Edition 2.0 of the PARSEME Shared Task on Multilingual Identification and Paraphrasing of Multiword Expressions

Manon Scholivet[1], Agata Savary[1], Carlos Ramisch[2], Eric Bilinski[1], Takuya Nakamura[1], Maria Mitrofan[3], Vasile Păiș[3]

[1]Paris-Saclay University, CNRS, LISN, Orsay, France,
[2]Aix Marseille Univ, CNRS, LIS, Marseille, France,
[3]RACAI, Romanian Academy, Romania

[1]first.last@universite-paris-saclay.fr, [2]first.last@lis-lab.fr, [3]first@racai.ro,

Abstract

Multiword expressions (MWEs) have been a major challenge in NLP for decades, and research on MWEs was driven notably by shared tasks, including those organized by the PARSEME community. We report the organisation and the results of edition 2.0 of the PARSEME shared task. For the first time, all syntactic categories are covered: verbal, nominal, adjectival, adverbial and functional. We rely on edition 2.0 of the PARSEME corpus, annotated for all these categories in 17 languages. We create a new dataset with paraphrases of sentences containing idioms in 14 languages, and define a new subtask dedicated to MWE paraphrasing. We extend our evaluation protocol by measuring both performance and diversity of systems, and including manual evaluation in paraphrasing. Ten systems participated in the MWE identification subtask and five in the paraphrasing subtask (baselines included). Results are promising, but known MWE identification challenges remain unsolved. Performance correlates positively with diversity in MWE identification, and negatively in MWE paraphrasing.

1 Introduction

Multiword expressions (MWEs) have been a major challenge in NLP for decades (Sag et al., 2002; Shwartz and Dagan, 2019). This is notably due to their prevalence in texts (Gross and Senellart, 1998; Candito et al., 2021), their partly regular and partly idiosyncratic behaviour (Gross, 1986, 1988; Savary et al., 2020), and their semantic non-compositionality (Mel'čuk, 2010). Many MWE tasks were addressed (Constant et al., 2017) and research has been boosted by SemEval shared tasks (Schneider et al., 2016; Tayyar Madabushi et al., 2022; Pickard et al., 2025; Arslan et al., 2026).

In this landscape, the PARSEME community has been carrying on long-standing efforts towards multilingual modelling of *verbal* MWEs, particularly challenging due to their morphosyntactic flexibility. The major outcomes have been verbal MWE annotation guidelines unified across 26 languages, manually annotated corpora for these languages (Savary et al., 2018, 2023) and 3 editions of a shared task on automatic identification of verbal MWEs (Savary et al., 2017; Ramisch et al., 2018, 2020).

However, MWEs come in all shapes and sizes (Baldwin and Kim, 2010). Beyond verbal MWEs (*to* ***pull*** *one's* ***leg***, *to* ***pay*** *a* ***visit***), there are numerous MWEs of other syntactic categories: functional (***by and large***, ***in that***, ***in spite of***), adjectival (***crystal clear***) adverbial (***by and large***) and nominal (***hot dog***).[1] Recently, PARSEME extended the scope of its guidelines to all these categories in the context of the UniDive COST Action (Savary et al., 2024).

This paper presents edition 2.0 of the PARSEME shared task building on these assets. It features two subtasks and represents a substantial extension of previous editions. The original contributions of this edition can be summarised as follows. First, the PARSEME 2.0 guidelines allowed the creation of MWE-annotated corpora in 17 languages, then used to evaluate systems on the identification of MWEs of *all categories* (subtask 1). Second, a corpus of sentences with idiomatic MWEs and their paraphrases has been created in 14 languages, allowing the evaluation of *paraphrasing systems* on verbal, nominal and adjectival idioms (subtask 2). Third, we propose metrics to assess the *diversity* of system results in both subtasks. Fourth, we rely on the *Codabench* platform to centralise evaluation. In this paper, we discuss related work (§2), the subtasks (§3), the underlying data (§4), the organisation (§5-6), the results (§7-8), and conclusions (§10).

2 Related Work

MWE Shared Tasks PARSEME 1.0 covered 18 languages and introduced the task of token-

[1] MWE examples follow the PMWE conventions (Markantonatou et al., 2021), enriched with colors and brackets.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 254–275
March 28, 2026 ©2026 Association for Computational Linguistics

level MWE identification. PARSEME 1.1 covered 20 languages, introduced the CUPT format and phenomenon-oriented evaluation metrics. PARSEME 1.2 covered 14 languages and focused on unseen MWEs, with controlled splits, full UD integration, and companion raw corpora. The three previous editions only cover *verbal MWEs*. In addition to these PARSEME shared tasks, other evaluation campaigns covered idiomaticity and MWEs.

The SemEval-2016 Task 10 (DiMSUM) for English tested the detection of minimal semantic units, including MWEs, and their meanings (Schneider et al., 2016). The underlying corpus contained online customer reviews, tweets, and TED talks, and was notably annotated for 2 MWE classes: strong or weak, according to their degree of idiomaticity.

In SemEval-2022 Task 2 for English, Portuguese, and Galician, systems competed for two tasks (Tayyar Madabushi et al., 2022). In task A, given a sentence containing a potentially idiomatic expression and its span, systems should decide if the expression was used idiomatically or literally, both in zero- and one-shot settings. In Task B, given a sentence with a MWE and two other sentences in which the MWE was replaced by its paraphrase and by a distractor (formally close but semantically distant), systems had to decide which pair of sentences was closest in meaning.

Finally, SemEval-2025 Task 1 (AdMIRe) for English and Brazilian Portuguese was dedicated to multimodal idiomaticity representation (Pickard et al., 2025). The task's goal was to align images depicting MWEs having more or less figurative meanings with sentences containing the same expressions used literally or idiomatically. The task was extended to 15 languages in AdMIRe 2 (Arslan et al., 2026), co-organised by UniDive jointly with our shared task.

Paraphrasing Shared Tasks Automatic processing of paraphrase also has a rich state of the art. Butnariu et al. (2009) test systems for accurate scoring of alignments between English noun compounds and their potential paraphrases, e.g. *sleeping pill* vs. *pill that induces sleeping*. Hendrickx et al. (2013) extend the previous task to automatically produce a ranked list of paraphrases for a given English noun compound, e.g. *air filter* → *filter for air*, *filter that cleans the air*. Evaluation measures are based on approximate n-gram matching between the system-generated paraphrases and those produced by human experts, with rank-based scaling.

Later, the scope of paraphrasing was extended to whole sentences, with 3 subtasks. ***Paraphrase identification*** consists in binary classification of pairs of sentences as being paraphrases or not (Xu et al., 2015; Lan et al., 2017). ***Semantic textual similarity*** is defined as assigning sentence pairs a similarity score from 0 to 1 (Agirre et al., 2015; Xu et al., 2015). Finally, ***paraphrase generation*** consists in reformulating a sentence to use a different wording or structure but preserve the original meaning (Zhou and Bhat, 2021). Evaluation relies on measures from machine translation (ROUGE, BLEU, METEOR or TER) or human scoring along multiple dimensions such as similarity, clarity, or fluency.

MWE Paraphrasing MWEs are particularly challenging for paraphrasing due to their non-compositional semantics. In related work, one of the motivations behind paraphrasing MWEs with their literal equivalents is eliminating idiomaticity prior to machine translation, as done by Santing et al. (2022) for English-German MT. Dedicated MWE-aware paraphrase datasets were built upon MWE definitions in lexicons (Pershina et al., 2015; Liu and Hwa, 2016), collected by crowdsourcing (Yimam et al., 2016), for English in both cases, or relied on machine (back-)translation (Qiang et al., 2023), for Chinese. Many verbal MWEs can be paraphrased by a single verb, as shown by Barančíková and Kettnerová (2018) for Czech. Tan and Jiang (2021) adapt paraphrase identification to idioms, a task similar to disambiguating literal from idiomatic MWE uses. Zhou et al. (2021) introduce 2 tasks: ***idiomatic sentence generation*** transforms a literal sentence into a sentence involving idioms; and ***idiomatic sentence paraphrasing*** 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). Evaluation metrics include ROUGE, BLEU, METEOR, GRUEN, BERT perplexity, as well as human judgements on semantics and fluency.

3 Task Definition and Metrics

Subtask 1: MWE Identification This historical PARSEME task focuses on token-level MWE identification in running text, as in previous editions. Systems are given as input a morphosyntactically analysed sentence in CoNLL-U format.[2] As output, they must group the tokens that belong to MWEs, assigning them a single label. For instance :

[2] https://universaldependencies.org/format

(1) **En plus**, ça **fait partie** du **centre ville** (fr)
In plus, it does part of.the centre city
'Moreover, it is part of the city centre.'

In the sentence above, the tokens belonging to the three MWEs (in bold) should be assigned unique labels, e.g. 1: (***En plus***), 2: (***fait partie***) and 3: (***centre ville***). Those not belonging to any MWE (*ça* and *du*) should not be assigned any label. Systems solving this task must address several challenges (Constant et al., 2017): discontinuities e.g. (fr) ***fait*** *toujours* ***partie*** 'is still part', morphological and syntactic variability as in (2), overlapping or nesting,as in (3), and idiomatic-literal ambiguity as in (4) vs. (5).

(2) a **da** un **sfat**, **sfaturi** au fost **date** (ro)
to give an advice, advices have been given

(3) **temos**$_{1,2}$ um **plano**$_{1}$ et uma **intenção**$_{2}$ (pt)
have.PL a plan and an intention
'We have a plan and an intention'

(4) この 問題 は **朝飯** **前** だ (ja)
this problem about breakfast before be
lit. 'This problem is before breakfast.'
'This problem is very easy.'

(5) 朝飯 前 に 会う (ja)
breakfast before LOC meet
'We meet before breakfast.'

The corpora are provided in CUPT format.[3] They are split into training, development, and test sets. The latter are available only during the evaluation phase (about 1 week) and gold annotations are not disclosed. Test corpora are completely new with respect to previous editions to prevent LLM-contamination (§ 5).

