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 will be taking place every **first Thursday of the month from 13:00 CEST** (for an hour). See the list of past and upcoming WG meetings.

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.