JOHN BENJAMINS
PUBLISHING COMPANY
Klaprozenweg 75G · P.O.Box 36224 · 1020 ME AMSTERDAM · The Netherlands · Tel. +31-20-6304747 · Fax +31-20-6739773

John Benjamins Publishing Company is pleased to have the privilege of publishing in its journal ***Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics***, the article / review entitled:
Parallel Corpora Spanish (PaCorES): A collection of multifunctional parallel corpora

by: Irene DOVAL REIXA y María Teresa SÁNCHEZ NIETO

Parallel Corpora Spanish (PaCorES): A collection of multifunctional parallel corpora

Corpus paralelos del español (PaCorES: una colección de corpus paralelos multifuncionales

IRENE DOVAL REIXA, University of Santiago de Compostela (Spain)
MARÍA TERESA SÁNCHEZ NIETO, University of Valladolid (Spain)*

Abstract

This paper introduces PaCorES, a compilation of parallel bidirectional corpora with Spanish as the focal language. It is currently comprised of four corpora in German/Spanish, English/Spanish, French/Spanish, and Chinese/Spanish, which are accessible online without restrictions.

The aim of this paper is to provide adequate documentation for the PaCorES_project and its contents for the corpus research community, while underscoring its multifunctionality. First, gaps are identified in existing parallel bi- and multilingual

corpora featuring Spanish as one of their components. Then, the criteria underlying the design and architecture of the PaCorES corpora are outlined. The main section of the paper is devoted to the different steps in the compilation of the corpora, with special emphasis on the alignment process and its review. Finally, attention is directed towards the three search levels available for querying the corpora and the display features. This renders PaCorES corpora a valuable resource for a wide range of applications and user groups across diverse linguistic fields, as well as language teachers, practitioners, and language learners alike.
Key words: parallel corpora, bidirectional corpora, corpus multifunctionality, corpus applications, corpus compilation, corpus alignment

- NOTE: Both authors of this article have contributed equally to its creation, and the order of authorship follows only an alphabetical criterion.

Resumen

El objetivo de este trabajo es proporcionar a los investigadores en el ámbito de la lingüística de corpus documentación adecuada del Proyecto PaCorES y los recursos que comprende, haciendo especial énfasis en la multifuncionalidad de estos. En primer lugar, se identifican las lagunas existentes en el panorama de los corpus paralelos bilingües y multilingües que incluyen el español como una de sus lenguas. A continuación, se exponen los criterios que han guiado el diseño y la arquitectura de los corpus de PaCorES. La sección principal del trabajo se dedica a las fases de la compilación de los corpus, con especial atención al proceso de alineado y de su revisión. Finalmente, se subraya la existencia de tres niveles de búsqueda disponibles para interrogar el corpus y se explica las funcionalidades de la interfaz y de la presentación de resultados. Estas últimas características hacen de los corpus de PaCorES un recurso valioso para un amplio rango de aplicaciones y tipos de usuarios: investigadores de diferentes ámbitos de la lingüística, docentes de lenguas, usuarios profesionales de la lengua o aprendices.

Palabras clave: corpus paralelos, corpus bidireccionales, multifuncionalidad de los corpus, aplicaciones de los corpus, compilación de corpus, alineamiento de corpus

1 Introduction

The compilation of extensive parallel corpora has been the key factor in the spectacular development of statistical (and more recently neural) machine translation systems and in other natural language applications, such as speech recognition, speech-to-text and text-to speech technologies, automatic document summarizing, and many others. On the other hand, parallel corpora have been also used for investigating linguistic and communicative phenomena, most notably languages in contrast and translation. Furthermore, parallel corpora can find uses in applied areas such as lexicography, language and translation teaching and learning. Of all the scenarios, the latter (i.e., translation and language learning and teaching with parallel corpora) relies the most on the availability and usability of parallel resources.

PaCorES is a collection of bilingual corpora with Spanish as the central language that can be easily and freely queried online. The roots of the PaCorES project date back to 2015 with the creation of the German<>Spanish Parallel Corpus (PaGeS)[1] (Doval et al., 2019, Doval & Jiménez Juliá, 2019, Doval & Sánchez Nieto, 2022).

At the time of writing, the English<>Spanish Parallel Corpus (PaEnS) (Author, year) and the French<>Spanish Parallel Corpus (PaFreS) have already been launched and can be accessed online. Moreover, the Chinese<>Spanish Parallel Corpus ( ) is currently being compiled and will be accessible soon. All these corpora draw on the same architecture and philosophy as the initial PaGeS corpus.

Considering that the creation of parallel corpora is a demanding task that consumes significant resources, our aim from the beginning was to design the PaCorES corpora to be multifunctional, capable of serving a wide range of applications and user groups, such as language learners and teachers, translators, translator trainers and educators, translator trainees and students, language, communication and translation scholars, and

[1]For the corpora's URLs see footnote 8.

lexicographers. It goes without saying that these corpora can be used as a reliable source of texts for training machine translation systems.

The purpose of this paper is to provide adequate documentation for the PaCorES project and its contents for the corpus research community. In section 2, PaCorES is mapped onto the landscape of those parallel bi- and multilingual corpora that are in one way or another comparable to its own. In Section 3, the criteria leading to the design and the make-up of the PaCorES corpora are explained. The tasks conducted before alignment are detailed in Section 4, while Section 5 delves into the alignment process. Section 6 introduces the features of the corpus search engine underlying all PaCorES corpora, offering insights into the corpus query system and results display. Finally, Section 7 wraps up the main conclusions of the paper.

2 Related work

In this section, we will provide a brief overview of a selection of parallel corpora that are comparable to the PaCorEs collection with the aim of discovering whether these resources are multifunctional as defined for the PaCorES corpora in Section 1.

The criteria for the inclusion of the resources in this overview are (i) being freely accessible and (ii) having Spanish as one of the included languages. As for corpus accessibility, a distinction can be made between corpora that can be accessed and queried online through a web-based corpus interface on the one hand, and corpora which are available for download online for further processing and elaboration on the other. As for the presence of the Spanish language in the selected corpora, attention must be paid to (a) whether the Spanish texts included are the result of either direct or indirect translation processes (i.e., as target texts translated from another language and subsequently parallelized to existing translations of the same text into further language/s); and to (b) whether the status of Spanish as either the source or target language is known.

Four selected parallel resources that meet the aforementioned criteria are briefly presented below.

The growing collection of bilingual and multilingual parallel corpora included in OPUS[2] (Tiedemann, 2012) offers not only the biggest collection of online searchable parallel resources, but also tools for uploading corpora to the collection and pretrained machine translation models. A number of the resources in OPUS are PoS-tagged (Tiedemann 2012, pp. 2216) and can be searched online with the help of multilingual concordance tools based on the query language used in the Corpus Workbench corpus manager (Tiedemann, 2012). External users can contribute their own parallel resources to the corpora collection (Aulamo and Tiedemann, 2019, p.1). At the time of writing, 81 resources can be searched and downloaded for the German/Spanish language pair, 122 for English/Spanish, 93 for French/Spanish and 19 for Chinese/Spanish.

Multilingwis[3] is one of the most successful attempts to make a parallel resource available through an online search interface. The system allows the user to query the parallel corpus of the European Parliament "for spotting translations of content words and multi-word units" (Multilingwis, 2023). The corpus contains 220 million words and word alignments for about 40 million words. The supported languages are English, Spanish, German, French, and Italian.

