Large Language Models as Multiword Expressions Annotators

Vasile Păiș, Maria Mitrofan, Verginica Barbu Mititelu, Dan Tufiș
Romanian Academy Research Institute for Artificial Intelligence, Bucharest, Romania
{vasile,maria,vergi,tufis}@racai.ro

Introduction

Multiword expressions (MWEs) are linguistic phenomena that are characterized by semantic non-compositionality (the defining feature of MWEs, according to many specialists), discontinuity (other words can occur between their components), ambiguity (the same word combination may be a MWE in some contexts, i.e., those in which it lacks semantic compositionality, while a perfectly compositional string in others), and variability (the components may vary lexically, morphologically, syntactically) (Constant et al., 2017), to mention only a few of their properties that makes them a pain in the neck for text processing (Sag et al., 2002). The notion of MWE encompasses a wide range of phenomena, including idioms (e.g. "kick the bucket"), collocations (e.g. "strong tea"), light verb constructions (e.g. "make a decision"), etc.

Although the number of MWEs is of the same order of magnitude as the number of single words in a speaker's lexicon (Jackendoff, 1997:156)), their frequency in corpora is quite reduced: e.g., the relative frequency of verbal MWEs in the PARSEME corpus v1.3 (Savary et al., 2023) is 1.3%. This further complicates the process of their identification, alongside their MWEs' linguistic idiosyncrasies. As NLP is shifting toward meaning-aware applications, detecting MWEs becomes essential to understand and generate fluent, correct text containing also such expressions, especially that are seen rarely in texts.

Important efforts have been invested in the field of MWE identification in corpora and progress has been made in this respect in the last decade. The PARSEME shared tasks on automatic MWEs identification (Savary et al., 2017; Ramisch et al., 2018; Ramisch et al., 2020) represent pivotal work in standardizing the task as a sequence labeling problem. They provided multilingual corpora, for over 20 languages, annotated with a unified typology (Savary et al., 2023). Their work has opened the door for further advances in the creation of MWEs resources (see Barbu Mititelu et al., 2025 for a presentation of the evolution of the development of such MWE-aware resources in the last decade).

MWEs identification has evolved from rule-based and statistical methods to neural architectures. Large Language Models (LLMs) represent the latest paradigm shift in MWEs identification, as well as in many other NLP and downstream tasks. Recently, the results of SemEval-2025 Task 1: AdMIRe (https://semeval2025-task1.github.io) on the extended evaluation set placed the average human annotator in the 5th position compared to the systems submitted to the shared task. The best-performing individual annotator outperformed all systems, while a pool of expert annotators matched the top-performing model.

This paper investigates the capabilities of LLMs for the task of MWEs identification. We explore Mistral and Llama models with different number of parameters considering two datasets: one is likely to have been seen by the models during training and a second dataset released recently, thus not seen by the models during their training.

LLMs are deep neural networks based on the Transformer architecture (Vaswani et al., 2017), pre-trained on a vast amount of text. Their core component, the self-attention mechanism, allows them to weigh the importance of all tokens in a context when processing a given token.

In the context of MWEs identification, to some extent, LLMs can:

- use the attention mechanisms to capture long-range dependencies and context, helping to distinguish between idiomatic and literal uses (Matarazzo and Torlone, 2025);
- learn to generalize across MWEs variants, recognizing both fixed and flexible forms (e.g., "give someone the cold shoulder" and "was given the cold shoulder") (De Leon et al., 2025);
- infer idiomatic meanings from context even with minimal training examples (Kim et al., 2025);
- perform semantic disambiguation, which is important for identifying whether a MWE is used literally or figuratively (Oh et al., 2025).

Generative LLMs, such as ChatGPT, Llama, or Gemini, have become a standard building block for large-scale NLP applications.

The Datasets

For our experiments we used two datasets: Wiki50 and CoAM. The Wiki50 (Vincze et al., 2011) corpus was created by annotating 50 randomly selected articles from the English Wikipedia, with each article chosen to contain at least 1,000 words excluding structured content like lists or tables. The resulting corpus contains 3,861 MWEs corresponding to 3,180 unique expression types. The authors report an inter-annotator agreement Kappa score of 0.6938. The CoAM corpus (Ide et al., 2024) is a multi-genre collection of texts curated for MWE analysis. Its composition is intentionally heterogeneous, drawing from professional news articles, public commentaries, and transcribed spoken data from TED talks and IWSLT workshops (Neubig et al., 2014; Cettolo et al., 2017). It contains 874 MWEs distributed across 1301 sentences. Wiki50 is an older dataset that was likely seen during LLM training (including the annotations), while CoAM is a recent dataset.

The Methodology

For this study, we used various LLMs to identify MWEs. In order to assess the importance of the system prompt, two different types of prompts were used: a very basic and a more detailed one. In the former case, the model is prompted for the task without additional instructions (assuming the model has encountered the term MWE during the pre-training phase). In the latter case, additional information regarding the MWEs identification task are provided.

