Dorien Nieuwenhuijsen

Notas sobre la aportación del análisis estadístico a la lingüística de corpus

1 Introducción

Hoy en día es un hecho que la lingüística histórica como disciplina académica está cambiada profunda y definitivamente, debido al desarrollo de la lingüística de corpus y las nuevas metodologías ofrecidas y hasta impuestas por los corpus diacrónicos digitales. Mientras que en el pasado se reunían los ejemplos leyendo minuciosamente una serie de textos o fragmentos de textos, actualmente es posible recopilar de manera más o menos automática cantidades considerables de ejemplos.

Una ventaja del método tradicional era que el investigador podía evaluar sobre la marcha la validez o invalidez de un ejemplo, procurando que su corpus de ejemplos fuera homogéneo y no contuviera ejemplos indebidos. Además, la lectura detenida de los textos que le servían de fuente le permitía, ya durante el proceso de la recolección de los datos, hacerse una idea de los contextos específicos sintácticos o pragmáticos en que aparecía la forma o construcción investigada. Al mismo tiempo, el investigador ya podía ir formulando hipótesis sobre los factores involucrados en la selección de la forma o construcción en cuestión. Este método, a la que Kabatek (2014) en un artículo reciente se refiere con el término de «lingüística empática», hacía que el investigador llegara a conocer muy bien su material, condición que, obviamente, es fundamental para un buen análisis lingüístico.[1]

No cabe duda de que hoy en día la mayor ventaja de un corpus digital es la disponibilidad de un banco de datos muy extenso, y la posibilidad de reunir un corpus de ejemplos mucho más amplio, y basado en muchos más textos, que con el método tradicional.[2] Sin embargo, con el estado actual de los corpus diacróni-

1 Kabatek (2014, 707) describe la lingüística empática como «la posibilidad, aun en el caso de la lengua de épocas remotas, de adquirir una cierta competencia lingüística y de desarrollar un ojo crítico que permite identificar fenómenos que eran de algún modo llamativos en la época».

2 Es interesante que Rojo (2012, 435) considere la gran cantidad de ejemplos que se puede reunir con un corpus digital justamente como una desventaja; compárense también las otras ventajas e inconvenientes que menciona Rojo en su trabajo de 2012 (435–436).

Dorien Nieuwenhuijsen: Universiteit Utrecht

cos digitales del español, hay temas lingüísticos que no se dejan estudiar fácilmente, es decir, no con una serie de búsquedas directas. Por ejemplo, en el ámbito de la morfología: el surgimiento del pronombre átono *os* como variante descuidada del átono *vos* (De Jonge/Nieuwenhuijsen 2009, 1629–1635). En el caso de *os*, se trata de una forma que originalmente surgió en posición enclítica, condición que constituye un obstáculo inseparable, dado que las posibles búsquedas o bien proporcionan más ejemplos de los que puede procesar el programa (*CORDE*), o bien únicamente ofrecen ejemplos impropios (*Corpus del español*; de aquí en adelante: *CdE*) (Nieuwenhuijsen 2009, 376–379). Asimismo, por poner un ejemplo en el ámbito de la sintaxis: el desarrollo de las oraciones yuxtapuestas (Nieuwenhuijsen 2013; 2014). Puesto que la yuxtaposición supone la unión de dos oraciones sin conjunción o nexo, el signo que se busca carece de presencia formal, característica que obviamente complica sumamente su análisis en un corpus digital.

Además, el trabajo con un corpus digital entraña el riesgo de que entre los ejemplos reunidos se encuentren casos indebidos, lo cual, obviamente, no solo contamina los datos sino también el análisis y los resultados. Buscando en *CORDE* las formas del imperfecto de subjuntivo en *–ra* y en *–se* de los verbos *ser* e *ir* para comparar su frecuencia relativa, rápidamente se obtienen las formas correspondientes, pero también la forma homónima del adverbio *fuera* y la preposición compuesta *fuera de* (cf. Rojo 2008, 167, nota 9; 2010, 34, nota 16). Una cala en *CORDE* muestra que la palabra *fuera* en el periodo 1900–1950 en España en todos los medios proporciona para los tres textos con más casos por texto, un total de 990 casos, de los que 377 ejemplos son casos del adverbio *fuera* o de la preposición *fuera de*. Esto implica que el 38 % de los casos recogidos no corresponde a la forma verbal.

En el *CdE* se pueden introducir categorías gramaticales, lo que permite, por ejemplo, la búsqueda de construcciones pasivas perifrásticas con el auxiliar *ser* y un complemento agente introducido por la preposición *de* o *por*: [ser] [VPS*] de/por [NP*]/[NN*]. No obstante, este tipo de búsqueda ofrece también casos como *fueron expulsados de España* y *primero sea arrastrado por las calles públicas*.

Mientras que la primera desventaja —la imposibilidad de buscar ciertas formas o construcciones—, que se sepa, por el momento no tiene remedio, la segunda se puede remediar con una revisión «manual» cuidadosa de todos los ejemplos seleccionados automáticamente por el programa. Al mismo tiempo, de esta manera el investigador puede acercarse a su material de estudio, comparable con el proceso por el que pasaba el lingüista tradicional.

Es evidente que una mayor cantidad de datos disponibles incide positivamente en la fiabilidad de los resultados. Sin embargo, más allá de la presentación de números absolutos y porcentajes, no cabe duda de que la fiabilidad de los

resultados puede aumentarse con un tratamiento estadístico de los datos, procedimiento todavía no muy común en la lingüística histórica de corpus.[3]

En el presente trabajo nos proponemos demostrar, a través de un caso concreto, que un análisis estadístico puede llevar a conocimientos más profundos sobre el tema lingüístico bajo estudio y que puede matizar conclusiones sacadas a base de los porcentajes calculados sobre los números absolutos de ejemplos. Para tal fin, estudiaremos la variación del modo indicativo y subjuntivo en oraciones interrogativas indirectas negadas que dependen del verbo *saber* (*no sé si/qué puedo/pueda)*, encabezadas por distintos sintagmas interrogativos en textos de procedencia peninsular y americana.

2 Modo verbal en las subordinadas interrogativas indirectas negadas dependientes del verbo *saber*

Las gramáticas, en general, suelen afirmar que en las oraciones interrogativas indirectas se utiliza el modo del indicativo en la subordinada. Esto puede ocurrir también si el verbo principal está negado, aunque entonces se admiten los dos modos, indicativo y subjuntivo, principalmente en caso de las llamadas dubitativas (Bello 1982, 335–336; Borrego Nieto/Gómez Asencio/Prieto 1987, 112; Fernández Álvarez 1987, 47; Gili Gaya 1981, 134–136; Matte Bon 1992, 64; Molho 1975, 416; RAE/ASALE 2010, 480; Sarmiento/Sánchez 1989, 269; Suñer 1999, 2184–2185).

