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
March 28, 2026 ©2026 Association for Computational Linguistics

(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*.