# Wikipedia as Multilingual Source of Comparable Corpora

**Pablo Gamallo Otero, Isaac González López**

University of Santiago de Compostela
Galiza, Spain
pablo.gamallo@usc.es, isaacjgonzalez@gmail.com

**Abstract**

This article describes an automatic method to build comparable corpora from Wikipedia using *Categories* as topic restrictions. Our strategy relies of the fact Wikipedia is a multilingual encyclopedia containing semi-structured information. Given two languages and a particular topic, our strategy builds a corpus with texts in the two selected languages, whose content is focused on the selected topic. Tools and corpora will be distributed under free linceses (General Public License and Creative Commons).

## 1. Introduction

Wikipedia is a free, multilingual, and collaborative encyclopedia containing entries (called "articles") for more than 300 languages. English is the more representative one with almost 3 million articles. As table 1 shows, the number of entries/articles for the most used languages in Wikipedia is so high that it could be considered a reliable multilingual resource. However, Wikipedia is not a parallel corpus as their articles are not translations from one language into another. Rather, Wikipedia articles in different languages are independently created by different users.

In accordance with fast growth of Wikipedia, many works have been published in the last years focused on its use and exploitation for multilingual tasks in natural language processing: extraction of bilingual dictionaries (Yu and Tsujii, 2009; Tyers and Pieanaar, 2008), alignment and machine translation (Adafre and de Rijke, 2006; Toms et al., 2001), multilingual retrieval information (Pottast et al., 2008). In addition, there exists theoretical work on the degree of comparability among the different multilingual versions of an entry/article in Wikipedia (Filatova, 2009). In particular, the author analyzes symmetries and asymmetries in multiple descriptions of multilingual entries.

In this paper, our main concern is the use of Wikipedia as a source of comparable corpora. The EAGLES - Expert Advisory Group on Language Engineering Standards Guidelines (see `\http://www.ilc.pi.cnr.it/EAGLES96/browse.html`) gives us the following definition for "comparable corpora":

> *A comparable corpus is one which selects similar texts in more than one language or variety.*

One of the main advantages of comparable corpora is their versatility to be used in many linguistic fields (Maia, 2003), like terminology extraction, Information Retrieval, and Knowledge Engineering. In addition, they can also be used as training corpus to improve statistic machine learning systems, in particular when parallel corpora are scarce for a given pair of languages. Another advantage concerns their availability. In contrast with parallel corpora, which require (not always available) translated texts, comparable corpora are easily retrieved from the web. It is much easier to find original texts on a particular subject than to find a pair consisting of the original and a good translation. Among the different web sources of comparable corpora, Wikipedia is likely the largest repository of similar texts in many languages. We only require the apropriate computational tools to make them comparable.

| Languages | number of articles |
|---|---|
| English | 2,826,000 |
| German | 888,000 |
| French | 786,000 |
| Polish | 593,000 |
| Italian | 576,000 |
| Japonese | 556,000 |
| Dutch | 528,000 |
| Portuguese | 470,000 |
| Spanish | 460,000 |
| Rusian | 376,000 |

Table 1: The top ten languages in Wikipedia ranked by number of articles (April 2009)

By taking into account multilingual potentialities of Wikipedia, the goal (and main contribution) of this paper is to describe a method to extract comparable corpora from this freely available encyclopedia, according with two parameters of variation: languages and topic. More precisely, given two languages and a particular topic, our strategy builds a corpus with texts in the selected languages, whose content is focused on the selected topic. Both the generated corpora and the tools used to generate them will be available under Creative Commons license in `http://gramatica.usc.es/pln`. Experiments will be performed with articles in English, Spanish, and Portuguese. As Table 1 shows, Spanish is the ninth most used language in Wikipedia, with 460 thousand articles, very close to Portuguese, which reaches 470 thousands.

This paper is organized as follows. Section (2.) describes how we convert the original Wikipedia into a new codified corpus, called "CorpusPedia". Section (3.) introduces different strategies to build comparable corpora from CorpusPedia. In Section (4.), we give some empirical data of CorpusPedia, as well as the results of some experiments per-

This work has been supported by the Galician Government, within the projects PGIDIT07PXIB204015PR and 2008/101.



*Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010*, pages 21–25
Malta, 22 May 2010

```
<page>
<title>Arqueoloxía</title>
<id>3</id>
<revision>
  <id>1310468</id>
  <timestamp>2009-10-06T02:42:14Z</timestamp>
  <contributor>
    <username>SieBot</username>
    <id>2109</id>
  </contributor>
  <minor />
  <comment>bot Engadido: [[ku:Arkeolojî]]</comment>
  <text xml:space="preserve">{{Historia en progreso}}

A '''arqueoloxía''' é a [[ciencia]] que estuda as [[arte|artes]],
[[monumento|monumentos]] e [[obxecto]]s da
[[antigüidade|antigüidade]], especialmente a través dos
seus restos. O nome ven do [[lingua grega|grego]]
"archaios", &quot;vello&quot; ou &quot;antigo&quot;, e
"logos", &quot;ciencia&quot;, &quot;saber&quot;.

[...]

[[zh: 考古学]]
[[zh-yue: 考古]]</text>
  </revision>
</page>
```

Figure 1: XML example of Wikipedia: excerpt of Galician entry "Arqueoloxía" (Archaeology)

formed using the strategies defined in (3.). The last section discusses future tasks we intend to implement in order to extend and improve our tools.

## 2. CorpusPedia

The first step of our method consists in converting the source files of Wikipedia to a set of files with a more friendly and easy-to-use XML structure: CorpusPedia. For this purpose, we developed tools aimed to automatically download Wikipedia in the required languages and then to apply the process of transforming the downloaded XML file into the new XML files of CorpusPedia. In the following, we will compare the structure of those two formats.

### 2.1. Format of Wikipedia

The whole Wikipedia is downloadable in XML files containing a great variety of metadata. Figure 1 shows an example of an article codified in this way. An entry/article is identified by the tag *page*, which contains a title, a data, an author, and the text of the article. The difference with regard to the most usual web markup languages (html or xhtml) is that the text of all articles is codified in *wiki format*, as the the tag *text* in Figure 1 illustrates. One of the main taks of CorpusPedia is to build a plain text version from that wiki text.

### 2.2. Format of CorpusPedia

The format of CorpusPedia also consists, essentially, in both the title and the text of each entry/article. Besides, further information is also provided using semi-structured data of Wikipedia and some conventions among editors (Clark et al., 2009). Figure 2 depicts the XML code used to markup the new format generated from Wikipedia. Tags *title*, *category*, *plaintext*, and *translations* are those required to generate comparable corpora.

The tag *category* is used to identify all the topics classifying the text content. In Wikipedia, each article is explicitly assigned to one or more categories representing different topics. This tag will allow us to extract those articles classified with similar categories and, then, with high degree of comparability. The tag *wikitext* contains the original format of Wikipedia. We keep this format since it can be useful for further extractions. based on semi-structured content. The tag *plaintext* (i.e., text without any codification) is generated from *wikitext* by applying a *wiki2plaintext* parser we developed for this purpose. Unlike other *wiki2plaintext* converters, we took into account specific semantic features of Wikipedia. The tag *translations* codifies a list of interlanguage links (i.e., links to the same articles in other languages). As we will explain in the next section, these links are useful to align article-by-article a comparable corpus if it is required by the user. The list of interlanguage links is always ranked in the same way (gl pt es en fr ca eu al it cs bg el). Besides, if there is no a specific interlanguage link, the symbol "#" is used to explicitly mark that the translation is not available. Other languages can be easily added to the list if they are required.

