Corpus PaGeS: A Multifunctional Resource for Language Learning, Translation and Cross-Linguistic Research

Irene Doval, Santiago Fernández Lanza, Tomás Jiménez Juliá, Elsa Liste Lamas, Barbara Lübke
University of Santiago de Compostela, Spain

Abstract: This chapter presents the bilingual parallel corpus PaGeS, compiled by the research group SpatiAlEs from the University of Santiago de Compostela. PaGeS currently amounts to nearly 20 million tokens and consists of texts originally written in German and in Spanish and their correspondent translations into the other language, as well as a small portion of German and Spanish translations from third languages. The present contribution introduces the main characteristics of the PaGeS corpus, focusing on its design and compilation. It first explains the criteria for the selection of the texts and the details of text pre-processing, automatic alignment and manual review. It then addresses the search and display features describing the server architecture and indexing process. Finally, the intended development of the PaGeS corpus is briefly discussed.

Keywords: parallel corpora, corpus alignment, corpus visualization, Spanish, German

1. Introduction

The project[1] to create the bilingual parallel corpus PaGeS[2] (**Pa**rallel Corpus of **Ge**rman and **S**panish) emerged as a result of the research demands of linguists studying the expression of spatial relations and motion events in German and Spanish. The corpus is meant to provide reliable data of written

[1] This project (2014-2020) is being carried out at the University of Santiago de Compostela by the research team SpatiAlEs, led by Prof. Irene Doval.

[2] <http://corpuspages.eu>

language use as the basis for contrastive analysis in this field of investigation.

In order to meet this purpose, PaGeS records samples of present-day German -including as well Swiss and Austrian texts- and Spanish -European and Latin American- from a variety of genres rich in lexical forms and grammatical patterns related to spatial notions. Source texts of both languages and their respective translations are carefully aligned, identifying corresponding sentences and smaller text segments in the original and translated versions. The query system allows monolingual as well as bilingual searches for words and sequences.

PaGeS has been conceived as a multi-modal and multi-functional language resource that will be publicly available. Among its multiple applications, it can be used in contrastive and general linguistics research in German and Spanish, translation studies, as well as in language teaching and language learning. Translators and learners at intermediate to advanced levels can use the corpus for finding multiple translation equivalents and contexts of use (Doval 2017b: 126).

Existing multilingual corpora and other language resources that include German and Spanish bitexts offer valuable information on this pair of languages, but they scarcely assemble all the features required to meet the previously mentioned purposes. Some limitations relevant to our aims of research and application motivated our decision to build the PaGeS corpus.

First of all, most of the openly available multilingual resources are restricted to samples of highly specific text-types and domains, representing mainly legal and administrative or technical language use. This is the case

of an important number of parallel corpora derived from source texts produced by the different institutions of the European Union (see Steinberger et al. 2014 for an overview). These are contained, for example, in the corpus included in *Multilingwis*[3] (Clematide et al. 2016), a search tool developed at the university of Zürich that covers the debates of the European Parliament in seven languages. The same applies to the major part of the large collection of parallel corpora included in the *Opus* project authored by Jörg Tiedemann at Uppsala University (Tiedemann 2009). Besides administrative documents from European institutions, *Opus*[4] covers journalistic texts and some minor collections from different online sources, as subtitles and technical documentation. Another huge multilingual parallel corpus is *InterCorp*[5], compiled at Charles University in Prague (Čermák, this volume). It covers 40 languages including German and Spanish and uses Czech as pivot language, that is all texts are aligned with a single Czech version (original or translation). A smaller part of the corpus, the so-called *core*, consists mostly of fiction, whereas the greater part, the so-called *collections*, covers political commentaries, legal texts from EU-institutions, proceedings of the European Parliament and subtitles of movies and TV series.

Text-types specialized in legal, administrative or technical usage are not likely to provide the empirical data needed for linguistic inquiry into the expression of space and motion, our primary field of investigation, since they exhibit little variation in lexical forms and grammatical structures

3 <https://pub.cl.uzh.ch/projects/sparcling/multilingwis2.demo>
4 <http://opus.lingfil.uu.se/>
5 <http://ucnk.korpus.cz/intercorp/?lang=en>

related to this semantic domain. Apart from linguistic research, the use of specialized corpora seems to be rather difficult in other fields of application, such as language teaching and language learning. As far as the fiction-oriented *core* part of *InterCorp* is concerned, users can search bilingual subcorpora for specific language pairs. In its current ninth release, the Spanish–German subcorpus contains 16.5 million tokens (Čermák, this volume). It has to be pointed out, however, that these include not only direct translations but also translations from third languages into Spanish and German[6]. Other subcorpora included in this large multilingual parallel corpus only comprise fictional texts written in German and their translations into Spanish. That is the case of the German–Spanish subcorpus included in COVALT[7]. Still, this subcorpus is rather small, as it only contains 834,161 words and it does not cover Spanish texts translated into German (Molés-Cases & Oster, this volume).

Another very important factor to our research and to translation-based contrastive linguistics in general is the identification of source and target language. Investigation may be based on the analysis of direct translations between two languages under study, or may include the comparison of target texts translated from a third language as well. In any case, source and target texts have to be clearly differentiated and the direction of translation should be taken into account in order to derive sound conclusions from contrastive analysis. Available parallel corpora, however,

---

6 According to Rosen (2016: 29; 31), in the 8th release of InterCorp the Spanish–German subcorpus in the core had a size of 11 million words but included only 587 thousand words in Spanish translations from German originals, and 901 thousand words in German translations from Spanish originals.

