

# Student Use of Large Corpora to Investigate Language Change

Mark Davies

Brigham Young University

## Abstract

*The use of corpora in historical linguistics courses is an idea whose time has come, but it is a topic that has received scant attention in previous studies. In this paper I examine the way in which students have used large corpora as a fundamental part of an online "History of the Spanish Language" course. These corpora include a parallel corpus of the entire Bible in Late Latin, Old Spanish, and Modern Spanish, which allows students to compare many different linguistic structures across these three languages. The main corpus used in the course is the recently-completed "Corpus del Español" – a web-based, 100 million word, fully-annotated corpus of Spanish texts from the 1200s-1900s. This corpus allows even beginning students of historical linguistics to quickly and easily extract data for a wide range of linguistic phenomena, and thus move beyond the simplistic memorization of "historical rules" that are found in many textbooks.*

## 1 Introduction

Most research on the use of corpora in the classroom deals with using corpora to provide non-native speakers with a database of authentic language data (see the articles from the TALC proceedings: Botley et al. 1996; Wichman et al. 1997; Burnard and McEnery 2000; Kettemann and Marko 2002). Because the goal deals with language learning by foreign speakers, the focus is obviously on the modern, synchronic stage of the language. In this study, however, I will discuss how language corpora can be used in quite a different sphere of teaching— that of historical linguistics.

The use of large, electronic corpora in teaching historical linguistics is still rather uncommon. Of course there have been many valuable applications of corpora methodology to examining problems of historical linguistics (for example, Rissanen 1992, 1993, 1997a,b for English, among many others). Nevertheless, a review of the literature shows only a handful of articles and presentations dealing with the pedagogical use of these materials in the classroom, such as Schmied (1996), Knowles (1997), Davies (2000), and Curzan (2000). This lack of research is unfortunate when one recognizes that the use of corpora in the teaching of historical linguistics can significantly enhance the learning process, as much (or more) as the use of corpora in learner-oriented and synchronically-based courses.

Traditionally, courses in historical linguistics focus on rather abstract rules governing changes in the phonetic, morphological, syntactic, or semantic

structure of the language in question. The students are responsible for memorizing a long list of rules, and perhaps supplying one or two samples of each type of linguistic change. For example, they might include one or two words that have undergone a particular phonetic shift, or one or two sample sentences showing the “before” and “after” stages of a grammatical shift in the language.

By using large corpora, however, the students can truly immerse themselves in the data and – by themselves – find new and interesting examples of linguistic change. Depending on the corpus they are using, it may be possible to extract hundreds or thousands of examples of a particular linguistic shift in a very short period of time. This large amount of data can then be used to model linguistic change much more precisely and accurately than had been done by even the best researchers, previous to the use of large electronic corpora. This is very empowering for the students, as they can easily and accurately use data to test the textbook rules for a particular linguistic shift. In essence, even advanced undergraduates or beginning graduate students can use the corpora to add valuable insight into what is known about the evolution of a particular language.

## 2 “History of the Spanish Language”

Previous studies such as Knowles (1997) and Curzan (2000) are in part “how to manuals,” discussing concrete ways in which corpora have been used in actual courses in historical linguistics. Both of these studies, however, deal just with English. In the present study, I will expand the focus somewhat and look at several different ways in which corpora have been used to teach a “History of the Spanish Language” course that has been offered by Illinois State University (<http://davies-linguistics.byu.edu/hispan>).

In addition to its strong reliance on corpus-based investigation, this “History of the Spanish Language” course is also unique in terms of its method of delivery. Although originally offered as an in-classroom course, since Spring 2000 it has been offered as an online course, and has been taught entirely via the Web. The lack of traditional classroom interaction was in fact one of the reasons for using large corpora. If the class had been offered in a traditional setting, we could have memorized the different types of linguistic change in Spanish, and the students would have been responsible for duplicating these on the test. There would have also been opportunity for the students to ask questions about the changes, and receive feedback from the professor in areas where clarification was needed or desired.

By teaching the class entirely via distance education, the dynamics of the class were altered dramatically. There would be much less opportunity for the traditional “give and take” of the classroom setting, which meant that the students themselves would be more responsible for internalizing the data. In addition, because the class is offered as a distance education course, there are problematic issues regarding the administration of tests and test security. For this reason, it was decided that student projects would form the basis of the evaluation.

Once the decision was made to focus on projects – rather than the rote memorization and recitation of rules – it was obvious that the students would need to have access to a well-built and highly usable database of historical texts, in order to extract the needed data. In subsequent sections, I will focus on the specific corpora that have been used in the class, and the way that they have been used by students to examine and model several different types of linguistic change. First, however, let us briefly consider the basic structure of the class.