Tanto Suñer (1999, 2185) como Ridruejo (1999, 3226) sostienen que el uso del subjuntivo en las oraciones interrogativas indirectas negadas era más frecuente en el español clásico que en la lengua actual. Keniston (1937, 348, 391, 392), para el siglo XVI, documenta tanto casos con indicativo como con subjuntivo en las interrogativas indirectas con un verbo de conocimiento, aunque añade que en este tipo de oraciones el indicativo es muy común, aún si el verbo principal está negado. Woehr (1977, 319) confirma este dato en un corpus que abarca los siglos XII a XVI.

Por otra parte, varios autores mencionan que el empleo del modo subjuntivo en esta clase de oraciones es dialectal. Así, Alarcos Llorach (1978, 247) caracteriza el uso como propio de hablantes no castellanos como los gallegos o asturianos. Asimismo, Suñer (1999, 2185) comenta que la alternancia entre subjuntivo o infinitivo (*no sé qué te diga/no sé qué decirte*) «ocurre en algunos dialectos

3 Cf. Torruella Casañas (2009, 100): «La utilización de técnicas estadísticas en la investigación en general y en la investigación en el campo de la lingüística histórica en particular, es hoy inevitable, puesto que abre la puerta a la justificación de las teorías existentes o a la argumentación de nuevas sobre bases analíticas».

hispanoamericanos como el peruano y el colombiano (entre otros)», dato confirmado en el *Manual* de la RAE y ASALE (2010, 480), donde se advierte que en el español americano, particularmente en México, Centroamérica, el Caribe y la zona andina, se suele usar el subjuntivo en expresiones como *no sé si te guste esta comida*. Además, un estudio de DeMello (1997) corrobora el carácter dialectal del empleo del subjuntivo en las interrogativas indirectas negadas, con mayor presencia en el español mexicano y menor presencia en el español chileno y frecuencias intermedias en el español colombiano y venezolano. Asimismo, Nieuwenhuijsen (2001), a partir de un corpus limitado, concluye que, mientras que en España se ha perdido la variación de modo en las subordinadas interrogativas negadas con el verbo *saber*, en México se conserva la misma.

De los estudios mencionados se puede desprender que antiguamente el uso del subjuntivo en las interrogativas indirectas negadas era más frecuente que hoy en día, si bien el indicativo siempre ha tenido una mayor frecuencia que el subjuntivo. Además, parece haber variación diatópica, dado que en varios trabajos se señala que el empleo del subjuntivo es más frecuente en el español americano que en el español peninsular. En España la variación se da sobre todo en hablantes no castellanos.

A continuación, en la sección 3 expondremos brevemente la recopilación y composición del corpus de ejemplos por medio del *CdE*; en la sección 4 analizaremos la evolución del empleo del indicativo y subjuntivo en el contexto sintáctico en cuestión. Además, en la sección 5 investigaremos la señalada variación diatópica distinguiendo entre ejemplos peninsulares y americanos. La sección 6 resume los resultados y considera la utilidad de los análisis estadísticos para la lingüística histórica de corpus.

3 El corpus

Como el *CdE* en gran parte está lematizado, es posible reunir un corpus de ejemplos amplio con un número de búsquedas muy reducido. La introducción de las secuencias:

> no [saber] si/[PQ*][VIP*]
> (no+forma verbal de saber+si/pronombre interrogativo+forma verbal del presente de indicativo)

y

> no [saber] si/[PQ*][VSP*]
> (no+forma verbal de saber+si/pronombre interrogativo+forma verbal del presente de subjuntivo)

resultó en ejemplos como los de (1) a (4). Para las subordinadas introducidas por *cómo*, *por qué* y *cuándo* se han realizado búsquedas aparte, dado que no salían automáticamente al buscar por la categoría de pronombre interrogativo (cf. los ejemplos (5) y (6)).

(1) son blancas de dentro, y el pescado de ellas, y muy sabrosas, no saladas, sino dulces y que han menester alguna sal, y dize que *no saben si naçen* en nácaras. (*Textos y documentos completos de Cristobal Colón*, siglo XV, *CdE*)
(2) ¿En qué me ejercitaré para agradaros? Gloria mía, yo *no sé qué haga*; decidmelo Vos, pues sabéis que deseo acertar a honraros y glorificaros. (*Epistolario*, siglo XVI, *CdE*)
(3) Aunque es verdad que la debo obligaciones, repara que ella *no sabe quién es*; y es bajeza y es infamia casarme yo con mujer… Clotaldo. (*La vida es sueño*, siglo XVII, *CdE*)
(4) Respuesta. – *No sé cuál sea* la estrategia que vaya a tomar el Pri, la verdad es un problema del Pri en este punto y nosotros estamos defendiendo nuestros puntos de vista de una iniciativa válida, meditada, bien hecha, (*Entrevista PAN*, siglo XX, *CdE*)
(5) Pues assí goze de mi alma, no se me ha quitado el mal de la madre; *no sé cómo pueda* ser. (*La Celestina*, siglo XV, *CdE*)
(6) *No sé por qué pasa* ni cómo explicarlo, pero sé que ocurre y que el público también lo siente. (*Entrevista ABC*, siglo XX, *CdE*)

El corpus así formado, por tanto, comprende subordinadas indirectas que dependen del verbo *saber* y están encabezadas por los sintagmas interrogativos *cómo*, *cuál*, *cuándo*, *cuánto*, *dónde*, *por qué*, *qué*, *quién* y *si*. Todos los ejemplos se han revisado «a mano», para quitar algunos casos dobles. En total se trata de 2202 ejemplos (1833 de indicativo, 369 de subjuntivo).

4 Frecuencia del modo verbal en las subordinadas interrogativas indirectas negadas con el verbo *sabe*

4.1 Frecuencias generales

Como primer paso se ha calculado la frecuencia del indicativo y subjuntivo en las subordinadas interrogativas indirectas negadas en general. Se ha utilizado el programa estadístico *SPSS*, con el cual se pueden realizar cómputos relativamente sencillos así como pruebas estadísticas relativamente complejas, como se verá más adelante.

Los resultados del primer cálculo se plasman en el gráfico 1, del que se desprende claramente que el indicativo, en general, es mucho más frecuente que el subjuntivo (el 83,24 % frente al 16,76 %).

