Datación por coeficientes de asociación[281]

Dating by correlation coefficients

Yoshifumi Kawasaki
Universidad de Tokio (JSPS investigador)

RESUMEN

En el presente trabajo tratamos de demostrar que *Dice-Sorenson*, *Jaccard*, *Phi convertido*, *Phi modificado* son los coeficientes de asociación más apropiados para nuestro método de datación propuesto en Kawasaki (en evaluación). Con *Phi modificado* más del 60% de los 538 documentos en el *Corpus de Documentos de Cancillería Real* (*CODCAR*) resultaron datados con un margen de error de ±5 años respecto a la fecha verdadera, y cerca de 80% con un margen de error de ±10 años. Por otra parte, hemos comprobado que el uso de n ($1 \leq n \leq 20$) documentos más semejantes con los que se realiza la datación de un texto no altera sustancialmente la capacidad predictiva. Además, hemos detectado una fuerte correlación negativa entre el promedio del margen de error absoluto y el número de documentos en el periodo dado.

Palabras clave: Datación, coeficiente de asociación (correlación), *Corpus de Documentos de Cancillería Real* (*CODCAR*).

ABSTRACT

This paper attempts to demonstrate that *Dice-Sorenson*, *Jaccard*, *Converted Phi*, *Adjusted Phi* are the most suitable correlation coefficients for the dating process proposed in my previous study. With the use of *Adjusted Phi*, more than 60% of the 538 medieval Spanish documents in the *Corpus de Documentos de Cancillería Real* (*CODCAR*) were successfully dated within the margin of error of ±5 years, and around 80% within ±10 years. On the other hand, the study confirmed that the predictability was not affected substantially by the use of n texts ($1 \leq n \leq 20$) showing the strongest correlation with the one to be dated. Moreover, a strong

281 Agradecemos a la profesora Nieves Sánchez González de Herrero de la Universidad de Salamanca por facilitarnos la transcripción paleográfica de los documentos incorporados al *Corpus de Documentos de Cancillería Real* (*CODCAR*). Nuestro agradecimiento es también para el profesor Hiroto Ueda de la Universidad de Tokio por sus valiosos comentarios. Para la revisión del artículo en español se ha contado con la ayuda de la profesora Ana Isabel García de la Universidad de Tokio. Este trabajo ha sido subvencionado por la *JSPS* (*Japan Society for the Promotion of Science*) KAKENHI Grant Number 13J03408 (This work was supported by JSPS KAKENHI Grant Number 13J03408).

negative correlation between the mean absolute error and the number of the documents in a given period was detected.

Keywords: Dating, correlation coefficient, *Corpus de Documentos de Cancillería Real* (*CODCAR*).

— o —

INTRODUCCIÓN

Este breve trabajo tiene como objetivo averiguar la predictibilidad que presentan una decena de coeficientes de asociación con el método de datación propuesto en Kawasaki (en evaluación). En el anterior estudio no nos servimos sino del coeficiente *Phi modificado* a la hora de medir la semejanza entre documentos, cuando existen múltiples coeficientes a nuestra disposición y no hay seguridad de que dicho coeficiente brinde la mejor capacidad predictiva sobre los demás. Por tanto, nos veíamos en la necesidad de comprobar si la predictibilidad se verá afectada o no por el tipo del coeficiente empleado.

Es de notar que nosotros no somos los primeros en ocuparnos de esta tarea. De hecho, el equipo del proyecto *DEEDS*[282] (*Documents of Early England Data Set*) de la Universidad de Toronto ha desarrollado un sistema estadístico bastante sofisticado para datar documentos (*charters*) medievales ingleses escritos en latín, que dispone como parámetros de todas las *k* palabras consecutivas (k-*shingle*) en el texto (Feuerverger *et al*. 2005, 2008; Fiallos 1997, 2000; Gervers 1997, 2000a, 2000b; Tilahun 2011; Tilahun *et al*. 2012). Ahora bien, la originalidad de nuestra investigación radica en la aproximación tanto filológica como estadística, que, por una parte, se propone establecer parámetros lingüísticos fundamentados en el conocimiento de la gramática histórica y, por otra, aplicarles un tratamiento matemático, pues lo que mayor interés presenta para los investigadores de la diacronía de la lengua es comprobar la posibilidad de realizar datación a partir de los rasgos lingüísticos, a saber, gráfico-fonéticos, morfosintácticos y léxicos, sin que recurramos al hecho extralingüístico anotado en el texto[283]. Por tanto, siendo idéntica la meta final de datación que es fechar correctamente textos antiguos para su indagación en campos relevantes, difiere nuestra aproximación de la del equipo canadiense en que resulta ser más restrictiva por contar solamente con unas 300 características lingüísticas que a día de hoy tenemos establecidas. Otro aspecto que de nuestra datación es digno de mencionar es el uso de la transcripción paleográfica realizada con rigor académico de acuerdo con los criterios establecidos por la red *CHARTA*[284] (*Corpus Hispánico y Americano en la Red: Textos Antiguos*), lo que permite minimizar la posible distorsión textual

282 http://deeds.library.utoronto.ca/

283 Sobre el análisis estadístico de los datos lingüísticos en el ámbito japonés, incluyendo la datación, véanse Jin (2009) y Murakami (1994, 2002 y 2006).

284 http://www.charta.es/criterios-de-edicion-/

por parte de editores modernos, porque incluso la abreviación puede encerrar cierta información sobre la circunstancia en la que se compuso el documento (Ueda 2013a).

Si nuestra datación no proporciona una capacidad predictiva tan elevada como lo hacen los estudios estadísticos del equipo de Toronto, será porque la variación lingüística no es de por sí lo suficientemente informativa como para precisar la procedencia cronológica del texto, lo que no es nada extraño si se piensa en la interdisciplinariedad que se requiere para una empresa tan ambiciosa como es la datación exacta.

