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.

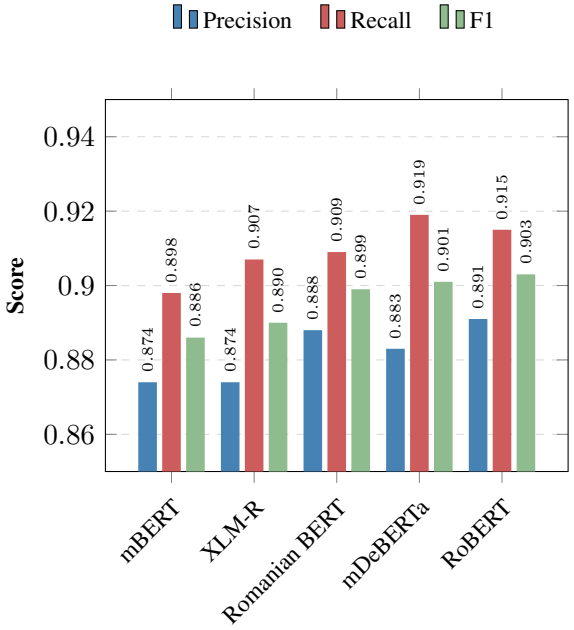

Figure 1: Comparative results of Transformer encoders sorted by F1 score

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.

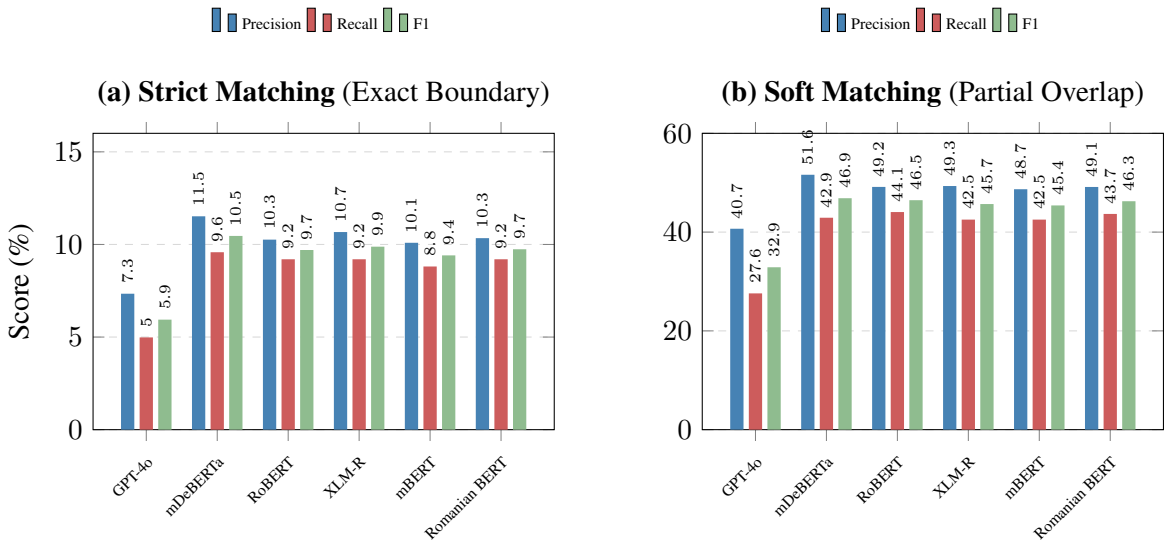

Figure 2: Global performance on our internal unseen subset (N=85). **(a)** Shows metrics under the Strict evaluation scenario (exact match), where all models struggle significantly. **(b)** Shows metrics under the Soft evaluation scenario, where fine-tuned encoders (like mDeBERTa and RoBERT) clearly outperform the few-shot GPT-4o baseline, particularly in Recall.