Irene Doval (University Santiago, Spain)

Chapter VIII The English–Spanish parallel corpus PaEnS

I. Introduction: The PaCorES project

The *En*glish-Spanish *Pa*rallel Corpus PaEnS is part of a major ongoing project, *PaCorES*, acronym for Spanish Parallel Corpora, which aims to collect a series of bilingual parallel corpora with Spanish as the central language. Parallel corpora contain translations of texts from a source language into one or more other languages with the translated elements linked or aligned across languages in units consisting of words, phrases, or sentences (McEnery & Hardie, 2012, p. 20).

Parallel corpora have become an essential resource for a wide range of applications. Five major fields of applications of parallel corpora can be distinguished, each with its specific users: (a) basic research in contrastive linguistics and translatology, (b) lexicography, (c) translation, (d) teaching of foreign languages and of translation, and (e) natural language processing, most prominently in the field of machine translation (McEnery & Xiao, 2007, p. 2).

In general research in contrastive linguistics, parallel corpora provide an easily accessible empirical basis, producing patterns of correspondence, allowing for the analysis of similarities and differences between two linguistic systems and providing quantitative data (Johansson, 2007). In translation studies, they are very useful for discovering collocational and syntactic patterns between two languages (Bernardini, 2004, p. 28). Moreover, in the case of bidirectional corpora (in which a given language can function as both the source and target language), they can also be used to compare monolingually the original language and the translated language and to test Baker's hypothesis (1996, pp. 175 ff) on the so-called "translation universals," typical phenomena exclusive to translated texts.

Heid (2008, pp. 137 ff) provides an overview of the applications of parallel aligned corpora for bilingual lexicography, highlighting their usefulness at various levels and providing syntagmatic data on word usage in both languages. The usefulness of parallel corpora in language teaching has been frequently addressed in the specific literature (Doval, 2018, pp. 65 ff; Johansson, 2009, pp. 33 ff). On the one hand, they can serve as a basis for the elaboration of teaching materials and reference grammars. Moreover, bilingual concordance systems can be used in addition to bilingual dictionaries, or even instead of them, as they provide

valuable contextual information in a multitude of translated usage examples (Doval, 2019, p. 116). Finally, parallel corpora are a fundamental resource for a wide range of natural language processing tasks (Doval, 2022, p. 187), such as multilingual terminology and information extraction and especially machine translation. As Wetzel and Bond (2012, p. 28) point out, large, high-quality parallel corpora are an indispensable resource for training statistical and neural machine translation systems.

Each of these applications has distinct user groups with very different starting premises (from specialists in linguistics and language technology to English or Spanish learners) and with specific requirements related to the corpus's design, the type of texts, their degree of annotation and the metadata to be stored. Considering that the creation of parallel corpora is a vast undertaking in terms of both time and effort—more time consuming and technically complex than that of monolingual corpora—it is intended that the corpora of the PaCorES Project can be exploited for multiple purposes. Thus, it is our intention that, in addition to linguistic and translation research, the corpora are useful to lexicographers, translators and NLP researchers. Special effort has been made to make them a useful and user-friendly tool in language and translation classrooms, so that the corpora can be used by intermediate to advanced learners of the given foreign language or Spanish to obtain a large number of translation suggestions shown in examples of usage.

Since existing parallel corpora with Spanish as the central language have major drawbacks (see section II) in order to be fully exploited by all the abovementioned groups of users, the project PaCorES is driven by the aim of addressing these shortcomings. Within the project, high-quality, sentence-aligned parallel corpora are being compiled for different languages that are paired with Spanish. For the time being the German-Spanish corpus (www.corpuspages.eu) and the English-Spanish corpus (www.corpuspaens.eu) are already available online. It is expected that the Spanish-Chinese and Spanish-French corpora will be added to these in the near future.

The remainder of this paper is structured as follows: Section II briefly describes other similar resources and their drawbacks for many of the applications listed above. Section III is devoted to the description of the design and composition of the English-Spanish PaEnS corpus. The remaining sections then deal with text preprocessing (IV), segmentation and alignment (V). The web presentation of the corpus and the search possibilities are then described (VI) and, finally, future work is outlined and a brief recapitulation of the distinctive features of the corpus is made (VII).

II. Related work

By far, the most important parallel language resources come from different institutions of the European Union. Steinberger et al. (2014) give a comparative overview of the different multilingual resources available, which include, among others, the Europarl corpus (see below), the Digital Corpus of the European Parliament (DCEP),[14] or the JRC Acquis.[15] In addition, the European Union supports projects aimed at the exploitation and use of these resources, and in general at the enhancement of multilingualism by means of parallel corpora, like the MultiParaCrawl[16] corpus, which is made up of parallel corpora from web crawls collected in the ParaCrawl project and further processed to make it a multi-parallel corpus by pivoting via English.

