RiCL Research in Corpus Linguistics

Review of Izquierdo, Marlén and Zuriñe Sanz-Villar eds. 2023. *Corpus Use in Cross-linguistic Research: Paving the Way for Teaching, Translation and Professional Communication*. Amsterdam: John Benjamins. ISBN: 978-9-027-21430-0. DOI: https://doi.org/10.1075/scl.113

Isabel Pizarro-Sánchez
University of Valladolid / Spain

The volume *Use in Cross-Linguistic Research: Paving the Way of Teaching Translation and Professional Communication*, edited by Izquierdo and Sanz-Villar, provides an in-depth overview of the diverse applications of corpora in cross-linguistic studies. The book presents a collection of 12 studies that, through illustrative examples, emphasize the importance of parallel and comparable corpora in various linguistic fields, including translation studies, language teaching, and natural language processing.

The opening section, authored by Izquierdo and Sanz-Villar, serves as an introduction to the volume. It contextualizes the subsequent analysis and underlines the importance of cross-linguistic research within the broader framework of contrastive linguistics and translation studies. The authors acknowledge the significant value of parallel and comparable corpora in conducting empirical studies within the field. Furthermore, they include a brief overview of the 12 chapters that follow, providing a detailed account of each individual contribution to the field.

In chapter 1, Marco and Bracho Lapiedra test, and empirically validate, the Gravitational Pull Hypothesis (GPH) on Light Verb Constructions (LVCs), adopting an innovative approach in which they formulate their hypothesis at the level of LVC types rather than individual constructions. Their study focuses on how emotional states and dynamic events are represented in translations between English, French, Catalan, and Spanish. The authors analyse the predicative nouns that collocate with the light verbs

fer (Catalan) and *dar* (Spanish) and their equivalents into English (*make* and *give*) and French (*faire* and *donner*). They firstly categorize the nouns as either emotional states or dynamic events, and then examine their frequency and salience in both translated and non-translated texts. The analysis is based on collocations extracted from the *Corpus Valencià de Literatura Traduïda* (COVALT),[1] to provide observable evidence. The results largely confirm the authors' predictions, indicating an under-representation of LVCs conveying emotional states and no significant differences for those conveying dynamic events, with positive results observed in five out of the eight language pair and LVC type combinations. This research contributes to the understanding of translation practices and illustrates the value of corpus-based studies in testing linguistic hypothesis and in raising contrastive awareness of the translated language features within the context of the translation classroom. While the scope of the study is limited to specific language pairs and LVC types, it establishes a valuable foundation for future research.

The second chapter, by Rabadán, explores the challenges and solutions of translating English LVCs into Spanish. Based on data extracted from the parallel corpus *P-ACTRES 2.0*,[2] which includes fictional and non-fictional material, Rabadán investigates how the semantic features and combinatorial capabilities of LVCs influence translation choices. Additionally, she delves into register-based variations between fictional and non-fictional texts. Through a systematic approach, the study examines a sample of the concordances of five English light verbs: *have*, *take*, *make*, *do*, and *give*. The selected sample is representative of the verbs and registers under study. The use of *P-ACTRES 2.0* provides robust empirical data, thus enabling a detailed examination of translation patterns and semantic features. The results reveal five recurrent translation patterns, with preference for full lexical verbs and single correlate verbs. This corroborates the hypothesis that the semantic features of LVCs significantly influence translation choices. Additionally, variations related to register are observed, indicating different translation strategies for fictional and non-fictional texts. These findings are interpreted in the context of translation studies and semantic theory, emphasizing the importance of understanding the semantic compatibility of LVCs in order to improve translation accuracy and consistency. Furthermore, the chapter discusses the implications for machine translation and bilingual writing support tools, outlining the potential applications of its results in enhancing machine translation systems, post-

[1] https://www.covalt.uji.es/en/
[2] https://actres.unileon.es/wp/parallel-corpora/

editing aids, and authoring support applications. Rabadán's rigorous analysis of LVCs and their translations is a significant contribution to the field, offering valuable insights into LVC semantic compatibility and translation strategies.