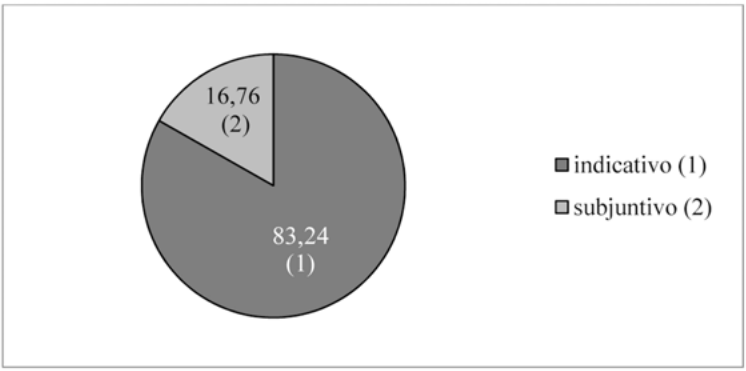

Gráfico 1: Frecuencia general (%) de *indicativo y subjuntivo* en subordinadas interrogativas indirectas negadas con *saber*

En el gráfico 2 se observa la frecuencia de los distintos tipos de interrogativas indirectas negadas. Es evidente que las interrogativas introducidas por *si* son las más frecuentes (35,2 %), seguidos por las que llevan *qué* (26,48 %) y *cómo* (13,03 %). Por otra parte, las interrogativas indirectas encabezadas por *cuándo* y *cuánto* ni siquiera llegan al 1 % en el corpus. Como las subordinadas con *cuándo* y *cuánto*, además, solo registran formas verbales en indicativo, los ejemplos correspondientes se han excluido de los demás cálculos del corpus.

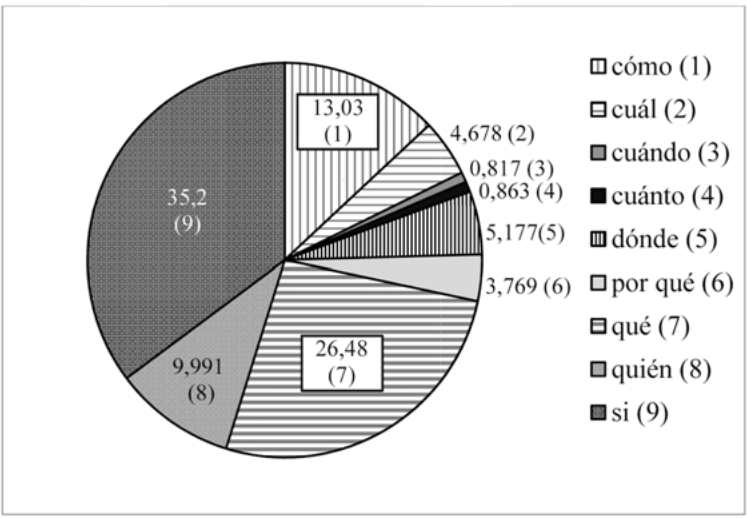

Gráfico 2: Frecuencia (%) de los distintos *sintagmas interrogativos* en interrogativas indirectas negadas con *saber*

El gráfico 1 presenta la frecuencia de los modos indicativo y subjuntivo en el corpus en su totalidad, pero de los estudios citados en la sección 2 se desprende que la distribución de ambos modos no siempre ha sido igual, es decir, en épocas anteriores el subjuntivo era más frecuente en las subordinadas interrogativas que en la actualidad. Por eso, se han hecho cortes para cada siglo presente en el corpus, calculando los porcentajes de uso de ambos modos. Los resultados del cálculo, presentados en la tabla 1, arrojan luz sobre la propagación de un modo frente al retroceso del otro a través de los siglos.

Tabla 1: Frecuencia de *indicativo* y *subjuntivo* en subordinadas interrogativas indirectas negadas con *saber* a través de los siglos

Siglo	Modo		Total
	INDICATIVO	SUBJUNTIVO	
XIII	3	0	3
	100 %	0 %	100 %
XIV	–	–	–
XV	1	3	4
	25 %	75 %	100 %
XVI	385	147	532
	72,4 %	27,6 %	100 %
XVII	395	111	506
	78,1 %	21,9 %	100 %
XVIII	144	32	176
	81,8 %	18,2 %	100 %
XIX	308	31	339
	90,9 %	9,1 %	100 %
XX	597	45	642
	93 %	7 %	100 %
Total	1833	369	2202
	83,2 %	16,8 %	100 %

Chi^2=123,233 (gl=6; p=0,000)

En los primeros siglos el uso del subjuntivo oscila considerablemente, pero es de notar que el corpus cuenta con muy pocos ejemplos para los siglos XIII y XV y no

registra ningún caso para el siglo XIV. Asimismo, los tres ejemplos del siglo XIII provienen de *Siete partidas*, un texto cuyo manuscrito original data del siglo XIII, pero que en *CORDE* aparece con la fecha de 1491, de acuerdo con la fecha de la primera edición conocida. A pesar de eso, en general se puede concluir que el empleo del subjuntivo en las subordinadas interrogativas indirectas negadas con *saber*, efectivamente, baja a lo largo de los siglos, del 27,6 % en el siglo XVI al 7 % en el siglo XX.

El desarrollo esbozado aquí se aprecia con mayor nitidez en el gráfico 3, que representa los datos de la tabla 1.

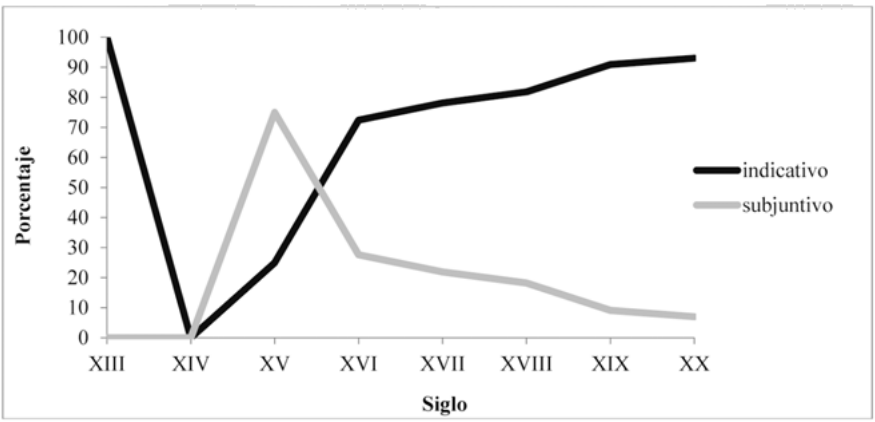

Gráfico 3: Frecuencia del empleo de *indicativo* y *subjuntivo* en subordinadas interrogativas indirectas negadas con *saber* a través de los siglos

4.2 La influencia del factor tiempo

Los porcentajes de la tabla 1 están calculados sobre los ejemplos del corpus reunido para este trabajo y, en tal calidad, dan información sobre la distribución de los dos modos verbales en el mismo. Sin embargo, dado que estudios anteriores advierten que el uso del subjuntivo en las interrogativas indirectas negadas era más frecuente en el español clásico, lo cual se comprueba en nuestro corpus, interesa conocer también el grado de influencia que ejerce el factor tiempo sobre la aparición del subjuntivo en esta clase de oraciones, es decir, interesa saber si existe una correlación entre el tiempo y el uso del subjuntivo en general.

Para investigar dicha correlación, se ha aplicado un test de regresión logística binaria, con el que se puede examinar si una o más variables independientes o predictoras tienen influencia sobre una variable dependiente y si el efecto es

positivo o negativo. En el siguiente cómputo, la variable dependiente es el modo verbal y la variable independiente la constituye el tiempo. El test se basa en todos los ejemplos del corpus y, a partir de su codificación en términos de siglo y modo, desarrolla un modelo que predice la probabilidad de que aparezca uno de los modos verbales. Asimismo, el test indica si los valores encontrados tienen significación estadística.

