

Un corpus anotado de 100.000.000 palabras del español histórico y moderno

Mark Davies

Illinois State University
4300 Foreign Languages
Normal, IL 61790-4300 USA
mdavies@ilstu.edu

Resumen: En <http://www.corpusdelespanol.org> se encuentra el Corpus del Español – 100.000.000 palabras en el primer corpus anotado del español histórico y moderno. A diferencia de otros corpus del español histórico, el “Corpus del Español” permite búsquedas por 35 categorías gramaticales, 20.000 lemas, y 30.000 grupos de sinónimos y antónimos, además de búsquedas por etimología, frecuencia, y por categorías semánticas y sintácticas creadas por el usuario mismo. Con todo esto, puede haber búsquedas tan complejas como “complemento directo pronominal + todas las formas de cualquier sinónimo de *querer* + infinitivo, que ocurre en el siglo XX pero no en los siglos XIII o XIX”. También se pueden producir fácilmente listados completos de colocaciones. La flexibilidad y el poder del corpus (juntos con la velocidad– menos de 2-3 segundos para casi todas las búsquedas) se deben a la arquitectura innovadora del corpus – varias bases de datos relacionales que están ligadas y que tienen anotación para los 45.000.000 n-grams distintos en el corpus.

Palabras clave: corpus, histórico, base de datos relacional

Abstract: The first annotated corpus of historical and modern Spanish – the 100,000,000 word Corpus del Español – is now online at <http://www.corpusdelespanol.org>. Unlike other corpora of historical Spanish, the “Corpus del Español” allows searches by 35 grammatical categories, 20,000 lemmata, and 30,000 groups of synonyms and antonyms, in addition to searches by etymology, frequency, and by user-defined semantic and syntactic categories. All of this allows searches as complex as “pronominal direct object + all forms of any synonym of *querer* + infinitive, which occurs in the 1900s but not in the 1700s or 1800s”. It is also possible to easily produce complete lists of collocations. The flexibility and power of the corpus (as well as the speed – 2-3 seconds for nearly all searches) are due to the innovative architecture of the corpus – several relational databases that are linked together and which contain annotation for the 45,000,000 distinct n-grams in the corpus.

Keywords: corpus, historical, relational database

1 Introducción

Han aparecido en los últimos cuatro o cinco años varios corpus que tienen gran utilidad para los lingüistas, incluso para los lingüistas históricos. Con estos corpus más amplios, es posible sacar miles de ocurrencias de ciertas palabras o frases en muy poco tiempo y así crear un modelo bastante exacto de los cambios históricos. Quizás el corpus histórico más conocido hasta la fecha es CORDE, de la Real Academia Española (<http://cronos.rae.es/cordenet.html>). Este corpus sirve muy bien para las búsquedas de palabras y frases exactas,

especialmente cuando se desea limitar las ocurrencias a géneros y periodos históricos exactos – por ejemplo, todas la ocurrencias de una sola palabra en obras de drama de 1620-1700.

Sin embargo, el motor de búsquedas que se emplea en CORDE también tiene unas limitaciones importantes. Por ejemplo, aunque es posible usar comodines para encontrar todas las palabras que comienzan con ciertas letras (p. ej. *desarroll** o *dix**), no es posible buscar las palabras que tienen cierta terminación (p. ej. **azo*, **ieran*) o que tienen un padrón específico en medio de la palabra (p. ej. **ísim**, *s_fr**).

Debido a estas limitaciones, CORDE no sirve muy bien para los estudios morfológicos – por ejemplo, para crear un listado de todas las formas que terminan en *-azo* en cierto siglo. También, como no se ha anotado el corpus para lema, no se pueden estudiar tampoco todas las formas de cierto sustantivo o verbo, por ejemplo la frecuencia histórica relativa de todas las formas de *hacer*.

