Grandes rasgos históricos de grafías españolas
Métodos de cuantificación y visualización de frecuencias de los datos de ALDICAM

Hiroto Ueda

Universidad de Tokio

1. INTRODUCCIÓN[1]

En enero de 2016 se inició un importante proyecto de investigación sobre la historia de la lengua española basada en los documentos emitidos en distintas localidades de Madrid. Se trata del proyecto ALDICAM («Atlas Lingüístico Diacrónico e Interactivo de la Comunidad de Madrid»)[2], en el que colaboramos en la parte informática de procesamiento de datos lingüísticos. En esta ocasión, brindada por el Instituto Cervantes de Tokio, vamos a explicar un estudio dedicado a los cambios históricos de las formas léxicas españolas observadas en los cien documentos preparados hasta el momento.

En la página web citada se ofrecen informaciones del proyecto ALDICAM divididas en las siguientes secciones: Qué es ALDICAM-CM / El equipo / Archivos utilizados / Historias en los documentos / Divulgación / Novedades / El proyecto día a día / Formación / Materiales de trabajo / Producción científica. Según ellas, el objetivo principal del proyecto es: «elaborar un mapa diacrónico dinámico de las particularidades lingüísticas de la Comunidad de Madrid. Su forma avanzada de visualización, al proyectar directamente sobre un mapa los resultados de las búsquedas, permite hacerse una idea inmediata del peso del factor geográfico en la variación lingüística histórica en dicho territorio desde el siglo XIII al XIX, ambos incluidos; y no solo para el léxico, sino para cualquier variante gráfica, fonética, morfosintáctica y léxica susceptible de búsqueda en un corpus que a tal efecto se formará». El proyecto es «absolutamente innovador» en el sentido de que

[1] Agradezco la ayuda prestada por Leyre Martín Aizpuru, Antonio Moreno Sandoval y Pedro Sánchez Prieto, en la preparación de este estudio. Este trabajo ha sido subvencionado para los proyectos de investigación: «Atlas Lingüístico Diacrónico y Dinámico de la Comunidad de Madrid» (ALDICAM-CM, Referencia H2015/HUM-3443, director Pedro Sánchez Prieto) y «Cronología relativa de los documentos antiguos españoles» (JSPS KAKENHI: 16K02657, director Hiroto Ueda).

[2] http://aldicam.blogspot.com/ [2 de octubre, 2018].

«nunca hasta ahora se había llevado a cabo un desarrollo semejante para ninguna otra lengua».



Fig. 1.1. http://aldicam.blogspot.com/

Su punto de partida es «la transcripción de un número importante de documentos elaborados (...) en diferentes localidades de la actual Comunidad de Madrid, procedentes del Archivo Regional de la CM, del de la Villa y de los diferentes archivos municipales, de instituciones eclesiásticas e incluso de particulares». Se ofrecen «(a) transcripción paleográfica, en la que se reflejan las grafías del documento;» y «(b) presentación crítica o interpretativa», en que se regularizan «diferencias gráficas no fonológicas» con «la unión y separación de palabras y puntuación; y «(c) imagen del documento».

Se aporta así un material de gran interés histórico, que contribuirá decisivamente a situar el habla de Madrid en un lugar relevante en la historia de la lengua española, de acuerdo con la hipótesis de que «el habla de Madrid contribuyó decisivamente a la forja del español moderno, cosa que los estudios realizados hasta ahora no han conseguido probar por acudir normalmente a textos literarios, más estandarizados y menos sensibles al componente geográfico que las fuentes documentales».

Antes de finalizar el proyecto trienal (2016 - 2019), ya contamos con

estudios en aspectos fonético, sintáctico y léxico en los tres trabajos de Sánchez-Prieto Borja y Vázquez Balonga (2016, 2018a, 2018b). Almeida Cabrejas (2016) se enfoca en los documentos escritos por *scriptores* con bajo nivel socioeducativos en Madrid del siglo XIX, época de la llegada de la ortografía académica a las escuelas. Por nuestra parte, en el arduo procesamiento de lematización de textos paleográficos y críticos[3], que es un trabajo manual y semiautomático de agrupamiento de distintas formas flexionadas en una forma representativa, hemos concebido la idea de ver todas las diferencias que hay entre las formas léxicas paleográficas y las modernas, bastante parecidas a las formas de la versión crítica, para seguir la cronología de cambios a lo largo de tiempo.