Annotated MWEs are assigned category labels (e.g. NID for nominal idiom, MVC for multi-verb construction). While previous editions covered only verbal MWE categories, the current edition covers all MWE categories, including nominal, verbal, adjectival, adverbial, and functional MWEs (see § 4). These category labels can guide system development, but they are not taken into account in evaluation metrics. Thus, systems need to group tokens belonging to the same MWE, but they do not have to tag the resulting MWE with a specific category.

The evaluation of this subtask is performed using two standard F-score variants: MWE-based and token-based (Savary et al., 2017). The former accounts for exact matches between all tokens of the predicted MWE and of the reference MWE, whereas the latter rewards partial matches, covering only part of the tokens. In addition, we report phenomenon-specific F-scores, focusing on discontinuous, single-token, variant and unseen MWEs (Ramisch et al., 2018). Edition 1.2 focused on unseen MWEs, that is, those whose multi-set of lemmas are annotated as MWEs at least once in the test corpus, but never in the training or development corpus (Ramisch et al., 2020). In the current edition, we propose and analyse diversity scores that also partly account for unseen/novel identified MWEs.

Subtask 2: MWE Paraphrasing This novel subtask is motivated by recent advances in text generation. We wish to challenge modern generative systems with idiomaticity-related problems in more advanced scenarios than done so far (§ 2). Paraphrasing may be a useful method for testing the ability of models to grasp the meaning of an MWE (Tayyar Madabushi et al., 2022; He et al., 2025). MWE paraphrasing may also help for text simplification.

First, we address paraphrase generation rather than binary detection or similarity scoring. Second, paraphrasing is not restricted to the MWE itself but, conversely, reformulation of other parts of the sentence is encouraged and rewarded by diversity metrics. Third, our gold paraphrases are produced by native speakers along unified guidelines for an unprecedented number of 14 languages. Finally, we use LM-driven evaluation (BERT-score) and show its good correlation with human evaluation.

The input for this task is a raw sentence containing exactly one verbal, nominal or adjectival idiom, not explicitly marked in text.[4] Systems must paraphrase the sentence so that the original MWE no longer occurs, but the meaning is kept. For instance, sentence (6) could be paraphrased as (7) or as (8).

(6) le **point de vue** de la réalisatrice … (fr)
the point of vue of the director …
'the director's point of view …'

(7) la perspective de la réalisatrice … (fr)
the perspective of the director …
'the director's perspective …'

(8) la vision du metteur en scène … (fr)
the vision of.the putter in scene …
'the stage director's vision …'

Additionally, to facilitate automatic evaluation, at least one of the lemmas of the original MWE should

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

[4] VID, NID or AdjID in the PARSEME typology (see § 4).

be totally absent from the paraphrase. For instance, (fr) ***peine de mort*** (lit. 'punishment of death') 'death penalty' should not be paraphrased as *peine consistant à causer la mort de la personne* 'punishment causing the death of the person'. We allow paraphrases to use MWEs, provided that they are different from the original one, as in (9)–(10).

(9) Dla nich świat **stanął w miejscu**. (pl)
For them world stood in place.
'For them the world stands still.'

(10) Dla nich świat przestał **się rozwijać**. (pl)
For them world stopped itself unroll.
'For them the world stopped developing.'

In subtask 2, only trial data in English and French is provided, but no training nor development data. The test data contains between 66 and 150 sentences per language. Like for subtask 1, the blind test data are made available to system authors for a week, and gold annotations are not disclosed. All test files are distributed in `.json` format.

Two evaluation measures are used. *Masked BERT-score* first checks if at least one of the MWE components was removed. If not, the score assigned to the paraphrase is 0. Otherwise, BERT-score (Zhang et al., 2020) is calculated between the system-generated paraphrase and up to two reference paraphrases: a minimal and a creative one (§ 4). The maximum of the two scores is retained. The second measure is *manual score*. For each sentence, native or near-native speakers are presented the paraphrases submitted by systems. In addition, annotators also see up to 2 reference paraphrases (minimal and creative), without knowing whether the paraphrase was generated by systems or by humans. This allows us to verify the quality of reference paraphrases with respect to system outputs. Annotators assign score 0 if the MWE is 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 is a weighted average of these 3 scores, with score (i) doubled, normalized to $[0, 100]$. Both masked BERT-score and manual score are averaged across all sentences, then macro-averaged across languages.

Diversity Metrics 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. In general, diversity is modelled as a property of *sets* whose *elements* can be apportioned into *categories*. It is here evaluated along two main dimensions: variety and balance (Stirling, 2007; Ramaciotti Morales et al., 2021; Estève et al., 2025). *Variety* relates to the number of categories, and *balance* to the evenness of the distribution of elements into categories. All other things being equal, the higher the variety, the higher the diversity, and the same holds for balance.

In our case, the sets evaluated for diversity are systems' predictions. In subtask 1, we follow Lion-Bouton et al. (2022), defining categories as *MWE types* and elements as their *occurrences* in text.[5] Only MWEs correctly identified by a system (i.e. true positives) are considered. Consider the toy test corpus (11)-(13), where MWE categories are boldfaced and bracketed, and a wrongly identified MWE (i.e. a false positive) is underlined:

(11) [**Me deparei**] [**cara a cara**] com… (pt)
Myself appeared face to face with…
'I found myself faced with…'

(12) [**Me dei mal**]: fiquei [**cara a cara**]… (pt)
Myself gave bad: got face to face…
'I was in a bad situation: I was facing…'

(13) Vendo a cara do pai, [**fez cara feia**] (pt)
Seeing the face of father, made face ugly
'Seeing her father's face, she frowned'

Suppose that system S_1 identified all these 6 expressions, i.e. 4 categories, 5 elements (true positives), and one false positive (ignored by diversity scores). System S_2, in turn, identified only the 3 categories and 3 elements from examples (12) and (13). We have $N_{S_1} = 4$ and $N_{S_2} = 3$, the number of categories of each system. As a measure of variety, we use *richness*, i.e. N_S. According to this measure, the predictions of S_1 are richer than those of S_2.

To assess balance, we use *Shannon evenness* (Smith and Wilson, 1996) defined by equation (14):

$$SE_S = \frac{SWE_S}{\ln(N_S)} \tag{14}$$

where SWE_S is the Shannon-Weaver entropy:

$$SWE_S = -\sum_{i=1}^{N_S} p_i * \ln(p_i) \tag{15}$$

[5] A MWE type is represented by the multiset of its components' lemmas. For instance, given the MWE (pt) ***cara a cara*** (lit. 'face to face') '(suddenly) facing', its multiset of lemmas, in lexicographic order, is {*a* 'to', *cara* 'face', *cara* 'face'}.

and p_i is the frequency of the ith category.[6] For S_1, $(p_1, p_2, p_3, p_4) = (\frac{1}{5}, \frac{2}{5}, \frac{1}{5}, \frac{1}{5})$ and $SWE_{S_1} = 1.33$, while for S_2, $(p_1, p_2, p_3) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $SWE_{S_2} = 1.1$. In eq. (14), entropy SWE is divided by its maximum value $\ln(N_S)$, so we have $SE_{S_1} = \frac{1.33}{1.39} = 0.96$ and $SE_{S_2} = \frac{1.1}{1.1} = 1.0$. Thus, the predictions of S_2 are more balanced than those of S_1.

Our last diversity measure is *Shannon-Weaver entropy* itself, from eq. (15). When not normalized, SWE_S is actually a hybrid metric, accounting for both variety *and* balance. According to SWE_S, the predictions of S_1 are more diverse than those of S_2.

For subtask 2, the same diversity measures are adapted by redefining categories as unique word types generated by a system and not present in the original sentence. For instance, given sentence (6), if systems S_3 and S_4 produced the outputs (7) and (8), then $N_{S_3} = 1$ (new words: *perspective*) and $N_{S_4} = 5$ (new words: *vision*, *du*, *metteur*, *en*, *scène*), $SWE_{S_3} = 0$, $SWE_{S_4} = 1.61$, $SE_{S_3} = 0$, and $SE_{S_4} = 1$. Thus, S_3 has less diverse predictions than S_4 according to the 3 measures.

4 Provided Data

Subtask 1. The dataset is a fruit of the PARSEME annotation campaign in which 17 teams took part, covering 10 previously covered languages – Modern Greek (el), Persian (fa), French (fr), Hebrew (he), Polish (pl), Portuguese (pt), Romanian (ro), Slovene (sl), Swedish (sv), Serbian (sr) – and 7 new ones – Egyptian (egy, ca. 2700-2000 BC), Ancient Greek (grc), Japanese (ja), Georgian (ka), Latvian (lv), Dutch (nl) and Ukrainian (uk). Human annotators worked on the corpus according to cross-linguistically unified guidelines composed of decision trees over elementary morphological, syntactic or distributional tests (Savary et al., 2026).[7]

Previous versions of the PARSEME corpora treated only verbal MWEs. Version 2.0 covers all MWE categories: verbal, nominal, adjectival-adverbial and functional. Some of those are subdivided into subcategories, such as:

- verbal idioms (VID): (nl) ***ijs breken*** 'break the ice';
- nominal idioms (NID): (ja) **一人相撲** (lit. 'one-person sumo') 'wrestling with oneself';
- adjectival idioms (AdjID): (sr) ***мртав пијан*** (lit. 'dead drunk') 'extremely drunk';
- conjunctive idioms (ConjID): (lt) ***kā arī*** (lit. 'as also') 'as well as'.

The test data have been synchronised between task 1 and 2. First, we identified all "unseen" sentences, e.g. those that have never been annotated in previous PARSEME editions (even partially), to avoid contamination (§5). Out of those, up to 150 sentences per language were randomly selected to meet the criteria for subtask 2, i.e. containing a single MWE (of category VID, NID or AdjID). For subtask 1, they were then completed with other randomly selected "unseen" sentences so as to reach roughly 500 annotated MWEs. As a result the test data for subtask 1 contains between 300 and 1,900 sentences per language.[8] The remaining sentences were split randomly (90%-10%) to create the training and development sets, which were provided to the participants, except in Ancient Greek, where not enough annotated data were available and only a test set exists.