El aspecto en que CORDE es quizás más limitado es con los estudios sintácticos. Tomemos el ejemplo de la evolución de la construcción causativa (*fizo que se fuessen, le hizo comer el pan*) (Davies 1995, 1996). Con el corpus de CORDE, sería casi imposible estudiarla. No habría manera de buscar a la vez todas las formas de *hacer* + un subjuntivo o un infinitivo, como no se pueden buscar las palabras que terminan en **ar/er/ir/*. De igual manera, no serviría para estudiar algo como la diacronía de la monta de clíticos (*(lo) quiero (lo) comprar(lo)*) (Davies 1998). Esta construcción se compone de un complemento pronominal + una forma de *querer/deber/poder*, etc. + un infinitivo. Otra vez, no se podrían buscar ni las varias formas de los verbos, ni los complementos pronominales (como grupo), ni los infinitivos. Estas construcciones, juntas con otras construcciones relacionado del infinitivo, se han investigado antes con otros corpus grandes, pero privados (p. ej. Davies 1995, 1996, 1997, 1998, 2000, 2002a, 2002b), pero con CORDE tales investigaciones serían imposibles. De modo que, aunque CORDE es útil para buscar palabras o frases exactas, habría grandes problemas (o sería imposible) usarlo para búsquedas más avanzadas.

Debido a estas limitaciones de CORDE, hace un año se decidió crear otro corpus del español histórico y moderno – el “Corpus del Español” – que tiene 100.000.000 palabras (véase <http://www.corpusdelespanol.org>). El corpus será terminado en agosto de 2002, pero ya se puede usar en su forma actual. Al crear el Corpus del Español, se ha puesto mucho esfuerzo en crear lo que CORDE no tiene, o lo que no puede hacer fácilmente.

Como veremos, a diferencia de CORDE, con el Corpus del Español es posible hacer búsquedas por lema, por categoría gramatical, y hasta por sinónimos, y de limitarlas a sólo las palabras y frases que ocurren con cierta frecuencia en los varios periodos históricos. Además, casi todas estas búsquedas (y combinaciones aun más complejas) se pueden

hacer en muy poco tiempo – por lo general menos de dos o tres segundos.

2 *La creación y arquitectura del corpus*

2.1 Antes de hablar del motor de búsquedas – lo que creemos que hace que el Corpus del Español sea innovador – primero comentaremos brevemente el corpus textual mismo. Se compone de 100.000.000 palabras en más de 10.000 textos y transcripciones del español hablado, del siglo XIII hasta el siglo XX. Estos textos vienen de varias fuentes, incluyendo ADMYTE (www.admyte.com), el “Hispanic Seminary of Medieval Studies” de los EEUU (www.hispanicsociety.org), la Biblioteca Virtual (www.cervantesvirtual.com), Comedia (www.coh.arizona.edu/spanish/comedia/escomedi.html), el “Proyecto filosofía en español” (www.filosofia.org), y varias fuentes para es español moderno – literatura (novelas, cuentos, obras de drama), textos orales (Habla Culta, el Corpus Oral [http://elvira.llif.uam.es/docs_es/corpus/corpus.html]), transcripciones de congresos, y entrevistas periodísticas), y miscelánea (enciclopedias, periódicos, etc).

Los 100.000.00 palabras de texto están en una base de datos relacional de SQL Server 7.0. La base de datos no tiene ninguna anotación – aparte de un código que indica la fuente de cada bloque de texto, pero sí está indexada con el “Full-Text Indexing” de SQL Server, lo cual hace que se pueda buscar rápidamente. Si ése fuera el único índice, sin embargo, sólo se permitirían más o menos las mismas búsquedas que CORDE, debido al hecho de que el motor de búsquedas (Microsoft Search) es básicamente el mismo. Es decir, sin otro índice u otra base de datos, el Corpus del Español tendría las mismas limitaciones que CORDE.

Para permitir búsquedas más complejas, por supuesto tiene que haber algún tipo de anotación. La mayoría de los corpus grandes que se han creado hasta la fecha emplean un sistema de anotación intertextual – es decir, la anotación es parte del corpus textual (Biber et al., 2000). Por ejemplo, en el British National Corpus, la anotación para lema y categoría gramatical es simplemente cuestión de prefijos o sufijos que se agregan a las formas individuales (Clear 1993, Berglund 1999).