Corpus de Documentos de Cancillería Real (CODCAR)

Hemos utilizado como corpus *CODCAR* (Corpus *de Documentos de Cancillería Real*) elaborado por el *GEDHYTAS*[285] (*Grupo de Estudio de Documentos Históricos y Textos Antiguos de la Universidad de Salamanca*), dirigido por la Profa. María Nieves Sánchez González de Herrero. Este corpus se compone de 538 piezas con *data chronica* explícita más una fechada no puntualmente, que se sitúan entre 1223 y 1311, coincidiendo el promedio con la mediana en 1272. La densidad documental, análoga a la popular, es de 6.18 (= 538/(1311-1223-1)), es decir, que se encuentran como promedio seis piezas cada año, aunque la distribución cronológica es desigual con la gran concentración de piezas en la segunda mitad del siglo XIII y menor cantidad en las demás épocas. Respecto a la variación geográfica, salta a la vista la preponderancia de documentos emitidos en Burgos, Sevilla y Valladolid (Figura 1). Es de destacar que el hecho de que todos los textos provinieran de la Cancillería, tal como indica el nombre del corpus, nos permite desatender hasta cierto punto la supuesta influencia ejercida por la variación tipológica.

	Asturias	Ávila	Badajoz	Burgos	Cádiz	Castellón	Ciudad Real	Córdoba	Cuenca	Granada	Guadalajara	Jaén	La Rioja	León	Madrid	Murcia	País Vasco	Palencia	Salamanca	Segovia	Sevilla	Soria	Toledo	Valencia	Valladolid	Vizcaya	Zamora	s.l.	Total
1221~1230																		1		1					1		1		4
1231~1240				7										1											3		1		12
1241~1250				2				2				2									10		1		3				20
1251~1260			2	21				2	1		2	1		2	1	9	2	1		8	51	4	15		16				138
1261~1270			1	2	12		1			2		6	2				2				49		2						79
1271~1280		1		22		1		3	2		3	1	1		1	13	7	4	3	7	7		4	2	15		8	1	106
1281~1290	1	1		22				1	2		3	3	2	3		1	3	5	4	6	6	4	11		15	1	6	1	101
1291~1300				9											2			8					2		27		2	1	51
1301~1310				10							4			1	2			1							7				25
1311~1320																									2				2
Total	1	2	3	95	12	1	1	8	5	2	12	13	5	7	6	23	14	20	7	22	123	8	35	2	89	1	18	3	538

Figura 1. Distribución espacio-temporal de los documentos en el corpus (número de piezas)

285 http://campus.usal.es/~gedhytas/

Coeficientes de asociación (correlación)

Algunos de los coeficientes de asociación más conocidos son *Correspondencia simple*, *Dice-Sorenson*, *Jaccard*, *Hamann*, *Phi*, *Phi modificado*, *Preferencia*, *Russel-Rao*, *Russel-Rao-3* y *Yule*, que se definen en las siguientes fórmulas (Ueda 2013b):

- *Correspondencia simple* = (a + d) / (a + b + c + d)
- *Dice-Sorenson* = 2a / (2a + b + c)
- *Jaccard* = a / (a + b + c)
- *Hamann* = {(a + d) - (b + c)} / {(a + d) + (b + c)}
- *Phi* = (a*d – b*c) / $\{(a + b)(c + d)(a + c)(b + d)\}^{1/2}$
- *Phi modificado* (*Ochiai*) = a / $\{(a + b)(a + c)\}^{1/2}$
- *Preferencia* = {2a – (b + c)} / {2a + (b + c)}
- *Russel-Rao* = a / (a + b + c + d)
- *Russel-Rao-3* = 3a / (3a + b + c + d)
- *Yule* = (a*d – b*c) / (a*d + b*c)

donde *a* representa el número de rasgos positivamente compartidos por A y B, *b* el número de rasgos presentes en A pero no en B, *c* el número de rasgos presentes en B pero no en A, y *d* el número de rasgos ausentes o negativamente compartidos en ambos. Nótese que *Correspondencia simple*, *Dice-Sorenson*, *Jaccard*, *Phi modificado*, *Russel-Rao* y *Russel-Rao-3* toman el valor entre cero y uno, mientras que *Hamann*, *Phi*, *Preferencia* y *Yule* entre menos uno y uno[286].

Cuanto mayor es el valor del coeficiente, o al acercarse a uno, mayor afinidad existe entre los dos, y cuanto menor, o al acercarse a cero o menos uno según el tipo del coeficiente empleado, menor afinidad entre los dos. El valor elevado del coeficiente se consigue cuando hay múltiples rasgos compartidos (*a* y, de ser el caso, *d*) y pocos rasgos no compartidos (*b* y *c*). En el caso que nos ocupa, damos por hecho que dos textos de análoga procedencia espacio-temporal presenten un valor mayor que dos de distinta procedencia crono-geográfica.

A modo de ejemplo, si deseamos saber cuál de los dos documentos B o C se asemeja más a A, compárese el valor del coeficiente de correlación entre A y B (A: B) con el que se da entre A y C (A: C).

---

286 Estos últimos pueden ser convertidos en la escala cero a uno, añadiéndole uno y luego dividiendo por dos, como se puede verificar en $-1 \leq x \leq 1 \leftrightarrow 0 \leq (x + 1)/2 \leq 1$.

		B	
		+	-
A	+	5 (= a)	4 (= b)
	-	3 (= c)	6 (= d)

		C	
		+	-
A	+	5 (= a)	7 (= b)
	-	6 (= c)	6 (= d)

	A: B	A: C
Correspondencia simple	0.611	0.458
Dice-Sorenson	0.588	0.435
Jaccard	0.417	0.278
Hamann	0.222 (0.611)	-0.083 (0.458)
Phi	0.224 (0.612)	-0.084 (0.458)
Phi modificado	0.589	0.435
Preferencia	0.176 (0.588)	-0.13 (0.435)
Russel-Rao	0.278	0.208
Russel-Rao-3	0.536	0.441
Yule	0.429 (0.714)	-0.167 (0.417)

Figura 2. Coeficientes de asociación[287]

Como se puede apreciar en la Figura 2, el uso de un coeficiente u otro no altera el mayor nivel de asociación entre A y B respecto a la que se da entre A y C, sino solamente el valor numérico. Así juzgaremos que el documento A es más afín a B que a C.

