スペイン文化シリーズ 11 号
Serie Cultura Hispánica, nº11

El uso del Corpus del Español y otros corpus en la investigación de la variación actual y los cambios históricos

Mark Davies

上智大学　イスパニア研究センター
Centro de Estudios Hispánicos, Universidad Sofía
2004, Tokio

本書の出版にあたってはスペイン教育文化スポーツ省のグラシアン基金より 2003 年度の助成を受けた。
La publicación de este libro ha sido subvencionada en 2003 por el Programa "Baltasar Gracián" del Ministerio de Educación, Cultura y Deporte de España.

ÍNDICE

PRESENTACIÓN

En este número 11 de la Serie Cultura Hispánica del Centro de Estudios Hispánicos de la Universidad Sofía ofrecemos el material usado en el Seminario *El uso del Corpus del Español y otros corpus en la investigación de la variación actual y los cambios históricos*, impartido por el profesor Mark Davies, del Departamento de Lingüística y Lengua Inglesa de la Universidad Brigham Young de Estados Unidos, durante los días 11 al 13 de junio de 2004. Este seminario está dirigido especialmente a estudiantes de postgrado y profesores e investigadores del español en general. Consta de una breve introducción teórica a la lingüística del corpus y se centra en los aspectos prácticos de diversas técnicas avanzadas de investigación lingüística mediante el uso de su macrocropus que el profesor ha puesto en su página web (http://www.corpusdelespanol.org) y que se puede consultar libremente.

El Prof. Davies, es un reconocido especialista de la lingüística del corpus y entre sus intereses se encuentra la investigación de la variación histórica y sintáctica del español, portugués e inglés. Graduado de la Universidad Brigham Young, obtuvo su doctorado en Sintaxis y Lingüística Histórica en la Universidad de Texas, Austin. También ha sido profesor de la Universidad Estatal de Illinois. Es autor de innumerables artículos relacionados con la lingüística del corpus.

Las técnicas de la lingüística del corpus se están valorando cada vez más como herramientas imprescindibles de investigación de casi todas las facetas del complejo fenómeno que es el lenguaje, ya sea en la enseñanza, la investigación del estilo, la lexicografía, o en ayudas a la traducción, etc. por lo que estamos convencidos de la utilidad de este seminario independientemente de la especialidad de los asistentes.

Quiero dejar constancia de nuestro profundo agradecimiento al profesor Mark Davies tanto por la meticulosa preparación de este seminario como por su amistad. Asimismo, esperamos los resultados de sus investigaciones de la

lingüística del corpus tanto del español como del inglés y portugués, así como el diccionario de frecuencias del español, cuya publicación está prevista para el año próximo por la prestigiosa editorial Routledge.

Antonio Ruiz Tinoco
Centro de Estudios Hispánicos
Universidad Sofía, Tokio

El uso del Corpus del Español y otros corpus en la investigación de la variación actual y los cambios históricos

Mark Davies
Department of Linguistics, Brigham Young University
mark_davies@byu.edu
http://davies-linguistics.byu.edu

1. La necesidad de un nuevo corpus histórico

1.0 En los últimos cuatro o cinco años han aparecido varios corpus que tienen gran utilidad para los lingüistas, incluso para los lingüistas históricos. Con estos corpus más amplios, es posible obtener 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 el CORDE, de la Real Academia Española (http://corpus.rae.es/cordenet.html). Este corpus sirve muy bien para las consultas 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_fi**). 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. Además, 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, ya que 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 subida 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. De nuevo, no se podrían buscar ni las varias formas de los verbos, ni los complementos pronominales (como grupo), ni tampoco los infinitivos. Estas construcciones, juntas con otras construcciones relacionadas con el infinitivo, se han investigado anteriormente 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 consultas más avanzadas.

El uso del Corpus del Español y otros corpus en la investigación de la variación actual y los cambios históricos

2. Un bosquejo de las estructuras del Corpus del Español

2.0 Debido a estas limitaciones del CORDE, hace un año se decidió crear otro corpus del español histórico y moderno – el "Corpus del Español" – que contiene 100.000.000 palabras (véase http://www.corpusdelespanol.org). El corpus se terminó en agosto de 2002, y está en funcionamiento actualmente. 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 consultas 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 consultas (y combinaciones aun más complejas) se pueden hacer en muy poco tiempo – por lo general menos de dos o tres segundos.

2.1 El Corpus del Español: los textos

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 proceden 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 [elvira.lllf.uam.es/docs_es/corpus/corpus.html], transcripciones de congresos, y entrevistas periodísticas), y miscelánea (enciclopedias, periódicos, etc).

La tabla siguiente nos da un bosquejo de los textos que se han utilizado en el corpus:

Tabla 1. Textos (bosquejo de lo histórico)

SIGLO	# PALABRAS	# TEXTOS	FUENTES
1200s	6,905,000	71	[HSMS] Electronic Texts and Concordances of the Madison Corpus of Early Spanish Manuscripts and Printings. Preparado por John ONeill. (Madison y New York, 1999). ADMYTE (*Archivo Digital de Manuscritos y Textos Españoles*). Vol 0 y 2. Biblioteca Virtual Gonzalo de Berceo: Obras Completas
1300s	2,820,000	50	
1400s	8,515,000	160	

1200s-1400s	18,240,000	281	
1500s	18,001,000	323	Biblioteca Virtual [1500s-1700s] COMEDIA (Univ. de Arizona) [1600s] Proyecto Filosofía en español [1700s]
1600s	12,746,000	499	
1700s	10,263,000	159	
1500s-1700s	41,010,000	981	
1800s	20,465,000	392 novelas	Biblioteca Virtual
1900s-Lit	6,750,000	850 novelas y cuentos	
1900s-Oral	6,800,000	2040+ entrevistas y transcripciones	
1900s-Misc	6,800,000	4770+ artículos (más detalles abajo)	
1800s-1900s	40.815.000	8052	
TOTAL	100.000.000	9314	

La tabla siguiente presenta más detalles sobre los textos del siglo XX. Nótese que para obtener un corpus equilibrado entre varios registros en los 20.000.000 de palabras, hemos incluido la tercera parte del español oral, la tercera parte de literatura (novelas y cuento cortos), y la tercera parte de textos no-literarios (p. ej. enciclopedias y periódicos).

Tabla 2. Textos: español moderno

	# palabras[1]	España	# palabras[1]	América Latina
Hablado	1.00	España Oral[2]	2.00	Habla Culta (diez países)
	0.35	Habla Culta (Madrid, Sevilla)		
3.35	1.35		2.00	
Transcripciones/ dramas	1.00	Transcripciones/ Entrevistas (congresos, ruedas de prensa, etc.)	1.00	Transcripciones/ Entrevistas (congresos, ruedas de prensa, etc.)
	0.27	Entrevistas en el periódico ABC		
	0.40	Dramas	0.73	Dramas
3.40	1.67		1.73	
Literatura	0.06	Novelas (BV[3])	1.60	Novelas (BV[3])
	0.00	Cuentos cortos (BV[3])	0.87	Cuentos cortos (BV[3])
	0.19	Tres novelas (BYU[4])	1.11	Doce novelas (BYU[4])
	2.17	Principalmente novelas, de LEXESP[5]	0.18	Cuatro novelas de Argentina[6]

			0.20	Tres novelas de Chile[7]
6.38	2.42		3.96	
Textos	1.05	Periódico ABC	3.00	Periódicos de seis países
	0.15	Ensayos in LEXESP[5]	0.07	Cartas de Argentina[6]
	2.00	Enciclopedia Encarta	0.30	Textos humanísticos (p.ej. filosofía e historia, de Argentina[6])
			0.30	Textos humanísticos (p.ej. filosofía e historia, de Chile[7])
6.87	3.20		3.67	
Total	8.64		11.36	

Notas:
1. Tamaño en millones de palabras
2. *Corpus oral de referencia de la lengua española contemporánea* (http://elvira.lllf.uam.es./docs_es/corpus/corpus.html)
3. *Biblioteca Virtual* (http://www.cervantesvirtual.com)
4. 15 novelas, adquiridas en forma electrónica del Humanities Research Center, Brigham Young University
5. *Léxico informatizado del español* (http://www.edicionsub.com/coleccion.asp?coleccion=90)
6. Del *Corpus lingüístico de referencia de la lengua española en argentina* (http://www.lllf.uam.es/~fmarcos/informes/corpus/coarginl.html)
7. Del *Corpus lingüístico de referencia de la lengua española en chile* (http://www.lllf.uam.es/~fmarcos/informes/corpus/cochile.html)

2.2. La arquitectura del corpus: una comparación con otros corpus grandes

Los 100.000.00 de palabras de texto están en una base de datos relacional 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 consultas 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 consultas 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 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 de "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.

Además, 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 aún 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.

2.3 La arquitectura del Corpus del Español: n-grams con frecuencias

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 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 consultas 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

[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 central es la categoría gramatical y el lema. Esta información se ha unido con los 45.000.000 n-grams distintos y su fuente es un diccionario que contiene 500.000 formas distintas, el cual se puede ver a continuación:

Tabla 3. Diccionario de formas / lemas / categoría

forma	lema	categoría
trabajaron	trabajar	v_pret
abuelas	abuelo	N

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:

Tabla 4. N-grams / frecuencia / lema / categoría[2]

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	...

Esta entrada para el 3-gram "*son las cosas*" es un ejemplo de los 33.000.000 de 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).

2.4 El acceso a la base de datos por medio del Web

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 insertan lo siguiente:

ser.* las *.n

y en el servidor se convierte en el 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

[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).

parcial de las 300+ frases, ordenado por frecuencias en los 1900s (muchas de las frases más frecuentes se han omitido)[3]:

Tabla 5: N-grams / frecuencias – resultados

#	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
...	...	...	...	...	...	...	...	...	...	...	...	...

La rapidez con que se realizan las consultas 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 obtener 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 consulta. Con nuestro sistema, son raras las consultas que toman más de dos o tres segundos.

Muchos sólo quieren ver el listado de palabras y frases que resultan de la consulta, pero los que quieren ver una palabra o frase en contexto pueden hacer clic 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 consulta también es muy rápida (~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 a continuación – pueden reordenar las ocurrencias por las palabras a la izquierda o a la derecha, y pueden hacer clic en cualquier ejemplo para ver más contexto (hasta un párrafo):

[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 por cada millón de palabras, o los dos a la vez.

Tabla 6. Palabras clave en contexto (KWIC)

#	=	MÁS	RE-ORDENAR POR: I-2 I-1	C	D-1 D-2
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 ...
5	...	...	...	...	...

3. Un bosquejo de las consultas con el Corpus del Español

3.0 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 consultas 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 consultas, 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 Los comodines y los padrones de palabras

Como en CORDE, en el Corpus del Español se pueden hacer consultas con comodines, por ejemplo *descri**. La diferencia es que con CORDE muchas de estas consultas terminan después de varios segundos sin producir resultados, porque el motor de búsquedas de CORDE en realidad no fue diseñado para consultas complejas con comodines. Con el Corpus del Español, sin embargo, esta consulta 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:

Tabla 7. Uso de comodines / frecuencia por siglos

	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
...	...	...	...	...	...	...	...	...	...	...

