresponding fusion weight, and summed. Candidates are then re-ranked by their aggregated Borda scores, providing a consistent mechanism for combining both modalities and model variants.

Zero-shot and few-shot transfer. To handle idioms and languages with limited training data, we support both zero-shot and few-shot strategies. When few-shot prompting is enabled, CLIP queries are augmented with hand-curated examples that illustrate literal and idiomatic usage (e.g., idiomatic *big fish* in a company vs. literal *big fish* in a fishing context), which are prepended to CLIP text queries to guide the vision–language encoder. For languages without training data, we rely on the multilingual capabilities of SigLIP2 and BGE M3: the system applies zero-shot classification directly to non-English sentences using the multilingual encoders, and when this fails, the system falls back to simple heuristics based on compound frequency in captions and basic lexical markers. This zero-shot setup yields reasonable performance (around 60% accuracy on PT) without language-specific training.

4 Results & Discussion

We evaluate our systems using the official AdMIRe 2.0 metrics, reporting Top-1 accuracy and NDCG@5. Experiments cover both subtasks and include baseline comparisons, strategy-specific variants, ablation studies, and the official blind-test evaluation on Codabench. We used a temperature of $\tau = 0.7$ for similarity scaling and Borda-count-based fusion as the default for final evaluation.

The CLIP-based baseline, which relies on frozen vision–language embeddings and heuristic sentence typing, achieves 26.7% Top-1 accuracy on the EN development set and 6.7% on the test set (see Table 1). This large drop highlights the difficulty of the task and confirms that naive multimodal retrieval is strongly biased toward literal interpretations when idiomatic compounds are present.

System	EN dev		EN test		PT dev (zero-shot)	
	Top-1 (%)	NDCG@5	Top-1 (%)	NDCG@5	Top-1 (%)	NDCG@5
CLIP baseline	26.7	0.655	6.7	0.607	–	–
+ Idiom replacement	60.0	0.800	–	–	–	–
+ LLM sentence typing	40.0	0.739	–	–	–	–
+ All improvements	60.0	0.797	–	–	60.0	0.822

Table 1: System performance across evaluation sets.

Introducing idiom synonym replacement yields the largest single improvement, raising Top-1 accuracy on the EN development set to 60.0% (Table 1). This confirms observations from AdMIRe 2024 that rewriting idiomatic expressions into compositional paraphrases substantially reduces literal bias in vision–language models. Literal-first sentence classification using an external LLM improves performance to 40.0% when used in isolation, indicating that sentence-type awareness is beneficial but insufficient without explicit paraphrasing. The best submission didn't use LLM classification. Combining sentence-type prediction, idiom replacement, and multimodal fusion in the improved ranker

System	Top$_1$ A	Top$_1$ B	ST Acc.
CLIP baseline (XLM-RoBERTa-ViT-H-14)	26.7	6.7	53.3
GPT-4o	*32.5*	*28.0*	*58.0*
Qwen3-32B	*31.0*	*26.5*	*56.5*
Llama3.1-70B	*29.5*	*25.0*	*55.0*
DeepSeek-v3	*28.0*	*23.5*	*53.5*
Mistral-large	*30.0*	*25.5*	*54.5*
Ensemble (Qwen3+GPT-4o)	***33.5***	***29.0***	***59.0***
Best submission	**35.0**	**32.0**	–

Table 2: Multilingual blind test results (15 languages). Best submission used SigLIP2 + idiom replacement with zero-shot classification.

maintains 60.0% Top-1 accuracy while providing greater robustness across prompt conditions and idiom types. Zero-shot transfer to PT also reaches 60.0% Top-1 and 0.822 NDCG@5, suggesting that the core strategies generalise across related languages.

Multilingual blind-test results. We submitted our systems to the official Codabench evaluation for both subtasks. On the multilingual blind test covering 15 languages, our submission achieved an average Top-1 accuracy of 0.35 and an NDCG of 0.73 for Subtask A, and 0.32 Top-1 accuracy with an NDCG of 0.71 for Subtask B. These scores correspond to a mid-range ranking among participating systems at the time of evaluation. Performance was consistent across languages, with comparable scores on Chinese and non-Chinese subsets, indicating stable cross-lingual behaviour in a fully zero-shot setting (Table 2).

For the blind-test phase, we apply the improved system unchanged to all 15 evaluation languages, using the official directory structure and file formats. The submitted configuration combines SigLIP2 with idiom replacement and zero-shot sentence-type classification. We verify full coverage for each language, with sample counts ranging from 48 to 363 instances (see Table 3 in the Appendix).

Ablation studies. Ablation experiments on the EN development set show that idiom replacement accounts for most of the performance gains, while sentence-type prediction mainly improves stability and reduces variance across idioms. Multimodal fusion consistently outperforms unimodal ranking, and Borda-style aggregation is more robust than alternative rank-combination strategies across hyperparameter settings. Temperature and fusion-weight sweeps show limited sensitivity around the chosen defaults, indicating that the improvements are not driven by fragile tuning.

Taken together, the results suggest that most of the attainable gains come from better framing of the vision–language matching problem rather than from stronger back-end language models (Radford et al., 2021). Idiom rewriting and multimodal fusion close much of the gap between the CLIP baseline and our best systems (Gao et al., 2024), while LLM-based sentence-type prediction yields smaller incremental gains at the cost of additional latency, API dependence, and prompt sensitivity (Jin et al., 2025). This trade-off is reflected in the final Codabench submission, which uses SigLIP2 with idiom replacement and zero-shot sentence typing, but omits LLM classification despite its slightly higher scores on the EN development set. The relatively modest drop from Subtask A to Subtask B and the stable NDCG values across languages indicate that caption-based reasoning can approximate image-based disambiguation when captions are informative, but also highlight persistent weaknesses on low-resource and culturally marked varieties such as Spanish Ecuador and Uzbek (Tschannen et al., 2025). Overall, the pattern of ablation results supports our design choice to prioritise frozen encoders and lightweight, transparent adaptations over heavy fine-tuning or tightly coupled LLM components (Xing et al., 2024).

Limitations and Future Work

A key limitation of our approach is that the idiom synonym replacement database is manually curated and currently covers only around 50 common English idioms, with limited coverage for the 14 non-English languages in the blind test. This restricts the effectiveness of idiom replacement for less frequent expressions and for many non-English instances. Future work will explore learned fusion weights as an alternative to our current fixed Borda-based scheme. Since our experiments so far rely mainly on decoder-style large language models, we also plan to incorporate encoder-based models (e.g., BERT) and encoder–decoder architectures (e.g., T5, mBART). These architectures may capture idiomaticity in different ways and provide complementary perspectives on representation learning for figurative language.

References

Doğukan Arslan, Rodrigo Wilkens, Wei He, Dilara Torunoğlu-Selamet, Thomas Pickard, Aline Villavicencio, Adriana S. Pagano, and Gülşen Eryiğit. 2026. MWE-2026 Shared Task 2: AdMIRe 2 - Advancing Multimodal Idiomaticity Representation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morocco. Association for Computational Linguistics.

Joanne Boisson, Jose Camacho-Collados, and Luis Espinosa Anke. 2022. Cardiffnlp-metaphor at semeval-2022 task 2: Targeted fine-tuning of transformer-based language models for idiomaticity detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 169–177.

Fabienne Cap, Manju Nirmal, Marion Weller, and Sabine Schulte Im Walde. 2015. How to account for idiomatic german support verb constructions in statistical machine translation. In *Proceedings of the 11th workshop on multiword expressions*, pages 19–28.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Tobias Domhan, Eva Hasler, Ke M Tran, Sony Trenous, Bill Byrne, and Felix Hieber. 2022. The devil is in the details: On the pitfalls of vocabulary selection in neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1861–1874.

Demeke Endalie, Getamesay Haile, and Wondmagegn Taye. 2023. Deep learning-based idiomatic expression recognition for the amharic language. *PLoS One*, 18(12):e0295339.

Michael Flor, Xinyi Liu, and Anna Feldman. 2025. A survey of idiom datasets for psycholinguistic and computational research. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, pages 90–100.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595.

Yash Jakhotiya, Vaibhav Kumar, Ashwin Pathak, and Raj Shah. 2022. Jarvix at semeval-2022 task 2: It takes one to know one? idiomaticity detection using zero and one-shot learning. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 165–168.

Yizhang Jin, Jian Li, Tianjun Gu, Yexin Liu, Bo Zhao, Jinxiang Lai, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xin Tan, and 1 others. 2025. Efficient multimodal large language models: A survey. *Visual Intelligence*, 3(1):27.

Muhammad Farmal Khan and Mousumi Akter. 2025. Evaluating large language models on urdu idiom translation. *arXiv preprint arXiv:2510.17460*.

Thanet Markchom, Tong Wu, Liting Huang, and Huizhi Liang. 2025. Uor-ncl at semeval-2025 task 1: Using generative llms and clip models for multilingual multimodal idiomaticity representation. *arXiv preprint arXiv:2502.20984*.

Dylan Phelps, Xuan-Rui Fan, Edward Gow-Smith, Harish Tayyar Madabushi, Carolina Scarton, and Aline Villavicencio. 2022. Sample efficient approaches for idiomaticity detection. *MWE 2022*, page 105.

Dylan Phelps, Thomas MR Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD)@ LREC-COLING 2024*, pages 178–187.

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, and Marco Idiart. 2025a. 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.

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, and Marco Idiart. 2025b. Semeval-2025 task 1: Admire–advancing multimodal idiomaticity representation. *arXiv preprint arXiv:2503.15358*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Dilara Torunoğlu-Selamet, Dogukan Arslan, Rodrigo Wilkens, Wei He, Doruk Eryiğit, Thomas Pickard, Adriana S. Pagano, Aline Villavicencio, Gülşen Eryiğit, Ágnes Abuczki, Aida Cardoso, Alesia Lazarenka, Dina Almassova, Amalia Mendes, Anna Kanellopoulou, Antoni Brosa-Rodríguez, Baiba Saulite, Beata Wojtowicz, Bolette Pedersen, and 59 others. 2026. A parallel cross-lingual benchmark for multimodal idiomaticity understanding. *Preprint*, arXiv:2601.08645.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer,

Ye Xia, Basil Mustafa, and 1 others. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

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 AAAI Conference on Artificial Intelligence*, volume 39, pages 25398–25406.

Jialu Xing, Jianping Liu, Jian Wang, Lulu Sun, Xi Chen, Xunxun Gu, and Yingfei Wang. 2024. A survey of efficient fine-tuning methods for vision-language models—prompt and adapter. *Computers & Graphics*, 119:103885.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

A Appendix

Table 3 summarises the official Codabench blind test results for our submission `nikoniko`. For Subtask A (image+text), our system achieved an average Top_1 accuracy of 0.35 and an NDCG of 0.73 across 15 languages. For Subtask B (text only), the corresponding scores were 0.32 Top_1 and 0.71 NDCG. In both subtasks, our run ranked fifth on the Codabench leaderboard at evaluation time. The per language breakdown shows clear variation across the multilingual blind test. For Subtask A, the best performing languages are Portuguese Brazil (PT-BR) (0.46), PT (0.42), and Norwegian (0.42), while Spanish Ecuador (ES-EC) (0.17), Uzbek (UZ) (0.29), and Georgian (KA) (0.27) emerge as the most challenging. Despite these differences in Top_1 accuracy, NDCG scores remain consistently high in the 0.69 to 0.76 range, with an overall average of 0.73.

Language	Subtask A (img+txt)		Subtask B (text)	
	Top_1 (%)	NDCG@5	Top_1 (%)	NDCG@5
Chinese (ZH)	35.0	0.72	28.0	0.69
Georgian (KA)	27.0	0.69	28.0	0.68
Greek (EL)	36.0	0.72	28.0	0.71
Igbo (IG)	33.0	0.73	29.0	0.69
Kazakh (KK)	33.0	0.74	28.0	0.72
Norwegian (NO)	42.0	0.75	41.0	0.75
Portuguese BR (PT-BR)	46.0	0.76	37.0	0.75
Portuguese (PT)	42.0	0.75	37.0	0.73
Russian (RU)	40.0	0.73	39.0	0.74
Serbian (SR)	39.0	0.74	31.0	0.71
Slovak (SK)	38.0	0.73	35.0	0.73
Slovenian (SL)	40.0	0.75	37.0	0.74
Spanish EC (ES-EC)	17.0	0.66	19.0	0.67
Turkish (TR)	34.0	0.71	32.0	0.70
Uzbek (UZ)	29.0	0.71	32.0	0.71
Macro average	35.4	0.72	32.1	0.72

Table 3: Per language performance on the multilingual blind test. Subtask A uses images and captions; Subtask B is text only.

IdiomRanker-X at MWE-2026 AdMIRe 2: Multilingual Idiom-Image Alignment via Low-Rank Adaptation of Cross-Encoders

Mehmet Utku Colak
Affiliation
colakme19@itu.edu.tr

Abstract

This paper describes the system submitted for the **MWE 2026 Shared Task** (AdMIRe 2.0 Subtask A). The submission focused on a text-centric approach, reframing the idiom-image alignment task as a sentence-pair classification problem using **mBERT** (Multilingual BERT). The submitted system relied on full fine-tuning using only the English training data, achieving a Top-1 Accuracy of approximately **0.30** on the blind test set.

Following the evaluation phase, significant limitations were identified in the cross-lingual generalization of the base model. In a post-evaluation study, the backbone was upgraded to **XLM-RoBERTa-Large-XNLI**, incorporating **Low-Rank Adaptation (LoRA)** and utilizing the full multilingual dataset with hard negative mining. These improvements boosted the accuracy to **0.41**, demonstrating the necessity of NLI-specific pre-training and parameter-efficient tuning for MWE-aware multimodal tasks.

1 Introduction

Idiomatic expressions—fixed phrases such as *"turn over a new leaf"* or *"spill the beans"*, or in Turkish, *"çürük elma"* (which corresponds to the English idiom *"bad apple"*)—pose a persistent challenge for Natural Language Processing (NLP) systems. Their meaning is often non-compositional, meaning it cannot be directly inferred from the sum of their constituent words (Pickard et al., 2025; Bobrow and Bell, 1973). While humans naturally integrate cultural knowledge and context to resolve this ambiguity, computational models frequently struggle, defaulting to literal interpretations that fail to capture the intended figurative semantics (Pan et al., 2025).

This limitation is particularly acute in current state-of-the-art Large Language Models (LLMs) and Vision-Language Models (VLMs). Despite their success on general benchmarks, these models exhibit a significant "literal bias," often failing to grasp the figurative nuance required for tasks like sentiment analysis, machine translation, and multimodal understanding (Mi et al., 2024; Phelps et al., 2024). For instance, a VLM prompted with *"eager beaver"* is more likely to generate or retrieve an image of an enthusiastic animal rather than an industrious person (Pickard et al., 2025).

To address this, the **MWE 2026 Shared Task: AdMIRe 2.0** was established to evaluate and improve the ability of models to interpret idioms in multimodal contexts (Arslan et al., 2026). The task utilizes a new parallel cross-lingual benchmark (Torunoğlu-Selamet et al., 2026) to shift the focus from simple classification to semantic alignment, requiring systems to rank images based on their relevance to a specific idiomatic sense within a context sentence.

1.1 Related Work

Text-Based Idiom Processing Early computational approaches to idiomaticity focused on supervised binary classification to distinguish between literal and figurative usage, often relying on syntactic patterns and lexical co-occurrences (Fazly et al., 2009). Subsequent research utilized word embeddings to predict compositionality, leading to benchmarks such as SemEval-2022 Task 2, which evaluated multilingual idiomaticity detection and sentence embedding (Tayyar Madabushi et al., 2022). However, concerns have been raised that text-only benchmarks may contain artifacts that allow models to perform well without achieving true semantic understanding (Boisson et al., 2023).

The Shift to Multimodal Representation Recent work has introduced the visual modality as a more rigorous test of semantic comprehension. Datasets like those introduced in the original AdMIRe (1.0) task build upon previous studies on

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 134–138

noun compound interpretation and paraphrase, incorporating both static images (Subtask A) and visual-temporal sequences (Subtask B) to capture the dynamic nature of certain expressions (Pickard et al., 2025). This multimodal setting has proven challenging for standard VLMs to handle, confirming the complexity of cross-modal idiomatic alignment (Yosef et al., 2023).

Approaches in Previous Iterations Participants in the previous AdMIRe shared task (SemEval-2025) explored various architectures to bridge the semantic gap. A common strategy involved **multimodal pipelines**, where visual features from Vision Transformers (ViTs) were fused with textual embeddings from models like BERT or XLM-RoBERTa to predict image relevance (Pan et al., 2025).

Alternative approaches leveraged the **in-context learning** capabilities of Generative LLMs (e.g., Llama-3, GPT-4). For example, Pan et al. (2025) demonstrated that while LLMs can achieve high performance in English through zero-shot prompting, alignment in lower-resource settings like Portuguese often benefits more from specialized fine-tuned encoders like XLM-RoBERTa combined with vision encoders. These findings highlight that while LLMs are powerful, they often require sophisticated prompting strategies or ensemble methods—such as Mixture-of-Experts (MoE)—to smooth over their inherent inconsistency in representing idiomaticity (Pickard et al., 2025).

2 Task Description and Dataset

The AdMIRe 2.0 Subtask A focuses on the challenge of **Multimodal Idiom Alignment**. The objective is to rank a set of candidate images based on their semantic correspondence to a specific expression within a context sentence. This requires the model to first implicitly determine whether the target phrase is being used in its **idiomatic** (figurative) sense or its **literal** (compositional) sense, and then select the visual representation that matches that specific meaning.

For example, consider the input sentence: *"The place got quite lively at one stage as a hen party moved in, with the bride-to-be in fancy dress with large balloons tied onto her."* Here, the phrase *fancy dress* functions as a Multiword Expression (MWE) referring to a costume, rather than a literal "formal dress" that is "fancy." The model must detect this non-compositional usage and prioritize images depicting costumes over those depicting formal evening wear. Figure 1 illustrates a sample prediction where the proposed system successfully identifies the correct figurative context.

Figure 1: Sample output generated by the proposed system. The model correctly ranks the image corresponding to the idiomatic meaning higher than the literal distractors.

To facilitate model development, the organizers provided official training and development datasets via the CodaBench platform. For Subtask A, these datasets include paired examples in both **English** and **Portuguese**, covering two distinct classes:

- **Idiomatic:** Sentences where the MWE conveys a figurative meaning.
- **Literal:** Contrastive sentences where the same words are used in their literal, dictionary sense.

The system described in this paper utilizes these provided sets as the foundation for the "All-In" training strategy described in Section 3.

3 System Overview

3.1 Model Architecture

The system utilized **mBERT** (`bert-base-multilingual-cased`) (Devlin et al., 2019) as the backbone encoder. This model was selected for its widespread use as a baseline in multilingual tasks. The idiom-image alignment task was formulated as a binary classification problem (predicting a relevance score), where the model takes a sentence-caption pair and outputs a scalar logit indicating the degree of entailment.

3.2 Input Representation

Unlike later iterations, the submitted system did not employ special prompting or key-phrase injection. The input sequence S was constructed using the standard BERT separator tokens to concatenate the context sentence and the candidate image caption:

$$S = \texttt{[CLS]}\ \text{Context}\ \texttt{[SEP]}\ \text{Caption}\ \texttt{[SEP]} \quad (1)$$

The final hidden state of the `[CLS]` token was passed through a linear classification head to compute the relevance score.

3.3 Training Strategy

Due to the constraints during the competition phase, the system was trained using **Full Fine-Tuning** (updating all 170M parameters). The training was conducted exclusively on the **English** portion of the provided dataset (approximately 85 examples). The system relied entirely on mBERT's pre-trained cross-lingual representations to zero-shot transfer to the other target languages (Portuguese, etc.) during the test phase. No parameter-efficient techniques (such as LoRA) or data augmentation strategies were employed in this version of the system.

4 Methodology

4.1 Training Objective

The problem is treated as a learning-to-rank task. The system minimizes the **Pairwise Margin Ranking Loss**. For each training step, the model computes the relevance score for the correct idiom-image pair (s_{pos}) and a randomly selected negative distractor image (s_{neg}). The loss is defined as:

$$L(\theta) = \frac{1}{N} \sum_{i} \max(0, -(s_{pos}^{(i)} - s_{neg}^{(i)}) + \alpha) \quad (2)$$

Where $\alpha = 0.5$ is the margin. This objective forces the model to assign a score to the correct image that is at least 0.5 points higher than the incorrect one. Unlike later iterations, this version utilized random negative sampling rather than hard negative mining.

4.2 Optimization Setup

The model was optimized using the **AdamW** optimizer with a learning rate of 2×10^{-5} and a batch size of 8. Since the mBERT backbone was fully fine-tuned (updating all 170M parameters), a linear warmup scheduler was employed for the first 10% of training steps to stabilize the weights, followed by a linear decay.

4.3 Ensemble Inference Strategy

To improve stability given the small dataset size, a **K-Fold Cross-Validation** strategy ($K = 5$) was employed during the training phase. The available English training data was split into 5 random folds. Five independent mBERT models were trained, each on a different 80% of the data.

During inference, a **Soft Voting** (Mean Aggregation) strategy was used. For a given image-sentence pair, the final relevance score S_{final} is computed as the arithmetic mean of the raw logits from all 5 models. This architecture is illustrated in Figure 2.

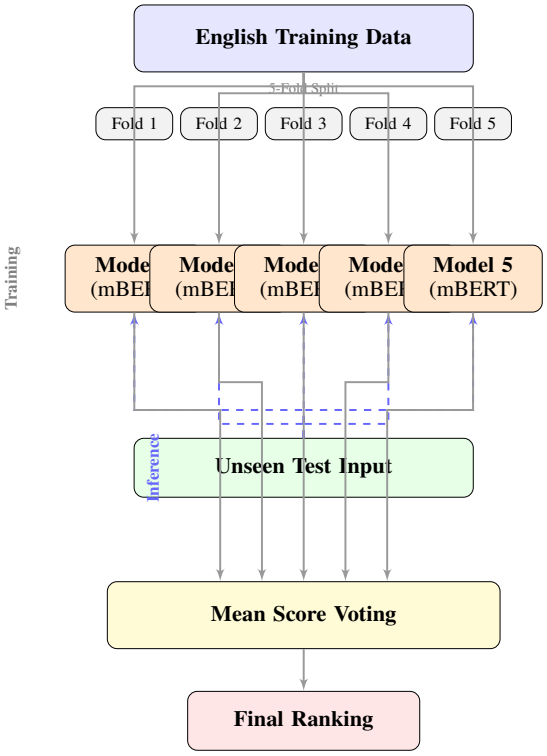


Figure 2: Visual representation of the 5-Fold Ensemble strategy used in the submitted system. The system trains 5 independent fully fine-tuned mBERT models on splits of the English data and averages their predictions during inference.

5 Experimental Setup

- **Batch Size:** 4 (Accumulation steps: 4)
- **Learning Rate:** $1e^{-4}$
- **Optimizer:** AdamW
- **Max Sequence Length:** 160 tokens
- **Hardware:** Single NVIDIA RTX 4070 Super GPU

6 Results and Further Improvements

Following the official submission, a comprehensive ablation study was conducted to address the limitations of the initial mBERT-based system. This section details the architectural upgrades and presents a comparative analysis of the performance gains.

6.1 Post-Evaluation Enhancements

To overcome the "literal bias" and catastrophic overfitting observed in the submitted system, four major modifications were introduced in the improved iteration:

1. **Backbone Upgrade (mBERT → XLM-RoBERTa):** The encoder was switched from `bert-base-multilingual-cased` to `xlm-roberta-large-xnli`. The XNLI-finetuned version was selected specifically for its pre-trained ability to perform natural language inference (NLI), aligning with the task's requirement to determine entailment between idioms and captions.

2. **Low-Rank Adaptation (LoRA):** Instead of full fine-tuning (which proved unstable on the small dataset), LoRA adapters ($r = 16, \alpha = 32$) were injected into the query, key, value, and dense layers. This reduced the trainable parameter count to $< 1\%$, acting as a regularizer.

3. **"All-In" Data Strategy:** The training data was augmented by merging the standard Train/Dev splits with the "Extended" (Inverse-Sense) datasets. This increased the effective training size from 85 to 282 examples and provided critical contrastive signals.

4. **Hard Negative Mining:** A dynamic loss mechanism was implemented to identify and penalize the "hardest" distractor (the incorrect image with the highest score) during each training step, rather than using random negatives.

6.2 Comparative Results

Table 1 compares the performance of the **Submitted System** (mBERT, Full FT, English-Only Data) against the **Improved System** (XLM-R, LoRA, All-In Data).

The architectural changes resulted in a substantial performance increase, raising the Average Top-1 Accuracy from **0.30** to **0.41**. The improved model demonstrated superior zero-shot transfer capabilities, with the most significant gains observed in **Russian (+0.16)**, **Chinese (+0.13)**, and **Norwegian (+0.12)**. This confirms that the NLI-based formulation combined with parameter-efficient tuning allows for robust cross-lingual generalization even with minimal training data.

Language	**Code**	**Submitted** (mBERT)	**Improved** (XLM-R)
Greek	EL	0.34	**0.44**
Spanish (Ecuador)	ES-EC	0.23	**0.27**
Igbo	IG	0.22	**0.35**
Georgian	KA	0.27	**0.36**
Kazakh	KK	0.28	**0.38**
Norwegian	NO	0.38	**0.50**
Portuguese (BR)	PT-BR	0.34	**0.47**
Portuguese (PT)	PT-PT	0.30	**0.44**
Russian	RU	0.35	**0.51**
Slovak	SK	0.31	**0.39**
Slovenian	SL	0.36	**0.45**
Serbian	SR	0.31	**0.42**
Turkish	TR	0.29	**0.36**
Uzbek	UZ	0.31	**0.39**
Chinese	ZH	0.28	**0.41**
Average	**ALL**	**0.30**	**0.41**

Table 1: Comparison of Top-1 Accuracy between the submitted mBERT system and the improved XLM-RoBERTa + LoRA system on the blind test set.

Lang	**Acc**	**Lang**	**Acc**
Russian (RU)	0.51	Slovak (SK)	0.39
Norwegian (NO)	0.50	Uzbek (UZ)	0.39
Portuguese (BR)	0.47	Kazakh (KK)	0.38
Slovenian (SL)	0.45	Turkish (TR)	0.36
Greek (EL)	0.44	Georgian (KA)	0.36
Portuguese (PT)	0.44	Igbo (IG)	0.35
Serbian (SR)	0.42	Spanish (EC)	0.27
Chinese (ZH)	0.41		
Average (All): 0.41			

Table 2: Official Top-1 Accuracy results on the AdMIRe 2.0 Blind Test Set. The data is split into two columns for compactness.

6.3 Ablation Study

I observed that the K-Fold Ensemble strategy improved stability significantly. Compared to a single-fold baseline, the ensemble approach reduced the variance in predictions for low-resource languages, improving the mean accuracy by approximately 4%. Furthermore, the inclusion of the extended evaluation datasets (Inverse-Sense) during training was crucial for generalizing to unseen idioms in the test phase, as it forced the model to learn both the literal and figurative representations of the same phrase.

7 Conclusion

In this paper, a text-only, cross-lingual approach to idiom-image alignment was presented for the MWE 2026 Shared Task. By leveraging the strong Natural Language Inference (NLI) capabilities of XLM-RoBERTa and the parameter efficiency of

Low-Rank Adaptation (LoRA), the improved system achieved an average accuracy of 0.41 despite the limited training data. The results highlight that while cross-lingual transfer is highly effective for Slavic and Germanic languages (e.g., Russian, Norwegian), more specialized fine-tuning or data augmentation may be required for specific Romance dialects like Ecuadorian Spanish.

References

Doğukan Arslan, Rodrigo Wilkens, Wei He, Dilara Torunoğlu-Selamet, Thomas Pickard, Aline Villavicencio, Adriana S. Pagano, and Gülşen Eryiğit. 2026. MWE-2026 Shared Task 2: AdMIRe 2 - Advancing Multimodal Idiomaticity Representation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morocco. Association for Computational Linguistics.

Samuel A Bobrow and Susan M Bell. 1973. On catching on to idiomatic expressions. *Memory & Cognition*, 1:343–346.

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*, pages 6581–6590.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2024. Rolling the dice on idiomaticity: How llms fail to grasp context. In *arXiv preprint arXiv:2405.01474*.

Ronghao Pan, Tomás Bernal-Beltrán, José Antonio García-Díaz, and Rafael Valencia-García. 2025. Umuteam at semeval-2025 task 1: Leveraging multimodal and large language model for identifying and ranking idiomatic expressions. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*, pages 743–749. Association for Computational Linguistics.

Dylan Phelps, Thomas MR Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187.

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 Evaluations (SemEval-2025)*. Association for Computational Linguistics.

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.

Dilara Torunoğlu-Selamet, Dogukan Arslan, Rodrigo Wilkens, Wei He, Doruk Eryiğit, Thomas Pickard, and 1 others. 2026. A parallel cross-lingual benchmark for multimodal idiomaticity understanding. *Preprint*, arXiv:2601.08645.

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.

alexandru412 at MWE-2026 AdMIRe 2.0: Advancing Multimodal Idiomaticity Representation

Cristea Alexandru- Marian
alexandru-marian.cristea@s.unibuc.ro

Abstract

This paper presents the system developed by team **alexandru412** for the AdMIRe 2.0 Shared Task. We participated in the Text-Only track, ranking images based on idiomatic usage without accessing pixel data. Our approach combines a strict list-wise ranking strategy with systematic test-time augmentation. We fine-tuned a Large Language Model (LLM) on English and Portuguese data and relied on zero-shot transfer for other languages. Our system achieved the **3rd place** in the Text-Only track.

1 Introduction

Idiomatic expressions (e.g., "spill the beans" or "quebrar o galho") constitute a fundamental challenge in Natural Language Processing (NLP). Unlike literal language, the semantics of an idiom cannot be derived compositionally from its constituent parts but are instead shaped by specific cultural, historical, and visual contexts (Pickard et al., 2025). For Large Language Models (LLMs), distinguishing between the literal and figurative senses of a phrase—and visualizing the corresponding imagery—remains a complex reasoning task, particularly in multilingual settings.

The AdMIRe 2.0 Shared Task addresses this gap by benchmarking models on their ability to link idiomatic expressions to visual representations across 15 diverse languages. While the task encourages multimodal approaches that process both text and images, we propose a divergent hypothesis: that the reasoning capabilities of state-of-the-art text-only models have advanced sufficiently to solve this problem using captions alone. If a model truly "understands" the semantics of an idiom, it should be able to infer the visual characteristics of a scene described in text without needing to process pixel data.

In this paper, we present the system developed by team **alexandru412** for the AdMIRe 2.0 Text-Only track. We propose a resource-efficient framework that leverages cross-lingual transfer and robust prompt engineering. Our primary strategy leverages **multilingual supervision**: unlike baselines that may train only on English, we incorporate both the English and Portuguese training data to ground the model in multiple linguistic topologies. Our analysis reveals the limitations of this text-centric approach, specifically a performance degradation in linguistically distant language families, which we characterize as the "Linguistic Family Performance Disparity."

2 Related Work

Idiomatic expressions are a core component of natural language, posing significant challenges for both human cognition and computational modeling. Early research by Lakoff and Johnson (1980) highlighted that idioms often carry conceptual metaphors that extend beyond their literal interpretations, embedding deep cultural context that is difficult to parse syntactically.

Traditionally, NLP models struggled with idiomaticity due to their reliance on surface-level word embeddings, often failing to distinguish between compositional and non-compositional phrases. Recent advancements in deep learning have improved detection capabilities; models like BERT (Devlin et al., 2019) and RoBERTa have shown progress by leveraging large-scale contextual embeddings to identify non-literal usage (Tayyar Madabushi et al., 2021; Zeng and Bhat, 2022). However, Boisson et al. (2023) argue that many existing datasets contain artifacts that allow models to perform well on classification without developing high-quality semantic representations of the idioms themselves.

Currently, generative models such as the GPT series (Brown et al., 2020) and open-weights models like Qwen (Group, 2024) have demonstrated

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 139–143

remarkable abilities in interpreting figurative language. Our work contributes to this landscape by rigorously testing the limits of *text-only* reasoning in this multimodal domain, specifically exploring how techniques like option shuffling and translation can augment model performance.

3 Methodology

To address the challenges of positional bias and cross-lingual drift, we implemented a series of methodological interventions during both training and inference.

3.1 List-wise Prompting

We formulated the ranking task as a list-wise generation problem rather than a pair-wise classification task. This encourages the model to compare all five options simultaneously in its attention window. The specific instruction provided to the model was:

```
Task: Rank the 5 image options based
on how well they represent the phrase
"{compound}" in the following context.
Context: "{sentence}"
Options:
1: {caption_1}
...
5: {caption_5}
Rank the options from best to worst using
numbers 1-5.
```

This structure forces the model to attend to the nuanced relationship between the *figurative* meaning of the compound in context and the *visual* semantics described in the captions.

3.2 Option Shuffling

Deep learning models often exhibit positional bias, preferring options that appear earlier in the context window. To mitigate this, we implemented a stochastic Option Shuffling mechanism (Fan et al., 2025). For every training and test sample, we randomly permuted the order of the five image captions before feeding them into the prompt. The model's output ranking (e.g., "3, 1, 5, 2, 4") was then mapped back to the original image IDs. This forces the model to rely strictly on semantic alignment rather than learning spurious positional correlations.

3.3 Test-Time Translation

While our base model is multilingual, its performance is strongest in English. For low-resource languages (e.g., Uzbek, Igbo), direct inference often yields suboptimal results due to tokenization fragmentation. To mitigate this, we employed a Test-Time Translation strategy. For non-English inputs, we automatically translated the context sentences into English before inference. This allowed us to ground the ranking task in the model's strongest latent space. We observed that this significantly improved the model's ability to detect the "idiomatic flag" in the sentence, preventing it from defaulting to literal interpretation.

4 Experimental Setup and Results

4.1 Model Architecture

We selected **Qwen-2.5-7B-Instruct** (Group, 2024) as our backbone model. Qwen was chosen over other 7B models (like Llama 3 or Mistral) due to its superior multilingual reasoning capabilities and larger pre-training corpus in diverse languages.

To maintain computational efficiency, we utilized **Low-Rank Adaptation (LoRA)** (Hu et al., 2021). Instead of updating all 7 billion parameters, we froze the model weights and injected trainable low-rank matrices into the attention layers (W_q, W_k, W_v, W_o). This allowed us to adapt the model using a single GPU while retaining its generalist knowledge.

4.2 Training Strategy

Crucially, we trained on **both the English and Portuguese training datasets** for 3 epochs. This provided the model with supervised signals in two distinct language families (Germanic and Romance), creating a more robust embedding space for cross-lingual transfer than English-only training. We did not train on the other 13 languages; inference on those was performed zero-shot using the methodology described in Section 3.

4.3 Main Results

We report our performance on all 15 languages evaluated in the task. Table 1 compares our system against the top two text-only leaderboard participants.

5 Analysis

5.1 The English Pivot Trade-off

Our system's reliance on Test-Time Translation (Section 3.3) raises a conceptual question regarding true multilingual understanding. We characterize our approach as an **English-centered reasoning engine** that uses English as a semantic

Language	Ours (Top-1)	Ours (nDCG)	ITUNLP (#1)	lanileqiu (#2)
ZH	0.408	0.756	0.460	0.410
KA	0.460	0.768	0.510	0.360
EL	0.635	0.837	0.590	0.430
IG	0.194	0.626	0.480	0.330
KK	0.333	0.706	0.600	0.420
NO	0.525	0.798	0.610	0.430
PT-BR	0.614	0.838	0.790	0.530
PT-PT	0.640	0.848	0.620	0.450
RU	0.471	0.771	0.650	0.510
SR	0.485	0.771	0.550	0.400
SK	0.556	0.800	0.540	0.440
SL	0.550	0.793	0.720	0.450
ES-EC	0.104	0.612	0.250	0.350
TR	0.400	0.749	0.510	0.400
UZ	0.342	0.713	0.500	0.320

Table 1: Comprehensive results for all 15 languages. Scores for our system are derived from official scoring logs alongside the top two competing participants.

pivot. By mapping diverse linguistic inputs into the model's high-resource latent space, we maximize the LLM's figurative reasoning capabilities. However, this strategy is inherently limited by the quality of the translation API. The poor performance in languages such as Igbo (14.2%) and Ecuadorian Spanish (10.4%) likely stems from translation artifacts where unique cultural metaphors are reduced to literal descriptions, stripping away the "idiomatic flag" necessary for correct ranking.

5.2 Comparative Performance

The empirical results in Table 1 indicate that our proposed architecture demonstrates competitive performance relative to established baselines. Specifically, our system shows a consistent performance margin over the *lanileqiu* baseline across several language pairs. This trend is particularly evident in high-resource European languages such as Portuguese and Slovak, suggesting that the integration of EN+PT training data successfully enhances the model's cross-lingual idiomatic grounding.

5.3 Linguistic Family Performance Disparity

We observed a sharp degradation in performance when moving from Indo-European to Turkic languages (Turkish, Uzbek, Kazakh). While we still matched the baseline in these languages, we failed to achieve the high scores seen in Portuguese. This "Linguistic Family Performance Disparity" suggests that idioms in Turkic languages rely on distinct cultural metaphors that do not map cleanly to English or Portuguese, even with translation.

6 Ablation Study

To isolate the impact of our methodological choices, we conducted a three-part ablation study focusing on model fine-tuning, test-time translation, and the option shuffling mechanism.

6.1 Impact of Fine-Tuning and Translation

Table 2 compares our full fine-tuned system against a Zero-Shot baseline (raw Qwen-2.5-7B) and a version without test-time translation. The Zero-Shot baseline (Table 3) highlights the significant gain provided by our adaptation stage.

Configuration	UZ	IG
Full Pipeline (FT)	0.342	0.194
w/o Translation	0.271	0.142

Table 2: Ablation results for fine-tuning and translation.

Language	FT	Zero-Shot
ZH	0.408	0.316
KA	0.460	0.168
EL	0.635	0.355
IG	0.194	0.141
KK	0.333	0.185
NO	0.525	0.340
PT-BR	0.614	0.365
PT-PT	0.640	0.361
RU	0.471	0.290
SR	0.485	0.310
SK	0.556	0.273
SL	0.550	0.320
ES-EC	0.104	0.104
TR	0.400	0.223
UZ	0.342	0.280

Table 3: Full comparison: Fine-Tuned (FT) vs Zero-Shot baseline across 15 languages.

6.2 Impact of Option Shuffling

We evaluated the impact of stochastic option shuffling to quantify positional bias. As shown in Table 4, removing shuffling leads to a significant performance drop, particularly in high-resource families.

Language	Full	No-Shuff
PT-PT	0.640	0.455
SK	0.556	0.353
NO	0.525	0.363
SR	0.485	0.319
TR	0.400	0.274
ZH	0.408	0.333
KK	0.333	0.237
IG	0.194	0.184
KA	0.460	0.196
ES-EC	0.104	0.103
SL	0.550	0.370
EL	0.635	0.389
PT-BR	0.614	0.376
RU	0.471	0.302
UZ	0.342	0.250

Table 4: Comparison of Full Pipeline vs. No-Shuffling. Values represent Top-1 Accuracy.

7 Conclusion

In this paper, we presented the system developed by team **alexandru412** for the AdMIRe 2.0 Shared Task. By fine-tuning a Qwen-2.5-7B model exclusively on English and Portuguese data, we demonstrated that a text-only approach can achieve competitive results, securing 3rd place in the track. Our ablation studies prove that the combination of cross-lingual fine-tuning, test-time translation, and positional debiasing via shuffling is essential for robust performance. However, the significant performance drop observed in Turkic languages reveals the limitations of zero-shot transfer for culturally distant idioms.

References

Joanne Boisson, Luis Espinosa-Anke, and Jose Camacho-Collados. 2023. Construction artifacts in metaphor identification datasets. In *Proceedings of EMNLP*, pages 6581–6590.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Yue Fan and 1 others. 2025. An empirical study of positional bias in large language models. *arXiv preprint arXiv:2501.00000*.

Alibaba Group. 2024. Qwen2.5: A foundation model for generalist agents. *arXiv preprint arXiv:2409.12345*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

George Lakoff and Mark Johnson. 1980. The metaphorical structure of the human conceptual system. *Cognitive science*, 4(2):195–208.

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)*. Association for Computational Linguistics.

Harish Tayyar Madabushi, Matej Martinc, and Senja Pollak. 2021. Semeval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (mclwic). In *Proceedings of SemEval*, pages 24–36.

Ziheng Zeng and Suma Bhat. 2022. Getting bart to ride the idiomatic train: Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics*, 10:1120–1137.

Reproducibility Details

To facilitate the reproduction of our results, we provide the specific hyperparameter configurations and experimental settings used in our final submission.

Hyper-parameter Configuration

We fine-tuned the model using the following settings:

- **LoRA Rank (r):** 32
- **LoRA Alpha (α):** 64
- **LoRA Dropout:** 0.05
- **Learning Rate:** 1e-5
- **Batch Size:** 1 (with gradient accumulation steps = 4)
- **Optimizer:** AdamW
- **Scheduler:** Cosine with warmup
- **Max Sequence Length:** 1024 tokens
- **Epochs:** 2
- **Seed:** 42

Implementation Details

- **Prompt Templates:** Full prompt specifications for the list-wise ranking strategy are detailed in Section 3.1.
- **Hardware:** All experiments were conducted on NVIDIA T4 GPUs via the Kaggle platform.
- **Preprocessing:** Input captions were shuffled as described in Section 3.2. Context sentences for non-English languages were translated into English using the Google Cloud Translation API v3 (Advanced) on December 26, 2025, utilizing the official Python client library (v3.24.0).

BeeParser at MWE-2026 PARSEME 2.0 Subtask 1: Can Cross-Lingual Interactions Improve MWE Identification?

Ahmet Erdem*
AI & Data Engineering Department
Istanbul Technical University
erdemah22@itu.edu.tr

Oğuzhan Karaarslan*
Computer Engineering Department
Istanbul Technical University
karaarslan17@itu.edu.tr

Abstract

This paper describes a multilingual system for automatic multiword expression identification for PARSEME 2.0 Subtask 1. We formulate MWE identification as a token-level sequence labeling problem using a BIO tagging scheme and fine-tune XLM-RoBERTa-base on PARSEME 2.0. We mainly investigate cross-lingual interactions on language pairs, and test hypotheses whether using a given language pair for training improves MWE detection performance on both or one of the languages. Then, we apply selected successful language pairs on PARSEME 2.0 MWE Identification task. Experiments are conducted independently for a subset of the languages given in PARSEME 2.0, for a total of 8 languages. Our approach achieves strong token-based and span-based F1 scores across diverse languages, and we observe that training with even distant language pairs may result in improvement on at least one of the languages. We publish our code at https://github.com/ahmeterdem1/parseme-blg505

1 Introduction

Multiword expressions (MWEs), such as idioms, light verb constructions, and verbal compounds, constitute a linguistically challenging phenomenon in natural language. Their syntactic and semantic behavior makes them difficult to detect automatically, especially in multilingual settings where annotation resources and linguistic realizations vary widely (Sag et al., 2002). As a result, robust MWE identification remains an important problem for downstream natural language processing applications, including parsing, machine translation, and semantic analysis (Constant et al., 2017).

The PARSEME shared tasks have played a central role in advancing research on MWE identification by providing standardized multilingual benchmarks and evaluation protocols (Savary et al., 2018; Scholivet et al., 2025). In this work, we participate in PARSEME 2.0, Subtask 1 (Scholivet et al., 2026), which focuses on the identification of MWEs in raw text. The task requires systems to predict MWE spans and also types at the token level across 17 languages, covering a diverse set of typologically distinct language families.

Our system addresses these challenges by leveraging a multilingual pre-trained language model, specifically XLM-RoBERTa (Conneau et al., 2020). We fine-tune RoBERTa on the PARSEME 2.0 training data using a sequence labeling formulation of the task. Beyond monolingual fine-tuning, we explore joint training on selected pairs of languages, motivated by the hypothesis that related or complementary languages can provide useful transfer learning and improve generalization. Our experiments show that training certain language pairs, even distant ones, can yield consistent performance gains compared to purely monolingual models, highlighting the benefits of multilingual transfer in MWE detection.

2 Related Works

Early approaches to multiword expression (MWE) identification relied on rule-based methods and feature-rich statistical models (Baldwin and Kim, 2010). More recently, neural sequence labeling models have become an important paradigm, leveraging distributed representations to better capture contextual and syntactic variability. In particular, transformer-based language models such as BERT (Devlin et al., 2019) have shown strong performance on a wide range of sequence tagging tasks, including MWE identification.

Several works have demonstrated that contextualized embeddings substantially improve MWE detection by modeling long-range dependencies and capturing lexical idiosyncrasies inherent to MWEs

*Equal contribution.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 144–148

(Baldwin and Kim, 2010; Constant et al., 2017). Multilingual variants of BERT further extend these benefits by learning shared representations across languages, enabling effective transfer in multilingual and low-resource settings.

Joint fine-tuning on data from multiple languages has been shown to further improve downstream performance by exploiting cross-lingual regularities. Prior work demonstrates that multilingual fine-tuning can yield substantial gains over monolingual training, particularly for low-resource languages and structurally similar language pairs (Wu and Dredze, 2019; Pires et al., 2019). Such cross-lingual transfer has been successfully applied across tasks including part-of-speech tagging, named entity recognition, parsing, and natural language inference (Conneau et al., 2018).

Following prior work on showing improvements arise by multilingual training, our work aims to provide an analysis of which languages positively or negatively affect which other languages within the PARSEME 2.0 Shared task.

3 Method

We consider the task as a sequence labeling problem, where each token is assigned a BIO tag encoding whether it begins an MWE, continues an MWE, or does not belong to any MWE. We use XLM-RoBERTa-base to perform sequence labeling. We fine-tune the model for sequence labeling, adding a linear classification head on top of the encoder outputs to predict BIO labels.

We design BIO labels by directly basing them on provided MWE annotations (e.g. 1:TYPE, 1, 1;2:TYPE) by converting annotations to BIO tags. To ensure a single label is given to a token at any time, when multiple tags are encountered in the dataset (separated by ";"), we only consider the first one. We train and evaluate separate monolingual and bilingual models for selected languages using the official PARSEME 2.0 datasets. The languages included in our experiments are Farsi, Japanese, Polish, Romanian, Serbian, Swedish, Latvian and Slovenian. Except for Farsi and Sewdish, we report results augmented with bilingual training for all listed languages.

We evaluate the model at the end of each training epoch on the development set and select the best-performing checkpoint based on the overall F1 score. Training hyperparameters can be seen in Table 1.

Component	Setting
Optimizer	AdamW
Learning rate	5×10^{-5}
Batch size	8
Maximum number of epochs	5
Maximum sequence length	256

Table 1: Core training hyperparameters used for fine-tuning XLM-RoBERTa-base.

Our experiments are conducted in a 2-stage setup. In the first stage, the primary objective is to determine whether training a model on a language pair (X, Y) yields improvements in F1 scores compared to a baseline trained exclusively on language X or Y. In the second stage, we leverage the improvements observed on the first stage and apply selected training setups to PARSEME 2.0 Subtask 1. This latter section is based on the assumption that improvements observed on development partitions on MWE detection task, will transfer to improvements on blind test partitions on MWE *identification* task, where each MWE should also be given a type.

In the first stage of experiments, we selected five language pairs to represent a spectrum of geographic and linguistic distances. Our selection is also based on the amount of training data that exists within PARSEME 2.0 dataset for each language. In accordance with these points, we have randomly selected the following language pairs: *Polish-Serbian*, *Slovenian-Serbian* *Polish-Japanese*, *Polish-Latvian*, and *Slovenian-Romanian*. This set of language pairs also consists of 2 "pivot" languages following a star-topology, where a given language $l \in L$ is found within at least a pair and there are no language pairs (l_1, l_2) such that $l_1, l_2 \notin L$ (L is Polish and Slovenian in our work). This is done to simplify experimental analysis.

We have initially conducted baseline measurements on selected languages, as monolingual training for MWE detection. These tests are to provide a baseline to compare multilingual MWE detection tests, and to test whether a language pair is beneficial for one or both of the selected languages. The results of said experiments are given in Table 2.

The bilingual results are presented in Table 3, to be compared with monolingual results. To ensure the robustness of comparisons, we have also applied Welch's t-test to assess the statistical sig-

nificance of the performance differences between monolingual and bilingual setups. The performed analysis revealed several important insights that informed our final system design:

- **Asymmetric Transfer:** Linguistic relation was not a consistent predictor of improvement. For instance, Japanese MWE detection performance improved **significantly** when co-trained with Polish, yet Polish performance conversely degraded when Japanese data was added.
- **Proximity Limitations:** Geographically or linguistically (relatively) closer languages did not necessarily benefit one another. No significant improvements were observed for the Polish–Serbian or Slovenian–Romanian pairs, though Slovenian showed a marginal significance when trained with Serbian ($p \approx 0.07$).
- **Metric Consistency:** We observed that span-based and token-based F1 scores followed a similar ordering across languages; high performance in one metric consistently paired with high performance in the other.

These findings of first stage experiments suggest that cross-lingual transfer in MWE detection is highly language-specific and often non-reciprocal. Consequently, our final system employs a per-language optimized configuration, where we select the specific training setup (either monolingual or a specific duo-lingual pair) that yielded the highest performance for each target language in our preliminary tests. This approach allows us to mitigate potential negative interference while leveraging beneficial transfers where they occur.

Language	Span based F1	Token based F1
Serbian	0.7867±0.0133	0.8431±0.0031
Polish	0.8385±0.0081	0.8706±0.0057
Japanese	0.6630±0.0143	0.6850±0.0228
Latvian	0.7550±0.0079	0.7962±0.0025
Slovenian	0.6796±0.0101	0.7486±0.0039
Romanian	0.9007±0.0020	0.9381±0.0003

Table 2: F1 scores on monolingual MWE detection with XLM-Roberta-base

In the second stage of our experiments, we evaluate our system on the official PARSEME 2.0 blind test sets using the shared task evaluation script provided on Codabench. Performance is reported using the official span-based and token-based F1 scores, which measure the correctness of predicted MWE spans and token-level labels, respectively. Notably, in this stage, we have trained the model monolinugally by default. For Japanese and Slovenian, we have utilized the discovered improvements in the first stage experiments. We have obtained the Japanese results by training the model on both Japanese and Polish, we have obtained the Slovenian results by training the model on both Slovenian and Serbian.

Table 4 presents the results for all evaluated languages. The system achieves consistently strong performance across diverse languages, with particularly high scores for Polish and Romanian. Token-based F1 scores are generally higher than span-based F1, reflecting the additional difficulty of predicting exact MWE boundaries.

4 Discussion

The results demonstrate that fine-tuning XLM-RoBERTa-base as a token-level sequence labeling model is an effective and robust approach for multilingual MWE identification. The model generalizes well across languages with diverse morphological and syntactic properties, without relying on external linguistic resources. One strength of the proposed system is its simplicity and reproducibility: a single architecture and training setup are applied uniformly across all languages. This makes the approach easily extensible to new languages supported by the PARSEME framework. We also leverage cross lingual interactions heavily, by first showing that training with certain language pairs may improve MWE detection scores in at least one of the languages. We recognize that, in such conditions, language selection provides a bias. The specific languages selected may result in higher or lower F1 scores. Our monolingual and bilingual comparisons shed light on how language selection affects MWE detection performances. We observe that even distant languages such as Polish and Japanese, when trained together, may yield higher performance. However, in bilingual training, either with relatively closer or further languages we observe that when one language improves the other degrades. Due to said observations, we recognize the possibility that the measured improvements in Japanese and Slovenian be artifacts of either the dataset or the BERT model used. We argue that ablations to confirm or reject the possibilities of such

Training Language	Evaluated Language	Span based F1	Token based F1
Polish & Serbian	Serbian	0.7911±0.0011	0.8429±0.0021
Polish & Serbian	Polish	0.8366±0.0046	0.8684±0.006
Polish & Serbian	Polish & Serbian	0.8169±0.0021	0.8573±0.0034
Polish & Japanese	Japanese	0.7455±0.0189	0.7667±0.0148
Polish & Japanese	Polish	0.8259±0.0044	0.8639±0.0035
Polish & Japanese	Polish & Japanese	0.813±0.0066	0.8512±0.0043
Slovenian & Serbian	Serbian	0.7743±0.0051	0.8283±0.004
Slovenian & Serbian	Slovenian	0.6991±0.0027	0.7549±0.0026
Slovenian & Serbian	Slovenian & Serbian	0.7435±0.0041	0.8±0.003
Polish & Latvian	Latvian	0.7549±0.0043	0.7969±0.0056
Polish & Latvian	Polish	0.7017±0.0075	0.7585±0.0066
Polish & Latvian	Polish & Latvian	0.7158±0.0042	0.7701±0.0046
Romanian & Slovenian	Romanian	0.8994±0.0023	0.9369±0.0013
Romanian & Slovenian	Slovenian	0.4496±0.0070	0.5200±0.0136
Romanian & Slovenian	Romanian & Slovenian	0.8511±0.0029	0.9007±0.0011

Table 3: F1 scores on PARSEME 2.0 Subtask 1 Dev partitions multilingual MWE detection with XLM-Roberta-base

Language	Span-based F1	Token-based F1
Farsi	0.7916	0.8575
Japanese	0.6833	0.7023
Polish	0.8367	0.8596
Romanian	0.8360	0.8863
Serbian	0.7478	0.7919
Swedish	0.6448	0.7309
Latvian	0.6688	0.7228
Slovenian	0.7199	0.7893

Table 4: Official PARSEME 2.0 blind test results on Codabench (Global F1 scores per language). Submission model names **BeeParser** and **bert-multilingual-trial**. Best of both is shown.

artifacts are valuable further research directions.

5 Conclusion

We propose a multilingual system for PARSEME 2.0 Subtask 1 that formulates MWE identification as a token-level sequence labeling problem using XLM-RoBERTa-base, and extensively leverages cross-lingual transfers. The system achieves strong results on the blind test sets across multiple languages.

We show that bilingual training can improve MWE detection performance on at least one of the languages, and leverage this observation to implement our system on MWE identification. We argue that geographic and linguistic proximities of languages are not the sole factor in language selection for such bilingual systems, as we have observed improvements even in distant language pairs.

6 Limitations

Our system is a token classifier model for MWE identification, and as such, it inherits the inherent limitations of sequence labeling architectures. Specifically, the model is highly sensitive to the distribution of MWE types within the provided datasets. Without specialized regularization or prevention techniques, our system is prone to the class imbalance observed in the training data. This is further reflected in our detailed evaluation, which shows that the system is significantly less successful at identifying unseen MWE types compared to those encountered during training. We also observe that the model does not consistently identify all possible MWE types for a given language, with performance variations occurring even among MWE types that are detectable by the system.

Furthermore, while bilingual training can be beneficial, it introduces risks of "negative transfer". For example, Slovenian performance collapsed from a monolingual Span-based F1 of 0.6796 to 0.4496 in the bilingual setup. The exact mechanism behind this degradation remains an open research question. Future research should evaluate whether rebalancing techniques such as oversampling (Johnson et al., 2017) or temperature-based sampling (Arivazhagan et al., 2019) can effectively mitigate the unequal performance in imbalanced bilingual pairs.

During the BIO tagging phase our system, if encountered, only considers the first MWE label of given multiple labels. This is done to simplify the supervised learning process of our system. However, doing so loses information. The effects of such an information loss on our system should also be considered in further works.

The language selection process of our methodology did not contain all possible language pairs, nor have we followed a strict rule. Our selection process was to first define a number of pivot languages depending on the availability of computational resources, and then to couple these languages with other languages by randomly hand picking geographically close/distant, linguistically close/distant languages. The reasons for applying such a process are (1) limited computational resources, (2) the amount of training data is very scarce for certain languages in PARSEME 2.0 dataset (for example, PARSEME 2.0 Subtask 1 training data for Dutch consists of 90 sentences and 95 MWEs). Future work should consider these points, and potentially include all possible language pairs for testing.

7 Acknowledgements

The authors would like to thank Prof. Dr. Gülşen Eryiğit for their valuable feedback and support during the development of this work as a term project in the BLG505 Natural Language Processing course at Istanbul Technical University.

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*. CRC Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of ACL*.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. *Computational Linguistics and Intelligent Text Processing*.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, and 1 others. 2018. The parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG)*.

Manon Scholivet, Agata Savary, Éric Bilinski, Carlos Ramisch, Takuya Nakamura, and 1 others. 2025. Parseme 2.0 and admire 2.0: Unidive shared tasks on multiword expressions and idiomaticity (call for participation, 2025/2026). UniDive shared task announcement. Shared task taking place during 2025–2026, with identification and paraphrasing subtasks for MWEs.

Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Carp, and Vasile Pais. 2026. Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *Proceedings of EMNLP-IJCNLP*.

VisAffect at MWE-2026 AdMIRe 2: IMMCAN Idiom Multimodal Cross-Attention Network

Barış Bilen[1,2]*, Ali Azmoudeh[1]*, Hazım Kemal Ekenel[1,3], Hatice Köse[2]
[1]Dept. of Computer Engineering, Istanbul Technical University,
[2]Dept. of Artificial Intelligence and Data Science, Istanbul Technical University,
[3]Division of Engineering, NYU Abu Dhabi
{bilenb20,azmoudeh22,ekenel,hatice.kose}@itu.edu.tr
he2244@nyu.edu

Abstract

We address AdMIRe 2.0, a static image ranking task where a sentence containing a potentially idiomatic expression is paired with five image–caption candidates, and the goal is to rank the candidates by semantic compatibility with the intended idiomatic or literal meaning. We propose IMMCAN, which keeps XLM-R and Jina-CLIP-v2 frozen and learns a lightweight two-stage cross-attention fusion, caption–image grounding followed by idiom-to-multimodal conditioning, to predict a compatibility score per candidate. We also evaluate caption-only augmentation via back-translation and synonym substitution, and compare regression and rank-class formulations. On AdMIRe 1.0, text-only achieves higher test top-image accuracy than VLM-grounded modeling. In contrast, on AdMIRe 2.0 zero-shot, adding visual patch grounding improves both accuracy and NDCG indicating better cross-lingual ranking transfer. Github.com/AliAZ98/IMMCAN

1 Introduction

Idiomatic expressions are central to natural language, yet their meanings are often non-compositional and cannot be inferred from the literal meanings of their component words, which makes them difficult for both human learners and natural language processing (NLP) systems, especially in multilingual and cross-lingual settings (Villavicencio et al., 2005; Zeng and Bhat, 2022; Madabushi et al., 2022). The AdMIRe 2.0 shared task (Arslan et al., 2026; Torunoğlu-Selamet et al., 2026) addresses this in a multimodal setup: given a context sentence containing a potentially idiomatic expression (PIE) and five candidate images with captions, systems must rank the images by how well they reflect the idiomatic or literal meaning intended in that sentence, with supervision provided as a relative ordering rather than a single gold image, capturing the graded nature of idiom–image compatibility (Pickard et al., 2025). This setting is challenging because idiomaticity is highly context-dependent, i.e., the same expression can be idiomatic in one sentence and literal in another (Madabushi et al., 2022; Haagsma et al., 2020). The task is multilingual and largely zero-shot beyond English, requiring transferable cross-lingual representations without language-specific supervision (Conneau et al., 2019; Madabushi et al., 2022). In addition, models must integrate textual and visual information. They align the sentence–idiom meaning with both the image content and the accompanying captions. The system then ranks fine-grained candidates (idiomatic, literal, related, and distractor). Because supervised data is limited, the risk of overfitting increases. Therefore, effective use of pretrained language and vision–language models is essential (Radford et al., 2021).

In this work, we present a multimodal system for AdMIRe 2.0 with three core contributions: (i) a multilingual XLM-RoBERTa idiomaticity detector trained on MAGPIE dataset where we tested on AdMIRe 1.0 English and Portuguese, (ii) we propose the Idiom Multimodal Cross-Attention Network (IMMCAN), which fuses frozen XLM-R idiom embeddings and frozen Jina-CLIP-v2 image–caption features via two-stage cross-attention to score and rank candidates, and (iii) caption-only augmentation (back-translation and synonym substitution), with zero-shot results showing consistent improvements of the multimodal model over text-only baselines in top-image accuracy and NDCG.

2 Related Work

Pickard et al. (Pickard et al., 2025) introduced the AdMIRe (SemEval-2025) shared task, which provides a multilingual dataset of English and Portuguese sentences labeled as idiomatic or literal and paired with multiple candidate images. Par-

*These authors contributed equally.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 149–153

ticipants are asked to rank the images according to how well they match the intended idiomatic or literal meaning of the sentence, a setting that is very close to our task. Within this framework, Pan et al. (Pan et al., 2025) propose a multimodal idiom ranking system that combines pretrained text encoders such as BERT or XLM-RoBERTa with a Vision Transformer for images, and then trains a regression model to rank images by semantic alignment with the idiomatic context. Khatoon et al. (Khatoon et al., 2025) also participate in AdMIRe and focus on vision–language modeling; they introduce an "Idiom Visual Understanding Dataset" and show that a CLIP-based model, which jointly embeds images and text, clearly outperforms text-only baselines for ranking images by idiomatic meaning. Their findings support the idea that integrating visual information with text can improve idiom interpretation, which aligns with our multimodal design.

Beyond AdMIRe, several works address idiomaticity from a mainly textual perspective. Chu et al. (Chu et al., 2022) treat idiom identification as a sentence-level classification problem and fine-tune a pretrained language model on sentences containing target multiword expressions, deciding whether each expression is used idiomatically or literally. Oh (Oh, 2022) proposes NEAMER, a model that exploits the similarity between idioms and named entities; which uses XLM-RoBERTa enriched with additional surface features and transfer learning from a named entity recognition (NER) task, and achieves strong multilingual idiom classification performance. At the benchmark level, Madabushi et al. (Madabushi et al., 2022) introduce SemEval-2022 Task 2, a multilingual idiomaticity task that includes English, Portuguese, and Galician data, with subtasks on idiom detection and semantic similarity in context. This line of work shows that multilingual encoders such as XLM-R can generalize idiom recognition across languages and provides an important foundation for our use of XLM-R-based components in a multimodal idiom-understanding setting.

3 System Overview

This work targets the AdMIRe 2.0 multimodal idiom task by scoring how well each of five image–caption options matches the idiomatic or literal meaning of a sentence that contains a compound idiom. Each dataset example includes the idiom,

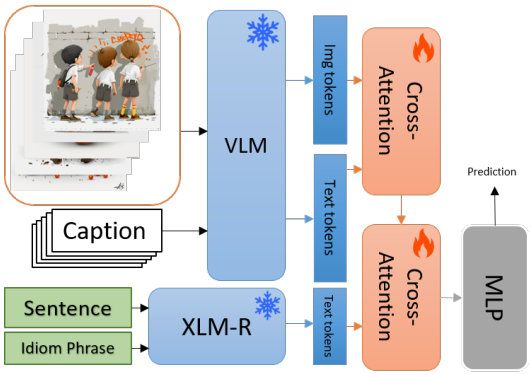


Figure 1: Overall representation of proposed IMMCAN

a sentence labeled as idiomatic or literal, and an expected ranking of the five images; the label is based on their relative semantic fit to the sentence meaning, not on visual correctness alone, so our dataloader turns each example into five text–image pairs and assigns numeric targets from this order. Our system uses a multimodal model with two frozen encoders: XLM-R produces contextual embeddings for the idiom in its sentence, and Jina-CLIP-v2 encodes each image with its caption into text tokens and visual patch features. We combine these signals with a two-step attention design: TextImageCoAttention links caption text to image features, and IMMCAN updates the idiom representation using this fused multimodal information. Finally, we pool the idiom-aware representation and use a regression or classification head to output the prediction per image, allowing the model to learn the fine-grained relationship between idiom meaning and the visual–textual candidates.

