PARSEME 2.0: a multilingual shared task proposal on identification, generation and understanding of multiword expressions

(12 November 2024)

Proposers: Manon Scholivet, Takuya Nakamura, Agata Savary, Carlos Ramisch,

PARSEME shared task series

PARSEME is an active research community which evolved from the homonymous COST action (2013-2017) focusing on **multiword expressions**. Multiword expressions (MWEs), are groups of words that must be treated as a unit at some processing level due to idiosyncratic composition (Baldwin & Kim, 2010). The **PARSEME shared task** series proposed an influential framework for the task of **MWE identification**. In this task, systems must identify tokens that belong to MWEs in running text.

Three editions of the PARSEME shared tasks took place, organised in conjunction with the annual MWE workshop in 2017 (edition 1.0, Savary et al., 2017), 2018 (edition 1.1, Ramisch et al., 2018) and 2020 (edition 1.2, Ramisch et al., 2020). The goal of these shared tasks was to stimulate the development of **multilingual** systems for the automatic identification of **verbal** MWEs in text. They were based on corpora annotated for verbal MWEs in 14 to 20 languages, split into training, development and test data and released on CLARIN/LINDAT afterwards. In addition, the corpora were released in a standalone version (with no associated shared task) in 2023, with full UD compatibility, and containing extensions and improvements, for a total of 9.3M tokens, 456K sentences, and more than 127K annotated verbal MWEs across all 26 languages that participated in at least one of the previous shared task editions (Savary et al. 2023).

In addition to the corpora and shared tasks, the PARSEME community also produced evaluation metrics and scripts for the MWE identification task, and a variety of systems addressing the task, ranging from dictionary-based matching to pre-trained language models fine-tuned to the task. One particularly relevant aspect of PARSEME 1.2 shared task was the proposal of a corpus splitting strategy with a controlled rate of “unseen” MWEs, that is, MWEs present in the test data that were not present in the training data. The goal was to focus on the robustness of the systems, assessing whether they could generalise across a diverse range of MWEs, and not only those that were memorised from the training data.

Goals of this proposal

This shared task has several objectives. The first will obviously be to ***evaluate the systems' ability to predict MWEs***. However, participants will be strongly encouraged to take an interest in the diversity of their predictions, since in addition to the usual F-score, the results will be **evaluated on the diversity of the MWEs predicted** by their systems.

The aim of this evaluation measure is to invite the participants not to overlook the rare MWEs, as well as those that have not been seen in training. We will then have to work harder and more creatively to find ways of making systems generalise even more effectively!

In addition to this task, one or more sub-tasks designed to **assess the ability of Large Language Models (LLMs) to understand and/or predict MWEs** will be carried out. The handling of the MWEs by LLMs is still unclear, and these sub-tasks will help to gain a better understanding of the issue.

Ongoing work on MWE annotation in Unidive WG1 (task 1.2)

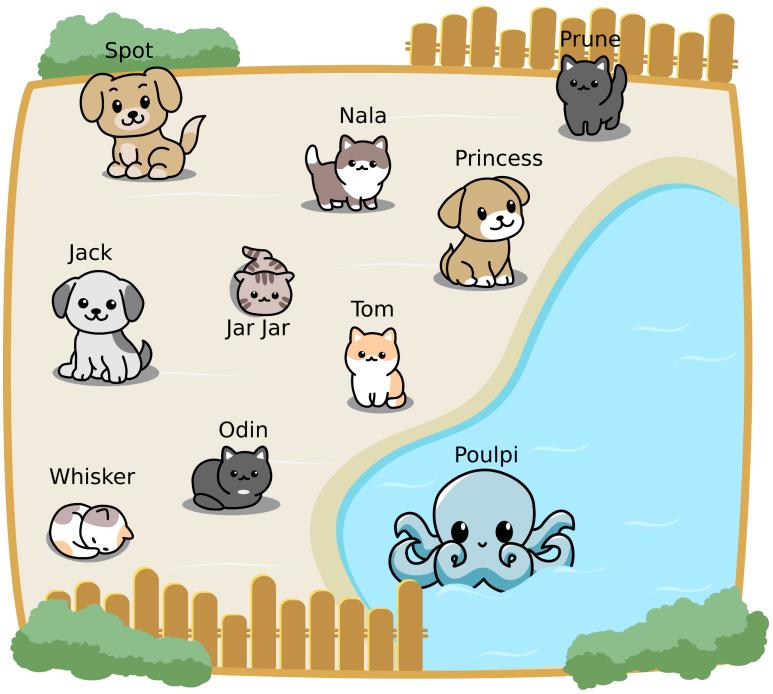
Unidive WG1 on corpus annotation features a task 1.2 on the annotation of multiword expressions of all categories **beyond verbal ones**. At the time of writing this proposal, task 1.2 has written guidelines for nominal, modifier and functional MWEs, covering a much more diverse range of phenomena. The guidelines are being discussed via gitlab issues, and a stable version is expected in the next few months. A pilot annotation is planned as part of the next Unidive General Meeting in Budapest in January 2025.