Table 1: Course topics

THE EARLIEST STAGES	QUESTIONS	PROJECTS
1. Introduction		0
2. Pre-romanic languages		0
3. Indo-European		0
4. Latin: External		0
5. Latin: Internal		0
6. Vulgar Latin and the Romance languages		0
7. The Visigoths		0
8. The Arabs		0
<b>LATIN &gt; MEDIEVAL SPANISH</b>		
9. Phonetic		0
10. Morphosyntax		0
11. Lexicon		0
<b>MEDIEVAL SPANISH</b>		
12. Medieval Spanish dialects		0
13. Medieval texts		0
14. The language c1250-1450		0
<b>MEDIEVAL &gt; MODERN SPANISH (INTERNAL)</b>		
15. Phonetic		0
16. Orthography		0
17. Morphology		0
18. Syntax		0
19. Lexicon		0
<b>MODERN SPANISH (EXTERNAL)</b>		
20. The language c1475-1700		0
21. Spanish in the Americas		0
22. Other modern dialects		0
23. The future of Spanish		0

### 3 Overview of Course Topics and Organization

The “History of the Spanish Language” course covers a wide range of topics, dealing both with language-internal as well as external factors. Table 1 shows the twenty-three topics that receive primary focus during the course. As can be seen in this table, there are two different types of activities in the course. For the topics that are more “language-internal” in nature, there are corpus-based projects. For the “external” topics, there are a number of activities that are somewhat more traditional in nature. These involve readings and selected essay-type questions, which are submitted and evaluated via the class website. Even with some of these topics, however, there is an attempt to use a simple corpus-based approach, wherever possible. For example, in the discussion of the medieval dialects, students are first presented with information on the major features distinguishing the dialects, and are then given a 200-300 word extracts from different “unlabeled” dialects and asked to identify the dialects, based on their linguistic features. Likewise, for the final topic – dealing with the present influence of other languages on Spanish – students are asked to use Google to find examples of English-based words in Spanish web pages.

In addition to these traditional “question and answer” activities, however, there are many corpus-based projects, and this is the focus of this paper. As we will see, the two major sets of corpora of historical Spanish are used to 1) investigate the relationship between different stages of the language, and 2) accurately model several different types of linguistic change in Spanish. In Section 4, I will discuss how the first goal was addressed in the use of the large parallel “Polyglot” Bible of Late Latin, Old Spanish, and Modern Spanish. In Sections 5 and 6, I will discuss the second goal, by considering the way in which large, multi-million word corpora of Spanish are used to map out linguistic change from one century to the next.

### 4 The Polyglot Bible

One of the difficulties in teaching a course in historical linguistics is the challenge of having students see the relationship between different stages of the language. One way to address this challenge is by having students study the same passage in a parallel corpus that contains the same text in different stages of the language. Perhaps the best text for this purpose is the Bible, which has been translated into most of the European languages several times since the Middle Ages. With this goal in mind, several years ago I placed online a “Polyglot Bible” that contains the entire Gospel of Luke (1150+ verses) in thirty different languages (see <http://davies-linguistics.byu.edu/polyglot>). In addition to the modern stages of many different Indo-European and non-Indo-European languages, it also contains older stages of English (Old English [1000s], Middle English [1300s], Early Modern English [1600s], and Present-Day English [1900s]) and Spanish (Old

Spanish [1200s], and Late Latin). Table 2 shows part of the story of the “Good Samaritan” (Luke 10:30-33) in the four stages of English:

Table 2: Polyglot/parallel corpus (stages of English)