InterCorp,[4] a part of the Czech National Corpus, is a multilingual parallel corpus whose 8th release includes about 1.4 billion words in 38 languages and 174 million words in Czech (Rosen, 2016, p. 22). It is comprised of a core corpus of mainly literary texts and collections of other texts, which are obtained from other multilingual, freely available resources. The texts are automatically aligned at the sentence level via Czech, which is the pivot language. The texts are also lemmatized and partly tagged (Rosen, 2016, p. 23) and can be freely queried on the internet.

Linguee[5] is an online multilingual dictionary combined with an engine that looks up words and word groups in billions of translated texts and returns aligned bilingual

---

[2]See all the OPUS resources at https://opus.nlpl.eu/ and the OPUS search interface at https://opus.nlpl.eu/bin/opuscqp.pl.26.

[3] https://pub.cl.uzh.ch/projects/sparcling/multilingwis/ and https://pub.cl.uzh.ch/projects/sparcling/multilingwis2.demo/.

[4] https://kontext.korpus.cz.

[5] https://www.linguee.de/.

examples. Equivalences can currently be generated in more than 20 languages (Hein, 2022, p. 325), including, of course, English, German, Spanish, Chinese, and French. Among the multilingual pages crawled by Linguee's multilingual corpus compilation system are also sites that rely on unsupervised machine translation (Alonso, 2013, pp. 22-23).

In summary, while we can conclude that all four of the aforementioned parallel resources undoubtedly have intrinsic value, they also bear certain disadvantages for both research in contrastive linguistics and translation, and for foreign language teaching and learning and thus cannot be considered multifunctional as described in section 1. Most of these resources are either limited to specific textual varieties (mainly legal, administrative, or technical language), or neither the original language nor the translation process carried out can be precisely determined. Finally, in some cases, the search and display features are not adapted to different user groups.

Multilingwis and many of the OPUS parallel corpora feature to a large extent examples of specialized communication. Business, administrative, and legal language, as well as medical or technical language, show little lexical and morphosyntactic variation and tend towards formulaic language use. These domains are thus unsuitable for teaching/learning for non-specific purposes, as the level of difficulty of the structures therein is too high. At the same time, Linguee provides many parallel texts for a number of language pairs. However, along with the linguistic resources produced by international institutions, such as the European Commision, Linguee also makes other materials available which have not been produced or revised by a human translator and the source language is not known, something that holds true for a number of the parallel resources in OPUS as well. Texts scraped from unverified websites via automatic extraction and alignment are likely to contain errors and misalignments. The InterCorp corpora have been compiled with Czech as the pivot language, so that the English/Spanish, German/Spanish etc. alignments are indirect translations; these are very useful in certain teaching and research contexts, but of more limited use in others. In the case of OPUS, the search language (CQL) in the multilingual search interface[6] requires prior training, which may discourage non-expert users that try on-the-fly

[6] https://opus.nlpl.eu/bin/opuscqp.pl.

queries on their own for the first time. The same might apply as well to certain non-expert users approaching the KonText interface to query InterCorp.[7]

In summary, it can be said that the PaCorES corpora have come to fill a gap within the landscape of bilingual corpora of the Spanish-speaking world. As the following sections will show, the strong focus on the quality of the content, the considerable size of the corpora (section 3) and their usability (see section 6) makes them amenable to a broad scope of potential users.

3 Composition of the PaCorES corpora

In this section we will describe in detail the different types of data included in the PaCorEs corpora. The existing corpora to date are as follows, in order of creation:

- The German<>Spanish Parallel Corpus, PaGeS
- The English<>Spanish Parallel Corpus, PaEnS
- The French<>Spanish Parallel Corpus, PaFreS, and
- The Chinese<>Spanish Parallel Corpus, PaCheS[8]

Given that this is an ongoing project that began several years ago, the different bilingual corpora currently find themselves at various stages of completion. As a result, they differ not only in size but also temporarily in terms of the types of indexed texts. In the case of the Chinese<>Spanish and French<>Spanish parallel corpora, the core corpora (see 3.1) are not yet online since the number of texts that have undergone review is still too small. Nevertheless, the general guidelines for compiling the different corpora remain consistent across all four.

In PaCorEs there are two large groups of well-differentiated data: the first consists of the core corpus, and the second is made up of the so-called “supplements”. Both groups differ in their features, the processing procedures they undergo, and the extent to which reviews are conducted during the various processing steps.

---

[7] https://www.korpus.cz/kontext/query?corpname=syn2020.

[8]The URLs in the order listed: www.corpuspages.eu, www.corpuspaens.eu, www.corpuspafres.eu, www.corpuspaches.eu.

3.1 PaCorES core corpora

Taking into consideration the purpose of the corpora and the different applications for which they are intended, we have drawn up a set of guidelines for the data collection for the core corpus as explained in sections 1 and 2. These include the following features:

- both the original data and the translations must have undergone verifiable quality controls
- there must be an unambiguous identification of the original and the translated language
- the translations must have been carried out directly -without a pivot language- from the original by human translators
- the texts should be as close as possible to the current standard variety of the language, and belong to a general, non-specialized register.

In addition to these criteria, it should be noted that compiling a parallel corpus, in contrast to a monolingual one, presents significant constraints in terms of data availability as the vast majority of texts are not translated (Zanettin, 2012, p. 41).

3.1.1 Corpus Design and Research Potential

This section will detail the design of the corpus and consider its potential as a research tool.

In order for the texts to meet the criteria outlined above, we have chosen to limit our scope to texts from reputable publishers, therefore ensuring that both the originals and the translations have undergone rigorous quality controls (Doval, 2018).

Each core corpus consists of two sets of texts:

- original texts in the foreign language (FL)[9] and their translations into Spanish
- original texts in Spanish and their translations into the foreign language.

Both sets are similar in size and the types of texts they contain. As a result, all the core corpora are bidirectional a nd balanced in terms of translation direction.

---

[9] With the term 'foreign language' (FL) we refer to the four languages included so far in the PacorES collection: German, English, French, and Chinese.

Regarding the latter, it should be noted that due to the regular incorporation of new works, slight discrepancies may occasionally occur.

This bidirectional and balanced design of the core corpora enables a broader range of analyses and applications compared to unidirectional corpora. (cf. Stieg Johansson, 2004, p. 61).

This design facilitates the comparison of original texts and translations across languages (Figure 1) and allows for bilingual analyses from both the FL to Spanish and Spanish to FL. This approach is essential for studying how patterns vary across languages. Bidirectionality is particularly crucial because, in many cases, translation equivalents are not biunivocal, i.e., the translation of item A (source text) into B (translation) does not necessarily imply that B (source) is translated back into A (translation). A bidirectional corpus allows for the measurement of the percentage of mutual correspondence (MC) between two specific items, ranging from 0% for no correspondence to 100% for complete correspondence. (Bengt Altenberg, 1999, p. 254).

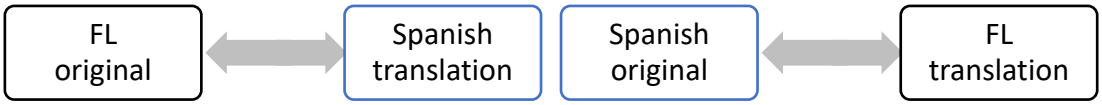

Figure 1: PaCorES as parallel corpora: original texts and translations FL and Spanish

Moreover, bidirectional corpora also enable comparisons between original and translated texts within each language (Figure 2) and between translated texts across languages (Figure 3). Hence, the corpora can be used for research into specific translation phenomena such as translationese or translation universals, i.e., recurring linguistic patterns or features observed in translated texts across different languages (Mona Baker 1993, pp. 742-746).