Sin embargo, varios corpus, como el COBUILD “Bank of English” (Clear et al., 1996), CETEMPUBLICO (Rocha et al. 2000;

Santos 2000), y otros creados con el IMS Corpus Workbench (Christ 1994) emplean otro sistema. En este caso, hay un índice de formas, que es separado del corpus textual, y que contiene una base de datos que indica la colocación de cada palabra en el corpus. Por ejemplo, la palabra <de> tendría millones de entradas, pero una palabra como <abrillantado> tendría solo cuatro o cinco entradas. Pero hay una diferencia importante entre el motor de búsquedas del IMS Corpus Workbench y el del “Microsoft Search” que se emplea en CORDE. La base de datos creada por el IMS CW está “abierta” y permite que haya otras columnas en la base de datos para indicar el lema y/o la categoría gramatical de las varias formas.

Pero es importante recordar que en este sistema, toda la anotación es todavía parte de una sola tabla en la base de datos. Lo que es aun más importante es que la base de datos sólo se puede acceder con el IMS Corpus Workbench. Esto significa que no se pueda integrar con otras bases de datos relacionales (diccionarios, sinónimos, etc.), y veremos que esto también es una limitación importante.

2.2 Nuestro método está relacionado ligeramente con el del IMS CW, por el hecho de que la base de datos es distinta del corpus textual y que contiene información sobre la anotación. Sin embargo, nuestro sistema usa las bases de datos relacionales de una forma más avanzada que casi cualquier otro corpus grande.

En el Corpus del Español, hay una base de datos que contiene cada 1, 2, y 3-gram (secuencias distintas de una, dos, y tres palabras) en todo el corpus – 900.000 1-grams, 11.000.000 2-grams, y 33.000.000 3-grams. También, para cada n-gram (1, 2, o 3-gram), hay información en la base de datos sobre la frecuencia en cada uno de los siglos desde el siglo XIII hasta el siglo XX, y en los tres registros distintos del español moderno (literatura, oral, y miscelánea)¹. Esta base de

¹ La información sobre todas las secuencias distintas, juntas con la frecuencia en cada periodo histórico y registro distinto, se creó por medio del programa WordList, que es parte de WordSmith (www.liv.ac.uk/~ms2928/WordSmith). Se crearon archivos con las secuencias y las frecuencias para más de veinte bloques de texto, se importaron en SQL Server, y después se unieron para crear tres tablas para las secuencias de 1, 2, y 3 palabras.

datos está ligada a muchas otras bases de datos (lema, categoría gramatical, sinónimos, etimologías, etc.), y la interacción entre estas bases de datos es lo que permite búsquedas tan complejas y poderosas.

Como se ha indicado, la base de datos central es la que contiene todos los n-grams distintos. La otra información que se encuentra en esta base de datos central es la categoría gramatical y el lema. Esta información se ha unido con las 45.000.000 n-grams distintas y su fuente es un diccionario que contiene 500.000 formas distintas, el cual se ve en lo siguiente:

forma	lema	categoría
trabajaron	trabajar	v_pret
abuelas	abuelo	N

Tabla 1: Diccionario de formas / lemas / categoría

El resultado de la unión de la base de datos de n-grams y de frecuencias y la de la categoría y del lema es algo parecido a la entrada que se ve en la tabla siguiente:

P1	L1	C1	P2	L2	C2	P3	L3	C3	12	13	14	15	16	...
son	ser	vp	las	lo	adef	cosas	cosa	n	38	16	77	67	16	...

Tabla 2: N-grams / frecuencia / lema / categoría²

Esta entrada para el 3-gram “*son las cosas*” es un ejemplo de las 33.000.000 entradas distintas en la tabla de los 3-grams, y se encuentran entradas semejantes en la tabla de los 2-grams (11.000.000 entradas) y los 1-grams (900,000 entradas).