A diferencia de CORDE y CREA, el Corpus del Español también permite el uso de comodines 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 consulta de *–azo,* ordenados por su frecuencia en el siglo XVI:

Tabla 8. Sufijos

#	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		
...	...	...	...	...	...	...	...	...	...	...	...	...

Obviamente, la capacidad 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.

La tabla siguiente presenta más detalles sobre la sintaxis de las consultas con comodines:

Tabla 9. Los comodines

Uso		Ejemplos	
_	una letra	**p_ra**	para, pera, pora, pura
*	varias letras	**p*ra**	para, postura, practicara
_ _	patrones más complejos	**comen_a_an**	comenzaban, començauan
* *		***esper***	inesperado, esperanza
_ *		**p_ra***	parajes, paran, purana

3.2 Las 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]:

Tabla 10. Las colocaciones

#	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		
...	...	...	...	...	...	...	...	...	...	...

Como se ha indicado antes, la base de datos también contiene información sobre la categoría y el lema, y se pueden incluir estos en la consulta – 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]:

Tabla 11. Las colocaciones con categoría gramatical

#	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	
...	...	...	...	...	...	...	...	...	...	...

La tabla siguiente presenta más detalles sobre la sintaxis de las consultas para encontrar las colocaciones más frecuentes:

Tabla 12. Las colocaciones (sintaxis)

Uso		Ejemplos	
* P2	ver las palabras que preceden	* suave	voz / viento suave
P1 * P3	ver las palabras en medio de otras	*tan * como*	*tan bueno / rápido como*
P1 P2 *	ver las palabras que siguen	dice que *	dice que hay / los / puede

3.3 El lema y la categoría gramatical

En el ejemplo anterior se ve cómo se ha usado la información sobre categoría gramatical como parte de la consulta. Ya que la base de dat contiene información sobre el lema, esto también puede formar parte de consulta – por ejemplo, una tabla que muestre todas las formas de *decir (di dize, dezir, dizen, dixieron*) o abuelo (*abuela, abuelo, abuelos, abueli* durante los últimos 800 años, junto con su frecuencia relativa en cada sig Esto también se puede combinar con la categoría gramatical para estudi construcciones bastante complejas como la de "Object to Subject Raisin ("Subida del Objeto") (Davies, 2002a). Estas construcciones tienen uno varios adjetivos (como *fácil* e *imposible*) + *de* + infinitivo (categor gramatical):

Tabla 13. La categoría gramatical y lema

	PALABRA(S)	...	15	16	17	18	19	Lit	Oral	Mis
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
...	...	...	...	...	...	...	...	...	...	...

Las dos tablas a continuación presentan más detalles sobre la sintax de las consultas que incluyen información sobre los lemas y la categor gramatical.

Tabla 14. Los lemas

Uso		Ejemplos	
palabra.*	todas las formas de la palabra indicada	***abuelo.****	abuelo, abuelita, auuelo
		decir.*	*dijo, dirá, dixeron*
		haber.*	habrá, havia, ouo

Tabla 15. La categoría gramatical

Uso		Ejemplos	
*.categoría	todas las palabras en esa categoría gramatical	***.prep***	con, hasta, par
	Todos los casos de *ver* + *el* + un sustantivo	***ver el *.n***	*ver el coche / libro / edificio*
	Todos los casos de una forma de	***difícil.* de***	difícil de

	difícil + de + un infinitivo	***.v_inf**	entender difíciles de leer

3.4 Los sinónimos

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 consulta 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 consulta 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', 'avispado', 'lúcido', 'capaz', 'ingenioso', 'versado', 'espabilado')

Esto produce resultados como el siguiente:

Tabla 16. Los sinónimos

#	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				
...	...	...								

Las siguientes dos tablas presentan más detalles sobre la sintaxis de los sinónimos y antónimos.

Tabla 17. Los sinónimos y antónimos (sintaxis)

Uso		Ejemplos	
!palabra	sinónimos	***!hablar***	decir, discutir, charlar
#palabra	*antónimos*	***#rico***	*pobre, indigente*

*!palabra.**	sinónimos – todas las formas	***!inteligente.****	astuto, astutas, lúcido

3.5 La 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 introducir lo siguiente en el formulario en el web:

Tabla 18. Limitar y ordenar por frecuencia

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

Esto saca 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 ordena según el número de ocurrencias en el siglo XV, p. ej:

Tabla 19. La frecuencia – resultados

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

Las tablas siguientes presentan más detalles sobre la sintaxis de las consultas que incluyen información sobre la frecuencia en los distintos periodos históricos o en los registros del español moderno.

Tabla 20. La frecuencia (sintaxis)

Escoger en "Límites"	Uso	Ejemplos
+siglo *-siglo*	Ocurre (o no ocurre) en ese siglo	**+1300s** ***-1800s***
siglo>X *siglo<X*	*Por lo menos X ocurrencias* *Menos de X ocurrencias*	***19lit>5*** ***19oral<2***
	No ocurre en los 1200s, 1300s, 1400s, pero ocurre más de dos veces en los 1500s	-1200s -1300s -1400s 1500s>2

La tabla siguiente presenta más ejemplos de este tipo de consultas:

Tabla 21. Más ejemplos de frecuencias

	Consulta	Resultado	Explicación
Comodines	*udo +1500s -1800s -1900s	*cabezudo, vedijudo, capilludo*	palabras con el sufijo [-udo], que aparecen en los 1500s, pero no en los 1800s o 1900s
Colocaciones	*.n duro.* +1900s -1500s -1600s	*línea, disco, cabeza,*	sustantivos que ocurren con una forma de [duro] en los 1900s, pero no en los 1500s o 1600s
Lema	haber.* 1500s>5 -1900s	*avia, uvo, obiese*	las formas de [haber] que ocurren por lo menos 5 veces en los 1500s, pero no ocurren en los 1900s
Categoría gramatical	*.v_inf +1900s -1800s	*detectar, frenar, intercambiar*	infinitivos que ocurren en los 1900s pero no en los 1800s (i.e. son verbos nuevos)
Sinónimos antónimos	!hablar.* +1900s -1700s +lema	*comentar, dialogar, platicar*	Sinónimos de *hablar* (*to speak*) que ocurren en los 1900s pero no en los 1700s, agrupados por lema
Combinaciones de las más sencillas	!mandar.* que *.v_subj_ra +1900s -1800s -1700s	*hizo que dijeran mandé que volviera*	Cualquier forma de cualquier sinónimo de *mandar* + *que* + el imperfecto del subjuntivo, que ocurre en los 1900s pero no en los 1700s o los 1800s

3.6 Las consultas más complejas

Por supuesto, el usuario también puede hacer cualquier combinación d consultas. Por ejemplo, la consulta puede incluir la categoría gramatical, lema, los sinónimos (que veremos en la próxima sección), y la frecuencia, sin embargo la consulta de las 100.000.000 palabras se realiza en poc segundos, y nos da más de 300 formas distintas (*le quiero decir, se des obtener, le apeteciera entrar*, etc.).

La tabla siguiente nos da unos ejemplos de estas consultas compleja

Tabla 21. Las consultas más complejas

Ejemplos		
me/le/les quiero/quieres decir	[*me* O *le* O *les*] y [*quiero* O *quieres*] y *decir*	me quieres decir le quie decir
le/les !querer.* *.v_inf	le or les + una forma de un sinónimo de *querer* + cualquier infinitivo	lo quiero compra le desear saludar
*se *.pn_obj ver.**	*se* + complemento pronominal + una forma de *ver*	se les vea se me viera
!mandar.* que *.v_subj_se	Una forma de un sinónimo de mandar + otra palabra + un subjuntivo en *–ra*	

3.7 Una arquitectura abierta

Aparte de la base de datos de sinónimos, también hay planes de crear unir otras que tienen información sobre etimologías y traducciones entre inglés y el español. Por último, el usuario mismo también puede crear s propias bases de datos en el momento de hacer la consulta – por ejempl campos semánticos especializados (la ropa, términos deportistas, etc) sintácticos (p. ej. verbos transitivos que toman cierto tipo de complement directo) – y después puede usarlas otro día como parte de otra consulta. L importante es que – debido a la arquitectura abierta del corpus – se pued

unir cualquier otra base de datos a la base de datos central y después usarla de manera muy fácil en la consulta.

4. El uso del Corpus del Español en la investigación de la variaci sintáctica actual

Sin duda, uno de los propósitos principales de usar los corpus es el describir (y quizás explicar) la variación entre los varios registros, q incluyen el español hablado, los textos novelísticos, y los textos n novelísticos (p. ej. los periódicos y las enciclopedias). Los siguient ejemplos de la variación actual entre los registros se basan en datos d Corpus del Español, y nos muestran cómo se pueden usar para investigar variación en el español actual.

Lo siguiente se da en forma muy esquemática – principalmente una ser de tablas, con un poco de explicación. Se basa en los 20.000.000 de palabr del siglo XX en el Corpus del Español. Viene en gran parte de un estud que realizo actualmente con Douglas Biber de Northern Arizona Universit el cual ha sido patrocinado por la National Science Foundation de los EEU El tema del proyecto es "un estudio multidimensional de la variaci sintáctica del español moderno e histórico", y emplea la misma metodolog sobre la variación sintáctica que ha usado Douglas Biber (1988, 1995, 200 en muchos otros estudios.

4.1 Las categorías léxicas

La siguiente tabla presenta los datos sobre la distribución de las part del habla en los tres registros. Obsérvese el alto porcentaje de sustantivos d los textos no-novelísticos (que principalmente presentan información). En español hablado, sin embargo, hay más verbos y adverbios, al ser este tipo d lenguaje es más "interactivo".

Tabla 22. Variación en la frecuencia de las categorías gramaticales

	Porcentaje			Porcentaje		
	Hablado	+Ficción	- Ficción	Hablado	+Ficción	- Ficc
sustantivo	19.5	24.7	32.4	279,269	272,723	608,6
verbo	19.4	18.6	12.0	276,820	205,883	225,5
adjetivo	4.0	4.5	7.2	56,651	50,099	134,9
adverbio	10.5	5.8	3.1	150,000	64,231	58,7
pronombre	9.3	7.2	3.1	133,317	80,086	58,1
conjunción	7.0	6.1	5.0	100,340	66,982	94,2
determinante	3.5	3.5	2.7	49,588	38,281	50,2
preposición	12.1	15.0	18.4	172,306	165,859	345,4
artículo	9.0	11.5	13.9	127,920	127,562	260,9
interrogativo	3.5	2.7	1.6	50,431	29,399	30,7

4.2 El tamaño de los sustantivos y la nominalización

Los datos otra vez nos indican que la complejidad de los sustantivos es más común en los textos no-novelísticos, que más que nada proporcionan información sobre cosas y procesos:

Tabla 23. Variación en el tipo de sustantivo

	Hablado	+ Ficción	- Ficción
Promedio del # de letras en los sustantivos	6.45	6.60	7.16
# nominalizaciones / cada millón de palabras	11211	8298	27384

4.3 La referencia pronominal

La siguiente tabla tiene que ver con la referencia pronominal e indica que la referencia a primera y segunda persona (yo, nosotros, tú, etc.) es mucho más común en la conversación [62% vs 5%] (lo cual no es tan sorprendente). También, hay muchos más sujetos [38.4] en la conversación que en los textos escritos [3.3%]. Por fin, hay más uso de los objetos indirectos en la conversación, porque se refiere a las personas, mientras que en los textos (especialmente los textos no-novelísticos) se refiere a las cosas y los procesos.