In Chapter 3, Molés-Cases investigates the translation from Spanish into German of manner-of-speaking expressions in narrative texts. Based on a subcorpus of ten contemporary novels from the *Parallel Corpus German Spanish* (PaGeS),[3] the research describes the translation techniques used for reporting verbs that introduce direct speech and examines the differences in translation when dealing with motion and speech domains. By analysing 1,571 bilingual concordances, the author explores how translators approach the typological differences between a verb-framed language (Spanish) and a satellite-framed language (German), with a particular focus on whether manner-of-speaking expressions are preserved or adapted in translation. The results of the research indicate that manner-of-speaking is largely maintained in translations, with transference being the predominant translation technique. Interestingly, the study also reveals that manner-of-motion is frequently added to German translations from Spanish. This practice, however, is notably less prevalent for manner-of-speaking, suggesting that typological differences do not appear to be a significant factor in the translation of the speech domain. The author also observes similar diversity in the use of manner-of-speaking verbs between Spanish and German versions, which is indicative of a unique behaviour of German within satellite-framed languages. The study offers valuable insights into the translation of reporting verbs and their lexical diversity in verb-framed and satellite-framed languages, despite being constrained to narrative texts and a specific language combination.

Sánchez Nieto's chapter deals with the translation of the German dative passive, a grammatical construction that is prevalent in German but does not exist in Spanish. The study examines the extent to which translators maintain the recipient perspective and the translation techniques employed in both German to Spanish and Spanish to German translations. Using a raw sample of texts from PaGeS, Sánchez Nieto aims to analyse frequencies, semantic roles, and translation techniques of the dative passive forms *bekommen*, *kriegen*, and *erhalten.* The sample is tagged and queried in *Sketch Engine* to retrieve paragraphs that include examples of the three variants of the dative passive.[4]

[3] https://www.corpuspages.eu/corpus/about/about?lang=en
[4] https://www.sketchengine.eu/

Following manual cleaning, the examples were imported into *ATLAS.ti* for marking and further qualitative analysis.[5] Results show that the *bekommen* passive is the most frequent, while the *erhalten* passive is not present. The data also reveal that in approximately two-thirds of the translations from German into Spanish the recipient perspective is maintained, with the remaining favouring the agent perspective. Simplification is identified as the most common translation technique and, in Spanish to German translations, the dative passive is typically employed when the recipient perspective is present in the source text. The results are interpreted in the light of the Thinking-for-Translating hypothesis, suggesting that translators adapt their strategies to align with the rhetorical style of the target language. The chapter's findings have practical implications for translation training and the development of contrastive competence.

Chapter 5, by Ramón, investigates the semantic differences between the near-synonyms English ingressive verbs *begin* and *start,* using translation corpora. This is achieved through a comprehensive cross-linguistic analysis of the parallel concordances of the lemmas *begin* and *start,* all retrieved from *P-ACTRES*, which includes 4.2 million words of English-Spanish translations across various registers. The analysis of the translations provides empirical evidence of the semantic differences between the two near-synonymous verbs. The results indicate that *begin* is more frequently followed by to-infinitive clauses and is associated with the initiation of actions. In contrast, *start* is more common with intransitive patterns and often implying the commencement of activities. In terms of their translations, the Spanish verbs *empezar* and *comenzar* are the most frequently used equivalents for both *begin* and *start.* However, the data indicate that *start* shows a wider range of translational equivalents using ingressive verbs in Spanish, in comparison to the near-synonym *begin*. This observation suggests a greater diversity of its sense relations, particularly in intransitive and transitive patterns. While considering the potential impact of target language factors, Ramón presents a compelling argument that systematic differences in translation equivalents may reveal subtle semantic differences in near-synonyms. The chapter's methodological rigour, combining quantitative analysis with qualitative interpretation of the translation choices, provides a solid basis for its conclusions.

In chapter 6, Labrador explores the complexities of core vocabulary and the use of

[5] https://atlasti.com

parallel corpora to gain insight into it, with the verb *run* serving as a case study. The English verb *run,* despite its apparent simplicity, is a fundamental tier-1 word in English with a large number of meanings and uses. This complexity presents a significant challenge for non-native speakers who intent to fully understand and employ it in their language production. The study analyses 926 occurrences of *run* in the *P-ACTRES* corpus, focusing on fictional and non-fictional texts. It applies an intra-linguistical approach to classify the uses of *run* in English and an inter-linguistic approach to analyse its Spanish translations. Factors such as syntactic structures, collocational patterns, and the expression of manner and path of motion events are considered. The findings reveal that *run* is more frequently used in fiction, particularly in expressions of motion, and it has a wide range of literal and metaphorical meanings, which are reflected in its Spanish translations. Different translation techniques are also identified, including crossed transposition, density change, and copying structure, each reflecting different aspects of the verb complexity. The chapter presents practical suggestions for implementing corpus-informed teaching methods, emphasizing the importance of teaching the different uses of core vocabulary in order to improve learners' communicative skills by producing more idiomatic and natural language.

