PARSEME/UniDive annotation campaign on multiword expressions

- **Event title**: PARSEME/UniDive annotation campaign (UniDive WG1 task 1.2)
- **Dates**: September 2023 – September 2025
- **Co-leaders**:
  - Voula Giouli, Aristotle University of Thessaloniki, Greece
  - Stella Markantonatou, Language and Speech Processing/ATHENA RC, Athens, Greece
  - Takuya Nakamura, Université Paris-Saclay, France
  - Carlos Ramisch, Aix-Marseille Université, France
  - Agata Savary, Université Paris-Saclay, France
  - Sara Stymne, Uppsala University, Sweden

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

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.

The current annotation campaign covers MWEs of **all syntactic types** (including nominal, adjectival, adverbial and functional MWEs). 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, submitted as a proposal for SemEval 2026.

Teams

Each language should be annotated by a team on **native annotators** (except when this is not possible, e.g. in the case of extinct languages like Ancient Greek or Egyptian). 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 2025 at latest). See the language teams from past and present annotation campaigns.

Each language team should have at least one **Language Leader**. See the call for Language Leaders.

Annotation work

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 2000 annotated MWEs** (so that each selected text is exhaustively annotated for all syntactic types of MWEs). A lower 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.

Conversions from other MWE annotation schemes are fine, if curated so as to fit the PARSEME guidelines.

Timeline

- ~~**[task leaders: 14 February]** Call for Language Leaders~~
- ~~**[language leaders: 27 February]** Expression of interest from Language Leaders~~
- ~~**[task leaders: late-February]** Creating FLAT accounts~~
- **[language leaders: mid-March]**
  - Reading the annotation guidelines 2.0
  - Reading the Language Leader's guide
  - Filling in MWE examples in the guidelines
  - Recruiting annotators
  - Selecting corpora
- ~~**[all: 28 March]** Pilot annotation, submitting issues~~
- ~~**[shared task leaders: 31 March]** SEMEVAL 2016 shared task proposal~~
- ~~**[SEMEVAL: 19 May]** Notification from SemEval about the selected shared tasks ⇒ rejected~~
- **[language teams: April-1 September]** Annotation for subtask 1 (PARSEME corpus)
  - Annotating the PARSEM Ecorpus with all syntactic types of MWEs
  - Double-annotating a sample for inter-annotator agreement estimation
  - Consistency checks
- **[task leaders: 15 September]** Preparing the data for subtask 2
- **[language teams: 1 October]** MWE paraphrasing for subtask 2
- **[task leaders: 30 October]** Consolidating and splitting the corpora for both subtasks
- **[task leaders: autumn]** Shared task proposal

Documents and tools

- PARSEME/UniDive annotation campaign master document
- PARSEME corpus wiki
- Annotation guidelines
  - PARSEME annotation guidelines 2.0
  - what's new in version 2.0
  - Gitlab issues from the guidelines
- Language Leader's guide

- FLAT annotation platform and FLAT User's Guide
- Minutes from task 1.2 co-leaders' meetings
- Minutes from Language Leaders' meetings

Language Leaders' meetings

Language Leaders meet weekly online during the annotation campaign. The timeline is the following:

- Tuesday 8 April 6 p.m. CEST
- Friday 18 April 9 a.m. CEST
- Friday 2 May 9 a.m. CEST
- Tuesday 6 May 6 p.m. CEST
- Friday 16 May 9 a.m. CEST
- Friday 30 May 9 a.m. CEST
- Tuesday 3 June 6 p.m. CEST
- Friday 13 June 9 a.m. CEST
- Friday 27 June ~~6~~ p.m. CEST
- ~~Tuesday 1 July 6 p.m. CEST~~
- Friday 25 July 9 a.m. CEST

We are using the recurrent zoom link.