PARSEME corpus 2.0 and shared task on automatic identification of multiword expressions

Agata Savary[1], Manon Scholivet[1], Carlos Ramisch[2], Takuya Nakamura[1], Eric Bilinski[1], Sara Stymne[3], Voula Giouli[4], Stella Markantonatou[5], Vasile Păiș[6], Maria Mitrofan[6], Louis Estève[1], Bruno Guillaume[7], Verginica Barbu Mititelu[6], Jaka Čibej[8], Roberto A. Díaz Hernández[9], Victoria Fendel[10], Polona Gantar[8], Olha Kanishcheva[11], Cvetana Krstev[12], Chaya Liebeskind[13], Irina Lobzhanidze[14], Aleksandra Marković[15], Gunta Nešpore-Bērzkalne[16], Adriana Pagano[17], Mehrnoush Shamsfard[18], Ranka Stanković[19], Vahide Tajalli[18], Carole Tiberius[20], Aakanksha Padhye[21]

[1]LISN, Paris-Saclay University, CNRS, France; [2]Aix Marseille Univ, CNRS, LIS, Marseille, France;
[3]Uppsala University, Sweden;
[4]Aristotle University of Thessaloniki and ILSP, ATHENA RC, Greece;
[5]ILSP and Archimedes Unit, ATHENA RC, Greece; [6]RACAI, Bucharest, Romania;
[7]Université de Lorraine, CNRS, Inria, LORIA, France; [8]University of Ljubljana, Slovenia;
[9]University of Jaen, Spain; [10]University of Oxford, United Kingdom;
[11]Heidelberg University, Germany; SET University, Ukraine;
[12]Jerteh - Language Resources and Technologies Society, Serbia;
[13]Jerusalem College of Technology, Israel; [14]Ilia State University, Tbilisi, Georgia;
[15]Institute for the Serbian Language SASA, Belgrade, Serbia;
[16]Institute of Mathematics and Computer Science, University of Latvia;
[17]Federal University of Minas Gerais, Brazil; [18]Shahid Beheshti University, Tehran, Iran;
[19]University of Belgrade, Serbia;
[20]Dutch Language Institute/Leiden University Centre for Linguistics, The Netherlands;
[21]Indian Institute of Technology Delhi, India

Relevant UniDive working groups: WG1, WG3, WG4

1 Introduction

Multiword expressions (MWEs), such as ***by and large***, or ***cut corners***, have been considered challenging in NLP for decades (Constant et al., 2017; Savary et al., 2019; Shwartz and Dagan, 2019; Miletić and Schulte im Walde, 2024) notably due to their semantic non-compositionality.

Despite several shared tasks dedicated to MWEs (Schneider et al., 2016; Tayyar Madabushi et al., 2022; Pickard et al., 2025), only the PARSEME and UniDive initiatives addressed understanding MWEs in highly multilingual settings (Savary et al., 2017; Ramisch et al., 2018, 2020; Scholivet et al., 2026; Arslan et al., 2026). We synthesize here part of the latest achievements of PARSEME: (i) the PARSEME 2.0 multilingual corpus annotated for MWEs (Savary et al., 2026),[1] (ii) the task of automatic identification of MWEs, included in the PARSEME 2.0 shared task (Scholivet et al., 2026). These efforts cover 17 languages: Dutch (nl), Egyptian (egy), French (fr), Georgian (ka), Ancient Greek (grc), 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).

2 Guidelines and corpus

Previous versions of the PARSEME corpora (1.0–1.3) only cover verbal MWEs. Version 2.0 is a considerable extension, covering all MWE categories, also including nominal, adjectival, adverbial, and functional MWEs.[2] The full tagset now includes 18 categories illustrated in Tab. 1.

