CREATING USEFUL HISTORICAL CORPORA: A COMPARISON OF CORDE, THE *CORPUS DEL ESPAÑOL*, AND THE *CORPUS DO PORTUGUÊS*

MARK DAVIES
Brigham Young University

1. Introduction

Many people mistakenly think that corpora are composed strictly of words and phrases, and that the corpus interface and architecture exist mainly as an "afterthought", to allow users to "look through many books and pages" to find words and phrases as quickly as possible. In this view, the best corpora are those that are the largest, which have texts from the widest range of genres and sources, and whose texts are the most accurate. However, this is an overly-simplistic view, and application of this approach may result in a corpus that is only of minimal value for many types of linguistic research.

As we will discuss at some length in this paper, a truly usable corpus is composed of at least two elements:

- The textual corpus (the texts in the corpus)
- The corpus architecture and interface

One can have a historical corpus that is composed of hundreds of millions of words of text from several different centuries, and which represent a wide range of genres. But without an adequate architecture and interface, this data is in essence "trapped", with little if any way of getting the data out. Users may be limited, for example, to just looking for specific words and phrases (such as with the Google interface), or to find the first occurrence of a word or phrase. If this is the extent of the complexity of the search, then essentially any architecture will work. But for more complex research on morphological, syntactic, lexical, or semantic change, this simplistic architecture may be completely inadequate. On the other hand, one can have the most advanced architecture and interface imaginable, but if it is built on top of a weak textual corpus, then its value is likewise questionable. For example, if the corpus is composed of just a million or so words, then there simply may not be enough data to answer the relevant questions.

In this paper, we will review two corpora that have been widely used for research on historical Spanish linguistics –CORDE (from the Real Academia Española) and the Corpus del Español. (For information on an earlier version of the Corpus del Español, see Davies 2002, 2005a, and 2005b. For information on the new architecture (from late 2007), see Davies 2008a and 2008b). In addition, we will briefly consider one corpus of Portuguese –the Corpus do Português– which has an architecture and interface that is exactly the same as that of the Corpus del Español. We will briefly consider the textual corpus for each of these three corpora. However, the main focus of this paper deals with the way in which the architecture and interface of the corpora either help or hinder research on a wide range of linguistic phenomena, particularly those dealing with language change and variation.

2. The textual corpus

CORDE was created in the late 1990s, and was the first large corpus of historical Spanish. It is composed of approximately 250 million words of text, with good representation across the different historical periods, and a nice balance between genres, including poetry, historical writings, literature, didactic materials, and so on.[1]

The Corpus del Español was completed in 2002, and underwent a major revision in late 2007. It is composed of about 100 million words from Old Spanish to the late 1990s, with about 18 million words from the 1200s-1400s, 42 million words from the 1500s-1700s, and about 40 million words from the 1800s-1900s. As with CORDE, it is also composed of texts from a wide range of genres, including more than five million words from transcripts of spoken conversation from the late 1900s. For the 1900s, they are evenly divided among spoken, fiction, newspaper, and academic. Complete details on each of the nearly 14,000 texts can be found via the "Texts" link at the corpus website, and users can download an Excel file listing all of the texts.

The Corpus do Português was completed in 2006, and was revised in early 2008. It is composed of about 45 million words of text, with about 15 million

[1] There are differing figures for the size of the CORDE corpus. Pascual and Domínguez (this volume) mention "over 300 million words". The page http://www.rae.es/rae/Noticias.nsf/Portada3?ReadForm&menu=3 suggests that it is 250 million, while at the CORDE website itself (see http://corpus.rae.es/ayuda_c.htm), it says 125 million words. Actual searches of the corpus suggest about 220-240 million. This is calculated by finding the frequency for common words like *de, que, en*, and then using a ratio to compare those frequencies from CORDE to the frequency in a corpus of a known size, such as the Corpus del Español.

from the 1200s-1400s, 10 million words from the 1500s-1700s, and 15 million from the 1800s-1900s. For the 1700s and later, the texts are evenly divided between Portugal and Brazil, and for the 1900s they are evenly divided among spoken, fiction, newspapers, and academic. As with the Corpus del Español, complete details on each of the more than 60,000 texts can be found via the "Texts" link at the corpus website, and users can download an Excel file listing all of the texts.

In summary, each of these three corpora is quite robust in terms of the textual composition, especially when they are compared to what is available for other languages. For example, the most widely-used corpus of historical English (the Helsinki Corpus) contains only about 1.6 million words from Old English to the early 1700s, and there are virtually no structured corpora of English from the 1700s-early 1900s (see http://davies-linguistics.byu.edu/personal/histengcorp.htm). So even the 45 million word Corpus do Português –the smallest of the three corpora that will be compared in this paper– is about thirty times as large as the main corpus of historical English. Both CORDE and the Corpus del Español, on the other hand, are more than sixty times as large as the Helsinki Corpus.

3. Using historical corpora to study a wide range of linguistic phenomena

Truly useful historical corpora should allow users to carry out research on phenomena like the following:

- Lexical: simple. At the most basic level, users can search for a word or phrase, find the first occurrence of the word or phrase, and see all occurrences in context.
- Lexical: more advanced. Users can easily see the frequency of a word or phrase over time, with normalized frequencies. (In other words, frequency per thousand or per million words of text, to account for the different corpus size in different historical periods.)
- Lexical: most advanced. Rather than having to tell the corpus what specific words or phrases to search for, the corpus can generate a list of words whose frequency matches certain criteria, such as nouns that entered the language in the 1600s, or all words that are used at least five times as much in the 1200s than in the 1300s.
- Morphological. Users should be able to search by prefixes, suffixes, and roots, and see the frequency of each matching form in the different historical periods, as well as the overall frequency of all forms in each historical period.

- Syntactic. In a truly useful corpus, the words will be tagged for part of speech and they will be lemmatized. This allows users to search for specific syntactic constructions, rather than having to search thousands of different exact phrases, with is the only option with an untagged and unlemmatized corpus.
- Semantic: simple. Users can find the most frequent collocates (nearby words) of a given word or phrase, which obviously provides very good insight into the meaning of the word. Virtually any corpus architecture and interface allows users to see the nearby words on a case by case basis, but truly useful corpora summarize all of this collocational information for all occurrences of a given word or phrase.
- Semantic: more advanced. Assuming the corpus can find collocates, it should be possible to compare these across historical periods or between different genres. Changes in collocates across historical periods often serve as markers of semantic change.
- Semantic: most advanced. Rather than just searching for words and phrases, users can search by semantic field. For example, if a thesaurus is integrated into the corpus, or if users can create customized lists of words, then they could create a search where any word in a semantic field is part of the query. An example of this might be [member of family] followed by [synonym of *pedir*] followed by [synonym of *limpiar*], or [time of day] near [synonym of *lúgubre*]. Likewise, they could compare the frequency for all of the words or phrases in an entire semantic field, and compare the frequency and distribution of each member over time.