Tabla 24. Variación en el uso pronominal

	Hablado	+ Ficción	- Ficción
1-SUBJ	15799	3108	89
1-OBJ	16193	8499	106
2-SUBJ	928	993	18
2-OBJ	4596	3388	108
3-SUBJ	6541	5717	1887
3-OD	9265	9555	3227
3-OI	7296	8010	1296
	60618	39270	6731
1/2 persona	61.9%	40.7%	4.7%
+SUBJ	38.4	16.2	3.3
% OI (3a persona)	44.1%	45.6	28.7

4.4 El tiempo y el aspecto

Los datos más importantes del cuadro siguiente son los siguientes:

- Se usa mucho más el presente en la conversación y los textos novelísticos, mientras que se usa mucho más el pasado en los te novelísticos (por ser narración).
- Se extiende esta distinción también al perfecto simple vs. pluscuamperfecto.
- Se usa el subjuntivo más en la ficción que en el español hablado, qu una función de la relación entre los deseos y la fuerza de un perso sobre otro personaje, etc. Sin embargo, el subjuntivo todavía es bast común en la conversación, pero menos en los textos no-novelísticos los que sólo se reportan los datos).
- Hay más uso de las construcciones progresivas en la conversación que tiene que ver con la cuestión de "¿qué me/nos pasa en momento?"

Tabla 25. Variación en el uso de los tiempos y el aspecto

	Porcentaje			# de frecuenci	
	Hablado	+Ficción	-Ficción	Hablado	+Ficción
Indicativo					
presente	61.3	33.6	45.8	132347	48233
pretérito	11.0	23.8	30.2	23790	34115
imperfecto	13.6	26.8	13.4	29367	38456
futuro	0.8	1.5	0.7	1740	2183
condicional	1.4	1.9	1.0	3008	2677
perfecto	3.9	1.4	3.1	8497	2001
pluscuamperfecto	0.7	2.8	1.4	1450	4080
Subjuntivo	5.8	7.4	4.3		
Presente	4.2	3.3	2.9	8960	4668
Imperfecto	1.3	3.6	1.3	2841	5105
Perfecto	0.1	0.1	0.1	251	76
pluscuamperfecto	0.2	0.6	0.1	414	795
Progresivo	1.4	0.7	0.2	3087	1019

4.5 El modo (el uso del subjuntivo) – con N, V, y ADV/CONJ

El cuadro siguiente indica con qué tipo de elemento gramatical (ver sustantivo, adverbio/conjunción, etc) se usa más el subjuntivo en los var registros. Los datos más importantes son los siguientes:

- Hay mucho más uso del subjuntivo con sustantivos de lo que quizás se esperaría (*busco alguien que me ayude*), especialmente cuando se considera la explicación que generalmente se da en los libros de texto
- Los textos no-novelísticos tienen casi el doble de frecuencia con sustantivos que con verbos. Esto resulta algo extraño, más que nada cuando se considera que el uso con verbos es supuestamente el uso "prototípico" del subjuntivo.

Tabla 26. Variación en el uso del subjuntivo

	Porcentaje			# de ocurrencias		
Qué produce	Hablado	+Ficción	-Ficción	Hablado	+Ficción	-Ficción
V	39.2	45.8	26.1	756	633	201
N	42.0	30.6	55.4	811	423	426
PREP	18.9	23.6	18.5	364	326	142
TOTAL				1931	1382	769

4.6 El modo -- verbos específicos

El cuadro siguiente indica con qué verbos es más común el subjuntivo en los distintos registros. Los datos más importantes son los siguientes:

- Hay más énfasis en verbos de [creencia] y [reacción emocional] en la conversación
- Hay más énfasis en verbos de [control personal] en la ficción que en la conversación
- Hay más énfasis en verbos de [control impersonal] en los textos no-novelísticos (es decir, se refiere a las condiciones que hacen que ocurra cierta cosa)

Tabla 27. Verbos que piden el subjuntivo

Hablado		+ Ficción		- Ficción	
98	querer	72	querer	37	hacer
95	creer	57	pedir	18	permitir
65	decir	31	decir	12	impedir
35	esperar	29	esperar	15	evitar
29	gustar	25	dejar	10	esperar
22	hacer	21	creer	4	lograr
16	dejar	15	parecer	5	exigir
12	pedir	13	proponer	4	ordenar
12	parecer	11	gustar	4	poder

12	poder	11	impedir	4	querer
11	importar	10	permitir	3	dejar
10	ver	10	temer	3	pedir
10	pensar	8	ordenar	2	temer
9	suponer	8	preferir	2	creer
5	necesitar	6	rogar	2	intentar
5	poder	5	conseguir	2	parecer
4	preferir	4	exigir	2	pensar
4	perdonar	4	servir	2	significar
4	aceptar	4	hacer	2	suponer

Otra vez, vemos dos cosas importantes con estos datos. Primero, h bastantes diferencias importantes entre la sintaxis de los distintos registr del español. Segundo, la capacidad de obtener los datos sobre la variaci semántica sólo se puede hacer con un corpus como el Corpus del Españ pero jamás se podría con un corpus no anotado, como CORDE o CREA de Real Academia Española.

5. El uso del Corpus del Español en la investigación de la variación léxica actual

5.0 Otro uso de un corpus grande es el de poder saber qué palabras son más comunes, para crear un "diccionario de frecuencias". Tal diccionario puede ayudar mucho a los principiantes, ya que podemos saber qué palabras ocurrirán más frecuentemente en la conversación o en la literatura. Así, pueden maximizar mejor su tiempo en aprender el vocabulario de otra lengua.

Desgraciadamente, no hay tal diccionario para el español. La siguiente es una lista de diccionarios que han aparecido hasta la fecha, pero cada uno tiene por lo menos un defecto grande:

- Buchanan, M.A. (1927). *A Graded Spanish Word Book.* Toronto: Univ. of Toronto Press.
- Eaton, H. (1940). *An English -French - German - Spanish Word Frequency Dictionary.* New York: Dover Publications.
- Rodriguez Bou, L. (1952) *Recuento de Vocabulario Español.* Rió Piedras : Universidad de Puerto Rico.
- García Hoz, V. (1953). *Vocabulario Usual, Vocabulario Común y Vocabulario Fundamental.* Madrid: CSIC.
- Juilland, A. & Chang-Rodríguez, E. (1964). *Frequency Dictionary of Spanish Words.* The Hague: Mouton.
- Alameda, J.R. & Cuetos, F., ------. (1995). *Diccionario de Frecuencias de las Unidades Lingüísticas del Castellano,* Oviedo: Universidad de Oviedo
- Sebastián, N., Martí, M.A., Carreiras, M.F. & Cuetos, F. ------. (2000), *LEXESP, Léxico Informatizado del Español.* Barcelona: Ediciones de la Universitat de Barcelona. (sólo en CD-ROM)

No hay ninguno que tenga todas las características siguientes:

- Se basa en el español escrito (tanto novelístico como no-novelístico) y el español hablado.
- Cubre el Español de España y de la América Latina.
- Se basa en un corpus grande -- de 10.000.000 palabras o más.
- El corpus está lematizado (p. ej. agrupación por verbos: *hacer = hacen, hará, hiciste*).

Por ejemplo, el mejor diccionario hasta la fecha es el de Juilland y Chang-Rodríguez (1964), que se basó en un corpus de solamente 1.000.000 palabras de solamente la literatura de España. Como el corpus es tan débil,

hay palabras como [duque] y [duquesa] que aparecen entre las 700 pala más frecuentes del español, cuando se sabe bien que casi nunca encontrarían tales palabras en la conversación normal.

Debido a estas debilidades en los diccionarios anteriores, he deci crear un nuevo Diccionario de Frecuencias del Español, que será publi por Routledge en 2005. Se basará en los 20.000.000 de palabras del s XX en el Corpus del Español, que incluye textos de conversación (la ter parte), textos novelísticos (la tercera parte), y textos no-novelísticos tercera parte). Los textos proceden tanto de España como de la Amé Latina, y han sido cuidadosamente lematizados para encontrar las pala más comunes. Los siguientes datos provienen de los que se han recogid para este diccionario.

5.1 El alcance del vocabulario

Un hecho interesante es que pocas palabras muy comunes representa gran mayoría de las palabras que se encuentran en la conversación literatura. Por ejemplo, las 1000 palabras más comunes en la conversa representan casi el 88% de todas las palabras en una conversación típica. un poco menos para los textos escritos (76-80%), al ser su vocabulario detallado y técnico. Otra cosa que indican los datos es que cada gr adicional de palabras (p. ej. #2001-3000 en frecuencia) proporciona me alcance adicional. Por ejemplo, al aprender 2000 palabras adicion después de las 1000 más comunes (es decir, #1001-3000), sólo sub alcance del 88% al 94%.

Tabla 28. El alcance del vocabulario en registros distintos

Grupo de palabras	- Ficción (periódicos y enciclopedias)	+ Ficción	Oral
1-1000	76.0	79.6	87.8
1001-2000	8.0	6.5	4.9
2001-3000	4.2	3.5	2.3
TOTAL: 1-3000	88.2	89.6	94.0

5.2 Unos datos más detallados sobre el alcance de palabras

Los tres cuadros siguientes presentan más detalles sobre el alcance. ejemplo, el cuadro siguiente se refiere a los textos orales. Indica conociendo los 2578 sustantivos más comunes en el sub-corpus del esp hablado (6.5000.000 palabras), se cubre el 88.8% de todas las ocurrencia sustantivos en este corpus. Los 231 verbos más frecuentes proporciona 90% de alcance, y hay 985 adjetivos y 57 adverbios que proporcio aproximadamente el 90% de alcance. Es decir, con aproximadamente 4

palabras en el español hablado, se reconocería más o menos el 90% de las palabras -- sustantivos, verbos, adjetivos, y adverbios -- en una conversación típica.

Tabla 29. El alcance del vocabulario: español hablado

		N			V			ADJ			ADV		
2%	258	50.4	140874	47	74.9	207378	66	43.4	24607	12	65.6	98467	3851
5%	645	67.5	188587	116	84.7	234407	165	61.6	34885	29	83.9	125910	
10%	1289	79.3	221496	231	90.6	250911	329	75.1	42537	57	93.1	139599	
20%	2578	88.8	248045	462	98.6	272908	657	85.9	48645	114	97.8	146725	
30%	3867	93.0	259611	693	99.1	274421	985	90.9	51493	171	99.0	148521	
40%	5155	95.3	266233	923	98.6	272908	1314	93.9	53192	227	99.4	149168	
50%	6444	96.9	270563	1154	99.1	274421	1642	95.9	54306	284	99.7	149497	
60%	7733	97.9	273355	1385	99.5	275346	1970	97.2	55075	341	99.8	149686	
70%	9021	98.6	275403	1615	99.7	275958	2299	98.3	55667	397	99.9	149804	
80%	10310	99.1	276692	1846	99.8	276359	2627	98.8	55995	454	99.9	149887	
90%	11599	99.5	277981	2077	99.9	276590	2955	99.4	56323	511	100.0	149944	
	12897		279269	2314		276820	3293		56651	568		150000	

Como se ve en los dos cuadros siguientes, hay que tener un vocabulario más amplio para reconocer el mismo porcentaje de palabras en los textos escritos, porque emplean un vocabulario más detallado y técnico. Por ejemplo, para los textos novelísticos, sube a más de 70000 palabras (Tabla 30), y para los textos no novelísticos, son casi 8000 palabras (Tabla 31).