Once the guidelines will be ready, we will coordinate an annotation campaign covering up to 29 languages. The goal is to annotate or extend the current PARSEME corpora with MWEs of all syntactic categories using the newly developed guidelines. We expect that this effort will generate corpora annotated with 1-10K MWEs in each language, which could be used as training and evaluation data in the planned shared task.

Diversity evaluation

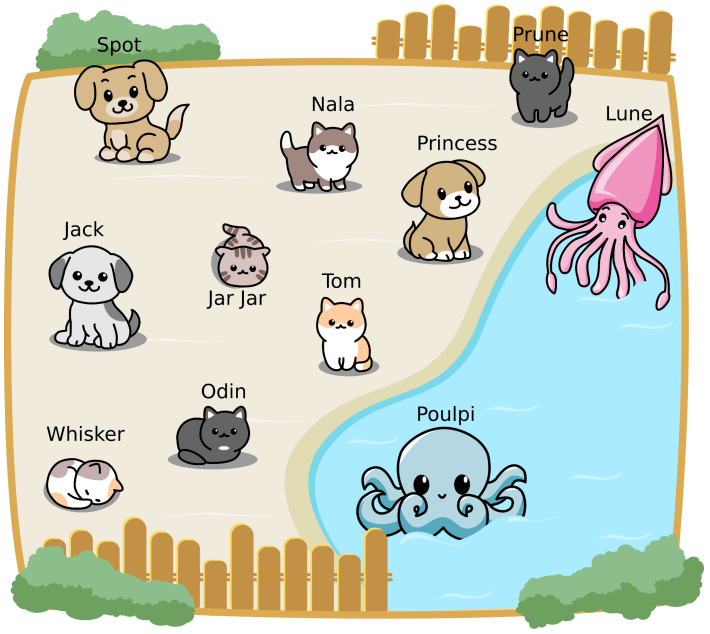
The question of the diversity of corpora and system predictions has been ignored for a long time. However, this issue seems essential in NLP in many contexts: a system training on a more diverse corpus will be able to generate more diverse phenomena. Also, a system that always predicts the most frequent phenomena, while ignoring rare phenomena, may have a good F-score. But do we really want to ignore rare phenomena?
This shared task will use diversity measures to evaluate the MWEs predicted by the systems. Inherited from the evaluation of unseen MWEs, the aim of these measures is to encourage participants to produce more diverse predictions (Lion-Bouton et al., 2022). But what exactly is the nature of these diversity measures?

The term **diversity** can refer to many different things. Here, we use a definition that is strongly inspired by the ecologists (Chao et al., 2014). Let's imagine, for example, that we have 3 *species* of animals on a piece of land near the sea (which we'll call BASE): cats, dogs and octopuses. There are *individuals* representing each *species* in the field. There are 3 dogs: Princess, Jack and Spot; 6 cats: Whisker, Tom, Nala, Prune, Odin and Jar Jar; and a single octopus: Poulpi.

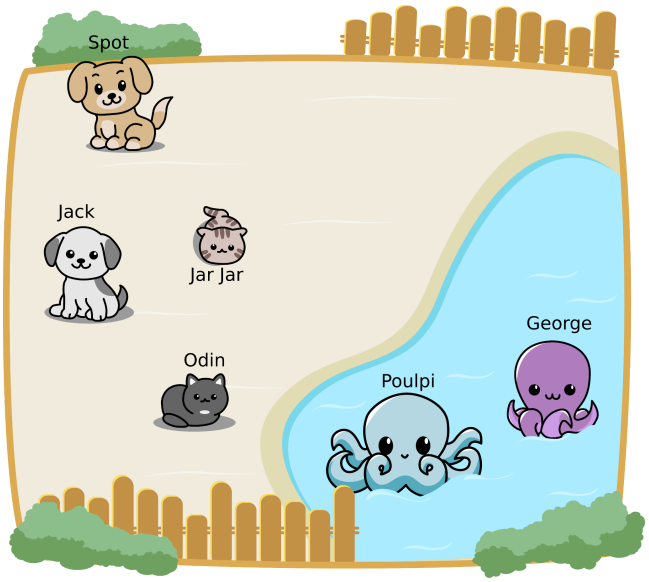


Three measures can be used to evaluate the diversity of this land:

- **Variety**
A variety measure is based on the number of species. The more species there are, the greater the variety. Here is an example of a plot with a greater variety than BASE, due to the presence of a new species, the squid Lune:



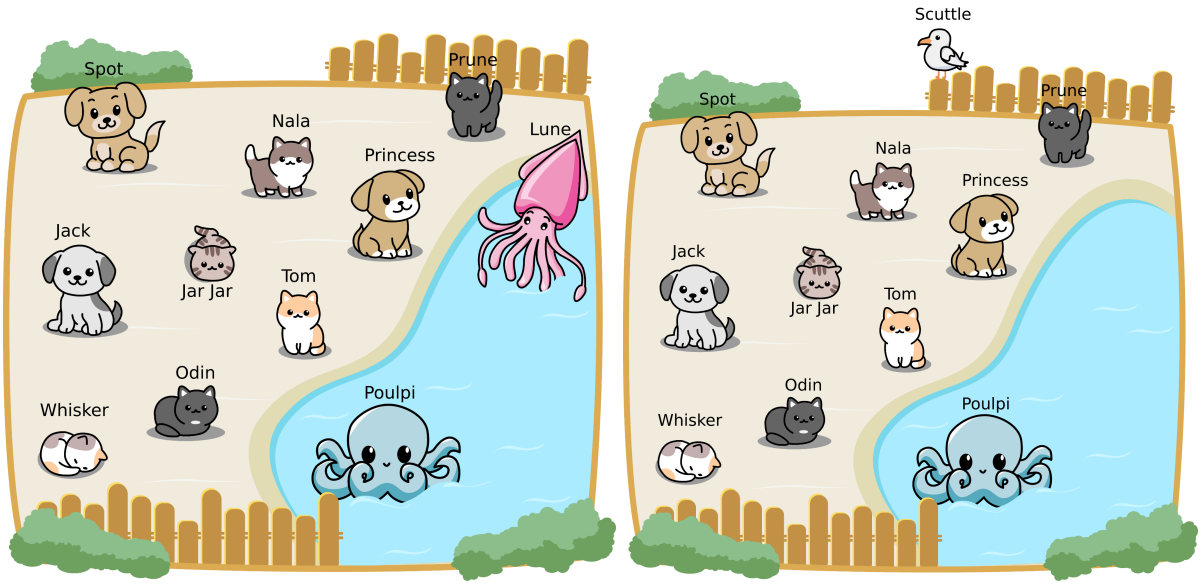
- **Balance**
A balance measure determines the extent to which the number of individuals of each species is even. Here is an example of a more balanced plot, where each species is represented by an equal number of *individuals*:



- **Disparity** (probably not used in this shared task)
A disparity measure estimates how related species are to each other. The more different the species, the greater the disparity.

In the first image below, we note the presence of the new individual, Lune, from the squid species. This species is very close to an existing one, the octopus. In the second image, the new individual, Scuttle, is a seagull. And it's the first bird in the picture! This species, which is very distant from all the others, increases the disparity more than Lune, the squid.

Disparity is a more unusual measure. It requires distance measurements between **Categories**, and further reflections are needed to find a potential efficient way to use it for MWEs. Therefore, for this shared task, we will mainly use measures of variety and balance. We will apply them to true positives only rather than to all predictions of a system, so as not to reward artificially diverse false positives.



In order to calculate the diversity, two elements are needed:
Species, which we will also call **Categories**, and individuals, which we will call **Elements** (Morales et al., 2021).

To move these considerations to the MWEs ground, let us consider the expression "*cry wolf*" (call for help when you don't need it) is much less frequent than the expression "*take place*", and may never have been encountered in the training corpus.

For this task, a possible *Category* would be the canonical form of a MWE, for example "*cry wolf*". The Elements (individuals) of this Category (species) would be the occurrences of this MWE. For example :

- "She ***cries wolf***"
- "Jar Jar often ***cries wolf***"
- "Jack has been ***crying wolf***"
- "You should never ***cry wolf***"

MWE identification systems are often good at identifying relatively frequent MWEs (e.g. "take place") but they miss many infrequent ones (e.g. "cry wolf"). In other words, their variety and balance are weak, despite good F-measure. Using diversity as an evaluation measure might encourage systems' authors to pay more attention to diverse although rare phenomena.