3.1 VLM Model

Our vision–language module uses Jina-CLIP-v2 (Koukounas et al., 2024), a multilingual dual-encoder that learns shared text–image representations by pairing a Jina XLM-RoBERTa text encoder (561M) with an EVA02-L/14 ViT vision encoder (304M), totaling 865M parameters. It supports 89 languages and long text inputs (up to 8,192 tokens) and processes images up to 512×512 using rotary positional embeddings and efficient attention. Both encoders produce 1024-d embeddings trained with a multi-task contrastive objective over text–text and text–image pairs, and we use the resulting token-level and global features in a frozen manner.

3.2 Idiomaticity Detection

A binary XLM-RoBERTa (Conneau et al., 2019) classifier has been trained to predict whether an ex-

pression is used idiomatically or literally, choosing XLM-R because its multilingual representations support zero-shot transfer to other languages even when training is done only on English. The model is trained on the MAGPIE dataset (Haagsma et al., 2020) using sentence–idiom pairs, where the full sentence and the target phrase are provided as separate segments to explicitly model idiom–context interaction. After training, we evaluate on the AdMIRe 1.0 English, Portuguese, and combined test sets (Table 1), and the strong Portuguese accuracy without additional supervision indicates meaningful zero-shot generalization for idiomaticity detection.

Lang.	Accuracy	Macro F1	Weighted F1
EN	0.7647	0.7607	0.7632
PT	0.6000	0.5746	0.5908
EN+PT	0.7077	0.7077	0.7077

Table 1: Model accuracy for English, Portuguese and combined on AdMIRe 1.0 dataset

3.3 IMMCAN

Figure 1 represents the proposed IMMCAN, which combines a pretrained language model and a pretrained vision–language model to build a multimodal representation for idiom understanding. The IdiomEmbedder (XLM-R) and JinaCLIPEmbedder are fully frozen and used only as feature extractors: XLM-R produces contextual idiom–sentence token embeddings, while the VLM provides caption tokens and image patch tokens from its transformer. These frozen features are fed to a small set of trainable modules, allowing us to leverage strong language and vision priors while updating few parameters. Fusion is performed with a two-step attention pipeline: TextImageCoAttention cross-attends between caption tokens and image tokens to form grounded multimodal features, which are pooled into global text and image vectors and concatenated. IMMCAN then conditions idiom tokens on this multimodal vector. Finally, a masked-mean pool produces a fixed-size embedding that an MLP head predicts the rank.

3.4 Data Augmentation

Data augmentation is used to improve robustness and reduce overfitting, but to keep idioms consistent we do not modify the idioms or their context sentences. Instead, we augment only the image captions using two methods, back-translation and synonym substitution.

Model Setup	Val	Test
TT-Reg-Base	0.4400	0.2143
TT-Reg-Aug	0.3600	0.3929
TT-Cla-Base	0.4800	0.4643
TT-Cla-Aug	**0.6400**	**0.6786**
VTT-Reg-Base	**0.4000**	**0.2857**
VTT-Reg-Aug	0.1200	0.1429
VTT-Cla-Base	0.3600	0.2500
VTT-Cla-Aug	0.0800	0.0714

Table 2: Validation and test performance of different models

For back-translation, each example's five captions are translated to an intermediate language and then translated back, using English→French→English for English captions and Portuguese→English→Portuguese for Portuguese captions, which varies syntax while keeping meaning. For synonym substitution, we rewrite captions by replacing randomly selected words at about a 15% rate, selecting candidate words by simple token filtering, e.g., minimum length, and retrieving English and Portuguese synonyms from WordNet.

4 Experimental Setup

AdMIRe 2.0 is a static image ranking task where each example provides a context sentence with a potentially idiomatic expression (PIE) and five image–caption candidates, and the system must rank the images by how well they match the intended meaning. For supervised learning, we use AdMIRe 1 (English and Portuguese) and expand each instance into five text–image candidates so the model outputs one score per candidate and recovers a full ranking at inference. Since labels are given as an ordering, we map them to either regression targets $\{1.0, 0.75, 0.50, 0.25, 0.0\}$ or classification labels $\{0, 1, 2, 3, 4\}$, and we evaluate with top image accuracy and NDCG@5 using relevance weights $[3, 1, 0, 0, 0]$. In the AdMiRe 1.0 dataset, there are 70 training, 15 validation, and 15 test samples in total. We do not use any AdMIRe 2.0 training or validation labels. All hyperparameters are tuned on AdMIRe 1.0 only, and AdMIRe 2.0 is used strictly for zero-shot evaluation.

We evaluate two variants under the same fusion design. The Vision-Text-Text (VTT) setting uses both VLM caption tokens and image patch tokens together with the idiom text stream, while Text-Text (TT) removes image patches and applies cross-attention only between caption tokens and idiom tokens to isolate the effect of visual grounding. For both settings, we compare a regression head

Lang.	TT-based				VTT-based			
	Reg-Base	Reg-Aug	Cla-Base	Cla-Aug	Reg-Base	Reg-Aug	Cla-Base	Cla-Aug
KA	0.283 (0.696)	0.327 (0.703)	0.310 (0.705)	0.248 (0.669)	0.336 (0.717)	0.106 (0.580)	**0.407 (0.747)**	0.381 (0.743)
PT-BR	0.298 (0.699)	0.303 (0.703)	0.307 (0.712)	0.237 (0.672)	0.298 (0.706)	0.044 (0.571)	0.342 (0.724)	**0.395 (0.738)**
UZ	0.325 (0.712)	0.250 (0.692)	0.258 (0.691)	0.233 (0.672)	**0.417 (0.749)**	0.117 (0.573)	0.342 (0.723)	0.383 (0.729)
ES-EC	0.167 (0.628)	0.167 (0.626)	0.229 (0.658)	0.188 (0.642)	0.333 (0.695)	0.188 (0.636)	**0.333 (0.727)**	0.333 (0.716)
NO	0.262 (0.689)	**0.302 (0.707)**	0.257 (0.705)	0.198 (0.665)	0.287 (0.698)	0.129 (0.606)	0.287 (0.705)	0.277 (0.706)
IG	0.322 (0.742)	0.374 (0.750)	**0.383 (0.739)**	0.296 (0.695)	0.296 (0.725)	0.035 (0.537)	0.365 (0.752)	0.339 (0.748)
SK	0.305 (0.712)	0.278 (0.704)	0.278 (0.706)	0.192 (0.656)	0.338 (0.723)	0.093 (0.579)	**0.411 (0.742)**	0.325 (0.716)
TR	**0.313 (0.714)**	0.308 (0.719)	0.269 (0.702)	0.264 (0.686)	0.308 (0.720)	0.099 (0.576)	0.308 (0.725)	0.280 (0.711)
RU	0.336 (0.721)	0.350 (0.725)	0.386 (0.746)	0.279 (0.673)	0.364 (0.732)	0.057 (0.572)	**0.400 (0.745)**	0.379 (0.731)
EL	0.250 (0.687)	0.245 (0.689)	0.260 (0.691)	0.236 (0.671)	0.274 (0.685)	0.120 (0.604)	**0.327 (0.712)**	0.269 (0.693)
SR	0.256 (0.695)	0.240 (0.690)	0.284 (0.703)	0.196 (0.666)	0.355 (0.718)	0.094 (0.583)	0.333 (0.722)	**0.358 (0.717)**
ZH	0.246 (0.678)	0.201 (0.661)	0.246 (0.685)	0.190 (0.640)	0.302 (0.714)	0.067 (0.580)	**0.363 (0.735)**	0.324 (0.719)
PT-PT	0.255 (0.691)	0.277 (0.694)	0.268 (0.691)	0.236 (0.674)	0.300 (0.706)	0.068 (0.576)	0.277 (0.696)	**0.318 (0.712)**
SL	0.246 (0.686)	0.254 (0.687)	0.300 (0.725)	0.238 (0.668)	0.283 (0.700)	0.083 (0.579)	**0.350 (0.726)**	0.329 (0.715)
KK	0.282 (0.706)	0.289 (0.709)	0.282 (0.705)	0.231 (0.671)	0.397 (0.739)	0.077 (0.565)	0.417 (0.752)	**0.442 (0.770)**
ALL	0.276 (0.697)	0.278 (0.697)	0.288 (0.704)	0.231 (0.668)	0.326 (0.715)	0.091 (0.578)	**0.350 (0.727)**	0.336 (0.721)

Table 3: AdMIRe Subtask A zero-shot results across TT-based and VTT-based models. Each cell reports Acc (NDCG).

(Reg) that predicts a scalar compatibility score and a classification head (Cla) that predicts rank classes, and we test two data conditions, Base captions and Aug captions generated with the method in Section 3.4.

For training, we use SmoothL1Loss with $\beta = 1.5$ for regression to reduce sensitivity to large errors, label-smoothed cross-entropy with $\epsilon = 0.3$ for VTT classification to avoid overconfident rank predictions, and a ListNet-style KL loss for TT to learn rankings in a list-wise manner. All models are trained with SGD (learning rate 0.003, weight decay 0.001), batch size 4, and an exponential learning-rate scheduler. Experiments are run on an NVIDIA RTX 3080 Ti GPU.

5 Results

Performance on AdMIRe 1.0 and multilingual zero-shot transfer results on AdMIRe 2.0 have been summarized here.

5.1 Test Results

The AdMIRe 1.0 validation/test split (see Table 2) indicated that the most reliable improvements have been obtained with the text-only classification. The highest test score has been achieved by TT-Cla-Aug implying that caption augmentation has been beneficial when a discrete rank-class target has been learned. For the multimodal variant, the best test result has been observed with VTT-Reg-Base, and performance has dropped sharply for augmented VTT configurations, which has shown that the current caption augmentation has not remained compatible with the visual-token fusion.

5.2 Zero-Shot Results

Zero-shot transfer results across AdMIRe 2.0 languages are summarized in Table 3: TT variants remain unchanged across settings, suggesting that text-only pairing provides limited cross-lingual transfer in our current setup, whereas VLM-based variants yield clearer gains across multiple languages, e.g., KA, UZ, RU, KK, indicating that explicit visual grounding improves both top-image selection and overall ranking quality. Caption augmentation shows mixed effects in zero-shot transfer, with Cla-Aug staying competitive and improving several languages, while Reg-Aug consistently degrades, which suggests that paraphrastic caption changes may add noise that harms regression-style scoring. Finally, we find that top image accuracy can be low while NDCG@5 remains high. This indicates that the model often places a highly relevant image near the top, even when it does not rank the single best image first. Because NDCG rewards near-correct ordering under graded relevance, it is less sensitive than top accuracy to swaps between the best and second-best candidates.

6 Conclusion

We presented IMMCAN for multimodal idiom understanding in AdMIRe 2.0, combining a zero-shot XLM-R idiomaticity detector with a lightweight fusion module over frozen XLM-R and Jina-CLIP-v2 features. Our results indicate that explicit visual grounding yields clearer gains in zero-shot transfer than text-only pairing, improving both top-image selection and NDCG@5 across several languages. We also found that caption augmentation has mixed effects: it tends to help when learning discrete rank classes, but it can degrade regression-based scor-

ing, likely due to added paraphrastic noise. Future work includes designing augmentation methods that preserve fine-grained ranking signals, exploring stronger list-wise objectives for multimodal settings, and training or adapting the fusion layers with more multilingual supervision to further improve cross-lingual generalization.

Acknowledgment

This work was partially funded by the ITU-Turkcell research scholarship.

References

Doğukan Arslan, Rodrigo Wilkens, Wei He, Dilara Torunoğlu-Selamet, Thomas Pickard, Aline Villavicencio, Adriana S. Pagano, and Gülşen Eryiğit. 2026. MWE-2026 Shared Task 2: AdMIRe 2 - Advancing Multimodal Idiomaticity Representation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morocco. Association for Computational Linguistics.

Zheng Chu, Ziqing Yang, Yiming Cui, Zhigang Chen, and Ming Liu. 2022. Hit at semeval-2022 task 2: Pre-trained language model for idioms detection. *arXiv preprint arXiv:2204.06145*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. arxiv. *arXiv preprint arXiv:1911.02116*, 10.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. Magpie: A large corpus of potentially idiomatic expressions. In *12th Language Resources and Evaluation Conference: LREC 2020*, pages 279–287. European Language Resources Association (ELRA).

Maira Khatoon, Arooj Kiyani, Tehmina Farid, and Sadaf Abdul-Rauf. 2025. Fjwu_squad at semeval-2025 task 1: An idiom visual understanding dataset for idiom learning. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1759–1765.

Andreas Koukounas, Georgios Mastrapas, Sedigheh Eslami, Bo Wang, Mohammad Kalim Akram, Michael Günther, Isabelle Mohr, Saba Sturua, Nan Wang, and Han Xiao. 2024. jina-clip-v2: Multilingual multimodal embeddings for text and images. *arXiv preprint arXiv:2412.08802*.

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. *arXiv preprint arXiv:2204.10050*.

Min Sik Oh. 2022. kpfriends at semeval-2022 task 2: Neamer–named entity augmented multi-word expression recognizer. *arXiv preprint arXiv:2204.08102*.

Ronghao Pan, Tomás Bernal-Beltrán, José Antonio García-Díaz, and Rafael Valencia-García. 2025. Umuteam at semeval-2025 task 1: Leveraging multimodal and large language model for identifying and ranking idiomatic expressions. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 743–749.

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, and Marco Idiart. 2025. Semeval-2025 task 1: Admire–advancing multimodal idiomaticity representation. *arXiv preprint arXiv:2503.15358*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Dilara Torunoğlu-Selamet, Dogukan Arslan, Rodrigo Wilkens, Wei He, Doruk Eryiğit, Thomas Pickard, Adriana S. Pagano, Aline Villavicencio, Gülşen Eryiğit, Ágnes Abuczki, Aida Cardoso, Alesia Lazarenka, Dina Almassova, Amalia Mendes, Anna Kanellopoulou, Antoni Brosa-Rodríguez, Baiba Saulite, Beata Wojtowicz, Bolette Pedersen, and 59 others. 2026. A parallel cross-lingual benchmark for multimodal idiomaticity understanding. *Preprint*, arXiv:2601.08645.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. Introduction to the special issue on multiword expressions: Having a crack at a hard nut.

Ziheng Zeng and Suma Bhat. 2022. Getting bart to ride the idiomatic train: Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics*, 10:1120–1137.

Sahara Tokenizers at MWE-2026 PARSEME 2.0 Subtask 1: Combining Contextual Embeddings with Structural Decoding for Multi-Word Expression Detection

Yunus Karatepe[1], Mert Sülük[1,2], Zeynep Tuğçe Kırımlı[1,3], Begüm Özbay[1,4]

[1]Istanbul Technical University
[2]Istanbul University
[3]Istanbul University-Cerrahpasa
[4]Yildiz Technical University
{karatepe22, suluk20, ozbaybe21}@itu.edu.tr, zeynep.kirimli@iuc.edu.tr

Abstract

Multi-Word Expressions (MWEs) pose a significant challenge for natural language processing systems due to their idiosyncratic semantic and syntactic properties. This paper describes our system for the PARSEME 2.0 Shared Task on automatic identification of verbal MWEs across 17 typologically diverse languages. Our approach combines multilingual BERT with explicit Part-of-Speech (POS) feature injection through a dual-head architecture that jointly performs BIO-based identification and category classification. We further investigate extensions, including Conditional Random Field (CRF) decoding for structured prediction, focal loss for addressing class imbalance, and model ensembling for improving discontinuous MWE detection. Our official submission achieves a global MWE-based F1 score of 48.39%, securing second place in the shared task. Ablation studies reveal a strong synergy between POS features and CRF decoding, with the combined approach yielding the best single-model performance. Furthermore, ensembling models trained with different objectives improves both overall F1 score and discontinuous MWE scores, demonstrating the importance of training diversity for capturing non-adjacent syntactic patterns.

1 Introduction

Multi-Word Expressions (MWEs) are idiosyncratic lexical units whose automatic identification is crucial for downstream NLP tasks (Baldwin and Kim, 2010). The PARSEME shared task series (Savary et al., 2017; Ramisch et al., 2020) addresses the core challenges of MWE detection: handling discontinuous surface realizations, managing severe class imbalance, and generalizing across typologically diverse languages.

In this work, we present a system for PARSEME 2.0 (Subtask 1) (Scholivet et al., 2026), which addresses the joint identification and classification of verbal MWEs. Our system is based on multilingual BERT (Devlin et al., 2019) augmented with explicit Part-of-Speech (POS) features. This approach employs a multi-task formulation with dual prediction heads for joint boundary identification (BIO) and category classification. Beyond the official submission, we investigate architectural extensions including Conditional Random Fields (CRF) for structured decoding, Focal Loss (Lin et al., 2017) for mitigating imbalance, and diverse ensembling strategies.

Our main contributions are as follows:

- We demonstrate that POS feature injection improves recall by 2.78 points, with strong gains on seen expressions (+1.64% F1).
- We identify a critical synergy between POS features and CRF decoding: while POS injection alone yields marginal gains, coupling it with structural constraints produces our best single-model result (70.60% F1).
- We show that ensembling models trained with diverse objectives (Cross-Entropy and Focal Loss) improves discontinuous MWE detection (+2.33 F1 in French).
- Our official submission achieves 48.39% global F1, ranking second in the Shared Task.

2 Related Work

Automatic identification of Multi-Word Expressions (MWEs) remains a core challenge in multilingual NLP due to their idiomaticity, non-

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 154–159

compositional semantics, discontinuity, and annotation sparsity. Foundational linguistic characterisation and formal language perspectives on alternating sequence computation and structured labeling complexity trace back to early formal studies such as alternation theory and lexical systematisation, which later informed NLP-oriented taxonomies for MWEs and decoding principles (Chandra et al., 1981; Baldwin and Kim, 2010). A broad survey of MWE processing highlights persistent issues, including sparse observations, idiomaticity, discontinuity, and cross-lingual variation, motivating architectures that combine contextual representations with explicit linguistic inductive biases.

A large body of work formulates MWE identification as structured sequence labeling, adopting token-level tagging schemes (e.g., BIO/BILOU) and log-linear structured decoders. BiLSTM-CRF models established strong baselines for enforcing tag consistency and segment boundaries in linear-chain CRF formulations with Viterbi/Viterbi-style inference (Lample et al., 2016; Ma and Hovy, 2016; Lafferty et al., 2001). Contextualized Transformer encoders, especially BERT, have since become the dominant representation backbone for sequence labeling tasks, including MWEs (Devlin et al., 2019). However, even with contextual encoders, tagging systems remain sensitive to (i) extreme class imbalance (most tokens are `O`), (ii) discontinuous MWEs, and (iii) recall and generalization to unseen expressions, particularly in multilingual blind-test scenarios.

The PARSEME shared task series created standardized multilingual corpora, annotation guidelines, and evaluation protocols for verbal MWEs (VMWEs), foregrounding both continuous and discontinuous expressions and enabling systematic cross-lingual evaluation (Savary et al., 2017). Later task editions emphasised generalization to unseen VMWEs through carefully constructed blind-test splits containing expressions not observed during training (Ramisch et al., 2020). Competitive neural systems commonly augment taggers with linguistic or structural signals: ERMI injects POS and dependency features in a BiLSTM–CRF architecture (Yirmibeşoğlu and Güngör, 2020), while MTLB-STRUCT frames VMWE identification under a multi-task paradigm by incorporating auxiliary syntactic structure and using a dual tagging head on multilingual BERT with CRF decoding, analyzing imbalance-aware objectives such as focal loss and the role of structured decoders for improved discontinuity resolution and unseen-expression recall (Taslimipoor et al., 2020). Explicit long-range relation modelling for bridging syntactic gaps was studied in discontinuity-focused settings (Rohanian et al., 2019), motivating dependency-based syntactic path reasoning to connect separated MWE components. Relational syntactic inference in multilingual pipelines was further shaped by deterministic and biaffine dependency parsers and benchmarks, along with trainable parsing frameworks such as UDPipe (Nivre, 2008; Dozat and Manning, 2017; Straka et al., 2016). Beyond syntactic biasing, contrastive and self-supervised sequence objectives such as those introduced by (Jaiswal et al., 2021; Gao et al., 2021) strengthened general sequence representations and multi-task optimisation dynamics were later systematized in predictive structure learning and representation surveys (Ando and Zhang, 2005; Ruder, 2017).

Our system aligns with the Transformer-based sequence labeling framework, enhanced by POS injection for syntactic bias, CRF decoding for structural consistency, and Focal Loss to mitigate class imbalance. Additionally, we leverage ensembling to capture diverse error profiles, which is crucial for robust discontinuous and unseen MWE detection.

3 Methodology

We frame MWE identification as a sequence labeling problem requiring the detection of continuous and discontinuous expressions under heavy class imbalance. Our approach utilizes a multi-task architecture predicting both identification (BIO) and classification (Category) labels.

3.1 Submitted System

Our official submission employs `bert-base-multilingual-cased` as a shared encoder. We derive two aligned supervision signals: (1) **BIO tags** for identification (`B`, `I`, `O`), and (2) **MWE categories** for classification.

POS Injection and Dual-Head. We augment contextual embeddings by concatenating them with learned POS tag embeddings. As shown in Figure 1, this fused representation $\tilde{h}_i$ feeds into two parallel linear heads. The BIO head predicts identification tags, while the Category head assigns MWE types.

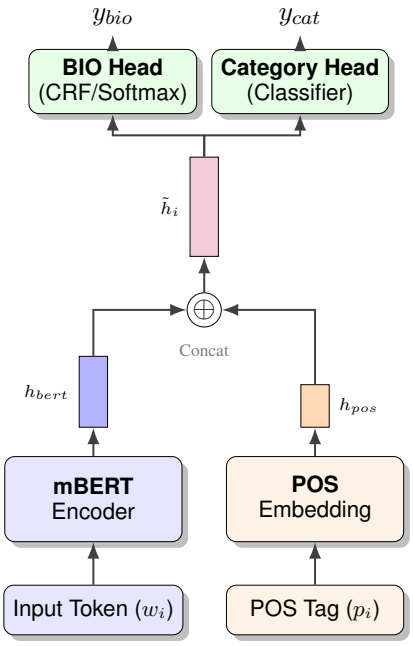


Figure 1: Architecture at a single time-step i. POS embeddings are concatenated with mBERT output to feed dual prediction heads.

3.2 Extensions

To improve robustness against multilingual interference and label imbalance, we investigate the following extensions.

3.2.1 CRF Decoding

To enforce valid label transitions, we replace the token-wise softmax with a Linear-Chain CRF. The model maximizes the score of the correct BIO sequence y, as defined in Equation 1.

$$\text{Score}(y) = \sum_{i=1}^{n} \left(A_{y_{i-1}, y_i} + s_{i, y_i}^{bio} \right), \qquad (1)$$

where A represents transition parameters. Inference is performed via Viterbi decoding.

3.2.2 Consistency via Masking

To align the two heads, we compute category loss only for tokens predicted as MWEs by the identification head. We apply a mask $m_i = \mathbb{I}(\hat{y}_i^{bio} \neq \texttt{0})$ to the category loss, ensuring that the classifier focuses only on valid MWE candidates and ignores the majority `0` class.

3.2.3 Focal Loss

To address the dominance of non-MWE tokens, we employ Focal Loss (FL) to down-weight easy negatives and focus training on hard examples, utilizing the formulation in Equation 2.

$$\text{FL}(p_t) = -\alpha(1 - p_t)^{\gamma} \log(p_t). \qquad (2)$$

We compare FL against standard Cross-Entropy (CE) to evaluate its impact on recall.

3.2.4 Ensemble Strategy

We construct ensembles by aggregating predictions from K diverse models (Base, POS-injected, and Focal Loss variants). For **category classification**, we simply average the probability distributions according to Equation 3.

$$\bar{p}_i^{cat} = \frac{1}{K} \sum_{k=1}^{K} p_{i,k}^{cat}. \qquad (3)$$

For **BIO identification**, we use a hybrid scheme: standard softmax models use probability averaging, while CRF-based models use **majority voting** on Viterbi sequences to preserve structural validity. This strategy combines the high recall of Focal Loss with the precision of standard baselines.

4 Experimental Results

We evaluate our proposed system on the PARSEME 2.0 blind test set, analyzing global metrics, language-specific performance, and error categories.

4.1 Submitted System Results

Table 1 compares our system against the multilingual BERT baseline.

Impact of POS Injection. Injecting POS features improved Global F1 to **48.39%**, driven by a gain in **Recall (+2.78 points)**. This indicates that explicit morphosyntactic information aids in recognizing valid candidates missed by the baseline.

Generalization & Memorization. We observe a performance disparity between *Seen* and *Unseen* MWEs. The POS-enhanced model excelled at **Seen MWEs** (73.70% F1, +1.64% gain), suggesting that POS tags reinforce confidence in learned syntactic templates (e.g., *Verb+Noun*). However, performance on **Unseen MWEs** remained low (∼20%), indicating that while explicit syntax aids pattern matching for known expressions, it offers limited benefit for zero-shot generalization.

Configuration	Global MWE-based			Global Token-based			Generalization (F1)	
	Prec.	Rec.	F1	Prec.	Rec.	F1	Seen	Unseen
Base BERT	**46.49**	48.55	47.50	**61.95**	53.83	57.61	72.06	**20.23**
POS Features + BERT	45.77	**51.33**	**48.39**	61.53	**57.12**	**59.24**	**73.70**	19.98

Table 1: Blind test set results. POS injection improves Global F1 and Recall, particularly for Seen MWEs.

Lang	Family	P	R	F1
High Performance				
FA	Indo-Iranian	70.67	77.29	73.83
JA	Japonic	75.92	70.00	72.84
RO	Romance	61.98	71.12	66.23
Mid Performance				
PL	Slavic	51.84	64.80	57.60
HE	Semitic	52.89	60.28	56.34
FR	Romance	52.60	50.50	51.53
Low Performance				
KA	Kartvelian	26.17	69.40	38.01
EGY	Semitic	33.67	13.20	18.97
GRC	Hellenic	8.81	6.01	7.14

Table 2: MWE-based F1 scores for representative languages.

4.1.1 Language-Specific Analysis

Table 2 details performance across diverse language families. High-resource languages with distinct syntactic markers (FA, JA) achieved the highest scores ($> 72\%$ F1). Conversely, low-resource or ancient languages (GRC, EGY) suffered from data sparsity. Notably, Georgian (KA) exhibited high recall but low precision (26.17%), suggesting systematic over-prediction likely driven by severe class imbalance and language-specific noise.

4.2 Ablation Studies: Extensions

We analyze extensions on the development set using a 90/10 split, focusing on two setups: **Monolingual** (French only) and **Multi-5** (FR, SV, EL, FA, JA).

4.2.1 Monolingual Case Study (French)

Table 3 highlights the interaction between linguistic features and decoding strategies.

Configuration	Global F1	Disc. F1
Single Models		
Base (mBERT)	67.59	55.65
`pos`	67.43	51.98
`pos_crf`	**69.28**	55.32
`crf`	68.06	52.94
`crf_focal`	67.61	53.28
`pos_crf_focal`	69.16	**56.41**
Ensemble Models		
`base+pos_crf+crf_focal`	69.50	54.55
`base+pos_crf+pos_crf_focal`	**69.92**	**57.98**

Table 3: Ablation on French (FR) development set.

Configuration	Global F1	Disc. F1
Single Models		
Base (mBERT)	69.84	47.16
`pos`	69.22	47.16
`pos_crf`	**70.60**	**47.67**
`crf`	69.45	45.29
`crf_focal`	69.39	46.19
`pos_crf_focal`	69.18	46.37
Ensemble Models		
`base+pos_crf+crf_focal`	**70.73**	**48.88**
`base+pos_crf+pos_crf_focal`	70.22	48.06

Table 4: Ablation on Multi-5 (FR, SV, EL, FA, JA) set.

POS-CRF Synergy. Injecting POS tags alone (`pos`) slightly degraded performance. However, coupling POS with CRF decoding (`pos_crf`) yielded the best single-model result (69.28%). This indicates that while POS tags provide valuable signals, the model requires the structured transition constraints of a CRF to utilize them effectively without overfitting.

Discontinuity via Ensembling. Single models struggled with discontinuous MWEs. However, the ensemble approach achieved a Discontinuous F1 of **57.98%** (+2.33 points over baseline). Averaging probability distributions from diverse models (Base + POS + Focal) effectively bridges syntactic gaps that single architectures miss.

4.2.2 Multilingual Analysis (Multi-5)

Table 4 summarizes the multilingual ablation results.

Synergy Consistency. Consistent with monolingual findings, `pos_crf` achieved the highest single-model F1 (70.60%), reversing the degradation seen with POS alone. This confirms that CRF constraints are essential for leveraging morphosyntactic cues across diverse languages.

Focal Loss & Diversity. While Focal Loss models underperformed in isolation, they were critical for the ensemble. The best system (Base + POS + Focal) reached **70.73% Global F1** and the highest **Discontinuous F1 (48.88%)**, confirming that diverse training objectives capture "hard" examples that standard models fail to detect.

5 Conclusion

We presented a multilingual MWE identification architecture combining mBERT with explicit POS features, which ranked second in the PARSEME 2.0 Shared Task (48.39% F1). Our experiments demonstrate a critical synergy between morphosyntactic features and structured decoding: while POS injection alone yields marginal gains, coupling it with a CRF layer effectively constrains the output space, achieving our best single-model performance. Furthermore, addressing the challenge of discontinuity, we showed that ensembling models trained with Focal Loss improves recall on non-adjacent expressions. Future work will further explore the integration of linguistic constraints into end-to-end training.

Limitations

Our study has several limitations that are important for interpreting the results and for guiding future improvements.

Reliance on POS quality. POS feature injection is beneficial when tags are accurate and consistent across languages; however, in low-resource or morphologically complex languages, tagging errors may propagate into MWE predictions and lead to unstable precision/recall trade-offs.

Unseen MWE generalization remains difficult. While our approach improves recall for seen expressions, performance on unseen MWEs remains a major bottleneck, suggesting the model still relies on distributional regularities observed during training rather than type-level constraints or composition-aware cues.

Token-level formulation and discontinuity. We use token-level BIO supervision and CRF decoding, which enforces local label consistency, but we do not explicitly model expression-level completeness or gap-aware structure. This can yield fragmented boundaries, particularly for discontinuous MWEs, where explicit gap modeling or syntactic integration may be necessary.

Compute and deployment cost. Ensembling improves robustness and discontinuous detection, yet increases inference time and memory. Distillation or lightweight diversity-preserving alternatives could make the approach more deployable.

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing*, pages 267–292. CRC Press.

Ashok K. Chandra and 1 others. 1981. Alternation and structured sequence complexity. *Journal of the ACM*, 28:114–133.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Timothy Dozat and Christopher Manning. 2017. Deep biaffine neural dependency parser. In *ICLR*, pages 1–12. OpenReview.

Tianyu Gao and 1 others. 2021. Simcse: Simple contrastive learning of sentence embeddings. *EMNLP*, arXiv. Contrastive sequence representation learning.

Ashish Jaiswal and 1 others. 2021. Survey on contrastive learning. *Journal of AI Research*, Survey. Systematization of self-supervised contrastive objectives.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML*, arXiv. Linear-chain CRF with Viterbi inference.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, Venice, Italy.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Joakim Nivre. 2008. Deterministic dependency parsing algorithms. In *ACL*, pages 513–520. Association for Computational Linguistics.

Carlos Ramisch and 1 others. 2020. Edition 1.2 of the parseme shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118. Association for Computational Linguistics.

Omid Rohanian, Shiva Taslimipoor, Samaneh Kouchaki, Le An Ha, and Ruslan Mitkov. 2019. Bridging the gap: Attending to discontinuity in identification of multiword expressions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2692–2698.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv*. Survey on multi-task learning and training diversity.

Agata Savary and 1 others. 2017. The parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions*, pages 31–47. Association for Computational Linguistics.

Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Carp, and Vasile Pais. 2026. Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

Milan Straka and 1 others. 2016. Udpipe: Trainable pipeline for multilingual dependency parsing. In *LREC*, pages 4290–4297. European Language Resources Association.

Shiva Taslimipoor, Sara Bahaadini, and Ekaterina Kochmar. 2020. MTLB-STRUCT @parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 142–148, online. Association for Computational Linguistics.

Zeynep Yirmibeşoğlu and Tunga Güngör. 2020. Ermi at parseme shared task 2020: A sequence labeling approach. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 140–144.

3K2T at MWE-2026 AdMIRe 2: CARIM– Category-Aware Reasoning for Idiomatic Multimodality

Kubilay Kağan KÖMÜRCÜ
Istanbul Technical University
komurcu17@itu.edu.tr

Tuğçe TEMEL
Istanbul Technical University
temel21@itu.edu.tr

Abstract

Idiomatic expressions pose a fundamental challenge for multimodal understanding due to their non-compositional semantics, while pretrained vision–language models tend to overrely on literal visual alignments. We address this issue in the context of the AdMIRe 2.0 multimodal idiomatic image ranking task (Arslan et al., 2026) by introducing CARIM (Category-Aware Reasoning for Idiomatic Multimodality), an inference-time framework that injects structured semantic reasoning without end-to-end retraining.Experiments on the official Codabench leaderboard demonstrate that CARIM achieves competitive Top-1 Accuracy and nDCG across multiple languages. Additional post-competition evaluation on the released test annotations further shows that CARIM maintains robust multilingual performance, highlighting the effectiveness of inference-time category-aware reasoning for multimodal idiomatic grounding.

1 Introduction

Idiomatic expressions are pervasive in natural language and pose a fundamental challenge for natural language processing due to their non-compositional semantics, where the figurative meaning cannot be inferred from the meanings of individual words. Correctly interpreting idioms is essential for downstream tasks such as machine translation, visual-language understanding, and language grounding.

While recent contextualized language models have improved idiom detection and interpretation by leveraging surrounding context, they remain largely text-centric and often struggle to distinguish figurative meaning from literal interpretations in ambiguous settings(Shwartz and Dagan, 2019; Garcia and García-Serrano, 2018). This limitation becomes more evident in multimodal scenarios, where models must associate an idiomatic expression with a visual representation that reflects its figurative meaning rather than its literal constituents (Liu et al., 2022; Ma et al., 2023).

Motivated by the strong imagery evoked by idiomatic language, recent work has explored grounding idioms in visual space using vision-language models (Liu et al., 2022). However, models pretrained on large-scale literal image–text pairs tend to align idioms with literal visual concepts, leading to incorrect or ambiguous mappings. This highlights the need for learning objectives that explicitly separate idiomatic and literal meanings across modalities (Ma et al., 2023).

In this work, we study idiom understanding from a multimodal perspective and formulate idiom-image association as an image-text ranking problem and our contributions are:

- We introduce CARIM, an inference-time category-aware framework for multimodal idiom–image ranking that separates literal and idiomatic visual evidence without retraining.
- We propose a structured category model and category-conditioned ranking rules that penalize misleading literal visual evidence under idiomatic usage.
- We evaluate CARIM on the AdMIRe 2.0 shared task and through a post-competition multilingual analysis, demonstrating robust performance across diverse languages.

2 Related Work

Idiomatic expressions pose a challenge for NLP due to their non-compositional semantics (Sag et al., 2002). Early work relied on lexical resources and rule-based identification of multiword expressions (Baldwin and Kim, 2010), but these approaches struggled with contextual variability.

Distributional methods showed that standard word embeddings are biased toward literal meanings and insufficient for modeling idioms (Salehi

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 160–164

et al., 2015). Contextualized language models such as BERT significantly improved idiom understanding by leveraging context, commonly framing the task as idiom detection or literal–figurative classification (Shwartz and Dagan, 2019; Garcia and García-Serrano, 2018). Several studies further explored idiom representation learning and paraphrase-based supervision (King and Cook, 2016; Peng and Feldman, 2018).

More recently, idiom learning has been extended to multimodal settings, motivated by the strong imagery associated with figurative language. Vision–language models have been used to align idiomatic expressions with figurative images while contrasting them against literal visual interpretations (Liu et al., 2022; Ma et al., 2023). These works typically formulate the problem as image–text ranking or contrastive learning, demonstrating that visual grounding provides complementary semantic signals beyond text-only models.

3 Methodology

In this section, we introduce CARIM, our inference-time category-aware reasoning model for multimodal idiomatic ranking.

3.1 CARIM Overview

At inference time, our system performs category-aware image ranking conditioned on the contextual usage of a compound expression. While the overall framework may incorporate learned components in earlier stages, the method described in this section operates exclusively at **prediction time** and does not update model parameters. Instead, it applies a structured reasoning procedure that encodes domain knowledge about idiomatic image categories directly into the inference process.

Given a compound expression c, a context sentence s, and a set of candidate images $\mathcal{I} = \{I_1, \ldots, I_5\}$ (with optional captions), the inference module outputs (i) a ranked list over the images and (ii) a compound-type label indicating whether c is used literally or idiomatically.

Inference Decomposition: The inference procedure is decomposed into two sequential steps:

1. **Compound-type inference**, which determines whether the compound is used literally or idiomatically in context.
2. **Category-aware image ranking**, which applies expert-defined ranking rules conditioned on the inferred sentence type.

This decomposition explicitly models the dependency between contextual meaning and visual relevance, reducing ambiguity compared to single-step ranking.

Sentence-Type Inference: In the first step, the model infers a sentence-type label

$$y \in \{\textsc{Literal}, \textsc{Idiomatic}\}$$

for the compound expression c as used in sentence s. A LITERAL usage denotes reference to the physical objects or properties described by the component words of c, whereas an IDIOMATIC usage denotes a figurative meaning that cannot be derived compositionally from the literal referents.

This decision is produced at inference time via a constrained prediction that conditions on c and s, and serves as a control signal for the subsequent ranking step.

Category Model: The ranking strategy is grounded in a category model that reflects a consistent structure in the candidate image sets. Each image is assumed to fall into one of five semantic categories relative to the compound expression:

1. **Literal (L):** depicts the literal referents of the component words of c.
2. **Literal-related (LR):** partially matches the literal meaning.
3. **Idiomatic (I):** directly depicts the figurative meaning of c.
4. **Idiomatic-related (IR):** loosely supports the figurative meaning.
5. **Distractor (D):** superficially related but incorrect for the intended meaning.

Literal object detection: To support category assignment, the compound is decomposed into its component words

$$c \rightarrow \{w_1, \ldots, w_m\}.$$

At inference time, the Chatgpt 5.1 evaluates whether an image contains salient visual evidence corresponding to these literal referents. This signal is used differently depending on the inferred sentence type.

Category-Aware Ranking Rules: The final ranking is produced by applying deterministic rules conditioned on y.

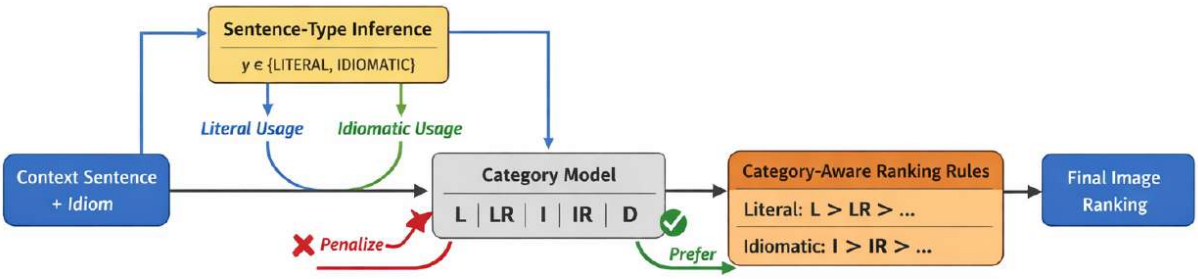


Figure 1: Overview of the proposed inference-time category-aware ranking framework, CARIM. Given a context sentence and a compound expression, the system first infers whether the expression is used literally or idiomatically. Based on this decision, candidate images are categorized into semantic types (L, LR, I, IR, D) and ranked using category-conditioned rules that penalize misleading literal visual evidence under idiomatic usage.

Literal usage. If $y = \textsc{Literal}$, images are ranked according to:

$$\mathbf{L} \succ \mathbf{LR} \succ \mathbf{IR} \succ \mathbf{I} \succ \mathbf{D}.$$

Images depicting the literal referents are preferred, while purely idiomatic depictions are down-ranked.

Idiomatic usage: If $y = \textsc{Idiomatic}$, the ranking order is:

$$\mathbf{I} \succ \mathbf{IR} \succ \mathbf{LR} \succ \mathbf{L} \succ \mathbf{D}.$$

Crucially, images with strong literal evidence (**L**) are explicitly penalized under idiomatic usage, as they typically correspond to incorrect surface-level interpretations.

Inference-Time Composition: The inference module combines the two steps as:

$$y = f_{\text{infer}}(c, s), \qquad \pi = f_{\text{rank}}(c, s, \mathcal{I} \mid y),$$

where f_{infer} denotes sentence-type inference and f_{rank} denotes category-aware ranking. Although parameter learning may occur elsewhere in the system, this module performs structured reasoning exclusively at inference time, interpretable and context-sensitive image ranking.

Illustrative Examples: For the compound *bad apple*, literal usage prioritizes images depicting a spoiled apple, whereas idiomatic usage prioritizes images depicting a corrupting individual and penalizes literal apples. Similarly, for *green fingers*, idiomatic usage favors gardening-related imagery while down-ranking images that merely depict green-colored fingers.

4 Results

We evaluate CARIM using the official Codabench leaderboard of the AdMIRe 2.0 shared task. Results are reported in terms of Top-1 Accuracy and nDCG, following the task evaluation protocol. Since our submission was evaluated on a subset of languages, we report results for Igbo (IG), Kazakh (KK), Turkish (TR), and Uzbek (UZ), together with the average across these languages. Full leaderboard results are shown in Table 1.

Overall, our system,CARIM achieves competitive performance across all evaluated languages without relying on additional end-to-end training or task-specific fine-tuning. This suggests that structured inference-time reasoning can be effective for multimodal idiomatic ranking even when operating on frozen backbone models.

Following the AdMIRe 2.0 shared task, we conducted a post-competition evaluation using the released test annotations to assess multilingual generalization. As shown in Table 2, CARIM achieves a macro-average Top-1 Accuracy of 57.2% and an average nDCG of %83.5 across all evaluated languages, without additional fine-tuning. These results demonstrate consistent multilingual performance of inference-time category-aware reasoning beyond the competitive setting.

4.1 Language-Specific Performance

As shown in Table 1, the proposed method performs strongest on Kazakh (KK), achieving its highest Top-1 Accuracy and nDCG among the evaluated languages. This indicates that the category-aware ranking strategy is particularly effective in settings where literal visual cues act as strong distractors, and where explicit penalization of literal imagery under idiomatic usage provides clearer

Participant	AVG		IG		KK		TR		UZ	
	Acc.	nDCG	Acc.	nDCG	Acc.	nDCG	Acc.	nDCG	Acc.	nDCG
ozgeumut	0.6075	0.86	0.56	0.84	0.70	0.89	0.65	0.88	0.52	0.83
ITUNLP	0.56	0.8375	0.43	0.78	0.61	0.84	0.68	0.90	0.52	0.83
kkkomurcu	**0.48**	**0.8025**	**0.41**	**0.76**	**0.56**	**0.84**	**0.54**	**0.83**	**0.41**	**0.78**
davidcotiga	0.4675	0.78	0.39	0.74	0.53	0.80	0.62	0.84	0.33	0.74
tiberiucarp	0.445	0.78	0.37	0.74	0.51	0.81	0.48	0.80	0.42	0.77
bilenbaris	0.3575	0.735	0.30	0.73	0.40	0.74	0.31	0.72	0.42	0.75
nikoniko	0.3225	0.7225	0.33	0.73	0.33	0.74	0.34	0.71	0.29	0.71
utkucolakitu	0.275	0.7025	0.22	0.69	0.28	0.71	0.29	0.70	0.31	0.71
akkurt_buzlu	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 1: Official Codabench leaderboard results for AdMIRe 2.0. Scores are reported as Top-1 Accuracy and nDCG, averaged across four evaluated languages (AVG) and for individual languages IG, KK, TR, and UZ .

Language	Acc.	nDCG
ZH	0.497	0.792
KA	0.531	0.808
EL	0.639	0.868
IG	0.409	0.761
KK	0.564	0.838
NO	0.644	0.873
PT-BR	0.864	0.942
PT-PT	0.636	0.866
RU	0.714	0.893
SR	0.579	0.826
SK	0.556	0.833
SL	0.708	0.887
ES-EC	0.292	0.727
TR	0.538	0.833
UZ	0.408	0.782
Average	**0.572**	**0.835**

Table 2: Top-1 Accuracy and nDCG scores of **CARIM** across languages on the AdMIRe 2.0 evaluation set (Torunoğlu-Selamet et al., 2026). The last row reports the macro-average over all languages.

separation.

Performance on TR, IG, and UZ language datsets remain competitive relative to other submissions on the leaderboard. While absolute scores vary across languages, the consistency between Top-1 Accuracy and nDCG suggests stable ranking behavior rather than isolated correct predictions. This aligns with the design goal of producing coherent, context-sensitive rankings instead of optimizing solely for the top-ranked image.

Notably, these results are obtained without modifying model parameters at inference time. In contrast to approaches that rely on contrastive retraining or language-specific adaptation, our method incorporates task knowledge through an explicit category model and deterministic ranking rules. This allows the system to systematically down-rank literal visual interpretations when idiomatic usage is inferred, addressing a common failure mode of pretrained vision-language models.

5 Conclusion

We presented an inference-time, category-aware ranking approach,CARIM for multimodal idiomatic understanding, motivated by the tendency of pretrained vision–language models to favor literal visual alignments. By explicitly decomposing inference into sentence-type prediction and category-conditioned image ranking, our method injects structured task knowledge without requiring end-to-end retraining or parameter updates. This design enables interpretable control over ranking behavior and directly addresses literal bias in idiom–image association.

Evaluation on the AdMIRe 2.0 Codabench leaderboard demonstrates that this lightweight reasoning framework achieves competitive performance across multiple languages, despite operating on frozen backbone models and limited supervision. The results suggest that explicit category reasoning at inference time can serve as an effective complement to learned multimodal representations, particularly in figurative language settings where surface-level visual similarity is misleading.

Acknowledgments

This work was conducted as part of the course BLG505-NLP at ITU. The authors would like to thank Prof. Gülşen Eryiğit for helpful comments and suggestions.

References

Doğukan Arslan, Rodrigo Wilkens, Wei He, Dilara Torunoğlu-Selamet, Thomas Pickard, Aline Villavicencio, Adriana S. Pagano, and Gülşen Eryiğit. 2026. MWE-2026 Shared Task 2: AdMIRe 2 - Advancing Multimodal Idiomaticity Representation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morocco. Association for Computational Linguistics.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of Natural Language Processing*.

Marcos Garcia and Ana García-Serrano. 2018. Towards deep learning of idiomaticity. In *Proceedings of EMNLP*.

Milton King and Paul Cook. 2016. Verifying semantic compositionality of multiword expressions. In *Proceedings of ACL*.

Fangyu Liu, Xiaowei Zhai, and Joyce Chai. 2022. Multimodal idiom understanding. In *Proceedings of ACL*.

Yukun Ma, Xiang Chen, and Zhiyuan Liu. 2023. Figurative language grounding with vision–language models. In *Proceedings of EMNLP*.

Jing Peng and Anna Feldman. 2018. Neural network models for idiom detection. In *Proceedings of LREC*.

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 Computational Linguistics and Intelligent Text Processing*.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of NAACL*.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. In *Proceedings of ACL*.

Dilara Torunoğlu-Selamet, Dogukan Arslan, Rodrigo Wilkens, Wei He, Doruk Eryiğit, Thomas Pickard, Adriana S. Pagano, Aline Villavicencio, Gülşen Eryiğit, Ágnes Abuczki, Aida Cardoso, Alesia Lazarenka, Dina Almassova, Amalia Mendes, Anna Kanellopoulou, Antoni Brosa-Rodríguez, Baiba Saulite, Beata Wojtowicz, Bolette Pedersen, and 59 others. 2026. A parallel cross-lingual benchmark for multimodal idiomaticity understanding. *Preprint*, arXiv:2601.08645.

PMI MWE Scorer at PARSEME 2.0 Subtask 1: identifying multi-word expressions using pointwise mutual information and universal dependencies

Syntax-Aware PMI for Multi-Word Expression Identification Using Universal Dependencies

Anna Bogdanova and Ileana Bucur
Eberhard Karls Universität Tübingen
Tübingen, Germany

Abstract

Multi-word expressions (MWEs) remain a challenge for NLP systems due to their syntactic variability and non-compositional semantics, that is why this issue was proposed as shared task within Unidive organization. With increasing popularity of large language models (LLM), it is important to continue researching alternative solutions. One of the classical approaches for identifying MWEs is calculating Pointwise Mutual Information (PMI), but this is a purely statistical approach that cannot reveal the links between words in natural text. To fix this issue, we propose this paper with a simple syntax-aware PMI method that leverages Universal Dependency (UD) trees (Nivre et al., 2016) to model co-occurrence between syntactically related words. By computing PMI over dependency-linked word pairs and aggregating these scores, we aim to improve surface-based methods. Unlike expectations, our experiment shows that the classical statistical approach gets better results in partially identifying MWEs. Still, this approach is aimed to find a balance between lightweight calculations as opposed to LLMs and precision in results.

1 Introduction

Multi-word expressions (MWEs), such as *in spite of*, *take place*, or *make sense*, are pervasive in natural language and pose long-standing challenges for NLP systems. Their meaning does not equal to the sum of meanings of its elements, and their surface forms may vary syntactically, making them difficult to identify using purely lexical or contextual cues.

PMI is calculated for each pair of adjacent words, which is not suitable for this kind of task because parts of multiword expressions can be located far away from each other in a sentence. It is possible to use a skip-gram, but if we have a look at an example sentence:

(1) *She turned the proposal that had been debated for months by several committees down.*

we would see that the phrasal verb and its particle are located 11 words away from each other, and classical PMI would not identify an MWE in this example sentence.

To deal with this issue as part of our participation in the shared task, we propose using dependency analysis based on Universal Dependencies (UD) as a basis for calculating the PMI score. As opposed to classical PMI, two words are considered a pair not if they are adjacent in a sentence, but only if they are connected by an edge of any type. In this scenario, turned and down form a pair because the particle depends on its verb.

UD provides cross-linguistically consistent dependency relations that explicitly encode syntactic relationships independent of surface adjacency. Furthermore, UD is language-independent. To mitigate unnecessary variability in PMI pairs for languages with rich morphology, we utilize UD lemmas rather than surface forms. In conclusion, our research provides insights into

- Analysis of the impact of UD on MWE detection precision
- Computation complexity difference between adjacency based PMI and UD-based PMI scoring.

2 Method

Our approach extends traditional PMI-based MWE identification by redefining word co-occurrence in terms of syntactic dependency relations rather than linear adjacency. While classical PMI relies on surface proximity to approximate lexical association, such an assumption is often violated by multi-word expressions whose components may be syntactically related but linearly distant. To address this limitation, we exploit Universal Dependency (UD)

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 165–169

parsers to define co-occurrence over syntactic structure instead of surface order, while preserving the lightweight and language-independent nature of PMI-based methods.

2.1 Pointwise Mutual Information

Pointwise Mutual Information (PMI) is a widely used association measure that quantifies the strength of co-occurrence between two lexical items relative to their independent occurrence probabilities. Formally, PMI between two words w_1 and w_2 is defined as:

$$PMI(w_1, w_2) = \log \frac{P(w_1, w_2)}{P(w_1)\,P(w_2)}$$

where $P(w_1, w_2)$ denotes the joint probability of the two words co-occurring under a given definition of co-occurrence, and $P(w_1)$ and $P(w_2)$ are their individual probabilities.

To mitigate sparsity and avoid undefined values caused by zero-frequency events (Jurafsky and Martin, 2025), we apply Laplace smoothing to all probability estimates, resulting in the following formulation:

$$\log \frac{P(w_1, w_2) + 1}{(P(w_1) + 1)(P(w_2) + 1)}$$

All counts are computed over lemmatized tokens in order to reduce morphological variation across languages. Punctuation symbols and stop-words are excluded from the computation, as they are unlikely to contribute to the identification of meaningful lexicalized expressions. We assume that few or no MWEs contain stop-words; thus, their exclusion results in a negligible number of false negatives.

2.2 Surface PMI Baseline

As a baseline, we compute PMI for adjacent word pairs occurring within a fixed window of size two. Under this setting, two lemmas are considered to co-occur if they appear consecutively in the corpus, and wheneer two lemmas appear together, the counter increments.

2.3 Dependency-Based PMI

In the dependency-based variant, the PMI formula remains unchanged, but the definition of co-occurrence is altered. Two lemmas are considered to form a co-occurring pair if they are directly connected by an edge in the Universal Dependency representation of a sentence, regardless of their linear distance.

In this manner, syntactically related words may form a pair even when separated by multiple intervening tokens. All dependency relations are treated uniformly, without restricting the computation to specific relation types.

Both the proposed method and classical PMI approaches are language-agnostic, but the dependency-based approach is considered to be more linguistically aware.

2.4 MWE Identification

To identify MWEs, all word pairs are ranked according to their PMI scores. Pairs exceeding a predefined threshold are marked as potential MWEs. Determining an optimal threshold is a significant challenge, as PMI values are not directly comparable across languages or datasets and are sensitive to corpus size and distributional properties. Rather than imposing a single threshold, we report results for multiple percentile thresholds.

3 Experimental Setup

3.1 Dataset

We evaluate our approach on a dataset provided by shared task organisers (Savary et al., 2023).

3.2 Evaluation Metrics

To evaluate our system, we adopted the official metrics of the original shared task.

The performance is assessed in terms of precision, recall, and F1-score at both the MWE level (exact match) and the token level (allowing partial matches). Diversity is measured through richness, Shannon evenness, and Shannon–Weaver entropy to capture the variety and balance of the identified MWE types. For each language, a distinct set of values is provided. In this study, we focus on the macro-averaged F1-score as our primary evaluation metric.

3.3 Baselines

We compare two PMI-based baselines that differ exclusively in how word co-occurrence is defined, while sharing the same scoring function, preprocessing pipeline, and thresholding strategy. This design allows us to isolate the effect of syntactic information on MWE identification, independently of other modeling choices.

4 Results and Analysis

The following are two tables Table 1 and Table 2 with F1 scores for every language for different thresholds. The first table contains scores for PMI approach enchanced with UD, and the second table presents scores for classical PMI.

Lang	Type	50	75	90	95	99
UK	MWE-based	0.0062	0.0046	0.0014	0.0003	0.0000
	Token-based	0.1173	0.0915	0.0500	0.0302	0.0055
EGY	MWE-based	0.0000	0.0000	0.0000	0.0000	0.0000
	Token-based	0.1006	0.0533	0.0155	0.0000	0.0000
EL	MWE-based	0.0089	0.0043	0.0022	0.0000	0.0000
	Token-based	0.0803	0.0594	0.0364	0.0302	0.0097
FA	MWE-based	0.0511	0.0446	0.0387	0.0288	0.0053
	Token-based	0.2062	0.1397	0.0771	0.0427	0.0041
FR	MWE-based	0.0108	0.0077	0.0033	0.0032	0.0013
	Token-based	0.1970	0.1231	0.0519	0.0257	0.0011
HE	MWE-based	0.0036	0.0022	0.0012	0.0004	0.0000
	Token-based	0.0916	0.0594	0.0330	0.0200	0.0051
JA	MWE-based	0.0714	0.0598	0.0387	0.0277	0.0081
	Token-based	0.2682	0.1844	0.0943	0.0534	0.0077
LV	MWE-based	0.0034	0.0021	0.0013	0.0011	0.0000
	Token-based	0.0394	0.0280	0.0199	0.0164	0.0056
NL	MWE-based	0.0730	0.0773	0.0000	0.0000	0.0000
	Token-based	0.2588	0.1939	0.0319	0.0089	0.0000
PL	MWE-based	0.0025	0.0017	0.0007	0.0003	0.0004
	Token-based	0.0879	0.0640	0.0415	0.0355	0.0288
PT	MWE-based	0.0033	0.0000	0.0000	0.0000	0.0000
	Token-based	0.0945	0.0693	0.0432	0.0107	0.0000
RO	MWE-based	0.0029	0.0020	0.0014	0.0009	0.0000
	Token-based	0.0885	0.0460	0.0134	0.0047	0.0014
SV	MWE-based	0.0015	0.0006	0.0000	0.0000	0.0000
	Token-based	0.0740	0.0481	0.0200	0.0091	0.0014
SR	MWE-based	0.0297	0.0160	0.0075	0.0034	0.0000
	Token-based	0.1781	0.1209	0.0633	0.0328	0.0046
KA	MWE-based	0.0443	0.0480	0.0265	0.0138	0.0027
	Token-based	0.1417	0.0960	0.0371	0.0161	0.0023

Table 1: F1 scores of the UD-based PMI approach across different thresholds for MWE-based and token-based evaluation.

From Tables 1 and 2 we suggest 3 observations:

- Scores for full MWE matches are extremely low for both approaches
- MWE-based results are slightly, but insignificantly better in UD approach
- Token-based results are better with lower threshold percentiles in classical PMI approach.

From observations above we can interfere the following:

- This suggests that while UD features may provide some benefit in helping capture MWEs, they do not dramatically improve full match accuracy. This could imply that syntactic information alone might not be sufficient to solve the full match problem and other factors might need to be considered.
- Stronger performance of classical PMI approach on lower thresholds suggests that raw statistics favour word pairs that overlap with MWEs, even if they do not recover the full expression.

As for the time consumed, on average for different languages UD approach takes 7% time more on the same machine. We expected UD approach to have drastically higher timings, and we are pleased to find that the increase in computing time is insignificant. Still, the results in UD approach did not pay off the time consumed.

5 Related Work

The identification of multi-word expressions (MWEs) has evolved from purely statistical association measures to complex neural architectures.

Statistical and Association Measures Traditional unsupervised methods rely on association measures (AMs) such as PMI or log-likelihood (Gries, 2018) to quantify lexical affinity. Evert (2005) provides a comprehensive foundational framework for these measures, noting their effectiveness for adjacent co-occurrences. However, Pecina (2008) demonstrated that no single AM is universal, suggesting that ranking performance varies significantly depending on the MWE category and language. While these methods are interpretable and cross-linguistic, they often fail to capture rare or syntactically flexible MWEs—a limitation we aim to address using syntactic graphs.

Syntax-aware Identification The transition from surface-based windows to syntactic adjacency was extensively explored by Seretan (2011), who argued that MWEs are primarily syntactic units and should be extracted from parsed corpora. By leveraging the Universal Dependencies (UD) framework (Nivre et al., 2016), researchers have sought to create cross-linguistically consistent identification pipelines. This is particularly relevant for shared tasks like PARSEME (Savary et al., 2023; Ramisch et al., 2018), where verbal MWEs often exhibit long-distance dependencies and interleaving components. To ensure high-quality syntactic input, modern pipelines often rely on robust parsers such as UDPipe 2.0 (Straka, 2018).

Neural and Hybrid Methods Contemporary research heavily utilizes Transformer-based models and contextual embeddings, such as Multilingual

Lang	Type	50	75	90	95	99
UK	MWE-based	0.0103	0.0057	0.0041	0.0035	0.0033
	Token-based	0.1279	0.1342	0.1376	0.1392	0.1406
EGY	MWE-based	0.0018	0.0000	0.0000	0.0000	0.0000
	Token-based	0.0986	0.0915	0.0911	0.0915	0.0919
EL	MWE-based	0.0049	0.0045	0.0027	0.0017	0.0011
	Token-based	0.0852	0.0903	0.0917	0.0915	0.0922
FA	MWE-based	0.0519	0.0478	0.0471	0.0470	0.0467
	Token-based	0.2571	0.2465	0.2433	0.2427	0.2419
FR	MWE-based	0.0188	0.0160	0.0151	0.0148	0.0146
	Token-based	0.2213	0.2207	0.2225	0.2229	0.2232
HE	MWE-based	0.0047	0.0029	0.0025	0.0024	0.0024
	Token-based	0.1196	0.1206	0.1217	0.1220	0.1223
JA	MWE-based	0.0287	0.0287	0.0306	0.0314	0.0313
	Token-based	0.2008	0.2017	0.2013	0.2007	0.1976
LV	MWE-based	0.0056	0.0022	0.0012	0.0012	0.0011
	Token-based	0.0627	0.0643	0.0696	0.0717	0.0731
NL	MWE-based	0.0457	0.0539	0.0435	0.0437	0.0506
	Token-based	0.2460	0.2624	0.2525	0.2566	0.2610
PL	MWE-based	0.0045	0.0033	0.0031	0.0031	0.0031
	Token-based	0.1200	0.1247	0.1277	0.1291	0.1305
PT	MWE-based	0.0154	0.0092	0.0083	0.0083	0.0083
	Token-based	0.0909	0.0876	0.0885	0.0890	0.0892
RO	MWE-based	0.0051	0.0031	0.0029	0.0028	0.0028
	Token-based	0.1490	0.1485	0.1490	0.1490	0.1490
SV	MWE-based	0.0269	0.0166	0.0079	0.0054	0.0036
	Token-based	0.1830	0.1806	0.1747	0.1724	0.1710
SR	MWE-based	0.0371	0.0401	0.0383	0.0374	0.0372
	Token-based	0.1534	0.1613	0.1656	0.1681	0.1711
KA	MWE-based	0.0011	0.0008	0.0007	0.0007	0.0007
	Token-based	0.0151	0.0128	0.0127	0.0128	0.0129

Table 2: F1 scores of the classical PMI approach (without UD) across different thresholds for MWE-based and token-based evaluation.

BERT (Avram et al., 2023), to capture semantic and contextual nuances. While these neural methods achieve high performance, Taslimipoor and Rohanian (2019) noted that they come with significant computational costs and a lack of interpretability compared to traditional statistical approaches.

Rule and Lexicon-Based Approaches Other paradigms include rule-based approaches (Cordeiro et al., 2016), which provide high precision for well-defined patterns but suffer from low recall and require intensive manual effort. Similarly, lexicon-based approaches (Mititelu et al., 2024) ensure reliable detection of known expressions by checking against idiom dictionaries, though their coverage is naturally limited by the dictionaries in use. Our work seeks a middle ground, maintaining the lightweight nature of PMI while introducing the linguistic awareness provided by UD trees.

6 Limitations

The approach relies heavily on the accuracy of dependency parsing. Parsing errors directly affect PMI estimates.

7 Conclusion

We presented a simple syntax-aware PMI method for MWE identification that leverages Universal Dependency trees. By redefining co-occurrence in terms of syntactic relations, our approach is meant to capture non-adjacent and syntactically flexible MWEs that surface-based methods miss. Despite the idea, in practice no significant difference was found.

In the future this work could be complemented with mixture of other approaches, for example rating PMI score higher if MWE follows a rule from rule-based approach or appears in a dictionary from lexicon-based approach

Acknowledgments

We would like to express our sincere gratitude to our supervisor, Prof. Çağrı Çöltekin, for their invaluable guidance and support throughout this research. We also extend our gratitude to the organizers of the PARSEME shared task for making the dataset available and for their efforts in coordinating the shared task.

References

Andrei Avram and 1 others. 2023. Multilingual BERT for multiword expression identification across 14 languages. *arXiv preprint arXiv:2306.10419*.

Ana Cordeiro and 1 others. 2016. Rule-based approaches for multiword expression identification. In

Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval), pages 1140–1145.

Stefan Evert. 2005. *The statistics of word cooccurrences: word pairs and collocations*. Stuttgart: University of Stuttgart.

Stefan Th. Gries. 2018. Multiword expressions: A corpus-driven approach. In *Proceedings of the Workshop on Multiword Expressions (MWE)*, pages 1–15.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released August 24, 2025.

Verginica Barbu Mititelu, Voula Giouli, Kilian Evang, Daniel Zeman, Petya Osenova, Carole Tiberius, Simon Krek, Stella Markantonatou, Ivelina Stoyanova, Ranka Stanković, and Christian Chiarcos. 2024. Multiword expressions between the corpus and the lexicon: Universality, idiosyncrasy, and the lexicon-corpus interface. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 147–153, Torino, Italia. ELRA and ICCL.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, and 1 others. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*.

