Creating and Using Multimillion-Word Corpora from Web-Based Newspapers

Mark Davies
Illinois State University

With the dramatic increase in the past few years of newspapers and magazines that are available on the Web, it has now become feasible for researchers to create multimillion-word corpora for their research. This chapter will first discuss two large corpora that I have created for Spanish and Portuguese (35 million and 26 million words, respectively) and how they complement other corpora that I have created for these two languages. I will then discuss—in a fairly practical and step-by-step fashion—the process by which researchers can create their own corpora, as well as some of the challenges they might face. Finally, I will discuss how large multimillion-word corpora can be used to complement more traditional corpora, specifically in the sense of studying newly emerging syntactic constructions in a particular language.

Two Multimillion-Word Web-Based Corpora

I have created a 35-million-word corpus of Modern Spanish and a 26-million-word corpus of Modern Portuguese. The Spanish corpus is composed of 20 million words from newspapers in Latin America and contains a total of 1 million words from at least two newspapers in each of the twenty Spanish-speaking countries. It also contains 15 million words from eight newspapers in Spain. (A list of the newspapers included in the Spanish corpus can be found in my appendix.) The Portuguese corpus is composed of 26 million words, including 15 million words from twelve newspapers in Brazil (representing the different geographical regions of the country), as well as 10 million words from eight newspapers in Portugal. One of the unique aspects of the Portuguese corpus is that there is an additional 1 million words of text from interviews extracted from these newspapers, including 625,000 words from five newspapers in Brazil and 390,000 words from five newspapers in Portugal. In summary, these corpora constitute some of the largest and most comprehensive corpora of Spanish and Portuguese in existence.

These corpora supplement a number of other, nonnewspaper Spanish and Portuguese corpora that I have created, which are summarized in tables 1 and 2 (more details are available at <http://mdavies.for.ilstu.edu/personal/texts.htm>). In table 1, I have divided the corpora into spoken and written texts. The distinction here is between those texts that are the transcripts of actual conversations (spoken), on the one hand, and, on the other, novels, short stories, and newspapers (written), which, although they may attempt

TABLE 1. Modern Spanish Corpora

Country	Corpus	Number of texts/ conversations	Number of words
Spoken			
Latin America	*Habla Culta* (Bogotá, Buenos Aires, Caracas, Havana, La Paz, Lima, Mexico City, San Jose [Costa Rica], San Juan [P.R.], Santiago [Chile])	385	2,193,000
Spain	*Habla Culta* (Madrid, Sevilla)	72	328,000
Spain	*Corpus oral de referencia de la lengua española contemporanea*	498	948,000
Written			
Latin America/ Spain	Novels	15	1,327,000
Latin America/ Spain	Short stories (same countries as those in the *Habla Culta* corpus)	356	1,054,000
Argentina	*Corpus lingüistico de referencia de la lengua española—Argentina*	22	1,913,400
Total		1350	7,763,000

TABLE 2. Modern Portuguese Corpora

County	Corpus	Number of texts/ conversations	Number of words
Brazil	*Linguagem Falada* (Recife, São Paulo, Rio de Janeiro, Salvador)	85 conversations	570,800
Brazil	Borba-Ramsey Corpus (essay, novel, journal, technical, drama)	102 blocks	1,670,300
Brazil	Short stories	26 authors	75,100
Portugal	Novels	11 novels	239,000
Total		224	2,555,200

to model the author's perception of native speech, are in fact inventions of the author.

Looking at the lists of corpora in tables 1 and 2, one can see that although the nonnewspaper corpora are fairly large (10 million words in both spoken and written texts between the two languages), the 60 million words from the Web-based newspapers represent much larger corpora and can thus provide some useful insight into certain syntactic constructions, a point that I will return to in the final section of this chapter.

Creating Multimillion-Word Web-Based Corpora

In what follows, I will discuss—step-by-step—some of the more important procedures in creating large multimillion-word corpora from Web-based newspapers and magazines. Although the discussion may appear a bit rudimentary at times, my purpose is to guide others through the process and hopefully help them to deal with some of the technical challenges that I faced when creating my two corpora.

Finding Newspapers

As most people who have spent much time on the Web are aware, there has recently been a virtual explosion in the number (and quality) of newspapers on the Web. While researchers of English have had access to large multimillion-word corpora for a number of years, these

Web-based newspapers constitute an important resource for those working in other languages or perhaps specialized varieties of English, such as obituaries, law reports, and editorials. In compiling my corpus of Spanish newspapers, it was encouraging to find that even small, less-developed countries, such as El Salvador and Paraguay, have three or four newspapers on-line. There are a number of excellent lists of international newspapers on-line, and a partial list of these lists can be found at <http://dir.yahoo.com/News>.