In chapter 7 Gutiérrez Lanza examines the process of synchronizing film dialogues for dubbing, with a focus on Conversational Makers (CMs). Her study deals with the adaptation process from draft translations to synchronized film scripts in the *Corpus of English-Spanish Cinema Scripts* (TRACEci),[6] comparing these with non-translated Spanish data retrieved from the *CORPES XXI subcorpus of guiones*.[7] The aim is to evaluate the influence of synchronization on CMs and determine whether the synchronization process results in a statistically significant overuse or underuse of CMs, thereby contributing to statistical dubbese. By analysing the frequency and use of CMs in different stages of translation and synchronization, Gutiérrez Lanza finds that CMs are overused in draft translations, and that there is a significant reduction in the use of CMs from draft translations to dubbed scripts, which reflects the adjustment made to meet lip-sync requirements. Some CMs such as *ehm*, *bueno*, *bien*, and *por supuesto* are overused in the dubbed scripts in comparison to the non-translated Spanish, confirming the presence of statistical dubbese (overuse). However, the overuse of certain CMs has been eliminated during the synchronization process, resulting in an overall improvement

[6] https://trace.unileon.es/es/fondos-trace/catalogos/textos-audiovisuales-cine-y-tv/
[7] https://www.rae.es/banco-de-datos/corpes-xxi

in translation quality. The challenge lies in maintaining the naturalness of the dubbed dialogues while simultaneously ensuring synchronization. These results are of interest to both the dubbing industry and translation training, as they contribute to a better understanding of the impact of CMs on translation quality.

In chapter 8, Hermosa-Ramírez investigates the linguistic characteristics of opera Audio Descriptions (AD) and Audio Introductions (AI) of opera scripts through corpus linguistics. By analysing scripts from the *Liceu Opera House* in Barcelona and the *Teatro Real* in Madrid, the research aims to situate opera AD and AI within the spoken-written language continuum. In order to achieve this objective, Hermosa-Ramírez analyses several linguistic measures, including lexical density, type-token ratio, mean word and sentence length, and the Flesch-Szigriszt readability index. The findings show that both AI and AD scripts share features with planned written language, particularly in terms of lexical density, and with spontaneous spoken language in lexical variation. However, AIs show longer mean sentence length and mean word length than ADs, which place them closer to written language. Readability scores provide further evidence to support this distinction, with AIs displaying more written language characteristics, whereas ADs, despite their formal structure, show a slight tendency towards the spoken language end due to the need for synchronization with visual elements. The findings underline the complex and multifaceted nature of these texts, which combine elements of both spoken and written language. The author concludes with valuable practical suggestions for applying her findings, including the potential for personalized AIs that may be adapted to diverse audiences. Hermosa-Ramírez's research makes a significant contribution to the audiovisual translation field, providing insights into the distinctive linguistic characteristics of opera AD and AI.

In chapter 9, Li describes the use of a multilingual parallel corpus, compiled for journalistic translation research, through a pilot case study that focuses on the national image construction in global news translation. The author employs the *New York Times Multilingual Parallel Corpus* (NYTMPC), which is a valuable resource for analysing how national images are constructed and reconstructed through the translation of English source news articles into Chinese and Spanish. The corpus consists of more than one million running words and 753 texts aligned at the paragraph level and manually annotated for headlines, leads (or subheadlines), publication date, news section, and translation shifts. To identify patterns in image construction, an analysis of

both the news headlines and the section labels is conducted. The research identifies several key topics that are more frequently associated with China in global news and contribute to the activation of specific national images, which are differently reconstructed across languages. These topics include international politics, COVID-19, economy, and technology. The work additionally analyses how these topics contribute to the activation of specific national images across languages by rephrasing news headlines and including or excluding particular news labels. This comprehensive analysis of the NYTMPC offers useful insights into the role of translation in shaping national images in global news. The findings have practical applications for the improvement of journalistic translation practice and the enhancement of the accuracy and objectivity of media representations.