Pavel Pecina. 2008. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop on Multiword Expressions*, pages 54–57.

Carlos Ramisch and 1 others. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Inflection (LaTeCH-CLFL-MWE)*, pages 222–240.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, and 9 others. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Violeta Seretan. 2011. *Syntax-based collocation extraction*. Springer Science & Business Media.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 shared task: Tagging in context, more syntactic features, and shared structures. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207.

Shiva Taslimipoor and Omid Rohanian. 2019. Investigating BERT for multilingual NLP tasks: The case of multiword expressions. In *Proceedings of the 15th Workshop on Multiword Expressions (MWE 2019)*, pages 157–163.

tiberiucarp at MWE-2026 AdMIRe 2: GLIMMER-Gloss-based Image Multiword Meaning Expression Ranker

Andrei Tiberiu Carp
Tomorrow University
ING Hubs Romania
tiberiucarp@gmail.com

Abstract

Multiword expressions (MWEs), particularly idioms, pose persistent challenges for vision-language systems due to their non-compositional semantics and culturally grounded meanings. This paper presents GLIMMER, a three-stage hybrid ranking system that evaluates how well images express the intended meaning of MWEs across 15 languages. Our approach uses LLM-generated semantic glosses as multilingual meaning anchors, combined with dual-path embedding scoring (textual captions and visual features), and LLM-based semantic verification. Evaluated on the ADMIRE shared task benchmark, GLIMMER achieves competitive performance across diverse languages without relying on parallel training data or language-specific resources. The results show that using glosses to anchor meaning helps match idioms with images across languages and modalities, and that combining retrieval with reasoning is more robust than using embeddings alone.

1 Introduction

Multiword expressions (MWEs), such as idioms, convey meanings that cannot be derived compositionally from their constituent words. Although large language models (LLMs) have shown improved handling of idiomaticity in text-only settings (Tayyar Madabushi et al., 2022; Tedeschi et al., 2022; Tian et al., 2023), multimodal understanding of idioms remains an open challenge. Vision-language models (VLMs) excel in literal and compositional grounding, but often fail when figurative meanings diverge from surface-level visual cues (Yuksekgonul et al., 2022; Akula et al., 2023).

This work, introduces GLIMMER (GLoss-based Image Multiword Meaning Expression Ranker), a hybrid system developed for the ADMIRE shared task (Arslan et al., 2026), which focuses on ranking images according to how well they express the intended (idiomatic or literal) meaning of an MWE in context. GLIMMER is designed around the observation that *idioms are meaning-level units*, and that explicit semantic representations can serve as stable anchors across modalities and languages. Our key contributions are:

1. **Gloss-based semantic anchoring** using LLM-generated contextual definitions as multilingual meaning pivots

2. **Hybrid retrieval-reasoning architecture** combining embedding-based similarity with LLM-based semantic verification

3. **Dual-modality scoring** using both image captions and raw visual features

GLIMMER was evaluated on the ADMIRE shared task benchmark (Torunoğlu-Selamet et al., 2026) and performs competitively in 15 typologically diverse languages without requiring parallel data or task-specific training, highlighting the effectiveness of gloss-centered multimodal reasoning.

The paper is structured as follows: Section 2 reviews related work; Section 3 describes the methodology; Section 4 presents the experimental setup; Section 5 reports the results and analysis; and Sections 6 and 7 discuss the limitations and conclusions, respectively.

2 Related Work

The detection of multilingual idiomaticity has emerged as a key challenge in NLP (Tayyar Madabushi et al., 2022; Tedeschi et al., 2022). While Transformer models encode idiomatic meanings differently from literal phrases (Tian et al., 2023), recent evaluations show that even LLMs struggle without explicit semantic cues (Phelps et al., 2024). In the visual domain, early vision–language

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 170–176

models (Li et al., 2023a; Huang et al., 2023) excel at compositional tasks but often fail on figurative grounding (Yuksekgonul et al., 2022; Akula et al., 2023; Saakyan et al., 2025). However, recent work demonstrates that textual explanations can act as semantic bridges for non-literal matching (Chakrabarty et al., 2023), motivating our gloss-based design.

Our architecture adapts retrieval-augmented generation (Borgeaud et al., 2022; Izacard et al., 2023) to the multimodal idiom domain. Hybrid pipelines combining dense retrieval with neural reasoning have proven effective for complex semantics (Ni et al., 2025; Mao et al., 2021). To enable zero-shot transfer, we leverage advances in multilingual sentence embeddings (Muennighoff et al., 2023; Li et al., 2023b; Duquenne, 2024). Finally, our approach aligns with findings in multimodal chain-of-thought reasoning (Achiam et al., 2023; Zhang et al., 2023), utilizing LLM-generated glosses as explicit semantic anchors to resolve ambiguity.

3 Methodology

3.1 Problem Formulation

Given a multiword expression e, context sentence s indicating usage type $t \in \{\text{idiomatic}, \text{literal}\}$, and a set of candidate images $\mathcal{I} = \{(I_1, c_1), \ldots, (I_n, c_n)\}$ where I_i is an image and c_i its caption, our goal is to rank images by how well they express the intended meaning of e in context s.

3.2 Three-Stage Pipeline

Stage 1: Gloss Generation We generate a contextual semantic gloss for each MWE using an instruction-tuned LLM (OpenAI GPT-5.1):

> *Given the expression "{e}" used in: "{s}"*
> *Is this idiomatic or literal usage?*
> *Provide a concise gloss explaining the meaning.*

The gloss g serves as a language-independent semantic anchor, cached for efficiency. For unlabeled test data, we infer usage type t via prompting. Despite potential generation noise, g provides a transparent intermediate representation that enhances downstream alignment.

Stage 2: Dual Embedding Scoring For each candidate (I_i, c_i), we compute two complementary similarity scores:

Text Path: Using multilingual sentence transformer (`paraphrase-multilingual-mpnet-base-v2`) (Reimers and Gurevych, 2020):

$$sim_{\text{text}}(i) = \cos(\text{embed}(c_i), \text{embed}(g)) \quad (1)$$

Vision Path: Using CLIP ViT-B-32 (Radford et al., 2021):

$$sim_{\text{clip}}(i) = \cos(\text{CLIP}_{\text{img}}(I_i), \text{CLIP}_{\text{text}}(g)) \quad (2)$$

Combined embedding score:

$$score_{\text{embed}}(i) = 0.6 \cdot sim_{\text{text}}(i) + 0.4 \cdot sim_{\text{clip}}(i) \quad (3)$$

Stage 3: LLM Semantic Verification Embedding-based similarity alone may conflate literal and idiomatic interpretations. We therefore use an LLM (OpenAI GPT-5.1) to perform fine-grained semantic verification, scoring whether a caption describes an image that expresses the gloss meaning:

> *Gloss: "{g}"*
> *Caption: "{c_i}"*
> *Does this caption match the gloss meaning?*
> *Rate 0-100.*

This yields $score_{\text{llm}}(i) \in [0, 100]$. Although this step relies on a proprietary LLM, the prompts are deterministic, and the gloss representations reusable, mitigating variability. We normalize and fuse the scores:

$$score_{\text{final}}(i) = 0.4 \cdot score_{\text{embed}}(i) + 0.6 \cdot \frac{score_{\text{llm}}(i)}{100} \quad (4)$$

The final ranking orders the images by descending $score_{\text{final}}$.

Figure 1 presents how a semantic gloss is generated to anchor the meaning (Stage 1), followed by dual-path embedding scoring (Stage 2) and fine-grained LLM verification (Stage 3).

3.3 Design Rationale

Why glosses? Glosses externalize meaning to enable zero-shot cross-lingual transfer without parallel data. They provide explicit semantic context for evaluation, serving as anchors for non-literal matching as validated in recent visual metaphor research (Chakrabarty et al., 2023).

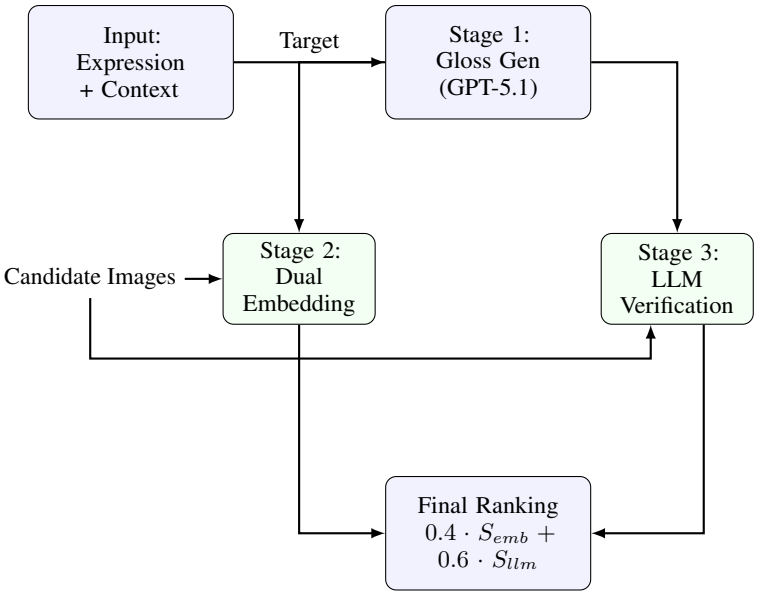


Figure 1: The GLIMMER architecture.

Why hybrid scoring? Fusion combines efficient embedding-based retrieval with precise LLM verification to balance scalability and reasoning depth (Ni et al., 2025). This approach mitigates the tendency of embeddings to conflate literal and idiomatic meanings while avoiding the computational costs of pure LLM scoring.

Why dual modality? Combining captions and images improves ranking robustness. This approach compensates for captions that omit visual cues and raw images that lack linguistic grounding, yielding more reliable retrieval.

Weight tuning We set Text/CLIP weights (60/40) to prioritize captions for abstract semantic clarity. Conversely, Embed/LLM weights (40/60) favor LLM verification to leverage nuanced reasoning for figurative language. These parameters were empirically validated via grid search on development data.

4 Experimental Setup

4.1 Dataset

We evaluate on the ADMIRE shared task dataset (Pickard et al., 2025), covering 15 languages: Chinese (ZH), Georgian (KA), Greek (EL), Igbo (IG), Kazakh (KK), Norwegian (NO), Portuguese-Brazil (PT-BR), Portuguese-Portugal (PT-PT), Russian (RU), Serbian (SR), Slovak (SK), Slovenian (SL), Spanish-Ecuador (ES-EC), Turkish (TR), and Uzbek (UZ).

For each language, the dataset provides:

- Multiword expressions with context sentences
- Sets of 5 candidate images per expression
- Image captions in the target language
- Usage type labels (idiomatic/literal) for training only

The test set contains expressions without usage type labels, requiring automatic inference. The evaluation metrics used are: Accuracy (Acc), Spearman Correlation, and Normalized Discounted Cumulative Gain (nDCG).

Metrics are computed per usage type (idiomatic/literal) and aggregated across languages.

4.2 Implementation

Our system is implemented in Python using SentenceTransformers and OpenAI libraries, with the following configuration:

- **LLM**: OpenAI GPT-5.1 via Responses API (temperature 0.7 for glosses, 0.0 for scoring)
- **Sentence Encoder**: paraphrase-multilingual-mpnet-base-v2
- **CLIP**: OpenAI ViT-B-32 with default preprocessing
- **Weight Parameters**: Text/CLIP $\alpha = 0.6$, Embedding/LLM $\alpha = 0.4$
- **Gloss Caching**: Enabled to reduce API calls (same expression $\rightarrow$ same gloss)

- **Text Normalization**: Language-specific handling (e.g., Uzbek apostrophe normalization)

The code is available at https://github.com/harapalb66/GLimmer.

5 Results and Analysis

5.1 Results

Table 1 shows aggregate results across all 15 languages. GLIMMER achieves 50.2% overall accuracy with strong ranking quality (nDCG: 0.804), placing fourth in the ADMIRE shared task competition (Pickard et al., 2025). Performance is higher for literal expressions (54.4%) than for idiomatic ones (46.3%), which is expected given that literal meanings are more directly grounded in visual evidence.

System	Acc ↑	ρ ↑	nDCG ↑
Shared Task Winner			
ITUNLP	**0.600**	—	**0.850**
GLIMMER (Our System)			
Overall	0.502	0.191	0.804
Idiomatic	0.463	0.187	0.778
Literal	0.544	0.194	0.835

Table 1: Aggregate results across 15 languages compared to the shared task winner.

Table 2 shows per-language performance. Portuguese-Brazil (66.7%), Russian (62.9%), and Slovenian (58.8%) achieve the highest accuracy, while Spanish-Ecuador (33.3%) and Igbo (37.4%) are most challenging.

Language	Acc	ρ	nDCG
Chinese	0.436	0.131	0.769
Georgian	0.496	0.135	0.791
Greek	0.543	0.310	0.831
Igbo	0.374	0.032	0.740
Kazakh	0.506	0.292	0.815
Norwegian	0.510	0.161	0.804
PT-BR	**0.667**	0.261	0.876
PT-PT	0.545	0.197	0.828
Russian	0.629	0.309	0.851
Serbian	0.479	0.161	0.784
Slovak	0.510	0.216	0.815
Slovenian	0.588	0.227	0.839
ES-EC	0.333	0.029	0.745
Turkish	0.484	0.118	0.800
Uzbek	0.425	0.289	0.774

Table 2: Per-language overall results. Best accuracy in bold.

5.2 Ablation Study

To assess the contribution of each component, we evaluate three variants on development data:

1. **Embed-only**: $score = score_{\text{embed}}$ (no LLM verification)
2. **Text-only**: $score_{\text{embed}} = sim_{\text{text}}$ (no CLIP)
3. **LLM-only**: Direct image-expression matching without gloss (no embeddings)

The following trends are observed across languages:

- CLIP vision path improves performance on visually distinctive cases
- LLM verification corrects embedding errors in subtle semantic distinctions
- Gloss-based grounding outperforms direct matching

Hyperparameter Tuning To determine the optimal fusion weights, we evaluated multiple configurations on development data. We found that assigning higher importance to textual captions ($\alpha = 0.6$) over visual features provided better semantic discrimination for abstract concepts. Similarly, prioritizing the LLM verification score ($\beta = 0.6$) over embedding similarity yielded the highest correlation across languages, as the reasoning capabilities of the LLM were crucial for correcting misalignments where embeddings conflated literal and idiomatic meanings.

5.3 Cross-Lingual Performance

We analyze performance across language families:

- **Romance** (PT-BR, PT-PT, ES-EC): Wide variance (33.3%-66.7%). Portuguese variants excel while Spanish-Ecuador struggles, possibly due to regional expression variations.
- **Slavic** (RU, SR, SK, SL): Strong overall (47.9%-62.9%), with Russian achieving the second-best accuracy. High Spearman correlations suggest good ranking quality.
- **Turkic** (KK, TR, UZ): Mixed results (42.5%-50.6%). Despite lower accuracy, Kazakh and Uzbek show surprisingly high correlations (0.289-0.292), indicating good relative ranking.

- **Other** (ZH, KA, EL, IG, NO): Greek performs exceptionally well (ρ=0.310, highest correlation), while Igbo is most challenging (37.4%), likely due to sparse representation in LLM pretraining corpora and culturally specific expressions with limited web image coverage.

The high nDCG scores (>0.74 across all languages), as computed via the official Codabench evaluation, demonstrate that GLIMMER produces reasonable rankings even when top-1 accuracy is modest, a valuable property for retrieval applications.

5.4 Error analysis

We examine a representative failure involving the Chinese compound 黑箱 ("black box"):

> 缺乏严格的程序性审查，导致很多加分通过黑箱操作等不正当的手段来获取。
>
> *"Lack of procedural review leads to bonus points obtained through black box operations and other **improper means**."*

Despite explicit corruption markers (不正当的手段, "improper means"), our system ranked images as shown in Table 3.

Table 3: Ranking failure for idiomatic

Image	System	Expected
Circuit cube	1st	–
Businessmen	5th	1st

Root cause: The LLM-generated gloss ("opaque operations") captured abstract semantics but missed pragmatic entailments: *human agency*, *institutional corruption*, *illicit gain*. This caused embeddings to prefer visually complex objects (circuits) over contextually appropriate scenes (businesspeople).

The *businessmen* image shows no lexical overlap with "opaque operations," yielding low text similarity and low CLIP similarity. Although LLM scoring (60% weight) recognized contextual fit, embedding scores (40%) had already created an insurmountable gap.

6 Limitations and Broader Impact

6.1 Limitations

Our approach depends on LLM-based gloss generation and verification. While gloss caching improves efficiency, future work will explore distilling gloss generation into smaller or open models. Additionally, gloss quality may vary for culturally specific idioms, and errors at this stage can propagate through the pipeline. Incorporating explicit uncertainty quantification mechanisms could improve system transparency and reliability (Ni et al., 2025). The dual-modality architecture requires both captions and images, limiting applicability to caption-free scenarios.

6.2 Broader Impact

Improved idiom grounding benefits multilingual retrieval, education, and cross-cultural communication tools. However, biases present in web imagery or LLM training data may influence rankings, particularly for under-resourced languages. Our system reflects patterns learned from existing data, which may not capture the full diversity of idiomatic usage across cultures.

7 Conclusion

We presented GLIMMER, a hybrid system for ranking images by multiword expression fit across 15 languages. Our gloss-based architecture provides a stable semantic anchor enabling cross-lingual transfer, while hybrid retrieval-reasoning scoring balances efficiency and semantic precision. Key findings indicate that i) gloss-based representations enable multilingual transfer without parallel data, ii) hybrid retrieval-reasoning architectures outperform embedding-only approaches, and iii) the integration of textual and visual modalities improves robustness to caption quality.

GLIMMER achieves 50.2% overall accuracy with strong ranking quality (nDCG: 0.804) across 15 typologically diverse languages, demonstrating that explicit semantic anchoring is an effective strategy for multimodal idiom understanding.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arjun R Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T Freeman, and 1 others. 2023. Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23201–23211.

Doğukan Arslan, Rodrigo Wilkens, Wei He, Dilara Torunoğlu-Selamet, Thomas Pickard, Aline Villavicencio, Adriana S. Pagano, and Gülşen Eryiğit. 2026. MWE-2026 Shared Task 2: AdMIRe 2 - Advancing Multimodal Idiomaticity Representation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morocco. Association for Computational Linguistics.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and 1 others. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. *arXiv preprint arXiv:2305.14724*.

Paul-Ambroise Duquenne. 2024. *Sentence Embeddings for Massively Multilingual Speech and Text Processing*. Ph.D. thesis, Sorbonne Université.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, and 1 others. 2023. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Ziheng Li, Shaohan Huang, Zihan Zhang, Zhi-Hong Deng, Qiang Lou, Haizhen Huang, Jian Jiao, Furu Wei, Weiwei Deng, and Qi Zhang. 2023b. Dual-alignment pre-training for cross-lingual sentence embedding. *arXiv preprint arXiv:2305.09148*.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4089–4100.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, and 1 others. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111.

Bo Ni, Zheyuan Liu, Leyao Wang, Yongjia Lei, Yuying Zhao, Xueqi Cheng, Qingkai Zeng, Luna Dong, Yinglong Xia, Krishnaram Kenthapadi, and 1 others. 2025. Towards trustworthy retrieval augmented generation for large language models: A survey. *arXiv preprint arXiv:2502.06872*.

Dylan Phelps, Thomas Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. *arXiv preprint arXiv:2405.09279*.

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, and Marco Idiart. 2025. Semeval-2025 task 1: Admire–advancing multimodal idiomaticity representation. *arXiv preprint arXiv:2503.15358*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

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.

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. *arXiv e-prints*, pages arXiv–2204.

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.

Ye Tian, Isobel James, and Hye Son. 2023. How are idioms processed inside transformer language models? In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)*, pages 174–179.

Dilara Torunoğlu-Selamet, Dogukan Arslan, Rodrigo Wilkens, Wei He, Doruk Eryiğit, Thomas Pickard, Adriana S. Pagano, Aline Villavicencio, Gülşen Eryiğit, Ágnes Abuczki, Aida Cardoso, Alesia Lazarenka, Dina Almassova, Amalia Mendes, Anna Kanellopoulou, Antoni Brosa-Rodríguez, Baiba Saulite, Beata Wojtowicz, Bolette Pedersen, and 59 others. 2026. A parallel cross-lingual benchmark for multimodal idiomaticity understanding. *Preprint*, arXiv:2601.08645.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

IPN at MWE-2026 PARSEME 2.0 Subtask 1: MWE Identification via Related Languages and Harnessing Thinking Mode

Anna Hülsing, Noah-Manuel Michael, Daniel Mora Melanchthon, Andrea Horbach
Kiel University, Germany
Leibniz Institute for Science and Mathematics Education, Kiel, Germany
huelsing@ipn.uni-kiel.de

Abstract

We present IPN, our system for Subtask 1 of the PARSEME 2.0 Shared Task, which targets the identification of MWEs in 17 languages. Overall, IPN outperformed a much larger-parameter baseline model, yet a performance gap to the top-performing systems remains. To better understand these results, we investigate QWEN3-32B's suitability for mono-, cross- and multilingual MWE identification. We also explore whether this model benefits from prepending automatically generated thinking data to the gold label during instruction-tuning. We find that target language data is vital for instruction-tuning. Prepending generated thinking data to a subset of the training data slightly improves performance for two out of three languages, but more detailed evaluation is required.

1 Introduction

This paper describes IPN, our system for Subtask 1 in the PARSEME 2.0 Shared Task (Scholivet et al., 2026), focusing on the identification of multiword expressions (MWEs) in 17 languages. An MWE is a combination of at least two lexemes that exhibits varying degrees of idiosyncrasy at least one linguistic level. Representative examples are *to pay a visit* (verbal MWE), *larger than life* (adjectival/adverbial), *hot dog* (nominal), or *every other* (as in *every other day*, functional).

In our system, we build upon recent developments in MWE detection by Ide et al. (2025) and use an open-weight model from the QWEN3 family (Qwen Team, 2025). We aim to maximize lexical diversity by sampling as many different MWE types as possible, and try to automatically enhance our training data to support the model's thinking process. For instruction-tuning, we use training data from multiple languages.

In this report, we describe our data preprocessing in Section 3 and our model in Section 4. We evaluate our experimental choices in Section 5 by addressing (1) the influence of the usage of synthetic thinking data for instruction-tuning and (2) the influence of the language(s) used for instruction-tuning on model performance. We do this by comparing four training data conditions: target language data, data only from languages related to the target language, data only from unrelated languages, and target data plus data from related languages.[1]

2 Background & Related Work

Instruction-Tuning. Savary et al. (2019) observe that unseen MWEs constitute the main source of errors and thus argue that MWE identification can be enhanced by large-coverage MWE lexicons. In a similar vein, Überrück-Fries et al. (2024) train a rule-based system that exploits MWEs extracted from a large online lexicon, and Tanner and Hoffman (2023) leverage encodings of WordNet sense definitions (Miller, 1994) in a BERT-based word-sense-disambiguation approach adapted to English MWE identification. However, Ide et al. (2025) outperform the system by Tanner and Hoffman (2023) by instruction-tuning QWEN2 on as little as 780 English training sentences. We examine whether QWEN is also effective for non-English languages.

Relatedness. Phylogenetic relatedness and geographical or cultural proximity are known to promote the sharing of MWEs across languages (Gluski, 1971; Strauss, 1994; Colson, 2008; Perepadia and Malakhova, 2023), although many expressions transcend these boundaries (Piirainen, 2012). This suggests potential for cross-lingual transfer. Swaminathan and Cook (2023) fine-tune BERT-based models on English and Portuguese and evaluate on Galician data from SemEval-2022 (Tayyar Madabushi et al., 2022). They find that fine-tuning on English does not surpass a majority-class

[1]The source code is available at https://github.com/AnHu2410/PARSEME_2_ipn.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 177–186

baseline, whereas fine-tuning on Portuguese – an Ibero-Romance language like Galician – does. Motivated by these findings, we examine how the relatedness of instruction-tuning data to the target language affects QWEN3's performance.

Thinking. Wang et al. (2025b) elicit critique from a teacher model for math problem solving and use this critique to fine-tune a student model, thereby improving its mathematical reasoning beyond standard SFT. Zhang et al. (2025) elicit explanations of why prompts are adversarial or benign and use these explanations to fine-tune a second model, achieving better adversarial-prompt classification. Similarly, we generate rule-based "thinking" data and use them to enhance MWE identification via instruction-tuning.

3 Data

To perform instruction-tuning efficiently while preserving as much coverage as possible, we apply the following filtering steps to reduce the size of the PARSEME 2.0 Shared Task training data: First, we collect all MWE types as sets of their lemmas. Then, for each MWE type we select one training sentence that contains it, while allowing a single sentence to serve as the representative example for multiple MWE types. This way, we sample all MWE types that occur in the training data at least once. Some occur more than once, if they happen to appear in sentences that were actually sampled for a different MWE. Additionally, we add up to 20 sentences that do not contain an MWE. The type-token ratio in the training datasets as well as the number of sentences that resulted from our filtering step is shown in Table 1. As for some languages there are substantially fewer sentences than for others, we decided to use the entire training data for languages where the filtering resulted in less than 400 sentences (EGY, EL, NL, PT).

Language	Tokens	Types	Sentences
EGY	117	66	79 (431)
EL	670	417	355 (1380)
FA	4415	1939	1010
FR	4604	2042	1230
HE	12286	4327	3480
JA	2664	1626	854
KA	1517	432	433
LV	2436	770	698
NL	95	88	75 (90)
PL	11946	3219	2915
PT	167	134	128 (421)
RO	57609	3263	2624
SL	5481	2027	1730
SR	8094	3457	2621
SV	2876	1118	957
UK	4929	2428	2021

Table 1: Number of MWE tokens, MWE types and number of sentences after filtering PARSEME 2.0 training data per language. For very small datasets, we used the entire training data (numbers of sentences in brackets). Included in *Sentences* are the sentences we added that did not contain MWEs.

4 Model

Ide et al. (2025) have demonstrated the effectiveness of instruction-tuning QWEN-2.5-72B-INSTRUCT for English MWE identification. We build on and extend their work by resorting to the more recent QWEN3 family, which covers 119 languages and dialects (Qwen Team, 2025; it is not mentioned which ones exactly). More precisely, we select QWEN3-32B, because in preliminary experiments this dense variant performed best under our computational constraints (one NVIDIA H100 with 18432 CUDA cores, 80GB GPU memory and 2.0TB/s memory bandwidth).

4.1 Model Input

Model input consists of the prompt in combination with one of the sentences from the filtered training data (see Table 1). The prompt is adapted from Ide et al. (2025), and consists of the following building blocks: a definition of what MWEs are, a congruency pointer instructing the model to make use of congruent MWEs (i.e. MWEs whose lexemes are one-to-one literal translations across two languages), and a format pointer. Each one of these blocks includes a reference example to guide the model (see complete prompt in Appendix A.1). The definition block differs slightly from the prompt from Ide et al. (2025). In particular, we removed the sentence "In other words, a semantically idiomatic [MWE] takes on a meaning that is unique to that combination of words", because we consider uniqueness of meaning to be a general property of lexical items and regard strict synonymy as rare. As the original prompt from Ide et al. (2025) was intended for MWE identification in English, we added the congruency pointer to exploit the potential of cross-lingually congruent MWEs. Finally, we opted for a lexeme-index format, e.g. "to and fro; 7,8,9 | break up; 12, 13". We deemed this to be more token-efficient than the format employed by Ide et al. (2025). The prompt-sentence combination was used as input for both instruction-tuning

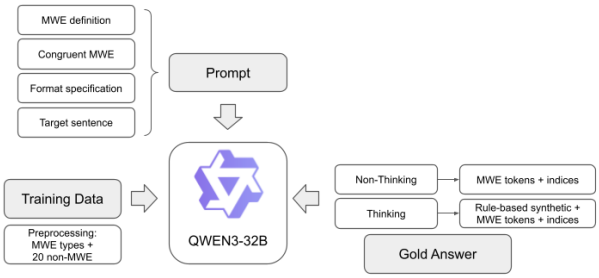


Figure 1: Overview over instruction-tuning procedure.

and inference.

4.2 Gold Thinking Content and Answer

Since QWEN3-32B supports thinking mode, we sought to exploit – or at least preserve – it's thinking capabilities by providing thinking content in the gold data used for instruction-tuning. To avoid the very time-consuming manual creation of thinking data, we employed a rule-based procedure to generate gold thinking content for each sentence in the filtered training data. We based this procedure on the thinking data that the model produced correctly without instruction-tuning (see Table 6 in Appendix A.2). Here, we identified the following Elements: 1) parsing of the sentence, 2) repetition of the MWE definition provided in the prompt, 3) discussion of potential MWE expressions, 4) repetition of the indices, and 5) declaration of the final answer.

Then, we mimicked the thinking of the non-instruction-tuned model for any given sentence in the training data: As a first step, we repeated the words and their indices from the gold data. In doing so, we combined the parsing of the sentence and the repetition of the indices (see Element 1 and 4). Secondly, we repeated the MWE definition from the prompt (see Element 2). As a third step, we discussed each MWE. To do this, we used a dictionary that we had constructed from all MWE categories occurring in the training data across all languages. For each category (e.g. AdvID), the dictionary stores a short paraphrase derived from the broad definitions in the PARSEME guidelines[2]. Examples of these paraphrases are given in Appendix A.5, Table 8. For each sentence, we retrieved the MWE category from the gold standard and, instead of providing a free-form explanation of why the expression is an MWE, we concatenated the lemmas of the MWE with the stored paraphrase of its category. After discussing all MWEs in the sentence we added the declaration of the final answer (Element 5). The complete thinking content was prepended to the gold answer, using QWEN3's formatting with thinking tags (<*think*> and </*think*>). For an example, see Appendix A.2.

This approach shows four main differences compared to the original thinking that would likely be avoided under manual annotation but arise in our automatic generation setup: Firstly, instead of mentioning which specific lexemes are e.g. non-compositional, the generated thinking mentions parts of speech such as "the verb is used non-compositionally". Secondly, our thinking does not explicitly refer to congruent MWEs; instead it gives examples of English MWEs of the same category from the PARSEME guidelines. These two differences could be optimized in a more elaborate setup. Thirdly, we do not discuss why certain words are *not* an MWE. As recall is often a problem with MWE identification (Ide et al., 2025), we deemed this acceptable. Fourthly, we did not include thinking tokens such as "Wait" as well as thought steps that are not needed for reaching the correct conclusion, as it was shown that these do not improve performance (Wang et al., 2025a; Li et al., 2025).

4.3 Instruction-tuning and Inference

Instruction tuning is a form of fine-tuning where, instead of optimizing on input–output pairs, one optimizes on instruction–input–output triples, so the model learns to respond to instructions rather than just map inputs to outputs. Figure 1 summarizes this triple – instruction (prompt), input (training data), and output (gold answer) – as used in our experiments and as described in the preceding sections. We instruction-tuned QWEN3-32B using LoRA adapters under 4-bit quantization. With the Unsloth library (Daniel Han and Unsloth Team, 2023), processing 100 sentences took approximately 3 minutes for the non-thinking and 4 minutes for the thinking variant. For VRAM-efficient inference, we used an 8-bit quantized version of QWEN3-32B. Inference for 100 sentences took approximately 30 minutes for the thinking variant and 3 minutes for the non-thinking variant. Inference with the non-instruction-tuned model took about 40 minutes for 100 sentences. Hyperparameters are shown in Appendix A.3.

[2]Version 2.0: https://PARSEMEfr.lis-lab.fr/PARSEME-st-guidelines/2.0/ , accessed 20.11.2025

4.4 Postprocessing Model Output

During inference, the model mostly followed the format that we asked for in the prompt. However, in some cases, it made slight changes, such as adding newlines to its prediction. Also, in rare cases, the thinking exceeded our token limit, and thus no prediction was generated. Therefore, we deleted noise such as line breaks and replaced all unfinished thinking by None. As the indices were not always correct, we align each predicted item to the original CUPT file as follows: If the lexeme occurs only once, we directly tag this lexeme in the input cupt file. If it occurs multiple times, we select the index closest to the predicted one. The resulting indices are used to map the predicted MWEs to lexemes in the CUPT file.

5 Experiments and Results

For our PARSEME 2.0 submission, we used the thinking variant for all languages except those with the largest blind test sets (KA and LV), for which we employed the non-thinking variant to maintain feasible inference times. Moreover, for languages belonging to the same branch of a language family, we jointly trained a single model on all its languages (Germanic, Slavic, Romance); for all other languages we trained one model per language. In order to validate these methods, we performed two experiments: Firstly, we investigated whether instruction-tuning the model with automatically generated thinking content in addition to the gold answer boosts performance. Secondly, we explored how MWE identification performance on a target language varies when training on target-language data only, on same-branch data, on data from unrelated languages, and on target+branch data.

Influence of Generated Thinking: For gauging the influence of thinking content on our model's performance, we instruction-tune three models. For the first one, the gold answer consists only of the tokens annotated as being part of an MWE in the training data plus their indices (non-thinking variant). For the second one this gold answer was appended to automatically generated thinking content (thinking variant, see Section 4.2). For the third one, 20% of the gold answers were presented in the thinking and 80% in the non-thinking variant (20/80-variant), in an attempt to mitigate potential overfitting to the wording of the thinking content. As test languages, we selected Romanian, Slovene, and Swedish because they represent different branches of the Indo-European language family (Romance, Slavic, and Germanic branch). Again, we jointly instruction-tuned a single model for all languages of each branch.

Evaluation was carried out on the PARSEME 2.0 dev sets, where the dev set for Romanian was cut to 2,500 sentences to maintain feasible inference times. To account for variability in model outputs arising from hardware differences and from the use of a non-zero temperature (required for optimal thinking), we trained two models with different random seeds and report the mean and standard deviation (SD) across inference runs (see Table 2). For Romanian, the non-thinking variant shows the best performance, while for Slovene and Swedish the 20/80-variant shows the best performance. The second best performance differs across languages, suggesting that the relative effectiveness of thinking varies in a language-specific manner. A limited amount of generated thinking appears beneficial for this task, whereas fully thinking-augmented instruction-tuning does not outperform either the non-thinking or the 20/80 configuration. Since these experiments were conducted after the Shared Task, the official results could, in principle, be improved by using either the 20/80 or the non-thinking variant.

	thinking	20/80	non-thinking
RO	30.3 (± 0.9)	32.95 (± 0.85)	**35.8** (±0.3)
SL	23.0 (± 2.3)	**25.6** (± 1.0)	19.1 (±2.1)
SV	22.9 (± 1.0)	**24.7** (± 1.65)	24.4 (±0.5)

Table 2: Mean MWE-based F1 scores (± SD) for thinking, non-thinking and a mixed training regime per language (Romanian, Slovene, and Swedish).

Influence of Relatedness: We next examined how performance on a target language degrades when instruction-tuning is performed only on target-language data, which in our exemplary cases is Romanian and Slovene (*target*), on data from other languages in the respective language family branch (Romance, Slavic; *branch*), on data from unrelated languages (*unrelated*), and on data from the target language plus data from the same branch (*target+*), with evaluation always on the target language.[3] For the first three scenarios, we instruction-tuned on 1500 sentences that we randomly selected

[3]We did not consider the Germanic branch as PARSEME 2.0 only covers Swedish and Dutch, with Dutch including only 90 training and 10 dev sentences. This renders a meaningful comparison with other language families impossible. For all other targets, there was only a single language per branch.

	RO			SL		
train	**precision**	**recall**	**F1**	**precision**	**recall**	**F1**
target	20.6 (±0.3)	**41.2** (±0.7)	27.4 (±0.1)	12.7 (±1.9)	**34.5** (±6.3)	**18.6** (±2.9)
branch	13.1 (±0.5)	22.6 (±0.3)	16.6 (±0.5)	7.7 (±0.6)	26.4 (±1.3)	11.9 (±0.9)
unrelated	11.1 (±0.3)	20.4 (±0.6)	14.4 (±0.4)	8.0 (±0.1)	22.3 (±4.3)	11.7 (±0.5)
target+	24.5 (±1.6)	40.2 (±2.9)	**30.4** (±0.4)	12.6 (±0.1)	34.4 (±4.3)	18.4 (±1.2)
none	**32.6** (±0.0)	22.1 (±0.0)	26.3 (±0.0)	**15.8** (±0.0)	19.1 (±0.0)	17.3 (±0.0)

Table 3: Mean MWE-based precision, recall, and F1 (± standard deviation) for Romanian (RO) and Slovene (SL) under different instruction-tuning data conditions.

from our filtered datasets (see Table 1). For *target+*, we combined 1500 sentences from *target* and 1500 from *branch*. The exact numbers for each scenario are found in Appendix A.4. We perform these experiments in the non-thinking variant, as inference times are reduced and thinking does not seem to introduce an unmitigated benefit. Again, we train two models with different random seeds and report mean and SD, and evaluate on the PARSEME 2.0 dev set (we reduce the Romanian dev set to 2,500 sentences). In addition, we report the scores of the non-instruction-tuned model (*none*; we ran the same model twice, which resulted in the same scores). The results are shown in Table 3. Instruction-tuning on *target-* plus *branch*-language data yields the best F1 for Romanian and the second-highest F1 for Slovene; only 0.2 points below the *target* model. Given the relatively large standard deviations, the two Slovene results are likely not meaningfully different. These findings corroborate the method employed for the Shared Task. The worst performance is seen when instruction-tuning on unrelated languages, as expected. Instruction-tuning on non-target language data shows a worse performance than using the untuned model, which highlights the necessity of target language data for instruction-tuning.

5.1 Results in the Overall PARSEME 2.0 Context

In the official PARSEME 2.0 ranking, IPN ranked third overall with an MWE-based mean F1-score of 28.4 across all 17 languages tested in the blind scenario. However, this score is clearly below the scores obtained by the two top-performing systems (MTLB-STRUCT: F1=48.4; Sahara-Tokenizers: F1=57.3). Except for Egyptian, Ancient Greek, Georgian, and Latvian, IPN outperforms the GPT-OSS-120B baseline in MWE-based F1 for all languages. This suggests that, in medium- to high-resource settings, instruction-tuning reverses the performance gap between the 32B and 120B models for this task. The four underperforming languages are used by 0.1% or less of the websites as content language, which shows that even though the QWEN3 family covers 119 languages and dialects, low-resource languages still pose a problem.[4] Our model is marked by a high recall, which is often competitive with the top-performing models (e.g. in the mean global token-based setting, recall is less than one point below the top-performing system), but also by low precision. The high recall is induced by instruction-tuning: as Table 3 shows, recall consistently improves when performing instruction-tuning (the only exception being RO, where recall decreases when instruction-tuning on unrelated languages), while precision declines.

6 Conclusion

We draw two main conclusions from our experiments. First, prepending generated thinking to the gold answer appears helpful for two out of three languages, and only when applied to a subset of training instances, rather than uniformly to all examples. At the same time, the thinking variant requires roughly an order of magnitude more inference time than the non-thinking model, yielding a poor cost–benefit ratio. Future work should evaluate the thinking variant more deeply, e.g. by investigating different settings (80/20, 50/50), and by conducting ablation studies to test whether removing specific components of the thinking data improves performance. Second, only instruction-tuning on data containing target-language data (*target* or *target+*) reliably outperforms the untuned baseline, underscoring the importance of in-language supervision for MWE identification with QWEN3. For most languages, IPN outperformed a much larger-parameter baseline model, yet a substantial performance gap to the top-performing systems remains.

[4] https://w3techs.com/technologies/overview/content_language; date of access: February 2, 2026.

Limitations

With QWEN3 being an open-weight but not an open-source model, code and pretraining data cannot be accessed. This means that we do not know whether the data was obtained in an ethically responsible way. Also, potential biases in the model cannot easily be detected and remedied.

In both the Shared Task and our subsequent experiments, the system prompt was included in the user prompt rather than being provided in a dedicated system role; this was discovered only after the Shared Task, so our reported results may underestimate the performance of a correct system–user prompt configuration.

References

Jean-Pierre Colson. 2008. Cross-linguistic phraseological studies: An overview. In Sylviane Granger and Fanny Meunier, editors, *Phraseology: An Interdisciplinary Perspective*, pages 191–206. John Benjamins Publishing Company, Amsterdam / Philadelphia.

Michael Han Daniel Han and Unsloth Team. 2023. Unsloth.

Jerzy Gluski, editor. 1971. *Proverbs: A Comparative Book of English, French, German, Italian, Spanish and Russian Proverbs with a Latin Appendix*. Elsevier Pub. Co., Amsterdam and New York.

Yusuke Ide, Joshua Tanner, Adam Nohejl, Jacob Hoffman, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. 2025. CoAM: Corpus of all-type multiword expressions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27004–27021, Vienna, Austria. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhamaneshi, Shishir G Patil, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. Language models can easily learn to reason from demonstrations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15979–15997, Suzhou, China. Association for Computational Linguistics.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Daria Perepadia and Yuliia Malakhova. 2023. Chinese and Japanese phraseologisms: A comparative aspect. *Philological Treatises*, 15(2):120–137.

Elisabeth Piirainen. 2012. *Widespread Idioms in Europe and Beyond: Toward a Lexicon of Common Figurative Units*, volume 5 of *International Folkloristics*. Peter Lang, New York.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91, Florence, Italy. Association for Computational Linguistics.

Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Carp, and Vasile Pais. 2026. Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions.

Emanuel Strauss. 1994. *Dictionary of European Proverbs*. Routledge, London and New York.

Raghuraman Swaminathan and Paul Cook. 2023. Token-level identification of multiword expressions using pre-trained multilingual language models. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 1–6, Dubrovnik, Croatia. Association for Computational Linguistics.

Joshua Tanner and Jacob Hoffman. 2023. MWE as WSD: Solving multiword expression identification with word sense disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 181–193, Singapore. Association for Computational Linguistics.

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.

Till Überrück-Fries, Agata Savary, and Agnieszka Dryjańska. 2024. Sailing through multiword expression identification with Wiktionary and linguse: A case study of language learning. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 248–262, Rennes, France. LiU Electronic Press.

Chenlong Wang, Yuanning Feng, Dongping Chen, Zhaoyang Chu, Ranjay Krishna, and Tianyi Zhou. 2025a. Wait, we don't need to "wait"! removing

thinking tokens improves reasoning efficiency. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7459–7482, Suzhou, China. Association for Computational Linguistics.

Yubo Wang, Ping Nie, Kai Zou, Lijun Wu, and Wenhu Chen. 2025b. Unleashing the reasoning potential of LLMs by critique fine-tuning on one problem. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3017–3027, Suzhou, China. Association for Computational Linguistics.

Yuyou Zhang, Miao Li, William Han, Yihang Yao, Zhepeng Cen, and Ding Zhao. 2025. Safety is not only about refusal: Reasoning-enhanced fine-tuning for interpretable LLM safety. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18727–18746, Vienna, Austria. Association for Computational Linguistics.

A Appendix

A.1 Prompt

Table 4 shows the complete prompt used in our experiments. Appended to the prompt is an example sentence from the training data.

A.2 Example for Original and Generated Thinking

Table 5 shows a complete example of the gold answer used for instruction-tuning in our thinking-variant, which consists of the automatically generated thinking and the solution. For the non-thinking variant, only the solution was shown to the model during instruction-tuning, and for the 20/80-variant, 20% of all instances were presentend with and 80% without the automatically generated thinking content.

Table 6 presents the original thinking for the sentence from Tables 4 and 5 of the pretrained, but non-instruction-tuned model. The following elements are found in both the original and the generated thinking: 1) parsing of the sentence and/or repetition of the indices ◯ , 2) repetition of the MWE definition provided in the prompt △, 3) discussion of potential MWE expressions □, and 4) declaration of the final answer ⊞.

A.3 Hyperparameters

LoRA adapters were applied with rank r=16, α=32, and dropout set to 0. Inference was then carried out with the default hyperparameters (Qwen Team, 2025) for thinking mode (temperature: 0.6, top-p value: 0.95), and for non-thinking mode (temperature: 0.7, top-p value: 0.8, presence penalty:1.5), each with a top-k value of 20 and a maximum of 5000 tokens using VLLM (Kwon et al., 2023).

A.4 Relatedness: Numbers of Training Sentences

Instruction-tuning was conducted under four data conditions, with the first three of them comprising 1500 sentences and the last one comprising 3000 sentences. The sources and compositions of the training data are summarized in Table 7.

A.5 MWE Description Dictionary

Table 8 shows three examples of MWE categories, how they are described in the PARSEME annotation guidelines (`https://PARSEMEfr.lis-lab.fr/PARSEME-st-guidelines/2.0`, date of access: November 20th, 2025), and how they were transformed into paraphrases. The paraphrases were used as building blocks for the thinking content, more precisely for the discussion of potential MWE expressions (see □ in Table 5).

definition	You are a helpful system for identifying multiple-word expressions (MWEs). Identify all MWEs in the given sentence, and output their surface forms. Each sentence is a string of words delimited by". An MWE is defined as a sequence that satisfies the following three conditions. 1. It consists of multiple words that are always realized by the same lexemes. The individual lexemes cannot be replaced by synonyms without distorting the meaning of the expression as a whole or violating language conventions. 2. It displays semantic, lexical, or syntactic idiomaticity. Semantic idiomaticity occurs when the meaning of an expression cannot be explicitly derived from its components. Lexical idiomaticity occurs when one or more components of an expression are not used as stand-alone words in standard English. Syntactic idiomaticity occurs when the grammar of an expression cannot be derived directly from that of its components. For example, semantically idiomatic MWEs include 'break up', the lexically idiomatic include 'to and fro', and syntactically idiomatic MWEs include 'long time no see'. 3. It is not a multi-word named entity, i.e., a specific name of a person, facility, etc.
congruency pointer	Remember that you can identify congruent MWEs across different languages. For example, you can identify the Romanian MWE 'pur și simplu' because you know the English MWE 'pure and simple'. Similarly, the Portuguese MWE 'ter lugar' is easy to identify because of the English MWE 'take place'. And the French MWE 'feu de circulation' is easy to identify, because it is almost congruent to the English MWE 'traffic lights'.
format pointer	Respond by providing all tokens of the MWE, and their indices. If no MWE occurs, output 'None'. If there are multiple MWEs, separate them by |, for example 'to and fro; 7,8,9 | break up; 12, 13'. Sentence: Macron planeja implementar a reforma por meio de decreto presidencial, o que significa que vai saltar os legisladores.

Table 4: Prompt used in our experiments with example sentence from Portuguese training data. The English translation of the sentence is: "Macron plans to implement the reform by means of a presidential decree, which means he will bypass lawmakers."

thinking	<think> ◯ Okay, let's tackle this query. The user wants me to identify all multiple-word expressions (MWEs) in the given sentence. The sentence is split into words by newlines, so first I need to parse each line as a token. Let me list them out with their indices to keep track. The words are: 1 Macron 2 planeja 3 implementar 4 a 5 reforma 6 por 7 meio 8 de 9 decreto 10 presidencial 11 , 12 o 13 que 14 significa 15 que 16 vai 17 saltar 18 os 19 legisladores 20 . △ Now we apply the MWE criteria. The first condition is that the expression consists of multiple words that are always realized by the same lexemes. The individual lexemes cannot be replaced by synonyms without distorting the meaning of the expression as a whole or violating language conventions. The second condition is that it displays semantic, lexical, or syntactic idiomaticity. Semantic idiomaticity occurs when the meaning of an expression cannot be explicitly derived from its components. Lexical idiomaticity occurs when one or more components of an expression are not used as stand-alone words in standard English. Syntactic idiomaticity occurs when the grammar of an expression cannot be derived directly from that of its components. For example, semantically idiomatic MWEs include 'break up', the lexically idiomatic include 'to and fro', and syntactically idiomatic MWEs include 'long time no see'. The third condition is that it is not a multi-word named entity. □ Looking at the words, let's check for possible MWEs. por meio de is a multiword expression. More precisely, it is a conjunction idiom or ConjID. It is comparable to the English ConjID "in that" in the following sentence: "she was fortunate in that she had friends to help her". The words cannot be realized by synonymous lexemes without altering the meaning of the construction as a whole, which satisfies condition 1. The given expression fulfills condition 2, as the expression displays syntactic idiomaticity. Also, the meaning of the expression is non-compositional. The expression is a conjunction and therefore does not refer to a person or place, so it also meets condition 3. ⊞ After considering all possibilities, my final answer is: por meio de; 6,7,8 </think>
solution	por meio de; 6,7,8

Table 5: Example for automatically generated thinking data for the sentence shown and translated in Table 4, and the gold solution.

original thinking	<think> ◯ Okay, let's tackle this query. The user wants me to identify all MWEs in the given sentence. The sentence is in Portuguese, and the example mentions that MWEs can be identified across languages if they are congruent with English ones. First, I need to parse the sentence. The words are: Macron, planeja, implementar, a, reforma, por, meio, de, decreto, presidencial, , , o, que, significa, que, vai, saltar, os, legisladores, . Looking at each word, I need to check if any sequences meet the three conditions of MWEs. △ Let's go through the conditions again. Condition 1: The MWE must consist of multiple words that are always realized by the same lexemes. They can't be replaced by synonyms without changing the meaning or violating conventions. Condition 2: It has semantic, lexical, or syntactic idiomaticity. Semantic idioms can't be derived from their components. Lexical idioms have components not used as standalone words. Syntactic idioms have grammar not derived from components. Condition 3: Not a named entity. □ Now, checking the Portuguese words. Let's see if any phrases are known MWEs. For example, "por meio de" is a common phrase in Portuguese, meaning "by means of." In English, it's "by means of" or "through," so that's a congruent MWE. Let's check if "por meio de" is a fixed expression. Yes, it's a prepositional phrase that's used as a single unit. The individual words "por" and "meio" can be standalone, but together they form an MWE. However, does it have semantic idiomaticity? The meaning of "por meio de" is "by means of," which is the same as the literal translation. So maybe not semantic idiomaticity. But syntactically, maybe? Wait, the structure "por meio de" is a fixed prepositional phrase. In Portuguese, it's a common collocation, so it might be considered an MWE. However, the user's example for English includes "take place" as a syntactically idiomatic MWE. But in this case, "por meio de" is a standard prepositional phrase. Hmm. Another possible MWE is "saltar os legisladores." Wait, "saltar" can mean "to jump over" or "to skip." But "saltar os legisladores" might be a specific expression. However, I'm not sure if that's a standard MWE. Alternatively, "vai saltar os legisladores" – does that mean skipping the legislators? Not sure. Maybe not a standard MWE. What about "decreto presidencial"? That's a noun phrase meaning "presidential decree." Each word is standard, so probably not an MWE. Looking for other possibilities. "por meio de" is a common phrase, but I'm not sure if it's an MWE here. The user's example for "take place" is a syntactically idiomatic MWE. But "por meio de" is a standard prepositional phrase, so maybe not. Alternatively, maybe "por meio de" is considered an MWE because it's a fixed collocation. But I need to check if the components can be replaced by synonyms. If you say "por meio de" vs. "por meio de" – replacing with synonyms might not be possible without changing meaning. However, in Portuguese, "por meio de" is a fixed expression, so maybe it's an MWE. Another possibility is "o que significa que" – but that's a common structure, not an MWE. "O que significa que" is like "which means that" in English. But it's a standard grammatical structure, not an idiom. ◯ So, the only possible MWE here might be "por meio de" (indices 6,7,8). Let me check the indices again. The words are: 1: Macron 2: planeja 3: implementar 4: a 5: reforma 6: por 7: meio 8: de 9: decreto 10: presidencial 11: , 12: o 13: que 14: significa 15: que 16: vai 17: saltar 18: os 19: legisladores 20: . So "por meio de" is at positions 6,7,8. If that's an MWE, then that's the only one. But I need to confirm if it's considered an MWE. Since the user's example included "take place" as an MWE, which is a congruent phrase with "ter lugar" in Portuguese, which is "take place" in English. So "por meio de" is congruent with "by means of" in English, which is a standard MWE. Therefore, "por meio de" is an MWE here. Another possible MWE could be "saltar os legisladores," but I don't think that's a standard MWE. "Saltar" in this context might mean "to skip over," but without more context, it's hard to tell. However, "saltar" as a verb can be used in different ways, so it's not a fixed expression. ⊞ So the answer would be "por meio de; 6,7,8". </think>

Table 6: Example for original thinking data for the sentence shown and translated in Table 4 generated by non-instruction-tuned QWEN3-32B.

Lang.	Condition	Training data
RO	**target**	1,500 Romanian sentences.
	branch	1,500 Romance sentences (1,079 French + 421 Portuguese).
	unrelated	1,500 sentences from Swedish, Dutch, Egyptian, Modern Greek, Farsi, Hebrew, Japanese, Georgian, Latvian, Ukrainian, Serbian, Slovene, and Polish; per-language contributions capped at ≈117–118 sentences, Dutch contributing its full 90.
	target+branch	3,000 sentences: 1,500 Romanian (as in **target**) + 1,500 Romance (as in **branch**).
SL	**target**	1,500 Slovene sentences.
	branch	1,500 Slavic sentences (500 Ukrainian + 500 Serbian + 500 Polish).
	unrelated	1,500 sentences from Romanian, French, Portuguese, Swedish, Dutch, Egyptian, Modern Greek, Farsi, Hebrew, Japanese, Georgian, and Latvian; per-language contributions capped at ≈128–129 sentences, Dutch contributing its full 90.
	target+branch	3,000 sentences: 1,500 Slovene (as in **target**) + 1,500 Slavic (as in **branch**).

Table 7: Instruction-tuning data conditions for Romanian (RO) and Slovene (SL).

Type	Guideline Description	Building Block
LVC.full	LVCs in which the verb is semantically totally bleached, EN: to give a lecture.	<*MWE lemmas*> is a multiword expression. More precisely, it is a full light verb construction (LVC.full) because it is formed by a verb and a noun, where the verb is semantically fully bleached. A similar bleaching occurs in the English MWE *(to) give (a) lecture*. The words cannot be realized by synonymous lexemes without altering the meaning of the construction as a whole, which satisfies condition 1. The given expression fulfills condition 2, as the verb displays semantic idiomaticity: the verb does not have its original meaning anymore, instead the meaning is fully bleached. The expression does not refer to a specific person or place, so it also meets condition 3.
NV.LVC.full	Deverbal nominal stemming from an LVC.full (NV.LVC.full), EN: a decision maker - deriving from the LVC.full to make a decision.	<*MWE lemmas*> is a multiword expression. More precisely, it is a deverbal nominal stemming from a full light verb construction (NV.LVC.full). NV describes the fact that a noun (N) is derived from a verb (V), whereas LVC.full refers to the full light verb construction which is the basis for this multiword expression. An English example is *decision maker*, which stems from the LVC.full *(to) make (a) decision*. The words cannot be realized by synonymous lexemes without altering the meaning of the construction as a whole, which satisfies condition 1. The given expression fulfills condition 2, as the verb displays semantic idiomaticity: the deverbal noun does not have its original meaning anymore, instead the meaning is fully bleached. The expression does not refer to a specific person or place, so it also meets condition 3.
AdvID	Adverbial idiom (AdvID) – a universal category, caracterized by lexical, morphological or syntactic irregularity, EN: by and large.	<*MWE lemmas*> is a multiword expression. More precisely, it is an adverbial idiom (AdvID), a universal category characterized by lexical, morphological, or syntactic irregularity. It is comparable to English AdvIDs such as “on the whole” or “by and large.” The lexemes and order are fixed; substituting synonyms disrupts acceptability or the idiomatic meaning, which satisfies condition 1. It fulfills condition 2 through semantic idiomaticity (the expression is non-compositional). The expression does not denote a specific named entity, so it also meets condition 3.

Table 8: Broad MWE category descriptions for three exemplary MWE categories from the PARSEME 2.0 guidelines, and the building blocks made out of the descriptions for the generation of thinking content.

Semantic Stars at MWE-2026 PARSEME 2.0 Subtask 2: Alternative Approaches for MWE Paraphrasing

Elif Bayraktar[1], Vedat Doğancan[1], Muhammed A. Gümüş[1], Nusret Ali Kızılaslan[1],

[1]Istanbul Technical University, Istanbul, Turkey,

Correspondence:{bayraktare24, dogancan23, gumus24, kizilaslan25 }@itu.edu.tr

Abstract

This paper describes the system submitted by Semantic Stars Team for Subtask 2 of the PARSEME 2.0 shared task (Paraphrasing Multiword Expressions). Our approach addresses the challenge of paraphrasing sentences containing MWEs such that the MWE is removed while the original meaning and grammatical structure are preserved. The paper describes multiple distinct approaches powered by open-weight Large Language Models (LLMs), each employing a combination of different techniques such as prompting, multi-agent pipelines and classical NLP methods. Four distinct methods are tested on the test data in French, including a fifth one combining the results from the first four. We tested with several different open-weight LLMs including Llama3.1:8b, Qwen3:8b and gpt-oss-120b and were able to achieve significant improvements over the baseline, securing the first place on the shared task leader board.

1 Introduction

Automatic paraphrasing of Multi-Word Expressions (MWEs), is a challenging task in Natural Language Processing (NLP). MWEs require precise identification of idiomatic boundaries (Savary et al., 2017) and the generation of semantic equivalents that preserve meaning and grammatical closeness. In French, with its discontinuous dependencies and strict morphological constraints, purely stochastic approaches often struggle to maintain grammatical and semantic integrity.

In this paper, we explore multiple paraphrasing methods to address these challenges within the PARSEME 2.0 shared task (Scholivet et al., 2026). We test different approaches, including detailed descriptive paraphrasing (Approach 1), minimal replacements (Approach 2), multi-stage LLM pipelines (Approach 3), and neuro-symbolic hybrid systems (Approach 4). Each method leverages distinct strategies to tackle the task, from rule-based validation to advanced LLM-driven generation. Additionally, we experimented with a combined approach that selects the best output from these methods, although it did not outperform the individual approaches. The results can be seen in Table 1. The following sections detail the architecture of each method and their respective performances.

2 Related Works

Multiword Expression (MWE) processing is a long-standing yet prevalent field in NLP. MWE processing encompasses challenging tasks such as discovery, identification, and paraphrasing of MWEs. It is clear that the academic literature on NLP has a continued interest in this topic. Surveys and review studies were conducted to explore the challenges, discover subtasks related to MWE processing and examine methods developed for modeling MWEs (Constant et al., 2017; Sag et al., 2002; Villavicencio and Idiart, 2019).

MWE paraphrasing, which is our focus, remains one of the most important tasks within the scope of MWE processing. For instance, Yimam et al. (2016) aimed to examine the importance of context for the paraphrase ranking task using a binary classification approach with K-Nearest Neighbors (kNN) and a learning-to-rank approach with the LambdaMART algorithm. Barančíková and Kettnerová (2018) used word embeddings to paraphrase Czech verbal MWEs using single verbs. A machine translation experiment was conducted comparing sentences from both versions, and the results showed that the performance of the approach is promising. Zhou et al. (2022) proposed an approach for the Idiomatic Sentence Paraphrasing (ISP) task without strong supervision. This study addressed the challenge of data scarcity in the MWE paraphrasing domain. Wada et al. (2023) proposed an unsupervised approach to MWE paraphrasing, which includes clustering sentences that

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 187–195

contain MWEs using the DBSCAN algorithm and utilizing pre-trained LLMs to generate paraphrases for MWEs. Barreiro and Mota (2023) addressed the problem of multilingual paraphrasing of MWEs and developed a tool named CLUE-Aligner to facilitate the alignment of non-contiguous multiword units. Liu et al. (2025) addressed the impact of MWEs on machine translation quality. Their results showed that the presence of MWEs negatively affects translation performance. To mitigate this, they proposed an LLM-based system that paraphrases MWEs into their literal counterparts using few-shot prompting before performing the machine translation task.

Ultimately, MWE paraphrasing for a wide range of languages remains a current and significant challenge in NLP. Following the objectives of the PARSEME 2.0 shared task, this study aims to develop a methodology to improve the MWE paraphrasing performance in French sentences.

3 Methodology

We developed 4 different methods to paraphrase the expressions. Here, we describe the two that achieved the highest BERTScores during automated evaluation, since these were the only ones submitted to the official leader board (Manon Scholivet, 2025). The other two are briefly mentioned in section 3.3. We also developed a fifth (combined) approach aimed at combining the best results from each of the methods.

3.1 Paraphraser 1

This approach addresses the challenge of removing idiomaticity while preserving meaning through a multi-agent pipeline powered by open-weight Local LLMs. The system employs a **"Generate-Validate-Fix"** architecture. It first generates a candidate paraphrase, then subjects it to a rigorous hybrid evaluation consisting of deterministic string-matching constraints and a semantic LLM judge. If a candidate fails, a specialized "Fixer" agent is activated.

We benchmarked several local models (including Llama3.2:3b, Gemma3:4b, Qwen3:8b and DeepSeek-r1:7b) and selected **Qwen3:8b** for its optimal balance of reasoning capability and resource efficiency as well as its multilingual abilities (Grattafiori et al., 2024; Team et al., 2025; Yang et al., 2025; Guo et al., 2025) .

The system moves beyond simple prompting by using an agentic workflow implemented in Python using the Ollama framework(Ollama Team) and LangChain Library(LangChain Team). The architecture consists of three distinct phases.

3.1.1 Phase 1: Initial Generation

The first stage employs a standard LLM agent prompted with strict constraints. We utilize a zero-shot prompting strategy. The prompt explicitly instructs the model to replace the MWE while keeping the rest of the sentence structure as close to the original as possible.

3.1.2 Phase 2: Hybrid Validation

Prior work has demonstrated that large language models often fail to follow negated instructions (e.g., "do not use word X") or strict negative constraints (Jang et al., 2023). Although LLMs excel at judging semantic content, we observed that they may still produce exactly the expression they are instructed to avoid. Therefore, we implement a hybrid validator:

Deterministic Validator (Classical NLP) We use Python-based string manipulation to enforce the "Elimination" criterion. We normalize both the MWE and the candidate prediction (lowercasing, punctuation removal). If the target MWE tokens appear in the prediction or if the prediction is identical to the source, the candidate is immediately marked as a failure ($Score = 0$) without consuming GPU resources for semantic checking.

Semantic Judge (LLM Agent) If the deterministic checks pass, a second LLM agent evaluates the candidate. This agent uses chain-of-thought and few-shot techniques to judge the results from the previous step. It is prompted to act as a "Linguistic Evaluator" checking two specific criteria: 1. **Meaning Preservation:** Does the paraphrase convey the exact sense of the original? 2. **Grammatical Closeness:** Does the paraphrase maintain the original morphological features (tense, number)?

3.1.3 Phase 3: The Fixer

If a candidate fails validation, it is passed to a "Fixer" agent. Unlike the initial generator, the Fixer receives the specific reason for failure (e.g., *"Failure: MWE (multiword expression) tokens still present"*). The Fixer uses detailed instructions as well as the data from the previous steps to correct the errors in the paraphrased sentence.

Example Expression Corrected by the Fixer While the original sentence was (fra) *Le **point de vue** de le réalisateur* (lit. 'The point of view of the director'), the initial generator produced a grammatically and idiomatically flawed substitute: (fra) *La **vue** de le réalisateur* (lit. 'The view of the director'). This was flagged by the *Semantic Judge* since, the use of *vue* (physical sight) is a non-idiomatic calque for "opinion" in this context. The *Fixer* agent successfully resolved the issue producing: (fra) *La **vision** du réalisateur* (lit. 'The vision of the director'). This result is more correct as it selects a semantically appropriate synonym (*vision*) and restores grammatical validity.

3.1.4 Design Decisions: Explainer Tool

We also experimented with a Retrieval Augmented Generation (RAG) approach using Wikipedia definitions, but it often introduced noise such as disambiguation errors. We found that 7B+ models performed better with internal knowledge. Although, there is potential to improve the RAG approach in future work, we chose to replace it with an internal "Explainer" mechanism in the Judge Agent (Section 3.1.2).

