PARSEME corpus and shared task on automatic paraphrasing of multiword expressions

Manon Scholivet[1], **Agata Savary**[1], **Carlos Ramisch**[2], **Takuya Nakamura**[1], **Eric Bilinski**[1], **Maria Mitrofan**[3], **Vasile Păiș**[3], **Marina Bagi**[4], **Verginica Barbu Mititelu**[3], **Astrid Berntsson Ingelstam**[5], **Jaka Čibej**[6], **Marijana Đukić**[4], **Polona Gantar**[6], **Voula Giouli**[7], **Olha Kanishcheva**[8], **Chaya Liebeskind**[9], **Irina Lobzhanidze**[10], **Aleksandra Marković**[4], **Gunta Nešpore-Bērzkalne**[11], **Adriana Pagano**[12], **Mehrnoush Shamsfard**[13], **Sara Stymne**[5], **Vahide Tajalli**[13]

[1]LISN, Paris-Saclay University, CNRS, France; [2]Aix Marseille Univ, CNRS, LIS, Marseille, France; [3]RACAI, Bucharest, Romania; [4]Institute for the Serbian Language SASA, Belgrade, Serbia; [5]Uppsala University, Sweden; [6]University of Ljubljana, Slovenia; [7]Aristotle University of Thessaloniki and ILSP, ATHENA RC, Greece; [8]Heidelberg University, Germany and SET University, Ukraine; [9]Jerusalem College of Technology, Israel; [10]Ilia State University, Tbilisi, Georgia; [11]Institute of Mathematics and Computer Science, University of Latvia; [12]Federal University of Minas Gerais, Brazil; [13]Shahid Beheshti University, Tehran, Iran

Relevant UniDive working groups: WG1, WG3, WG4

1 Introduction

Paraphrasing is a semantically-oriented task with a rich bibliography. It has been tackled in different settings: (i) as scoring of alignments between noun phrases and their potential paraphrases (Butnariu et al., 2009), (ii) as paraphrase identification in entire sentences (Xu et al., 2015; Lan et al., 2017), (iii) as semantic text similarity task, i.e. assigning similarity scores to sentence pairs (Agirre et al., 2015; Xu et al., 2015), (iv) as paraphrase generation, i.e. reformulating a sentence with different wording or structure but preserving the original meaning (Zhou and Bhat, 2021).

In this context, multiword expressions (MWEs), such as *a* ***hot dog*** or ***to pull*** *one's* ***leg***, pose special challenges due to their semantic non-compositionality. Dedicated MWE-aware paraphrase datasets relied on MWE definitions from existing lexicons (Pershina et al., 2015; Liu and Hwa, 2016), dedicated lexicographic projects (Barančíková and Kettnerová, 2018), crowdsourcing (Yimam et al., 2016), or machine back-translation (Qiang et al., 2023). The paraphrase identification task was adapted to idioms (Tan and Jiang, 2021) and MWE-specific tasks were introduced (Zhou et al., 2021): *idiomatic sentence generation*, which transforms a literal sentence into a sentence involving idioms, and *idiomatic sentence paraphrasing*, which simplifies sentences so as to replace idioms with literal expressions. In the latter, the aim is to paraphrase only the MWE, leaving the rest of the sentence unchanged (Wada et al., 2023; Qiang et al., 2023). These MWE-related efforts have been dedicated to few languages and focused on paraphrasing the MWE themselves.

We summarize here a considerable extension of this state of the art by the PARSEME community: (i) a multilingual dataset of natural sentences containing idioms, with minimal and creative paraphrases constructed by native human experts, (ii) a shared task on automatic paraphrasing of MWEs (Scholivet et al., 2026). These efforts cover 14 languages: French (fr), Georgian (ka), Modern Greek (el), Hebrew (he), Japanese (ja), Latvian (lv), Persian (fa), Polish (pl), Portuguese (pt), Romanian (ro), Slovene (sl), Swedish (sv), Serbian (sr) and Ukrainian (uk).

Potential applications include text simplification (Zhang et al., 2024), eliminating idiomaticity prior to machine translation (Santing et al., 2022) and testing the ability of models to grasp idiomatic meaning (Tayyar Madabushi et al., 2022; He et al., 2025).

2 Dataset

To construct the paraphrasing dataset, we used the PARSEME 2.0 corpus annotated for MWEs in 17 languages (Savary et al., 2026). Based on the test set of this corpus, we extracted up to 150 sentences per language, containing exactly one MWE of one of the three types: verbal idiom (VID), like (fr) ***voir le jour*** (lit. 'see the day') 'to be born', nominal idiom (NID), e.g. (pt) ***tiro no pé*** (lit. 'shot in the foot') 'a

self-inflicted mistake' or adjectival idiom (AdjID), e.g. (ro) ***de bază*** (lit. 'of base') 'basic'. Given a sentence with a highlighted idiom, native human experts were to provide two paraphrases: a *minimal* and a *creative* one. The former was obtained by modifying as few tokens as possible among those that do not belong to the MWE. When creating the latter, conversely, significant changes were encouraged, both lexical (adding, deleting or replacing words) and grammatical (e.g. changing the word order or transforming active to passive voice), as long as the meaning of the original sentence was maintained. For example, sentence (1) received the minimal paraphrase (2) and the creative one (3):

(1) PDPA ხელისუფლებამ მოქმედებაში
PDPA the.government in.action
მოიყვანა სოციალისტური **დღის**
brought socialist of.the.day
წესრიგი. (ka)
order.
'The PDPA government put into action a socialist agenda.'