La tabla 2 recoge los datos más importantes del test de regresión logística binaria.

Tabla 2: Probabilidad y valor de significación del empleo del *subjuntivo* con la variable independiente de *siglo*

	Wald	gl	Sig.	Exp(B)
siglo	103,531	1	,000	,666
constante	64,376	1	,000	266,195

Chi^2=116,329 (gl=1; p=0,000)

En la columna izquierda se encuentra la variable independiente de tiempo (siglo). En esta misma columna figura también la constante o intersección, que indica la probabilidad estimada de que aparezca el subjuntivo si el valor de todas las variables independientes es de 0. En general, el valor de la constante no tiene relevancia independiente. La columna titulada 'Wald' da el resultado de la prueba de Wald, un test con el que se determina si los datos de la columna derecha (Exp(B), el exponencial del coeficiente u *odds ratio*) son significativos. El Exp(B), por su parte, indica la probabilidad de que aparezca la variable dependiente con la variable independiente. Un valor de Exp(B) mayor de 1 quiere decir que la probabilidad de que la variable dependiente aparezca con la variable independiente aumenta (si la variable independiente sube con un punto). Un Exp(B) mayor de 1, por tanto, apunta a un efecto positivo. En cambio, con un Exp(B) < 1 la misma probabilidad disminuye, en cuyo caso, por consecuencia, se detecta un efecto negativo. La columna titulada 'gl' indica los grados de libertad, o sea el número de variables involucradas en el cálculo, determinado automáticamente por el programa. Por último, en la columna 'Sig.' se da el valor *p*, que señala la significatividad estadística del efecto dado bajo Exp(B). Si el valor *p* es < 0,05 se entiende que tiene significatividad estadística y que la diferencia encontrada probablemente no se deba al azar. En cambio, con un valor de p > 0,05 no hay significación estadística y no se puede descartar la posibilidad de que se trate de una asociación casual.

De la tabla 2 se puede concluir que la probabilidad de que el verbo aparezca en subjuntivo (frente al indicativo) en la interrogativa indirecta negada disminuye

a través de los siglos (Exp(B) < 1, =0,666) y que el efecto tiene significación estadística (Sig.=0,000). Esto encaja con los resultados de la tabla 1, si bien en los siglos XIII y XV el corpus proporcionaba un panorama relativamente irregular (el 0 % en el siglo XIII y el 75 % en el siglo XV).

Al presentar los resultados del test estadístico de manera gráfica, se observa que la probabilidad de que aparezca el subjuntivo disminuye constantemente a través de los siglos, con una decadencia muy marcada en los primeros siglos (gráfico 4).

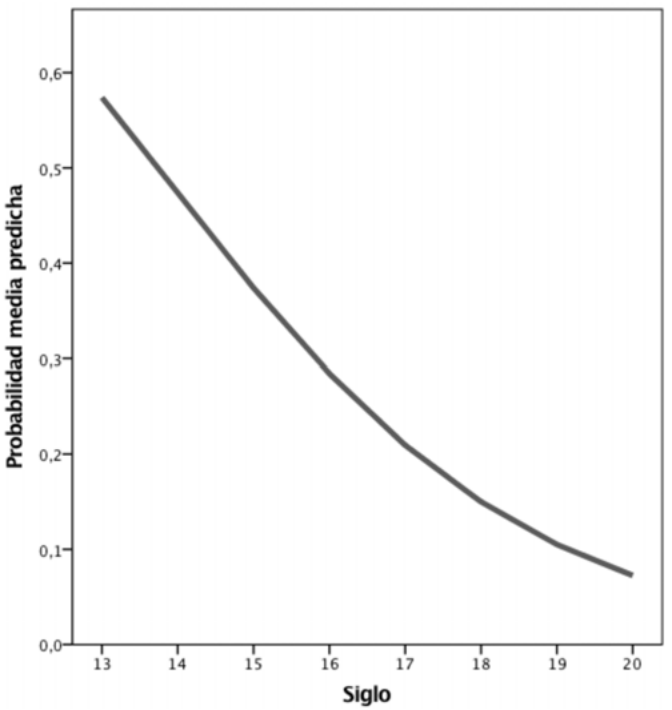

Gráfico 4: Probabilidad media predicha del empleo del *subjuntivo* en subordinadas interrogativas indirectas negadas con *saber* a través de los siglos

4.3 Los distintos sintagmas interrogativos y el modo verbal

En los apartados anteriores se han presentado datos acerca del uso del indicativo y subjuntivo a través de los siglos sin distinguir entre la clase de interrogativa indirecta, es decir, sin distinguir entre los sintagmas interrogativos que encabezan las subordinadas. Sin embargo, no se puede descartar la posibilidad de que el uso

del modo verbal sea diferente según el sintagma interrogativo específico que introduzca la oración subordinada; ya se ha señalado que con los sintagmas *cuándo* y *cuánto* el corpus únicamente registra el modo indicativo.

El gráfico 5 presenta la frecuencia de los dos modos verbales con los distintos sintagmas interrogativos.

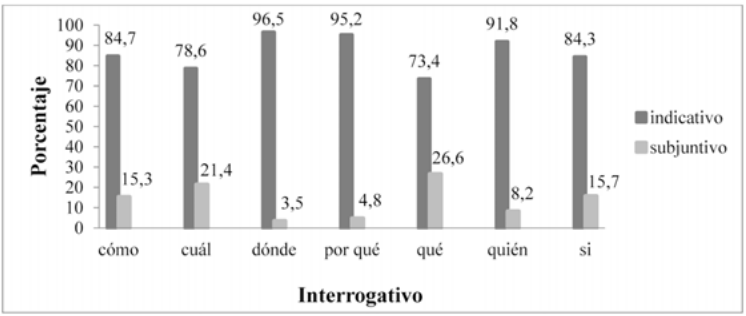

Gráfico 5: Frecuencia del empleo de *indicativo* y *subjuntivo* en subordinadas interrogativas indirectas negadas con *saber* con distintos *sintagmas interrogativos*

Del gráfico 5 se desprende que el subjuntivo es más frecuente en las subordinadas interrogativas indirectas encabezadas por *qué* (26,6 %), seguido por *cuál* (21,4 %), *si* (15,7 %) y *cómo* (15,3 %) respectivamente. Los demás sintagmas presentan porcentajes del subjuntivo por debajo del 10 %.

Si bien el empleo del subjuntivo, en general, disminuye a lo largo de los siglos (cf. la tabla 1), los distintos sintagmas interrogativos presentan desarrollos divergentes, tal como se puede apreciar en el gráfico 6.

