Preface

Stella Markantonatou
Institute for Language and Speech Processing, Athena RIC, Greece

Carlos Ramisch
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

Agata Savary
University of Tours, LIFAT, France

Veronika Vincze
University of Szeged, Hungary

In this introductory chapter we present the rationale for the volume at hand. We explain the origin and the selection process of the contributing chapters, and we sketch the contents and the organization of the volume. We also describe notational conventions put forward for citing and glossing multilingual examples of multiword expressions. We finally acknowledge the efforts which paved the way for setting up this book project, ensuring its quality and publication.

Multiword expressions (MWEs) belong to those language phenomena which pose the hardest challenges both in linguistic modelling and in automatic processing. This is due notably to their semantic non-compositionality, that is, the impossibility to predict their meaning from their syntactic structure and from the semantics of their component words in a way deemed regular for the given language. But MWEs also exhibit unpredictable behaviour on other levels of language modelling such as the lexicon, morphology and syntax, and call, therefore, for dedicated procedures in natural language processing (NLP) applications.

These challenges have been addressed by an ever-growing and increasingly multilingual community gathering at the Multiword Expressions Workshop, organized yearly since 2003, often jointly with major NLP conferences. The 13th

Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze. 2018. Preface. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, iii–xiii. Berlin: Language Science Press.
DOI:10.5281/zenodo.1469549

edition of the Workshop, co-located with the EACL 2017 conference in Valencia, Spain, saw a major evolution of the topics and methods addressed by the community. This evolution resulted notably from the efforts coordinated by PARSEME, a European research network dedicated to parsing and MWEs, gathering, since 2013, researchers from 31 countries and working on as many languages.[1]

One of PARSEME's main outcomes was a corpus in 18 languages annotated for verbal MWEs (VMWEs), based on a unified methodology and terminology, and published under open licenses. This considerable collective and inclusive effort mobilized experts from different linguistic traditions, triggered cross-language and cross-domain discussions, and brought convergence to modelling and processing of MWE-related phenomena. The availability of this new open resource also made it possible to organize the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions, i.e. a competition of VMWE identification tools, whose culminating event was hosted by the MWE 2017 workshop in Valencia. The 7 participating systems covered jointly all 18 languages represented in the corpus. They also offered a large panorama of VMWE identification techniques. These assets, as well as some other contributions published in the main track of the MWE 2017 workshop, showed a growing awareness by the MWE community of specific challenges posed, in particular, by verbal MWEs, such as their discontinuity and high morpho-syntactic flexibility. The workshop programme addressed a large variety of MWE-dedicated tasks such as: lexical and grammatical encoding, annotation, tokenization, extraction, identification, classification, variation study, parsing, compositionality prediction, and translation. Finally, it testified that MWE research has reached a highly multilingual stage.

1 Organization and contents of the volume

This volume is a collection of selected extended papers from the MWE 2017 workshop in Valencia: 8 of them from the main track, and 5 from the shared task track. They address 19 languages from 9 language families, as shown in Figure 1. The chapter selection process was initiated by an open call, addressed to all coauthors of the workshop papers. The call included the requirement of extending the original contributions by at least 30% with unpublished content. An international programme committee reviewed 15 submissions and selected 14 of them for publication. One of the selected chapters was further withdrawn. As a result, the volume consists of 13 chapters covering a large variety of aspects related to MWE representation and processing, with a particular focus on verbal MWEs.