---

287 Entre paréntesis aparece el coeficiente convertido en una escala de cero a uno. Nótese la coincidencia entre *Hamann convertido* (*Hamann conv.*) y *Correspondencia simple* y la que se da entre *Preferencia convertida* (*Preferencia conv.*) y *Dice-Sorenson*, algo lógico si se piensa en la manipulación aducida en la nota 286. En lo que sigue, se suprimen *Hamann conv.* y *Preferencia conv.* por *Correspondencia simple* y *Dice-Sorenson*, respectivamente:

$$Hamann\ conv. = \frac{(Hamann + 1)}{2} = \frac{\{(a+d)-(b+c)\}+\{(a+d)+(b+c)\}}{2\{(a+d)+(b+c)\}}$$
$$= \frac{a+d}{(a+d)+(b+c)} = Correspondecia\ simple$$

$$Preferencia\ conv. = \frac{(Preferencia + 1)}{2} = \frac{\{2a-(b+c)\}+\{2a+(b+c)\}}{2\{2a+(b+c)\}}$$
$$= \frac{2a}{2a+(b+c)} = Dice\text{-}Sorenson$$

$$Yule\ conv. = \frac{(Yule + 1)}{2} = \frac{(a*d-b*c)+(a*d+b*c)}{2(a*d+b*c)} = \frac{a*d}{a*d+b*c}$$

$$Phi\ conv. = \frac{(Phi + 1)}{2} = \frac{(a*d-b*c)+\sqrt{(a+b)(c+d)(a+c)(b+d)}}{2\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

El cómputo ha sido realizado por el programa informático NUMEROS.xlsm para el análisis de datos numéricos, elaborado por el Prof. Hiroto Ueda de la Universidad de Tokio, que se puede obtener gratis en su página web http://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo/index.html.

DATACIÓN[288]

La datación se realiza como sigue. En primer lugar, detectamos en cada uno de los documentos la presencia o ausencia de los parámetros lingüísticos establecidos en Kawasaki (en evaluación). En el CODCAR 001, por ejemplo, encontramos los siguientes parámetros en orden de aparición: *sepades* por *sepáis*, *mio* por *mi*, *nengun* por *ningun*, *pora* o *por a* por *para*, *no<n>* por *non* o *no*, *sos* por *sus*, *logar* por *lugar*, *parte* por *part* apocopado, *portadgo* por *portazgo*, los cuales van marcados en negrita:

{1} [lat. Ferr<andus>. d<e>i gr<ati>a Rex Cast<e>lle & Tol<e>ti. om<n>ib<us> hominib<us> regni sui hanc cartam uidentib<us> sal<u>t & {2} gra<tia>m.] **Sepades** que yo falle por pesquisa que **mio** Auuelo mando que **nengun** ganado {3} ni otra cosa **nenguna pora** uender en razon de mercadura **no<n>** passe taio en puente ni {4} en barco fuera por la puente de Toledo. & de Alfariella. & de zorita. Et pues que esta {5} pesquisa fallo yo mando que ni ganado ni otra cosa **nenguna** que **por a** uender sea fuera {6} conducho cada uno **pora** sus c*sas & a **sos** g*nados & **no<n> pora** uender **no<n>** passe taio fuera {7} por estos tres **logares**. & si los fraires lo fallassen en otra **parte** passando mando que lo pre<n>dan {8} por descaminado. Otro si fallo por pesquisa que los de Ocanna de todo lo que passaren {9} an adar **portadgo**. fuera de pan & de uino & de sal q<ue> passen **pora** sus casas & **pora sos** ganados {10} & **no<n> pora** uender. [lat. F<ac>ta carta ap<u>d Collar Reg<e> exp<rimente>.xi. die Julij. ERA M<a>. CC<a>. Lx. p<ri>ma {11} Anno regni mei sexto] (CODCAR 001, Año 1223[289], Cancilleresco, Cuéllar, Segovia, Castilla la Vieja).

La detección automática de los parámetros por medio del programa de nuestra elaboración, se traduce en la matriz de datos en la que '1' y '0' significa presencia y ausencia del parámetro dado respectivamente (Figura 3).

ID	Año	Provincia	Región	Tipología	NINGUN	NENGUN	PORA	PARA	PART	PARTE	NO<N>	NON	NO	LOGAR	LUGAR
CODCAR 001	1223	Segovia	CV	C	0	1	1	0	0	1	1	0	0	1	0
CODCAR 002	1228	Valladolid	CV	C	1	0	0	0	0	0	1	1	0	0	0
CODCAR 003	1230	Palencia	CV	C	0	0	0	0	1	1	0	0	1	0	0
CODCAR 004	1231	Zamora	LE	C	0	1	0	0	0	0	1	0	0	0	0
CODCAR 005	1231	Burgos	CV	C	1	0	0	0	0	0	1	1	0	0	0

Figura 3. Muestra de la matriz de presencia o ausencia de parámetros[290]

[288] Por razones de espacio no abordamos aquí la datación geográfica.

[289] Sustráiganse 38 años a la Era Hispánica para convertirla en la cristiana, de ahí 1261 (M<a>. CC<a>. Lx. p<ri>ma) - 38 = 1223. El comienzo de la Era Hispánica, el 1 del enero del año 38 a. C., correspondería a la conquista y la pacificación de la Península Ibérica por los romanos (Riesco Terrero 2004: 291).

