Axolotl: a Web Accessible Parallel Corpus for Spanish-Nahuatl

Ximena Gutierrez-Vasques, Gerardo Sierra, Isaac Hernandez
GIL IINGEN
Instituto de Ingeniería, UNAM, Mexico City, Mexico
xim@unam.mx, {gsierram,ihernandezpo}@iingen.unam.mx

Abstract

This paper describes the project called Axolotl which comprises a Spanish-Nahuatl parallel corpus and its search interface. Spanish and Nahuatl are distant languages spoken in the same country. Due to the scarcity of digital resources, we describe the several problems that arose when compiling this corpus: most of our sources were non-digital books, we faced errors when digitizing the sources and there were difficulties in the sentence alignment process, just to mention some. The documents of the parallel corpus are not homogeneous, they were extracted from different sources, there is dialectal, diachronical, and orthographical variation. Additionally, we present a web search interface that allows to make queries through the whole parallel corpus, the system is capable to retrieve the parallel fragments that contain a word or phrase searched by a user in any of the languages. To our knowledge, this is the first Spanish-Nahuatl public available digital parallel corpus. We think that this resource can be useful to develop language technologies and linguistic studies for this language pair.

Keywords: Parallel corpus, Low-resource languages, Search tool

1. Introduction

Parallel corpora are a rich linguistic resource which comprises bodies of text in parallel translation, i.e., a set of texts in different languages which are translations from each other. This type of corpus is one of the most valuable resources in the development of several Natural Language Processing (NLP) applications. For instance, parallel corpora provides indispensable training data for the statistical machine translation systems (Brown et al., 1993) and it is also useful for some other applications like multilingual text retrieval and automatic bilingual lexical acquisition (Widdows et al., 2002; Guinovart, 2012).

Parallel texts are also a useful resource for aiding human translators. For instance, computer assisted translation tools or parallel corpus search interfaces, allow users to search for words and expressions in bilingual texts and see how other people translated an expression and in which context a certain translation is used (Volk et al., 2014). Additionally, parallel corpora can be used in the linguistics field, since it is helpful for performing contrastive and translation studies (Johansson, 2007).

The most common sources for gathering large amounts of parallel data include specialized domain texts such as parliamentary proceedings, religious texts and software manuals. Additionally, the World Wide Web represents a good source for finding large-size and balanced parallel text (Resnik and Smith, 2003). On the Web it is possible to find bilingual or multilingual websites which can be useful to extract readily available parallel text in several domains and language pairs. There are several examples of websites that offer their contents in several languages, for instance: international institutions, universities, touristic services, etc.

However, only for few language pairs there exist large amounts of readily available parallel data. We face a scarcity problem specially when one or both of the languages are low-resourced, i.e., the cases in which a language have a small amount of digital documents due to a small density of speakers or due to technological and other reasons.

In this work, we focus on the language pair Spanish-Nahuatl. These two languages are spoken in the same country (Mexico) but they are distant from each other, they belong to different linguistic families: Indo-European and Uto-Aztecan. Nahuatl is an indigenous language with around 1.5M speakers and it is a language with scarcity of monolingual and parallel corpora. Spanish has a much more bigger amount of speakers and available digital resources, however, we are dealing with a low-resource setting since there are not large amounts of Spanish-Nahuatl parallel texts and it is difficult to obtain them using the traditional sources.

We present a Spanish-Nahuatl parallel corpus that was gathered mainly from non-digital sources, our sources were books from a variety of domains. In addition, we implemented a search interface that could offer to the user an efficient and useful way to exploit the gathered parallel corpora. To our knowledge, it did not exist a digital publicly available parallel corpus for this language pair.

The structure of the paper is as follows: Section 2 contains a description of the parallel corpus and the compilation process of the parallel documents. In section 3, we describe the search interface that allows to make queries in the corpus. Section 4 contains a brief overview of the applications of our work. Finally, section 5 contains the conclusions and a discussion of the future work.

2. The Spanish-Nahuatl parallel corpus

2.1. Compilation of the documents

As we have mentioned before, Spanish and Nahuatl are distant languages, i.e., they do not share many orthographic, morphological or syntactic similarities. Table 1 contains examples of Nahuatl-Spanish parallel sentences that illustrate some differences between the languages. Both languages have rich morphology but Nahuatl is a highly agglutinative language while Spanish is a fusional language. We can see that one single Nahuatl word can correspond to several words in the other language. In these examples the syntactic order is not always the same in both languages. Furthermore, the last Nahuatl sentence is written using a different orthography compared to the previous Nahuatl sentences.

tinechcaquiznequi (Nahuatl) me quieres oir (Spanish) *you want to hear me*
In cihuamizton ipan ahcopechtli ca.(Nahuatl) La gata estaba encima de la mesa. (Spanish) *The (female) cat is on the table*
pejke san motlajtlachiliyaj (Nahuatl) empezaron a mirarse nada mas (Spanish) *they started to just look at each others*

Table 1: Examples of Nahuatl-Spanish parallel sentences

When we were gathering sources of parallel texts for the Spanish-Nahuatl language pair, it was not easy to obtain parallel content from the typical web sources. Nahuatl does not have a web presence or text production comparable to Spanish, it is not possible to find many multilingual websites as in the case of several indo-european languages. Despite the fact that Nahuatl is the second most spoken native language in Mexico, governmental, touristic and other websites do not offer their content in this language.

