

Working Group 1: Corpus Annotation

- **Leader:** [Bruno Guillaume](#) (France)
- **Vice-leader:** [Kaja Dobrovoljc](#) (Slovenia)

Workplan

Annotated corpora constitute the Action's major operational tools for NLP-applied universality. Therefore, WG1 will be dedicated to the following activities:

1. **Studies** and community **discussions** in language typology and language universals at the level of morphology, syntax and semantics, with special attention paid to idiosyncrasy at all these levels;
2. Unification and enhancement of cross-lingual annotation **guidelines** for morpho-syntax and MWEs:
 - defining the division of labour between morpho-syntactic and semantic annotation,
 - addressing hard or weakly covered syntactic phenomena (syntactically irregular structures, relative clauses, coordination, pronoun inclusivity, etc.),
 - covering new MWE categories (nominal, adjectival and functional MWEs),
 - paving the way for unified annotation guidelines for idiosyncratic constructions;
3. Coordinate the development and maintenance of centralized **software** for universality-based corpus construction:
 - online spaces for community discussion and editing annotation guidelines,
 - tools for automatic pre-annotation, annotation transfer and manual annotation of corpora,
 - tools for corpus merging, validation, curation, statistics, conversion and release. The software development itself will be funded at national levels;
4. Defining **file formats** for corpora annotated according to the unified guidelines;
5. Construction of annotated **corpora**:
 - adapting the existing corpora to the enhanced guidelines,
 - creating new annotated corpora following the enhanced guidelines.

Members and organisation

- [List](#) of current WG1 members
- Activities are currently structured around four primary [WG1 tasks](#) detailed below, but proposals for new activities are always welcome.

Upcoming meetings

- WG Meeting 8 (online) - 11 April 2024, 09:00 **CEST**
- WG Meeting 9 (online) - 11 June 2024, 13:30 **CEST**

Minutes of past meetings

- WG1 Meeting 7 (Naples, Italy) - 7 February 2024: co-located with the [2nd General Meeting](#) in

Naples on 8-9 Feb 2024, [\[short report\]](#)

- WG1 Meeting 6 (online) - 17 January 2024: Presentation of the WG1 activities in Naples [\[minutes\]](#)
- WG1 Meeting 5 (online) - 20 December 2023: Updates on WG1 tasks and discussion of the activities proposed for Naples [\[minutes\]](#)
- WG1 Meeting 4 (online) - 27 November 2023: Updates on WG1 wiki, WG1 task activities and general [\[minutes\]](#)
- WG1 Meeting 3 (online) - 25 October 2023: Updates on WG1 tasks activities [\[minutes\]](#)
- WG1 Meeting 2 (online) - 13 September 2023: launching WG1 tasks [\[minutes\]](#)
- WG1 Meeting 1 (Paris-Saclay University, France) - 16-17 March 2023: [brainstorming](#) topics and [slides](#) - co-located with [UniDive 1st general meeting](#)

WG1 Tasks

- **Task 1.1: Linguistic typology and multilingual corpus annotation**
 - [Minutes](#) from the task meetings
 - [Agenda](#) and [report](#) from the Naples 2024 meeting
- **Task 1.2 on MWE annotation guidelines and UD-PARSEME unification**
 - [Leaders / Contacts](#): Agata Savary, Voula Giouli, Stella Markanotatou, Sara Stymne, Carlos Ramisch
 - [Objectives](#): Model and annotate multiword expressions in a way which is unified across many languages. Make UD and PARSEME initiatives converge in this respect.
 - [Workplan](#): After performing pilot annotation based on the [draft guidelines for nominal MWEs](#) during WG1 Day in Naples, we are currently working on transforming the annotator feedback into Gitlab issue discussions. We plan to have a consolidated version of the guidelines for all MWEs (verbal, nominal, modifier, functional) ready by end of 2024, so as to conduct a large-scale annotation campaign in early 2025. The results will be used for organizing a shared task on automatic MWE identification.
 - [How can I contribute](#): <TBA>
 - [Documents](#)
 - [Minutes](#) from the Task 1.2 meetings
 - White paper proposition of the [roadmap for UD/PARSEME unification](#)
- **Task 1.3: Extensions and updates to morphosyntactic annotation guidelines**
 - [Minutes](#) from the task meetings
 - [Agenda](#) and [report](#) from the Naples 2024 meeting
- **Task 1.4: Sharing tools, formats, and infrastructure**
 - [Agenda](#) and [report](#) from the Naples 2024 meeting

Training

- [UniDive webinar](#) for newcomers to Universal Dependencies, PARSEME and/or Grew-match
- [1st UniDive training school](#) will take place in Chişinău, Moldova on 8-12 July 2024

Channels

- [WG1 mailing list](#) for general announcements and proposals

- [WG1 Telegram group](#) for special announcements and discussions
- [WG1 GitHub repository](#) for collaborative surveys and information sharing

From:

<https://unidive.lisn.upsaclay.fr/> - **Universality, diversity and idiosyncrasy
in language technology
CA21167 COST Action**

Permanent link:

<https://unidive.lisn.upsaclay.fr/doku.php?id=wg1:wg1&rev=1712581503>

Last update: **2024/04/08 15:05**