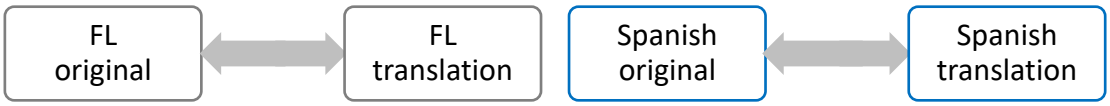

Figure 2: PaCorES as parallel corpora: original texts and translations in each language

Figure 3: PaCorES as parallel corpora: translations across languages

Finally, the bidirectional configuration of the corpora offers the additional benefit of providing a comparable corpus (Figure 4). This is achieved by comparing original texts in different languages, provided they share similar properties, such as genre, text type or domain, as is the case with the PaCorES corpora.

Figure 4: PaCorES as comparable corpora: original texts across languages

3.1.2 Composition of the core corpora

The core corpora are composed predominantly of fiction spanning a number of genres, with a smaller percentage dedicated to non-fiction works (essays, self-help, journalistic articles, etc.).

Regarding the geographical origin of the works, we have aimed for diversity by incorporating literature from British, American, Irish, and Australian writers for English, and from Austrian, Swiss, and German authors for German. In the case of Spanish, we've included works from both European and Latin-American writers. Our primary objective is to compile a contemporary text corpus, with the majority of the material stemming from recent decades, placing particular emphasis on works from the 21st century.

As previously mentioned, the different corpora are currently at varying stages of completion. Presently, the French and Chinese core corpora have not yet reached our minimum size requirement (5 million words) to be put online.

The core corpora currently online consist of 198 works for PaGeS (108 German originals and 90 Spanish originals) and 148 works (78 original English texts and 70 original Spanish texts) for PaEnS, along with their translations. These works can be accessed on the respective websites where they are listed by author.[10] Table 1 displays the size of the core German and English corpora as of November 2023.

[10] The full list of authors and works can be found here: https://rb.gy/n8d83 (PaGeS) and https://rb.gy/e0z7g. (PaEnS).

Table 1: Composition of the PaGeS (v 2.1) and PaEnS (v 2.0) core corpora (November 2023)

Language	Tokens	Bisegments	Language	Tokens	Bisegments
German Original	10.313.588	662.307	English Original	10.747.688	631.927
Spanish Translation	10.829.682		Spanish Translation	11.276.785	
Spanish Original	8.687.999	484.590	Spanish Original	7.267.961	395.791
German Translation	8.787.456		English Translation	7.518.400	
Total PaGeS	**38.618.725**	**484.590**	**Total PaEnS**	**36.810.834**	**395.791**

Due to copyright restrictions on both the originals and the translations, and because we have been unable to obtain explicit permission from all copyright holders, the works are not included in their entirety, but only in fragments.

Not only are the texts included in the core corpora of a high quality, but they also provide a higher level of lexical diversity. Lexical diversity refers to "the ratio between the number of unique word forms (types) and the number of running words (tokens) in the corpus" (Zanettin, 2012, pp. 14-15). The higher the ratio, the greater the lexical diversity of the text. In the case of PaCorES, which consists of corpora of very different lengths, to calculate the lexical diversity we have used the mean segmental type-token ratio (MSTTR), which consists of the average TTR for each non-overlapping segment of equal size (for every 1000 tokens).[11] When compared to the supplements (see Table 5), we can observe that the texts of the core corpora present the highest ratios, with the only exception being PaGeS for Spanish, which is surpassed by fictional works as well, albeit translated.

Table 2: Lexical diversity (MSTTR) of the core corpora of PaGeS and PaEnS

PaGeS		PaEnS	
German	0,572	English	0,522
Spanish	0,538	Spanish	0,537

[11] https://wordcruncher.com/pdf/Phrase%20Compare%20TTR%20Stat%20formulas.pdf.

3.2 PaCorES supplements

The second distinct group of texts in PaCorES, whose inclusion was not initially planned but came about for various reasons, is known as the supplements. For specific users and applications, such as studies in Natural Language Processing, Machine Translation, and lexicography, among others, these data have proven to be a valuable resource, despite not meeting all the aforementioned criteria. Additionally, the expansion of the core corpora progresses slowly due to the highly time-consuming manual processing involved. As a result, we offer a separate group of data materials with an adequate level of quality and textual variety, in many cases already aligned, even though we may not always be able to verify the original language, or the alignment may not have undergone manual review. These supplements serve as an additional option for users to consider during their consultation.

In these texts, the original language is not always specified; when such information is available, it has been indexed along with the other metadata. Only the alignment of the Ted-Talks has been manually reviewed as described in Section 5.2. For all other supplemental texts, no manual alignment check has been conducted. In order to minimize misalignments as much as possible, highly unequal bisegments (see Section 5.2) have been excluded as these are more prone to misalignments. Moreover, segments over 350 characters regardless of language were filtered out to avoid overly lengthy segments (see Section 5.1).

The main collections that currently comprise the supplements are as follows:

1. Europarl v7: This subcorpus compiles the proceedings ('Verbatim Reports of Proceedings') of the European Parliament from 1996 to 2011.[12] It is comprised of verbatim transcripts of each speaker's utterances during the plenary meetings of the European Parliament, along with translations into languages of other member states.

[12] The European Parliament ceased translating the proceedings into all EU languages in the second half of 2011. The Verbatim reports of plenary sessions are still available on the European Parliament's website but speeches are only published in the languages in which they were delivered.

The subcorpus was extracted by Philipp Koehn (2005), with the release in 2012 being the most recent version available.[13]

2. Ted-Talks:[14] This subcorpus collects transcriptions (mostly original English) and their translations into Spanish, German, French, and Chinese from talks spanning 2006 to 2020.
3. Global-Voices:[15] This subcorpus consists of texts written by an international, multilingual, primarily volunteer community of writers, translators, academics, and human rights activists. A group of Lingua volunteers make the stories available in dozens of languages.
4. Part of the United Nations Parallel Corpus v1.0[16] is included in the Chinese PaCheS corpus. It comprises official records and other parliamentary documents of the United Nations. The current version of the corpus contains content that was produced and manually translated between 1990 and 2014, including sentence-level alignments (Ziemski, Junczys-Dowmunt & Pouliquen, 2016).
5. A collection of fictional works: In these subcorpora—only available in PaGeS and PaFreS— both texts are translations from a third original language. In the case of PaFreS, five original works in French from the 19th century have been added.[17]
6. For the Chinese corpus there are two additional resources: Wikimatrix, a collection of automatically extracted parallel sentences from the content of Wikipedia articles,

---

[13] http://www.statmt.org/europarl. In PaCorES the cleaned and corrected version CoStEP Corpus (Graën / Batinic / Volk 2014) is used as well as its metadata <http://pub.cl.uzh.ch/purl/costep/>.

[14] https://www.ted.com/talks. We use the version provided by the Web Inventory of Transcribed and Translated Talks https://wit3.fbk.eu/ (Cettolo / Girardi / Federico 2012). Thanks are due to Mr. Mauro Cettolo, who kindly made the talks from 2018 to 2021 available to the corpus.

[15] https://globalvoices.org/.