In the sections that follow, we will provide concrete examples of how the three corpora – CORDE, the Corpus del Español, and the Corpus do Português –can (or cannot) be used to research the wide range of phenomena listed above. As we do so, some readers may begin to gain an entirely new perspective on what can be done with historical corpora. If they have used corpora with limited architectures and interfaces, they may be used to just doing queries to find the occurrences of a specific word or phrase. Once a person has used a corpus that allows a wide range of queries like these, however, they suddenly realize that there are hundreds and thousands of topics in historical linguistics that could be studied with a full-featured corpus.

4. Lexically-oriented searches: basic

As was mentioned above, the most basic thing that a corpus should allow one to do it to search for a word or phrase, find the first occurrence of the word or

phrase, and see all occurrences in context. The programs to allow such searches are plentiful, and (because of the simple search), all should be quite fast –1-2 seconds for even a 100 million word corpus.

CORDE is of course able to do these basic searches, and it does them quite well. For example, suppose that the user wants to find all occurrences of the word *braueza*. After submitting the search, the user sees that there are 273 tokens in 86 documents. Clicking on "Obtención de Ejemplos", the user then sees Keyword in Context (KWIC) entries like the following:

TABLE 1
Keyword in Context display with CORDE

1	los buenos que los malos, sino que á en ellos **braueza** e espántanse ante las águilas. E los ruuios s	1250	Toledo, Abraham de
2	o buen fecho uos dara mas su amor que non uuestra **braueza**. Et sepades que los non abredes a uuestro man	c. 1250	Anónimo
3	que demostro Saturno. de tempradas maneras entre **braueza** & mansedat. de fermosa voz. de buen conseio	1254-1260	Anónimo
4	trar al Rey deue lo ffazer omjldosa mjente & ssin **braueza**. Et otrossi non deue denostar njn Amenazar A	a1260	Anónimo

One disadvantage of the CORDE interface, however, is that it limits users to seeing the word in context, only when the word occurs less than 1000 times in the corpus. For thousands of words, then, there is no easy way to see the words in context.

A basic search for simple words and phrases works in a similar manner with the Corpus del Español and the Corpus do Português. With the Corpus del Español, for example, after submitting the search, the user will see the following:

TABLE 2
Frequency listing with the Corpus del Español

	PALABRA	**TOT**	**s13**	**s14**	**s15**	**s16**	**s17**	**...**
1	BRAUEZA	144	**48**	**32**	**27**	32	5	
			6.10	10.80	2.78	1.62	0.31	

This shows the raw frequency of the word in each century, the occurrences per million words as well. Unlike CORDE, however, the user can easily see the frequency per century. In addition, one can see the keyword in context for any word, not just those with a low frequency (as is the case with CORDE).

In the Corpus del Español, clicking on the numbers in any column will show the Keyword in Context display for that century, or one can see all entries at one time by clicking on TOTal.

TABLE 3
Keyword in Context display with CORDE

20	General Estoria IV	& quando estos consules uieron que ninguna yente non se podie enffestar contra la **braueza** de belino & de brennio. & ouieron estos consules so conseio con sos senadores
21	General Estoria V	por vos non detener ca pieça ha que quedaron los vientos contrarios & la **braueza** dela mar. / las primeras estrellas del cielo paresçen ya & yua el sol
24	Judizios de las estrellas	fuere en Gemini. es conlo que demostro Saturno. de tempradas maneras entre **braueza** & mansedat. de fermosa voz. de buen conseio. & de bien fablar
25	Siete partidas I	por todo esto no serie derecha si la diesse con sanna o con **braueza** por malquerencia que ouiesse contra el. E por esto dixo santyago en su epistola

To this point, then, the searches in the two corpora are quite similar. CORDE has one advantage –in that it is a large corpus– while the Corpus del Español (and the Corpus do Português) have the advantage of showing the frequency in each century and in showing the keyword in context for all words, regardless of frequency.

5. Lexically-oriented searches: more advanced

In addition to just displaying all occurrences of a given word or phrase, however, users often want to know how frequent a word was in different centuries or other historical periods. It is at this point that CORDE begins to exhibit some serious weaknesses. For example, after searching for *braueza* and then selecting "Ver estadística", the user sees:

TABLE 4
Frequency (by year) with CORDE

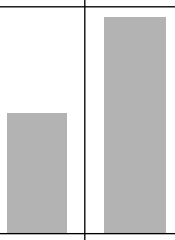
Año	%	Casos
1627	20.95	35
1547	20.35	34
1610	17.96	30
1632	8.98	15
1566	4.79	8
1622	2.39	4
…	…	…
Otros	14.97	25

This table tells us the specific *years* in which the word or phrase is most common, but it is impossible to see the frequency by decade or by century. It does little good to show that the word was the most frequent in 1627, if in fact the word is much less common in the 1600s than in the 1200s or 1300s. The other serious problem is that the figures are not normalized. In other words, we see the raw frequency per year, but a word or phrase may be more common in that year simply because there are more words for that year in the corpus. Any serious comparison of frequency requires that the results be "normalized" across historical periods, so that we can take into account the differing sizes of the corpus in different historical periods, and see how frequent the word or phrase is per million words.

The Corpus del Español and the Corpus do Português allow such types of searches quite easily. For example, with *braueza* in the Corpus del Español, we can either see the "table display" (as in Table 2 above), or a chart display (Figure 1).

This shows us the raw frequency (e.g., 32 tokens in the 1300s), as well as the important normalized frequency (por millones), which takes into account the size of the section, in millions of words. For example, there are 32 tokens in the 3.0 million words from the 1300s, or 10.8 tokens per million words. A chart such as this is the only way to really see shifts in the frequency of a word, phrase, or construction, and it is only possible with the Corpus del Español (and the Corpus do Português).

FIGURE 1
Corpus del Español: Frequency of word and phrase by century

SECCIÓN	s13	s14	s15	s16	s17	s18	s19
							
POR MILLONES	6.1	10.8	2.8	1.6	0.3	0.0	...
TAMAÑO	7.0	3.0	9.7	19.7	14.8	11.5	...
OCURRENCIAS	48	32	27	32	5	0	...

