Student use of large, annotated corpora to analyze syntactic variation

Mark Davies
Brigham Young University, USA

This study discusses the way in which advanced language learners have used several large corpora of Spanish to investigate a wide range of phenomena involving syntactic variation in Spanish. The course under discussion is taught via the Internet, and is designed around a textbook that contains both descriptive and prescriptive rules of Spanish syntax. The students carry out studies for those syntactic phenomena in which there is supposedly variation – either between registers, between dialects, or where there is currently a historical change underway. The three main corpora used by the students are the 100 million word Corpus del Español (which I have created), the CREA corpus from the Real Academia Española, and searches of the Web via Google. The students have found that each of the three large corpora has its own weaknesses, and that the most effective strategy is to leverage the strengths of each corpus to find the desired data. The students also learn strategies for comparing results across different corpora, and even within components of the same corpus – such as the frequency of occurrences on the Web from different countries.

1. Introduction

A goal of many language learners is to more fully understand the range of syntactic variation in the second language and thus move beyond the simplistic rules that are presented in many textbooks. This effort can be aided by large corpora of the second language, which allow users to easily and quickly extract hundreds or thousands of examples of competing syntactic constructions from millions of words of text in different dialects and registers. Using this data, the teachers and students can then have a more realistic picture of how the constructions in question vary from one country to another, whether they are

more common in formal or informal speech, and whether their use is increasing or decreasing over time.

This paper provides an overview of the way in which students in a recent online course used three very large sets of corpora – involving hundreds of millions of words of text – to study "Variation in Spanish Syntax". We will examine the goals of the course, how the students carried out their research using several corpora, the way in which they analyzed competing structures from the corpora, and how they were ultimately successful in describing syntactic variation in Spanish.

The issue of how students can be trained as corpus researchers to learn about a foreign language has been the focus of a number of recent articles, including Bernardini (2000, 2002), Davies (2000), Osborne (2000), Kennedy and Miceli (2001, 2002), Kirk (2002), and Kübler and Foucou (2002). Although the issue of "student as researcher" is the focus of our study as well, it differs from many of the previous studies in several ways – the corpus used in this course is much larger (hundreds of millions of words of text), the students are more advanced (graduate students, with many of them language teachers themselves), the focus is on variation rather than norms, and the linguistic phenomena studied in the course deal with complex syntactic constructions, rather than simple collocations. Yet the additional data from our course should help to provide a more complete picture of the different ways in which students can use large corpora to study and analyze the grammar of the second language.

In terms of the organization of the course, the "Variation in Spanish Syntax" class that we will be discussing was offered for the first time in 2002, and was taught online to twenty language teachers from throughout the United States (see http://davies.linguistics.byu.edu/sintaxis). Each week the students would examine two to four chapters in *A New Reference Grammar of Modern Spanish* (Butt and Benjamin, 2000), which is an extremely complete reference grammar of Spanish. Table 1 below lists the primary topics for each week's assignments.

After reading the assigned chapters from the reference grammar, each student would identify a particular syntactic construction from among the topics for that week, for which Butt and Benjamin indicated there was some type of variation – between geographical dialects, speech registers, or an overall increase or decrease in the use of the construction. Once the students had identified their topic of study, they would then spend the week using three different sets of corpora or web-based search engines to search for data on the

Table 1. Topics in the "Variation in Spanish Syntax" course

Week	Topic	Week	Topic
1	Morphology: gender and plurals	8	Progressive, gerund, participles
2	Articles, adjectives, numbers	9	Subjunctive, imperative, conditionals
3	Demonstratives, *lo*, possessives	10	Infinitives, auxiliaries
4	Prepositions, conjunctions	11	*Ser/estar*, existential sentences
5	Pronouns: subject and objects	12	Negation, adverbs, time clauses
6	Pronominal verbs, passives, impersonals	13	Questions, relative pronouns and clauses
7	Indicative verb tenses	14	Cleft sentences, word order

constructions in question. The corpora were the Corpus del Español (www.corpusdelespanol.org), the CREA and CORDE corpora from the Real Academia Española (www.rae.es), and the web through the Google and Google Groups search engines (www.google.com, groups.google.com) – all of which will be discussed below.

