Johanna Monti, Gloria Corpas Pastor, Ruslan Mitkov & Carlos Manuel Hidalgo-Ternero (eds.): *Recent Advances in Multiword Units in Machine Translation and Translation Technology*. Amsterdam, Philadelphia: John Benjamins Publishing Company, 2024. 261 pp. ISBN 9027246386 / E-ISBN 9789027246387

The prolific age of automation has significantly impacted multiple language domains and disciplines, reshaping both the workplace and academia. However, the phraseological richness that underpins language still poses major challenges to modern technological systems, which have been the subject of study by numerous scholars for several decades. Driven by this increasing research interest in computational phraseology, the edited volume *Recent Advances in Multiword Units in Machine Translation and Translation Technology* aims at exploring some of the latest developments in the field of computational treatment of multiword expressions (MWEs).

The present volume, edited in 2024 by Johanna Monti, Gloria Corpas Pastor, Ruslan Mitkov, and Carlos Manuel Hidalgo-Ternero, leading scholars in the field of computational linguistics, corpus linguistics, natural language processing, and related disciplines (see, for example, Corpas Pastor and Colson 2020; Dell'Orletta, Monti and Tamburini 2020; Corpas Pastor, Bautista Zambrana and Hidalgo-Ternero, 2021; Mitkov 2022). This volume is part of a series of books edited by the same specialists, continuing their long-standing and influential contributions to computational phraseology and linguistics (see Monti, Mitkov, Corpas Pastor and Seretan 2013; Corpas Pastor, Monti, Seretan and Mitkov 2016; Monti, Corpas Pastor and Seretan 2018). The 261-page book, published by John Benjamins Publishing Company, includes groundbreaking chapters that cover multiple specialised topics and innovative approaches in phraseology, with special emphasis on machine translation (MT). Therefore, due to its interdisciplinary nature, this book will certainly be of great interest to language professionals and researchers alike, for example, translators, interpreters, linguists, lexicographers, and language teachers.

After a comprehensive table of contents, a detailed preface, and acknowledgements authored by the editors, the book contains thirteen chapters. It is divided into two main sections: the first includes five chapters and the second the remaining eight. The first section, titled "Computational treatment of multiword units", examines general and specialised phraseological translation as rendered by notorious neural machine translation (NMT) systems. It also proposes innovative methods to effectively improve the quality of these systems when translating both multiword units (MWUs) and multiword terms (MWTs). The second section, "Corpus-based and linguistic studies in phraseology", studies MWUs and MWTs in specialised domains or under different linguistic perspectives. It also presents cutting-edge methods for the NMT-oriented treatment of MWEs, as well as proposals for new models designed for phraseological applied purposes.

The first contribution in the book is by Jean-Pierre Colson, titled "Multi-word units in neural machine translation: Why the tip of the iceberg remains problematic". Colson compares a corpus in French with a corpus of translations made by DeepL and Google Translate from English into French to automatically evaluate the translated phraseology. Although NMT quality has significantly improved, the author concludes that NMT systems cannot yet fully capture complex contextual features when translating MWUs; however, he also emphasises the versatility of the methodology presented, highlighting its multiple applications in the professional world.

In the second chapter, "ReGap: A text-preprocessing algorithm to enhance MWE-aware neural machine translation systems", Carlos Manuel Hidalgo-Ternero and Gloria Corpas Pastor undertake a study with flexible verb-noun idiomatic construction using the algorithm ReGap to test the performance of DeepL when dealing with discontinuous MWEs in the Spanish>English and Spanish>Germain language pairs. Their findings prove that ReGap enhances NMT's performance for the somatisms in both languages examined, raising awareness of the challenges posed by discontinuous MWEs and the potential for optimising other NMT systems.

In the third contribution, "Evaluating the Italian-English machine translation quality of MWUs in the domain of archaeology", Giulia Speranza and Johanna Monti manually assess the quality of 100 machine-translated MWUs rendered by three NMT systems: Google Translate, DeepL, and Microsoft Bing Translator, both out-of-context and in-context. The error analysis of NMT output shows that these systems do not often deliver satisfactory results; mistranslations mostly happen when the internal component of the MWU term is a common or polysemous word. However, NMT systems perform better when the MWU is in-context.

Following research on specialised domains, Natalie Kübler, Hanna Martikainen, Alexandra Mestivier, and Mojca Pecman authored the fourth contribution, titled "Post-editing neural machine translation in specialised languages: The role of corpora in the translation of phraseological structures". With the help of specialised corpora, they conduct thorough research into the three typical errors committed by translation students when post-editing phraseology in the English-French language pairs. This experimental research achieves several teaching purposes, helping learners to identify errors while also highlighting the limitations and strengths of MT and corpora in translating and post-editing phraseology.

In the last chapter of the first section, titled "Evaluating a bracketing protocol for multiword terms", Pilar León-Araúz and Melania Cabezas-García propose multiple bracketing protocols applied to a set of three-constituent MWTs for querying corpora in order to reduce translation difficulties, notably disambiguation. Their comprehensive study unravels the most productive bracketing rules for these purposes: adjacency, longer MWTs, dependency, shortening, and paraphrases. In addition, it highlights that domain-specific corpora are more efficient for these queries, concluding that corpus size is not that relevant, except for computational purposes.

The second section begins with the chapter titled "Suggestions for a new model of functional phraseme categorization for applied purposes", in which Anna Fankhauser highlights the need among language practitioners and English as a Foreign Language (EFL) learners related to phraseme categorisation. Thus, the author proposes a new model for this purpose based on a spoken British and American English corpora: BEspoken corpus and AEspoken corpus. Her comprehensive classification model describes relevant MWEs, highlighting its applications not only in educational spheres but also in professional environments.