[16] https://shorturl.at/eyDS9

[17] For the lists of works see https://www.corpuspages.eu/corpus/search/listofworks?lang=en&format=&search_context=0 (German) and https://www.corpuspafres.eu/corpuspafres/about/about?lang=en&format= (French)

and Paracrawl[18], created automatically by crawling multilingual websites that are then aligned at the sentence level.

The following tables show the number of tokens and bisegments in the different collections that make up the supplemental corpora.

Table 3: Composition of PaGeS (DE-ES) and PaEnS (EN-ES) in supplements

Bisegments	Tokens	Lang	Collection	Lang	Tokens	Bisegments
1.586.374	39.726.336	DE	Europarl	EN	39.481.818	1.536.548
	43.662.223	ES		ES	41.476.923	
310.968	5.599.587	DE	TED-Talks	EN	8.240.550	431.095
	5.805.812	ES		ES	7.860.866	
152.077	2.463.109	DE	Fiction. Works	EN	14.020.955	667.838
	2.479.765	ES	Global Voices	ES	15.039.034	
152077	**99736832**		**TOTAL**		**126120146**	**667838**

Table 4: Composition of PaFreS (FR-ES) and PaCheS (ZH-ES) in supplements

Bisegments	Tokens	Lang.	Collection	Lang.	Characters/ Tokens*	Bisegments
1.944.439	59.651.196	FR	Europarl	ZH	24.031.462	1.219.488
	53.583.854	ES	UNO	ES	9.103.395	
50.285	1.179.414	FR	Global Voices	ZH	15.235.788	498.145
	1.097.488	ES	Paracrawl	ES	7.534.822	
57.733	1.399.200	FR	Fiction. Works	ZH	12.546.473	360.968
	1.282.476	ES	Wikimatrix	ES	6.028.866	
254.222	5.197.553	FR	TED-Talks	ZH	3.695.097	98.713
	4.686.514	ES		ES	1.740.554	
254222	**128077695**		**TOTAL**		**79916457**	**98713**

*For Chinese, counting is done in characters, since one character corresponds more closely to one token: ratio ranges from 2 to 3 between Spanish tokens and Chinese characters.

Table 5 shows the lexical diversity (see section 3.1) of the various collections and corpora. The highest values of each corpus are in bold. As a general trend it can be observed that the fictional works present a higher lexical diversity than the other types of text, although they do not reach the values of the core corpora. (see above).

Table 5: Lexical diversity (MSTTR) by corpora and collections in the supplements

Collection	PaFreS		PaGeS		PaEnS	
Europarl	French	0,496	German	0,542	English	0,485
	Spanish	0,482	Spanish	0,481	Spanish	0,465
TED	French	0,531	German	0,543	English	0,476
	Spanish	0,534	Spanish	0,506	Spanish	0,506

[18] For more detailed information, see https://arxiv.org/pdf/1907.05791.pdf (Wikimatrix) and https://aclanthology.org/2020.acl-main.417.pdf (Paracrawl).

Global Voices	French	0,497			English	**0,518**
	Spanish	0,489			Spanish	**0,528**
Fictional Works	French	**0,545**	German	0,545		
	Spanish	**0,537**	Spanish	**0,569**		

4 Text preprocessing, textual mark-up, and metadata

This section deals with those steps that the texts for the different PaCorEs core corpora (see 3.1.) undergo before they are segmented and aligned (Section 5). These relate to the areas of text normalization, annotation of textual divisions, and annotation of metadata.

Prior to text normalization, both versions of the work are assessed to see if they consistently match each other or if there are differences between them (i.e., chapters or parts of chapters missing). If these mismatches are significant, the work is discarded.

Text normalization includes the following operations:

- encoding the plain text files in the UTF-8 standard, the format recommended by Tony McEnery and Richard Xiao (2005) for corpus construction (Zanettin, 2012, p. 74)
- removing any element that isn't part of the work itself. That includes (i) paratextual elements that are not parallel, i.e., those elements that appear in only one of the language versions of the works, such as informative paratexts (editorial information), title pages, authors' acknowledgments and dedications, translators' forewords, translators' notes and their calling signs, translators' glossaries, translators' acknowledgments and dedications; (ii) epitextual elements such as table contents, promotional blurbs, promotional references to other works by the same author or by the same publishing house, biographical notes, synopses, and bibliographical references
- standardizing txt files by cleaning them of (i) any extralinguistic elements that might remain after pdf to txt conversion, such as page numbers or end-of-line hyphenation (both seldom needed), line breaks inherited from the pdf conversion; (ii) ornamental text division signs such as clusters of three stars (* * *); tab marks (either individual or groupings); several paragraph marks together, which are replaced by only one paragraph mark, and standardizing the ellipsis mark (three individual dots are replaced with the standardized character «…».

All these normalization tasks are executed with the help of regular expressions. The pieces of software used for this purpose are either the free software Notepad++ or the licensed software EmEditor.[19]

As for the structural annotation, parts and chapters of a work are signaled with specific tags. The tags for the parts of a work are placed between curly brackets ({}) and tags for chapters are placed between angular brackets (<>); arabic numbers are used. These tags are useful at a later stage of the processing: during the alignment process, the aligning algorithm will use them as anchors for recognizing the beginning and the end of parallel text sections.

The extralinguistic annotation or text metadata are collected in a structured manner via a spreadsheet that allows for a conversion of the relevant information into xml format. The collected metadata can be divided into two different groups: text-level metadata, and project-level metadata.

Text-level metadata include, firstly, the nature of the text (original or translation), the status of its Spanish version (either original or translation), and language of the original version. These metadata result in a five-number code identifying (i) the corpus to which the text belongs (PaGeS, PaEnS, PaChEs, PaFreS), (ii) the translation direction, and (iii) the text itself. Further text-level metadata are the text's dialectal variant (resulting in a four-letter code, e.g., *eses* for European Spanish, *esch* for Chilean Spanish, etc.), [20] its genre/subgenre, and reviewer (team member[s] in charge of the preprocessing tasks for that text and the manual revision of the alignment).

Project-level metadata include information about the text material that results from the processing tasks: character, token, and word number for the foreign language and for the Spanish text.

5 Segmentation and alignment

5.1 Segmentation of the data

[19] https://notepad-plus-plus.org/ and https://www.emeditor.com/.

[20] See the full list at https://www.corpuspages.eu/corpus/help/help?lang=en .

Segmentation involves breaking up a string of written text into smaller processing units. The type of units (document, paragraph, sentence, or word) determines the type of alignment that can be subsequently established. Proper segmentation is essential for the performance of the automatic alignment process since erroneous segmentation often leads to a number of misalignments, as Jörg Tiedemann (2011, p. 9) indicates.

As in most parallel corpora, in the PacorES project the basic unit of alignment is the sentence, which is described in detail below. The sentence segmentation task involves identifying sentence boundaries between different sentences (Palmer, 2000, p. 11). The criteria for a sentence boundary are a punctuation mark such as a period, exclamation point, question mark (.!?), or occasionally a semicolon, followed by whitespace. But this task poses a few challenges, primarily because these marks are not unambiguous. So, a period may denote in Spanish, like in many other European languages, in addition to the end of a sentence, an abbreviation, an ellipsis, or a decimal point, among others. To prevent the incorrect assignment of sentence boundaries, the final punctuation rules accommodate those exceptions.