The Results

The results of the experiments are presented in table 1. For each model we present both the results for the basic and for the detailed prompting.

Model	**Template**	**Dataset**	**P**	**R**	**F1**
Llama3.3-70b	Basic	CoAM-test	0.1169	0.4697	0.1872
Llama3.3-70b	Detailed	CoAM-test	0.1480	0.3984	0.2159
Llama3.3-70b	Basic	Wiki50	0.1787	0.6255	0.2780
Llama3.3-70b	Detailed	Wiki50	0.2191	0.4789	0.3006
Mistral-7b	Basic	CoAM-test	0.0233	0.1108	0.0385
Mistral-7b	Detailed	CoAM-test	0.0299	0.1214	0.0479
Mistral-7b	Basic	Wiki50	0.0629	0.2466	0.1003
Mistral-7b	Detailed	Wiki50	0.0628	0.2182	0.0975
Mistral-Nemo-12b	Basic	CoAM-test	0.0560	0.1425	0.0804
Mistral-Nemo-12b	Detailed	CoAM-test	0.0400	0.0501	0.0444
Mistral-Nemo-12b	Basic	Wiki50	0.0940	0.1904	0.1258
Mistral-Nemo-12b	Detailed	Wiki50	0.0800	0.0898	0.0847
Mistral-Large-123b	Basic	CoAM-test	0.1345	0.3720	0.1976
Mistral-Large-123b	Detailed	CoAM-test	0.1713	0.3193	0.2230
Mistral-Large-123b	Basic	Wiki50	0.1249	0.2820	0.1732
Mistral-Large-123b	Detailed	Wiki50	0.1212	0.1534	0.1354

Table 1: Results from prompting different LLMs for MWE identification.

In order to improve the results presented in Table 1, we also used the chain-of-thought prompting method with a single system prompt and multiple user prompts with associated responses. This mechanism resembles a dialogue with the LLM, where the user begins with a general problem and, through a sequence of queries, gradually guides the model toward a specific area of interest. The final response is obtained after the model has been presented with the entire conversation history (the system prompt, user prompts 1..n, and responses 1..n-1). For the purposes of this work, the responses obtained were refined with a new series of prompts, one for each of the identified MWEs.

The experiments considered Llama 3.3 (70 billion parameter model) and the Mistral family of models (7b,12b, and 123b parameters). The family of Mistral models was chosen to illustrate the impact of the chain of thought prompting on models with different numbers of parameters from the same family (the same architecture).

As shown in both Table 1 and Table 2, the detailed prompt has a benefic influence on models with larger number of parameters, particularly Llama 3.3 (70b parameters) and Mistral-Large. Llama 3.3 has similar performance to

Model	Template	Dataset	P	R	F1
Llama3.3-70b	Basic	CoAM-test	0.1840	0.3826	0.2485
Llama3.3-70b	Detailed	CoAM-test	0.2981	0.2823	0.2900
Llama3.3-70b	Basic	Wiki50	0.2223	0.4049	0.2870
Llama3.3-70b	Detailed	Wiki50	0.2759	0.1911	0.2258
Mistral-7b	Basic	CoAM-test	0.0258	0.0923	0.0404
Mistral-7b	Detailed	CoAM-test	0.0346	0.0765	0.0477
Mistral-7b	Basic	Wiki50	0.0702	0.1943	0.1031
Mistral-7b	Detailed	Wiki50	0.0768	0.1372	0.0985
Mistral-Nemo-12b	Basic	CoAM-test	0.0758	0.0712	0.0735
Mistral-Nemo-12b	Detailed	CoAM-test	0.0704	0.0264	0.0384
Mistral-Nemo-12b	Basic	Wiki50	0.1297	0.1008	0.1134
Mistral-Nemo-12b	Detailed	Wiki50	0.1123	0.0404	0.0594
Mistral-Large-123b	Basic	CoAM-test	0.1848	0.3588	0.2439
Mistral-Large-123b	Detailed	CoAM-test	0.2796	0.2929	0.2861
Mistral-Large-123b	Basic	Wiki50	0.1621	0.2466	0.1956
Mistral-Large-123b	Detailed	Wiki50	0.1711	0.1172	0.1391

Table 2: Results for MWE identification using chain of thought prompting technique.

Mistral Large in spite of the reduced number of parameters. This can be seen for both the basic prompt and the detailed prompt. Moreover, the chain-of-thought approach achieved better results.

Conclusion and future work

In conclusion, the refined prompting methods produce better results, but they cannot be compared with the state-of-the-art (SOTA) results in the field yet. Our approach may be used as a starting point to identify the best LLMs in a zero-shot approach. However, to compete with SOTA results, it is necessary to further fine-tune the language model or employ additional techniques. Moreover, MWEs resources have been developed for languages other than English too (Lornegaard, 2016; Barbu Mititelu et al, 2025). In this paper, we focused on English and corresponding LLMs, while in the future we will consider other languages as well. While at this moment the work focused on out-of-the-box understanding of MWEs by LLMs, additional techniques such as retrieval augmented generation (RAG) are of interest and will be explored.