Aprovechamos la ocasión para presentar algunas funciones desarrolladas por nosotros en el conjunto de programas de análisis lingüísticos y estadísticos denominado LYNEAL («Letras y Números en Análisis Lingüísticos»), ahora disponible y abierto a todos los interesados[4]:

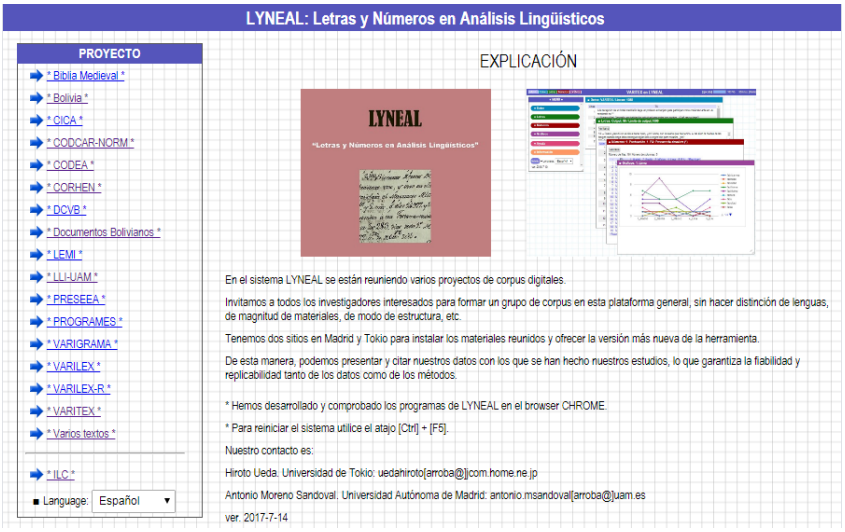


Fig. 1.2. http://shimoda.lllf.uam.es/ueda/lyneal/

[3] El trabajo de lematización ha sido realizado en colaboración con Pedro Sánchez-Prieto.

[4] En Tokio: https://lecture.ecc.u-tokyo.ac.jp/~cueda/lyneal/ [2 de octubre, 2018]. En Madrid: http://shimoda.lllf.uam.es/ueda/lyneal/ [2 de octubre, 2018].

2. PREPARACIÓN DE DATOS

Para explicar nuestro método de aparejamiento de formas paleográficas con las modernas, nos permitimos utilizar la foto y textos paleográfico y crítico que ofrece la página web del proyecto ALDICAM, citada en la sección anterior. La misma página presenta la siguiente foto:

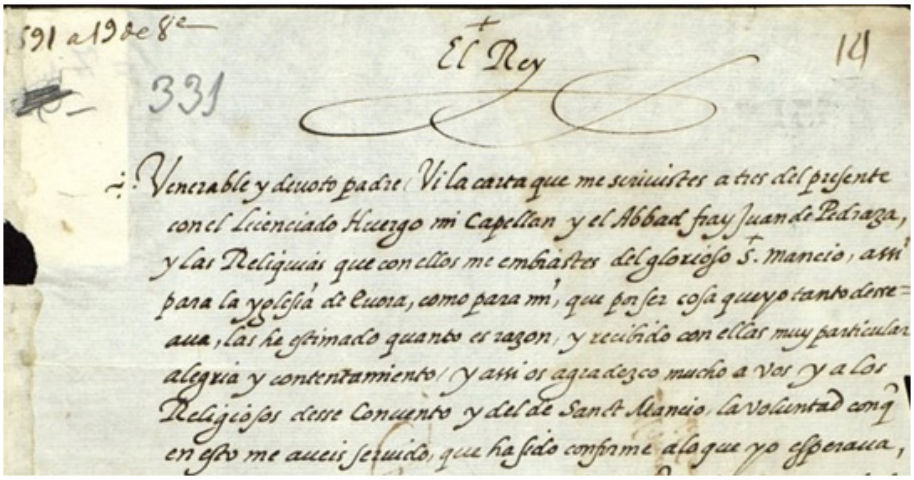


Fig. 2.1. Carta de agradecimiento de Felipe II,
San Lorenzo de El Escorial, 1591

La foto está acompañada de la cabecera del documento:

Origen: Archivo Histórico Nacional | Clero, León, carpeta 946, nº 3

Fecha: 19/10/1591 (España, Madrid, San Lorenzo de El Escorial, —)

Letra: Humanística cursiva

Regesto: Carta de agradecimiento de Felipe II al licenciado Huergo y al abad fray Juan de Pedraza por las reliquias de San Mancio que le han enviado.