Once they had extracted sufficient data for the construction, they would then write a short summary for that week's project. In this summary they would point out the four most important findings from the data, summarize how their findings confirmed or contradicted the claims of Butt and Benjamin regarding variation, and then briefly discuss some of the possible motivations for this syntactic variation. The final step for each project, which was completed the following week, was to then review the projects of three other students.

By the end of the semester, each student had carried out fairly in-depth research on fifteen different syntactic constructions involving variation in Modern Spanish, and in addition each student had reviewed another forty-five studies by other students. Based on the quality of their projects, it seems clear that these corpus-based activities were extremely valuable in helping the students to move beyond simplistic textbook descriptions of Spanish grammar, and to acquire a much better sense of the actual variation in contemporary Spanish syntax.

2. The corpora

The corpora were the foundation for the entire course, and therefore an understanding of the composition and features of each corpus is fundamental to understanding how the students carried out their research.

2.1 The Corpus del Español

The primary corpus for the course was the 100 million word Corpus del Español that I have created, and which was placed online shortly before the start of the semester. The Corpus del Español has a powerful search engine and unique database architecture that allow the wide range of queries shown in Table 2. These include pattern matching (1), collocations (2), lemma and part of speech for nearly 200,000 separate word forms (3), synonyms and antonyms for more than 30,000 different lemmas (4), more complex searches using combinations of the preceding types of searches (5), queries based on the frequency of the construction in different historical periods and registers of Modern Spanish (6), and queries involving customized, user-defined lists (7). Note also that it would take only about 1–2 seconds to run any of these queries against the complete 100 million word corpus.

In short, the Corpus del Español is richly annotated and allows searches for many types of linguistic phenomena, which made it extremely useful for the wide range of constructions that were studied in the "Variation in Spanish Syntax" course.

2.2 CREA / CORDE and Google (Groups)

In addition to the 100 million word Corpus del Español, the students used two other sets of Spanish corpora. The first set is the CREA (Modern Spanish) and CORDE (Historical Spanish) corpora from the Real Academia Española, which contain a combined total of about 200 million words of text. The second set are the Google and Google Groups search engines. While these search engines are of course not limited just to web pages in Spanish, the main Google index covers more than 100 million words of text in Spanish language web pages, while the Google Groups search engine contains millions or tens of millions of words of text in messages to Spanish newsgroups.

Table 2. Range of searches possible with the "Corpus del Español"

1	est_b* *ndo	estaba cantando, estábamos diciendo
2	lo * possible	"as * as possible" *lo mejor posible, lo antes posible, lo máximo posible*
3	poder.* *.v_inf	forms of *poder* ("to be able") + infinitive *puede tener, pudiera escapar*
4	!dificil.*	all forms of all synonyms of *difícil* "difficult" *imposible, duros, compleja, complicadas, . . .*
5	estar.* !cansado.* *.prep *.v_inf	any form of *estar* + any form of any synonym of *cansado* ("tired") + preposition + infinitive *estoy harto de vivir, estaba cansada de escuchar*
6	*.adv {19misc>5 19oral=0}	all adverbs that occur more than five times in newspapers or encyclopedias from the 1900s, but not in spoken texts from the 1900s *inversamente, clínicamente*
7	le/les [Bill.Jones:causative] *rse	*le* or *les* + a customized list of [causative] verbs created by [Bill.Jones] + words ending in [-rse] *le mandaban ponerse, les hace sentirse*

The main advantage that CREA/CORDE and Google (Groups) have over the Corpus del Español is their ability to limit searches to specific countries. As we will see in Section 3.3, this is useful for students who want to look at the relative frequency of constructions in different geographical dialects. In addition, CREA and CORDE allow users to compare the relative frequency of constructions in several different registers, beyond just the three divisions of the Corpus del Español (literature, spoken, newspaper/encyclopædias).