CH:V	OE (1000s)	ME (1300s)	EME (1600s)	PDE (1900s)
10:30	<p>þa cweþ se hæled hine up besoende; Sum man ferde fram hierusalem to hiericho and becom on þa sceadan. þa hine bereafodon; and fintregodon hine; and forleton hine samucene:</p>	<p>sobli Jesus byholdende vp seide, sum man cam down fro ierusalem to Jericho, &amp; fel in to þeues, þe which also robbeden hym, &amp; woundis put in, wenten away, þe man left half quic</p>	<p>And Jesus answering said, A certain [man] went down from Jerusalem to Jericho, and fell among thieves, which stripped him of his raiment, and wounded [him], and departed, leaving [him] half dead.</p>	<p>In reply Jesus said: "A man was going down from Jerusalem to Jericho, when he fell into the hands of robbers. They stripped him of his clothes, beat him and went away, leaving him half dead.</p>
10:31	<p>þa gebyrode hit þæt sum sacerd ferde on þam yþcan wege and þa he þæt geseah he hine forbeh.</p>	<p>forsope it befel þæt sum prest cam down in þe same weite, &amp; hym seen, passede forþ</p>	<p>And by chance there came down a certain priest that way; and when he saw him, he passed by on the other side.</p>	<p>A priest happened to be going down the same road, and when he saw the man, he passed by on the other side.</p>
10:32	<p>and eallswa se diacon. þa he was wið þa stowe and þæt geseah he hinc eac forbeah;</p>	<p>Also forsope &amp; a dekne when he was biside þe place &amp; say hym, passede forþ</p>	<p>And likewise a Levite, when he was at the place, came and looked [on him], and passed by on the other side.</p>	<p>So too, a Levite, when he came to the place and saw him, passed by on the other side.</p>
10:33	<p>þa ferde sum samaritanisc man wið hine; þa he hine geseah þa wearð he mid mildheortnesse of er hine astyred</p>	<p>forsope sum samaritan makende iourney, cam biside þe weite, &amp; he seende hym, is stirid bi mercy</p>	<p>But a certain Samaritan, as he journeyed, came where he was: and when he saw him, he had compassion [on him].</p>	<p>But a Samaritan, as he traveled, came where the man was; and when he saw him, he took pity on him.</p>

The parallel text is a useful tool, in that it allows students and other users to see exactly the same text in different historical periods, and thus see quite clearly how the language has changed. A function of the usefulness of the online “Polyglot Bible” is the fact that the historical English corpus is currently being used as part of a number of “History of the English Language” courses throughout the world.

In the case of Spanish, the parallel text is not just for the 1150-verse Gospel of Luke, but rather it contains the text for nearly all of the Old and New Testaments – nearly 15,000 verses (see <http://davies-linguistics.byu.edu/span3>). Table 3 is a small selection, containing part of the “Good Samaritan” (Luke 10:30-33) in the three stages of Latin and Spanish.

Table 3: Polyglo/parallel corpus (stages of Latin/Spanish)

CH:V	LATIN	OLD SPANISH	MODERN SPANISH
10:30	suscipiens autem Iesus dixit homo quidam descendebat ab Hierusalem in Hiericho et incidit in latrones qui etiam despoliaverunt eum et plagis impositis abierunt semivivo relicto	Catando Ihesu Christo a suso, dixo: un ombre decendie de Iherusalem a Iherico, e cayo en ladrones, e despoiaron le, e firieron le; de hy dexaron le medio uiuo e fueron se.	Respondiendo Jesús dijo: --Cierta hombre descendía de Jerusalén a Jericó y cayó en manos de ladrones, quienes le despojaron de su ropa, le hirieron y se fueron, dejándole medio muerto.
10:31	accidit autem ut sacerdos quidam descenderet eadem via et viso illo praeteriuit	Acaceto que aquel mismo día un sacerdot passaua por aquella misma carrera, e quando l uiuo, passos e fue su uia.	Por casualidad, descendía cierto sacerdote por aquel camino; y al verte, pasó de largo.
10:32	similiter et Levita cum esset secus locum et videret eum pertransiit	E otrosi un leuita que passo cab el, quando l uiuo, fuesse adelant.	De igual manera, un levita también llegó al lugar; y al ir y verte, pasó de largo.
10:33	Samaritanus autem quidam iter faciens venit secus eum et videns eum misericordia motus est	E un samaritano que passaua por alli, quando l uiuo, fue mouido de piedat;	Pero cierto samaritano, que iba de viaje, llegó cerca de él; y al verte, fue movido a misericordia.

In addition to the inherent advantages of presenting the same text in parallel format, the online corpus also has the advantage of being searchable, and this allows students to perform a number of useful queries of the data. For example, one of the projects in the course is to find evidence for seven or eight of the major morphosyntactic changes from Late Latin to Old Spanish, such as the loss of nominal case, the creation of articles, the maintenance of specific verbal inflexions, the loss of others (e.g., future and passive), the creation of others (e.g., analytic perfect tenses), and negation. In this case, a student might investigate the disappearance of the synthetic Latin future (*facient*; “3PL will make”) and the emergence of the analytic Romance future (VL *facere habent* > OSP. *fazer* (*h*)an > ModSp *harán*). In examining this shift, students can search for a Modern Spanish form (e.g., *harán*), and in less than half a second they retrieve the 33 matching hits in the 15,000 verses of text (see, for example, Table 4).