Subtask 2. The dataset for subtask 2 is totally new and provided for 14 languages: the same as in subtask 1, except Egyptian, Ancient Greek, and Dutch.[9] Based on the test set of subtask 1, we extracted up to 150 sentences, as described above, meeting the criteria for subtask 2. We selected sentences containing only 1 MWE to simplify the definition and evaluation of subtask 2. We focus on VID, NID and AdjID because their degree of non-compositionality seems overall the highest, avoiding notoriously hard discussions about partial compositionality, as in (ro) ***ia măsuri*** 'take action', (ro) ***în cadrul*** (lit. 'in frame') 'in the framework (of)'.

Given a sentence with a highlighted idiom, native human experts were to provide at least one of 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 (16) received the minimal paraphrase (17) and the creative one (18):

[6] E.g. for S_1 predictions, p_1 is the frequency of the category *me deparei*, p_2 the frequency of *cara a cara*, ...

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

[8] With one outlier, Georgian, having around 40K sentences.

[9] Egyptian and Ancient Greek are skipped because they are non-spoken languages, while paraphrasing can be performed reliably only by native speakers. Dutch is skipped due to the unavailability of the annotators at the time of the paraphrase corpus construction.

(16) ხელისუფლება PDPA-მ მოქმედებაში
The.government PDPA in.action
მოიყვანა სოციალისტური **დღის**
brought socialist of.the.day
წესრიგი. (ka)
order.

‘The PDPA government put into action a socialist agenda.’

(17) ხელისუფლება PDPA-მ მოქმედებაში
The.government PDPA in.action
მოიყვანა სოციალისტური გეგმა. (ka)
brought socialist plan.

‘The PDPA government brought into action a socialist plan.’

(18) სოციალისტური წყობის რეალიზება
socialist set.up realization
ხელისუფლება PDPA-მ მოქმედებაში
the.government PDPA in.action
მოახერხა. (ka)
succeeded.

‘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.

5 Running a Shared Task in the LLM era

So far, the PARSEME corpus in all 4 versions, as well as the system results from editions 1.1. and 1.2 of the PARSEME shared tasks, have been made publicly available under open licenses on the CLARIN/LINDAT infrastructure[10] and in public Gitlab repositories.[11] The most recent versions of the data being annotated have regularly been uploaded to public Gitlab language repositories, and made available by consistency checking web pages, to the best benefit of the research community.

Most of these practices have recently been jeopardised by aggressive scraping policies of some AI companies. Their bots scan the Internet, strongly targeting open source community infrastructures,[12] ignoring conventions such as `robots.txt` files.[13]

The PARSEME infrastructure is also concerned. Particularly intrusive is GPTBot, which scrapes data to train OpenAI’s products. For instance, it sent almost 3 million queries in April–December 2025 to two of our servers, with up to 14,000 queries per day per server, likely acting in “distributed denial of service” mode to remain anonymous. OpenAI is known to violate the licenses under which data and software are distributed (Mueller, 2025). Last but not least, data contamination (Deng et al., 2024), particularly frequent due to LLMs, occurs when test data are included in the training phase, which leads to inflated performance scores.

This last risk drove major challenges in our data annotation and publication policy. Texts previously published with MWE annotations could no longer be used as test data. Thus, for languages from previous PARSEME corpus editions, we had to add significant amounts of new data annotated for all MWE categories from scratch. This prevented us from applying random or custom train/dev/test splits, used for estimating performance, notably on unseen MWEs (Ramisch et al., 2020). We also had to make private our public git repositories, used for everyday corpus development, and to hide consistency checking pages behind secret URLs, burdening legitimate users with new procedures. Even so, these changes do not preclude data contamination, since corpus or system developers may inadvertently store copies of test data in their own public spaces.

6 Implementation on Codabench

For the first time, the participants’ submissions to PARSEME shared tasks were evaluated on the Codabench platform, an online framework designed for running machine learning competitions (Xu et al., 2022). Codabench allows benchmarks to run in a stable and a less error-prone environment based on docker, not subject to library version changes. Benchmarks are easily reproducible, and

[10] E.g. `http://hdl.handle.net/11372/LRT-5124`

[11] `https://gitlab.com/parseme/sharedtask-data`

[12] `https://next.ink/186593/les-crawlers-des-ia-menacent-les-sites-scientifiques`

[13] `https://arstechnica.com/ai/2025/03/devs-say-ai-crawlers-dominate-traffic-forcing-blocks-on-entire-countries/`

their scores are available online for participants right away. They occur on the leaderboard, which provides permanent links for lasting access.

Two competitions were established, each corresponding to a specific subtask.[14] A comprehensive bundle was prepared and uploaded to the platform to calculate the scores of the submissions. The bundle comprises gold data, the scoring tool and its dependencies, as well as the participant instructions that elucidate the competition's goals and the submission process. The bundle also includes a configuration file that specifies the paths of all the data, the docker container image, start and end dates of the competitions, and the maximum number of submissions per participant (here: 10).

During the competition, participants submitted a zip file with one directory per language, each containing their system's predictions. The scoring program returned numerical scores, displayed on the leaderboard, where the performances of participants were compared. For subtask 1, the leaderboard ranking was based on the global MWE-based F-score, but global and token-based precision and recall were also displayed, with an additional link to detailed results per language. For subtask 2, the ranking was based on average masked BERT-score.

The competitions and the results are now frozen, but a copy was created so that new participants can continuously propose new solutions.[15] They will no longer have to wait for a new evaluation to quickly and accurately assess their systems under the same conditions as in the shared task.

7 Systems

Subtask 1. This subtask features 10 participating systems: 9 submissions plus the baseline (Tab. 1). Among them, 5 are based on pre-trained encoder transformer models, fine-tuned for the task using BIO-style tags. `MTLB-STRUCT` relies on `bert-base-multilingual-cased` (Taslimipoor et al., 2020), with no auxiliary parsing task.[16] `Sahara-Tokenizers` (Karatepe et al., 2026) relies on the same pre-trained model, but introduces (a) explicit part-of-speech injection and (b) multi-task objective for joint BIO-style tagging and category classification. `Bert-multilingual-trial` and `BeeParser` (Erdem and Karaarslan, 2026) fine-tune `XLM-RoBERTa-base` on single languages, but also on language pairs, studying cross-lingual transfer. Finally, `romanian-bert` (Roscan and Nisioi, 2026) fine-tunes the language-specific `RoBERT-base` model after after comparing several models on challenging data subsets. One system, `pmi-mwe-scorer` (Bogdanova and Bucur, 2026), proposes a method based on syntax-aware pointwise mutual information (PMI) that leverages UD trees. Two systems rely on generative language models: `IPN` (Hülsing et al., 2026) applies instruction fine-tuning to `Qwen3-32B`, while `MorphoFiltered-Gemini` (Moise and Nisioi, 2026) relies on `gemini-2.0-flash-lite` with a lightweight morphological filter to remove unlikely outputs.

Subtask 2. We received 5 submissions listed in Tab. 2, including the baseline. All of them are based on LLMs: `GPT-CREATIVE` (Roscan and Nisioi, 2026) relies on prior MWE identification (with `romanian-bert` of subtask 1) followed by `GPT-4o` queries using category-oriented prompts. `Star-Paraphrasing-Cosine` (Bayraktar et al., 2026) and `Multiagent` are variants: `Cosine` tries to substitute a pre-identified MWE by single-word alternatives weighted by cosine similarity, while `Multiagent` is based on a combination of LLMs that generate, validate, and fix the paraphrase. Finally, `MISP` (Ciminari and Barrón-Cedeño, 2026) relies on `Qwen3-4B-Instruct` and cross-lingual transfer, fine-tuning the model on synthetic MWE paraphrases in Portuguese.

Baselines The baseline was implemented in Java as an API client for LLMs. It allows communication with both cloud-based APIs and locally hosted LLMs. To produce the baseline results for the test sets, we used the `gpt-oss-20b` model through a local Ollama installation. This prevented data leakage to cloud-based solutions (§ 5). The baseline system allows specifying a dataset and custom templates to be used as the system and user prompts. Each sample in the dataset is converted into a LLM call by filling the prompt templates. Templates indicate to the LLM the need to produce output that can be parsed by the system. For the first subtask, the LLM identifies MWEs at the entire sentence level, which are then mapped back into the tokenized form. For the second subtask, the LLM directly produces the new sentence. We also conducted preliminary experiments with `Llama-4-Scout` and `gpt-oss-20b`

[14] Subtask 1: https://www.codabench.org/competitions/12003/, subtask 2: https://www.codabench.org/competitions/12002/

[15] Subtask 1: https://www.codabench.org/competitions/13186/, subtask 2: https://www.codabench.org/competitions/13192/

[16] https://github.com/shivaat/MTLB-STRUCT/

however, gpt-oss-20b consistently achieved better performance than the other two models. The prompts[17] used for both subtasks were intentionally simple and served only to guide the model toward the expected output format. Aiming at a baseline contribution, we did not focus on heavily refining the prompts in order to obtain highly competitive results with language-specific elements.

8 Performance Results

Subtask 1. Out of the 10 participating systems, 5 systems cover all 17 languages for which data were available, 2 systems cover 16 languages, 2 systems cover 6 languages, and 1 system covers only one language (Romanian). In this section, systems are referred to by their names on the leaderboard.

Tab. 1 presents the general ranking of subtask 1, including the baseline. The number of languages covered by each system is shown in column *#Langs*. Then, we report the global MWE-based and token-based precision (P), recall (R) and F-scores (F1), with results macro-averaged over languages. Macro-average calculation ranges over all 17 languages. If a system did not submit results for a given language, this language is included in macro-average calculation as having P=R=$F1$=0. Systems are ranked by decreasing F-scores. Phenomenon- and language-specific results are shown in App. A .