[1] http://www.parseme.eu

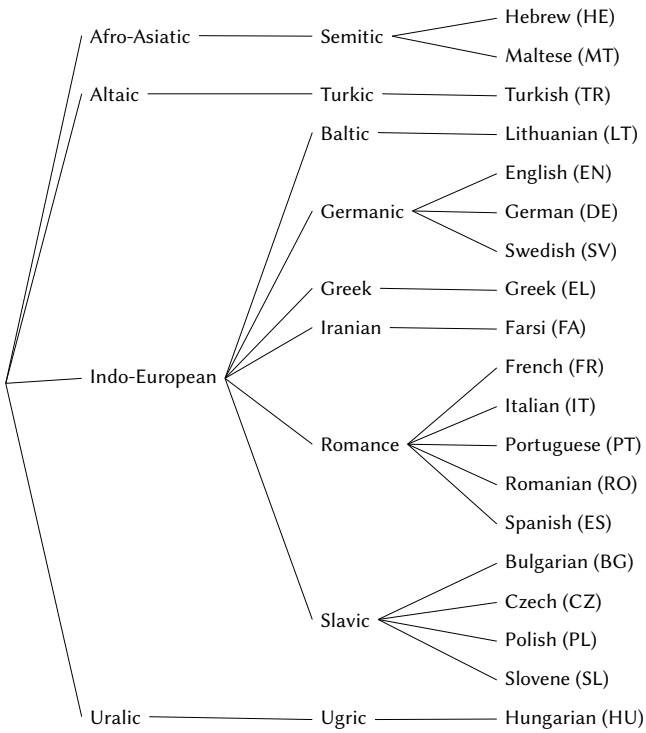

Figure 1: Languages addressed in the chapters of this volume, together with their two-letter language code, language families and genera (middle columns), according to WALS (World Atlas of Language Structures, Dryer & Haspelmath 2013).

Chapters 1 to 3 address outstanding linguistic properties of VMWEs and their automatic assessment. Geeraert et al. (2018 [this volume]) discuss idiomatic *variation* of several types of English verbal idioms. They draw on multimodal evidence, namely acceptability rating and eye-tracking measurements, to investigate comprehension mechanisms. Barančíková & Kettnerová (2018 [this volume]) deal with light-verb constructions and verbal idioms in Czech, and explore their paraphrasability by single verbs. They also propose a lexicographic scheme for VMWE paraphrase encoding, and show its usefulness in machine translation. Bhatia et al. (2018 [this volume]) focus on English verb-particle constructions, and estimate their compositionality degree, so as to further compute the semantics of sentences containing verb-particle constructions on the basis of lexical, grammatical and ontological resources.

Chapters 4 to 8 are dedicated to the PARSEME shared task on automatic identification of verbal MWEs. Savary et al. (2018 [this volume]) describe the PARSEME multilingual VMWE-annotated corpus underlying the shared task. In a first step, the annotation guidelines and methodology are presented, then the properties of the annotated corpus are analysed across the 18 participating languages. Maldonado & QasemiZadeh (2018 [this volume]) offer a critical analysis of the shared task organization and of its results across languages and participating systems. Chapters 6 to 8 are dedicated to three of the seven VMWE identification systems participating in the shared task. They show a representative panorama of recent techniques used to address the VMWE identification task. Moreau et al. (2018 [this volume]) model the task as sequence labelling with reranking. Al Saied et al. (2018 [this volume]) present a dedicated transition-based dependency parser, which jointly predicts a syntactic dependency tree and a forest of VMWEs. Finally, Simkó et al. (2018 [this volume]) rely on a generic dependency parser trained on a corpus with merged syntactic and VMWE labels.

Chapters 9 to 11 further discuss MWE identification issues in various settings and scopes. Brooke et al. (2018 [this volume]) show how comparing various annotations of the same MWE in an English corpus can help correct annotation errors, enhance the consistency of corpus annotation, and consequently increase the quality of downstream MWE identification systems. Scholivet et al. (2018 [this volume]) address identification of French continuous MWEs via sequence labelling, compare its results to more sophisticated parsing-based approaches, and show that feature engineering based on external lexical data (whether handcrafted or automatically extracted) systematically enhances the tagging performance. Taslimipoor et al. (2018 [this volume]), conversely, advocate modelling MWE identification as classification rather than tagging. They exploit word embeddings as classification features in Italian, Spanish and English, and put forward a MWE-specific methodology of train vs. test corpus split.

The last two chapters of the book, 12 and 13, are dedicated to multilingual MWE-oriented applications. Garcia (2018 [this volume]) describes automatic extraction of bilingual collocation equivalents in English, Spanish, and Portuguese, using syntactic dependencies, association measures and distributional models. Finally, Salehi et al. (2018 [this volume]) predict the compositionality of English and German MWEs on the basis of their translations extracted from highly multilingual lexical resources.

2 Conventions for citing and glossing multilingual MWE examples

As mentioned above, this volume addresses a large number of languages, particularly in the chapters related to the PARSEME corpus and shared task. Therefore, the editorial effort around this volume includes putting forward notational conventions which might become a standard for citing and glossing multilingual MWE examples. We illustrate the proposed conventions by the *numbered examples* (1) to (4). Each numbered example contains:

(i) a sample use of the VMWE, followed by the 2-letter ISO 639-1 language code (cf. Figure 1),

(ii) a transcription, if the language of the example is written with a script different from the one used for the main text,[2]

(iii) a gloss following the Leipzig Glossing Rules,[3]

(iv) a literal translation, followed by an idiomatic translation in single quotes.