[290] Debido al espacio limitado, presentamos únicamente la parte correspondiente a los primeros cinco documentos de la matriz compuesta en total de unas 540 filas (para documentos) y unas 300 columnas (para parámetros).

De esta matriz creamos otra del coeficiente de asociación que en el presente caso computamos con *Phi modificado* (Figura 4). Ni que decir tiene que la correlación con sí mismo es de uno.

ID	Año	Provincia	Región	Tipología	CODCAR 001	CODCAR 002	CODCAR 003	CODCAR 004	CODCAR 005
CODCAR 001	1223	Segovia	CV	C	1.000	0.136	0.105	0.504	0.096
CODCAR 002	1228	Valladolid	CV	C	0.136	1.000	0.000	0.154	0.471
CODCAR 003	1230	Palencia	CV	C	0.105	0.000	1.000	0.000	0.274
CODCAR 004	1231	Zamora	LE	C	0.504	0.154	0.000	1.000	0.109
CODCAR 005	1231	Burgos	CV	C	0.096	0.471	0.274	0.109	1.000

Figura 4. Muestra de la matriz del coeficiente de asociación Phi modificado

Al ordenar en orden descendente esta matriz de acuerdo con el valor del coeficiente que presenta cada texto respecto al que es objeto de datación, podemos identificar documentos más afines a este. La premisa metodológica es que los documentos que presenten mayor afinidad en el uso lingüístico provengan de una fecha poco alejada. La Figura 5 muestra las diez piezas más parecidas al CODCAR 001.

ID	Año	Provincia	Región	Tipología	CODCAR 001
CODCAR 001	1223	Segovia	CV	C	1.000
CODCAR 026	1247	Sevilla	AN	C	0.612
CODCAR 221	1271	Murcia	MU	C	0.603
CODCAR 088	1255	Burgos	CV	C	0.569
CODCAR 004	1231	Zamora	LE	C	0.504
CODCAR 181	1266	Sevilla	AN	C	0.500
CODCAR 007	1233	Burgos	CV	C	0.471
CODCAR 029	1248	Sevilla	AN	C	0.471
CODCAR 085	1254	Burgos	CV	C	0.458
CODCAR 034	1250	Sevilla	AN	C	0.444
CODCAR 017	1241	Córdoba	AN	C	0.433

Figura 5. Los diez documentos más afines al CODCAR 001

Finalmente, la fecha estimada de un documento se obtiene calculando el promedio ponderado de la *data chronica* de *n* (≥ 1) documento(s) que con este mismo presenta(n) la mayor semejanza, cuya fórmula se expresa de la siguiente manera. Téngase presente que para nuestro método de datación el coeficiente debe tomar valor positivo. En el presente estudio, fijamos *k* (≥ 1) a 1:

$$\sqrt[k]{\sum_{i=1}^{n}\left((fecha_i)^k \times \frac{coeficiente_i}{\sum_{i=1}^{n} coeficiente_i}\right)}$$

Siguiendo esta fórmula (n = 4), el CODCAR 001 resulta fechado en 1252, predicción desacertada a raíz, supuestamente, de un reducido número de parámetros presentes en el texto tan corto como este[291]:

$$1247 \times \frac{0.612}{2.287} + 1271 \times \frac{0.603}{2.287} + 1255 \times \frac{0.569}{2.287} + 1231 \times \frac{0.504}{2.287} \cong 1252$$

Con el programa informático de elaboración propia del autor, hemos datado los 583 documentos en el corpus dando por desconocida la fecha de redacción verdadera. A continuación reproducimos la variabilidad en el promedio (Figura 6), desviación estándar (Figura 7) y mediana (Figura 8) del margen de error absoluto medido por año respecto a la *data chronica* anotada (cuanto menor, mejor) conforme al uso de *n* documentos más semejantes al que intentamos datar. Se aprecia poca sensibilidad al distinto valor de *n* (= 1, 2, 3,..., 10, 15, 20).

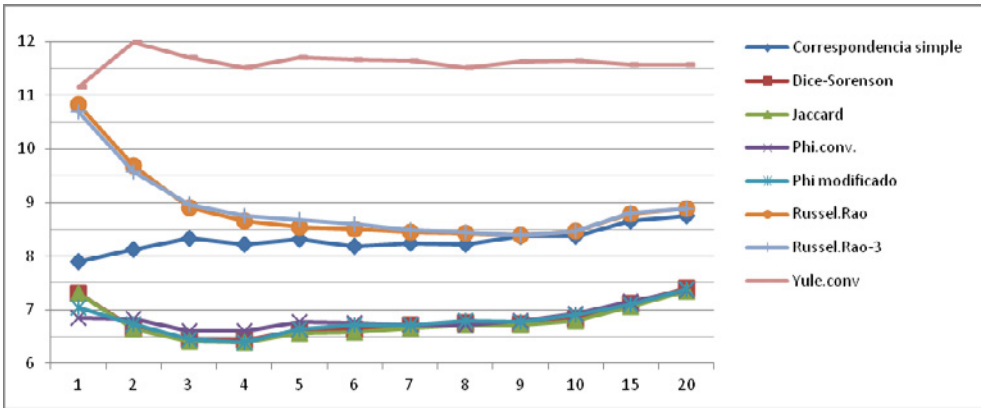


Figura 6. Promedio del margen de error absoluto según el uso de *n* documentos más semejantes