3.1.5 Parallelization

To optimize the performance, we implemented a thread-based parallelization strategy. We enabled concurrent request processing (`OLLAMA_NUM_PARALLEL=2`) and managed VRAM usage by dynamically limiting the context window.

The next approach consists of three main phases: MWE extraction, semantic replacement generation with similarity validation, and sentence transformation. We employ a pipeline that leverages LLMs (Qwen3:8b, gpt-oss-120b (OpenAI et al., 2025)) for linguistic processing and embedding models for semantic similarity verification.

3.1.6 Phase 1: MWE Extraction

The first phase involves identifying and extracting the target MWE from the input sentence. In the provided dataset, MWEs are marked using double square brackets (e.g., `[[expression here]]`). We employ regular expression pattern matching to extract these marked expressions from the raw text. This extraction step isolates the multiword expression that will subsequently be replaced with a semantically equivalent shorter form.

3.1.7 Phase 2: Single-Word Equivalent Generation with Semantic Validation

The second phase constitutes the core of our methodology: generating a semantically appropriate single-word replacement for the extracted MWE. This phase employs an iterative refinement process guided by semantic similarity measures.

Initial Replacement Generation We prompt the LLM to generate a single-word French equivalent for the given MWE. The model is instructed through a system prompt to analyze the multiword expression and output only a single-word replacement that preserves the original meaning. Few-shot examples are provided to guide the model's behavior.

Semantic Similarity Verification To ensure that the generated replacement maintains semantic closeness to the original MWE, we employ cosine similarity as our semantic equivalence metric. This verification step is crucial for filtering out replacements that, while grammatically valid, may drift from the intended meaning of the original expression.

Cosine Similarity Computation. We compute the cosine similarity between two vectors as:

$$\text{sim}(\mathbf{e}_{\text{mwe}}, \mathbf{e}_{\text{rep}}) = \frac{\mathbf{e}_{\text{mwe}} \cdot \mathbf{e}_{\text{rep}}}{\|\mathbf{e}_{\text{mwe}}\| \times \|\mathbf{e}_{\text{rep}}\|} \quad (1)$$

This score ranges from -1 to 1, with higher values indicating greater similarity.

Threshold-Based Validation. We establish an empirically determined similarity threshold of $\tau = 0.50$ to determine whether a proposed replacement is semantically acceptable. If $\text{sim}(\mathbf{e}_{\text{mwe}}, \mathbf{e}_{\text{rep}}) \geq \tau$, the replacement is accepted. Otherwise, the system initiates the iterative refinement process described in the following paragraph. This threshold was chosen empirically, taking into account the typical length of a multi-word expression relative to the surrounding sentence and the fact that BERTScore will be used for evaluation. Since BERTScore compares predictions to both a "minimal" (close to the source) and a "creative" (more divergent) reference and selects the closest match, using cosine similarity in our system primarily ensures that replacements stay close to the minimal reference, with iterative refinement applied only when similarity falls below the threshold.

Table 1: Automated & Manual evaluation scores for PARSEME 2.0 Subtask 2 French Test Set.

Approach	Ave. BERTScore	Manual Score	Diversity of New Words		
			Richness	Evenness	Entropy
Paraphraser 2	**93.90**	64.82	236	0.83	4.54
Paraphraser 1	89.46	**79.25**	**456**	0.90	5.48
Paraphraser 4	88.76	–	345	0.91	5.33
Combined	87.10	–	415	0.94	5.64
Paraphraser 3	85.72	–	399	**0.95**	**5.72**
Baseline	77.55	72.70	326	0.92	5.33

Note: The baseline model is gpt-oss-120b. Evenness and Entropy refer to Shannon Evenness and Shannon-Weaver Entropy, respectively. Manual evaluation was omitted for three approaches as they were not officially submitted to the leaderboard.

Iterative Refinement Process If the initial similarity score falls below the threshold, we initiate an iterative refinement process. The system maintains a blacklist of previously rejected candidates along with their similarity scores. In subsequent iterations, the LLM is prompted to generate alternative replacements while explicitly excluding blacklisted expressions. This process continues for a maximum of three iterations. Throughout this process, we track the best candidate encountered (i.e., the one with the highest similarity score). If no candidate exceeds the threshold after all iterations, we select the best candidate observed during the refinement process.

3.1.8 Phase 3: Sentence Transformation

The final phase transforms the original sentence by substituting the MWE with the validated single-word equivalent. We leverage the LLM, providing it with:

- The original sentence containing the MWE
- The identified multiword expression
- The validated single-word replacement

The model is instructed to perform the substitution while maintaining grammatical correctness. The prompt explicitly requests only the transformed sentence without additional explanations or formatting.

3.2 Combined Approach

Paraphrasers 3 and 4 utilize different strategies for paraphrasing. **Approach 3** is a multi-stage LLM-based pipeline that combines prompt engineering, POS-based constraints, and rule-based validation to generate and score paraphrase candidates. **Approach 4** integrates morpho-syntactic analysis and semantic enrichment with LLM generation and post-processing to preserve grammatical structure and figurative meaning.

In addition to these methods, we implemented a combined approach that aimed to leverage the strengths of each strategy. This approach used an LLM (gpt-oss-120b) as a selector to choose the optimal output from the candidates generated by the other methods. However, despite this, the combined approach performed worse, revealing some biases in the selection process (such as favoring or punishing approaches, depending on the order in which they are presented in the prompt). This suggests that the selection mechanism has room for improvement.

4 Results

We evaluated our system using the official PARSEME 2.0 metrics alongside a manual evaluation by language experts. The automated metrics include BERTScore (Zhang et al., 2019), Richness, Shannon Evenness, and Shannon-Weaver Entropy for diversity.

As we can see in Table 1, all of our experimental alternatives exceeded the BERTScore baseline of 77.55. On the official PARSEME 2.0 shared task leaderboard for French, our Paraphraser 2 approach (93.90) ranked first among all participants in the automated evaluation, while our Paraphraser 1 approach (79.25) secured first place overall in the manual scoring track. This divergence between automated and manual success is one of the key findings of our study, highlighting the importance of metric sensitivity in MWE paraphrasing.

Qualitative Analysis and Strategy Divergence Through manual inspection, we found that this divergence is rooted in the distinct paraphrasing strategies employed. Paraphraser 2 (Minimal) ex-

cels at lexical compression and identifying direct synonyms, which yields a high BERTScore due to its proximity to the source text. However, despite utilizing a much larger 120B parameter backbone ($gpt - oss - 120$), it is prone to contextual overgeneralization, for instance, reducing the technical NMWE *système d'information* (lit.'information system') to the overly broad *informatique* (lit.'IT').

In contrast, Paraphraser 1 (Explanatory) utilizes a multi-agent pipeline powered by a smaller 8B model ($Qwen3$) to generate descriptive, context-aware expansions. While this strategy results in a slightly lower BERTScore and higher verbosity, manual evaluation confirms it is superior in the majority of the complex cases. It preserves domain-specific precision and idiomaticity (e.g., *année dédiée à l'international* (lit.'the international dedicated year')) where the Minimal approach often produces literal calques or incorrect substitutions.

These empirical findings, further illustrated via comparative traces in Appendices B–E, suggest that the Explanatory approach is better suited for maintaining the semantic non-compositionality inherent in complex MWEs. Although our initial attempts at a combined ensemble did not yield immediate performance gains, the complementary nature of these strategies(lexical brevity versus semantic precision) suggests that a more refined categorical selection process based on MWE type remains a promising area for future optimization.

5 Conclusion

We presented several strategies for the PARSEME 2.0 French paraphrasing task, each integrating external validation—such as string matching and rule-based filtering—to overcome the limitations of LLMs in adhering to strict linguistic constraints. This hybrid approach, merging generative power with symbolic verification, proved essential for maintaining semantic precision and output quality.

Paraphraser 2 ranked first in automated scoring (93.90), yet Paraphraser 1 secured first place in the manual track (79.25). This "metric sensitivity gap" indicates that while automated scores favor lexical compression, human experts prioritize the semantic precision and explanatory depth necessary for non-compositional MWEs. The superior manual performance of the 8B approach over the 120B model suggests that targeted multi-agent architectures outperform raw parameter scale for complex linguistic tasks.

Although our evaluation focused on French, the majority of our pipeline components are language-agnostic, suggesting potential applicability to other languages addressed by the PARSEME 2.0 shared task. Future work could explore the cross-lingual transfer of these methods.

Discussion

Recent mechanistic studies have identified an "ironic rebound" effect in LLMs, where explicitly naming a forbidden word paradoxically primes the model to generate that very token (Mann et al., 2025). This phenomenon (where the instruction's activation of a concept is systematically stronger than its suppression signal) is a primary cause of negative constraint failure (Rana, 2026). While our methodology naming the target MWE within prompt delimiters (e.g., [[...]]) theoretically risks this priming, our system mitigates this through external string-matching constraints and iterative verification.

Furthermore, the task of paraphrasing itself may offer a **generalizable solution to the ironic rebound problem**. Future work could explore whether paraphrasing architectures such as ours could provide a generalizable solution to the ironic rebound problem. By transforming a negative constraint into a positive generative goal (a "generative pivot"), hybrid pipelines that combine neural generation with classical NLP methods may offer a robust pathway for bypassing the internal priming failures inherent in current LLM architectures.

Limitations

The combined approach underperformed due to biases in the LLM selection process, affecting its ability to consistently choose the best paraphrase. Additionally, the LLM-based multi-agent framework, unlike traditional rule-based NLP methods, can struggle with stability and precision, as it relies on the quality of underlying models, which may introduce errors or inconsistencies.

Acknowledgments

The authors used AI-assisted editing tools, to improve spelling, grammar, clarity, and readability during part of the manuscript preparation. After using the tools, the authors carefully reviewed and edited the text and take full responsibility for the final content.

References

Petra Barančíková and Václava Kettnerová. 2018. Paraphrases of verbal multiword expressions: The case of czech light verbs and idioms. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, volume 2, page 35. Language Science Press.

Anabela Barreiro and Cristina Mota. 2023. A multilingual paraphrasary of multiwords. In *Proceedings of the 1st International Workshop on Multilingual, Multimodal and Multitask Language Generation*, pages 47–56.

Matthieu 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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can large language models truly understand prompts? a case study with negated prompts. In *Proceedings of The 1st Transfer Learning for Natural Language Processing Workshop*, volume 203 of *Proceedings of Machine Learning Research*, pages 52–62. PMLR.

LangChain Team. Langchain. `https://www.langchain.com/`. Framework for developing applications powered by large language models.

Linfeng Liu, Saptarshi Ghosh, and Tianyu Jiang. 2025. Evaluating the impact of verbal multiword expressions on machine translation. *arXiv preprint arXiv:2508.17458*.

Logan Mann, Nayan Saxena, Sarah Tandon, Chenhao Sun, Savar Toteja, and Kevin Zhu. 2025. Don't think of the white bear: Ironic negation in transformer models under cognitive load. *ArXiv*, abs/2511.12381.

Agata Savary Eric Bilinski Carlos Ramisch Manon Scholivet, Takuya Nakamura. 2025. PARSEME 2.0: Shared task on idiomaticity and multiword expressions.

Ollama Team. Ollama. `https://ollama.com/`. Local inference framework for large language models.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b and gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Shailesh Rana. 2026. Semantic gravity wells: Why negative constraints backfire. *Preprint*, arXiv:2601.08070.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *International conference on intelligent text processing and computational linguistics*, pages 1–15. Springer.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and 1 others. 2017. The parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th workshop on multiword expressions (MWE 2017)*, pages 31–47.

Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Carp, and Vasile Pais. 2026. Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Aline Villavicencio and Marco Idiart. 2019. Discovering multiword expressions. *Natural Language Engineering*, 25(6):715–733.

Takashi Wada, Yuji Matsumoto, Timothy Baldwin, and Jey Han Lau. 2023. Unsupervised paraphrasing of multiword expressions. *arXiv preprint arXiv:2306.01443*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Seid Muhie Yimam, Héctor Martínez Alonso, Martin Riedl, and Chris Biemann. 2016. Learning paraphrasing for multi-word expressions. In *MWE 2016-Multiword Expression Workshop 2016*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Jianing Zhou, Ziheng Zeng, Hongyu Gong, and Suma Bhat. 2022. Idiomatic expression paraphrasing without strong supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11774–11782.

A System Trace Example

The following trace demonstrates Paraphraser 1 ("Generate-Validate-Fix") resolving a combined lexical and grammatical failure.

Original Sentence: (fra) *Le **point de vue** de le réalisateur*
(lit. 'The point of view of the director')
'The director's perspective'.

Initial Prediction: (fra) *La **vue** de le réalisateur*
(lit. 'The view of the director').

Validator Feedback: *"Failure: The prediction contains a grammatical error in 'de le réalisateur' which should be 'du réalisateur'. Additionally, 'vue' is non-idiomatic for this context."*

Final Prediction: (fra) *La **vision** du réalisateur*
(lit. 'The vision of the director')
'The director's vision'.

B Comparative Trace: Expansion vs. Overgeneralization

This example illustrates the risk of overgeneralization in the minimalist approach (Approach 2) compared to the context-aware generation of Approach 1.

Original Sentence: (fra) *Malone L' année 1996 a été... d' **année internationale** pour...*
(lit. '...of the international year for...')

Paraphraser 1 (Explanatory): (fra) *...d'**année dédiée à l'international** pour...* (lit. '...of the international dedicated year for...')
Verdict: Passed first try. The agent expanded the MWE into an idiomatic descriptive phrase that preserves the UN's formal register.

Paraphraser 2 (Minimal): (fra) *...d'**Internationalisation** pour...*
(lit. '...of internationalization for...')
Verdict: Semantic failure. *Internationalisation* is a process, not a designation for a specific year, leading to a loss of core meaning.

C Comparative Trace: Stylistic Convergence

This example demonstrates a case where both the Minimalist and Explanatory approaches produce semantically accurate and idiomatic results.

Original Sentence: (fra) *...avec comme **personnage principal** le meilleur matador...*
(lit. '...with as main character the best matador...')

Paraphraser 1 (Explanatory): (fra) *...avec comme **figure centrale** le meilleur matador...*
(lit. '...with the best matador as the central figure...')
Verdict: High semantic similarity. The choice of *figure centrale* is idiomatic and provides a smooth, literary flow.

Paraphraser 2 (Minimal): (fra) *...avec comme **protagoniste** le meilleur matador...*
(lit. '...with as protagonist the best matador...')
Verdict: High semantic similarity. This shows the strength of Paraphraser 2's iterative search when a direct, single-word synonym is available. Note: capitalization was manually corrected from the system output.

D Comparative Trace: Granularity and Specificity

This case illustrates the difference between descriptive expansion and collective abstraction.

Original Sentence: (fra) *...issus de les **familles nobles** de la ville.*
(lit. '...descended from the noble families of the city.')

Paraphraser 1 (Explanatory): (fra) *...issus des **familles de la noblesse** de la ville.* (lit. '...descended from the noble families (or families of the nobility) of the city.')
Verdict: High similarity. It maintains the plural focus on family units.

Paraphraser 2 (Minimal): (fra) *...issus de **la Noblesse** de la ville.*
(lit. '...descended from the Nobility of the

city.')
Verdict: Highly idiomatic but less specific. By substituting a collective noun for a plural MWE, it shifts the focus from individual families to the social class as a whole.

E Comparative Trace: Technical and Domain-Specific Precision

This trace demonstrates the failure of minimalist constraints when handling specialized terminology.

Original Sentence: (fra) *Le Forbin a été équipé d' un **système d' information**...*
(lit. 'The Forbin was equipped with an information system...')

Paraphraser 1 (Explanatory): (fra) *...équipé d'un **système de gestion d'information**...* (lit. '...equipped with an information management system...')
Verdict: High semantic closeness. While slightly more descriptive than the original, it preserves the technical and operational register required for the maritime context.

Paraphraser 2 (Minimal): (fra) *...équipé d'**Informatique**...*
(lit. '...equipped with IT...')
Verdict: Low semantic closeness. *Informatique* is too broad; it fails to capture the notion of a specific operational management system, rendering the sentence vague and non-idiomatic.

F Prompt Templates

This appendix provides the prompt templates used for Methods 1 and 2.

Paraphraser 1

Paraphraser 1 utilizes a multi-agent approach to ensure semantic depth and grammatical correctness through a feedback loop.

`generation_template`

```
"""
You are a linguistic expert.
Task: Paraphrase the sentence to remove the
    idiomatic nature of the specified
    expression.

Input Sentence: "{sentence}"
Expression to replace: "{mwe}"

Instructions:
1. Replace the expression '{mwe}' with a
    literal, non-idiomatic equivalent.
2. Keep the rest of the sentence structure as
    close to the original as possible.
3. Ensure the output is in the SAME LANGUAGE as
     the input.
4. Output ONLY the new sentence text with no
    introduction or quotes.
"""
```

`judge_template`

```
"""
You are a Linguistic Evaluator.
Task: Evaluate a Paraphrase based on strict
    criteria.

Original: "{original}"
Prediction: "{prediction}"
MWE Removed: "{mwe}"

Evaluate against these criteria:
1. MEANING: What is the meaning of the
    multiword expression (MWE) in the sentence?
     Is the sense preserved?
2. GRAMMATICALITY: Is the prediction
    grammatically correct (no spelling/grammar
    errors)?
3. GRAMMATICAL CLOSENESS: Does it keep the
    exact Tense, Mood, and Number of the
    original?
   - Example Fail: Original is "He was walking"
     (Past Continuous), Prediction is "He walks
    " (Present). -> FAIL.

Respond in JSON only.
{format_instructions}
"""
```

`fixer_template`

```
"""
You are a Linguistic Correction Expert.
A system attempted to paraphrase a sentence but
     failed evaluation.

Data:
- Original Sentence: "{original}"
- MWE to Remove: "{mwe}"
- Failed Prediction: "{prediction}"
- Failure Reason: "{reason}"

Instructions:
1. Write a NEW paraphrase.
2. ELIMINATION: Do NOT use the MWE "{mwe}".
3. MEANING: Keep the exact meaning.
4. GRAMMATICAL CLOSENESS: You MUST match the
    Tense, Mood, and Number of the original
    sentence.
5. You MUST preserve the exact meaning of the
    original
6. Output ONLY the new sentence.
7. Original sentence MUST BE PRESERVED other
    than PARAPHRASING of the indicated MWE (
    multi word expression).
"""
```

Paraphraser 2

Paraphraser 2 focuses on high-precision lexical substitution using a larger model and an iterative blacklist mechanism.

FRENCH_MWE_TRANSLATION

```
"""
You are a tool that converts French multiword
    expressions into single-word equivalents.
    Briefly analyze the multiword expression
and convert them into a single word. You should
     ONLY OUTPUT THE NEW EXPRESSION. NOTHING
    ELSE. Here are some examples:

Input: Gravir les chelons
Output: Progresser

Input: Faire attention
Output: Surveiller

DO NOT RETURN THE INPUT ITSELF! YOU SHOULD GIVE
     A NEW EXPRESSION THAT HAS THE SAME MEANING
     AS THE ORIGINAL ONE. IT SHOULD BE A SINGLE
     WORD
"""
```

new_mwe_prompt

```
def new_mwe_prompt(blacklist: dict):
    blacklist_str = ""

    for word, similarity in blacklist.items():
        blacklist_str += f"\n- {word}"

    return f"""
    You are a tool that converts French
     multiword expressions into single-word
     equivalents. Briefly analyze the multiword
     expression
    and convert them into a single word
     expression. You should ONLY OUTPUT THE NEW
     EXPRESSION. NOTHING ELSE. Here are some
     examples:

    Input: Gravir les chelons
    Output: Progresser

    Input: Faire attention
    Output: Surveiller

    DO NOT RETURN THESE EXPRESSIONS:
    {blacklist_str}
    """
```

minimal_sentence_prompt

```
def minimal_sentence_prompt(sentence, mwe,
    new_exp):
    return f"""
You are a French text transformation tool. Your
     task is to replace a multiword expression
    in a sentence with a new expression.

INSTRUCTIONS:
- Replace the specified multiword expression
    with the new expression
- Keep the grammer correctly.
- Output ONLY the transformed sentence
- Do not include explanations, quotes, or
    additional text

EXAMPLE:
Sentence: Les lves de la classe de CE1 de l'
    cole Notre-Dame ont pris part  les
    festivits.
Multiword expression: pris part
New expression: particip
Output: Les lves de la classe de CE1 de l'cole
    Notre-Dame ont particip aux festivits.

YOUR TASK:
Sentence: {sentence}
Multiword expression: {mwe}
New expression: {new_exp}

JUST OUTPUT THE TRANSFORMED SENTENCE ONLY:
"""
```

MorphoFiltered-Gemini at MWE-2026 PARSEME 2.0 Subtask 1: Tackling LLM Overgeneration via Universal POS-based Constraints

Irina Moise* **and Sergiu Nisioi***
Human Language Technologies Research Center
Faculty of Mathematics and Computer Science
University of Bucharest
moiseirina42@gmail.com
sergiu.nisioi@unibuc.ro

Abstract

This paper describes **MorphoFiltered-Gemini**, a system submitted to the PARSEME 2.0 Shared Task (Scholivet et al., 2025), subtask 1 on MWE identification, covering all 17 target languages. The system combines LLM-based predictions generated via the Gemini API with a morphological post-filter designed to reduce false positives. Rather than optimizing peak performance on individual languages, our approach prioritizes cross-lingual stability and precision. As a result, the system exhibits a balanced performance across languages and MWE categories, achieving the highest Shannon evenness score among all submissions.

1 Introduction

Multiword expressions (MWEs) are "word combinations that exhibit lexical, syntactic, semantic, pragmatic, and/or statistical idiosyncrasies" (Baldwin and Kim, 2010) . For example, *a big fish* refers to an important person. Similar phenomena appear across languages, including light verb constructions (to *grant rights*), adjectival idioms (to be *on cloud nine*), fixed adpositional phrases (*on behalf of*), pronoun idioms (*each other*) and many others. The constituent words of multi-word expressions are common, and their combination behaves as a single unit.

PARSEME 2.0 addresses the automatic identification of MWEs in running text in a multilingual setting. Unlike previous PARSEME shared tasks, which focused only on verbal MWEs, this edition (Savary and Ramisch, 2025) extends the task to all syntactic types (verbal; nominal; adjectival and adverbial; functional). Systems must identify MWEs across 17 languages: Dutch, Egyptian (ca. 2700-2000 BC), French, Georgian, Greek (Ancient), Greek (Modern), Hebrew, Japanese, Latvian, Persian, Polish, Brazilian Portuguese, Romanian, Serbian, Slovene, Swedish, and Ukrainian.

MWEs may be continuous or discontinuous, and they can overlap, further complicating automatic detection. Systems are evaluated both at the token level and at the MWE level, where a correct prediction requires identifying all components of an expression. As a result, partial matches are penalized, making the task sensitive to recall and span boundary errors.

While morphologically constrained approaches often achieve high precision, they typically suffer from poor recall and limited generalization. In contrast, large language models generalize well but tend to accept too many candidates, leading to many false positives.

In this work, we explore a hybrid approach that combines the strengths of both paradigms. MorphoFiltered-Gemini[1] relies on Gemini 2.0 Flash-Lite (DeepMind, 2025) to generate MWE predictions using prompting, followed by morphological post-processing, applied selectively. The goal is not to maximize performance on specific languages, but to study whether minimal linguistic constraints can stabilize LLM behavior.

Our contributions are threefold: (i) we propose a LLM-based pipeline for multilingual MWE identification, (ii) we investigate the effects of few-shot prompting across diverse languages, and (iii) we introduce a morphological filtering strategy that improves precision without relying on language-specific training.

2 Experiments

We evaluated several strategies to understand the trade-offs between recall-oriented LLM predictions and precision-oriented filtering.

As an initial baseline, we implemented a purely rule-based system relying on morphological patterns. Although this approach achieved a token-based F1 of 0.152 and performed reasonably

*Corresponding authors.

[1] https://github.com/irinamoise/PARSEME

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 196–202

well for a small number of languages (e.g. Romanian and Persian), it failed to generalize, yielding extremely low scores for others such as Georgian. This confirmed the poor recall of rule-based methods in a multilingual setting.

We then applied Gemini as a post-processing validator on those candidates. Although the LLM consistently rejected around 30% of proposed MWEs, this strategy led to negligible improvements. Due to the very low recall of the rule-based system, only a small fraction of true MWEs were ever presented to the LLM, making it impossible to recover missing predictions.

Subsequently, Gemini was adopted as the primary predictor. Using separate prompts for detection and classification improved recall. Few-shot prompting with examples from the training sets and negative constraints yielded substantial gains for some languages, but caused severe degradation for others, including complete failure for Egyptian.

Finally, we introduced a lightweight morphological post-filter applied to LLM-generated predictions. This filter consistently lowers the number of false positives and leads to more stable performance across languages.

3 System Architecture and Methodology

The overall architecture of our system is illustrated in Figure 1 (see Appendix A).

The system follows a five-stage pipeline:

1. Preprocessing: Conversion of PARSEME CUPT annotations to BIO tags.
2. MWE Detection: Span prediction using Gemini 2.0 Flash Lite via batch prompting.
3. MWE Classification: Assignment of MWE categories to detected spans.
4. Morphological Post-Processing: Removal of unlikely or low-confidence MWEs using POS-based and single-token filters.
5. Format Conversion: Reconstruction of CUPT-formatted outputs from BIO predictions.

3.1 Preprocessing and BIO Representation

PARSEME annotations are provided in the CUPT format (Ramisch, 2018), an extension of Universal Dependencies (UD) (Nivre, 2020). While CUPT offers rich annotation capabilities, it is not directly suitable for LLM-based processing. We therefore convert annotations into standard BIO tags, which provide a simplified, token-level representation of MWE spans. This conversion simplifies span reconstruction from LLM outputs.

A limitation of this representation is that it does not adequately capture discontinuous MWEs or cases where tokens participate in multiple overlapping MWEs. As a consequence, such expressions are often misinterpreted or collapsed into incomplete spans during BIO-based processing, which leads to zero scores for discontinuous MWEs in the official evaluation. These structural conflicts, including overlapping and discontinuous configurations, are not yet resolved in the current system.

3.2 LLM-based MWE Detection and Classification

Batch Prompting for Detection MWE detection is performed using Gemini 2.0 Flash-Lite in batches of ten sentences. The prompt includes a general definition of MWEs, language-specific examples, and token-indexed sentences to ensure span identification (see appendix B.1). For a subset of languages (EL, FA, FR, SV), the prompts uses an improved strategy: the dictionary exemples are substituted by 10 full phrases with MWE examples extracted from the training data. An additional prompt containing negative constraints (anti-examples) explicitly defines undesirable outputs, thereby narrowing the solution space and ensuring the model adheres to specific stylistic and structural boundaries.

The output format is strictly constrained to token indices corresponding to predicted MWE spans, or a special NONE marker when no MWE is detected.

MWE Category Classification Detected MWEs are passed to a second prompt that assigns MWE categories based on generic category descriptions (see appendix B.2). Separating detection and classification makes the model focus on span identification independently of category semantics, which improved the results.

Few-Shot Prompting Strategies We experimented with few-shot prompting (Brown et al., 2020) using 4-5 examples of MWEs from a small dictionary and with full sentences from the training sets that contained MWEs (see appendix B.3). While the second strategy produced good results for a few languages, it led to severe degradation for others. These observations motivated the final

system design, which applies the last method selectively rather than uniformly across all languages.

3.3 Morphological Post-Processing

POS Pattern Filtering LLM predictions frequently include false positives. To adjust this effect, we applied a morphological filter that removes unlikely POS patterns, such as (DET, NOUN) or (ADJ, NOUN), while preserving high-precision constructions such as (VERB, NOUN), (ADP, NOUN). See appendix C and D.

This filter is applied to 11 languages (EGY, EL, FA, FR, GRC, NL, PL, RO, SR, SV, UK) for which empirical evaluation showed consistent precision gains (see appendix F).

Single-Token MWE Filtering Gemini occasionally predicts single tokens as MWEs, despite PARSEME annotations treating most single-token expressions as non-MWEs. To address this issue, we introduced a single-token filter that removes isolated MWE predictions (see appendix E). This filter proved beneficial for EL, FA, GRC, KA, and LV, and was therefore retained for these languages in the final system.

Language-Specific Exclusions For Hebrew, Japanese, and Slovene, morphological filtering consistently degraded performance. These languages were therefore excluded from postprocessing. This suggests that POS-based constraints are less reliable for languages with complex morphology, logographic writing systems, or ambiguous UD tagsets.

3.4 Caching and Seen/Unseen Analysis

To reduce redundant LLM calls, the system implements a caching mechanism that stores previously predicted MWE spans. This mechanism primarily benefits MWEs that appear multiple times across training and development data.

Evaluation results confirm a substantial performance gap between MWEs seen during training/development and unseen expressions. For MWEs identical to those in the training/development data, the system achieves high precision (P=75.06, R=15.91, F1=26.25). Variant forms of seen MWEs also benefit from caching and contextual similarity (P=54.54, R=12.00, F1=19.67). In contrast, unseen MWEs exhibit much lower scores (P=8.59, R=9.21, F1=8.89). When aggregating all seen-in-traindev MWEs, performance remains substantially higher (P=73.08, R=14.86, F1=24.70), highlighting the role of memorization and context reuse in LLM-based systems.

4 Results

Overall Performance In the official evaluation (Ramisch, 2025), **MorphoFiltered-Gemini** was submitted for all 17 languages. The system did not achieve top overall F1 scores, but it ranks third in token-based precision and exhibited a clear gap between token-based and MWE-based performance, as shown in Table 1. This discrepancy reflects the conservative behavior induced by morphological filtering and the strict nature of MWE-level evaluation, where partially correct predictions are penalized.

Token-Based vs MWE-Based Evaluation High token-based precision indicates accurate boundary detection for individual tokens, even when full MWEs are not perfectly reconstructed. In contrast, MWE-based evaluation requires exact span matches, penalizing conservative systems and those unable to represent discontinuous expressions.

Diversity Metrics The system achieves the highest Shannon evenness score among all submissions, as shown in Table 2, indicating balanced performance across languages and MWE categories. Unlike systems exhibiting strong performance peaks for a small subset of languages, **MorphoFiltered-Gemini** avoids extreme failures and maintains a stable cross-lingual behavior.

4.1 Error Samples

We discuss a few prediction errors in French, Portuguese and Romanian with the goal of presenting the current limitations of our system.

In French, the system frequently misses light verb constructions (LVC.full) when they are discontinuous. For example, in *"Éric Halphen reçoit à son cabinet un coup de fil anonyme"*(Éric Halphen receives an anonymous phone call at his office), the MWE *reçoit ... coup de fil* is not detected because the verb and the noun are separated by other words. Similar issues occur for adverbial idioms (AdvID) like *avec vigueur* in *"avec une vigueur accrue"*(with increased vigor), which are often interpreted as free prepositional phrases.

False positives mostly involve grammatical constructions that resemble MWEs but are actually compositional. In Romanian, a peposition and an infinitive marker *de a* are incorrectly predicted as

System	#Langs	Global MWE-based				Global Token-based			
		P	R	F1	Rank	P	R	F1	Rank
MorphoFiltered-Gemini	17/17	20.95	14.50	17.14	7	34.14	24.20	28.32	5

Table 1: General Ranking

System	#Langs	Richness		Shannon-Weaver Entropy		Shannon-Evenness	
		Value	Rank	Value	Rank	Value	Rank
MorphoFiltered-Gemini	17/17	56.53	7	3.71	5	0.97	1

Table 2: Diversity Ranking

an AdpID in *"șansa de a vedea clar"* (the chance to see clearly), although it is just a syntactic pattern. In Portuguese, standard contractions such as *de o* (of the) and *em o* (usually *in the*, here meaning *at the*) are over-generated as MWEs , as in *"No final do quarto ano"* (At the end of the fourth year), even though they reflect regular morphology rather than idiomatic usage. All in all, false positives tend to arise from surface patterns that look fixed, while false negatives are mainly caused by discontinuity, reflexive clitics, and complex adpositional structures.

5 Conclusion

The results suggest that combining general-purpose LLM predictions with minimal linguistic post-processing yields balanced evaluation outcomes, but with limitations. Future work includes extending the representation to support discontinuous MWEs and exploring adaptive filtering strategies that better balance recall and precision across languages.

Acknowledgements

This work was supported by the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology), and by the Romanian National Research Council (CNCS) through the Executive Agency for Higher Education, Research, Development and Innovation Funding (UEFISCDI) under grant PN-IV-P2-2.1-TE-2023-2007 InstRead.

References

Timothy Baldwin and Su Nam Kim. 2010. *Multiword Expressions*. Taylor and Francis.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Google DeepMind. 2025. Gemini 2.0 flash-lite: Cost-efficient multimodal reasoning. Google Developers Blog. Released February 5, 2025.

Joakim Nivre. 2020. Universal Dependencies v2: An Evergreen Corpus for Cormpus-based Linguistics and NLP. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043.

Carlos Ramisch. 2018. The CUPT format for MWE annotation in Universal Dependencies. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Information Extraction (LAW-MWE-2018)*, pages 220–231.

Carlos Ramisch. 2025. PARSEME 2.0 shared task subtask 1: Detailed results. https://gitlab.com/parseme/sharedtask-data/-/blob/master/2.0/subtask1/Detailed_results.md.

Agata Savary and Carlos Ramisch. 2025. PARSEME 2.0 shared task guidelines.

Manon Scholivet, Takuya Nakamura, Agata Savary, Éric Bilinski, and Carlos Ramisch. 2025. Parseme 2.0 shared task on identification and paraphrasing of multiword expressions. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*.

A System Pipeline

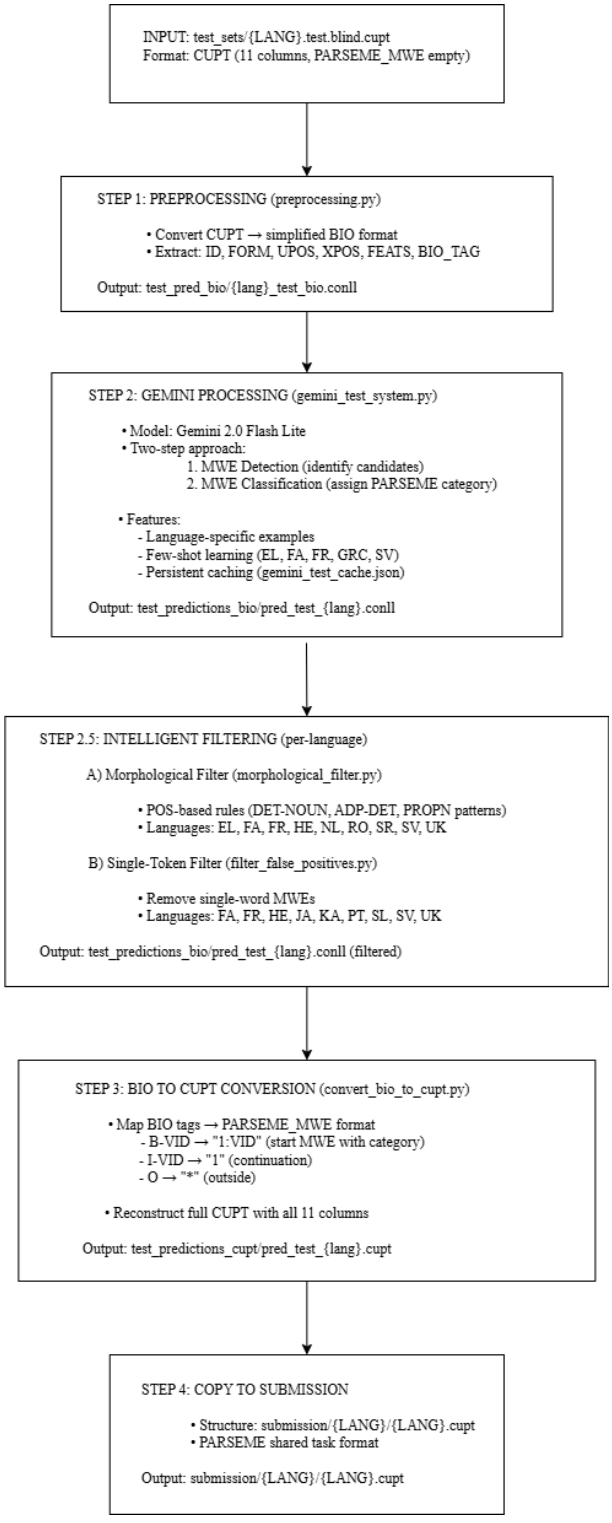


Figure 1: Detailed MWE Prediction Pipeline

B Prompting

B.1 Detailed MWE Detection Prompt

You are an expert annotator for the PARSEME shared task on multiword expression (MWE) identification. {strategy_note}

Task: Extract multiword expressions (MWEs) from sentences in {lang_info['name']}.

A MWE is a sequence of 2+ tokens that form a linguistic unit. Include:

- verbal MWEs
- nominal MWEs
- adjectival and adverbial MWEs
- functional MWEs

{few_shot_text}
{negative_examples}

CRITICAL SPAN RULES (follow exactly):

1. *Be MINIMAL: prefer SHORTER spans over longer ones.*
2. *For IAV: ONLY verb + preposition, NEVER include objects.*
   CORRECT: "depend on" | WRONG: "depend on him"
3. *For VPC: ONLY verb + particle, NEVER include objects.*
   CORRECT: "give up" | WRONG: "give up hope"
4. *DROP trailing punctuation ALWAYS.*
   CORRECT: "look up" | WRONG: "look up ."
5. *Use EXACT surface forms from input (copy verbatim).*

Format: [N]: MWE1 | MWE2 | NONE
Now analyze these NEW sentences:

Data Serialization and Batching Strategy: Sentences are processed in batches of 10 to ensure contextual consistency. Each sentence is serialized as a space-separated string of tokens, prefixed by a unique numerical identifier `[i]`. Then, the prompt concludes with a structural constraint:*"Output (one line per sentence, list only MWE surface forms separated by |)"*.

B.2 MWE Category Classification Prompt

Classify the category of this multiword expression (MWE) in {language}.

MWE: "{mwe_text}"

Categories (choose the MOST specific match):

VERBAL CATEGORIES (verb is present):

- VID: Verbal Idiom - idiomatic/figurative meaning
- LVC.full: Full Light Verb Construction - verb has little meaning, noun carries semantics
- LVC.cause: Causative Light Verb - causes state/action
- IAV: Inherently Adpositional Verb - verb REQUIRES preposition/particle (VERB+PREP only, 2-3 tokens)
- IRV: Inherently Reflexive Verb
- VPC.full/cause/semi: Verb-Particle Constructions
- MVC: Multi-Verb Construction

NON-VERBAL ID CATEGORIES (no verb):

- NID: Nominal Idiom

- AdvID: Adverbial Idiom (HINT: If single compound word functioning as adverb → AdvID)
- AdjID: Adjectival Idiom

FUNCTIONAL MWES:

- AdpID: Prepositional Idiom
- DetID: Determiner Idiom
- ConjID
- PronID, IntjID: Other idiom types

COMPOSITE (mix of categories):

- NV.VID: Nominal + Verbal Idiom
- AV.: Adverbial + Verbal constructions*

Respond with ONLY the category code (e.g., "VID", "LVC.full", "IAV", "AdvID"). Use lowercase for compound categories: "LVC.full" not "LVC.FULL".

B.3 Contextual Guidance: Few-Shot and Language Examples

To stabilize the LLM's predictions across diverse languages, the system injects language-specific examples and negative constraints into the prompt.

B.3.1 Positive Few-Shot Examples from Training Sets

For selected languages (SV, EL, FA, FR), the prompt includes full-sentence examples to illustrate span boundaries and category assignments. Examples:

Swedish (SV):
Sentence: *Disibodenbergklostret upplöstes och förföll i ruiner till följd av reformationen .*
MWEs: ***upplöstes | till följd av***
Categories: VPC.semi, AdpID

Portuguese (PT):*
Sentence: *A relatoria caiu com o ministro Gilmar Mendes , por meio de sorteio eletrônico .*
MWEs: ***por meio de***
Categories: AdpID
**Note: While PT did not use the full "improved" strategy in the final submission, it served as a development baseline.*

B.3.2 Negative Constraints (Anti-Examples)

To combat over-generation and the LLM's tendency to label compositional phrases as MWEs, the following explicit exclusions are included in the prompt:

IMPORTANT - NOT MWEs (do NOT extract these):

Simple noun phrases: "the book", "big house", "my friend"
Adjective + noun (compositional): "red car", "happy person"
Verb + full object: "read the book", "eat an apple"
Preposition + full noun phrase: "in the morning", "on the table"
Determiner + noun: "a cat", "the dog"

B.3.3 Language-Specific Prototype Examples

A small dictionary with a few examples for almost every language was created. For languages where full few-shot sentences were not used, the system provides prototypical MWE examples from that dictionary:

Dutch (NL):
"plaats vinden" (to take place) – light verb construction
"van tevoren" (beforehand) – fixed adverbial
"in orde" (in order) – fixed expression
"op de hoogte" (informed) – idiomatic phrase

Romanian (RO):
"de asemenea" (also/furthermore) – fixed adverbial
"în sfârșit" (finally) – temporal idiomatic expression
"cu toate că" (although) – compound conjunction
"a avea loc" (to take place) – verb with participle
"pe de altă parte" (on the other hand) – adverbial phrase

C Universal POS Filtering Patterns

The following POS-based constraints are applied to the LLM outputs to eliminate sequences that are unlikely to be Multiword Expressions. These rules prioritize precision by filtering common compositional or functional patterns:

`(DET, NOUN / ADJ / VERB):` Filters standard noun phrases and nominalized adjectives (e.g., *"the house"*).

`(ADJ, NOUN / NOUN, ADJ):` Eliminates simple compositional adjective-noun combinations unless previously seen as lexicalized.

`(DET, NUM / NUM, DET):` Removes articles combined with cardinal numbers.

`(PRON, VERB / VERB, PRON):` Filters subject-verb or verb-object pairs lacking idiomatic or reflexive properties.

`(CCONJ, NOUN / VERB):` Prunes spans starting or ending with coordinating conjunctions.

`(PUNCT, ANY / ANY, PUNCT):` Removes spans containing leading or trailing punctuation noise.

`(AUX, NOUN / ADJ):` Eliminates copular constructions (e.g., *"is good"*) that do not constitute MWEs.

D High-Precision Preservation Patterns

To maintain recall for core MWE categories, the filter is configured to bypass ("whitelist") the following high-confidence linguistic structures:

`(ADP, NOUN) and (ADP, DET, NOUN)` : Reliable prepositional idioms.

`(NOUN, ADP, NOUN)` : Common nominal compounds (e.g., *"horário de folga"*).

`(VERB, NOUN / ADP)` : Core verbal MWE types such as Light Verb Constructions (LVCs) and Inherently Adpositional Verbs (IAVs).

`(ADV, ADV / ADP)` : Multi-word adverbs and compound prepositions.

E Structural Pruning Rules

In addition to POS tagging, the system enforces three rigid structural constraints to refine the span boundaries:

1. **Length Check**: Predictions with fewer than 2 tokens are discarded (unless cached).
2. **Punctuation Only**: Spans consisting solely of non-alphanumeric characters are removed.
3. **Lexical Density**: Spans composed entirely of function words (`DET`, `PRON`) are eliminated.

F Ablation Study: Impact of Morphological Filtering

To quantify the impact of the POS-based constraints and single-token filters, we conducted an ablation study the development data at some point. We have since fixed some logic errors, therefore some languages who had bad scores are now benefiting from morphological filtering and vice versa. We compare the *Baseline LLM Pipeline* (prompting only) against the *Filtered Pipeline* (LLM + Morphological Post-processing).

	LLM		Filtered Pipeline	
Lang	**MWE-F1**	**Token-F1**	**MWE-F1**	**Token-F1**
EGY	0.1053	0.1026	0.1053	0.1026
EL	0.1452	0.3183	**0.2222**	0.3170
FA	0.1831	0.2898	**0.3470**	**0.4855**
FR	0.1056	0.1794	0.0594	0.1283
HE	0.0101	0.0134	0.0058	0.0095
JA	0.0603	0.2700	0.0571	0.2657
KA	0.0362	0.0531	0.0285	0.0488
NL	0.1721	0.3048	**0.3636**	**0.4762**
PL	0.0451	0.0584	**0.0452**	**0.0586**
PT	0.0857	0.1500	**0.0896**	**0.1558**
RO	0.0149	0.0218	0.0146	0.0215
SL	0.0431	0.0849	0.0370	0.0819
SR	0.0382	0.0525	0.0382	0.0525
SV	0.1792	0.2992	**0.1821**	**0.3002**
UK	0.0795	0.1026	0.0794	0.1027

Table 3: Scores of the filtering strategy

Analysis of Results:
A delicate aspect of our final system configuration was deciding to apply morphological filtering even to languages with marginal performance decreases during the abation study. The peformance is highly sensitive to the outputs of the LLM for that particular run (identical prompts can yeld varying levels of noise across different executions. We believe linguistic stability provided by the filters is more valuable for overall system robustness than hyper-optimizing for a limited data set.

LST at MWE-2026 AdMIRe 2: Advancing Multimodal Idiomaticity Representation

Le QIU and Yu-Yin HSU and Emmanuele CHERSONI
The Hong Kong Polytechnic University
11 Yuk Choi Rd, Hung Hom, Hong Kong SAR

Abstract

This paper presents our methods for the AdMIRe 2.0 shared task, which addresses multilingual and multimodal idiom understanding. Our submission focuses on the text-only track. Specifically, we employ an ensemble of three large language models (LLMs) to directly perform the presented image ranking task. Each model independently produces a ranking of the candidate images, and we aggregate their outputs using a hard voting strategy to determine the final prediction. This ensemble learning framework, by leveraging the complementary strengths of different LLMs, provides a training-free and robust solution to the AdMIRe 2.0 task and places our method in the second position on the leaderboard.

1 Introduction

The AdMIRe 2.0 Shared Task (Arslan et al., 2026; Torunoğlu-Selamet et al., 2026) is an expanded continuation of its precedent, Subtask A of SemEval-2025 Task 1: AdMIRe – Advancing Multimodal Idiomaticity Representation (referred to as AdMIRe 1.0 for distinction). In AdMIRe 1.0, Subtask A is formulated as a static image ranking task (Pickard et al., 2025): Given a Potentially Idiomatic Expression (PIE), specifically a nominal compound (NC), its surrounding context sentence, and a set of five images each accompanied by a descriptive caption, the system is required to rank the images according to how accurately they depict the meaning of the NC in the provided context. A mono-modal track of the task allows participants to perform the ranking using only the textual captions. The images are not randomly generated; rather, each is deliberately associated with the PIE either figuratively or literally, with the fifth a distractor. A demonstration is provided in Figure 1. AdMIRe 1.0 covers two languages: English and Portuguese. Building upon this, AdMIRe 2.0 broadens the scope of the task by extending the dataset to a substantially larger set of 15 languages. Importantly, during the training phase, only the AdMIRe 1.0 data (in English and Portuguese) are available, and participants are not informed of which additional languages will appear in the evaluation. The specific test languages become known only when the test phase begins and the test data are released, with no labeled training data provided for those languages.

Based on the observations from the AdMIRe 1.0 system reports, where the top-performing submissions consistently relied on large language models (LLMs) — for instance, You et al. (2025) in the bimdoal track and Fan et al. (2025) in the text-only track —- we adopted a similar strategy for AdMIRe 2.0. For better performance, robutness and to reduce model-specific biases, we employed multiple LLMs and aggregated their predictions through a hard voting scheme, which formed our final submission for the text-only track.

Although our official submission focuses solely on the text-only setting — and can be viewed as a relatively direct, shortcut-style application of LLM capabilities — it nevertheless achieved a strong result, ranking second on the official leaderboard.

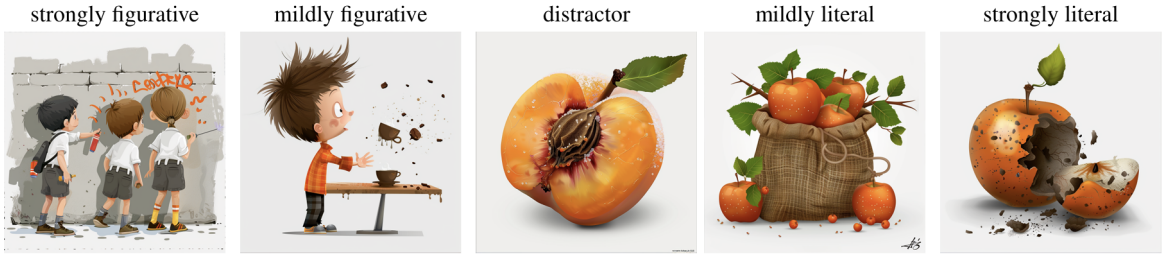
2 Methods and Results

Our method, which is entirely based on prompting, is motivated by both empirical observations from AdMIRe 1.0 and practical considerations specific to AdMIRe 2.0. Given the limited amount of available data[2], the restricted development time, and the fact that AdMIRe 2.0 spans a wide range of languages—including several low-resource ones such as Georgian, Igbo, Kazakh, and Uzbek — we aimed to approach the task in a training-free, zero-shot manner. This was one of the primary motivations for relying on LLMs.

[1]Example source: https://semeval2025-task1.github.io/

[2]Subtask A in AdMIRe 1.0 provides only 70 training instances in English and 32 in Portuguese.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 203–207
March 28, 2026



(a) The image depicts three children standing in front of a gray, textured wall...

(b) The image depicts a cartoon-style illustration of a young boy standing at a table...

(c) The image depicts a halved peach with a detailed and realistic appearance...

(d) The image depicts a rustic, burlap sack filled with several bright orange apples...

(e) The image depicts an orange-colored apple that appears to be decomposing or decaying...

Figure 1: Illustration of the Static Image Ranking Task. Consider the NC *bad apple* as in *We have to recognize that this is not the occasional bad apple but a structural, sector-wide problem*, the expected ordering should be (a), (b), (e), (d), and (c).[1]

Our final submission is therefore based entirely on prompting and ensemble learning: we directly provided the textual captions to each model, asked them to perform the ranking, and then combined the predictions from the three LLMs using a hard voting strategy. The captions are cleaned, re-indexed, and concatenated with the prompt prefix below before being fed into the chatbots.

> *Consider the expression [NC] in the [LANG] language, and its meaning in context [SENT]. Rank the following captions based on how well they reflect that meaning. Return only a list of reordered indices, e.g., [1, 3, 2, 4, 5], from most to least similar. Do not include any explanation.*

We found that explicitly indicating the language category in the prompt (i.e., [*LANG*]) has the potential to improve accuracy. This is probably because LLMs internally maintain separate multilingual semantic subspaces. Providing the language label helps the model activate the correct linguistic and cultural knowledge and constrains the reasoning space, thereby improving zero-shot predictions.

Our ensemble consists of three LLMs — GPT-4o (Achiam et al., 2023), Qwen-Plus (Team, 2025), and DeepSeek (Liu et al., 2024). The first two models were selected in that we conducted preliminary experiments on AdMIRe 1.0 Subtask A using a similar prompting strategy and found that their performance was close to that of the top systems reported (see Table 1). A third model was included to facilitates a stable majority-vote ensemble. Also, these models have been referred to in AdMIRe 1.0 system reports, are currently among the most capable and popular multilingual LLMs, and, due to their scale and training diversity, are more likely to provide broad language coverage — probably including for the low-resource languages present in AdMIRe 2.0. The results across all languages are presented in Table 2, which place us in the second position on the leaderboard.

Also, we explored whether incorporating type information (see Table 1 for reference) could provide additional benefits to the final results. To this, we adopted a simple bubble-sorting procedure. Since a PIE used idiomatically in the given context has the opposite usage type of literal, and vice versawe first perform pairwise comparisons to identify the caption most similar to the target usage, followed by the second most likely one. We then conduct another round of pairwise comparisons to determine the caption most similar to the opposite usage and the second most likely one. The remaining caption is treated as the distractor, composing the final ranking.

Due to resource constraints, we only experimented on the Chinese subset using a prompt prefix below:

> *The expression [NC] can be used idiomatically or literally in the LANG language. You are provided with 2 captions, each describing an image that may or may not be related to this expression. Decide which caption most likely relate to*

		Test Set			Extended Evaluation Set		
		Top 1 Acc	DCG	Type Acc	Top 1 Acc	DCG	Type Acc
English	CTYUN-AI	0.64	3.10		0.87	3.51	
	DeepSeek	0.58	3.07	0.89	0.73	3.24	0.80
	GPT-4o	0.59	3.03	0.76	0.60	3.07	0.80
Portuguese	CTYUN-AI	0.92	3.43		0.56	2.97	
	DeepSeek	0.77	3.31	0.77	0.64	3.07	0.76
	GPT-4o	0.77	3.35	0.54	0.42	2.77	0.76

Table 1: Results on the test set and the extended evaluation set of AdMIRe 1.0 Subtask A (text-only track), compared with the best-performing team — CTYUN-AI. Although not required for submission, the AdMIRe tasks expect systems to predict the usage type of a given PIE in its context sentence (*idiomatic* vs. *literal*). Taking the instance in Figure 1 as an example, the system is expected to predict that the PIE *bad apple* in the given sentence is used idiomatically. We therefore also report the type prediction accuracy as *Type Acc* in addition to *Top 1 Acc* (Top Image Accuracy) and *DCG* (NDCG@5). All scores are reported on a scale of 1.

	Top 1 Acc	DCG
Chinese	0.36	0.74
Georgian	0.4	0.74
Greek	0.43	0.76
Igbo	0.33	0.71
Kazakh	0.42	0.76
Norwegian	0.43	0.77
Portuguese-Brazil	0.53	0.81
Portuguese-Portugal	0.45	0.77
Russian	0.51	0.79
Serbian	0.40	0.74
Slovak	0.44	0.78
Slovenian	0.45	0.78
Spanish-Ecuador	0.35	0.73
Turkish	0.40	0.74
Uzbek	0.32	0.73

Table 2: Official results on the test set of AdMIRe 2.0.

> *the [TYPE] meaning or use of this expression. Only output the chosen caption name, such as (A) or (B), do not include any analysis.*

The type information (i.e., *[TYPE]*) is also obtained using a hard-voting strategy. The results are shown in Table 3. Overall, incorporating type information as guidance leads to a slight decrease in performance compared with directly ranking captions using the LLMs (see Table 2). The Top Image Accuracy remains unchanged, while the DCG score drops by 0.01. A closer look at the type-specific scores reveals that performance on literal metrics decreases while on idiomatic metrics improves. This outcome is likely due to the fact that idiomatic usage tends to be more abstract linguistically. Providing explicit type information may therefore help the model identify relevant cues in the captions and align them with the intended usage, therefore improving the accuracy. In contrast, literal interpretations could depend more on visual grounding (again, the *bad apple* example), which is unavailable in the text-only track.

3 Related Work

Studies have shown that language models — ranging from basic BERT to larger generative models such as ChatGPT — continue to exhibit limitations in interpreting idiomatic expressions (IEs) (Shwartz and Dagan, 2019; Wu et al., 2024; Raunak et al., 2023).

Typical solutions to IE representation learns phrase embedding directly from contextual co-occurrence (Mikolov, 2013; Yin and Schütze, 2014, 2016). This is effective for frequent expressions but struggles with sparse IEs. Alternatively, compositional approaches derive phrase embeddings by combining the embeddings of individual components (Mitchell and Lapata, 2010; Yu and Dredze, 2015), but they often fail to capture the semantic opacity characteristic of IEs. More recent work leverages pre-trained language models (PLMs) for IE representation through adaptive modules and contrastive learning, such as Zeng and Bhat (2022); He et al. (2024); Wu et al. (2024). It has also been found that external knowledge, such as synonyms and definitions can enhance model performance (Long et al., 2020; Wang et al., 2020;

	Top 1 Acc	DCG	Literal Acc	Idiomatic Acc	Literal DCG	Idiomatic DCG
1	0.36	0.74	0.46	0.29	0.78	0.70
2	0.36 (0)	0.73 (-0.01)	0.41 (-0.05)	0.33 (+0.04)	0.75 (-0.03)	0.87 (+0.17)

Table 3: Results on the Chinese subset. The first row reports the ranking results obtained directly from LLMs, and the second row presents the results produced by incorporating type information through pairwise comparison. In addition to the overall Top 1 ACC and DCG scores, we also report type-specific performance (*literal* vs. *idiomatic*). The values in parentheses in the second row indicate the increments relative to the first row.

Sha et al., 2023).

Such modelling stratgies have also been noticed in the AdMIRe shared task (1.0), although participation in the text-only track has been relatively limited, likely because images provide stronger cues for the ranking task. Nonetheless, Petersen et al. (2025) fine-tuned SBERT for the task and augmented captions using GPT-4-generated descriptions. The top-performing team (Fan et al., 2025) also applied extensive data augmentation — including synonym substitution and back-translation — for the multilingual setting, and fine-tuned Qwen (Team, 2025) models for the task.

4 Conclusion

This report presents our work on the ADMIRE 2.0 task under the text-only setting. We adopt an ensemble learning framework for the caption ranking task, which places our methods in the second position on the leaderboard. Nevertheless, the results clearly show that ADMIRE 2.0 is substantially more challenging than ADMIRE 1.0, with noticeably lower scores. Even strong LLMs struggle to reliably distinguish literal from idiomatic representations.

Limitations

First, we experimented with only a small subset of available LLMs, which restricts the breadth of our evaluation. Second, our approach relies solely on zero-shot inference combined with a hard-voting ensemble, without exploring more sophisticated or innovative modeling strategies. Due to time and resource constraints, we were unable to investigate alternative architectures, training paradigms, or more creative methods that might further improve performance. We leave these directions for future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Doğukan Arslan, Rodrigo Wilkens, Wei He, Dilara Torunoğlu-Selamet, Thomas Pickard, Aline Villavicencio, Adriana S. Pagano, and Gülşen Eryiğit. 2026. MWE-2026 Shared Task 2: AdMIRe 2 - Advancing Multimodal Idiomaticity Representation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morocco. Association for Computational Linguistics.

Yuming Fan, Dongming Yang, Zefeng Cai, and Binghuai Lin. 2025. CTYUN-AI at SemEval-2025 task 1: Learning to rank for idiomatic expressions. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 16–19, Vienna, Austria. Association for Computational Linguistics.

Wei He, Marco Idiart, Carolina Scarton, and Aline Villavicencio. 2024. Enhancing idiomatic representation in multiple languages via an adaptive contrastive triplet loss. *arXiv preprint arXiv:2406.15175*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Siyu Long, Ran Wang, Kun Tao, Jiali Zeng, and Xinyu Dai. 2020. Synonym knowledge enhanced reader for chinese idiom reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3684–3695.

Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Wiebke Petersen, Lara Eulenpesch, Ann Piho, Julio Julio, and Victoria Lohner. 2025. Transformer25 at SemEval-2025 task 1: A similarity-based approach. In *Proceedings of the 19th International Workshop on*

Semantic Evaluation (SemEval-2025), pages 2311–2317, 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.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023. Do GPTs produce less literal translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada. Association for Computational Linguistics.

Ying Sha, Mingmin Wu, Zhi Zeng, Xing Ge, Zhongqiang Huang, and Huan Wang. 2023. A prompt-based representation individual enhancement method for chinese idiom reading comprehension. In *International Conference on Database Systems for Advanced Applications*, pages 682–698. Springer.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Dilara Torunoğlu-Selamet, Dogukan Arslan, Rodrigo Wilkens, Wei He, Doruk Eryiğit, Thomas Pickard, Adriana S. Pagano, Aline Villavicencio, Gülşen Eryiğit, Ágnes Abuczki, Aida Cardoso, Alesia Lazarenka, Dina Almassova, Amalia Mendes, Anna Kanellopoulou, Antoni Brosa-Rodríguez, Baiba Saulite, Beata Wojtowicz, Bolette Pedersen, and 59 others. 2026. A parallel cross-lingual benchmark for multimodal idiomaticity understanding. *Preprint*, arXiv:2601.08645.

Xinyu Wang, Hongsheng Zhao, Tan Yang, and Hongbo Wang. 2020. Correcting the misuse: A method for the chinese idiom cloze test. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 1–10.

Mingmin Wu, Yuxue Hu, Yongcheng Zhang, Zeng Zhi, Guixin Su, and Ying Sha. 2024. Mitigating idiom inconsistency: A multi-semantic contrastive learning method for chinese idiom reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19243–19251.

Wenpeng Yin and Hinrich Schütze. 2014. An exploration of embeddings for generalized phrases. In *Proceedings of the ACL 2014 Student Research Workshop*, pages 41–47.

Wenpeng Yin and Hinrich Schütze. 2016. Discriminative phrase embedding for paraphrase identification. *arXiv preprint arXiv:1604.00503*.

Runyang You, Xinyue Mei, and Mengyuan Zhou. 2025. PALI-NLP at SemEval 2025 task 1: Multimodal idiom recognition and alignment. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1211–1216, Vienna, Austria. Association for Computational Linguistics.

Mo Yu and Mark Dredze. 2015. Learning composition models for phrase embeddings. *Transactions of the Association for Computational Linguistics*, 3:227–242.

Ziheng Zeng and Suma Bhat. 2022. Getting bart to ride the idiomatic train: Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics*, 10:1120–1137.

UniBO at MWE-2026 PARSEME 2.0 Subtask 2: A Cross-lingual Approach to Multiword Expression Paraphrasing

Debora Ciminari and Alberto Barrón-Cedeño
DIT, University of Bologna
{debora.ciminari2 , a.barron}@unibo.it

Abstract

This paper describes MISP (Multilingual Idiomatic Sentence Paraphrasing), a system submitted to the PARSEME 2.0 Multilingual Shared Task on Identification and Paraphrasing of Multiword Expressions (MWEs). We participated in Subtask 2 on MWE paraphrasing and developed our system based on Qwen3-4B-Instruct fine-tuned on synthetic Portuguese MWE paraphrases. We applied MISP not only to Portuguese, but also to French and Romanian, aiming to leverage cross-lingual transfer within related languages, with ours being the only submission for Portuguese. Our results indicate that MISP struggles to generate paraphrases that both rephrase and preserve the original meaning of the MWE. Additionally, instruction fine-tuning does not appear to improve performance. Overall, our findings highlight the challenges of paraphrasing MWEs, particularly in a cross-lingual setting.[1]

1 Introduction

Multiword expressions (MWEs) are a major area of interest within the field of natural language processing (NLP). Different definitions have been proposed that emphasise either their formulaic nature (Wray, 2002) or their treatment as units rather than as sequences of individual words (Calzolari et al., 2002; Sag et al., 2002). Other definitions highlight the idiosyncratic nature of MWEs on different levels (Baldwin and Kim, 2010; Calzolari et al., 2002). We adhere to the definition provided by Baldwin and Kim (2010, p. 269), which is the one adopted by the PARSEME 2.0 Multilingual Shared Task on Identification and Paraphrasing of Multiword Expressions. MWEs are described as "lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity". This definition does not equal idiomaticity to non-compositionality (as traditionally argued), but to "markedness" or "deviation" from the features of the MWE's components.

MWEs constitute a large part of language and are estimated to be comparable in number to single words in a speaker's lexicon (Jackendoff, 1997). They also come in different forms and exhibit a high level of heterogeneity. MWEs can be placed on an idiomaticity continuum, from highly idiomatic to more compositional expressions (Moon, 1998), as well as on a fixedness continuum, spanning from fixed expressions to flexible combinations (Sag et al., 2002). Additionally, MWEs are attested in a wide range of languages, in which they are constructed differently (Baldwin et al., 2004).

The processing of MWEs is central to the development of more linguistically precise NLP systems. Efficiently handling MWEs can have great utility for various downstream tasks, such as parsing, machine translation, information extraction and sentiment analysis (Mititelu et al., 2025).