7 <http://www.covalt.uji.es>

do not always provide the relevant data. As far as documents of EU institutions are concerned, Steinberger et al. (2014: 6) observe:

> The source language for most documents produced by the EU institutions is no longer known. This information is not part of the explicit meta-information available for the documents. […] It is likely that at least some documents were translated via an intermediate language, that is that there are translations of translations.

With respect to resources compiled automatically from bilingual web sites, such as *Linguee*[8], it appears to be even more difficult to determine the source language of the included texts.

On the whole, we found that direct translations between German and Spanish are rather scarce or cannot be safely identified in available multilingual parallel language resources.

Through the compilation of PaGeS we hope to compensate for the problems of existing resources and provide an extensive base of empirical data suitable for contrastive research and bilingual information on German and Spanish. In the following sections, we present the characteristics of the corpus focusing on its composition (Section 2), the text preprocessing, (Section 3), the automatic alignment and its manual review (Section 4), the search and display features (Section 5) and the server architecture and indexing process (Section 6). Finally, Section 7 provides a brief summary and our outlook on the further development of the corpus.

8 <http://www.linguee.com>

2. Components and Content

At the present stage (July 2017), the corpus contains 19.2 million tokens and 655.321 bisegments, that is pairs of aligned text chunks (sentences or smaller segments) from 104 sources. The German original texts and their translations into Spanish account for 46.1%, and the Spanish original texts together with their German translations make up 37.6% of the total tokens. A smaller part of the corpus (16%) contains translations into German and Spanish from other languages (so far English, French, Italian and Swedish). We think these texts provides useful data especially for translation studies since it allows the comparison of German and Spanish translations done from a third language source text. Table 1 gives a detailed account of numbers of tokens and sources at the present stage. Ongoing enlargement of the corpus is aimed at improving the balance of German and Spanish source language and at establishing a proportion of 10% of texts from a third language.

Table 1. Composition of PaGeS

Components	Sources	Tokens	Percent
German original text	54	4.253.900	22.4
German translation from Spanish	38	3.564.688	18.7
Spanish original text	38	3.584.908	18.9
Spanish translation from German	54	4.507.832	23.7
German translation from 3rd languages	12	1.577.794	8.0
Spanisch translation from 3rd languages	12	1.528.715	8.3
Total	**104**	**19.017.837**	

The compilation of the corpus has been restricted, for obvious reasons, to texts being available in both versions, original and direct translation, or, for a minor part of components, to being available in translation of a third-language source text into German and Spanish. For

technical reasons, texts available in digital format were preferred. Apart from these conditions, the quality of the material has been a decisive criterion of text-selection. For this reason, only published editions, of original texts as well as translations, which had passed the control of established publishing houses, have been included. On the other hand, as mentioned before, we preferred text-types and genres that were expected to present a wide range of expressions related to spatial relations and motion events.

The vast majority of the corpus texts (95%) belong to samples of narrative fiction, dating, with very few exceptions, from the last 50 years (1967 – 2014); the bulk of texts are from the first decades of the current century. Although comprising a considerable range of subgenres from literary as well as from genre fiction, a significant part of the samples comes from crime, historical fiction, children's literature and young adult fiction, that is from genres particularly dense and varied in the description of spatial configurations and motion events. The corpus also covers a small proportion of non-fiction samples (5 %) from genres like self-help books, travel diaries and biographies.

3. Text Preprocessing, Textual Mark-Up and Metadata

After having selected the texts according to the criteria mentioned in Section 2, these have to be preprocessed and prepared for the alignment process. For this purpose, all texts are stored in a .txt format, using the common character encoding UTF-8. The first step of preprocessing aims at reducing the noise and achieving as much parallelism as possible between source and target

text in order to achieve better results with the alignment software. First of all, passages that do not belong to the body text are removed, that is, front matter, such as title, colophon, frontispiece, dedication, epigraph, contents, foreword, preface, and back matter, such as appendix, bibliography, author's and/or translator's notes, glossaries, etc. Moreover, chapters epigraphs, pictures and captions are discarded. Afterwards, the texts are carefully proofread in order to detect and amend errors caused by the digitalization or conversion process. However, we do not correct spelling mistakes and do not adapt the German and Spanish original texts to the current spelling conventions.

Then, textual mark-up corresponding to the internal structure of the texts, that is the divisions in parts, chapters and pages, is tagged. Smaller units, such as paragraphs, and further information concerning text formatting have not been considered yet.

Finally, descriptive metadata, that is "tags which encode descriptions of the corpus and its constituent texts" (Wynn 2008: 714 -715) are collected and stored. The metadata list aims to be as detailed as possible, since we think this information could be extremely valuable for the conducting of linguistic and/or translational studies. For pairs of texts originally written in German and Spanish and then translated into Spanish and German, the metadata includes the following information: author's and translator's name, original and translated title, publication date and publisher of the original and translated version, original language and language version, and genre. Moreover, information about copyright, basic statistics (number of tokens, words and bisegments), and the name(s) of the alignment reviewer(s) are

included. For texts originally written in another language and translated into German and Spanish, the metadata list additionally includes the original title and the publication year of the original version. This information is held in a separate database in tsv format and is linked to the single bitext files after the alignment process described in the following section.