Evaluating Newspapers

As might be expected, there are differences between the various on-line newspapers in terms of how easily they can be incorporated into a corpus. One issue is the thematic content of the newspaper. While some have extensive sections for national and international news and for society, sports, and finance news, other newspapers are monothematic, focusing on one topic, such as sports or finance. For researchers wanting a wider range of styles and registers, newspapers with many different thematic sections are of course preferable. A second issue is the server on which the newspaper is located and the speed at which articles can be downloaded, especially when corpus creation involves downloading hundreds of thousands of articles.

A more important issue concerns the archives of past editions. Many newspapers (including some of the best ones) only provide links to the editions for the past week, which makes it difficult to create a multimillion-word corpus, unless one revisits the site over a number of weeks. Other newspapers, however, provide an archive of issues for the entire past month, and some provide every issue for the past two or three years.

The final issue, and perhaps the most important one, deals with the amount of redundancy in articles from day to day. If the same article appears two or three days in a row, automated downloading of the contents for those days means that tokens from that article will appear multiple times in the corpus, which is of course problematic. The best solution is to avoid newspapers with a high degree of redundancy, but another solution (if the newspaper has extensive archives) is to download just one day a week for the past two or three years and hope that there is little overlap. One final possibility is to take care of matters at the level of text retrieval, using a program or algorithm to account for redundancy at that point.

Locating Past Editions

The next stage is to develop a sense of the structure of the Web site, so that the individual articles can be downloaded as efficiently as possible. Let us suppose that the "home page" for a particular issue has the address <http://www.excelsior.com.mx/9807/980704/>; after the domain (www.excelsior.com.mx) is the date, with the format */yymm/yymmdd/*. In our word processor, we can copy this address, or URL, for as many days as we want to download, then use macros to modify the "base" URL for all of the particular dates that we want to download (e.g., /9710/971005/ [October 5, 1997], /9805/980522/ [May 22, 1998]). Finally, we can save this list of links in **HTML** format (<A HREF="<link>"><title></A>).

Employing Automated Downloading

It would of course be impractical to manually download tens and hundreds of thousands of articles. Fortunately, there is software that will do this for us. Using an **off-line browser,** such as Grab-A-Site (<http://www.bluesquirrel.com/products/grabasite/grabasite.html>; see the related software at <http://tucows.pdnt.com/offline95. html>), we indicate the starting URL and how many levels down we want to "crawl." For example, the Web page containing the list of URLs (from the previous section) would be level 1. The links to the home pages for the different issues of the newspaper would be level 2. The links on those home pages to the various sections (national news, society news, sports news, etc.) would be level 3. The actual articles would be level 4, so we would set the "crawling" level to "4." With such programs as Grab-A-Site, we can also direct the crawler to limit downloads to the Web site for the newspaper (rather than following links to ads or URLs for companies mentioned in the articles), as well as limiting downloads to just Web pages (HTML files) rather than graphics and other multimedia. If we have a fairly fast network connection, it is possible to configure the "crawler" for downloads and then return in the morning with several thousand articles from a particular newspaper.

Working with Directories and Files

Most "crawler" software allows us to mirror the directory structure of the Web site, so that all of the articles for a particular date will be in the same directory. After scanning through the file names, we can delete files that are clearly redundant (e.g., Web pages with subscription information) or not useful (e.g., stock quotes, where there is little usable text), and we can delete these files recursively in child directories with just one command. In some cases, simple **batch files** may help to automate the process.

In cases where the directories on the Web site organize articles by date, all of the articles for the same date will be together, and we can use a simple command (e.g., "copy *.html 0502.htm" for articles in the May 2 directory) to concatenate all of the separate files for one day into just one file, which is easier to use in subsequent stages. There is a problem, however, when different articles for a particular date are spread among different directories (e.g., /sports/1998/0502,/finance/1998/0502). In such cases, we sometimes have to resort to fairly complex batch files to get all of the files for one date into the same directory, so that we can concatenate them into one file.

Converting from HTML Files to Text

At this stage, the Web pages will usually still be in HTML format, with all of the style and formatting tags that entails. If we want simple text files for the corpus, we will need to convert all of the tens of thousands of files, and it would of course be impractical to do this manually. The process can be automated using a program, such as HTMASC32 (<http://www.bitenbyte.com/htmasc.htm>). This particular software can very quickly convert to text tens of thousands of HTML files, and it can also handle files as large as 10 to 20 megabytes, for those cases in which we have created large concatenated files.