According to both MWE-based and Token-based F1, the top-3 systems are MTLB-STRUCT, Sahara-Tokenizers and IPN. In terms of MWE-based F1, the best system MTLB-STRUCT beats the second best Sahara-Tokenizers by almost 9 points. The difference between the second and third ranks is even larger, reaching 19.9 MWE-based F1 points and 23.3 Token-based F1 points. MTLB-STRUCT favours precision, whereas the two other systems favour recall. On the other hand, while 4 systems overcome the baseline, 5 of them fail to do so, among which 3 cover between 16 and 17 languages.

These average results ignore inter-language variability. For instance, romanian-bert is the best system for Romanian, reaching 85.65 MWE-based F1. MTLB-STRUCT has the highest MWE-based F1 for 8 languages, but Sahara-Tokenizers beats it in 3 languages, while bert-multilingual-trial is the best in 2 languages, IPN is the best for Dutch, BeeParser is the best for Serbian, and the baseline has the highest MWE-based F1 on Ancient Greek.

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). No training or development data was provided for Ancient Greek, while Egyptian and Dutch have the smallest training and development corpora, with 103 and 133 annotated MWEs in total.

Phenomenon-specific scores (App. A, Tab. 4-7) confirm that the main challenges in MWE identification, studied since edition 1.1, remain unsolved, especially for unseen MWEs. We emphasize that the best system from edition 1.2, MTLB-STRUCT, still gains the upper hand, suggesting that little progress is achieved in modern LLMs concerning MWE identification, despite their progress in other tasks.[18]

Subtask 2. Four teams submitted predictions of 5 systems (including the baseline), shown in Tab. 2, together with their global macro-average scores and ranks for the automatic (global masked BERT-score) evaluation. Only the baseline covered all 14 languages. The 4 other systems jointly covered 4 languages: French (fr), Georgian (ka), Portuguese (pt) and Romanian (ro). We performed manual evaluation only for these 4 languages, therefore we do not report the respective global ranking. The coverage of a low number of languages explains the low scores for the 4 last systems. This is why per-language scores, given in App. A are more interesting to analyse. Tab. 26–29 show that automatic evaluation (with masked BERT-score) nicely correlates with manual evaluation. More precisely, Pearson and Spearman correlation between the automatic and the manual scores for these 4 languages amount to 0.92 and 0.90, respectively. This indicates that (masked) BERT-score is a promising measure for MWE paraphrasing, despite its known weaknesses (Hanna and Bojar, 2021; Sun et al., 2022).

Tab. 26 and 29 in App. A show that systems specialised in one language (Star-Paraphraser in French, GPT-CREATIVE in Romanian) largely outperform the baseline. In French, Star-Paraphraser-Cosine has a high automatic score, but the manual score downgrades it to the third position.

Tab. 3 shows inter-annotator agreement for manual evaluation. We use Krippendorff's α Artstein and Poesio (2008) well suited for numerical scores,

[17] https://github.com/racai-ai/mwe_baseline/tree/master/templates

[18] One caveat is that the shared task's particularly tight schedule may have prevented the development of complex systems.

System	#Langs	Global MWE-based				Global Token-based			
		P	R	F1	rank	P	R	F1	rank
MTLB-STRUCT	17/17	62.21	53.09	57.29	1	70.55	54.75	61.65	1
Sahara-Tokenizers	17/17	45.77	51.33	48.39	2	61.53	57.12	59.24	2
IPN	17/17	21.37	42.32	28.40	3	26.32	56.66	35.94	3
BeeParser	6/17	26.62	25.84	26.22	4	29.26	27.03	28.10	6
baseline-gpt-oss-120b	17/17	17.44	34.86	23.25	5	23.59	52.93	32.64	4
bert-multilingual-trial	6/17	21.69	20.30	20.97	6	26.78	23.26	24.90	7
MorphoFiltered-Gemini	17/17	20.95	14.50	17.14	7	34.14	24.20	28.32	5
romanian-bert	1/17	5.35	4.76	5.04	8	5.55	4.82	5.16	10
Pattern-Based-MWE-Identifier	16/17	2.25	12.69	3.82	9	13.15	50.07	20.83	8
pmi-mwe-scorer	16/17	0.97	2.59	1.41	10	7.15	22.66	10.87	9

Table 1: Subtask 1 results – number of languages covered by systems (#Langs); then macro-averaged MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks.

System	#Langs	gm-BS	Autom. rank
baseline-gpt-oss-120b	14/14	71.62	1
MISP	4/14	14.21	2
Star-Par.-Cosine	1/14	6.71	3
Star-Par.-Multiagent	1/14	6.39	4
GPT-CREATIVE	1/14	6.38	5

Table 2: Subtask 2 results – number of languages covered by systems (#Langs); global masked BERT-score (gm-BS) and the associated rank.

Language	A_1-A_2	A_1-*Adj*	A_2-*Adj*
French (fr)	–	96.50	97.37
Portuguese (pt)	80.30	87.56	91.14
Romanian (ro)	75.99	90.37	87.42

Table 3: Manual evaluation of subtask 2, inter-annotator agreement (Krippendorff's α, interval difference): pairwise scores between annotators (A_1, A_2) and adjudicator (Adj). Values multiplied by 100 for better readability.

with disagreements between s_1 and s_2 weighted proportionally to $(s_1 - s_2)^2$. Georgian is omitted because there was only one annotator. For the other 3 languages, a third adjudicator Adj unified annotations, which were then used to assess system's performances. For French there were no overlapping items between annotators A_1 and A_2. Agreement between A_1 and A_2 ranges from 75 to 80, whereas it is greater than 85 with respect to Adj. Thus, our evaluation protocol seems reproducible, although assessing meaning similarity is usually a hard task.

9 Diversity Results

Subtask 1. Previous editions of PARSEME have shown a strong correlation between the number of unseen MWEs in the test sets and the overall performance of systems. The diversity measures we propose in this new edition are a continuation of these reflections. We therefore began by measuring the correlation between the performance scores of systems and their diversity scores. Detailed diversity scores are available in App. A. We calculated a correlation score for each language, which we then averaged to obtain an overall view (Tab. 25). This gave us a Pearson correlation of 0.72 and a Spearman correlation of 0.76 between MWE-based F1 and the hybrid variety-balance measure. These results, confirming those obtained by Lion-Bouton et al. (2022), once again highlight the importance of predicting diverse MWEs in order to obtain high-quality predictions.

However, when we look at the correlation between performance and not entropy, but variety and balance individually, we see a significant difference in behaviour. Variety is correlated with performance at 0.81 (Pearson and Spearman), while balance is correlated at -0.39 (Pearson) and -0.46 (Spearman). It would therefore appear that entropy is more impacted by variety than by balance, and that a more balanced system would have a negative impact on performance, unlike a varied system.

Subtask 2. Per-language diversity scores are reported in App. A. Conversely to subtask 1, we see the so-called performance-diversity trade-off typical for generation scenarios (Ippolito et al., 2019; Zhang et al., 2021). For instance in French (Tab. 26), the higher the performance, i.e. the quality of the generated paraphrases, the lower the diversity, and vice-versa. One exception is Star-Paraphraser-Cosine. It has the lowest diversity, which is likely why it obtains high BERT-scores (the generated paraphrases resemble the original sentence). However, it does not achieve the highest manual score, which means that in reality it does not perform particularly well. Notable is also MISP in Romanian (Tab. 29), which shows a particularly high lexical creativity (richness), but low manual scores.

Overall, balance scores are rather high across both subtasks. In subtask 1, this might result from a high number of infrequent MWEs. In subtask 2, the newly introduced vocabulary items (not appearing in the original sentence) might also often be hapaxes. More insight into these results will be gained from future analyses, and from new editions of the shared task, with more systems and languages.

10 Conclusions and Future Work

The data and systems discussed here are only the beginning of a deeper study of MWE identification and paraphrasing in the LLM era. In addition to traditional metrics, human evaluation (subtask 2) and diversity scores provide complementary views on the results. The overall trend in subtask 1 indicates that pre-trained encoder models and BIO encoding are still competitive. The results of subtask 2 are an initial step towards MWE paraphrasing, that we intend to generalise cross-lingually.

Limitations

The use of a baseline based on an LLM entails very large processing costs in subtask 1, especially in Georgian (including times and machine requirements).

The Georgian (ka) dataset is extremely large compared to other languages, and is very sparse, containing few annotations. This raises questions about the guidelines and its interpretation, which may vary depending on the language. The size of the Georgian corpus therefore involves very long processing times.

The amount of data for all languages is not balanced. Some languages have small training corpora (e.g. Dutch) and in particular Ancient Greek (grc) has no training nor development data available.

The use of BERT-score for automatic evaluation is known to have numerous weaknesses (Hanna and Bojar, 2021; Sun et al., 2022), sometimes ranking participants imperfectly compared to manual evaluation.

We do not calculate nor report statistical significance in the rankings: some small observed differences may be due to chance. Further analyses such as bootstrapped p-values are required to establish the robustness of our results (Ramisch et al., 2023).

Ethical Considerations

Despite the concerns raised in Section 5, the baseline for the shared task is based on an LLM. Although we chose to use the one with the highest level of openness available to us, the model is not completely open source, as the weights of the model are available, but we do not know the exact training corpus used to train this LLM.

Furthermore, the languages participating in the shared task are predominantly Indo-European languages, which are not particularly low resourced. The addressed languages do not include the 3 lowest levels of the Joshi et al. (2020) resourcedness scale.

Acknowledgements

This work received support from the CA21167 COST action UniDive, funded by the European Union via COST (European Cooperation in Science and Technology). Further support came from (i) the French Agence Nationale pour la Recherche, via the SELEXINI project (ANR-21-CE23-0033-01), (ii) the Romanian Ministry of Research, Innovation and Digititalization - UEFISCDI, project number PN-IV-P8-8.2-EUD-2025-0061, within PNCDI IV, (iii) the LLMS4EU project 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.