The remaining tags of CorpusPedia provide further useful relations with other articles in Wikipedia. This way, the tag *related* adds those articles that are somehow related to the current one and which have been explicitly marked in Wikipedia. Finally, *links* introduce the set of links to other

<article>
<title>Arqueoloxía</title>
<category>Arqueoloxía</category>
<related>Antropoloxía, Arqueoloxía industrial, Arqueoloxía submarina</related>
<links>ciencia, arte|artes, monumento|monumentos, obxecto, antigüidade|antigüidade, lingua grega|grego, cultura, estudo, psicolóxico, condutistas, antropoloxía, idade de pedra, Idade Media, Arqueoloxía industrial, Antropoloxía, Arqueoloxía industrial, Arqueoloxía submarina</links>
<translations># Arqueologia Arqueología Archaeology Archéologie Arqueologia Arkeologia # Archeologia Archeologie Археология Αρχαιολογία</translations>
<plaintext>A arqueoloxía é a ciencia que estuda as artes, monumentos e obxectos da antigüidade, [...] o que se coñece como Arqueoloxía industrial.</plaintext>
<wikitext>{{Historia en progreso}}

A '''arqueoloxía''' é a [[ciencia]] que estuda as [[arte|artes]], [[monumento|monumentos]] e [[obxecto]]s da [[antigüidade|antigüidade]], [...]
[...]
[[yi:ארכעאלאגיע]]
[[zh: 考古学]]
[[zh-yue: 考古]]</wikitext>
</article>

Figure 2: XML example of CorpusPedia: excerpt of Galician entry "Arqueoloxía" (Archaeology)

articles that were explicitly mentioned within the text (also called "interlinks").

## 3. Strategies to Elaborate Wikipedia-Based Comparable Corpora

Given the information structure of CorpusPedia, it is possible, not only to easily collect articles about the same topic in the same language, but also to put them in relation with articles about the same topic in other languages. It means the structure of CorpusPedia enables to easily build comparable corpora. For this purpose, we developed three tools aimed to extract corpora with different degrees of comparability. These tools, which correspond to three strategies, are described in the following subsections.

### 3.1. Not-Aligned Comparable Corpora

This strategy extracts those articles in two languages having in common the same topic, where the topic is represented by a category and its translation (for instance, the english-spanish pair "Archaeology-Arqueología"). The algorithm used to extract not-aligned comparable corpora from CorpusPedia is the following:

```
Given two languages, L1 and L2, and
two bilingual categories, C1 and C2,
where C2 is the translation of C1 in
L2:
    (1) extract those articles in L1
    containing C1 within the section
    <category> ;
    (2) Repeat the same process in L2,
    using C2.
```

It results in a not-aligned comparable corpora, consisting of texts in two languages ($L1$ and $L2$) sharing the same topic: $C1$-$C2$. We called it "not-aligned" because the version of an article in one language may have not its corresponding version in the other language. In technical terms, it means articles extracted from L1 will contain both empty and not empty interlanguage links to articles in L2.

### 3.2. Strong Alignment

The corpus resulting of the previous process can be considered as being too heterogeneous, since it may contain articles in one language that have not their corresponding versions in the other one. For instance, we can find an English article with the title "Australian archaeology" that has not any interlanguage link in Spanish, i.e., that has not a Spanish version with the title "Arqueología australiana" in the Wikipedia. To build an aligned corpus at the level of articles, we define a strategy to extract only those articles that have interlanguage links to the target language. The algorithm of this strategy is the following:

```
Given two languages, L1 and L2, and
two bilingual categories, C1 and C2,
where C2 is the translation of C1 in
L2:
    (1) extract those articles in L1
    with the following properties:
        - C1 is within the section
        <category>
        - there is a interlanguage
        link to an article in L2
        containing C2 in the section
        <category>
```

```
    (2) Repeat the same process from
    L2 and remove inconsistencies.
```

We obtain a comparable corpus constituted by the same articles in both languages. The strategy used to align article by article is very restrictive and then has very low coverage. In fact, no only each article in one language must have its corresponding article in the other one, but also both articles must share the same categorial restriction. Let's note that we have to automatically remove inconsistencies in annotations, such as for instance ill-defined interlangue links. These annotations problems were inherited from the source file.