6. Lexical: most advanced

Rather than having to tell the corpus what specific words or phrases to search for, a well-designed corpus architecture and interface would generate a list of words whose frequency matches certain criteria. For example, it might find all nouns that entered the language in the 1600s, or all words that are used at least five times as much in the 1200s than in the 1300s. Such as query is completely impossible with CORDE. All it can do is search for specific words and phrases. If it does "know" the frequency of all words and phrases in all historical periods, it certainly does not allow researchers to use that information as part of the query.

With the Corpus del Español and the Corpus do Português, on the other hand, such queries are quite simple. For example, with the Corpus del Español, one can simply search for [nn*] (nouns) and select [s. XIII-s. XIV] (1200-1499) "SECción 1" to compare to [s. XIX-s. XX] (1800-1999) "SECción 2". Within one or two seconds, the user sees the following list. (Note that in the version on the web, there are frequencies (raw and normalized) for each word, as well as links to see the word in context, as shown in Table 6. In Table 5, we have simplified the display.)

Obviously, only some of the words in this list are meaningful. Many words are simply spelling variants, and others are proper nouns that might occur in a handful of texts in one century but not the other.

Although the focus of this paper is on older stages of Spanish, perhaps it might be useful to see an example from Modern Spanish. The following table

TABLE 5
Corpus del Español: comparison of word frequency by century
(all words at one time)

Siglo XIII	**Siglo XIV**
capitolo, ascendente, ladeza, saturnus, orizon, morauedis, roque, xaque, acendent, armella, murcia, baldouin, ascendent, segonda, gudufre, significador, fferrando, corualan, alffil, zonte, dond, iudga, tiemplo, boymonte, catamiento, infortunas, uinie, caput, camyaron, sacrifficio, declinacion, sacrifficios, yguador, juppiter, algarue, linnas, deillos, hierusalem, decima,	armadas, osso, ome, avia, paris, ahe, collado, camjno, hercoles, .ley, çima, rrayzes, armada, elena, falcon, yuierno, verano, gonçales, encarnaçion, pase, bjen, venja, avras, falcones, jnfante, façer, puerco, ynfanta, ynfante, ynperio, vyno, venjdo, sembrar, fojas, ençima, talante, mjel, menalao, syenpre, dolençia, ssiete, avedes, castilla, muria, aujdo, peça, arroyo, vuas, çiençia, termjno, tenjan

shows (to the left) nouns that are common in the 1900s but not the 1800s and (to the right) those that are common in the 1800s but not the 1900s. (SEC 1 and PM1 gives the raw frequency and normalized frequency (per million) in Section 1 (with SEC 2 and PM 2 for Section 2), as well as the ration between the two.) For example, *control* occurs 3059 times in the 1900s but only 3 times in the 1800s, and so the normalized frequency is 1,033 times greater in the 1900s than the 1800s. Likewise, *aposento* occurs 1174 times in the 1800s but only 44 times in the 1900s, and the normalized frequency is about 26 times more frequent in the 1800s than in the 1900s. (Note that a frequency of 0 is assigned a value of 0.1 to avoid division by zero.)

Due to the architecture of the Corpus del Español and the Corpus do Português, where the corpus "knows" the frequency of each word and phrase in each historical period, such comparisons are quite simple. But with CORDE, where the corpus apparently does not know the frequency of words and phrases in each section (until they are searched for, one specific word at a time), such a listing would be completely impossible.

7. Morphological

Ideally, users should be able to move beyond exact words and phrases and search by prefixes, suffixes, and roots. This will allow them to see the frequency of each

TABLE 6
Corpus del Español: comparison of word frequency (nouns in 1800s/1900s)

+ 1900s / – 1800s						+ 1800s / – 1900s					
PALABRA	SEC1	SEC2	PM 1	PM 2	RATIO	PALABRA	SEC1	SEC2	PM 1	PM 2	RATIO
SECTOR	2540	0	111.29	0.00	1,112.95	VENTURA	1237	25	53.47	1.10	48.81
CONTROL	3059	3	134.04	0.13	1,033.67	APOSENTO	1174	44	50.74	1.93	26.32
PELÍCULAS	942	1	41.28	0.04	954.94	HONRA	1652	63	71.40	2.76	25.87
TELEVISIÓN	2119	0	92.85	0.00	928.48	MENESTER	1545	60	66.78	2.63	25.40
INICIO	838	1	36.72	0.04	849.51	ERMITAÑO	553	23	23.90	1.01	23.72
DIRIGENTE	795	1	34.83	0.04	805.92	DUQUESA	632	27	27.32	1.18	23.09
CAPITALIZACIÓN	705	1	30.89	0.04	714.68	CONDESA	1370	59	59.22	2.59	22.91
SECTORES	1360	0	59.59	0.00	595.91	PRESIDIO	618	29	26.71	1.27	21.02
CONTAMINACIÓN	550	1	24.10	0.04	557.55	MARQUÉS	1676	86	72.44	3.77	19.22
FÚTBOL	1208	0	52.93	0.00	529.31	SAZÓN	1133	60	48.97	2.63	18.63
LÍDER	1119	0	49.03	0.00	490.31	VIZCONDE	588	33	25.42	1.45	17.58
EMERGENCIA	452	1	19.81	0.04	458.21	PRECEPTO	386	23	16.68	1.01	16.56

matching form in the different historical periods, as well as the overall frequency of all forms in each historical period.

CORDE has serious problems in terms of morphologically-oriented searches, because the search engine was not designed to be used for linguistically-oriented research. In the best of cases, the corpus produces results, although they are not overly useful. For example, suppose that a user searches for des*m?ento in the 1200s-1400s. The corpus indicates that there are "797 casos en 81 documentos". One can then page through all of the 797 tokens –one by one– and manually count up the total for each different form (*desfazimiento, desaffiamiento*, etc.) to see how frequently each one occurs. This would, however, take an hour or two. One could select "Recuperar/agrupaciones" to see the most frequent two, three, and five word strings (*destroymiento de*, *destruimiento de*, etc), which might only take an hour or so to find the most frequent words that match this pattern. And these searches only work when the total number of tokens for a given form occurs 1000 times or less in the corpus. For a search like *azo (*puñetazo, portazo*, etc.) the corpus simply states that "No se pueden ver estadísticas. Demasiados documentos". Again, this is because their search engine was designed to allow users to find and read entire documents (like with Google), and (in this case, at least) it is inadequate for linguistic research.