Using Macros to Clean Up Texts

Most newspaper articles contain redundant information that we probably do not want as part of the corpus, such as headers or footers containing links to other articles, subscription information, and links to ads. Using macros in our word processor, we can strip out as much of this information as we would like. For example, we can create a macro to search for certain text that indicates the beginning of the footer on the Web pages. We then block the material between that text and the title of the following article (to eliminate the header from that article). The macro then deletes all of the blocked material, eliminating the

headers and footers. However, to determine which textual elements we will want to delete on every page, we need to obtain some sense of the layout of the pages for a particular newspaper.

Assuming we have a fairly powerful computer, we can have the macro run on hundreds of files (e.g., one for each day's issue) and strip out this text from hundreds or thousands of articles in each file. I offer one caveat, however—the faster the word processor, the better. When dealing with hundreds of thousands of articles, we do not want a complicated word processor that is overloaded with features or has an overly complex graphical interface. In my work, I have used WordPerfect 5.1 (DOS) or TextPad (<http://www.textpad.com>), both of which are quick, allow macros, and have worked quite well.

Locating and Extracting Special Kinds of Texts

In some cases, we might want to select certain types of articles (based on thematic or subgenre characteristics) and create a special subcorpus of these materials. For example, I was aware that my 25-million-word Portuguese corpus contained over a thousand transcripts from oral interviews, and I wanted to create a special "spoken" corpus of just these materials. Again, using macros, this should be feasible. We first find the textual markers that reoccur in all of the articles that interest us. For example, all of the interviews in one of the Brazilian newspapers had a hard return and then the word *Estado* (the name of the newspaper) and a hyphen at the beginning of the interviewer's comments. Once we know what these markers are, we can write a simple macro to search for that marker, then block the entire article and place it in a separate file that contains the specialized corpus. By running the macro iteratively on multiple files, we can easily and quickly extract all of the desired articles.

Using a Text Retrieval and Analysis Program

The final stage in creating the corpus is to do whatever processing is necessary to get the files in a form that can be used by our concordance or text retrieval and analysis program. In my research dealing with dialectal and historical variation in Spanish syntax, I have used WordCruncher (DOS). Once I have the text files, I simply insert markers to identify newspaper, date, and article name, then the WordCruncher indexing program can create an every-word index for the

corpus. Although I do not use tagged corpora, in WordCruncher I can create and save part-of-speech lists (e.g., all infinitives, prepositions, or pronouns) that can be retrieved and used time after time. For example, in looking in the corpus of Latin American Newspapers for cases of the newly emerging construction "lexical subject of infinitive" (e.g., "es difícil **para él hacer**lo" [it's hard **for him to do** it]), I can search for a preposition (2,400,000+ tokens), followed by a personal pronoun (26,000+ tokens), followed by an infinitive (540,000+ tokens). Once the search has been set up, it takes fewer than two seconds to find the 80 to 90 tokens, and it would still take fewer than three seconds even if there were 5,000 to 10,000 tokens.

Using Multimillion-Word Web-Based Corpora in Research

With a few exceptions, most of the corpora currently available from works of literature or transcripts of conversations are much smaller than the Web-based corpora that I have discussed to this point; they are often in the range of 100,000–5,000,000 words. As Biber notes (1990, 1993), these smaller corpora are usually sufficient for most kinds of linguistic research. However, with emerging constructions that are still quite rare, the sheer size of the corpora of Web-based newspapers makes them a welcome addition to more traditional corpora by providing data on constructions that will appear in sufficient numbers only in very large corpora. In the sections that follow, I will focus on just my Spanish corpus and discuss how the 35-million-word corpus of Web-based newspapers has helped to complement the other corpora of Modern Spanish that I have created (see table 1) and how it has helped to provide important insight into the nature of syntactic variation and change.

Before examining cases in which a large Web-based corpus has been of value in examining emerging syntactic constructions, let us briefly consider three kinds of limitations that even corpora of this size have. Each limitation will be illustrated here by examples.

