General Analysis of an Online Language Corpus

Kerwin A. Livingstone, Universidade do Porto
General Analysis of an Online Language Corpus

El análisis general de un corpus de lengua en línea

Kerwin A. Livingstone

Universidade do Porto
profesordelenguasmodernas@yahoo.es

Abstract: Corpus-based research is rapidly gaining ground in the field of Applied Linguistics. More interesting is the evidence of many online language corpora which can be easily accessed, with just the click of the mouse. A quick navigation of the Web will produce different kinds of corpora in a vast number of language areas. Given the need to find new and exciting ways to improve the language learning and teaching process, corpus linguistics does have potential for generating significant learner experiences. Taking into consideration the above-mentioned, this paper deals with the general analysis of an online language corpus. The specific corpus chosen is the Corpus del Español, one from within the Brigham Young University (BYU) Corpora, created by Professor Mark Davies in 2002 of BYU. An analysis of the corpus is done, highlighting its implications for language learning and teaching.

Keywords: corpus, language corpus, applied linguistics, language learning and teaching, Spanish, corpus-based research, corpus linguistics.

Resumen: La investigación basada en corpus está ganando terreno rápidamente en el ámbito de la lingüística aplicada. Interesante aún es la evidencia de muchos córpora de lengua en línea los cuales pueden accederse fácilmente, con un clic del ratón. Una búsqueda rápida de la Web producirá distintos tipos de córpora, en un gran número de áreas de la lengua. Dada la necesidad de encontrar nuevas y emocionantes maneras de mejorar el proceso de enseñanza-aprendizaje de lenguas, la lingüística de corpus sin duda tiene el potencial para engendrar experiencias de aprendizaje significativas. De acuerdo con lo anteriormente mencionado, este artículo versa sobre el análisis general de un corpus de idioma en línea. El corpus específico seleccionado es el Corpus del Español, uno de los córpora de la Universidad Brigham Young, creado por el profesor Mark Davies en 2002, profesor en esta universidad. Se lleva a cabo un análisis del corpus, destacando sus implicaciones para la didáctica de lenguas.

Palabras clave: corpus, corpus de lengua, lingüística aplicada, enseñanza-aprendizaje de lenguas, español, investigación basada en corpus, lingüística de corpus.
1. Introduction

Substantial collections of language texts in electronic form have been available to scholars for almost forty years, and they offer a view of language structure that has not been available before. While much of it confirms and deepens our knowledge of the way language works, there is also a fascinating area of novelty and unexpectedness – ways of making meaning that have not previously been taken seriously (Sinclair, 2004; Biber, Conrad & Reppen, 2004; Frankenberg-Garcia, Flowerdew & Aston, 2011). The afore-mentioned affirmation clearly speaks to the fact that bodies of linguistic/language texts have been accessible for a while now. These bodies of texts can be referred to as corpus (singular) or corpora (plural).

A corpus can be defined as a systematic collection of naturally occurring texts of both written and spoken language (Nasselhauf, 2004). It is systematic for two reasons: (1) the structure and contents of the corpus follows certain extralinguistic principles (the principles on the basis of which the texts included were chosen); (2) the information on the exact composition of the corpus is available to the researcher, including the number of words in each category, and in the whole corpus, how the texts included in the corpus were sampled, and the like.

Maia (2014: 3) postulates that “the main purpose of a corpus is to verify a hypothesis about language – for example, to determine how the usage of a particular sound, word, or syntactic construction varies”. Such an assertion is important for research in languages and linguistics, since it implies that empirical studies employing quantitative studies and analyses can be conducted, with the support of language corpus. Additionally, important to note also is that even though a corpus can refer to any systematic text collection, it is commonly used in a narrower sense today, and is frequently only used to refer to systematic text collections that have been computerised.

Bearing in mind the afore-mentioned, the term Linguistics has to do with the scientific study of language. Fusing the two terms together gives way to Corpus Linguistics. Corpus Linguistics can be described as “a linguistic methodology which is founded on the use of electronic collections of naturally occurring language, viz. corpora” (Granger 2002: 3). In other words, it is a method for carrying out linguistic analyses,
with the help of a computer with specialised software, and which takes into account the frequency of the phenomenon being investigated. Nasselhauf (2004) establishes that corpus linguistics has transformed itself into one of the most ubiquitous methods of linguistic research in recent years.

