

Working Group 2: Lexicon-corpus interface

- **Leader:** [Verginica Barbu Mititelu](#) (Romania)
- **Vice-leader:** [Voula Giouli](#) (Greece)

Workplan

In the context of a quest for diversity, electronic lexica are complementary to corpora because they aim at holistic language modelling, describing possibly many linguistic objects, whereas in corpora many phenomena occur rarely or never (§1.1.1.2). Lexica can also be useful in unifying terminologies, e.g., when a category can be described as a closed word list. In this context WG2 will be dedicated to:

1. Cross-language **unification of lexical features**:
 - harmonizing the definition of a “syntactic word” across languages,
 - harmonizing lemmatization rules (for words and MWEs) and lexical features across languages,
 - standardizing lists of lexemes for auxiliaries, pronouns and determiners;
2. **Design** of a lexicon-corpus **interface** aiming at:
 - interlinking MWE lexicon entries with their occurrences in corpora,
 - cross-lingually unified lexicography of idiosyncratic **constructions**;
3. Proof-of-concept lexical encoding of MWEs following the above design.

Organization

The monthly online meetings of WG2 are announced on the dedicated email list.

Current Subtasks

- **Task 2.1:** Cross-language unification of lexical features [co-leaders: Kilian Evang, Dan Zeman, Petya Osenova]
- **Task 2.2:** Design of a lexicon-corpus interface [co-leaders: Simon Krek, Carole Tiberius, Jaka Čibej]
- **Task 2.3:** Proof-of-concept lexicon encoding of MWEs [co-leaders: Stella Markantonatou, Ivelina Stoyanova, Christian Chiacos, Ranka Stanković]
- **Task 2.4:** Universal Canonical form of MWE,
- **Task 2.5:** Extensions of UD-PARSEME annotation for constructions and constructions.

Documents

- [WG2 Meeting 12 minutes](#), **28 June 2024** (online)
- [WG2 Meeting 11 minutes](#), **31 May 2024** (online)
- [WG2 Meeting 10 minutes](#), **26 April 2024** (online)
- [WG2 Meeting 9 minutes](#), **8-9 February 2024**, University of Naples L'Orientale, Italy
- [WG2 Meeting 8 minutes](#), **19 December 2023** (online)

- [WG2 Meeting 7 minutes](#), **6 November 2023** (online)
- [WG2 Meeting 6 minutes](#), **5 September 2023** (online)
- [WG2 Meeting 5 minutes](#), **6 July 2023** (online)
- [WG2 Meeting 4 minutes](#), **1 June 2023** (online)
- [WG2 Meeting 3 minutes](#), **4 May 2023** (online)
- [WG2 Meeting 2 minutes](#), **6 April 2023** (online)
- [WG2 Meeting 1 minutes](#), **16-17 March 2023**, Paris-Saclay University, France (co-located with UniDive 1st general meeting,
- Martin Haspelmath's paper draft on [defining the notion of the word](#).

Presentations

Jan Odijk, Canonical form of MWEs - [short presentation](#), [long presentation](#)

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=wg2:wg2&rev=1733424124>

Last update: **2024/12/05 19:42**