(2) PDPA ხელისუფლებამ მოქმედებაში
PDPA the.government in.action
მოიყვანა სოციალისტური გეგმა. (ka)
brought socialist plan.
'The PDPA government brought into action a socialist plan.'

(3) სოციალისტური წყობის რეალიზება PDPA
socialist set.up realization PDPA
ხელისუფლებამ მოქმედებაში მოახერხა.
the.government in.action succeeded.
(ka)
'The PDPA government managed to realize a socialist set-up.'

In each paraphrase, we asked annotators to remove at least one component of the MWE. New MWEs were allowed in the creative paraphrase, but not in the minimal one. The use of LLMs was prohibited for annotators, but online dictionaries and synonym lists were allowed. It was possible to provide more than two paraphrases, and then the two best ones had to be indicated (for system evaluation). Occasionally, it happened that a minimal or a creative paraphrase was not possible, then only one paraphrase was given. In rare problematic cases, the original sentence was totally discarded.

The resulting dataset contains from 66 (Swedish) to 150 (Georgian) original sentences per language. In total, there are 1,742 original sentences, with 726 VIDs, 863 NIDs and 153 AdjIDs, as well as 1,670 minimal and 1,618 creative paraphrases.

3 Paraphrasing task

Based on the paraphrasing dataset, PARSEME's shared task 2.0 (Scholivet et al., 2026) included a subtask dedicated to automatic paraphrasing of MWEs. Only trial data in English and French were provided in the system training phase, but no training or development data.

In this task, for every language, systems were given a raw sentence containing exactly one VID, NID or AdjID, not explicitly marked in text. On output, they were expected to paraphrase the sentence so that the original MWE no longer occurred, but the meaning of the original sentence was kept. Additionally, to facilitate automatic evaluation, at least one of the components (in any inflected form) of the original MWE should be totally absent from the paraphrase. We allowed paraphrases to use MWEs, provided that they were different from the original one.

The competition was implemented on the Codabench platform[1] and later cloned as an everlasting benchmark.[2]

Performance was evaluated with two measures. As automatic score we used *masked BERT-score*, which first checked if at least one of the MWE components had been removed. If not, the score was 0, otherwise the maximum BERT-score (Zhang et al., 2020) between the system-generated paraphrase and the two gold paraphrases was calculated. For the manual score, annotators assigned score 0 if the MWE was not removed, and, otherwise, three scores from 0 to 3 for keeping: (i) the sense of the removed MWE, (ii) the sense of the rest of the sentence, and (iii) grammaticality and naturalness. The final manual score was a weighted average of these 3 scores, with (i) doubled, normalized to $[0, 100]$. Both the masked BERT score and the manual score were averaged across all sentences in a language, then macro-averaged across languages.

We also measured the diversity of the predictions in terms of novel words, i.e., those present in the prediction but not in the original sentences. Three diversity measures were: variety (number of distinct novel words), balance (evenness of the distribution of novel words) and variety/balance hybrid (entropy of novel words).

The task received 5 submissions, including our own baseline. Only the baseline covered all 14 languages. The 4 other systems jointly covered

[1] https://www.codabench.org/competitions/12002/
[2] https://www.codabench.org/competitions/13192/

French, Georgian, Portuguese, and Romanian. All systems were based on generative LLMs, with additional mixtures of techniques including MWE pre-identification, category-oriented prompts, single-word synonyms for MWEs, cross-lingual transfer or a combination of specialised LLMs.

We noted a strong positive correlation between the automatic and the manual scores[3], as well as a high inter-evaluator agreement in manual evaluation[4]. We also observed the so-called performance-diversity trade-off typical for generation scenarios (Ippolito et al., 2019; Zhang et al., 2021): the higher the performance, i.e. the quality of the generated paraphrases, the lower the diversity, and vice-versa. These observations should be handled with care, given the relatively small number and coverage of the participating systems.

For detailed results of the paraphrasing subtask, see (Scholivet et al., 2026).

In the future, the paraphrasing dataset should be expanded to provide both train/dev and test subsets and to cover new languages, ideally including some of the very low-resourced ones. The shared task should then be organized with a relaxed timeline to encourage broad participation and extensive experimentation prior to the submission of results.

4 Acknowledgements

We are grateful to all the annotators having participated in the project.