Particular mention should be made of OPUS[17] a huge, constantly-growing collection of freely available multilingual text collections compiled by a crawler, which automatically searches the web for bilingual web pages, and also includes already aligned resources, such as some of those mentioned above (Tiedemann, 2012, pp. 2214 ff.). All preprocessing is done automatically. No manual corrections have been carried out. It also provides tools for data processing, as well as multiple search systems for corpus queries. This vast collection is maintained by Jörg Tiedemann and currently consists of more than 90 languages (cf. Tiedemann, 2016). Obviously, not all language pairs are equally represented. The English/Spanish language pair consists of 36 million aligned bisegments (Tiedemann, 2012, pp. 2214 ff.). The main fields covered by OPUS are legislative and administrative, as a large part of the texts come from the European Union or other international institutions. There are also, although to a lesser extent, other

14 The DCEP is a multilingual sentence-aligned parallel corpus in 23 official EU languages (253 language pairs) consisting of European Parliament texts produced between 2001 and 2012 and containing over 1.3 billion words. It excludes the documents already available in the Europarl corpus to avoid overlapping. https://wt-public.emm4u.eu/Resources/DCEP-2013/DCEP-Download-Page.html.

15 JRC-Acquis is a collection of legislative texts from the European Union and is currently comprised of selected texts written between the 1950s and the present day (https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis).

16 The last version ParaCrawl Corpus release v9 (March 2022) includes 705 parallel corpora and 41 languages. It contains for the Spanish-English language pair 269,394,967 aligned sentences. https://paracrawl.eu/.

17 http://opus.lingfil.uu.se/.

text types such as subtitles, journalistic texts and some other smaller collections from various Internet sources, such as subtitles and technical documentation

We will now turn to the shortcomings, given that these corpora have laid the groundwork for the creation of the PaCorES corpus collection and their extension to new language pairs. In the corpora created automatically from the web by means of a web crawler, it is impossible to identify the direction of the translation. Therefore, they lack the identification of the source and the translated language (Doval, 2017). Closely related to this point is the fact that it is not known whether the translations have been made directly between the two working languages, or rather indirectly as independent translations from a third language acting as the pivot language. In many of the European Union resources, source texts are translated into English in the first step, from which they are then translated into other languages. While these issues might not pose a major problem for NLP applications, they are nevertheless key drawbacks for translation and cross-linguistic research.

On the other hand, to make the corpora a suitable tool for use in the language and translation classroom, both source and target texts must be of high quality, and the translations must be unambiguously carried out by professional human translators. This aspect cannot be verified by corpora automatically produced by gathering web pages, where it is not known if the texts have undergone any quality control.

All in all, perhaps the main drawback of these corpora is, despite their huge size, the poor variety of their texts, as they are limited to very specific domains, mainly administrative, legal and commercial language (cf. Steinberg et al., 2014, p. 4) where the general language is not well represented. Therefore, these corpora are poorly suited for different user groups, and especially for use in translation and second-language learning.

Finally, it should be noted that users without specific training have difficulties to use these resources due to their lack of a web interface. This makes it difficult for the common occasional user to access and use the system. This is the case with the resources from the European Union, which, although they can be downloaded in XML format, do not offer a web interface to be consulted online. The project PaCorES aims at making the resources easily available to researchers and students in different fields, even if they lack programming skills.

III. Design and composition of PaEnS

The PaEnS corpus consists of two major well-differentiated parts: the core corpus and the supplements. The core corpus was entirely developed within the project

and will be the focus of this paper. The supplements are texts of different origins, which have been added subsequently and should be considered as mere complements to the core corpus.

3.1. The core corpus

Here we will describe the data of the PaEnS core corpus in terms of size, text types, language varieties, number and publication date of the texts.

A corpus is not just an arbitrary collection of electronic texts, but they must be "selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language" (Sinclair, 1996, p. 4) in order to ensure that the corpus is suitable for the intended purpose. In the case of translation corpora, however, an important limitation regarding the available material must be kept in mind. The vast majority of texts are not translated at all and, if translated, only a tiny proportion has passed any quality control. Hence, the basic criterion for gathering texts for a parallel corpus must be an opportunistic one, that is, it is necessary to resort to what is available (Doval, 2017). For this reason, the requirements placed on monolingual corpora in terms of representativeness and balance (Biber, 1993; Egbert et al., 2022) representing a relevant range of subjects and registers are hardly applicable here (Nádvorníková, 2017).

In addition to this major constraint, another one is the requirement of the quality of the data (originals and translations). Apart from the aforementioned institutional language resources, the only way to ensure text quality is to make use of written texts from reputable publishers, where both original texts and translations are subjected to strict quality control.

Therefore, the PaEnS core corpus contains original texts in English and Spanish and their published translations. So far it comprises a collection of 75 works[18] and their translations. The corpus is bidirectional and quite balanced regarding translation direction, where English originals are only slightly more prevalent (26 %) vs. Spanish Originals (23 %). Figure 1 displays the structure of the corpus. The arrows between the boxes indicate the types of studies this structure enables, which include, among others: contrastive and translation studies based on original texts and their translations (solid horizontal arrows) or based on parallel original texts (dotted vertical arrows) and general features of translated texts (dashed diagonal arrows).