Another major issue in the segmentation process is that in literary texts, certain authors or works, due to their specific literary style,[21] may have long stretches of text without any final punctuation marks. In this case, automatic segmentation leads to very long segments that would be unsuitable for the purposes of the project. The shorter the bisegments the search engine returns, the easier it is for the user to spot equivalences for a given phenomenon. For this reason, both segments are manually subdivided after automatic alignment. The maximum number of characters allowed for a segment in either language has been set to 300. Segments exceeding this length are split by inserting manual break marks in suitable places in both the L1 and the L2. The average segment length (ASL) of a work is calculated by dividing the total number of characters by the number of bisegments and multiplying by 1000.

$$ASL = \frac{ChN(L_1) + ChN(L_2)}{BisegN * 1000}$$

[21] For example, authors like Javier Marías or Leonardo Padura stand out for writing long pieces of text without final punctuation marks.

The average segment length (ASL) ranges from 0 to 1, with a lower score indicating that fewer segments will need to be manually cut. The ASL can vary significantly from work to work depending on the author's style.[22] The average ASL for all texts in the core corpus after subsegmentation is 0.161. In the overall manual review, this task is the most time-consuming.

5.2 Alignment and manual review

Sentence alignment is the task of identifying correspondences between sentences in each half of a bitext. The first proposals for automatic sentence alignment date back to the early 1990s and can be categorized into length-based and lexical matching approaches (Tiedemann, 2011, p. 38). To address the limitations of these methods, hybrid approaches such as Hunalign[23] (Varga et al., 2007) and Gargantua (Braune and Fraser, 2010) were later introduced, combining both length-based and lexical matching strategies.

In the PacorES project the hybrid open-source software LF-Aligner, that relies on the Hunalign algorithm, and occasionally Gargantua [24] are used because they have achieved the best results in several tests. [25] LF-Aligner produces not only one-to-one alignments, but many-to-one and one-to-many alignment pairings as well, which are needed to ensure that all the input sentences are actually aligned. (Xu et al., 2015, p. 5).

---

[22] To mention two extreme examples, Leonardo Padura's work *The Man Who Loved Dogs* (11048) has an ASL of 0.556, while Louise Penny Louise's *The Beautiful Mistery* (10062) has an ASL of 0.11.

[23] This tool is representative of the current state-of-the art in sentence alignment and has been widely used in multilingual projects, like Europarl (https://www.statmt.org/europarl/), Opus corpora (https://opus.nlpl.eu/tools.php), InterCorp (https://intercorp.korpus.cz/?lang=en), etc.

[24] LF-Aligner: https://sourceforge.net/projects/aligner/,
Gargantua: https://github.com/braunefe/Gargantua.

[25]The following aligners were tested, among others: ABBYY Aligner 2.0 (https://abbyy.store/aligner-corporate/), Youalign (https://youalign.com/) and WinAlign (https://www.trados.com/solutions/translation-alignment/), commercial tools for aligning parallel texts and creating Translation Memory databases; Champollion http://sourceforge.net/projects/champollion/, a lexicon-based sentence aligner; bitext https://sourceforge.net/projects/bitext2tmx/.

The output is a sequence of bilingual sentence pairs (bisegments) in tab delimited format.

The accuracy of alignment depends largely on the type of texts. It is well known that all systems performed poorly on literary texts compared to other types of texts such as administrative or technical (Xu et al., 2015, p. 1; Zanettin, 2012, p. 155). Moreover, within fictional texts, the degree of correspondence of the bitext varies depending on the author, the translator, the texts themselves, and the direction of the translation.[26]

After the text has been automatically aligned, each bisegment receives a unique ID, composed of the 5 numbers of the work (see above) and 7 sequential numbers, separated by a hyphen, which correspond to the bisegment itself. This ID is crucial for the manual review of the alignment, as it allows the segments to be sorted without losing the original order of the text and recalled throughout the different steps that will be described in the subsequent paragraphs of this subsection.

After subsegmentation (see Section 5.1), the mapping between the corresponding sentences is checked. The first step is to check for automatic correspondences of 1:0 or 0:1. These alignments may be due to omissions or additions in the target text (L2) with respect to the source text (L1). In both cases, the following is indicated: if a piece of text is omitted in L2, it is marked as [n_t_s] (non-translated segment); if it has been added in L1, it is marked as [a_s_t] (text added in the translation). This can be seen in Table 6. These marks, which might seem uninteresting to some users, play a crucial role in analyzing specific instances of translator intervention (Munday, 2007).

Table 6: Omission and addition in the translation (L2)

01076-0000299	Ponía enormes cantidades de grasa de oca en todo, al estilo de su Belgrado natal,	Sie benutzte riesige Mengen Gänseschmalz, wie sie es aus ihrer serbischen Heimat kannte,

[26] Particularly challenging is the alignment of Chinese literary texts. The above-mentioned aligners have here a much lower level of accuracy. Since it requires very laborious and time-consuming correction work, there are very few texts in the core corpus of PaCheS, so it is not offered online for the time being. To go deeper into this issue would go far beyond the scope of this paper. We simply mention that we are testing with non-monotonic software, specifically with Vecalign (Thompson /Koehn 2019) It employs sentence embeddings to estimate alignment and it has shown promising results in aligning sentences with challenging cases where the sentence structure differs significantly, as in the case of Chinese.

01076-0000300	y yo ya había engordado un par de kilos.	[n_t_s]
11041-0002935	—¿No la viste, anoche?	“Didn’t you see her last night?”
11041-0002936	[a_s_t]	Oscar asked.
11041-0002937	—¡Claro que la vi!	“Of course I saw her!”

In other cases, alignments of 1:0 and 0:1 are due to misalignments resulting from a sentence having been split into more than one sentence in the target text (correspondence 1:<1, Table 7) or, inversely, several sentences have been merged into one (correspondence <1:1, Table 8). Here the necessary corrections are made.

Table 7: Incorrect alignment 1:2

11048-0006152	—¿Te va a dar una rabieta? —la voz de Tom destilaba ironía—.	“Are you going to have a tantrum?”
11048-0006153	No te voy a repetir lo que ya sabes.	Tom’s tone was sarcastic.
11048-0006154		“I’m not going to repeat what you already know.

Table 8: Incorrect alignment 2:1

11048-0002714	—Mi amigo murió…	
11048-0002715	Y cuando yo muera, y cuando muera la otra única persona que, según sé, conoce casi todos los detalles, esa historia se perderá.	My friend died… and when I die, and when the only other person dies who, as far as I know, is familiar with all the details, that story will be lost.

Finally, those unbalanced bisegments are reviewed that are more likely to be misaligned in terms of the number of characters. Since we have found that the relevance of the imbalance depends on the number of characters, the less characters the bisegment has the less relevant the imbalance is. To measure the level of imbalance (IL) we apply the following ratio, where *x* represents the total number of characters in the L1 segment and *y* in the L2 segment. The formula uses logarithms to reduce the effect of the number of characters.

$$IL = 1\frac{log(x) - log\ (y)}{max\left(log(x), log(y)\right)}$$

The closer the values are to 1, the more balanced the two segments are and the more unlikely a misalignment is. In the manual review we focus on the segments with a value below <0.5.

Table 9: Unbalanced incorrect alignment

00081-0004868	Bajó las ventanillas y condujo así, suspirando confusa y salpicando el interior del coche con las gotas que llevaba adheridas por fuera.	Verstört seufzte sie auf und ließ die Fensterscheiben herunter.	0,36
00081-0004869	El sonido del teléfono, que reposaba en el asiento del copiloto, interrumpió un hilo de pensamientos oscuros.	Die Tropfen, die außen daran gehaftet hatten, spritzten ins Wageninnere. In diesem Augenblick klingelte das Handy, das auf dem Beifahrersitz lag.	0,42

Note that this ratio can give both false positives and false negatives. **Table** 10 presents examples of false positives, i.e., unbalanced segments that nevertheless are correctly aligned.