Table 4: Searching the parallel corpus to compare constructions (Lat/OSP/MSP)

Text	LATIN	OLD SPANISH	MODERN SPANISH
Deut 25:2	sin autem peccavit dignum viderint prosterment coram se verberari pro peccati modus	eum qui aqueill lo prosterment se <u>facient</u> tender lan & ante <u>fazer lo an</u> acotar segunt erit et plagarum	mas si ellos vieren que aqueill lo prosterment se <u>facient</u> tender lan & ante <u>fazer lo an</u> acotar segunt erit et plagarum
			Sucedará que si el delincuente merece ser azotado, el juez lo hará recostar en el suelo y lo <u>harán azotar</u> en su presencia. El número de azotes será de acuerdo al delito.

Likewise, the assignment might require the student to find evidence for a particular linguistic shift from Old Spanish to Modern Spanish. For example, Modern Spanish often uses [ir + a + INF] to express the future (*va a cantar* “3SG is going to sing), whereas this was still very infrequent in Old Spanish. A student can therefore look for cases like [va a \*r], and will retrieve several examples like in Table 5.

Table 5: Searching the parallel corpus to compare constructions (OSP/MSP)

Text	OLD SPANISH	MODERN SPANISH
Rev 2:10	Non temas ninguna destas cosas por que as de passar. Euas que el diablo <u>metra</u> de uos en carcel . . .	No tengas ningún temor de las cosas que has de padecer. He aquí, el diablo <u>ya a echar</u> a algunos de vosotros en la cárcel. . .
1 Sam 10:27	Mas los hijos de belial dixieron que non le trayeron dones et eill fazie semblant que no lo oye	Pero unos perversos dijeron: “¿Cómo nos <u>podra defender</u> : Desdennaron lo nos <u>ya a librar</u> éste?” Ellos le & non le trayeron dones et eill fazie semblant que no lo oye presente. Pero él calló.

In summary, the parallel corpora can help students to find an unknown form in a different stage of the language, simply by working from the stage with which they already feel the most comfortable.

## 5 The Original “Corpus del Español” (3 million words; unannotated)

The parallel text “Polyglot Bible” that has just been described allows students to easily compare equivalent structures in different stages of the language, and to actually see the contrasting structures in context. However, this corpus would not allow students to see how a particular form or construction developed over a number of centuries (i.e. in the period between the three or four specific stages that appear in the polyglot text). For this type of research, students would need access to a comprehensive corpus of many different texts. In the case of Spanish, this would include texts from each of the centuries from the 1200s to the 1900s.

Fortunately, before the “History of the Spanish Language” course was taught on the web for the first time, I had already developed such a corpus of historical Spanish texts. Table 6 shows the composition of the corpus, which contained more than three million words in nearly 200 texts:

Table 6: Composition of the original 3,000,000 word corpus

Historical		Modern Spanish	
CENTURY	# texts	# words	(#) texts
1200	14	250,000	13
1300	10	250,000	14
1400	15	250,000	Habla Culta, 250,000 Esp Oral
1500	19	250,000	Novels, Short 250,000 Stories
1600	16	250,000	Habla Culta + 250,000
1700	17	250,000	Novels, Short 250,000 Stories

As can be imagined, because there are at least a quarter of million words from each century from the 1200s-1900s, the students are able to use the corpus to very accurately describe several different types of language change. As was shown in Table 1, Units 15-19 of the course require students to show evidence from the corpus for specific linguistic changes in terms of the sound system, orthography, morphology, syntax, and the lexicon, and the three million word corpus of historical Spanish texts allow them to provide extensive data for these changes.

In fact, the range of linguistic phenomena that the students are able to study is both quite broad as well as quite in-depth. The following table provides just a sampling of some of the shifts that the students have to map out and describe for two of these areas of language change – morphology and syntax – and comparable lists are given for phonetic, orthographical, and lexical changes. In each case, the information given in parenthesis after the shift (e.g., C 213) refers to the book and page number that describes the shift. The task of the students is to use the data from the corpus to verify whether the information in the textbook is in fact correct.

Let us examine a concrete example of how the students carry out their research. In #5 of the “Pronouns” section above, it mentions that pronouns in the [indirect]+[direct] sequence changed from [gelo] in Old Spanish to [se lo] in Modern Spanish (e.g., *se lo di* “to-him it I-gave”). Students studying this shift would simply enter [gelo] or [se lo] into a web-based search form, and select the centuries for which they wanted to retrieve data.