18 The full list of authors and works can be found here: shorturl.at/hjoy2.

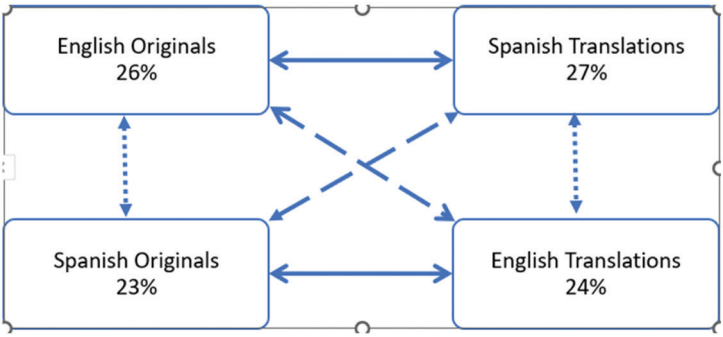

Figure 1. Structure of the PaEnS corpus.

Eighty percent of the works are fiction (fantasy, thriller, historical fiction, children's and young adult literature among others) and 20 % are non-fiction of various genres (essays, advice literature, biographical and popular science texts). This type of texts not only guarantees high quality but also a greater lexical diversity and, especially in children's and young adult literature, a fictional spoken informal language, which is a very important resource for a bilingual corpus, since it is a very scarce register in the translated language.

In regard to the geographical provenance of the works, a certain dialectal diversity has been pursued by including works from British, American, Irish, and Australian writers for English, and European and Latin-American writers for Spanish. As our aim is to build up a corpus of present-day texts, all of them were published after 1960, with special emphasis on twenty-first-century works, as shown in Figure 2.

The works were not included in their entirety but in excerpts in order to comply with the constraints for copyrighted publications. Moreover, consequently this leads to a greater variety of texts. These omissions in the text flow are indicated according to an ellipsis in square brackets […]. Table 1 gives an overview of the current composition of PaEnS, that contains nearly 17 million tokens and over a half million bisegments, that is, pairs of aligned text chunks (sentences or smaller segments).

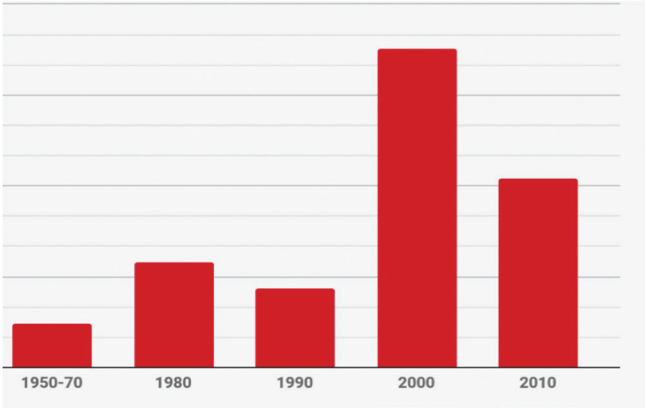

Figure 2. Texts by time period.

Table 1. Number of Works, Tokens and Bisegments by Language and Original Text in the PaEnS Core Corpus (July 2022)

Language	Works	Tokens	Bisegments
English Original	37	4.446.050	279.624
Spanish Translation		4.521.373	
Spanish Original	38	3.755.859	235.866
English Translation		3.948.366	
TOTAL	75 (x2)	16.671.648	515.490 (x2)

3.2. The supplements

As its name implies, this part of the corpus was designed only as a supplement to the core corpus in order to cover certain, such as administrative and legal language, which the corpus' core collection lacked. Unlike the core corpus, the texts had previously been automatically aligned at the sentence level without manual review. The texts have subsequently undergone different automatic control and cleaning procedures. So large segments above 350 characters (in Spanish or English, respectively) were excluded as well as unbalanced bisegments or any others considered to lack linguistic value. These omissions are not indicated. Currently,

this supplementary part includes three collections: Europarl, TED Talks and Global Voices. We will now briefly explain the key features of each of them.

Europarl,[19] created by Philipp Koehn, is a parallel corpus containing the minutes of the European Parliament (Verbatim reporting). It was obtained by automatically gathering the proceedings of the European Parliament from its website (Koehn, 2005). The proceedings contain the edited and revised transcripts of all speeches delivered in plenary debates by the members of the European Parliament, usually in their mother languages, as well as the translations of the transcripts into all other official languages of the European Union (Bernardini, Ferraresi & Miličević, 2016, pp. 68–69). However, a certain number of statements have no information about the original language. The latest version (release v7) contains the transcripts between 1996 and 2011 in 21 official EU languages. PaEnS uses the cleaned and structurally enriched version of the CoStEP[20] Corpus as well as its metadata. The English-Spanish portion comes to over 40 million words per language after the aforementioned cleaning procedures (s. Table 2).

