EPiC Series in Language and Linguistics
Volume 1, 2016, Pages 88–96
CILC2016. 8th International Conference on Corpus Linguistics

PaGeS: Design and Compilations of a Bilingual Parallel Corpus German Spanish

Compilation of bilingual corpora for linguistic research

Irene Doval

University of Santiago de Compostela, Spain
i.doval@usc.es

Abstract

In this paper it will reflect on the specific needs of the linguistic research regarding the construction of bilingual parallel corpora and primarily on the conclusions to be drawn for their design, compilation and domains. A research group of the university in Santiago is currently building a bilingual parallel corpus (Corpus PaGeS) consisting of original texts in German and Spanish together with their translations into the other language, as well as German and Spanish translations from a third language. This corpus was originally intended for linguistic research purposes, specifically, the analysis of the expression of the spatial relations. Initially a brief survey of some significant existing related corpora is performed, and their limitations for linguistic studies are outlined. The different issues that were taken into account for the design of the corpus will be explained, such as type of texts, domains, regional language variety or quality and direction of translations. After describing the manual preparation process of the texts to make the documents suitable for further processing it is explained the manual and automatic annotation procedure: the metadata, and the automatically linguistic annotation. Then the process of sentence alignment and the manual review of the alignment are described and finally the next steps of future work are outlined

1 Introduction

The creation of PaGes (www.corpuspages.eu), a bilingual parallel corpus presented here, is part of a larger research project which aims at studying and analyzing spatial relations in Spanish and German (www.usc.es/spatiales). Instead of using a small *ad hoc* created corpus the research team came to the conclusion that it made more sense to create a large bilingual corpus with enough representativeness to draw sound conclusions. Even though this corpus will be used within our research team for the aforementioned purpose, efforts are being made with regard to interoperability and standardization of the corpus resource. The main idea behind this effort is that once the corpus is available it can be exploited for multiple purposes. These applications could include general research in contrastive linguistics: making text-based contrastive analysis of linguistic features possible, yielding patterns of correspondences and providing quantitative data. Closely related to this use is the application in translation studies: helping translators to find translational equivalents between German and Spanish, specific uses of lexical items and collocational and syntactic patterns as well as training automatic translation systems. The corpus will also be useful for German or Spanish learners at intermediate to advanced level for getting a great number of translation suggestions shown in usage examples.

Commonly accepted terminology (Mcenery/Zhonghua, 2007: 2-3) makes a distinction between, on the one hand, comparable corpora and, on the other, parallel corpora. Comparable corpora are composed of monolingual texts in different languages sharing subject matter, genre, text type, register and having a similar origin and extension, such as weather broadcasts, job offers, journal articles, etc. The texts of a comparable corpus are not translations of each other but they are selected according to common selection criterions (McEnery, 2003: 450). One of the most well-known comparable corpora is the *Aarhus Corpus*, a collection of contract law texts written in Danish, French and English. Parallel corpora, to the contrary, contain the same collection of texts in more than a language, one of those is the original, the other ones the translation. One of the first and best known parallel corpora is the *Hansard Corpus*, acts of the Canadian Parliament published in English and French. Parallel corpora can be bilingual or multilingual, i.e. they consist of texts of two or more languages. They can be either unidirectional (e.g. Spanish texts translated into German), bidirectional (e.g. Spanish texts translated into German and vice versa), or multidirectional (e.g. an English text such as an EU regulation translated into different languages).

In the next section we present a brief survey of related research work and we explain the motivation to create the PaGeS corpus, followed by a description of the compilation process: the different aspects taken into consideration during the design and planning, the texts used and the issues that had to be dealt with due to copyright restrictions. Next we explain the different steps of the construction of the corpus up until now: the manual pre-processing of the texts, markup, the automatic sentence alignment and the manual checking.

2 Motivation and related work

Most parallel corpora, including those in Spanish and German, are multilingual corpora. An exception to this is the small parallel corpus of documents from the Technical Regulations Information System for German-Spanish (v, 0.2), designed and compiled by Carla Parra Escartín. It is based on the European database of Technical Regulations Information System (TRIS) and contains 70,648 aligned sentence pairs and 1,563,000 words. (Parra Escartín, 2012: 2199).