Table 7: Examples of specific types of phenomena investigated by the students

Morphological shifts, 1200s-1900s	Syntactic shifts, 1200s-1900s
<b>Nouns</b>	<b>Pronouns</b>
1. Gender (C 213, 243) (S 101-2)	1. placement (C 245) (S 119-20, 170-1)
2. la + -o (C 243)	2. mesoclitic future: cantar lo (h)an (S 114-5)
3. -issimo (C 213)	3. “redundant” DO/IO (C 245)
	4. impersonal se (C 246)
<b>Determiners / pronouns</b>	5. gelo / se lo (C 246)
1. vos(tus) / os (C 214)	6. vos / tú / usted (C 214, 244) (S 167-8)
2. la tu / tu (C 246)	7. omne = se (S 106) (L 402-3)
3. nosotros/vosotros (C 214)	
4. los/les (C 214, 245) (S 103, 201-2)	<b>Meaning and use of verb forms</b>
5. mio/mi, sos/sus, etc (C 215)	1. ser / estar (C 218) (S 127-8, 204) (L 400-1)
6. alguien:quien, nadie:otrie (C 215)	2. haber / tener (C 249) (S 127)
7. gelo / se lo (C 244) (S 103-4)	3. haber / ser + PP (C 249) (S 126, 169)
	4. haber / hacer (S 127)
<b>Verbs</b>	5. subjunctive(C 217, 248) (S 169)
1. -zco (verbs) (C 215-6)	6. infinitives (C 217) (S 123)
2. irregular past participles (C 216)	
3. irregular preterites (S 113-4)	
4. imperfect in -ie / ia (C 216) (S 112-3)	
5. irregular future tense (C 216) (S 115)	

They would then see the frequency of the construction in each historical period, as shown in Table 8.

Table 8: 3,000,000 word corpus – search interface and frequency listings

Word/phrase	<input type="text" value="gelo"/>	<input type="button" value="Submit"/>	<input type="button" value="Reset"/>
Time period	<input type="checkbox"/> 1200s <input type="checkbox"/> 1300s <input type="checkbox"/> 1400s <input type="checkbox"/> 1500s <input type="checkbox"/> 1600s <input type="checkbox"/> 1700s <input type="checkbox"/> 1800s <input type="checkbox"/> 1900s		
<b>Search string</b>	<b>1200s</b>	<b>1300s</b>	<b>1400s</b>
gelo	36	30	23
se lo	4	2	3
	54	56	31
	80	80	70

By comparing the two sets of data, the student can clearly see that it was about the 1500s that the new [se lo] form became the norm. For more precision, the students can click on the numbers indicating the frequency of any form in any century, and see the examples in context. Because this KWIC display shows the exact date of each text, it would be possible to describe the period of greatest change even more precisely.

Similar queries and investigations for any of the other morphological or syntactic shifts could be (and are) carried out in like fashion. Students can easily map the emergence or disappearance of a given word, the variation in the use of a

particular verbal conjugation, or the changes in the spelling (and perhaps also pronunciation) of a certain subset of words. Because of the design of the corpus, even relatively inexperienced students are able to quickly and easily extract large amounts of useful data. In fact, in many cases the descriptions that they give for different types of linguistic change are more detailed (in terms of the historical trajectories) than the descriptions given in the textbooks that we use in the class, which were written by experts with much more experience. All of this is very "empowering" to the students, in helping them to discover data that no one else had ever seen before.

## 6 The Present "Corpus del Español" (100 million words; richly annotated)

The three million word corpus that has just been described was the corpus that was used the first time that the course was offered online in Spring 2000. Although it was quite useful in its own right, it also had a number of limitations, which made certain types of linguistic investigations quite difficult. For example, the search engine for the corpus (Microsoft Search) does not allow much in the way of wildcard searches, which would have been quite useful for examining sound and spelling changes. More importantly, there was really no way to annotate the corpus. This meant that it was impossible to search by lemma (e.g., all of the forms of a particular verb) or by grammatical category.

Table 9: Composition of the newer, NEH-funded 100,000,000 word corpus

CENTURY	# WORDS	# TEXTS	CENTURY	# WORDS	# TEXTS
1200s	6,905,000	71	1800s	20,465,000	392 novels
1300s	2,820,000	50	1900s-Lit	6,750,000	850 novels/ stories
1400s	8,515,000	160	1900s-Oral	6,800,000	2040+ transcripts
1200s-1400s	18,240,000	281	1900s-Misc	6,800,000	4770+ articles
1500s	18,001,000	323	<b>1800s-1900s</b>	<b>40,815,000</b>	<b>8052</b>
1600s	12,746,000	499			
1700s	10,263,000	159			
<b>1500s-1700s</b>	<b>41,010,000</b>	<b>981</b>	<b>TOTAL</b>	<b>100,000,000</b>	<b>9314</b>

In order to address these shortcomings, a new corpus has been created, and this will now serve as the main database for the class. The new corpus was funded by a grant from the national Endowment for the Humanities, and was created between April 2001 and July 2002. It contains 100 million words of text, including 20 million from the 1200s-1400s, 40 million for the 1500s-1700s, and

40 million for the 1800s-1900s. Table 9 provides more details on the composition of the corpus.

