Multilingual MWE resources

WG1

Survey about Multilingual MWE resources

Introduction

Losnegaard, G., Sangati, F., Parra, C., Savary, A., Bargmann, S., and **J. Monti**. (2016): "*PARSEME Survey on MWE Resources*", in the *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, 23-28 May 2016, Portorož, Slovenia (poster). (PARSEME country: France, Germany, Italy, Norway, Spain)

Type	Count	%
Multilingual MWE lists (4.1.1)	4	26%
Multilingual MWE lexicons (4.1.2)	10	66%
Others (4.3)	1	0.6%

Table 2: Types of Multilingual MWE Resources

Introduction

- Multilingual MWE resources are hard to find.
- They are not pure MWE resources but MWEs are in most cases only part of the data
- Definition of the different types of resources are not homogeneous

Definition

- **A multilingual MWE resource** is a large and structured set of linguistic data concerning MWEs. It may contain multiword expressions in two or more languages. These resources can be used as stand-alone resources or can be integrated into NLP applications of different nature (Machine Translation, CAT tools, CLIR applications, etc.).

Parallel and comparable resources

According to the type of relationship between the languages represented in the resource, multilingual MWE resources can be of two types:

- **Parallel MWE resources** which have been specially formatted for side-by-side comparison and contain MWE data in two or more languages aimed at translation purposes. They are usually monodirectional, from a source language to two or more target languages.
- **Comparable MWE resources** the MWE data are of the same kind and cover the same content, but they are not translations of each other.

Generic and specialised resources

According to the content of the resource (coverage?), we can have:

- **generic multilingual MWE resources**: they include MWE which belong to a standard language
- **specialised MWE multilingual MWE resources**: they include MWE which belong to a particular sublanguage or to a restricted application domain (for instance, journalism, medicine, law, children, …)

Type of modality: spoken/written/multimedia,

Types of multilingual MWE resources

According to different types of linguistic resources, we can have:

1. **Multilingual MWE list**: a list which contains (exclusively or not) MWE data in digital form and in two or more languages and can be accessed through a number of different media. They usually contain a list of words or terms in alphabetical order with some limited additional information such as definitions, synonyms, examples.

Example: multilingual MWE glossaries,

Types of multilingual MWE resources

2. Multilingual MWE terminological databanks: a databank which contains (exclusively or not) MWE terms in digital form and in two or more languages and can be accessed through a number of different media.
Example: IATE

Types of multilingual MWE resources

3. Multilingual MWE electronic dictionary: a dictionary which contains (exclusively or not) MWE data in digital form and in two or more languages and can be accessed through a number of different media. Electronic dictionaries can be found in several forms , including (i) as dedicated handheld devices, (ii) as apps on smartphones and tablet computers or computer software, (iii) as a function built into an E-reader, (iv) as CD-ROMs and DVD-ROMs, (v)as free or paid-for online products.

Example: English-Italian online collocation dictionary (http: //oxforddictionary.so8848.com/)

Types of multilingual MWE resources

4. **Multilingual MWE machine-readable dictionary**: a dictionary stored as machine (computer) data which contains (exclusively or not) MWE and can be loaded in a database for NLP applications, such as MT, CAT, CLIR applications. A multilingual MWE MRD may be a dictionary with a proprietary structure that is queried by dedicated software (for example online via internet) or it can be a dictionary that has an open structure and is available for loading in computer databases and thus can be used via various software applications. A MWE MRD contains MWE entries with various annotations (POS, morpho-syntactic, semantic, translation annotations) which may differ according to the NLP application it is used for. Example: ….

Types of multilingual MWE resources

5. Multilingual MWE annotated corpora: they contain text in electronic format annotated with MWE data. They can be divided into:

5.1. Parallel MWE annotated corpora (or translation corpora): they contains texts in two languages and the text (target text) in one language is the translation of the other one (source text). The source and the target texts are aligned, i.e. equivalent text segments (phrases or sentences) in the source and the target text. These segments are linked so that they are formatted for side-by-side comparison. They may contain (exclusively or not) alignments or annotations of MWE data. Example: MWE-TED corpus

Types of multilingual MWE resources

5.2. Comparable MWE annotated corpora: they contain texts in two or more languages of the same kind and cover the same content, but they are not translations of each other. They may contain (exclusively or not) annotations of MWE data. Example: ??

6. **Multilingual MWE annotated treebank**: a parsed text corpus that annotates syntactic or semantic sentence structure with MWE data annotations in a multilingual perspective
Example: ??

Types of multilingual MWE resources

7. Multilingual MWE thesaurus: a reference work that lists MWE grouped together according to similarity of meaning (containing synonyms and sometimes antonyms) in a multilingual perspective.
Example: ??

8. Multilingual MWE taxonomy: classification of MWE in a multilingual perspective
Example: ??

Types of multilingual MWE resources

9. **Multilingual MWE ontology**: a model for describing MWE in a multilingual perspective that consists of a set of types, properties, and relationship types.
Example: ??