The main disadvantage of CREA/CORDE and Google (Groups) is that they are not annotated in any way, which makes it impossible to search for syntactic constructions using lemma or part of speech. Even the wildcard features of both sets of search engines are rather limited, which means that it is also impossible to look for morphologically similar parts of speech or lemma. Both CREA/CORDE and Google (Groups) are really only useful in searching for exact phrases. However, as we will see in Section 3.3, when they are used in

conjunction with the highly annotated Corpus del Español, the overall collection of corpora does permit students to carry out detailed searches on syntactic variation for a very wide range of constructions in Spanish.

3. Student outcomes: Examining syntactic variation through corpus use

In this section, we will consider some of the challenges that the students faced in using the corpora effectively, how they overcame these obstacles, and how they were ultimately successful in carrying out more advanced research in Spanish and thus moving beyond the simplistic rules of many introductory and intermediate level textbooks. In the sections that follow, we will use as an example the question of clitic placement, in which the clitic can either be pre-posed or post-posed (*lo quiero hacer* vs. *quiero hacerlo*; "I want to do it"), and which exhibits variation that is dependent on dialect, register, and several functional factors (cf. Davies 1995, 1998).

3.1 Learning to use the corpora

The first challenge facing the students was simply learning to use the different corpora successfully, in order to limit searches and extract the desired information. To help the students, during the first week of class they spent several hours completing two rather lengthy "scavenger hunt" quizzes using the Corpus del Español, the CREA and CORDE corpora, and Google (Groups). Each question would outline the type of data that they would look for from a particular corpus. For example, one of the questions dealing with the Corpus del Español asked them to look for the most common phrases involving any object pronoun followed by any form of any synonym of *querer* ("to want") followed by an infinitive.

The students would be responsible for looking at the help file to see how to search for lemma and parts of speech, and would then hopefully use the correct search syntax to query the corpus, in this case [*.pn_obj !querer.* *.v_inf].

As a hint to make sure that the students were on the right track in their search, they were told what the first two entries in the results set would be (e.g., *te quiero decir*, *le quiero decir*), and they could use this to check their results. They were then responsible for providing the third entry from the list (in this case, *me quiere decir*). By the time they had answered sixty such questions for the different corpora during the first week, they were ready to use the corpora to extract data on syntactic constructions of their choosing.

By examining Table 1 we can see that the topics in the course were arranged so that the students would start with constructions that were relatively easy to find, and that as the semester progressed became gradually more abstract and difficult. For example, at the beginning of the semester students started at the word-internal level (morphological variation for gender and number), and then moved to constructions involving adjacent words (e.g., demonstratives and possessives), then semantically more complex localized constructions (e.g., pronominal verbs), then even less well-defined structures (e.g., subjunctives), and ending up with fairly abstract and less localized constructions (e.g., cleft sentences and word order).

3.2 Formulating the research question

One of the hardest parts of carrying out linguistic research is knowing how to frame the question, and setting up the actual search of the corpora. To help the students in the course, during the first four weeks of the course I required that they send me a *Plan de Trabajo* (Work Plan) before they started the research in earnest. In a short paragraph they would first briefly outline what type of variation was described in the reference grammar. They would then indicate which corpora would be most useful to examine the variation, and show exactly what type of queries would be run against these corpora to extract the data.

Sometimes there were problems with the general research question – for example, the topic was much too wide or too narrow. Returning to the clitic placement construction, for example, they might propose to look at clitic placement with all main verbs (too wide) or with just three or four exact phrases (too narrow). Sometimes they intended to use a corpus that was not the best one for the question at hand, or else they had the wrong search syntax. In all such cases, I would help them frame the search correctly before they started. Once they had received this feedback, they were then ready to start the search itself. This procedure seemed to prevent a lot of wasted time and frustration on the part of the students as they were learning to use the corpora. By the end of the fourth week of using corpora, however, most of the students had sufficient experience in framing the research questions and setting up the queries, and it then became optional to submit a work plan before consulting the corpora.