With the Corpus del Español and the Corpus do Português, however, morphologically-oriented searches are both easy and fast. For example, suppose that a user wants to find the most frequent forms for *des*m?ento* in the 1200s-1400s. Within about one second, s/he will see the following (Table 7).

In this case, TOT is the total count for the word in all centuries (only the 1200s-1600s are shown in the table above, but all are seen in the web interface). The interface then shows the frequency of each form in each century (e.g., 54 cases of *desterramiento* in the 1200s), as well as the total for the selected centuries (in this case the 1200s-1400s) in the rightmost column. Users can select whichever forms and whichever centuries are of interest, and then click to see the words in context.

In addition to seeing the individual frequencies for all matching forms (as in the table above), it is also possible to see the aggregate total for all matching forms in each century, as in Figure 1 above. Finally, as was described in Section 6, we can also compare the frequency of forms across different sections of the corpus. For example, suppose that a user of the Corpus do Português wants to see which words ending in *çar* are more common in the 1300s and the 1400s, respectively. In less than one second, he would see the following (Table 8).

This shows, for example, that *escabeçar* is found 11 times in the 1300s but none in the 1400s, and that *percalçar* is found 68 times in the 1400s, but only once in the 1300s (and is there about 44 times as common per million words in the 1400s). The ability to compare word forms across different centuries is a

TABLE 7
Corpus del Español: frequency of word forms (des*m?ento in 1200s-1400s)

	PALABRA	TOT		s13	s14	s15	s16	s17	…		SEC 1
1	DESTRUYMIENTO	121		89	17	15					121
2	DESTROYMIENTO	97		94	2	1					97
3	DESTRUYMJENTO	87		14	45	28					87
4	DESFALLESÇIMJENTO	71				71					71
5	DESTERRAMIENTO	67		54	10	2					66
6	DESCOMULGAMIENTO	42		41		1					42
7	DESTERRAMJENTO	42		4	14	24					42
8	DESAFIAMIENTO	31		19	10						29
9	DESPRECIAMIENTO	26		25			1				25
10	DESEREDAMIENTO	24		22		2					24
11	DESFAZIMIENTO	22		13	3	6					22

TABLE 8
Corpus do Português: comparison of word forms (*çar in 1300s / 1400s)

+ 1300s / – 1400s						+ 1400s / – 1300s					
PALAVRA	OCCOR1	OCCOR2	PM 1	PM 2	PROP	PALAVRA	OCCOR1	OCCOR2	PM 1	PM 2	PROP
ESCABEÇAR	11	0	5.98	0.00	59.83	LAMÇAR	46	0	16.17	0.00	161.71
RENUÇAR	5	0	2.72	0.00	27.19	GAANÇAR	21	0	7.38	0.00	73.82
DESPEDAÇAR	4	0	2.18	0.00	21.76	ALCÁÇAR	18	0	6.33	0.00	63.28
TERÇAR	3	0	1.63	0.00	16.32	LLAMÇAR	13	0	4.57	0.00	45.70
POSFAÇAR	3	0	1.63	0.00	16.32	PERCALÇAR	68	1	23.90	0.54	43.95
ENDERËÇAR	10	1	5.44	0.35	15.47	ADERENÇAR	9	0	3.16	0.00	31.64
APREÇAR	4	1	2.18	0.35	6.19	EXALÇAR	9	0	3.16	0.00	31.64

powerful feature of the Corpus del Español and the Corpus do Português, but is not possible with CORDE.

8. Morphology: lemmatization

Because Spanish and Portuguese have rich morphology, it would be very useful to be able to search for all of the forms of a word at one time, rather than having to search for each form individually –one after another. With CORDE, however, lemma-based queries are impossible. The corpus does not "know" the forms of *saber* (*sé*, *sabemos*, *supieran*, etc), *blanco* (*blanco*, *blancas*, etc), nor any other word of Spanish.

With the Corpus del Español and the Corpus do Português, however, lemma-based searches are quite easy. For example, if a user of the Corpus do Português wants to see the most frequent forms of *fazer* in the 1300s-1500s, he would simply enter [fazer] (the brackets indicate lemma). Within less that a second, he would see a chart like the following (Table 9).

(In this case, TOT refers to the total frequency of all tokens in the entire corpus, and SEC 1 refers to the overall frequency in the selected sections of the search form, in this case the 1200s-1400s).

As with other types of searches, users can also compare the frequency across time periods in the corpus. For example, if a user wanted to see which forms of *ter* (= Spanish *tener*) occurred more in the 1300s than in the 1400s, he would simple enter [ter], select SECTION 1 = [1300s] and SECTION 2 = [1400s]. In less than a second, he would see the following (Table 10).

This shows, for example, that there are 171 tokens of *tehudo* in the 1300s, but only 3 in the 1400s, making it 88 times as common (per million words) in the 1300s. Although it is useful to be able to map out the changes in forms for a given word over time, the real power of lemmatization, however, is that it allows us to carry out complex syntactically-oriented searches, as we will discuss in the following section.

Before discussing syntax, however, let us briefly consider what is involved in creating a lemmatized corpus. For the modern stages of the language, there are complications such as whether a form like *limpia* belongs to the verb *limpiar* or the adjective *limpio* in a given case (this is done by looking at the context of the word as it is being lemmatized). For older stages of the language, however, it is much more complex. Not only do we have to take into account the context, but also all of the variant spellings of the particular form. For example, in Old Spanish there are about 72 theoretically possible forms for the single Modern Spanish form *hubiese*, depending on whether there is an initial [h], whether the first

TABLE 9
Corpus do Português: lemmatization (forms of *fazer* in 1300s-1700s)

	PALABRA	TOT		s13	s14	s15	s16	s17	…	SEC 1
1	FAZER	47858		**3741**	**5908**	**7508**	3602	2750		**17157**
2	FEZ	24728		**1582**	**3241**	**5117**	1738	1351		**9940**
3	FAZ	20287		**552**	**2023**	**2547**	1940	2002		**5122**
4	FEITO	11120		**879**	**884**	**1595**	750	387		**3358**
5	FAZEM	8257		**86**	**866**	**1705**	819	974		**2657**
6	FAZIA	10024		**305**	**803**	**1423**	639	364		**2531**
7	FEITA	6302		**581**	**572**	**750**	561	170		**1903**
8	FAÇA	4566		**384**	**738**	**689**	414	355		**1811**
9	FAZENDO	7029		**231**	**529**	**922**	440	294		**1682**
10	FEYTO	1349		**531**	**373**	**338**	9	96		**1242**
11	FEZESSE	1061		**585**	**391**	**76**		9		**1052**