First, it may not make much sense to use very large corpora to study some phenomena, such as two alternate constructions that appear very frequently in the language, since these are sufficiently represented in much smaller corpora. Davies 1995a, a study of clitic climbing in Modern Spanish, shows that climbing (the movement of the unstressed object pronoun to a position in front of the main verb)

is dependent on the particular verb. The data, based on more than 15,000 tokens, show that clitic climbing is more common with simple verbs, such as *querer*, "to want" (47 percent of cases with climbing in the spoken register, 15 percent in the written register), than with semantically more complex verbs, such as *desear*, "to desire" (20 percent and 4 percent, respectively), or *esperar*, "to hope" (0 percent in both registers). For an explanation of the parentheses in examples 1–3, see <http://www.mdavies.for.ilstu.edu/personal/texts.htm>.

1. a. *y qué* **me** *quiere decir* (Bogotá M14:186)
   "and what he wants to tell me"
   b. *y qué día* **lo** *desea realizar* (ARG Cartas: Carta 855)
   "and when he wants to achieve it"
   c. *porque no esperaba encontrar***lo** *allí* (Gazapo 41)
   "because he didn't expect to find it there"

Since there are already many thousands of tokens for these constructions in the smaller, traditional corpora, there is little to be gained from the larger Web-based corpora. In the Latin American and Spanish newspapers from the Web, the frequency of clitic climbing was not significantly different with the various verbs than it was in the other corpora; it was 22 percent with *querer*, 6 percent with *desear*, and 2 percent with *esperar* (based on 8,400 tokens with these three verbs, as opposed to 1,320 tokens in the original corpus).

Second, in cases where smaller corpora already provide evidence for emerging constructions, an increased number of examples from larger corpora are not particularly insightful. Davies 1992, 1995b, and 1996 discuss the evolution of causative constructions in Spanish and Portuguese and show that there has been a shift toward "biclausal" constructions. One piece of evidence for the newer structure is the fact that with many causative verbs and verbs of perception, it is now possible to have "reflexive" embedded verbs, whereas this was not attested in older stages of the language.

2. a. *cuando lo vio balancear***se** *por primera vez* (Hombres 75:1)
   "when she saw him rock back and forth the first time"
   b. *Oliveira lo dejó ir***se** (Rayuela 377:1)
   "O. let her go away"

However, one verb with which the reflexive verbs are still quite uncommon in Modern Spanish is the prototypical causative *hacer*, "to make." Although Finneman 1982 reports that some speakers allow reflexives with *hacer*, there are no cases in the 600,000-word corpus in Davies 1992. In the expanded 7,700,000-word corpus of Modern Spanish (see table 1), however, there are more than 40 examples from both written and spoken Spanish.

3. a. *acaba uno el examen y lo hacen retirar***se** (Bogotá M12:165)
   "you take the test and then they make you leave"
   b. *lo hace . . . despreocupar***se** *de lo que está diciendo* (Santiago M41:231)
   "it makes you not pay attention to what you're saying"

Not surprisingly, in the larger 35-million-word corpus of Web-based Latin American and Spanish newspapers, there are even more tokens—46 from Latin America and 39 from Spain. In this case, then, the critical increase in corpus size was not from 7,700,000 to 35 million words (and from 40 to 85 tokens) but from 600,000 to 7,700,000 words (and from 0 to 40 tokens).

Third, even large corpora are unable to provide examples of certain emerging linguistic phenomena. There are cases where (based on historical trends) one might expect certain constructions to appear in larger corpora, yet they are still unattested. Although Davies 1995a and 1998 show that there is a definite shift toward clitic climbing in Modern Spanish, there are still no cases of clitic climbing with certain verbs, such as *hay que*, "one has to" (480 tokens), *soñar con*, "to dream of" (just one token), and *insistir en*, "to insist on" (6 tokens); in all cases, clitic placement follows the infinitive.

4. a. *¡hay que abrir***la***!* (Cuba: *CubaNet*)
   "you've got to open it!"
   b. *soñaba con conocer***lo** *personalmente* (Costa Rica: *Prensa Libre*)
   "she dreamed of meeting him in person"
   c. *sus detractores insisten en acusar***lo** (El Salvador: *Diario*)
   "his detractors insist on accusing him"

In the Web-based corpora, however, there are still no cases of clitic climbing with these verbs (1,375 tokens with *hay que*, 16 with *soñar con*, and 28 with *insistir en*). In consequence, although there is a historical shift toward clitic climbing, it is apparently still unacceptable with certain verbs, and even the larger Web-based corpora do not (yet) show evidence of a shift.

In the examples just considered, the larger corpus of Web-based

newspapers simply confirmed what might have already been proven in the smaller corpora—regarding the relative frequency of constructions, the fact that they have now become an established part of the grammar, or (conversely) the fact that certain constructions are still unacceptable. In the examples that follow, however, we will see that sometimes a larger corpus does in fact provide the crucial evidence for emerging constructions in a language—data that is unavailable in smaller corpora. In each of these examples, we will be comparing the smaller 7,700,000-word corpus of spoken and written Spanish (see table 1) with the larger 35-million-word corpus of Web-based Spanish newspapers.