Like in previous edition, guidelines 2.0 are based on the linguistic hypothesis that semantic non-compositionality is hard to test directly but correlates with lexical and morpho-syntactic inflexibility (Gross, 1986, 1988; Nunberg et al., 1994). For instance, in (fr) ***à la place de*** *Luc* (lit. 'at the place of Luc') 'instead of Luc', the expression in bold

[1]http://hdl.handle.net/11372/LRT-6123.

[2]https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/

VMWE	VID	verbal idiom	(ka) **შიშს ჭამს** (šišs čams | lit. 'he eats horror') 'he panics'
	LVC.full	light-verb construction; bleached verb	(nl) *een* ***toespraak houden*** (lit. 'to hold a speech') 'to make a speech'
	LVC.cause	light-verb construction; causal verb	(sl) ***narediti konec*** *nečemu* (lit. 'to make an end to sth') 'to end sth'
	IRV	inherently reflexive verb	(sr) **бојати се** (bojati se | lit. 'fear oneself') 'to be afraid'
	IVPC.full	idiomatic verb-particle construction	(el) **βάζω μπρος** (vazo bros | lit. 'to put forward') 'to start'
	IVPC.semi	semi-idiomatic verb-particle constr.	(sv) ***fråga ut*** (lit. 'to ask out') 'to interrogate / to invite out'
	MVC	multi-verb construction	(egy) 𓇋𓏏𓈖 (**šm.t iw.t** nsw | lit. 'going coming king') 'king paying a visit'
	IAV	inherently adpositional verb	(uk) **вилився в** те (vylyvsya v te | lit. 'spilled into this') 'resulted in this'
NMWE	NID	nominal idiom	(mr) **हातपाय** (hātapāya | lit. 'hand-feet') 'limbs'
	PronID	pronominal idiom	(ja) **何て事** (lit. 'what be even') 'whatever'
	NV	deverbal nominal MWE	(lv) ***kāju atstiepšana*** (lit. 'stretching of one's legs') 'dying'
AMWE	AdjID	adjectival idiom	(grc) **καλοὶ κἀγαθοὶ** (lit. 'good and beautiful') 'physically and morally excellent'
	AdvID	adverbial idiom	(pl) *zrobić coś* ***raz dwa*** (lit. 'to do sth one two') 'to do sth quickly'
	AV	deverbal adjectival/adverbial MWE	(fa) **دل‌نشین** (del neshin | lit. 'sitting on the heart') 'pleasant'
FuncMWE	DetID	determiner idiom	(ro) ***tot felul de*** *demersuri* (lit. 'all kinds of steps') 'various steps'
	AdpID	adpositional idiom	(fr) ***hors de*** *danger* 'out of danger'
	ConjID	conjunction idiom	(he) **כמו כן** (kmo ken | lit. 'as so') 'likewise'
	IntjID	interjection idiom	(pt) ***aí está*** (lit. 'there is') 'here you are'

Table 1: PARSEME typology of MWEs with multilingual examples

does not allow *place* to be in plural (*aux places de Luc* 'at the places of Luc' can only be interpreted literally), although nouns regularly inflect for number in French. Atomic tests like this one are organized in decision diagrams, for the sake of reproducibility, with a specific branch for each syntactic category. There are 76 atomic tests in total, 23 of which are language-specific. This shows PARSEME's contribution to universalist modeling of language.

Some challenges still remain to be solved in the guidelines. Those include notably the choice between AdpID and AdvID categories for examples like ***in the absence (of)*** *someone* vs. ***in*** *someone's* ***absence***. Some tests are also hard to apply to non-spoken languages (Ancient Greek and Egyptian), which do not benefit from native speakers.

The corpus covers 17 languages (Sec. 1). In total, it contains 253,643 sentences, 4,884,449 tokens, and 141,570 annotated MWEs. The most numerous are verbal and functional MWEs (50,107 and 37,255 occurrences, respectively). By far the largest language is Romanian, the smallest are Ancient Greek, Dutch, and Egyptian. Georgian is an outlier in terms of its very low MWE density.