3.3 Extracting data from multiple corpora

The students soon learned that the most productive research was that which incorporated searches from all three sets of corpora, by using each of the corpora for those purposes for which they were most useful. Typically, the students would start searching with the Corpus del Español, because it is the only one of the three that was annotated. For example, if they were examining variation in clitic placement, they could search for all cases of pre-positioning ("clitic climbing") (1a) or postpositioning (1b) with all of the synonyms of a particular verb:

(1a) [*.pn_obj !querer.* *.v_inf] lo quiero hacer, me preferían hablar
(1b) [!querer.* *.v_inf_cl] quiero hacerlo, preferían hablarme
"I want to do it, he preferred to talk to me"

The students would see all of the matching phrases for both constructions and could easily compare the relative frequency across historical periods – to see whether one construction or the other is increasing over time – and they could also compare the relative frequency in the three general registers of literature, spoken, and newspapers / encyclopaedias.

In order to carry out even more detailed investigations of register or dialectal variation, however, the students often turned to the CREA/CORDE or Google (Groups) corpora. Because these corpora only allow searches of exact words and phrases, however, the student would need to select individual phrases from the lists generated in the Corpus del Español (e.g., *lo quiero hacer* vs. *quiero hacerlo*; "I want to do it"), and then search for these individual phrases one by one. Although somewhat cumbersome, this would allow students to compare the relative frequency of specific phrases in more than twenty Spanish-speaking countries and (in the case of CREA/CORDE), in a much wider range of register subdivisions than in the Corpus del Español. The two-step process – starting with the Corpus del Español and then using its data (when necessary) to search for individual phrases in CREA/CORDE and Google (Groups) – consistently yielded the best results for the students.

3.4 Organizing the data

After running the queries on the different corpora, the next step was to organize the data so that it would confirm or deny Butt and Benjamin's claims about syntactic variation in Spanish. At the beginning, this was rather difficult for some students. They might examine two competing syntactic constructions in different geographical dialects of Spanish – for example [**clitic** + main verb + infinitive] ("Type A") vs. [main verb + infinitive + **clitic**] ("Type B"). In their attempt to show whether Type A or Type B was more common in different dialects, they might discover that CREA or Google had many more examples of Type A from Spain than from Mexico or Argentina. Inexperienced students might interpret this to mean that Type A was more common in Spain than in Mexico or Argentina, without realizing that there are more examples from Spain simply because the textual corpus from Spain is so much larger than that of other countries. Of course, the issue is not the relative frequency of Type A in Spain vs. the relative frequency of Type A in Mexico or Argentina, but rather the relative frequency of Type A vs. Type B in each of these three countries. Once students learned to correctly use relative frequencies to compare geographical dialects, different registers, or different historical periods, they were on much firmer footing as regards making valid judgments about the data.

3.5 Drawing conclusions regarding variation

Once the data had been organized correctly, students were responsible for summarizing the most important findings from the data, and for suggesting whether the data confirmed or denied the original hypothesis regarding variation with the particular syntactic construction. During the first two or three weeks, students found it very difficult to clearly and concisely summarize the findings, and instead preferred to hope that quantity equaled quality. Therefore, starting in the third week I limited them to only four short sentences to explain the major findings from the data. In addition, they were asked to include a two or three sentence conclusion, which showed whether the four points just mentioned confirmed or denied the hypothesis from Butt and Benjamin regarding syntactic variation. My sense was that this "bottom line assessment" was very useful in helping them to organize their data collection and written summary.