The TED Talks are a collection of transcribed and translated talks published by the TED Conference website.[21] Since 2007, the TED Conference has been posting all video recordings of its talks together with subtitles in English and their translations into more than 80 languages (Cettolo et al., 2012, p. 261). TED Talks are mostly originally held in English and their videos are available through the TED website together with subtitles. The talks have been translated by volunteers into different languages. The version used in PaEnS is that provided by the Web Inventory of Transcribed and Translated Talks.[22] It is comprised of the Spanish translations and the original English transcripts of 4,043 TED Talks from 2006 to 2020 aligned at the sentence level. The automatic alignment was manually reviewed within the project.

19 http://www.statmt.org/europarl.

20 http://pub.cl.uzh.ch/purl/costep/ It was made freely available as the Corrected & Structured Europarl Corpus (CoStEP) in order to further enhance the usefulness of Europarl and to compensate for some of its drawbacks (Graen et al. 2014).

21 http://www.ted.com.

22 https://wit3.fbk.eu/>. For more information, see Cettolo et al. (2012). Special thanks are due to Mr. Cettolo, who kindly made the talks from 2018 onwards available to the PaEnS corpus.

The most recent collection added to the corpus is Global Voices,[23] a corpus of texts written by an international, multilingual, primarily volunteer community of writers, translators, academics, and human rights activists. A group of volunteers translates the stories into dozens of languages. The texts, taken from release v2018q4, are sentence aligned and include data up to December 2018.

The summary statistics for all three Supplements included so far in PaEnS are presented in the Table 2.

Table 2. Number of Tokens and Bisegments in the Supplements (July 2022)

Resource	Language	Tokens	Bisegments
Europarl	English	42.178.712	1.550.421
	Spanish	44.128.158	
TED Talks	English	8.676.842	430.667
	Spanish	8.338.726	
Global Voices	English	15.285.853	680.530
	Spanish	16.361.642	
TOTAL	English	66.141.407	2.661.618
	Spanish	68.828.526	

The following sections describe the steps that have been carried out for the creation of the PaEnS core corpus.

IV. Data preprocessing

This section describes the data preprocessing steps that prepare the incoming texts for further processing: alignment (section V) and linguistic annotation. The data preprocessing involves three tasks: text normalization, annotation of textual divisions and annotation of metadata.

Each work is assigned a unique ID, the texts are converted to txt-format and the characters are encoded to UTF-8. Then both versions are tested for

23 https://globalvoices.org/ The parallel corpus included here was compiled and provided by Casmacat (http://casmacat.eu/corpus/global-voices.html) and it was adjusted for OPUS (Tiedemann, 2012).

discrepancies and non-corresponding text fragments are removed. Essentially the aim is to achieve as much parallelism as possible between the source and target data to obtain the best alignment results. First, all text fragments that are not part of the body of the text are removed, such as bibliographic information, dedications, tables of contents, tables, diagrams, indexes, footnotes, headers and footers, and author's or translator's notes. Similarly, any appendix with no equivalent in the other version is deleted.

Second, both sets of data (original and translation) are checked for errors associated with the digitization process. Typical errors include the insertion of a space or a hyphen within a word, the deletion of spaces between words, or the occasional confusion of certain characters.

Most works include some sort of segmentation in terms of parts or chapters that are tagged as divisions to make it easier to refer back to the source. Further subdivisions such as page breaks in the source text, paragraphs or lines are not marked, given that most of the texts were primarily digital. Typographical highlighting (italics or bold) is not marked.

The metadata is used to capture relevant information about the source texts to retrieve it from the corpus. Each of the works included in the PaEnS corpus is provided with a metadata list containing, among other things: author and translator, title, year of publication and other bibliographic information, original language and version language, genre, manual reviewer, and information on the basic statistics (number of characters, tokens, and bisegments) of the documents. These additional metadata tags are attached to the individual text files and stored locally with each text document.

V. Segmentation and alignment

Corpus alignment is a central task in the construction and exploitation of a parallel corpus. The type of alignment is determined by the previous type of segmentation of the text into different segments, such as paragraphs, sentences or words.

Tiedemann (2011) defines alignment as "the process of linking corresponding parts with each other" (p. 123). Alignment is the task of making this correspondence explicit, by linking segments of the bitext (source and target text) that are equivalent and assigning the segments of the translation to corresponding segments of the original. These aligned segment pairs, one from each half of the bitext, are called bisegments (Tiedemann, 2011, p. 24).

Based on this segmentation a parallel corpus can be aligned at different levels. For the PaEnS corpus, it was decided to focus on sentence alignment, since this

level of alignment is the most established for parallel corpora (see among others Tiedemann, 2011; Volk et al., 2014).

