

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”.

¹The implementation and dataset of MISP are available at <https://github.com/TinFoil/unibo-misp/>.

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 Task² (Scholivet et al., 2026), organised within the “Universality, diversity and idiosyncrasy in language technology” (UniDive) CA21167 COST Action³, 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,

²https://unidive.lisn.upsaclay.fr/doku.php?id=other-events%3Aparseme-admire-st-call#call_for_participation

³<https://unidive.lisn.upsaclay.fr>

(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 *AStitchInLanguage-Models* dataset (Tayyar Madabushi et al., 2021). The results, presented in Section 4, suggest that MISIP 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-120b⁶ (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 MISIP 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 MISF 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 MISF 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.

5 Conclusions

In this paper, we described MISF, 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 analy-

Elimination – MWE not deleted
O Une déviation en <i>voie rapide</i> a été inaugurée en 1993.
P Une modification sur la <i>voie rapide</i> a été mise en place en 1993
Meaning – original meaning not preserved
O Effets négatifs de la <i>rétenion d’information</i>
P Les effets néfastes de la <i>conservation d’information</i>
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 <i>bed and breakfast</i> .
P Après quelques années de résidence à Ringerike, ils s’installent dans le quartier Nordstrand d’Oslo, où ils reprendent une <i>auberge de jeunesse</i>
Grammatical Closeness – change in mood, time, number
O Nelson Downs n’a jamais <i>pris de cours</i> 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 <i>suivi de cours</i> 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’ <i>équipe réserve</i> du Saturn Ramenskoïe en troisième division.
P De 2005 à 2007, il participe à l’ <i>équipe de réserve</i> 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.

sis suggests that the baseline from gpt-oss-120b achieves significantly higher BERTScores across all languages, whereas MISF 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, MISF’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, MISF, appear more able to generate paraphrases that exhibit higher linguistic diversity.

Error analysis in French shows that MISF 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 with out 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, Antonio 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, Morocco. 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 <i>fast track</i> was implemented in 1993
Meaning – the original meaning was not preserved	
O	Negative effects of <i>information retention</i> .
P	The adverse effects of <i>information conservation</i>
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 <i>bed and breakfast</i> .
P	After living in Ringerike for a few years, they moved to the Nordstrand district of Oslo, where they took over a <i>youth hostel</i> .
Grammatical Closeness – the paraphrase contains a change in mood, time or number	
O	Nelson Downs never <i>took</i> magic <i>lessons</i> ; he was self-taught and, at a very young age, developed remarkable skills in manipulating cards and coins.
P	Nelson Downs never <i>had</i> magic <i>lessons</i> ; 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 <i>reserve team</i> in the third division.
P	From 2005 to 2007, he took part in Saturn Ramenskoye’s <i>reserve team</i> , which competed in the third division.

Table 5: English translations of examples of **Original** and **Paraphrase** 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.