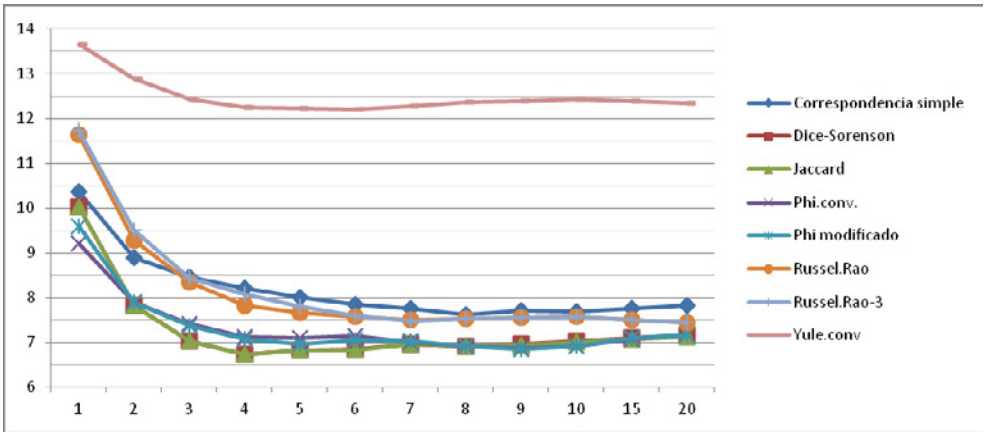


Figura 7. Desviación estándar en el margen de error absoluto según el uso de n documentos más semejantes

291 Existen varios factores desfavorables para la datación precisa, de los que trataremos en un futuro estudio.

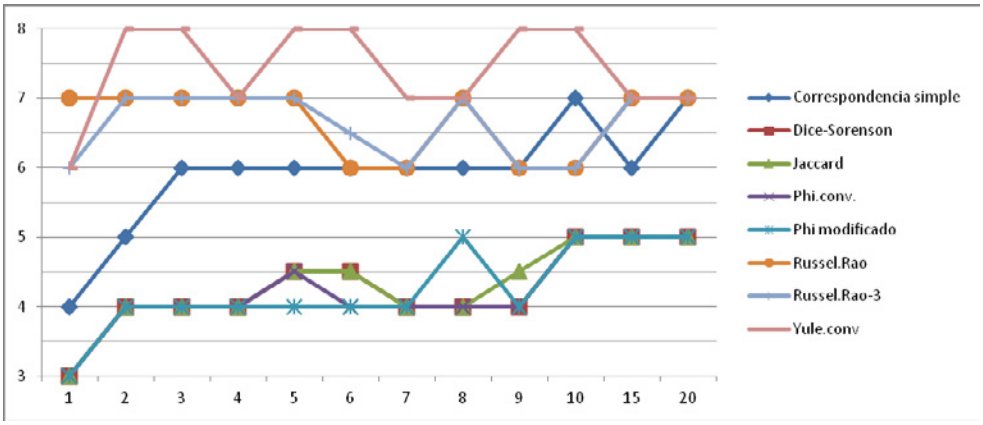


Figura 8. Mediana en el margen de error absoluto según el uso de *n* documentos más semejantes

Como se observa en estas figuras, *Dice-Sorenson*, *Jaccard*, *Phi conv.* y *Phi modificado* presentan un valor inferior a los demás respecto al promedio, desviación estándar y mediana del margen de error absoluto, de ahí que resulten igualmente más apropiados para nuestro objetivo que es la datación. Conviene hacer notar que el valor mayor o menor del margen de error absoluto se debe únicamente al distinto valor proporcionado por cada coeficiente (o coeficiente$_i$), porque, como hemos señalado con anterioridad, los *n* documentos más análogos al documento en cuestión (o fecha$_i$) siguen siendo idénticos. Además, se puede apreciar que cuando *n* es 4 obtenemos el rendimiento mayor.

En la Figura 9 reproducimos la predecibilidad de la datación (*n* = 4). Por razones de espacio, nos limitamos a *Dice-Sorenson*, *Jaccard*, *Phi conv*. y *Phi modificado*, que proporcionan la mayor capacidad predictiva.

	Dice-Sorensen		Jaccard		Phi conv.		**Phi modificado**	
Error: ±5	312/538	58.0%	312/538	58.0%	311/538	57.8%	**324/538**	**60.2%**
Error: ±10	419/538	77.9%	417/538	77.5%	412/538	76.6%	**421/538**	**78.3%**
Error: ±20	512/538	95.2%	512/538	95.2%	510/538	94.8%	**509/538**	**94.6%**
Error: ±30	535/538	99.4%	535/538	99.4%	534/538	99.3%	**533/538**	**99.1%**
Error: ±40	537/538	99.8%	537/538	99.8%	537/538	99.8%	**537/538**	**99.8%**
Error: ±50	538/538	100.0%	538/538	100.0%	538/538	100.0%	**538/538**	**100.0%**
Error: ±100	538/538	100.0%	538/538	100.0%	538/538	100.0%	**538/538**	**100.0%**
Promedio	6.43		6.40		6.61		**6.39**	
Desviación estándar	6.74		6.73		7.12		**7.08**	
Mediana	4.00		4.00		4.00		**4.00**	
Max. (Error)	44		44		43		**43**	
Min. (Error)	0		0		0		**0**	

Figura 9. Predictibilidad por distintos coeficientes de asociación (*n* = 4)

Con estos coeficientes, cerca del 60% de los documentos fueron datados con un margen de error de ±5 años respecto a la fecha verdadera, más del 75% con un margen de error de ±10 años y casi el 95% con un margen de error de ±20 años. Si tomamos como punto de referencia el promedio del margen de error absoluto o bien el número de textos correctamente datados con un margen de error de ±5 o 10 años, *Phi modificado* se considera el más eficaz en este caso, aunque los otros tres le son prácticamente comparables.