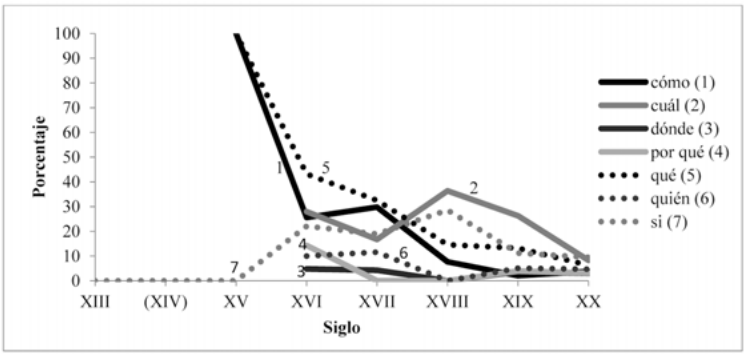

Gráfico 6: Frecuencia del empleo de *subjuntivo* en subordinadas interrogativas indirectas negadas con *saber* encabezadas por distintos *sintagmas interrogativos* a través de los siglos

A partir del gráfico 6 se puede concluir que, si bien en todos los sintagmas interrogativos el empleo del modo subjuntivo disminuye a través de los siglos, *cuál* y *si* presentan un aumento en el siglo XVIII, aumento que también se observa en *cómo* en el siglo XVII. La fuerte caída observada en *cómo* y *qué* del siglo XV a XVI (100 % a 25,6 % y 43,2 % respectivamente), no es fidedigna, dado que se trata de 1 y 2 casos de subjuntivo en total.

4.4 Los sintagmas interrogativos y el uso del subjuntivo a través del tiempo

De cada sintagma interrogativo también se ha calculado la probabilidad de que aparezca con una forma verbal de subjuntivo en la subordinada, además de la interacción entre las dos variables independientes, es decir, entre cada uno de los sintagmas interrogativos y el tiempo. Los resultados del test se dan en la tabla 3.

Tabla 3: Probabilidad y valor de significación del empleo del *subjuntivo* con las variables independientes de *siglo* y *sintagma interrogativo* e interacción entre *siglo* y *sintagma interrogativo* en subordinadas interrogativas indirectas negadas con *saber*

	Wald	df	Sig.	Exp(B)
siglo	57,061	1	,000	,546
qué	23,257	6	,001	
cómo	,002	1	,965	1,127
cuál	5,309	1	,021	,001
dónde	2,991	1	,084	,000
por qué	,808	1	,369	,005
si	16,379	1	,000	,001
quién	4,748	1	,029	,001
qué por siglo	20,616	6	,002	
cómo por siglo	,075	1	,785	,958
cuál por siglo	5,201	1	,023	1,512
dónde por siglo	2,041	1	,153	1,634
por qué por siglo	,417	1	,518	1,238
si por siglo	14,274	1	,000	1,462

Tabla 3: (continuada)

	Wald	df	Sig.	Exp(B)
quién por siglo	3,150	1	,076	1,421
constante	48,069	1	,000	12875,463

Chi^2=206,254 (gl=13; p=0,000); ref.=qué

El test estadístico revela varios resultados interesantes. Por una parte, se confirma el hecho de que el tiempo influye en la aparición del subjuntivo en las interrogativas indirectas, en el sentido de que el empleo de este modo disminuye a través de los siglos (Exp(B)=0,546; Sig.=0,000). Por otra parte, la probabilidad de que ocurra el subjuntivo en la subordinada encabezada por *cuál*, *dónde*, *por qué*, *si* y *quién* disminuye en comparación con la subordinada introducida por *qué* (Exp(B) < 1), en tanto que con *cómo* dicha probabilidad aumenta moderadamente (Exp(B)=1,127) comparada con *qué*. Sin embargo, el efecto solo es significativo con *si* (Sig.=0,000). Asimismo, a través del tiempo la probabilidad de que aparezca el subjuntivo aumenta con todos los sintagmas interrogativos en comparación con *qué*, menos con *cómo*, pero, de nuevo, únicamente en el caso de *si* el resultado tiene significación estadística (Sig.=0,00). El gráfico 7, creado a base de los resultados del test de regresión, presenta la influencia del factor tiempo sobre la ocurrencia del subjuntivo con los distintos sintagmas interrogativos.

Como se puede ver, la probabilidad de que el verbo de la subordinada esté en subjuntivo disminuye a lo largo del tiempo independientemente del sintagma interrogativo, puesto que con todos los sintagmas interrogativos la curva desciende. No obstante, las pérdidas más substanciales a través de los siglos se dan con los interrogativos *cómo* y *qué*, y, aunque en menor grado, también con *si*. Por otra parte, con *dónde*, *por qué* y *quién* la probabilidad siempre ha sido relativamente baja, de manera que su papel en la disminución del uso del subjuntivo en esta clase de oraciones subordinadas en perspectiva diacrónica es insignificante. Asimismo, si bien en el caso de *cuál* la probabilidad de que el verbo aparezca en subjuntivo ha bajado a lo largo del tiempo, comparable con la caída de *si* a partir del siglo XVI, son las interrogativas encabezadas por *cuál* las que en el siglo XX tienen mayor probabilidad de aparecer con subjuntivo.

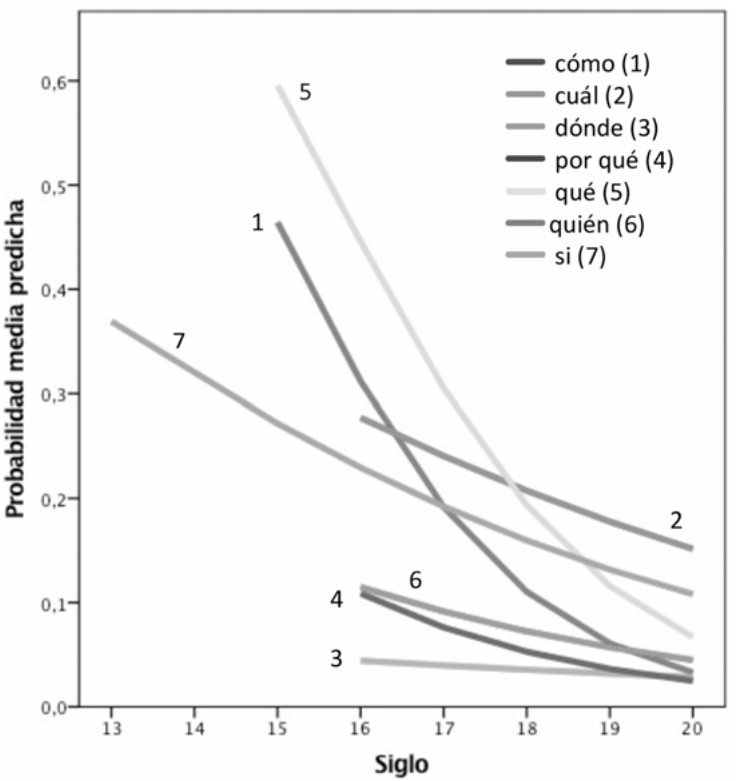

Gráfico 7: Probabilidad media predicha del empleo del *subjuntivo* en subordinadas interrogativas indirectas negadas con *saber* introducidas por distintos *sintagmas interrogativos* a través de los siglos

