Léxico Hispanoamericano (1493-1993) (review)

Mark Davies

La corónica: A Journal of Medieval Hispanic Languages, Literatures, and Cultures, Volume 33, Number 1, Fall 2004, pp. 261-266 (Review)

Published by La corónica: A Journal of Medieval Hispanic Languages, Literatures, and Cultures
DOI: *https://doi.org/10.1353/cor.2004.0043*

Boyd-Bowman, Peter. Léxico Hispanoamericano (1493-1993). *Eds. Ray Harris-Northall and John Nitti. CD-ROM. Hispanic Seminary of Medieval Studies, 2003. ISBN 0900411287.*

Introduction

The *Léxico Hispanoamericano* (LHA) CD-ROM provides information on the frequency, distribution, and use of tens of thousands of lexical items in Latin America in different geographic regions and in different historical periods from 1493 to 1993. This database is the end product of work that was carried out by Peter Boyd-Bowman from the 1960s through the 1980s. Since 1967, Boyd-Bowman had collected, on citation slips, entries for thousands of words from several million words of text. These citation slips formed the basis for a microfilm edition of the LHA, which was subsequently released by the Hispanic Seminary of Medieval Studies. Five sets of microfiche were produced between 1982 and 1994 covering each of the centuries from the 1500s to the 1900s. Although the microfilm editions were quite valuable in their own right, it was suggested that the material would be more accessible in searchable electronic form. In 1994 the National Endowment for the Humanities provided funding for such a project, and the current CD-ROM is the result of this effort.

Technical information

In terms of system requirements and installation information, it should be noted that the CD-ROM runs only on the Windows platform, but can be installed under any Windows operating system from Windows 95 to Windows XP. The database requires approximately 750 MB of hard drive space, and in our experience the installation program worked very well, on both a Windows 2000 Server and Windows XP Pro machine. In terms of technical issues, there is one fairly serious bug in the program. On both machines, the first time the program attempted to export a report to MS Word, it caused a problem with one of the registry values for Microsoft Office (in both versions, XP and 2003). Subsequently, each time Word, Excel, or any MS Office program was run or we attempted to Copy or Paste text, there was a 5-10 second delay as the program searched for a missing file. Once Word was re-registered, however, ([winword /r] from the Command Prompt), the problem disappeared.

The database

The citation slips which form the basis for the database come from 387 texts in approximately 105 locations, ranging from Arequipa and Asunción to the Yucatán and Zacatecas. A detailed bibliographic listing of all the source texts can be obtained from the Indexes/Sources entry on the main menu, and a simple listing of all of the localities can be obtained via Indexes/Locations. As noted, all of the citations come from texts from 1493-1993, and a listing of all of the dates can be found under Indexes/Chronologies.

One basic piece of information regarding the database that is lacking is the overall size of the textual corpus from which it was compiled. With the appearance of more and more large corpora of historical Spanish, it would be useful to have a general idea of the overall size of the LHA corpus, so that the same query in different corpora allows us to compare "apples to apples". Yet by using some simple ratios of word frequency in comparable corpora, such as CORDE from the Real Academia Española (www.rae.es), we can estimate the size of the LHA corpus. While the ratio of frequencies with different words provides differing results, we can estimate that the textual corpus was probably somewhere between one and ten million words in size.

More specific information is also lacking regarding the distribution of the corpus between different texts, locations, and historical periods. For example, what it the size of the corpus from Perú or from Chile, and how many words are there from the 1600s or the 1800s? In the original Boyd-Bowman introductions for each of the five volumes for the 1500s through the 1900s, there is some indication of the number of pages that were consulted for each of the historical periods, but this is not translated into number of words. Without knowing the relative size of each historical period, it is very difficult to determine accurately whether a word has increased or decreased in frequency from one period to another. Likewise, we can never determine whether a word was more common in one country than in another (per million words), because we don't know the relative size of the corpus from these two countries.

Queries

The LHA interface allows five different types of searches: by headword, chronologies, context words, locations, and sources. When they are used individually, however, the chronologies, context words, locations, and sources searches are not overly useful. For example, if we specify [1720-1730] for the chronologies search, we are given 4,081 entries like the following:

[1720 Chile] se habían entibiado (en la fe) [MIC 616]

[1721 México] dicha cantidad en mi poder a disposición del gobernador [FSP 71]

[1722 Guatemala] un indio principal nombrado cangrejo [FSP 44]

Yet it is not immediately apparent what one would do with these 4000 lines of text, since they are not sorted or grouped by headword or location. Likewise, if we do a search by "sources" and select the text CNG (*Crónicas de la conquista de Nueva Galicia en territorio de la Nueva España*), we obtain 1,317 lines like the following:

[1532 México] a cabo de dos días (...) partimos [CNG 226]

[1532 México] a él hirieron en un muslo y en una pantorrilla [CNG 125]

[1532 México] a ella le pesaba no podello estorbar [CNG 38]

Certainly, however, we would want to bring this data into another database program or spreadsheet, in order to make use of it.

