Optimizing Multilingual MWE Identification: From Morphologically-Filtered LLMs to Pure Transformer Architectures

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

Relevant UniDive working groups: WG1, WG3

1 Introduction

This project was submitted to the PARSEME 2.0 Shared Task and it was presented at the 22nd Workshop of Multiword Expessions (2026), colocated with the EACL.

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 in different 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.

Our current research evaluates the performance of Large Language Models (LLMs) on this task and proposes a transition toward more specialized, transformer-based architectures to improve the detection of complex linguistic patterns.

*Corresponding authors.

2 Current Methodology

Our primary contribution to date is the development of a hybrid system designed to identify MWEs across 17 languages. The system combines LLM-based predictions generated via the Gemini API with a morphological post-filter (Rambow et al., 2003) designed to reduce false positives. Rather than optimizing peak performance on individual languages, our approach prioritizes cross-lingual stability and precision. The architecture (see appendix A) consists of two main stages:

- **LLM Prediction**: We utilize **Gemini 2.0 Flash-Lite** (DeepMind, 2025) via a few-shot prompting strategy (Brown et al., 2020). The model is tasked with identifying potential MWEs and labeling them using the BIO (Beginning, Inside, Outside) tagging scheme.

- **Morphological Filtering**: A known issue with LLMs in this domain is "overgeneration", meaning the tendency to label random word sequences as idioms. To diminish this, we implemented a post-processing filter based on **Universal Parts of Speech (UPOS)** tags. This filter automatically rejects candidates that do not follow valid grammatical patterns for the specific language.

3 Results

Overall Performance In the official evaluation, our system was submitted for all 17 languages. It 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.

#Langs	Global MWE-based				Global Token-based			
	P	R	F1	Rank	P	R	F1	Rank
17/17	20.95	14.50	17.14	7	34.14	24.20	28.32	5

Table 1: General Ranking

#Langs	Richness		Shannon-Weaver Entropy		Shannon-Evenness	
	Value	Rank	Value	Rank	Value	Rank
17/17	56.53	7	3.71	5	0.97	1

Table 2: Diversity Ranking

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 (see appendix B).

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 other systems exhibiting strong performance peaks for a small subset of languages, ours avoids extreme failures and maintains a stable cross-lingual behavior.

3.1 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 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.

4 Future Work: Transitioning to Transformers

The BIO (Beginning, Inside, Outside) tagging scheme is insufficient for discontinious and overlapping of MWEs. We propose moving to a **BIOES+** (Beginning, Inside, Outside, End, Single, plus category) scheme (Reimers and Gurevych, 2017), since it provides explicit markers for the end of an expression and for single-token MWEs. The "+" denotes the inclusion of specific PARSEME categories (e.g., B-LVC, E-VID).

In future work, we plan to transition from the current generative API-based pipeline to a dedicated, local Transformer architecture like XLM-RoBERTa (Conneau et al., 2020) to address the structural complexities of multi-word expressions.

To solve the problem of "illegal" tag transitions, we are thinking of adding a Conditional Random Field (CRF) layer (Lafferty et al., 2001), which learns a transition matrix to ensure mathematically valid sequences.

Furthermore, we intend to move beyond post-filtering by using POS injection, where morphological embeddings are fed directly into the Transformer's self-attention mechanism (Vaswani et al.,

2017). This architectural shift is specifically designed to leverage the self-attention property of Transformers, allowing the model to bridge the gap in discontinuous MWEs by linking distant components (e.g., verb-particle splits) across the entire sentence.

5 Conclusion

The results suggest that combining general-purpose LLM predictions with minimal linguistic post-processing yields balanced evaluation outcomes, but with limitations. The next step in multilingual MWE identification requires syntax-aware, fine-tuned models. The proposed shift to a Transformer-CRF framework aims to combine the cross-lingual fairness of our current system with the structural rigor needed to solve complex, non-local linguistic dependencies.

References

Timothy Baldwin and Su Nam Kim. 2010. *Multiword Expressions*. Taylor and Francis.

Tom Brown et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*.

Alexis Conneau et al. 2020. Unsupervised cross-lingual representation learning at scale. *Proceedings of ACL*. Foundational paper for XLM-R.

Google DeepMind. 2025. Gemini 2.0 flash-lite: Cost-efficient multimodal reasoning. Google Developers Blog. Released February 5, 2025.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*.

Owen Rambow et al. 2003. Rule-based and bootstrapping approaches for named entity recognition. *HLT-NAACL*.

Nils Reimers and Iryna Gurevych. 2017. Optimal hyperparameters for deep LSTM-CRF sequence labeling systems. In *Proceedings of EMNLP*.

Agata Savary and Carlos Ramisch. 2025. PARSEME 2.0 shared task guidelines.

