Tutorial on annotating multiword expressions

Verginica Barbu Mititelu
Romanian Academy
Research Institute for Artificial Intelligence
vergi@racai.ro

Ivelina Stoyanova
Institute for Bulgarian Language
Bulgarian Academy of Sciences
iva@dcl.bas

Abstract

This tutorial is meant to familiarize the audience with the principles and guidelines for annotating (mainly verbal and nominal) multiword expressions in a corpus. The notion of "multiword expression" will be introduced with a view to the challenges this phenomenon raises in automatic processing of texts and the importance of creating corpora annotated with multiword expressions. A general picture of the preoccupations with multiword expressions will be created, and then we will zoom into the activities within the previous PARSEME and current UniDive COST Actions and their efforts to offer a consistent, uniform treatment of such expressions in various languages, with an eye to capturing its universality as well as accommodating the specificities of various languages. The annotation guidelines will be explained and illustrated and then the audience will have the opportunity of testing them themselves and discuss the observations during the hands-on sessions.

Keywords: multiword expressions, corpus annotation, guidelines.

1 General presentation

The **aim** of the proposed tutorial is to familiarize the audience with the guidelines for annotating multiword expressions (MWEs) of various parts of speech (with a focus on verbal and nominal ones), so as to enlarge the number of annotators of this phenomenon in corpora for various languages (including Bulgarian).

MWEs are linguistic phenomena (covering a wide spectrum of linguistic entities, from idioms to collocations and named entities) having idiomaticity as their common property. Idiomaticity may manifest at various linguistic levels (from lexical to pragmatic), and even statistical (Baldwin and Kim, 2010). Although there is no total agreement among specialists with respect to the limits of this phenomenon, its existence has never been denied and it has attracted increasing interest from experts, especially in the last decades, with PARSEME[1] and UniDive[2] COST Actions being fora in which a large network of specialists (from several fields, such as theoretical and computational linguistics, language diversity, language technology) discuss the topic with an eye to its universality, as well as to language-specific characteristics.

The importance of creating corpora annotated with MWEs is motivated by several factors, among which we highlight:

- it offers a clearer image of the extension of the phenomenon in the language;
- it ensures training, tuning and testing material for systems developed to automatically identify or discover MWEs in texts;
- it offers language specialists a set of examples of the behaviour of MWEs in real data, for better describing them in general or dedicated resources.

Our focus will be the PARSEME guidelines for annotating verbal MWEs (Khelil et al., 2020) and UniDive guidelines for annotating nominal MWEs (under development). We chose to work with these because of their characteristics: not being English-centred, trying to capture the universality, the quasi-universality and the specificity of the phenomenon with respect to the languages included in the actions.

We present the PARSEME multilingual corpora (Savary et al., 2023), the annotation scheme, and the applications for MWE recognition, cross-lingual studies, etc. The tutorial will offer hands-on sessions as well, which will provide an opportunity

[1] https://gitlab.com/parseme/corpora/
[2] https://unidive.lisn.upsaclay.fr/

to experience annotation and the challenges of the identification and classification of MWEs.

On the other hand, the treatment of MWEs in corpora is bound to rely on lexical resources representing their lexical, morphological and semantic properties (Savary et al., 2019). Thus, their annotation will be presented in view of their representation and treatment in monolingual as well as multilingual resources. Moreover, we pose the question which categories of MWEs need to be included in lexicons.

We focus on the following properties of MWEs with the aim of studying irregularities in their realization:

- lexical – lemma and representation of the MWEs;
- morphosyntactic – internal structure, forms, word order, etc.;
- syntactic – behavior of MWEs in a sentence;
- semantic – decomposability and idiosyncrasy;
- statistical – frequency of appearance as compared to other phrases of similar meaning.

2 Target audience

The expected audience of this tutorial is represented by upper-level undergraduates, graduates and post-graduates, and linguistically knowledgeable participants. Previous annotation skills are not required, though they will be helpful.

Expected knowledge includes familiarity with linguistic notions, language features, language transformations, word inflection, lexico-semantic relations.

Experience in documenting and working with corpora would also be beneficial.

3 Objectives

- To offer a top-down perspective on the MWE phenomenon;
- To offer a description of MWEs, both verbal and nominal, with view to their identification in text;
- To use well-defined criteria in order to acquire or improve the skills of identifying MWEs in text;
- To use the new skills in annotating manually a short text with verbal and nominal MWEs.

The outcomes for the attendees will include:

- theoretical – better understanding of the phenomenon of MWEs and their characteristics; considerations about the language-specific nature of the phenomenon; the notion of decomposability;
- practical – skills for dealing with MWE identification, annotation and analysis in text; adhering to annotation guidelines, analysing annotation, interannotators' agreement, etc.;
- applied – use of annotated resources for linguistic analysis, for NLP applications, etc.

