New Directions in Spanish and Portuguese Corpus Linguistics

Mark Davies
Brigham Young University

Abstract

Within the past decade several large, freely-available online corpora of Spanish and Portuguese have become available. With these new corpora, researchers of Spanish and Portuguese can now carry out the same type of corpus-based research that has been done for other languages (such as English) for years. This includes advanced research on morphological and syntactic variation (thanks to full functionality with substring searches, part of speech tagging, and lemmatization), semantics and pragmatics (via collocates, synonyms, customized word lists, and word comparisons), and historical changes and synchronic register variation (via architectures and interfaces that allow easy comparisons of frequency in different sections of the corpus).

1. Introduction

This paper will provide an overview of recent corpus development and corpus-based research in Spanish and Portuguese. In Section 2 we will provide a very brief discussion of the methodology of corpus-based linguistics, and the interplay of this paradigm with the formalist or Chomskyan paradigm during the past three or four decades. Sections 3 and 4 provide an overview of what type of data one might obtain from a corpus, and how a corpus should be designed to answer research questions in the fields of lexical research, morphology, syntax, semantics, and historical linguistics. Sections 5-8 provide an overview of corpora and corpus-like resources that are currently available for Spanish and Portuguese, including offline resources, large online text archives, and online corpora. Section 9 considers how two of these corpora can be used to address a wide range of research questions. Finally, in Section 10 we provide a brief overview of recent corpus-based research in Spanish and Portuguese.

It may seem that we have focused too much on what constitutes a good corpus and what corpora are available, to the exclusion of what research has actually been carried out with these corpora. There are two reasons for this focus. First, most corpus-based research of Spanish and Portuguese that was published more than four or five years ago was based on small, proprietary corpora, many of which are now outdated. The majority of this research could, and probably should, be re-done with

Studies in Hispanic and Lusophone Linguistics
Volume 1, Issue 1
Spring 2008

the large, publicly-accessible corpora that are now available, and it is only by understanding the "state of the art" in terms of these corpora that such research can be carried out. Second, we recognize that many of the readers of this article may not have carried out (many) corpus-based studies themselves. By focusing on what is possible with some of the state of the art corpora that have recently become available, this will hopefully provide ideas for possible research by an entire new generation of corpus-based researchers of Spanish and Portuguese.

2. Corpus linguistics methodology

Before we discuss what one can do with a corpus, we should first consider a more basic issue – how the general methodology of corpus linguistics differs from that of Chomskyan or formalist linguistics, which is the model with which many readers will be most familiar. Perhaps the best overview of these methodological differences is found in McEnery & Wilson (2001; especially Chapter 1), and good discussions are also found in other introductory books on corpus linguistics such as Biber, Conrad & Reppen (1998), Kennedy (1998) and McEnery, Xiao & Tono (2006). In this section, we will briefly summarize some of the more important points from the overview in McEnery & Wilson (2001), focusing on the historical interplay between the empirical/corpus approach and the intuitive/formalist approach.

McEnery & Wilson first note that previous to the publication of Chomsky's *Syntactic Structures* in 1957, linguistics was very data-oriented and even corpus-oriented. Due to the early influence of linguists like Boas and Sapir, many linguists worked with "exotic" Native American languages, for which they, of course, had no native speaker intuitions. The only way to work with these languages was to obtain and carefully organize as much data as possible. In the 1930s-1950s, there was also strong influence on linguistics from positivism and behaviorism. The idea was that if linguists could collect enough data to model the input of a language learner (L1 or L2) they could also predict with some certainty the language development of that speaker. So again, collecting large amounts of data was quite important for most linguists up through the late 1950s.

This all changed with the Chomskyan revolution. Chomsky forcefully attacked many of the methodological underpinnings of previous corpus-based and empirically-based research. One of his main criticisms of corpus linguistics and similar models was that the linguistic databases that linguists had created were much too small to be of value. He showed that even in a one million word corpus, the data for many linguistic phenomena would be too sparse to be of interest or to provide insight into actual linguistic processing. He also argued that in many cases, corpora provided trivial insights into facts related to the real world (such as why *New York* would appear more frequently than *Dayton (Ohio)* in a corpus), but little insight into how humans produce and process linguistic data. Perhaps most

importantly, he argued strongly for the value of introspection. He said that it made little sense to go to the trouble of creating a corpus of a given language, when it would be possible to just sit down with a native speaker and quickly and easily get the relevant data – or to probe one's own intuitions, if the linguist was a native speaker of the language.

Due to Chomsky's critiques of certain types of empirical linguistics, data-based and corpus linguistics essentially went underground for the next 20-30 years. While there were occasional achievements such as the *Brown Corpus** (1961) and the *LOB corpus** (1970s) of English, most work in the field of corpus linguistics disappeared until the early to mid-1980s.[1] And yet this situation did not change as much for research on Spanish and Portuguese as it did for research on English, where the Chomskyan paradigm was more influential. Perhaps two brief examples will suffice. First, in the field of quantitative lexicography there was, of course, no alternative to using corpora, and this led to (what were for their time) very well-constructed frequency dictionaries, such as Juilland & Chang-Rodríguez (1964). More importantly – especially in terms of their relationship to later corpora – there was work on (what were for their time) large and very well-constructed corpora of Spanish and Portuguese, especially the *Habla Culta* project. This project, which was started in the early 1970s and which was spearheaded by Juan Lope Blanch (e.g. Lope Blanch 1993a, 1993b), led to the development of comparable corpora of "learned Spanish" from twelve different countries in Latin America and Spain, and similar corpora for Brazil were developed as part of the *Linguagem Falada* project.

And yet it was mainly in the 1980s that there was a resurgence of interest in corpus linguistics. Let us briefly consider some of the factors that led to this renewed interest. First, Chomsky was probably right in arguing that small, one million word corpora were often of little value. By the 1980s, however, advances in technology meant that it was now possible to create much larger corpora, such as 100+ million word *Bank of English** and the *British National Corpus**, and so his objection from 20-30 years earlier was not as relevant. In terms of his second objection – that corpus data tells us more about the real world than about linguistic knowledge – researchers were showing that these supposedly trivial real world insights were actually not very trivial at all, especially when it came to programming computers to process natural languages. Third, increasing numbers of researchers were arguing for a more nuanced model of grammar, in which stark binary judgments on grammaticality were replaced by models in which grammar was more an issue of tendencies. This fit in well with the quantitative approach of corpus linguistics, where real data is typically more "messy" than rarified data created from a linguist's own mind.

This leads us to the fourth – and perhaps most important – factor, which is the generative emphasis on the natural primacy of linguistic intuitions. For two or three decades, researchers had been arguing that complex theories were often built on questionable empirical data. Many researchers would simply shrug off problems

with their data, even when the data showed the theory to be fundamentally flawed. In many cases, it became a "he said / she said" scenario, in which opposing researchers could (and would) argue that in *their* dialect (or often their own personal idiolect), the data was as they claimed it to be – all in the attempt to defend their particular theory (see Fillmore 1992). As time went on, it became apparent that there needed to be some type of publicly-available standard – some common linguistic database – that could be used as a check on introspection. These publicly-available databases were in most cases the type of corpora that came to the fore in the 1980s and 1990s.

3. What can one do with a corpus?

Having very briefly considered the historical development of the field of corpus linguistics, let us turn now to the question of exactly what type of data one can obtain from a corpus. We will look at data as it relates to the lexicon, syntax and morphology, stylistics, language change, and language acquisition. This is just a brief overview – in subsequent sections, we will give more detailed examples from several fields of Spanish and Portuguese corpus linguistics.

When one thinks of corpus data, often the first thing that comes to mind are word counts. Assuming that the corpus is an adequate representation of the language as a whole – a point that we will later consider in more detail – we can use a corpus to determine which words are the most frequent in a given language. This data can then be used for a number of purposes, such as the development of materials for language teaching, or for use by computational linguists in helping machines to acquire (in a very reduced sense, of course) the knowledge of a native speaker. In terms of word frequency, corpora can show the distribution of words and phrases in different genres or registers, which can be useful information for lexicographers and foreign language teachers in helping to identify specialized vocabulary for particular domains such as medicine, legal language, or journalistic writing. In addition to frequency and distributional information, corpora can also provide data on collocates (i.e. which words occur most frequently with others) and this can provide valuable insight into the mental lexicon, semantics, and pragmatics. Finally, we can compare the collocates for related words (such as *pelo* and *cabello*, or *comenzar*, *empezar*, and *iniciar*) to tease out differences in meanings between these words.

Corpora can also be useful for research on syntax and morphology. On the one hand, researchers might take a large corpus of one genre or register and compare it to other genres – such as spoken, fiction, newspapers, and academic – and then attempt to determine why a whole range of constructions are more common in one set of registers than in another (cf. Biber, Johansson, Leech, Conrad & Finegan 1999). For example, why is the passive typically more common in academic texts, or the progressive in spoken language, or the perfect tenses in fiction? A different

approach is to view a corpus as a large "bag of words or sentences" that serve as a proxy for the entire language, and then search for one particular syntactic construction in this large linguistic database. Research on grammatical variation also interfaces with other types of data, such as the lexical information. For example, a large, robust corpus could show us which verbs occur mainly with either the imperfect or the preterit, which adjectives occur mainly with *ser* or *estar*, or which verbs or adjectives are the main subjunctive triggers in Spanish or Portuguese. Finally, corpora can provide useful morphological information, such as the gender distribution of words ending in –z (e.g. *la luz*, *el pez*), which words ending in *–tud* or beginning with *des-* are the most common (over time or in different genres), or which forms with particular roots (such as *-temp-* or *-cont-*) are most common (which, of course, ties back into the lexicon).

Corpus data is also of great importance in terms of understanding language change. Of course, we usually do not have access to the intuitions of native speakers from Spanish in the 1400s-1500s or Portuguese speakers from the 1600s-1700s, to know what was happening in the language in those periods. All we have available to us are corpora – texts from the particular periods. This diachronic data can be used to look at many different types of linguistic change – lexical, morphological, grammatical, stylistic, and so on. In terms of the lexicon, a corpus can show us, for example, which words are neologisms in the language or which have increased or decreased in frequency most during the last two hundred years. In terms of syntax, we can investigate in detail the historical trajectory of a given construction like clitic climbing (e.g. *lo quiere hacer* / *quiere hacerlo*), causatives (e.g. *fizola sentar__ > hizo que se sentara* or *la hizo sentarse*), or subject raising (e.g. *parece que María está enferma > María parece estar enferma*) (see Davies 2005a). We can also investigate semantic, stylistic, or cultural shifts in the language by examining changes in collocates, such as which words occur with *negro* or *duro* in different periods, or which adjectives are used with *mujer* in the 1600s-1700s compared to the 1900s-2000s.

Finally, corpus data can be used for language teaching. On the one hand, researchers can create what are known as learner corpora, which are corpora of learners of Language X by speakers of Languages A, B, and C, for example. Data from these corpora can help researchers move beyond anecdotal evidence to see precisely what types of errors or interlanguage are produced by the language learners. On the other hand, corpora can be very useful in helping to create teaching materials for a given language. The frequency and distributional data on words and phrases can be employed to create more useful textbooks and dictionaries, and materials for learning a particular type of language, such as medical Spanish or journalistic Portuguese.