In the seventh contribution, "Verb collocations and their semantics in the specialised language of science", Eva Lucía Jiménez-Navarro studies the verb collocations found in an ad hoc specialised corpus of research articles. After semantically classifying the data, the author compares it with her previous study on noun collocations. The similar semantic frames (55 in total) between noun and verb collocations in scientific English are presented in her findings. In addition, the study suggests that the different methodologies followed are both complementary and necessary for the advancement of this specialised topic.

The eighth chapter, titled "Negative–positive adjective pairing in travel journalism in English, Italian, and Polish", delves into language patterns of specialised phraseology in several language pairs. For this purpose, David Brett, Antonio Pinna and Barbara Loranc carefully examine the ADJ+but+ADJ pattern in an English corpus of tourism domain, compiling two corpora in Italian and Polish. Their quantitative analyses conclude that the ADJ+but+ADJ construction is typically found in travel journalism, although the data might be interpreted differently depending on the corpus. The Negative-Positive sequence is the most common pattern in all three languages.

In the ninth contribution, "The middle construction and some machine translation issues: Exploring the process of compositional cospecification in quality-oriented middles", Macarena Palma Gutiérrez investigates how the *qualia* structure works differently in middle constructions with the English verbs *cut* and *drive* when machine-translated. The study examines more than 500 instances that follow the Adverb+Verb collocation with quality-oriented adjuncts and several entities. The author's findings enhance the computational handling of machine-translated MWEs by exploring different patterns.

In the context of specialised translation, Juan Rojas-Garcia explores the semantic analysis of predicative-argument structure from the perspective of Frame-based Terminology in named rivers expressed as both single and multiword terms. The chapter is titled "Semantic annotation of named rivers and its application for the prediction of multiword-term bracketing", and it investigates the influence of the semantic information encoded in specific sentences on the bracketing of a three-component MWT. The study evidences a correlation between predicates in the same lexical domain and their tendency to combine with identical or closely related semantic categories, suggesting that propositional representations may have practical applications in semantic equivalence for future MT applications.

In the eleventh chapter, "Irony in American-English tweets: A cognitive and phraseological analysis", Beatriz Martín-Gascón carries out an in-depth cognitive and qualitative analyses of ironic utterances found in Twitter, comparing American-English users with Spanish users. The author highlights the difficulty of perceiving irony in these texts and the importance of integrating both non-linguistic and linguistic cues. The results show the potential of the analysis as a pedagogical resource to raise awareness in future second language learners of the mental mechanisms and linguistic representations involved in irony.

In the twelfth chapter, "A comprehensive Japanese MWE lexicon: JMWEL", Masahito Takahashi, Toshifumi Tanabe, Jack Halpern, and Kosho Shudo introduce the newest version of their NLP-oriented syntactic lexicon of Japanese MWEs, JMWEL. The paper provides an in-depth explanation of the organisation of JMWEL and carefully describes the selection of the MWEs and the information encoded. Finally, the authors highlight the remarkable features of the JMWEL, such as its large variety of MWE categories, its detailed morpho-syntactic structures, and the orthographic and syntactical variety of each MWE.

In the last chapter, titled "Ontology-based formalisation of Italian clitic verbal MWEs: An approach for supporting machine translation", Maria Pia di Buono, Johanna Monti and Valeria Caruso suggest a bilingual resource, specifically an ontology-based bilingual lexicographic resource composed of Italian clinic Verbal MWEs. Due to the inherent ambiguity of VMWE, the study displays the main translation challenges in the Italian>English language pair. In their conclusions, the authors suggest that this resource could be applied to the automatic translation of clinic verbs, thereby supporting MT and NLP of MWEs.

In short, the book brings together a perfect combination of qualitative and quantitative studies, covering multiple languages, topics, and approaches that truly address current and pressing phraseological challenges. In addition, the methodologies presented are rigorously followed, demonstrating the high level of quality and scientific excellence of the contributions. Since the chapters are multidisciplinary, the book can be very intellectually stimulating for both professionals and scholars in multiple language fields. Undoubtedly, this volume not only gathers insights of recognised scholars in the field but also paves the way for further study on the computational treatment of MWEs, alongside corpus-based and linguistic research on phraseology.

References

Corpas Pastor, Gloria, Bautista Zambrana, María del Rosario & Carlos Manuel Hidalgo-Ternero. 2021. *Sistemas fraseológicos en contraste: enfoques computacionales y de corpus*. Granada: Comares.

Corpas Pastor, Gloria, Johanna Monti, Violeta Seretan & Ruslan Mitkov (eds.). 2016. *Proceedings of the workshop on multi-word units in machine translation and translation technology*, Malaga, Spain, 1–2 July 2015. Genève: Tradulex.

Dell'Orletta, Felice, Johanna Monti & Fabio Tamburini (eds.). 2020. *Proceedings of the seventh Italian conference on computational linguistics CLiC-it 2020*. Torino: Accademia University Press. https://doi.org/10.4000/books.aaccademia.8203

Mitkov, Ruslan. 2022. *The Oxford handbook of computational linguistics*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199573691.001.0001

Mitkov, Ruslan, Johanna Monti, Gloria Corpas Pastor & Violeta Seretan (eds.). 2018. *Multiword units in machine translation and translation technology* (Current issues in linguistic theory, 341). Amsterdam & Philadelphia: John Benjamins.

Monti, Johanna, Ruslan Mitkov, Gloria Corpas Pastor & Violeta Seretan (eds.). 2013. *Proceedings of the 14th machine translation summit: Workshop on multi-word units in machine translation and translation technologies*. Allschwil: European Association for Machine Translation.

Laura Noriega-Santiáñez
Correspondence address: laura.noriega@uma.es