The PARSEME 2.0 Shared Task[2] (Scholivet et al., 2026), organised within the "Universality, diversity and idiosyncrasy in language technology" (UniDive) CA21167 COST Action[3], focuses on the identification (Subtask 1) and paraphrasing (Subtask 2) of MWEs. Both subtasks adopt a multilingual perspective and involve 17 and 14 languages, respectively.

We focused on Subtask 2 and developed MISP (Multilingual Idiomatic Sentence Paraphrasing), a system for MWE paraphrasing in French (fr), Georgian (ka), Portuguese (pt), and Romanian (ro). For this task, based on a sentence containing a MWE, the system should generate a paraphrase that (*i*) completely or partially removes the MWE,

[1] The implementation and dataset of MISP are available at https://github.com/TinfFoil/unibo-misp/.

[2] https://unidive.lisn.upsaclay.fr/doku.php?id=other-events%3Aparseme-admire-st-call#call_for_participation

[3] https://unidive.lisn.upsaclay.fr

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 208–216

(*ii*) preserves the original meaning, and (*iii*) exhibits linguistic diversity. In this paper, we describe our approach for three Romance languages —French, Portuguese, and Romanian— based on cross-lingual transfer. We selected languages from the same language subgroup to assess whether their linguistic kinship could be leveraged to benefit cross-lingual transfer. We instruction-tuned `Qwen3-4B-Instruct-2507` (Qwen Team, 2025)[4] on a dataset of synthetically generated paraphrases in Portuguese created from the *AStitchInLanguageModels* dataset (Tayyar Madabushi et al., 2021). The results, presented in Section 4, suggest that `MISP` is not capable of generating paraphrases that rephrase the MWE of the original sentence and that retain its original meaning. Instruction fine-tuning does not necessarily improve the model's performance, and the effects of cross-lingual transfer appear limited or even negative.

The rest of the paper is distributed as follows. Section 2 provides an overview of previous studies and existing datasets for MWE paraphrasing. Section 3 describes the experimental setup, from data creation to fine-tuning. Then, Section 4 presents the results and discussion of the experiments, while Section 5 draws some conclusions and underscore possible future developments.

2 Related Work

NLP applications have been focusing on two major issues associated with MWEs: identification and interpretation (Baldwin and Kim, 2010). These issues are related to the idiomaticity or markedness characterising MWEs, which calls for the need to disentangle their intrinsic ambiguity, especially on the semantic and the syntactic levels. Multilinguality adds another layer of complexity since languages license different ways of creating MWEs, and the criteria to define them can vary based on language (Villavicencio et al., 2004). As exemplified by Baldwin et al. (2004), while in English only expressions made of multiple whitespace-delimited lexemes are considered, in German this criterion is more flexible, since the language is rich in compound nouns. While multilinguality introduces a great degree of variability, it can also serve as an aid for identifying MWEs by examining translations between languages (Melamed, 1997). For instance, the lexemes *balance* and *sheet* are most often translated as *équilibre* and *feuille* in French, while the expression *balance sheet* is translated as *bilan*, which might indicate that *balance sheet* is semantically idiomatic (Baldwin and Kim, 2010).

Since MWEs constitute a major interest in NLP, various tasks have been explored. For example, Zhou et al. (2021a) introduce the task of idiomatic sentence paraphrasing (ISP), consisting in paraphrasing a sentence containing an idiomatic expression by replacing them with their literal meaning. As they point out, ISP aims to improve the readability of an idiomatic sentence, possibly having benefits on tasks such as machine translation. For this purpose, they release the Parallel Idiomatic Expression (PIE) dataset, which includes sentence-level mappings between idiomatic sentences and their paraphrases, and BIO tags to signal the MWE. PIE is the first parallel dataset of idiomatic sentence paraphrases and represents a significant contribution to the study of MWEs in NLP. However, datasets to address ISP remain relatively scarce, even more so for other languages.

Tayyar Madabushi et al. (2021) release an annotated idiom-related dataset in English and Portuguese, which is further extended with Galician in the SemEval2022 Task 2 dataset (Tayyar Madabushi et al., 2022). Given the scarcity of data, studies have attempted to develop unsupervised or weakly supervised systems, such as those by Zhou et al. (2021b), who propose an unsupervised system that leverages the meaning and the POS tags of an idiomatic expression to generate an accurate paraphrase of the whole sentence. Yet this system is outperformed by their weakly supervised method based on a high-quality parallel dataset and which draws on back translation.

Considering the limited data for ISP, we developed a system based on the synthetic generation of paraphrases from existing data in Portuguese, and we attempted to exploit cross-lingual transfer between Portuguese, French, and Romanian to leverage their common genealogical root.

3 Experiments

3.1 Creation of Synthetic Data in Portuguese

To carry out the instruction fine-tuning, we first needed paraphrases of sentences containing MWEs. To this end, we used *AStitchinLanguageModels* (Tayyar Madabushi et al., 2021), a dataset of naturally occurring sentences containing potentially idiomatic MWEs in both English and Portuguese.

[4] `https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507`

This dataset provides different annotations of the MWEs, including whether their usage is literal or idiomatic and a paraphrase for both meanings. For example, the expression `big fish` is associated with the paraphrases "large fish" (literal) and "important person" (idiomatic). We used the Portuguese subset and extracted (*i*) the MWEs, (*ii*) the sentences with an idiomatic usage of MWEs, and (*iii*) the corresponding paraphrases.

Based on these, we synthetically generated three paraphrases of each sentence through prompts entirely written in Portuguese, following existing findings suggesting that instructions in English do not necessarily yield better results (e.g., Enomoto et al., 2025; Phelps et al., 2024). We used `Apertus-8B-Instruct-2509` (Hernández-Cano et al., 2025)[5] for the generation. Appendix A shows an example of the prompts and the paraphrases generated by the model. The prompt is designed to provide the model with the meaning of the MWE extracted from the dataset: in the example from Appendix A, the MWE `mercado negro` (black market) is defined as `mercado ilegal` (illegal market), which gives the model extra information about the meaning of the MWE. Additionally, the prompt comprises practical guidelines to generate a paraphrase to help the model better follow the instruction. Since the model is asked to follow a specific output format, we were able to extract the paraphrase from the whole model's response by using regular expressions.

We also attempted to filter the data based on the BERTScore (Zhang et al., 2019) between original and the paraphrased sentence. Synthetic paraphrases having a BERTScore lower than 0.7 were filtered out. However, since the filtered dataset exhibited a significant drop in terms of diversity, the filtering stage was omitted.

Following this methodology, we created a dataset of 3,537 synthetic paraphrases in Portuguese, which we used in the fine-tuning.

3.2 Instruction Fine-Tuning

Based on these generated data, we conducted the instruction fine-tuning on `Qwen3-4B-Instruct`. We used QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al., 2023), which combines 4-bit quantization with LoRA (Hu et al., 2021) to save computational resources while maintaining model performance. The training process runs for 5 epochs using a batch size of 4 and the 8-bit paged AdamW optimizer.

4 Results and Discussion

After fine-tuning, we applied the model to the test data in Portuguese, French and Romanian. The test sets comprise 158 sentences in Portuguese, 95 in French and 138 in Romanian. We used prompts entirely written in the target language, consistently with the approach adopted during fine-tuning.

4.1 Automatic Evaluation

We carried out the automatic evaluation of our system following the shared task's approach. First, BERTScore is computed only for the predictions completely or partially deleting the MWE. This score is designed to assess the semantic similarity between such predictions and two manually crafted paraphrases: a "minimal" one, which is closer to the source sentence, and a "creative" one, which greatly differs. Out of these two scores, the higher one is chosen as the final BERTScore. BERTScore is not computed on the original sentence, but on gold standard paraphrases, since it might struggle to capture the original idiomatic, non-compositional semantics and give inaccurate scores. Besides semantic similarity, the linguistic diversity of the system's paraphrasing is assessed through three metrics, namely richness, Shannon evenness, and Shannon–Weaver entropy (Shannon, 1948). Richness measures how varied the system's linguistic choices are, indicating the diversity of its vocabulary for expressing similar meanings, without resorting to the same patterns. Shannon evenness captures how balanced these choices are in the system's output, ensuring that all are equally opted for. Finally, Shannon–Weaver entropy is computed to measure the unpredictability of the system's paraphrasing behaviour.

Table 1 shows the obtained results. Baseline results are generated with `gpt-oss-120b`[6] (OpenAI, 2025), a 117-billion parameter open-weight model released by OpenAI in 2025. `Qwen3-4B-Instruct` results reflect the performance of the model prior to fine-tuning, whereas `MISP` indicates the performance of the fine-tuned model. For comparison, we include the scores obtained by the top participants of the shared task for French and Romanian (ours is the only run submitted for Portuguese).

[5] https://huggingface.co/swiss-ai/Apertus-8B-Instruct-2509

[6] https://huggingface.co/openai/gpt-oss-120b

	BERTScore	Richness	Evenness	Entropy
French				
gpt-oss-120b	77.55	326	0.92	5.33
Top PARSEME	**93.90**	236	0.83	4.54
Qwen3-4B-Instruct	48.31	**630**	0.91	5.88
MISP	49.53	564	**0.93**	**5.89**
Portuguese				
gpt-oss-120b	**80.21**	619	0.92	5.93
Top PARSEME*	–	–	–	–
Qwen3-4B-Instruct	66.16	**1,048**	0.91	**6.34**
MISP	58.59	789	**0.93**	6.20
Romanian				
gpt-oss-120b	74.74	742	0.93	6.14
Top PARSEME	**89.25**	235	**0.98**	5.36
Qwen3-4B-Instruct	67.68	**1,148**	0.91	**6.42**
MISP	57.01	1,096	0.91	6.36

* No other team submitted runs for Portuguese

Table 1: BERTScores, richness, evenness, and entropy for baseline, Qwen3-4B-Instruct, and MISP in French, Portuguese, and Romanian. For comparison, the scores obtained by the top participant of the shared task are included. Best values in bold.

As far as the BERTScore is concerned, MISP is generally outperformed by the other models. Our model fails to approximate the baseline across all three languages. While for Portuguese and Romanian the difference reaches approximately 20 points, for French the gap between baseline and MISP is nearly 30 points. This gap widens when considering the top participant's scores for French and Romanian: 93.90 and 89.25, respectively. This suggests that these models are substantially more capable of generating paraphrases that retain the meaning of the original sentence. By contrast, MISP appears to struggle to preserve the sense of the original sentence across all three languages. Its highest BERTScore is observed for Portuguese (58.59) but is still relatively low.

Focusing on the effectiveness of instruction fine-tuning, the findings show that for Portuguese and Romanian fine-tuning does not lead to significant improvements but instead results in poorer performance, with BERTScores approximately 10 points lower. This suggests that instruction fine-tuning not only degraded performance in Romanian but also in Portuguese, the language used for fine-tuning. Further, the results might indicate that cross-lingual transfer between these two languages did not occur or had a negative impact on the model's performance. For French, fine-tuning does not lead to a deterioration in performance and yields a BERTScore that is approximately 1-point higher. Although this improvement is modest, it may indicate that fine-tuning on Portuguese had a small beneficial effect on French data, suggesting that cross-lingual transfer might have taken place at some degree.

When interpreting these results, however, we need to consider that gpt-oss-120b is much larger in size than Qwen3-4B-Instruct, which might partially explain the difference in performance. Additionally, our fine-tuning approach presents two notable limitations. First, it relies on synthetically generated data, which might be noisy and of variable quality. Second, the size of the training dataset is relatively small.

As for linguistic diversity, the baseline is generally outperformed by both MISP and the other models across all languages and metrics. The most notable difference is observed for richness, which reaches a value of 1,148 for Romanian with Qwen3-4B-Instruct. Additionally, both Qwen3-4B-Instruct and MISP outperform the top participant's models across all three metrics (with the exception of "richness" for Romanian). The results suggest that both Qwen3-4B-Instruct and MISP employ a more diverse and balanced vocabulary when generating paraphrases and avoid using the same repetitive patterns. However, it should be noted that BERTScores remain relatively modest,

and a manual evaluation would be necessary to determine if the increased linguistic diversity reflects correct and appropriate paraphrases. The findings in Table 1 might indicate that striking a balance between meaning preservation and diversity is a true challenge.

4.2 Error Analysis on French

In addition, we conducted an error analysis of the paraphrases generated by MISP for French. We followed the scheme adopted by the shared task for the manual evaluation of the submitted systems' predictions. Five criteria are evaluated. (*i*) Elimination indicates whether the MWE has been partially or completely removed and replaced by a paraphrase. (*ii*) Meaning assesses whether the original meaning is preserved. (*iii*) Grammaticality checks for spelling or grammar errors. (*iv*) Grammatical closeness evaluates whether grammatical features such as time, mood, and number are retained. (*v*) Other includes additional aspects such as naturalness and register. Criteria (*ii*) to (*v*) are only assessed for paraphrases where the MWE has been eliminated.

We compared 95 paraphrases in French against the original sentences. The error analysis shows that the model managed to completely or partially remove the MWE in only 39 out of 95 cases. Among these 39 cases, the meaning is preserved in 21 sentences (54%), while in others the model fails to retain it. As for grammaticality, correctness is preserved in 37 paraphrases, and grammatical closeness is retained in 35 cases. Finally, in 3 cases the paraphrases are correct both in meaning and grammar, but lack naturalness. Some examples of each error type are provided in Table 2, while the corresponding translations into English and some statistics are reported in Appendix B.

The error analysis suggests that MISP struggles to paraphrase the MWE contained in the original sentence and retains it in the majority of cases. Similarly, it often fails to generate paraphrases that maintain the meaning of the original sentence. In contrast, only minor issues emerge regarding language quality.

	Elimination – MWE not deleted
O	Une déviation en ***voie rapide*** a été inaugurée en 1993.
P	Une modification sur la ***voie rapide*** a été mise en place en 1993
	Meaning – original meaning not preserved
O	Effets négatifs de la ***rétention d'information***
P	Les effets néfastes de la ***conservation d'information***
	Grammaticality – contains a grammar error
O	Après quelques années passées à Ringerike, ils s'installent dans le quartier Nordstrand d'Oslo, où ils reprennent un *bed and breakfast*.
P	Après quelques années de résidence à Ringerike, ils s'installent dans le quartier Nordstrand d'Oslo, où ils **reprendent** une *auberge de jeunesse*
	Grammatical Closeness – change in mood, time, number
O	Nelson Downs n'a jamais *pris* de *cours* de magie, il était autodidacte et très jeune il **développa** une remarquable habileté pour la manipulation de cartes et de pièces.
P	Nelson Downs n'a jamais *suivi* de *cours* de magie, il était autodidacte et, très jeune, il **a développé** une remarquable capacité à manipuler des cartes et des pièces
	Other – issues with naturalness, register, other
O	Entre 2005 et 2007, il joue pour l'*équipe réserve* du Saturn Ramenskoïe en troisième division.
P	De 2005 à 2007, il **participe** à l'*équipe de réserve* du Saturn Ramenskoïe, qui compétitionne en troisième division

Table 2: Examples of **O**riginal and **P**araphrase evaluation errors in different categories. MWEs are shown in italics, while errors are highlighted in bold.

5 Conclusions

In this paper, we described MISP, a multilingual system for MWE paraphrasing submitted to Subtask 2 of the PARSEME 2.0 Shared Task. We reported the findings related to three Romance languages: French, Portuguese, and Romanian. Our analysis suggests that the baseline from gpt-oss-120b achieves significantly higher BERTScores across all languages, whereas MISP often fails to preserve the original meaning. Instruction fine-tuning on Portuguese and Romanian did not yield improvements, and in some cases it led to degraded performance, suggesting limited or negative cross-lingual transfer. However, MISP's performance on French appears slightly improved compared to Qwen3-4B-Instruct, which may indicate that fine-tuning was beneficial and that some degree of cross-lingual transfer took place and helped the model better perform on French data. On the other hand, Qwen3-4B-Instruct and its fine-tuned version, MISP, appear more able to generate paraphrases that exhibit higher linguistic diversity.

Error analysis in French shows that MISP often fails to rephrase MWEs and confirms that it struggles to maintain meaning, although grammatical correctness and closeness are generally preserved. Overall, our findings highlight the intrinsic difficulty of paraphrasing MWEs across languages and underscore the need for task-specific multilingual data.

Limitations

Our study has limitations that might account for the observed results. First, the fine-tuning dataset is relatively small, which may not help the model learn effective paraphrasing in a significant way. Second, the dataset used for the fine-tuning is synthetically generated, introducing potential noise and variability that could have negatively affected the model's performance. As for the error analysis for French, it was conducted by a non-native speaker, which may have reduced the reliability of some judgments.

Ethics Statement

In this paper, we deliberately employ two relatively small language models (8B and 4B parameters). This choice was partly driven by practical constraints, such as available hardware resources. Additionally, the use of smaller models reflects an ethical consideration regarding the more resource efficiency of such models, with the aim of supporting more sustainable as well as reproducible NLP practices.

References

Timothy Baldwin, Emily M. Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. 2004. Road-testing the English Resource Grammar over the British National Corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing*. CRC Press, Taylor and Francis Group.

Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Taisei Enomoto, Hwichan Kim, Zhousi Chen, and Mamoru Komachi. 2025. A fair comparison without translationese: English vs. target-language instructions for multilingual LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 649–670, Albuquerque, New Mexico. Association for Computational Linguistics.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, and 82 others. 2025. Apertus: Democratizing Open and Compliant LLMs for Global Language Environments. https://arxiv.org/abs/2509.14233.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Ray Jackendoff. 1997. Twistin' the Night Away. *Language*, 73(3):534.

I. Dan Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. In *Second Conference on Empirical Methods in Natural Language Processing*.

Verginica Barbu Mititelu, Voula Giouli, Gražina Korvel, Chaya Liebeskind, Irina Lobzhanidze, Rusudan Makhachashvili, Stella Markantonatou, Alexandra Markovic, and Ivelina Stoyanova. 2025. The challenges of syntactic descriptions of multiword expressions in electronic lexicography. In *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography. Proceedings of the eLex 2025 conference.*

R. Moon. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford Studies in Lexicography and Lexicology. Clarendon Press.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Dylan Phelps, Thomas Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. *Preprint*, arXiv:2405.09279.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In Gerhard Goos, Juris Hartmanis, Jan Van Leeuwen, and Alexander Gelbukh, editors, *Computational Linguistics and Intelligent Text Processing*, volume 2276, pages 1–15. Springer Berlin Heidelberg, Berlin, Heidelberg.

Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Carp, and Vasile Pais. 2026. Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

C. E. Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(4):623–656.

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.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aline Villavicencio, Timothy Baldwin, and Benjamin Waldron. 2004. A multilingual database of idioms. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Alison Wray. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021a. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48, Online. Association for Computational Linguistics.

Jianing Zhou, Ziheng Zeng, Hongyu Gong, and Suma Bhat. 2021b. Idiomatic expression paraphrasing without strong supervision. *CoRR*, abs/2112.08592.

A Prompt for the Synthetic Data Generation of Data and Prompt for Instruction Fine-tuning

Table 3 shows an example of the prompt used to synthetically generate paraphrases in Portuguese, while Table 4 illustrates the prompt used for the instruction fine-tuning.

Prompt
Instruções: Você é falante nativo de português brasileiro. A frase a seguir contém a expressão composta "mercado negro", que significa "mercado ilegal". Elabore 3 reformulações da frase seguindo os passos abaixo: - Leia a frase atentamente e certifique-se de compreender todo o seu significado. - Escreva a ideia principal da frase com suas próprias palavras: use sinônimos, altere a estrutura da frase e a estrutura gramatical. - Compare sua reformulação com a frase original para garantir que o significado foi preservado. Formate sua resposta desta maneira: As paráfrases são: 1) 2) 3) **### Frase:** Os medicamentos estão sendo vendidos no mercado negro e os resultados dos testes estão demorando dias. **### Resposta:**
Synthetically Generated Paraphrases
1) Os medicamentos estão sendo comercializados ilegalmente e os resultados dos exames estão levando dias para serem divulgados. 2) Medicamentos estão sendo vendidos no mercado ilegal, e os resultados dos testes estão demorando dias para serem apresentados. 3) Os medicamentos estão sendo vendidos no mercado clandestino e os resultados dos exames estão levando dias para serem divulgados.

Table 3: Example of prompt in Portuguese and the three paraphrases generated by the model according to the instructions.

Prompt
Instruções: Você é falante nativo de português brasileiro. A frase a seguir contém a expressão composta "{idiom}". Elabore uma reformulação da frase seguindo os passos abaixo: - Leia a frase atentamente e certifique-se de compreender todo o seu significado. - Escreva a ideia principal da frase com suas próprias palavras: use sinônimos, altere a estrutura da frase e a estrutura gramatical. - Compare sua reformulação com a frase original para garantir que o significado foi preservado. Formate sua resposta desta maneira: A paráfrase é: 1) **### Frase:** {sentence}. **### Resposta:** 1) {paraphrase}

Table 4: Prompt in Portuguese for the instruction fine-tuning.

B Results of Error Analysis

Table 5 shows the English translations of the error examples reported in Section 4.2. Table 6 shows the statistics related to the error types occurring in MISP generated paraphrases.

Elimination – the MWE was not deleted	
O	A deviation to the **fast track** was inaugurated in 1993.
P	A change to the ***fast track*** was implemented in 1993
Meaning – the original meaning was not preserved	
O	Negative effects of ***information retention***.
P	The adverse effects of ***information conservation***
Grammaticality – the paraphrase contains a grammar error	
O	After a few years in Ringerike, they moved to the Nordstrand district of Oslo, where they took over a *bed and breakfast*.
P	After living in Ringerike for a few years, they moved to the Nordstrand district of Oslo, where they **took over** a y*outh hostel*.
Grammatical Closeness – the paraphrase contains a change in mood, time or number	
O	Nelson Downs never *took* magic *lessons*; he was self-taught and, at a very young age, **developed** remarkable skills in manipulating cards and coins.
P	Nelson Downs never *had* magic *lessons*; he was self-taught and, at a very young age, **has developed** remarkable skills in manipulating cards and coins.
Other – the paraphrase has issues regarding naturalness, register or other aspects	
O	Between 2005 and 2007, he **played** for Saturn Ramenskoye's *reserve team* in the third division.
P	From 2005 to 2007, he **took part** in Saturn Ramenskoye's *reserve team*, which competed in the third division.

Table 5: English translations of examples of **O**riginal and **P**araphrase evaluation errors in different categories. MWEs are shown in italics, while errors are highlighted in bold.

Meaning	**Retained**	**Not Retained**
	54%	46%
Grammaticality	**Correct**	**Incorrect**
	95%	5%
Grammatical Closeness	**Retained**	**Not Retained**
	90%	10%
Other	**No Issues**	**Issues**
	95%	5%

Table 6: Percentages of paraphrases with retained meaning, correct grammaticality, and retained grammatical closeness.

DCSN-NLP at MWE-2026 AdMIRe 2: Bridging Literal and Figurative Meaning Through Hierarchical Multimodal Reasoning

David Cotigă* **and Sergiu Nisioi***
Human Language Technologies Research Center
Faculty of Mathematics and Computer Science
University of Bucharest
cotigadavid@gmail.com
sergiu.nisioi@unibuc.ro

Abstract

This paper presents our system for the MWE-2026 AdMIRe 2.0 shared task, which aimed to advance multimodal idiomatic understanding across 15 languages. We address the task of selecting, from a set of five images, the one that best represents either the literal or idiomatic meaning of a given compound in context. Our approach follows a multi-step pipeline: a large language model (LLM) first determines whether the compound is used literally or idiomatically and generates auxiliary text, consisting of an idiomatic meaning explanation and a visual description of the literal meaning. An ensemble of three CLIP models then identifies the two images most semantically similar to the appropriate generated text via a voting mechanism. Finally, the LLM selects the best image from these two candidates.

1 Introduction

Idiomatic expressions pose a fundamental challenge for language understanding systems because their meanings are not compositionally derivable from their constituent words (Dankers et al., 2022). Rather than being inferable from surface lexical cues, idioms often encode abstract, culturally grounded semantics that require contextual reasoning and access to shared world knowledge (Sag et al., 2002).

For large language models (LLMs), this difficulty is further amplified by inherent ambiguity and distributional bias in training data. Because many idioms occur predominantly in their figurative sense, models tend to strongly associate a given compound with its idiomatic meaning, often underrepresenting or overlooking its literal interpretation.

For example, when prompting GPT-5 (Singh et al., 2025) with the sentence "Some cat's eyes were installed on the newly built road for the safety of the drivers" and asking whether "cat's eyes" is used literally or idiomatically, the model incorrectly interprets it as literal, despite the term's metaphorical usage referring to reflective road studs.

The AdMIRe 2.0 task is designed to test these limitations by evaluating models' ability to understand idiomatic expressions across both linguistic and visual modalities. Given an idiomatic compound, a contextual sentence, and a set of candidate images, systems must identify the image that best matches the intended interpretation of the expression (Arslan et al., 2026).

To address this challenge, we propose a multi-step multimodal disambiguation pipeline that combines linguistic reasoning with structured visual grounding. At a high level, our approach first determines whether an expression is used idiomatically or literally in context, then leverages contrastive text–image representations to progressively narrow the space of candidate interpretations before making a final decision.

2 Task and Dataset

The first edition of the task (Pickard et al., 2025) evaluated systems on a dataset comprising two languages: English and Brazilian Portuguese. The second edition significantly expands this scope, introducing a substantially larger multilingual dataset covering 15 different languages (Torunoğlu-Selamet et al., 2026) which are composed as shown in Appendix C.

Our pipeline was originally developed and validated using the English subset from the first edition. Additionally, at the time of writing, ground-truth annotations for the second edition dataset have not yet been released. For these reasons, we restrict our experiments to the English dataset and focus exclusively on this language for the remainder of the paper.

*Corresponding authors.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 217–225

The English dataset is divided into training, development, test, and extended evaluation subsets, as shown in Table 1. The extended evaluation subset is formed by concatenating the 3 subsets previously mentioned, using different contextual sentences.

Data	# instances
English Train	70
English Dev	15
English Test	15
English Extended	100

Table 1: Dataset statistics for the English language

For each item, the system receives an idiomatic expression (e.g. "bad apple"), a contextual sentence (e.g."The team's efforts were spoiled by one bad apple") and a set of five images (see Figure 1). The system must choose the image which best represents the literal or the idiomatic meaning, depending on which is used in the contextual sentence. In this example, the expected response is the first image. Each set of five images is constructed according to a fixed blueprint: one image that strongly reflects the idiomatic meaning, one that weakly reflects the idiomatic meaning, one image that strongly reflects the literal meaning, one that weakly reflects the literal meaning, and one image that is unrelated to either interpretation. In what follows, these images are referred to as strong idiomatic, weak idiomatic, strong literal, weak literal, and distractor, respectively.

3 Pipeline Overview

This section provides a high-level overview of the proposed multimodal disambiguation pipeline. First, a LLM is provided with a compound expression and its surrounding sentence and tasked with classifying the usage as either idiomatic or literal. Next, for each compound, two textual representations are generated: one explicitly describing the idiomatic meaning, and another describing the literal meaning in visual terms (e.g., for "love triangle": "three hearts connected at the corners, forming a triangle").

These textual descriptions, together with the set of five candidate images, are then processed by three distinct CLIP-based models (Radford et al., 2021) to produce joint text–image embeddings. For each image, similarity scores with respect to both textual descriptions are computed, and an ensemble voting mechanism selects the two most semantically aligned candidates. Finally, the LLM performs a comparison between these two finalists and makes the ultimate disambiguation decision (see Figure 2).

Conceptually, our system can be viewed as a hierarchy in which the LLM serves as the final decision-maker, while the preceding multimodal stages act as a structured filtering process that progressively reduces the ambiguity space to a minimal set of competing interpretations.

4 System Description

4.1 Baselines

First, it is important to acknowledge both the capabilities and limitations of large language models in the context of multimodal idiomatic understanding. When prompted (see Appendix A) to directly select the correct image from a set of five candidates, OpenAI's GPT-4o (OpenAI et al., 2024) achieves an accuracy of approximately 75%. These findings align with results presented in the first edition of the task (Alfter, 2025), where an accuracy of up to 81% was obtained using an LLM-only approach augmented with additional idiom detection and explanation steps.

When it comes to human evaluation, the paper describing the first edition of the task mentions the performance and evaluation method of human annotators, who averaged a precision of 71%, with the best scoring individual obtaining a score of 86%.

4.2 Binary Classification

As a first step, we use the LLM to classify each expression-context pair as either idiomatic or literal. We employ a reasoning-oriented prompting strategy that provides the model with specific evaluation criteria to ground its decision, while strictly constraining the generated output to a single binary label. (The exact prompts are presented in Appendix A). This step is critical, as the resulting classification determines which textual representation is used in the subsequent multimodal similarity stages.

Table 2 reports the classification accuracies obtained using two large language models, GPT-5 and GPT-4o, under two prompting strategies: simple prompting and the reasoning-oriented prompting strategy. For the stronger model (GPT-5), this prompting strategy yields only a modest improvement, as the model already performs near ceiling

Strongly idiomatic | Weakly idiomatic | Distractor | Weakly literal | Strongly literal

Figure 1: Data example for *bad apple*

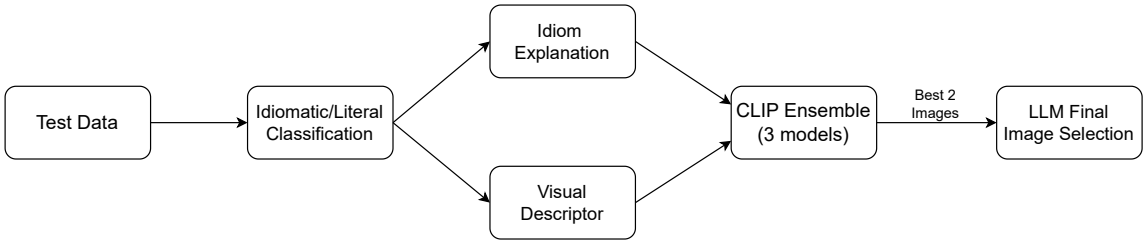


Figure 2: Overview of the proposed multimodal pipeline.

with simple prompting (96%), reaching 99% when guided reasoning is encouraged. In contrast, for the less capable GPT-4o model, the same prompting strategy produces a substantial performance gain, increasing accuracy from 91% to 99%. This effectively brings GPT-4o to near-perfect performance and largely closes the performance gap to GPT-5.

Model	Reasoning	Simple
gpt-5	99%	96%
gpt-4o	99%	91%

Table 2: Performance comparison between Reasoning-Oriented and Simple prompting.

4.3 Auxiliary Text

Certain compounds (e.g., "love triangle", "eye candy") are overwhelmingly used in idiomatic contexts and rarely occur with a literal interpretation in either natural language or visual data. Consequently, multimodal models such as CLIP, which are trained on large-scale web corpora, tend to encode these expressions primarily according to their idiomatic meaning. To ease the burden on the encoding models and encourage more interpretable representations, we generate two auxiliary textual descriptions: one explaining the idiomatic meaning and another providing a visual description of the literal interpretation. In what follows, we will refer to these as "synonym" and "visual", respectively. Also, we will refer to the original expression as "compound". Such examples are presented in the appendix (Table 6).

Table 3 illustrates the issue discussed above using the love triangle example (see Figure 3): the three encoding models assign a higher average similarity score to the strongly idiomatic image rather than to the strongly literal one. The visual descriptor used for the literal interpretation: "three hearts connected at the corners, forming a triangle" does not exactly match the image; nevertheless, it guides the models toward a more faithful representation of the intended literal meaning.

Image Category	Compound	Synonym	Visual
Strong Literal	0.169	0.207	**0.187**
Distractor	0.072	0.137	0.015
Weak Idiomatic	0.135	0.211	0.086
Weak Literal	0.107	0.097	0.091
Strong Idiomatic	**0.188**	**0.256**	0.079

Table 3: Average cosine similarity scores between the three text types (Compound, Synonym, Visual) and the five image categories. The highest score per column is bolded.

However, the outputs of the LLM when used with the default temperature vary considerably. Lower decoding temperatures reduce variability by producing more concentrated token distributions and more deterministic outputs. While this can improve consistency, it also limits exploration of alternative reasoning paths, making the generation more sensitive to early token choices and increasing the

Setting	1st run	2nd run	3rd run
Temp 1.0	82%	90%	87%
Temp 0.3	84%	83%	84%
MBR 0.7	88%	90%	91%
MBR 1.0	90%	88%	89%

Table 4: Performance comparison across text generating strategies.

likelihood that initial errors influence subsequent reasoning.

Minimum Bayes Risk (MBR) decoding explicitly explores multiple plausible generations and selects the solution that is consistent across samples. This allows MBR to recover from individual generation errors and better approximate the model's underlying posterior, resulting in improved performance despite higher computational cost. Recent work has shown that MBR-style decoding can substantially improve robustness by selecting outputs that minimize disagreement across multiple generations rather than relying on a single sample (Heineman et al., 2024).

In our implementation, we apply MBR decoding by generating 20 independent responses at a temperature of 1.0 and 0.7, respectively, for each input. Each response is embedded using a Sentence-Transformers model (all-mpnet-base-v2) (Reimers and Gurevych, 2019). The final prediction is selected as the one whose embedding exhibits minimal average distance to the others, effectively choosing the most representative or consensus solution among the candidates. As shown in Table 4, this strategy consistently outperforms both default-temperature decoding and low-temperature decoding across all evaluated metrics, confirming that increased diversity combined with consensus-based selection yields better performance than increased determinism alone.

4.4 CLIP Ensemble

To encode both images and text, we use an ensemble of vision–language models. Table 5 reports results for seven models released by LAION (Schuhmann et al., 2022), Google (Zhai et al., 2023), and OpenAI (Radford et al., 2021), evaluated under ground-truth idiomaticity classification, thereby isolating the image–text matching performance from upstream classification errors. The three selected models are highlighted in the table.

We report performance using six metrics. Top-1 (T1) and Top-2 (T2) accuracy measure whether the correct image is ranked first or among the top two candidates, respectively. Evaluation is conducted over three complementary sub-tasks: Syn, which assesses idiomatic cases by matching images against the synonym-based generated text; Lit, which evaluates literal cases using the original compound expression; and Vis, which measures performance on literal cases using the visual descriptor of the compound.

As part of the candidate selection process, we introduce an exclusion step in which the image whose embedding is closest to the visual descriptor is removed from consideration. This is intended to prevent this image from being selected as a finalist. Without this exclusion, the system achieves an average accuracy of 83%.

The performance gain introduced by this step is observed exclusively in idiomatic cases, as literal interpretations are generally easier for the models to detect. Applying the same exclusion strategy to literal cases—by removing the most idiomatically aligned image—leads instead to a decrease in overall accuracy.

4.5 LLM Final Choice

For literal and idiomatic interpretations, the distinction between the image that best represents the meaning and the one that more vaguely resembles it is often subtle and can exceed the discriminative capacity of vision–language embedding models alone. The following example illustrates this, as the synonym text generated for "piece of cake" ("Something that is very easy to do or accomplish, often used to describe a task or activity that requires little effort or skill.") produced an embedding more similar to the weak idiomatic, instead of the expected image (see Table 7).

In such cases, similarity-based ranking may fail to reliably separate near-identical candidates. To address this limitation, we introduce a final LLM-based decision stage.

After the CLIP ensemble reduces the candidate pool to the two most semantically aligned images, the LLM is prompted to perform a direct comparison between these finalists, conditioned on the original compound, its context sentence, and the relevant auxiliary text.

Empirically, this final decision step yields a substantial improvement in accuracy, increasing performance from 74% achieved by the CLIP ensemble alone, to 90%. This result highlights the comple-

Model	Syn T1	Syn T2	Lit T1	Lit T2	Vis T1	Vis T2
google/siglip-so400m-patch14-384	61	82	77	92	88	98
laion/CLIP-ViT-H-14-laion2B-s32B-b79K	68	83	81	94	83	96
openai/clip-vit-large-patch14	70	85	83	98	77	96
laion/CLIP-ViT-g-14-laion2B-s12B-b42K	72	87	81	98	84	100
google/siglip-large-patch16-384	4	32	26	51	22	52
laion/CLIP-ViT-B-32-laion2B-s34B-b79K	55	78	81	96	81	96
google/siglip-base-patch16-256	25	44	34	62	30	54

Table 5: Performance comparison of CLIP and SigLIP models.

mentary strengths of embedding-based retrieval and generative reasoning: while the former effectively narrows the search space, the latter excels at fine-grained semantic discrimination.

5 Results

The model accurately predicted the correct image in 53% of tests across all 15 languages, achieving second place in the overall ranking for the image and text subtrack. The exact results for each language are displayed in the appendix. On the English dataset, which was not part of the official evaluation, it achieved an accuracy of around 90% on average. The observed discrepancy between performance on the English subset and the multilingual dataset may be attributed to several factors, including differences in data availability, language-specific idiomatic usage, and the predominantly English-centric training and prompting of the underlying models, as well as a possible difference in the quality of the data.

Considering the variant that was used in the official evaluation and which is described in this paper: there are sources of error at each step of the pipeline, from the binary classification to the final LLM choice. The testing data, once augmented and published, will be a valuable asset that could further improve the system, especially by enabling development focused on multiple languages.

6 Conclusion

In this paper, we presented a multi-step, multimodal system for the MWE-2026 AdMIRe 2.0 shared task. Our approach combines auxiliary text generation, a CLIP model ensemble, and a final LLM-based decision stage to address the challenges of multimodal idiomatic understanding.

Our analysis shows that auxiliary textual representations significantly improve image–text alignment, that Minimum Bayes Risk decoding yields more robust generations than temperature-based control, and that LLMs are most effective when used as high-level decision-makers rather than direct classifiers. Despite the increased computational cost, the resulting gains justify the design choices in settings where accuracy is critical. Because the system relies heavily on LLM-based reasoning, model selection directly impacts both performance and cost. Stronger models improve accuracy but introduce significant computational and economic overhead, necessitating a trade-off between effectiveness and efficiency.

Future work includes extending the system to additional languages, exploring fine-tuned multimodal encoders, and investigating more principled risk functions for MBR decoding. We hope our findings contribute to a deeper understanding of idiomaticity in multimodal language processing.

7 Limitations

We identify several limitations that warrant further investigation:

- The experiments were conducted exclusively on the English language; therefore, the results are strongly influenced by the quality of the models in multilingual settings.
- A more extensive ablation study would be necessary to fully understand the contribution of each component; at present, each component yields a direct and gradual increase in accuracy.
- A more in-depth analysis of the models' familiarity with Brazilian Portuguese data is needed, as it may indicate potential data contamination from the first edition of the task.

Acknowledgments

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology). This research is also supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416.

References

David Alfter. 2025. daalft at SemEval-2025 task 1: Multi-step zero-shot multimodal idiomaticity ranking. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 127–140, Vienna, Austria. Association for Computational Linguistics.

Doğukan Arslan, Rodrigo Wilkens, Wei He, Dilara Torunoğlu-Selamet, Thomas Pickard, Aline Villavicencio, Adriana S. Pagano, and Gülşen Eryiğit. 2026. MWE-2026 Shared Task 2: AdMIRe 2 - Advancing Multimodal Idiomaticity Representation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morocco. Association for Computational Linguistics.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

David Heineman, Yao Dou, and Wei Xu. 2024. Improving minimum Bayes risk decoding with multi-prompt. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22525–22545, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. pages 1–15.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Preprint*, arXiv:2210.08402.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Dilara Torunoğlu-Selamet, Dogukan Arslan, Rodrigo Wilkens, Wei He, Doruk Eryiğit, Thomas Pickard, Adriana S. Pagano, Aline Villavicencio, Gülşen Eryiğit, Ágnes Abuczki, Aida Cardoso, Alesia Lazarenka, Dina Almassova, Amalia Mendes, Anna Kanellopoulou, Antoni Brosa-Rodríguez, Baiba Saulite, Beata Wojtowicz, Bolette Pedersen, and 59 others. 2026. A parallel cross-lingual benchmark for multimodal idiomaticity understanding. *Preprint*, arXiv:2601.08645.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

A LLM Prompts

Baseline Prompt
You are given 5 images, the compound '{compound}' and the context: {sentence}. If the compound is used in the context idiomatically, choose the image that best represents the IDIOMATIC MEANING, and if it is used literally, choose the image that best represents the LITERAL MEANING. Respond with ONLY "1", "2", "3", "4" or "5" - nothing else.

Reasoning-Oriented Prompt for Classification
Here is the idiom: {compound} Although this idiom is usually used idiomatically, it can be used, in rare cases, literally. Think about such an example, then look at this sentence: {sentence} Think about both possible interpretations: literal and idiomatic. Then decide which meaning matches the compound in this sentence. Your response must be exactly one word: literal OR idiomatic

Simple Prompt for Classification
In the following sentence, is the compound {compound} used idiomatically or literally: {sentence} Your response must be exactly one word: literal OR idiomatic.

Prompt for Synonym Text Generation
Define this idiom in 30-40 words, as it would appear in a dictionary: {compound} Write ONLY the definition itself in english. Do NOT include the idiom phrase in your response. Start directly with the meaning, like: "Dishonest or mischievous behavior..." not "Monkey business means..."

Prompt for Visual Description Text Generation
Generate a generic, SHORT visual description for the literal meaning of: {compound} Write ONLY the visual description itself in english. Do NOT include the compound phrase in your response. Start directly with the meaning, like: "Multiple hens having a party..." not "Women having a party..." for hen party and "Very angry aunt..." NOT "Caring woman sending letters..." for "agony aunt" REMEMBER: I WANT THE LITERAL MEANING, NOT IDIOMATIC OR METAPHORICAL

Prompt for Final 2-Candidate Choice
Choose which image better shows the {LITERAL/IDIOMATIC} meaning of: {compound} Respond only with 1 or 2.

B Examples

Type	Generated Text
	Compound: Ghost town
Synonym	A deserted town or area that was once populated or active, often characterized by abandoned buildings and a lack of human activity, typically resulting from economic decline or natural disasters.
Visual	A spectral figure wandering through deserted streets and empty buildings.

Table 6: Examples of generated auxiliary text for the compound "Ghost town".

Image Category	Avg. Similarity
Strong Idiomatic	0.155
Weak Idiomatic	**0.197**
Distractor	0.095
Weak Literal	0.101
Strong Literal	0.093

Table 7: Average cosine similarity scores between the 5 images and the synonym text, for the "piece of cake" example

Strongly idiomatic | Weakly idiomatic | Distractor | Weakly literal | Strongly literal

Figure 3: Data example for *love triangle*

Strongly idiomatic | Weakly idiomatic

Figure 4: Data example for *piece of cake*

C Test Data

Language	# Records	# Compounds
Chinese	179	57
Georgian	113	32
Greek	208	52
Igbo	115	42
Kazakh	156	51
Norwegian	202	51
Portuguese (Brazil)	228	57
Portuguese (Portugal)	220	58
Russian	140	39
Serbian	363	95
Slovak	151	42
Slovenian	240	60
Spanish (Ecuador)	48	13
Turkish	180	54
Uzbek	120	42

Table 8: Dataset composition by language. #Records denotes the total number of instances (rows) available for each language, while #Compounds indicates the number of distinct idiomatic expressions.

Language	**Acc_{all}**	**Acc_{lit}**	**Acc_{id}**	**$Corr_{all}$**	**$Corr_{lit}$**	**$Corr_{id}$**	**DCG_{all}**	**DCG_{lit}**	**DCG_{id}**
Brazilian Portuguese	0.80	0.89	0.71	0.19	0.23	0.15	0.91	0.95	0.87
Russian	0.68	0.90	0.50	0.24	0.59	-0.03	0.87	0.95	0.82
Slovenian	0.67	0.77	0.58	0.23	0.27	0.20	0.87	0.91	0.83
Turkish	0.62	0.75	0.53	0.15	0.24	0.08	0.84	0.90	0.81
European Portuguese	0.57	0.68	0.46	0.12	0.17	0.07	0.81	0.86	0.77
Greek	0.57	0.67	0.46	0.20	0.29	0.10	0.84	0.89	0.80
Kazakh	0.53	0.70	0.44	0.04	0.21	-0.05	0.80	0.89	0.75
Norwegian	0.52	0.72	0.32	0.13	0.30	-0.04	0.80	0.89	0.71
Slovak	0.48	0.54	0.42	0.19	0.21	0.17	0.78	0.80	0.77
Georgian	0.47	0.50	0.44	0.18	0.27	0.10	0.75	0.77	0.74
Chinese	0.45	0.43	0.47	0.11	0.21	0.03	0.76	0.74	0.77
Serbian	0.45	0.55	0.36	0.10	0.14	0.06	0.76	0.81	0.72
Ecuadorian Spanish	0.42	0.18	0.62	0.16	0.16	0.17	0.81	0.78	0.84
Igbo	0.39	0.55	0.33	-0.01	0.00	-0.01	0.74	0.80	0.72
Uzbek	0.33	0.47	0.26	-0.00	0.25	-0.11	0.74	0.80	0.71

Table 9: Per-language evaluation results for AdMIRe 2.0 Subtask A, sorted by decreasing overall accuracy (Acc_{all}).

ITUNLP at MWE-2026 AdMIRe 2: A Zero-Shot LLM Pipeline for Multimodal Idiom Understanding and Ranking

Atakan Site*, **Oğuz Ali Arslan, Gülşen Eryiğit**
Department of Artificial Intelligence and Data Engineering
Istanbul Technical University
{site21, arslanog20, gulsenc}@itu.edu.tr

Abstract

This paper presents our system for AdMIRe 2 (Advancing Multimodal Idiomaticity Representation), a shared task on multilingual multimodal idiom understanding. The task focuses on ranking images according to how well they depict the literal or idiomatic usage of potentially idiomatic expressions (PIEs) in context, across 15 languages and two tracks: a text-only track, and a multimodal track that uses both images and captions. To tackle both tracks, we propose a hybrid zero-shot pipeline built on large vision–language models (LVLMs). Our system employs a chain-of-thought prompting scheme that first classifies each PIE usage as literal or idiomatic and then ranks candidate images by their alignment with the inferred meaning. A primary–fallback routing mechanism increases robustness to safety-filter refusals, while lightweight post-processing recovers consistent rankings from imperfect model outputs. Without any task-specific fine-tuning, our approach achieves 55.9% Top-1 Accuracy in the text-only track and 60.1% in the multimodal (text+image) track, ranking first overall on the official leaderboard. These results suggest that carefully designed zero-shot LVLM pipelines can provide strong baselines for multilingual multimodal idiomaticity benchmarks.

1 Introduction

Idioms constitute a subclass of multi-word expressions (MWEs) and remain a challenging problem even for state-of-the-art large language models (LLMs). The core difficulty stems from the fact that idiomatic meaning is non-compositional and often unpredictable from the constituent words; that is, it cannot be reliably inferred by composing the semantics of the individual words. For instance, the expression "bad apple" rarely refers to a defective piece of fruit; instead, it typically denotes a person whose negative behavior can corrupt or undermine a group. Idioms can also introduce ambiguity between a literal reading suggested by their surface form and the intended idiomatic interpretation. These properties make idioms a valuable probing ground for examining how NLP models represent and compose meaning.

In recent years, progress has been made toward modeling idiomatic meaning (Umut et al., 2025; Kim et al., 2025; Khoshtab et al., 2025). Nevertheless, language models still struggle with figurative and abstract meaning, often failing to go beyond surface-level lexical cues and relying on shallow lexical associations (Mi et al., 2025; Leon et al., 2025). This weakness has practical implications for downstream tasks that require meaning beyond straightforward compositional semantics. In particular, failures in idiom understanding can lead to incorrect reasoning in natural language inference (Stowe et al., 2022), mismatches in retrieval and semantic similarity (Tayyar Madabushi et al., 2022), and erroneous decisions in question-answering systems (Rakshit and Flanigan, 2022). Improving idiom understanding is therefore an important step toward more robust language understanding, motivating targeted benchmarks and analyses of idiomatic language to better assess model generalization.

Idiom processing has been evaluated with idiom-specific datasets and benchmarks at both token and sentence level, covering idiomaticity detection and representation learning (Cook et al., 2008; Haagsma et al., 2020; Saxena and Paul, 2020; Tayyar Madabushi et al., 2022; Tedeschi et al., 2022). More recently, some datasets have incorporated visual information to create multimodal evaluation settings, revealing that grounded figurative and idiomatic understanding can be substantially more challenging for LVLMs than text-only benchmarks (Saakyan et al., 2025; Yosef et al., 2023). SemEval-2025 Task 1 (AdMIRe: Advancing Multimodal Idiomaticity Representation) frames idiomaticity through image-based meaning representation and

*Corresponding author.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 226–236

prediction, providing a benchmark for grounding idiom interpretation beyond text-only cues (Pickard et al., 2025). Building on this foundation, AdMIRe 2 further expands the evaluation to a broader multilingual and multimodal setting centered on potentially idiomatic expressions (PIEs). The shared task is run in two tracks, a text-only track and an image+text track, which enables direct measurement of how visual grounding affects literal–idiomatic disambiguation (Arslan et al., 2026).

In this paper, we propose a hybrid zero-shot system for both tracks that combines two LVLMs through a primary–fallback mechanism and uses efficient prompting strategies to produce robust image rankings.

We show that our proposed system ranks first on the official leaderboard in terms of average performance across tracks. Our pipeline achieves 55.9% accuracy in the text-only track and 60.1% accuracy in the multimodal (text+image) track, without requiring any task-specific fine-tuning.

All the code is available on our GitHub[1].

2 Background

This section reviews prior work on LLMs and LVLMs, with a particular focus on their applications to idiom processing and figurative language understanding.

Early work on idioms and PIEs predates large-scale LLMs and focuses on building dedicated resources and supervised models. Classical idiom datasets target token-level usage labeling and sentence-level idiomaticity, and have shown that idioms are systematically harder for distributional models than compositional expressions (Cook et al., 2008; Saxena and Paul, 2020). More recent corpora such as MAGPIE scale this line of work up to large, annotated collections of PIEs drawn from the British National Corpus, with tens of thousands of instances labeled as literal or idiomatic across more than 1,700 expressions (Haagsma et al., 2020). Other idiom-focused resources extend this direction to naturally occurring English and Portuguese sentences containing multiword expressions, annotated with fine-grained sense labels to evaluate idiom usage detection and sentence-level representation learning (Madabushi et al., 2021). SemEval-2022 Task 2 further consolidates this direction by casting multilingual idiomaticity detection and idiom-aware sentence embeddings as a shared task in English, Portuguese, and Galician, and by providing a common evaluation protocol for idiom-sensitive representations (Tayyar Madabushi et al., 2022). In parallel, large language models have also been explored as tools for idiom corpus construction, generating synthetic idiom corpora across multiple languages and assessing their value for idiomaticity detection (Arslan et al., 2025). Taken together, these efforts establish idiom and PIE processing primarily as a supervised, text-only classification and representation problem, and highlight both the usefulness of idiom-focused resources and the difficulty of encoding idiomatic meaning even for strong pretrained transformers.

With the emergence of LLMs, idiom processing has increasingly been revisited through prompt-based evaluation, particularly in zero-shot and few-shot settings. Recent work constructs controlled contrastive datasets of minimal sentence pairs where the same idiom is used in either a literal or a figurative context, and shows that LLMs often fail precisely when disambiguation requires careful use of contextual cues rather than surface-level associations (Mi et al., 2025). Another line of work investigates LLMs as classifiers for multiword expressions and PIEs, finding that carefully engineered prompts can match supervised baselines on some idiom and MWE identification benchmarks, but that performance does not generalize reliably across datasets and is highly sensitive to annotation choices (Hashiloni et al., 2025). Complementary work evaluates conversational LLMs on challenging idiom detection test suites and reports systematic errors, including over-predicting idiomatic readings in literal contexts and difficulty dealing with polysemous expressions (De Luca Fornaciari et al., 2024). Overall, these studies show that, despite notable progress, state-of-the-art LLMs still rely on shallow lexical cues and frequency statistics when processing idioms and PIEs, and often fail to recover the intended non-compositional meaning from context.

Figurative language has also been studied more broadly, beyond idioms, as a testbed for LLMs' semantic and reasoning capabilities. A natural language inference benchmark for figurative language frames the task as recognizing entailment between around nine thousand premise–hypothesis pairs covering various figurative phenomena, each annotated with an entailment label and a human-written explanation (Chakrabarty et al., 2022). Experiments on such benchmarks show that even

[1] https://github.com/oguzaliarslan/idiom-nlp

strong sequence-to-sequence models fine-tuned on the data exhibit substantial gaps in both prediction accuracy and explanation quality, indicating that figurative language remains challenging even in text-only settings.

More recently, research has begun to explore figurative language in multimodal settings, where images provide additional grounding. The IRFL dataset pairs idioms, metaphors and similes with both figurative and literal candidate images and defines recognition tasks that require models to identify which image best reflects the figurative meaning (Yosef et al., 2023). State-of-the-art LVLMs achieve only around 22% accuracy on IRFL, compared to 97% for humans, underscoring the difficulty of multimodal figurative understanding. V-FLUTE extends this direction by framing visual figurative language understanding as an explainable visual entailment task, covering metaphors, similes, idioms, sarcasm, and humor. Given an image and a caption, a model must decide whether the image entails the caption and provide a textual explanation (Saakyan et al., 2025). Experiments on V-FLUTE reveal that LVLMs struggle to generalize from literal to figurative meaning, particularly when figurative cues are primarily present in the visual modality, and often produce hallucinated or incomplete explanations.

Within this broader figurative-language landscape, multimodal idiom understanding has only recently become a dedicated research focus. IRFL includes an idiom subset, but treats idioms alongside other figurative phenomena in a unified recognition setting, without explicitly modeling potentially idiomatic expressions or literal–idiomatic ambiguity. In contrast, SemEval-2025 Task 1: AdMIRe (Advancing Multimodal Idiomaticity Representation) offers a more targeted benchmark centered on idioms and PIEs in multilingual, multimodal contexts (Pickard et al., 2025). AdMIRe introduces datasets where nominal compounds with both literal and idiomatic readings are embedded in context sentences and paired with images generated to depict either the literal or idiomatic interpretation, across English and Brazilian Portuguese. The task description reports that the strongest participating systems rely on mixtures of pre-trained LLMs and LVLMs, multi-query prompting, and reranking strategies, yet performance still varies considerably across languages, idiom types and sense (literal vs idiomatic), indicating that robust multimodal idiom grounding remains an open challenge.

Participant papers from AdMIRe and related multimodal shared tasks broadly converge on a similar pattern: rather than training models from scratch, systems typically start from powerful proprietary or open-source LVLMs and focus on designing task-specific prompts, scoring functions and ensembling schemes for each benchmark (You et al., 2025). This line of work demonstrates that careful prompt engineering can substantially improve idiom-related performance, but it also highlights the engineering cost and limited generality of heavily task-tuned pipelines. Based on these studies, our work investigates how far a purely zero-shot LVLM-based system can go on AdMIRe 2, using efficient prompting strategies to address both the text-only and image+text tracks. By directly comparing performance across tracks in a shared multilingual PIE setting, we provide complementary evidence on the strengths and remaining limitations of LVLMs for grounded idiomaticity.

3 Methodology

3.1 Proposed Architecture

We propose a zero-shot inference pipeline designed to rank associated images based on the literal or idiomatic usage of PIEs. Our approach utilizes the reasoning capabilities of the state-of-the-art LVLMs through a structured chain-of-thought (CoT) prompting strategy.

As depicted in Figure 1, our architecture leverages both candidate images and their captions to support two tracks: **text-only** and **multimodal** (image + text). In the text-only track, captions proxy visual content, whereas the multimodal track additionally allows reasoning over visual cues absent from text.

3.2 Chain-of-Thought Prompting

Our system employs a five-step CoT prompting strategy[2] that guides the model through explicit reasoning stages before producing a final image ranking.

Step 1: Usage Type Classification First, we analyze the context sentence to determine whether the PIE is used literally or idiomatically. To achieve this, we provide explicit definitions for both categories: literal usage describes physical, photographable scenarios, whereas idiomatic us-

[2]Complete prompt templates for each step are provided in Appendix A.

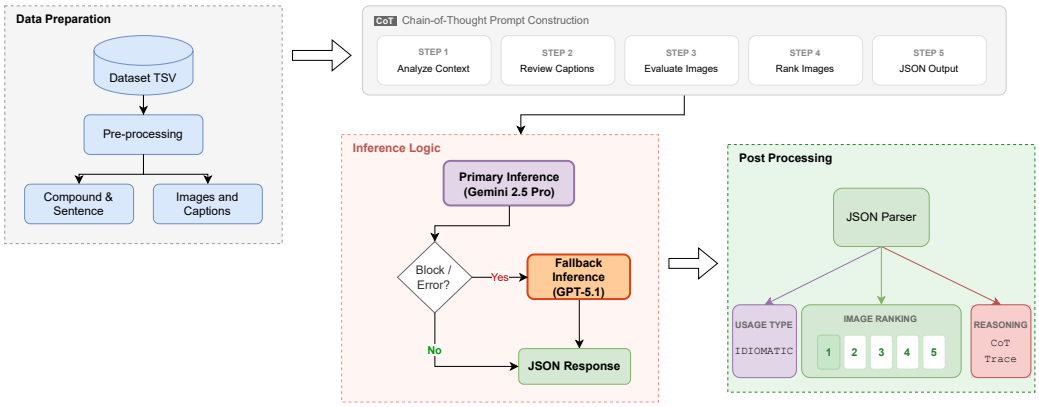


Figure 1: Overview of the proposed system. The system takes a PIE, context sentence, and five candidate images (with captions) as input. A chain-of-thought prompting strategy guides the model through usage classification, image evaluation, and ranking stages to produce the final output.

age conveys figurative meaning distinct from the surface-level interpretation.

Step 2: Reviewing Images and Captions The model examines the five candidate images to understand what each represents. For the text-only track, we analyze the provided image captions to interpret the scenery. In the multimodal track, both the raw images and their corresponding captions are passed to the model.

Step 3: Image Evaluation Following the review of images and captions, the model evaluates how well each of the five candidate images represents the usage type (literal or idiomatic) identified in Step 1. A critical design choice is to prioritize evaluation of how well each image represents the expression's literal or idiomatic meaning, rather than how closely it matches the context sentence. For instance, if the expression "night owl" is used idiomatically in the context sentence, the model is instructed to prefer images that depict the figurative meaning over images that match the literal description of the PIE. For each candidate image, the model generates a quality rating, supported by a rationale that quotes the cues and information utilized for the inference.

Step 4: Image Ranking Based on the evaluations from Step 3, the model ranks all five candidate images from best to worst according to how well they represent the identified meaning of the PIE. We instruct the model to compare lower-ranked candidates with higher-ranked ones, using the ratings from Step 3, to ensure that the ranking is grounded in the preceding analysis rather than generated arbitrarily.

Step 5: Output Generation Finally, the model produces a structured JSON output containing all components of its analysis. The output includes the identified usage type, the reasoning behind this classification, evaluations for each candidate image, and the final ranking of the candidate images with supporting statements.

3.3 Model Inference Logic

Our inference pipeline implements a hybrid LVLM architecture that routes instances between two models via a primary-fallback mechanism. We use Gemini 2.5 Pro (Comanici et al., 2025) as our primary model and GPT-5.1 (OpenAI, 2025) as the fallback.

For each PIE in the dataset, we construct the CoT prompt and send it to the primary model. A major challenge while processing PIEs across different languages is that certain expressions can be flagged by the API's safety filters, particularly idioms involving sensitive terms. We monitor API responses for refusal signals and route to a fallback model when the primary model refuses. This ensures valid rankings for every instance without manual intervention.

3.4 Post Processing

In this stage, we parse the model's JSON output directly, with a regex-based fallback to extract rankings when parsing fails due to syntax errors. This ensures valid rankings even from malformed outputs.

3.5 Evaluation

The system's performance on both tracks was evaluated using two official metrics: Top-1 Accuracy, measuring whether the gold-standard best image is ranked first, and NDCG@5 with relevance weights [3, 1, 0, 0, 0], capturing overall ranking quality. Leaderboard rankings were determined by average Top-1 Accuracy across languages.

4 Experiments

4.1 Dataset

The AdMIRe 2 shared task provides a dataset covering PIEs across 15 diverse languages, listed in Table 2. The construction and statistics of this multilingual resource are described in the accompanying dataset paper (Torunoğlu-Selamet et al., 2026). Each data instance consists of a PIE, a context sentence in which the PIE is used, and five candidate images along with their generated captions.

4.2 Text-only Track

For the text-only track, Table 1 presents our system's performance across all 15 languages. Our approach achieves 55.9% average Top-1 accuracy, with a NDCG@5 of 0.831. Brazilian Portuguese leads with 78.9% accuracy, followed by Slovenian 72.5% and Russian 65.0%, while Ecuadorian Spanish presents the most challenging case at only 25.0% accuracy.

4.3 Multimodal Track

The right half of Table 1 presents our hybrid system's performance on the multimodal track for all languages. Incorporating images leads to clear gains: average Top-1 Accuracy improves to 60.1% and NDCG@5 increases to 0.849. Overall, our system shows 4.2% better performance on multimodal track.

The benefit of visual information is especially observed in several languages. Turkish shows the largest improvement, from 50.5% to 67.8%, followed by Brazilian Portuguese, Serbian, Norwegian, Slovak, and Slovenian. Even for the hardest language, Ecuadorian Spanish, multimodality yields a modest gain.

4.4 Literal vs. Idiomatic Asymmetry

Across both tracks, we observe a consistent asymmetry between literal and idiomatic usages. In the text-only setting, our system attains 61.1% average accuracy on literal uses, compared to 51.3% on idiomatic ones, indicating that idiomatic readings are substantially harder to capture. When images are added, overall performance improves for both usage types: average accuracy increases to 66.9% for literal cases and 54.8% for idiomatic ones. However, the literal–idiomatic gap does not disappear; in fact, it slightly widens, suggesting that current LVLMs leverage visual cues more effectively for concrete, photographable meanings than for abstract figurative interpretations. This pattern is particularly evident in languages such as Greek and Norwegian, where literal cases consistently dominate idiomatic ones, and aligns with prior findings that idioms remain challenging even for strong multimodal models.

4.5 Comparison between Text-only and Multimodal Track

Comparing columns in Table 1, 14 out of 15 languages benefit from adding images, confirming that visual grounding generally helps the model align PIE interpretations with the correct image. Notably, Igbo is the only exception where text-only setting outperforms the multimodal setting by 4%, which may be due to noisier captions or weaker image-text alignment for this language. This suggests that while multimodal information is beneficial in most cases, its impact can be uneven across languages and data conditions.

4.6 Model Comparison

To understand the contribution and performance of the underlying LVLMs, we analyze the performance of our primary model against the fallback model. Detailed breakdowns are provided in Table 3 (text-only) and Table 4 (text + image) in the Appendix.

Gemini 2.5 Pro outperforms GPT-5.1 across both tracks, achieving 55.9% vs. 55.0% on text-only and 59.8% vs. 57.3% on the multimodal track. In the text-only track, the two models perform similarly, with less than one percentage point difference, demonstrating the strong linguistic capabilities of both models. However, the gap widens considerably when images are included, showing Gemini 2.5 Pro's stronger vision-language capacity. Gemini gains 3.9% points when moving from text-only to text+image, compared to GPT-5.1's 2.3% improvement. This disparity is most pronounced in Turkish, where Gemini's accuracy jumps from 50.5% to 67.6% while GPT-5.1 shows a more modest increase from 52.7% to 58.2%.