By far the most important parallel corpora are the multilingual parallel language resources of the European Union. Steinberg et al. (2014) give a comparative overview of the different multilingual resources provided here. Koehn (2005) released the EuroParl sentence-aligned texts in 11 languages. That was followed by several other EU corpora: The Digital Corpus of the European Parliament (DCEP) [*] contains the documents published on the European Parliament's official website and constitutes the largest release of documents by a European Union institution. It comprises a variety of document types, from press releases to session and legislative documents. The current version of the corpus contains documents produced between 2001 and 2012. The JRC-Acquis[†] is a large parallel

[*] https://ec.europa.eu/jrc/en/language-technologies/dcep https://ec.europa.eu/jrc/en/language-technologies/dcep. For more details Hajlaoui, Najeh et al. DCEP -Digital Corpus of the European Parliament 3164-3171, http://www.lrec-conf.org/ proceedings /lrec2014/ pdf/943_ Paper.pdf.

[†] https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis

corpus consisting of multilingual legislative texts and currently comprises selected texts written between the 1950s and now. The corpus is currently aligned at document level and work is on-going to sentence-align it for all language pairs.

Worthy of special mention is OPUS (http://opus.lingfil.uu.se/) , probably the largest collection of freely available multilingual parallel corpora.·It is a growing resource and also provides tools for processing parallel data as well as several interfaces for searching the data. The language pair with the largest amount of parallel data is Spanish-English with about 36 million parallel sentences. The largest domains covered by OPUS are legislative and administrative texts, mostly from the European Union and associated institutions. There are also a substantial amount of newspaper texts and some other smaller collections from various on-line sources, such as data from the European Central Bank, subtitles and technical documentation (Tiedemann, 2012).

Today, an increasing number of bilingual corpora that are automatically compiled by scraping data from bilingual Internet sites are being released. One good example for this type of corpus is Linguee (www.linguee.com), a collection of pairs of bilingual web sites that covers some 25 languages. After identifying bilingual web sites through a search engine, the web sites and the sentences in the paired documents are aligned using dynamic programming. Obviously the best represented pairs are the ones with English. Although it includes a certain variety of texts, most of the texts are linked to the administrative or commercial text type.

As indicated above, the original motivation to create the PaGeS corpus was our specific cross-linguistic research objectives: the analysis of motion and localization event expressions and their semantic and syntactic properties in German and Spanish. For this purpose the aforementioned existing corpora entail important limitations. On the one hand corpora in sufficient data sizes are only available for language pairs, where English is one of the pair languages. Obviously neither Spanish nor German are under-resourced languages, but bilingual parallel corpora between these languages, as previously described, cover only some specialized domains, mainly administrative and commercial language, in which motion expressions do not appear in a significant/sufficient number. On the other hand, we needed direct translations between German and Spanish, where it is clear which text is the source text and which one is the translation. In the corpora above the direction of the translation is mostly unclear. For the European Union texts it can be supposed that in most cases the texts were primarily written in English and then translated into different languages. In the case of corpora compiled by scraping bilingual websites such as Linguee, the websites could have been written first in the native language of the website and then translated into different languages, but it lacks any evidence to confirm that supposition. For this reason it can be assumed that in the multilingual existing corpora a direct translation between German and Spanish is probably somewhat uncommon. In addition in most cases these corpora cannot assure quality standards for the original texts or for the translations, as they have not been in any way quality controlled. Precisely this is one of the major issues of our research, to create a solid empirical basis with an assured quality of original and translated texts.

3 Corpus Data Compilation

In the previous section we presented the existing corpora and showed why they do not fit our purposes. Clearly, a corpus is not just a collection of electronic texts, but rather these texts should be collected according to specific criteria linked to the purpose for which the corpus is created. In this section we will describe the design of the PaGeS corpus, such as type of corpus, sizes of the text samples to be included, range of language varieties and the time period to be sampled.

The PaGeS corpus has been designed as a bilingual corpus between German and Spanish, although the possibility of a further multilingual expansion of the corpus has not been precluded. From the beginning we emphasized quality with regard to content and translation. In order to assure this quality the only option is to use written texts published by prestigious publishing houses, where texts and translations pass through challenging quality control. Our plan is to compile a core corpus consisting of narrative texts, written after 1960, with a special focus on works from the last two decades. As for genre, we chose both fiction and non-fiction texts but the majority are fictional, since they constitute the greatest part of the available resources. We plan to include some 120 original entire texts and their translations, targeting an overall size of about 25 million words. In order for the corpus to be balanced according to the direction of translation, the following distribution is planned: 40% of the texts will be originally Spanish, 40% originally German and the remaining 20% will be translations from a third language. Table 1 gives an overview of the composition of the corpus at the current stage (April 2016).