Let us take as the first example the object-to-subject raising (OSR) in Modern Spanish. One difference between Spanish and English is that in Spanish, there is a much more restricted range of adjectives that allow the construction. While Spanish allows OSR with many of the synonyms of *fácil* and *difícil* ("easy" and "hard") and with the adjectives *posible* and *imposible*, the common view is that OSR in Spanish does not occur with adjectives meaning "nice," "fun," "important," and "interesting" (Reider 1993).

5. a. **ese coche es* ***divertido de manejar***
   "that car is fun to drive"
   b. **es una película* ***interesante de mirar***
   "it's an interesting movie to watch"

There are no cases of OSR with these adjectives in the 5,300,000-word corpus of historical Spanish texts, and there are likewise no cases of OSR with these adjectives in the smaller 7,700,000-word corpus of spoken and written Modern Spanish. In the larger corpus of Web-based newspapers, however, there are cases with each of these adjectives (*importante*, 4 in Latin America; *agradable*, "nice," 2 in Spain, 1 in Latin America; *divertido*, "fun," 1 in Spain; and *interesante*, 1 in Latin America), which may suggest that the range of adjectives permitting OSR in Spanish is increasing.

6. a. *lo promuevan como un destino* ***importante de visitar*** (Nicaragua: *Prensa*)
   "they promote it as an important place to visit"
   b. *un espacio distendido . . . [que] resulte* ***agradable de ver*** (Spain: *El Mundo de Baleares*)
   "a stretch of land that is nice to look at"

   c. *[el libro] resulta ameno y* ***divertido de leer*** (Spain/Barcelona: *Vanguardia*)
   "the book is nice and fun to read"
   d. *cuyos efectos sin duda serían muy* ***interesantes de estudiar*** (Guatemala: *Prensa Libre*)
   "whose effects would without doubt be very interesting to study"

In addition to the fact that OSR can occur with a wider range of adjectives than previously thought, there is also some evidence that there is a semantic reanalysis of the construction taking place, in which the fronted NP is coreferenced in some way with the subject position of the embedded clause. Evidence for this from the larger corpus is the fact that there are scattered cases of OSR involving passives, agentive phrases, and the "reflexive" marker *se*.

7. a. *las propiedades eran difíciles de* ***ser estudiadas*** *experimentalmente* (Madrid: *País*)
   "the properties were hard to study experimentally" (passive)
   b. *servicios que son imposibles de disfrutar* ***por la mayoría de la gente*** (Cuba: *CubaNet*)
   "services that are impossible for most people to enjoy" (agentive)
   c. *un desorden institucional difícil de solucionarse* (Venezuela: *Universal*)
   "an institutional mess [that is] hard to fix" (reflexive)

While each of these constructions is either nonexistent or occurs just once in the smaller corpus of written and spoken Spanish, there are somewhat more examples from the Web-based corpus. For example, the reflexive *se* construction is not found in the smaller corpus but appears nine times in the newspaper corpus. In summary, the large corpus of Web-based materials provides crucial evidence for several aspects of the Modern Spanish OSR construction, evidence that is not available in any of the smaller corpora.

In papers dealing with the subject-to-subject raising (SSR) in both historical Spanish and Modern Spanish (Davies 1997a, 1997b), I briefly discuss an aspect of the construction that had previously not been studied. These are the cases of partial raising. In nonraised structures, the subject of the embedded clause stays within the lower clause (see example 8a). In full raising, the subject of the embedded

clause raises to the main clause and triggers agreement only in that clause (see example 8c). In partial raising, the embedded subject raises to the main clause (and triggers agreement there) but also leaves some type of "trace" in the embedded clause, which triggers agreement in the embedded clause as well (see example 8b).

8. a. __ ***parece*** *[que* ***saben*** *la respuesta]* No Raising
   "it seems that they know the answer"
   b. *ellos* ***parecen*** *[que* __ ***saben*** *la respuesta]*Partial Raising
   "they seem as if they know the answer"
   c. *ellos* ***parecen*** *[*__ ***saber*** *la respuesta]* Full Raising
   "they seem like they know the answer"

I have hypothesized that partial raising was an important construction in the historical development of subject raising in Spanish. In Old Spanish, the subject of the embedded clause could not raise to the main clause (as in example 8a). In Modern Spanish, full raising (as in example 8c) is quite common, but partial raising (as in example 8b) is rather uncommon. During the 1400s–1600s, at the very moment that raising was becoming possible in Spanish, partial raising was quite common, but then it decreased sharply once full raising became common in the 1700s (Davies 1997b). I have suggested that the language moved from no raising to full raising via intermediate partial raising (active in the 1400s–1600s) and that once this construction had "played its part" (so to speak), it then died out.

