Call for expressions of interest in PARSEME/UniDive annotation campaign on multiword expressions

The UniDive COST action (task 1.2) and the PARSEME community are happy to announce the upcoming **multilingual corpus annotation campaign** dedicated to multiword expressions (MWEs).

We call for expression of interest from current or future Language Leaders, who wish to propose a language team. If you are interested, please, fill in the EoI form, best before **27 February 2025**.

More details about the role of the Language Leader can be found in the PARSEME Language Leader guide.
Feel free to contact us for any questions you might have.

UniDive task 1.2 co-leaders
 Voula Giouli, Stella Markantonatou, Carlos Ramisch, Agata Savary, Sara Stymne

The current annotation campaign will cover MWEs of **all syntactic types**. It follows the spirit of **universality**. Namely, the annotation guidelines are unified across all participating languages, whenever possible, still leaving room for truly language-specific phenomena. This approach is expected to promote meaningful cross-language comparisons. The resulting corpus will be used in a **PARSEME/UniDive shared task** on identifying and understanding MWEs, to be proposed for SemEval 2026.

For the languages already present in the PARSEME corpus, the agenda is to:

- Re-annotate the existing corpus with MWEs other than verbal. Annotating only part of the existing corpus is an option. In this case we recommend a minimum of 3500 annotations (so that each selected text is exhaustively annotated for all syntactic types of MWEs). More number of annotations can do but the system results are expected not to be representative.
- Add some new texts annotated from scratch (to counterbalance language model contamination from previously published data)

For new languages, corpora will be annotated for all syntactic types at once.

A language team should consist of **at least 2 annotators** (including the Language Leader), for the sake of inter-annotator agreement estimation. It is possible to start annotating alone and recruit more annotators at a later stage (May at latest).

Centralized
https://gitlab.com/parseme/corpora/-/wikis/PARSEME-Language-Leader-Guidedocumentation and tools
(including the online FLAT annotation platform) are available.

We propose the following timeline:

- [language leaders: 27 February] Expression of interest from Language Leaders
- [task leaders: 1st March] Meeting, FLAT architect
- [language leaders: mid-March] Reading guidelines, reading the Language Leader guide, filling in MWE examples, terminology, selecting corpora
- [all: March] Pilot annotation
- [task leaders: 1 April] SEMEVAL shared task proposal
- [language teams: April-May] Annotation (including a double-annotated sample for inter-annotator agreement estimation)
- [SEMEVAL: 19 May] Notification about the selected shared task
- [language leaders: June] Consistency checks and inter-annotator agreement estimation
- [shared task leaders: 15 July] Sample data ready
- [task leaders: July-August] Consolidating and splitting the corpora
- [WG3 shared task leaders: 1 September] Training data for SEMEVAL

Three past PARSEME annotation campaigns were dedicated exclusively to verbal MWEs (VMWEs) and resulted in 4 editions of the PARSEME corpus, which jointly covers **26 languages**. Three PARSEME shared tasks on automatic identification of VMWEs have been organized on the basis of this corpus and set the state of the art in the task.