The process of carrying out queries with the newer 100,000,000 corpus is fairly similar to the older 3,000,000 word corpus. With the new corpus there are more options as far as limiting the query by frequency in different centuries, how the results will be grouped (word form or lemma), how the results will be sorted, etc. But the only field that is required is the [SEARCH] field itself. For example, suppose that a student wants to search for cases of an object pronoun + any form of *querer* "to want" + an infinitive (e.g., *lo quiero hacer* "it I-want-to-do"). Suppose also that the students want to limit the strings only to those that occur at least once in the 1900s, and that they want to sort the results by the frequency of the string in the 1900s. The students would enter the following into the search form, and then see the following results:

Table 10: 100,000,000 word corpus – query interface and frequency listings

	SEARCH	SORT		LIMITS		GROUP		RESET				
		*.pn	obj	querer.*	v	inf	[1900s]	+1900s	FORMS	SUBMIT		
#	PHRASE(S)	12	13	14	15	16	17	18	19	Lit	Oral	Misc.
1	te quiero decir				17	10	1	8	49	11		38
4	me quiero ir				32	10	1	2	23	7		16
19	Le quiere dar	9	1	1	7	6	2	6	4	3		1
22	Te quiero contar				1	11	3		4		4	

Once they are presented with the frequency listing of all matching forms, users can then use the checkboxes to select which phrase(s) to see in context and in which historical period(s). After selecting these phrases, they then see a "keyword in context" display, in which the example sentences can be re-sorted by left and right contextual words, or see a more expanded block of text. (Note: in Table 11 the examples are truncated, unlike on the web).

Table 11: 100,000,000 word corpus – KWIC display

TIME	TEXT	RE-SORT BY: L-2	L-1	C	R-1	R-2
12	Libro de los..	tiene gela forçada.	Et non	le quiere dar	lo que a tomado & en lugar de	
15	La Serrana de..	desdicha el desengaño.	No	me quiero casar,	padre, que creo que mientras no	
19_L	Follaje en..	¡Haré lo que quiera,	no	me quiero ir!	Ya soy grande y sé hacer de	
19_O	EspOral:CO..	a mi madre y a mi padre.	Te	quiero decir	que es una cosa que yo - y mis	

Even more important than the size of the corpus is its annotation scheme and search engine, which provide capabilities for a wider range of searches than almost any other large corpus in existence. The corpus uses a unique relational database architecture – which I have designed especially for this corpus – which allows searching by substring (advanced wildcard queries), subqueries, lemma, part of speech, synonyms, and user-defined features. In addition, the queries on the corpus are very fast. Even the most complex queries only take three or four seconds to return data from the 100 million word corpus. In the sections that follow, I will discuss very briefly how the new corpus can meet the needs of students in the “History of the Spanish Language” course, in terms of mapping out in very detailed fashion a wide range of linguistic shifts.

First, the substring function allows students to investigate sound change and shifts in spelling. Examples of the types of queries allowed by the search engine are given in the Table 12, where the three columns refer to the student input, examples of the output, and an explanation of the search.

Table 12: Examining sound/spelling changes

s_fr*	<i>sofryr, sufrye, sufriendo</i>	words relating to the root [s-fr] “to suffer” ([o] in Old Spanish, [u] in Modern Spanish)
*mbre 1200s>5 1900s<5	<i>ombre, fambre, ombre</i>	words ending in -mbre, which occur at least five times in the 1200s, but less than five times in the 1900s (i.e. forms from Old Spanish)
*aua* +1200s *aba* +1900s	<i>jablaua, caualteros, claua hablaba, caballeros, claba</i>	words with the pattern *aua* in the 1200s, which have an equivalent with *aba* in the 1900s (resulting from a spelling change in the 1700s)

Second, the corpus can be used to examine morphological change and variation. This is due to the wildcard searches (just mentioned), as well as the fact that the word forms are annotated for lemma (= lemma.\*):

