Review of Monti, Corpas Pastor, Mitkov, and Hidalgo-Ternero (Eds.) (2024): *Recent Advances in Multiword Units in Machine Translation and Translation Technology*. Amsterdam and Philadelphia: John Benjamins Publishing Company. 264pp.

Simon Copet

University of Mons | Université libre de Bruxelles

The book *Recent Advances in Multiword Units in Machine Translation and Translation Technology* highlights the particular interest of multiword units (or multiword expressions) in automatic language processing, and hence in machine translation. The book also shows the importance of conducting research in this area, which is proving useful for professionals in the sector, be they translators, lexicographers or language teachers. This book has been edited by Johanna Monti (L'Orientale University of Naples), Gloria Corpas Pastor (IUITLM, University of Malaga), Ruslan Mitkov (Lancaster University) and Carlos Manuel Hidalgo-Ternero (IUITLM, University of Malaga), all renowned researchers in the field of automatic language processing. The 13-chapter book is divided into two main sections: 'Computational treatment of multiword units' (5 chapters) and 'Corpus-based and linguistic studies in phraseology' (8 chapters).

In the inaugural chapter (Chap. 1: *Multi-word units in neural machine translation Why the tip of the iceberg remains problematic*), Jean-Pierre Colson analyzes phraseology in *Google Translate* and *DeepL* translations via corpus methods, showing that neural machine translation (NMT) systems produce phraseological units at rates rivaling or exceeding human output. While quantitatively impressive, qualitative flaws in contextual accuracy reveal limitations, suggesting NMT's competence in handling multiword expressions remains superficially effective but contextually flawed. The decision to position this article at the

commencement of the book is particularly perspicacious, as Colson methodically guides the reader through an exploration of the concept of MWE, while concurrently providing a comprehensive exposition of machine translation systems. It is noteworthy that this chapter could be seamlessly integrated into courses pertaining to translation or NLP.

The second article (Chap. 2: *ReGap A text-preprocessing algorithm to enhance MWE-aware neural machine translation systems*), by Carlos Manuel Hidalgo-Ternero and Gloria Corpas Pastor, describes a text-preprocessing algorithm designed to automatically detect and convert discontinuous multiword expressions (MWEs) into their canonical (continuous) form to enhance NMT systems. They evaluate ReGap's effectiveness using verb-noun idiomatic constructions (VNICs) in Spanish, testing its impact on *DeepL*'s performance for Spanish-to-English and Spanish-to-German translations. The promising results demonstrate that ReGap improves MWE recognition in NMT, suggesting new strategies for developing systems better equipped to handle linguistically complex, discontinuous expressions. In doing so, this contribution highlights the interest of this work for more technically savvy readers, although it remains accessible to the wider audience and offers a comprehensive bibliography that will prove useful to all backgrounds.

The third chapter (Chap. 3: *Evaluating the Italian-English machine translation quality of MWUs in the domain of archaeology*), by Giulia Speranza and Johanna Monti, analyze the challenges multiword units (MWUs) pose for NMT systems, focusing on the specialized domain of archaeology for the Italian-English language pair. Using a Gold Standard of 100 domain-specific MWUs, they manually evaluate out-of-context and in-context translations from *Google Translate*, *DeepL*, and *Microsoft Bing Translator*, comparing them to human reference translations. Findings reveal persistent issues in translating terminology-rich MWUs, though contextual input occasionally improves accuracy, suggesting that NMT systems may adapt better to domain-specific phraseology with enriched contextual cues. This

study underscores both current limitations and context-dependent potential in NMT's handling of complex linguistic constructs.

In chapter four (Chap. 4: *Post-editing neural machine translation in specialised languages The role of corpora in the translation of phraseological structures*), Natalie Kübler, Hanna Martikainen, Alexandra Mestivier and Mojca Pecman examine phraseology in specialized texts as well as challenges faced by students in post-editing NMT outputs through a corpus-based methodological framework. By analyzing recurrent student errors linked to phraseology, the authors propose a descriptive error typology to develop targeted pedagogical materials. The research aims to demonstrate that systematic training in corpus querying and data interpretation enhances students' ability to improve translations and post-edited texts, particularly in specialized domains. The findings advocate for integrating corpus literacy into training programs to address phraseological complexities in both human translation and machine translation post-editing workflows. This contribution is truly enriching in that it demonstrates a genuine practical application of their corpus to translator training, in particular by highlighting the most frequent errors, which echoes a great deal of research in the field.