In this alignment process, two tasks are combined: First, segmentation of the monolingual texts into sentences is performed, and then the English and Spanish sentence segments are linked to each other, as shown in Figure 3:

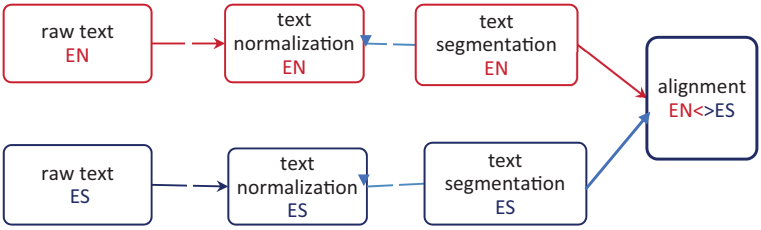

Figure 3. PaEnS Core Corpus: Workflow.

In the PaEnS corpus, the open-source program LF-Aligner[24] is used for sentence alignment because it has achieved the best results in several tests. It is based on Hunalign (Varga, 2012, pp. 92–119), a very common alignment software for multilingual corpora. Hunalign combines both a length-based and a lexical matching approach and is therefore a so-called hybrid algorithm (Tóth et al., 2008; Varga et al., 2005).

Alignment is performed in four steps:

(a) The texts are segmented based on punctuation. A sentence ends after a full-stop, question-mark or exclamation-mark (.?!)
(b) Then they are aligned using a modified version of Brown et al.'s (1991) sentence length-based model.
(c) The program creates an automatic dictionary based on this initial alignment, and
(d) finally, it refines the alignment in a second run using the automatic dictionary.

AQ2 Citation Brown et al. (1993) has been changed to Brown et al. (1991) as per the reference list. Please confirm.

The accuracy of alignment depends, first on the type of texts; specifically, fictional texts, as in the case of the PaEnS corpus, present greater problems than other types of texts, such as administrative or technical texts, as pointed out by Zanettin (2012, p. 155). Moreover, within fictional texts, the degree of correspondence

AQ3 Please note that the cross-reference Zanettin (2012) has not been provided in the reference list. Please provide the same.

24 http://sourceforge.net/projects/aligner/.

of the bitext varies depending on the author, the translator, the texts themselves and the direction of the translation.

Sentence alignment would be trivial if a sentence were translated into exactly one sentence, but, obviously, a one-to-one correspondence is not always possible since during the translation process, sentences in the target text can be omitted (Table 3) or inserted (Table 4), the translator can split a sentence (correspondence 1:2, Table 5), merge two sentences (correspondence 2:1, Table 6), or rearrange them to produce a natural translation in the target language (Varga, 2012, p. 94).

Table 3. Omission of a Segment in the Target Text (3101, chap. 1)

Junto a los dólares había dos pasaportes: el suyo y el del Güero.	**With the dollars were two valid US passports—hers and Guero's.**
Los dos tenían visas norteamericanas vigentes.	[n_t_s]
Contempló un momento la foto del Güero:	She studied his photo:

Table 4. Addition of a Segment in the Target Text (3007, part 3, chap. 7)

"The Ministry's not too happy."	**Al Ministerio no le ha hecho ninguna gracia.**
[a_s_t]	—El trébol es el símbolo de Irlanda.

Table 5. Misalignment due to 1:2 Correspondence (3132, part 5, chap. 9)

—He llamado a Héctor casi todos los días, *saben que estoy preocupado.*	**"I've been calling Héctor almost every day.**
	They know I'm worried."

Table 6. Misalignment due to 2:1 Correspondence (3012, chap. 4)

You decide that you want a better life, in a manner that will also make the life of your family better.	**Decides que quieres una vida mejor que al mismo tiempo suponga una mejora para la vida de tu familia,** *o la de tu familia y la de tus amigos, o la de todos ellos y también los desconocidos que los rodean.*
Or the life of your family, and your friends, and the strangers who surround them.	

In the project an effective procedure was developed to manually review and validate the results of the automatic alignment of the selected bitexts, and hence to improve its quality. This is done in three steps. First segments of more than 350 characters are split by inserting manually breaks at appropriate places in both segments. In a second step, the empty alignments are identified. As previously stated, they may be due to misalignments or to deletions or insertions in the translated text. If the segment is misaligned, the necessary corrections are made. If the segment has been omitted in the translation, the mark [n_t_s] (=non-translated segment) is inserted in the empty cell (Table 3). If the segment has been added in the translated text, the mark [a_s_t] (=text added in the translation) is inserted (Table 4). Finally, to reduce manual work, we find unbalanced bisegments in terms of the number of characters in English and Spanish. These unbalanced bisegments are more likely to be misaligned. We apply a ratio, comparing the number of characters in both, the English and the Spanish segment, order the bisegments by the value of this ratio and focus on the range 10:- 10 (Doval et al., 2019). This way manual checking of the automatic alignment is done more efficiently and takes less time. This procedure is a compromise to minimize the tedious work involved in review, while ensuring a high level of accuracy.

VI. Search and display

As Dörk and Knight (2015) assess, many existing corpora are "aimed mainly at people with expertise with linguistics" (p. 84). Thus, there is less reflection on the decisions involved in designing search and visualization of corpora and corpus analysis for non-expert users.