This work received support from the CA21167 COST action UniDive, funded by the European Union via the COST (European Cooperation in Science and Technology). Further support came from: (1) French Agence Nationale pour la Recherche, via the SELEXINI project (ANR-21-CE23-0033-01), (2) Swedish national research infrastructure Språkbanken, jointly financially supported by the Swedish Research Council (2025–2028; grant 2023-00161) and the 10 participating partner institutions, (3) Brazilian National Council for Scientific and Technological Development (CNPq 404722/2024- 5; 313103/2021-6) and Minas Gerais State Agency for Research and Development (FAPEMIG), (4) Latvian Council of Science via the project "Advancing Latvian computational lexical resources for natural language understanding and generation"(LZP2022/1-0443), (5) Slovenian Research and Innovation Agency (research core funding No. P6-0411 *Language Resources and Technologies for Slovene* and No. P6-0215 *Slovene Language – Basic, Contrastive, and Applied Studies*), (6) the Ministry of Science, Technological Development and Innovation, Republic of Serbia (GRANT 451-03-33/2026-03/200174) and the Science Fund of the Republic of Serbia #7276, *Text Embeddings-Serbian Language Applications, TESLA*; (7) the "Large Language Models for the European Union (LLMs4EU)", project no. 101198470, call DIGITAL-2024-AI-B-06-LANGUAGE, funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them, (8) a grant of the Ministry of Research, Innovation and Digititalization - UEFISCDI, Romania, project number PN-IV-P8-8.2-EUD-2025-0061, within PNCDI IV, (9) NATO Science for Peace and Security Programme under grant id. G8648, project DeepNewsDef, (10) a grant of the Ministry of Education and Research, CCCDI - UEFISCDI, Romania, project number PN-IV-P8-8.2-NATO-SPS-2025-0005, within PNCDI IV.

[3]Pearson and Spearman correlation between the automatic and the manual scores (for the 4 languages for which system scores were submitted) amount to 0.92 and 0.90, respectively.

[4]Krippendorff's α Artstein and Poesio (2008) of 80.30 and 75.99 for Portuguese and Romanian, respectively.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Petra Barančíková and Václava Kettnerová. 2018. Paraphrases of verbal multiword expressions: The case of Czech light verbs and idioms. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 35–59. Language Science Press., Berlin.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2009. SemEval-2010 task 9: The interpretation of

noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 100–105, Boulder, Colorado. Association for Computational Linguistics.

Wei He, Tiago Kramer Vieira, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2025. Investigating idiomaticity in word representations. *Computational Linguistics*, 51:505–555.

Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762, Florence, Italy. Association for Computational Linguistics.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234, Copenhagen, Denmark. Association for Computational Linguistics.

Changsheng Liu and Rebecca Hwa. 2016. Phrasal substitution of idiomatic expressions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 363–373, San Diego, California. Association for Computational Linguistics.

Maria Pershina, Yifan He, and Ralph Grishman. 2015. Idiom paraphrases: Seventh heaven vs cloud nine. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 76–82, Lisbon, Portugal. Association for Computational Linguistics.

Jipeng Qiang, Yang Li, Chaowei Zhang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2023. Chinese idiom paraphrasing. *Transactions of the Association for Computational Linguistics*, 11:740–754.

Lukas Santing, Ryan Sijstermans, Giacomo Anerdi, Pedro Jeuris, Marijn ten Thij, and Riza Batista-Navarro. 2022. Food for thought: How can we exploit contextual embeddings in the translation of idiomatic expressions? In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 100–110, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Agata Savary, Manon Scholivet, Carlos Ramisch, Takuya Nakamura, Eric Bilinski, Sara Stymne, Voula Giouli, Stella Markantonatou, Vasile Păiș, Maria Mitrofan, Louis Estève, Bruno Guillaume, Verginica Barbu Mititelu, Jaka Čibej, Roberto A. Díaz Hernández, Victoria Fendel, Polona Gantar, Olha Kanishcheva, Cvetana Krstev, Chaya Liebeskind, Irina Lobzhanidze, Aleksandra Marković, Gunta Nešpore-Bērzkalne, Adriana Pagano, Mehrnoush Shamsfard, Ranka Stanković, Vahide Tajalli, and Carole Tiberius. 2026. PARSEME 2.0 multilingual corpus of multiword expressions. In *Proceedings of LREC 2026*, Palma di Mallorca, Spain.

Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Mitrofan, 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)*, pages 254–275, Rabat, Marocco. Association for Computational Linguistics.

Minghuan Tan and Jing Jiang. 2021. Does BERT understand idioms? a probing-based empirical study of BERT encodings of idioms. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1397–1407, Held Online. INCOMA Ltd.

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.

Takashi Wada, Yuji Matsumoto, Timothy Baldwin, and Jey Han Lau. 2023. Unsupervised paraphrasing of multiword expressions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4732–4746, Toronto, Canada. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. SemEval-2015 task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 1–11, Denver, Colorado. Association for Computational Linguistics.

Seid Muhie Yimam, Héctor Martínez Alonso, Martin Riedl, and Chris Biemann. 2016. Learning paraphrasing for multiword expressions. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 1–10, Berlin, Germany. Association for Computational Linguistics.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33, Online. Association for Computational Linguistics.

Jingshen Zhang, Xinglu Chen, Xinying Qiu, Zhimin Wang, and Wenhe Feng. 2024. Readability-guided idiom-aware sentence simplification (RISS) for Chinese. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1183–1200, Taiyuan, China. Chinese Information Processing Society of China.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021. 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.