Transcripción: Rocío Díaz Moreno / Bautista Horcajada Diezma / María Jesús Torrens Álvarez

Seguidamente veamos su transcripción paleográfica[5]:

[encabezamiento: El Rey]
{1} Venerable y **deuoto** padre, Vi la carta que me **scriuistes** a tres del presente
{2} con el Licenciado Huergo mi Capellan y el **Abbad** fray Juan de Pedraza,
{3} y las Reliquias que con ellos me **embiastes** del glorioso **Sanct.** mancio**, assi**
{4} para la **yglesia** de **Euora**, como para **mi**, que por ser cosa **que-yo** tanto **d-esse+**
{5}+**aua**, las **he** estimado **quanto** es **razon**, y recibido con ellas muy particular
{6} alegria y contentamiento, y **assi** os agradezco mucho a vos y a los
{7} **Religiosos d-esse Conuento** y del de **Sanct** Mancio, la voluntad **con-que**
{8} en esto me **aueis seruido**, que ha sido conforme a lo que yo **esperaua**.

Fig. 2.2. Transcripción paleográfica

Se destacan las formas marcadas en el texto anterior en comparación con la siguiente presentación crítica:

{h. 1r} El Rey.
{1} Venerable y **devoto** padre, vi la carta que me **escrivistes** a tres del presente
{2} con el licenciado Huergo, mi capellán, y el **abad** fray Juan de Pedraza,
{3} y las reliquias que con ellos me **embiastes** del glorioso **Sant** Mancio, **assí**
{4} para la **iglesia** de Évora como para **mí**, que por ser cosa **que yo** tanto **desseava**
{5} las é estimado **cuanto** es **razón** y recibido con ellas muy particular
{6} alegría y contentamiento, y **assí** os agradezco mucho a vós y a los
{7} **religiosos d»esse** convento y del de **Sant** Mancio la voluntad **con que**
{8} en esto me **avéis servido**, que á sido conforme a lo que yo **esperava**.

Fig. 2.3. Presentación crítica

[5] El cambio de línea y letras negritas y subrayadas son nuestros. Hemos marcado las palabras que no coinciden perfectamente con las formas que vienen en la presentación crítica. La parte subrayadas indican las diferencias concretas.

Estos dos textos son útiles para elaborar una edición moderna con la ortografía académica actual[6]:

{h. 1r} El Rey.
{1} Venerable y **devoto** padre, vi la carta que me **escribisteis** a tres del presente
{2} con el licenciado Huergo, mi capellán, y el **abad** fray Juan de Pedraza,
{3} y las reliquias que con ellos me **enviasteis** del glorioso **San** Mancio, **así**
{4} para la **iglesia** de Évora como para **mí**, que por ser cosa **que yo** tanto **deseaba**
{5} las **he** estimado **cuanto** es **razón** y recibido con ellas muy particular
{6} alegría y contentamiento, y **así** os agradezco mucho a vos y a los
{7} **religiosos de ese convento** y del de **San** Mancio la voluntad **con que**
{8} en esto me **habéis servido**, que ha sido conforme a lo que yo **esperaba**.

Fig. 2.4. Edición moderna

A modo de ejemplo de correspondencia de las tres formas, paleográfica, crítica y moderna, preparamos una lista parcial de comparación:

Forma paleográfica	Forma crítica	Forma moderna
deuoto	*devoto*	*devoto*
scriuistes	*escrivistes*	*escribisteis*
Abbad	*abad*	*abad*
embiastes	*embiastes*	*enviasteis*
Sanct	*Sant*	*San*
assi	*assí*	*así*

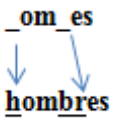
Fig. 2.5. Ejemplos de forma paleográfica, crítica y moderna

[6] Se trata de formas léxicas, que no hacemos la adaptación sintáctica ni léxica al español actual.