References

[1] Applebaum, A. (2013). *Does Eastern Europe Still Exist?*. Prospect Magazine, 20.

[2] Sag, I. A.,et al. (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 (pp. 1–15). Springer.

[3] Jackendoff, Ray (1997). *The Architecture of the Language Faculty*, Cambridge, MA: MIT Press.

[4] Savary, A., Ben Khelil, C., Ramisch, C., Giouli, V., Barbu Mititelu, V., Hadj Mohamed, N., Krstev, C., Liebeskind, C., Xu, H., Stymne, S., Güngör, T., Pickard, T., Guillaume, B., Bejček, E., Bhatia, A., Candito, M., Gantar, P., Iñurrieta, U., Gatt, A., Kovalevskaite, J., Lichte, T., Ljubešić, N., Monti, J., Parra Escartín, C., Shamsfard, M., Stoyanova, I., Vincze, V., and Walsh, A. (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.

[5] Barbu Mititelu, V., Giouli, V., Korvel, G., Liebeskind, C., Lobzhanidze, I., Makhachashvili, R., Markantonatou, S., Markovic, A., and Stoyanova, I. (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.

[6] Matarazzo, A., & Torlone, R. (2025). *A Survey on Large Language Models with some Insights on their Capabilities and Limitations*. arXiv preprint arXiv:2501.0

[7] De Leon, F. L., Madabushi, H. T., & Lee, M. G. (2025). *Evaluating Large Language Models on Multiword Expressions in Multilingual and Code-Switched Contexts*. arXiv preprint arXiv:2504.20051.

[8] Kim, J., Shin, Y., Hwang, U., Choi, J., Xuan, R., & Kim, T. (2025). *Memorization or Reasoning? Exploring the Idiom Understanding of LLMs*. arXiv preprint arXiv:2505.16216.

[9] Oh, S., Huang, X., Pink, M., Hahn, M., & Demberg, V. (2025). *Tug-of-war between idiom's figurative and literal meanings in LLMs*. arXiv preprint arXiv:2506.01723.

[10] Constant, M., Eryigit, G., Monti, J., Van Der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). *Multiword expression processing: A survey*. Computational Linguistics, 43(4), 837–892.

[11] Savary, A., Ramisch, C., Hwang, J., Schneider, N., et al. (2017). *The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions*. In Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017) (pp. 31–47). EACL.

[12] Ramisch, C., Cordeiro, S., Savary, A., Vincze, V., Barbu Mititelu, V. et al. (2018). Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, Santa Fe, United States, pp.222 - 240.

[13] Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Barbu Mititelu, V., Bhatia, A., Iñurrieta, U., Giouli, V., et al. (2020). *Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions*. In The Joint Workshop on Multiword Expressions and Electronic Lexicons, 13 December 2020, Online (pp. 107–118).

[14] Barbu Mititelu, V., & Mitrofan, M. (2019). *Leaving no stone unturned when identifying and classifying verbal multiword expressions in the Romanian WordNet*. In Proceedings of the 10th Global Wordnet Conference (pp. 10–15).

[15] Mi, W., Villavicencio, A., & Moosavi, N. S. (2024). *Rolling the DICE on Idiomaticity: How LLMs Fail to Grasp Context*. arXiv preprint arXiv:2410.16069.

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is All You Need*. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems, 30. Curran Associates, Inc.

[17] Vincze, V., Nagy, I., & Berend, G. (2011). *Multiword expressions and named entities in the Wiki50 corpus*. In Proceedings of the International Conference Recent Advances in Natural Language Processing 2011 (pp. 289–295).

[18] Ide, Y., Tanner, J., Nohejl, A., Hoffman, J., Vasselli, J., Kamigaito, H., & Watanabe, T. (2024). *CoAM: Corpus of All-Type Multiword Expressions*. arXiv preprint arXiv:2412.18151.

[19] Barbu Mititelu, V., Cristescu, M., Mitrofan, M., Zgreabăn, B.-M., & Bărbulescu, E.-A. (2022). *A Romanian Treebank Annotated with Verbal Multiword Expressions*. In Proceedings of the 5th International Conference on Computational Linguistics in Bulgaria (CLIB 2022) (pp. 137–145).

[20] Leseva, S., Barbu Mititelu, V., & Stoyanova, I. (2020). *It Takes Two to Tango—Towards a Multilingual MWE Resource*. In Proceedings of the 4th International Conference on Computational Linguistics in Bulgaria (CLIB 2020) (pp. 101–111).

[21] Smørdal Losnegaard, G., Sangati, F., Parra Escartín, C., Savary, A., Bargmann, S. and Monti, J. (2016). PARSEME Survey on MWE Resources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2299–2306, Portorož, Slovenia. European Language Resources Association (ELRA).