Original Language	Works	Words Original	Words Translation
German	45	3550803	3686384
Spanish	36	4380704	4007941
Others	16	1789324	1863901
Total	97	9,720,831	9,558,226

Table 1

As stated, the texts are all recent texts covering a wide range of writers, from very famous authors such as García Márquez or Vargas Llosa to young writers such as Joël Dicker. Therefore all works are protected by copyright. This is one major issue in building a corpus and a factor limiting its availability. For this reason many researchers choose to use only out-of-date or copyright-free data. But we intend to research the contemporary use of language and using only texts older than 70 years, when the copyright expires, would not fulfill this goal. For the texts selected, we wrote to the copyright holders – usually publishers - asking for permission to include the texts in our corpus and use them for the purposes of linguistic and translation research. Only texts for which we receive explicit permission will be included in the corpus. As of the time of writing, all publishers with one exception have responded positively. To be allowed to use their texts, we have been subjected to following conditions: The corpus can only be used for research and it is also limited to scholars and students of research and/or educational institutions with previous registration. No commercial use is permitted.

Nevertheless, the greatest problem regarding permission is that a good part of copyright holders did not reply at all, despite the numerous attempts to make contact. To what extent their silence can be taken as implying consent is not clear. Thus, we often have to work with texts for which the permission has not yet been secured, with the attendant risk of discarding them later, in case that permission is refused. Therefore to clarify copyright issues is proving to be a challenging and frustrating task causing a lot of issues.

4 Text pre-processing and corpus markup

After we have selected and digitalized the texts, these must undergo a manual process in order to prepare them for the alignment. This consists basically of reducing the noise and achieving as much parallelism as possible between source and target text in order to obtain the best results in the alignment. This implies three tasks: removal of non-corresponding texts, bad characters and pictures, proofreading, mark up and annotation of metadata.

All texts not part of the body texts are removed, such as bibliographic information, dedications, and notes from the author or translator. In the same way, every appendix without a corresponding one in the other version is removed. Even if long texts were untranslated they are removed.[‡]

On the other hand, both versions (original and translation) are accurately proofread and corrected. It happens sometimes that the digitalization process causes some mistakes such as the insertion of a space within a word, deletion of spaces between words, or occasional character confusions. Only the mistakes linked to the digitalization process are removed, but not the mistakes in our basis edition. For the same reason we don't adapt the texts to the current German or Spanish orthographic norms in case they differ from them. At this stage, the characters and words (without spaces) in the original and in the translation were counted.

Corpus markup is a system of standard codes inserted into a document to provide information about the text itself. This metadata is used to annotate text documents or whole corpora with additional information in order to retrieve relevant information from the corpus. Each of the texts included in the corpus is given a unique ID which is used as a file name and provides information relating to the original language and the version language.

Markup for major divisions of the texts such as parts, chapters or subchapters, are inserted manually. The main goal of this procedure is to facilitate the localization of the text string within the work. These internal divisions are always constant in every edition and in the source and target text. In most cases the page number is omitted for different reasons: in some cases the basis texts were not the printed but the digitalized edition and did not have this mark up. In addition, the page number would only be useful for the localization of a text string if the searcher had the same print edition.

In addition, particular information is provided for each text in the PaGeS-corpus. These additional metadata tags are individually attached to the single text files and these metadata attachments are locally stored together with each individual text document. The metadata list includes information on the text: author and/or translator, title, date, publishing information, original language and version language, gender, as well as information on copyright and on basic statistics (number of tokens, words and bisegments) and on the manual reviewer. We follow the guidelines of the widely used markup scheme TEI (Text Encoding Initiative P5, Version 2.9.1: http://www.tei-c.org/Guidelines/P5/), which is an application of the Extensible Markup Language (XML) for the mark-up of the texts.

After proofreading and marking them up, the texts are saved in a common encoding scheme UTF-8 and they are lemmatized using Freeling (http://nlp.lsi.upc.edu/freeling, Padró 2011) for the Spanish texts and Treetagger for the German ones.