4. Defining useful corpora

As we have seen, there are many different uses for corpora; and yet it is no simple matter to create corpora that can provide researchers with this wide range of data. In this section we will consider three aspects of corpus design and corpora that are integrally related to the type and value of data that they can yield – representativity, annotation, and architecture. In Sections 5-8, we will use these three aspects to critique the major corpora that are presently available for Spanish and Portuguese.

The first goal of corpus design is to ensure that the corpus is representative of what it purports to model. For example, if a corpus is advertised as a corpus of "general, synchronic" Spanish or Portuguese, then we would assume that it has adequate coverage of the major genres or registers, such as spoken, fiction, newspapers, and academic. (These are the four main registers used in Biber et al. 1999 and many other studies). Sometimes corpus creators forget this basic concept, and suggest that size can compensate for poor corpus design. A 200 million word corpus of Spanish newspapers may tell us a great deal about journalistic Spanish prose, but it may tell us little or nothing about spoken Spanish or literary Spanish from works of fiction. As with investment in stocks, it is also wise to diversify our corpus, so that a handful of texts do not skew the results. For example, if we create an ad-hoc historical corpus that is composed of only five or six texts from two different centuries, and if the data from just one of those texts is not representative of the historical period being studied, then our data and our findings will be of little value.

Even the best textual corpus, however, is of little value if the needed data cannot be extracted from the corpus. The second goal of good corpus design, then, is to annotate the corpus so that we will have access to the needed linguistic data. One basic type of annotation is document-level annotation. To the degree that the individual texts include information about author, time period, genre, and so on, we can potentially use that information as part of the query. For example, suppose that a researcher of Portuguese wants to find the most frequent collocates of a word like *mulher* or *verdade*, to see what insight these collocates provide into pragmatic and cultural influences in Portuguese. Even with a 100 million word corpus of Portuguese, the researchers could examine historical shifts and genre variation only if the individual texts were categorized accurately, and if the user could use that information as part of the search query. Without the necessary document-level annotation, a corpus is just a large blob of information, and provides little insight into the complexities of the language.

Equally as important, or perhaps even more so, than document-level annotation is word and phrase level annotation. This would involve, for example, marking up the text for part of speech and lemma (e.g. *decir* = *dice*, *dijeron*, *dirás*, etc.). Suppose that a corpus of Spanish were composed strictly of words and strings, with no linguistic annotation, and that researchers wanted to look for cases of the clitic

climbing construction (e.g. *quiere hacerLO* vs. *LO quiere hacer*). Without annotation, the researchers would have to perform a series of queries – for each combination of a form of a matrix verb like *querer* or *tener que* (*quiero*, *queríamos*, *tuviera que*, *tendrán que*, etc.) followed by an infinitive (*cantar*, *irse*, *derrumbarlo*, etc.). As one can imagine, because the words are not marked for lemma or part of speech, the researchers would have to carry out tens of thousands of individual queries to study this phenomenon, and this would take weeks or months. If the corpus is linguistically annotated for part of speech and lemma, on the other hand, this research would take perhaps five or ten seconds.

The third important aspect of corpus design is the architecture and interface. One can create the best textual corpus imaginable, and accurately annotate it at the document and word level, but the corpus might still be seriously limited by using the wrong software to index, organize, and provide access to the corpus (see Davies 2005b). A good example of this would be a typical text archive for a newspaper or magazine (such as the *Washington Post* or *Time* magazine) or even the Web (via a search engine like *Google*). With these archives, users can typically only search for words and phrases, and then see the entire article as part of the results set. On the other hand, users typically cannot search (either efficiently, or at all) for substrings (e.g. **tud* or **temp**), which means that the corpus cannot really be used to study morphology. If the corpus is not tagged for part of speech and lemma, or if the interface does not allow users to access this information, then it will be very difficult (if not impossible) for researchers to use the corpus for research on syntax. If the search algorithms cannot efficiently find the collocates for a given word (particularly in a large corpus), then it will be very hard to carry out pragmatic or semantically-based research. Without the right architecture and interface, they cannot compare across different lexical items (e.g. collocates of *esquina* compared to those of *rincón*) or across different time periods or registers (e.g. collocates of *agudo* in fiction and academic, or collocates of *pobre* in the 1600s and the 1900s). In summary, without the necessary architecture and interface, users are limited to just the most basic types of queries, and cannot carry out more meaningful research on lexical, pragmatic, semantic, morphological, and syntactic processes in the language.

5. An introduction to Spanish and Portuguese corpora and text archives

In the sections that follow, we will provide an overview of Spanish and Portuguese corpora and text archives that are available as of August 2007. As part of this survey, we will pay close attention to how easily the corpora can be used to address the range of research questions posed in the previous sections.

In these sections, we will focus on three types of resources. In Section 6 we will consider texts that can be downloaded from the Web or purchased on CD-ROM or DVD. Some of these texts are simply collections of literature, journalistic material,

or government documents, while others take a more systematic approach to creating a representative corpus of Spanish or Portuguese. The downside of these materials, however, is that they must be processed on one's own computer, with specialized software for corpus analysis. In Section 7 we will consider online text archives. In addition to being available online, these resources have the advantage of typically being very large (tens or hundreds of millions of words). The downside, however, is that the web interface and the architecture of text archives are typically so poor that they make it very hard to carry out useful linguistically-oriented research. In Section 8 we will consider online corpora, which potentially have the best of both worlds in terms of the two types of resources just mentioned. Like text archives, they are online and they are often quite large. Like downloadable corpora, they are more representative than a simple text archive. In addition, the best online corpora have architectures and interfaces that rival or surpass those used to process downloaded texts.

We should mention that – especially in the case of the downloadable texts and the text archives – the resources that we present are just a sampling. Certainly, by the time that this article is published, other valuable resources will already have become available. Yet while this list of corpora is not exhaustive, we believe that we have included the vast majority of the more well-known and useful resources that are presently available. Finally, we should note that the focus in this section has been on textual corpora, rather than collections of audio files for Spanish or Portuguese. Linguists who are interested in audio files may wish to obtain these from sources like the Linguistic Data Consortium* (LDC; see especially CALLHOME, CALLFRIEND, and the Spanish Broadcast News Speech materials), or consult the extensive *Linguateca** list of resources for Portuguese.

6. Offline resources

The first type of resources are texts that can be downloaded from the web or purchased on CD-ROM or DVD, and then processed on one's own computer with specialized software. These resources can be further divided into three groups. The first are collections of Spanish or Portuguese texts whose range is quite narrow and whose original use was for some other purpose (such as copies of newspaper or magazine articles). A second and related group of resources are multilingual or parallel texts (typically governmental or technical documents), which can be used to compare Spanish and/or Portuguese to other languages (perhaps for use in translation). The third and final group contains texts that form a more well-rounded corpus, and which may have in fact been created expressly for the purpose of creating corpora that could be used for linguistic analysis.

6.1 Collections of texts

All of the resources mentioned in this section can be downloaded from the Web (or purchased on CD-ROM or DVD), and then processed on one's own computer. In addition, a new trend is the creation of very simple web-based interfaces for some of these resources, much like the simple word and phrase searching that one can do with *Google*. These resources are considered separately from text archives (see Section 7), however, in that they *can* be processed on one's own computer, to perform more advanced linguistic analysis.

The first set of downloadable resources (or available on CD-ROM or DVD) are texts from earlier stages of Spanish and Portuguese. Perhaps the largest archive of older books from Spanish is the *Biblioteca Virtual**, which contains more than 30,000 books, most from the 1500s-1800s. In total, this archive probably contains several hundred million words of text. In addition to the *Biblioteca Virtual*, there are other large archives with Spanish and Portuguese books, including (both Spanish and Portuguese) *Project Gutenberg**, the *Oxford Text Archive*,* *Wordtheque**, and *Google Books**. For Spanish, there are also ADMYTE* (medieval texts), *Textos Lemir** and the Association for Hispanic Classical Theater*. For Portuguese, there are also *Projecto Vercial**, the *Biblioteca Digital da Porto Editora**, the *Acervo Digital** from the National Library of Brazil, and the *Biblioteca Digital de Literatura**. In terms of more modern texts, there are also collections of newspapers and magazines that are available on CD-ROM or DVD from organizations such as the LDC, the Evaluations and Language Resources Distribution Agency* (ELDA), and Elsnet*.

6.2 Multilingual texts

A second set of text archives contains text in Spanish and/or Portuguese and (typically) several other languages. These archives typically come from large governmental agencies like the European Union or the United Nations or from technical manuals, where the same document is translated into several languages. In the best case scenario, the texts are arranged in parallel format, where the text is aligned at the paragraph or even the sentence level, which makes them very useful as a database for translation. Some of the best multilingual corpora are the *JRC-Acquis Multilingual Parallel Corpus** and *EuroParl** (30+ million words of both Spanish and Portuguese from European Union documents), UN Parallel text (from the LDC), the PHP* and CRATER* archives (technical manuals), and the Par-C (Portuguese/English parallel corpus) and Comp-C (Portuguese/English comparable corpus) from the *Lácio-Web** project. In addition, it would also make sense to check the listings for the LDC, ELDA, and also the *Linguateca* website for additional resources.

6.3 More organized corpora

The previous two sections referred to texts that were not originally designed or created for linguistic analysis, but which might nevertheless be used for this purpose. This section discusses the handful of downloadable corpora that were designed specifically for linguistic analysis.

For Portuguese, these corpora include the *Corpus Informatizado do Português Medieval** and the *Tycho Brahe Parsed Corpus of Historical Portuguese**. In both cases, an attempt was made to create a corpus that sampled a wide variety of text types in each historical period. In addition, in the case of the *Tycho Brahe* corpus, nearly half of the texts have been tagged for part of speech, to further enable linguistic analysis. In addition to being available from the URLs provided, these two corpora are also part of the *Corpus do Português**, where they have been tagged and lemmatized, and where they are incorporated into the advanced architecture and user interface.

For modern Portuguese, downloadable corpora include *Lácio-Ref* (a very well designed corpus of nearly ten million words of Brazilian Portuguese that is part of the *Lácio-Web* project) and several downloadable corpora from the *Linguateca* website, including the *Floresta Sintáctica** (one million words of tagged Portuguese from both Portuguese and Brazilian newspapers), and the massive *CETEMPúblico** corpus (180 million words from the *O Público* newspaper in Portugal). These corpora from *Linguateca* can be downloaded and processed on one's own computer and are also searchable via the AC/DC portal at the *Linguateca* website.