The one complication, however, is that there are still a handful of cases of partial raising in the smaller corpus of Modern Spanish.

9. a. *me* ***parecen*** *que no* ***eran*** *los correctos* (La Paz M22:229)
   "they seem to me like they're not the right ones"
   b. *qué problemas te* ***parecen*** *que* ***son*** *los más import*antes (Sevilla Popular M7:176)
   "which problems seem to you like they're the most important ones"

It is difficult to explain the fact that eight of nine cases are in the corpus of spoken Spanish, even though nearly all of the more than 100 native speakers who were asked about this construction in a survey quickly rejected it. I have hypothesized that the few cases of the construction in spoken Spanish might be due to some kind of "garden path" phenomena, in which the speakers start into the SSR construction (thus the fronted NP) but then, for a number of functional reasons, "backtrack" into the nonraised construction (Davies 1997a). Since this is due to production constraints, it nearly always occurs in the spoken language. Unfortunately, the data from the larger corpus of Web-based newspapers does not allow any such "processing" or "performance" explanation. There are eight additional cases of the construction (6 from Latin America and 2 from Spain), and they are all from the written register; none of these tokens are taken from interviews or reported speech.

10. a. *me* ***parecen que*** *no* ***conducen*** *a nada* (Mexico: *Yucatán*)
    "it seems to me like they don't lead anywhere"
    b. *hasta la que* ***parecen que*** *se* ***van*** *de home run* (Panama: *Siglo*)
    "until they look like they're making a home run"
    c. ***parecen que*** *no lo* ***son*** *tanto* (Spain/Oviedo: *Comercio*)
    "it looks like they aren't that way as much"

This, then, is an example of a construction that was assumed to have more or less died out by Modern Spanish (judging again by the universal rejection of the construction by native speakers), at least in the written register, where speakers have the time to carefully craft the sentences. Therefore, the cases of partial raising in journalistic prose of the Web-based newspapers raises some questions that will need to be addressed in future research.

Perhaps the clearest example of the value of the multimillion-word corpus of Web-based newspapers concerns the construction "lexical subject of infinitive." While English easily allows subjects of nonfinite verbs (e.g., "after Bill's leaving", "it's nice for you to say that"), these are quite uncommon in Spanish. Nearly all researchers to date have commented that these constructions are restricted to the Caribbean dialects and that they are mainly a feature of informal, colloquial spoken Spanish (Suñer 1986; Lipski 1991). DeMello 1995 shows, however, that the construction has now spread to the spoken register in a number of cities of Latin America and Spain, and this is verified in my collection of texts from the *Habla Culta* project, which contains transcripts of conversations with native speakers from ten different countries.

11. a. *por el hecho* ***de él haber*** *sido eliminado de la presidencia* (Santiago M49:395)
"by reason of his having been eliminated from the presidency"
b. *es la consecuencia* ***de yo haberme*** *parado trescientas veces* (Madrid M13:219)
"it's the result of my having stopped three hundred times"

In my 4,300,000-word corpus of written Spanish, however, the construction occurs in written Spanish only four times, and all of these cases are from the Caribbean dialects.

12. a. *había tierra fértil antes* ***de yo nacer*** (Puerto Rico 1:243)
"there was a lot of fertile land before I was born"
b. *cualquiera es buena* ***para tú encontrarme*** (Venezuela 1:179)
"any of them are fine for you to get [them] for me"

Therefore, there is still no evidence from the smaller corpus that this emerging construction has now spread to written texts from other countries beyond the Caribbean zone. In the larger corpus of Web-based newspapers, however, there is clear evidence that the construction has in fact spread throughout Latin America (22 examples) and even to Spain (16 examples).