5 Sentence Alignment

A crucial step in the construction and exploitation of a parallel corpus is the alignment. Tiedemann (2011:123) defines alignment "as a process of making symmetric correspondences explicit in order to enable further processing of parallel resources". This correspondences set of two texts, one of which is the translation of the other one, a so-called bitext. The units of alignment of the bitext depend on the levels of segmentation granularity that are considered: paragraphs, sentences or words. Currently sentence alignment is standard for most parallel corpora. Therefore, we have focused on the sentence as the basic alignment unit.

In the alignment process two tasks are combined, a prior tokenization and segmentation into sentence segments and the proper linking of those sentence segments to corresponding ones. The tokenization and segmentation is done monolingually and alignment is done based on this step.

For sentence alignment numerous algorithms have been proposed in the literature, which can be grouped in three main classes of sentential alignment methodologies: length-based approaches, lexical

[‡] It happens more often than expected that for a variety of reasons some text passages are omitted in the translation.

matching approaches and hybrid approaches. On the one hand, length-based approaches exploit sentence length in terms of characters (Gale/ Church 1993) or words (Brown et al. 1993) to evaluate likeliness of an alignment of some number of sentences in source language to some number of sentences in target language. On the other hand, lexical matching approaches (Kay / Roscheisen 1993) proposed aligning the sentences by using lexicon based method; it identifies sure anchor points for the alignment using bilingual dictionaries or surface similarities of word forms. Finally, hybrid methods combine both former approaches and are the most recent, state-of-the art approaches to the problem (Braune / Fraser, 2010).

In the PaGeS corpus we used LF-Aligner (http://sourceforge.net/projects/aligner/)[§], since in several tests this alignment tool achieved the highest accuracy. It relies on Hunalign (Varga, et al., 2007), a common choice among creators of multilingual parallel corpora. It uses both sentence length and lexical correspondences to derive the final alignment, but since the lexical correspondences are themselves derived automatically, it does not require an externally supplied lexicon. The alignment occurs in a three-step process: (1) Its input is tokenized and sentence-segmented text in two languages; (2) Then the corpus is aligned using a modified version of Brown et al.'s sentence-length-based model; (3) Finally, it builds an automatic dictionary based on this alignment and then it realigns the text in a second pass, using the automatic dictionary. LF-Aligner offers a different output format: TMX, tab-delimited text or XLS. For various reasons we chose the tab-delimited format. These files are imported later to google spreadsheets to be edited.

Sentence alignment would be trivial if one sentence were always translated into exactly one sentence. During the translation process sentences might be split, merged, deleted, inserted or reordered by the translator in order to create a natural translation in the target language. Sometimes original paragraphs are summarized, rather than translated. All these issues are prominent challenges for an automatic sentence alignment.

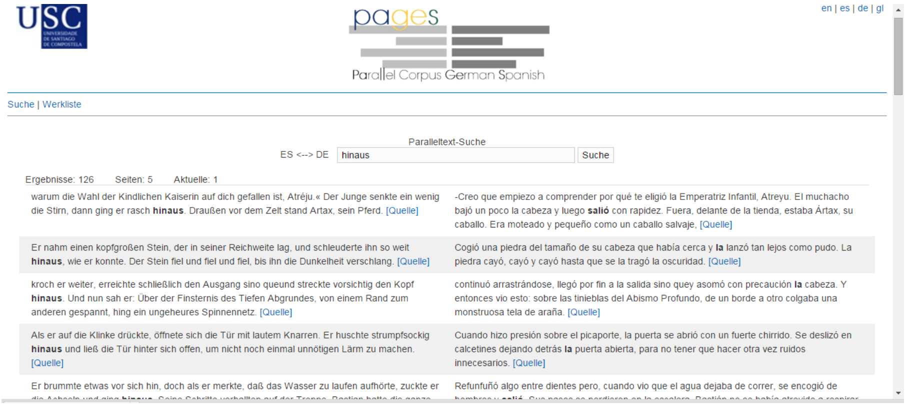
The accuracy of the autoalignment depends entirely on the quality of the source material and the manner of translation. Thus, LF-Aligner achieves in PaGeS corpus in some works an accuracy percentage of 98%, but in other works it can descend to levels lower than 90%. In the latter case we have discarded the works, as the huge amount of manual checking would not have been worthwhile. This is because the degree of correspondence between source and target texts varies significantly depending on the texts themselves, on the translators and on the direction of the translation. The texts in which the German and the Spanish version are translations from a third language (up until now only from English or French) are particularly challenging in this aspect, since they have undergone two independent translation processes.

