

Acquisition of a US-Spanish Newspaper Corpus

Petra Prochazkova* Peter Kolb†

October 9, 2006

Abstract

The aim of this paper is to present a simple method of gathering newspaper data in order to acquire a US-Spanish corpus.

1 Introduction

Corpora are important linguistic resources. Although there exist corpora of Iberian and American Spanish, there is no large corpus of *US-Spanish* as of now. The focus of our project is to collect the daily news on the web in order to build a **US-Spanish News Corpus** (USSN). The main goal of such a corpus is to provide a basis for different kinds of linguistic analyses. In particular to investigate the differences of Spanish used in the U.S. and in Spain, and the influence of US-English over US-Spanish.

2 Acquisition

2.1 Method

First we compiled a list of Spanish news sources in the U.S. For every newspaper we determined the permanent URLs of all existing topics (e.g. politics, economy, and science).

Currently, we start new crawls on a daily

basis. We extract all new links from every topic page that lead to news articles. By doing that, we can download topic-classified articles. The URLs of the downloaded pages are stored in a hash table. This has two reasons: firstly, it allows accessing the original article on the web later. Secondly, it prevents crawling one article twice and therefore guarantees the acquirement of exclusively new articles. Hence we don't have to perform complicated date extraction and we cause very low traffic at the site.

In our project we use the perl modules **LWP**, **HTML** and **URI** for collecting the newspaper data.

2.2 Pre-processing and Annotation

Before pre-processing starts, we have to identify the character encoding of every HTML page. This is easy when the language is known. If necessary, we convert the file to UTF-8 (Unicode).

For extracting the textual content from the downloaded HTML pages (so called *boilerplate stripping*) we developed a tool that counts the numbers of HTML tags and words per line. On account of that, it acquires merely lines consisting of more words than tags. This filters out menus, lists of links, and advertisements.

*Department of Romance Studies, Humboldt University of Berlin, petra.prochazkova@gmail.com

†Institute for Linguistics, University of Potsdam, kolb@ling.uni-potsdam.de

As a further step, we convert HTML to text, i.e., we filter out HTML tags and substitute entities with their Unicode equivalents. That means presently we do not retain the textual structure (headings, emphasizes, etc.).

Finally, the news articles are tagged and lemmatised with the Spanish version of the TreeTagger (*Schmid* 1994).

Every article is stored as a separate file. Source, topic, and date are combined into a unique file name.

3 Corpus Data

We found 22 usable online newspapers that are published for Spanish-speaking people in the USA:

Newspaper	news	tokens
<i>Univision</i>	113.4	122,391
<i>Aldiatx</i>	28.9	13,778
<i>El Diario La Prensa</i>	70	34,539
<i>Atlanta Latino</i>	5	2,958
<i>Diario San Diego</i>	12.7	5,107
<i>El Nuevo Herald</i>	257	135,983
<i>El Sentinel</i>	46.3	21,687
<i>El Tiempo Latino</i>	3.3	1,689
<i>El Vocero Hispano</i>	11.5	3,575
<i>Houston Chronicle</i>	155.9	55,541
<i>Hoy Internet</i>	25.7	13,295
<i>IBLNews</i>	26.1	10,053
<i>La Estrella</i>	47.3	20,688
<i>La Opinión</i>	82	22,425
<i>La Raza</i>	22.1	8,013
<i>MSN</i>	110	48,367
<i>Rumbo</i>	14.9	5,522
<i>Vida en el Valle</i>	2.7	329
<i>El Hispano News</i>	1.4	439
<i>El Mensajero</i>	1	84
<i>El Paso Times</i>	4.7	2,279
<i>El Argentino</i>	5	1,046

The table shows the amount of news articles and tokens averagely downloaded per day.

Among the 22 newspapers there are only six that provide their news in English as well. It is not clear whether the news are translated from one language to the other, or if they are completely new articles with no relation to the Spanish version.

It should be possible to classify the news as translations or independent versions using the method proposed in *Bernardini et al.* 2006.

3.1 Period of acquisition

We started our crawlers on 9/1/2006 on a daily basis. The corpus will grow every day. From the listed average of daily news, we expect to build a corpus of 367,000 news articles containing 193,000,000 tokens within one year.

4 In Use

The USSN Corpus is already used by students of Spanish and in the project *Latin-US* (*Knauer* 2006) in the Department of Romance Studies at the Humboldt-University of Berlin to investigate the characteristics of the US-Spanish language on different linguistic levels (vocabulary, phonetic, syntax), particularly in the public communications domain. We also used the USSN Corpus to compute distributional word similarities using the program DISCO that was developed by Peter Kolb at the University of Potsdam (*Kolb* 2003).

5 Corpus in the Web

In order to browse the collected news corpus, a web interface based on Corpus Workbench (*Christ* 1994) and *cqp* (*Christ, and Schulze*

1995) was developed by Petra Prochazkova (*Prochazkova* 2006) at the Humboldt-University of Berlin. To overcome copyright issues the free access to the corpus is restricted to show only a narrow context (5 words to the right and to the left) of the search keyword so the displayed results are not protected by copyright. This limits in no way the utility, since the full text can always be received from the original source site by a single mouse click. (Keep in mind: we store the URL of every downloaded page.)

6 References

References

- Bernardini, S., Baroni, M., and Evert, S. (2006): “A new approach to the study of translationese: Machine-learning the difference between original and translated text”. In: *Literary and Linguistic Computing*, **21**(3), 259–274.
- Christ, O. (1994): “A Modular and Flexible Architecture for an Integrated Corpus Query System”. In: *Proceedings of COMPLEX '94: 3rd Conference on Computational Lexicography and Text Research, Budapest*.
- Christ, O., and Schulze, B. M. (1995): “Ein flexibles und modulares Anfragesystem für Textcorpora”. In: *Tagungsbericht des Arbeitstreffen Lexikon und Text*.
- Knauer, G. (2006): “Latinus: Spanish in American Public Communication, Humboldt-university”.
- Kolb, P. (2003): “Distributionelle Semantik”. Magisterarbeit, Universität Potsdam.
- Prochazkova, P. (2006): “Latinus Web Interface, <http://rom99.sprachen.hu-berlin.de/latinus/korpora/login.php>”.
- Schmid, H. (1994): “Probabilistic Part-of-speech Tagging Using Decision Trees”. In: *International Conference on New Methods in Language Processing*.