As mentioned at the beginning of this chapter, to make the corpus a real multipurpose tool, useful for very different user groups, from cross-linguistics and

translation researchers to lexicographers and NLP researchers to occasional or regular users, as well as English or Spanish learners, it is essential to provide an adequate interface for displaying and retrieving data, including corpus texts, metadata and linguistic annotation. To achieve this aim, the web interface of PaEnS presents relevant functionalities.

It has a fast and user-friendly search: the search engine allows queries to be carried out very quickly through large amounts of data. The basic query language is very simple and an advanced, more complex, query language is only displayed if required. Therefore, it was designed as a three-level search. The first level is the simple or standard search. In this case, to search for a concordance (all occurrences of a given word displayed in context) the user need only enter the search term (a word or a phrase) in English or Spanish in the search field. Figure 4 provides an example of the standard search menu and some of its features.

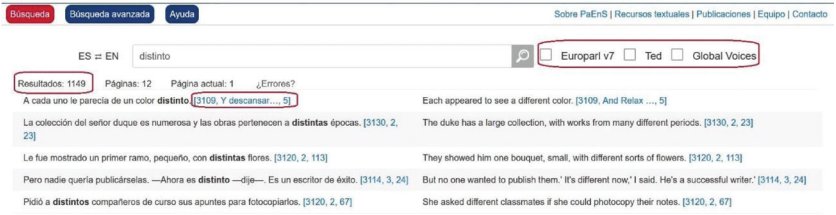

Figure 4. Standard search in the PaEnS corpus.

In these types of queries, lemmatization is applied by default. With multiword queries, all search words within a specific distance are found. The search habits of Internet users are exploited, emulating the google search language as the basic model. So, if the term or phrase is enclosed between quotation marks (“ “), the search only returns concordances that exactly match the entered word form or phrase (s. Figure 6). To expand the text search, wildcards (*?) can be used to search for characters that do not exactly match the search criteria.

The search results are displayed in an easy-to-read format. The matching segments are displayed side-by-side. We discarded the most common concordance format, KWIC (Key Word in Context)—the node word occupies a central position with all lines vertically aligned around the node—because this presentation of the results for bilingual corpora is not user-friendly (Doval et al., 2019), since it does not allow the original and the translation to be displayed side-by-side. For this reason, we decided on a two-column html table for the display of the query results, where the left column corresponds to concordances of the search term in the source texts

(English or Spanish). The corresponding segment with the translation is displayed in the same row in the right column. Occurrences of the search term in the translations are shown afterward in the right column. The search term or phrase is displayed with some context and highlighted in bold (Figure 4). With each query, information on the number of hits, the total number of pages, as well as the current page number is shown.

At the bottom of the table, a set of links are available to allow the user to navigate through the pages and to download (feature only available for registered user) the query results in two formats: ODS and CSV. In each cell, information concerning the text ID, the corresponding part or chapter are displayed in blue square brackets. By clicking on the work ID, the user can select a larger language context. Moreover, this screen shows detailed information concerning the bibliographic information, as shown in Figure 5.

Figure 5. Context with and bibliographic information in the PaEnS corpus.

The second level of search is the advanced one (Figure 6). At this level, the user can control and restrict the scope of the search by applying the following drop-down search filters: text ID, author, publication year, original or translated text, genre and dialectal variety. Moreover, here bilingual searches can be performed at the same time, returning concordances where a given word in English corresponds to a specific word in Spanish (Figure 7).

Figure 6. Advanced search in the PaEnS corpus.

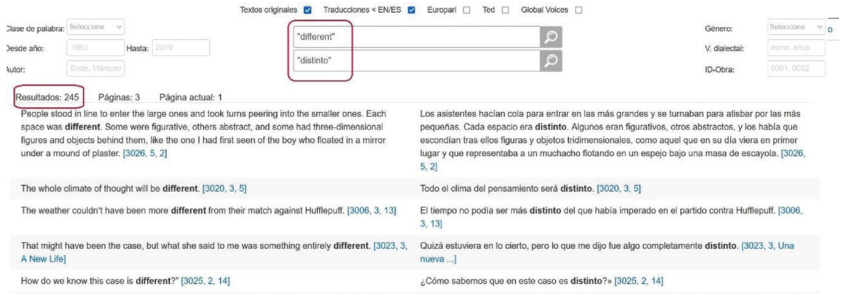

Figure 7. Advanced bilingual search in the PaEnS corpus.

The last search level is the most complex and it gives the user full command of the powerful query syntax used in the underlying query tool: Solr. (Version 7.5.0).[25] This supports searches using regular expressions (RegEx),[26] that expands considerably the power of the query language. The search term has to be preceded by [SS] (Solr Search). The search expression must be constructed word by word, and for each word, it will be possible to specify several parameters. Figure 8 shows a formal search. At this level the search system allows for complex queries across multiple layers of linguistic annotation, combining text string and PoS-tags searches.