### 3.3. Soft Alignment

The strong alignment algorithm is not able to extract some relevant articles, in particular those that, having interlanguage links to the target languange, do not fill the categorial restriction. For instance, there may be articles categorized in the English Wikipedia by means of the term "Archaeology"[1], which have not been categorized in the Spanish Wikipedia with the corresponding term "Arqueología". However, these Spanish articles can be considered as being indirectly classified by the English category. In fact, Spanish Wikipidia is less categorized as the English one (Spanish editors tend to use fewer categories by article). Similarly, the Portuguese Wikipedia is still less categorized as the Spanish one. This categorial asymmetry is responsible for the low coverage reached by the previous strategy (strict alignment). To solve this problem, we propose a less rigid alignment. The goal is to extract pairs of bilingual articles related by interlanguage links if, at least, one of both contains the required category. The algorithm is the following:

```
Given two languages, L1 and L2, and
two bilingual categories, C1 and C2,
where C2 is the translation of C1 in
L2:
    (1) extract those articles in L1
    with the following properties:
        - C1 is within the section
        <category>
        - there is an interlanguage
        link to an article in L2
    (2) extract those articles in L2
    with an interlanguage link to
    the articles in L1 which have
    been already extracted, and
    remove inconsistencies.
```

It results in a corpus that has also been alingned article by article, but using a technique not so restrictive as in the previous method.

## 4. Experiments and Results

### 4.1. Size of CorpusPedia

In the last version of CorpusPedia, the plaintext in English contains about 1,2 billion token words, 180 million in Spanish, and 120 million in Portuguese. Notice that the Spanish version contains more words than the Portuguese one. However, the Portuguese Wikipedia contains a larger number of articles, as is shown in Table 1. It follows the plaintext content of Portuguese articles tends to be smaller than that of the Spanish version.

[1]The structured list of categories in the English Wikipedia avoids language variation. In this case, the normalized term is the British Archaelogy instead of Archeology.

| strategy | size (in words) | number of articles |
|---|---|---|
| en/es not-aligned | 738,000 / 344,000 | 1120 / 462 |
| en/pt not-aligned | 738,000 / 64,000 | 1120 / 100 |
| es/pt not-aligned | 344,000 / 64,000 | 462 / 100 |
| en/es strong-align | 34,000 / 23,000 | 34 / 34 |
| en/pt strong-align | 29,000 / 11,000 | 16 / 16 |
| es/pt strong-align | 27,000 / 11,000 | 19 / 19 |
| en/es soft-align | 220,000 / 134,000 | 191 / 191 |
| en/pt soft-align | 161,000 / 60,000 | 124 / 124 |
| es/pt soft-align | 132,000 / 64,000 | 119 / 119 |

Table 2: Comparable corpora in english-spanish, english-portuguese, and spanish-portuguese. They were obtained using category "Archaelogy-Arqueología-Arqueologia" and three strategies.

### 4.2. Size of Comparable Corpora Generated with the Three Strategies

Taking CorpusPedia as input source, we performed several experiments to build comparable corpora (english-spanish, spanish-portuguese, and english-portuguese) containing texts on the same topic, namely Archaeology. We used the three strategies described in the previous section. The specific topic in both Spanish and Portuguese was selected with the corresponding translations of "Archaeology", that is: "Arqueología" in Spanish and "Arqueologia" in Portuguese. Table 2 summarizes the quantitative description of all generated corpora.

The table shows there are significant differences in size among the three language. As it was expected, the baseline strategy without alignment yields an English corpus with 730 thousand words in contrast to only 64 thousand in Portuguese. However the difference between Spanish and Portuguese (344 against 64 thousand) is less expected since Wikipedia contains more articles in Portuguese. Two reasons explains such a difference. First, the system found 420 Spanish articles sharing the category "Arqueología" against only 100 in Portuguese. This is in accordance with the fact that Portuguese articles tend to contain fewer categories than the Spanish ones. Second, the plaintext size of Spanish articles is larger than in Portuguese. This is easily confirmed by the results obtained using alignment techniques: given the same number of extracted articles (19 with strong alignment and 119 with soft alignment), the size of the Spanish corpus is about twice larger than in Portuguese. The same tendency is verified between English and Spanish. So, it follows the size of English articles is almost three times larger than that found in Portuguese. Yet, this is true only concerning aligned articles. Those English articles that were not aligned (i.e., without their corresponding versions in Spanish or Portuguese) are much smaller than those aligned with their Spanish and Portuguese versions. All those significant asymmetries should be taken