The inter-annotator agreement is measured by F-measure and varies from below 0.4 for Ancient Greek and Hebrew to over 0.9 for Farsi, French, and Dutch.[3] More insight into these discrepancies is needed. We also measure the diversity of the corpus in terms of variety (number of distinct MWEs) and balance (evenness of the distribution of these distinct types). We notably observe that richness grew in PARSEME 2.0 wrt. the previous editions in Persian, Hebrew, Polish, Romanian, Slovene, Serbian, and Swedish. Compared with another multilingual MWE corpus (Tedeschi et al., 2022), PARSEME is substantially richer, despite its substantially smaller size.

3 Identification task

Based on this new release of the corpus, PARSEME organized a new edition of its shared task, with subtask 1 dedicated to automatic identification of MWEs of all syntactic types in running text. The corpora for all 17 languages were split into train/dev/test subsets (except for Ancient Greek, for which only a test set was available). Particular challenges in data annotation and splitting were due to aggressive scraping policies by some AI companies, most notably OpenAI, and to data leakage to large language models (LLMs).

In the MWE identification task, for every language, systems are given a `.cupt`[4] file with UD-compatible morpho-syntactic annotations but without MWE annotations. On output, they return the same file completed with MWE annotations. The competition ran on the Codabench platform[5] and was later cloned as an everlasting benchmark.[6]

Performance is evaluated with the traditional measures: the (strict) MWE-based and the (fuzzy) token-based precision, recall and F-score, as well as their variants dedicated to unseen, discontinu-

[3] See (Savary et al., 2026, Tab. 1) for the detailed IAA scores.

[4] https://gitlab.com/parseme/corpora/wikis/CUPT-format

[5] https://www.codabench.org/competitions/12003/

[6] https://www.codabench.org/competitions/13186/

ous, variable, and single-token MWEs. All scores are given both per language and as cross-lingual macro-averages (Savary et al., 2017).

A novel evaluation dimension in this shared task is diversity. The idea is that the quality of a system's results should possibly go hand in hand with their diversity. Like for the corpus itself (Sec. 2), we evaluate the diversity of the systems' results along the dimensions of variety and balance of MWEs correctly identified by the system (true positives). We also use a hybrid between variety and balance, namely Shannon-Wiener entropy.

The task received 10 submissions, including our own baseline (based on a local installation of the `gpt-oss-20b` model with no fine-tuning and a relatively generic prompt). The leaderboard was ranked according to the MWE-based macro-average F-score. 5 systems covered all 17 languages, 2 covered 16 languages, 2 others covered 6 languages, and 1 system covered only Romanian.

The winner is `MTLB-STRUCT` (Taslimipoor et al., 2020), the very same system which also won edition 1.2 of the PARSEME shared task in 2020. It relies on a fine-tuned encoder (mBERT), like 2 other systems which also outperformed the baseline: `Sahara-Tokenizers` (Karatepe et al., 2026), mBERT-based, and `BeeParser` (Erdem and Karaarslan, 2026), XLM-RoBERTa-based. The 4th system outperforming the baseline, `IPN` (Hülsing et al., 2026), uses a generic generative model (`Qwen3-32B`). The best language-specific scores are reached in Farsi, Japanese, Romanian, and Polish (MWE-based F1$\geq$ 80), followed by Serbian, Slovenian, and Latvian (MWE-based F1$\geq$70). In Egyptian, Ancient Greek, and Dutch, the best systems reach the lowest scores (MWE-based F1$<$30). Note that these results are not directly comparable to those from previous editions of the PARSEME shared task, notably edition 1.2 (Ramisch et al., 2020), for many reasons. First, the scope of annotation is much larger (extended to MWEs of all syntactic types, not only verbal ones), in some languages the corpus is extended or replaced, the train/dev/test corpus splits follow different principles,[7] etc.