4. Alignment[9]

Needless to say, the choice of an alignment tool constitutes one of the most fundamental decisions in the construction of a parallel corpus. In fact, the alignment and its accuracy play a crucial role for both the building and the exploitation of the corpus. For the corpus PaGeS, we decided to focus on sentence alignment systems, since this level of alignment is the most established for parallel corpora (see among others Tiedemann 2011: 37; Volk et al. 2014).

The choice of the alignment tool for PaGeS based on several tests conducted on five German and Spanish original texts and their corresponding translations using the following alignment tools: ABBYY Aligner[10], bitext2tmx[11], cwb-align[12], LF-Aligner, based on Hun-Align[13], and

[9] The terminology used in this section is mostly based on Tiedemann (2011) and Zanettin (2012).

[10] Available on https://www.abbyy.com/en-eu/aligner/. The PELCRA Polish-Russian parallel corpus was aligned with ABBY Aligner (Łaziński & Kuratczyk 2016).

[11] Available under http://bitext2tmx.sourceforge.net/.

[12] Included in the IMS CWB Open Corpus Workbench (Evert & CWB Development Team 2016).

[13] LF Aligner is available under https://sourceforge.net/projects/aligner/. *Hun-Align* was used, for example, for the alignment of Intercorp (Čermák & Rosen 2012) and in the platform Multilingwis (Clematide et al. 2016). It is also part of the Uplug-tool used for the building of the OPUS corpora.

Vanilla aligner[14]. LF Aligner achieved the highest alignment accuracy and was therefore chosen.

LF Aligner is a GUI wrapper that includes the hunalign sentence alignment system (Varga 2012: 199). 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). The alignment process runs in three main steps[15]: (1) hunalign "builds alignments using a simple similarity measure" (Steinberger et al. 2006: 5), which is based "on sentence lengths and the ratio of identical words" (Steinberger et al. 2006: 5); (2) it builds a bilingual lexicon based on this first alignment; (3) it returns the alignment also taking into consideration the lexical similarity by means of the created dictionary (see Varga et al. 2005; Steinberger et al. 2006 and Varga 2012).

The alignment accuracy mainly depends on the degree of correspondence between source and target text. Obviously, a one-to-one correspondence is not always possible since during the translation process sentences can or have to be split, merged or be reordered and the translator may choose to insert or omit sentences or whole text passages (see among others Tiedeman 2011: 9; Varga 2012: 94). The genre also plays a very important role and in this regard, the alignment of literary texts may be more difficult than technical ones (Zanettin 2012: 155). Moreover, within literary texts, the degree of correspondence varies depending on the author,

[14] Described in Danielsson & Ridings (1997) and used for the alignment of the Dutch Parallel Corpus (Macken et al. 2007).

[15] For a detailed description of the algorithm and of how it runs if a bilingual lexicon is provided from the beginning, see Varga (2012: 92-119).

the translator, the texts themselves and on the direction of translation. Furthermore, as mentioned in Section 2, PaGeS also includes bitexts in which both the German and the Spanish bitext halves are translations from a third language and these "are particularly challenging in this aspect [alignment], since they have undergone two independent translation processes" (Doval 2016: 93).

The inclusion of the aligned bitexts in the corpus PaGeS is based on two critieria. The first one is the percentage of empty alignments, for instance, a segment in one bitext half that has been linked to an empty segment in the other bitext half ($S_{src} > \emptyset_{trg}$ or $\emptyset_{src} > S_{trg}$). The second one is the percentage of segments containing more than 350 characters[16]. We generally discard bitexts showing more than 10% of empty alignments and/or 20% of segments longer than 350 characters, since their manual review and processing would be too time-consuming. However, we also take into consideration the location of the empty alignments and the long segments in the bitexts. If they are concentrated in concrete passages (e.g. long passages omitted, added or free translated), these are eliminated in both the source and the target bitext halves and the rest of the bitext is included in the corpus. Up until now, 17 of 134 bitexts have been definitively excluded. Further, we developed an effective procedure to manually review and validate the results of the alignment of the selected bitexts, and hence to improve its quality. For this purpose, we export them in Excel or Google

[16] We think that segments longer than 350 characters would hinder the results' visualization and the identification of the equivalent of a searched term in the other bitext half.

Spreadsheets. We then count the number of characters of each segment separately and calculate the ratio for each bisegment.

After splitting those segments longer than 350 characters, we identify empty alignments, which are of two types in our bitexts. On the one hand, a segment in the source bitext half may have no correspondence in the other half because of an omission or an addition. Table 2 and 3, respectively, illustrate these cases of one-to-zero correspondence and zero-to-one correspondence:

Table 2. Omission of a segment in the target text

German source text	Spanish target text
Und es wäre eine Erklärung dafür, dass Goldbergs Sohn keine vierundzwanzig Stunden, nachdem wir die Leiche seines Vaters gefunden haben, mit einer ganzen Streitmacht auftaucht, um uns an weiteren Ermittlungen zu hindern.	Es más, eso explicaría también que, después de que encontráramos el cadáver de su padre, el hijo no haya tardado ni veinticuatro horas en presentarse aquí con las fuerzas armadas al completo para impedirnos seguir con las investigaciones.
Entweder Goldberg junior oder jemand anderes hat beste Beziehungen und ein Interesse daran, die sterblichen Überreste seines Vaters so schnell wie möglich verschwinden zu lassen.	
Goldbergs Geheimnis sollte geheim bleiben.	El secreto de Goldberg tenía que seguir oculto.