En lo que sigue analizaremos con detenimiento la capacidad predictiva de *Phi modificado*. La Figura 10 revela de nuevo la poca influencia que sobre la precisión ejerce el uso de *n* documentos más semejantes, con el resultado de que, en las líneas generales, la mayor predictibilidad se obtiene cuando *n* es 4.

	n = 1		n = 2		n = 3		**n = 4**		n = 5		n = 6	
Error: ±5	328/538	61.0%	318/538	59.1%	319/538	59.3%	**324/538**	**60.2%**	311/538	57.8%	318/538	59.1%
Error: ±10	410/538	76.2%	417/538	77.5%	425/538	79.0%	**421/538**	**78.3%**	409/538	76.0%	412/538	76.6%
Error: ±20	475/538	88.3%	501/538	93.1%	501/538	93.1%	**509/538**	**94.6%**	511/538	95.0%	509/538	94.6%
Error: ±30	522/538	97.0%	527/538	98.0%	531/538	98.7%	**533/538**	**99.1%**	532/538	98.9%	534/538	99.3%
Error: ±40	535/538	99.4%	535/538	99.4%	537/538	99.8%	**537/538**	**99.8%**	536/538	99.6%	536/538	99.6%
Error: ±50	536/538	99.6%	538/538	100.0%	538/538	100.0%	**538/538**	**100.0%**	538/538	100.0%	538/538	100.0%
Error: ±100	538/538	100.0%	538/538	100.0%	538/538	100.0%	**538/538**	**100.0%**	538/538	100.0%	538/538	100.0%

	n = 7		n = 8		n = 9		n = 10		n = 15		n = 20	
Error: ±5	313/538	58.2%	307/538	57.1%	299/538	55.6%	294/538	54.6%	290/538	53.9%	278/538	51.7%
Error: ±10	408/538	75.8%	404/538	75.1%	408/538	75.8%	401/538	74.5%	404/538	75.1%	401/538	74.5%
Error: ±20	510/538	94.8%	510/538	94.8%	513/538	95.4%	511/538	95.0%	506/538	94.1%	505/538	93.9%
Error: ±30	533/538	99.1%	533/538	99.1%	533/538	99.1%	534/538	99.3%	532/538	98.9%	532/538	98.9%
Error: ±40	537/538	99.8%	537/538	99.8%	537/538	99.8%	537/538	99.8%	536/538	99.6%	535/538	99.4%
Error: ±50	538/538	100.0%	538/538	100.0%	538/538	100.0%	538/538	100.0%	538/538	100.0%	538/538	100.0%
Error: ±100	538/538	100.0%	538/538	100.0%	538/538	100.0%	538/538	100.0%	538/538	100.0%	538/538	100.0%

Figura 10. Predictibilidad de *Phi modificado* según el uso de *n* documentos más semejantes

Ahora bien, la Figura 11 ilustra gráficamente la correspondencia entre la fecha estimada y la fecha verdadera de los 538 documentos datados con *Phi modificado*. La gran mayoría se sitúa exitosamente sobre o poco alejada de la línea y = x. Sin embargo, observamos una marcada tendencia a la sobrestimación respecto a los documentos de composición temprana y, por el contrario, a la baja estimación respecto a los de composición tardía, situación que ha sido causada por la naturaleza de extremidad (*the one-sided nature of nearest neighbors at the edges*) que entraña este método (Tilahun *et al.* 2011). En realidad, existe una fuerte correlación negativa (-0.626) entre la fecha verdadera y el margen de error (Figura 12).

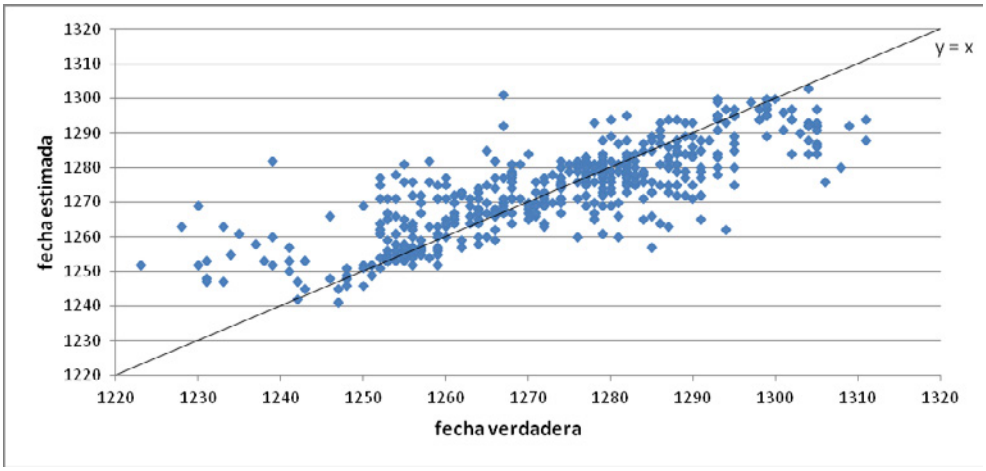


Figura 11. La fecha estimada contra la fecha verdadera

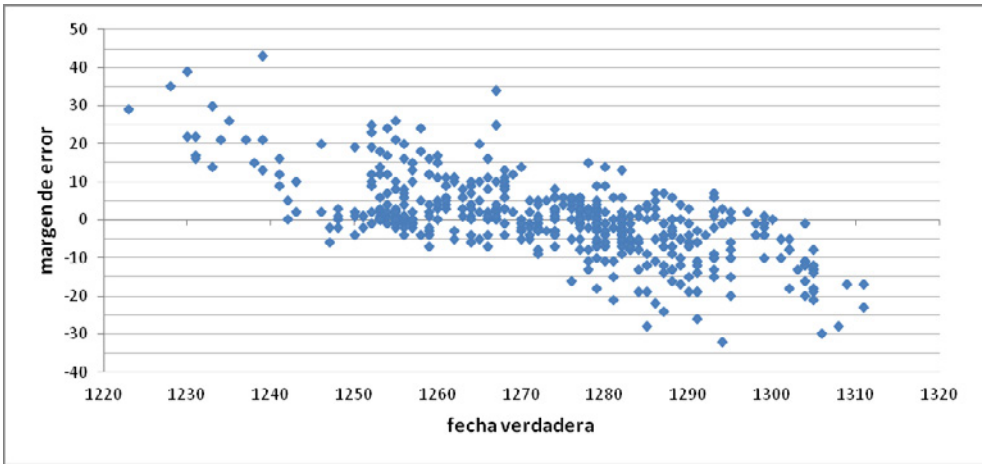


Figura 12. El margen de error contra la fecha verdadera

Por último, no olvidemos señalar otra fuerte correlación negativa (-0.751) que se da entre el número de piezas existentes y el promedio del margen de error absoluto en el periodo dado (Figura 13).