For English examples, items (ii)–(iv) are irrelevant or optional but idiomatic translation might sometimes be useful to ease the comprehension by non-native readers. For right-to-left languages (e.g. Farsi or Hebrew), item (i) is spelled right-to-left, item (iv) left-to-right and items (ii)–(iii) left-to-right within components, and right-to-left from one component to another. Lexicalized components of the VMWE, i.e. those which are always realized by the same lexeme (cf. Savary et al. 2018 [this volume], §2, p. 92) are highlighted in bold face.

(1) She reluctantly **took on** this task. (EN)
'She reluctantly agreed to be in charge of this task.'

(2) *Ida* ***skriva glavo v pesek.*** (SL)
Ida hide.3.SG head in sand
Ida hides her head in the sand. 'Ida pretends not to see a problem.'

[2]For instance, transcription is needed for Bulgarian, Greek, Farsi and Hebrew examples in this volume. Conversely, examples in English, or any other language using Latin script, would require transcriptions in texts written in Cyrillic, Greek, Arabic or Hebrew script.
[3]https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf

(3) *H Zωή* ***παίρνει*** *μία* ***απόφαση***. (EL)
i Zoi perni mia apofasi
the.FEM.SG Zoe.FEM.SG take.3.SG a decision
Zoe takes a decision. 'Zoe makes a decision.'

(4) دیده است خواب من برای کافی قدر به (FA)
ast **dide khab** man **baraye** kafi qadre be
is seen sleep me for enough quantity to
He had enough sleep for me. 'He has many plans for me.'

In-line examples, used for brevity, are preceded by the 2-letter language code, contain items (i), (ii) if relevant, and (iv) only, and the idiomatic translation (if any) is introduced by a double arrow '⇒'. For instance, an in-line example corresponding to numbered example (2) would be the following: (SL) *Ida* ***skriva glavo v pesek*** 'Ida hides her head in the sand' ⇒ 'Ida pretends not to see a problem'. If the language under study is written with a non-Latin alphabet, the inline example should not be in italics, and the transliteration should be included in parentheses, e.g. (EL) Η Ζωή **παίρνει** μία **απόφαση** (I Zoi perni mia apofasi) 'The Zoe takes a decision' ⇒ 'Zoe makes a decision'. To keep such examples reasonably short, the first item can be omitted and only the transliterated example is kept: (EL) I Zoi perni mia apofasi 'The Zoe takes a decision' ⇒ 'Zoe makes a decision'. The literal or the idiomatic translation are sometimes superfluous or too verbose, and can be skipped, as in: (EL) I Zoi perni mia apofasi 'Zoe makes a decision'.

The typesetting commands for both numbered and in-line examples for LaTeX can be found in the GitHub repository containing the source codes of this volume, accessible from its webpage.[4]

3 Acknowledgements

Huge collective efforts paved the way towards the publication of this volume.

The PARSEME Shared Task on Automatic Identification of VMWEs was made possible by the European research network PARSEME[5], funded by the IC1207 COST[6] Action in 2013–2017. The PARSEME core group and eighteen language teams (cf. acknowledgements in Savary et al. 2018 [this volume]) prepared the annotation guidelines and tools, and created the multilingual corpus underlying

[4]http://langsci-press.org/catalog/book/204
[5]http://www.parseme.eu
[6]http://www.cost.eu/

the shared task. We are also grateful to the COST Officials, notably to Ralph Stübner, for their continuous support in the scientific and financial administration of the Action.

The 13th Workshop on Multiword Expressions (MWE 2017)[7] was organized and sponsored by PARSEME jointly with the Special Interest Group on the Lexicon (SIGLEX)[8] of the Association for Computational Linguistics. An international Programme Committee of over 80 researchers from 27 countries reviewed the workshop submissions and ensured a high-quality paper selection.

Our warm acknowledgements go also to the Editorial Staff of Language Science Press, and in particular to Sebastian Nordhoff, for his expert and friendly editorial assistance. We are also grateful to the editors of the *Phraseology and Multiword Expressions* book series for their support. In particular, Yannick Parmentier played the role of the editor-in-charge of the volume, and offered advice on technical editorial issues.