Table 3. Addition of a segment in the target text

Spanish source text	German target text
Llegó septiembre.	Es wurde September.
	Mit Arnau ging es allmählich aufwärts.
Bernat ya había visto sonreír y gatear a su hijo por la cueva y sus alrededores.	Bernat hatte seinen Sohn bereits lächeln sehen und er machte auf allen vieren Ausflüge durch die Höhle und in die nähere Umgebung.

In these cases, the empty segment is tagged according to the translation direction. Empty segments resulting from an omission in the target text are tagged as n_t_s (i.e. *not translated segment*) and those resulting from an addition are tagged as a_s_t (i.e. *added segment in*

translation). In the case of bitexts derived from third languages, the empty segment is tagged as […] since both the German and the Spanish bitext halves are translations and it is impossible to determine whether the empty segment is of type n_t_s or a_s_t.

On the other hand, an empty alignment is not always the result of a one-to-zero or zero-to-one correspondence and can be due to a misalignment, for instance, a segment in one bitext half that the aligner did not match to its correspondence in the other half. These misalignments are generally occasioned by one-to-two or two-to-one correspondences, as shown in Table 4 and 5, respectively.

Table 4. Empty alignment due to a misalignment (one-to-two correspondence)

La azafata le rozó para desensimismarle.	Die Stewardess tippte ihm auf die Schulter und riß ihn aus seiner Versunkenheit.
	Sie deutete mit einem Lächeln ihres vollen, gesunden Gesichtes auf den Sicherheitsgurt.
Le indicó el cinturón con una sonrisa llena de carne sana y rouge enmarcada por una cabellera castaña, casi pelirroja, de las que no se encuentran en España.	Sie hatte Rouge auf den Wangen, und ihr langes Haar war von einem Kastanienbraun, das ins Rötliche spielte, wie man es in Spanien nicht findet.

Table 5. Empty alignment due to a misalignment (two-to-one correspondence)

Man werde sich einigen, sagte der Herzog.	Llegarían a un acuerdo, dijo el duque.
Ein Professorentitel sei möglich.	
Wenn auch nicht bei doppelten Bezügen.	La cátedra era posible, aunque sin doble sueldo.

After reviewing empty alignments, we focus on bisegments that despite having been paired do not correspond to each other or at least partially, as shown in Table 6.

Table 6. Empty alignment due to a misalignment

Saladin hat dem Tempelritter das Leben geschenkt.	Saladino le perdonó la vida al templario, fue un buen acto, un mizwa, como tú lo llamas, que propició otro, pues al templario

	debes agradecerle que tu querida hija siga con vida, no lo olvides.
Das war eine gute Tat, eine Mizwa, wie du es nennst, die sogleich eine andere Mizwa nach sich gezogen hat, denn diesem Tempelritter hast du zu verdanken, dass deine geliebte Tochter noch lebt, vergiss das nicht.« Recha öffnete den Mund, um ihr zu widersprechen, doch Daja legte ihr begütigend die Hand auf den verbundenen Arm und das Mädchen presste die Lippen zusammen und schwieg.	Recha abrió la boca para replicar, pero Daja le puso una mano tranquilizadora en el brazo vendado y la niña apretó los labios y calló.

To identify them, we look at the ratio and focus on those bisegments with a value between 7 and -7, since we discovered that these tend to display misalignment and have therefore to be checked and realigned as necessary. Beside the degree of correspondence between bitext halves, segmentation is another aspect that should be taken into consideration, as pointed out by Tiedemann (2011: 9 -11). We are currently examining to what extent improvements in the texts' automatic splitting could improve the alignment results. For instance, the sentence splitter does not consider some punctuation marks as sentence boundaries, such as "…" both in German and Spanish, "?" and "!" before "—", and ";" in Spanish, as shown in Table 7.

Table 7. Automatic splitting of sentences containing … and ? —

Segmentation of the German text	Segmentation of the Spanish text
Man wollte eine Papierfabrik in Krakau bauen, die Metallindustrie in Riga auf Vordermann bringen, eine Zementfabrik in Tallinn errichten und so weiter.	Se abrió una fábrica papelera en Cracovia, se reformó una industria metalúrgica en Riga, una fábrica de cemento en Tallin… **La dirección del CADI, compuesta por pesos pesados del mundo de la banca y de la industria suecas, repartió el dinero.**
Die Gelder wurden von den Vorständen des SIB verteilt, lauter einflussreichen Persönlichkeiten aus der Welt der Banken und der Großindustrie.«	—¿Te refieres al dinero de los contribuyentes**? —Alrededor del cincuenta por ciento provenía de subvenciones estatales; el resto lo pusieron los bancos y la industria.**
»Steuergelder also?«	

»Ungefähr 50 % staatliche Zuschüsse, den Rest steuerten die Banken und die Unternehmen selbst bei.	