Because the information is already in a MicroSoft Access database, it is somewhat surprising that the interface does not allow users to make use of cross-categorial comparisons. We know that the database can determine which headwords appear in texts from different historical periods and different geographic regions. Therefore, it should be possible to get a listing of all words, for example, that occur in texts from Honduras but not in texts from Nicaragua or Costa Rica. Likewise, we should be able to find all words that occurred 1700-1730, which had fallen out of use by the 1750s. With the right database architecture, such queries can be done quite easily. For example, in the Corpus del Español (www.corpusdelespanol.org) one can find all words ending in [-dura] that occur in the 1600s but not the 1700s or 1800s, or all verbs that occur for the first time in the 1900s. The ability to define sub-corpora and then compare frequencies across these sub-corpora allows for some of the most interesting and useful queries possible.

Where the historical and geographical information in the LHA is more useful, in terms of its searches, is with its "Combined Searches", as seen in Figure 1. Once we combine Location, Chronologies, Context Words, and Source information with the Headword information, we can then do some interesting queries.

For example, we can search for all words ending in [-dor] that occur in México in the first half of the 1700s, and we see 36 entries for 30 different lexical items, like the following:

[c. 1746 México] ... especialmente de los **armadores** del buseo de perlas [VST 2, 385]

[c. 1746 México] un tomín del seis por ciento del **cobrador** [VST 1, 55]

[1722 México] ...más de dos mil pesos se asignara al factor o al **ministrador** [FSP 81]

Likewise, we can search for all forms of *deslizar* in thirty different regions of Mexico from [1800-1899], and we will see the one occurrence:

[c. 1880 México] **deslizándose** como una culebra [PBR 252]

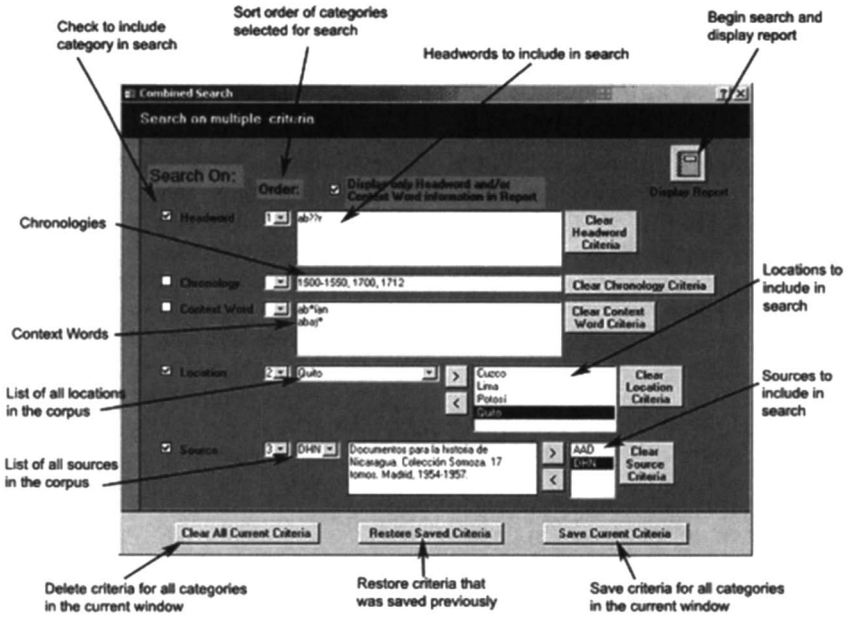

Figure 1. Combined searches

As a final example, we can search for all lemma starting with [dorm-] (*dormir, dormán, dormería, dormido, dormidero*) that occur in the three countries of Argentina, Uruguay and Paraguay during the 400 years from 1500-1899, and we will see the two occurrences in the corpus:

[1833 Argentina] **dormiendo** y comiendo [RBN 3, 111]

[1840 Uruguay] me le **dormí** al fletecillo [HAP 76]

The preceding query reveals an important point about the LHA, that it is lemmatized, meaning that it can find different forms for the same lemma –*i.e., duerme, dormían, dormiendo* (as above), y *dormí* for the lemma [*dormir*]. This is clearly better than having to search individually for 40-plus separate forms of [*dormir*], which is the strategy that one would have to follow with CORDE, since it is not lemmatized (although wildcards may help some in this case). The only other corpus of historical Spanish that is lemmatized is the Corpus del Español, which is also annotated for part of speech.