We thank Louis Estève for adapting the diversity metrics scripts to the shared task. We are grateful to all the language leaders and annotators of the PARSEME community who contributed to created the PARSEME 2.0 datasets: `https://gitlab.com/parseme/corpora/-/wikis/home#language-teams`. We would also like to thank the annotators who contributed to the manual evaluation of subtask 2: Gabriela Berndt de Souza, Mihaela Cristescu, Letícia Guedes Guimarães, Irina Lobzhanidze, Verginica Mititelu, Adriana Pagano and Carmen Vasile. We are grateful to the Codabench support team, the AdMIRe 2 shared task organisers, the MWE 2026 workshop chairs, the members of the UniDive COST Action, and the MWE Section of SIGLEX for their continued feedback and support.

We dedicate this work to the memory of Federico Sangati and Silvio Ricardo Cordeiro.

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.

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)*.

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

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of natural language processing*, volume 2, pages 267–292. CRC Press, Boca Raton, USA.

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.

Elif Bayraktar, Vedat Doğancan, Muhammed A. Gümüş, and Nusret Ali Kızılaslan. 2026. Semantic Stars at PARSEME 2.0 Subtask 2:Alternative Approaches for MWE Paraphrasing. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

Anna Bogdanova and Ileana Bucur. 2026. PMI MWE Scorer at PARSEME 2.0 Subtask 1: identifying multiword expressions using pointwise mutual information and universal dependencies. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

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.

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier, and Silvio Cordeiro. 2021. A french corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*, 8(2).

Debora Ciminari and Alberto Barrón-Cedeño. 2026. MISP at PARSEME 2.0 Subtask 2: A Cross-lingual Approach to Multiword Expression Paraphrasing. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

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.

Chunyuan Deng, Yilun Zhao, Yuzhao Heng, Yitong Li, Jiannan Cao, Xiangru Tang, and Arman Cohan. 2024. Unveiling the spectrum of data contamination in language model: A survey from detection to remediation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16078–16092, Bangkok, Thailand. Association for Computational Linguistics.

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.

Louis Estève, Marie-Catherine de Marneffe, Nurit Melnik, Agata Savary, and Olha Kanishcheva. 2025. A survey of diversity quantification in natural language processing: The why, what, where and how. *Preprint*, arXiv:2507.20858.

Gaston Gross. 1988. Degré de figement des noms composés. *Langages*, 90:57–72.

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

Maurice Gross and Jean Senellart. 1998. Nouvelles bases statistiques pour les mots du français. In *Proceedings of JADT'98, Nice 1998*, pages 335–349.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. 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.

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. SemEval-2013 task 4: Free paraphrases of noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume*

2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 138–143, Atlanta, Georgia, 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.

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.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. 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.

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.

Adam Lion-Bouton, Yagmur Ozturk, Agata Savary, and Jean-Yves Antoine. 2022. Evaluating diversity of multiword expressions in annotated text. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3285–3295, Gyeongju, Republic of Korea. International Committee on 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.

Stella Markantonatou, Carlos Ramisch, Victoria Rosén, Mike Rosner, Manfred Sailer, Agata Savary, and Veronika Vincze. 2021. PMWE conventions for examples containing multiword expressions. Technical report, Phraseology and Multiword Expressions – book series at Language Science Press.

Igor Mel'čuk. 2010. La phraséologie en langue, en dictionnaire et en TALN. In *Actes de la 17ème Conférence sur le Traitement Automatique des Langues Naturelles 2010*, Montréal, Canada.

Irina Moise and Sergiu Nisioi. 2026. MorphoFiltered-Gemini at PARSEME 2.0 Subtask 1: Tackling LLM Overgeneration via Universal POS-based Constraints. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

Florian Mueller. 2025. First copyright ruling against OpenAI worldwide: music rights collecting society wins German injunction over song lyrics —to be appealed now. Accessed on 01.01.2026.

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.

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.

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.

Pedro Ramaciotti Morales, Robin Lamarche-Perrin, Raphaël Fournier-S'Niehotta, Rémy Poulain, Lionel Tabourier, and Fabien Tarissan. 2021. Measuring diversity in heterogeneous information networks. *Theoretical Computer Science*, 859:80–115. Publisher: Elsevier.

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, and 6 others. 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.

Carlos Ramisch, Abigail Walsh, Thomas Blanchard, and Shiva Taslimipoor. 2023. A survey of MWE identification experiments: The devil is in the details. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 106–120, Dubrovnik, Croatia. Association for Computational Linguistics.

Rares-Alexandru Roscan and Sergiu Nisioi. 2026. Archaeology at PARSEME 2.0 Subtasks 1 and 2: Parsing is for Encoders, Paraphrasing is for LLMs. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

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, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, and 9 others. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, and 3 others. 2018. PARSEME multilingual corpus of verbal multiword expressions. 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 87–147. Language Science Press., Berlin.

Agata Savary, Simon Petitjean, Timm Lichte, Laura Kallemeyer, and Jakub Waszczuk. 2020. Object-oriented lexical encoding of multiword expressions: Short and sweet. *Lexique*, 27:87–120.

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, and 9 others. 2026. PARSEME 2.0 multilingual corpus of multiword expressions. In *Submitted to LREC 2026, under review.*

Agata Savary, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz, and Alina Wróblewska. 2024. UniDive: A COST action on universality, diversity and idiosyncrasy in language technology. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 372–382, Torino, Italia. ELRA and ICCL.

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.

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.

Benjamin Smith and J. Bastow Wilson. 1996. A Consumer's Guide to Evenness Indices. *Oikos*, 76(1):70–82. Number: 1 Publisher: [Nordic Society Oikos, Wiley].

Andy Stirling. 2007. A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4(15):707–719. Number: 15 Publisher: Royal Society.

Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. BERTScore is unfair: On social bias in language model-based metrics for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739, Abu Dhabi, United Arab Emirates. 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.

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.

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.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

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.

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.

A Detailed shared task results

This appendix presents detailed results for subtasks 1 and 2: cross-lingual macro-averages per phenomenon (subtask 1) and per-language results, including diversity scores per language.

In the phenomenon-specific rankings of subtask 1, a MWE is considered seen if a MWE with the same multi-set of lemmas was annotated at least once in the training corpus or in the development corpus. This definition impacts four MWE-based evaluation metrics and rankings: unseen-in-traindev, seen-in-traindev, variant-of-traindev and identical-to-traindev.

Please, interpret cross-lingual macro-averages carefully, as some scores depend on the dataset size, and the size of the underlying datasets varies across languages. These results are also published on the shared task git repository:

- Subtask 1: `https://gitlab.com/parseme/sharedtask-data/-/blob/master/2.0/subtask1/Detailed_results.md`
- Subtask 2: `https://gitlab.com/parseme/sharedtask-data/-/blob/master/2.0/subtask2/Detailed_results.md`

		Performance							
System	#Langs	Discontinuous MWE-based				Continuous MWE-based			
		P	R	F1	Rank	P	R	F1	Rank
MTLB-STRUCT	17/17	45.57	32.46	37.91	1	63.35	55.25	59.02	1
BeeParser	6/17	20.62	21.31	20.96	2	27.11	26.18	26.64	4
bert-multilingual-trial	6/17	15.83	15.46	15.64	3	22.00	20.58	21.27	6
IPN	17/17	12.70	20.01	15.54	4	22.27	45.53	29.91	3
baseline-gpt-oss-120b	17/17	12.68	4.08	6.17	5	17.42	40.32	24.33	5
romanian-bert	1/17	4.62	3.75	4.14	6	5.46	4.92	5.18	8
pmi-mwe-scorer	16/17	0.03	0.10	0.05	7	1.71	2.98	2.17	10
MorphoFiltered-Gemini	17/17	0.00	0.00	0.00	8	20.95	16.93	18.73	7
Pattern-Based-MWE-Id.	16/17	0.00	0.00	0.00	8	2.25	14.56	3.90	9
Sahara-Tokenizers	17/17	0.00	0.00	0.00	8	45.78	59.63	51.80	2

Table 4: Subtask 1 results – phenomenon-specific scores for discontinuous vs. continuous. Macro-averaged MWE-based (exact match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks.

		Performance							
System	#Langs	Unseen-in-traindev MWE-based				Seen-in-traindev MWE-based			
		P	R	F1	Rank	P	R	F1	Rank
MTLB-STRUCT	17/17	26.72	21.38	23.75	1	84.25	71.74	77.49	1
Sahara-Tokenizers	17/17	16.68	24.92	19.98	2	83.57	65.91	73.70	2
IPN	17/17	9.62	36.14	15.20	3	76.16	44.71	56.34	3
baseline-gpt-oss-120b	17/17	7.99	29.15	12.54	4	76.72	35.23	48.29	4
BeeParser	6/17	11.13	14.01	12.41	5	34.13	29.64	31.73	5
MorphoFiltered-Gemini	17/17	9.83	11.18	10.46	6	73.08	14.86	24.70	8
bert-multilingual-trial	6/17	8.59	9.21	8.89	7	27.93	24.72	26.23	7
romanian-bert	1/17	0.74	1.31	0.95	8	5.71	4.89	5.27	9
pmi-mwe-scorer	16/17	0.54	3.72	0.94	9	61.66	2.15	4.16	10
Pattern-Based-MWE-Id.	16/17	0.26	2.72	0.47	10	60.75	19.98	30.07	6

Table 5: Subtask 1 results – phenomenon-specific scores for unseen-in-traindev vs. seen-in-traindev. Macro-averaged MWE-based (exact match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks.