It is worth mentioning that there exist some online Nahuatl resources like the Nahuatl Wikipedia. However, we discarded this resource since its texts do not constitute a parallel corpus (the articles are no exact translations), Wikipedia is more properly a comparable corpus. On the other hand, many of the contributors are not Nahuatl native speakers and the orthography can significantly change from article to article. We were not sure what would be the impact of this kind of texts in our parallel corpus, so we did not take them into account in this first compilation of parallel texts.

Most of our parallel texts come from non digital books, we searched for books with parallel content in several libraries and then we digitized them. As we have mentioned before, when we work with low-resource languages, traditional sources and methods does not work exactly the same, in our case, we faced difficulties when digitizing the texts. We used an Optical Character recognition (OCR)[1] software but it made several mistakes in the task of automatically recognizing Nahuatl text.

We identified that these mistakes were mainly associated with the fact that the OCR could not properly identify the language Nahuatl. Since the software had a lack of experience in processing Nahuatl, it tried to adapt character patterns corresponding to other languages and often make false corrections. Some other mistakes were related to phonological marks and typography that were difficult to recognize by the software and also to the fact of having more than one language mixed in the same page.

We had to perform a manual correction of the texts after getting them recognized by the software. The parallel documents belong to different domains, e.g., history, literature, didactic material, short stories, recipes. Beside the different domains, the documents of the parallel corpus are not quite homogeneous in the sense that there is dialectal, diachronic and even orthographical variation. Nahuatl is a language that has many dialects, moreover, nowadays it does not exist a general agreement regarding to the appropriate way to write the language.

So far, we have digitized and corrected 31 books, we have also added parallel texts that were already in a digital format (texts found on the web or given to us by collaborators). In total, we have 35 different sources of parallel texts. The current total size of the parallel corpus is around 1,186,662 tokens, i.e, taking into account the documents of both languages combined.

2.2. Corpus information

The figure 1 shows a rough classification of the document genres in the parallel corpus. In addition, we made a very general classification of the dialect in which the Nahuatl texts are written (Table 2).

The older texts of our corpus are written in Classical Nahuatl, i.e., the dialect in which Nahuatl texts were written when a latin alphabet writing system was first adapted to the language. Classical Nahuatl was mainly used for religious, chronicles, and legal texts around 16th and 17th centuries. In the modern Nahuatl classification, we included the several modern linguistic variants found in our corpus.

We only show the dialect distribution for Nahuatl, since in this sense Spanish texts are more or less uniform across the parallel corpus.

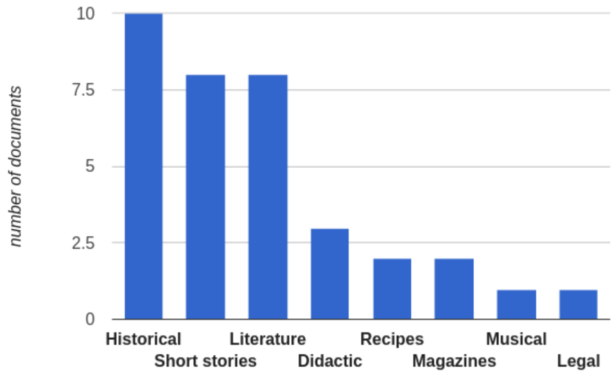

Figure 1: General genre classification of the parallel documents

Nahuatl dialect	Percentage of documents
Classical	45.7%
Modern	54.3%

Table 2: General distribution of Nahuatl dialects in the parallel corpus

One important aspect to allow the explotation of the bilingual lexical information contained in a parallel corpus is the alignment. Alignment is the process of pairing bilingual correspondences at an specific level, i.e., document (Braschler and Scäuble, 1998), paragraph (Gelbukh and Sidorov, 2006), sentence (Brown et al., 1991; Gale and

[1] http://www.abbyy.com/

Church, 1993) or word level (Brown et al., 1993). We performed alignment at the sentence level. We used several methods depending on the type of document because it is not always a straightforward task when we deal with distant languages and with very different sources. In some texts, it was possible to use traditional statistical methods (Brown et al., 1991; Gale and Church, 1993) that are based on sentence length. For some type of documents, we performed semi-automatic alignment, i.e., we took advantage of several markers contained in the documents indicating the corresponding translations. And in some cases we had to perform manual alignment.