Tabla 30. El alcance del vocabulario: escrito / +ficción

		N			V			ADJ			ADV		
2%	314	41.4	112961	61	57.1	117649	83	37.8	18954	15	61.2	39286	7043
5%	784	58.1	158489	152	70.5	145228	208	53.4	26764	37	83.0	53317	
10%	1568	71.6	195378	304	80.8	166441	415	66.1	33104	74	93.7	60213	
20%	3135	84.1	229326	607	89.8	184926	830	78.8	39484	148	97.0	62306	
30%	4702	90.1	245845	910	94.1	193715	1245	85.9	43021	221	98.0	62977	
40%	6269	93.6	255389	1214	96.4	198529	1660	90.4	45280	295	98.7	63384	
50%	7836	95.8	261345	1517	97.8	201383	2074	93.5	46822	368	99.1	63650	
60%	9404	97.2	265221	1820	98.7	203226	2489	95.7	47931	442	99.4	63842	
70%	10971	98.3	268022	2124	99.3	204399	2904	97.3	48761	516	99.6	63990	
80%	12538	98.9	269589	2427	99.6	205139	3319	98.3	49270	589	99.8	64084	
90%	14105	99.4	271156	2730	99.9	205580	3734	99.2	49685	663	99.9	64158	
			272723			205883			50099			64231	

Tabla 31. El alcance del vocabulario: escrito / -ficción

		N			V			ADJ			ADV		
2%	601	52.4	319013	51	47.9	108047	87	41.0	55303	12	63.3	37206	7870
5%	1502	70.8	431095	127	63.7	143597	217	58.2	78474	29	79.6	46794	
10%	3003	82.6	502863	254	77.2	174197	433	72.6	97974	58	88.8	52177	
20%	6005	90.9	553108	508	89.6	202030	866	85.4	115171	115	94.8	55727	
30%	9008	94.2	573372	762	94.6	213411	1299	91.2	123071	173	97.2	57101	
40%	12010	96.0	584484	1015	97.1	218937	1732	94.6	127604	230	98.3	57756	
50%	15013	97.2	591460	1269	98.3	221837	2164	96.7	130394	288	98.9	58138	
60%	18015	98.0	596600	1523	99.1	223477	2597	97.9	132140	345	99.3	58386	
70%	21018	98.5	599603	1776	99.5	224430	3030	98.8	133274	403	99.6	58551	
80%	24020	99.0	602605	2030	99.7	225004	3463	99.4	134043	460	99.8	58655	
90%	27023	99.5	605608	2284	99.9	225316	3896	99.7	134476	518	99.9	58713	
			608610			225569			134908			58770	

En el Diccionario de Frecuencias del Español (Routledge, 2005), tendremos entre 5000-6000 palabras, lo que nos dará aproximadamente el 90% de alcance (según el registro específico)

5.3 La importancia de la "distribución"

Al crear (y usar) un diccionario de frecuencias, hay que fijarse no solamente en la frecuencia básica, sino que también es importante la "distribución" -- es decir, en cuántas secciones del corpus ocurre la palabra. Por ejemplo, supongamos que dividimos el corpus en 100 secciones, con la misma cantidad de palabras en cada una. Supongamos que una palabra tiene buena frecuencia -- ocurre muchas veces, pero solamente en 8 de las 100 secciones (por ejemplo, sólo los textos que se refieren a la ciencia, o ficción que habla de la realeza). ¿Incluiríamos tales palabras en nuestro diccionario, o sólo las que tienen mejor distribución?

El cuadro siguiente es ejemplo de la diferencia en la distribución. Todas las palabras tienen más o menos la misma frecuencia -- entre 50-70 ocurrencias por cada millón de palabras. Sin embargo, la distribución es muy diferente. Las palabras a la izquierda ocurren en por lo menos 60 de las 100 secciones del corpus, mientras que las que están a la derecha ocurren en menos de 15 secciones. Como se puede ver, las palabras que tienen mejor distribución probablemente serían mejores candidatas para el diccionario que las que solamente ocurren en un tipo de textos muy limitado.

Tabla 32. La relación entre la frecuencia y la distribución

distribución > 60 (/100)			distribución <= 15 (/100)		
distribución		frecuencia	distribución		frecuencia
68	notable	66	7	verbo	69
67	falta	63	7	cromosoma	58
66	introducción	69	7	neutrón	56
65	preocupación	65	9	bailarín	65
64	propósito	69	9	sonata	62
64	disposición	69	10	cirugía	56
64	empleo	68	11	galaxia	54
64	peligro	67	12	enciclopedia	54
64	difusión	62	13	glándula	61
64	duda	57	13	fármaco	59
62	protección	69	13	filo	58
62	clave	60	14	orquesta	58
62	reconocimiento	59	14	jazz	57
62	precedente	58	14	corán	53
62	complete	54	15	turbina	61
61	impacto	55	15	enzima	54
61	margen	53			

El registro también importa en la distribución. El cuadro siguiente nos muestra que puede haber buena distribución en cierto registro, pero no en los otros. Estas palabras ocurren en casi todas las secciones del sub-corpus de los textos orales, pero en muy pocas del sub-corpus de enciclopedias. Como es más probable que los estudiantes necesiten saber el vocabulario de la conversación, se daría preferencia a estas palabras comunes y con buena distribución de ese registro.

Tabla 33. La diferencia de distribución entre los registros (+oral/-[-ficción])

distribución			verbo	frecuencia - (x/1.000.000 palabras)		
diferencia	oral	enciclopedia		diferencia	oral	enciclopedia
90	99	9	gustar	1296	1301	5
85	97	12	preguntar	266	272	6
82	88	6	meter	315	318	3
75	94	19	imaginar	247	258	11
68	71	3	encantar	161	162	1
65	86	21	casar	309	324	15
64	96	32	tocar	296	356	60
62	77	15	echar	177	184	7
62	85	23	faltar	140	157	17
62	89	27	comprar	331	355	24

60	79	19	costar	149	161	12
60	89	29	mandar	212	240	28
60	96	36	mirar	615	640	25
58	96	38	oír	483	524	41
57	85	28	andar	201	230	29
54	86	32	comer	328	370	42
49	87	38	acordar	315	373	58
47	79	32	olvidar	92	115	23
46	82	36	valer	170	192	22
45	71	26	bajar	116	135	19

El cuadro siguiente indica lo opuesto. Estas palabras son muy com y tienen buena distribución en los textos descriptivos (enciclopedi periódicos), pero mucho menos en la conversación. Quizás no s candidatas tan buenas para el diccionario.

Tabla 34. La diferencia de distribución entre los registros (-oral/+[-ficci

distribución			verbo	frecuencia - (x/1.000.000 pala		
diferencia	enciclopedias	oral		diferencia	enciclopedias	
87	100	13	denominar	455	468	
82	93	11	contener	314	326	
80	86	6	añadir	135	142	
80	97	17	situar	326	346	
78	93	15	obstar	230	244	
77	88	11	sustituir	148	160	
76	87	11	combinar	163	177	
76	92	16	proporcionar	229	239	
75	89	14	mediar	187	196	
73	97	24	componer	209	240	
73	93	20	contribuir	110	140	
73	92	19	adoptar	282	308	
73	90	17	poseer	171	187	
72	100	28	constituir	325	359	
72	97	25	introducir	230	256	
72	98	26	extender	307	330	
71	85	14	caracterizar	182	194	

Al contrario de lo que quizás se pudiera esperar, también hay difere entre la literatura y la conversación. Las palabras en el cuadro siguient comunes y tienen buena distribución en la literatura, pero no conversación.

Tabla 35. La diferencia de distribución entre los registros (escrito/± ficción)

distribución			verbo	frecuencia – (x/1.000.000 palabras)		
diferencia	literatura	oral		diferencia	literatura	oral
78	84	6	sonreír	281	288	7
76	78	2	murmurar	232	233	1
70	90	20	soltar	210	241	31
70	93	23	gritar	396	420	24
69	72	3	acariciar	182	184	2
68	75	7	besar	200	206	6
68	75	7	oler	157	163	6
67	81	14	encender	204	216	12
66	78	12	asomar	139	148	9
66	78	12	envolver	135	143	8
63	71	8	alzar	231	237	6
63	77	14	apretar	214	225	11
61	93	32	reír	376	417	41
61	76	15	colgar	169	183	14
60	76	16	arrancar	146	160	14

Es decir, para crear un buen diccionario de frecuencias (como hemos intentado nosotros) hay que tener (o tener en cuenta) lo siguiente:

- Un corpus grande (el nuestro tiene 20.000.000 palabras)
- Textos de muchos registros (el nuestro tiene de textos orales, de literatura, y de textos no-novelísticos)
- Un corpus bien lematizado y etiqueteado
- Hay que tomar en cuenta la frecuencia absoluta, pero también la distribución (y decidir qué registros importarán más -- si así es -- en la lista de palabras)

5.4 Un bosquejo el diccionario final

Nuestro diccionario tendrá varios índices -- y con todos estos se espera que se pueda tener una buena idea de la distribución de las palabras más comunes en el español -- desde varios puntos de vista. Primero, el listado fundamental tendrá las 5000-6000 palabras más comunes en orden de frecuencia (las más comunes primero). Cada entrada tendrá:

1) la palabra (lema)
2) su categoría gramatical
3) una traducción al inglés
4) un ejemplo de la palabra en contexto -- sacado del Corpus del Español

5) la frecuencia absoluta, y
6) (optativamente) una indicación de la distribución de la palabra, s mucho más común en cierto registro, o dialecto geográfico (Espa América Latina)

El cuadro siguiente es ejemplo de este listado:

Tabla 36. El diccionario de frecuencias: índice por frecuencia

1500 asociación *n* 'association' *en estos países no existen las asociaciones de socorro* 1199	1509 salón n 'room, hall' com llegara a un salón de clases 1193
1501 perfectamente *adv* 'perfectly' *sabiendo perfectamente lo que andan diciendo* 1199 s	1510 cifra *n* 'figure, number' *grandes cifras macroeconómica nuestro país* 1192
1502 zapato *n* 'shoe' *se limpió el polvo de los zapatos* 1199	1511 hueso *n* 'bone' *era osteomielitis del hueso frontal* 11
1503 manejar *v* 'to handle, manage, drive' *aprendí a manejar aquél coche* 1197 L/s	1512 monte *n* 'mountain' *él atra aquellos montes y llanuras* 1189
1504 brillante *adj* 'brilliant' *ha sido uno de los alumnos más brillantes* 1196 w	1513 tribunal *n* 'court' *en jurisprudencia del Tribunal Supr* 1188 E
1505 procedimiento *n* 'procedure' *no aguantaba los procedimientos judiciales* 1195 w	1514 desconocer 'to be unawar *se desconoce qué es lo que hay cajón* 1187
1506 rama *n* 'branch' *saltaba desde la rama de un árbol* 1195	1515 mensaje *n* 'message' *mensaje en la contestadora* 118(
1507 comprobar *v* 'to prove, check' *comprobó que su pistola estaba sin seguro* 1195 w	1516 moneda *n* 'coin, currenc *propia moneda se convierte capital* 1185
1508 contribuir *v* 'to contribute' *una habilidad que contribuye mucho al regocijo* 1194	1517 relato *n* 'story, report *narrador lo dirige a través del r* 1184 L