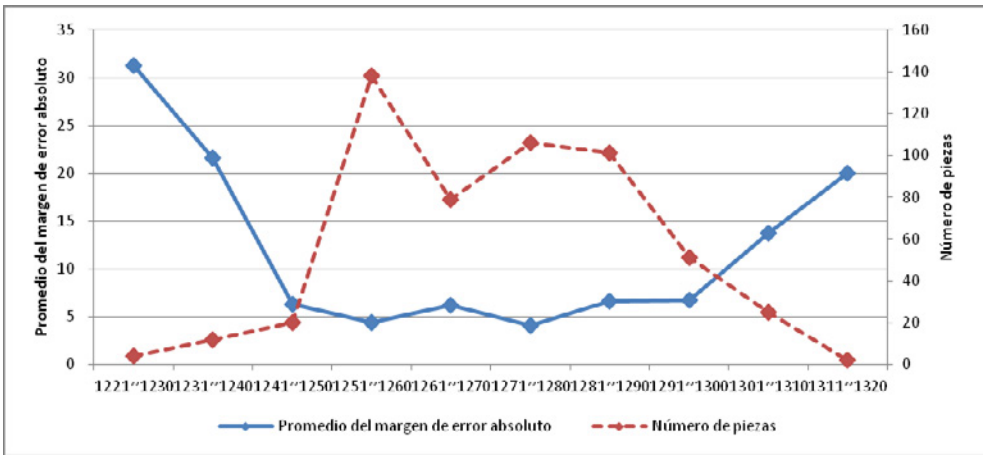


Figura 13. Relación entre el promedio del margen de error absoluto y el número de piezas en cada periodo

Terminaremos este apartado con la datación del único documento sin *data chronica* explícita en el corpus, CODCAR 521, en el que hallamos los siguientes parámetros en orden de aparición: *Castill<a>* por *Castiella*, *en<n>a* por *en la*, *onor* por *honor*, *sal<u>t* por *salud*, *conçejo* por *conçello*, *om<n>es* por *hombres*, *t<ie>rra* por *tierra*, *oujero<n>* por *hubieron*, *costunbrado* por *costumbrado*, *juyzo* por *juizio*, *juyz* por *juez*, *agora* por *ahora*, *mie<n>tre* por *mente*, *lles* por *les*, *non* por *no<n>* o *no*, *ffazer* por *fazer* o *hazer*, *ssin* por *sin*, *derech<o>* por *dereito*, *rreys* por *rreyes*, *fforo<n>* por *fueron*, *ssos* por *sus*, *ssodes* por *sois*, *pedio* por *pidió*, *merçed* por *merçet*, *toujesse* por *tuujesse*, *onde*, *assi* por *así*, *husado* por *usado*, *ffasta* por *fasta* o *hasta*, *ffeziestes* por *fizisteis*, *diz<er>* por *dezir*, *conçello* por *conçejo*, *uirdes* por *uieredes* y *dies* por *días*, los cuales van marcados en negrita:

{1} Don Alffon<so>. por la gr<aci>a de dios Rey de **Castill<a>** de Leo<n> ** {2} Toledo de Gallizia. de Seuill<a>. de Cordoua de Murç<ia> de {3} Jah<e>n. & de Lalgarb<e>. Auos Pedro pel<ae>z m<er>jno. por don ma<n>rriq<ue> gil {4} **en<n>a onor** de uillar mazari**. **sal<u>t** & gr<aci>a. alffon<so>. m<artine>z. p<er>son<er>o del {5} **Conçejo** de Leo<n>. se me q<ue>rello * ***. q<ue>los **om<n>es** desa **t<ie>rra ouje**{6}**ro<n>** & an usado. & **costunbrado** de uenir A**juyzo** Ante los **Juyzes** de {7} Leo<n>. por lliurar sus pl<e>itos; & q<ue> uos **Agora** nueua **mie<n>tre** q<ue>les {8} cotades & **lles** deffendedes q<ue> **non** uayan AAuer **Juyzo**. y. Et {9} esto q<ue>lo **ffazedes ssin** rrazo<n> & **ssin derech<o>.** & cont<ra> ffuero & cont<ra> pri{10}uilegios q<ue> an delos **rreys** q<ue> **fforo<n>** An[* **] mj & co<n> ffirmados por {11} mj. & q<ue> p<ri>ndades alos **ssos** uezinos. por q<ue> traen Armas por aq<ue>lla {12} **t<ie>rra**. donde uos **ssodes** merino. & q<ue>**lles ffazedes** o tros tuertos & ma{13}les como **no<n>** deuedes. Et **pedio** me **merçed** q<ue> mandasse y. lo q<ue> **touj**{14}**esse** por bien. **Onde** uos ma<n>do. uista esta mj carta. q<ue> dexedes al{15}os **om<n>es** dela **t<ie>rra**. yr A**juyzo** Ante los **juyzes** de Leo<n>. **Assi** como {16} **oujero<n> husado ffasta** Aq<ui>. & q<ue>**lles** en me<n>dedes las fforçias & los {17} tuertos q<ue> dizen q<ue>**lles**

ffeziestes. P<er>o si cont<ra> esto algo q<ui>sieren **diz<er>** {18} o rrazonar uenit o en ujat u<uest>ro. p<er>son<er>o. co<n> el p<er>son<er>o del **co<n>çello** {19} [* * *] mj. del dia q<ue> esta mj carta **uirdes** A q<ui>nze **dies**. & yo oyre {20} las [* * *] delos tres & mandare aq<ue>llo q<ue> **touier** por bien. (CODCAR 521, Año 1255-1284, Cancilleresco, s.l.).

