Preface

Yannick Parmentier
University of Orléans
University of Lorraine

Jakub Waszczuk
University of Tours
University of Düsseldorf

In this introductory chapter, we first present the topic and context of this volume. We then summarize its contributions, which have been collected through an open call for submissions and a peer-reviewing process.

1 Introduction

While Multiword Expressions (MWEs), i.e. sequences of words with some unpredictable properties such as *to count somebody in* or *to take a haircut*, have been attracting attention for a long time because of these idiosyncratic properties which go beyond word boundaries, they remain a challenge for both linguistic theories and natural language (NL) applications.

Indeed, most of these theories and applications admit an (explicit or implicit) division of language phenomena into clear-cut levels: (i) tokens (indivisible text units, roughly words), (ii) morphology (properties of words e.g. number, gender, etc.), (iii) syntax (structural links between words, e.g. number/gender agreement), (iv) semantics (meaning of words and sentences). However, human languages frequently show a high degree of ambiguity and fuzziness with respect to this layer-oriented model. In particular, MWEs are placed on the frontier between these levels due to their idiosyncratic properties on the one hand, and their morphological, syntactic and semantic variations on the other hand. For instance, their meaning is often non-compositional as in *to take a haircut* (i.e. *to suffer a*

Yannick Parmentier & Jakub Waszczuk. 2019. Preface. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, iii–ix. Berlin: Language Science Press. DOI:10.5281/zenodo.2579031

serious financial loss), although they admit some syntactic variation similarly to many other expressions (*take/takes/have taken/has taken/took a serious/70% haircut*). Strictly layer-oriented language models fail to reflect this specificity, and thus yield erroneous text processing results (e.g. word-to-word translations of idioms). Although the quantitative importance of MWEs is well known (they cover up to 30% of all words in human language utterances, and are much more numerous in lexicons than single words), the achievements in their formal representation and automatic processing are still largely unsatisfactory.

In this context, an international and multilingual consortium of researchers recently took part in the European PARSEME COST Action[1] (2013–2017), which aimed at better understanding the nature of MWEs in order to improve their support in natural language applications. Two main challenges were considered: LINGUISTIC PRECISION (how to account for the highly heterogeneous nature of MWEs in linguistic resources and treatments?) and COMPUTATIONAL EFFICIENCY (how to deal with MWEs' idiosyncratic properties within reliable applications?).

To contribute to meeting these two challenges, PARSEME was based on four Working Groups (WGs):

- WG1 focused on the Grammar/Lexicon interface and the design of interoperable MWE lexicons,
- WG2 aimed at developing parsing techniques for MWEs,
- WG3 studied hybrid (e.g. symbolic and/or statistical) NL applications dealing with MWEs (e.g. MWE detection, machine translation, etc.),
- WG4 was concerned with the annotation of MWEs within treebanks.

This book has been created within WG2. It consists of contributions related to the definition, representation and parsing of MWEs. These contributions were collected via an open call for chapters. Each Chapter proposal was reviewed by 2 members of the editorial board. Out of this reviewing, 10 proposals were selected. They reflect current trends in the representation and processing of MWEs. They cover various *categories* of MWEs such as verbal, adverbial and nominal MWEs, various *linguistic frameworks* (e.g. tree-based and unification-based grammars), various *languages* including English, French, Modern Greek, Hebrew, Norwegian), and various *applications* (namely MWE detection, parsing, automatic translation) using both symbolic and statistical approaches.

[1] http://www.cost.eu/COST_Actions/ict/IC1207

2 Outline of the book

The book is organized as follows.

Part 1: MWE representations

The first part of the volume (Chapters 1 to 5) is dedicated to the study of MWE properties and representations.

In Chapter 1, Lichte et al. (2019 [this volume]) discuss the representation of MWEs within lexicalised formalisms. In particular, they show how the eXtensible MetaGrammar (XMG2) formalism offers a natural encoding of MWEs, which allows us to account for the fact that irregularities exhibited by MWEs are a matter of scale rather than binary properties.