		Performance							
System	#Langs	Variant-of-traindev MWE-based				Identical-to-traindev MWE-based			
		P	R	F1	Rank	P	R	F1	Rank
MTLB-STRUCT	17/17	79.35	53.68	64.04	1	85.22	77.84	81.36	1
Sahara-Tokenizers	17/17	77.16	40.15	52.82	2	85.01	76.55	80.56	2
IPN	17/17	61.65	37.94	46.97	3	82.56	46.83	59.76	3
baseline-gpt-oss-120b	17/17	58.31	21.04	30.92	4	80.81	41.22	54.59	4
BeeParser	6/17	32.84	26.52	29.34	5	34.52	31.05	32.69	6
bert-multilingual-trial	6/17	27.26	21.05	23.76	6	28.19	26.40	27.27	7
MorphoFiltered-Gemini	17/17	54.54	12.00	19.67	7	75.06	15.91	26.25	8
Pattern-Based-MWE-Id.	16/17	40.13	12.73	19.33	8	72.43	22.54	34.38	5
romanian-bert	1/17	5.16	3.31	4.03	9	5.77	5.13	5.43	9
pmi-mwe-scorer	16/17	43.91	1.62	3.12	10	72.46	2.27	4.40	10

Table 6: Subtask 1 results – phenomenon-specific scores for variant-of-traindev vs. identical-to-traindev. Macro-averaged MWE-based (exact match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks.

		Performance							
System	#Langs	Single-token MWE-based				Multi-token MWE-based			
		P	R	F1	Rank	P	R	F1	Rank
MTLB-STRUCT	17/17	24.89	49.32	33.08	1	64.93	51.30	57.32	1
IPN	17/17	15.49	35.60	21.59	2	20.44	42.37	27.58	3
Sahara-Tokenizers	17/17	12.53	52.91	20.26	3	62.06	49.96	55.36	2
BeeParser	6/17	12.64	19.79	15.43	4	27.29	24.81	25.99	4
bert-multilingual-trial	6/17	8.77	20.04	12.20	5	22.63	19.05	20.69	6
baseline-gpt-oss-120b	17/17	10.68	8.81	9.66	6	16.99	38.03	23.49	5
MorphoFiltered-Gemini	17/17	6.13	2.14	3.17	7	21.56	15.99	18.36	7
pmi-mwe-scorer	16/17	1.38	12.17	2.48	8	0.66	1.57	0.93	10
Pattern-Based-MWE-Id.	16/17	0.98	11.83	1.81	9	6.58	11.52	8.38	8
romanian-bert	1/17	0.49	1.96	0.78	10	5.49	4.77	5.10	9

Table 7: Subtask 1 results – phenomenon-specific scores for single-token vs. multi-token. Macro-averaged MWE-based (exact match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks.

	Performance								Diversity (of identified MWEs)					
System	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
Sahara-Tokenizers	33.67	13.20	18.97	1	47.40	17.45	25.51	2	2.53	4	17.00	3	0.89	3
MTLB-STRUCT	31.07	12.80	18.13	2	45.11	15.92	23.53	3	2.45	5	16.00	4	0.88	4
MorphoFiltered-Gemini	12.00	3.60	5.54	3	15.79	4.60	7.13	6	2.55	2	14.00	5	0.97	1
baseline-gpt-oss-120b	3.60	6.20	4.56	4	7.89	13.71	10.02	5	2.54	3	18.00	2	0.88	4
IPN	2.53	8.00	3.85	5	10.36	32.02	15.65	4	2.99	1	24.00	1	0.94	2
Pattern-Based-MWE-Id.	1.98	8.20	3.19	6	19.28	40.46	26.11	1	2.03	6	11.00	6	0.85	5

Table 8: Egyptian (egy) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = -0.03$, $\rho = -0.03$); Richness: ($r = -0.03$, $\rho = 0.09$); SE: ($r = -0.21$, $\rho = 0.32$).

	Performance								Diversity (of identified MWEs)					
System	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
MTLB-STRUCT	67.70	39.40	49.81	1	74.10	37.02	49.38	1	3.92	3	103.00	4	0.85	5
Sahara-Tokenizers	36.86	37.60	37.23	2	55.89	40.93	47.25	2	3.91	4	104.00	3	0.84	6
IPN	15.77	35.80	21.90	3	21.97	56.06	31.57	4	4.51	2	130.00	2	0.93	3
MorphoFiltered-Gemini	20.34	19.40	19.86	4	32.69	34.58	33.61	3	4.51	2	93.00	5	1.00	1
baseline-gpt-oss-120b	12.76	36.00	18.84	5	17.03	52.40	25.70	5	4.74	1	139.00	1	0.96	2
pmi-mwe-scorer	0.47	2.40	0.78	6	6.17	23.43	9.77	7	2.48	5	12.00	6	1.00	1
Pattern-Based-MWE-Id.	0.28	2.80	0.50	7	6.86	38.57	11.64	6	1.97	6	9.00	7	0.89	4

Table 9: Modern Greek (el) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.58$, $\rho = 0.34$); Richness: ($r = 0.66$, $\rho = 0.50$); SE: ($r = -0.63$, $\rho = -0.54$).

	Performance								Diversity (of identified MWEs)					
System	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
MTLB-STRUCT	83.57	82.07	82.81	1	87.97	84.88	86.40	1	5.65	1	320.00	1	0.98	2
BeeParser	79.64	78.69	79.16	2	86.51	84.99	85.75	2	5.62	2	307.00	2	0.98	2
Sahara-Tokenizers	70.67	77.29	73.83	3	83.77	83.68	83.73	3	5.61	3	306.00	3	0.98	2
IPN	37.52	41.04	39.20	4	46.32	57.94	51.48	4	5.14	4	180.00	4	0.99	1
baseline-gpt-oss-120b	32.00	30.28	31.12	5	42.48	53.23	47.25	5	4.69	5	121.00	5	0.98	2
MorphoFiltered-Gemini	41.54	22.51	29.20	6	57.97	39.43	46.94	6	4.58	6	102.00	6	0.99	1
Pattern-Based-MWE-Id.	8.19	16.33	10.91	7	33.26	52.03	40.58	7	4.11	7	65.00	7	0.98	2
pmi-mwe-scorer	4.67	5.58	5.09	8	14.65	16.43	15.49	8	3.23	8	26.00	8	0.99	1

Table 10: Persian (Farsi) (fa) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.93$, $\rho = 1.00$); Richness: ($r = 0.99$, $\rho = 1.00$); SE: ($r = -0.52$, $\rho = -0.51$).

	Performance								Diversity (of identified MWEs)					
System	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
MTLB-STRUCT	73.55	53.29	61.81	1	83.82	57.42	68.15	1	5.34	1	226.00	1	0.98	3
Sahara-Tokenizers	52.60	50.50	51.53	2	75.76	58.33	65.91	2	5.28	2	212.00	2	0.98	3
IPN	37.74	42.71	40.07	3	48.43	58.00	52.79	3	5.14	3	184.00	3	0.99	2
baseline-gpt-oss-120b	32.03	32.73	32.38	4	42.92	53.08	47.47	4	4.91	4	144.00	4	0.99	2
MorphoFiltered-Gemini	14.17	6.99	9.36	5	35.36	18.42	24.22	5	3.26	5	29.00	5	0.97	4
Pattern-Based-MWE-Id.	1.11	4.79	1.80	6	15.22	51.08	23.46	6	3.06	6	22.00	6	0.99	2
pmi-mwe-scorer	0.48	1.00	0.64	7	14.64	22.83	17.84	7	1.61	7	5.00	7	1.00	1

Table 11: French (fr) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.92$, $\rho = 1.00$); Richness: ($r = 0.99$, $\rho = 1.00$); SE: ($r = -0.32$, $\rho = -0.54$).

	Performance								Diversity (of identified MWEs)					
System	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
baseline-gpt-oss-120b	8.55	15.92	11.12	1	15.23	29.89	20.18	1	3.41	1	35.00	1	0.96	3
MorphoFiltered-Gemini	43.33	3.90	7.16	2	59.09	4.98	9.19	4	2.10	4	9.00	4	0.95	4
Sahara-Tokenizers	8.81	6.01	7.14	3	14.19	8.17	10.37	3	2.45	3	14.00	3	0.93	5
IPN	3.95	6.91	5.02	4	7.42	20.82	10.94	2	3.01	2	21.00	2	0.99	1
MTLB-STRUCT	3.31	1.20	1.76	5	9.87	3.83	5.52	6	0.56	6	2.00	6	0.81	6
pmi-mwe-scorer	0.74	1.80	1.05	6	5.45	12.01	7.49	5	1.56	5	5.00	5	0.97	2

Table 12: Ancient Greek (grc) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.81$, $\rho = 0.71$); Richness: ($r = 0.83$, $\rho = 0.71$); SE: ($r = 0.36$, $\rho = -0.09$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
MTLB-STRUCT	67.38	62.67	64.94	1	69.68	61.73	65.47	1	5.26	1	226.00	1	0.97	3
Sahara-Tokenizers	52.89	60.28	56.34	2	57.93	60.23	59.06	2	5.20	2	214.00	2	0.97	3
IPN	14.36	40.52	21.20	3	15.75	46.56	23.54	4	5.02	3	170.00	3	0.98	2
baseline-gpt-oss-120b	14.64	34.53	20.56	4	17.53	42.51	24.82	3	4.79	4	135.00	4	0.98	2
MorphoFiltered-Gemini	12.42	3.79	5.81	5	17.67	4.71	7.44	7	2.80	6	17.00	6	0.99	1
Pattern-Based-MWE-Id.	1.28	10.78	2.29	6	9.24	54.85	15.81	5	3.71	5	46.00	5	0.97	3
pmi-mwe-scorer	0.22	2.20	0.39	7	4.25	42.13	7.72	6	2.27	7	10.00	7	0.99	1

Table 13: Hebrew (he) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.80$, $\rho = 0.96$); Richness: ($r = 0.92$, $\rho = 0.96$); SE: ($r = -0.66$, $\rho = -0.59$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
Sahara-Tokenizers	75.92	70.00	72.84	1	79.03	68.41	73.34	1	5.50	1	281.00	1	0.97	3
MTLB-STRUCT	80.36	63.00	70.63	2	84.32	58.44	69.03	3	5.40	2	252.00	3	0.98	2
bert-multilingual-trial	74.64	63.00	68.33	3	82.30	61.25	70.23	2	5.39	3	253.00	2	0.97	3
BeeParser	69.25	59.00	63.71	4	78.14	58.06	66.62	4	5.31	4	235.00	4	0.97	3
IPN	31.05	46.20	37.14	5	32.89	57.16	41.76	5	5.19	5	200.00	5	0.98	2
baseline-gpt-oss-120b	20.97	26.80	23.53	6	29.60	55.37	38.57	6	4.85	6	129.00	6	1.00	1
Pattern-Based-MWE-Id.	6.77	18.20	9.86	7	21.67	48.21	29.90	7	4.01	7	64.00	7	0.96	4
MorphoFiltered-Gemini	14.22	6.20	8.64	8	32.44	24.81	28.12	8	3.39	8	30.00	8	1.00	1
pmi-mwe-scorer	3.58	4.00	3.78	9	15.58	28.77	20.22	9	2.83	9	18.00	9	0.98	2

Table 14: Japanese (ja) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.89$, $\rho = 1.00$); Richness: ($r = 0.98$, $\rho = 0.98$); SE: ($r = -0.38$, $\rho = -0.36$).