| English articles | Spanish Articles |
|---|---|
| Adena culture | Cultura Adena |
| Afanasevo culture | Cultura Afanasevo |
| Alalakh | Alalakh |
| Alexandria National Museum | Museo Nacional de Alejandría |
| Amazons | Amazona (mitología) |
| Ancient footprints of Acahualinca | Huellas de Acahualinca |
| Antiguo Oriente | Antiguo Oriente |
| Antikenmuseum Basel und Sammlung Ludwig | Museo de arte antiguo de Basilea y colección Ludwig |
| Apadana | Apadana |
| Archaeological Museum of Asturias | Museo Arqueológico de Asturias |
| Archaeological Museum of Granada | Museo Arqueológico y Etnológico de Granada |
| Archaeological Survey of India | Servicio arqueológico de la India |
| Archaeology | Arqueología |
| Archaeology of the Americas | Prehistoria de América |
| Archaic period in the Americas | Periodo arcaico de América |

Table 3: Sample of titles extracted from en/es soft-alignment.

into account to build, not only homogeneous corpora according to a specific topic, but also balanced resources with regard to corpus size.

Finally, Table 3 shows a sample of bilingual titles english-spanish representing some of the articles extracted using the soft alignment strategy. Lists of bilingual pairs as those depicted in 3 allow us to observe the degree of comparability between texts in both languages. A large-scale automatic evaluation of quantitative features will be the goal of further experiments.

## 5. Conclusions and Future Work

The emergence of multilingual resources, such a Wikipedia, make it possible to design new methods and strategies to compile corpus from the web, methods that are more efficient and powerful than the traditional ones. In particular, the semi-structured information underlying Wikipedia turns out to be very useful to build comparable corpora. On the one hand, editors classify articles with categories corresponding to topics or genders and, on the other, a network of interlanguage links enables to create bilingual relations between articles.

Our current research is focused on how to improve the strategies by extending coverage (more articles) without losing accuracy (the same topic). For this purpose, we are testing and evaluating two techniques to expand categories using a list of similar terms: those tagged as *related* in CorpusPedia and those identified as hyponyms or co-hyponyms of the source category. In order to find hyponyms and co-hyponyms of a term, it is required to make use of an ontology well suited to encyclopedic knowledge. One of our current tasks is to build an ontology of categories using the semi-structured information of Wikipedia (Chernov et al., 2006).

Finally, in future work, we will define an evaluation protocol to measure the degree of comparability between texts. For this purpose, we will make use of techniques described in (Saralegui and Alegria, 2007).

## 6. References


S.F. Adafre and M. de Rijke. 2006. Finding similar sentences acorss multiple languages in wikipedia. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69.

Sergey Chernov, Tereza Iofciu, Wolfgang Nejdl, and Xuan Zhou. 2006. Extracting semantic relationships between wikipedia categories. In *SemWiki2006 - From Wiki to Semantics*, Budva, Montenegro.

M. Clark, Ian Ruthven, and Patrick O'Brian Holt. 2009. The Evolution of Genre in Wikipedia. In *Proceedings of JLCL 2009*, volume 24, pages 1–22.

Elena Filatova. 2009. Directions for Exploiting Asymmetries in Multilingual Wikipedia. In *CLEAWS3*, pages 30–37, Colorado.

Belinda Maia. 2003. What Are Comparable Corpora. In *Workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives*, pages 27–34, Lancaster, UK.

M. Pottast, B. Stein, and M. Anderka. 2008. A wikipeda-based multilingual retrieval model. In *Advances in Information Retrieval*, pages 522–530.

X. Saralegui and I. Alegria. 2007. Similitud entre documentos multilnges de carcter cientfico-tcnico en un entorno Web. In *Procesamiento del Lenguaje Natural*, page 39.

J. Toms, J. Bataller, and F. Casacuberta. 2001. Mining Wikipedia as a Parallel and Comparable Corpus. In *Language Forum*, volume 1, page 34.

M.F. Tyers and J.A. Pieanaar. 2008. Extracting Bilingual Word Pairs from Wikipedia. In *LREC 2008, SALTMIL Workshop*, Marrakech, Marroco.

Kun Yu and Junichi Tsujii. 2009. Bilingual dictionary extraction from wikipedia. In *MT Summit XII*, Ottawa, Canada.