	Text-Only Track						Multimodal Track					
	Accuracy			NDCG@5			Accuracy			NDCG@5		
Language	All	Lit	Id	All	Lit	Id	All	Lit	Id	All	Lit	Id
Chinese	.458	.556	.378	.774	.806	.749	**.497**	**.568**	**.439**	**.799**	**.811**	**.789**
Georgian	.513	.600	.444	.791	.828	.762	**.531**	**.620**	**.460**	**.808**	**.840**	**.783**
Greek	.591	.702	.481	.856	.917	.795	**.639**	**.740**	**.538**	**.874**	**.925**	**.822**
Igbo	**.478**	**.606**	**.427**	**.784**	.836	**.764**	.435	.545	.390	.777	**.843**	.751
Kazakh	.603	.593	**.608**	.838	.849	**.833**	**.609**	**.667**	.578	**.844**	**.879**	.825
Norwegian	.614	.780	.451	.853	.909	.799	**.673**	**.790**	**.559**	**.881**	**.922**	**.841**
Portuguese (Brazil)	.789	.860	.719	.917	.951	.884	**.855**	**.868**	**.842**	**.939**	**.956**	**.922**
Portuguese (Portugal)	.618	.670	**.570**	.863	.872	**.855**	**.641**	**.717**	**.570**	**.865**	**.882**	.848
Russian	.650	.742	**.577**	.871	.903	**.846**	**.686**	**.855**	.551	**.891**	**.947**	**.846**
Serbian	.551	.599	**.505**	.819	.849	**.792**	**.617**	**.740**	.500	**.837**	**.897**	.779
Slovak	.543	.691	.422	.836	.889	.793	**.596**	**.721**	**.494**	**.854**	**.902**	**.815**
Slovenian	.725	.792	.658	.893	.925	.860	**.779**	**.833**	**.725**	**.911**	**.940**	**.882**
Spanish (Ecuador)	.250	.045	**.423**	.723	**.658**	.777	**.271**	**.091**	**.423**	**.726**	.650	**.791**
Turkish	.505	.514	.500	.823	.826	.821	**.676**	**.722**	**.645**	**.895**	**.912**	**.884**
Uzbek	.500	.417	**.536**	.816	.811	.818	**.517**	**.556**	.500	**.834**	**.855**	**.824**
Average	**.559**	**.611**	**.513**	**.831**	**.855**	**.810**	**.601**	**.669**	**.548**	**.849**	**.877**	**.827**

Table 1: Results across both tracks for 15 languages. Accuracy and NDCG@5 are reported for all instances (All), literal usage (Lit), and idiomatic usage (Id). The multimodal track outperforms text-only by 4.2% in overall accuracy.

5 Conclusion

In this paper, we presented the solution developed by the ITUNLP group for the AdMIRe 2 shared task on multilingual multimodal idiom understanding. The proposed approach tackles the task in a purely zero-shot setting. Our system yields promising results, achieving 1^{st} place on the official leaderboard in terms of average performance across both tracks.

For future work, we plan to evaluate open-source LVLMs within our proposed system to gain a broader perspective on model performance. Furthermore, since the AdMIRe 2 dataset currently covers only 15 languages, we aim to apply our system to additional languages and datasets, further expanding its applicability and testing its robustness.

References

Doğukan Arslan, Hüseyin Anıl Çakmak, Gulsen Eryigit, 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 Torunoğlu-Selamet, Thomas Pickard, Aline Villavicencio, Adriana S. Pagano, and Gülşen Eryiğit. 2026. MWE-2026 Shared Task 2: AdMIRe 2 - Advancing Multimodal Idiomaticity Representation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morocco. 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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The vnc-tokens dataset.

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.

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 23782–23801, Suzhou, China. Association for Computational Linguistics.

Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2025. Comparative study of multilingual idioms and similes in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8680–8698, Abu Dhabi, UAE. Association for Computational Linguistics.

Jisu Kim, Youngwoo Shin, Uiji Hwang, Jihun Choi, Richeng Xuan, and Taeuk Kim. 2025. Memorization or reasoning? exploring the idiom understanding of LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21689–21710, Suzhou, China. Association for Computational Linguistics.

Frances Laureano De Leon, Harish Tayyar Madabushi, and Mark G. Lee. 2025. Evaluating large language models on multiword expressions in multilingual and code-switched contexts. *Preprint*, arXiv:2504.20051.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. Astitchinlanguagemodels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. *Preprint*, arXiv:2109.04413.

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.

OpenAI. 2025. Gpt-5.1 instant and gpt-5.1 thinking system card. `https://cdn.openai.com/pdf/4173ec8d-1229-47db-96de-06d87147e07e/5_1_system_card.pdf`. Technical report.

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.

Geetanjali Rakshit and Jeffrey Flanigan. 2022. FigurativeQA: A test benchmark for figurativeness comprehension for question answering. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 160–166, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2025. Understanding figurative meaning through explainable visual entailment. *Preprint*, arXiv:2405.01474.

Prateek Saxena and Soma Paul. 2020. Epie dataset: A corpus for possible idiomatic expressions. *Preprint*, arXiv:2006.09479.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

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.

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.

Dilara Torunoğlu-Selamet, Dogukan Arslan, Rodrigo Wilkens, Wei He, Doruk Eryiğit, Thomas Pickard, Adriana S. Pagano, Aline Villavicencio, Gülşen Eryiğit, Ágnes Abuczki, Aida Cardoso, Alesia Lazarenka, Dina Almassova, Amalia Mendes, Anna Kanellopoulou, Antoni Brosa-Rodríguez, Baiba Saulite, Beata Wojtowicz, Bolette Pedersen, and 59 others. 2026. A parallel cross-lingual benchmark for multimodal idiomaticity understanding. *Preprint*, arXiv:2601.08645.

Özge Umut, Atakan Site, Doğukan Arslan, and Gülşen Eryiğit. 2025. Exploring Turkish idiomaticity with large language models. In *2025 10th International Conference on Computer Science and Engineering (UBMK)*, pages 533–538.

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.

Runyang You, Xinyue Mei, and Mengyuan Zhou. 2025. PALI-NLP at SemEval 2025 task 1: Multimodal idiom recognition and alignment. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1211–1216, Vienna, Austria. Association for Computational Linguistics.

Appendix

A Generation Prompts

A.1 Task Setup and Context

Research Context & Task

You are an expert linguist conducting academic research on potentially idiomatic expressions (PIEs) for a linguistic analysis study. This is a scholarly research task analyzing language usage patterns.
Research Context: Academic linguistic analysis of expression usage patterns.
Expression: `{compound}`
Context Sentence: `{sentence}`
Research Task:

1. Determine if `{compound}` is used LITERALLY or IDIOMATICALLY in this sentence.
2. Rank 5 images by how well they represent the LITERAL MEANING or IDIOMATIC MEANING of `{compound}` (based on Step 1).

CRITICAL: Rank by how well images represent the literal/idiomatic MEANING of the expression, NOT by how well they match the specific context sentence. If it's literal, rank images closest to the literal meaning highest. If it's idiomatic, rank images closest to the idiomatic meaning highest. You must work through this problem step-by-step. Complete each step fully before moving to the next.

A.2 Chain-of-Thought Reasoning Steps

Step 1: Analyze Context and Determine Usage Type

Determine whether `{compound}` is used **LITERALLY** or **IDIOMATICALLY** in this sentence.

- **LITERAL:** Words describe their actual physical meaning; you could photograph exactly what they describe (e.g., "night owl" = actual owl).
- **IDIOMATIC:** Expression has figurative meaning different from literal interpretation; describes abstract concepts (e.g., "night owl" = person who stays up late).

Output: State "LITERAL" or "IDIOMATIC" with 2-3 sentence reasoning referencing specific context clues.

Step 2: Review Image Captions

Below are the 5 image captions. Examine the caption text to understand what each image represents:
[System Note: Image captions inserted here dynamically]

- **Image 1 (`{img_name}`):** Caption: "`{caption}`"
- ...

Note what each caption describes and how it relates to literal vs figurative representation.

Step 3: Evaluate Each Image Against Usage Type

For each image (based on its caption), evaluate how well it represents the LITERAL MEANING or IDIOMATIC MEANING of `{compound}` (based on Step 1).
CRITICAL: Evaluate how well the image (based on its caption) represents the literal/idiomatic MEANING of the expression, NOT how well it matches the specific context sentence.
Evaluation criteria:

- **If LITERAL:** Does the caption suggest the image shows the physical/literal meaning of `{compound}`? Rank images that best represent what `{compound}` literally means (the actual physical thing/action), regardless of whether they match the specific context scene.
- **If IDIOMATIC:** Does the caption suggest the image represents the figurative/idiomatic meaning of `{compound}`? Rank images that best represent the idiomatic meaning (the abstract concept), regardless of whether they match the specific context.

For each image (4-5 sentences):

- Quote specific caption phrases
- Explain how well the caption suggests the image represents the literal/idiomatic MEANING of `{compound}`
- Do NOT evaluate based on context matching - only evaluate meaning representation
- Match quality: EXCELLENT / GOOD / MODERATE / POOR / VERY POOR

Step 4: Rank Images

Rank all 5 images from BEST to WORST based on how well their captions suggest they represent the literal/idiomatic MEANING of `{compound}`.
CRITICAL RANKING RULE: Rank by meaning representation, NOT context matching:

- **If LITERAL:** Rank images that best represent the LITERAL MEANING of `{compound}` highest (captions suggesting the actual physical thing/action). The top 2 images should be closest to the literal meaning, regardless of context.
- **If IDIOMATIC:** Rank images that best represent the IDIOMATIC MEANING of `{compound}` highest (captions suggesting the figurative/abstract concept). The top 2 images should be closest to the idiomatic meaning, regardless of context.

For #1 image (5-6 sentences):

- Start by restating WHY the expression is literal/idiomatic (from Step 1)
- Quote caption phrases
- Explain why this image's caption best represents the literal/idiomatic MEANING of `{compound}`
- Focus on meaning representation, NOT context matching
- Briefly compare to lower-ranked images

A.3 Final Output Specification

Step 5: Final Output (JSON)

Provide your complete analysis as VALID JSON. Ensure all 5 images are included in the ranking. **CRITICAL:** Use the ACTUAL image filenames listed above, NOT placeholder names like "image1.png".

JSON structure (use actual filenames from the images listed above):

```
{
  "usage_type": "literal" or "idiomatic",
  "usage_reasoning": "2-3 sentences with context clues",
  "image_evaluations": {
      "{image_names[0]}": "EXCELLENT - quote caption...",
      "{image_names[1]}": "GOOD - quote caption...",
      "{image_names[2]}": "MODERATE - quote caption",
      "{image_names[3]}": "POOR - quote caption",
      "{image_names[4]}": "VERY POOR - quote caption"
  },
  "reasoning": "5-6 sentences: restate WHY literal/idiomatic...",
  "ranking": ["{image_names[0]}", "{image_names[1]}", "{image_names[2]}",
              "{image_names[3]}", "{image_names[4]}"]
}
```

Requirements:

- Use the EXACT image filenames from Step 2 (e.g., `{image_names[0]}`)
- Quote captions, connect to usage type
- Escape quotes as
  " in JSON strings
- Output ONLY JSON.

B Language Codes

Table 2: List of languages and their corresponding codes.

Language	Code	Language	Code
Chinese	zh	Russian	ru
Georgian	ka	Serbian	sr
Greek	el	Slovak	sk
Igbo	ig	Slovenian	sl
Kazakh	kk	Spanish (Ecuador)	es-EC
Norwegian	no	Turkish	tr
Portuguese (Brazil)	pt-BR	Uzbek	uz
Portuguese (Port.)	pt-PT		

C Detailed Per-Language Results on Text-Only and Multimodal Tracks

D Inference Parameters

	Gemini 2.5 Pro						GPT 5.1					
	Accuracy			NDCG@5			Accuracy			NDCG@5		
Language	All	Lit	Id	All	Lit	Id	All	Lit	Id	All	Lit	Id
Chinese	**.458**	**.556**	.378	.774	**.806**	.749	.436	.444	**.429**	**.776**	.774	**.778**
Georgian	.513	.600	.444	.791	.828	.762	**.531**	.600	**.476**	**.804**	**.834**	**.781**
Greek	**.591**	**.702**	.481	**.856**	**.917**	.795	**.591**	.673	**.510**	.848	.896	**.800**
Igbo	**.478**	**.606**	**.427**	**.784**	**.836**	**.764**	.391	.454	.365	.766	.824	.742
Kazakh	**.603**	.593	**.608**	.838	.849	.833	.577	**.630**	.549	**.850**	**.880**	**.834**
Norwegian	**.614**	**.780**	**.451**	**.853**	**.909**	**.799**	.554	.700	.412	.838	.892	.785
Portuguese (Brazil)	**.789**	**.860**	**.719**	**.917**	**.951**	**.884**	.741	.816	.667	.905	.946	.865
Portuguese (Portugal)	**.618**	.670	**.570**	**.863**	.872	**.855**	.595	**.717**	.482	.849	**.902**	.800
Russian	.650	.742	.577	.871	.903	.846	**.664**	**.758**	**.590**	**.876**	**.913**	**.848**
Serbian	**.551**	**.599**	**.505**	**.819**	**.849**	**.792**	.543	.588	.500	.809	.842	.777
Slovak	**.543**	**.691**	.422	**.836**	**.889**	.793	.530	.574	**.494**	.824	.840	**.811**
Slovenian	**.725**	**.792**	.658	.893	**.925**	.860	.704	.742	**.667**	**.894**	.915	**.873**
Spanish (Ecuador)	.250	.045	.423	.723	.658	.777	**.354**	**.090**	**.576**	**.739**	**.661**	**.806**
Turkish	.505	.514	.500	.823	.826	.821	**.527**	**.556**	**.509**	**.826**	**.831**	**.823**
Uzbek	.500	.417	.536	.816	**.811**	.818	**.508**	.417	**.548**	.816	.803	**.822**
Average	**.559**	**.611**	.513	**.831**	**.855**	**.810**	.550	.584	**.518**	.829	**.855**	.817

Table 3: Detailed performance comparison of Gemini 2.5 Pro and GPT 5.1 on the **text only track**. Accuracy and NDCG@5 are reported for all instances (All), literal usage (Lit), and idiomatic usage (Id).

	Gemini 2.5 Pro						GPT 5.1					
	Accuracy			NDCG@5			Accuracy			NDCG@5		
Language	All	Lit	Id	All	Lit	Id	All	Lit	Id	All	Lit	Id
Chinese	.480	**.617**	.367	.789	**.825**	.759	**.497**	.568	**.439**	**.799**	.811	**.789**
Georgian	**.531**	**.620**	**.460**	**.808**	**.840**	**.783**	.469	.560	.397	.792	.835	.757
Greek	**.639**	.740	**.538**	**.874**	.925	**.822**	.635	**.760**	.510	.868	**.930**	.807
Igbo	.426	.485	**.402**	.777	.828	**.757**	**.435**	**.545**	.390	.777	**.843**	.751
Kazakh	**.609**	.667	**.578**	**.844**	.879	**.825**	.564	**.722**	.480	.838	**.915**	.797
Norwegian	**.673**	**.790**	**.559**	**.881**	**.922**	.841	.658	.770	.549	.874	.906	**.844**
Portuguese (Brazil)	**.855**	.868	**.842**	**.939**	.956	**.922**	**.855**	**.921**	.789	.938	**.969**	.907
Portuguese (Portugal)	**.641**	**.717**	**.570**	**.865**	.882	**.848**	.614	.708	.526	.856	**.893**	.822
Russian	**.686**	**.855**	**.551**	**.891**	**.947**	.846	.650	.790	.538	.871	.931	.823
Serbian	.595	.661	**.532**	.829	.867	**.792**	**.617**	**.740**	.500	**.837**	**.897**	.779
Slovak	**.596**	**.721**	.494	**.854**	**.902**	**.815**	.583	.676	**.506**	.844	.897	.800
Slovenian	**.779**	**.833**	**.725**	**.911**	**.940**	**.882**	.758	.817	.700	.907	.933	.880
Spanish (Ecuador)	**.271**	**.091**	**.423**	**.726**	**.650**	**.791**	.208	.045	.346	.721	.649	.782
Turkish	**.676**	**.722**	**.645**	**.895**	**.912**	**.884**	.582	.681	.518	.835	.874	.810
Uzbek	**.517**	**.556**	**.500**	**.834**	**.855**	**.824**	.475	.444	.488	.815	.824	.811
Average	**.598**	**.663**	**.539**	**.848**	**.875**	**.826**	.573	.650	.512	.838	.874	.811

Table 4: Detailed performance comparison of Gemini 2.5 Pro and GPT 5.1 on the **text+image track**. Accuracy and NDCG@5 are reported for all instances (All), literal usage (Lit), and idiomatic usage (Id).

Model	**Max Output Tokens**	**Temperature**	**Reasoning Effort**
Gemini 2.5 Pro	8,192	1.0	Default
GPT-5.1 (fallback)	2,500	Default	Medium

Table 5: Inference parameters used for the primary and fallback models.

Archaeology at MWE-2026 PARSEME 2.0 Subtasks 1 and 2: Parsing is for Encoders, Paraphrasing is for LLMs

Rareș-Alexandru Roșcan* **and Sergiu Nisioi***
Human Language Technologies Research Center
Faculty of Mathematics and Computer Science
University of Bucharest
roscanrares@gmail.com
sergiu.nisioi@unibuc.ro

Abstract

This paper presents our approach to the PARSEME 2.0 Shared Task on Romanian, covering both Identification (Subtask 1) and Paraphrasing (Subtask 2). While Large Language Models (LLMs) excel at semantic generation, we hypothesize that they lack the structural precision required for MWE identification, leading to "boundary hallucinations" that compromise downstream simplification. Our Rank 1 results on Romanian confirm this: a specialized encoder (RoBERT) using standard sequence labeling outperforms both few-shot LLMs and complex structural parsers (MTLB-STRUCT). This justifies our proposed pipeline: using encoders as precise "pointers" to guide the generative power of LLMs.

1 Introduction

Text Simplification (TS) aims to reduce the linguistic complexity of text to make it more accessible to diverse audiences, including non-native speakers (Bunparit and Riabroi, 2025) and individuals with cognitive impairments (Guidroz et al., 2025). Although modern TS approaches have achieved significant success (Alva-Manchego et al., 2025), Multiword Expressions (MWEs) pose persistent challenges (Barbu Mititelu et al., 2025). A TS system that fails to identify them correctly may simplify them literally, compromising the original meaning (Agrawal and Carpuat, 2024).

However, performance drops substantially in low-to-medium resource settings (Zhong et al., 2026). Languages like Romanian expose the limitations of massively multilingual models, where the scarcity of native supervision amplifies Anglocentric inductive biases. Despite their scale, such architectures often "think in English" (Etxaniz et al., 2024), effectively overriding local morpho-syntactic constraints with dominant English patterns.

Emerging benchmarks for Romanian (Anghel et al., 2025) reveal a critical trend: linguistically grounded baselines consistently outperform purely generative models, which struggle with morphological precision.

We hypothesize that reliable MWE paraphrasing is currently suboptimal when performed end-to-end, particularly in low-resource languages. It requires the decoupling of two distinct capabilities: precise structural analysis to anchor the expression, and context-aware generation to rewrite it. We validate this modular architecture within the context of the PARSEME 2.0 Shared Task (Scholivet et al., 2026), proposing a pipeline where Subtask 1 serves as the structural scaffolding for Subtask 2:

- **Subtask 1 (Identification):** we utilize a fine-tuned encoder (`RoBERT-base`) to strictly localize expressions. we demonstrate that encoder-only architectures offer the boundary precision necessary to prevent the "over-simplification" of surrounding context a precision that generative models currently lack in Romanian.

- **Subtask 2 (Paraphrasing):** we leverage the Few-shot capabilities of `GPT-4o`, but strictly constrain its input to the spans identified by the encoder. This hybrid approach ensures that the LLM's creativity is channeled exclusively into the MWE's simplification, minimizing hallucinations and preserving the sentence's original meaning.

This approach highlights that effective TS isn't about massive scale, but about using specialized encoders to guide the generative power of LLMs.

2 Related Work

MWEs have famously been characterized as a "pain in the neck" for Natural Language Processing (Sag

*Corresponding authors.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 237–247

et al., 2002). Particularly idioms, these constructions pose longstanding challenges due to their idiosyncrasy and non-compositionality, functioning as single semantic units despite their variable syntax (Baldwin and Kim, 2010). Fixed phrases like the Romanian "a spăla putina" (lit. *to wash the barrel* → to run away) exemplify this strict non-compositionality, rendering literal interpretation strategies entirely ineffective.

Following the inclusion of Romanian in the inaugural PARSEME Shared Task (Savary et al., 2017), Barbu Mititelu et al. (2019) consolidated these efforts into the first large-scale, open-access corpus of Romanian Verbal MWEs. This work highlighted the language's specific challenges: high morphological richness and relatively free word order.

While the global PARSEME leaderboards have historically been dominated by massive multilingual architectures such as MTLB-STRUCT (Taslimipoor et al., 2020), which leverage cross-lingual transfer to mitigate data scarcity, this paradigm has shown limitations for Romanian. Recent retrospective studies (Avram et al., 2023) challenged this multilingual dominance, demonstrating that a standard, fine-tuned monolingual `RoBERT` model significantly outperforms complex multilingual baselines (including MTLB-STRUCT) by better capturing the language-specific morphosyntactic nuances required for resolving discontinuity.

This syntactic flexibility makes Romanian MWEs significantly more elusive. Unlike fixed English idioms (e.g., *kick the bucket*), Romanian expressions exhibit extreme morphosyntactic elasticity, allowing for extensive interpolation that defies the contiguity biases of Anglocentric models.

Even within the Romance family, Romanian displays a higher degree of word order freedom, allowing components to be separated by arbitrarily long sequences (e.g., *"o luase [fără să se uite înapoi] la sănătoasa"* – lit. *"he took it [without looking back] to the healthy"*, meaning *"he fled"*). Consequently, transfer learning from high-resource languages often fails to capture these specific discontinuity patterns, necessitating dedicated monolingual architectures.

3 Task Description

The PARSEME 2.0 Shared Task addresses the end-to-end processing of MWEs.

Subtask 1: Identification and Categorization Systems are required to detect MWE spans and assign them fine-grained categories. The taxonomy covers a broad spectrum of expressions, ranging from Verbal MWEs (e.g., Idioms – `VID`, Light Verb Constructions – `LVC`) to Nominal (`NID`), Adjectival (`AdjID`), and Adverbial (`AdvID`) constructions. The primary challenge in Romanian is handling discontinuity and nesting. Due to the relatively free word order of Romanian, MWE components are frequently separated by intervening tokens of arbitrary length.

Subtask 2: Paraphrasing This subtask targets the semantic substitution of MWEs. Systems must replace the identified spans with semantically equivalent words or phrases, regardless of whether the output remains idiomatic or becomes literal.

4 System Description

Our experiments are conducted on the Romanian corpus[1] provided by the PARSEME 2.0 Shared Task, distributed in the `.cupt` format (an extension of CoNLL-U). The dataset contains annotations for various categories of MWEs, adhering to the universal guidelines of the PARSEME network.

Our approach is grounded in the observation that MWE identification and MWE paraphrasing require fundamentally different processing capabilities. Consequently, we treat Subtask 1 as a strictly structural sequence labeling problem, while Subtask 2 is treated as a semantic generation problem.

4.1 MWE Identification (Subtask 1)

We frame MWE identification as a sequence labeling task. The training data provided in the CUPT format contains complex, often nested annotations. To make this compatible with standard transformer-based classifiers, we linearized the structures using the **BIO (Begin, Inside, Outside)** tagging scheme.

Formally, for a sentence $S = \{w_1, w_2, ..., w_n\}$, each token w_i is assigned a label $y_i \in \mathcal{L}$, where $\mathcal{L}$ represents the set of MWE categories prefixed with positional tags: specifically, the scheme assigns **B-CAT** to the initial token of a category *CAT* (e.g., `B-VID`), **I-CAT** to subsequent components within the expression, and **O** to all tokens outside any MWE.

Preprocessing: we addressed the legacy encoding inconsistencies common in Romanian by normalizing all diacritics to the standard comma-below

[1] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/

form (*ș, ț*) prior to tokenization, ensuring vocabulary alignment with the pre-trained models.

We evaluated five Transformer-based architectures divided into two categories: Multilingual Models, including `bert-base-multilingual-cased`, `xlm-roberta-base`, and `mdeberta-v3-base`; and Romanian Monolingual Models, specifically `RoBERT-base` and `bert-base-romanian-cased-v1`.

LLM Benchmarking Setup: since fine-tuned encoders already handle standard cases efficiently, demonstrating that a computationally expensive LLM can replicate this performance offers no practical added value. Therefore, we designed a targeted stress test focused exclusively on the encoders' known failure modes (e.g., rare or unseen MWEs). Our goal was to determine if `GPT-4o` provides genuine gains in these "hard" scenarios where the baseline struggles, rather than redundantly evaluating it on easy instances.

4.2 LLMs vs Encoders

To construct a representative evaluation subset ($N = 170$), we implemented a deterministic sampling script that filters the validation data through a hierarchical cascade. We enforced strict quotas to mitigate frequency bias and ensure balanced coverage of structural complexity.

We categorized MWE types into three priority tiers based on preliminary `RoBERT-base` F1 scores as detailed in Appendix Table 5: *Rare/Hard* ($F1 < 0.75$), *Medium* ($0.75 \leq F1 \leq 0.89$), and *Frequent* ($F1 > 0.89$).

1. **Rare/Hard Instances (50):** High-priority categories (e.g., `NV.VID`).
2. **Structural Complexity (40):** Discontinuous MWEs selected from the *remaining* pool, explicitly sorting to prioritize **Rare/Medium** types over frequent ones.
3. **Density Stress-Test (20):** The remaining sentences with the highest expression count (≥ 3).
4. **Balanced Baseline (60):** A random sample from the final remainder to complete the set.

To accommodate the boundary inconsistencies typical of generative models, we employed two scoring levels based on token index set operations. Note that for GPT-4o, which outputs raw text, we implemented a heuristic alignment step to map generated phrases back to source token indices before evaluation. Let $P_{indices}$ denote the set of token indices predicted by the system and $G_{indices}$ the set of ground truth indices.

- **Strict F1:** Enforces a rigid criterion where both the MWE category and the set of token indices must exactly match ($P_{indices} = G_{indices}$).
- **Soft F1:** Adopts a relaxed matching strategy to account for "boundary hallucinations". A prediction is considered a true positive if the category matches and the predicted span has a non-empty intersection with the gold span ($P_{indices} \cap G_{indices} \neq \emptyset$).

To investigate whether LLMs can be leveraged to bridge the coverage gaps of supervised encoders on novel data, we constructed a second evaluation subset ($N = 85$) consisting exclusively of Unseen MWEs. Our objective was to determine if the extensive pre-training of LLMs enables them to resolve idiomatic instances that are entirely absent from the fine-tuning curriculum.

We define an MWE as "unseen" if its lexical signature is entirely absent from the training partition. The selection process involved a strict set-difference operation between the validation and training lexicons: $S_{unseen} = \{(L, C) \mid (L, C) \in \mathcal{D}_{dev} \wedge (L, C) \notin \mathcal{D}_{train}\}$ where $\mathcal{D}_{train}$ and $\mathcal{D}_{dev}$ denote the sets of unique MWE instances found in the training and development partitions respectively, C is the MWE category, and L represents the set of component lemmas.

Crucially, our extraction algorithm utilizes order-agnostic lemma matching. By representing each MWE as a sorted tuple of lemmas ((e.g., <*"a", "decizie", "lua"*> for the canonical expression *"a lua o decizie" – lit. "to make a decision"*), we ensure that syntactic variations of training examples are not mistakenly classified as novel. This subset, therefore, tests the model's true few-shot capability on new idiomatic combinations.

4.3 MWE Paraphrasing (Subtask 2)

We frame paraphrasing as a two-stage pipeline to mitigate the target ambiguity inherent in unconstrained generation. We observed that without explicit span boundaries, LLMs frequently target in-

cidental collocations rather than the ground truth, causing unintended text modifications.

To prevent this, `RoBERT-base` acts as a target anchor, strictly localizing `GPT-4o`'s input to the identified span. Subsequently, a Category-Aware Prompting strategy ensures the paraphrase preserves the syntactic structure dictated by the encoder's predicted category.

Specifically, we leverage the fine-grained category predicted by the encoder to construct dynamic Few-Shot Prompts tailored to each MWE type (`VID`, `NID`, `AdjID`). This categorization enables us to retrieve and inject contextually relevant examples, demonstrating, for instance, how to preserve tense in verbal idioms versus how to handle gender agreement in adjectival constructions. This optimization is intrinsically dependent on the structural scaffolding provided by Subtask 1; without the encoder's precise categorization, the system would be forced to rely on generic instructions, effectively forfeiting the performance gains derived from grammatically targeted demonstrations.

To emulate the semantic flexibility inherent in human paraphrasing, and drawing from the annotation guidelines, we designed two distinct prompt variants for each MWE category. The Minimal Strategy enforces strict lexical substitution to maintain the original sentence structure and semantic fidelity, whereas the Creative Strategy encourages broader structural reformulations to explore the model's generative flexibility beyond simple lexical substitution. The prompts were engineered in Romanian to minimize cross-lingual interference (see Appendix C for the complete list).

5 Results

We present the evaluation of our system in three phases: (1) a comparative benchmark of Transformer encoders to select the optimal backbone for Subtask 1, (2) a focused investigation into the capabilities of LLMs versus specialized encoders and (3) the official results obtained on the blind test set for both Identification and Paraphrasing.

5.1 Encoder Selection

Our initial experiments on the validation set compared five Transformer architectures to determine the optimal backbone for sequence labeling. Figure 1 presents the comparative results sorted by F1 score.

`RoBERT-base` secured the top F1 (0.903), validating monolingual specialization for Romanian. However, unlike older baselines like `XLM-RoBERTa` (0.890) which lagged significantly, the modern `mDeBERTa-v3` effectively closed the performance gap (F1: 0.901). As detailed in Figure 1, `mDeBERTa-v3` achieved the benchmark's highest Recall (0.919), demonstrating that modern architectural advancements can compensate for the lack of language specificity.

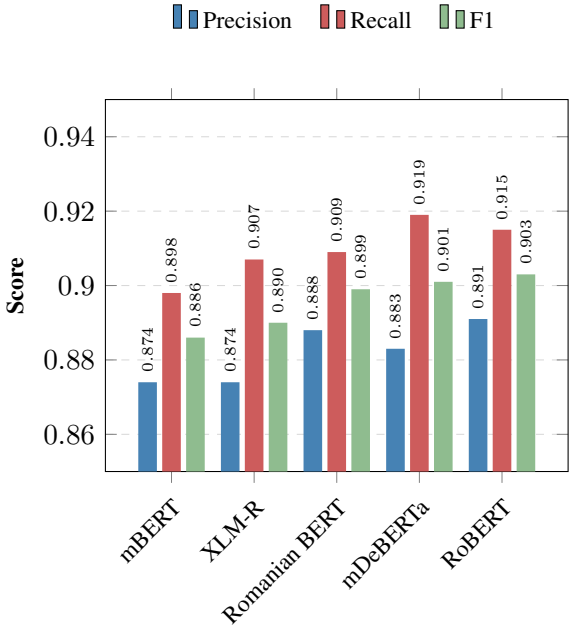


Figure 1: Comparative results of Transformer encoders sorted by F1 score

5.2 Few-Shot LLM vs. Fine-Tuned Encoders (Stratified Subset)

We evaluated the models on the stratified subset ($N = 170$) designed to stress-test specific linguistic phenomena. As detailed in Appendix Table 5, the specialized encoders consistently outperformed the few-shot LLM approach.

Fine-tuned encoders consistently dominated the benchmark, with all evaluated architectures achieving Soft F1 scores exceeding **54.4%**, significantly surpassing GPT-4o's **31.3%**. The gap is even more pronounced in the Strict evaluation, where every encoder maintained performance above **17.2%** compared to the LLM's **7.0%**. This disparity confirms that while the LLM captures the underlying semantics, it lacks the precision required for exact token-level extraction.

Memorization vs. Generalization The breakdown by category reveals a critical limitation of the

LLM. On "Rare" idioms (e.g., *LVC.cause*), GPT-4o failed completely (F1: 0.0%), whereas encoders maintained a baseline capability (F1: 10.0%). GPT-4o's performance spiked only on "Frequent" expressions (F1: 40.5%), suggesting it relies on memorizing common collocations rather than detecting the underlying syntactic structure of the MWE.

5.3 Unseen Test Set

The system's generalization capability was validated on our internal unseen subset ($N = 85$), as shown in Figure 2.

Consistent with the stratified analysis, the fine-tuned encoders maintained their superiority. Under the Soft evaluation, `mDeBERTa-v3` and `RoBERT-base` achieved F1 scores of **46.9%** and **46.5%** respectively, whereas GPT-4o trailed significantly at **32.9%**.

The breakdown in Figure 2 indicates that the LLM's primary weakness lies in Recall (27.6% Soft), missing less salient MWEs. However, its Precision remains competitive (40.7%), suggesting that identified expressions are generally correct, albeit with imprecise boundaries.

5.4 Official Shared Task Results

Although our system was specialized exclusively for Romanian, it was evaluated in the global PARSEME 2.0 leaderboard alongside massive multilingual systems. We focus our analysis on the Romanian language track, where our approach demonstrated decisive superiority.

State-of-the-Art on Romanian As presented in Table 1, our system (`romanian-bert`) secured the **Rank 1** position for Romanian, achieving a Global MWE-based F1 score of **85.65%**. We outperformed complex multilingual architectures such as `BeeParser` (F1: 83.60%) and `MTLB-STRUCT` (F1: 82.27%). This result serves as a strong validation for language-specific pretraining, confirming that a classic, fine-tuned Romanian encoder is sufficient to outperform highly engineered multilingual parsers without the need for structural complexity.

Handling Discontinuity A major challenge in Romanian is the high frequency of discontinuous expressions (e.g., intervening syntactic constituents). Our BIO-based linearization strategy proved highly effective for this structural complexity. In the global rankings, our system placed higher on *Discontinuous MWEs* (Rank 6) than on *Continuous MWEs* (Rank 8) relative to other participants. This suggests that explicit boundary encoding (B/I-tags) is particularly adept at bridging long-distance dependencies, a capability often diluted in generalist multilingual parsers.

Generalization to Unseen Data Crucially, our system also secured **Rank 1** on the "Unseen" subset for Romanian (F1: 16.00%), significantly surpassing the next best BeeParse (F1: 9.88%), and MTLB-STRUCT (F1: 4.82). This validates that the model learned compositional patterns rather than merely memorizing the training lexicon.

Table 2 shows the trade-off between semantic adherence and lexical novelty.

The **Minimal** strategy validated its role for strict simplification, achieving the highest Semantic Fidelity (Avg. BERTScore: **89.25**).

In contrast, the **Creative** strategy successfully induced stylistic variation, evidenced by a massive surge in Richness (+167% unique terms) and higher Entropy (6.13). This confirms that the model generated more diverse and unpredictable formulations. However, this freedom comes with a trade-off: a 4.5-point drop in BERTScore, reflecting the natural semantic drift inherent in structural reformulation.

6 Conclusion

This paper presented the "Archaeology" system for PARSEME 2.0, focusing on the processing of MWEs in Romanian. Our work highlights a fundamental dichotomy in NLP architecture: the need for structural rigidity in identification versus semantic fluidity in paraphrasing.

For Subtask 1, we demonstrated that fine-tuned Transformer encoders remain the optimal solution for token-level extraction. Our specialized monolingual model (`RoBERT-base`) achieved the top rank for Romanian (F1: 85.65%), proving particularly effective at resolving discontinuous dependencies where generative baselines struggled. Furthermore, our benchmarks reveal that while modern multilingual encoders like `mDeBERTa-v3` are effectively closing the performance gap, few-shot LLM prompting still lacks the precision required for strict boundary detection.

For Subtask 2, we showed that LLMs (GPT-4o) can be effectively harnessed through a constrained pipeline. By anchoring generation to encoder-predicted spans and employing a multi-tiered prompting strategy, we successfully balanced semantic fidelity with lexical diversity.

System	P	R	F1	Rank
Ours (RoBERT-base)	**91.03**	80.88	**85.65**	**1**
BeeParser	84.98	**82.27**	83.60	2
MTLB-STRUCT	83.71	80.88	82.27	3
Sahara-Tokenizers	61.98	71.12	66.23	4

Table 1: Official results on the Romanian test set on Subtask 1(Global MWE-based)

Prompt Strategy	Semantic Fidelity Avg. BERTScore	Lexical Diversity Richness	Lexical Diversity Evenness	Lexical Diversity Entropy
GPT-Minimal	**89.25**	235	**0.98**	5.36
GPT-Creative	84.73	**628**	0.95	**6.13**

Table 2: Official results on the Romanian test set on Subtask 2

Limitations and Ethical Considerations

While our hybrid architecture establishes a new state-of-the-art for Romanian, we acknowledge specific constraints. Structurally, the sequence labeling component exhibits a performance gap in detecting Single-Token MWEs. We attribute this to a combination of factors: the loss of lexical co-occurrence signals (which reduces the task to unassisted Word Sense Disambiguation) and a frequency bias inherent in the training distribution, where multi-token spans are overwhelmingly dominant. Furthermore, the system suffers from acute data sparsity in long-tail categories, where the encoder lacks sufficient supervision to generalize beyond memorized instances (see Appendix 6).

Ethically, we recognize the environmental costs and reproducibility challenges associated with large proprietary models like GPT-4o. However, our experimental design deliberately utilized GPT-4o solely to establish a theoretical upper bound for identification, demonstrating that even massive architectures fail at structural precision without guidance. Crucially, our proposed decoupling of identification from generation effectively lowers the reasoning barrier for the paraphrasing subtask. By offloading the structural heavy lifting to a lightweight encoder, our pipeline enables the future deployment of smaller, open-source, and environmentally efficient models for generation. Thus, the system is designed to reduce reliance on massive compute, as it no longer requires the LLM to function as a structural parser, but merely as a controlled rewriter.

Acknowledgements

This work was supported by the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology), and by the Romanian National Research Council (CNCS) through the Executive Agency for Higher Education, Research, Development and Innovation Funding (UEFISCDI) under grant PN-IV-P2-2.1-TE-2023-2007 InstRead.

References

Sweta Agrawal and Marine Carpuat. 2024. Do text simplification systems preserve meaning? a human evaluation via reading comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448.

Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025)*, pages 116–130, Suzhou, China. Association for Computational Linguistics.

Fabian Anghel, Cristea Petru-Theodor, Claudiu Creanga, and Sergiu Nisioi. 2025. RALS: Resources and baselines for Romanian automatic lexical simplification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31481–31492, Suzhou, China. Association for Computational Linguistics.

Andrei Avram, Verginica Barbu Mititelu, and Dumitru-Clementin Cercel. 2023. Romanian multiword expression detection using multilingual adversarial training and lateral inhibition. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 7–13, Dubrovnik, Croatia. Association for Computational Linguistics.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing*.

Verginica Barbu Mititelu, Mihaela Cristescu, and Mihaela Onofrei. 2019. The Romanian corpus annotated with verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 13–21.

Verginica Barbu Mititelu, Voula Giouli, Gražina Korvel, Chaya Liebeskind, Irina Lobzhanidze, Rusudan Makhachashvili, Stella Markantonatou, Aleksandra Markovic, and Ivelina Stoyanova. 2025. The challenges of syntactic descriptions of multiword expressions in electronic lexicography. In *eLex 2025: Electronic Lexicography in the 21st Century*, pages 1–20. Lexical Computing CZ s.r.o. Paper ID: 17; 16–18 Nov 2025, Brno, Czech Republic.

Chutima Bunparit and Pennapa Riabroi. 2025. Effects of narrow reading on the reading comprehension, vocabulary acquisition, and perceptions of l2 students in an esp classroom. *LEARN Journal: Language Education and Acquisition Research Network*, 18(2):571–593.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Theo Guidroz, Diego Ardila, Jimmy Li, Adam Mansour, Paul Jhun, Nina Gonzalez, Xiang Ji, Mike Sanchez, Sujay Kakarmath, Mathias MJ Bellaiche, Miguel Ángel Garrido, Faruk Ahmed, Divyansh Choudhary, Jay Hartford, Chenwei Xu, Henry Javier Serrano Echeverria, Yifan Wang, Jeff Shaffer, Eric, and 8 others. 2025. Llm-based text simplification and its effect on user comprehension and cognitive load. *Preprint*, arXiv:2505.01980.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Carp, and Vasile Pais. 2026. Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions.

Shiva Taslimipoor, Sara Bahaadini, and Ekaterina Kochmar. 2020. MTLB-STRUCT @parseme 2020: Capturing unseen multiword expressions using multitask learning and pre-trained masked language models. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 142–148, online. Association for Computational Linguistics.

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Weihang You, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, Junhao Chen, and Tianming Liu. 2026. Opportunities and challenges of large language models for low-resource languages in humanities research. *Preprint*, arXiv:2412.04497.

A Experimental Setup

Subtask 1 We fine-tuned the pre-trained encoder models using the Hugging Face library. Given the structural complexity of the task specifically the need to resolve discontinuous MWEs via BIO tagging we adopted a training horizon of 10 epochs. This extended schedule, coupled with a linear learning rate decay, allowed the model to converge on low-frequency MWE classes that typically require more gradient updates than standard continuous entities.

We utilized a batch size of 16 and a learning rate of 2×10^{-5} with a 10% warmup period. The optimization was performed using AdamW. The complete set of hyperparameters is detailed in Table 3.

Hyperparameter	Value
Encoder Architecture	RoBERT-base
Optimizer	AdamW
Learning Rate	2×10^{-5}
LR Scheduler	Linear Decay
Batch Size	16
Training Epochs	10
Warmup Ratio	0.1
Max Sequence Length	256

Table 3: Hyperparameters used for fine-tuning the sequence labeling models (Subtask 1)

Subtask 2 For the generation phase, we utilized the `GPT-4o` model via the OpenAI API, specifically targeting the snapshot `gpt-4o-2024-08-06` to ensure reproducibility. We fixed the sampling temperature at $\tau = 0.6$. We employed the 'system' role to enforce the linguistic persona constraints. The configuration details are summarized in Table 4.

Parameter	Value
Model ID	`gpt-4o-2024-08-06`
API Endpoint	Chat Completions
Temperature (τ)	0.6
Top-p	1.0 (default)
Frequency Penalty	0.0 (default)
System Prompt	Enabled

Table 4: Configuration details for the Generative Component (Subtask 2)

B Discussion and Error Analysis

Our decision to prioritize a standard `RoBERT-base` architecture is empirically supported by Avram et al. (2023). They demonstrated that for Romanian MWEs, specialized monolingual fine-tuning outperforms complex techniques like multilingual adversarial training. However, our experiments with `mDeBERTa-v3` add a modern nuance to this finding. While previous multilingual baselines (e.g., XLM-R) lagged behind, we observe that modern architectures have effectively closed the specialization gap, achieving performance nearly identical to `RoBERT-base` without requiring auxiliary losses. Consequently, we directed our efforts towards integrating these robust detectors into a generative pipeline for Subtask 2.

The discrepancy between our top-tier performance on discontinuous MWEs and lower performance on specific sub-categories warrants a deeper analysis of the underlying data distribution and tagging limitations.

The Single-Token Bottleneck. While the BIO scheme excelled at capturing multi-token spans, it showed limitations when predicting isolated idiomatic tokens (Global Rank 10). This suggests that the model relies heavily on the "contextual width" of an expression. In the absence of a multi-token span, the encoder loses the strong structural signal usually provided by the attention mechanism across multiple phrase components.

This is best exemplified by the token *"varză"* (lit. cabbage). When part of a longer idiom (e.g., *"a face varză"* - to mess up), the verb *face* acts as a contextual anchor. However, when *"varză"* appears alone (meaning *chaotic*), the model must rely purely on subtle semantic cues. Without structural reinforcement, the distinction between the literal and idiomatic senses becomes blurry for the encoder.

Impact of Data Scarcity A granular analysis of the Romanian test results reveals that errors are heavily concentrated in "long-tail" categories. For instance, the model achieved **0.0 F1** on `NV.VID`. This correlates directly with extreme data scarcity: `NV.VID` appears only 98 times in the entire training corpus of 1.27 million tokens (approx. 0.007% frequency). In contrast, frequent categories like `AdpID` (16,509 training examples) were detected with **90.0% F1**. This confirms that the encoder's performance is strictly bounded by the density of category representation in the fine-tuning data.

We propose that future research should explicitly prioritize discontinuous MWEs and unseen expressions, identifying them as the primary barriers currently limiting system robustness. Mastering the syntactic elasticity of discontinuous structures and the compositional reasoning required for unseen data rather than optimizing for memorized, continuous spans is the necessary step to bridge the gap between simple sequence labeling and true language understanding.

Table 7 presents the exact templates used for the *Minimal* and *Creative* strategies across different MWE categories.

Due to space constraints, we present the structural template common to all prompts and contrast the specific instructions used for the two paraphrasing strategies. The full prompts for all categories (VID, NID, AdjID) follow this architectural pattern.

Data Usage Our experiments rely exclusively on the PARSEME 2.0 Shared Task dataset, which is a publicly available, anonymized corpus. No private or personally identifiable information was processed or generated during this study.

C Prompt Templates

Although the experiments were conducted using prompts strictly engineered in Romanian (to prevent cross-lingual artifacts), we present here the English translations of the structural templates and key instructions for clarity.

C.1 The Anatomy of a Prompt

To ensure consistent parsing and minimize hallucinations, all prompts share a rigid architectural skeleton comprising five enforced components:

1. **Persona Definition:** Establishes the role of an expert linguist specialized in semantics.

2. **Task Specification:** Defines the specific MWE category (VID/NID/AdjID) and input format markers (double brackets).

3. **Strategy Constraints:** Specific rules for the paraphrasing style. This block controls the divergence between system behaviors (as detailed in Table 7).

4. **Negative Constraints (Critical):** Explicit penalties for hallucinating boundaries or retaining original tokens (e.g., *"If ALL bracketed tokens appear in output → Score 0"*).

5. **Few-Shot Examples:** A set of 4-5 input-output pairs demonstrating the desired transformation logic.

C.2 Strategy Differentiation

While the skeleton remains constant, the divergence between the *Minimal* and *Creative* behaviors is controlled exclusively via the instruction block (Component 3). The contrasting instructions are presented in Table 7.

Few-Shot Formatting Example To guide the model's reasoning, we provided examples following a specific "Input → Token Identification → Output" format. To ensure clarity for non-Romanian speakers, English translations are provided below in parentheses.

Ex. 1 (VID Minimal):
Input: Ion [[a dat ortul popii]] ieri dimineață.
(En: Ion [[kicked the bucket]] yesterday morning.)
Tokens MWE: {a, dat, ortul, popii}
Parafraza: Ion a murit ieri dimineață.
(En: Ion died yesterday morning.)

Ex. 2 (AdjID Creative):
Input: Fratele meu este mereu [[cu capul în nori]].
(En: My brother is always [[with his head in the clouds]].)
Tokens MWE: {cu, capul, în, nori}
Parafraza: Fratele meu este mereu dus cu pluta, parcă trăiește pe altă planetă.
(En: My brother is always spaced out, as if he lives on another planet.)

Model	Overall		Rare		Medium		Frequent		Discont		Dense	
	Strict	**Soft**	**Strict**	**Soft**	**Strict**	**Soft**	**Strict**	**Soft**	**Strict**	**Soft**	**Strict**	**Soft**
Fine-tuned Encoders												
RoBERT-base	17.5	55.6	6.7	10.0	12.5	22.7	19.4	66.5	0.4	16.1	17.3	58.3
Romanian BERT	17.4	55.8	6.7	10.0	12.6	22.8	19.2	66.6	0.4	16.1	17.2	58.6
mBERT	17.2	55.0	6.7	10.0	11.7	21.1	19.3	66.3	0.5	17.0	17.0	57.5
XLM-R	17.2	55.2	6.7	10.0	11.6	21.6	19.3	66.3	0.4	16.3	17.1	58.0
mDeBERTa-v3	**17.6**	54.4	**6.7**	**10.0**	11.5	20.7	**19.9**	65.7	0.4	**16.4**	**17.6**	57.1
Generative LLM (Few-shot)												
GPT-4o	7.0	31.3	0.0	0.0	0.8	13.1	9.8	40.5	**0.6**	11.3	7.0	33.8

Table 5: Complete performance breakdown on the Stratified Stress Test (N=170). Scores are reported as F1 (%). The *Strict* metric requires exact boundary matching, while *Soft* allows for partial overlap. Fine-tuned encoders demonstrate significantly higher recall on specific MWE categories, whereas the LLM struggles with Rare and Medium idioms in Romanian.

MWE Category	MWE-based			Token-based		
	P	**R**	**F1**	**P**	**R**	**F1**
AV.IAV	100.0	66.67	80.00	100.0	66.67	80.00
AdjID	100.0	79.25	88.42	100.0	79.82	88.78
AdpID	97.30	83.72	90.00	98.05	84.83	90.96
AdvID	80.95	76.12	78.46	81.94	76.13	78.93
ConjID	84.38	93.10	88.52	87.14	93.85	90.37
DetID	100.0	100.0	100.0	100.0	100.0	100.0
IAV	86.49	57.14	68.82	94.59	54.69	69.31
IRV	84.00	85.71	84.85	86.00	86.87	86.43
IntjID	100.0	100.0	100.0	100.0	100.0	100.0
LVC.full	75.00	60.00	66.67	100.0	70.00	82.35
NID	95.00	88.79	91.79	96.14	89.96	92.95
NV.LVC.cause	100.0	100.0	100.0	100.0	100.0	100.0
NV.VID	0.00	0.00	0.00	0.00	0.00	0.00
PronID	100.0	100.0	100.0	100.0	100.0	100.0
VID	90.91	75.00	82.19	97.53	75.24	84.95

Table 6: Official evaluation results for the Shared Task on the blind test set. The table reports Precision (P), Recall (R), and F1-scores for both MWE-based (strict per-expression) and Token-based (per-token) evaluation metrics across all MWE categories.

Minimal Strategy (Translation)	**Creative Strategy (Translation)**
• Modify as few words as possible outside the MWE span.	• Reorganize the sentence structure completely (e.g., active $\leftrightarrow$ passive).
• Strictly preserve verb tense, person, number, and voice.	• Use distinct metaphors or idioms if contextually appropriate.
• Do NOT use Light Verb Constructions (LVCs) as replacements.	• Change the narrative perspective or add explanatory context.
• **Constraint:** Do not replace the MWE with another MWE.	• **Goal:** Maximize lexical diversity and structural novelty.

Table 7: Contrastive instructions injected into the system prompt. The *Minimal* set enforces substitution, while the *Creative* set encourages rewriting.

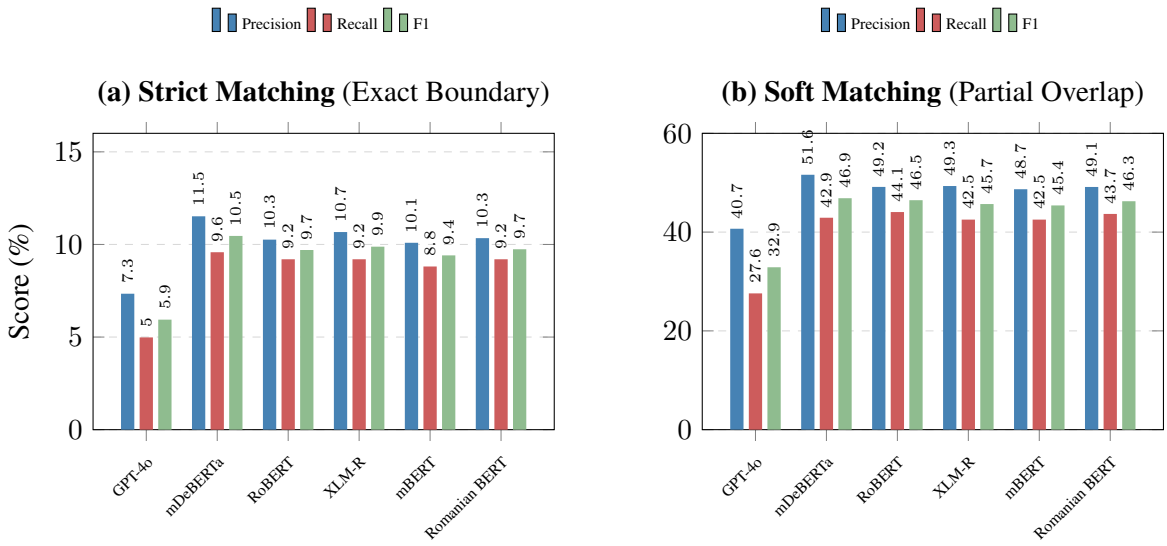


Figure 2: Global performance on our internal unseen subset (N=85). **(a)** Shows metrics under the Strict evaluation scenario (exact match), where all models struggle significantly. **(b)** Shows metrics under the Soft evaluation scenario, where fine-tuned encoders (like mDeBERTa and RoBERT) clearly outperform the few-shot GPT-4o baseline, particularly in Recall.

ITUNLP2 at MWE-2026 AdMIRe 2: Modular Zero-Shot Pipelines for Multimodal Idiom Grounding and Ranking

Özge Umut and Bora Şenceylan
Istanbul Technical University
Türkiye
{umut24,senceylan19}@itu.edu.tr

Abstract

We describe a zero-shot system for AdMIRe 2, a shared task on multimodal understanding of potentially idiomatic expressions (PIEs) (Arslan et al., 2026). Given a context sentence with a PIE and five candidate images, the system predicts whether the usage is literal or idiomatic and ranks the images based on how well they match the intended meaning. We use closed-source large multimodal models and compare prompting pipelines from direct one-step ranking to modular multi-step pipelines that separate sense prediction, PIE-focused image semantics, and final ranking. All steps produce constrained JSON outputs to enable deterministic parsing and composition. In the official AdMIRe 2 evaluation on CodaBench, our best pipeline achieves an average Top-1 accuracy of 0.52 and an average nDCG score of 0.70 across the 12 languages we submitted. We obtain the best score among submitted systems in 10 of these languages (cod, 2026).

1 Introduction

Idioms are hard because the intended meaning is often not compositional. In AdMIRe 2, each example provides a context sentence containing a potentially idiomatic expression (PIE) and five candidate images; systems must decide whether the PIE is literal or idiomatic and rank the images by semantic match (Arslan et al., 2026). This setting links sense disambiguation to grounded retrieval, and errors can come from either the meaning decision or the fine-grained ordering among several plausible images.

Recent prompting work shows that forcing intermediate steps can improve reliability on tasks that require multi-step reasoning (Wei et al., 2022; Kojima et al., 2022), and similar ideas have been applied to multimodal reasoning (Zhang et al., 2023). Following this direction, we build modular pipelines where the model first commits to a PIE sense, then produces PIE-centered image semantics, and only then produces a final ranking. We evaluate multiple pipeline variants with OpenAI-o3 and Gemini-3-pro-preview, and we analyze when ordering is the main remaining failure mode. Complete code for our pipeline variants is publicly available [1] .

Our contributions are:

- a set of zero-shot modular pipelines for multimodal idiom grounding, implemented with structured JSON interfaces;
- an ablation-style comparison across pipeline depths and aggregation strategies;
- official multi-language evaluation results from the AdMIRe 2 CodaBench phase (cod, 2026).

2 Related Work

Idiomaticity has been studied widely in text-only settings, including idiom corpus construction tasks (Eryigit et al., 2021), (Arslan et al., 2025) and multilingual idiom detection shared tasks (Tayyar Madabushi et al., 2022). AdMIRe extends this line by requiring multimodal grounding: a PIE in context must be matched to images that represent either idiomatic or literal meaning (Pickard et al., 2025), and many strong systems use multi-stage designs rather than a single direct ranking.

Several AdMIRe system papers explicitly separate sense disambiguation from visual ranking. AlexUNLP-NB predicts literal vs. idiomatic usage, derives a literalized meaning signal, and then performs retrieval-style ranking (Badran et al., 2025). PALI-NLP refines image descriptions with PIE-relevant details and applies a revision step before ranking (You et al., 2025). Other approaches strengthen the retrieval backbone with fine-tuning

[1] GitHub repository for pipeline codes

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 248–253

or ensembling (Wang et al., 2025). For the vision–language similarity component, CLIP (Radford et al., 2021) and newer contrastive objectives such as SigLIP (Zhai et al., 2023) are common choices, but performance still depends strongly on the text representation used for matching. Our work stays in a training-free setting and focuses on how step-wise prompting and intermediate semantic signals affect ranking quality.

3 Methodology

3.1 Dataset

To design and evaluate our pipelines, we used the SemEval-2025 Task 1 (AdMIRe) dataset (Pickard et al., 2025) which includes examples in both English and Portuguese. Each example consists of a context sentence containing a potentially idiomatic expression (PIE) and 5 candidate images encompassing : Idiomatic Synonym, Idiomatic Related, Literal Synonym, Literal Related and a Distractor. The English subset comprises 200 sentences covering 100 unique PIEs, while the Portuguese subset contains 110 sentences with 55 unique PIEs. After finalizing the pipeline, final evaluation was conducted using the AdMIRe 2 blind test set via CodaBench (cod, 2026), which extends the task to 15 diverse languages (Torunoğlu-Selamet et al., 2026).

3.2 System Overview

For addressing the challenge of ranking images based on their relevance to a context sentence containing a PIE, we employed a zero-shot framework. Analysis of earlier SemEval-2025 Task 1: AdMIRe - Advancing Multimodal Idiomaticity Representation results indicated that closed-source multimodal models demonstrated strong performance. Therefore, we used Gemini-3-pro-preview and OpenAI-o3 for our experiments. As it was demonstrated in recent research into Large Multimodal Models (LMMs) (Khot et al., 2023; Khan et al., 2023), decomposing complex tasks into modular sub-tasks significantly improves performance in both text-only and vision-language settings. Consistent with these studies, most participating systems in SemEval Task 1 adopted multi-step reasoning pipelines. Our methodology builds upon this principle of modular reasoning by decomposing the reasoning process of the models into discrete, structured steps iteratively, improving image-text matching accuracy. We developed and evaluated seven distinct pipelines, progressively increasing the granularity of the reasoning process. For each step of the pipelines, all model outputs were constrained to JSON format with pre-defined fields depending on its sub-task to ensure reproducible data extraction and pipeline integration. Additionally, for each sub-step, models were prompted to provide explanations of its reasoning for analytical purposes. The explanation of each pipeline is given below:

3.2.1 Pipeline A: One Step Direct Ranking (Baseline)

In this single-step approach, the model is given the context sentence, the PIE, and the five candidate images simultaneously and asked to directly output a ranked list and an explanation of its reasoning. The results of this pipeline are utilized to assess the impact of breaking the task into sub-tasks on performance.

3.2.2 Pipeline B: Two-step (Sense Prediction + Ranking)

We next introduce a two-step pipeline to explicitly see how model interprets the PIE in the context sentence. In the first step model is asked to predict whether the PIE is used literally or idiomatically in the given context sentence. In the second step, model is asked to rank the images utilizing the predicted sense from step 1. In each step model is prompted to provide a brief explanation on its reasoning.

When these explanations are inspected in detail, it was noticed that during image interpretation, models tend to over-analyze other aspects of the sentence, rather than PIE itself. For example, in the sentence “Lyn says that her relationship with Paul is ancient history, but Steph thinks that she should go and see him to give him a chance to apologize", the model ranked an image depicting a couple holding hands last (as distractor), justifying the decision by noting that it showed an ongoing relationship. This reasoning reflects that model paid attention to the general relationship theme more than PIE "ancient history". To mitigate this, we introduced next three-stage modular pipeline:

3.2.3 Pipeline C: Three-step (Sense Prediction + Image Semantics + Ranking)

As in pipeline B, the model was asked to identify literal or idiomatic usage. In step 2, with the aim of the model to focus only on the PIE, given each im-

age and their captions, we ask the model to categorize them into five predefined categories: Idiomatic Synonym, Idiomatic Related, Literal Synonym, Literal Related, Distractor; provided only the PIE, not the context sentence. In the third step the model receives the outputs from steps 1 and 2 and asked to produce the final ranking. This stage acts as an aggregator, weighing the specific PIE-to-image classification against the broader sentence context.

3.2.4 Pipeline D: Three-step 2 (Sense Prediction + Image Semantics + Manual Ranking)

This pipeline extends the three-step approach by replacing model-based ranking with a rule-based manual ranking strategy in the final stage. The first and second steps are the same as the previous pipeline, in the third step images are ranked deterministically based on following rule: If the sentence is predicted as literal, images are ordered as: literal synonym, literal related, idiomatic related, idiomatic synonym, distractor. If the sentence is predicted as idiomatic, images are ordered as: idiomatic synonym, idiomatic related, literal related, literal synonym, distractor. The objective of this pipeline is to determine whether aggregating the outputs of the first two steps using explicit rule yields better performance than relying on an LLM to perform the final ranking.

3.2.5 Pipeline E: Four-Step 1 (Caption Refinement and Extension)

This pipeline introduces an explicit image caption refinement stage prior to image classification step to test whether richer visual descriptions improve the image semantic classification. Models were asked to expand the original caption into a more detailed description and add concrete visual details. The rest of the steps are the same as pipeline C.

3.2.6 Pipeline F: Four-Step 2 (Explicit PIE–Image Grounding)

This pipeline focuses on strengthening explicit semantic grounding between the PIE and each image before ranking. In step 1, model is asked to predict whether the PIE is used literally or idiomatically, in step 2 model explicitly explains the relation between the PIE and each image, in step 3, using the explanations generated in the previous step, each image is assigned to its semantic category. In the last step, the model is asked to rank the images based on the information from previous steps.

3.2.7 Pipeline G: Four-Step 3 (Explicit Grounding + Manual Ranking)

This pipeline mirrors the previous four-step approach but replaces model-based ranking with manual rule-based aggregation as described in pipeline D.

3.3 Evaluation Metrics

Each pipeline is evaluated using Top-1 Accuracy, Exact Match Accuracy, Top-2 Accuracy and Discounted Cumulative Gain (DCG) (Järvelin and Kekäläinen, 2002). Top-1 Accuracy measures the proportion of instances in which the most relevant image is predicted correctly, Top-2 Accuracy measures how often the images predicted in the first and second positions are correct. Exact-Match Accuracy requires the entire predicted ranking of images to exactly match the ground-truth order. It measures how well the model orders all images correctly. Normalized Discounted Cumulative Gain used with weights: [3, 1, 0, 0, 0] to assess overall ranking quality and break ties.

4 Results

The experimental results for our different pipeline configurations in the English portion of the SemEval-2025 Task 1 (AdMIRe) dataset (Pickard et al., 2025) with OpenAI-o3 and Gemini-3-pro-preview are summarized in Tables 1 and 2. Overall, the results show that multi-step pipelines consistently outperform simpler 1-step and 2-step approaches, supporting the previous research demonstrating decomposing complex tasks into modular sub-tasks significantly improves performance in vision-language settings.

OpenAI-o3 Results: For the OpenAI-o3 model (Table 1), the 1-step baseline (Pipeline A) achieved a Top-1 Accuracy of 0.82. This was exceeded by all modular pipelines, with the Pipeline F: 4-step 2 approach reaching the highest Top-1 score of 0.90, followed closely by Pipeline C: 3-step 1 at 0.89. A similar trend appeared in the Discounted Cumulative Gain (DCG) scores, which rose from 2.83 (Pipeline A: 1-step) to a peak of 3.16 (Pipeline C: 3-step 1).

Due to the small performance difference between the 3-step and 4-step pipelines, and considering the significantly higher computational cost and latency of 4-step pipelines, we focused our experiments with Gemini-3-pro-preview on 1-step and 3-step configurations.

Table 1: Zero-Shot Evaluation Results for OpenAI-o3 Model

Metric	Pipeline A	Pipeline B	Pipeline C	Pipeline D	Pipeline E	Pipeline F	Pipeline G
Top-1 Acc.	0.82	0.83	0.89	0.84	0.88	**0.90**	0.87
Exact-Match Acc	0.09	0.09	0.235	**0.63**	0.21	0.14	0.61
Top-2 Acc.	0.60	0.60	**0.76**	0.75	0.74	0.72	0.75
Average DCG	2.83	2.87	**3.16**	2.99	3.11	3.15	3.06

Gemini-3-pro-preview Results: As shown in Table 2, the performance improvements from multi-step decomposition are even more pronounced for Gemini than for OpenAI-o3; the Pipeline C: 3-step 1 configuration improved Top-1 Accuracy from 0.81 to 0.92. Overall, Gemini outperformed o3 in most metrics, except for exact-match accuracy in specific 3-step settings.