Con *Phi modificado* (*n* = 4), se le ha atribuido 1278 (Figura 14), fecha que va en consonancia con la asignada por las circunstancias extralingüísticas o los años de reinado de Alfonso X, 1255-1284.

ID	Año	Provincia	Región	Tipología	CODCAR 521
CODCAR 521	s.a.	s.l.	s.l.	C	1.000
CODCAR 242	1274	Zamora	LE	C	0.686
CODCAR 202	1268	Sevilla	AN	C	0.629
CODCAR 343	1287	Burgos	CV	C	0.597
CODCAR 335	1286	León	LE	C	0.583

Figura 14. Los cuatro documentos más afines al CODCAR 521

FUTURAS INVESTIGACIONES

Por el momento tenemos previsto introducir un par de nuevos tratamientos matemáticos. Uno consistirá en cuantificar el grado de precisión en cuanto a la fecha estimada, pues algunos textos resultan más difíciles de fechar a raíz de su poca extensión, un reducido número de parámetros presentes, poca concentración de documentos en algunas épocas y la intervención de más de un escribano, entre otros motivos[292]. El otro será asignar un peso distinto a los parámetros conforme a la importancia que poseen, lo cual se hace a través del cálculo de la matriz ponderado a partir del patrón de la distribución de los parámetros.

BIBLIOGRAFÍA

CHARTA (*Corpus Hispánico y Americano en la Red: Textos Antiguos*): http://www.charta.es/criterios-de-edicion-/

DEEDS (*Documents of Early England Data Set*): http://deeds.library.utoronto.ca/

Feuerverger, Andrey, Peter Hall, Gelila Tilahun y Michael Gervers. 2005. Distance measures and smoothing methodology for imputing features of documents. *Journal of Computational and Graphical Statistics*, 14 (2): 255-262.

Feuerverger, Andrey, Peter Hall, Gelila Tilahun y Michael Gervers. 2008. Using statistical smoothing to date medieval manuscripts. En N. Balakrishnan, E. Pena, & M. J. Silvapulle (Eds.), *Beyond parametrics in interdisciplinary research: Festshrift in honor of professor Pranab K. Sen*, Vol. 1: 321-331.

292 Debemos al Prof. Hiroto Ueda la valiosa sugerencia al respecto.

Fiallos, Rodolfo. 1997. Procedure for dating undated documents using a relational database. En J. Brown, & W. P. Stoneman (Eds.), *A distinct voice: medieval studies in honor of Leonard E. Boyle*, O. P. Notre Dame (Indiana): 480-504.

— 2000. An overview of the process of dating undated medieval charters: latest results and future developments. En M. Gervers (Ed.), *Dating undated medieval documents*. Suffolk: Boydell & Brewer: 37-48.

Gervers, Michael. 1997. The dating of medieval English private charters of the twelfth and thirteenth centuries. En J. Brown, & W. P. Stoneman (Eds.), *A distinct voice: medieval studies in honor of Leonard E. Boyle*, O. P. Notre Dame (Indiana): 455-480.

Gervers, Michael. (Ed.). 2000a. *Dating undated medieval charters*. Suffolk: Boydell & Brewer.

Gervers, Michael. 2000b. The DEEDS project and the development of a computerised methodology for dating undated English private charters of the twelfth and thirteenth centuries. En M. Gervers, *Dating undated medieval charters*. Suffolk: Boydell & Brewer: 13-35.

Jin, Mingzhe. 2009. *Tekisuto-deeta no toukei-kagaku nyuumon* (*Introducción al análisis estadístico de los textos*). Tokio: Iwanami shoten. (en japonés).

Kawasaki, Yoshifumi. En evaluación. Datación crono-geográfica de documentos notariales medievales, *Scriptum digital*. http://scriptumdigital.org/

Murakami, Masakatsu. 1994. *Shingan no kagaku: Keiryoo-bunkengaku nyuumon* (*Introducción a la estilometría*). Tokio: Asakura shoten. (en japonés).

— 2002. *Bunka wo hakaru: bunka-keiryoogaku zyosetsu* (*Introducción a la culturametría*). Tokio: Asakura shoten. (en japonés).

— (Ed.). 2006. *Bunka-zyoohoogaku nyuumon* (*Culture and information science*). Tokio: Bensei shuppan. (en japonés).

Riesco Terrero, Ángel. (2004). *Introducción a la paleografía y la diplomática general*. Madrid: Editorial Síntesis.

Sánchez González de Herrero, María Nieves (dir.). *CODCAR* (*Corpus de documentos de Cancillería Real*). http://campus.usal.es/~gedhytas/

Tilahun, Gelila. 2011. *Statistical methods for dating collections of historical documents*. Tesis de doctorado, Universidad de Toronto. Inédita.

Tilahun, Gelila, Andrey Feuerverger y Michael Gervers. (2012). Dating medieval English charters. *The Annnals of applied statistics*, Vol. 6 (4): 1615-1640.

Ueda, Hiroto. 2013a. Pautas y frecuencias grafotácticas de formas abreviadas: Su utilización para la datación de los documentos notariales del siglo XIII al XVII, Comunicación oral en el *III Congreso Internacional Tradición e Innovación: nuevas perspectivas para la edición, la investigación y el estudio de documentos antiguos*, Salamanca, 5-7 de junio de 2013.

— 2013b. Una nota sobre el método de taxonomía cuantitativa de grandes datos: Coeficientes de asociación aplicados a las variantes del *Diccionario de americanismos*, *Dialectología*. Special issue, IV: 221-235.

Ueda, Hiroto. NUMEROS.xlsm (programa informático para análisis de datos numéricos), http://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo/index.html.