TABLE 10
Corpus do Português: comparison of lemma forms (*ter* in 1300s / 1400s)

+ 1300s / – 1400s							+ 1400s / – 1300s						
	PALAVRA	OCCOR1	OCCOR2	PM 1	PM 2	PROP		PALAVRA	OCCOR2	OCCOR1	PM 2	PM 1	PROP
1	TEVERON	33	0	17.95	0.00	179.48	1	TẼ	117	0	41.13	0.00	411.30
2	TÏJNÃ	18	0	9.79	0.00	97.90	2	TEREM	50	0	17.58	0.00	175.77
3	TEHUDO	171	3	93.00	1.05	88.19	3	TIUER	44	0	15.47	0.00	154.68
4	TIJÑA	16	0	8.70	0.00	87.02	4	TEẼR	41	0	14.41	0.00	144.13
5	TẼN	16	0	8.70	0.00	87.02	5	TEẼS	35	0	12.30	0.00	123.04
6	TERRÍA	15	0	8.16	0.00	81.58	6	TYNHAM	35	0	12.30	0.00	123.04
7	TÏJNA	15	0	8.16	0.00	81.58	7	TÏIA	33	0	11.60	0.00	116.01

vowel is [o] or [u], whether there is a single or a double [s], and so on. In the corpus, 56 of these 72 theoretically possible forms actually appear, including the following (with their associated frequency):

TABLE 11
Corpus del Español: forms of *hubiese*

hubiese 2776, ouiesse 1961, ouiese 1898, oujese 558, oviese 423, oviesse 397, uviesse 392, oujesse 359, huviesse 337, houiesse 316, hobiese 186, huuiesse 134, hubiesse 60, ouyesse 54, uviese 53, houiese 44, ubiese 24, huujesse 19, habuisse 18, huviese 17, obiese 12, ovjese 12, huuiasse 11, hobiesse 8, obiesse 7, ouisse 4, ouyese 4, ovyese 4, houjese 3, hoviesse 3, huujase 2, huuisse 2, huuiese 2, hoviese 2, ouesse 2, ovisse 2, ovyesse 2, ubiesse 2, uvisse 2, uuiese 2, huuiose 1, huujasse 1, huujese 1, houiosse 1, oujsse 1, ouiosse 1, objese 1, huviessse 1, huvíesse 1, uvíesse 1, uvyesse 1, uuiesse 1, ovjesse 1

Imagine the difficulty in lemmatizing a historical corpus, with such a high degree of spelling variation. The 56 forms just given are for just one form (*hubiese*) of just one verb (*haber*). If we multiply this by all of the possible forms for each lemma, and all of the possible lemmas, we see that there are more than a million distinct forms in a 50-100 million word historical corpus, which need to be lemmatized. In the case of the Corpus do Português, there is nearly complete lemmatization of all forms, and it is somewhat less complete for the (older stages of the) Corpus del Español. But both of these corpora have much more lemmatization than CORDE, which has none at all.

9. Syntax

Perhaps the best example of the difference between a corpus architecture that was designed to do linguistic research and one that was not, is in the area of syntax. Let us briefly consider a syntactic construction or two, and see how researchers would study the constructions using CORDE, the Corpus del Español, and the Corpus do Português. Let us start with the causative construction, which is composed of a form of *hacer* followed by an infinitive (*fizo llamar*, *haze venir*, *fizieron escriuir*, etc) (see Davies 1994, 1995a, 1995b, 1996a, 1996b, 2000).

How would one study this construction with CORDE? Remember that CORDE is not lemmatized, so we would have to search for each possible form of *hacer* individually. In addition, it is not tagged for part of speech, so it does not know what a verb, or noun, or infinitive is. In other words, the possible combinations are all of the forms of *hacer* (possibly 200 or more, if we consider variant historical spellings) plus about 10,000 unique forms for infinitives. (There are about 5400 unique forms for infinitives in the 1800s-1900s portion of the Corpus del Español, and we probably have to double that for all forms back to Old Spanish). This would result in perhaps two million (200 x 10,000) unique potential two-word strings. Obviously, it would take a very, very long time to do two million searches.

One might think that it would be possible to take an alternate approach with CORDE, and simply use wildcards to search for a syntactic string. For example, even though CORDE doesn't know what the forms of *hacer* are, or what an infinitive is, it should be possible to search for something like *f?z* *r* (*fizjesse estar*, *fazen dezir*, etc). Yet this does not work either. Because CORDE can't really handle substrings, a query such as this causes the corpus to grind away for 4-5 minutes, before "timing out" and producing an error. In other words, there really is no way to do syntactic research with CORDE, without searching for thousands or millions of unique strings.

With the Corpus del Español and the Corpus do Português, things would be considerably easier. Using the Corpus del Español, for example, the researcher would simply enter *[hacer] [vr*]*, and in less than three second he would see all of the relevant forms, such as (Table 12).

Let us take a second example. In the case of passives, we want to find cases of a form of *ser* followed by a past participle (*fueron llevados*, *sera destruido*, etc). With CORDE, we would have to look for each possible spelling variant of each conjugation of *ser* (perhaps 200 or more), followed by each possible past participle (possibly 4000-5000 or so). This may not be as time-consuming as with the causative, but it would still take a year or so of non-stop work. With the Corpus del Español or the Corpus do Português, however, we would simply enter [ser] [vk*]. In less than four seconds, a user of the Corpus do Português would see the following, which are the most common passives in the 1300s and 1400s (Table 13).

As we can see, there is a huge difference between a corpus architecture that is designed around searching for exact words and phrases (which is what CORDE does quite nicely) and one that can include linguistic annotation such as lemmatization and part of speech. In the first case, it is either very difficult or impossible to do serious research on syntax and syntactic change. In the second case, the corpus architecture makes such searches both quick and easy.