Finally, we thank the following reviewers, who provided insightful reviews to the chapters submitted to this volume:

- Dimitra Anastasiou (Luxembourg Institute of Science and Technology, Luxembourg)
- Doug Arnold (University of Essex, UK)
- Timothy Baldwin (University of Melbourne, Australia)
- Eduard Bejček (Charles University in Prague, Czech Republic)
- António Branco (University of Lisbon, Portugal)
- Marie Candito (Paris Diderot University, France)
- Fabienne Cap (Uppsala University, Sweden)
- Matthieu Constant (Université de Lorraine, France)
- Paul Cook (University of New Brunswick, Canada)
- Lucia Donatelli (Georgetown University, USA)
- Silvio Ricardo Cordeiro (Aix-Marseille University, France)

[7] http://multiword.sourceforge.net/mwe2017
[8] http://siglex.org/

- Scott Piao (Lancaster University, UK)
- Martin Riedl (University of Stuttgart, Germany)
- Manfred Sailer (Goethe-Universität Frankfurt am Main, Germany)
- Nathan Schneider (Georgetown University, USA)
- Sabine Schulte Im Walde (University of Stuttgart, Germany)
- Ruben Urizar (University of the Basque Country, Spain)
- Aline Villavicencio (Federal University of Rio Grande do Sul, Brazil)
- Jakub Waszczuk (University of Tours, France)
- Shuly Wintner (University of Haifa, Israel)

We hope that the quality of this volume will be a valuable reward to all these contributors, and a source of information and inspiration for the international MWE community.

References

Al Saied, Hazem, Marie Candito & Matthieu Constant. 2018. A transition-based verbal multiword expression analyzer. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 209–226. Berlin: Language Science Press. DOI:10.5281/zenodo.1469561

Barančíková, Petra & Václava Kettnerová. 2018. Paraphrases of verbal multiword expressions: The case of Czech light verbs and idioms. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 35–59. Berlin: Language Science Press. DOI:10.5281/zenodo.1469553

Bhatia, Archna, Choh Man Teng & James F. Allen. 2018. Identifying senses of particles in verb-particle constructions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 61–86. Berlin: Language Science Press. DOI:10.5281/zenodo.1469555

Brooke, Julian, King Chan & Timothy Baldwin. 2018. Semi-automated resolution of inconsistency for a harmonized multiword-expression and dependency-parse annotation. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 245–262. Berlin: Language Science Press. DOI:10.5281/zenodo.1469565

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/, accessed 2018-7-25. Accessed on.

Garcia, Marcos. 2018. Comparing bilingual word embeddings to translation dictionaries for extracting multilingual collocation equivalents. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 319–342. Berlin: Language Science Press. DOI:10.5281/zenodo.1469571

Geeraert, Kristina, R. Harald Baayen & John Newman. 2018. "Spilling the bag" on idiomatic variation. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 1–33. Berlin: Language Science Press. DOI:10.5281/zenodo.1469551

Maldonado, Alfredo & Behrang QasemiZadeh. 2018. Analysis and Insights from the PARSEME Shared Task dataset. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 149–175. Berlin: Language Science Press. DOI:10.5281/zenodo.1469557

Moreau, Erwan, Ashjan Alsulaimani, Alfredo Maldonado, Lifeng Han, Carl Vogel & Koel Dutta Chowdhury. 2018. Semantic reranking of CRF label sequences for verbal multiword expression identification. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 177–207. Berlin: Language Science Press. DOI:10.5281/zenodo.1469559

Salehi, Bahar, Paul Cook & Timothy Baldwin. 2018. Exploiting multilingual lexical resources to predict MWE compositionality. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 343–373. Berlin: Language Science Press. DOI:10.5281/zenodo.1469573

Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon

Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87–147. Berlin: Language Science Press. DOI:10.5281/zenodo.1469555

Scholivet, Manon, Carlos Ramisch & Silvio Cordeiro. 2018. Sequence models and lexical resources for MWE identification in French. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 263–297. Berlin: Language Science Press. DOI:10.5281/zenodo.1469567

Simkó, Katalin Ilona, Viktória Kovács & Veronika Vincze. 2018. Identifying verbal multiword expressions with POS tagging and parsing techniques. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 227–243. Berlin: Language Science Press. DOI:10.5281/zenodo.1469563

Taslimipoor, Shiva, Omid Rohanian, Ruslan Mitkov & Afsaneh Fazly. 2018. Identification of multiword expressions: A fresh look at modelling and evaluation. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 299–317. Berlin: Language Science Press. DOI:10.5281/zenodo.1469569