It is important to mention that due to the variety of texts and to the different techniques used for the alignment, the obtained alignments sometimes correspond to larger units, like paragraphs.

3. The search interface

Once we built the parallel corpus, we were interested in developing an application that could make this resource easily available. We designed a web querying interface, named Axolotl, that allows to search within the corpus for words or phrases in both languages.

The idea of the system is to perform queries in the parallel texts and to display those parallel fragments that contain the searched word in one of the languages. In our web interface, users are able of making queries in Spanish or Nahuatl and to retrieve parallel fragments of our parallel corpus containing the results. It is important to mention, that this type of search is possible since the parallel corpus is aligned at sentence level.

Our web system is similar to other online parallel corpus query systems like Linguee[2] and the OPUS Corpus Query (Tiedemann, 2012), just to mention some.

The search results displayed by the Axolotl system does not only contain textual fragments but also information about the source of these fragments and a preview of the PDF document from where the text was extracted.

Axolotl's search engine provides flexibility in the queries. Beside the standard search of a word or a phrase, the users can make more complex queries in order to obtain more accurate results. For instance, they can use binary operators (AND, OR) and they can use proximity matching in order to find words that are within a specific distance away, just to mention some.

3.1. Basic architecture of the system

The basic architecture of the system comprises several information technologies. We tried to design the system in a way that could manage large amounts of parallel text, even though our corpus is relative small. The aligned parallel content is stored in a MongoDB[3] database that allows to write and read the data in a fast way. We stored the documents in the database in a way that it was indicated the correspondence between parallel sentences.

In order to be able to search and retrieve efficiently the user's queries, a web search engine must be used. In our case, we implemented it by using Lucene/Solr[4] which are widely known information retrieval tools. Using this engine we indexed the parallel texts previously stored in the database. This indexing allows to perform queries in the corpus. The search engine is an essential component, it allows to retrieve the parallel sentences ordered by relevance according to the user's query. It also allows to use operators to make complex queries, among other functionalities.

Regarding to the web interface, Solr provides APIs for several programming languages, we chose Ruby on Rails[5] for the web implementation.

Figure 2 illustrates the basic architecture of our system. The system is based in a model–view–controller (MVC) architecture, where a controller recieves the parameters that the user sends from the view. After the petition is processed, the controller sends back the results to the view, so the user can visualize them in the web interface.

The diagram shows how the user, through the search view in the web interface, selects the language and provides a word or phrase to search. These parameters are processed by the Rails controller and sent through the method *#search*. This method gives to Solr the parameters in order to perform the query.

Solr search engine retrieves from the database the content that fulfills the search parameters. Once the search results are obtained, they are stored in a variable. This variable stores only a first set results, which are sent to the view (show view) where the user can choose to see the next set of results (pagination).

3.2. Resource availability

The parallel corpus search interface Axolotl is freely available on the Web[6] . Our aim is that this web service become part of the LRE MAp and the ”Sharing LRs” initiative.

Regarding to the copyrights of the texts, we consulted the national law valid in Mexico. Not the same legal constraints apply to all the texts. For instance, many ancient Nahuatl texts are considered as public domain but in some cases, the diffusion of the Spanish associated translation can be protected if it is a recent translation. In sum, these legal constraints prevent in many cases the diffusion of an entire digitized book but they allow to show fragments, specially if it is for educational or research purposes.

Our current version of Axolotl allows to make searches through the whole parallel corpus, but, due to these restrictions, it does not allow to download entire books and their scanned version, instead it shows a preview of the scanned book.

4. Applications

Regarding to the application of our work, the Spanish-Nahuatl parallel corpus and its querying interface, could encourage the generation of several language technologies for this language pair. We believe that this is especially important in a country with a vast linguistic diversity but very few, or none, language technologies developed for the regional indigenous languages.