In Chapter 2, Sheinfux et al. (2019 [this volume]) study a specific type of MWEs (namely verbal MWEs), focusing mostly on Hebrew, and show that unlike what previous work suggests, flexibility of verbal MWEs is not a discrete concept but rather a continuous property. They propose a new classification of MWEs which is based on semantic notions.

In Chapter 3, Dyvik et al. (2019 [this volume]) present the analysis of MWEs in an LFG grammar for Norwegian, NorGram, which is used in the construction of NorGramBank, a treebank of parsed sentences. The chapter describes how classes of MWEs are analysed by means of LFG templates, which capture the lexical and syntactic properties of MWEs in a succinct way.

In Chapter 4, Markantonatou et al. (2019 [this volume]) present a grammar of Modern Greek in the LFG formalism. Their grammar has been implemented with the Xerox Linguistic Engine (XLE), a grammar editor which also includes a parsing engine. In their Chapter, the authors pay a particular attention to the use of a pre-processor to detect and annotate MWEs prior to parsing.

In Chapter 5, Angelov (2019 [this volume]) presents the Grammatical Framework, a description language for developing NLP multilingual resources, and its application to some classes of MWEs. In particular, the author shows how to define MWE-aware multilingual grammars, which can be used for instance for in-domain machine translation.

Part 2: MWE parsing

The second part of the volume (Chapters 6 to 8) focuses on MWE parsing, that is, on the automatic construction of deep representations of the syntax of MWEs. Two main approaches to parsing coexist: the data-driven approach aims at extracting syntactic information from corpora using Machine Learning techniques

and is discussed in Chapter 6. The knowledge-based approach relies on the encoding of linguistic properties of MWEs within lexical entries, which are used by a parsing algorithm to compute the expected syntactic structure. The impact of MWE detection on such parsing algorithms is discussed in Chapters 7 (for a categorial parser) and 8 (for an attachment-rule-based parser).

In Chapter 6, Constant et al. (2019 [this volume]) give a detailed overview of various ways to extend statistical parsing with MWE identification, either during parsing or as a pre- or post-processing step. These extensions are compared and their evaluation discussed.

In Chapter 7, de Lhoneux et al. (2019 [this volume]) extend a CCG parsing architecture for English with a module for detecting MWEs and pre-process them. The effect of this pre-processing is evaluated in terms of parsing accuracy when (i) the parser is trained on pre-processed data (so-called training effect) and (ii) the parser uses information from pre-processed data (so-called parsing effect).

In Chapter 8, Foufi et al. (2019 [this volume]) investigate the extension of a knowledge-based parser with collocation identification. They apply this extension to the description of MWEs for various languages (including English and Greek), and show how it improves parsing efficiency in terms of percentages of complete analyses.

Part 3: Multilingual NL applications for MWEs

Finally, in the third part of the volume (Chapters 9 and 10), multilingual MWE acquisition techniques are presented.

In Chapter 9, Semmar et al. (2019 [this volume]) present three techniques for word alignment between parallel corpora and their application to MWEs. The bilingual MWE lexicons built using these techniques are then evaluated according to their effect on phrase-based statistical machine translation. The authors empirically show that MWE-aware lexicons improve translation quality.

Finally, in Chapter 10, Jacquet et al. (2019 [this volume]) present an architecture which allows for the identification of multiword entities (organizations, medical terms, etc.) within large collections of texts, together with the linking of monolingual variants of a given multiword entity, and of groups of variants accross multiple languages. Their architecture is evaluated against data from *Wikipedia*.

3 Acknowledgments

We are grateful to the COST framework of the European Union for their support for the PARSEME Action.

We would like to warmly thank Agata Savary and Adam Przepiórkowski, respectively chair and vice-chair of PARSEME, for their commitment to this action. They made it a dynamic environment, where researchers can have fruitful discussions and exchange ideas, leading to long-term collaborations.