5 Frecuencia del modo verbal en las subordinadas interrogativas indirectas negadas con el verbo *saber + si* en textos peninsulares y americanos

El corpus reunido para este estudio permite distinguir entre ejemplos procedentes de textos peninsulares, por una parte, y ejemplos procedentes de textos americanos, por otra parte.[4] Para saber si, de hecho, se documenta una variación

4 Somos conscientes de que la clasificación en términos de «América» y «americano» no corresponde con la rica variación lingüística en el continente americano. Sin embargo, dado el

diatópica en el uso del indicativo y subjuntivo en las subordinadas interrogativas negadas con *saber*, se han categorizado todos los ejemplos en términos de procedencia de manera manual, ya que el *CdE* no procesa esta etiquetación automáticamente. Es de notar que para este cómputo nos hemos limitado a los casos de *si*, la clase de subordinada interrogativa que ofrece el mayor número de ejemplos en el corpus (véase el gráfico 2).

En la tabla 4 se presentan los porcentajes del indicativo y subjuntivo por siglo y por origen de los ejemplos. Los signos de interrogación indican que no se conoce el origen, porque el autor es anónimo.

Tabla 4: Frecuencia de *indicativo* y *subjuntivo* en subordinadas interrogativas indirectas negadas con *saber* introducidas por *si* en textos *peninsulares* y *americanos* a través de los siglos

Siglo	Origen	Modo		Total	Significación
		INDICATIVO	SUBJUNTIVO		
XIII	España	3	0	3	
		100 %	0 %	100 %	
XV	España	1	0	1	
		100 %	0 %	100 %	
XVI	??	7	1	8	$Chi^{2=}1{,}055$; gl=2; p=0,590
		87,5 %	12,5 %	100 %	
	América	2	0	2	
		100 %	0 %	100 %	
	España	104	31	135	
		77 %	23 %	100 %	
XVII	??	3	1	4	$Chi^{2=}1{,}422$; gl=2; p=0,491
		75 %	25 %	100 %	
	América	6	3	9	
		66,7 %	33,3 %	100 %	
	España	124	27	151	
		82,1 %	17,9 %	100 %	

bajo número de ejemplos procedentes de textos no peninsulares (para América: 208 ejemplos en total, solo 35 de subjuntivo) no ha sido posible hacer una subclasificación por país o zona dialectal americana.

Tabla 4: (continuada)

Siglo	Origen	Modo		Total	Significación
		INDICATIVO	SUBJUNTIVO		
XVIII	América	0	2	2	$Chi^{2=}5{,}147$; gl=1; p=0,023
		0 %	100 %	100 %	
	España	50	18	68	
		73,5 %	26,5 %	100 %	
XIX	??	1	0	1	$Chi^{2=}2{,}203$; gl=2; p=0,332
		100 %	0 %	100 %	
	América	10	3	13	
		76,9 %	23,1 %	100 %	
	España	76	8	84	
		90,5 %	9,5 %	100 %	
XX	América	155	27	182	$Chi^{2=}15{,}641$; gl=1; p=0,000
		85,2 %	14,8 %	100 %	
	España	111	1	112	
		99,1 %	0,9 %	100 %	
Total	??	11	2	13	$Chi^{2=}0{,}252$; gl=2; p=0,8810
		84,6 %	15,4 %	100 %	
	América	173	35	208	
		83,2 %	16,8 %	100 %	
	España	469	85	554	
		84,7 %	15,3 %	100 %	

Esta tabla muestra que del siglo XVII en adelante el empleo del subjuntivo siempre es más alto en ejemplos procedentes de textos americanos que en ejemplos procedentes de textos peninsulares. En el siglo XX solo se documenta un caso de subjuntivo en España, que resulta ser una cita literal de unas palabras pronunciadas por un historiador del siglo XVII.

En el gráfico 8, que representa los porcentajes del subjuntivo de la tabla 4, se aprecia claramente la mayor preferencia por el subjuntivo en ejemplos americanos en comparación con ejemplos peninsulares. Asimismo, mientras que en

España el empleo del subjuntivo oscila ligeramente a través de los siglos, con mayor uso en el siglo XVIII (26,5 %), en América se observa un aumento del uso del subjuntivo muy marcado entre el siglo XVII y el XVIII, aunque se basa en escasos ejemplos.

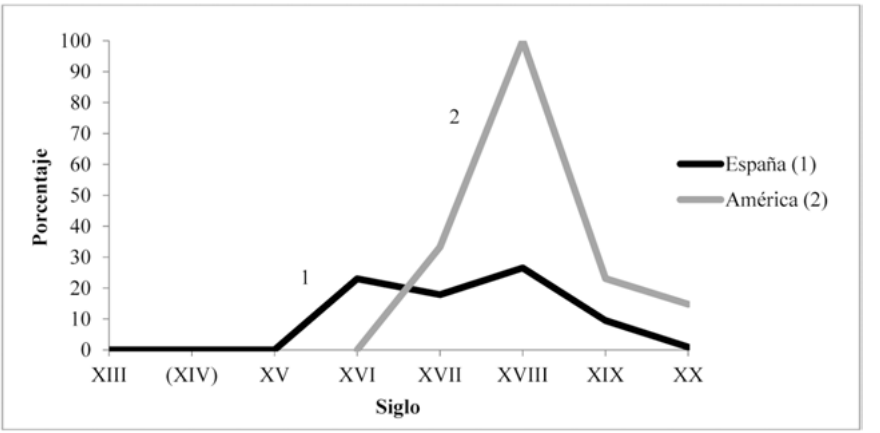

Gráfico 8: Frecuencia del empleo del *subjuntivo* en subordinadas interrogativas indirectas negadas con *saber* introducidas por *si* en textos *peninsulares* y *americanos* a través de los siglos

5.1 La influencia de los factores origen y tiempo

Si bien el gráfico 8 presenta un panorama general del uso del subjuntivo en interrogativas indirectas negadas a través de los siglos, el test de regresión logística binaria es capaz de medir la influencia del origen de los ejemplos sobre la aparición del subjuntivo, es decir es capaz de predecir la probabilidad de que el subjuntivo aparezca en ejemplos americanos frente a ejemplos peninsulares. Además, con el mismo test se puede medir la interacción de las dos variables independientes de origen y tiempo, o sea que se puede medir si el efecto del tiempo para España es distinto del efecto para América. La tabla 5 recoge los resultados más importantes del test.