For Spanish, the downloadable corpora include three very useful resources that were created at the Universidad Autónoma de Madrid in the early 1990s: the *Corpus Oral de Referencia del Español Contemporáneo** (one million words of spoken Spanish from Spain), the *Corpus de Referencia de la Lengua Española en Chile** (one million words of written Spanish from Chile) and the *Corpus de Referencia de la Lengua Española en la Argentina** (one million words of written Spanish from Argentina). In addition, one of the best collections of Spanish corpora are the several *Habla Culta* corpora. These corpora contain nearly 2,600,000 words from conversations by native speakers from numerous Spanish-speaking countries (Cuba, Puerto Rico, Mexico, Costa Rica, Venezuela, Colombia, Peru, Chile, Bolivia, Argentina, and Spain). At one time these corpora were available on CD-ROM from the Universidad de la Canarias, but apparently no longer. However, all of the *Habla Culta* materials and the three *Corpus de Referencia* corpora mentioned above are available as part of the *Corpus de Referencia del Español Actual** (CREA), from the *Real Academia Española*, as well as the *Corpus del Español**, where they have been tagged and lemmatized, and where they are incorporated into the advanced architecture and user interface. Similar materials are available from the *Linguagem Falada* corpus from Brazil, which includes texts from São Paulo, Rio de Janeiro, Recife and other cities.

6.4 Using downloaded texts and corpora

All of the resources mentioned in the previous three sections can be downloaded and processed one one's own computer. There are a number of useful tools that are available to use to analyze these texts. First, one may wish to annotate the corpus so that it can be searched by part of speech or lemma. There are many public-domain taggers and lemmatizers for both Spanish and Portuguese, which run on Windows and/or Unix/Linux machines. Rather than providing here a list of these tools (which would certainly be outdated within a year or two), it is suggested that users search through recent postings on listservs such as *CORPORA** to find such tools (search for *taggers* and *Spanish*, for example). Second, researchers will probably want to use a concordancer to search through and organize the data. For Unix machines, perhaps the best option is either the *Natural Language Toolkit**, or the *IMS Corpus Workbench** for more advanced searches of much larger corpora. For Windows, the most widely-used program is probably *WordSmith**, although *WordCruncher** is also a very good option and offers many features not available in *WordSmith*. For those with a background in programming, and who want to do much more advanced analysis of the data, *R* (see *R Project* in the appendix) is also a good option (available both for Windows and Unix). A fairly comprehensive listing of these concordancing and text retrieval programs can be found online at David Lee's website*.

7. Online text archives

The two main advantages of online text archives are that they can be very large, and they also tend to be very recent. It is common to have ten or a hundred million words of text from just the past five or ten years. The downside to using these resources is that they were not designed for linguistic research, and it is often very difficult to retrieve the needed data.

Perhaps the largest text archives are the archives of hundreds of millions or billions of words of text from recent newspapers, magazines, and academic journals. For example, *Lexis-Nexis** (menu: Academic Search / Non-English News / Spanish (or Portuguese)) has more than one hundred newspapers and magazines that can be searched at one time, and some of these individual sources on *Lexis-Nexis* (such as *CNNenEspanol* or *Folha de São Paulo*), have tens of millions or perhaps even hundreds of millions of words of text. Another database is *Fuente Académica** from EBSCO. This database, as the description states, "provides full text for 260 scholarly Spanish language journals. This multi-disciplinary database offers full text content to many academic areas including business and economics, medical sciences, political science, law, computer science, library and information sciences, literature, linguistics, history, philosophy and theology." Similar databases are available from ProQuest (*Research Library**) and Wilson Web (*Humanities Full*

*Text**). In addition to these text archives, there are, of course, also hundreds of Spanish and Portuguese newspapers and magazines available online, most of which can be searched at the individual websites for these publications.

7.1 Problems in using text archives for linguistic analyses

While the size and currency of the text archives is certainly an advantage, there are also at least three important limitations in terms of their use. First, virtually none of the text archives are lemmatized or tagged for part of speech, so it will be necessary to look for hundreds or thousands of exact words and phrases to do work on syntax. Second, the search engine for most of these archives do not allow substring searches (or else organize the output of such searches very poorly), which means that they cannot be used well for work on morphology. Third, virtually none of the interfaces for these text archives allow users to find collocates for a given word, which means that they cannot easily be used for semantic research. In summary, the text archives can be useful for counts of exact words and phrases, but are often problematic for anything more advanced than that.

Another important type of annotation that is often not made available to the researcher using a text archive is document-level annotation. For example, if a researcher is searching through 50 million words in a text archive of Portuguese magazines, it may not be possible to view the results by year or by geographical region (e.g. Portugal vs. Brazil), or register (e.g. financial reporting vs. popular culture). Without that fine-grained information, there is no way to know whether certain linguistic phenomena are on the increase or decrease, or how they are distributed across different dialects, or whether they represent formal or informal tendencies in the language. Even if it is possible to limit the queries to a particular time, dialect, or register, with a text archive it may be necessary to carry out the query many times – in each part of the archive – and then organize and analyze the results.

7.2 Web as corpus

In terms of the usefulness of text archives, let us briefly consider the issue of web as corpus. In a sense, the web is perhaps the ultimate text archive, and as such it has both the very best and the very worst in terms of the characteristics of text archives. First, it is, of course, extremely large. Exactly how large the web is in terms of Spanish and Portuguese texts is problematic. The best way to measure size is by finding the frequency of a word on the Web via *Google* and then comparing those frequencies to the frequencies from corpora of known sizes – such as CREA or the *Corpus del Español* (for Spanish) or *CETEMPúblico* or the *Corpus do Português* (for Portuguese). Using this method, we can give a rough, conservative estimate of

about 75 billion words for Spanish and about 45 billion words for Portuguese, both of which, of course, dwarf any available corpus.

But even more than a typical text archive, there is virtually no linguistic annotation, nor is there the ability to search by substrings. In addition, it is impossible to know how the web-based data might relate to register differences in the language, since there is usually no clear mapping between web domains and standard registers like spoken, fiction, newspaper, and academic. Finally, an often-overlooked aspect of searching the web for linguistic data is that the frequency data that is given to us by the search engine is often wildly off the mark in terms of actual occurrences (e.g. Véronis 2005). This is particularly true of search strings involving more than one single word, where *Google* essentially guesses the frequency, based on certain statistical tendencies. Thus it would be most unwise to uncritically use the web (perhaps via *Google*) as the basis for many types of research into linguistic phenomena in Spanish and Portuguese.

7.3 An example of the difficulty in using online text archives and the web as corpora

We can summarize the differences between a true corpus and a text archive by looking at a concrete example – clitic climbing in Spanish (e.g. *quiero hacerLO* vs. *LO quiero hacer*). With a tagged corpus (such as the *Corpus del Español*, the *Corpus do Português*, *Sketch Engine**, or the AC/DC and VISL/*CorpusEye** corpora) we can search for something like |||[querer] [vr]|| (i.e. any form of *querer* + the infinitival form of any verb).[2] In less than ten seconds, we find the 1000 most frequent matching strings (*les quiero decir*, *les querían recordar*, etc.). The corpus interface will sort and limit the hits, and it will show us the frequency and distribution of clitic climbing over time, between different registers, and between different verbs.

How would we carry out similar queries with the search interface for an online text archive? Because text archives are not lemmatized or tagged for part of speech, we would have to search for each form of *querer* (more than 40 forms) + all possible infinitives + clitics. These tens of thousands of queries would take weeks or months to carry out, and we would then have to start a new round of searches for *ir+a* and *esperar*. In perhaps a little less than one year, we would have the relevant data. We might take a shortcut, and look for just two or three forms of each matrix verb (e.g. *ir+a*, *querer*, *esperar*) followed by just three or four infinitival verbs (maybe *cantar*, *tener*, *dibujar*, etc.). Yet we would never know whether this small subset of verbs accurately reflects the entire range of actual strings in the text archive.

Even if we take the shortcut of using a handful of exact strings as proxies for all of the actual relevant strings, we still face serious problems with the text archive. They likely would not indicate whether the phenomenon is increasing or decreasing

over time, and whether there is a difference between different text types. In order to obtain this data, we would have to carry out separate searches in different historical periods and for different text types, and this might now involve several years worth of queries – simply to get the data that we could obtain from a well-organized corpus in less than one minute.

In summary, text archives (and the web) initially appear to be a very attractive option for corpus-based investigations. They are typically very large, and they often contain current texts. Yet they are so seriously limited in terms of representativity and the lack of annotation that they are of only marginal use for many types of linguistic research.

8. Online corpora

For many researchers who are just beginning to do research with corpora (as well as more advanced researchers), online corpora present a valuable type of research. First, like text archives they are easily accessible via the web – there is no need to download, install, and learn to use complicated concordancing tools or text retrieval programs. However, unlike text archives, some of these online corpora allow researchers to carry out advanced linguistically-oriented research on the texts. Rather than just searching for specific words and phrases and then being directed to a long list of matching documents, some online corpora have an architecture and interface that allow them to be used for a wide range of research involving morphology, syntax, and semantics.

In the section that follows, we will return to the four criteria for useful corpora already mentioned, three of which we presented in some detail in Section 4. To summarize, these are:

- size: useful corpora typically contains tens of millions of words of text
- representativity: the best corpora will contain texts from a wide range of genres
- annotation: the texts will be lemmatized and will be tagged for part of speech
- architecture and interface: it will be possible to search by substring (for morphology), lemma and part of speech (for syntax), collocates and synonyms (for semantics), and frequency in different historical periods and registers (for lexical research and for stylistics and historical linguistics)

8.1 Spanish

There are probably only a handful of online corpora that meet at least two of the four criteria mentioned above. They are:

- CREA
- CORDE
- VISL/*CorpusEye*
- *Leeds Collection of Internet Corpora* / *Sketch Engine*
- *Corpus del Español*

CREA and CORDE* (*Corpus Diacrónico del Español*) were created in the late 1990s and are freely-available from the *Real Academia Española*. CREA is the corpus of contemporary Spanish. It contains approximately 125 million words from 1975-1999, and about 90% of the corpus comes from written texts, while 10% comes from transcripts of spoken Spanish. In terms of geographical coverage, the corpus is evenly divided between Spain and Latin America. CORDE is the historical corpus. It also contains about 125 million words of text, with 21% of the texts from 1100s-1492, 28% from 1493-1713, and 51% from 1714-1974.

CREA and CORDE certainly fulfill the requirements of size, as well as representativity and breadth. For someone who wants to see concordance lines or distributional statistics (i.e. frequency in different countries, historical periods, and text types) for specific words and phrases from the largest online corpora of Spanish, these are arguably the best corpora available. Unfortunately, however, the creators of these two corpora chose as the basis of their search engine an extremely limited architecture. The interface and architecture do not allow for substrings (in any real sense) or collocates (although it is possible to see the most frequent word in one particular "slot" to the left or right of a given word). It is not possible to find all words or phrases that have certain frequencies in different historical periods or in different registers. Finally, the texts are not lemmatized or tagged for part of speech. In summary, the range of searches is more or less what one would find with *Google* Advanced Search or with online text archives. It is unfortunate that such a well-designed textual corpus has been coupled with the rudimentary search interface and corpus architecture, which greatly limits its usefulness.