We are grateful to Manfred Sailer, who, as a member of the editorial board of the *Phraseology and Multiword Expressions* series, accompanied us throughout the publication process.

We would like to thank the reviewers of this volume:

- Doug Arnold, University of Essex, UK
- Gosse Bouma, University of Groningen, the Netherlands
- Svetla Koeva, Bulgarian Academy of Sciences, Bulgaria
- Cvetana Krstev, University of Belgrade, Serbia
- Ana R. Luís, University of Coimbra, Portugal
- Stella Markantonatou, Institute for Language and Speech Processing/Athena RIC, Greece
- Petya Osenova, Bulgarian Academy of Sciences, Bulgaria
- Carla Parra Escartín, Dublin City University, ADAPT Centre, Ireland
- Victoria Rosén, University of Bergen, Norway
- Michael Rosner, University of Malta, Malta
- Manfred Sailer, University of Frankfurt am Main, Germany
- Agata Savary, University of Tours, Blois, France
- Veronika Vincze, University of Szeged, Hungary
- Shuly Wintner, University of Haifa, Israel

We are grateful for their valuable evaluations, comments and feedback, and to the proofreaders for their thorough work. Without their help, this book would not exist.

Special thanks go to Language Science Press (especially Sebastian Nordhoff and Stefan Müller for their continuous help and their engagement in the promotion of high-quality peer-reviewed open-access publication).

Yannick Parmentier and Jakub Waszczuk, Feb. 2019

References

Angelov, Krasimir. 2019. Multiword expressions in multilingual applications within the Grammatical Framework. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 127–146. Berlin: Language Science Press. DOI:10.5281/zenodo.2579041

Constant, Mathieu, Gülşen Eryiğit, Carlos Ramisch, Mike Rosner & Gerold Schneider. 2019. Statistical MWE-aware parsing. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 147–182. Berlin: Language Science Press. DOI:10.5281/zenodo.2579043

de Lhoneux, Miryam, Omri Abend & Mark Steedman. 2019. Investigating the effect of automatic MWE recognition on CCG parsing. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 183–215. Berlin: Language Science Press. DOI:10.5281/zenodo.2579045

Dyvik, Helge, Gyri Smørdal Losnegaard & Victoria Rosén. 2019. Multiword expressions in an LFG grammar for Norwegian. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 69–108. Berlin: Language Science Press. DOI:10.5281/zenodo.2579037

Foufi, Vasiliki, Luka Nerima & Eric Wehrli. 2019. Multilingual parsing and MWE detection. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 217–237. Berlin: Language Science Press. DOI:10.5281/zenodo.2579047

Jacquet, Guillaume, Maud Ehrmann, Jakub Piskorski, Hristo Tanev & Ralf Steinberger. 2019. Cross-lingual linking of multi-word entities and language-dependent learning of multi-word entity patterns. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 269–297. Berlin: Language Science Press. DOI:10.5281/zenodo.2579049

Lichte, Timm, Simon Petitjean, Agata Savary & Jakub Waszczuk. 2019. Lexical encoding formats for multi-word expressions: The challenge of "irregular" regularities. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 1–33. Berlin: Language Science Press. DOI:10.5281/zenodo.2579033

Markantonatou, Stella, Niki Samaridi & Panagiotis Minos. 2019. Issues in parsing MWEs in an LFG/XLE framework. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 109–126. Berlin: Language Science Press. DOI:10.5281/zenodo.2579039

Semmar, Nasredine, Christophe Servan, Meriama Laib, Dhouha Bouamor & Morgane Marchand. 2019. Extracting and aligning multiword expressions from parallel corpora. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 239–268. Berlin: Language Science Press. DOI:10.5281/zenodo.2579047

Sheinfux, Livnat Herzig, Tali Arad Greshler, Nurit Melnik & Shuly Wintner. 2019. Verbal multiword expressions: Idiomaticity and flexibility. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 35–68. Berlin: Language Science Press. DOI:10.5281/zenodo.2579035