2.3 Ya veremos cómo se tiene acceso a la base de datos por medio del interfaz en el web. Por ejemplo, para buscar los casos de cualquier forma de *ser + las + N* (*eran las casas, serán las circunstancias*), los usuarios meten lo siguiente:

ser.* las *.n

y en el servidor esto se convierte al comando SQL:

```
select * from [table] where L1 = 'ser' and w2 = 'las' and c3 = 'n'
```

En menos de un segundo, se muestran más de 300 3-grams, ordenados por frecuencia en el siglo que se desee. En la tabla siguiente se ve un listado parcial de las 300+ frases, ordenado

² En esta tabla ‘P1/2/3’ = la palabra, ‘L1/2/3’ = lema, ‘C1/2/3’ = la categoría gramatical, y 12-19 = el siglo (1200s, 1500s, 1900s-literatura, etc).

por frecuencia en los 1900s (muchas de las frases más frecuentes se han omitido)³:

#	PALABRA(S)	12	13	14	15	16	17	18	19	Lit	Oral	Misc
1	son las cosas	38	16	77	67	16	19	33	45	22	19	4
16	eran las palabras	2		1	4	2		5	8	6	1	1
29	ser las cosas	4	1	6	9		1	2	6	4	2	1
73	sean las circunstancias							5	3		2	1
...

Tabla 3: N-grams / frecuencias – resultados

La rapidez con que se realizan las búsquedas se debe a la arquitectura de la base de datos. Cada una de las columnas tiene su propio índice, así que es mucho más rápido pasarle un comando SQL a la base de datos y sacar las frases que corresponden, que atravesar todo el corpus de 100.000.000 de palabras, lo cual (aunque fuera posible), llevaría varios minutos para cada búsqueda. Con nuestro sistema, son raras las búsquedas que toman más de dos o tres segundos.

Muchos sólo quieren ver el listado de palabras y frases que resulta de la búsqueda. Pero los que quieren ver una palabra o frase en contexto pueden hacer click en la palabra o frase para verla en contexto en todos los siglos, o pueden limitarlo a solamente ciertos siglos. De la misma manera, pueden escoger solamente ciertas palabras (o frases) y ciertos siglos para verlos en formato KWIC. Esta parte de la búsqueda también es muy rápido (~1000 resultados en menos de 2-3 segundos) debido al “Full Text Indexing” del corpus textual.

Una vez que tienen los resultados KWIC – como se ve en lo siguiente – pueden re-ordenar las ocurrencias por las palabras a la izquierda o a la derecha, y pueden hacer click el cualquier ejemplo para ver más contexto (hasta un párrafo):

# = MÁS	1 / 4 >	RE-ORDENAR POR: I-2 L-1	C	D-1 D-2	AYUDA
1	13	Tratado de cetrería. ... vn poco E ve Acaçar E sepas en verdat que estas	son las cosas	con que omne puede fazer caçar los falcones // El ...	
2	13	Sevillana medicina. ... cuerpos que han olor / y poresta razon	son las cosas	callentes de mayor olor que non las frias / y si las ...	
3	14	Esopete ystoriado. ... & asi fechas como se cuentan. Argumentos	son las cosas	que non fueron fechas. mas pueden ser fechas: asi ...	
4	14	Libro de los olios ... & vale este olio a otras muchas cosas E estas	son las cosas	que entran en el / Reçipe olio comun tras ...	
...

Tabla 4: Palabras clave en contexto (KWIC)

³ En este ejemplo y en los siguientes, se debe notar que además de mostrar las frecuencias absolutas, también se puede indicar el número de ocurrencias en cada millón de palabras, o los dos a la vez.

3 Búsquedas más avanzadas

Como se puede ver, el Corpus del Español tiene mucho en común con CORDE. El tamaño del corpus para los siglos XIII-XIX es básicamente el mismo (~80.000.000 palabras), y en las búsquedas de palabras y frases exactas, la funcionalidad de los dos corpus es más o menos igual. La ventaja del Corpus del Español, sin embargo, se encuentra en la variedad y complejidad de las búsquedas, lo que hace que sea muy útil no sólo para las investigaciones léxicas, pero más que nada para las investigaciones morfológicas y sintácticas.