The VISL/*CorpusEye** materials for Spanish combine a useful search engine with a somewhat eclectic set of corpora. In total, the Spanish corpora are 51 million words in size. They include 27 million words in the Spanish texts from *EuroParl* (government documents from the European Union), 22 million words from the Spanish version of *Wikipedia*, and a little more than one million words from the Spanish newspaper *El Diario Sur* (1991-92). Like CREA and CORDE, the corpus interface provides basic concordancing features – left and right context, ability to re-sort by contextual words, and frequency information for the words in particular slots to the left and right of the node word (i.e. the word searched for). What sets it apart from the Real Academia Española corpora, however, is that it is lemmatized and tagged for part of speech, allowing queries like |||[pos="N"] [lex="perfecto"]|||, which would yield *pareja perfecta*, *cuadrados perfectos*, etc. Unlike CREA and CORDE, however, there is no really useful distributional information (i.e. frequency

in different sections of the corpus), which probably makes sense, since the textual corpus itself is mainly a collection of disparate public-domain texts.

The corpus for the *Leeds Collection of Internet Corpora** and *Sketch Engine* is essentially the same corpus. It is composed of approximately 117 million words of text from hundreds of thousands of web pages. The corpus has been annotated using *TreeTagger**, and has been lemmatized and tagged for part of speech. In both cases, the corpus uses the *IMS Corpus Workbench* as its architecture. The *Leeds* interface allows complex searching by substring, part of speech, and lemma, and the *Sketch Engine* interface allows an even wider range of queries. With *Sketch Engine*, users can find the frequency of all words and strings that match a certain pattern, they can search for collocates (any number of words to the left or right of a given word), and they can see a "word sketch", which shows, for example, the most common subjects, objects, adjectives, or adverbs for a given lemma. Finally, users can see a word sketch difference display that compares the word sketches for any two lemma. This can be very useful for language learners, for example, as they attempt to learn the differences between two competing words.

In many respects, CREA/CORDE and the *Leeds*/*Sketch Engine* corpora are polar opposites. CREA and CORDE are wonderful textual corpora, and cover extremely well the range of text types in different registers and historical periods of Spanish. Yet they are seriously limited because of an overly-simplistic architecture and interface. The *Leeds*/*Sketch Engine* corpora, on the other hand, are annotated and they employ a very powerful interface and architecture. Their drawback, however, is the textual corpus – it is composed strictly of web pages. To the degree that fiction, or spoken, or academic Spanish is different from what would show up on a typical web page, these differences will not be represented in this corpus.

The final online corpus of Spanish that we will consider is the *Corpus del Español*, which has been freely available online since 2002. In terms of the criteria that we mentioned above in Section 4, it is large (100 million words), it contains a wide range of text types and genres (for example, the 20 million words from the 1900s are evenly divided between spoken, fiction, newspaper, and academic texts), and it is well annotated for part of speech and lemma..

The corpus also has a very robust architecture and interface. As we will see in the following section, the *Corpus del Español* allows users to search for words, phrases, and substrings (for lexical research and morphology), part of speech and lemma (for research on syntax), collocates, synonyms, customized lists, and word comparisons (for research on semantics and pragmatics). In addition, due to the unique relational database architecture employed by the corpus, it is possible to use frequency as part of the query. As a result, users can find all words, phrases, substrings, and so on that have a given frequency in different historical periods or in different registers of modern Spanish. This allows users to easily examine register variation and historical change (e.g. different word senses in different registers or historical periods).

8.2 Portuguese

As with Spanish, there are probably only a handful of Portuguese corpora that meet at least two of the four criteria mentioned above (size, representativity, annotation, and architecture and interface). These corpora are:

- AC/DC corpora
- VISL/*CorpusEye*
- *Corpus de Referência do Português Contemporâneo*
- *Sketch Engine*
- *Leeds Collection of Internet Corpora*
- *Corpus do Português*

One of the best sites for Portuguese corpora is the AC/DC corpus interface at the *Linguateca* project. The vast majority of the corpora – more than 97% – (as measured by size in words) come from newspapers. These include the huge *CETEMPúblico* corpus (192 million words from *O Público* (Pt), 1991-98), CHAVE (90m words, *O Público* (Pt), 1994-95; *A Folha* (Br) 1994-95), *NILC*/*São Carlos* (30m, mainly newspapers from Brazil), *Avante!* (6m, *Avante!* (Pt), 1997-2002), *DiaCLAV* (6m, regional newspapers from Portugal, 1999-2000), *Natura*/*Público* (6m, first two paragraphs from *Público* articles), CONDIVport (2m, sports articles from Portugal and Brazil), and *Natura*/*Minho* (1.6m, *Diário do Minho* (Pt)). The 3% of the textual material that comes from non-newspaper sources are from literary texts from previous centuries – *Vercial* (8.4 million words from *Projecto Vercial*; mainly literary texts from the 1800s-1900s), and *Clássicos LP* (1.3 million words from *Porto Editora*; mainly literary texts from the 1600s-1800s).

Probably the two best features of this site are the size of these texts and the fact that they are lemmatized and tagged for part of speech, which makes them useful for research on syntax. For example, one can search for [pos="N"] [lema="perfeito"] and see results like *passes perfeitos*, *forma perfeita*, etc. The output display can be concordance (perhaps hundreds or thousands of results, with node word surrounded by limited context to the left and right), frequency lists by exact word form, lemma, part of speech, morphological information (e.g. person, number), and frequency in parts of the corpus (e.g. different years of a given newspaper).

As with every corpus (or corpus architecture and interface), there are some limitations with the AC/DC materials. The first is the corpus itself. As mentioned, it is nearly all newspaper language, with very little fiction, no academic, and no spoken. Therefore, one might get a skewed view of the overall language from this one narrow genre. In addition, each one of the corpora has to be searched separately. There is really no way to compare the frequency in different dialects or registers without running the same queries one after another in the different corpora, and then tabulating the results in another program.

While the word-level annotation is very good (e.g. part of speech and lemma), the search interface only partially makes use of this information. In the results set for a multi-slot query (e.g. the two slots in |||[pos="N"] [lema="perfeito"]||, only the most frequent words for one of the two slots/words is returned. More problematic is the fact that it is not possible to click on an entry in a rank-ordered listing of forms and then see those forms in a keyword in context (KWIC) display. One can either see frequency listings or search for exact strings, but these are two separate, unconnected queries. In addition, it is not possible to search for collocates or synonyms (thus limiting semantically-oriented queries) or to limit queries by frequency in different sections of the corpus.

Another corpus site that is very similar to the AC/DC site at *Linguateca* is VISL/*CorpusEye*, and this is due to the fact that Eckhard Bick (the creator of VISL/*CorpusEye*) is also the person who tagged the materials for AC/DC. In most cases, then, they are the same corpora with different interfaces. As with the Spanish VISL/*CorpusEye* corpora, the Portuguese corpora include *EuroParl* (EU documents) and *Wikipedia* in Portuguese. They also contain about 200 million words from *CETEMPúblico* and about 25 million words from the *Folha de São Paulo* newspaper. As with the Spanish corpora, one has access to basic concordancing features and can also search by lemma and part of speech (e.g. |||[pos="N"] [lex="perfeito"]|| = *crime perfeito*, *crianças perfeitas*, etc.). In addition, with Portuguese it is possible to search a "treebank" of about 170,000 words to find more complex syntactic structures, including things such as noun phrases, different types of subjects, intransitive verbs, and so on. In summary, for someone who is interested in syntactic research on very large Portuguese corpora, and who doesn't mind being limited primarily to one particular register (i.e. newspaper Portuguese) the VISL/*CorpusEye* materials are a good option.

There are also multiple corpora from the *Corpus de Referência do Português Contemporâneo** project, which are available from a search engine at that site. These corpora include the untagged RL Corpus and the ELAN corpora (8.7 million and 2.8 million words, respectively, from newspapers, magazines, and technical books), a small 500,000 word tagged portion of the preceding corpora, and a three million word, untagged corpus of Portuguese from Lusophone countries in Africa. Unfortunately, the corpus architecture and search interface are extremely limited (being related as they are to the same ones used for CORDE and CREA). Users are limited mainly to KWIC displays for a given exact word or exact phrase, and (unlike CORDE and CREA) they cannot see distributional information for the frequency in different parts of the corpus.

The Portuguese corpora that are available from the *Leeds Collection of Internet Corpora* and the *Sketch Engine* corpus are similar to what is available for Spanish. In both cases, the architecture and interface are exactly the same as for the Spanish corpora; the only difference is in the textual corpus itself. In the case of the *Leeds* corpus, it comes entirely from web pages. While no mention is made on their

website of the size of the corpus, some preliminary frequency probes suggest that it is probably 50-60 million words. In the case of the *Sketch Engine* corpus for Portuguese, it is not based on the *Leeds* web materials (as with Spanish), but rather it is based on the *CETEMPúblico* and *Cetenfolha* materials from the *Linguateca* project, which have been tagged by Eckhard Bick (see above).

The final online Portuguese corpus that we will consider is the *Corpus do Português*, which has been freely-available online since 2006. In terms of size, it contains 45 million words, including 20 million words from the 1900s, 10 million from the 1800s, and 15 million from earlier centuries. For the 20 million words from the 1900s, it has 2 million words from spoken, 6 million from fiction, 6 million from newspapers and magazines, and 6 million from academic. In addition, it is divided evenly between texts from Portugal and Brazil, both overall and for each of the four registers just mentioned. As with the *Corpus del Español*, the *Corpus do Português* is fully annotated, having been lemmatized and tagged for part of speech. In addition, because its interface and architecture are almost exactly the same as the *Corpus del Español*, it allows the same wide range of queries, including word, phrase, substring, part of speech, lemma, synonyms, customized lists, word comparisons, collocates, and frequency-based queries (by historical period, dialect, and register).

9. Examples of possible queries and research with full-featured corpora

In Sections 3 and 4 we discussed some of the types of research that should be possible with well-designed corpora. In Sections 5-8 we provided an overview of the major corpora that are available for Spanish and Portuguese. We focused on both the textual corpus (whether they were mainly newspapers, or whether they had texts from a wide range of genres) and on the architecture and interface (which relates to the range and types of queries).

As we have seen, only the *Corpus del Español* and the *Corpus do Português* have a wide range of genres (as well as historical texts) and also allow a wide range of queries. In addition, they are freely-available online, which might be important for those who want to retrieve data quickly and easily and do not want to create their own corpora from scratch. In this section, then, we will focus on how these two corpora can be used to address the wide range of research questions that we posed in Sections 3 and 4.

Although these two corpora will serve as a point of departure for most of our discussion, we will at times consider how similar data might be obtained from other corpora. Finally, we should remember that because the architecture and interface (and even the corpus design) of the *Corpus del Español* and the *Corpus do Português* are very similar, nearly every query that is possible with one is possible with the other. To simplify our discussion here, however, the majority of the examples will come from the *Corpus del Español*.