Table 13: Examining morphological changes

*iere +1200s +1300s -1900s	<i>fiziere, naciere, toutiere</i>	word ending in -iere that occur at least once in the 1200s and 1300s but not the 1900s. This would retrieve many forms of the future subjunctive, a verbal form that has essentially died out by Modern Spanish
*simo +1400s -1300s	<i>santisimo, altisimo, grandisimo</i>	words ending in -[i]simo (a marker of the superlative), which do not occur in the 1300s but which do occur for the first time in the 1400s
decir.* +1200s -1500s -1900s	<i>dize, dixiere, dezir</i>	forms of <i>decir</i> “to say” that occur in the 1200s, but not in the 1500s or 1900s (i.e. forms of the verb from Old Spanish, which have subsequently disappeared)

Third, it is possible to carry out advanced syntactic analysis on the corpus, due to the fact that the corpus is annotated for part of speech (= \*.pos):

Table 14: Examining syntactic changes

*.v_inf -1700s -1800s +1900s	<i>detectar, liberar, programar</i>	infinitives that occur in the 1900s, but which do occur in the 1700s or 1800s (i.e. new verbs that have entered into the language)
poder.* lo/la/los/las *.v_inf +1200s	<i>puede lo fazer, podemos las fazer</i>	forms of <i>poder</i> “to be able” + object pronouns (e.g. <i>lo/la/los/las</i> ) + an infinitive (common word order in Old Spanish)
estar.* cansado.* de *.v_inf	<i>estoy harto de vivir, estaba cansada de escuchar</i>	any form of <i>estar</i> “to be” + any form of any adjective of <i>cansado</i> (“tired”) + <i>de</i> + infinitive

Fourth, the corpus can be used to investigate semantic change. Two features of the corpus make this possible. The first is the possibility of using collocations to see what other words occur with a given word in different historical periods. If the words that co-occur have changed significantly over time, then that may indicate that the word in question has also changed its meaning. Second, the corpus has a built-in thesaurus for more than 30,000 words. This allows users to see which synonyms of a given word have increased or decreased in frequency over time.

Table 15: Examining semantic changes

romper 1900s>10 1700s<5	<i>irrumpir, incumplir, escacharar</i>	Synonyms of <i>romper</i> “to break” that occur at least ten times in the 1900s, but less than five times in the 1700s
*.n_suaue +1900s -1800s -1700s	<i>sabor suaue, pelaje suaue</i>	Nouns that occur with <i>suaue</i> “soft” in the 1900s, but not in the 1800s or 1700s. May indicate recent shifts in the meaning of <i>suaue</i> .

In addition to all of the types of searches shown previously, it is also possible to create “customized” lists of words that can be re-used in subsequent searches. These lists can include items that are semantically, syntactically, or morphologically related, such as parts of clothes, temporal adverbs, or words ending in -azo (a suffix that sometimes refers to a strike or blow made with an object, e.g. *puerta* > *portazo* = “to hit with a door”). The students simply create the list of words via a simple form in the search interface, and they can later modify the list and use it as part of the search syntax. For example, suppose that a student named [susana.rubio] has created lists called [ropa] “clothes” and [azo] “strikes/blows with an X” with the following items:

ropa: sombrero, pantalón, camisa, zapato, cinturón  
 azo: puñetazo, portazo, manotazo, latigazo, collazo

Later that day, or even weeks later, this student could then re-use this list in a search, as shown in Table 16.

Table 16: User-defined lists

poner.* ella/los/las [susana.rubio:ropa].*	ponerse los pantalones, puso el sombrero	any form of poner ("to put") + definite article (lo/la/los/las) + any form of any word in the [ropa] list
dar.* un [susana.rubio:azo]	dé un portazo, da un puñetazo, dio un codazo	any form of dar ("to give") + un ("a") + any word in the [azo] list

In summary, the Corpus del Español that I have created offers a wider range of searches than is possible with any other historical corpus of any language. This allows students in the online "History of the Spanish Language" course to investigate and describe an ever wider range of linguistic phenomena than has been possible in the past. All of this suggests that the time has past when students needed to memorize long lists of overly-abstract rules of linguistic change from textbooks. Using state-of-the-art corpora of the type that I have described, the students themselves are now in control of extracting the data, and can by themselves find evidence for and describe a wide range of historical changes in the language.