3.1 Comodines / padrones de palabras

Como en CORDE, en el Corpus del Español se pueden hacer búsquedas con comodines, por ejemplo *descri**. La diferencia es que con CORDE muchas de estas búsquedas terminan después de varios segundos sin producir resultados, porque el motor de búsquedas de CORDE en realidad no fue diseñado para búsquedas complejas con comodines. Con el Corpus del Español, sin embargo, esta búsqueda producirá más de 300 palabras distintas (*descripción, descritas, describas, describió,* etc) en menos de un segundo. Se ve en una sola ventana la frecuencia de cada palabra en cada siglo, lo cual facilita mucho la comparación entre las formas:

	PALABRA(S)	...	15	16	17	18	19	Lit	Oral	Misc
1	descripción	...	272	212	380	289	335	83	65	187
2	describe	...	66	39	70	148	264	39	35	190
3	describir	...	15	19	51	246	243	52	47	144
4	descrito	...	16	6	29	130	140	23	25	92
...

Tabla 5: Uso de comodines / frecuencia por siglos

Además del uso de comodines al final de la palabra – algo que también puede hacer CORDE – el Corpus del Español también los permite al comienzo o en medio de la palabra, y puede distinguir entre una letra [_] y más de una letra [*]. Por ejemplo, se pueden buscar todas las palabras con el padrón:

[s_fr*] *sufrir, sufriendo, sofrimento, sofre*

[*azo] *puñetazo, portazo, latigazo, manotazo*

La tabla siguiente es un listado muy parcial de las 300+ palabras que resultan de la búsqueda de *-azo*, ordenados por su frecuencia en el siglo XVI:

#	PALABRA(S)	12	13	14	15	16	17	18	19	Lit	Oral	Misc
13	flechazo				36	13	8	24	18	10	4	4
14	arcabuzazo				28	13		4				
25	garrotazo				6	1	3	28	10	10		
44	puntillazo				2	3	1	1	1	1		
...

Tabla 6: Sufijos

Obviamente, la habilidad de buscar sufijos ayuda mucho con las investigaciones morfológicas. También, es muy útil ver un listado completo de todas las palabras que resultan de cierto padrón, junto con su frecuencia en los varios siglos y registros.

3.2 Colocaciones

Además de omitir letras, también se pueden omitir palabras. Lo útil de esto es que así se pueden ver fácilmente las colocaciones más comunes de cierto entorno, por ejemplo:

tan * como :

tan pronto / importante / grande / bien como

* suave :

voz / viento / luz / tono suave

La siguiente tabla es un listado parcial de los resultados con [* suave]:

#	PALABRA(S)	...	15	16	17	18	19	Lit	Oral	Misc
10	voz suave	...	10	6	2	13	13	11		2
15	viento suave	...		2		3	7	6		1
21	tono suave	...		1	1	7	3	3		
35	mano suave	...				4	2	2		
...

Tabla 7: Colocaciones

Como se ha indicado antes, la base de datos también tiene información sobre la parte del habla y el lema, y se pueden incluir estos en la búsqueda – algo que no sería posible con CORDE. Por ejemplo, se puede obtener el listado de todas las frases con los siguientes padrones, otra vez ordenado por frecuencia en cualquier periodo histórico:

hasta que *.v_subj_ra :

hasta que llegara / vinieran / muriera / hubiera

qué *.adj:

qué bueno / lindo / raro / interesante

La tabla siguiente muestra un listado parcial con [qué *.adj]:

#	PALABRA(S)	...	15	16	17	18	19	Lit	Oral	Misc
4	qué bueno	...	20	48	3	50	86	24	62	
7	qué raro	...	1	3	3	8	52	26	26	
14	qué difícil	...	1	3	2	4	28	14	12	2
20	qué horrible	...	2	8	7	48	20	10	10	
...