Diversity can also be quantified interlingually, when Elements are languages and Categories are language genera or families. We will experiment with measures that reward systems producing more diverse outcomes in more diverse languages, especially low-resourced languages.

Multiword expressions in LLMs

Since transformers and LLMs have revolutionised the world of NLP, we can look at their performance on MWE-related tasks (Miletić and Schulte im Walde, 2024). Two potential ideas of sub-tasks for evaluating them are submitted below for discussion with the community:

Idea 1: Idiomatic language generation by LLMs

The idea here is to test the capacity of LLMs to generate MWEs. Given a one paragraph context and a sentence truncated on the start of a MWE, what will the LLM do ?
Several levels of difficulty could be set up, depending on the compositionality of the MWE elements. For example :

- PROMPT: “It’s ***raining cats and*** …”
- PROMPT: “Jack has been ***crying*** …”

The first example is really often seen in the context of the MWE “raining cats and dogs”, while “crying” might be followed by “wolf”, but might as well be followed by “loudly”. The second one will probably be harder to predict than the first one.

The evaluation of this task would consist of checking whether or not the expected MWE was predicted. One advantage of this task is that the data can be extracted directly from the test data.

Idea 2: Idiomatic language comprehension by LLMs

Here, the idea is to assess to what extent LLMs are able to **understand** idiomatic expressions and distinguish them from literal or coincidental co-occurrence (Savary et al. 2019). This idea has already been explored on a smaller scale in a SemEval-2022 shared task (Madabushi et al. 2022). We will try to extend this idea to many languages and diverse MWEs in the context of PARSEME.

Therefore, we propose two adaptations of the traditional MWE identification task. First, we could prompt the model with a sentence containing a potentially idiomatic expression, and ask whether it should be interpreted literally or idiomatically, for example:

- PROMPT: In “[...] who is going to be ***pulling the strings*** in the Europe of tomorrow”, should we interpret “pulling the strings” as an idiom, it is the sentence literally about “pulling” and “strings”?
- PROMPT: In “[...] the artist was ***pulling the strings*** to move the marionette”, should we interpret “pulling the strings” as an idiom, it is the sentence literally about “pulling” and “strings”?

In this case, the data from the PARSEME corpora could be automatically converted into textual prompts, exploiting both MWE annotations and sentences containing non-annotated literal and coincidental co-occurrences. Evaluation would be accuracy of the response.

Second, we could ask the LLM to explain the meaning of the potentially idiomatic word combination. For example:

- PROMPT: In “[...] who is going to be ***pulling the strings*** in the Europe of tomorrow”, what is the meaning of “pulling the strings”?

In that case, data generation would not be 100% automatic, but would require the existence of some lexicon containing MWE definitions. Nonetheless, we could implement an automated procedure to generate definitions from an open lexicon (e.g. Wiktionary) and then manually select those prompt-reply pairs that are correct. Moreover, the similarity between gold and predicted definitions would need to be assessed using some imperfect metric (e.g. BERTSCore). Given the more expensive and exploratory nature of this variant, we would probably be able to provide less annotated examples for evaluation.

In all ideas described above, examples of the task could be provided in the prompt, either with the same lemmas or with different lemmas. The amount and nature of the provided context could provide different evaluation metrics that could distinguish LLMs based on their ability to learn and generalise from prompt context.

These exploratory sub-tasks might be applied to only a subset of languages.

Practicalities

- In order to ensure the long-term future of this shared task, the aim will be to use the CodaBench platform. This platform enables participants to update the leaderboard even after the shared task has ended, so that we always know who is at the state of the art on the different tasks.
- To ensure good visibility, this shared task will be submitted to SemEval 2026.
- Finally, this shared task is the first step towards a diversity benchmark. The aim of this benchmark will be to encourage the NLP community to take a closer look at the issue of diversity, and the importance of not neglecting rare phenomena in many tasks.

Tentative timeline

- 29 January: WG3 meeting presentation
- February 2025 : WG3 notification

The following dates depend on the WG3 notification :

- January - August 2025 : PARSEME corpus annotation for MWEs of all syntactic types
- Spring 2025 : SemEval 2026 shared task proposal
- Autumn 2025 : Training data publication
- Winter 2026 : Systems evaluation
- Summer 2026 : SemEval Workshop