

# 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 the **Working Group 2** (lexicon-corpus interface) will be taking place every **first Thursday of the month from 13:00 CEST** (for an hour)

### Upcoming WG2 meetings

- **4th May 2023**, 13:00 CEST
- every first Wednesday in a month at 1 p.m. CEST

### Past WG2 meetings

- **6th April 2023**, 13:00 CEST: [minutes](#)

## Documents

- [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](#).

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=1680783319>

Last update: **2023/04/06 14:15**