## References

- Botley S., J. Glass, T. McEnery, and A. Wilson (eds) (1996), *Proceedings of teaching and language corpora 1996*, Lancaster: University Centre for Computer Corpus Research on Language Technical Papers 9 (Special Issue).
- Burnard L. and T. McEnery (eds) (2000), *Rethinking language pedagogy from a corpus perspective: Papers from the Third International Conference on Teaching and Language Corpora*, Frankfurt: Peter Lang.
- Curzan, A. (2000), English historical corpora in the classroom: The intersection of teaching and research, *Journal of English Linguistics*, 28: 77-89.
- Davies, M. (2000), Using multi-million word corpora of historical and dialectal Spanish texts to teach advanced courses in Spanish linguistics, in L. Burnard and T. McEnery (eds), *Rethinking language pedagogy from a corpus perspective: Papers from the Third International Conference on Teaching and Language Corpora*, Frankfurt: Peter Lang, pp. 173-186.
- Kettenmann, B. and G. Marko (eds) (2002), *Teaching and learning by doing corpus analysis: Proceedings of the Fourth International Conference on Teaching and Language Corpora*, Amsterdam: Rodopi.

Knowles, G. (1997), Using corpora for the diachronic study of English, in A. Wichmann, S. Fligelstone, T. McEnery, and G. Knowles (eds), *Teaching and language corpora*, London: Longman, pp. 195-210.

Rissanen, M. (1992), *History of Englishes: New methods and interpretations in historical linguistics*, Berlin: Mouton de Gruyter.

Rissanen, M. (1993), *Early English in the computer age: Explorations through the Helsinki corpus*, Berlin: Mouton de Gruyter.

Rissanen, M. (1997a), *Grammaticalization at work: Studies of long-term developments in English*, Berlin: Mouton de Gruyter.

Rissanen, M. (1997b), *English in transition: Corpus-based studies in linguistic variation and genre styles*, Berlin: Mouton de Gruyter.

Schmied, J. (1996), Encouraging students to explore language and culture in Early Modern English pamphlets, Unpublished presentation given at TALC 96 (Lancaster University).

Wichmann A., S. Fligelstone, T. McEnery, and G. Knowles (eds) (1997), *Teaching and language corpora*, London: Longman.



## Contents

### Preface

iii

### Analyzing Spoken and Written Language Corpora

- 'Like the Wise Virgins and All that Jazz': Using a Corpus to Examine Vague Categorisation and Shared Knowledge  
*Anne O'Keefe* 1
- A Corpus-driven Analysis of the Use of Intonation to Assert Dominance and Control  
*Martin Warren* 21
- // → FRIENDS // ↘ ↗ LADIES and GENTLEMEN //: Some Preliminary Findings from a Corpus of Spoken Public Discourses in Hong Kong  
*Winnie Cheng* 35
- A Corpus Linguistic Investigation of Vocabulary-based Discourse Units in University Registers  
*Douglas Biber, Eniko Csomay, James K. Jones and Casey Keck* 53
- Formulating Writer Stance: A Contrastive Study of EFL Learner Corpora  
*JoAnne Neff, Francisco Ballesteros, Emma Dafouz, Francisco Martinez, Juan-Pedro Rica, Mercedes Díez and Rosa Prieto* 73
- Using Corpus Linguistics to Investigate Class, Ideology, and Discursive Practices in Online Political Discussions  
*Susana M. Sotillo and Julie Wang-Gempp* 91
- Pedagogical Applications of Corpora**
- Computer Learner Corpus Research: Current Status and Future Prospects  
*Sylviane Granger* 123
- Concordancing and Corpora for K-12 Teachers: Project MORE  
*Boyd Davis and Lisa Russell-Pinson* 147
- Units of Meaning, Parallel Corpora, and their Implications for Language Teaching  
*Wolfgang Teubert* 171

Online access is included in print subscriptions:  
see [www.todopi.nl](http://www.todopi.nl)

The paper on which this book is printed meets the requirements of "ISO 9706:1994, Information and documentation - Paper for documents - Requirements for permanence".

ISBN: 90-420-1922-0 (bound)

©Editions Rodopi B.V., Amsterdam - New York, NY 2004  
Printed in The Netherlands

---

Making the Web More Useful as a Source for Linguistic Corpora <i>William H. Fletcher</i>	191
Student Use of Large Corpora to Investigate Language Change <i>Mark Davies</i>	207
The Montclair Electronic Language Database Project <i>Eileen Fitzpatrick and M. S. Seegmiller</i>	223
Bridging the Gap between Applied Corpus Linguistics and the Reality of English Language Teaching in Germany <i>Joybrato Mukherjee</i>	239
Top-down and Bottom-up Approaches to Corpora in Language Teaching <i>John Osborne</i>	251
Towards an Instrument for the Assessment of the Development of Writing Skills <i>Pieter de Haan and Kees van Esch</i>	267