2. General Analysis of the Corpus del Español

2.1 Description of the Corpus

The *Corpus del Español* contains 100 million words in Spanish, from the 1200s to the 1900s. The interface of the corpus seems very organised and easy to follow. On the top left hand corner of the corpus, there are a number of sections, for example “Display”. In this section, one can either select the option ‘list’, ‘chart’, ‘key word in context (KWIC)’, or ‘compare’. The other sections are “Search String”, “Sections”, “Sorting and Limits”, and “Click to See Options” which has other subcategories for specific kinds of searches. With a click of any of these options, the researcher can begin to investigate whatever it is that needs to be investigated.

One notable characteristic of this corpus is that in the section called “Sections”, one has the option of choosing the period in which the word was used, between the 1200s – 1900s. Note-worthy is that there are four contexts from which the corpus has been constructed: *oral, fiction, news* and *academic*. Interesting to note is that under the section called “Search String”, the subcategory ‘Pos List’ allows one to select the syntactic class of the lexical item that is to be studied. Another fundamental characteristic of the corpus is that it can be accessed in English and Spanish. In other words, the instructions in English may very well be for the novice users who are not proficient in Spanish. Similarly, the instructions in Spanish are naturally directed to those who are more linguistically competent.

2.2 Using the Corpus

By way of seeing how this corpus functions, two lexical items (quantifiers), sometimes confusing in Spanish, were selected: *bastante* and *muy*. These were tested only in the period 1900s.

In the case of *bastante* (rather), a search was done to list various contexts in which the word appeared. For the period, 1900s, the frequency of this quantifier was 3,894.
This took only 0.523 seconds. On clicking the word ‘bastante’ at the top centre of the page, a vast number of examples of use appeared, together with the context in which it was used (oral, fiction, news, or academic). Upon clicking one context in which the word appears, “Estoy bastante satisfecho de mi vida”, the source information was immediately displayed, providing the date, title, source and author of the information. Additionally, an expanded context was provided to show exactly how the phrase/sentence was used.

On selecting the option ‘Chart’ in the “Display” section, an actual chart appeared which provided the frequency of the quantifier ‘bastante’ from the 1200s – 1900s. The frequency in each of the four contexts was also made ostensible. Figure 1 presents this information:

<table>
<thead>
<tr>
<th>SECTION</th>
<th>1200s</th>
<th>1300s</th>
<th>1400s</th>
<th>1500s</th>
<th>1600s</th>
<th>1700s</th>
<th>1800s</th>
<th>1900s</th>
<th>ACAD</th>
<th>NEWS</th>
<th>FICT</th>
<th>ORAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>FREQ</td>
<td>4</td>
<td>8</td>
<td>155</td>
<td>1013</td>
<td>1039</td>
<td>1945</td>
<td>3682</td>
<td>3894</td>
<td>492</td>
<td>422</td>
<td>759</td>
<td>2221</td>
</tr>
<tr>
<td>PER MIL</td>
<td>0.60</td>
<td>3.00</td>
<td>18.99</td>
<td>59.47</td>
<td>84.14</td>
<td>198.13</td>
<td>190.80</td>
<td>170.62</td>
<td>98.40</td>
<td>85.02</td>
<td>159.12</td>
<td>524.68</td>
</tr>
</tbody>
</table>

Figure 1. Frequency of use of the word ‘bastante’ from 1200s-1900s.

A check of the same word in the category ‘KWIC’ revealed the word highlighted, and the contexts in which it appeared. An example is provided below in Table 1:

| 82 | 19-OR Habla Culta: Mexico: M2 | A | B   | C         | pues, al otro día | con una aparición | bastante | 2.62 | 2.96 | 2.08 | 2.00 |

Table 1. Example of the key word in context

In relation to the above information, the same process was repeated for the quantifier muy. For the period 1900s, a total frequency of 35,393 was evidenced in 0.332 seconds. Clicking the word revealed the four contexts in which they were utilised, together with a plethora of examples of use. One selected phrase was “Es un camino muy largo el que nos queda por recorrer”. Similar to the other quantifier above, the source information and expanded context were presented.

