Using multi-million word corpora of historical and dialectal Spanish texts to teach "Advanced Spanish Syntax"

Mark Davies, Dept. of Foreign Languages, Illinois State University

During the past six or seven years I have created a 5,300,000 word corpus of 118 Spanish texts from 1200-1900 (see list online at [http://138.87.135.33/~mdavies/espanol.htm]), as well as a 7,700,000 word corpus of spoken and written Modern Spanish from 11 different countries (see list online at [http://138.87.135.33/~mdavies/texts.htm]). These represent some of the largest and most diverse corpora currently available for Spanish.

In the past three years, I have begun to use these corpora in an advanced syntax course at Illinois State University that deals with "Approaches to Spanish Syntax". I believe that the corpora have been very useful for the students in helping them to accurately model historical shifts in Spanish syntax, as well determine the extent and nature of dialectal and register variation in Modern Spanish.

In this course, we look at 8-10 different syntax phenomena (causatives, subject raising, clitic climbing, use of the "reflexive" marker *se*, etc.) from within a number of competing syntactic theories (classical TG, GB, Relational Grammar, typological-functional grammar, etc.) In addition to introducing the students to "theory-based" analyses of the syntactic phenomena, I also ask them to carry out original "corpus-based" research on these phenomena to see whether the data are as "clean" and "predictable" as the theory-based studies suggest they are.

To allow the students to carry out this original research, I allow them to use the historical and modern Spanish corpora that I have created, which are based on my NT server but have been made accessible to the students via the departmental LAN. All of the texts in these corpora have been converted to electronic form (most of them by scanning in the texts), and then have been converted to Word Cruncher format. This creates an "every word" index of the corpora, which then allow complex proximity and Boolean searches of the texts. I also create and place on the server a limited number of Word Cruncher format "retrieved lists" that contain exhaustive indices of certain parts of speech or certain lexical items (eg. 240,000 cases of infinitives in the Modern Spanish corpus, which would be time-consuming for the student themselves to create).

Using these corpora, the students are then able to map quite nicely historical shifts in Spanish syntax and variation between dialects and registers in Modern Spanish. For example, the last time the course was taught the students extracted and analyzed data on some of the following topics, among others:

v the emergence of reflexive markers with causative constructions (*lo dejaron lavarse* "they let him bathe himself") in the last 300-400 years, an indicator of a shift towards "biclausal" structure

v the rise of verbs with "inherent" reflexive markers (e.g. *jactarse, arrepentirse* de *"to boast, to repent"), and how this relates to a "unified theory" of pronominal verbs in Spanish

v whether there are any cases of *gustar* (to be pleasing) with reflexive pronouns either in the historical or Modern Spanish corpus (*Juan se gusta* "John pleases himself"), and how the presence or absence of such constructions relates to the underlying structure of "psych verbs"

v historical shifts and dialectal and register variation in Modern Spanish in the use of "multiple clitics" (*me lo, te las, me te, nos le*, etc.), as well as clitic doubling (*le* dimos el libro a María "we gave the book to Mary"), and how these two phenomena relate to the hypothesis that clitics are becoming less "word-like" and more like "affix-like" in Spanish

Because of the every-word index that Word Cruncher creates and because of its ability to carry out complex proximity and Boolean searches, the students are able to very easily extract large amounts of data in a relatively short period of time. For example, it takes less that two seconds to find 6600+ likely cases or the verb *gustar* "to like" in the Modern Spanish corpus, another second or two to find the 105,000+ cases of the reflexive pronoun *se*, and then just 2/10 of a second to combine these two lists to find the cases of e.g. *se gusta* "he pleases himself" (examples which do exist and which present a challenge to the theory-based analysis of these constructions). In addition, because Word Cruncher allows distributional analysis of the results, the students can easily determine how the data is mapped out over the past 800 years, and how it varies in Modern Spanish, depending on the particular Spanish-speaking country and whether the text is from the spoken or written register.

Of course extracting the data is only the first step. The students then have to determine whether the data is in accordance with the predictions of the "theory-based" analyses, and how the theory may need to be modified to account for new data. This is especially true in the case of the historical plane, since many of the "theory-based" analyses are oriented much more towards the Modern Spanish data, but have problems in accounting for and explaining historical change.

When I teach the course again in Fall 1998, I plan on incorporating into the "corpus linguistics" part of the course one additional resources that is in the public domain. I will have them perform searches on 100,000,000 past newgroup posts (only a small portion of these being in Spanish, of course) using the extremely powerful search engine at DejaNews (www.dejanews.com). While it would not be possible to perform searches my grammatical category or variant forms of a given lexical item (e.g. all infinitives or all forms of a given verb), it will still be possible to carry out proximity and Boolean searches on this extremely large corpus of (often informal) written Spanish.

In summary, the electronic corpora allow advanced Spanish students to extract large amounts of data from the 13,000,000+ words of text in the historical and Modern Spanish corpora in a relatively quick and simple manner. This allows them to model historical changes and current syntactic variation much easier than would have been thinkable even 10-15 years ago.