Introducing a boundary before the two sentences marked in bold made in this case a difference in the alignment results, as shown in Table 8 and 9, respectively:

Table 8. Alignment based on the original sentence splitting

German target text	**Spanish target text**
Man wollte eine Papierfabrik in Krakau bauen, die Metallindustrie in Riga auf Vordermann bringen, eine Zementfabrik in Tallinn errichten und so weiter.	Se abrió una fábrica papelera en Cracovia, se reformó una industria metalúrgica en Riga, una fábrica de cemento en Tallin… La dirección del CADI, compuesta por pesos pesados del mundo de la banca y de la industria suecas, repartió el dinero.
Die Gelder wurden von den Vorständen des SIB verteilt, lauter einflussreichen Persönlichkeiten aus der Welt der Banken und der Großindustrie.« »Steuergelder also?«	—¿Te refieres al dinero de los contribuyentes? —Alrededor del cincuenta por ciento provenía de subvenciones estatales; el resto lo pusieron los bancos y la industria.
»Ungefähr 50 % staatliche Zuschüsse, den Rest steuerten die Banken und die Unternehmen selbst bei.	

Table 9. Alignment based on the improved sentence splitting

German target text	**Spanish target text**
Man wollte eine Papierfabrik in Krakau bauen, die Metallindustrie in Riga auf Vordermann bringen, eine Zementfabrik in Tallinn errichten und so weiter.	Se abrió una fábrica papelera en Cracovia, se reformó una industria metalúrgica en Riga, una fábrica de cemento en Tallin…
Die Gelder wurden von den Vorständen des SIB verteilt, lauter einflussreichen Persönlichkeiten aus der Welt der Banken und der Großindustrie.«	La dirección del CADI, compuesta por pesos pesados del mundo de la banca y de la industria suecas, repartió el dinero.
»Steuergelder also?«	—¿Te refieres al dinero de los contribuyentes?

These types of differences led to a hypothesis that considering more punctuation marks as sentence-boundary markers and exploring systematic differences in the German and Spanish punctuation systems could imply improvements in the alignment accuracy.

Overall, although being rather laborious and time-consuming, the procedures described in this section ensure a very high degree of alignment accuracy, which is essential for our research purposes. Moreover, and crucially, the manual reviewing process will also contribute to improving alignment accuracy of new bitexts, since the manually validated bitexts will be incorporated into a training corpus by means of which the aligner tool will be regularly retrained.

5. Search and Display Features

As Dörk and Knight (2015: 84) assess, many existing corpora are "aimed mainly at people with expertise with linguistics". 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, we do not want to limit the target users of the corpus PaGeS to expert ones. Therefore, and in order to make the corpus a really multipurpose tool, that is 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 German 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 corpus should provide the following functionalities (Doval 2017b: 191ff):

a. *Fast search*: Given that a significant increase in size of the corpus PaGeS is planned, the search engine must allow searches in a quick and efficient way through large amounts of language data.

b *User-friendly search*: The query language must be as simple as possible. An advanced, more complex, query language is only displayed if required. In addition, the search habits of users on the internet should be exploited, and thus, the query language of Google should serve as a model.

c. *Multi-level search*: The search system must allow queries across multiple layers of linguistic annotation, such as lemmatization and POS tagging.

d. *Display*: The search results must be displayed in an easy-to-read format. The matching segments have to be displayed side-by-side, and both the search word or phrase and its potential equivalent have to be highlighted.

To match the requirements of the above-mentioned users we have designed a three-level search. The first one, whose provisional web interface is shown in Figure 1, is the simple or standard search. In this case, the user only has to enter in the search field the search term (a word or a phrase) in German or Spanish. In these types of queries, lemmatization is applied by default. With multiword queries, all search words within a specific distance are found. Similar to Google search, if the term is enclosed between quotation marks " ", the search only returns results that exactly match the entered word form or phrase.

As Wynne (2008: 706) points out, "the most popular way to display the results of a search in a corpus is in the form of a concordance" and the most common concordance format is the KWIC (Key Word in Context) concordance, that is the node word is in a central position with all lines vertically aligned around the node. This presentation of the results, properly

sorted, is very useful in monolingual corpora for visualizing patterns of use. However, in bilingual corpora, the KWIC format cannot be considered user-friendly, since one of the main applications of these corpora is to quickly find possible equivalents of a search term.

For this reason, we decided for the visualization of the query results in a two-column html table, where one column corresponds to source texts and the other one to target ones. The search term is displayed in a cell with some context and is highlighted in bold. Depending on whether the search term is located in a source or a target text, it is shown in one column or in the other.

Occurrences in source texts are displayed in the left column, while those in target texts are shown in the right one. The correspondent segment is displayed in the same row but in a cell of the other column. With each query, information on the number of hits, the total number of pages, as well as the current page number is shown. The number of segments per page is fixed at 30, but this could be easily changed in the web application configuration file. Figure 1 provides an example of the standard search menu and some of its features:

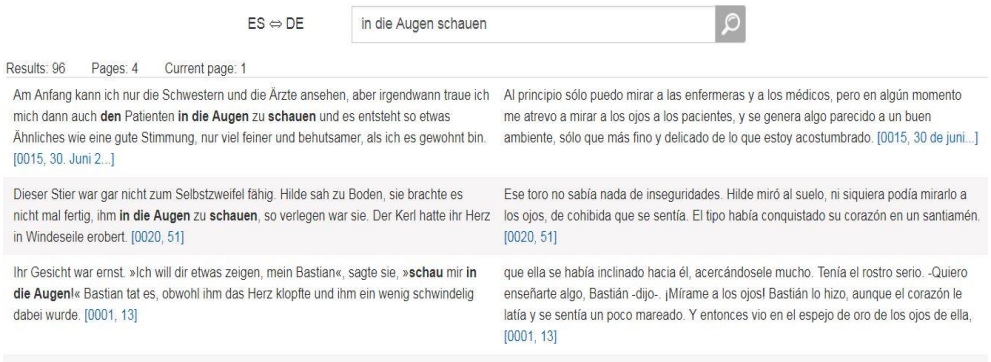


Figure 1. Standard search menu

At the bottom of the table, a set of links are available to allow the user to navigate through the pages and to download the query results in three formats: Excel, ODS and CSV (only available for registered user).