Table 10: False positive correct alignment

01076-0000764	Un escalofrío me atenazó la espalda.	Mich schauderte.	0,29
10043-0010055	Far from it.	Las cosas no eran así, ni mucho menos.	0,21

Occasionally sentences are reordered when translating, but Hunalign, like most sentence aligners, is monotonic, meaning that the relative order of sentences is preserved and it is unable to come up with crossing alignments, i.e., segments A and B in one language corresponding to segments B' A' in the other language. In these cases, a simple operation of merging the sentences into larger units must be performed to find a monotonic mapping for the corresponding parts, even at the cost of longer segments, as in the following example:

Table 11: Crossing alignment

11048-000283	*Demasiado tarde comprendió que había menospreciado la inteligencia* del ex seminarista georgiano, y no había sido capaz de valorar su genio para la intriga, su desvergüenza para mentir y armar componendas.	Incapable of appreciating the Georgian ex-seminarian's genius for intrigue, his shamelessness in lying and putting together shady deals, Lev Davidovich *understood too late that he had underestimated his intelligence*,

This method allows for more efficient manual revision of the automatic alignment, saving time compared to a comprehensive manual review of each bisegment. It strikes a balance between minimizing tedious review work and maintaining a high level of accuracy.

Figure 5 outlines the PaCorES workflow. Once the texts have been aligned, they are PoS-tagged in each individual language before being rejoined and indexed. It should be noted here that due to technical difficulties, only the untagged versions are currently indexed despite having undergone PoS-tagging. Therefore, we do not address PoS-tagging in this paper.

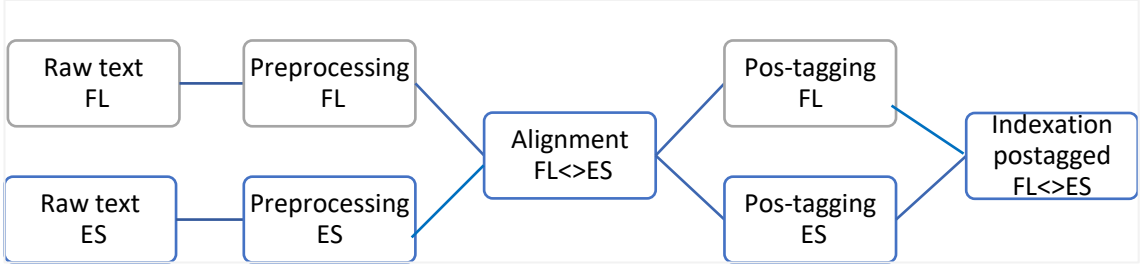

Figure 5: Workflow for the creation of the PaCorES core corpora

6 Search and display features

Once the texts are indexed, the resulting database can be queried with the help of the SolR search engine (Version 7.5.0).[27] In this section, attention is drawn to the ways the user can query the corpus and the kind of information included in the search results. In both aspects we have tried to meet the needs of a wide range of users, including experts and non-experts.

6.1 Search features

The PaCorES corpora can be queried at three different levels: the first level corresponds to the standard search interface, the second to the advanced search interface, and the third to an "expert level". Figure 6 illustrates how the user can alternate between the standard and the advanced search modes in the case of the PaEnS corpus.

[27] http://lucene.apache.org/solr/ and https://solr.apache.org/docs/7_5_0/.

Figure 6: Toggling between the standard and advanced search modes in the PaCorEs corpora (PaEnS)

At the first level (standard search interface), searches, by default, are performed:

- based on lemma (i.e., the character chains are interpreted as forms of a lemma and any segment containing forms of that lemma is added to the search results). If querying the form *back*, the user will be presented with both segments containing *back* as a preposition and segments containing back as a noun. To search for an exact word, the search term must be placed in quotation marks
- on the core corpus, although the user can choose to include the supplements in the query by ticking the respective boxes (see Figure 6). This allows the user to minimally select the kind of language he or she wants to search for (usually more general language in the core corpus vs. specialized language in the case of the Europarl, TED, or Global Voices supplements, which contain instances of administrative, popular science, and journalistic language, respectively)
- without differentiating between original and translated texts.

At the second level (advanced search interface), the user can choose between the following options (see Figure 7):

- querying specific works or groups of works by selecting the specific ID numbers, authors, publication years, genres, or varieties — US English, Mexican Spanish, Austrian German, etc.
- in the case of the core corpus, performing the queries on original works, translated works or both. In the case of the Europarl supplement, this can be done by selecting the corresponding language variety/varieties in the field “dialectal version”
- searching for a specific expression in one language which has (or does not have, see below) a specific equivalent in the other language, by placing each expression in the corresponding language field.

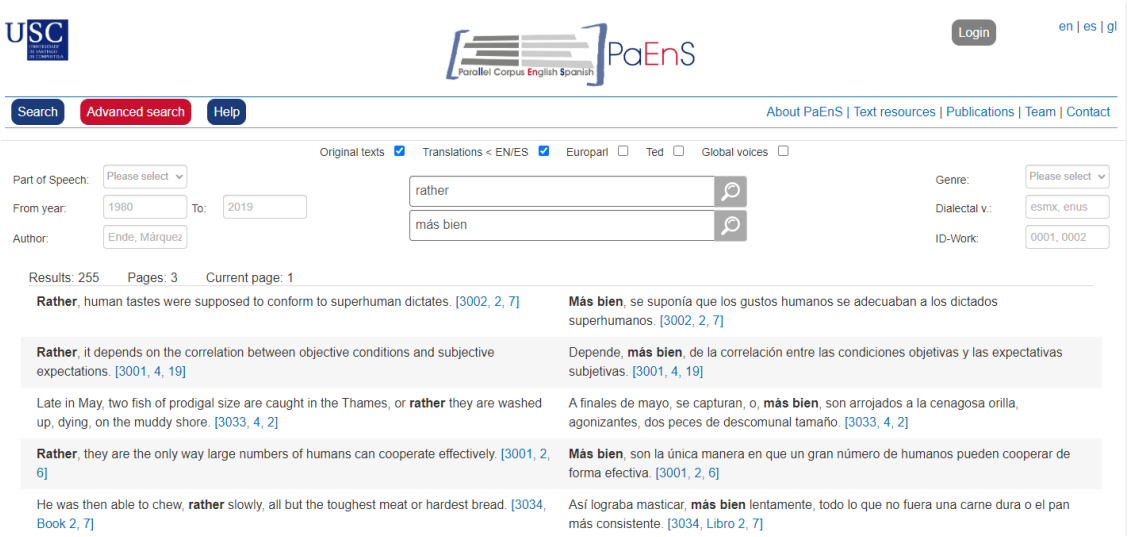

Figure 7: Advanced search interface in the PaCorES corpora (PaEnS)

The third or "expert" level does not have a specific search interface. Expert level queries can be performed both in the standard and in the advanced search interface. At this level, the user (typically a linguistics, modern language, or translation studies scholar) can take full advantage of the corpora by using the powerful query syntax of the underlying query tool SolR (Version 7.5.0), and if needed, in combination with regular expressions (Regex). Some clarifications are pertinent here:

– Expert searches can of course be combined with the possibilities of fine-tuned queries offered by the advanced search interface.
– The [SS] operator (meaning "SolR search") must precede any query chain at the expert level.
– The expert search operators are also used in browsers and bibliographic databases, so the expert user may already be acquainted with them. They include the wildcards * (for multiple characters) and ? (for any single character), the tilde ~ (either as the fuzzy search operator and as part of the distance operator ~1, ~2, etc.), as well as the Boolean operators NOT and OR (the Boolean operator AND underlies any search input consisting of two or more character-chains separated by a whitespace).
– On the Help page for each corpus a number of simple-word, multi-word and bilingual sample queries are displayed and explained in different tables for the user

to copy and paste into the search fields. He or she can then observe the results and build his or her own queries by adapting or enlarging the samples.[28]

6.2 Display features

Display features are also oriented to make interacting with the PaCorES corpora amenable to different user groups, as will be explained in this subsection.