También habrá un índice alfabético, como el siguiente:

Tabla 37. El diccionario de frecuencias: índice alfabético

labio *n* lip 1081	lástima *n* pity, shame 2574	legal *adj* legal 149
labor *n* work 1404	lateral *adj* side, lateral 3723	legislativo *adj* legislative 2801
laboratorio *n* laboratory 1751	latín *n/adj* latin 2915	lejano *adj* distant, off 1533
lado *n* side 221		

lago *n* lake 1715	lavar *v* to wash 2527	lejos *adv* far (away) 624
lágrima *n* tear(drop) 954	lazo *n* tie, bond 3412	lengua *n* tongue, language 486
laguna *n* lagoon, gap, lapse 3155	le *pron* to him/her (IO) 27	lenguaje *n* language 1125
lamentable *adj* regrettable 4191	leal *adj* loyal 4602	lentamente *adv* slowly 2045
lamer *v* to lick 5954	lealtad *n* loyalty 4325	lente *n* lens 4641
lámpara 3887	lección *n* lesson 3026	letra *n* letter 974
lana *n* wool 3551	leche *n* milk 706	levantar *v* to raise, lift 408
lanzar *v* to throw, launch 1229	lecho *n* (river) bed 2596	leve *adj* slight, light 2890
lápiz *n* pencil 4829	lectura *n* reading (material) 1449	ley *n* law 121
largo *adj* long 185	leer *v* to read 394	
lector *n* reader 1756	lento *adj* slow 1539	
	leña *n* (fire)wood 4670	
	león *n* lion 1624	

Por fin, habrá un índice por categoría gramatical. Esto ayudará a los que quieren fijarse en los sustantivos, los verbos, u otra categoría gramatical:

Tabla 38. El diccionario de frecuencias: índice por categoría gramatical

Adjetivo		Sustantivo		Verbo	
.....		.....		.....	
3659	apasionado	3626	interrupción	3312	ahogar
3662	delicioso	3630	evaluación	3313	fingir
3679	cerebral	3631	serpiente	3317	registrar
3684	templado	3632	capricho	3324	suspender
3686	marítimo	3633	desaparición	3343	dictar
3695	repentino	3634	furia	3349	anticipar
3714	decorativo	3636	revisión	3355	burlar
3719	ardiente	3641	adorno	3362	expulsar
		3644	carreta	3363	distribuir
Adverbio		3645	cohete	3388	resaltar
.....				3389	retroceder
3857	ligeramente	Preposición		3398	sumar
3881	suavemente	.....		3399	abarcar
3930	francamente	15	por	3400	fluir
4007	enteramente	17	con	3403	instar
4071	continuamente	21	al	3414	ajustar
4162	puramente	49	sin	3418	tropezar
4342	frecuentemente	53	sobre	3431	retrasar

6. El uso del Corpus del Español en la investigación de los ca históricos

6.0 Como se puede observar fácilmente, el Corpus del Español no solamente para estudiar la variación actual, sino también los ca históricos. Tiene más de 80.000.000 palabras de los siglos XIII-XIX puede enfocar en cualquiera de estos siglos. También, se muestra clara la cronología exacta de los cambios – en qué siglo surgió cierta fo construcción o cuándo desapareció otro fenómeno. En esta sección, pı consideramos otros corpus y recursos que he creado para el estudio cambios históricos, y cómo se han usado éstos para enseñar la "Historia Lengua Española". Después, consideramos cómo se puede usar el C del Español para estudiar los cambios fonéticos, ortográficos, morfoló y léxicos. Al final, estudiamos cómo se pueden llevar a cabo estudio detallados sobre los cambios sintácticos, léxicos, y semánticos.

6.1 Otros corpus y el curso de la "Historia de la Lengua Española"

He tenido oportunidad de impartir un curso de "Historia de la L Española" por medio del Internet, y este curso se ha basado casi por en los corpus que he creado. Esto resulta muy bien, porque por lo ger los estudiantes les gusta encontrar evidencia de los cambios lingüístic sí mismos, en vez de sólo leer explicaciones en los libros de texto. Co puede ver en el sitio de web del curso (http://d linguistics.byu.edu/hisspan/) hay varios temas en el curso:

Tabla 39. Temas en el curso de "Historia de la Lengua Española"

TEMA	Preguntas	Proye
LAS ETAPAS MÁS TEMPRANAS		
1. Introducción	O	
2. Las lenguas prerromanas	O	
3. El indoeuropeo	O	
4. El latín: externo	O	
5. El latín: interno	O	O
6. El latín vulgar y las lenguas romances	O	
7. Los visigodos	O	
8. Los árabes	O	

LATÍN > EL ESPAÑOL ANTIGUO		
9. Fonética		O
10. Morfosintaxis		O
11. Léxico		O
EL ESPAÑOL ANTIGUO		
12. Los dialectos hispánicos antiguos	O	O
13. Los textos medievales	O	O
14. La lengua c1250-1450	O	
EL ESPAÑOL ANTIGUO > ESPAÑOL MODERNO		
15. Fonética		O
16. Ortografía		O
17. Morfología		O
18. Sintaxis		O
19. Léxico		O
EL ESPAÑOL MODERNO		
20. La lengua en 1475-1700	O	
21. El español de las Américas	O	
22. Otros dialectos modernos	O	
23. El futuro del español	O	O

Casi todos los temas que tienen un [Proyecto] se basan en varios corpus que he creado. El primero es una "Biblia Políglota", que contiene casi toda la Biblia en el latín vulgar, el español antiguo, y el español moderno (http://davies-linguistics.byu.edu/span3/). El siguiente cuadro presenta un ejemplo de cuatro versículos del corpus:

Tabla 40. El corpus paralelo del latín y el español antiguo y moderno

C:V	LATIN VULGAR	ESPAÑOL ANTIGUO	ESPAÑOL MODERNO
10:30	suscipiens autem Iesus dixit homo quidam descendebat ab Hierusalem in Hiericho et incidit in latrones qui etiam despoliaverunt eum et plagis inpositis 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 Je dijo: --Cierto ho descendía de Jerusalén a Jeric cayó en manos d ladrones, quiene despojaron de su ropa, le hirieron fueron, dejándol medio muerto.
10:31	accidit autem ut sacerdos quidam descenderet eadem via et viso illo praeterivit	Acaecio que aquel mismo dia un sacerdot passaua por aquella misma carrera, e quandol uio, passos e fue su uia.	Por casualidad, descendía cierto sacerdote por aq camino; y al verl pasó de largo.
10:32	similiter et Levita cum esset secus locum et videret eum pertransiit	E otrosi un leuita que passo cab el, quandol uio, fuesse adelant.	De igual manera levita también ll al lugar; y al ir y verle, pasó de la
10:33	Samaritanus autem quidam iter faciens venit secus eum et videns eum misericordia motus est	E un samaritano que passaua por alli, quandol uio, fue mouido de piedat;	Pero cierto samaritano, que de viaje, llegó ce de él; y al verle, movido a misericordia.

Lo bueno del corpus es que se puede estudiar la evolución de palabra o una construcción gramatical durante los últimos 1500 años, pc el mismo versículo en tres etapas de la lengua – una al lado de la otra. ejemplo, el cuadro siguiente es un ejemplo de la construcción futura en tardío (de hace 1500 años), el español antiguo (hace 750 años) y el es moderno. Se ve claramente el cambio del futuro sintético (una sola pal en el español antiguo al futuro analítico (dos palabras o más) en el es moderno:

Tabla 41. El uso del corpus paralelo para comparar el uso del futuro

VERS	LATIN VULGAR	ESPAÑOL ANTIGUO	ESPAÑOL MODERNO
Apoc 2:10	nihil horum timeas quae passurus es ecce missurus est diabolus ex vobis in carcerem ...	Non temas ninguna destas cosas por que as de passar. Euas que el diablo metra de uos en carcel . . .	No tengas ningún temor de las cosas que has de padecer. He aquí, el diablo va a echar a algunos de vosotros en la cárcel. . .
1 Sam 10:27	filii vero Belial dixerunt num salvare nos poterit iste et despexerunt eum et non adtulerunt ei munera ille vero dissimulabat se audire	Mas los fijos de belial dixieron Como nos podra deffender: Desdennaron lo & non le trayeron dones et eill fazie semblant que no lo oye	Pero unos perversos dijeron: "¿Cómo nos va a librar éste?" Ellos le tuvieron en poco y no le llevaron un presente. Pero él calló.

El siguiente cuadro presenta otro ejemplo, esta vez de la colocación de los pronombres átonos. Se ve claramente el cambio de la colocación media en el latín tardío y el español antiguo a la colocación al final del verbo no conjugado en el español moderno:

Tabla 42. El uso del corpus paralelo para comparar el uso del los clíticos

LATIN VULGAR	ESPAÑOL ANTIGUO	ESPAÑOL MODERNO	LATIN VULGAR
Rom 16:25	ei autem qui potens est vos confirmare iuxta evangelium meum et praedicationem Iesu Christi secundum revelationem mysterii temporibus aeternis taciti	Aquel aya onra e gloria que a poder de uos afirmar en el mio euangelio y en la preigacion de Ihesu Christo, segund el descobrimiento de la incarnacion que siempre fue encubierta,	Y al que puede haceros firmes -- según mi evangelio y la predicación de Jesucristo; y según la revelación del misterio que se ha mantenido oculto desde tiempos eternos,

6.2 Los cambio fonéticos y ortográficos

Como se ha visto en la sección anterior, a veces es útil ver el texto en varios periodos históricos para estudiar un cambio histórico. quizás es más útil ver la misma palabra o construcción en una serie de (p.ej. el siglo XIII al XX) y ver exactamente cuándo ocurrió cierto c En esta sección, veremos algunos ejemplos concretos de esto, con resp los cambios fonéticos y ortográficos.

Primero, se debe recordar que se pueden usar los comodines consultas, y esto ayuda mucho en el estudio de los cambios fonético ejemplo, hubo bastante variación con las vocales átonas en el e medieval y hasta el español renacentista, por ejemplo *sofrir / sufri* puede hacer la consulta siguiente en el Corpus del Español, y se siguiente, lo cual indica que el mayor cambio ocurrió en los años 1500s.

[PALABRA/FRASE] s_fr*

Tabla 43. El cambio fonético [o/u]

palabra	12	13	14	15	16	17	18
sofrir	529	335	514	125	4	8	4
sofrido	62	22	32	3	3	4	0
sofria	5	16	17	0	0	0	0
sufre	109	51	279	602	296	167	506
sufra	27	28	28	126	126	40	100
sufrir	74	23	137	1410	792	531	1234
sufren	22	22	59	149	58	80	277

Otro ejemplo es el cambio de [-AUA] a [-AVA] a [-ABA]. Se hacer la consulta siguiente en el Corpus del Español, y se verá lo sig lo cual indica que [-AUA] era lo común en los años 1200s-1400s, AVA] era común en los años 1400s-1600s, y que [-ABA] ha sido l común desde los años 1500s.