Table 2: Zero-Shot Evaluation Results for Gemini-3-pro-preview

Metric	Pipeline A	Pipeline C	Pipeline D
Top-1 Acc.	0.81	**0.92**	0.91
Exact-Match Acc.	0.08	0.04	**0.75**
Top-2 Acc.	0.60	0.78	**0.83**
Avg. DCG	2.79	3.26	**3.27**

Hybrid Evaluation: To investigate whether the two models provide complementary strengths, we also evaluated a hybrid zero-shot configuration (Table 3), using Gemini-3-pro-preview for sense prediction and image classification and OpenAI-o3 for final ranking. While this achieved a competitive Top-1 Accuracy of 0.904, it did not outperform the standalone Gemini 3-step pipeline.

Table 3: Hybrid Zero-Shot Results (Gemini + o3)

Metric	Score
Top-1 Acc.	0.90
Exact-Match Acc.	0.20
Top-2 Acc.	0.72
Avg. DCG	3.17

Cross-Lingual Transfer: To assess cross-lingual generalization, we applied the same 3-step pipeline to a Portuguese language split without modifying the overall approach. As shown in Table 4, Gemini again achieves higher Top-1 accuracy and DCG, while OpenAI-o3 performs slightly better on exact-match accuracy. The overall gap between models is relatively small, suggesting that ranking quality, rather than language-specific understanding, remains the primary issue.

The "Ordering" Challenge: Model vs. Heuristic Ranking: When overall results were analyzed for both models, it is observed that the exact-match accuracy is the most challenging metric across all experiments, as it requires the entire image ranking to be correct. In fully model-driven pipelines, exact-match remains low for both models. However, when manual rule-based ranking is introduced (Pipeline D: 3-step 2 and Pipeline G: 4-step 3), exact-match accuracy increases substantially. For example, OpenAI-o3 improves from 0.09 in the 1-step baseline to 0.63 in the 3-step 2 pipeline, while Gemini reaches 0.75 in the same setting. This sharp improvement indicates that many remaining errors arise not from incorrect PIE interpretation or image classification, but from ordering decisions among multiple partially relevant images.

Table 4: Zero-Shot Pipeline C: 3 step 1 Results on Portuguese Data

Metric	Gemini	o3
Top-1 Acc.	**0.88**	0.84
Exact-Match Acc.	0.05	**0.11**
Top-2 Acc.	**0.62**	0.49
Avg. DCG	**3.06**	2.86

Final System Selection: Based on these findings, we selected Gemini-3-pro-preview with the Pipeline C: 3-step 1 pipeline as our final system for generating predictions on the AdMIRe 2 competition test set via CodaBench. This configuration offered the best trade-off between performance and computational cost. In our official submission, we evaluated 12 languages; Greek, Serbian, and Slovak were not included because our final three-step setup requires three separate LLM calls per instance, and completing the full blind test across all 15 languages was not feasible within the submission deadline given API latency and rate limits. Our final submission achieved 3rd place overall in the competition and ranked 1st in 10 out of the 12 languages we participated in, indicating that the proposed modular reasoning setup performs reliably across languages. The complete results are presented in Table 5. Performance varies notably across languages. The system performs strongest

Table 5: Competition Results Across Languages

Metric	AVG	ZH	KA	IG	KK	NO	PT-BR	PT-PT	RU	SL	ES-EC	TR	UZ
Top-1 Accuracy	0.52	0.53	0.56	0.56	0.70	0.82	0.88	0.72	0.71	0.82	0.40	0.65	0.52
nDCG Score	0.70	0.81	0.82	0.84	0.89	0.94	0.96	0.91	0.91	0.94	0.79	0.88	0.83

on Brazilian Portuguese (PT-BR), Norwegian (NO), Slovenian (SL), where both Top-1 accuracy and nDCG exceed 0.80, suggesting robust cross-lingual generalization for these languages. In particular, PT-BR achieves the highest scores (Top-1: 0.88, nDCG: 0.96), indicating effective semantic alignment in this language. Worst performance is observed in Ecuadorian Spanish (ES-EC). The limited data for this language with fewer than 50 samples may explain its position as the lowest performer in the dataset.

5 Conclusion

We presented a zero-shot, modular prompting system for multimodal idiom grounding in AdMIRe 2 (Arslan et al., 2026). Across two closed-source multimodal models, multi-step pipelines improve Top-1 accuracy and ranking quality compared to direct one-step ranking. Manual aggregation experiments show that a large part of the remaining gap comes from ordering decisions among partially relevant images, not only from sense disambiguation. In the official evaluation phase on CodaBench, the selected pipeline obtained strong cross-lingual results across the 12 languages we entered (cod, 2026).

Limitations

Our approach relies entirely on closed-source large multimodal models in a zero-shot setting. As a result, the reproducibility of our pipelines depend on the availability, behavior, and pricing policies of external APIs. Differences in model versions, hidden system prompts, or parameter updates may affect performance over time.

The modular design increases the number of calls per instance. This improves performance in our experiments, but it also increases cost and latency, and it becomes harder to run large batches under API rate limits. For the official evaluation, these constraints affected our submission, and we evaluated 12 languages instead of all 15.

References

2026. Admire 2.0 shared task: Codabench competition results. https://www.codabench.org/competitions/10547/#/results-tab. Accessed: 2026-01-06.

Doğukan Arslan, Hüseyin Anıl Çakmak, Gulsen Eryigit, 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 Torunoğlu-Selamet, Thomas Pickard, Aline Villavicencio, Adriana S. Pagano, and Gülşen Eryiğit. 2026. MWE-2026 Shared Task 2: AdMIRe 2 - Advancing Multimodal Idiomaticity Representation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morocco. Association for Computational Linguistics.

Mohamed Badran, Youssef Nawar, and Nagwa El-Makky. 2025. AlexUNLP-NB at SemEval-2025 task 1: A pipeline for idiom disambiguation and visual representation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 546–550, Vienna, Austria. Association for Computational Linguistics.

Gülsen Eryigit, Ali Sentas, and Johanna Monti. 2021. Gamified crowdsourcing for idiom corpora construction. *CoRR*, abs/2102.00881.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Zaid Khan, Vijay Kumar BG, Samuel Schulter, Manmohan Chandraker, and Yun Fu. 2023. Exploring question decomposition for zero-shot VQA. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

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)*.

Dilara Torunoğlu-Selamet, Dogukan Arslan, Rodrigo Wilkens, Wei He, Doruk Eryiğit, Thomas Pickard, Adriana S. Pagano, Aline Villavicencio, Gülşen Eryiğit, Ágnes Abuczki, Aida Cardoso, Alesia Lazarenka, Dina Almassova, Amalia Mendes, Anna Kanellopoulou, Antoni Brosa-Rodríguez, Baiba Saulite, Beata Wojtowicz, Bolette Pedersen, and 59 others. 2026. A parallel cross-lingual benchmark for multimodal idiomaticity understanding. *Preprint*, arXiv:2601.08645.

Yanan Wang, Dailin Li, Yicen Tian, Bo Zhang, Jian Wang, and Liang Yang. 2025. dutir914 at SemEval-2025 task 1: An integrated approach for multimodal idiomaticity representations. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1198–1203, Vienna, Austria. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Runyang You, Xinyue Mei, and Mengyuan Zhou. 2025. PALI-NLP at SemEval-2025 task 1: Multimodal idiom recognition and alignment. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1211–1216, Vienna, Austria. Association for Computational Linguistics.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*.

Zhuosheng Zhang, Aston Zhang, and Mu Li. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Edition 2.0 of the PARSEME Shared Task on Multilingual Identification and Paraphrasing of Multiword Expressions

Manon Scholivet[1], Agata Savary[1], Carlos Ramisch[2], Eric Bilinski[1], Takuya Nakamura[1], Maria Mitrofan[3], Vasile Păiș[3]

[1]Paris-Saclay University, CNRS, LISN, Orsay, France,
[2]Aix Marseille Univ, CNRS, LIS, Marseille, France,
[3]RACAI, Romanian Academy, Romania

[1]first.last@universite-paris-saclay.fr, [2]first.last@lis-lab.fr, [3]first@racai.ro,

Abstract

Multiword expressions (MWEs) have been a major challenge in NLP for decades, and research on MWEs was driven notably by shared tasks, including those organized by the PARSEME community. We report the organisation and the results of edition 2.0 of the PARSEME shared task. For the first time, all syntactic categories are covered: verbal, nominal, adjectival, adverbial and functional. We rely on edition 2.0 of the PARSEME corpus, annotated for all these categories in 17 languages. We create a new dataset with paraphrases of sentences containing idioms in 14 languages, and define a new subtask dedicated to MWE paraphrasing. We extend our evaluation protocol by measuring both performance and diversity of systems, and including manual evaluation in paraphrasing. Ten systems participated in the MWE identification subtask and five in the paraphrasing subtask (baselines included). Results are promising, but known MWE identification challenges remain unsolved. Performance correlates positively with diversity in MWE identification, and negatively in MWE paraphrasing.

1 Introduction

Multiword expressions (MWEs) have been a major challenge in NLP for decades (Sag et al., 2002; Shwartz and Dagan, 2019). This is notably due to their prevalence in texts (Gross and Senellart, 1998; Candito et al., 2021), their partly regular and partly idiosyncratic behaviour (Gross, 1986, 1988; Savary et al., 2020), and their semantic non-compositionality (Mel'čuk, 2010). Many MWE tasks were addressed (Constant et al., 2017) and research has been boosted by SemEval shared tasks (Schneider et al., 2016; Tayyar Madabushi et al., 2022; Pickard et al., 2025; Arslan et al., 2026).

In this landscape, the PARSEME community has been carrying on long-standing efforts towards multilingual modelling of *verbal* MWEs, particularly challenging due to their morphosyntactic flexibility. The major outcomes have been verbal MWE annotation guidelines unified across 26 languages, manually annotated corpora for these languages (Savary et al., 2018, 2023) and 3 editions of a shared task on automatic identification of verbal MWEs (Savary et al., 2017; Ramisch et al., 2018, 2020).

However, MWEs come in all shapes and sizes (Baldwin and Kim, 2010). Beyond verbal MWEs (*to* ***pull*** *one's* ***leg***, *to* ***pay*** *a* ***visit***), there are numerous MWEs of other syntactic categories: functional (***by and large***, ***in that***, ***in spite of***), adjectival (***crystal clear***) adverbial (***by and large***) and nominal (***hot dog***).[1] Recently, PARSEME extended the scope of its guidelines to all these categories in the context of the UniDive COST Action (Savary et al., 2024).

This paper presents edition 2.0 of the PARSEME shared task building on these assets. It features two subtasks and represents a substantial extension of previous editions. The original contributions of this edition can be summarised as follows. First, the PARSEME 2.0 guidelines allowed the creation of MWE-annotated corpora in 17 languages, then used to evaluate systems on the identification of MWEs of *all categories* (subtask 1). Second, a corpus of sentences with idiomatic MWEs and their paraphrases has been created in 14 languages, allowing the evaluation of *paraphrasing systems* on verbal, nominal and adjectival idioms (subtask 2). Third, we propose metrics to assess the *diversity* of system results in both subtasks. Fourth, we rely on the *Codabench* platform to centralise evaluation. In this paper, we discuss related work (§2), the subtasks (§3), the underlying data (§4), the organisation (§5-6), the results (§7-8), and conclusions (§10).

2 Related Work

MWE Shared Tasks PARSEME 1.0 covered 18 languages and introduced the task of token-

[1]MWE examples follow the PMWE conventions (Markantonatou et al., 2021), enriched with colors and brackets.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 254–275

level MWE identification. PARSEME 1.1 covered 20 languages, introduced the CUPT format and phenomenon-oriented evaluation metrics. PARSEME 1.2 covered 14 languages and focused on unseen MWEs, with controlled splits, full UD integration, and companion raw corpora. The three previous editions only cover *verbal MWEs*. In addition to these PARSEME shared tasks, other evaluation campaigns covered idiomaticity and MWEs.

The SemEval-2016 Task 10 (DiMSUM) for English tested the detection of minimal semantic units, including MWEs, and their meanings (Schneider et al., 2016). The underlying corpus contained online customer reviews, tweets, and TED talks, and was notably annotated for 2 MWE classes: strong or weak, according to their degree of idiomaticity.

In SemEval-2022 Task 2 for English, Portuguese, and Galician, systems competed for two tasks (Tayyar Madabushi et al., 2022). In task A, given a sentence containing a potentially idiomatic expression and its span, systems should decide if the expression was used idiomatically or literally, both in zero- and one-shot settings. In Task B, given a sentence with a MWE and two other sentences in which the MWE was replaced by its paraphrase and by a distractor (formally close but semantically distant), systems had to decide which pair of sentences was closest in meaning.

Finally, SemEval-2025 Task 1 (AdMIRe) for English and Brazilian Portuguese was dedicated to multimodal idiomaticity representation (Pickard et al., 2025). The task's goal was to align images depicting MWEs having more or less figurative meanings with sentences containing the same expressions used literally or idiomatically. The task was extended to 15 languages in AdMIRe 2 (Arslan et al., 2026), co-organised by UniDive jointly with our shared task.

Paraphrasing Shared Tasks Automatic processing of paraphrase also has a rich state of the art. Butnariu et al. (2009) test systems for accurate scoring of alignments between English noun compounds and their potential paraphrases, e.g. *sleeping pill* vs. *pill that induces sleeping*. Hendrickx et al. (2013) extend the previous task to automatically produce a ranked list of paraphrases for a given English noun compound, e.g. *air filter* → *filter for air*, *filter that cleans the air*. Evaluation measures are based on approximate n-gram matching between the system-generated paraphrases and those produced by human experts, with rank-based scaling.

Later, the scope of paraphrasing was extended to whole sentences, with 3 subtasks. ***Paraphrase identification*** consists in binary classification of pairs of sentences as being paraphrases or not (Xu et al., 2015; Lan et al., 2017). ***Semantic textual similarity*** is defined as assigning sentence pairs a similarity score from 0 to 1 (Agirre et al., 2015; Xu et al., 2015). Finally, ***paraphrase generation*** consists in reformulating a sentence to use a different wording or structure but preserve the original meaning (Zhou and Bhat, 2021). Evaluation relies on measures from machine translation (ROUGE, BLEU, METEOR or TER) or human scoring along multiple dimensions such as similarity, clarity, or fluency.

MWE Paraphrasing MWEs are particularly challenging for paraphrasing due to their non-compositional semantics. In related work, one of the motivations behind paraphrasing MWEs with their literal equivalents is eliminating idiomaticity prior to machine translation, as done by Santing et al. (2022) for English-German MT. Dedicated MWE-aware paraphrase datasets were built upon MWE definitions in lexicons (Pershina et al., 2015; Liu and Hwa, 2016), collected by crowdsourcing (Yimam et al., 2016), for English in both cases, or relied on machine (back-)translation (Qiang et al., 2023), for Chinese. Many verbal MWEs can be paraphrased by a single verb, as shown by Barančíková and Kettnerová (2018) for Czech. Tan and Jiang (2021) adapt paraphrase identification to idioms, a task similar to disambiguating literal from idiomatic MWE uses. Zhou et al. (2021) introduce 2 tasks: ***idiomatic sentence generation*** transforms a literal sentence into a sentence involving idioms; and ***idiomatic sentence paraphrasing*** simplifies sentences so as to replace idioms with literal expressions. In the latter, the aim is to paraphrase only the MWE, leaving the rest of the sentence unchanged (Wada et al., 2023; Qiang et al., 2023). Evaluation metrics include ROUGE, BLEU, METEOR, GRUEN, BERT perplexity, as well as human judgements on semantics and fluency.

3 Task Definition and Metrics

Subtask 1: MWE Identification This historical PARSEME task focuses on token-level MWE identification in running text, as in previous editions. Systems are given as input a morphosyntactically analysed sentence in CoNLL-U format.[2] As output, they must group the tokens that belong to MWEs, assigning them a single label. For instance :

[2] https://universaldependencies.org/format

(1) **En plus**, ça **fait partie** du **centre ville** (fr)
In plus, it does part of.the centre city
'Moreover, it is part of the city centre.'

In the sentence above, the tokens belonging to the three MWEs (in bold) should be assigned unique labels, e.g. 1: (***En plus***), 2: (***fait partie***) and 3: (***centre ville***). Those not belonging to any MWE (*ça* and *du*) should not be assigned any label. Systems solving this task must address several challenges (Constant et al., 2017): discontinuities e.g. (fr) ***fait*** *toujours* ***partie*** 'is still part', morphological and syntactic variability as in (2), overlapping or nesting,as in (3), and idiomatic-literal ambiguity as in (4) vs. (5).

(2) a **da** un **sfat**, **sfaturi** au fost **date** (ro)
to give an advice, advices have been given

(3) **temos**$_{1,2}$ um **plano**$_1$ et uma **intenção**$_2$ (pt)
have.PL a plan and an intention
'We have a plan and an intention'

(4) この 問題 は **朝飯 前** だ (ja)
this problem about breakfast before be
lit. 'This problem is before breakfast.'
'This problem is very easy.'

(5) 朝飯 前 に 会う (ja)
breakfast before LOC meet
'We meet before breakfast.'

The corpora are provided in CUPT format.[3] They are split into training, development, and test sets. The latter are available only during the evaluation phase (about 1 week) and gold annotations are not disclosed. Test corpora are completely new with respect to previous editions to prevent LLM-contamination (§ 5).

Annotated MWEs are assigned category labels (e.g. NID for nominal idiom, MVC for multi-verb construction). While previous editions covered only verbal MWE categories, the current edition covers all MWE categories, including nominal, verbal, adjectival, adverbial, and functional MWEs (see § 4). These category labels can guide system development, but they are not taken into account in evaluation metrics. Thus, systems need to group tokens belonging to the same MWE, but they do not have to tag the resulting MWE with a specific category.

The evaluation of this subtask is performed using two standard F-score variants: MWE-based and token-based (Savary et al., 2017). The former accounts for exact matches between all tokens of the predicted MWE and of the reference MWE, whereas the latter rewards partial matches, covering only part of the tokens. In addition, we report phenomenon-specific F-scores, focusing on discontinuous, single-token, variant and unseen MWEs (Ramisch et al., 2018). Edition 1.2 focused on unseen MWEs, that is, those whose multi-set of lemmas are annotated as MWEs at least once in the test corpus, but never in the training or development corpus (Ramisch et al., 2020). In the current edition, we propose and analyse diversity scores that also partly account for unseen/novel identified MWEs.

Subtask 2: MWE Paraphrasing This novel subtask is motivated by recent advances in text generation. We wish to challenge modern generative systems with idiomaticity-related problems in more advanced scenarios than done so far (§ 2). Paraphrasing may be a useful method for testing the ability of models to grasp the meaning of an MWE (Tayyar Madabushi et al., 2022; He et al., 2025). MWE paraphrasing may also help for text simplification.

First, we address paraphrase generation rather than binary detection or similarity scoring. Second, paraphrasing is not restricted to the MWE itself but, conversely, reformulation of other parts of the sentence is encouraged and rewarded by diversity metrics. Third, our gold paraphrases are produced by native speakers along unified guidelines for an unprecedented number of 14 languages. Finally, we use LM-driven evaluation (BERT-score) and show its good correlation with human evaluation.

The input for this task is a raw sentence containing exactly one verbal, nominal or adjectival idiom, not explicitly marked in text.[4] Systems must paraphrase the sentence so that the original MWE no longer occurs, but the meaning is kept. For instance, sentence (6) could be paraphrased as (7) or as (8).

(6) le **point de vue** de la réalisatrice … (fr)
the point of vue of the director …
'the director's point of view …'

(7) la perspective de la réalisatrice … (fr)
the perspective of the director …
'the director's perspective …'

(8) la vision du metteur en scène … (fr)
the vision of.the putter in scene …
'the stage director's vision …'

Additionally, to facilitate automatic evaluation, at least one of the lemmas of the original MWE should

[3] https://gitlab.com/parseme/corpora/-/wikis/CUPT-format

[4] VID, NID or AdjID in the PARSEME typology (see § 4).

be totally absent from the paraphrase. For instance, (fr) ***peine de mort*** (lit. 'punishment of death') 'death penalty' should not be paraphrased as *peine consistant à causer la mort de la personne* 'punishment causing the death of the person'. We allow paraphrases to use MWEs, provided that they are different from the original one, as in (9)–(10).

(9) Dla nich świat **stanął w miejscu**. (pl)
For them world stood in place.
'For them the world stands still.'

(10) Dla nich świat przestał **się rozwijać**. (pl)
For them world stopped itself unroll.
'For them the world stopped developing.'

In subtask 2, only trial data in English and French is provided, but no training nor development data. The test data contains between 66 and 150 sentences per language. Like for subtask 1, the blind test data are made available to system authors for a week, and gold annotations are not disclosed. All test files are distributed in `.json` format.

Two evaluation measures are used. *Masked BERT-score* first checks if at least one of the MWE components was removed. If not, the score assigned to the paraphrase is 0. Otherwise, BERT-score (Zhang et al., 2020) is calculated between the system-generated paraphrase and up to two reference paraphrases: a minimal and a creative one (§ 4). The maximum of the two scores is retained. The second measure is *manual score*. For each sentence, native or near-native speakers are presented the paraphrases submitted by systems. In addition, annotators also see up to 2 reference paraphrases (minimal and creative), without knowing whether the paraphrase was generated by systems or by humans. This allows us to verify the quality of reference paraphrases with respect to system outputs. Annotators assign score 0 if the MWE is not removed, and, otherwise, three scores from 0 to 3 for keeping: (i) the sense of the removed MWE, (ii) the sense of the rest of the sentence, and (iii) grammaticality and naturalness. The final manual score is a weighted average of these 3 scores, with score (i) doubled, normalized to $[0, 100]$. Both masked BERT-score and manual score are averaged across all sentences, then macro-averaged across languages.

Diversity Metrics A novel evaluation dimension in this shared task is diversity. The idea is that the quality of a system's results should possibly go hand in hand with their diversity. In general, diversity is modelled as a property of *sets* whose *elements* can be apportioned into *categories*. It is here evaluated along two main dimensions: variety and balance (Stirling, 2007; Ramaciotti Morales et al., 2021; Estève et al., 2025). *Variety* relates to the number of categories, and *balance* to the evenness of the distribution of elements into categories. All other things being equal, the higher the variety, the higher the diversity, and the same holds for balance.

In our case, the sets evaluated for diversity are systems' predictions. In subtask 1, we follow Lion-Bouton et al. (2022), defining categories as *MWE types* and elements as their *occurrences* in text.[5] Only MWEs correctly identified by a system (i.e. true positives) are considered. Consider the toy test corpus (11)-(13), where MWE categories are boldfaced and bracketed, and a wrongly identified MWE (i.e. a false positive) is underlined:

(11) [**Me deparei**] [**cara a cara**] com… (pt)
Myself appeared face to face with…
'I found myself faced with…'

(12) [**Me dei mal**]: fiquei [**cara a cara**]… (pt)
Myself gave bad: got face to face…
'I was in a bad situation: I was facing…'

(13) Vendo a cara do pai, [**fez cara feia**] (pt)
Seeing the face of father, made face ugly
'Seeing her father's face, she frowned'

Suppose that system S_1 identified all these 6 expressions, i.e. 4 categories, 5 elements (true positives), and one false positive (ignored by diversity scores). System S_2, in turn, identified only the 3 categories and 3 elements from examples (12) and (13). We have $N_{S_1} = 4$ and $N_{S_2} = 3$, the number of categories of each system. As a measure of variety, we use *richness*, i.e. N_S. According to this measure, the predictions of S_1 are richer than those of S_2.

To assess balance, we use *Shannon evenness* (Smith and Wilson, 1996) defined by equation (14):

$$SE_S = \frac{SWE_S}{\ln(N_S)} \tag{14}$$

where SWE_S is the Shannon-Weaver entropy:

$$SWE_S = -\sum_{i=1}^{N_S} p_i * \ln(p_i) \tag{15}$$

[5] A MWE type is represented by the multiset of its components' lemmas. For instance, given the MWE (pt) ***cara a cara*** (lit. 'face to face') '(suddenly) facing', its multiset of lemmas, in lexicographic order, is {*a* 'to', *cara* 'face', *cara* 'face'}.

and p_i is the frequency of the ith category.[6] For S_1, $(p_1, p_2, p_3, p_4) = (\frac{1}{5}, \frac{2}{5}, \frac{1}{5}, \frac{1}{5})$ and $SWE_{S_1} = 1.33$, while for S_2, $(p_1, p_2, p_3) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $SWE_{S_2} = 1.1$. In eq. (14), entropy SWE is divided by its maximum value $\ln(N_S)$, so we have $SE_{S_1} = \frac{1.33}{1.39} = 0.96$ and $SE_{S_2} = \frac{1.1}{1.1} = 1.0$. Thus, the predictions of S_2 are more balanced than those of S_1.

Our last diversity measure is *Shannon-Weaver entropy* itself, from eq. (15). When not normalized, SWE_S is actually a hybrid metric, accounting for both variety *and* balance. According to SWE_S, the predictions of S_1 are more diverse than those of S_2.

For subtask 2, the same diversity measures are adapted by redefining categories as unique word types generated by a system and not present in the original sentence. For instance, given sentence (6), if systems S_3 and S_4 produced the outputs (7) and (8), then $N_{S_3} = 1$ (new words: *perspective*) and $N_{S_4} = 5$ (new words: *vision*, *du*, *metteur*, *en*, *scène*), $SWE_{S_3} = 0$, $SWE_{S_4} = 1.61$, $SE_{S_3} = 0$, and $SE_{S_4} = 1$. Thus, S_3 has less diverse predictions than S_4 according to the 3 measures.

4 Provided Data

Subtask 1. The dataset is a fruit of the PARSEME annotation campaign in which 17 teams took part, covering 10 previously covered languages – Modern Greek (el), Persian (fa), French (fr), Hebrew (he), Polish (pl), Portuguese (pt), Romanian (ro), Slovene (sl), Swedish (sv), Serbian (sr) – and 7 new ones – Egyptian (egy, ca. 2700-2000 BC), Ancient Greek (grc), Japanese (ja), Georgian (ka), Latvian (lv), Dutch (nl) and Ukrainian (uk). Human annotators worked on the corpus according to cross-linguistically unified guidelines composed of decision trees over elementary morphological, syntactic or distributional tests (Savary et al., 2026).[7]

Previous versions of the PARSEME corpora treated only verbal MWEs. Version 2.0 covers all MWE categories: verbal, nominal, adjectival-adverbial and functional. Some of those are subdivided into subcategories, such as:

- verbal idioms (VID): (nl) ***ijs breken*** 'break the ice';
- nominal idioms (NID): (ja) ***一人相撲*** (lit. 'one-person sumo') 'wrestling with oneself';
- adjectival idioms (AdjID): (sr) ***мртав пијан*** (lit. 'dead drunk') 'extremely drunk';
- conjunctive idioms (ConjID): (lt) ***kā arī*** (lit. 'as also') 'as well as'.

The test data have been synchronised between task 1 and 2. First, we identified all "unseen" sentences, e.g. those that have never been annotated in previous PARSEME editions (even partially), to avoid contamination (§5). Out of those, up to 150 sentences per language were randomly selected to meet the criteria for subtask 2, i.e. containing a single MWE (of category VID, NID or AdjID). For subtask 1, they were then completed with other randomly selected "unseen" sentences so as to reach roughly 500 annotated MWEs. As a result the test data for subtask 1 contains between 300 and 1,900 sentences per language.[8] The remaining sentences were split randomly (90%-10%) to create the training and development sets, which were provided to the participants, except in Ancient Greek, where not enough annotated data were available and only a test set exists.

Subtask 2. The dataset for subtask 2 is totally new and provided for 14 languages: the same as in subtask 1, except Egyptian, Ancient Greek, and Dutch.[9] Based on the test set of subtask 1, we extracted up to 150 sentences, as described above, meeting the criteria for subtask 2. We selected sentences containing only 1 MWE to simplify the definition and evaluation of subtask 2. We focus on VID, NID and AdjID because their degree of non-compositionality seems overall the highest, avoiding notoriously hard discussions about partial compositionality, as in (ro) ***ia măsuri*** 'take action', (ro) ***în cadrul*** (lit. 'in frame') 'in the framework (of)'.

Given a sentence with a highlighted idiom, native human experts were to provide at least one of two paraphrases: a *minimal* and a *creative* one. The former was obtained by modifying as few tokens as possible among those that do not belong to the MWE. When creating the latter, conversely, significant changes were encouraged, both lexical (adding, deleting or replacing words) and grammatical (e.g. changing the word order or transforming active to passive voice), as long as the meaning of the original sentence was maintained. For example, sentence (16) received the minimal paraphrase (17) and the creative one (18):

[6] E.g. for S_1 predictions, p_1 is the frequency of the category *me deparei*, p_2 the frequency of *cara a cara*, …

[7] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0

[8] With one outlier, Georgian, having around 40K sentences.

[9] Egyptian and Ancient Greek are skipped because they are non-spoken languages, while paraphrasing can be performed reliably only by native speakers. Dutch is skipped due to the unavailability of the annotators at the time of the paraphrase corpus construction.

(16) ხელისუფლება PDPA-მ მოქმედებაში
The.government PDPA in.action
მოიყვანა სოციალისტური **დღის**
brought socialist of.the.day
წესრიგი. (ka)
order.
'The PDPA government put into action a socialist agenda.'

(17) ხელისუფლება PDPA-მ მოქმედებაში
The.government PDPA in.action
მოიყვანა სოციალისტური გეგმა. (ka)
brought socialist plan.
'The PDPA government brought into action a socialist plan.'

(18) სოციალისტური წყობის რეალიზება
socialist set.up realization
ხელისუფლება PDPA-მ მოქმედებაში
the.government PDPA in.action
მოახერხა. (ka)
succeeded.
'The PDPA government managed to realize a socialist set-up.'

In each paraphrase, we asked annotators to remove at least one component of the MWE. New MWEs were allowed in the creative paraphrase, but not in the minimal one. The use of LLMs was prohibited for annotators, but online dictionaries and synonym lists were allowed. It was possible to provide more than two paraphrases, and then the two best ones had to be indicated (for system evaluation). Occasionally, it happened that a minimal or a creative paraphrase was not possible, then only one paraphrase was given. In rare problematic cases, the original sentence was totally discarded.

The resulting dataset contains from 66 (Swedish) to 150 (Georgian) original sentences per language. In total, there are 1,742 original sentences, with 726 VIDs, 863 NIDs and 153 AdjIDs, as well as 1,670 minimal and 1,618 creative paraphrases.

5 Running a Shared Task in the LLM era

So far, the PARSEME corpus in all 4 versions, as well as the system results from editions 1.1. and 1.2 of the PARSEME shared tasks, have been made publicly available under open licenses on the CLARIN/LINDAT infrastructure[10] and in public Gitlab repositories.[11] The most recent versions of the data being annotated have regularly been uploaded to public Gitlab language repositories, and made available by consistency checking web pages, to the best benefit of the research community.

Most of these practices have recently been jeopardised by aggressive scraping policies of some AI companies. Their bots scan the Internet, strongly targeting open source community infrastructures,[12] ignoring conventions such as `robots.txt` files.[13]

The PARSEME infrastructure is also concerned. Particularly intrusive is GPTBot, which scrapes data to train OpenAI's products. For instance, it sent almost 3 million queries in April–December 2025 to two of our servers, with up to 14,000 queries per day per server, likely acting in "distributed denial of service" mode to remain anonymous. OpenAI is known to violate the licenses under which data and software are distributed (Mueller, 2025). Last but not least, data contamination (Deng et al., 2024), particularly frequent due to LLMs, occurs when test data are included in the training phase, which leads to inflated performance scores.

This last risk drove major challenges in our data annotation and publication policy. Texts previously published with MWE annotations could no longer be used as test data. Thus, for languages from previous PARSEME corpus editions, we had to add significant amounts of new data annotated for all MWE categories from scratch. This prevented us from applying random or custom train/dev/test splits, used for estimating performance, notably on unseen MWEs (Ramisch et al., 2020). We also had to make private our public git repositories, used for everyday corpus development, and to hide consistency checking pages behind secret URLs, burdening legitimate users with new procedures. Even so, these changes do not preclude data contamination, since corpus or system developers may inadvertently store copies of test data in their own public spaces.

6 Implementation on Codabench

For the first time, the participants' submissions to PARSEME shared tasks were evaluated on the Codabench platform, an online framework designed for running machine learning competitions (Xu et al., 2022). Codabench allows benchmarks to run in a stable and a less error-prone environment based on docker, not subject to library version changes. Benchmarks are easily reproducible, and

[10] E.g. http://hdl.handle.net/11372/LRT-5124
[11] https://gitlab.com/parseme/sharedtask-data
[12] https://next.ink/186593/les-crawlers-des-ia-menacent-les-sites-scientifiques
[13] https://arstechnica.com/ai/2025/03/devs-say-ai-crawlers-dominate-traffic-forcing-blocks-on-entire-countries/

their scores are available online for participants right away. They occur on the leaderboard, which provides permanent links for lasting access.

Two competitions were established, each corresponding to a specific subtask.[14] A comprehensive bundle was prepared and uploaded to the platform to calculate the scores of the submissions. The bundle comprises gold data, the scoring tool and its dependencies, as well as the participant instructions that elucidate the competition's goals and the submission process. The bundle also includes a configuration file that specifies the paths of all the data, the docker container image, start and end dates of the competitions, and the maximum number of submissions per participant (here: 10).

During the competition, participants submitted a zip file with one directory per language, each containing their system's predictions. The scoring program returned numerical scores, displayed on the leaderboard, where the performances of participants were compared. For subtask 1, the leaderboard ranking was based on the global MWE-based F-score, but global and token-based precision and recall were also displayed, with an additional link to detailed results per language. For subtask 2, the ranking was based on average masked BERT-score.

The competitions and the results are now frozen, but a copy was created so that new participants can continuously propose new solutions.[15] They will no longer have to wait for a new evaluation to quickly and accurately assess their systems under the same conditions as in the shared task.

7 Systems

Subtask 1. This subtask features 10 participating systems: 9 submissions plus the baseline (Tab. 1). Among them, 5 are based on pre-trained encoder transformer models, fine-tuned for the task using BIO-style tags. `MTLB-STRUCT` relies on `bert-base-multilingual-cased` (Taslimipoor et al., 2020), with no auxiliary parsing task.[16] `Sahara-Tokenizers` (Karatepe et al., 2026) relies on the same pre-trained model, but introduces (a) explicit part-of-speech injection and (b) multi-task objective for joint BIO-style tagging and category classification. `Bert-multilingual-trial` and `BeeParser` (Erdem and Karaarslan, 2026) fine-tune `XLM-RoBERTa-base` on single languages, but also on language pairs, studying cross-lingual transfer. Finally, `romanian-bert` (Roscan and Nisioi, 2026) fine-tunes the language-specific `RoBERT-base` model after after comparing several models on challenging data subsets. One system, `pmi-mwe-scorer` (Bogdanova and Bucur, 2026), proposes a method based on syntax-aware pointwise mutual information (PMI) that leverages UD trees. Two systems rely on generative language models: `IPN` (Hülsing et al., 2026) applies instruction fine-tuning to `Qwen3-32B`, while `MorphoFiltered-Gemini` (Moise and Nisioi, 2026) relies on `gemini-2.0-flash-lite` with a lightweight morphological filter to remove unlikely outputs.

Subtask 2. We received 5 submissions listed in Tab. 2, including the baseline. All of them are based on LLMs: `GPT-CREATIVE` (Roscan and Nisioi, 2026) relies on prior MWE identification (with `romanian-bert` of subtask 1) followed by `GPT-4o` queries using category-oriented prompts. `Star-Paraphrasing-Cosine` (Bayraktar et al., 2026) and `Multiagent` are variants: `Cosine` tries to substitute a pre-identified MWE by single-word alternatives weighted by cosine similarity, while `Multiagent` is based on a combination of LLMs that generate, validate, and fix the paraphrase. Finally, `MISP` (Ciminari and Barrón-Cedeño, 2026) relies on `Qwen3-4B-Instruct` and cross-lingual transfer, fine-tuning the model on synthetic MWE paraphrases in Portuguese.

Baselines The baseline was implemented in Java as an API client for LLMs. It allows communication with both cloud-based APIs and locally hosted LLMs. To produce the baseline results for the test sets, we used the `gpt-oss-20b` model through a local Ollama installation. This prevented data leakage to cloud-based solutions (§ 5). The baseline system allows specifying a dataset and custom templates to be used as the system and user prompts. Each sample in the dataset is converted into a LLM call by filling the prompt templates. Templates indicate to the LLM the need to produce output that can be parsed by the system. For the first subtask, the LLM identifies MWEs at the entire sentence level, which are then mapped back into the tokenized form. For the second subtask, the LLM directly produces the new sentence. We also conducted preliminary experiments with `Llama-4-Scout` and `gpt-oss-20b`

[14] Subtask 1: https://www.codabench.org/competitions/12003/, subtask 2: https://www.codabench.org/competitions/12002/

[15] Subtask 1: https://www.codabench.org/competitions/13186/, subtask 2: https://www.codabench.org/competitions/13192/

[16] https://github.com/shivaat/MTLB-STRUCT/

however, gpt-oss-20b consistently achieved better performance than the other two models. The prompts[17] used for both subtasks were intentionally simple and served only to guide the model toward the expected output format. Aiming at a baseline contribution, we did not focus on heavily refining the prompts in order to obtain highly competitive results with language-specific elements.

8 Performance Results

Subtask 1. Out of the 10 participating systems, 5 systems cover all 17 languages for which data were available, 2 systems cover 16 languages, 2 systems cover 6 languages, and 1 system covers only one language (Romanian). In this section, systems are referred to by their names on the leaderboard.

Tab. 1 presents the general ranking of subtask 1, including the baseline. The number of languages covered by each system is shown in column *#Langs*. Then, we report the global MWE-based and token-based precision (P), recall (R) and F-scores (F1), with results macro-averaged over languages. Macro-average calculation ranges over all 17 languages. If a system did not submit results for a given language, this language is included in macro-average calculation as having P=R=$F1$=0. Systems are ranked by decreasing F-scores. Phenomenon- and language-specific results are shown in App. A .

According to both MWE-based and Token-based F1, the top-3 systems are MTLB-STRUCT, Sahara-Tokenizers and IPN. In terms of MWE-based F1, the best system MTLB-STRUCT beats the second best Sahara-Tokenizers by almost 9 points. The difference between the second and third ranks is even larger, reaching 19.9 MWE-based F1 points and 23.3 Token-based F1 points. MTLB-STRUCT favours precision, whereas the two other systems favour recall. On the other hand, while 4 systems overcome the baseline, 5 of them fail to do so, among which 3 cover between 16 and 17 languages.

These average results ignore inter-language variability. For instance, romanian-bert is the best system for Romanian, reaching 85.65 MWE-based F1. MTLB-STRUCT has the highest MWE-based F1 for 8 languages, but Sahara-Tokenizers beats it in 3 languages, while bert-multilingual-trial is the best in 2 languages, IPN is the best for Dutch, BeeParser is the best for Serbian, and the baseline has the highest MWE-based F1 on Ancient Greek.

The best language-specific scores are reached in Farsi, Japanese, Romanian, and Polish (MWE-based F1$\geq$ 80), followed by Serbian, Slovenian and Latvian (MWE-based F1$\geq$ 70). In Egyptian, Ancient Greek, and Dutch, the best systems reach the lowest scores (MWE-based F1$<$ 30). No training or development data was provided for Ancient Greek, while Egyptian and Dutch have the smallest training and development corpora, with 103 and 133 annotated MWEs in total.

Phenomenon-specific scores (App. A, Tab. 4-7) confirm that the main challenges in MWE identification, studied since edition 1.1, remain unsolved, especially for unseen MWEs. We emphasize that the best system from edition 1.2, MTLB-STRUCT, still gains the upper hand, suggesting that little progress is achieved in modern LLMs concerning MWE identification, despite their progress in other tasks.[18]

Subtask 2. Four teams submitted predictions of 5 systems (including the baseline), shown in Tab. 2, together with their global macro-average scores and ranks for the automatic (global masked BERT-score) evaluation. Only the baseline covered all 14 languages. The 4 other systems jointly covered 4 languages: French (fr), Georgian (ka), Portuguese (pt) and Romanian (ro). We performed manual evaluation only for these 4 languages, therefore we do not report the respective global ranking. The coverage of a low number of languages explains the low scores for the 4 last systems. This is why per-language scores, given in App. A are more interesting to analyse. Tab. 26–29 show that automatic evaluation (with masked BERT-score) nicely correlates with manual evaluation. More precisely, Pearson and Spearman correlation between the automatic and the manual scores for these 4 languages amount to 0.92 and 0.90, respectively. This indicates that (masked) BERT-score is a promising measure for MWE paraphrasing, despite its known weaknesses (Hanna and Bojar, 2021; Sun et al., 2022).

Tab. 26 and 29 in App. A show that systems specialised in one language (Star-Paraphraser in French, GPT-CREATIVE in Romanian) largely outperform the baseline. In French, Star-Paraphraser-Cosine has a high automatic score, but the manual score downgrades it to the third position.

Tab. 3 shows inter-annotator agreement for manual evaluation. We use Krippendorff's α Artstein and Poesio (2008) well suited for numerical scores,

[17] https://github.com/racai-ai/mwe_baseline/tree/master/templates

[18] One caveat is that the shared task's particularly tight schedule may have prevented the development of complex systems.

System	#Langs	Global MWE-based				Global Token-based			
		P	R	F1	rank	P	R	F1	rank
MTLB-STRUCT	17/17	62.21	53.09	57.29	1	70.55	54.75	61.65	1
Sahara-Tokenizers	17/17	45.77	51.33	48.39	2	61.53	57.12	59.24	2
IPN	17/17	21.37	42.32	28.40	3	26.32	56.66	35.94	3
BeeParser	6/17	26.62	25.84	26.22	4	29.26	27.03	28.10	6
baseline-gpt-oss-120b	17/17	17.44	34.86	23.25	5	23.59	52.93	32.64	4
bert-multilingual-trial	6/17	21.69	20.30	20.97	6	26.78	23.26	24.90	7
MorphoFiltered-Gemini	17/17	20.95	14.50	17.14	7	34.14	24.20	28.32	5
romanian-bert	1/17	5.35	4.76	5.04	8	5.55	4.82	5.16	10
Pattern-Based-MWE-Identifier	16/17	2.25	12.69	3.82	9	13.15	50.07	20.83	8
pmi-mwe-scorer	16/17	0.97	2.59	1.41	10	7.15	22.66	10.87	9

Table 1: Subtask 1 results – number of languages covered by systems (#Langs); then macro-averaged MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks.

System	#Langs	Autom.	
		gm-BS	rank
baseline-gpt-oss-120b	14/14	71.62	1
MISP	4/14	14.21	2
Star-Par.-Cosine	1/14	6.71	3
Star-Par.-Multiagent	1/14	6.39	4
GPT-CREATIVE	1/14	6.38	5

Table 2: Subtask 2 results – number of languages covered by systems (#Langs); global masked BERT-score (gm-BS) and the associated rank.

Language	A_1-A_2	A_1-Adj	A_2-Adj
French (fr)	–	96.50	97.37
Portuguese (pt)	80.30	87.56	91.14
Romanian (ro)	75.99	90.37	87.42

Table 3: Manual evaluation of subtask 2, inter-annotator agreement (Krippendorff's α, interval difference): pairwise scores between annotators (A_1, A_2) and adjudicator (Adj). Values multiplied by 100 for better readability.

with disagreements between s_1 and s_2 weighted proportionally to $(s_1 - s_2)^2$. Georgian is omitted because there was only one annotator. For the other 3 languages, a third adjudicator Adj unified annotations, which were then used to assess system's performances. For French there were no overlapping items between annotators A_1 and A_2. Agreement between A_1 and A_2 ranges from 75 to 80, whereas it is greater than 85 with respect to Adj. Thus, our evaluation protocol seems reproducible, although assessing meaning similarity is usually a hard task.

9 Diversity Results

Subtask 1. Previous editions of PARSEME have shown a strong correlation between the number of unseen MWEs in the test sets and the overall performance of systems. The diversity measures we propose in this new edition are a continuation of these reflections. We therefore began by measuring the correlation between the performance scores of systems and their diversity scores. Detailed diversity scores are available in App. A. We calculated a correlation score for each language, which we then averaged to obtain an overall view (Tab. 25). This gave us a Pearson correlation of 0.72 and a Spearman correlation of 0.76 between MWE-based F1 and the hybrid variety-balance measure. These results, confirming those obtained by Lion-Bouton et al. (2022), once again highlight the importance of predicting diverse MWEs in order to obtain high-quality predictions.

However, when we look at the correlation between performance and not entropy, but variety and balance individually, we see a significant difference in behaviour. Variety is correlated with performance at 0.81 (Pearson and Spearman), while balance is correlated at -0.39 (Pearson) and -0.46 (Spearman). It would therefore appear that entropy is more impacted by variety than by balance, and that a more balanced system would have a negative impact on performance, unlike a varied system.

Subtask 2. Per-language diversity scores are reported in App. A. Conversely to subtask 1, we see the so-called performance-diversity trade-off typical for generation scenarios (Ippolito et al., 2019; Zhang et al., 2021). For instance in French (Tab. 26), the higher the performance, i.e. the quality of the generated paraphrases, the lower the diversity, and vice-versa. One exception is Star-Paraphraser-Cosine. It has the lowest diversity, which is likely why it obtains high BERT-scores (the generated paraphrases resemble the original sentence). However, it does not achieve the highest manual score, which means that in reality it does not perform particularly well. Notable is also MISP in Ro-

manian (Tab. 29), which shows a particularly high lexical creativity (richness), but low manual scores.

Overall, balance scores are rather high across both subtasks. In subtask 1, this might result from a high number of infrequent MWEs. In subtask 2, the newly introduced vocabulary items (not appearing in the original sentence) might also often be hapaxes. More insight into these results will be gained from future analyses, and from new editions of the shared task, with more systems and languages.

10 Conclusions and Future Work

The data and systems discussed here are only the beginning of a deeper study of MWE identification and paraphrasing in the LLM era. In addition to traditional metrics, human evaluation (subtask 2) and diversity scores provide complementary views on the results. The overall trend in subtask 1 indicates that pre-trained encoder models and BIO encoding are still competitive. The results of subtask 2 are an initial step towards MWE paraphrasing, that we intend to generalise cross-lingually.

Limitations

The use of a baseline based on an LLM entails very large processing costs in subtask 1, especially in Georgian (including times and machine requirements).

The Georgian (ka) dataset is extremely large compared to other languages, and is very sparse, containing few annotations. This raises questions about the guidelines and its interpretation, which may vary depending on the language. The size of the Georgian corpus therefore involves very long processing times.

The amount of data for all languages is not balanced. Some languages have small training corpora (e.g. Dutch) and in particular Ancient Greek (grc) has no training nor development data available.

The use of BERT-score for automatic evaluation is known to have numerous weaknesses (Hanna and Bojar, 2021; Sun et al., 2022), sometimes ranking participants imperfectly compared to manual evaluation.

We do not calculate nor report statistical significance in the rankings: some small observed differences may be due to chance. Further analyses such as bootstrapped p-values are required to establish the robustness of our results (Ramisch et al., 2023).

Ethical Considerations

Despite the concerns raised in Section 5, the baseline for the shared task is based on an LLM. Although we chose to use the one with the highest level of openness available to us, the model is not completely open source, as the weights of the model are available, but we do not know the exact training corpus used to train this LLM.

Furthermore, the languages participating in the shared task are predominantly Indo-European languages, which are not particularly low resourced. The addressed languages do not include the 3 lowest levels of the Joshi et al. (2020) resourcedness scale.

Acknowledgements

This work received support from the CA21167 COST action UniDive, funded by the European Union via COST (European Cooperation in Science and Technology). Further support came from (i) the French Agence Nationale pour la Recherche, via the SELEXINI project (ANR-21-CE23-0033-01), (ii) the Romanian Ministry of Research, Innovation and Digititalization - UEFISCDI, project number PN-IV-P8-8.2-EUD-2025-0061, within PNCDI IV, (iii) the LLMS4EU project funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

We thank Louis Estève for adapting the diversity metrics scripts to the shared task. We are grateful to all the language leaders and annotators of the PARSEME community who contributed to created the PARSEME 2.0 datasets: `https://gitlab.com/parseme/corpora/-/wikis/home#language-teams`. We would also like to thank the annotators who contributed to the manual evaluation of subtask 2: Gabriela Berndt de Souza, Mihaela Cristescu, Letícia Guedes Guimarães, Irina Lobzhanidze, Verginica Mititelu, Adriana Pagano and Carmen Vasile. We are grateful to the Codabench support team, the AdMIRe 2 shared task organisers, the MWE 2026 workshop chairs, the members of the UniDive COST Action, and the MWE Section of SIGLEX for their continued feedback and support.

We dedicate this work to the memory of Federico Sangati and Silvio Ricardo Cordeiro.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Dogukan Arslan, Rodrigo Wilkens, Wei He, Dilara Torunoğlu-Selamet, Thomas Pickard, Aline Villavicencio, Adriana S. Pagano, and Gülşen Eryiğit. 2026. MWE-2026 Shared Task 2: AdMIRe 2 - Advancing Multimodal Idiomaticity Representation. In *Proceedings of the 22nd International Workshop on Multiword Expressions (MWE-2026)*.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of natural language processing*, volume 2, pages 267–292. CRC Press, Boca Raton, USA.

Petra Barančíková and Václava Kettnerová. 2018. Paraphrases of verbal multiword expressions: The case of Czech light verbs and idioms. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 35–59. Language Science Press., Berlin.

Elif Bayraktar, Vedat Doğancan, Muhammed A. Gümüş, and Nusret Ali Kızılaslan. 2026. Semantic Stars at PARSEME 2.0 Subtask 2:Alternative Approaches for MWE Paraphrasing. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

Anna Bogdanova and Ileana Bucur. 2026. PMI MWE Scorer at PARSEME 2.0 Subtask 1: identifying multiword expressions using pointwise mutual information and universal dependencies. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2009. SemEval-2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 100–105, Boulder, Colorado. Association for Computational Linguistics.

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier, and Silvio Cordeiro. 2021. A french corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*, 8(2).

Debora Ciminari and Alberto Barrón-Cedeño. 2026. MISP at PARSEME 2.0 Subtask 2: A Cross-lingual Approach to Multiword Expression Paraphrasing. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

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

Chunyuan Deng, Yilun Zhao, Yuzhao Heng, Yitong Li, Jiannan Cao, Xiangru Tang, and Arman Cohan. 2024. Unveiling the spectrum of data contamination in language model: A survey from detection to remediation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16078–16092, Bangkok, Thailand. Association for Computational Linguistics.

Ahmet Erdem and Oguzhan Karaarslan. 2026. Cross Lingual BERT at PARSEME 2.0 Subtask 1: Can Cross-Lingual Interactions Improve MWE Identification? In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

Louis Estève, Marie-Catherine de Marneffe, Nurit Melnik, Agata Savary, and Olha Kanishcheva. 2025. A survey of diversity quantification in natural language processing: The why, what, where and how. *Preprint*, arXiv:2507.20858.

Gaston Gross. 1988. Degré de figement des noms composés. *Langages*, 90:57–72.

Maurice Gross. 1986. Lexicon-grammar: The representation of compound words. In *Proceedings of the 11th Coference on Computational Linguistics*, COLING '86, pages 1–6. Association for Computational Linguistics.

Maurice Gross and Jean Senellart. 1998. Nouvelles bases statistiques pour les mots du français. In *Proceedings of JADT'98, Nice 1998*, pages 335–349.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

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

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. SemEval-2013 task 4: Free paraphrases of noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume*

2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 138–143, Atlanta, Georgia, USA. Association for Computational Linguistics.

Anna Hülsing, Noah-Manuel Michael, Daniel Ignacio Mora Melanchthon, and Andrea Horbach. 2026. IPN at PARSEME 2.0 Subtask 1: MWE Identification via Related Languages and Attempts at Harnessing Thinking Mode. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762, Florence, Italy. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Yunus Karatepe, Mert Sülük, Begüm Özbay, and Zeynep Tuğçe Kırımlı. 2026. Sahara Tokenizers at PARSEME 2.0 Subtask 1: Combining Contextual Embeddings with Structural Decoding for Multi-Word Expression Detection. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234, Copenhagen, Denmark. Association for Computational Linguistics.

Adam Lion-Bouton, Yagmur Ozturk, Agata Savary, and Jean-Yves Antoine. 2022. Evaluating diversity of multiword expressions in annotated text. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3285–3295, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Changsheng Liu and Rebecca Hwa. 2016. Phrasal substitution of idiomatic expressions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 363–373, San Diego, California. Association for Computational Linguistics.

Stella Markantonatou, Carlos Ramisch, Victoria Rosén, Mike Rosner, Manfred Sailer, Agata Savary, and Veronika Vincze. 2021. PMWE conventions for examples containing multiword expressions. Technical report, Phraseology and Multiword Expressions – book series at Language Science Press.

Igor Mel'čuk. 2010. La phraséologie en langue, en dictionnaire et en TALN. In *Actes de la 17ème Conférence sur le Traitement Automatique des Langues Naturelles 2010*, Montréal, Canada.

Irina Moise and Sergiu Nisioi. 2026. MorphoFiltered-Gemini at PARSEME 2.0 Subtask 1: Tackling LLM Overgeneration via Universal POS-based Constraints. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

Florian Mueller. 2025. First copyright ruling against OpenAI worldwide: music rights collecting society wins German injunction over song lyrics —to be appealed now. Accessed on 01.01.2026.

Maria Pershina, Yifan He, and Ralph Grishman. 2015. Idiom paraphrases: Seventh heaven vs cloud nine. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 76–82, Lisbon, Portugal. 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.

Jipeng Qiang, Yang Li, Chaowei Zhang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2023. Chinese idiom paraphrasing. *Transactions of the Association for Computational Linguistics*, 11:740–754.

Pedro Ramaciotti Morales, Robin Lamarche-Perrin, Raphaël Fournier-S'Niehotta, Rémy Poulain, Lionel Tabourier, and Fabien Tarissan. 2021. Measuring diversity in heterogeneous information networks. *Theoretical Computer Science*, 859:80–115. Publisher: Elsevier.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, and 6 others. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica

Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Carlos Ramisch, Abigail Walsh, Thomas Blanchard, and Shiva Taslimipoor. 2023. A survey of MWE identification experiments: The devil is in the details. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 106–120, Dubrovnik, Croatia. Association for Computational Linguistics.

Rares-Alexandru Roscan and Sergiu Nisioi. 2026. Archaeology at PARSEME 2.0 Subtasks 1 and 2: Parsing is for Encoders, Paraphrasing is for LLMs. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Lukas Santing, Ryan Sijstermans, Giacomo Anerdi, Pedro Jeuris, Marijn ten Thij, and Riza Batista-Navarro. 2022. Food for thought: How can we exploit contextual embeddings in the translation of idiomatic expressions? In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 100–110, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, and 9 others. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, and 3 others. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 87–147. Language Science Press., Berlin.

Agata Savary, Simon Petitjean, Timm Lichte, Laura Kallemeyer, and Jakub Waszczuk. 2020. Object-oriented lexical encoding of multiword expressions: Short and sweet. *Lexique*, 27:87–120.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Agata Savary, Manon Scholivet, Carlos Ramisch, Takuya Nakamura, Eric Bilinski, Sara Stymne, Voula Giouli, Stella Markantonatou, Vasile Păiş, Maria Mitrofan, Louis Estève, Bruno Guillaume, Verginica Barbu Mititelu, Jaka Čibej, Roberto A. Díaz Hernández, Victoria Fendel, Polona Gantar, Olha Kanishcheva, Cvetana Krstev, and 9 others. 2026. PARSEME 2.0 multilingual corpus of multiword expressions. In *Submitted to LREC 2026, under review.*

Agata Savary, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz, and Alina Wróblewska. 2024. UniDive: A COST action on universality, diversity and idiosyncrasy in language technology. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 372–382, Torino, Italia. ELRA and ICCL.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559, San Diego, California. Association for Computational Linguistics.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Benjamin Smith and J. Bastow Wilson. 1996. A Consumer's Guide to Evenness Indices. *Oikos*, 76(1):70–82. Number: 1 Publisher: [Nordic Society Oikos, Wiley].

Andy Stirling. 2007. A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4(15):707–719. Number: 15 Publisher: Royal Society.

Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. BERTScore is unfair: On social bias

in language model-based metrics for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Minghuan Tan and Jing Jiang. 2021. Does BERT understand idioms? a probing-based empirical study of BERT encodings of idioms. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1397–1407, Held Online. INCOMA Ltd.

Shiva Taslimipoor, Sara Bahaadini, and Ekaterina Kochmar. 2020. MTLB-STRUCT @parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 142–148, online. Association for Computational Linguistics.

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.

Takashi Wada, Yuji Matsumoto, Timothy Baldwin, and Jey Han Lau. 2023. Unsupervised paraphrasing of multiword expressions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4732–4746, Toronto, Canada. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. SemEval-2015 task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 1–11, Denver, Colorado. Association for Computational Linguistics.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

Seid Muhie Yimam, Héctor Martínez Alonso, Martin Riedl, and Chris Biemann. 2016. Learning paraphrasing for multiword expressions. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 1–10, Berlin, Germany. Association for Computational Linguistics.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48, Online. Association for Computational Linguistics.

A Detailed shared task results

This appendix presents detailed results for subtasks 1 and 2: cross-lingual macro-averages per phenomenon (subtask 1) and per-language results, including diversity scores per language.

In the phenomenon-specific rankings of subtask 1, a MWE is considered seen if a MWE with the same multi-set of lemmas was annotated at least once in the training corpus or in the development corpus. This definition impacts four MWE-based evaluation metrics and rankings: unseen-in-traindev, seen-in-traindev, variant-of-traindev and identical-to-traindev.

Please, interpret cross-lingual macro-averages carefully, as some scores depend on the dataset size, and the size of the underlying datasets varies across languages. These results are also published on the shared task git repository:

- Subtask 1: `https://gitlab.com/parseme/sharedtask-data/-/blob/master/2.0/subtask1/Detailed_results.md`
- Subtask 2: `https://gitlab.com/parseme/sharedtask-data/-/blob/master/2.0/subtask2/Detailed_results.md`

System	#Langs	Performance							
		Discontinuous MWE-based				Continuous MWE-based			
		P	R	F1	Rank	P	R	F1	Rank
MTLB-STRUCT	17/17	45.57	32.46	37.91	1	63.35	55.25	59.02	1
BeeParser	6/17	20.62	21.31	20.96	2	27.11	26.18	26.64	4
bert-multilingual-trial	6/17	15.83	15.46	15.64	3	22.00	20.58	21.27	6
IPN	17/17	12.70	20.01	15.54	4	22.27	45.53	29.91	3
baseline-gpt-oss-120b	17/17	12.68	4.08	6.17	5	17.42	40.32	24.33	5
romanian-bert	1/17	4.62	3.75	4.14	6	5.46	4.92	5.18	8
pmi-mwe-scorer	16/17	0.03	0.10	0.05	7	1.71	2.98	2.17	10
MorphoFiltered-Gemini	17/17	0.00	0.00	0.00	8	20.95	16.93	18.73	7
Pattern-Based-MWE-Id.	16/17	0.00	0.00	0.00	8	2.25	14.56	3.90	9
Sahara-Tokenizers	17/17	0.00	0.00	0.00	8	45.78	59.63	51.80	2

Table 4: Subtask 1 results – phenomenon-specific scores for discontinuous vs. continuous. Macro-averaged MWE-based (exact match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks.

System	#Langs	Performance							
		Unseen-in-traindev MWE-based				Seen-in-traindev MWE-based			
		P	R	F1	Rank	P	R	F1	Rank
MTLB-STRUCT	17/17	26.72	21.38	23.75	1	84.25	71.74	77.49	1
Sahara-Tokenizers	17/17	16.68	24.92	19.98	2	83.57	65.91	73.70	2
IPN	17/17	9.62	36.14	15.20	3	76.16	44.71	56.34	3
baseline-gpt-oss-120b	17/17	7.99	29.15	12.54	4	76.72	35.23	48.29	4
BeeParser	6/17	11.13	14.01	12.41	5	34.13	29.64	31.73	5
MorphoFiltered-Gemini	17/17	9.83	11.18	10.46	6	73.08	14.86	24.70	8
bert-multilingual-trial	6/17	8.59	9.21	8.89	7	27.93	24.72	26.23	7
romanian-bert	1/17	0.74	1.31	0.95	8	5.71	4.89	5.27	9
pmi-mwe-scorer	16/17	0.54	3.72	0.94	9	61.66	2.15	4.16	10
Pattern-Based-MWE-Id.	16/17	0.26	2.72	0.47	10	60.75	19.98	30.07	6

Table 5: Subtask 1 results – phenomenon-specific scores for unseen-in-traindev vs. seen-in-traindev. Macro-averaged MWE-based (exact match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks.

System	#Langs	Performance							
		Variant-of-traindev MWE-based				Identical-to-traindev MWE-based			
		P	R	F1	Rank	P	R	F1	Rank
MTLB-STRUCT	17/17	79.35	53.68	64.04	1	85.22	77.84	81.36	1
Sahara-Tokenizers	17/17	77.16	40.15	52.82	2	85.01	76.55	80.56	2
IPN	17/17	61.65	37.94	46.97	3	82.56	46.83	59.76	3
baseline-gpt-oss-120b	17/17	58.31	21.04	30.92	4	80.81	41.22	54.59	4
BeeParser	6/17	32.84	26.52	29.34	5	34.52	31.05	32.69	6
bert-multilingual-trial	6/17	27.26	21.05	23.76	6	28.19	26.40	27.27	7
MorphoFiltered-Gemini	17/17	54.54	12.00	19.67	7	75.06	15.91	26.25	8
Pattern-Based-MWE-Id.	16/17	40.13	12.73	19.33	8	72.43	22.54	34.38	5
romanian-bert	1/17	5.16	3.31	4.03	9	5.77	5.13	5.43	9
pmi-mwe-scorer	16/17	43.91	1.62	3.12	10	72.46	2.27	4.40	10

Table 6: Subtask 1 results – phenomenon-specific scores for variant-of-traindev vs. identical-to-traindev. Macro-averaged MWE-based (exact match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks.

System	#Langs	Performance							
		Single-token MWE-based				Multi-token MWE-based			
		P	R	F1	Rank	P	R	F1	Rank
MTLB-STRUCT	17/17	24.89	49.32	33.08	1	64.93	51.30	57.32	1
IPN	17/17	15.49	35.60	21.59	2	20.44	42.37	27.58	3
Sahara-Tokenizers	17/17	12.53	52.91	20.26	3	62.06	49.96	55.36	2
BeeParser	6/17	12.64	19.79	15.43	4	27.29	24.81	25.99	4
bert-multilingual-trial	6/17	8.77	20.04	12.20	5	22.63	19.05	20.69	6
baseline-gpt-oss-120b	17/17	10.68	8.81	9.66	6	16.99	38.03	23.49	5
MorphoFiltered-Gemini	17/17	6.13	2.14	3.17	7	21.56	15.99	18.36	7
pmi-mwe-scorer	16/17	1.38	12.17	2.48	8	0.66	1.57	0.93	10
Pattern-Based-MWE-Id.	16/17	0.98	11.83	1.81	9	6.58	11.52	8.38	8
romanian-bert	1/17	0.49	1.96	0.78	10	5.49	4.77	5.10	9

Table 7: Subtask 1 results – phenomenon-specific scores for single-token vs. multi-token. Macro-averaged MWE-based (exact match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks.