In each cell, information concerning the text ID, the corresponding part or chapter name and/or number are displayed in blue square brackets. By clicking on the work ID, the user can see a larger linguistic context and select the number of segments above and/or below they will see. Moreover, this screen shows detailed information concerning the bibliographic information, as shown in Figure 2.

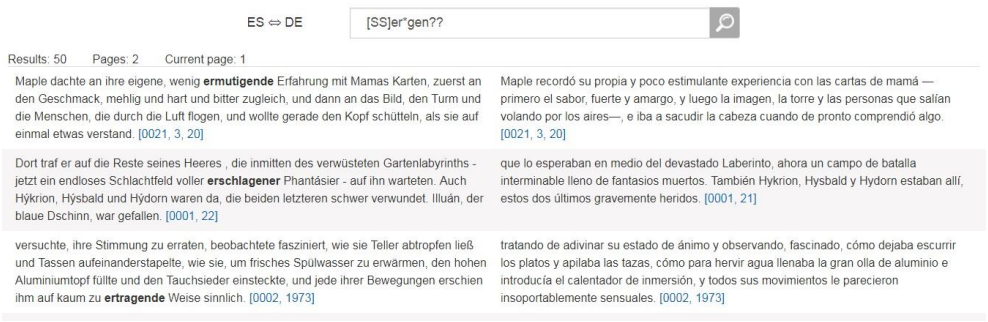
Figure 2. Scope of context and bibliographic information

The second level of search, as of yet not implemented, is in the advanced one. 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, translation from third languages and part-of-speech-tags. The occurrences are expected to be sorted by various criteria.

The last search level is the most complex one, and is currently partially available over the standard search interface. This level gives the user full command of the query syntax used in the underlying query tool:

Solr. (Version 5.0.0)[17]. This supports searches using regular expressions (RegEx[18]). The search term has to be preceded by [SS] (Solr Search). The search expression has to be constructed word by word, and for each word, it will be possible to specify several parameters. Figure 3 shows a formal search.

Finally, this level will allow the combination search of words or phrases and POS tags (see Section 7). This formal search is naturally very powerful in executing precise search queries, but not particularly user-



friendly. It is intended for more demanding users, such as researchers in contrastive linguistics or translation, who usually need a very specific subset of results, only possible with complex queries including a large set of parameters.

Figure 3. Formal search menu

6. Server Architecture and Publishing Data

The PaGeS corpus platform is mainly composed by a search engine and a web application. Both components are installed in an Ubuntu virtual machine hosted at Amazon[19]. The search engine is a Solr server that

[17] <http://lucene.apache.org/solr/> (12 February 2018)
[18] https://www.regular-expressions.info/ (12 February 2018).
[19] <https://aws.amazon.com/>(12 February 2018).

contains all corpus indexed data and a lemma dictionary for the lemmatized search. The web application is developed with java Grails framework[20] and deployed in a Tomcat[21] server. This web application contains a component including Java methods, that calls the *solr-solrj* library[22]. This library is a Solr java client that communicates the PaGeS Web Application with the Solr server.

After the alignment review described in Section 4, some processes have to be performed in order to publish in a Solr server, all the information of the aligned bitexts. The result of the alignment manual review is a set of text files with the following tsv-format:

Text_Segment_ID<TAB>Text_Segment_Lang_1<TAB>Text_Segment_Lang_2

First, it is necessary to generate a lemma dictionary for each language in order to perform the lemmatized search[23]. This kind of dictionary consists of sets of word clusters, whereas each cluster groups all words appearing in the corpus under the same lemma. The lemma information was obtained through TreeTagger[24] for German and FreeLing[25] for Spanish, the tools we have chosen for POS tagging after several performance tests[26]. Both dictionaries are stored in a configuration file of the

---

20 <https://grails.org/>(12 February 2018).

21 <http://tomcat.apache.org/> (12 February 2018).

22 <https://wiki.apache.org/solr/Solrj> (29 August 2017)

23 The lemmatizer of the TreeTagger is not able to correctly lemmatize verbs when they occur with verb stem and particle split. Volk et al. (2016) have developped an algorithm to re-attach the separated prefix to the verb. This script was kindly made available to the corpus PaGeS by Prof. Volk and it is intended to implement it soon.

24 http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (29 August 2017)

25 http://nlp.lsi.upc.edu/freeling/node/1 (29 August 2017).

26 However, in the future we intend to test again the TreeTager with the Spanish texts after having trained it on a hand tagged subcorpus of our texts. The mentioned tests were carried out with the standard parameter files distributed with TreeTagger, without any additional manually training corpus (see Doval 2017a).