After the automatic alignment, as said, we manually validate the alignment. Only in this way can the results we are striving for –an error rate of under 0.5 %- be achieved. We proceed in three phases. First we select the segments longer than 350 characters, because they have to be split in order to be processed. To do this we manually insert break marks <br> in suitable places in the text of both segments. Later the segments are automatically split where the breaks were placed. In a second step we locate empty alignments, i.e. the unpaired segments in the source or target text. In this case it can be a misalignment or can correspond to deletions or insertions in the translated text. If the segment is misaligned we do any necessary corrections. If the segment was not translated we insert into the empty cell the mark [n_t_s] (=non translated segment). If the segment was added in the translation text the mark [a_s_t] (=added text in translation) is inserted. Finally in order to minimize the amount of manual checking we focus on bisegments where because of the length correlations between source and target segment it is more likely that mistakes might occur. To identify them we calculate the quotient of the sum of characters of the bisegment and the difference of characters between source and target segment. Then we apply this ratio to order the bisegments. Mistakes tend to occur in the

[§] LF Aligner is a freeware, open source, multiplatform alignment tool written by Farkas András. We thank him for the support received.

bisegments where the value range of the ratio is -5 - 5. Manual checking of alignment results can be done in this way more efficiently, the process is less labor intensive and less time-consuming. This procedure is a compromise between what would be desirable and what is feasible, and it is able to assure a high level of accuracy.

At the current stage of the corpus we have 506,316 segments automatically aligned, of which some 150,000 bisegments have been manually reviewed. The manually validated texts are added to the training corpus and the aligner tool is regularly retrained to improve accuracy. Figure 1 shows the provisional user interface



Ergebnisse: 126 Seiten: 5 Aktuelle: 1

warum die Wahl der Kindlichen Kaiserin auf dich gefallen ist, Atréju.« Der Junge senkte ein wenig die Stirn, dann ging er rasch **hinaus**. Draußen vor dem Zelt stand Artax, sein Pferd. [Quelle]	-Creo que empiezo a comprender por qué te eligió la Emperatriz Infantil, Atreyu. El muchacho bajó un poco la cabeza y luego **salió** con rapidez. Fuera, delante de la tienda, estaba Ártax, su caballo. Era moteado y pequeño como un caballo salvaje, [Quelle]
Er nahm einen kopfgroßen Stein, der in seiner Reichweite lag, und schleuderte ihn so weit **hinaus**, wie er konnte. Der Stein fiel und fiel und fiel, bis ihn die Dunkelheit verschlang. [Quelle]	Cogió una piedra del tamaño de su cabeza que había cerca y **la** lanzó tan lejos como pudo. La piedra cayó, cayó y cayó hasta que se la tragó la oscuridad. [Quelle]
kroch er weiter, erreichte schließlich den Ausgang sino queund streckte vorsichtig den Kopf **hinaus**. Und nun sah er: Über der Finsternis des Tiefen Abgrundes, von einem Rand zum anderen gespannt, hing ein ungeheures Spinnennetz. [Quelle]	continuó arrastrándose, llegó por fin a la salida sino quey asomó con precaución **la** cabeza. Y entonces vio esto: sobre las tinieblas del Abismo Profundo, de un borde a otro colgaba una monstruosa tela de araña. [Quelle]
Als er auf die Klinke drückte, öffnete sich die Tür mit lautem Knarren. Er huschte strumpfsockig **hinaus** und ließ die Tür hinter sich offen, um nicht noch einmal unnötigen Lärm zu machen. [Quelle]	Cuando hizo presión sobre el picaporte, la puerta se abrió con un fuerte chirrido. Se deslizó en calcetines dejando detrás **la** puerta abierta, para no tener que hacer otra vez ruidos innecesarios. [Quelle]
Er brummte etwas vor sich hin, doch als er merkte, daß das Wasser zu laufen aufhörte, zuckte er die Achseln und ging **hinaus**. Seine Schritte verhallten auf der Treppe. Bastian hatte die ganze	Refunfuñó algo entre dientes pero, cuando vio que el agua dejaba de correr, se encogió de hombros y **salió**. Sus pasos se perdieron en la escalera. Bastián no se había atrevido a respirar