The fifth chapter (Chap. 5: *Evaluating a bracketing protocol for multiword terms*), written by Pilar León-Araúz and Melania Cabezas-García, addresses the challenge of parsing multiword terms (MWTs) in scientific texts, particularly three-constituent units with opaque internal relations. Existing NLP models for dependency analysis (bracketing) remain insufficient, prompting the authors to develop a hybrid protocol merging multiple approaches. The method evaluates disambiguation rules by their effectiveness in resolving structural ambiguities and their corpus retrieval reliability, while also assessing how corpus characteristics (size, text type) influence bracketing accuracy. The research advances computational strategies for interpreting complex MWTs, which are critical for enhancing technical text processing.

The sixth chapter (Chap. 6: *Suggestions for a new model of functional phraseme categorization for applied purposes*), by Anna Fankhauser, critiques existing phraseme categorization models in applied linguistics (translation, teaching, lexicography) for inadequately addressing practitioners' and learners' needs. Proposing a functionally driven classification framework, the author grounds her model in a corpus analysis of spoken British and American English. By prioritizing functional properties over traditional criteria, the approach aims to generate a systematic inventory of phrasemes directly applicable to real-world language tasks (e.g. production, and comprehension). The empirical, corpus-based design seeks to maximize practical utility, bridging gaps between theoretical categorization and the demands of language pedagogy and professional practice.

The seventh article (Chap. 7: *Verb collocations and their semantics in the specialized language of science*), by Eva Lucía Jiménez-Navarro, examines verb collocations in scientific discourse, analyzing their semantic roles in evoking domain-specific topics. Using a specialized corpus, collocations are extracted, manually refined, and categorized. Compared to prior work (Jiménez-Navarro 2019) on noun collocations, findings show shared semantic frameworks but methodologically distinct yet complementary insights, underscoring verbs' unique role in structuring scientific knowledge. While confirming the accessible and novel/inspiring nature of these contributions, readers would also have benefitted from further details into the composition/compilation of the corpus for this study.

In the eighth chapter (Chap. 8: *Negative–positive adjective pairing in travel journalism in English, Italian, and Polish*), David Brett, Antonio Pinna and Barbara Loranc explore the ADJ+but+ADJ pattern in tourism discourse, analyzing its use across English, Italian, and Polish travel journalism corpora. Building on prior research highlighting adjectives' role in tourism phraseology, the experiment investigates whether this contrastive structure (e.g. "remote but accessible") is language-specific or a universal discourse strategy

within the register. Findings reveal the pattern's cross-linguistic prevalence, suggesting that it functions as a shared rhetorical tool in travel texts, transcending linguistic boundaries to balance positive and negative descriptors effectively.

The ninth chapter (Chap. 9: *The middle construction and some machine translation issues Exploring the process of compositional cospecification in quality-oriented middles*), by Macarena Palma Gutiérrez, delves into verb collocations in scientific texts, focusing on how their semantic structure reflects and shapes domain-specific topics. The author constructs a specialized corpus of research articles, extracts collocation candidates algorithmically, refines them manually, and classifies them semantically. A comparison with noun collocations from earlier research reveals shared semantic frameworks (e.g. cause-effect and process-description), but methodological divergences–verbs emphasize dynamic interactions, while nouns highlight conceptual entities. The study demonstrates that verb collocations offer complementary insights into scientific discourse, revealing how linguistic patterns encode and organize domain knowledge. This methodological workflow bridges corpus analysis and semantic theory, underscoring the role of collocations in domain-specific knowledge structuring.

In the tenth chapter (Chap. 10: *Semantic annotation of named rivers and its application for the prediction of multiword-term bracketing*), Juan Rojas-Garcia investigates knowledge acquisition in specialized translation by focusing on two interconnected aspects. It applies frame-based terminology to semantically annotate predicate-argument structures involving named rivers, uncovering cognitive frameworks that govern their use in specialized discourse. Concurrently, it examines how contextual semantic data can predict the bracketing of three-component multi-word terms, successfully developing machine-learning models for this task. The research underscores the value of integrating semantic analysis into

terminological resources and computational tools, demonstrating its potential to refine the representation and processing of specialized phraseology in translation workflows.