9.1 Basics of the corpus interface

There are three parts to the interface for the *Corpus del Español* and the *Corpus do Português*: the search form, the frequency and distribution window, and the KWIC window. By selecting [CHART] in the search form, users will see a bar chart indicating the frequency of a given word, phrase, substring, or grammatical construction in each century (12-19 in Figure 1; 1200s-1900s), and in each register from the 1900s (AC, NW, FC, SP in Figure 1 for academic, newspaper, fiction, and spoken; more complete abbreviations are used on the web interface). The chart also shows the number of tokens (i.e. occurrences) in each section, the size of that section (in millions of words), and the normalized frequency of tokens per million words. For example, suppose that we enter ||des*miento|| into the search form of the *Corpus del Español*. The resulting bar chart shows a gradual increase in the overall frequency of these forms since the 1300s, and it shows that these morphologically complex forms are more common in the more formal registers of Spanish.

CENTURY	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**		**AC**	**NW**	**FC**	**SP**
													
PER MIL	**78**	**33**	**29**	**43**	**40**	**54**	**56**	**76**		**135**	**92**	**50**	**26**
SIZE (MW)	7.9	3	9.7	20	14.8	11.5	23.1	22.8		5.7	5.7	5.7	5.7
FREQ	617	99	289	867	599	623	1312	1745		776	529	289	151

Figure 1. Frequency bar chart from the *Corpus del Español*

With these bar charts, users can click on any of these bars to see the most frequent forms from that century or register. For example, by clicking on the bar for the 1700s, users would see that the most frequent forms from the 1700s are *descubrimiento*, *desabrimiento*, *descaecimiento*, *desfallecimiento*, *desvanecimiento*, etc.

In the preceding query, we arrived at the frequency table, which shows the frequency of each individual form, by clicking on part of the frequency chart (Figure 1). More commonly, however, users will simply select [TABLE] in the search form to see this table directly. For example, by re-doing the query ||des*miento|| and selecting [TABLE], users will see the frequency of each matching string (*descubrimiento*, *desfallecimiento*, etc.) in each century from the 1200s-1900s,

as well as in each of the four registers from the 1900s (Table 1). We can also rank order the results by the frequency in a set of time periods or registers, and we can also see subtotals for the frequency of each form in a given set of centuries and/or registers.[3]

	WORD	12	13	14	15	16	17	18	19	AC	NW	FC	SP	TOT
1	DESCUBRIMIENTO			13	515	336	444	406	936	495	247	123	71	2650
2	DESENVOLVIMIENTO						1	253	31	4	17	1	9	285
3	DESABRIMIENTO			2	80	78	33	78						271
4	DESCONOCIMIENTO		1	12	15	7	3	88	84	7	41	16	20	210
5	DESPLAZAMIENTO							1	196	111	46	26	13	197
6	DESPRENDIMIENTO					1	9	87	55	12	22	16	5	152
7	DESVANECIMIENTO			2	15	60	12	47	16	1	2	13		152
8	DESTRUYMIENTO	89	17	15										121

Table 1. Frequency table from the *Corpus del Español*

To see the actual KWIC listing, users simply click on the word or phrase in the column to the left, or, to see just the entries from a particular century or register, they would click on the number in the appropriate column. For example, a user who clicks on ||desplazamiento|| from the 1900s-Newspapers would be able to page through the 46 KWIC entries like those in Table 2. Users can click on any title to see even more context, and to get detailed information on the text. As far as we are aware, the *Corpus del Español* and the *Corpus do Português* are the only two corpora of Spanish or Portuguese that allow users to seamlessly move from frequency and distribution charts to KWIC, as shown above. With all other corpora, one can see a frequency listing, but it would then be necessary to re-do the search to see the word(s) in context.

1	19-N	Arg:Prensa:75_ESCR	mantenidas por el rectorado de la Uba para evitar el **desplazamiento** del doctor Oscar Shuberoff, por iniciativa de varias facultades encabezadas por Veterinaria
2	19-N	Bolivia:ERBOL:04...	huelgas, bloqueos, etc. Por lo tanto, **desplazamiento** de la fuerza ejercida por las multitudes. La riqueza de lo
3	19-N	Col:Semana:821	fenómeno han concluido que los paramilitares son los generadores del **desplazamiento** de campesinos a los grandes centros urbanos. Y algo peor: que
4	19-N	Col:Semana:826	del país. Lo hicieron vía aérea. Para el **desplazamiento** no se dejó nada al azar. Los guerrilleros fueron divididos en cuatro

Table 2. KWIC (Keyword in Context) from the *Corpus del Español*

9.2 Lexical

Having now briefly considered the corpus interface, let us turn to a particular area of corpus-based research – the distribution and frequency of lexical items. The *Corpus del Español* and the *Corpus do Português* provide summaries (via tables and bar charts) that show the normalized frequency (per million words) in each century from the 1200s-1900s, as well as in each of the four main registers (i.e. spoken, fiction, newspaper, and academic) from the 1900s. In addition, with the *Corpus do Português*, it is possible to compare the frequency of usage in the two main dialects (i.e. Portugal and Brazil) and this will soon be added for the *Corpus del Español* as well (i.e. Spain and Latin America). This frequency information can be of value in creating frequency dictionaries for language learners, such as Davies (2005c) and Davies & Preto-Bay (2007).

As an example, take *lúgubre* (or *lugubre*) in the *Corpus del Español*. After entering the word and selecting [CHART], the user sees a chart like Figure 1 above. The chart would show that the frequency of *lúgubre* increased from .31 occurrences per million words in the 1400s (3 tokens total), to .46 in the 1500s (9), 1.28 in the 1600s (19), 3.31 in the 1700s (38) and hit its peak at 14.09 in the 1800s (326). In the 1900s it is back down at the 1700s level – 3.02 tokens per million words (or 69 tokens in the 20 million words from the 1900s). Comparing the different registers/genres from the 1900s, we see that *lúgubre* is used much more in fiction than in any other register. It occurs 10.7 times per million words in fiction (61 tokens), but only .87 in newspapers (5), .53 in academic (3), and it does not occur at all in the five million words from spoken.

In addition to showing the frequency and distribution of a given word or set of words, one of the real strengths of the *Corpus del Español* (and the *Corpus do Português*) vis-à-vis other corpora – in terms of lexical data – is the ability to use frequency information as part of the query, which is something that we will return to in Section 9.6. The *Corpus del Español* and the *Corpus do Português* are the only two corpora – among all of those that we have surveyed – that allow one to find all words that occur only in a given date range or in a given text type, or with a given frequency in different parts of the corpus (e.g. words that decreased significantly from the 1500s-1600s, or which occur much more in fiction than in academic). These two corpora can produce such a list, because the relational database "knows" the frequency of each word and phrase in the different time periods and registers, and can use this as part of the query. Because of the non-relational database architecture of the other corpora, however, this is not possible. These other corpora can only provide tokens for the particular word(s) or string(s) that the user manually inputs into the search interface.

Aside from the *Corpus del Español* and the *Corpus do Português*, the only other corpora that show the frequency of use for a given word or phrase in a number of different genres and registers (or historical periods) are CREA and CORDE from the *Real Academia Española*. To repeat the query shown above, we can input the word *lúgubre* in the CREA corpus (1975-present) and then see its distribution over the past decade or two. The data show, for example, that the year with the most tokens is 1995 (24 tokens, or 7.79% of all tokens), then 1991 (21; 6.81%), and then 1996 (21; 6.81%), and so on. One can also obtain data on the frequency in different countries. For example, CREA shows that 59% of all tokens (or 201 tokens) are in texts from Spain, followed by Argentina (30 tokens; 8.8%) and México (25 tokens; 7.4%). Finally, one can see the distribution in different text types. For example, 65.2% (227 tokens) are from fiction, 14.5 (51 tokens) from *ciencias sociales*, *creencias y pensamiento*, and so on.

The problem with the distributional data for date, country, and text type, however, is that it is not normalized. In order for the number of tokens to be meaningful, we need to know how big the corpus is by date, country, and text type. There may be more tokens of *lúgubre* from 1995, or in texts from Spain, or in fiction, just because each of those categories have more words overall, and so we would expect to find more of *anything* (including *lúgubre*) in that sub-corpus. The strange thing is that this information on corpus size by date, country, and text type is available in one of the help files on the corpus website. But it is not included as part of the web interface and results display, which makes comparisons between dates, countries, and text types either difficult or meaningless. Where this information is more useful, however, is in the ability to limit queries to just a particular year, country, or genre/text type. For example, if we wanted to, we could limit the query for *lúgubre* in CREA to fiction from countries in South America from 1975-1985, and this would correctly limit the query to show the 26 tokens.

9.3 Morphology

With the *Corpus del Español*, the *Corpus do Português*, and *Sketch Engine*, and most of the AC/DC corpora, it is possible to search by substring. For example, with the *Corpus del Español* and the *Corpus do Português*, one can search for ||*tud|| (to find all words ending in *-tud*) or ||la *z|| (the most frequent feminine nouns ending in *–z,* as measured by –z nouns that are preceded by the definite article *la*). With these two corpora, even more complex queries are possible, such as ||des*.[vr]|| (i.e. infinitival forms of verbs that begin with *des-*: *descubrir*, *despertar*, *desarrollar*, *descansar*) or ||re*able.[j*]|| (i.e. adjectives that begin with *re-* and end with *able*: *respetable*, *responsable*, *rentable*, *recomendable*, etc.). We can also carry out lemma-oriented queries, such as |||[cair]|| in the *Corpus do Português*, which would show the frequency for all forms of the verb *cair*: *cair*, *caiu*, *cai*, *caindo*, *caía*, etc. Finally, on the morphology-syntax interface, we could search for the most common forms of a given part of speech, such as |||[vii@3p]|| (i.e. third-person plural forms of the imperfect: *eran*, *habían*, *estaban*, *tenían*, *iban*, *podían*, etc.). As far as we are aware, only the *Corpus del Español* and the *Corpus do Português* allow queries like this, which combine word form, part of speech, and lemma.

9.4 Syntax

With the *Corpus del Español*, the *Corpus do Português*, *Sketch Engine*, and the AC/DC and VISL/*CorpusEye* corpora, it is possible to search by part of speech. As an example of the value of tagged corpora, let us briefly return once more to the clitic climbing construction (e.g. *LO quiero hacer*, *ME pueden hablar*). In the *Corpus del Español*, the query string |||[ppo] [querer] [vr]|| (i.e. object pronoun + form of *querer* + infinitive) finds the results in just five or six seconds: *te quiero decir*, *me quiere decir*, *me quiero ir*, *le quiero decir*, *lo quiero decir*, etc. As we have already mentioned, with untagged corpora like CREA and CORDE, syntax-oriented queries like this are either very difficult or even impossible. Because these corpora do not know what words are forms of *querer* and because they do not know what a pronoun or an infinitive is, users would have to search – one by one – for all combinations of a pronoun followed by a form of *querer* followed by any one of thousands of infinitival forms of verbs, which would take weeks or months to carry out.