Tabla 8: Colocaciones con categoría gramatical

3.3 Lema y categoría gramatical

En el ejemplo anterior se ve cómo se ha usado la información sobre la categoría gramatical como parte de la búsqueda. Ya que la base de datos contiene información sobre el lema, esto también puede formar parte de la búsqueda – por ejemplo, una tabla que muestra todas las formas de *decir* (*dixo, dize, dezir, dizen, dixieron*) o abuelo (*abuela, abuelo, abuelos, abuelita*) durante los últimos 800 años, junto con su frecuencia relativa en cada siglo. Esto también se puede combinar con la categoría gramatical para estudiar construcciones bastante complejas como la de “Object to Subject Raising” (Davies, 2002a), que tiene ciertos adjetivos (lemas como *fácil* e *imposible*) + *de* + infinitivo (categoría gramatical):

	PALABRA(S)	...	15	16	17	18	19	Lit	Oral	Misc
1	difícil de explicar	...	0	0	2	11	20	11	8	1
2	difícil de entender	...	4	1	6	3	16	4	9	3
4	difíciles de encontrar	...	0	0	1	2	11	1	3	7
...

Tabla 9: Categoría gramatical y lema

3.4 Frecuencia

Ya nos hemos referido al hecho de que se puede usar la información en la base de datos que indica la frecuencia en los varios periodos históricos y en los registros del español moderno. Por ejemplo, se puede meter lo siguiente en el formulario en el web:

BUSCAR	ORDENAR	LIMITAR
decir.*	1400s	+1300s +1400s -1500s -1600s

Tabla 10: Limitar y ordenar por frecuencia

Esto sacaría de la base de datos todas las formas de *decir* que aparecen en el siglo XIV y el siglo XV pero que ya han desaparecido para los siglos XVI y XVII, y las ordenaría por el número de ocurrencias en el siglo XV, p. ej:

	PALABRA(S)	12	13	14	15	16	...
1	dexiemos	2214	115	316	0	0	...
2	deçir	7	44	132	0	0	...
3	dixiese	324	137	112	0	0	...
4	dixeronle	73	55	59	0	0	...
...

Tabla 11: Frecuencia – resultados

3.5 Búsquedas más complejas

Por supuesto, el usuario también puede hacer cualquier combinación de búsquedas, por ejemplo:

clítico + todas las formas de los sinónimos de *querer* + infinitivo, que ocurren por lo menos una vez en los 1900s pero no en los 1800s

Esta búsqueda incluye categoría gramatical, lema, sinónimos (que veremos en la próxima sección), y frecuencia, y sin embargo la búsqueda de las 100.000.000 palabras se realiza en pocos segundos, y nos da más de 300 formas distintas (*le quiero decir, se desea obtener, le apeteciera entrar*, etc).

3.6 Sinónimos / antónimos

Una de las ventajas más grandes de usar la arquitectura abierta de bases de datos de SQL Server es que se pueden unir otras bases de datos relacionales a la base de datos central, que contiene información sobre los n-grams, frecuencia, lema, y categoría gramatical. Por ejemplo, hemos creado otra base de datos que tiene 30.000 grupos de sinónimos y antónimos. El usuario simplemente realiza una búsqueda empleando [!] para referirse a los sinónimos de una palabra:

hombre/mujer !inteligente

y el “script” crea un comando SQL que saca los sinónimos de una base de datos y los usa como parte de la búsqueda de la segunda base de datos (la de los n-grams y frecuencias):

```
select top 300 * from [table] where P1 in ('hombre', 'mujer') and w2 in ('inteligente', 'astuto', 'instruido', 'perspicaz', 'despierto', 'vivo', 'agudo', 'listo', 'despejado', 'avisado', 'lúcido', 'capaz', 'ingenioso', 'versado', 'espabilado')
```

Esto produce resultados como el siguiente:

#	PALABRA(S)	...	15	16	17	18	19	Lit	Oral	Misc
1	hombre inteligente	...	4	1	3	14	13	11	1	1
5	mujer capaz	...				6	4			
6	hombre astuto	...	5	5	5	8	2			2
15	hombre agudo	...	3	3		2				
...								