VII. Concluding remarks and future work

The corpus PaEnS is part of a larger project, PaCorES, started at the University Santiago de Compostela in 2014, that intends to create a large collection of Spanish bilingual parallel corpora of high quality and considerable size. So far there are two corpora available online: the German-Spanish corpus, PaGeS, with nearly 40 million tokens, and the English-Spanish corpus, PaEnS, with more than 16 million tokens (core corpus), to which this paper is devoted. As has also been previously stated, the stated purpose of the project is to build qualitative multifunctional parallel corpora so that they become useful tools for a multiplicity of users.

This paper describes the different steps we have completed in the construction of PaEnS. Since this is an ongoing project, we want to continue and complete

25 http://lucene.apache.org/solr/.
26 https://www.regular-expressions.info/.

it in several dimensions. On the one hand, we plan to add new texts to the nuclear corpus to reach some 50 million words, 25 per language. We also intend to include new collections in the supplements to complement the domains of the existing ones. On the other hand, we are implementing two new features: PoS tagging and word alignment. The English texts have already been tagged with Treetagger and the Spanish texts with Freeling and both tagsets have been mapped to the universal part-of-speech tagset.[27] However, the tagged texts are not yet indexed, but they are expected to be available online soon. Regarding word alignment, several tools (Berkeley Aligner, Giza++, NaTools and Nile)[28] are being tested, but a final resolution has not yet been reached.

Despite the existence of other Spanish-English parallel corpora, PaEnS has a number of distinctive features, which make it stand out from similar resources.

The typology of texts included is very diverse: it ranges from fiction and non-fiction texts in the core corpus to the administrative and legal texts of the Europarl, journalistic texts (Global Voices) and oral texts covering a wide variety of topics (TED Talks). It includes domains like the fictional spoken language for which few resources exist.

It is bidirectional and well-balanced enabling not only comparison between the original language and the translation, but as well between originals and translations in the same language or between translations into two languages. The data is complemented with metadata that provides relevant information about the source text.

The high quality of the source texts and translations of the core corpus is assured by using only texts already published by renowned publishers. This makes it a very reliable data source. Moreover, a quality control system for each step of the corpus construction was performed. The corpus has been checked manually at different levels, including compilation, preprocessing, sentence splitting, and alignment. Finally, PaEnS is equipped with a user-friendly web interface allowing for efficient and refined searches tailored to the needs of each user.

---

27 TreeTagger: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/, Freeling: https://nlp.lsi.upc.edu/freeling/index.php/node/1, Universal PoS Tags: https://universaldependencies.org.

28 Berkeley Aligner: https://github.com/mhajiloo/berkeleyaligner, Giza++: http://www2.statmt.org/moses/giza/GIZA++.html, NaTools: http://linguateca.di.uminho.pt/natools/, Nile: https://jasonriesa.github.io/nile/.

All these features make PaEnS a very useful resource for multiple applications, from contrastive research, through translation studies to foreign language learning and teaching.

Acknowledgments

The research reported in this paper has been developed as part of the project *PaGeS 2.0 Optimization of a multipurpose resource* (2017–2022, FFI2017-85938-R) and the project *PaCorES: Online Spanish Parallel Corpora* (2022–2026, PID2021-125313OB-I00), funded by the State Research Agency (AEI) of Spanish Ministry of Science, Innovation and Universities.

References

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers (Ed.), *Terminology, LSP and translation* (pp. 175–186). Benjamins.

Bernardini, S. (2004). Corpora in the classroom. In J. Sincalir (Ed.), *How to use corpora in language teaching* (pp. 15–36). John Benjamins.

Bernardini, S., Ferraresi, A., & Miličević, M. (2016). From EPIC to EPTIC. Exploring simplification in interpreting and translation from an intermodal perspective. *Target: International Journal of Translation Studies*, *28*(1), 61–86.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257. https://doi.org/10.1093/llc/8.4.243

Brown, P.F., Lai, J.C., & Mercer, R.L. (1991). Aligning sentences in parallel corpora. In Proceedings of the 29th Annual Meeting on Association for Computational Linguistics, ACL '91*, Stroudsburg, PA* (pp. 169–176). ACL.

Cettolo, M., Girardi, C., & Federico, M. (2012). *WIT3: Web inventory of transcribed and translated talks*. Proceedings of EAMT, Trento, Italy, pp. 261–268.

Dörk, M., & Knight, D. (2015). WordWanderer: A navigational approach to text visualisation. *Corpora*, *10*(1), 83–94.

Doval, I. (2017). La construcción de un corpus paralelo bilingüe multifuncional. *Moenia. Revista lucense de lingüística y literatura*, *27*, 125–141.