[2] http://www.linguee.com/

[3] https://www.mongodb.org/

[4] https://lucene.apache.org/

[5] http://rubyonrails.org/

[6] http://www.corpus.unam.mx/axolotl

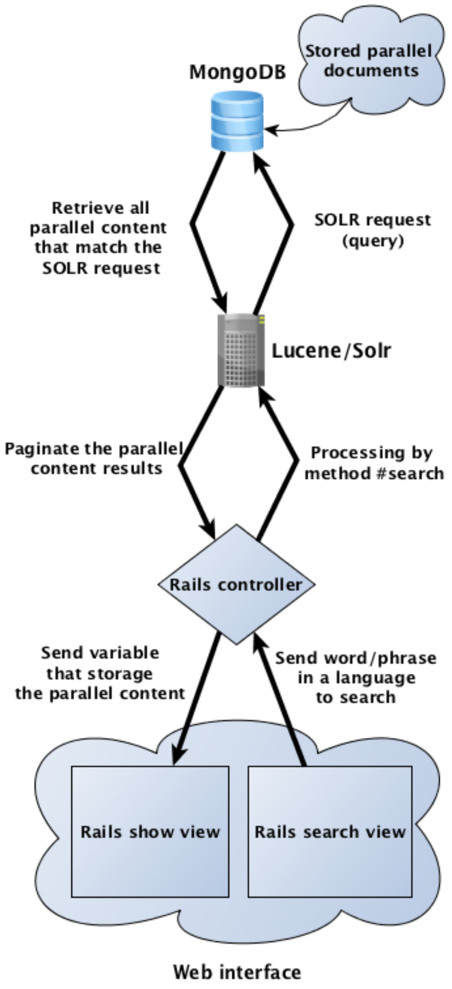
Stored parallel documents
MongoDB
Retrieve all parallel content that match the SOLR request
SOLR request (query)
Lucene/Solr
Paginate the parallel content results
Processing by method #search
Rails controller
Send variable that storage the parallel content
Send word/phrase in a language to search
Rails show view
Rails search view
Web interface

Figure 2: Basic architecture of Axolotl's parallel corpus search interface

This parallel corpus is being currently used in a automatic bilingual lexicon extraction task (Gutierrez-Vasques, 2015).

Although our parallel corpus is small compared to the sizes usually required by statistical machine translation systems, it could represent a first step in the development of automatic translation technology which is currently not available for this language pair.

In general, the parallel corpus could be an useful resource for the development of multilingual language technologies and also for performing more linguistic studies such as contrastive and historical linguistics and also translation studies.

5. Conclusions and future work

In this work, we have presented a parallel corpus for the language pair Spanish-Nahuatl. We described the parallel corpus compilation process which comprised digitizing most of the texts, manual corrections and a sentence-level alignment.

Since the parallel texts come from very different sources there is dialectal and diachronic variation. Furthermore, there is a lack of orthographic normalization. Despite that the lack of an orthographic norm and other types of variation can represent "noise" for many of the NLP statistical methods, we decided to keep all the parallel documents due to the data scarcity.

We developed a public available web interface, Axolotl, that allows to make queries in the whole parallel corpus. The system shows the parallel sentences that contain the word or phrase searched by a user. Additionally, the system shows from where the text was taken and a preview of the original document.

Nahuatl is an indigenous low-resource language, this parallel corpus could encourage the creation of several language technologies and linguistic studies for the Spanish-Nahuatl language pair.

We realize that the different types of variation in the parallel texts, imply many challenges when building a parallel corpus and its querying interface. As a future work, it may be necessary to use richer annotation schemes to specify the dialect in which a text is written and also to perform some text normalization. This could be helpful to retrieve more accurate results when a search is performed and also it could provide more information, for instance, in the cases in which a user is interested in performing some linguistic analysis like dialectology.

Currently, we keep adding some digital and non-digital sources to the parallel corpus, in some cases these texts have to pass through a manual review. In the future, we would like to allow users to contribute with the parallel corpus by uploading parallel texts through the web interface.

6. Acknowledgements

This work is supported by the Mexican Council of Science and Technology (CONACYT), funds 370713 and CB-2012/178248. Thanks to the reviewers for their valuable comments

7. Bibliographical References

Braschler, M. and Scäuble, P. (1998). Multilingual information retrieval based on document alignment techniques. In *Research and Advanced Technology for Digital Libraries*, pages 183–197. Springer.

Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 169–176. Association for Computational Linguistics.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

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

Gelbukh, A. and Sidorov, G. (2006). Alignment of paragraphs in bilingual texts using bilingual dictionaries and dynamic programming. In *Progress in Pattern Recognition, Image Analysis and Applications*, pages 824–833. Springer.

Guinovart, F. J. G. (2012). A hybrid corpus-based approach to bilingual terminology extraction. In *Encoding* *the past, decoding the future: corpora in the 21st Century*, pages 147–175. Cambridge University Press.

Gutierrez-Vasques, X. (2015). Bilingual lexicon extraction for a distant language pair using a small parallel corpus. *NAACL-HLT 2015 Student Research Workshop (SRW)*, page 154.

Johansson, S. (2007). *Seeing through Multilingual Corpora: On the use of corpora in contrastive studies*, volume 26. John Benjamins Publishing.

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *LREC*, pages 2214–2218.

Volk, M., Graën, J., and Callegaro, E. (2014). Innovations in parallel corpus search tools. In *LREC*, pages 3172–3178.

Widdows, D., Dorow, B., and Chan, C.-K. (2002). Using parallel corpora to enrich multilingual lexical resources. In *LREC*, pages 240–245. Citeseer.