Tabla 12: Sinónimos

3.7 Una arquitectura abierta

Aparte de la base de datos de sinónimos, también hay planes de crear y unir otras que tienen información sobre etimologías y

traducciones entre el inglés y el español. Por último, el usuario mismo también puede crear sus propias bases de datos en el momento de hacer la búsqueda – por ejemplo, campos semánticos especializados (la ropa, términos deportistas, etc) o sintácticos (p. ej. verbos transitivos que toman cierto tipo de complemento directo) – y después puede usarlas otro día como parte de otra búsqueda. Lo importante es que – debido a la arquitectura abierta del corpus – se puede unir cualquier otra base de datos a la base de datos central y después usarla de manera muy fácil en la búsqueda.

4 Conclusión

Sólo hemos podido resumir muy brevemente la organización del corpus y los varios tipos de búsquedas que permite. Tampoco hemos descrito algunos de los problemas que nos han enfrentado (y que todavía nos enfrentan) al crear el corpus y las bases de datos. Pero se espera que con esta pequeña introducción al Corpus del Español, hayamos podido demostrar el valor y el poder de usar una base de datos relacional para crear el motor de búsquedas para un corpus grande. También hemos intentado demostrar que nuestro enfoque utiliza una arquitectura abierta que permite añadir muchas otras bases de datos a la base de datos principal, para permitir aun más tipos de búsquedas. Por último, esperamos que se haya visto cómo un corpus como el nuestro puede ayudar a los investigadores a estudiar muchos fenómenos lingüísticos que jamás se podrían estudiar con cualquier otro corpus del español que se haya creado hasta la fecha.

Bibliografía

- Berglund, Y. 1999. Exploiting a Large Spoken Corpus: An End-User's Way to the BNC. *International Journal of Corpus Linguistics*, 4:29-52.
- Biber, D., S. Conrad, R. Reppen. 1998. *Corpus linguistics: investigating language structure and use*. Cambridge : Cambridge University Press.

- Christ, O. 1994. A modular and flexible architecture for an integrated corpus query system. En *Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research*, páginas 23-32, Budapest.
- Clear, Jeremy H. 1993. The British National Corpus. En *The Digital Word: Text-Based Computing in the Humanities*, páginas 163-87, MIT Press (Cambridge).
- Clear, J. G. Fox, G. Francis, R. Krishnamurthy, R. Moon. 1996. COBUILD: The State of the Art. *International Journal of Corpus Linguistics*, 1:303-14.
- Davies, M. 1995. The Evolution of the Spanish Causative Construction. *Hispanic Review*, 63:57-77.
- . 1996. The Diachronic Interplay of Finite and Nonfinite Verbal Complements in Spanish and Portuguese. *Bulletin of Hispanic Studies* (Glasgow), 73:137-58.
- . 1997. The History of Subject Raising in Spanish with Parecer. *Bulletin of Hispanic Studies* (Liverpool), 74:399-411.
- . 1998. The Evolution of Spanish Clitic Climbing: A Corpus-Based Approach. *Studia Neophilologica*, 69:251-63.
- . 2000. Syntactic Diffusion in Spanish and Portuguese Infinitival Complements. En *New Approaches to Old Problems: Issues in Romance Historical Linguistics*, páginas 109-27, John Benjamins (Philadelphia).
- . 2002a. "Esto es ligero de fazer": Object to Subject Raising in Medieval and Early Modern Spanish. En *Papers from the 4th Hispanic Linguistics Symposium*, Cascadilla (Somerville, MA). To appear.
- . 2002b. Diachronic Shifts and Register Variation with the Lexical Subject of Infinitive Construction (*Para yo hacerlo*). En *Papers from the 5th Hispanic Linguistics Symposium*, Cascadilla (Somerville, MA). To appear.
- Rocha, P., D. Santos. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. En *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)*, páginas 131-140, Atibaia (São Paulo).
- Santos, D., E. Bick. 2000. Providing Internet access to Portuguese corpora: the AC/DC project. En *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, páginas 205-210. Institute for Language and Speech Processing (Athens).