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

- 4 December 2024 at 4pm CET

Minutes of past meetings

- WG1 Meeting 10 (online) - 11 October 2024: Updates on WG1 tasks [minutes]
- WG1 Meeting 9 (online) - 11 June 2024: Updates on WG1 tasks [minutes]

- WG1 Meeting 8 (online) - 11 April 2024: Naples de-briefing and updates on WG1 tasks [minutes]
- 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**
  - Leaders / Contacts: André Coneglian, A. Seza Doğruöz, Flavio Massimiliano Cecchini
  - Objectives: Discuss how Linguistic Typology, Corpus Annotation and Universal Dependencies can be brought into mutual relevance. More specifically, discuss how Linguistic Typology and Corpus Annotation can be explored to provide insights into the development of treebanks for new languages not in the UD database.
  - Work plan:
    1. Determine ways in which linguistic typology can help in the trade-off between universality and language specific phenomena in corpus annotation (Systematic overview of problematic (or difficult) phenomena for annotation (e.g., noun incorporation, and others) in order to evaluate what solutions what been proposed and whether or not unification is possible).
    2. Take into account less-resourced languages in corpus annotation so as to create new annotated corpora
    3. More broadly, assess how annotated treebanks (particularly UD treebanks) can figure in typological research
  - How can I contribute: <TBA>
  - Documents / Links:
    - 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 / Links

        - 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**
- Leaders / Contacts: Atul Kr. Ojha, Daniel Zeman
- Objectives: The general objective of the task is to improve the linguistic part of annotation activities, focusing in particular on the Universal Dependencies treebanks (because MWE annotation guidelines are the focus of Task 1.2). In order to not duplicate efforts, most of the work is done directly in the UD infrastructure.
    - Subtask **A:** Issues in the UD Issue Tracker. Select issues of interest, discuss them from the perspective of the languages we know (either during a meeting or directly in the issue tracker, including non-members of UniDive), propose a solution. If UniDivers have issues that are not yet present in the issue tracker, create a new issue there.
    - Subtask **B:** Construction-oriented guidelines. The UD website is relatively good as a reference manual, with separate pages for individual part-of-speech tags, morphological features and relations in individual languages. It is not so good in providing the big picture with a wholistic solution for individual constructions and strategies, although there is a growing number of documentation pages that attempt to close this gap. Since 2018, there is also an incubator for construction-oriented documentation. We can help this documentation grow, again with a typologically varied perspective, given the joint language expertise present in UniDive.
- How can I contribute:
    - Join the ongoing discussions on GitHub (UD issue tracker, see the link above).
    - If you can write part of the construction-oriented documentation, get in touch with the task leaders.
- Documents / Links:
    - Expression of interest
    - Minutes from the task meetings
    - Agenda and report from the Naples 2024 meeting
    - Next meeting: 2024-06-26 on Zoom

- **Task 1.4: Sharing tools, formats, and infrastructure**
    - Leaders / Contacts: František Forgáč, Bruno Guillaume
    - Objectives: The general objective of the task is to improve the technical part of annotation activities, focusing on tools, file formats and storage infrastructures. We are currently focusing on two more specific objectives:
        - Subtask **A**: Provide an overview of existing software and/or tools that support manual linguistic annotation
        - Subtask **B**: Evaluate the pros and cons of tabular formats (such as CoNNL-U) currently used in the UD and Parseme projects
    - Workplan:
        - Subtask **A**: The specific objective is to create a comparison table of available manual annotation tools morpho-syntactic and multiword expression annotations. A survey will be propose in the upcoming weeks, to collect feedback adn to produce the final version of the table.
        - Subtask **B**: Conduct a detailed analysis of the advantages and disadvantages of the tabular annotation formats, specifically CoNLL-U, as utilized in the Universal Dependencies (UD) and PARSEME projects. A first draft of an evolution of the formats currently used will be proposed for discussions and for testing.
    - How can I contribute?
        - Join to the ongoing discussions on GitHub (links above)

        - Stay tuned for the call to complete the survey
        - Join the task co-leaders team
    - Documents
        - Comparison table (WIP)
        - GitHub discussions about the comparison table and about file formats
        - Document used in the Task 1.4 session at the WG1 meeting in Naples (February 2024): Slides and Agenda

- **Task 1.5: Annotation of Spoken data**
    - Leaders / Contacts: Sylvain Kahane, Kaja Dobrovoljc
    - Objectives: To identify and discuss issues related to spoken data annotation, propose harmonized guidelines for morphosyntactic annotation of speech-specific phenomena, and promote the awareness of spoken data in other WG1 activities.
    - How can I contribute?
        - Join us at the first T1.5 (online) meeting on **October 2nd at 16:30 CEST** , where we will discuss the main goals and the possible ways to achieve them.
    - Documents:
        - Minutes of the meetings and presentations so far

WG1 session in Budapest

Here is the detail schedule of the WG1 session, on Thursday 30 January, 9:00 - 10h3, during the 3rd general UniDive meeting in BudaPest:

* 9:00-9:45: First part, plenary

- 15 min task 1.5: presentation of the task and the survey, discussion with the room
- 10 min task 1.1: presentation of the survey about UD treebanks annotation practises
- 10 min task 1.2: presentation of new guidelines focusing on functionals (new since Naples)
- 10 min task 1.4: presentation of the survey about annotation tools

* 9:45-10:30: Second part, 3 parallel task sessions

- task 1.1 **Typology and corpus annotation**. In this session we will further discuss the results of the survey about annotation practices in Universal Dependencies launched between December 2024 and January 2025 among UD and UniDive's communities. We will open a discussion about how a better theoretical and methodological uniformity can be reached inside UD, under the light of the issues, experiences and expectations of the participants and of the community at large. We will debate about how typological awareness can be increased, and how lingusitic typology can help reach this goal. Some points include: the implementation of a specific methodology associated with typological research, the distributional method, challenges provided by the annotated data, and accessibility of UD documentation.

- task 1.2 **MWE annotation**. Following the announcement of the UniDive MWE annotation campaign and the accompanying call for expressions of interest sent out prior to the Budapest meeting, this session will focus on:
    - Outlining the plans for the UniDive MWE annotation campaign scheduled for spring 2025, including its goals, timeline, and recommendations for getting started. [link TBA]
    - Presenting the current state of the new MWE guidelines, which now encompass verbal, nominal, modifier, and functional MWEs. [link TBA]
    - Demonstrating the use of GitLab issues to report challenges or propose amendments

encountered during the annotation process.
    - Presenting the tests for modifier and functional MWEs and enriching them with multilingual examples through onsite brainstorming with participants. [link TBA]

- task 1.4 **Tools and formats**. TBA

Training

- UniDive webinar for newcomers to Universal Dependencies, PARSEME and/or Grew-match
- 1st UniDive traininig school will take place in Chișinău, 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