System	Performance								Diversity (of identified MWEs)					
	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
MTLB-STRUCT	39.08	63.00	48.24	1	42.18	63.29	50.62	1	3.05	5	61.00	3	0.74	6
Sahara-Tokenizers	26.17	69.40	38.01	2	28.88	70.14	40.91	2	3.27	3	76.00	1	0.76	5
MorphoFiltered-Gemini	2.64	34.00	4.90	3	3.00	35.96	5.54	3	3.40	1	62.00	2	0.82	2
baseline-gpt-oss-120b	0.68	36.20	1.33	4	0.67	40.85	1.32	5	3.32	2	60.00	4	0.81	3
Pattern-Based-MWE-Id.	0.42	36.40	0.83	5	1.40	61.78	2.73	4	2.24	6	33.00	6	0.64	7
IPN	0.29	26.00	0.57	6	0.40	36.24	0.79	6	3.23	4	56.00	5	0.80	4
pmi-mwe-scorer	0.01	1.80	0.02	7	0.16	26.38	0.31	7	2.20	7	9.00	7	1.00	1

Table 15: Georgian (ka) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.29$, $\rho = 0.46$); Richness: ($r = 0.53$, $\rho = 0.86$); SE: ($r = -0.30$, $\rho = -0.43$).

	Performance								Diversity (of identified MWEs)					
System	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
MTLB-STRUCT	81.89	62.53	70.91	1	87.69	61.71	72.45	1	4.59	3	151.00	2	0.91	3
bert-multilingual-trial	73.27	61.52	66.88	2	84.79	62.99	72.28	2	4.57	4	149.00	3	0.91	3
Sahara-Tokenizers	64.69	59.12	61.78	3	75.25	60.71	67.20	3	4.52	5	143.00	4	0.91	3
baseline-gpt-oss-120b	12.39	58.52	20.45	4	13.53	67.82	22.57	5	4.66	2	163.00	1	0.91	3
MorphoFiltered-Gemini	23.40	17.64	20.11	5	29.87	22.79	25.85	4	4.03	6	67.00	6	0.96	2
IPN	7.00	40.68	11.94	6	9.52	57.61	16.34	7	4.79	1	149.00	3	0.96	2
Pattern-Based-MWE-Id.	2.58	28.46	4.73	7	9.45	63.90	16.46	6	3.98	7	79.00	5	0.91	3
pmi-mwe-scorer	0.37	3.21	0.67	8	3.25	27.62	5.82	8	2.69	8	15.00	7	0.99	1

Table 16: Latvian (lv) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.54$, $\rho = 0.50$); Richness: ($r = 0.62$, $\rho = 0.67$); SE: ($r = -0.62$, $\rho = -0.67$).

	Performance								Diversity (of identified MWEs)					
System	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
IPN	24.81	30.41	27.33	1	38.61	53.20	44.75	2	4.78	1	123.00	1	0.99	2
baseline-gpt-oss-120b	23.94	24.65	24.29	2	38.13	50.38	43.41	3	4.48	2	94.00	3	0.99	2
MTLB-STRUCT	36.79	17.97	24.15	3	67.86	20.63	31.64	5	4.11	3	68.00	4	0.98	3
Sahara-Tokenizers	19.02	26.04	21.98	4	51.29	41.04	45.60	1	4.48	2	95.00	2	0.98	3
MorphoFiltered-Gemini	20.65	8.76	12.30	5	57.97	27.25	37.08	4	3.64	4	38.00	5	1.00	1
Pattern-Based-MWE-Id.	1.88	4.38	2.63	6	18.35	24.43	20.96	6	2.06	6	9.00	7	0.94	4
pmi-mwe-scorer	1.19	2.30	1.57	7	12.40	19.54	15.17	7	2.30	5	10.00	6	1.00	1

Table 17: Dutch (nl) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.97$, $\rho = 0.90$); Richness: ($r = 0.95$, $\rho = 0.86$); SE: ($r = 0.27$, $\rho = -0.11$).

	Performance								Diversity (of identified MWEs)					
System	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
bert-multilingual-trial	83.33	84.00	83.67	1	87.13	84.82	85.96	1	5.49	1	286.00	1	0.97	4
MTLB-STRUCT	83.65	79.80	81.68	2	86.86	80.93	83.79	3	5.42	3	268.00	3	0.97	4
BeeParser	80.51	82.60	81.54	3	84.86	84.54	84.70	2	5.44	2	276.00	2	0.97	4
Sahara-Tokenizers	51.84	64.80	57.60	4	75.36	74.57	74.96	4	5.21	4	220.00	5	0.97	4
IPN	28.58	72.20	40.95	5	29.74	79.60	43.30	5	5.42	3	260.00	4	0.98	3
baseline-gpt-oss-120b	18.15	38.80	24.73	6	25.22	61.76	35.82	6	4.84	5	146.00	6	0.97	4
MorphoFiltered-Gemini	22.07	18.80	20.30	7	35.22	30.65	32.78	7	4.32	6	80.00	7	0.99	2
Pattern-Based-MWE-Id.	2.83	17.40	4.87	8	14.63	70.40	24.23	8	4.17	7	70.00	8	0.98	3
pmi-mwe-scorer	0.29	1.40	0.48	9	5.50	20.87	8.70	9	1.95	8	7.00	9	1.00	1

Table 18: Polish (pl) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.78$, $\rho = 0.95$); Richness: ($r = 0.93$, $\rho = 0.97$); SE: ($r = -0.75$, $\rho = -0.82$).

	Performance								Diversity (of identified MWEs)					
System	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
Sahara-Tokenizers	40.00	49.60	44.29	1	50.51	53.67	52.04	1	4.73	3	147.00	3	0.95	4
MTLB-STRUCT	50.17	30.20	37.70	2	64.22	31.42	42.19	2	4.13	4	88.00	4	0.92	5
IPN	17.74	52.20	26.48	3	23.79	69.92	35.50	3	4.98	2	182.00	2	0.96	3
baseline-gpt-oss-120b	15.93	51.60	24.34	4	19.96	76.50	31.66	4	5.06	1	192.00	1	0.96	3
MorphoFiltered-Gemini	8.37	13.80	10.42	5	19.45	26.50	22.43	5	3.92	5	56.00	5	0.97	2
Pattern-Based-MWE-Id.	0.61	7.00	1.13	6	6.73	43.00	11.63	6	1.74	7	8.00	6	0.84	6
pmi-mwe-scorer	0.22	1.20	0.38	7	4.93	17.92	7.73	7	1.79	6	6.00	7	1.00	1

Table 19: Brazilian Portuguese (pt) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.80$, $\rho = 0.64$); Richness: ($r = 0.72$, $\rho = 0.68$); SE: ($r = 0.10$, $\rho = -0.41$).

	Performance								Diversity (of identified MWEs)					
System	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
romanian-bert	91.03	80.88	85.65	1	94.35	81.98	87.73	2	5.39	2	275.00	2	0.96	4
BeeParser	84.98	82.27	83.60	2	92.36	85.19	88.63	1	5.43	1	280.00	1	0.96	4
MTLB-STRUCT	83.71	80.88	82.27	3	91.97	83.76	87.68	3	5.39	2	272.00	3	0.96	4
Sahara-Tokenizers	61.98	71.12	66.23	4	84.06	79.48	81.71	4	5.23	3	236.00	4	0.96	4
IPN	37.12	50.80	42.89	5	42.83	62.09	50.69	5	5.05	4	187.00	5	0.97	3
baseline-gpt-oss-120b	26.09	38.25	31.02	6	34.87	60.84	44.33	6	4.82	5	145.00	6	0.97	3
MorphoFiltered-Gemini	27.02	15.34	19.57	7	42.82	26.58	32.80	7	4.17	6	68.00	7	0.99	2
Pattern-Based-MWE-Id.	3.06	14.34	5.05	8	18.15	68.69	28.71	8	3.77	7	54.00	8	0.95	5
pmi-mwe-scorer	0.12	0.40	0.19	9	7.63	22.48	11.39	9	0.69	8	2.00	9	1.00	1

Table 20: Romanian (ro) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.78$, $\rho = 0.97$); Richness: ($r = 0.98$, $\rho = 0.98$); SE: ($r = -0.55$, $\rho = -0.46$).

	Performance								Diversity (of identified MWEs)					
System	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
MTLB-STRUCT	76.98	68.06	72.25	1	83.88	70.84	76.81	2	5.00	3	193.00	3	0.95	4
bert-multilingual-trial	71.15	72.85	71.99	2	80.95	77.00	78.93	1	5.06	2	205.00	1	0.95	4
Sahara-Tokenizers	44.72	53.29	48.63	3	73.07	64.40	68.46	3	4.84	4	159.00	4	0.95	4
IPN	21.70	62.67	32.24	4	24.41	72.95	36.58	4	5.10	1	202.00	2	0.96	3
baseline-gpt-oss-120b	14.67	42.71	21.84	5	18.12	58.60	27.68	6	4.74	5	138.00	5	0.96	3
MorphoFiltered-Gemini	24.67	14.97	18.63	6	38.95	23.83	29.57	5	4.02	6	61.00	6	0.98	2
Pattern-Based-MWE-Id.	1.51	12.38	2.69	7	10.88	65.13	18.65	7	3.58	7	41.00	7	0.96	3
pmi-mwe-scorer	0.09	0.60	0.16	8	3.33	17.30	5.58	8	1.10	8	3.00	8	1.00	1

Table 21: Slovenian (sl) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.71$, $\rho = 0.83$); Richness: ($r = 0.86$, $\rho = 0.88$); SE: ($r = -0.73$, $\rho = -0.90$).