Figure 1

6 Future work

This is an on-going project and we are now continuing, on the one hand, to add new works to the corpus to arrive at the amount and balance mentioned. On the other hand, we are implementing other features such as PoS tag-sets and word alignment. We are developing a common PoS tag set on the basis of the sets in Treetagger for German and Freeling for Spanish. To increase the quality of the linguistic annotations, about 10% of the processing will be manually verified. Regarding the word alignment, several tools (Giza++, Nattools, OPUS WordAlign) are being tested but a final decision has not yet been taken.

A big challenge lies in accessing the corpus texts and its linguistic information in the best possible way for use by various user groups. Figure 1, above, shows the provisional web interface we are working on. It presents a prototype implementation of a corpus management system incorporated into the general purpose XML-aware search engine Solr. With Solr, it is possible to carry out single and multi-word searches, lemmatized searches and to sort the results in a desirable way. The results are by default contextualised: in addition to the sentence containing the words searched for, the preceding and the following sentences are also displayed, but the context can be expanded by 5 sentences before and after the searched word. The Solr search engine has, however, limitations that sometimes prevent it from meeting the most demanding requirements of cross-linguistic research. Such detailed requests can only be processed by a specific query system that allows for queries over linguistic annotation levels such as PoS tags or morphology labels. This is the reason why we are exploring the use of Corpus Workbench, one of the best known open-source tools for managing and querying corpora with linguistic annotations, developed at the IMS of the University of Stuttgart.

7 Conclusions

This paper describes and discusses the different steps we have completed up until the time of writing in the construction of a bilingual parallel corpus created primarily for linguistic research purposes. Despite the availability of other existing parallel corpora, the corpus PaGeS presents a number of distinguishing features:

1) Balanced composition: The texts are not restricted to a particular administrative, legislative or commercial domain and the PaGeS covers text types (fiction and non-fiction) that represent the current usage of language.

2) Quality of the source texts and translations is assured by using solely texts already published by renowned publishers

3) Quality control of the process: a quality control system for each step in compiling and aligning the corpus is performed. The corpus is being checked manually at different levels, including compilation, pre-processing, sentence splitting, and alignment.

4) Level of annotation: the PaGeS corpus is aligned and lemmatized; it will also be tagged at the part-of-speech level.

5) Availability. We plan for the corpus to be made available through a user-friendly web interface. Due to copyright restrictions, the corpus can only be used for research by scholars and students of research institutions.

All these special features make PaGeS very suitable not only for cross-linguistic research but also for translation studies and language learning and teaching.

Acknowledgments

The PaGeS-Corpus project is carried out within the research project SpatiAlEs (FFI2013-42571-P), which is funded by the Spanish Ministry of Economy and Competitiveness.

References

Carlisle, D. (2010, April). *graphicx: Enhanced support for graphics.* Retrieved from http://www.ctan.org/tex-archive/ help/Catalogue/entries/graphicx.html

Voronkov, A. (2004). *EasyChair conference system*. Retrieved from easychair.org

Braune, F., Fraser, A. (2010). Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora. In Coling 2010: 81–89. Beijing, China:

Brown, P.F., Pietra, S. A. D., Della Pietra, V. J. D., Mercer, R.L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics* 19/2: 263–311.

Gale, W., Church, K. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19/1: 75–102.

Kay, M., Röscheisen, M. (1993). Text-Translation Alignment. *Computational Linguistics* 19/1: 121–142.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit*, pp. 79–86. [http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf]

Mcenery, A., Zhonghua, X. (2007). *Parallel and comparable corpora: What are they up to?* [http://eprints.lancs.ac.uk/59/1/corpora_and_translation.pdf]

Padró, Lluís (2011). Analizadores Multilingües en FreeLing. *Linguamatica* 3/ 2: 13—20.

Parra Escartín, C. (2012). *Design and compilation of a specialized Spanish-German parallel corpus,* LREC 5: 2199-2206.

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

Tiedemann, J. (2011). Bitext Alignment. Toronto: Morgan & Claypool.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012) ELRA 2012: 2214-2218.

Varga, D. et al. (2005). Parallel corpora for medium density languages. In Proceedings of the RANLP 2005: 590–596.