On the one hand, the following display features are oriented towards the general user (an example of these features can be seen in Figure 7, above):

- Matches are shown in a side-by-side display that is easier to process visually than the KWIC display typical of monolingual corpora (Doval & Jiménez Juliá, 2019: 314-315).
- The left column corresponds to the original text and the right column corresponds to the translated text, be they English or Spanish (for example in the case of PaEnS);
- Matches appearing in the original texts are always shown first.
- The queried expression(s) appears in bold, regardless of whether it appears in the original texts, in the translated texts, or in both.
- Matches are presented in groups of 100; to avoid over-representation of a specific work/author in the query results, the matches are randomized and shuffled for each query. That means that querying the same expression twice or more will present the user with different results each time.

On the other hand, the following display features might be of interest for specific, more advanced user groups:

- Users involved in scholarly writing who include examples in their texts obtained from querying the PaCorES corpora will find that for each match the bibliographical reference of both the original and the translated part of the bisegments (author, original title, translated title, place of publication of both original and translation, and the part and/or chapter to which the match belongs) are available. This information can be accessed in a new window via a short link placed after the

[28] See, for example, the Help page at the PaGeS corpus: https://www.corpuspages.eu/corpus/help/help?lang=en&format=.

matched text (see Figure 8). Moreover, this window features a context selector widget, offering the user the ability to choose between three context widths.

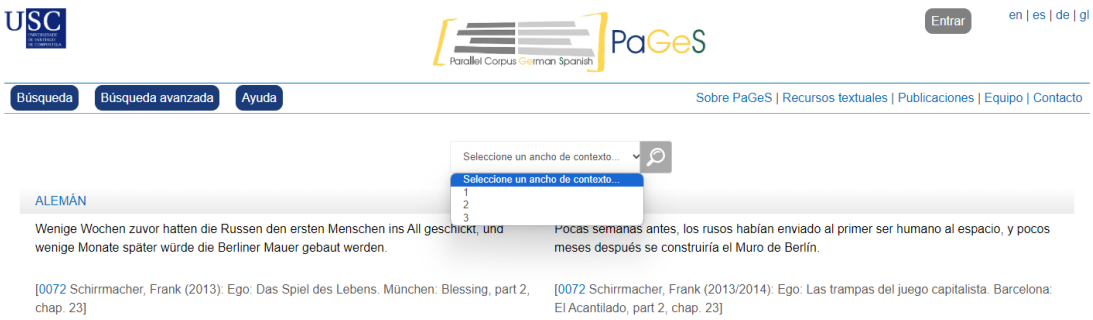

Figure 8: Bibliographic information and context width selector (PaGeS)

- In combination with the advanced search, the bibliographical information is of special interest for expert users performing dialectally and temporally focused searches.
- Registered users can download a CSV file with search results for further processing.[29]
- For those users interested in Translation Studies, signs of translator intervention have been preserved with the help of the labels denoting additions and omissions (see Section 5); similarly, the lack of immediate (preceding or following) context is signaled as well with the label n_i_t (non-included text).

7 Summary and outlook

In this article, we have provided an overview of the PaCorES collection (core corpora and supplements) offering a glimpse into its features and the possibilities it holds for various user groups. Now we will summarize the specificities that set it apart within the broader landscape of parallel corpora.

PaCorES is not a multilingual corpus but a collection of bilingual corpora with Spanish as the central language. To the best of our knowledge, it is the only project with

[29] Registration is free. Unregistered users can only view the first 10 results. For more information about the terms of use and copyright see https://www.corpuspages.eu/corpus/about/privacyterms?lang=en, https://www.corpuspaens.eu/corpuspaens/about/privacyterms?lang=null.

Spanish as its core language. Moreover, it offers a particular language pair, Chinese/Spanish, which is a very uncommon resource and holds great promise in terms of the potential number of users.

On the other hand, the PaCorES corpora are fully accessible and stable. Accessibility is due to the fact that the corpora can be freely queried online on a browser-independent basis. Stability is warranted as the PaCorES corpora are successively released in clearly identified versions, so that researchers can keep track of the source of their findings.

As for the central focus of the project, the core corpora, PaCorES includes a sizeable collection of contemporary prose texts (fiction and non-fiction), which are seldom found in the online parallel corpora currently available. Nor do available online parallel corpora usually include fictional informal spoken language, a kind of register that is particularly frequent in children's and young adult literature. Thus, these resources represent a highly valuable asset in our bilingual corpora. All texts in the core corpora offer proven quality and provide reliable human translations. Additionally, all text processing, including alignment, has been subjected to a comprehensive revision.

The core corpora are annotated with rich meta information, documenting not only the full source of the texts, but also other data such as translation direction, directness, and translator intervention.

PoS-tagged texts have not been implemented in the search yet, although we plan to do so in the near future. This will enable the user to search by morpho-syntactic categories. Additionally, we plan to offer several sorting options for the results display, specifically by the preceding and following context.

Finally, not only the PaCorES components but the project itself are designed with flexibility in mind, as potentially new language pairs can be added within the same corpus architecture and new texts can be added to the individual components.

All these features make the PaCorES corpora a truly multifunctional resource that caters to a diverse range of users. They meet the needs of linguistic specialists across various fields such as NLP, lexicography, contrastive linguistics, translation studies, or language and translation teaching. Moreover, the user-friendliness of its search and display features along with the speed of recall enable the PaCorES collection to be used as a teaching resource in the language and translation classroom. Here, intermediate to

advanced learners can discover a plethora of translation suggestions for a given item, directly presented through reliable examples of use.

Acknowledgement

We wish to acknowledge the financial support provided by the State Research Agency (AEI) of Spanish Ministry of Science, Innovation and Universities for the research project Spanish Parallel Corpora PaCorES (PID2021-125313OB-I00). Its funding significantly contributed to the successful completion of this research endeavor.

References

Alonso, E. (2013). 'Linguee' y las nuevas formas de traducir. *Skopos: Revista Internacional de Traducción e Interpretación, 2*, 5–28. https://dialnet.unirioja.es/servlet/extart?codigo=4567491 [retrieved: 09.05.2024].

Altenberg, B. (1999) Adverbial connectors in English and Swedish: semantic and lexical correspondences. In H. Hasselgård & S. Oksefjell (Eds.), *Out of corpora. Studies in honour of Stig Johansson* (pp. 249–268). Rodopi.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–250). John Benjamins.

Braune, F. & Fraser, A. (2010). Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora. In C.-R. Huang & D. Jurafsky (Eds.), *Coling 2010: Posters.* Coling 2010 Organizing Committee (pp. 81–89). https://aclanthology.org/C10-2010.pdf [retrieved: 09.05.2024].