[PALABRA/FRASE] *aua*, *ava*, *aba*

Tabla 44. El cambio fonético [-v-]

palabra	12	13	14	15	16	17	18
cantauan	48	11	43	4	2	0	0
cantauas	0	0	1	0	0	0	0
cantaua	24	9	37	5	5	0	0
cantavas	0	0	1	2	1	0	0

cantavan	0	0	15	13	6	2	0	0
cantava	3	2	21	33	9	1	0	0
cantaba	2	0	0	143	100	48	334	275
cantaban	7	0	1	126	55	28	175	114
cantabas	0	0	0	4	1	2	1	7

6.3 Los cambios morfológicos

En el curso de la "Historia de la Lengua Española", los alumnos estudian los temas siguientes, que están relacionados con los cambios morfológicos. Después, llevan a cabo sus propios estudios sobre tres o cuatro de estos temas, basándose en los datos del Corpus del Español.

Tabla 45. Los cambios morfológicos en "La historia de la Lengua Española"

Sustantivos	Determinantes / pronombres
1. Género: el cuchar > la cuchara	1. vos(uos) / os
2. la + -o/u: la/el tribu, el/la mar	2. DET + POS: la tu / tu
3. –íssimo: común ya en siglo XV	3. nosotros / vosotros
	4. los/les
Verbos	5. mio/mi, sos/sus, etc
1. -zco (verbos) (pareço)	6. alguien:quien, nadie:otrie, etc.
2. participios pasados (conesçudo)	7. gelo / se lo
3. pretéritos irregulares (tove)	
4. imperfecto en -ié / ía (salié)	
5. futuro irregular (combré)	

Vamos a ver tres ejemplos concretos: la pérdida del futuro del subjuntivo, el crecimiento del uso del superlativo "latino" en –ÍSIMO, y la pérdida de varias conjugaciones verbales con [*decir*]. El cuadro siguiente presenta un resumen de estas consultas:

Tabla 46. Los cambios morfológicos

palabra	límites	ejemplos	explicación
*iere	+1200s +1300s –1900s	*fiziere,* *naciere,* *tosiere*	futuro del subjuntivo
*simo	+1500s +1400s -1300s	*santísimo,* *altísimo,* *grandisimo*	superlativo de [-ísimo] (que comenzó h. el siglo XV)

decir.*	+1200s –1500s –1900s	*dize, dixiere, decir*	formas verbales de [decir], y perdidas

Los tres cuadros siguientes presentan más detalles. El primer c presenta los datos de la primera consulta de la Tabla 46, e indica q futuro del subjuntivo se pierde en gran parte alrededor de los años 1500s

Tabla 47. El cambio morfológico: el futuro del subjuntivo

palabra	12	13	14	15	16	17	18
ouiere	2361	374	489	7	0	0	0
fiziere	1221	469	508	21	0	0	0
tosiere	575	93	261	1	0	0	0
dixiere	537	156	89	4	0	1	0
viniere	244	25	1	3	0	0	0
saliere	226	34	160	93	35	34	14
entendiere	185	111	63	63	13	14	3
acaeciere	176	141	172	0	0	0	0
podiere	167	66	106	0	0	7	1

El segundo cuadro presenta los datos del uso del superlativo, y se en la segunda consulta de la Tabla 46. Indica que el uso del superl "latino" en [-SIMO] aumenta mucho en los años 1400s-1600s. Nótes hay muchas otras formas en –ÍSIMO (en vez del –ÍSSIMO, común e años 1400s-1500s) que han seguido hasta el español moderno:

Tabla 48. El cambio morfológico: el superlativo

palabra	12	13	14	15	16	17	18
serenissimo			68	5	2	4	
grandissimo	1		36	11	1	4	2
dulcissimo			25	1			
altissimo	2		16	2	2	6	2
santissimo			15	10	2	6	
excellentissimo			15	1	2	2	
grandíssimo			14	156	5	9	
tricesimo	1		13	1			2
altisimo			13	2		1	

grandisimo	12	1		8
clarissimo	12	2	1	1
potentissimo	68	5	2	4

Por último, se puede usar el corpus para ver qué formas de un verbo (o sustantivo) se han perdido o han surgido durante la historia del español. Por ejemplo, el cuadro siguiente presenta los datos de la última consulta de la Tabla 46, e indica qué formas de [*decir*] ocurrían hasta aproximadamente 1500, pero perdidas desde entonces:

Tabla 49. El cambio morfológico: las formas de [decir]

palabra	12	13	14	15	16	17	18	19
dixiemos	2214	115	316	0	0	2	0	0
dixol	1216	180	5	0	0	1	0	0
dizien	1128	77	14	0	0	0	0	0
dixiese	324	137	112	0	0	0	0	0
dezjr	284	487	739	0	0	0	0	0
dizian	212	95	551	0	0	0	0	0
dixieren	167	40	28	0	0	0	0	0
dizia	160	150	337	0	0	0	0	0
dixieronle	86	50	34	0	0	0	0	0
dixieran	80	7	7	0	0	0	0	0
dixeronle	73	55	59	0	0	0	0	0
dixiestes	63	18	1	0	0	1	0	0
dixemos	55	10	12	0	0	0	0	0

6.4 Los cambios léxicos

El Corpus del Español nos permite indicar cuántas veces ocurre una palabra en los diferentes siglos. Esto nos permite hacer unas consultas muy fáciles y rápidas para ver qué palabras han entrado en la lengua o se han perdido de la lengua entre un par de siglos. Por ejemplo, el siguiente cuadro indica qué palabras (en varias categorías gramaticales) han entrado en el español durante los últimos 100-200 años. Vamos a hacer mucho más con esto durante el taller.

Tabla 50. Palabras nuevas: 1800s>1900s

*.v_inf -1700s -1800s +1900s	*detectar, liberar, programar, conectar, potenciar, intercambiar, presionar*	verbos nuevos en lengua
*.n -1700s -1800s +1900s	*televisión, sector, proteína, fútbol, foto, encuesta, aeropuerto, grabación*	sustantivos nuev lengua
*.adj -1700s -1800s +1900s	*estadounidense, estatal, nuclear, espacial, global, básico, electrónico, genético*	adjetivos nuevos lengua

6.5 Los cambios sintácticos

En el curso de la "Historia de la lengua española", usamos los para encontrar datos sobre muchos cambios sintácticos, como los sigu

Tabla 51. Los cambio sintácticos en "La Historia de la Lengua Espa

Pronombres
1. colocación: quiero lo fazer
2. futuro mesoclítico: cantar lo (h)an
3. OD/OI "redundante": e dixo a el las cosas
4. se impersonal se: si se habla de
5. omne = se: sy omne tjene …
6. gelo / se lo: quando ge lo vendimos
7. vos / tú / usted: vos tenéis / tu tienes / Ud. tiene

El significado de las formas verbales
1. ser / estar: el era en esse lugar
2. haber / tener: yo he muchas cosas
3. haber / ser + PP: que son venidos a España
4. haber / hacer: ha mucho que salimos
5. infinitivo: cf. los artículos por Davies

(levantamiento de sujeto/objeto, causativos, etc)

Ya daremos unos ejemplos muy breves de cómo se pueden llevar unas consultas rápidas y fáciles sobre tales cambios sintácticos fenómenos que se presentan aquí son del tipo que he estudiado en los 10-15 años, y se relacionan (en gran parte) con cambios históricos construcciones del infinitivo en español. Se debe recordar que ning estas consultas sería posible con otros corpus como el CORDE, que n

etiqueteados para la categoría gramatical. En único corpus que puede estudiar tales construcciones gramaticales es el Corpus del Español.

El primer ejemplo es del infinitivo con sujeto explícito (cf. Davies 2003). Ocurrían bastante en los años 1400-1600s, después hay pocos casos en los años 1600s-1900s, pero ya han surgido otra vez, más que nada en el habla y en la zona caribeña. La consulta siguiente sacaría los datos del Corpus del Español y el cuadro siguiente nos presenta estos datos:

[palabra/frase] *.prep *.pn_subj *.v_inf

Tabla 52. Construcción: el infinitivo con sujeto explícito

frase	12	13	14	15	16	17	18	19	19L	19O	19M
de yo haber	0	0	0	0	0	0	2	3	0	3	0
para él ser	0	0	0	1	0	0	1	3	0	3	0
de él haber	0	0	0	1	1	0	0	2	0	2	0
para él tener	0	0	0	1	0	0	1	2	1	1	0
para yo tener	0	0	0	0	0	0	0	2	0	1	1
para nosotros lograr	0	0	0	0	0	0	0	2	0	2	0
para nosotros tener	0	0	0	0	0	0	0	2	0	2	0

Otro ejemplo del cambio sintáctico ocurre con los sujetos "experimentadores" con el levantamiento del sujeto, que se ha estudiado en Davies (1997a, 1997b). Sólo han sido comunes desde aproximadamente los 1500s. La consulta siguiente sacaría los datos del Corpus del Español y el cuadro siguiente nos presenta estos datos:

[palabra/frase] le/les parecer.* *.v_inf

Tabla 53. Construcción: el levantamiento de sujeto

frase	12	13	14	15	16	17	18	19	19L	19°	19M
le pareció ser	0	0	0	16	10	4	0	1	1	0	0
le pareciese convener	0	0	0	2	3	0	0	0	0	0	0
le pareció estar	0	0	0	5	3	1	0	1	1	0	0
le pareció haber	0	0	0	4	3	2	2	1	1	0	0
le parecía ser	0	0	0	11	3	2	5	2	2	0	0
les pareció ser	0	0	0	9	3	1	0	0	0	0	0
le pareció avisar	0	0	0	0	2	0	0	0	0	0	0
le pareció aguardar	0	0	0	0	2	0	0	0	0	0	0

Utilizando los sinónimos, se pueden llevar a cabo consultas complejas. Por ejemplo, la consulta siguiente encuentra lo: decausativos (p. ej. *se hundió el barco, se rompieron los vasos*) e verbo es un sinónimo de [*romper*] (cf. Davies, 2004a). El cuadro ir ha habido casos desde el español medieval, pero que han au paulatinamente desde entonces:

[palabra/frase] se !romper.*

Tabla 54. Construcción: los verbos decausativos

SE +	12	13	14	15	16	17	18	19	19L	1
romper	5	5	63	96	103	114	241	295	123	1
quebrar	2	9	35	102	65	26	52	65	47	
violar	0	0	0	1	0	9	8	23	2	
destrozar	0	0	0	0	0	5	14	17	12	
despedazar	0	0	0	11	7	7	28	13	10	
dividir	0	0	0	0	3	1	3	7	2	
rajar	0	1	0	1	0	3	3	13	7	

Otro ejemplo del uso de los sinónimos y la categoría gramatic la construcción llamada "subida del objeto" (cf. Davies 2002). E siguiente indica que otra vez ha habido un aumento paulatino c 1500s-1600s:

[palabra/frase] !difícil.* de *.v_inf [límites] +1800s

Tabla 55. Construcción: levantamiento del objeto

	12	13	14	15	16	17	18	19	19L	1
duro de pelar	0	0	0	0	0	0	14	1	0	
difícil de explicar	0	0	0	0	0	2	13	26	14	9
imposible de describir	0	0	0	0	0	0	12	3	2	
difícil de hacer	0	0	0	3	0	2	10	3	1	
difícil de comprender	0	0	0	1	2	4	9	8	3	2
difícil de vencer	0	0	0	1	2	2	9	5	3	
imposible de realizar	0	0	0	0	0	0	9	0	0	

Aparte de las bases de datos que contienen información categoría gramatical y los sinónimos, los usuarios mismos pueden propias listas de palabras, almacenarlas, y después usarlas como par consulta. Por ejemplo, supongamos que alguien quiere estudiar la e de las construcciones "causativas", que ocurren con los verbos cau los verbos de percepción (*me hizo pensar, la dejó salir, nos viero*

(cf. Davies 1995c, 1996a, 1996b). En este punto del estudio, el investigador quiere enfocarse en sólo los verbos de percepción. Primero, crea una lista con varios verbos de percepción (*oír, ver, sentir, mirar*, etc). Después, el usuario (supongamos que se llama [lópez]) puede usar esta lista directamente como parte de la consulta, por ejemplo:

[palabra/frase] le/les [lopez:percepción].* *.v_inf

Veremos que otra vez ha habido ocurrencias desde el español medieval, pero que ha crecido el uso poco a poco hasta el español moderno:

Tabla 55. Construcción: los verbos de percepción + infinitivo

mayúsculas = todas las formas del verbo	12	13	14	15	16	17	18	19	19L	19º	19M
le OÍR decir	0	0	0	20	17	6	32	23	17	6	0
le VER entrar	0	1	0	4	8	5	26	1	1	0	0
le OÍR hablar	1	0	0	1	2	2	20	3	2	1	0
le VER llegar	0	0	0	3	1	2	17	5	5	0	0
le VER pasar	0	0	0	2	2	0	13	1	1	0	0
le VER salir	3	0	1	6	5	0	12	0	0	0	0
le VER venire	2	1	3	5	3	1	11	0	0	0	0

Vamos a ver un ejemplo más de cómo se pueden usar las listas que han sido creadas por los usuarios. Supongamos que el usuario [garcía] quiere estudiar los decausativos (hundir, perder, romper, cerrar, mover, etc). Primero, crea tal lista (llamada [decausativo]), y después la puede usar para ver todos los casos de la construcción con [se] y objeto indirecto con cualquier forma de los verbos en esta lista:

[palabra/frase] se le/les [garcía:decausativo].*

Tabla 56. Construcción: objeto indirecto + verbo decausativo

	12	13	14	15	16	17	18	19	19L	19º	19M
se le perder	1	0	8	49	26	8	17	34	27	7	0
se me perder	0	0	1	14	13	0	4	28	19	9	0
se le abrir	1	0	0	17	8	5	16	22	15	3	4
se le mover	1	1	1	4	0	2	4	20	19	1	0
se le enredar	0	0	0	0	0	1	10	19	12	6	1
se le hundir	0	0	0	11	0	1	9	16	14	2	0
se le romper	0	0	1	11	4	4	17	16	8	7	1
se me romper	0	0	0	3	0	1	5	16	9	6	1
se me cerrar	0	0	0	4	0	1	5	14	10	4	0
se le cerrar	0	0	2	20	5	3	31	11	4	6	1

7. Los cambios semánticos

7.0 Mientras que tradicionalmente ha sido difícil a veces estu cambios sintácticos, por lo general ha sido aun más difícil estu cambios semánticos. ¿Cómo se puede saber si una palabra ha cam significado?

Quizás la manera más fácil de discernir el cambio de signif estudiando qué palabras co-ocurren con esa palabra. Se ha dich puede saber mucho de una palabra por medio de las otras palabra que ocurre. Para extender un poco la analogía, en la vida humana decir que alguien ha cambiado de personalidad si comienza a pas tiempo con un nuevo tipo de amigos. Es lo mismo con las Entonces la mejor manera de ver si el significado de una pa cambiado es comparando las colocaciones de esa palabra en un per las de otro periodo.

Afortunadamente, el uso de nos n-grams en el Corpus del permite el análisis fácil y rápido de las colocaciones, para ver qué ocurren más con otras. Además, el Corpus del Español permite bases de datos que contienen sinónimos y las tablas que los usuario crean. Todo esto hace que el usuario pueda hacer muchas cons serían completamente imposibles con los corpus como CREA o con las que sólo se pueden buscar frases exactas.

7.1 El uso de los n-grams en el estudio de los cambios semántico

Un ejemplo del uso de los n-grams para estudiar el signific siguiente. En esta tabla, vemos qué sustantivos ocurren más adjetivos *suave* y *duro* en el español moderno. Los sustantivos a la ocurren mucho más con *suave*, y las que están a la derecha ocurrer *duro*:

Tabla 57. Comparar la frecuencia de dos palabras opuestas

ADJ	+suave	-duro	ADJ	+duro
voz	22	2	cara	10
viento	10	0	pan	10
músico	6	0	tiempo	9
sabor	5	0	vida	8
invierno	5	1	palabra	8
brisa	5	1	hombre	7
tono	4	1	ojo	6

Ahora podemos extender este tipo de análisis a los periodos históricos. La siguiente consulta producirá los datos sigui muestran las colocaciones más comunes con *duro* en los 1500s y lo

[palabra/frase] *.n duro.* [límites] +1500s -1900s / -1500s +1900s

Tabla 58. Comparar la frecuencia a través del tiempo

ADJ	+1500s	-1900s	ADJ	+1900s	-1500s
suerte	17	0	línea	13	0
hierro	11	0	disco	12	0
muerte	10	0	roca	12	1
pena	10	0	cuello	11	0
bronce	6	0	cara	10	0
cuero	6	1	pan	10	2

Otro ejemplo aun más claro de los cambios semánticos son los sustantivos que ocurren con *hacer* en los 1200s y los 1900s. Se puede notar la cantidad de frases en los años 1200s que se refieren al honor, las relaciones humanas, etc.

[palabra/frase] hacer.* *.n [límites] 1200s> 20 1900s<3

Tabla 59. Frases con [hacer]: +1200s / -1900s

HACER +	1200s	1900s	SUSTANTIVO	1200s	1900s
emienda	207	0	limosna	40	1
adulterio	121	0	honra	34	0
duelo	77	0	postura	31	0
pleito	76	1	altar	25	0
engaño	65	1	adiós	23	2

Si lo hacemos al revés, podemos ver qué sustantivos ocurren más con *hacer* en los años 1900s que en los años 1200s. Se nota que hay más frases que se refieren a conceptos abstractos:

[palabra/frase] hacer.* *.n [límites] 1900s> 20 1200s<3

Tabla 60. Frases con [hacer] : -1200s / +1900s

HACER +	1200s	1900s	SUSTANTIVO	1200s	1900s
falta	0	880	frío	6	115
tiempo	5	456	cargo	0	99
caso	0	318	calor	0	84
referencia	0	206	uso	1	80
rato	0	177	política	0	64

7.2 El uso del los sinónimos

Aparte del uso de los n-grams, también se puede utilizar el hec el Corpus del Español nos permite buscar los sinónimos de un: Podemos combinar esto con la habilidad de limitar los resultad frecuencia en diferentes periodos históricos. Por ejemplo, la siguiente busca los sinónimos de hablar que ocurren poco en los a pero sí ocurren con frecuencia en los años 1900s. Es decir, éstos s que han entrado en la lengua en los últimos 200-300 años, y que : al concepto de "*hablar*":

[palabra/frase] !hablar.* [límites] 1900s>10 170

Tabla 61. Los sinónimos de hablar, 1700s-1900s

VERB	12	13	14	15	16	17	18
dialogar	0	0	2	1	0	0	10
susurrar	0	0	2	0	2	2	22
enunciar	0	0	0	0	0	2	25
cuchichear	0	0	0	0	0	0	4
musitar	0	0	0	0	0	2	3
balbucir	0	0	0	0	0	0	14
disertar	0	0	0	0	1	3	21

Otro ejemplo un poco más complejo del uso de los sinón siguiente. En este caso, buscamos los sinónimos de *tener* y de para encontrar frases relacionadas a *tener problemas*. Además, lim ocurrencias a las que han subido en uso en los últimos 200 años:

[palabra/frase] !tener.* !problema.* [límites] 1900s>10 1:

Tabla 61. Las frases relacionadas con [tener problema], 1700s-19(

	17	18	19	19L	19°
tener problema	1	0	296	61	212
haber problema	0	4	270	51	203
tener dificultad	5	0	43	17	15
tener duda	0	3	43	27	13
haber duda	0	2	17	7	4
haber dificultad	0	2	4	1	2

7.3 El uso de las especializadas (creadas por el usuario)

Una de las ventajas más importantes del Corpus del Españ que no es posible con ningún otro corpus grande – es la de poder especializadas del vocabulario de cierto campo semántico, y de po más tarde en otras consultas. Por ejemplo, supongamos que alg

estudiar el uso y la distribución de palabras relacionadas con las emociones negativas (*enojarse, deprimirse, fastidiarse*, etc) o ciertas frases idiomáticas relacionadas con la ropa (*zapatos, sombrero, pantalones*, etc). En estos casos, el usuario crea una lista de todas las palabras deseadas de ese campo semántico. Después, puede usar esta lista directamente como parte de otras consultas.

Por ejemplo, vamos a suponer que el usario [lópez] quiere estudiar las expresiones relacionadas al concepto de [hacerse daño al cuerpo]. Primero, puede crear una lista de las partes del cuerpo, llamada [cuerpo]. Después, usa esta lista en la consulta siguiente, para ver qué palabras (que se refieren a las partes del cuerpo) ocurren con una forma de *romper*, y ve lo siguiente:

[palabra/frase] romper.* el/la/los/las [lópez:cuerpo].*

Tabla 62. Un campo semántico: ROMPER + las partes del cuerpo

ROMPER +	12	13	14	15	16	17	18	19	19L	19O	19M
cabeza	0	0	0	4	15	24	39	21	15	6	0
pierna	0	0	0	0	1	2	2	8	6	2	0
boca	0	0	0	1	1	0	0	6	5	1	0
nariz	0	0	0	0	0	2	3	6	4	2	0
dedo	0	0	0	0	0	0	0	4	3	1	0
ojo	0	0	0	0	0	0	0	3	1	2	0
pie	0	0	1	0	0	1	1	2	1	1	0

Otro ejemplo del poder de las listas creadas por el usuario es el siguiente. Primero, el usuario busca los sustantivos que ocurren con *morir de*.