Tabla 5: Probabilidad y valor de significación del empleo del *subjuntivo* con las variables independientes de *siglo* y *origen* e interacción entre *siglo* y *origen* en subordinadas interrogativas indirectas negadas con *saber* introducidas por *si*

	Wald	gl	Sig.	Exp(B)
siglo	19,568	1	,000	,679
América	,000	1	,999	,993
América por siglo	,052	1	,819	1,051
constante	11,211	1	,001	156,596

Chi^2=24,767 (gl=3; p=0,000); ref.=España

De esta tabla se deprende, otra vez, que el empleo del subjuntivo en estas interrogativas indirectas disminuye a través del tiempo. El Exp(B) de esta variable es < 1 (0,679), y el efecto tiene significatividad estadística (Sig.=0,000).

En caso del origen, se observa que esta variable no influye de manera significativa en la aparición del subjuntivo. Es verdad que la probabilidad de que el subjuntivo ocurra en ejemplos americanos en comparación con la aparición de ese modo verbal en ejemplos peninsulares disminuye (Exp(B)=0,993), pero la diferencia no es significativa en absoluto (Sig.=0,999).

Lo mismo se observa con la interacción del origen y tiempo. La probabilidad de que el subjuntivo aparezca en textos americanos (frente a textos peninsulares) aumenta en cada siglo sucesivo (Exp(B)=1,051), pero el efecto no alcanza la significatividad estadística (Sig.=0,819).

A continuación se plasman los resultados del test de regresión logística binaria de manera gráfica (gráfico 9).

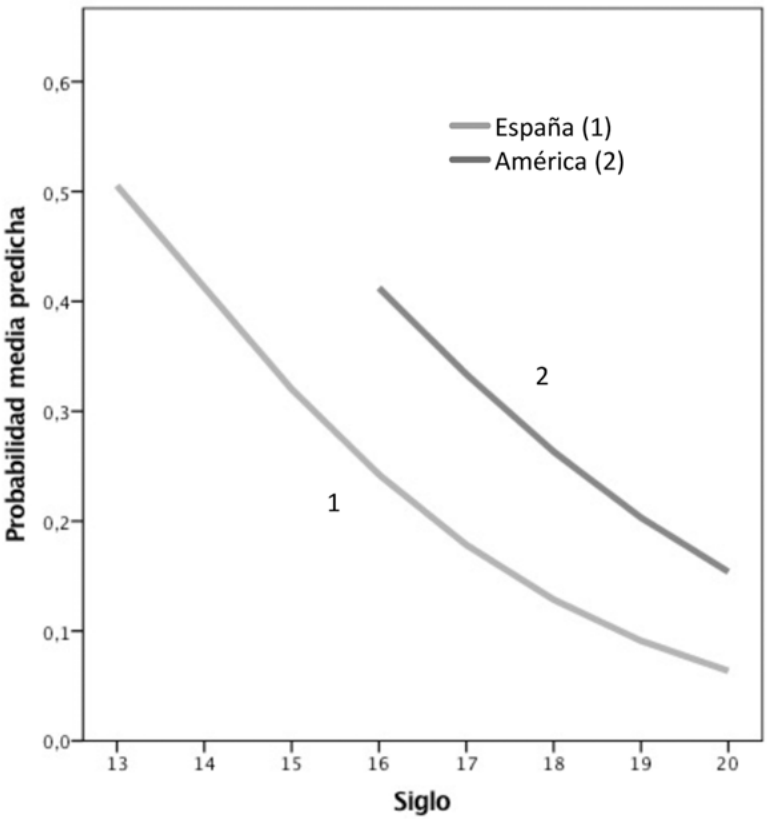

Gráfico 9: Probabilidad media predicha del empleo del *subjuntivo* en subordinadas interrogativas indirectas negadas con *saber* introducidas por *si* en textos *peninsulares* y *americanos* a través de los siglos

Por una parte, el gráfico 9 muestra que la probabilidad de que el modo subjuntivo ocurra en interrogativas indirectas negadas introducidas por *si* en todo el período es más alta en ejemplos de origen americano que en ejemplos de origen peninsular, lo que corresponde con las observaciones de varios estudiosos sobre del tema. Por otra parte, si bien en el corpus se ha observado un aumento del empleo del subjuntivo en ejemplos americanos entre los siglos XVI y XVIII (gráfico 8), el test estadístico predice un descenso continuo y gradual para el uso del subjuntivo en América. De hecho, el gráfico 9 sugiere que la disminución se ha producido y se está produciendo por igual en España y en América, puesto que las dos líneas descienden de manera muy similar.

6 Conclusiones

Con la creación de los corpus digitales diacrónicos en línea y la disponibilidad de grandes cantidades de datos, se ha hecho casi imprescindible el uso de un programa estadístico para trabajar estos datos. Teóricamente, el cálculo de la frecuencia de cierta forma lingüística se puede hacer de manera manual y el provecho de un programa estadístico radica en tal caso, sobre todo, en la mayor comodidad, rapidez y corrección al realizar los cálculos. Dichos cálculos relativamente sencillos son esenciales para llegar a conocer las tendencias generales del material bajo estudio y para saber cómo se desarrolla cierta forma o construcción en el corpus de ejemplos. Los gráficos 1 a 3 y la tabla 1 dan cuenta de las tendencias generales en el corpus acerca del uso del modo verbal en oraciones interrogativas indirectas negadas que dependen del verbo *saber*.

Por otra parte, el tratamiento estadístico de los datos permite calcular probabilidades del empleo de cierta forma o construcción, además de medir la influencia de cierto factor o variable independiente sobre la aparición de una forma o variable dependiente. Estos cálculos son importantes para comprobar o refutar ciertas tendencias observadas en el corpus. El gráfico 3 del presente trabajo sugiere un aumento del uso del subjuntivo en las interrogativas indirectas negadas con el verbo *saber* en el siglo XV y una marcada decaída en el siglo XVI; no obstante, el test de regresión logística muestra que la probabilidad de que el verbo aparezca en subjuntivo en esta clase de oraciones va disminuyendo continuamente a través de los siglos y no comprueba ninguna de las oscilaciones sugeridas por las frecuencias porcentuales.

El mismo tipo de refutación se observa en los cómputos del empleo de subjuntivo con los distintos sintagmas interrogativos, ya que en el gráfico 7, que plasma las probabilidades de ocurrencia del subjuntivo con cada sintagma interrogativo, se aprecia, otra vez, que en perspectiva diacrónica el subjuntivo constantemente pierde terreno en las interrogativas indirectas negadas con el verbo *saber* y que ninguno de los aumentos sugeridos por el gráfico 6 se confirma con el test estadístico.

Además, el gráfico 7 muestra que la propagación del indicativo a expensas del subjuntivo se produce, principalmente, en las oraciones interrogativas encabezadas por *cómo*, *qué* y *si*, y que a través de los siglos el subjuntivo se ha mantenido más firme en las interrogativas encabezadas por *cuál*. Con el test de regresión logística, por tanto, se puede capturar el cambio lingüístico en curso, dado que los resultados del mismo predicen con cierto detalle cómo habrá transcurrido la disminución del uso del subjuntivo a través de los siglos, es decir en qué contextos y en qué momento este modo verbal habrá perdido más terreno.