Choosing the subcategory ‘Chart’ showed the following for the word ‘muy’ in Figure 2:
Using the ‘KWIC’ method, the following example was highlighted, as in Table 2:

2.3 Comparison of the two quantifiers ‘bastante’ and ‘muy’

By selecting the option ‘Compare’ in the “Display” section of the corpus, some of the many examples presented are displayed in Table 3:

As can be seen from the above, each of the individual quantifiers is placed in tables with an exhausted list of words with which they can be used. Word 1 is ‘bastante’ and Word 2 is ‘muy’. In each of the individual categories, contextual comparisons are made with Word 1 and Word 2. Table 4 presents examples of these data.
2.4 Analysis of the Corpus

The Corpus del Español presents an ample range of linguistic text in Spanish, dealing with all aspects of the language. It is a very user-friendly, useful corpus, and very easy to manipulate. To continue using the corpus, the researcher has to register with the system, providing valid information which, when approved, allows for continued, uninterrupted use. This is notable.

The various contexts from which the corpus has gathered data – oral, fiction, news and academic – present information gathered from the different Spanish-speaking countries across North America, Central America, South America, Europe, and the Caribbean Sea, thus providing a plethora of examples of rich, diverse linguistic text. This is very important for language pedagogy, since providing rich inputs fosters language acquisition (Krashen, 1981).

Experimenting with the corpus with other lexical items, phrases and grammar concepts have shown that the linguistic data have been carefully compiled to include almost all aspects of language production. This is note-worthy. It is an excellent tool for both language teachers and learners, and it is recommended that it be used to complement pedagogical practices, with specific reference to the learning-teaching of Spanish as a first language (L1), second language (L2), or foreign language (FL).

One recommendation would be to find a way to include recorded speech in this corpus as a means of listening comprehension and pronunciation exercises for language students. In other words, a specific section of the corpus could be reserved for listening/pronunciation drills only.
3. Implications for Language Learning and Teaching

As mentioned in the above-paragraph, this corpus has implications for Spanish language learning and teaching:

1. Too often in the language classroom, pedagogic materials do not always meet the needs of students. They are either too abstract, or very often ignore cultural aspects. Using this corpus can correct that deficiency, since language teachers will now be able to provide their students with rich, authentic examples of language use from native speakers in real contexts.

2. This corpus, extending itself from the 1200s to the 1900s, lends itself very well for various types of Spanish language study: linguistics, history (the development and use of language over time), literature (the various linguistic styles during literary periods), dialectology (comparisons of dialectal use), translation and interpreting, sociolinguistics, pragmatics, syntax, semantics, language change and variation, among others.

3. This corpus is vital for confirming a hypothesis about language (Maia, 2014). Conducting quantitative comparisons and analyses of a wide range of linguistic features in this corpus, representing different varieties of the Spanish language, can show how different features cluster together in distinctive distributional patterns, effectively creating different text types. In other words, it can create much better descriptions of many of the different language registers (informal conversation, formal speech, academic writing, among others) and dialects of native Spanish speakers (Peru, Mexico, Spain, Venezuela, and the like).

To complement the use of corpus in quantitative studies, Livingstone (2014), in a quasi-experimental longitudinal study of Spanish L2 students, conducted in 2008, made use of recorded speech of participants, which was then transcribed and stored in the computer. From this corpus of recorded speech, the author was able to quantitatively determine effectiveness of a mixed methodology for improving the linguistic skills, and those language areas that needed further study, testing, and eventual improvement.
4. Conclusion

This exercise of conducting a general analysis of the Corpus del Español has shed light on its potentialities and implications for language learning and teaching. Corpus-based studies, throughout the years, since its genesis, have opened many doors for detailed documentation of, and research about, language use. In fact, corpus-based research has become increasingly popular in applied linguistics. Corpus linguistics will continue to play a defining role in all areas of study. It is now left up to researchers, teachers, educators, and even students, to decide if they wish to be a part of this process, which can allow for improved pedagogical practices and significant educational experiences.

References