13. a. ***Para yo hablar*** *de política tengo que estar en Colombia* (Ecuador: *Vistazo*)
"for me to talk about politics I have to be in Colombia"
b. *que se vaya Siles* ***para él habilitarse*** *como candidato* (Bolivia: *ERBOL*)
"for S. to go away so he can get in shape as a candidate"
c. *sería mejor* ***para ella salir*** *de Washington* (Barcelona: *Periódico*)
"it would be better for her to leave Washington"
d. ***para ellos subcontratar*** *y sacar su beneficio* (Oviedo: *Comercio*)
"for them to subcontract things out and get the benefit"

The spread of this construction to the written Spanish of nearly all of Latin America and Spain is not a trivial matter. Lexical subjects of infinitives should not occur at all in Spanish due to several supposedly universal syntactic constraints, and its appearance in spoken Caribbean Spanish is supposedly due to related morphological and syntactic features of this dialect and register (see Suñer 1986; Lipski

1991). The evidence from the 35-million-word corpus of Web-based newspapers, however, shows that the theory will somehow need to be modified to allow for the numerous cases in this corpus, which have not appeared in smaller, less comprehensive corpora.

Appendix: List of Newspapers in the Spanish Corpus

The following 51 newspapers comprise my 35-million-word corpus of Spanish-language Web-based newspapers. There are 15 million words from the seven newspapers in Spain and 1 million words from each of the Spanish-speaking countries in the Americas (for a total of 20 million words in this corpus). All of the URLs in parentheses are correct as of October 1999; those in square brackets were no longer functioning as of this date, and no functioning URL could be found.

Argentina
La Nueva Provincia (www.lanueva.com.ar)
InterVoz (www.intervoz.com.ar)

Bolivia
Los Tiempos (www.lostiempos.com)
Agencia de Noticias ERBOL (jaguar.pg.cc.md.us/diario.html)

Chile
El Mercurio (www.mercurio.cl)
Hoy [www.hoy.web.cl]

Colombia
El Espectador (www.elespectador.com)
La Semana (www.semana.com.co)

Costa Rica
La Nación (www.nacion.co.cr)
Prensa Libre (www.prensalibre.co.cr/plibre.html)

Cuba
CubaNet [www.netpoint.net/~cubanet]
Trabajadores (www.trabajadores.cubaweb.cu)
Granma [206.130.183.236]

Ecuador
Estadio [www.telconet.net/estadio]
El Vistazo (www4.vistazo.com.ec)
Universo (www.eluniverso.com)

El Salvador
El Diario (www.elsalvador.com)
La Prensa Grafica (www.laprensa.com.sv)

Guatemala
Prensa Libre (www.prensalibre.com.gt)
La Gerencia (www.nortropic.com/gerencia)
La Hora (www.lahora.com.gt)

Honduras
La Tribuna (www.latribuna.hn)
La Prensa (www.laprensahn.com)

Mexico
Excelsior (www.excelsior.com.mx)
Diario Yucatán (www.yucatan.com.mx)

Nicaragua
La Tribuna [www.latribuna.com.ni]
La Prensa [www.sgc.com.ni/laprensa]

Panama
El Siglo (www.elsiglo.com)
La Prensa (www.prensa.com)

Paraguay
ABC (www.una.py/sitios/abc)
Diario Noticias Online (www.diarionoticias.com.py)

Perú
Caretas (www.caretas.com.pe)
El Tiempo (www.eltiempo.com.pe)

Puerto Rico
Noticentro (noticentro.coqui.net/pp.htm)
El Nuevo Día (www.endi.com)

República Dominicana
Ultima Hora (www.ultimahora.com.do)
Listin Digital (www.listin.com.do)

Spain
ABC (abc.es)
El Pais (www.elpais.es)
Diario AS (www.diario- as.es)
El Periodico (www.elperiodico.es)
La Vanguardia (www.vanguardia.es)
El Mundo de Baleares (www.el-mundo.es)
Diario Sur (www.diariosur.es)
El Comercio (www7.uniovi.es/noticias)

United States
Miama Herald (www.elherald.com)
Diaro de las Americas (www.diariolasamericas.com)

Uruguay
Diario El Pais (www.diarioelpais.com)
Brecha [www.chasque.apc.org/brecha]

Venezuela
El Universal (Digital (www.eud.com))
El Carabobeno (www.el-carabobeno.com)

References

Biber, Douglas. 1990. Methodological Issues regarding Corpus-Based Analyses of Linguistic Variation. *Literary and Linguistic Computing* 5:257–69.

———. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing* 8:243–57.

Davies, Mark. 1992. The Diachronic Evolution of Causative Constructions in Spanish and Portuguese. Ph.D. diss., University of Texas at Austin.

———. 1995a. Analyzing Syntactic Variation with Computer-Based Corpora: The Case of Modern Spanish Clitic Climbing. *Hispania* 78:370–80.