As mentioned above in Section 9.1 (and as shown in Figure 1) with the *Corpus del Español* and the *Corpus do Português* it is also possible to see the overall frequency of any item. This can be particularly useful in looking at the frequency and distribution of syntactic constructions. For example, after selecting [CHART], one could see a bar chart (cf. Figure 1) showing the overall frequency of the "*ser* + passive" construction in Spanish (i.e. all forms of *ser* + a past participle: *fueron descubiertos*, *fue llamado*, etc.) with the query string |||[ser] [vps*]||. The chart

would show that the use of the passive has increased every century since at least the 1500s and (perhaps more importantly) that the passive is most common in the more formal registers, where the agent is de-emphasized and the emphasis is placed on the process that was carried out (e.g. *el experimento fue llevado a cabo por …*). Nothing like this charting feature (to compare the overall frequency of a construction in different time periods and in different registers) is possible with any other corpus of Spanish or Portuguese.

In terms of useful corpora for studying Spanish syntax, we should also note an alternative interface for the *Corpus del Español**. This architecture and interface is based on the results from a large study to examine register variation in Spanish, which was carried out in 2002-2004 (see Biber, Davies, Jones & Tracy-Ventura 2006, Davies 2007). There are two main types of queries available for users of the site. With the first type, users can select any one of 120 different syntactic features (e.g. passive, relative pronoun, clefting, QU-question) and then see the rank-ordered frequency of this feature in twenty different registers (informal conversation, interviews, fiction, editorials, business letters, etc.).

Alternatively, users can select any two of the twenty registers, and then see which of the 120 syntactic constructions exhibit the most difference between the two registers. For example, a user can use the search interface to compare [Informal Conversation] and [Encyclopedias], and s/he would see that tag questions, yes/no questions, first person pronouns, and exclamations are much more common in conversations. None of these are probably too surprising, since these two registers are very different from each other. But even with more similar registers, there are interesting differences. For example, compared to [Formal Conversation], the [Debate] register has about twice as many cases (per thousand words) of [haber+de/que] (e.g. *hay que hacerlo*), [ser+ADJ+que+INDIC] (e.g. *es obvio que no saben*), or [el+que+SUBJUNC] (e.g. *los que lo critiquen no lo conocen*), all of which in fact do represent more debate-oriented syntactic phenomena.

9.5 Semantics

9.5.1 Collocates

There is a saying in corpus linguistics that "you can tell a lot about a word by the other words that it hangs out with." It is therefore important that the corpus architecture and interface allow users to find the most frequent collocates for a particular word. The only corpora that allow this are the *Corpus del Español*, the *Corpus do Português*, and *Sketch Engine*.

There are two ways of searching for collocates in the *Corpus del Español* and the *Corpus do Português*. First, users can search for the most frequent words in a particular position before or after the node word. For example, the query ||[n*] [verde]|| will find the most frequent nouns immediately before a form of the lemma

verde, and the results would be *ojos*, *color*, *partido*, *luz*, *terciopelo*, *pájaro*, *hojas*, *algas*, and *revolución*. Likewise ||[rostro] [j*]|| will find the most common adjectives immediately after a form of *rostro*: *pálido*, *encendido*, *humano*, *lívido*, *hermoso*, *descompuesto*, *serio*, *bello*, and *oculto*.

Often, however, in order to understand the meaning and use of a word, it makes more sense to extend the "collocation window" out to five or ten words to the left and to the right of the node word. For example, if we search for exact strings for ||[n*] [hondo]|| (i.e. all nouns immediately before a form of *hondo*), the only exact string that occurs five times or more in the 1800s and 1900s is *tapete hondo*. If we select [CONTEXT] in the search form, however, and extend the collocation window to five words to the left and to the right, then there are 123 different nouns that occur five times or more in this window, with the most common being *raíces*, *alma*, *pecho*, *corazón*, *suspiro*, *abismo*, *silencio*, and *impresión*. Similarly, the most common adjectives within five words of ||rostro|| are *pálido*, *encendido*, *hermoso*, *bello*, *lívido*, and *humano*.

When no part of speech for the collocates is selected, the most common collocates are, of course, the most common words in the language. For example, the most frequent collocates of *rostro* (discussed above) are *el*, *de*, *y*, *su*, and *en* – which don't provide us with much information about the meaning of *rostro.* In this case, users can choose to have the results sorted by Mutual Information score (MI), which takes into account the overall frequency of the collocates in the corpus and tends to eliminate the high frequency "noise" words. In the case of *rostro*, the MI-ranked collocates would then be *angel*, *maquillado*, *desencajado*, *cubriéndose*, *atezado*, *inexpresivo*, *ovalado*, and *coloreó*. In the case of *hondo*, the MI-ranked collocates would be *respirar*, *quiebras*, *cañada*, *pozos*, *raíces*, *tapete*, and *cavernas*.

Collocates can also be very useful in terms of understanding the meaning and use of a given word in different registers. For example, in English the collocates of *chair* in fiction would be words like *table*, *back*, *leg*, whereas in academic texts they would be *committee* or *professor*. In *Sketch Engine* it is difficult to search for collocates within a particular register, but it is quite easy with the *Corpus del Español* and the *Corpus do Português*. Let us look at one brief example from the *Corpus do Português*. In Portuguese, *fundo* means *deep*, *bottom*, *end*, or *fund*. In the interface for the *Corpus do Português*, a user can enter ||[fondo]|| as the node word, choose [Fiction] for the first section, [Academic] for the second section, and then select [CONTEXT] to see the surrounding words. The user then sees the most frequent collocates in fiction but not in academic (i.e. *casa*, *sala*, *olhos*, *corridor*, *quintal*, and *rua*, which show the meaning of *deep* or *bottom*) and academic but not fiction (i.e. *desenvolvimento*, *radiação*, *empréstimo*, *país*, *dólares*, *milhões*, and *participação*, which show the meaning of *funds*).

Finally, collocates can also show changes in meaning or usage over time. For example, in English the word *engine* might appear most commonly near *car* or *train* before the mid-1990s but near *search* or *Google* in the 2000s, which shows that

engine is being used with a different sense in more recent texts. With the *Corpus del Español* and the *Corpus do Português*, we can easily compare the collocates in two competing sets of historical periods. For example, returning to the example with *fundo* in the *Corpus do Português*, we can select [1900s-All] for the first section and then [1600s, 1700s, 1800s] for the second section. We then see that *braças*, *fogo*, *rapaz*, *natividade*, *cortina*, and *mariquinhas* are used much more in the 1600s-1800s than in the 1900s, and these collocates refer primarily to the meaning of *deep*, *bottom*, or *end*. In the 1900s, on the other hand, the most frequent collocates are *estabilização*, *garantia*, *recursos*, *participação*, and *coesão*, which refer primarily to the meaning of *funds*.

In addition to semantic change, a change in the collocates with a certain word can serve as important indicators of cultural or social shifts. For example, if we compare the collocates of *mulheres* in the *Corpus do Português* in the 1600s-1800s and compare them to the collocates from the 1900s, we see that the meaning of *mulheres* has not changed – in both cases it means *women*. In the 1600s-1800s, however, frequent collocates include *infeliz*, *desgraçada*, *abafada*, *honradas*, *constantes*, and *escandalosos*, most of which refer to the moral character (or sometimes emotional state) of women. In the 1900s, on the other hand, the adjectives tend to be more objective, such as *grávidas*, *jovens*, *social*, *equilibrada*, *negros*, and *mundial* (but also *nuas*, which reflects other cultural shifts). In summary, the use of collocates with the *Corpus del Español* and the *Corpus do Português* can be used as powerful tools to find evidence for semantic, pragmatic, cultural, and societal shifts in the language.

9.5.2 Synonyms

In addition to collocates, there are a number of other types of searches that are available via the *Corpus del Español* and the *Corpus do Português* that can provide valuable data for semantic analysis. First, with these two corpora (but not with any others) it is possible to find the frequency and distribution of the synonyms of more than 70,000 words. This information on synonyms is very useful, since a typical thesaurus does not provide information on which synonyms are most frequent, which are coming into or leaving the language, or which are used in different styles of speech (e.g. spoken or academic) and so on. With the *Corpus del Español* and the *Corpus do Português*, one simple query will provide all of this information.

For example, the query ||[=templado]|| will show the frequency and distribution of all of the synonyms of *templado*, and rank these in order of frequency: *suave*, *agradable*, *tranquilo*, *sereno*, *blando*, *benigno*, *moderado*, *tibio*, etc.[4] It is also possible to click on a link after any of the words in the list to find the synonyms for that word, and then click on a word in the new list to find the synonyms for that word, and so on. In this way, students can easily follow a "chain of synonyms" to discover the relationship between many different words.

This semantic information from the synonyms database can also be incorporated into more complex queries. For example, if a student wanted to see which nouns are used with all forms of different synonyms of *templado*, s/he would simply enter ||[n*] [[=templado]]||, and s/he would then see a results set with strings like *agua tibia*, *clima cálido*, *partido moderado*, *voz suave*, *noche serena*, *cielo sereno*, *palabras blandas*, *ojos serenos*, *mar tranquilo*, *luz serena*, *vida tranquila*, and *cama blanda*. Another example would be ||[=desastre] [j*]||, which would find all synonyms of *desastre* followed by an adjective, such as *tragedia griega*, *ruina total*, *derrota electoral*, *incendio forrestal*, *calamidad pública*, *perdición eterna*, and *caos confuso*. Of course, not all of these strings have words that are exact synonyms of *templado* or *desastre* in that particular case; such automatic semantic disambiguation is nearly impossible. But it does provide users with a useful way to compare the range of uses of the synonyms of a given word.

9.5.3 Customized word lists

Related to the ability to include semantic information from the synonym sets is the ability to create and re-use customized lists with the *Corpus del Español* and the *Corpus do Português*. Via the web-based interface, users can create lists of words for semantic category – such as military terms, color, clothing, or emotions – and then re-use this list as part of the query. For example, a user [Ferreira] might create in the *Corpus do Português* a list called [emoçoes] with words related to emotions (e.g. *cansado*, *contente*, *arrependido*, *satisfeito*, *seguro*, *triste*, *persuadido*, *ansioso*, *feliz*, *enganado*, *louco*). S/he could then use this list as part of another query, like ||estou [Ferreira@emoçoes]||, to retrieve matching strings like *estou contente*, *estou cansado*, *estou satisfeito*, *estou arrependido*, *estou feliz*, *estou enganado*, *or estou confidante*. Another example would be a user [Gómez] of the *Corpus del Español*, who creates a list [colores] of 15-20 different colors and a list of 30-40 types of clothing (*vestido*, *zapato*, *bufanda*, etc.). After creating these lists (which can be easily modified at any point in the future), s/he could then submit a query like ||[Gómez:ropa] [Gómez:colores]|| to retrieve the most frequent matching strings, like *traje negro*, *pañuelo blanco*, *vestido blanco*, *guante blanco*, *sombrero gris*, *chaleco blanco*, or *corbata azul.*

9.5.4 Word comparisons

The last type of semantically-based query that is possible with the *Corpus del Español* and the *Corpus do Português* (as well as *Sketch Engine*) are word comparisons. A typical thesaurus will show, for example, that *pelo* and *cabello* or *iniciar*, *empezar*, and *comenzar* mean essentially the same thing. Yet by comparing the collocates of contrasting words, we can tease out differences in meaning that are far beyond the capacity of even some of the best thesauruses or dictionaries. For

example, the simple query with ||pelo|| as the first word in a comparison and ||cabello|| as the second word (along with selecting [CONTEXT]) will find all of the collocates of *pelo*, all of the collocates of *cabello*, group them together, and then compare them to each other. This query shows, for example, that *pelo* is used almost exclusively with *mechones*, *capa*, and *verde*, as well as animals like *gato* and *camello*. *Cabello*, on the other hand, is used (vis-à-vis *pelo*) more than half of the time with *oscuro*, *mirada*, *ensortijado*, *rojizo*, *desorden*, and *ondas*. It may take the intuitions of a native speaker to sort out why certain collocates occur with one word and not with another. (For example, is *caían largos mechones de* [*cabello*] (vs. *pelo*) *áspero* possible in Spanish?) Yet tools such as these corpora allow researchers to go far beyond the thesaurus to look at intricate patterning of contrastive lexical items.