4 Content

The tutorial will describe the current state of identifying MWEs in text. Previous results and current activities for corpora annotation will be presented, with a focus on low-resourced languages, including Bulgarian. The perspective on the phenomenon is as presented within the PARSEME and UniDive COST Actions, as mentioned above.

Topics to be covered:

- why MWEs are of interest and a challenge for language processing;
- definition and characteristics of the phenomenon;
- delimiting the phenomenon;
- decision tree for verbal MWEs;
- decision tree for nominal MWEs;
- hands-on experience in identification of MWEs in text.

5 Duration

The duration of the tutorial will be 6 sessions, where each session is 45 minutes.

Schedule of the sessions:

Session 1: The phenomenon of MWE: definition, examples, characteristics. Lexical resources of MWEs. Tutor: Verginica Barbu Mititelu

Session 2: SOTA of MWEs identification in text. Previous results in MWE annotation in corpora. Tutor: Ivelina Stoyanova

Session 3: Decision tree for verbal MWEs. Tutor: Verginica Barbu Mititelu

Session 4: Annotation of verbal MWEs in a sample corpus. Hands-on activity. Part I. Tutor: Ivelina Stoyanova

Session 5: Annotation of verbal MWEs in a sample corpus. Hands-on activity. Part II. Tutor: Verginica Barbu Mititelu

Session 6: Findings from the annotation activity. Conclusions and discussion. Tutor: Ivelina Stoyanova

6 Prerequisites

The following are recommended but not compulsory:

- General knowledge about the MWE phenomena in language;
- Familiarity with linguistic description, language feature and inflection of MWEs;
- Experience or general knowledge about corpus compilation and annotation.

7 Recommended reading

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. **2002**. *Multiword Expressions: A Pain in the Neck for NLP*. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics* (CICLing-2002), pages 1–15, Mexico City, Mexico.

Timothy Baldwin and Su Nam Kim. **2010**. *Multiword expressions*. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, Second Edition, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 9781420085921.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. **2017**. *The PARSEME shared task on automatic identification of verbal multiword expressions*. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE 2017), pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejcek, Fabienne Cap, Slavomír Ceplo, Silvio Ricardo Cordeiro, Gulsen Eryigit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaite, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova, and Veronika Vincze. **2018**. *PARSEME multilingual corpus of verbal multiword expressions*. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 87–147. Language Science Press., Berlin.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaite, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. **2018**. *Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions*. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions* (LAW-MWE-CxG-2018), pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. **2020**. *Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions*. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

8 Instructors' presentation

Verginica Mititelu is a linguist working as a senior researcher for the Romanian Academy Research Institute for Artificial Intelligence. She performed her Master studies at and received her PhD in Philology in 2010 from the University of Bucharest. She has constantly been preoccupied with and involved in the development of language resources, especially for Romanian, applying up-to-date annotation schemes and adjusting them to the characteristics of the language under study. She has also been concerned with standardizing the resources developed, especially using Linked Data principles of representation, and with the registration of their metadata in international data repositories. She is the language leader for Romanian in annotation corpora for MWEs in PARSEME and UniDive COST Actions. In the latter, she also serves as a leader of the Working Group on Lexicon-Corpus Interface.

Ivelina Stoyanova works at the Department of Computational Linguistics at the Institute for Bulgarian Language, Bulgarian Academy of Sciences. She has a master's degree in Bulgarian Studies from Sofia University and a bachelor's degree in Computer Science and Mathematics from the University of Bath, UK. She obtained her PhD degree from the Institute for Bulgarian Language in 2012 and her thesis was on the topic of MWE recognition and tagging in Bulgarian. She works actively on many national and international projects on developing language resources and applications for language processing. She is the language leader for Bulgarian in the annotation task of PARSEME COST Action.

9 Available materials

Materials that will be available after the tutorial include slides, bibliography, sample corpora, annotation guidelines.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second edition*, pages 267–292. CRC Press, Boca Raton.

Chérifa Ben Khelil, Archna Bhatia, Claire Bonial, Marie Candito, Fabienne Cap, Silvio Cordeiro, Vassiliki Foufi, Polona Gantar, Voula Giouli, Najet Hadj Mohamed, Carlos Herrero, Uxoa Iñurrieta, Mihaela Ionescu, Iskandar Keskes, Alfredo Maldonado, Verginica Mititelu, Johanna Monti, Joakim Nivre, Mihaela Onofrei, Viola Ow, Carla Parra Escartín, Manfred Sailer, Carlos Ramisch, Renata Ramisch, Monica-Mihaela Rizea, Agata Savary, Nathan Schneider, Ivelina Stonayova, Sara Stymne, Ashwini Vaidya, Veronika Vincze, Abigail Walsh, and Hongzhi Xu. 2020. Parseme corpora of multiword expressions - version 1.2 (2020), annotation guidelines. https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze, and Abigail Walsh. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91, Florence, Italy. Association for Computational Linguistics.