———. 1995b. The Evolution of the Spanish Causative Construction. *Hispanic Review* 63:57–77.

———. 1996. The Diachronic Evolution of the Causative Construction in Portuguese. *Journal of Hispanic Philology* 17:261–92.

———. 1997a. A Corpus-Based Analysis of Spanish Subject Raising in Modern Spanish. *Hispanic Linguistics* 9:33–63.

———. 1997b. The Evolution of Subject Raising in Spanish. *Bulletin of Hispanic Studies* (Liverpool) 74:399–411.

———. 1998. The Evolution of Spanish Clitic Climbing: A Corpus-Based Approach. *Studia Neophilologica* 69:251–63.

DeMello, George. 1995. Preposicion + sujeto + infinitivo: "Para yo hacerlo." *Hispania* 78:825–36.

Finnemann, David. 1982. Aspects of the Spanish Causative Construction. Ph.D. diss., University of Minnesota.

Lipski, John M. 1991. In Search of the Spanish Personal Infinitive. In *New Analyses in Romance Linguistics*, ed. Dieter Wanner et al., 201–20. Amsterdam: Benjamins.

Reider, Michael. 1993. On Tough Movement in Spanish. *Hispania* 76:160–70.

Suñer, Margarita. 1986. Lexical Subjects of Infinitives in Caribbean Spanish. In *Studies in Romance Linguistics*, ed. Osvaldo Jaeggli et al., 189–203. Dordrecht: Foris.

Contents

Introduction: North American Perspectives on Corpus Linguistics at the Millennium 1
Rita C. Simpson and John M. Swales

Part 1. Corpus Building and Tools

The International Corpus of English: Progress and Prospects 17
Charles F. Meyer

Collaboration between Corpus Linguists and Digital Librarians for the MICASE Web Search Interface 32
Christina Powell and Rita C. Simpson

Representing Spoken Language in University Settings: The Design and Construction of the Spoken Component of the T2K-SWAL Corpus 48
Douglas Biber, Randi Reppen, Victoria Clark, and Jenia Walter

Creating and Using Multimillion-Word Corpora from Web-Based Newspapers 58
Mark Davies

Concordance Programs for Corpus Linguistics 76
Susan Hockey

Part 2. Corpus-Based Analyses and Applications

Using Corpus-Based Methods to Investigate Grammar and Use: Some Case Studies on the Use of Verbs in English 101
Douglas Biber

Discovering the Usual with Corpora: The Case of *Remember* 116
Hongyin Tao

Discourse Management and New-Episode Flags in MICASE 145
John M. Swales and Bonnie Malczewski

Reflexive Academic Talk: Observations from MICASE 165
Anna Mauranen

Rethinking French Grammar for Pedagogy: The Contribution of Spoken Corpora 179
Aaron Lawson

The Lexical Phrase as Pedagogical Tool: Teaching Disagreement Strategies in ESL 195
Stephanie Burdine

Writing Development among Elementary Students: Corpus-Based Perspectives 211
Randi Reppen

Glossary for Part 1 227

Contributors 233

Subject Index 235

Author Index 239

Introduction: North American Perspectives on Corpus Linguistics at the Millennium

Rita C. Simpson and John M. Swales
University of Michigan

Corpus linguistics is essentially a technology, but like many technologies, it may have, at least potentially, considerable consequences. After all, the telescope transformed astronomy, the X-ray machine radicalized medicine, the tape recorder impelled the advance of sociolinguistics and the study of oral discourse, the video recorder advanced the study of small-group interactions, and the spectrograph (and similar devices) consolidated the development of instrumental phonetics. Corpus linguistics technology requires a computer that can store a collection of text files (the **corpus**) and then apply software to those files to produce frequency lists, lists of key words, and, most importantly, strings of words showing which words co-occur (or **collocate**) with others. The text files in a corpus may consist entirely of written texts (as in the Helsinki Corpus of English Texts), entirely of transcriptions of speech (as in MICASE—the Michigan Corpus of Academic spoken English), or of both (as in the Bank of English).[1] These corpora are typically constructed on certain principles that lead to appropriate sampling, and they can vary greatly in size. The Bank of English corpus, an earlier stage of which underpinned the important corpus-based *COBUILD English Language Dictionary*,[2] is huge and, at the time of writing, rapidly approaching 400 million words; small specialized corpora, especially those devoted to single genres, such as research articles or university lectures, can be orders of magnitude smaller. The pros and cons of large diffuse corpora and small narrow ones is a matter of current debate.

The range of corpus-based research currently taking place is both