9.6 Frequency-based queries (historical and register variation)

As mentioned previously, the *Corpus del Español* and the *Corpus do Português* are the only two corpora that allow users to search for all words, phrases, synonyms, word forms, or strings of particular grammatical construction, which have a given frequency in different historical periods or in different registers of the modern language. This information can be very useful to study historical change and contemporary variation in the two languages. Let us briefly provide a few examples of these frequency-based queries, from among an unlimited number of possibilities.

First, in terms of morphology, it would be possible to find in the *Corpus do Português* all forms with the pattern ||f?z*|| (i.e. f + any letter + z + anything else), which occur at least ten times in the 1300s-1500s, but which are not frequent (or are non-existent) in the 1900s. With one simple query, users would see a list like *fezesse*, *fezera*, *fezerom*, *fezo*, *fezerõ*, *fizerão*, *fezer*, *fezessem*, and *fazião* – all of which are older forms for *fazer*. If so desired, this list of 40-50 word forms would then be used to create a customized word list, which could then be used as a group in subsequent queries. In terms of the lexicon, one could find in the *Corpus do Português* verbs that are more common in Brazil than in Portugal (e.g. *planejar*, *liberar*, *registrar*, *bancar*, *placar*, *checar*), or more common in Portugal than in Brazil (e.g. *registar*, *calhar*, *cozer*, *rematar*, *albergar*, *clarificar*). With the *Corpus del Español*, one could find those adjectives that occur much more in the 1800s than in the 1900s (e.g. *aleve*, *importuno*, *impío*, *desventurado*, *gallardo*, *aventajado*, *medroso*) or more in the 1900s than in the 1800s (e.g. *estadounidense*, *mundial*, *básico*, *similar*, *genético*, *estatal*, *impresionante*).

In terms of semantics, one could use the *Corpus del Español* to find the nominal collocates that precede *duro* (or ||[n*] [duro]||) and that occur more in academic than in fiction (e.g. *maderas duras*, *gestación dura*, *roca dura*, *línea dura*, *disco duro*) or more in fiction than in academic (e.g. *mirada dura*, *cuello duro*, *cabeza dura*, *cara dura*, *palabras duras*). In terms of semantics, with one simple query a user could find in the *Corpus del Español* the synonyms of *decir* (or

||[=decir]||) that are more common in fiction than in academic (e.g. *insinuar*, *largar*, *proferir*, *confesar*, *parlotear*, *opinar*) or more common in academic than in fiction (e.g. *significar*, *expresar*, *enunciar*, *mencionar*, *declarar*), or synonyms of *gritar* which occur more in the 1900s than the 1800s (e.g. *ulular*, *chillar*, *patalear*) or more in the 1800s than the 1900s (e.g. *exclamar*, *prorrumpir*, *vocear*, *bramar*, *clamar*). In all of these cases, these frequency-based searches – which are possible only with the *Corpus del Español* and the *Corpus do Português* – are done with just two or three clicks of the mouse, and it takes just two or three seconds to carry out these complex queries in a large 100 million word corpus.

10. Recent research with Spanish and Portuguese corpora

As we mentioned in the introduction, we will give just a brief overview of the corpus-based research that has been carried out in Spanish and Portuguese. As was mentioned previously, this section is limited by the fact that most research has been carried out with small, proprietary, and overall fairly limited corpora. In most cases, it would be impossible to go back and replicate any of these studies or check the data, since the corpora were ad-hoc collections of texts that were created to answer a single research question. In addition, the architecture of many of these corpora (which has unfortunately persisted even into more recent corpora) means that the data was collected in ways that now look quite rudimentary, and which would now probably be carried out with a full-featured corpus.

A second reason for including just a summary sketch of past research is perhaps a more political one. With more than 1200 entries in the *Linguistics and Language Behavior Abstracts** (LLBA) with the terms [*Spanish* or *Portuguese*] and [*corpus* or *corpora* or *corpus-based*], how can we select the thirty or forty most influential articles for review in this section? This would obviously lead to some researchers feeling that they had been slighted. Since virtually all researchers have access to the LLBA, we would suggest that they use the tool themselves to find the relevant articles for a particular area of interest. In this section, we simply provide an overview of the main topics related to Spanish and Portuguese corpus linguistics, without focusing on specific books and articles.

Finally, it would probably not be possible to provide a comprehensive survey of previous research, due to the fundamental problem in defining just what a corpus is. For example, as we searched the LLBA for articles that had [*Spanish* or *Portuguese*] and [*corpus* or *corpora* or *corpus-based*] in any of the search fields, we found many articles that used the term *corpus* in the article abstract. Yet upon closer investigation, we found that these "corpora" included such things as ten questions given to Spanish language learners, six radio ads in Portuguese, a collection of Spanish *refranes* from the 1700s, and twenty poems by a long-forgotten Bolivian poet. Therefore, it would have been impossible in any case to correctly delimit the

range of articles to those that actually deal with a true linguistic corpus, without actually reading through all 1200+ of these articles.

10.1 Methodology

In spite of the problems in reviewing past research, we will provide the results of queries of the LLBA database, to see some overall trends in research with Spanish and Portuguese corpora during the past ten or fifteen years. In terms of the methodology, as we have mentioned, we searched the LLBA using the terms [*Spanish* or *Portuguese*] and [*corpus* or *corpora* or *corpus-based*] and limiting the hits to articles published since 1990. After retrieving the list, we then exported it to RefWorks, where we further processed the entries, and then imported them into a database, where we analyzed the results. Using this approach, we were able to create a rank-ordered list of all of the subject and topic descriptors mentioned in the entries for the 1266 articles.

10.2 Language and dialect

There were 1115 articles that mentioned either Spanish or Portuguese or a dialect of one of these languages as a descriptor. Of these, 70% (787 articles) dealt with Spanish, while 30% (328 articles) dealt with Portuguese. For Spanish, the dialects or countries that were most frequently mentioned were Spain (43), Argentina (28), Venezuela (21), Chile (20), Colombia (18), Mexico (14) and Costa Rica 14. Notice that all of these countries have texts from the *Habla Culta* project (which has been available for most countries since the 1980s), which may explain why more corpus-based studies have been done on the Spanish from these countries. For Portuguese, more than twice as many articles deal with Brazilian Portuguese (232) as those for European Portuguese (96). This may reflect the much larger population of Brazil, or the fact that many of the *Linguagem Falada* materials (the Portuguese equivalent to the *Habla Culta* project) have been available for Brazil for more than twenty years. About one third (392) of the articles mention another language besides Spanish or Portuguese, which shows that the focus of many of the articles was a comparison of structures between languages, a focus on translation, or a focus on language acquisition. Of these languages, English was by far the most commonly listed (157), followed by French (63) and German (26).

10.3 Language structures and processes

We divided the language structures or processes that were discussed in these 1266 articles into seven categories: phonology, morphology, syntax, semantics, discourse/pragmatics, the lexicon, and language change. The goal was to determine which areas of linguistic investigation were most heavily influenced by corpus-

based studies. In terms of frequency, these seven areas of research are ranked as follows, with the most frequent sub-topics for each area also listed (along with the number of articles for each of these sub-topics)[5]:

420 pragmatics: discourse analysis (66), discourse strategies (60), pragmatics (57), text structure 41
368 syntax: verbs (55), pronouns (53), syntactic structures (52), word order (42), subject (24)
273 lexicon: lexicon (71), lexicography (53), borrowing (50), terminology (47), dictionaries (39)
161 language change: language history (69), historical text analysis (57), language change (20)
139 semantics: semantic analysis (53), syntax-semantics relationship (22), modality (21)
126 phonology: intonation (27), phonological analysis (21), suprasegmentals (18)
58 morphology: word formation (16), derivation (14), inflection (11)

It is probably not surprising that the majority of the articles refer to syntax/grammar and the lexicon, which are common topics of interest for corpus-based studies of many different languages. It is somewhat surprising, however, that so many articles deal with pragmatics. It is notoriously difficult to do high-level pragmatic research with corpora, since context is so important (but is typically not coded well in corpora), and since the phenomena are less localized than specific syntactic constructions or particular words. This suggests that perhaps many of these articles dealt more informally with particular pragmatic markers, or (more likely) that they dealt mainly with the stylistics of loosely-defined corpora. Finally, there are, of course, few studies that deal with phonology, since the corpora are text-based, and there are relatively few dealing with morphology, perhaps because until recently there were no corpora that easily allowed substring-based queries.

10.4 Language in its textual, communicative and social context

In addition to the structures and processes just discussed, we should also look at how corpus-based studies of Spanish and Portuguese have related to language in its textual and social context, as well as language in the context of L1 or L2 acquisition. The category [Textual context] is the most frequent, with 365 articles in topics such as discourse/text genres (54 articles), newspapers (43), scientific technical language (41), oral language (36), written language (28), colloquial language (20), advertisements (20), legal language (19) and Internet (19). It is probably not surprising that textual context is the most common area of research (since corpora are primarily text-based), but it is interesting to see the range of genres and registers

covered in these studies. Finally, since it is only recently that corpora with a wide range of genres have become available (such as CREA, CORDE, the *Corpus del Español* and the *Corpus do Português*), we can presume that most of these articles deal with the linguistic aspects of one or two genres or registers, rather than studies across a wide range of genres.

There are also 165 articles for the category [Social context and demographic variables], including social functions of language (28), social factors (27), code switching (25), sex differences (22), interpersonal communication (20), college students (19), age differences (16), language attitudes (14), sociolinguistics (14), socioeconomic status (11) and interpersonal relationships (11). A quick analysis of these articles suggests that many of them deal with data from either the *Habla Culta* and *Linguagem Falada* projects or the *Corpus Oral de Referencia del Español Contemporáneo* corpus, and the attempt to match up linguistic features with demographic variables.

Finally, the category [Language acquisition] has 174 articles, including language acquisition (41), child language (28), error analysis (language) (23), written language instruction (22), English as a second language learning (18), second language instruction (18) and elementary school students (17). A survey of these articles suggests that most actually deal with English, as in speakers of Spanish or Portuguese learning English. There are, however, a few ad-hoc "corpora" of other speakers learning Spanish or Portuguese. The limited, unsystematic approach of most of these articles is due to the fact that there is still no large learner corpus of Spanish or Portuguese (i.e. material produced by learners of these languages), as there is for English, which has the *International Corpus of Learner English**.