TABLE 12
Corpus del Español: *fazer* + infinitive in 1200s-1500s

	PALABRA	TOT	s13	s14	s15	s16	…	SEC 1
1	FIZO FAZER	241	**107**	**61**	**61**	12		**229**
2	FAZER SABER	205	**78**	**26**	**67**	34		**171**
3	FIZO MATAR	113	**14**	**23**	**74**			**111**
4	FAZER PERDER	92	**51**	**24**	**16**	1		**91**
5	FIZO LLAMAR	64	**14**	**17**	**32**	1		**63**
12	HIZO HAZER	64	**21**	**4**	**19**	20		**44**
13	FIZO VENIR	74	**17**	**5**	**18**	32		**40**
14	FIZO ESCRIUIR	38	**36**	**2**				**38**
15	FIZO LEUAR	36	**7**	**11**	**18**			**36**
16	FIZO PRENDER	33	**9**	**8**	**15**	1		**32**
17	HAZER SABER	56	**10**	**1**	**20**	24		**31**
18	FIZO TORNAR	30	**9**	**8**	**11**	2		**28**
26	FIZO VENJR	24	**2**	**17**	**5**			B

TABLE 13
Corpus do Português: passive (*ser* + past participle) in 1200s-1400s

	PALABRA	TOT	s13	s14	s15	s16	s17	...	SEC 1
1	SEER FEITA	191	**141**	**48**	2				**189**
2	FOY FEITO	211	**87**	**93**	25	6			**180**
3	FOR MOSTRADA	188	**18**	**151**	15		4		**169**
4	HE DITO	191	**53**	**104**	27		7		**157**
5	FOY FEITA	148	**79**	**57**	8	4			**136**
6	FOY POSTO	160	**89**	**46**	14	6	5		**135**
7	HE CHAMADO	170	**29**	**102**	28	2	9		**131**
8	FOSSE DADO	144	**94**	**14**	9				**108**
9	FOY DADO	142	**51**	**51**	36		4		**102**
10	HE FEITA	136	**36**	**65**	20		12		**101**
11	HE DADO	134	**35**	**64**	19	4	10		**99**

10. Semantics: simple

As corpus linguists are fond of saying, "you can tell a lot about a word by the other words that it hangs out with". Sometimes, the collocates (nearby words) simple confirm what we already know. For example, the most common nominal collocates (nearby words) for *selva* are *árboles*, *vegetación*, *sierra*, *bosque*, etc. For a less concrete word, they are often more insightful. For example, the most common nouns occurring with forms of *lúgubre* are *acento*, *voz*, *silencio, noche*, *eco*, *gemido*, etc. And for foreign language learners, collocates can easily help them to see the difference between two words in the foreign language, which would both be translated as one single word in the native language. For example, native speakers of Spanish have a good sense of the difference between *blando* and *suave*, but this is difficult for native speakers of English, where both mean "soft". However, by seeing the collocates (*blando = lecho*, *tejidos*, *maderas*, *cera*, *cama*, *créditos*, while *suave = música*, *inviernos*, *melancolía*, *pelo*, *aire*, *temperaturas*) the language learner can begin to acquire the same type of intuitions regarding meaning, which the native speaker already had.

The key to meaning, then, is often found in collocates, or the nearby words. Virtually any corpus architecture and interface allows users to search for a word, and then see that word in context. The corpus user can always go one by one through the examples, making notes about common nearby words, and then trying to use this to discern meaning. However, this can be extremely time-consuming for common words. A much better approach would be to have the corpus find all of the collocates by itself, and then present them to the user in order of frequency.

Let us briefly consider how CORDE, the Corpus del Español, and the Corpus do Português allow users to find and process collocates, to gain insight into word meaning. Turning first to CORDE, suppose that we want to examine the nearly 38,000 collocates of all forms of *duro* (*dura*, *duros*, etc). Assuming that it takes a user about 20 seconds to find each occurrence in context and write down (what he or she assumes to be) the relevant nearby words, it would take about 26 hours (at eight hours a day) to go through all of the relevant examples. And this assumes that the user does not then decide to change the width of the "collocates window", or search for a different type of collocate, in which case s/he would have to spend another month or so.

CORDE does allow users to see "agrupaciones" for a given word, such as those for the single form *duro* from the 1200s (Table 14).

But such a listing is of little value. Because CORDE does not have any "built-in" way of knowing which words are relevant, it gives us phrases like *dura la*, *dura en*, etc (since *la*, *en*, etc. occur frequently with almost any word), but these

TABLE 14
CORDE: *duro* + collocates

		%	casos
duro	la	6.69	30
duro	el	6.02	27
duro	en	5.35	24
duro	fasta	3.79	17
duro	esta	3.34	15
duro	e	3.34	15
duro	&	2.23	10
duro	de	2.23	10
duro	mucho	2.23	10
Otros		64.73	290

phrases provide little if any insight into the meaning of the word. At any rate, it only lists the nine most frequent collocates, which is not enough to be meaningful anyway.

Things are much easier with the Corpus del Español and the Corpus do Português. A user simple enters the "node word" (e.g., *duro*, or *lúgubre*, or *selva*), optionally selects the part of speech of the collocates, and within about 2-3 seconds s/he has all of the collocates, in order. For example, suppose that a user wants to find collocates relating to the concept *duro* in Old Spanish. After entering [=duro] (*duro*, *duras*, etc) and waiting about two seconds, the user then sees a list like the following (Table 15).

This table shows the frequency of each collocate in each century (here just the 1200s-1700s are shown). For example, *ceruiz* occurs 11 times near [*duro*] in the 1200s and 4 more times in the 1400s. There are 140 total occurrences of *ceruiz* in the 1200s-1400s, and so the 15 cases near [*duro*] are about 10.7% of all tokens. This translates into a Mutual Information score of 6.22, which shows the relationship between the two words to be significant. Hence with the Corpus del Español and the Corpus do Português (which works exactly the same), we can do in 2-3 seconds what would take a month or more to do with CORDE.

TABLE 15
Corpus del Español: collocates of *duro* in 1200s-1400s

PALABRA	TOT	s13	s14	s15	s16	s17	s18	…	SEC1		TODOS	PERC	InfMut
DUREO	3			**3**					**3**		10	30.00	7.71
GOR	3			**3**					**3**		10	30.00	7.71
FORANOS	4			**4**					**4**		16	25.00	7.44
BLASFEMADOR	3			**3**					**3**		16	18.75	7.03
SOLIDOS	7			**7**					**7**		46	15.22	6.73
CALLOSA	4			**3**			1		**3**		20	15.00	6.71
PAPADO	5	**5**							**5**		34	14.71	6.68
MATER	39			**33**	2	1	2		**33**		258	12.79	6.48
ALCAZ	4	**4**							**4**		32	12.50	6.44
OBSTINADO	10			**4**	3	1	1		**4**		34	11.76	6.36
YERTO	7	**3**		**3**		1			**6**		52	11.54	6.33
GUJJARROS	4			**4**					**4**		36	11.11	6.27
CERUIZ	15	**11**		**4**					**15**		140	10.71	6.22

11. Semantics: more advanced

If we have a corpus architecture and interface that allows us to easily find collocates (as do the Corpus del Español and the Corpus do Português), we can then use this information in ingenious ways to examine semantic change. The basic idea is that if the words "nearby" a given word change over time, it may be because the word itself has changed meaning (or is at least being used in a different way). For example, the following table shows (to the left) the nouns that occur with [duro] in the 1900s but which are not very common in the 1800s, and (to the right) in the 1800s but not the 1900s (Table 16).