Cettolo, M., Girardi, C. & Federico, M. (2012): WIT3: Web Inventory of Transcribed and Translated Talks. In M. Cettolo, M. Federico, L. Specia & A. Way (Eds.), *Proceedings of the 16th Annual Conference of the European Association for Machine Translation* (EAMT 2012) (pp. 261-268). Fondazoine Bruno Kessler. https://cris.fbk.eu/bitstream/11582/104409/1/WIT3-EAMT2012.pdf [retrieved: 09.05.2024].

Doval, I., Fernández Lanza, S., Jiménez Juliá, T. E., Liste Lamas, E. & Lübke, B. (2019): Corpus PaGeS: A Multifunctional Resource for Language Learning, Translation and Cross-Linguistic research. In I. Doval & M. T. Sánchez Nieto (Eds.), *Parallel Corpora for Contrastive and Translation Studies: New resources and applications* (pp. 103–121). John Benjamins.

Doval, I., & Jiménez Juliá, T. E. (2019). Multifuncionalidad de los corpus paralelos, ejemplificada con el corpus alemán/español PaGeS. In M. Blanco Domínguez, H. Olbertz & V. Vázquez Rozas (Eds.), *Corpus y construcciones: perspectivas hispánicas* (pp. 303–320). Universidade de Santiago de Compostela: Servizo de Publicacións e Intercambio Científico.

Doval, I., & Sánchez Nieto, M. T. (2019). Parallel Corpora in Focus: An Account of Current Achievements and Challenges. In I. Doval & M. T. Sanchez Nieto (Eds.), *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications* (pp. 1–15). John Benjamins. https://doi.org/10.1075/scl.90.01dov.

Doval, I., Sánchez Nieto, M. T. (2022). Das Deutsch-Spanische Parallelkorpus Pages: Aufbau und Nutzungsmöglichkeiten. In B. de la Fuente Marina & I. Holl (Eds.), *La traducción y sus meandros: diversas aproximaciones en el par de lenguas alemán-español* (pp. 319–341). Ediciones Universidad de Salamanca. https://doi.org/10.14201/0AQ0320319341 [retrieved: 09.05.2024].

Doval, I. (2023). The English-Spanish Parallel Corpus Paens. In I. C. Santos, M. Torrado Cespón, J. M. Díaz Lage, S. López Pérez (Eds.), *Current Trends on Digital Technologies and Gaming for Teaching and Linguistics* (pp. 145-164). Peter Lang.

Graën, J., Batinic, D., and Volk, M. (2014). Cleaning the Europarl Corpus for Linguistic Applications. In J. Ruppenhofer & G. Faaß (Eds.), *Proceedings of the 12th Edition of the KONVENS Conference* (pp. 222–227). Universitätsverlag Hildesheim. https://hildok.bsz-bw.de/files/265/p040.pdf [retrieved: 09.05.2024].

Hein, M. (2022). La competencia fraseológica y la enseñanza de fraseología en las carreras universitarias de traducción en la República Argentina. (Doctoral

dissertation, Universitat d'Alacant-Universidad de Alicante). https://rua.ua.es/dspace/handle/10045/123684 [retrieved: 09.05.2024].

Johansson, S. (2004). Multilingual corpora: models, methods, uses. *TradTerm 10*, 59–82. https://www.revistas.usp.br/tradterm/article/view/47044/50767 [retrieved: 09.05.2024].

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers* (pp. 79–86). Asia-Pacific Association for Machine Translation (AAMT). https://aclanthology.org/2005.mtsummit-papers.11.pdf [retrieved: 09.05.2024].

McEnery, T., & Xiao, R. (2005). Character Encoding in Corpus Construction. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 47-58) Oxbow Books. https://eprints.lancs.ac.uk/id/eprint/60/1/character_encoding.pdf [retrieved: 09.05.2024]

Moore, R. C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas: Technical Papers* (pp.135–144). https://link.springer.com/chapter/10.1007/3-540-45820-4_14 [retrieved: 09.05.2024].

Multilingwis (2023). Multilingwis: Finding Translation Variants in Multilingual Corpora. https://pub.cl.uzh.ch/projects/sparcling/multilingwis2.demo/ [retrieved: 09.05.2024].

Munday, J. (2007). *Translation as Intervention*. Continuum.

Palmer, David D. (2000). Tokenisation and Sentence. In R. Dale, H. Moisl & H. L. Somers (Eds.), *Handbook of Natural Language Processing* (pp. 11-33). Marcel Dekker.

Rosen, A. (2016). InterCorp – A Look Behind the Façade of a Parallel Corpus. In E. Gruszczyńska & A. Leńko-Szymańska (Eds.), *Polish-language Parallel Corpora* (pp. 21–40). Institute of Applied Linguistics WLS UW. https://depot.ceon.pl/bitstream/handle/123456789/13397/02_Rosen.pdf [retrieved: 09.05.2024].

Tiedemann, J. (2011). *Bitext Alignment. Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In N. Calzolari et al. (Eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC'2012) (pp. 2214–2218). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf [retrieved: 09.05.2024].

Tiedemann, J., Aulamo, M., Bakshandaeva, D., Boggia, M., Grönroos, S.-A., Nieminen, T., Raganato, A., Scherrer, Y., Vazquez, R., & Virpioja, S. (2022). Democratizing Neural Machine Translation with OPUS-MT. *Springer Nature* 2021, 1-45. http://arxiv.org/abs/2212.01936 [retrieved: 09.05.2024].

Varga, D. et al. (2007). Parallel corpora for medium density languages. In N. Nicolov et al. (Eds.), *Recent Advances in Natural Language Processing* IV: *Selected papers from RANLP 2005* (pp. 247-258). John Benjamins. https://hlt.bme.hu/media/pdf/ranlp05parallel.pdf [retrieved: 09.05.2024].

Xu, Y., Max A., Yvon, F. (2015). Sentence Alignment for Literary Texts. *Linguistic Issues in Language Technology 12*, 1-25. https://hal.science/hal-01634995/document [retrieved: 09.05.2024].

Zanettin, F. (2012). *Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies*. St. Jerome.

Ziemski, M., Junczys-Dowmunt, M. & Pouliquen, B. (2016). The United Nations Parallel Corpus. In Calzolari, N. et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 3530-3534). European Language Resources Association (ELRA). https://aclanthology.org/L16-1561.pdf [retrieved: 09.05.2024].

Primary literature

[01076] Gómez-Jurado, Juan (2014). *El paciente.* Planeta.

German: *Zerrissen.* DTV.

[11041] Cabrera Infante, Guillermo (2013). *Mapa dibujado por un espía*. Galaxia Gutemberg.

English: *Map drawn by a spy*. Archipielago Books.

[11048] Padura, Leonardo (2009/ 2013). *El hombre que amaba a los perros*. Tusquets.

English: *The man who loved Dogs*. Bitter Lemon Press.

Address for correspondence:

María Teresa SÁNCHEZ NIETO

Universidad de Valladolid

Facultad de Traducción e Interpretación

Campus Universitario Duques de Soria, s/n

E- 42003 SORIA (Spain)

mariateresa.sanchez.nieto@uva.es

OrcidID: 0000-0002-0378-3720

Co-author:

Irene DOVAL REIXA

Universidade de Santiago de Compostela

Facultad de Filología

Avda. de Castelao, s/n,

E-15782 SANTIAGO DE COMPOSTELA (Spain)

i.doval@usc.es

OrcidID: 0000-0002-9050-3183