The eleventh article (Chap. 11: *Irony in American-English tweets A cognitive and phraseological analysis*), by Beatriz Martín-Gascón, provides a compelling analysis of verbal irony on Twitter through a cognitive linguistics framework, comparing American-English and Spanish speakers. Using Ruiz de Mendoza's echoic theory (Ruiz de Mendoza 2017), the research examines 495 tweets from a vast dataset, identifying key contextual markers like hashtags, emojis, and punctuation that aid irony detection. Findings reveal that American-English speakers predominantly use explicit-echoic irony with clear markers, while Spanish tweets incorporate features like vowel elongation and metaphorical mappings. Political irony, particularly about Trump, is the most frequent. The study's innovative combination of cognitive linguistics and big data offers valuable insights into the cultural and cognitive dimensions of irony, with promising applications for second-language instruction. A deeper exploration of cultural influences and expanded pedagogical strategies could further enhance its impact.

In the twelfth chapter (Chap. 12: *A comprehensive Japanese MWE lexicon JMWEL*), Masahito Takahashi, Toshifumi Tanabe, Jack Halpern, and Kosho Shudo present the latest version of the Japanese Multiword Expressions Lexicon (JMWEL), a comprehensive resource addressing the challenges of Japanese MWEs, including idioms, collocations, and proverbs. Developed over decades, JMWEL contains approximately 160,000 lemmas with detailed syntactic and morphological annotations for use in linguistics, computational linguistics, and machine translation. By organizing MWEs into thematic sub-lexicons and encoding dependency structures, the lexicon enhances parsing, semantic analysis, and NLP tasks, particularly machine translation. This well-structured and accessible resource meets a critical need in Japanese lexicography and is valuable even for non-Japanese speakers.

The final chapter of this volume (Chap. 13: *Ontology-based formalisation of Italian clitic verbal MWEs An approach for supporting machine translation*), written by Maria Pia di Buono, Johanna Monti and Valeria Caruso, is devoted to an ontology-based formalization of Italian clitic verbal multiword expressions (VMWEs). The authors begin their study with a brief but effective contextualization of developments in the performance of machine translation systems and highlight the difficulties still posed by MWEs, particularly for Italian clitics, due to their ambiguity. The main aim of their study is to propose 'a resource for this language-specific category according to the OntoLex-Lemon model' (p. 244), the function of which is to improve the performance of NMT systems through a reinjection process. To develop this lexicographic resource, they first analyze the features of clitic verb classes and evaluate the translation challenges faced by existing MT systems, such as *Google Translate*. Subsequently, they describe the morphological, syntactic, and semantic aspects to be included in the ontology-based formalization. Finally, they propose a formalization of clitic VMWEs accompanied by usage examples and translations, leveraging the OntoLex-Lemon model to construct a bilingual lexicographic resource. This contribution is of particular interest, innovative and worthy of further investigation, given the scarcity of literature on this specific subject. Furthermore, the relevant avenues for further research proposed by the authors provide compelling justification for further investigation into this topic.

All in all, the only drawback of *Recent Advances in Multiword Units in Machine Translation and Translation Technology* is perhaps the restrictive scope of its title, as its content and applications go far beyond the world of translation technology. Indeed, the sections ‘Computational treatment of multiword units’ and ‘Corpus-based and linguistic studies in phraseology’ cover a wide range of disciplines, such as pragmatics, semantics, corpus linguistics, computational linguistics, machine translation and many others. However, the wide range of fields covered in this book does not prevent it from being an extremely

valuable reference source for any researcher interested in MWUs, but also for language learners, teachers and professionals.

References

Jiménez-Navarro, E.L. 2019. “Nominal collocations in scientific English: A frame-semantic approach”. In *Computational and corpus-based phraseology*, edited by G. Corpas Pastor and R. Mitkov, 187-199. Cham: Springer Nature Switzerland AG.

Ruiz de Mendoza, F.J. 2017. “Cognitive modeling and irony”. In *Irony in language use and communication*, edited by H. Colson and A. Athanasiadou, 179–200. Amsterdam: John Benjamins.

Simon Copet

simon.copet@ulb.be