Multiword Expressions between the Corpus and the Lexicon: Universality, Idiosyncrasy and the Lexicon-Corpus Interface

Verginica Barbu Mititelu, Voula Giouli, Kilian Evang, Daniel Zeman,
Petya Osenova, Carole Tiberius, Simon Krek, Stella Markantonatou,
Ivelina Stoyanova, Ranka Stankovic, Christian Chiarcos

Romanian Academy Research Institute for Artificial Intelligence; Institute for Language and Speech Processing, Athena Research Centre; Heinrich Heine Univ. Düsseldorf; ÚFAL MFF, Charles Univ.; Institute of Information and Communication Technologies, BAS; Leiden Univ.; Jožef Stefan Institute; Institute for Language and Speech Processing, Athena Research Centre; Institute for Bulgarian Language, BAS; Univ. of Belgrade; Univ. of Augsburg

Funded by the European Union

Goals:

- proof-of-concept lexical encoding of MWEs in lexica
- creation of an ecosystem of interlinked MWE-dedicated lexica and annotated corpora (already in place within PARSEME COST Action)
- identification of MWEs - within UniDive COST Action (WG2)

Challenges:

(a) the harmonisation of corpora and lexica by also accounting for universality and diversity,
(b) the efficient encoding of MWEs of all grammatical categories cross-linguistically, and
(c) the adoption of the appropriate mechanisms and tools for linking lexica and corpora

MWEs in computational lexica: SOTA

- 72% of the resources are aimed for NLP use
- > 40 languages and dialects are represented (esply Indoeuropean
- 70.7% of the resources are monolingual, 18.7% bilingual and 10.6% multilingual.
- Most datasets were acquired manually or semi-automatically.
- Only 24% of the resources are linked to a (usually small) corpus and 12% are linked to other resources.
- 45% of the resources provide comprehensive description of MWEs.

Capturing Universality

- Definition of the notion of “word” - running survey based on Martin Haspelmath’s definition
- Definition of the notion of “lemma”
  - Challenges raised by words:
    - Pronouns
    - Doublet verbs
    - Numbers
    - Negated words
    - diminutives
  - Challenges raised by MWEs:
    - Compounding
    - Quasi-reflexive verbs

Linking MWE lexicon entries with their occurrences in corpora

- ELEXIS-WSD parallel sense annotated corpus enhanced with new languages and upgrading the annotation to enable linking MWE lexicon entries with their occurrences in the corpora;
- published as Linked Data (using NLP Interchange Format - NIF) to facilitate linking with the sense repository of the corpus
- OntoLex vocabulary: the core module Lemon and MWE relevant modules: Decomp, Morph, FrAC
- Linked Data enhances accessibility, interoperability, semantic enrichment, community collaboration, and the promotion of open science

Proof-of-concept lexical encoding of MWEs: minimal requirements:

- a definition of the notion of “word” that is as universal as possible,
- a shared understanding of MWEs that can be annotated in corpora and then linked with lexicon entries (both the MWE as a whole and its components), including all types of MWEs,
- centralised guidelines for lexicon encoding regarding, i.e., the notions of lemma, canonical form, lexical features, etc.,
- a uniform representation of the syntactic properties of MWEs, and
- tools and mechanisms for linking MWE entries with their occurrences in corpora.

25 May 2024
Turin, Italy, Joint MWE-UD Workshop