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
Sahara-Tokenizers	33.67	13.20	18.97	1	47.40	17.45	25.51	2	2.53	4	17.00	3	0.89	3
MTLB-STRUCT	31.07	12.80	18.13	2	45.11	15.92	23.53	3	2.45	5	16.00	4	0.88	4
MorphoFiltered-Gemini	12.00	3.60	5.54	3	15.79	4.60	7.13	6	2.55	2	14.00	5	0.97	1
baseline-gpt-oss-120b	3.60	6.20	4.56	4	7.89	13.71	10.02	5	2.54	3	18.00	2	0.88	4
IPN	2.53	8.00	3.85	5	10.36	32.02	15.65	4	2.99	1	24.00	1	0.94	2
Pattern-Based-MWE-Id.	1.98	8.20	3.19	6	19.28	40.46	26.11	1	2.03	6	11.00	6	0.85	5

Table 8: Egyptian (egy) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = -0.03$, $\rho = -0.03$); Richness: ($r = -0.03$, $\rho = 0.09$); SE: ($r = -0.21$, $\rho = 0.32$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
MTLB-STRUCT	67.70	39.40	49.81	1	74.10	37.02	49.38	1	3.92	3	103.00	4	0.85	5
Sahara-Tokenizers	36.86	37.60	37.23	2	55.89	40.93	47.25	2	3.91	4	104.00	3	0.84	6
IPN	15.77	35.80	21.90	3	21.97	56.06	31.57	4	4.51	2	130.00	2	0.93	3
MorphoFiltered-Gemini	20.34	19.40	19.86	4	32.69	34.58	33.61	3	4.51	2	93.00	5	1.00	1
baseline-gpt-oss-120b	12.76	36.00	18.84	5	17.03	52.40	25.70	5	4.74	1	139.00	1	0.96	2
pmi-mwe-scorer	0.47	2.40	0.78	6	6.17	23.43	9.77	7	2.48	5	12.00	6	1.00	1
Pattern-Based-MWE-Id.	0.28	2.80	0.50	7	6.86	38.57	11.64	6	1.97	6	9.00	7	0.89	4

Table 9: Modern Greek (el) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.58$, $\rho = 0.34$); Richness: ($r = 0.66$, $\rho = 0.50$); SE: ($r = -0.63$, $\rho = -0.54$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
MTLB-STRUCT	83.57	82.07	82.81	1	87.97	84.88	86.40	1	5.65	1	320.00	1	0.98	2
BeeParser	79.64	78.69	79.16	2	86.51	84.99	85.75	2	5.62	2	307.00	2	0.98	2
Sahara-Tokenizers	70.67	77.29	73.83	3	83.77	83.68	83.73	3	5.61	3	306.00	3	0.98	2
IPN	37.52	41.04	39.20	4	46.32	57.94	51.48	4	5.14	4	180.00	4	0.99	1
baseline-gpt-oss-120b	32.00	30.28	31.12	5	42.48	53.23	47.25	5	4.69	5	121.00	5	0.98	2
MorphoFiltered-Gemini	41.54	22.51	29.20	6	57.97	39.43	46.94	6	4.58	6	102.00	6	0.99	1
Pattern-Based-MWE-Id.	8.19	16.33	10.91	7	33.26	52.03	40.58	7	4.11	7	65.00	7	0.98	2
pmi-mwe-scorer	4.67	5.58	5.09	8	14.65	16.43	15.49	8	3.23	8	26.00	8	0.99	1

Table 10: Persian (Farsi) (fa) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.93$, $\rho = 1.00$); Richness: ($r = 0.99$, $\rho = 1.00$); SE: ($r = -0.52$, $\rho = -0.51$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
MTLB-STRUCT	73.55	53.29	61.81	1	83.82	57.42	68.15	1	5.34	1	226.00	1	0.98	3
Sahara-Tokenizers	52.60	50.50	51.53	2	75.76	58.33	65.91	2	5.28	2	212.00	2	0.98	3
IPN	37.74	42.71	40.07	3	48.43	58.00	52.79	3	5.14	3	184.00	3	0.99	2
baseline-gpt-oss-120b	32.03	32.73	32.38	4	42.92	53.08	47.47	4	4.91	4	144.00	4	0.99	2
MorphoFiltered-Gemini	14.17	6.99	9.36	5	35.36	18.42	24.22	5	3.26	5	29.00	5	0.97	4
Pattern-Based-MWE-Id.	1.11	4.79	1.80	6	15.22	51.08	23.46	6	3.06	6	22.00	6	0.99	2
pmi-mwe-scorer	0.48	1.00	0.64	7	14.64	22.83	17.84	7	1.61	7	5.00	7	1.00	1

Table 11: French (fr) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.92$, $\rho = 1.00$); Richness: ($r = 0.99$, $\rho = 1.00$); SE: ($r = -0.32$, $\rho = -0.54$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
baseline-gpt-oss-120b	8.55	15.92	11.12	1	15.23	29.89	20.18	1	3.41	1	35.00	1	0.96	3
MorphoFiltered-Gemini	43.33	3.90	7.16	2	59.09	4.98	9.19	4	2.10	4	9.00	4	0.95	4
Sahara-Tokenizers	8.81	6.01	7.14	3	14.19	8.17	10.37	3	2.45	3	14.00	3	0.93	5
IPN	3.95	6.91	5.02	4	7.42	20.82	10.94	2	3.01	2	21.00	2	0.99	1
MTLB-STRUCT	3.31	1.20	1.76	5	9.87	3.83	5.52	6	0.56	6	2.00	6	0.81	6
pmi-mwe-scorer	0.74	1.80	1.05	6	5.45	12.01	7.49	5	1.56	5	5.00	5	0.97	2

Table 12: Ancient Greek (grc) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.81$, $\rho = 0.71$); Richness: ($r = 0.83$, $\rho = 0.71$); SE: ($r = 0.36$, $\rho = -0.09$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
MTLB-STRUCT	67.38	62.67	64.94	1	69.68	61.73	65.47	1	5.26	1	226.00	1	0.97	3
Sahara-Tokenizers	52.89	60.28	56.34	2	57.93	60.23	59.06	2	5.20	2	214.00	2	0.97	3
IPN	14.36	40.52	21.20	3	15.75	46.56	23.54	4	5.02	3	170.00	3	0.98	2
baseline-gpt-oss-120b	14.64	34.53	20.56	4	17.53	42.51	24.82	3	4.79	4	135.00	4	0.98	2
MorphoFiltered-Gemini	12.42	3.79	5.81	5	17.67	4.71	7.44	7	2.80	6	17.00	6	0.99	1
Pattern-Based-MWE-Id.	1.28	10.78	2.29	6	9.24	54.85	15.81	5	3.71	5	46.00	5	0.97	3
pmi-mwe-scorer	0.22	2.20	0.39	7	4.25	42.13	7.72	6	2.27	7	10.00	7	0.99	1

Table 13: Hebrew (he) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.80$, $\rho = 0.96$); Richness: ($r = 0.92$, $\rho = 0.96$); SE: ($r = -0.66$, $\rho = -0.59$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
Sahara-Tokenizers	75.92	70.00	72.84	1	79.03	68.41	73.34	1	5.50	1	281.00	1	0.97	3
MTLB-STRUCT	80.36	63.00	70.63	2	84.32	58.44	69.03	3	5.40	2	252.00	3	0.98	2
bert-multilingual-trial	74.64	63.00	68.33	3	82.30	61.25	70.23	2	5.39	3	253.00	2	0.97	3
BeeParser	69.25	59.00	63.71	4	78.14	58.06	66.62	4	5.31	4	235.00	4	0.97	3
IPN	31.05	46.20	37.14	5	32.89	57.16	41.76	5	5.19	5	200.00	5	0.98	2
baseline-gpt-oss-120b	20.97	26.80	23.53	6	29.60	55.37	38.57	6	4.85	6	129.00	6	1.00	1
Pattern-Based-MWE-Id.	6.77	18.20	9.86	7	21.67	48.21	29.90	7	4.01	7	64.00	7	0.96	4
MorphoFiltered-Gemini	14.22	6.20	8.64	8	32.44	24.81	28.12	8	3.39	8	30.00	8	1.00	1
pmi-mwe-scorer	3.58	4.00	3.78	9	15.58	28.77	20.22	9	2.83	9	18.00	9	0.98	2

Table 14: Japanese (ja) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.89$, $\rho = 1.00$); Richness: ($r = 0.98$, $\rho = 0.98$); SE: ($r = -0.38$, $\rho = -0.36$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
MTLB-STRUCT	39.08	63.00	48.24	1	42.18	63.29	50.62	1	3.05	5	61.00	3	0.74	6
Sahara-Tokenizers	26.17	69.40	38.01	2	28.88	70.14	40.91	2	3.27	3	76.00	1	0.76	5
MorphoFiltered-Gemini	2.64	34.00	4.90	3	3.00	35.96	5.54	3	3.40	1	62.00	2	0.82	2
baseline-gpt-oss-120b	0.68	36.20	1.33	4	0.67	40.85	1.32	5	3.32	2	60.00	4	0.81	3
Pattern-Based-MWE-Id.	0.42	36.40	0.83	5	1.40	61.78	2.73	4	2.24	6	33.00	6	0.64	7
IPN	0.29	26.00	0.57	6	0.40	36.24	0.79	6	3.23	4	56.00	5	0.80	4
pmi-mwe-scorer	0.01	1.80	0.02	7	0.16	26.38	0.31	7	2.20	7	9.00	7	1.00	1

Table 15: Georgian (ka) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.29$, $\rho = 0.46$); Richness: ($r = 0.53$, $\rho = 0.86$); SE: ($r = -0.30$, $\rho = -0.43$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
MTLB-STRUCT	81.89	62.53	70.91	1	87.69	61.71	72.45	1	4.59	3	151.00	2	0.91	3
bert-multilingual-trial	73.27	61.52	66.88	2	84.79	62.99	72.28	2	4.57	4	149.00	3	0.91	3
Sahara-Tokenizers	64.69	59.12	61.78	3	75.25	60.71	67.20	3	4.52	5	143.00	4	0.91	3
baseline-gpt-oss-120b	12.39	58.52	20.45	4	13.53	67.82	22.57	5	4.66	2	163.00	1	0.91	3
MorphoFiltered-Gemini	23.40	17.64	20.11	5	29.87	22.79	25.85	4	4.03	6	67.00	6	0.96	2
IPN	7.00	40.68	11.94	6	9.52	57.61	16.34	7	4.79	1	149.00	3	0.96	2
Pattern-Based-MWE-Id.	2.58	28.46	4.73	7	9.45	63.90	16.46	6	3.98	7	79.00	5	0.91	3
pmi-mwe-scorer	0.37	3.21	0.67	8	3.25	27.62	5.82	8	2.69	8	15.00	7	0.99	1

Table 16: Latvian (lv) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.54$, $\rho = 0.50$); Richness: ($r = 0.62$, $\rho = 0.67$); SE: ($r = -0.62$, $\rho = -0.67$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
IPN	24.81	30.41	27.33	1	38.61	53.20	44.75	2	4.78	1	123.00	1	0.99	2
baseline-gpt-oss-120b	23.94	24.65	24.29	2	38.13	50.38	43.41	3	4.48	2	94.00	3	0.99	2
MTLB-STRUCT	36.79	17.97	24.15	3	67.86	20.63	31.64	5	4.11	3	68.00	4	0.98	3
Sahara-Tokenizers	19.02	26.04	21.98	4	51.29	41.04	45.60	1	4.48	2	95.00	2	0.98	3
MorphoFiltered-Gemini	20.65	8.76	12.30	5	57.97	27.25	37.08	4	3.64	4	38.00	5	1.00	1
Pattern-Based-MWE-Id.	1.88	4.38	2.63	6	18.35	24.43	20.96	6	2.06	6	9.00	7	0.94	4
pmi-mwe-scorer	1.19	2.30	1.57	7	12.40	19.54	15.17	7	2.30	5	10.00	6	1.00	1

Table 17: Dutch (nl) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.97$, $\rho = 0.90$); Richness: ($r = 0.95$, $\rho = 0.86$); SE: ($r = 0.27$, $\rho = -0.11$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
bert-multilingual-trial	83.33	84.00	83.67	1	87.13	84.82	85.96	1	5.49	1	286.00	1	0.97	4
MTLB-STRUCT	83.65	79.80	81.68	2	86.86	80.93	83.79	3	5.42	3	268.00	3	0.97	4
BeeParser	80.51	82.60	81.54	3	84.86	84.54	84.70	2	5.44	2	276.00	2	0.97	4
Sahara-Tokenizers	51.84	64.80	57.60	4	75.36	74.57	74.96	4	5.21	4	220.00	5	0.97	4
IPN	28.58	72.20	40.95	5	29.74	79.60	43.30	5	5.42	3	260.00	4	0.98	3
baseline-gpt-oss-120b	18.15	38.80	24.73	6	25.22	61.76	35.82	6	4.84	5	146.00	6	0.97	4
MorphoFiltered-Gemini	22.07	18.80	20.30	7	35.22	30.65	32.78	7	4.32	6	80.00	7	0.99	2
Pattern-Based-MWE-Id.	2.83	17.40	4.87	8	14.63	70.40	24.23	8	4.17	7	70.00	8	0.98	3
pmi-mwe-scorer	0.29	1.40	0.48	9	5.50	20.87	8.70	9	1.95	8	7.00	9	1.00	1

Table 18: Polish (pl) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.78$, $\rho = 0.95$); Richness: ($r = 0.93$, $\rho = 0.97$); SE: ($r = -0.75$, $\rho = -0.82$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
Sahara-Tokenizers	40.00	49.60	44.29	1	50.51	53.67	52.04	1	4.73	3	147.00	3	0.95	4
MTLB-STRUCT	50.17	30.20	37.70	2	64.22	31.42	42.19	2	4.13	4	88.00	4	0.92	5
IPN	17.74	52.20	26.48	3	23.79	69.92	35.50	3	4.98	2	182.00	2	0.96	3
baseline-gpt-oss-120b	15.93	51.60	24.34	4	19.96	76.50	31.66	4	5.06	1	192.00	1	0.96	3
MorphoFiltered-Gemini	8.37	13.80	10.42	5	19.45	26.50	22.43	5	3.92	5	56.00	5	0.97	2
Pattern-Based-MWE-Id.	0.61	7.00	1.13	6	6.73	43.00	11.63	6	1.74	7	8.00	6	0.84	6
pmi-mwe-scorer	0.22	1.20	0.38	7	4.93	17.92	7.73	7	1.79	6	6.00	7	1.00	1

Table 19: Brazilian Portuguese (pt) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.80$, $\rho = 0.64$); Richness: ($r = 0.72$, $\rho = 0.68$); SE: ($r = 0.10$, $\rho = -0.41$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
romanian-bert	91.03	80.88	85.65	1	94.35	81.98	87.73	2	5.39	2	275.00	2	0.96	4
BeeParser	84.98	82.27	83.60	2	92.36	85.19	88.63	1	5.43	1	280.00	1	0.96	4
MTLB-STRUCT	83.71	80.88	82.27	3	91.97	83.76	87.68	3	5.39	2	272.00	3	0.96	4
Sahara-Tokenizers	61.98	71.12	66.23	4	84.06	79.48	81.71	4	5.23	3	236.00	4	0.96	4
IPN	37.12	50.80	42.89	5	42.83	62.09	50.69	5	5.05	4	187.00	5	0.97	3
baseline-gpt-oss-120b	26.09	38.25	31.02	6	34.87	60.84	44.33	6	4.82	5	145.00	6	0.97	3
MorphoFiltered-Gemini	27.02	15.34	19.57	7	42.82	26.58	32.80	7	4.17	6	68.00	7	0.99	2
Pattern-Based-MWE-Id.	3.06	14.34	5.05	8	18.15	68.69	28.71	8	3.77	7	54.00	8	0.95	5
pmi-mwe-scorer	0.12	0.40	0.19	9	7.63	22.48	11.39	9	0.69	8	2.00	9	1.00	1

Table 20: Romanian (ro) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.78$, $\rho = 0.97$); Richness: ($r = 0.98$, $\rho = 0.98$); SE: ($r = -0.55$, $\rho = -0.46$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
MTLB-STRUCT	76.98	68.06	72.25	1	83.88	70.84	76.81	2	5.00	3	193.00	3	0.95	4
bert-multilingual-trial	71.15	72.85	71.99	2	80.95	77.00	78.93	1	5.06	2	205.00	1	0.95	4
Sahara-Tokenizers	44.72	53.29	48.63	3	73.07	64.40	68.46	3	4.84	4	159.00	4	0.95	4
IPN	21.70	62.67	32.24	4	24.41	72.95	36.58	4	5.10	1	202.00	2	0.96	3
baseline-gpt-oss-120b	14.67	42.71	21.84	5	18.12	58.60	27.68	6	4.74	5	138.00	5	0.96	3
MorphoFiltered-Gemini	24.67	14.97	18.63	6	38.95	23.83	29.57	5	4.02	6	61.00	6	0.98	2
Pattern-Based-MWE-Id.	1.51	12.38	2.69	7	10.88	65.13	18.65	7	3.58	7	41.00	7	0.96	3
pmi-mwe-scorer	0.09	0.60	0.16	8	3.33	17.30	5.58	8	1.10	8	3.00	8	1.00	1

Table 21: Slovenian (sl) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.71$, $\rho = 0.83$); Richness: ($r = 0.86$, $\rho = 0.88$); SE: ($r = -0.73$, $\rho = -0.90$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
BeeParser	72.87	76.80	74.78	1	79.00	79.38	79.19	1	5.62	1	312.00	1	0.98	2
MTLB-STRUCT	71.89	76.20	73.98	2	76.35	77.58	76.96	2	5.61	2	308.00	2	0.98	2
Sahara-Tokenizers	53.46	63.40	58.01	3	70.03	68.97	69.49	3	5.43	3	258.00	3	0.98	2
IPN	35.99	57.80	44.36	4	37.88	64.90	47.84	4	5.40	4	247.00	4	0.98	2
baseline-gpt-oss-120b	20.36	34.00	25.47	5	27.00	52.32	35.62	6	4.91	5	147.00	5	0.98	2
MorphoFiltered-Gemini	19.70	8.00	11.38	6	30.62	13.53	18.77	8	3.45	8	34.00	7	0.98	2
pmi-mwe-scorer	2.08	7.60	3.27	7	11.12	37.56	17.16	9	3.60	6	37.00	6	1.00	1
Pattern-Based-MWE-Id.	1.84	9.40	3.08	8	13.96	61.49	22.76	7	3.49	7	37.00	6	0.97	3
bert-multilingual-trial	0.35	0.60	0.44	9	44.36	38.69	41.33	5	1.10	9	3.00	8	1.00	1

Table 22: Serbian (sr) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.84$, $\rho = 0.95$); Richness: ($r = 0.98$, $\rho = 0.95$); SE: ($r = -0.38$, $\rho = -0.27$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
bert-multilingual-trial	65.97	63.05	64.48	1	75.72	70.65	73.09	1	5.14	1	207.00	1	0.96	3
MTLB-STRUCT	67.19	60.04	63.41	2	73.96	67.52	70.60	3	5.13	2	204.00	2	0.97	2
BeeParser	65.35	59.84	62.47	3	76.62	67.30	71.66	2	5.14	1	204.00	2	0.97	2
Sahara-Tokenizers	47.18	55.42	50.97	4	64.89	66.41	65.64	4	5.07	3	191.00	3	0.97	2
IPN	21.16	45.38	28.86	5	24.65	64.40	35.65	5	5.01	4	176.00	4	0.97	2
MorphoFiltered-Gemini	20.17	34.14	25.35	6	26.00	50.56	34.34	6	4.72	6	131.00	6	0.97	2
baseline-gpt-oss-120b	17.27	39.96	24.12	7	21.90	69.87	33.34	7	4.81	5	144.00	5	0.97	2
Pattern-Based-MWE-Id.	1.53	9.04	2.62	8	11.10	49.33	18.12	8	3.47	7	36.00	7	0.97	2
pmi-mwe-scorer	1.39	6.22	2.27	9	4.99	22.99	8.21	9	3.34	8	29.00	8	0.99	1

Table 23: Swedish (sv) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.86$, $\rho = 0.95$); Richness: ($r = 0.93$, $\rho = 0.98$); SE: ($r = -0.61$, $\rho = -0.73$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
MTLB-STRUCT	59.29	49.50	53.95	1	69.44	53.83	60.65	1	5.05	3	179.00	3	0.97	3
Sahara-Tokenizers	37.60	45.53	41.19	2	58.67	54.53	56.52	2	5.04	4	174.00	4	0.98	2
IPN	25.97	60.04	36.25	3	32.39	73.69	45.00	3	5.39	1	241.00	1	0.98	2
baseline-gpt-oss-120b	22.43	45.53	30.05	4	28.95	60.63	39.19	4	5.13	2	185.00	2	0.98	2
MorphoFiltered-Gemini	29.37	14.71	19.60	5	45.54	22.21	29.86	5	4.23	5	70.00	5	1.00	1
Pattern-Based-MWE-Id.	2.44	15.90	4.23	6	13.35	57.84	21.70	6	4.08	6	65.00	6	0.98	2
pmi-mwe-scorer	0.54	2.39	0.88	7	7.50	26.92	11.73	7	2.48	7	12.00	7	1.00	1

Table 24: Ukrainian (uk) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.82$, $\rho = 0.71$); Richness: ($r = 0.84$, $\rho = 0.71$); SE: ($r = -0.70$, $\rho = -0.78$).

	SWE r	SWE ρ	Richness r	Richness ρ	SE r	SE ρ
Egyptian (EGY)	-0.03	-0.03	-0.03	0.09	-0.21	0.32
Modern Greek (EL)	0.58	0.34	0.66	0.50	-0.63	-0.54
Persian (Farsi) (FA)	0.93	1.00	0.99	1.00	-0.52	-0.51
French (FR)	0.92	1.00	0.99	1.00	-0.32	-0.54
Ancient Greek (GRC)	0.81	0.71	0.83	0.71	0.36	-0.09
Hebrew (HE)	0.80	0.96	0.92	0.96	-0.66	-0.59
Japanese (JA)	0.89	1.00	0.98	0.98	-0.38	-0.36
Georgian (KA)	0.29	0.46	0.53	0.86	-0.30	-0.43
Latvian (LV)	0.54	0.50	0.62	0.67	-0.62	-0.67
Dutch (NL)	0.97	0.90	0.95	0.86	0.27	-0.11
Polish (PL)	0.78	0.95	0.93	0.97	-0.75	-0.82
Brazilian Portuguese (PT)	0.80	0.64	0.72	0.68	0.10	-0.41
Romanian (RO)	0.78	0.97	0.98	0.98	-0.55	-0.46
Slovenian (SL)	0.71	0.83	0.86	0.88	-0.73	-0.90
Serbian (SR)	0.84	0.95	0.98	0.95	-0.38	-0.27
Swedish (SV)	0.86	0.95	0.93	0.98	-0.61	-0.73
Ukrainian (UK)	0.82	0.71	0.84	0.71	-0.70	-0.78
Mean	0.72	0.76	0.81	0.81	-0.39	-0.46

Table 25: Subtask 1 – Pearson (r) and Spearman (ρ) correlations between Global MWE-based F1 score and diversity of systems' true positives – SWE: Shannon-Weaver entropy, SE: Shannon evenness.

System	Performance				Diversity (of new vocabulary)					
	Automatic eval.		Manual eval.		Variety		Balance		Var.-bal.	
	gmBS	Rank	Manual score	Rank	Richness	Rank	SE	Rank	SWE	Rank
Star-Paraphraser-Cosine	93.90	1	64.82	3	236.00	4	0.83	4	4.54	4
Star-Paraphraser-Multiagent	89.46	2	79.25	1	456.00	2	0.90	3	5.48	2
baseline-gpt-oss-120b	77.55	3	72.70	2	326.00	3	0.92	2	5.33	3
MISP	49.53	4	29.25	4	564.00	1	0.93	1	5.89	1

Table 26: French (fr) subtask 2 results – gm-BS: global masked BERT-score, SWE: Shannon-Weaver entropy, SE: Shannon evenness.

System	Performance				Diversity (of new vocabulary)					
	Automatic eval.		Manual eval.		Variety		Balance		Var.-bal.	
	gmBS	Rank	Manual score	Rank	Richness	Rank	SE	Rank	SWE	Rank
baseline-gpt-oss-120b	63.99	1	24.22	1	804.00	2	0.98	1	6.54	1
MISP	33.75	2	3.39	2	971.00	1	0.89	2	6.08	2

Table 27: Georgian (ka) subtask 2 results – gm-BS: global masked BERT-score, SWE: Shannon-Weaver entropy, SE: Shannon evenness.

System	Performance				Diversity (of new vocabulary)					
	Automatic eval.		Manual eval.		Variety		Balance		Var.-bal.	
	gmBS	Rank	Manual score	Rank	Richness	Rank	SE	Rank	SWE	Rank
baseline-gpt-oss-120b	80.21	1	55.88	1	619.00	2	0.92	2	5.93	2
MISP	58.59	2	38.23	2	798.00	1	0.93	1	6.20	1

Table 28: Brazilian Portuguese (pt) subtask 2 results – gm-BS: global masked BERT-score, SWE: Shannon-Weaver entropy, SE: Shannon evenness.

System	Performance				Diversity (of new vocabulary)					
	Automatic eval.		Manual eval.		Variety		Balance		Var.-bal.	
	gmBS	Rank	Manual score	Rank	Richness	Rank	SE	Rank	SWE	Rank
GPT-CREATIVE	89.25	1	77.31	1	235.00	3	0.98	1	5.36	3
baseline-gpt-oss-120b	74.74	2	46.17	2	742.00	2	0.93	2	6.14	2
MISP	57.01	3	22.66	3	1096.00	1	0.91	3	6.36	1

Table 29: Romanian (ro) subtask 2 results – gm-BS: global masked BERT-score, SWE: Shannon-Weaver entropy, SE: Shannon evenness.

MWE-2026 Shared Task: AdMIRe 2 Advancing Multimodal Idiomaticity Representation

Doğukan Arslan[1], Rodrigo Wilkens[2], Wei He[2], Dilara Torunoğlu Selamet[1], Thomas Pickard[3], Aline Villavicencio[2,3], Adriana Pagano[4], Gülşen Eryiğit[1]
[1] Istanbul Technical University, Türkiye
[2] University of Exeter, UK
[3] University of Sheffield, UK
[4] Federal University of Minas Gerais, Brazil
{arslan.dogukan, torunoglud, gulsen.cebiroglu}@itu.edu.tr, tmrpickard1@sheffield.ac.uk
{w.he, r.wilkens, a.villavicencio}@exeter.ac.uk, apagano@ufmg.br

Abstract

Idiomatic expressions present a unique challenge in NLP, as their meanings are often not directly inferable from their constituent words. Despite recent advancements in large language models, idiomaticity remains a significant obstacle to robust semantic representation. We present datasets and task results for MWE-2026 Shared Task 2: Advancing Multimodal Idiomaticity Representation 2 (AdMIRe 2), which challenges the community to assess and improve models' ability to interpret idiomatic expressions in multimodal contexts across multiple languages. Participants competed in an image ranking task in which, for each item, systems receive a context sentence containing a potentially idiomatic expression (PIE) and five candidate images. Participating systems are required to predict the sentence type (i.e., idiomatic vs. literal) for the given context and rank the images by how well they depict the intended meaning in that context. Among the participating systems the most effective methods include pipelines utilizing closed-source commercial models such as Gemini 2.5 and GPT-5, and employing chain-of-thought reasoning strategies. Methods to mitigate language models' bias towards literal interpretations and ensembles to smooth out variance were common.

1 Introduction

Idioms constitute a class of multiword expressions (MWEs) that remains challenging for state-of-the-art language models, as their meanings are often not predictable from the meanings of their constituent words (Dankers et al., 2022; Villavicencio et al., 2005). For instance, the expression *devil's advocate* is not typically used with literal denotation derived from its component words, but rather construes the meaning of someone who presents a contentious opinion in order to test an opposing argument or provoke debate. Idiomatic expressions may further give rise to ambiguity between a literal, compositional interpretation and an idiomatic, non-compositional one (He et al., 2024). These characteristics make idioms a particularly informative testbed for investigating how current language models represent and process meaning. While large language models (LLMs) perform well on general benchmarks, it is still unclear to what extent they consistently exhibit good understanding of figurative language (Mi et al., 2025; Phelps et al., 2024), even for well-resourced languages such as English.

These challenges have recently been highlighted in shared evaluations. For instance, the first edition of this task (SemEval-2025 Task 1: AdMIRe; Pickard et al., 2025) focused on two languages, English and Portuguese, to assess models' ability to interpret idiomatic expressions in multimodal contexts and the PARSEME 2.0 shared task (Scholivet et al., 2026) proposed two multilingual challenges targetting MWEs: (a) their identification and (b) their paraphrasing. In addition, there are several benchmark datasets dedicated to the processing of idiomatic expressions in text (e.g. Chakrabarty et al., 2022; Haagsma et al., 2020; Tedeschi et al., 2022; Tayyar Madabushi et al., 2021; Garcia et al., 2021; Mi et al., 2025; Arslan et al., 2025). While current models may display competitive performance on some of these datasets, it is unclear to what extent they actually require that language models possess good representations of idiom meaning (Boisson et al., 2023; He et al., 2024), or whether models are benefiting from other artifacts to address these tasks.

Moreover, even if the addition of a visual modality (alongside text) to idiom processing could lead to more informative clues being available to disambiguate and interpret potentially idiomatic expressions, it is not certain whether models benefit from the additional information. Indeed, performance on datasets like IRFL (Yosef et al., 2023) and V-FLUTE (Saakyan et al., 2025) indicates that idiomaticity processing is more difficult for vision-language models (VLMs) to perform.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 276–287

In this edition of the AdMIRe shared task, in an attempt to determine the multilingual coverage and generalizability of the results obtained by available models, we expand the number of languages, adding new evaluation instances for Chinese, Georgian, Greek, Igbo, Kazakh, Norwegian, Portuguese (Portugal), Portuguese (Brazil), Russian, Serbian, Slovak, Slovenian, Spanish (Ecuador), Turkish, and Uzbek to the existing English and Portuguese (Brazil) training data. These languages provide variation in terms of language families and scripts, and also in terms of NLP resources. Two variants of one of the languages are also included: Brazilian Portuguese and European Portuguese. We incorporate visual (§2) modalities for all 15 languages in an effort to promote the construction of higher-quality semantic representations of idioms. Our dataset incorporates items in both English (EN) and Brazilian Portuguese (PT-BR) as part of the training data, while the other languages are available only at test time, as unseen items. As in AdMIRe 1, we use nominal compounds and verbal idioms having interpretations in literal and idiomatic senses which are both plausible and imageable. This paper presents the task (§3), participating systems and results (§4) and finishes with discussions (§5) conclusions, limitations and future work (§6).

2 Dataset

Following the first edition of the AdMIRe shared task, Torunoğlu-Selamet et al. (2026) recently introduced a cross-lingual benchmark for multimodal idiomaticity understanding, as an initiative under the UniDive COST Action (Savary et al., 2024). The paper followed the same data creation strategy from Pickard et al. (2025) and introduced data for a large number of languages.[1]

For each language, the dataset contains around 60 potentially idiomatic expressions (PIEs), expressions that can be interpreted idiomatically, whether or not they are used that way in context. Annotators select a subset of the English PIEs from the dataset used in first version of the AdMIRe shared task and provide their counterparts in each target language. For the text modality, each of these is provided with at least two context sentences where the PIE is used with either its idiomatic or literal meaning. This allows verifying how well and how consistently models can distinguish these two uses for each of the target languages. The context sentences originate from diverse sources, including naturally occurring corpus data and sentences produced through expert construction or large language models. In addition, for the visual part, 5 images were machine-generated using manually-written prompts and validated by the language experts. These images cover a spectrum from fully literal to fully idiomatic interpretations of the expression, along with a semantically unrelated distractor (i.e., strongly figurative, mildly figurative, mildly literal, strongly literal, and distractor). Figure 1 provides an example set of images for the expression *green fingers*. Additionally, auto-generated captions are provided for each image in the text-only track and as part of the textual information available to the models. That means that for each language and for each PIE the models have available:

- 2 manually validated context sentences (one literal, one idiomatic)
- 5 automatically generated then manually validated images
- 5 automatically generated captions

3 Task Description

Given a context sentence containing a PIE and a set of five images, the task is to rank the images based on how well they depict the meaning of the PIE used in that sentence. A variation of the task (i.e., text-only) also allows for unimodal settings, where given a sentence and five text captions (each describing the content of one of the images) the goal is to rank the image captions on how accurately they capture the meaning of the PIE.

Publicly available training data from the first edition of the task was provided to shared-task participants for English and Brazilian Portuguese only, while no training data was released for the remaining languages. AdMIRe 2 excluded English from the set of test languages and introduced newly-created test sets for Portuguese, enabling evaluation across two language variants: a new unseen set for Brazilian Portuguese and a new set for European Portuguese.

3.1 Evaluation

We set an expected rank ordering of the 5 images following the sense in which the expression is used

[1] The full resource contains data for 34 different languages; however, due to the strict and unified formatting requirements for the shared task, only 15 languages were fully prepared at the time of the AdMIRe 2 shared task data release and were included in the evaluation.

(a) The image depicts an elderly man engaged in gardening activities. He is crouching down next to a small garden bed...

(b) The image depicts an elderly man with gray hair and a mustache, wearing a light-colored shirt and dark pants...

(c) The image depicts a close-up of a human hand with the palm facing upwards and fingers slightly curled...

(d) The image depicts a cartoon-style illustration of a young girl with brown hair tied back in a ponytail...

(e) The image depicts an elderly man playing the piano. He is seated on a black stool with orange accents...

Figure 1: Data example for *green fingers*. Images generated using Midjourney. Captions are displayed partially. (Torunoğlu-Selamet et al., 2026)

in the context sentence. The image strongly associated with the target sense is expected to be ranked first, followed by the mildly associated one. The images for the other senses and the 'distractor' image can follow interchangeably. For instance, for an idiomatic use of *green fingers* in a context sentence, strongly figurative and mildly figurative are expected to be ranked first. For the images in Figure 1, this would produce, for instance, $[a, b, d, e, c]$.

Performance for the task is assessed with two key metrics: a) Top-1 Image Accuracy, which measures only the correct identification of the **most** representative image and b) Normalized Discounted Cumulative Gain (nDCG) (Järvelin and Kekäläinen, 2002), which was also adopted in the first edition of AdMIRe, as it is an established information retrieval metric that not only captures the fraction of retrieved relevant information but also takes into account their correct ordering.

Normalized Discounted Cumulative Gain (nDCG) is defined as

$$\text{nDCG} = \frac{DCG_n}{iDCG_n} = \frac{\sum_{i=1}^{n} \frac{rel_i}{\log_2(i+1)}}{\sum_{i'=1}^{n} \frac{rel^*_{i'}}{\log_2(i'+1)}},$$

where n is the number of items considered, rel_i is the relevance score (gain) of the i-th item in the system's ranking, $rel^*_{i'}$ is the relevance score of the i'-th item in the ideal ranking, and $iDCG$ is DCG of the ideal ordering of results.

Because our expected order of images is somewhat arbitrary (for a literal instance of a given expression, the idiomatic depictions are essentially no more relevant than the distractor), after experimentation we adopt relevance scores ($rel^*_{i'}$) of $[3, 1, 0, 0, 0]$ for the five image positions; this allows the metric to capture some of the relevant semantics beyond the top image accuracy without penalising systems which permute the order of the low-relevance images. The maximum (ideal) DCG score obtainable is therefore 3.631 and nDCG is bounded between 0 and 1, with higher values reflecting better ranking quality.

Competition rankings for the task are based on top image accuracy, with nDCG breaking ties.

4 Participating Systems and Results

The AdMIRe shared task competitions[2] were configured using the Codabench platform (Xu et al., 2022), attracting 27 registered participants in the images & text track and 22 registered participants in the text-only track. Users were allowed to submit multiple times during the competition, and their best result was used for evaluation. Submissions during the test phase (which determined the final leaderboard position) were limited to 10 in order to discourage 'gaming' the system while allowing participants to evaluate more than one approach if desired.

Once the competition ended, teams were asked to complete a brief questionnaire outlining their approach and enabling us to link CodaBench usernames with team names in their system description papers. Only teams who submitted a system description paper are included in the official task leaderboards. A total of 10 official team submissions were received.

[2] Available at https://www.codabench.org/competitions/10547/ and https://www.codabench.org/competitions/10548/

Team	Rank	Top-1 Accuracy			nDCG Score		
		Overall	Literal	Idiomatic	Overall	Literal	Idiomatic
ITUNLP	1	0.60 ± 0.1	0.67 ± 0.2	0.55 ± 0.1	0.85 ± 0.1	0.88 ± 0.1	0.83 ± 0.0
DCSN-NLP	2	0.53 ± 0.1	0.62 ± 0.2	0.46 ± 0.1	0.81 ± 0.0	0.85 ± 0.1	0.77 ± 0.0
ITUNLP2	3	0.52 ± 0.3	0.53 ± 0.3	0.52 ± 0.3	0.70 ± 0.4	0.70 ± 0.4	0.70 ± 0.4
tiberiucarp	4	0.50 ± 0.1	0.54 ± 0.2	0.46 ± 0.1	0.80 ± 0.0	0.84 ± 0.1	0.78 ± 0.0
PolyFrame	5	0.35 ± 0.1	0.57 ± 0.1	0.16 ± 0.0	0.73 ± 0.0	0.85 ± 0.0	0.62 ± 0.0
VisAffect	6	0.33 ± 0.0	0.13 ± 0.1	0.47 ± 0.1	0.72 ± 0.0	0.59 ± 0.0	0.81 ± 0.0
IdiomRanker-X	7	0.30 ± 0.2	0.48 ± 0.3	0.13 ± 0.1	0.58 ± 0.3	0.69 ± 0.4	0.49 ± 0.3
3K2T	8	0.13 ± 0.2	0.13 ± 0.2	0.13 ± 0.2	0.21 ± 0.4	0.21 ± 0.4	0.21 ± 0.4

Table 1: Leaderboard results for the image and text track. Macro-averaged Top-1 Accuracy and nDCG scores are reported overall and separately for literal and idiomatic sentences, together with their standard deviations. Teams are ranked by overall Top-1 Accuracy.

Team	Rank	Top-1 Accuracy			nDCG Score		
		Overall	Literal	Idiomatic	Overall	Literal	Idiomatic
ITUNLP	1	0.56 ± 0.1	0.61 ± 0.2	0.51 ± 0.1	0.83 ± 0.0	0.86 ± 0.1	0.81 ± 0.0
LST	2	0.41 ± 0.1	0.58 ± 0.1	0.28 ± 0.1	0.76 ± 0.0	0.85 ± 0.1	0.68 ± 0.0
alexandru412	3	0.32 ± 0.2	0.33 ± 0.2	0.29 ± 0.3	0.59 ± 0.3	0.60 ± 0.3	0.57 ± 0.3
PolyFrame	4	0.32 ± 0.1	0.48 ± 0.1	0.19 ± 0.1	0.71 ± 0.0	0.81 ± 0.1	0.63 ± 0.0

Table 2: Leaderboard results for the text-only track. Macro-averaged Top-1 Accuracy and nDCG scores are reported overall and separately for literal and idiomatic sentences, together with their standard deviations. Teams are ranked by overall Top-1 Accuracy.

4.1 Results

The team's ranking is shown in Tables 1 and 2, where the former includes both modality types and the latter only text. The tables report the macro-averaged mean and standard deviation of accuracy and nDCG scores for literal and idiomatic items, as well as the overall performance. Systems that do not support a given language are assigned a score of zero for that language when computing the macro-average.

In this version of the AdMIRe shared task, teams were challenged with 15 different languages. Most teams tested their solutions in 15 languages, except ITUNLP2 (Umut and Şenceylan, 2026), IdiomRanker-X (Çolak, 2026), and alexandru412 (Alexandru-Marian, 2026) (12 languages), and 3K2T (Kömürcü and Temel, 2026) (3 languages). Although we observe performance variation across the different languages, in general, performance within each language is fairly consistent, as illustrated in Figure 2, which shows the average performance and standard deviation for each language[3]. Detailed performance figures by language can be seen in the Appendix.

[3]ISO 639-1 codes have been used to represent languages.

Finally, the overall results obtained by the participating teams consistently display better accuracy for literal than for idiomatic items (Tables 1 and 2). The exception is the system by VisAffect (Bilen et al., 2026), which got better accuracy for idiomatic items in both Image and Text and Text only tracks.

4.2 Popular Approaches

Model Types Participating teams employed a variety of approaches to the AdMIRe 2 task, predominantly relying on large generative language models (LLMs) and vision-language models (VLMs). For text generation and reasoning, teams frequently utilized the GPT series (specifically GPT-4o and GPT-5; OpenAI, 2024, 2025), Qwen (versions 2.5 and 3) (Bai et al., 2023), DeepSeek models (DeepSeek-AI, 2025) and Gemini 2.5 Pro (Google, 2025) and 3 Pro Preview (DeepMind, 2025). For embeddings and vision-language alignment, there was a shift towards newer architectures; while standard CLIP (Radford et al., 2021) variants remained popular (used by DCSN-NLP (Cotigă and Nisioi, 2026) and tiberiucarp (Carp, 2026)), some teams also adopted SigLIP2 (Tschannen et al., 2025) and Jina-CLIP-v2 (Koukounas et al., 2024); i.e. PolyFrame (Hosseini-

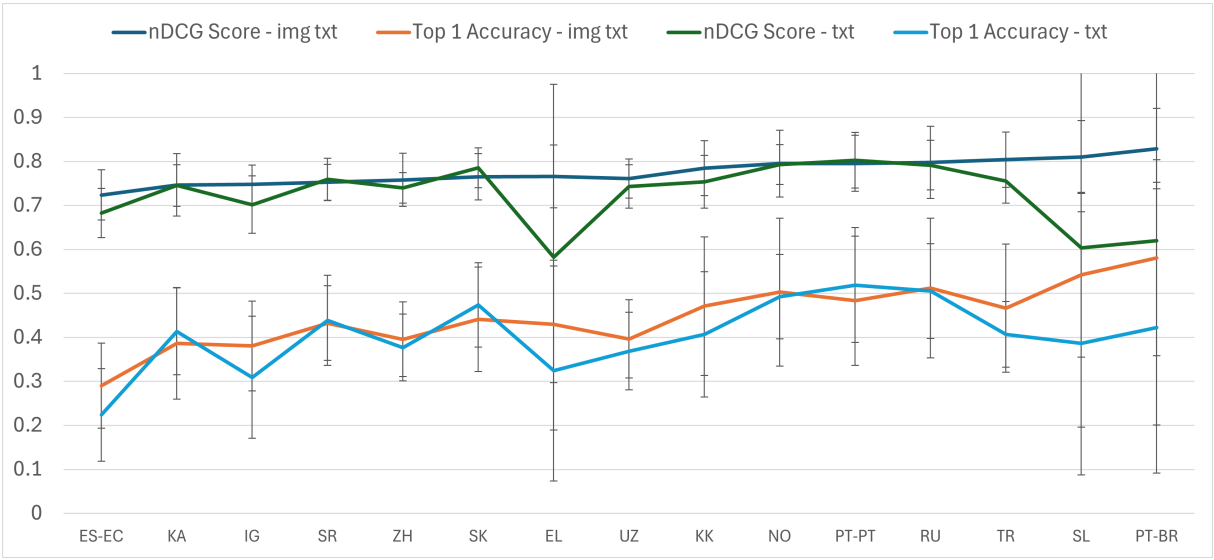


Figure 2: Average (and standard deviation) of evaluation metrics by language

Kivanani, 2026) and VisAffect (Bilen et al., 2026) respectively. Dense retrieval models like BGE-M3 (Chen et al., 2024) and multilingual encoders like XLM-RoBERTa (Conneau et al., 2020) were also widely used for text-specific ranking components.

Pipeline components Most teams implemented multi-stage pipelines rather than end-to-end solutions. A dominant pattern involved an initial binary classification of the context sentence as literal or idiomatic, triggering specialized downstream processing. To bridge the modality gap, several teams used LLMs to generate auxiliary text—such as semantic glosses or visual descriptions—to better guide the vision models, while others employed hybrid architectures that fused scores from direct vision-language matching and text-caption retrieval.

Ensembles and Fusion Robustness was often achieved through ensemble techniques. DCSN-NLP utilized a voting mechanism across an ensemble of three CLIP models (ViT-H-14, ViT-L-14, ViT-g-14). Polyframe adopted weighted Borda rank aggregation to fuse outputs from vision and text streams. LST (QIU et al., 2026) employed a large-model ensemble, aggregating outputs from GPT-4, Qwen-Plus, and DeepSeek-V3 prompting strategies.

Bias Mitigation Teams actively addressed the "literal bias" of LLMs (Phelps et al., 2024; Mi et al., 2025). PolyFrame implemented "idiom synonym replacement" for idiomatic instances, replacing the target expression with a non-figurative synonym to prevent the model from grounding visual features in the literal constituent words. Alexandru412 introduced stochastic option shuffling during inference to mitigate positional bias in multiple-choice ranking.

Data Augmentation and Cross-Lingual Strategies Zero-shot transfer was critical for handling the 15 target languages. alexandru412 utilized a test-time translation strategy (converting context sentences to English) to leverage an English-fine-tuned Qwen model. IdiomRanker-X employed dynamic prompting with "focus markers" to guide attention. PolyFrame relied on the inherent multilingual capabilities of SigLIP2 and BGE-M3 for zero-shot ranking without language-specific fine-tuning.

4.3 Most Effective Approaches

4.3.1 Text & Images Methods

The top-performing system from ITUNLP (Site et al., 2026) achieved the highest accuracy across both the Multimodal (Text + Image) and Text-Only leaderboards. Their approach leveraged a "hybrid LVLM pipeline" that combined the reasoning capabilities of GPT-5.1 with the multimodal understanding of Gemini 2.5 Pro. By delegating the initial semantic analysis to a strong reasoning model and the visual grounding to a specialized VLM, they effectively mitigated the noise often seen in end-to-end zero-shot inference.

The second-placed team, DCSN-NLP , introduced a "Hierarchical Multimodal Reasoning"

strategy to align abstract idioms with visual features. Instead of matching images directly to the idiomatic expression, their pipeline first used an LLM (such as GPT-5 or GPT-4o) to generate auxiliary text—specifically, a visual description of the literal meaning and an explanation of the idiomatic meaning. These generated descriptions were then used to query an ensemble of three CLIP models. While a voting mechanism was used to narrow the selection to the top two candidates, the final selection was performed by the LLM, which compared the finalists to make the ultimate decision.

As mentioned in the previous section, to address the "literal bias" present in vision-language models, PolyFrame (rank 5) implemented a targeted transformation step. For sentences classified as idiomatic, they replaced the target expression with a non-figurative synonym before passing the text to the vision encoder. This simple yet effective substitution prevents the VLM (in their case, SigLIP2) from grounding the literal objects and focuses the ranking on the semantic payload of the expression.

4.3.2 Text-Only Methods

The Text-Only track demonstrated that strong language models can rival multimodal systems by exploiting caption semantics. LST secured second place (Top-1 Accuracy: 0.41) using an ensemble of GPT-4, Qwen-Plus, and DeepSeek, effectively reasoning over captions without accessing pixel data. For teams participating in both tracks, the contribution of the visual modality varied. The winning team, ITUNLP, saw a drop in accuracy (0.60 to 0.56) when removing images, confirming the value of their VLM pipeline. Furthermore, PolyFrame observed a minimal performance gap (0.35 vs. 0.32) as well.

5 Discussion

We were pleased to see that most teams covered all 15 languages in the shared task. The results obtained confirm that idiomatic processing is still challenging for models, and that they are still more accurate when processing literal than idiomatic instances. Moreover, visual data seems to be helping disambiguation for both literal and idiomatic items. However, more in-depth analyses of the specific causes of error in each of the languages is still needed, and will be left for future work.

6 Conclusions

The AdMIRe tasks provide a particularly original approach to assessing models' idiom understanding by grounding figurative meaning in both textual and visual contexts. AdMIRe 2 establishes a challenging and carefully designed benchmark for multilingual and multimodal idiomaticity understanding. By combining textual contexts with visually grounded representations that span the idiomatic–literal continuum, the task enables a fine-grained evaluation of models' ability to disambiguate figurative language across languages and modalities. This shared task paves the way for further cross-lingual analyses and provides a valuable benchmark for systematically assessing the capabilities of large language models and vision–language models in idiom understanding. While the top-performing system attains an nDCG score of 85%, the task remains challenging for today's systems, leaving clear room for improvement.

Limitations

Zero-shot setting In AdMIRe 2, while training data was provided for English and Portuguese (inherited from the previous iteration; Pickard et al., 2025), the shared task introduced additional languages for which no labelled training examples were released. This experimental design forced systems to rely on zero-shot cross-lingual transfer or static pretrained knowledge rather than learning from task-specific examples. Consequently, models could not be fine-tuned to capture the specific cultural and linguistic nuances of idioms in these new set of languages, making performance heavily dependent on the coverage and biases of the underlying LLMs or VLMs rather than their ability to adapt to the specific task distribution.

Cultural background The datasets were constructed by creators who work in academic settings, and who are native speakers of the language that they study. The second edition replicates some of the limitations of the first edition, simply by means of adopting the same protocol. The language experts worked independently on their languages, and the examples they selected are also impacted by the constraints of the shared task and the parallel effort. Although the language specific examples have been carefully curated by these language experts, a subsequent independent crosslingual validation would require native speaker knowledge of the target lan-

guages and is left for future work. Our backgrounds and experiences will certainly have influenced the idiomatic expressions and context sentences we selected, the visual representations we favoured and so on.

AI tools used As for the first edition, the datasets are likely to reflect biases and limitations present in the tools used to construct them, especially the image generation and captioning models. For instance, efforts to introduce diversity in the images depicted depended on the quality of the image generation tools employed, and were not always successfully achieved.

Acknowledgments

The authors would like to acknowledge the contributions of the language leader collaborators who made this work possible. We would also like to thank the MWE 2026 Workshop Chairs, the members of the UniDive COST Action (Savary et al., 2024) and the SIGLEX-MWE Special Interest Group for their continued feedback and support. We dedicate this work to the memory of Federico Sangati and Silvio Ricardo Cordeiro.

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology). This work was also partly supported by the UKRI AI Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications [UKRI grant number EP/S023062/1]. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

References

Cristea Alexandru-Marian. 2026. alexandru412 at MWE-2026 AdMIRe 2: Dominating text-only idiom understanding via cross-lingual transfer and augmentation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. Association for Computational Linguistics.

Doğukan Arslan, Hüseyin Anıl Çakmak, Gulsen Eryigit, 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.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Barış Bilen, Ali Azmoudeh, Hazım Kemal Ekenel, and Hatice Kose. 2026. VisAffect at MWE-2026 AdMIRe 2: IMMCAN idiom multimodal cross-attention network. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. 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*, pages 6581–6590, Singapore. Association for Computational Linguistics.

Andrei Tiberiu Carp. 2026. tiberiucarp at MWE-2026 AdMIRe 2: GLIMMER-Gloss-based image multiword meaning expression ranker. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. 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.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

David Cotigă and Sergiu Nisioi. 2026. DCSN-NLP at MWE-2026 AdMIRe 2: Bridging literal and figurative meaning through hierarchical multimodal reasoning. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. Association for Computational Linguistics.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Google DeepMind. 2025. Gemini 3 system card. Technical report, Google.

DeepSeek-AI. 2025. DeepSeek-V3 technical report. *Preprint*, arXiv:2412.19437.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2730–2741, Online. Association for Computational Linguistics.

Google. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the 12th language resources and evaluation conference*, pages 279–287, Marseille, France. European Language Resources Association.

Wei He, Tiago Kramer Vieira, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2024. Investigating idiomaticity in word representations. *Computational Linguistics*, pages 1–48.

Nina Hosseini-Kivanani. 2026. Polyframe at MWE-2026 AdMIRe 2: When words are not enough: Multimodal idiom disambiguation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. Association for Computational Linguistics.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Andreas Koukounas, Georgios Mastrapas, Sedigheh Eslami, Bo Wang, Mohammad Kalim Akram, Michael Günther, Isabelle Mohr, Saba Sturua, Nan Wang, and Han Xiao. 2024. jina-clip-v2: Multilingual multimodal embeddings for text and images. *arXiv preprint arXiv:2412.08802*.

Kubilay Kağan Kömürcü and Tuğçe Temel. 2026. 3K2T at MWE-2026 AdMIRe 2: CARIM– category-aware reasoning for idiomatic multimodality. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. 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.

OpenAI. 2024. GPT-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. GPT-5 system card. Technical report, OpenAI.

Dylan Phelps, Thomas M. R. Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL.

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.

Le QIU, Yu-Yin Hsu, and Emmanuele Chersoni. 2026. Lst at AdMIRe 2: Advancing multimodal idiomaticity representation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

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, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz, and Alina Wróblewska. 2024. UniDive: A COST action on universality, diversity and idiosyncrasy in language technology. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 372–382, Torino, Italia. ELRA and ICCL.

Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Carp, and Vasile Pais. 2026. Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morocco. Association for Computational Linguistics.

Atakan Site, Oğuz Ali Arslan, and Gülşen Eryiğit. 2026. ITUNLP at MWE-2026 AdMIRe 2: A Zero-Shot LLM pipeline for multimodal idiom understanding and ranking. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Dilara Torunoğlu-Selamet, Dogukan Arslan, Rodrigo Wilkens, Wei He, Doruk Eryiğit, Thomas Pickard, Adriana S. Pagano, Aline Villavicencio, Gülşen Eryiğit, Ágnes Abuczki, Aida Cardoso, Alesia Lazarenka, Dina Almassova, Amalia Mendes, Anna Kanellopoulou, Antoni Brosa-Rodríguez, Baiba Saulite, Beata Wojtowicz, Bolette Pedersen, Carlos Manuel Hidalgo-Ternero, Chaya Liebeskind, Danka Jokić, Diego Alves, Eleni Triantafyllidi, Erik Velldal, Fred Philippy, Giedre Valunaite Oleskeviciene, Ieva Rizgeliene, Inguna Skadina, Irina Lobzhanidze, Isabell Stinessen Haugen, Jauza Akbar Krito, Jelena M. Marković, Johanna Monti, Josue Alejandro Sauca, Kaja Dobrovoljc, Kingsley O. Ugwuanyi, Laura Rituma, Lilja Øvrelid, Maha Tufail Agro, Manzura Abjalova, Maria Chatzigrigoriou, María del Mar Sánchez Ramos, Marija Pendevska, Masoumeh Seyyedrezaei, Mehrnoush Shamsfard, Momina Ahsan, Muhammad Ahsan Riaz Khan, Nathalie Carmen Hau Norman, Nilay Erdem Ayyıldız, Nina Hosseini-Kivanani, Noémi Ligeti-Nagy, Numaan Naeem, Olha Kanishcheva, Olha Yatsyshyna, Daniil Orel, Petra Giommarelli, Petya Osenova, Radovan Garabik, Regina E. Semou, Rozane Rebechi, Salsabila Zahirah Pranida, Samia Touileb, Sanni Nimb, Sarfraz Ahmad, Sarvinoz Nematkhonova, Shahar Golan, Shaoxiong Ji, Sopuruchi Christian Aboh, Srdjan Sucur, Stella Markantonatou, Sussi Olsen, Vahide Tajalli, Veronika Lipp, Voula Giouli, Yelda Yeşildal Eraydın, Zahra Saaberi, and Zhuohan Xie. 2026. A parallel cross-lingual benchmark for multimodal idiomaticity understanding. *Preprint*, arXiv:2601.08645.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Özge Umut and Bora Şenceylan. 2026. ITUNLP2 at MWE-2026 AdMIRe 2: Modular zero-shot pipelines for multimodal idiom grounding and ranking. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. Association for Computational Linguistics.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech & Language*, 19(4):365–377. Special issue on Multiword Expression.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

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.

Mehmet Utku Çolak. 2026. Idiomranker-x at AdMIRe 2: Multilingual idiom-image alignment via low-rank adaptation of cross-encoders. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*. Association for Computational Linguistics.

Appendix

A Image & Text Results

Team	Rank	Top-1 Acc.	nDCG
ITUNLP2	1	0.53	0.81
ITUNLP	2	0.50	0.80
DCSN-NLP	3	0.45	0.76
tiberiucarp	4	0.44	0.77
PolyFrame	5	0.35	0.72
VisAffect	6	0.30	0.71
IdiomRanker-X	7	0.28	0.70

Table 3: Chinese (ZH) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP2	1	0.56	0.82
ITUNLP	2	0.53	0.81
tiberiucarp	3	0.50	0.79
DCSN-NLP	4	0.47	0.75
VisAffect	5	0.34	0.72
PolyFrame	6	0.27	0.69
IdiomRanker-X	7	0.27	0.70

Table 4: Georgian (KA) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.64	0.87
DCSN-NLP	2	0.57	0.84
tiberiucarp	3	0.54	0.83
PolyFrame	4	0.36	0.72
IdiomRanker-X	5	0.34	0.72
VisAffect	6	0.27	0.69

Table 5: Greek (EL) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP2	1	0.56	0.84
ITUNLP	2	0.43	0.78
3K2T	3	0.41	0.76
DCSN-NLP	4	0.39	0.74
tiberiucarp	5	0.37	0.74
PolyFrame	6	0.33	0.73
VisAffect	7	0.30	0.73
IdiomRanker-X	8	0.22	0.69

Table 6: Igbo (IG) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP2	1	0.70	0.89
ITUNLP	2	0.61	0.84
3K2T	3	0.56	0.84
DCSN-NLP	4	0.53	0.80
tiberiucarp	5	0.51	0.81
VisAffect	6	0.40	0.74
PolyFrame	7	0.33	0.74
IdiomRanker-X	8	0.28	0.71

Table 7: Kazakh (KK) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP2	1	0.82	0.94
ITUNLP	2	0.67	0.88
DCSN-NLP	3	0.52	0.80
tiberiucarp	4	0.51	0.80
PolyFrame	5	0.42	0.75
IdiomRanker-X	6	0.38	0.74
VisAffect	7	0.29	0.70

Table 8: Norwegian (NO) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP2	1	0.88	0.96
ITUNLP	2	0.86	0.94
DCSN-NLP	3	0.80	0.91
tiberiucarp	4	0.67	0.88
PolyFrame	5	0.46	0.77
IdiomRanker-X	6	0.34	0.73
VisAffect	7	0.30	0.71

Table 9: Braz. Portuguese (PT-BR) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP2	1	0.72	0.91
ITUNLP	2	0.64	0.86
DCSN-NLP	3	0.57	0.81
tiberiucarp	4	0.55	0.83
PolyFrame	5	0.43	0.76
VisAffect	6	0.30	0.71
IdiomRanker-X	7	0.30	0.72

Table 10: Euro. Portuguese (PT-PT) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP2	1	0.71	0.91
ITUNLP	2	0.69	0.89
DCSN-NLP	3	0.68	0.87
tiberiucarp	4	0.63	0.85
PolyFrame	5	0.40	0.73
VisAffect	6	0.36	0.73
IdiomRanker-X	7	0.35	0.74

Table 11: RU – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.62	0.84
tiberiucarp	2	0.48	0.78
DCSN-NLP	3	0.45	0.76
PolyFrame	4	0.39	0.74
VisAffect	5	0.36	0.72
IdiomRanker-X	6	0.31	0.71

Table 12: Serbian (SR) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.60	0.85
tiberiucarp	2	0.51	0.82
DCSN-NLP	3	0.48	0.78
PolyFrame	4	0.38	0.73
VisAffect	5	0.34	0.72
IdiomRanker-X	6	0.31	0.72

Table 13: Slovak (SK) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP2	1	0.82	0.94
ITUNLP	2	0.78	0.91
DCSN-NLP	3	0.67	0.87
tiberiucarp	4	0.59	0.84
PolyFrame	5	0.41	0.75
IdiomRanker-X	6	0.36	0.75
VisAffect	7	0.28	0.70

Table 14: Slovenian (SL) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
DCSN-NLP	1	0.42	0.81
ITUNLP2	2	0.40	0.79
tiberiucarp	3	0.33	0.74
VisAffect	4	0.33	0.69
ITUNLP	5	0.27	0.73
IdiomRanker-X	6	0.23	0.69
PolyFrame	7	0.17	0.66

Table 15: Ecuadorian Spanish (ES-EC) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.68	0.90
ITUNLP2	2	0.65	0.88
DCSN-NLP	3	0.62	0.84
3K2T	4	0.54	0.83
tiberiucarp	5	0.48	0.80
PolyFrame	6	0.34	0.71
VisAffect	7	0.31	0.72
IdiomRanker-X	8	0.29	0.70

Table 16: Turkish (TR) – Image and Text

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.52	0.83
ITUNLP2	2	0.52	0.83
tiberiucarp	3	0.42	0.77
VisAffect	4	0.42	0.75
3K2T	5	0.41	0.78
DCSN-NLP	6	0.33	0.74
PolyFrame	7	0.32	0.72
IdiomRanker-X	8	0.31	0.71

Table 17: Uzbek (UZ) – Image and Text

B Text-only Results

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.46	0.77
alexandru412	2	0.41	0.76
LST	3	0.36	0.74
PolyFrame	4	0.28	0.69

Table 18: Chinese (ZH) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.51	0.79
alexandru412	2	0.46	0.77
LST	3	0.40	0.74
PolyFrame	4	0.28	0.68

Table 19: Georgian (KA) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.59	0.86
LST	2	0.43	0.76
PolyFrame	3	0.31	0.71

Table 20: Greek (EL) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.48	0.78
LST	2	0.33	0.71
PolyFrame	3	0.29	0.69
alexandru412	4	0.14	0.63

Table 21: Igbo (IG) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.60	0.84
LST	2	0.42	0.76
alexandru412	3	0.33	0.71
PolyFrame	4	0.28	0.72

Table 22: Kazakh (KK) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.61	0.85
alexandru412	2	0.52	0.80
LST	3	0.43	0.77
PolyFrame	4	0.41	0.75

Table 23: Norwegian (NO) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.79	0.92
LST	2	0.53	0.81
PolyFrame	3	0.37	0.75

Table 24: Brazilian Portuguese (PT-BR) – Text Only

Team	Rank	Top-1 Acc.	nDCG
alexandru412	1	0.64	0.85
ITUNLP	2	0.62	0.86
LST	3	0.45	0.77
PolyFrame	4	0.37	0.73

Table 25: European Portuguese (PT-PT) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.65	0.87
LST	2	0.51	0.79
alexandru412	3	0.47	0.77
PolyFrame	4	0.39	0.74

Table 26: RU – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.55	0.82
alexandru412	2	0.48	0.77
LST	3	0.40	0.74
PolyFrame	4	0.31	0.71

Table 27: Serbian (SR) – Text Only

Team	Rank	Top-1 Acc.	nDCG
alexandru412	1	0.56	0.80
ITUNLP	2	0.54	0.84
LST	3	0.44	0.78
PolyFrame	4	0.35	0.73

Table 28: Slovak (SK) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.72	0.89
LST	2	0.45	0.78
PolyFrame	3	0.37	0.74

Table 29: Slovenian (SL) – Text Only

Team	Rank	Top-1 Acc.	nDCG
LST	1	0.35	0.73
ITUNLP	2	0.25	0.72
PolyFrame	3	0.19	0.67
alexandru412	4	0.10	0.61

Table 30: Ecuadorian Spanish (ES-EC) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.51	0.82
LST	2	0.40	0.74
alexandru412	3	0.40	0.75
PolyFrame	4	0.32	0.70

Table 31: Turkish (TR) – Text Only

Team	Rank	Top-1 Acc.	nDCG
ITUNLP	1	0.50	0.82
alexandru412	2	0.34	0.71
LST	3	0.32	0.73
PolyFrame	4	0.32	0.71

Table 32: Uzbek (UZ) – Text Only

Author Index