Con el test de regresión logística también se ha podido medir la influencia del factor origen y la interacción entre este factor y el tiempo. Aunque es cierto que la probabilidad de que el subjuntivo aparezca en esta clase de subordinadas encabezadas por *si* es más alta en textos de origen americano que en textos peninsulares, tanto en el período antiguo como en la actualidad, el presente estudio también demuestra que se ha producido un descenso constante del empleo del subjuntivo tanto en América como en España y que la pérdida de terreno del subjuntivo en este contexto sintáctico diatópicamente es muy similar.

A pesar del gran valor de los test estadísticos para la lingüística histórica, no queremos abogar aquí por la supresión de los análisis tradicionales y la sustitución completa de los análisis tradicionales por las pruebas estadísticas. El análisis cuantitativo sigue siendo imprescindible para formarse una idea global de la frecuencia y desarrollo de una forma o construcción sintáctica. Asimismo, el análisis cualitativo, es decir el detenido estudio de ejemplos específicos en su contexto, permite identificar posibles factores que hayan influido en la evolución del cambio lingüístico. El tratamiento estadístico, en cambio, constituye una herramienta complementaria muy potente, que sirve para comprobar la validez de las conclusiones sacadas en ambos tipos de análisis y para medir la posible influencia de distintos factores identificados en el material estudiado.[5]

7 Corpus

CdE – Davies, Mark, *Corpus del español*, <http://www.corpusdelespanol.org> [última consulta: junio de 2014].

CORDE – Real Academia Española, *Corpus diacrónico del español*, <http://www.rae.es> [última consulta: junio de 2014]

8 Bibliografía

Alarcos Llorach, Emilio, *Estudios de gramática funcional del español*, Madrid, Gredos, 1978.

Bello, Andrés, *Gramática de la lengua castellana*, Madrid, EDAF, 1982.

Borrego Nieto, Julio/Gómez Asencio, José J./Prieto, Emilio, *El subjuntivo. Valores y uso*, Madrid, SGEL, 1987.

5 Cf. Torruella Casañas (2009, 100): «La estadística, sin embargo, debe ser considerada sólo como un puro instrumento para la investigación, nunca como su finalidad. Un instrumento que ha de ayudar a la consecución de dos objetivos: en primer lugar, describir y resumir los datos y, en segundo lugar, hacer estimaciones de significación y de fiabilidad».

De Jonge, Robert/Nieuwenhuijsen, Dorien, *Formación del paradigma pronominal y formas de tratamiento*, in: Company Company, Concepción (dir.), *Sintaxis histórica de la lengua española. Segunda parte: La frase nominal*, vol. 2, México D.F., Universidad Nacional Autónoma de México y Fondo de Cultura Económica, 2009, 1593–1671.

DeMello, George, *Tense and mood after No sé si*, Hispanic Review 63:4 (1995), 555–573.

Fernández Álvarez, Jesús, *El subjuntivo*, Madrid, Edelsa, 1987.

Gili Gaya, Samuel, *Curso superior de sintaxis española*, Barcelona, Bibliograf, 1981.

Kabatek, Johannes, *Lingüística empática*, RILCE, Revista de Filología Hispánica 30:3 (2014), 705–723.

Keniston, Hayward, *The Syntax of Castilian Prose. The Sixteenth Century*, Chicago, The University of Chicago Press, 1937.

Matte Bon, Francisco, *Gramática comunicativa del español*, vol. 1, Madrid, Difusión, 1992.

Molho, Mauricio, *Sistemática del verbo español*, Madrid, Gredos, 1975.

Nieuwenhuijsen, Dorien, *Modo verbal en las oraciones interrogativas indirectas*, Nueva Revista de Filología Hispánica XLIX:2 (2001), 339–362.

Nieuwenhuijsen, Dorien, *El rastreo del desarrollo de algunos pronombres personales en español: (im)posibilidades de los corpus diacrónicos digitales*, in: Enrique-Arias, Andrés (ed.), *Diacronía de las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus*, Madrid/Fráncfort, Iberoamericana/Vervuert, 2009, 365–384.

Nieuwenhuijsen, Dorien, *Yuxtaposición y tradiciones discursivas en el español antiguo*, La corónica 41:2 (2013), 135–172.

Nieuwenhuijsen, Dorien, *Oraciones yuxtapuestas*, in: Company Company, Concepción (dir.), *Sintaxis histórica de la lengua española. Tercera parte: Adverbios, preposiciones, conjunciones. Relaciones interoracionales*, vol. 2, México D.F., Universidad Nacional Autónoma de México y Fondo de Cultura Económica, 2014, 387–436.

Real Academia Española/Asociación de Academias de la Lengua Española, *Nueva gramática de la lengua española. Manual*, Madrid, Espasa, 2010.

Ridruejo, Emilio, *Modo y modalidad. El modo en las subordinadas sustantivas*, in: Bosque, Ignacio/Demonte, Violeta (edd.), *Gramática descriptiva de la lengua española*, vol. 2, Madrid, Espasa-Calpe, 1999, 3209–3251.

Rojo, Guillermo, *De nuevo sobre la frecuencia de las formas llegara y llegase*, in: Albrecht, Jörn/Harslem, Frank (edd.), *Heidelberger Spätlese. Ausgewählte Tropfen aus verschiedenen Lagen der spanischen Sprach- und Übersetzungswissenschaft. Festschrift anlässlich des 70. Geburtstages von Prof. Dr. Nelson Cartagena*, Bonn, Romanistischer Verlag, 2008, 161–182.

Rojo, Guillermo, *Sobre codificación y explotación de corpus textuales: Otra comparación del Corpus del español con el CORDE y el CREA*, Lingüística 24 (2010), 11–50.

Rojo, Guillermo, *El papel de los corpus en el estudio de la historia del español*, in: Montero Cartelle, Emilio (ed.), *Actas del VIII Congreso Internacional de Historia de la Lengua Española (Santiago de Compostela, 2009)*, vol. 1, Santiago de Compostela, Meubook, 2012, 433–444.

Sarmiento, Ramón/Sánchez, Aquilino, *Gramática básica del español. Norma y uso*, Madrid, SGEL, 1989.

Suñer, Margarita, *La subordinación sustantiva: la interrogación indirecta*, in: Bosque, Ignacio/Demonte, Violeta (edd.), *Gramática descriptiva de la lengua española*, vol. 2, Madrid, Espasa-Calpe, 1999, 2149–2195.

Torruella Casañas, Joan, *Bases científicas en la investigación a partir de corpus: el caso del* Corpus Informatitzat del català antic, in: Enrique-Arias, Andrés (ed.), *Diacronía de las*

lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus, Madrid/Fráncfort, Iberoamericana/Vervuert, 2009, 95–115.

Woehr, Richard, *Syntactic atrophy and the indirect interrogative in Spanish*, Studia Neophilologica 49:2 (1977), 311–326.