For example, *críticas* occurs near [*duro*] 27 times in the 1900s, but no times in the 1800s. *Ley*, on the other hand, occurs 23 times with [*duro*] in the 1800s, but only one time in the 1800s. Assuming *críticas* does occur in the corpus in the 1800s and *ley* does occur in the 1900s (and both are true), why does their frequency as a collocate with [*duro*] change so much from one century to another? Is it because the meaning of [*duro*] itself has changed slightly in some way?

To take another example, the following is a partial list of the adjectival collocates of *mujer* (and *mujeres*) in the 1900s and the 1800s (Table 17).

Notice how the adjectives from the 1800s (to the right) refer to the "moral virtues" of women, whereas these are almost completely absent in the 1900s. In the 1900s, on the other hand, they are much more prosaic, and simple refer to classifications that refer to nationality, employment, and so on. In this case, the corpus data provides interesting insight into the changing view of women in these two centuries.

Applied to Old Spanish and Old Portuguese, one could take a similar approach. Using the interface for the Corpus del Español or the Corpus do Português, one simply indicates what words or concepts are of interest, specifies the type of collocate (noun, verb, etc., if applicable), and then clicks once or twice more to show which two historical periods should be compared. Within two or three seconds, all of the relevant data is gathered and summarized. Using CORDE, on the other hand, searches like this would be either very difficult or impossible, since the CORDE architecture does not know how to find collocates.

12. Semantics: most advanced

With the right corpus architecture, it would be possible for users to search by semantic fields, rather than just searching for words and phrases. For example, in the case of the Corpus del Español and the Corpus do Português, powerful thesauruses are integrated into the corpus architecture. At the most basic level, this

TABLE 16
Corpus del Español: comparison of collocates of *duro*, 1800s / 1900s

+ 1900s / – 1800s						+ 1800s / – 1900s					
PALAVRA	OCCUR1	OCCOR2	PM 1	PM 2	PROP	PALAVRA	OCCUR2	OCCOR1	PM 2	PM 1	PROP
CRÍTICAS	27	0	1.18	0.00	11.83	MEDIO	30	1	1.30	0.04	29.59
LUCHA	10	1	0.44	0.04	10.14	LEY	23	1	0.99	0.04	22.69
MADERAS	19	0	0.83	0.00	8.33	NECESIDAD	23	1	0.99	0.04	22.69
CUELLO	8	1	0.35	0.04	8.11	PESETAS	18	1	0.78	0.04	17.76
LÍNEA	18	0	0.79	0.00	7.89	MILES	30	0	1.30	0.00	12.97
COMPETENCIA	15	2	0.66	0.09	7.60	SUERTE	13	1	0.56	0.04	12.82
MUNDO	7	1	0.31	0.04	7.10	ACENTO	12	1	0.52	0.04	11.84
ECONOMÍA	7	1	0.31	0.04	7.10						

TABLE 17
Corpus del Español: comparison of collocates of *mujer*, 1800s / 1900s

+ 1900s / – 1800s						+ 1800s / – 1900s					
PALAVRA	OCCUR1	OCCOR2	P/M1	P/M2	PROP	PALAVRA	OCCUR2	OCCOR1	P/M2	P/M1	PROP
MADURA	15	1	0.66	0.04	15.21	INFELIZ	68	2	2.94	0.09	33.54
MAYORES	11	1	0.48	0.04	11.15	SANTA	26	1	1.12	0.04	25.65
DIFERENTES	10	1	0.44	0.04	10.14	HONRADA	57	0	2.46	0.00	24.64
GORDA	19	2	0.83	0.09	9.63	GRIS	24	1	1.04	0.04	23.67
MARAVILLOSA	9	1	0.39	0.04	9.12	DIVINA	21	1	0.91	0.04	20.72
NACIONAL	9	1	0.39	0.04	9.12	DÉBIL	56	3	2.42	0.13	18.41
ARGENTINA	8	1	0.35	0.04	8.11	CELOSA	16	1	0.69	0.04	15.78
NORMAL	8	1	0.35	0.04	8.11	DESGRACIADA	15	1	0.65	0.04	14.80
TRABAJADORAS	8	1	0.35	0.04	8.11	VULGAR	15	1	0.65	0.04	14.80
INTERNACIONAL	17	0	0.74	0.00	7.45	DÉBILES	14	1	0.61	0.04	13.81
DISTINTAS	7	1	0.31	0.04	7.10	HONRADAS	31	0	1.34	0.00	13.40
CUBANAS	15	0	0.66	0.00	6.57	PIADOSA	12	1	0.52	0.04	11.84

TABLE 18
Corpus del Español: comparison of synonyms of *oscuro*

SINÓNIMO	TOT	…	s15	s16	s17	s18	s19	s20		ACAD	PER	FIC	ORAL		SEC1
OSCURO	2188		32	187	170	160	**863**	**749**		128	84	462	75		**1612**
HUMILDE	4503		69	933	1504	509	**1199**	**284**		25	60	143	56		**1483**
CERRADO	2068		43	408	284	163	**544**	577		78	186	184	129		**1121**
MISTERIOSO	1137			20	104	87	**692**	**234**		28	50	143	13		**926**
MODESTO	1038		25	79	122	126	**482**	**201**		19	57	93	32		**683**
CONFUSO	2082		73	441	670	204	**530**	**147**		16	23	98	10		**677**
SOMBRÍO	699			29	28	29	**519**	**94**		7	22	65			**613**
PARDO	1251		52	212	206	142	**287**	**275**		99	101	63	12		**562**
INCOMPRENSIBLE	494			49	26	45	**268**	**106**		7	26	63	10		**374**
DUDOSO	1091		40	267	278	188	**249**	**66**		8	21	29	8		**315**
NOCTURNO	409		6	30	39	29	**83**	**216**		34	28	103	51		**299**
INCIERTO	755		14	179	141	149	**148**	**124**		21	42	55	6		**272**
APAGADO	344		6	17	20	41	**151**	**106**		9	14	71	12		**257**
PESIMISTA	171						**66**	**105**		29	21	12	43		**171**
OPACO	190		4	7	12	29	**53**	**85**		16	18	45	6		**138**

allows us to find the historical frequency of all words relating to a particular concept. For example, a user might enter [=oscuro], and he could then see the frequency of all synonyms over time (and in different genres from the 1900s), e.g. (Table 18).