The results show that an identification task like ours is still more accurately addressed by supervised models based on fine-tuned encoders than by generative LLMs. The fact that a 6-year-old system outperforms more recent methods might corroborate the hypotheses put forward by Savary et al. (2019). There, we suggested that MWE identification is a fundamentally lexical task, with limited potential for generalization, and that its results primarily depend on how many MWEs from the test corpus were seen in the training corpus. Therefore, much simpler (and fully interpretable) rule/lexicon-based approaches might also remain competitive wrt. deep-learning models, as was the case with the Seen2Seen system (Pasquer et al., 2020) in edition 1.2. While such systems did not compete in edition 2.0, reintroducing them in future experiments is left to future work.

The diversity scores show a strong positive correlation with performance in terms of variety and the variety/balance hybrid, and a moderate negative correlation with balance alone. All detailed results are available in the reference paper (Scholivet et al., 2026).

In the future, more attention should be paid to very low-resourced languages, and to the known challenges which remain unsolved, notably unseen MWEs (Ramisch et al., 2020).

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 Develop-

[7] In edition 1.2, the test corpus for each language was chosen so as to contain a fair number of MWEs unseen in the train/dev corpus. Such a split was not possible in edition 2.0 due to data leakage. Namely, a test corpus could only contain annotations that have never been published before.

ment 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 Stichting Taaltechnologie Utrecht University, (8) 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, (9) a grant of the Ministry of Research, Innovation and Digitalization - UEFISCDI, Romania, project number PN-IV-P8-8.2-EUD-2025-0061, within PNCDI IV, (10) NATO Science for Peace and Security Programme under grant id. G8648, project DeepNewsDef, (11) 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.

References

Dogukan Arslan, Rodrigo Wilkens, Wei He, Dilara Torunoğlu-Selamet, Thomas Pickard, Aline Villavicencio, Adriana S. Pagano, and Gülşen Eryiğit. 2026. MWE-2026 Shared Task 2: AdMIRe 2 - Advancing Multimodal Idiomaticity Representation. In *Proceedings of the 22nd International Workshop on Multiword Expressions (MWE-2026)*.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Ahmet Erdem and Oguzhan Karaarslan. 2026. Cross Lingual BERT at PARSEME 2.0 Subtask 1: Can Cross-Lingual Interactions Improve MWE Identification? In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

Gaston Gross. 1988. Degré de figement des noms composés. *Langages*, 90:57–71. Paris : Larousse.

Maurice Gross. 1986. Lexicon-grammar: The representation of compound words. In *Proceedings of the 11th Conference on Computational Linguistics*, COLING '86, pages 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anna Hülsing, Noah-Manuel Michael, Daniel Ignacio Mora Melanchthon, and Andrea Horbach. 2026. IPN at PARSEME 2.0 Subtask 1: MWE Identification via Related Languages and Attempts at Harnessing Thinking Mode. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

Yunus Karatepe, Mert Sülük, Begüm Özbay, and Zeynep Tuğçe Kırımlı. 2026. Sahara Tokenizers at PARSEME 2.0 Subtask 1: Combining Contextual Embeddings with Structural Decoding for Multi-Word Expression Detection. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

Filip Miletić and Sabine Schulte im Walde. 2024. Semantics of multiword expressions in transformer-based models: A survey. *Transactions of the Association for Computational Linguistics*, 12:593–612.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70:491–538.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020. Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3333–3345, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, and Marco Idiart. 2025. SemEval-2025 task 1: AdMIRe - advancing multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2597–2609, Vienna, Austria. Association for Computational Linguistics.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of 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-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the*

Joint Workshop on Multiword Expressions and Electronic Lexicons, pages 107–118, online. Association for Computational Linguistics.

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91, Florence, Italy. Association for Computational Linguistics.

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.

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.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559, San Diego, California. Association for Computational Linguistics.

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.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Shiva Taslimipoor, Sara Bahaadini, and Ekaterina Kochmar. 2020. MTLB-STRUCT @parseme 2020: Capturing unseen multiword expressions using multi-task 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.

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.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.