Ashish Vaswani et al. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

A System Pipeline

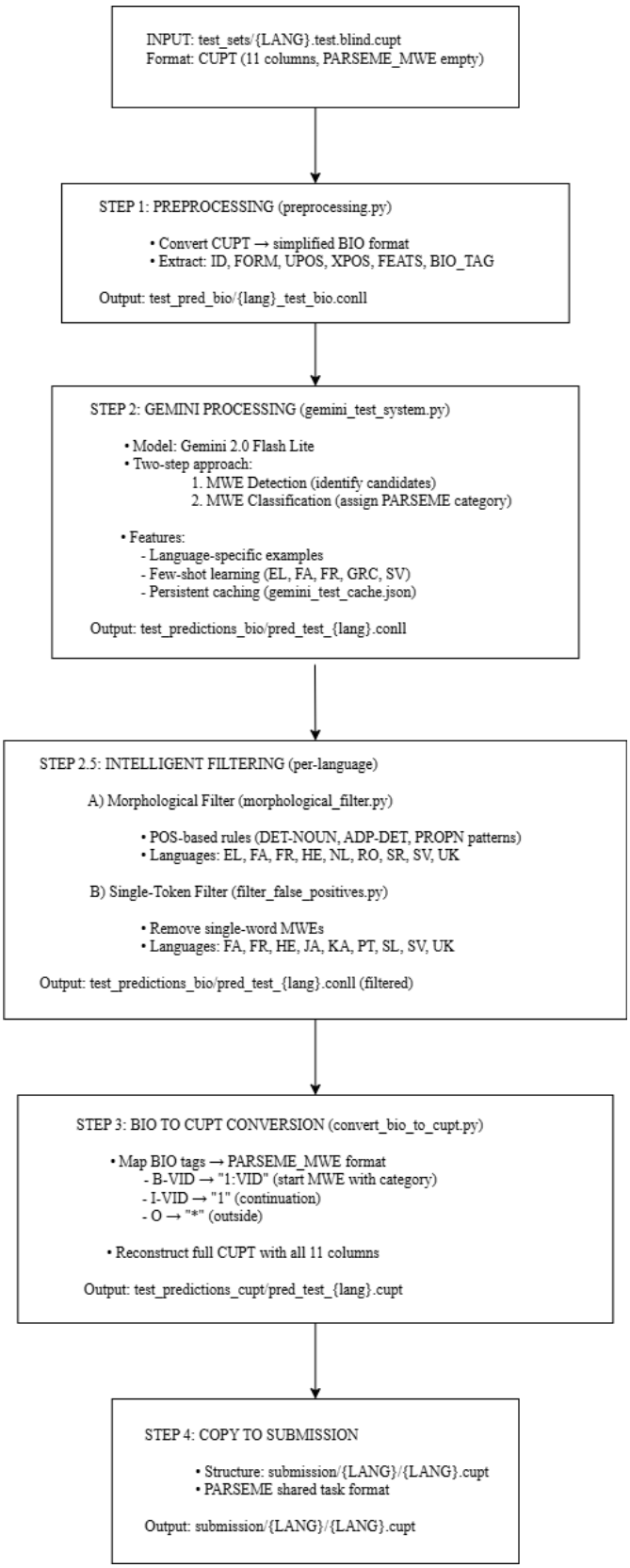

Figure 1: Detailed MWE Prediction Pipeline

B Performance Gap

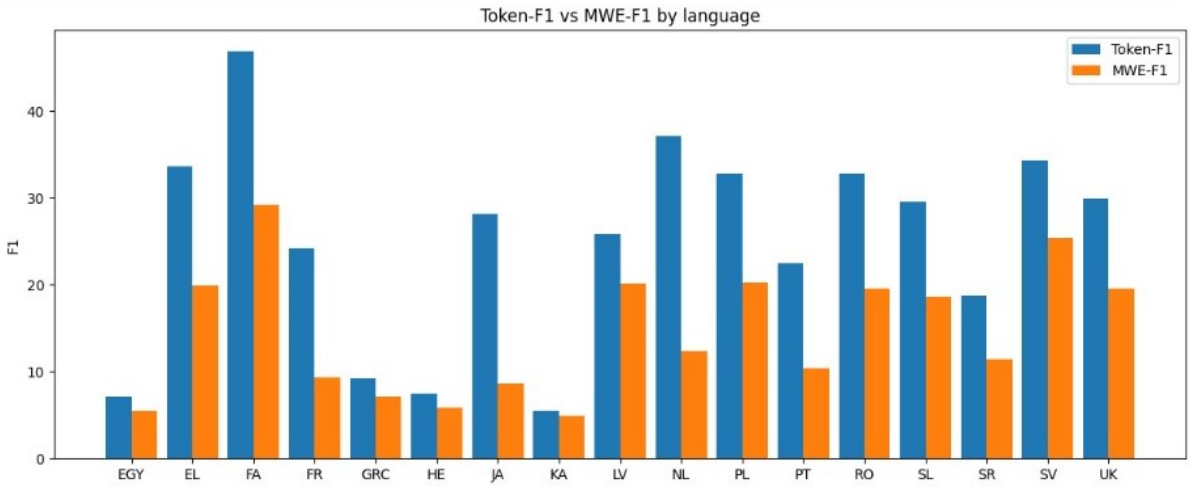

Figure 2: A pronounced performance gap reveals that while the model successfully identifies constituent tokens, it fails to accurately delineate expression boundaries, underscoring a persistent challenge in exact boundary reconstruction for multiword expressions. A main reason for this behavior is the usage of BIO representation: 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.