Solr server and are used for query expansion at searching time. Each dictionary contains a comma-separated word cluster per line, according to the following format:

word_1, word_2, …

...

The next step consists of adding the metadata mentioned in Section 3. This information is actually stored in a text file (in "field=value" format). Each reviewed bitext (stored as .tsv file) has a corresponding metadata document. Finally, all information (aligned segments and metadata) is indexed at the Solr search engine. Solr provides a fast indexing and quick searching tool for the corpus. As Solr is a general purpose search engine, data and information can be respectively structured and encoded without many restrictions.

In order to deal with this goal, a Tsv2Solr java program was implemented. This software calls the above-mentioned solr-solrj library again and then reads the content of tsv aligned text files and metadata files, opens a new Solr server session, eventually deletes all previous information and adds the new data. Figure 4 shows the PaGeS architecture and both the components and technologies used.

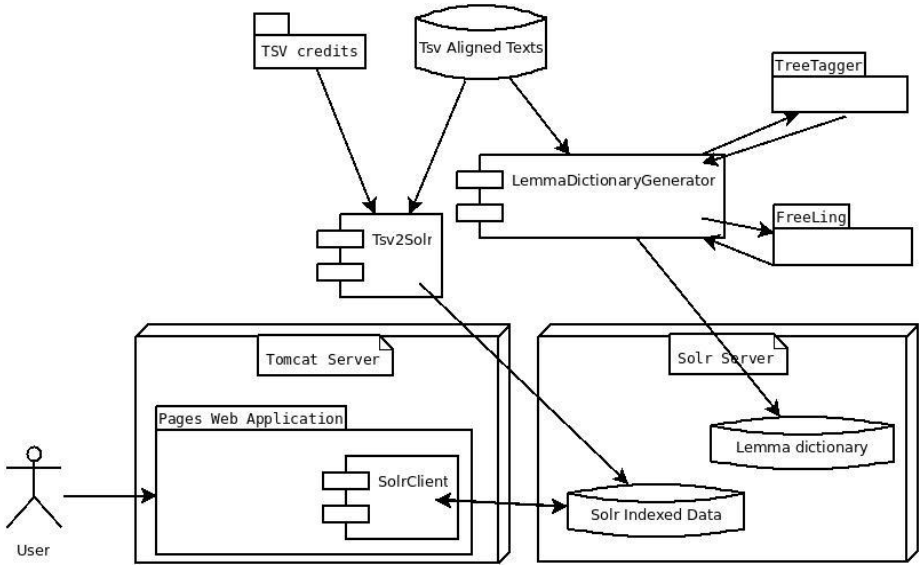


Figure 4. PaGeS Platform Architecture

7. Summary and Outlook

In the previous sections, we have presented the main characteristics of the corpus PaGeS and the processes involved in its compilation and its indexation. Starting with very specific research objectives (Section 1), we have created a valuable parallel corpus German/Spanish, which includes texts carefully selected (Section 2) and offers a very high alignment quality (Section 4). The three type of search levels (Section 5) are directed to very different users and thus will make the corpus PaGeS a multifunctional resource with an enormous potential. Its concrete applications in contrastive linguistics and language learning are currently being exploited within our research group (Doval 2018, Lübke & Liste Lamas 2018).

We are of course aware of some of the limitations of our corpus and intend to introduce new functionalities as soon as possible. As reported in Section 5, we first intend to implement an interface for advanced searches. Moreover, since our bitexts are already POS tagged, this information will be available soon. (6). The inclusion of POS tag information should allow much more complex and precise queries. In a later stage of development, we also intend to add word alignment. This alignment level is highly useful, since it allows the identification at a glance of the equivalent of a search term. Two word aligners are currently being tested: Giza++[27] and NATools[28].

[27] <https://github.com/moses-smt/giza-pp>
[28] <http://linguateca.di.uminho.pt/natools/>

We are also very aware of some current shortcomings of our search engine, such as the non-availabiliy of KWIC format or the inflexibility of the results sorting. It should soon be possible to sort the segments according to different criteria, like alphabetical order, publication date, text type or equivalent of the search word or expression. For all these reasons, we are considering whether it will be possible to continue using a general search engine such as Solr or it would be preferable to switch to another more specific system, such as Corpus Workbench[29].

Acknowledgement

The research project PaGeS Corpus has been funded by the Spanish Ministry of Economy and Competitiveness, Agencia Estatal de Investigación (FFI2013-42571-P, FFI2017-85938-R) and by the Galician Government (GI-1954 LitLinAl).

References

Aijmer, Karin. 2008. Parallel and comparable corpora. In *Corpus linguistics. An International Handbook*, Anke Lüdeling & Merja Kytö (eds), 275-292. Berlin: Walter de Gruyter.

Čermák, Petr. This volume. InterCorp. Parallel corpus of 40 languages. In *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications*, Irene Doval & M. Teresa Sánchez (eds). Amsterdam: John Benjamins.

Čermák, František & Rosen, Alexandr. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 13(3): 411-427.

Clematide, Simon, Graën, Johannes & Volk, Martin. 2016. Multilingwis - A multilingual search tool for multi-word units in multiparallel corpora. In *Computerised and Corpusbased Approaches to Phraseology: Monolingual and Multilingual Perspectives - Fraseología computacional y basada en*

[29] <http://cwb.sourceforge.net/>

corpus: perspectivas monolingües y multilingües, Gloria Corpas Pastor (ed), 447-455. Geneva: Tradulex.