10.5 General focus and methodology of the articles

There were several descriptors for the corpus-based article that did not fit in well with the categories in Sections 10.2-10.4 above, and in fact many of these refer to the overall methodology of these articles. Not surprisingly, the most common descriptor was corpus linguistics, which was listed for 223 of the 1266 articles. This was followed by language usage (163), text analysis (95), computerized corpora (46), databases (41), text structure (41), computer generated language analysis (39), statistical analysis (37), natural language processing (29), variationist linguistics (27), computational linguistics (22), research design (22), computer applications (16), computer software (13), information structure (13), data collection (13), information content (12) and statistical analysis of style (10).

This list is actually more insightful than it might appear at first glance. It is interesting that only about one fifth of all of the articles that mention Spanish or Portuguese corpora in the abstract are classified as being a corpus linguistics article. This shows, perhaps, the informal nature of these articles in terms of corpus use.

Notice also that descriptors that refer to a more formal or systematic use of corpora (like *statistical analysis*, *natural language processing*, *computational linguistics*, *research design*, *computer applications*, etc.) are much less common. In fact, they are much less common for articles dealing with Spanish and Portuguese than they are for articles dealing with English (adjusted for the overall number of corpus-related articles for each language). This again suggests the fairly informal nature of corpus linguistics for Spanish and Portuguese that has held sway until quite recently. This may in turn be related to the fact that – until four or five years ago – there were no large, publicly-accessible corpora of Spanish and Portuguese that had architectures and interfaces to support research that consisted of more than relatively simplistic counts of specific words and phrases. In the case of English, on the other hand, such corpora (such as the *Bank of English* and the *British National Corpus*) have been available since the early 1990s, and they have led to hundreds of very well-researched papers since that time. The hope is that the recent introduction of such corpora for Spanish and Portuguese will help to create even more insightful corpus-based studies for these two languages.

Acknowledgements

I would like to thank Marc Carmen of BYU for the significant input he provided in compiling the list of Spanish and Portuguese corpora.

Notes

1 Please see the appendix for URLs for these and other resources indicated in this paper by an asterisk following the name of the resource at its first mention.
2 In this and subsequent examples, the material between the sets of vertical lines (||example||) is what the user would input into the web-based search engine.
3 Table 1 shows the raw frequency for each century and register. It is also possible to see the normalized frequency (tokens per million words), or a combination of the two numbers. Note, however, that the table is color coded, and this is a function of the normalized frequency, no matter which number is displayed in the cell.
4 To find all forms of all synonyms, users simply use (an extra level of) square brackets, as with other lemma-based searches, e.g. || [[=templado]] ||.
5 The total would actually be somewhat higher than each of these numbers. This is because we only exported from the database those descriptors that occurred ten times or more in the database (about 200 descriptors in all). We then grouped these descriptors into the seven categories shown here, and then queried the database again to see how many articles had at least one of the sub-topics in that category. However, this ignores articles that may have belonged to one of these seven categories, but none of the sub-topics occurred more than

nine times, and so they were not exported and grouped into the seven main categories. Note also that there was overlap and multiple listings for these categories in the descriptors, so that an article could be listed as *syntax* and *historical* or *discourse analysis* and *semantics*.

Appendix

URLs for corpora and related materials

Acervo Digital (National Library of Brazil)	http://www.bn.br
ADMYTE	http://www.admyte.com
Association for Hispanic Classical Theater	http://www.trinity.edu/org/comedia/textlist.html
Bank of English (COBUILD)	http://www.collins.co.uk/Corpus/CorpusSearch.aspx
Biblioteca Digital da Porto Editora	http://www.portoeditora.pt/bdigital/
Biblioteca Digital de Literatura	http://alecrim.inf.ufsc.br/bdnupill/
Biblioteca Virtual	http://www.cervantesvirtual.com/catalogo/index.jsp
British National Corpus	http://www.natcorp.ox.ac.uk/ See also http://corpus.byu.edu/bnc for a free interface to the BNC, with many features not available with other BNC interfaces.
Brown Corpus	http://icame.uib.no/brown/bcm.html
CETEMPúblico (Linguateca)	http://acdc.linguateca.pt/cetempublico/
CORPORA listserv	http://listserv.linguistlist.org/archives/corpora.html
Corpus de Referencia de la Lengua Española en Chile	http://www.lllf.uam.es/~fmarcos/informes/corpus/cochile.html
Corpus de Referencia de la Lengua Española en la Argentina	http://www.lllf.uam.es/~fmarcos/informes/corpus/coarginl.html
Corpus de Referencia del Español Actual (CREA; Real Academia)	http://corpus.rae.es/creanet.html
Corpus de Referência do Português Contemporâneo	http://www.clul.ul.pt/sectores/linguistica_de_corpus/projecto_crpc.php
Corpus del Español	http://www.corpusdelespanol.org
Corpus del Español (Interface for comparison of registers)	http://www.corpusdelespanol.org/registers
Corpus Diacrónico del Español (CORDE; Real Academia)	http://corpus.rae.es/cordenet.html
Corpus do Português	http://www.corpusdoportugues.org
Corpus Informatizado do Português Medieval	http://cipm.fcsh.unl.pt/

Corpus Oral de Referencia del Español Contemporáneo	http://www.lllf.uam.es/corpus/corpus_oral.html
CRATER archive	http://www.comp.lancs.ac.uk/computing/research/ucrel/corpora.html#crater
David Lee (list of corpora)	http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/software.htm
Elsnet	http://www.elsnet.org/resources/eciCorpus.html
EuroParl	http://people.csail.mit.edu/koehn/publications/europarl/
Evaluations and Language Resources Distribution Agency (ELDA)	http://www.elda.org/
Floresta Sintáctica	http://www.linguateca.pt/Floresta/
Fuente Académica (EBSCO)	http://www.ebsco.com
Google Books	http://books.google.com
Humanities Full Text (Wilson)	http://www.hwwilson.com/Databases/humani.htm
IMS Corpus Workbench	http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/
International Corpus of Learner English	http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm
JRC-Acquis Multilingual Parallel Corpus	http://langtech.jrc.it/JRC-Acquis.html
Lácio-Web	http://www.nilc.icmc.usp.br/lacioweb/corpora.htm Especially Par-C (Eng-Pt parallel corpus) and Comp-C (Eng-Pt comparable corpus)
Leeds Collection of Internet Corpora	http://corpus.leeds.ac.uk/internet.html
Lexis-Nexis (Spanish newspapers)	http://www.lexisnexis.com
Linguateca (list of resources)	http://www.linguateca.pt/corpora_info.html
Linguistics and Language Behavior Abstracts (LLBA)	http://www.csa.com/factsheets/llba-set-c.php
Linguistic Data Consortium (LDC)	http://www.ldc.upenn.edu Especially spoken data (CALLHOME, CALLFRIEND, Spanish Broadcast News Speech) and newspapers (Gigaword, Spanish News Text, and Spanish Newswire Text)
LOB Corpus	http://khnt.hit.uib.no/icame/manuals/lob/index.htm
Natural Language Toolkit	http://nltk.sourceforge.net/
Oxford Text Archive	http://ota.ahds.ac.uk/
PHP archive	http://stp.ling.uu.se/opus/php.html
Project Gutenberg	http://www.gutenberg.org/browse/languages/es
Projecto Vercial	http://alfarrabio.di.uminho.pt/vercial/programas.htm

R Project	http://www.r-project.org/
Research Library (ProQuest)	http://www.proquest.com
Sketch Engine	http://www.sketchengine.co.uk/
Textos Lemir	http://parnaseo.uv.es/Lemir/Textos/index.htm
TreeTagger	http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
Tycho Brahe Parsed Corpus of Historical Portuguese	http://www.ime.usp.br/~tycho/corpus/index.html
VISL/CorpusEye	http://corp.hum.sdu.dk/cqp.es.html (Spanish) http://corp.hum.sdu.dk/corpustop.pt.html (Portuguese) http://visl.hum.sdu.dk/visl/pt/ (Portuguese)
WordCruncher	http://www.hlanalysis.com/WordCruncher/WC.aspx
WordSmith	http://www.lexically.net/wordsmith/index.html
Wordtheque	http://www.logos.it/literature/literature.html

References

Biber, Douglas, Susan Conrad & Randi Reppen. 1998. Corpus linguistics: Investigating language structure and use. Cambridge: Cambridge University Press.

Biber, Douglas, Mark Davies, James Jones & Nicole Tracy-Ventura. 2006. Spoken and written register variation in Spanish: A multi-dimensional analysis. Corpora 1, 1-37.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. Longman grammar of spoken and written English. New York: Longman.

Davies, Mark. 2005a. Advanced research on syntactic and semantic change with the *Corpus del Español*. In Claus D. Pusch, Johannes Kabatek & Wolfgang Raible (eds.), Romance corpus linguistics II: Corpora and diachronic linguistics, 203-214 Tübingen: Gunter Naar Verlag.

Davies, Mark 2005b. The advantage of using relational databases for large corpora: Speed, advanced queries, and unlimited annotation. International Journal of Corpus Linguistics 10, 301-28.

Davies, Mark. 2005c. A frequency dictionary of Spanish: Core vocabulary for learners. London: Routledge

Davies, Mark. 2007. Towards the first comprehensive survey of register variation in Spanish. In Eileen Fitzpatrick (ed.), Corpus linguistics beyond the word: Corpus research from phrase to discourse, 73-86. Amsterdam: Rodopi.

Davies, Mark & Ana Raposo Preto-Bay. 2007. A frequency dictionary of Portuguese: Core vocabulary for learners. London: Routledge

Fillmore, Charles. 1992. “Corpus linguistics” or “Computer-aided armchair linguistics”. In Jan Svartvik (ed.), Directions in corpus linguistics: Proceedings Nobel Symposium 82 Stockholm, 4-8 August 1991, 35-60. Berlin: Mouton de Gruyter.

Juilland, Alphonse G. & Eugenio Chang-Rodríguez. 1964. Frequency dictionary of Spanish words. The Hague: Mouton.

Kennedy, Graeme D. 1998. An introduction to corpus linguistics. London: Longman.

Lope Blanch, Juan. 1993a. Ensayos sobre el español de América. México City: UNAM.

Lope Blanch, Juan. 1993b. Nuevos estudios de lingüística hispánica. México City: UNAM.

McEnery, Tony & Andrew Wilson. 2001. Corpus linguistics: An introduction. (2nd edition). Edinburgh: Edinburgh University Press.

McEnery, Tony, Richard Xiao & Yukio Tono. 2006. Corpus-based language studies: An advanced resource book. London: Routledge.

Véronis, Jean. 2005. Web: Google’s missing pages: Mystery solved? Technologies du Langage: Actualités-Commentaires-Réflexions (February 8). http://aixtal.blogspot.com/2005/02/web-googles-missing-pages-mystery.html