Doval, I. (2018). Corpus paralelos en la enseñanza de lenguas extranjeras: un ejemplo de aplicación basado en el corpus PaGeS. *CLINA*, *4*(2), 65–82. https://doi.org/10.14201/clina2018426582

Doval, I. (2022). Parallelkorpora: Ergänzung oder Ersatz bilingualer Wörterbücher? Möglichkeiten und Grenzen der didaktischen Nutzung von Parallelkorpora vs. bilingualen Wörterbüchern für den Fremdsprachenunterricht.

In I. Leibrandt & K. Jahn (Eds.), *Arbeitswelten von gestern bis heute* (pp. 185–210). Lang.

Doval, I., Fernández Lanza, S., Jiménez Juliá, T., Liste Lamas, E., & Lübke, B. (2019). Corpus PaGeS: A multifunctional resource for language learning, translation and cross-linguistic research. In I. Doval & M.T.S. Nieto (Eds.), *Parallel corpora for contrastive and translation studies: New resources and applications* (pp. 103–121). John Benjamins.

Egbert, J., Biber, D., & Gray, B. (2022). *Designing and evaluating language corpora: A practical framework for corpus representativeness*. Cambridge University Press. https://doi.org/10.1017/9781316584880

Graën, J., Batinic, D., & Volk, M. (2014). Cleaning the Europarl corpus for linguistic applications. In *Konvens*. Stiftung Universität Hildesheim.

Heid, U. (2008). Corpus linguistics and lexicography. In A. Lüdeling & M. Kyto (Eds.), *Corpus linguistics. An international handbook* (Vol. 1, pp. 131–153). Walter de Gruyter.

Johansson, S. (2007). Using corpora: From learning to research. In E. Hidalgo, L. Quereda, & J. Santana (Eds.), *Corpora in the foreign language classroom* (pp. 17–30). Rodopi.

Johansson, S. (2009). Some thoughts on corpora and second-language acquisition. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 33–44). John Benjamins.

Koehn, P. (2005). *EuroParl: A parallel corpus for statistical machine translation*. Proceedings of the Machine Translation Summit, Phuket, Thailand, pp. 79–86. http://www.statmt.org/europarl/

McEnery, T., & Hardie, A. (2012). *Corpus linguistics*. Cambridge University Press.

McEnery, T., & Xiao, R. (2007). Parallel and comparable corpora: What are they up to?. In G. James & G. Anderman (Eds.), *Incorporating corpora: Translation and the linguist*. Multilingual Matters. http://someya-net.com/104-IT_Kansai_Initiative/corpora_and_translation.pdf

Nádvorníková, O. (2017). Pièges méthodologiques des corpus parallèles et comment les éviter. *Corela HS-21*. http://corela.revues.org/4810

Sinclair, J. (1996, mayo). Preliminary recommendations on corpus typology (EAGLES Document EAG-TCWG- CTYP/P). Expert Advisory Group on Language Engineering Standards. http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html

Steinberger, R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Schlüter, P., Przybyszewski, M., & Gilbro, S. (2014). An overview of the European Union's highly multilingual parallel corpora. *Language Resources and Evaluation Journal (LRE)*. https://doi.org/10.1007/s10579-014-9277-0

Tiedemann, J. (2011). *Bitext alignment*. Morgan & Claypool.

Tiedemann, J. (2012). *Parallel data, tools and interfaces in OPUS*. Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012). http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

Tiedemann, J. (2016). OPUS—parallel corpora for everyone [Special issue]. *Baltic Journal of Modern Computing (BJMC)*, *4*(2). Proceedings of the 19th Annual Conference of the European Association of Machine Translation (EAMT). https://www.bjmc.lu.lv/fileadmin/user_upload/lu_portal/projekti/bjmc/Contents/4_2_28_Products.pdf

Tóth, K., Farkas, R., & Kocsor, A. (2008). Sentence alignment of Hungarian–English parallel corpora using a hybrid algorithm. *Acta Cybern*, *18*, 463–478.

Varga, D. (2012). *Natural language processing of large parallel corpora* [PhD dissertation, Eötvös Loránd University].

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., & Trón, V. (2005). *Parallel corpora for medium density languages*. Proceedings of the RANLP, pp. 590–596.

Volk, M., Graën, J., & Callegaro, E. (2014). *Innovations in parallel corpus search tools*. Proceedings LREC 2014 (Language Resources and Evaluation Conference), pp. 3172–3178.

Wetzel, D., & Bond, F. (2012). *Enriching parallel corpora for statistical machine translation with semantic negation rephrasing*. Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation, pp. 20–29.

Primary sources

[3007] Rowling, J.K. (2000). *Harry Potter and the Globet of Fire*. Bloomsbury Publishing. *Harry Potter y el cáliz de fuego*. Salamandra.

[3012] Peterson, J.B. (2018). *12 Rules for Lif*e. Allen Lane. *12 Reglas para Vivi*r. Planeta.

[3101] Pérez Reverte, A. (2002). *La reina del sur*. Alfaguara. *The Queen of the South*. Picador.

[3132] Pen, P. (2011). *El aviso*. RBA Libros. *The Warning*. Amazon Crossing.