This partial listing of results shows, for example, that *modesto* and *sombrío* have decreased since the 1800s (per million words), whereas *pesimista* and *opaco* increased from the 1800s to the 1900s. In terms of the medieval periods, what would obviously be needed is some type of "historical thesaurus", which may of course never be available. But to the degree that it was, the corpus architecture could easily accommodate it.

In addition to looking for the frequency of single words, semantic information from thesauruses or user-defined, customized wordlists can be integrated directly into the query syntax. For example, in the Corpus del Español and the Corpus do Português, it is possible for users to create (via the web interface) customized lists of words, which refer to a particular semantic field of interest. Examples might be naval terms, words relating to emotions, a list of terms relating to family structure, or a list of words relating to a particular theological concept. This customized list of words can then be used as part of the query syntax. For example, if a user [andrés.gómez] creates a list of 100 words relating to "emotions" in Old Spanish, as well as another list of 70 words relating to "family relations" (*padre*, *hermanastro*, *nuera*, etc), he could then find every occurrence where a word in List 1 occurs nearby List 2. In this way, powerful semantically-oriented searches can be carried out on the corpus.

The Corpus del Español and the Corpus do Português are able to accommodate these types of semantically-oriented queries, because of the underlying architecture of the corpora, which is based on relational databases. With relational databases, it is possible to add any number of new datasets (thesauruses, user-defined wordlists, etc), and then integrate them in seamlessly into the query syntax. The architecture for CORDE, on the other hand, is not "open", and cannot be integrated into other datasets. Only single words or phrases can be searched for, but nothing approaching an entire semantic field or anything similar.

13. Conclusion

As has already been mentioned, those with a traditional training in philology might think of a corpus as simply a large collection of texts. In this view, the greatest care needs to be taken to ensure that the best texts have been selected, and that they are transcribed accurately. And in this view, once the textual corpus is completed, so is corpus as a whole –the architecture and interface are just an

afterthought. However, once one uses a full-featured corpus, one sees that this model is only half right– there is still so much to do after the texts have been assembled, to make sure that they are usable for a wide range of linguistically-oriented queries.

As we have seen, CORDE uses an older, now-outdated corpus architecture. This architecture relies on "off-the-shelf" Microsoft Indexing Services technology from the late 1990s, which was never designed for –and is completely inadequate for– most types of linguistic research. This technology was designed primarily to allow users to search for an exact word or phrase, and then view the entire document (as if one were reading a book or a series of documents). In this one particular type of search, CORDE does its job quite well. It is able to find all occurrences of a specific word or phrase, and then display these in context.

CORDE cannot, however, show whether the word or phrase was increasing or decreasing from one time period to another, or where it was the most frequent (which is quite useful information for philologists). Moving beyond simple word and phrase searches, we also find that it cannot search for all of the words and phrases that have increased or decreased between two time periods, which limits its usefulness for lexicographical research. It cannot search (well) by substring, which limits its useful for morphological research. It cannot search by lemma or by part of speech, which seriously limits its usefulness for syntactic research. And it cannot find collocates and it cannot incorporate information from other databases (such as thesauruses or user-defined wordlists), which seriously limits its usefulness for semantically-oriented research.

Although all of these queries are impossible with CORDE, they are both quick and easy with the Corpus del Español and the Corpus do Português. This is due to the fact that these corpora were designed with linguistic research in mind, rather than just as an afterthought.

Certainly, there are ways in which the Corpus del Español and the Corpus do Português can and should be improved. Although the architecture is "state of the art", in the case of the Corpus del Español particularly, there is certainly some work on the textual corpus that can and should be done to correct a few problematic texts (and the same is certainly true of CORDE as well). Some researchers may be aware that the Corpus del Español was created by just one person in less than a year and a half, and with very limited funds. It would therefore be a welcome change to have collaboration with other researchers with the philological expertise to help correct a handful of problematic texts. Yet even with this caveat, one should not ignore or minimize the value of the Corpus del Español and the Corpus do Português. As more than 120,000 unique users over the past six years have discovered, with these two corpora researchers can examine an extremely wide range of linguistic shifts in ways that are not possible with any other historical corpus.

References

DAVIES, Mark. (1994): "Parameters, Passives, and Parsing: Explaining Diachronic Shifts in Spanish and Portuguese", in: Beals, K., et al. (ed.): *Variation and Linguistic Theory*. Chicago: CLS. Vol 2, 46-60.

— (1995a): "The Evolution of Causative Constructions in Spanish and Portuguese", in: Amastae, John, et al. (ed.): *Current Research in Romance Linguistics.* Philadelphia: John Benjamins, 1995, 105-122.

— (1995b): "The Evolution of the Spanish Causative Construction", in: *Hispanic Review* 63, 57-77.

— (1996a): "The Diachronic Interplay of Finite and Nonfinite Verbal Complements in Spanish and Portuguese", in: *Bulletin of Hispanic Studies* (Glasgow) 73, 137-58.

— (1996b): "The Diachronic Evolution of the Causative Construction in Portuguese", in: *Journal of Hispanic Philology* 17, 261-92.

— (2000): "Syntactic Diffusion in Spanish and Portuguese Infinitival Complements",in: Dworkin, Steven/Wanner, Dieter (eds.): *New Approaches to Old Problems: Issues in Romance Historical Linguistics*, Amsterdam; Philadelphia: John Benjamins, 109-27.

— (2002): "Un corpus anotado de 100.000.000 palabras del español histórico y moderno", in: *SEPLN 2002* (Sociedad Española para el Procesamiento del Lenguaje Natural). (Valladolid), 21-27.

— (2005a): "Advanced research on syntactic and semantic change with the Corpus del Español", in: Pusch, Claus/Kabatek, Johannes/Raible, Wolfgang (eds.): *Romance Corpus Linguistics II: Corpora and Diachronic Linguistics*. Gunter Narr, 203-14.

— (2005b): "The advantage of using relational databases for large corpora: speed, advanced queries, and unlimited annotation", in: *International Journal of Corpus Linguistics* 10, 301-28.

— (2008a): "Relational databases as a robust architecture for the analysis of word frequency", in Archer, Dawn (ed.): *AHRC ICT Methods Network: Expert Seminar on Linguistics: Word Frequency and Keyword Extraction*. London: Ashgate.

— (2008b): "Spanish and Portuguese Corpus Linguistics", in: *Studies in Hispanic and Lusophone Linguistics* 1, 149-86.