The issue of selectivity

Those who have used other corpora –such as CORDE or CREA, or the Corpus del Español– may initially be somewhat surprised by the way in which the LHA allows searches of particular lexical items. In a traditional

corpus, the entire text is full-text searchable, which means that one can search for any given word in the corpus and find all of the matching occurrences. The LHA, on the other hand, is a selective database. As Boyd-Bowman was creating the citation slips on which the searchable LHA corpus is based, he created slips just for those words that he felt would be most useful to subsequent researchers. This was the natural course of action, for the time in which he was working (1960-1980), when without computers it would have been impossible to manually create a separate slip for each of the one to ten million words of text. The result, then, is that the entries for "less common" words are quite complete, as with *candelerazo* (one occurrence; 1792-Lima), *tendezuela* (two occurrences; 1580s in Mexico and 1850s in Bogotá), or *zabullón* (one occurrence; 1890s in Medellín). For more common words like *dormir* or *tiempo* or *lindo*, however, the results are much more selective – just a fraction of the actual occurrences in the original textual corpus. Nevertheless, this selectivity should in no way diminish the magnitude of Boyd-Bowman's achievement in single-handedly creating tens of thousand (perhaps hundreds of thousands) of entries for different words.

The issue of size

Since the late 1980s, the size of corpora has increased significantly. In English, for example, the largest single corpus before the mid-1980s was the million-word Brown corpus and other one-million word corpora (such as the LOB) that were modeled on the Brown corpus. At the present time, however, "mega-corpora" such as the British National Corpus (100 million words) and Cobuild (400-plus million words) are becoming increasingly common. The same has occurred in Spanish. In the mid-1980s (the time in which the LHA corpus was essentially complete), there were no Spanish corpora larger than one million words. By the present time, however, there are several corpora of Spanish that are at least 100 million words in size. Therefore, researchers should not expect to find similar results in the LHA corpus as they would in the 200-plus million-word CORDE corpus or the 100 million-word Corpus del Español. It would be unfair to expect a corpus that was initiated in the 1960s to have the range of the more modern corpora. Yet as we will see in the following section, even the smaller LHA corpus can be very profitably used with the larger, newer resource tools.

The role of the *Léxico Hispanoamericano* vis-à-vis other corpora

With the advent of larger, perhaps more accessible corpora such as CORDE, CREA, and the Corpus del Español, how can a corpus such as LHA be used most effectively? Perhaps the best approach to using corpora for historical and variationist research is to leverage the strong points of

several different corpora, and then use them in conjunction with each other. For example, we know that CORDE is much larger than the LHA, and that it allows full-text searching. The same is true of the Corpus del Español, which (unlike CORDE and CREA) also allows queries based on wildcards, collocations, lemma, part of speech, synonyms, frequency, and user-defined lists. So how is it that the LHA can be used most effectively with these more modern corpora of historical Spanish?

First, we should remember that while it is smaller than the other corpora, the LHA has a more diverse geographic range and more detailed chronological coverage than either the Latin American component of CORDE or the Corpus del Español. For someone who wants to map the distribution and use of a particular lexical item in Latin America during the past five hundred years, the LHA is a valuable addition. In terms of leveraging different corpora, however, once we have used the LHA to search for lexical items at the "micro" level, we might then use larger corpora to search for additional occurrences at the "macro" level.

Second, as we have already mentioned, the LHA is annotated quite nicely for lemma, the different forms of a given verb, noun or adjective. This is an important point in terms of historical corpora, where there is such a wide range of spelling variation. To provide just one example, there are at least 72 possible spellings of the one verb form [*hubiese*] –*huvyese, obiesse, uujesse*, etc.–and 56 of the 72 were actually found in the Corpus del Español texts as the corpus was being lemmatized. In the LHA, one entry would find all cases of *hubiese* and potentially hundreds of other forms of *haber*. With a corpus such as CORDE or CREA, however, each of the hundreds of different forms would have to be searched individually. In terms of leveraging the advantages of different corpora, then, we might first use the LHA to find variant forms, and then use a larger corpus to find additional occurrences of these forms.

The LHA occupies an important place in the world of Spanish corpora. Because of its inception in the 1960s, it is somewhat more limited in terms of size and the range of lexical entries, as compared to subsequent "mega-corpora" such as CORDE, CREA, or the Corpus del Español. Yet because of the extremely careful way in which the database was created by Boyd-Bowman, it has a degree of accuracy and precision that the mega-corpora sometimes lack. By using these different resources in conjunction with each other –and leveraging the advantages of each– researchers can now carry out a range of studies that would have been unthinkable even ten years ago.

Mark Davies
Brigham Young University