Danielsson, Pernilla & Ridings, Daniel. 1997. Practical presentation of a Vanilla Aligner. In *TELRI Workshop in alignment and exploitation of texts*, *Ljubljana, Slovenia.* <http://www.kfben.com/dfilea/3122035922vanilla/ljubljana.pdf> (30 May 2017).

Dörk, Marian & Knight, Dawn. 2015. WordWanderer: A navigational approach to text visualisation. *Corpora* 10(1): 83-94.

Doval, Irene. 2016. PaGeS: Design and compilation of a bilingual parallel corpus German Spanish. *Epic Series in Languages and Linguistics* 1: 88-96.

Doval, Irene. 2017a**.** POS-tagging a bilingual parallel corpus: methods and challenges. *Research in Corpus Linguistics* 5: 35-46.

Doval, Irene. 2017b. La construcción de un corpus paralelo bilingüe multifuncional. *Moenia* 27: 125-141.

Doval, Irene. 2018. Das PaGeS-Korpus, ein Parallelkorpus der deutschen und spanischen Gegenwartssprache. *Revista de Filología Alemana* 26: 181-197.

Macken, Lieve, Trushkina, Julia, Paulussen, Hans, Rura, Lidia, Desmet, Piet & Wandeweghe, Wily. 2007. Dutch Parallel Corpus: a multilingual annotated corpus. In *Proceedings of the fourth Corpus Linguistics conference*, *University of Birmingham.* <http://ucrel.lancs.ac.uk/publications/CL2007/paper/173_Paper.pdf> (12 April 2017).

Molés-Cases, Teresa & Oster, Ulrike. This volume. Indexation and analysis of a parallel corpus using CQPweb: the COVALT PAR_ES corpus (EN/FR/DE>ES). In *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications*, Irene Doval & M. Teresa Sánchez (eds). Amsterdam: John Benjamins.

Łaziński, Marek & Kuratczyk, Magdalena. 2016 Korpus Polsko-Rosyjski Uniwersytetu Warszawskiego / The University of Warsaw Polish-Russian Parallel Corpus. In *Polskojęzyczne korpusy równoległe - Polish-language Parallel Corpora*, Ewa Gruszczyńska & Anieszka Leńko-Szymańska (eds), 83-95. Warszawa: Instytut Lingwistyki Stosowanej WLS, Uniwersytet Warszawski.

Lübke, Barbara & Liste Lamas, Elsa. 2018. Raumrelationen im Deutschen: Kontrast, Erwerb und Übersetzung. Tübingen: Stauffenburg.

Lüdeling, Anke & Kytö, Merja (eds). 2008. *Corpus Linguistics. An International Handbook.* Volume 1. Handbücher zur Sprach- und Kommunikationswissenschaft. Berlin: Walter de Gruyter.

Rosen, Alexandr. 2016. InterCorp - a look behind the façade of a parallel corpus. In *Polskojęzyczne korpusy równoległe - Polish-language Parallel Corpora*, Ewa Gruszczyńska & Anieszka Leńko-Szymańska (eds), 21-40. Warszawa:

Instytut Lingwistyki Stosowanej WLS, Uniwersytet Warszawski.

Steinberger, Ralf et al. 2006. The JRCAcquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. <https://arxiv.org/ftp/cs/papers/0609/0609058.pdf> (12 October 2017).

Steinberger, Ralf et al. 2014. An overview of the European Union's highly multilingual parallel corpora. *Language Resources and Evaluation* 48(4): 679-707.

Tiedemann, Jörg. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing* (vol. V), Nicolas Nicolov, Kalina Bontcheva, Galia Angelova & Ruslan Mitkov (eds), 237-248. Amsterdam: John Benjamins.

Tiedemann, Jörg. 2011. *Bitext Alignment*. Morgan & Claypool Publishers.

Tóth, Krisztina, Farkas, Richárd & Kocsor, András. 2008. Sentence Alignment of Hungarian–English Parallel Corpora Using a Hybrid Algorithm. *Acta Cybern* 18: 463-478.

Varga, Dániel, Németh, László, Halácsy, Péter, Kornai, András, Trón, Viktor & Nagy, Viktor. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP 2005*, 590-596.

Varga Dániel. 2012. *Natural Language Processing of Large Parallel Corpora.* Ph.D Dissertation. Budapest: Eötvös Loránd University.

Volk, Martin, Graen, Johannes & Callegaro, Elena. 2014. Innovations in parallel corpus search tools. In *Proceedings of LREC, Reykjavik*. <http://www.zora.uzh.ch/id/eprint/97282 /1/Volk_Graen_Callegaro_LREC_2014_v06.pdf> (13 May 2017)

Volk, Martin, Clematide, Simon, Graen, Johannes, Ströbel, Phillip. 2016. Bi-particle adverbs, pos-tagging and the recognition of German separable prefix verbs. In *Proceedings of the 13th Conference on Natural Language Processing* (KONVENS 2016), 296-305.

Wynne, Martin 2008. Searching and concordancing. In *Corpus linguistics. An International Handbook*, Anke Lüdeling & Merja Kytö (eds), 706-737. Berlin: de Gruyter.

Zanettin, Federico. 2012. *Translation-driven Corpora.* London & New York: Routledge.