	Performance								Diversity (of identified MWEs)					
System	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
BeeParser	72.87	76.80	74.78	1	79.00	79.38	79.19	1	5.62	1	312.00	1	0.98	2
MTLB-STRUCT	71.89	76.20	73.98	2	76.35	77.58	76.96	2	5.61	2	308.00	2	0.98	2
Sahara-Tokenizers	53.46	63.40	58.01	3	70.03	68.97	69.49	3	5.43	3	258.00	3	0.98	2
IPN	35.99	57.80	44.36	4	37.88	64.90	47.84	4	5.40	4	247.00	4	0.98	2
baseline-gpt-oss-120b	20.36	34.00	25.47	5	27.00	52.32	35.62	6	4.91	5	147.00	5	0.98	2
MorphoFiltered-Gemini	19.70	8.00	11.38	6	30.62	13.53	18.77	8	3.45	8	34.00	7	0.98	2
pmi-mwe-scorer	2.08	7.60	3.27	7	11.12	37.56	17.16	9	3.60	6	37.00	6	1.00	1
Pattern-Based-MWE-Id.	1.84	9.40	3.08	8	13.96	61.49	22.76	7	3.49	7	37.00	6	0.97	3
bert-multilingual-trial	0.35	0.60	0.44	9	44.36	38.69	41.33	5	1.10	9	3.00	8	1.00	1

Table 22: Serbian (sr) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.84$, $\rho = 0.95$); Richness: ($r = 0.98$, $\rho = 0.95$); SE: ($r = -0.38$, $\rho = -0.27$).

	Performance								Diversity (of identified MWEs)					
System	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
bert-multilingual-trial	65.97	63.05	64.48	1	75.72	70.65	73.09	1	5.14	1	207.00	1	0.96	3
MTLB-STRUCT	67.19	60.04	63.41	2	73.96	67.52	70.60	3	5.13	2	204.00	2	0.97	2
BeeParser	65.35	59.84	62.47	3	76.62	67.30	71.66	2	5.14	1	204.00	2	0.97	2
Sahara-Tokenizers	47.18	55.42	50.97	4	64.89	66.41	65.64	4	5.07	3	191.00	3	0.97	2
IPN	21.16	45.38	28.86	5	24.65	64.40	35.65	5	5.01	4	176.00	4	0.97	2
MorphoFiltered-Gemini	20.17	34.14	25.35	6	26.00	50.56	34.34	6	4.72	6	131.00	6	0.97	2
baseline-gpt-oss-120b	17.27	39.96	24.12	7	21.90	69.87	33.34	7	4.81	5	144.00	5	0.97	2
Pattern-Based-MWE-Id.	1.53	9.04	2.62	8	11.10	49.33	18.12	8	3.47	7	36.00	7	0.97	2
pmi-mwe-scorer	1.39	6.22	2.27	9	4.99	22.99	8.21	9	3.34	8	29.00	8	0.99	1

Table 23: Swedish (sv) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.86$, $\rho = 0.95$); Richness: ($r = 0.93$, $\rho = 0.98$); SE: ($r = -0.61$, $\rho = -0.73$).

	Performance								Diversity (of identified MWEs)					
System	Global MWE-based				Global Token-based				Var.-bal.		Variety		Balance	
	P	R	F	Rank	P	R	F	Rank	SWE	Rank	Richness	Rank	SE	Rank
MTLB-STRUCT	59.29	49.50	53.95	1	69.44	53.83	60.65	1	5.05	3	179.00	3	0.97	3
Sahara-Tokenizers	37.60	45.53	41.19	2	58.67	54.53	56.52	2	5.04	4	174.00	4	0.98	2
IPN	25.97	60.04	36.25	3	32.39	73.69	45.00	3	5.39	1	241.00	1	0.98	2
baseline-gpt-oss-120b	22.43	45.53	30.05	4	28.95	60.63	39.19	4	5.13	2	185.00	2	0.98	2
MorphoFiltered-Gemini	29.37	14.71	19.60	5	45.54	22.21	29.86	5	4.23	5	70.00	5	1.00	1
Pattern-Based-MWE-Id.	2.44	15.90	4.23	6	13.35	57.84	21.70	6	4.08	6	65.00	6	0.98	2
pmi-mwe-scorer	0.54	2.39	0.88	7	7.50	26.92	11.73	7	2.48	7	12.00	7	1.00	1

Table 24: Ukrainian (uk) subtask 1 results – MWE-based (exact match) and token-based (partial match) scores: precision (P), recall (R), F-score (F1) and F1-based ranks. SWE: Shannon-Weaver entropy, SE: Shannon evenness – Pearson (r) and Spearman (ρ) correlation between global MWE-based F-score and SWE: ($r = 0.82$, $\rho = 0.71$); Richness: ($r = 0.84$, $\rho = 0.71$); SE: ($r = -0.70$, $\rho = -0.78$).

	SWE r	SWE ρ	Richness r	Richness ρ	SE r	SE ρ
Egyptian (EGY)	-0.03	-0.03	-0.03	0.09	-0.21	0.32
Modern Greek (EL)	0.58	0.34	0.66	0.50	-0.63	-0.54
Persian (Farsi) (FA)	0.93	1.00	0.99	1.00	-0.52	-0.51
French (FR)	0.92	1.00	0.99	1.00	-0.32	-0.54
Ancient Greek (GRC)	0.81	0.71	0.83	0.71	0.36	-0.09
Hebrew (HE)	0.80	0.96	0.92	0.96	-0.66	-0.59
Japanese (JA)	0.89	1.00	0.98	0.98	-0.38	-0.36
Georgian (KA)	0.29	0.46	0.53	0.86	-0.30	-0.43
Latvian (LV)	0.54	0.50	0.62	0.67	-0.62	-0.67
Dutch (NL)	0.97	0.90	0.95	0.86	0.27	-0.11
Polish (PL)	0.78	0.95	0.93	0.97	-0.75	-0.82
Brazilian Portuguese (PT)	0.80	0.64	0.72	0.68	0.10	-0.41
Romanian (RO)	0.78	0.97	0.98	0.98	-0.55	-0.46
Slovenian (SL)	0.71	0.83	0.86	0.88	-0.73	-0.90
Serbian (SR)	0.84	0.95	0.98	0.95	-0.38	-0.27
Swedish (SV)	0.86	0.95	0.93	0.98	-0.61	-0.73
Ukrainian (UK)	0.82	0.71	0.84	0.71	-0.70	-0.78
Mean	0.72	0.76	0.81	0.81	-0.39	-0.46

Table 25: Subtask 1 – Pearson (r) and Spearman (ρ) correlations between Global MWE-based F1 score and diversity of systems' true positives – SWE: Shannon-Weaver entropy, SE: Shannon evenness.

	Performance				Diversity (of new vocabulary)					
System	Automatic eval.		Manual eval.		Variety		Balance		Var.-bal.	
	gmBS	Rank	Manual score	Rank	Richness	Rank	SE	Rank	SWE	Rank
Star-Paraphraser-Cosine	93.90	1	64.82	3	236.00	4	0.83	4	4.54	4
Star-Paraphraser-Multiagent	89.46	2	79.25	1	456.00	2	0.90	3	5.48	2
baseline-gpt-oss-120b	77.55	3	72.70	2	326.00	3	0.92	2	5.33	3
MISP	49.53	4	29.25	4	564.00	1	0.93	1	5.89	1

Table 26: French (fr) subtask 2 results – gm-BS: global masked BERT-score, SWE: Shannon-Weaver entropy, SE: Shannon evenness.

	Performance				Diversity (of new vocabulary)					
System	Automatic eval.		Manual eval.		Variety		Balance		Var.-bal.	
	gmBS	Rank	Manual score	Rank	Richness	Rank	SE	Rank	SWE	Rank
baseline-gpt-oss-120b	63.99	1	24.22	1	804.00	2	0.98	1	6.54	1
MISP	33.75	2	3.39	2	971.00	1	0.89	2	6.08	2

Table 27: Georgian (ka) subtask 2 results – gm-BS: global masked BERT-score, SWE: Shannon-Weaver entropy, SE: Shannon evenness.

	Performance				Diversity (of new vocabulary)					
System	Automatic eval.		Manual eval.		Variety		Balance		Var.-bal.	
	gmBS	Rank	Manual score	Rank	Richness	Rank	SE	Rank	SWE	Rank
baseline-gpt-oss-120b	80.21	1	55.88	1	619.00	2	0.92	2	5.93	2
MISP	58.59	2	38.23	2	798.00	1	0.93	1	6.20	1

Table 28: Brazilian Portuguese (pt) subtask 2 results – gm-BS: global masked BERT-score, SWE: Shannon-Weaver entropy, SE: Shannon evenness.

	Performance				Diversity (of new vocabulary)					
System	Automatic eval.		Manual eval.		Variety		Balance		Var.-bal.	
	gmBS	Rank	Manual score	Rank	Richness	Rank	SE	Rank	SWE	Rank
GPT-CREATIVE	89.25	1	77.31	1	235.00	3	0.98	1	5.36	3
baseline-gpt-oss-120b	74.74	2	46.17	2	742.00	2	0.93	2	6.14	2
MISP	57.01	3	22.66	3	1096.00	1	0.91	3	6.36	1

Table 29: Romanian (ro) subtask 2 results – gm-BS: global masked BERT-score, SWE: Shannon-Weaver entropy, SE: Shannon evenness.