[palabra/frase] morir.* de *.n [límites] +1900s -1800s, o
[límites] +1800s -1900s

Tabla 63. Un campo semántico: las palabras negativas con [morir de]

MORIR DE	+18	-19	MORIR DE	+19	-18
pena	35	7	ganas	22	0
dolor	20	5	calor	9	1
vergüenza	24	13	cáncer	9	0
amor	35	14	miedo	23	11
tristeza	13	4	risa	30	19
celos	8	3	pulmonía	5	0
pesadumbre	7	0	aburrimiento	5	1

Nótese que en los 1800s se refiere más al amor trágico (*morir vergüenza, amor, tristeza, celos, pesadumbre*), mientras que en el es o más literal (*calor, cáncer, pulmonía*) o metafórico (*ganas, m aburrimiento*). Después de encontrar estas palabras, el usuario puede meter estas palabras "negativas" en una lista nueva. De usar esta lista para ver qué otras palabras o frases ocurren más par conceptos negativos.

7.4 Conclusión

En fin, la única limitación en el uso del Corpus del Esp imaginación. Se pueden buscar fácilmente palabras y frases exa que con CREA y CORDE). Pero también – a diferencia de estos – se puede usar la categoría gramatical, los lemas, la frecuencia periodos históricos o en ciertos registros, los sinónimos, y las list por el usuario mismo. Todo esto permite que el usuario pueda lle un sinfín de investigaciones sobre la variación actual y los cambios – mucho más que con cualquier otro corpus del español.

Obras Citadas

Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. 1995. Dimensions of register variation: A cross-linguistic comparison. Cambridge: Cambridge University Press.

Conrad, S., and D. Biber (eds.). 2001. *Variation in English: Multi-Dimensional studies*. London: Longman.

Davies, Mark. (To appear) "Student use of large, annotated corpora to analyze syntactic variation". In Guy Aston, et al (eds). *Proceedings from the Fifth International Conference on Teaching and Language Corpora*. Rodopi, 2004.

------. (To appear) "Vocabulary Range and Text Coverage: Insights from the Forthcoming *Routledge Frequency Dictionary of Spanish"*. In David Eddington, et al (eds). *Proceedings from the 7th Hispanic Linguistics Symposium*. Cascadilla, 2004.

------. (To appear) "Student use of large corpora to investigate language change". In Thomas Upton, et al (eds). *Applied Corpus Linguistics: A Multidimensional Perspective*. Rodopi, 2004.

------. (To appear) Review of "Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching (Sylvaine Granger, et al). In *Modern Language Journal*, 2004.

------. (To appear) "On diachronic shifts with Spanish *se*: preliminary evidence from large electronic corpora." In Claus Pusch, et al (eds). *Proceedings of the Second Freiburg Workshop on Romance Corpus Linguistics*. Tübingen: Gunter Narr Verlag, 2005.

------. (To appear) "Advanced research on syntactic and semantic change with the Corpus del Español". In Claus Pusch, et al (eds). *Proceedings of the Second Freiburg Workshop on Romance Corpus Linguistics*". Tübingen: Gunter Narr Verlag, 2005.

------. (2003) "Diachronic Shifts and Register Variation with the "Lexical Subject of Infinitive" Construction. (Para yo hacerlo)". In Silvina Montrul and Francisco Ordóñez, *Linguistic Theory and Language Development in Hispanic Languages*. Somerville, MA: Cascadilla Press. 13-29.

------. (2003) "Annotation without lexicons: an alternative to the standard bootstrapping approach". In Paul Rayson, et al. *Proceedings from Corpus Linguistics 2003* (Lancaster, England, March 2003). 174-83.

------. (2003) "Relational n-gram databases as a basis for unlimited annotation on very large corpora". In Kiril Simov, ed. *Proceedings from the Workshop on Shallow Processing of Large Corpora* (Lancaster, England, March 2003). 23-33.

------. (2002) "Un corpus anotado de 100.000.000 palabras del e histórico y moderno". *SEPLN 2002 (Sociedad Española Procesamiento del Lenguaje Natural*). (Valladolid). 21-

------. (2002) "'*Esto es ligero de fazer*: Object to Subject Raisin; Medieval and Early Modern Spanish". In James F. Lee, *Structure, Meaning, and Acquisition of Spanish*. Somerv Cascadilla Press. 19-31.

------. (2001) "Review of *Construcciones causativas en el espai* by Milagros Alfonso Vega". *Revista Canadiense de Estu Hispánicos* 25:329-30.

------. (2001) "Creating and using multi-million word corpora fi based newspapers". In *Corpus Linguistics in North Ame* Rita C. Simpson and John M. Swales. Ann Arbor: U Mic 75.

------. (2000) "Using multi-million word corpora of historical a Spanish texts to teach advanced courses in Spanish lingu *Rethinking Language Pedagogy from a Corpus Perspect* Burnard and Tony McEnery. Frankfurt am Main; New Y Lang. 173-85.

------. (2000) "Syntactic Diffusion in Spanish and Portuguese I Complements". In *New Approaches to Old Problems: Is Romance Historical Linguistics*, eds.Steven Dworkin an Wanner. Amsterdam; Philadelphia: John Benjamins. 10

------. (1999) "The Historical Development of Subject Raising Portuguese: A Corpus-Based Approach". *Neuphilologisc Mitteilungen* 100:95-110.

------. (1999) "A Computer Corpus-Based Study of Subject Ra Modern Portuguese". *Lingvisticae Investigationes* 21:37

------. (1998) "The Evolution of Spanish Clitic Climbing: A Co Approach." *Studia Neophilologica* 69:251-63.

------. (1997) "A Corpus-Based Approach to Diachronic Clitic Portuguese." *Hispanic Journal* 17: 93-111.

------. (1997) "Using Large Computer-Based Corpora as a Philo An Analysis of Four Medieval Spanish Bibles." *Dactylu*

------. (1997) "The History of Subject Raising in Spanish". *Bul Hispanic Studies* (Liverpool) 74: 399-411.

------. (1997) "A Corpus-Based Analysis of Subject Raising in Spanish." *Hispanic Linguistics* 9: 33-63.

------. (1996a) "The Diachronic Interplay of Finite and Nonfini Complements in Spanish and Portuguese." *Bulletin of Hi Studies* (Glasgow) 73:137-58.

------. (1996b) "The Diachronic Evolution of the Causative Construction in Portuguese." *Journal of Hispanic Philology* 17:261-92.
------. (1995a) "The Evolution of Causative Constructions in Spanish and Portuguese." In *Current Research in Romance Linguistics*, ed. John Amastae, et al. Philadelphia: John Benjamins, 1995. 105-122.
------. (1995b) "Omnipage and WordCruncher: Tools for Creating and Searching Digitalized Text Corpora." *La Corónica* 23:111-115.
------. (1995c) "The Evolution of the Spanish Causative Construction." *Hispanic Review* 63:57-77.
------. (1995d) "Analyzing Syntactic Variation with Computer-Based Corpora: The Case of Modern Spanish Clitic Climbing". *Hispania* 78:370-380.
------. (1994) "Parameters, Passives, and Parsing: Explaining Diachronic Shifts in Spanish and Portuguese". In *Variation and Linguistic Theory*, ed. K. Beals, et al. Chicago: CLS. Vol 2. 46-60.
------. (1992) "A Tentative Bibliography of Historical Spanish Syntax." *Hispanic Linguistics* 5:279- 351.

Algunas de las actividades a realizar durante el seminario

1. La frecuencia y distribución (por registro) de cierto(a):
   - prefijo — des-, en-, re-
   - raíz — -cans-, -valor-
   - sufijo — -dor, -mente, -ción

2. Las colocaciones
   - cualquier palabra — tan * como
   - ADJ con cierto N — mujer.* *.adj
   - N con ciertos ADJ — *.n sucio.*
   - N con ciertos V — borrar.* el/la *.n

3. Distribución léxica
   - N más común en textos — *.n — 19misc>10 19oral
   - V más común en oral — *.v_pres — 19oral>10 19misc
   - ADJ más común en lit — *.adj — 19lit>10 19oral<3

4. Cambios fonéticos y morfológicos
   - canta_a — cantaua, cantava, cantaba
   - s_fr* — sofren, sufría
   - formas verbales antiguas — haber.* — +1200s -1500s -1
   - futuro del subjuntivo — *iere — +1200s -1500s -1

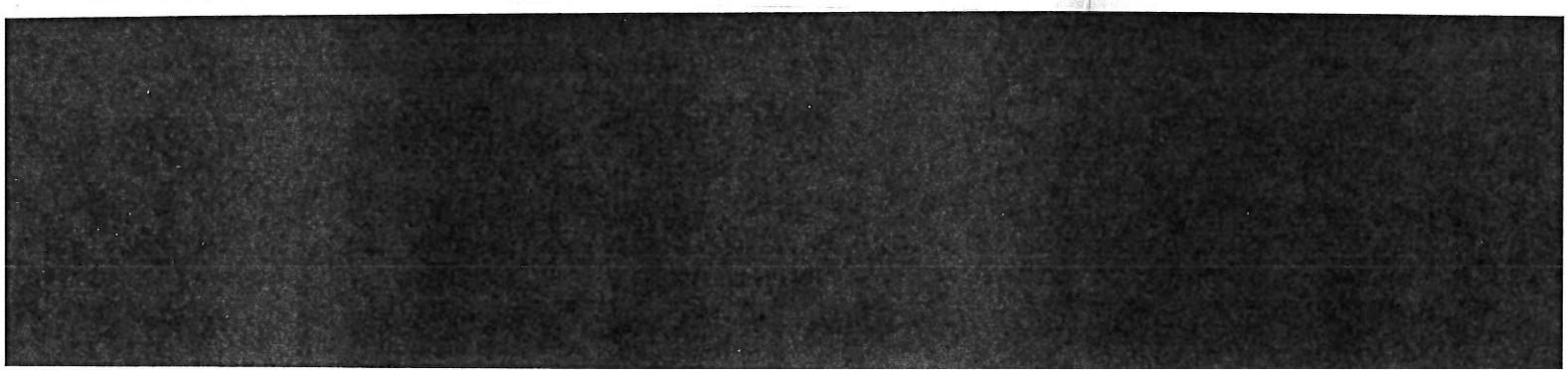
5. Cambios léxicos
   - N nuevos — *.n — +1900s -1800s -1
   - ADJ antiguos — *.adj — +1800s -1900s
   - V nuevos — *.v — +1900s -1800s -1

6. Sinónimos/antónimos
   - sinónimos de [hablar] — !hablar
   - todas las formas — !hablar.*
   - antónimos — #rico
   - historia — !hablar.* — +1900s -1800s -1

スペイン文化シリーズ 11 号
El uso del Corpus del Español y otros corpus en la investigación de la variación actual y los cambios históricos del español

発行日　2004 年 6 月 3 日
発行者　上智大学　イスパニア研究センター
〒102-8554　東京都千代田区紀尾井町 7-1
TEL.03-3238-3533
印　刷　株式会社　ピーコム
Edita: Centro de Estudios Hispánicos
Universidad Sofía
Kioicho 7-1, Chiyoda-ku, Tokio
102-8554
3 de junio de 2004

ISBN4-921193-11-8



7. Construcciones gramaticales
    - lemma lema.*
    - categoría gramatical *.pos
    - alternatives a/b/c/d
    - ejemplo: le/les querer.* *.v_inf

8. Listas personales
    - Hacer clic en "Listas personales"
    - Meter tu nombre y el nombre de la lista
    - Añadir unas palabras
    - Usar la lista en una consulta

        poner.* el/la [davies:ropa]
        estar.* tan [garcía:emociones]