Chapter 10, by Contarino and De Camillis, presents a comprehensive study on domain-adapting and assessing a machine translation engine for the unique variety of the German used in South Tyrol public institutions. The distinctive linguistic features of this variety demand the use of specialized translation tools. Previous studies on Machine Translation (MT) performance for South Tyrolean German point to significant challenges in accurately translating its legal terminology, thus, the need for adapting an MT system like ModernMT (MMT). In order to adapt the MMT, the *LEXB* corpus is used:[8] this is a parallel corpus of bilingual legal-administrative texts and Italian laws and codes translated into German. Additionally, the authors developed a customized automatic terminology evaluation tool to assess the MT quality of South Tyrolean legal terminology. The findings reveal significant improvements in the overall translation quality following the domain adaptation, as assessed by the standard quality metrics BLEU and chrF3. However, they also point to the persistence of difficulties in accurately translating specific legal terms, which suggests the limitations of on-the-fly adaptation for domains with limited parallel data. The authors conclude that further research is required to refine MT systems and terminology evaluation tools, particularly for low-resource language pairs and specialized domains. Both *LEXB* and the automatic terminology evaluation tool are accessible to the scientific community.

Chapter 11, by Politova, Bonetskaya, Dolgov, Frolova, and Pyrkova, describes the particular difficulties of aligning lexical units between two typologically distant languages such as Russian and Chinese, for which no gold standard was previously

[8] https://www.sketchengine.eu/eur-lex-parallel-corpus/

available. In particular, the authors present a rigorous methodology for the creation of a gold standard dataset for the Russian-Chinese word alignment. They provide a detailed account of their alignment guidelines and rules, based on previous research and adapted for this specific language pair. These guidelines, supported by clear and illustrative examples, address a range of linguistic phenomena, including punctuation, pronouns, classifiers, Chinese particles, and speech figures. The evaluation section introduces a comprehensive testing methodology, where two different machine learning models are used to compare their performance. The models were trained on the *Russian-Chinese Parallel Corpus* (RuZhCorp),[9] and fine-tuned on a manually annotated gold dataset. The findings demonstrate that the best results were achieved with *LaBSE*, and that fine-tuning the models on a gold dataset improves the performance of the algorithms. In addition to its specific focus on Russian and Chinese, this work provides a valuable model for the development of alignment guidelines and gold datasets for other typologically distant language pairs. The practical implications of this research are significant, potentially improving machine translation systems and enhancing corpus-based linguistic studies.

Finally, Ortego Antón's chapter presents the methodology used in the development of *GEnerador de Fichas de EMbutidos* (GEFEM),[10] a corpus-based writing tool designed for Spanish professionals to facilitate the composing dried meat product cards in English. The author establishes a prototypical rhetorical structure for both English and Spanish, classifying each rhetorical move and step according to their occurrence frequencies, from compulsory to occasional. She also identifies the model lines for each rhetorical element and creates a bilingual terminological database on dried meats. This is achieved by identifying usage patterns and extracting linguistic data from a unidirectional Spanish-English parallel corpus and a comparable English-Spanish corpus of dried meat product cards, both of which were specifically compiled for this purpose. Thus, GEFEM was developed on the basis of these three corpus-based research elements. Its user-friendly interface guides technical writers through the writing process, using colour-coded buttons and offering terminological suggestions from a database. The terms in the database are categorized by semantic fields, such as ingredients and allergens, thereby ensuring consistent and appropriate usage of terminology. Then, users can preview and download the final product card in docx

[9] http://ruzhcorp.ruscorpora.ru/en/
[10] https://actres.unileon.es/demos/generadores/applications.html#generatorsSection

format. In sum, GEFEM is an illustrative case study of the transfer of knowledge from corpus-based linguistic research to the agri-food industry. The application of this knowledge can facilitate the expansion of companies into international markets and enhance the productivity of technical writers.

The principal strength of the volume is its comprehensive coverage of the theoretical and practical aspects of the use of corpora in cross-linguistic research and the inclusion of a diverse range of domains and languages. The editors have brought together a diverse range of studies that evidence the versatility and applicability of corpora. Each chapter is based on rigorous empirical research, providing valuable insights and practical solutions for real-world linguistic issues and applications. In addition, the volume stands out for its attention to under-researched languages and domains, thus addressing an important gap in corpus linguistics. While it has notable strengths and each chapter is valuable in its own right, it could benefit from a more cohesive thematic structure. In addition, some chapters address complex technical issues that may be challenging for readers lacking expertise in corpus linguistics or statistical methods. In conclusion, Izquierdo and Sanz-Villar have succeeded in creating a valuable resource that not only advances academic knowledge in the field of cross-linguistic research, but also offers practical solutions that can be applied in language teaching, translation, and professional communication. This book is a highly recommended reference for anyone interested in the latest developments in corpus-based cross-linguistic research.

Reviewed by
Isabel Pizarro-Sánchez
University of Valladolid
Plaza del Campus Universitario s/n
Departamento de Filología Inglesa
Facultad de Filosofía y Letras
47011 Valladolid
Spain
E-mail: isabel.pizarro@uva.es