3.6 "Explaining" the syntactic variation

If students were able to accomplish all of the proceeding tasks, they were judged to have been successful in carrying out the research. Starting in about the fifth week, however, they were presented with an additional goal, which was to suggest possible motivations for the syntactic variation – whether it was geographical, register-based, or diachronic in nature. The second textbook that was required for the course – in addition to the reference grammar – was *Spanish-English Contrasts* (Whitley 2002), which is an overview of recent research on a wide range of syntactic constructions in Spanish. To the extent possible, the students were asked to use any of the more theory-based explanations in Whitley for the particular construction in question, and see whether this might be useful in helping to "explain" variation. For example, with the clitic placement construction, they might realize that placement is a function of the semantics of the main verb, with semantically light verbs allowing clitic climbing more often (cf. Davies 1995, 1998). In some cases students were able to identify possible causal factors, but in other cases it was simply sufficient to point out the actual variation and leave it at that. Even in these cases, however, the data that they presented was often more complete than that found in previous studies by much more accomplished researchers, simply because of the size and power of the corpora that were available to the students in the course.

4. Conclusions

As was explained in the introduction, there has been recent interest in the way in which students can be "trained" as corpus linguists to extract data from the foreign language. Many of these studies have focused on intermediate level students looking for "correct rules" for simple constructions in relatively small corpora. In the course that we have described, however, the graduate-level students focused on variation from the norm with rather complex constructions in hundreds of millions of words of data.

In spite of the differences between this course and those described in previous studies, the hope is that our experience might provide insight into how students can use corpora to perform advanced research on the foreign language. As we have seen, if there is proper guidance and feedback, even students who are relatively inexperienced in syntactic research can be trained to use the

corpora, formulate research questions and search strategies, organize the data, confirm or deny previous claims about language variation, and perhaps even begin to find motivations for this variation. In accomplishing these goals, these students have been successful in moving beyond the simplistic, prescriptivist rules found in many textbooks, and have begun to use corpora to acquire a much more accurate view of the syntactic complexity of the foreign language.

References

Bernardini, S. 2000. "Systematising serendipity: Proposals for large-corpora concordancing with language learners". In Burnard and McEnery, 225–234.

Bernardini, S. 2002. "Exploring new directions for discovery learning". In Ketteman and Marko, 165–182.

Burnard, L. and McEnery, T. (eds). 2000. *Rethinking Language Pedagogy from a Corpus Perspective* [Łódź Studies in Language, Vol. 2]. Frankfurt am Main: Peter Lang.

Butt, J. and Benjamin, C. 2000. *A New Reference Grammar of Modern Spanish*. 3rd edition. Chicago: McGraw-Hill.

Davies, M. 1995. "Analyzing syntactic variation with computer-based corpora: The case of modern Spanish clitic climbing". *Hispania* 78:370–380.

Davies, M. 1998. "The evolution of Spanish clitic climbing: A corpus-based approach". *Studia Neophilologica* 69:251–263.

Davies, M. 2000. "Using multi-million word corpora of historical and dialectal Spanish texts to teach advanced courses in Spanish linguistics". In Burnard and McEnery, 173–186.

Kennedy, C. and Miceli, T. 2001. "An evaluation of intermediate students' approaches to corpus investigation". *Language Learning and Technology* 5:77–90.

Kennedy, C. and Miceli, T. 2002. "The CWIC project: Developing and using a corpus for intermediate Italian students". In Ketteman and Marko, 183–192.

Kettemann, B. and Marko, G. 2002. *Teaching and Learning by Doing Corpus Analysis (Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19–24 July, 2000)*. Amsterdam: Rodopi.

Kirk, J. 2002. "Teaching critical skills in corpus linguistics using the BNC". In Ketteman and Marko, 155–164.

Kübler, N. and Foucou, P-Y. 2002. "Linguistic concerns in teaching with language corpora. Learner corpora". In Ketteman and Marko, 193–203.

Osborne, J. 2000. "What can students learn from a corpus? Building bridges between data and explanation". In Burnard and McEnery, 165–172.

Whitley, M.S. 2002. *Spanish-English Contrasts*. Washington, D.C: Georgetown University Press.