Para detectar los cambios formales entre la palabra paleográfica y la moderna, hemos elaborado un programa de función, Diff (*x, y*), que al recibir las dos formas, *x* y *y*, devuelve la diferencia que hay entre ellas:

Diff(«ome», «hombre») → «- > h-br-»

lo que significa que la diferencia entre «ome» y «hombre» está en «- > h-br-»:



Este procesamiento se ha efectuado en todas las palabras separadas por un espacio en la presentación crítica. El cuadro siguiente es una parte inicial del primer documento de ALDICAM[7]:

Forma paleográfica	Forma crítica	Forma moderna	Lema	C.G.	Dif.PM
Conoscuda	*Conoçuda*	*conocida*	*conocer*	**vb.**	**-s-u- > -i-**
cosa	*cosa*	*cosa*	*cosa*	**sus.**	**-**
sea	*sea*	*sea*	*ser*	**vb.**	**-**
a	*a*	*a*	*a*	**prep.**	**-**
todos	*todos*	*todos*	*todo*	**adj.**	**-**
los	*los*	*los*	*el*	**art.**	**-**
omes	*omes*	*hombres*	*hombre*	**sus.**	**- > h-br-**
que	*que*	*que*	*que*	**conj. relat.**	**-**
esta	*esta*	*esta*	*este*	**demos.**	**-**
carta	*carta*	*carta*	*carta*	**sus.**	**-**
uieren	*vieren*	*vieren*	*ver*	**vb.**	**u- > v-**

C.G.: Categoría gramatical;
Dif.PM: Diferencia entre la forma paleográfica y la moderna

Fig. 2.6. parte inicial del primer documento de ALDICAM

[7] 1250 octubre 4 (s.l. [Madrid]) / AHN, Ordenes Militares, carpeta 86, nº 7/ Carta por la que Rodrigo Yeñéguez, maestre de la orden de la caballería de Santiago, y Rui Bueso, comendador, conceden a Pedro Ruiz durante su vida una heredad en Val del Puerco. Transcripción: Rocío Díaz Moreno / Pedro Sánchez-Prieto Borja.

En los cien documentos actualmente preparados, se cuentan 43 420 palabras separadas por espacio, de las cuales 35 517 (81.8 %) son idénticas entre la transcripción paleográfica y la presentación crítica, mientras que 7903 (18.2 %) formas presentan algunas diferencias gráficas entre ellas.

3. ANÁLISIS CUANTITATIVO

3.1. Ratio de variación

Los lemas que no modifican su forma gráfica y que presentan más de 100 ocurrencias en 100 documentos son: *el* (art.), *de* (prep.), *en* (prep.), *a* (prep.), *se* (pron. clit.), *por* (prep.), *su* (poses.), *este* (demos.), *con* (prep.), *para* (prep.), *él* (pron. pers.), *otro* (indef.), *todo* (adj.), *como* (conj.), *día* (sus.), *dos* (num.), *don* (sus.), *más* (adv.), *poder* (sus.), *tierra* (sus.), *parte* (sus.), en orden descendente de frecuencia. Estas palabras no cambian de forma gráfica más de 5 % (frecuencia de forma desigual / frecuencia total * 100)[8].

En cambio, contamos con los 17 lemas muy frecuentes que cuentan con más de un 20 % de variación. Para extraer estos lemas variantes hemos utilizado la ratio de variación (RV) en forma de:

RV = Frecuencia de forma antigua variante / Frecuencia total * 100

Por ejemplo, el lema «y» posee las formas antiguas <&> y <e> (frecuencia: 862) y la forma moderna <y> (2182), en total, 3044. De acuerdo con nuestro interés, la variación, nos enfocamos en las formas antiguas variantes <&> y <e> (frecuencia: 862) dentro de la totalidad de frecuencia (3044): RV = 862 / 3044 * 100 = 28.3 (%). La siguiente tabla representa la ratio de variación (RV) de los 17 lemas variantes más frecuentes, con más de 100 ocurrencias:

N	Lema	Forma antigua	F.a.	Forma moderna	F.m.	Total	RV (%)
1	*y (conj.)*	*<&>, <e>*	862	*<y>*	2 182	3 044	28.3
2	*un (art.)*	*<v>n*	234	*<u>n*	196	430	54.4

[8] Se trata de formas gráficas. Las mismas formas pueden variar en el uso de mayúsculas, acento, formas abreviada y unión y separación de palabras, que merecen un estudio independiente.

3	*villa (sus.)*	*<u>illa*	86	*<v>illa*	244	330	26.1
4	*haber (vb.)*	*<_>aber*	230	*<h>aber*	95	325	70.8
5	*hacer (vb.)*	*<f>acer*	137	*<h>acer*	103	240	57.1
6	*señor (sus.)*	*se<nn>or*	67	*se<ñ>or*	132	199	33.7
7	*año (sus.)*	*a<nn>o*	41	*a<ñ>o*	137	178	23.0
8	*Juan (n. prop.)*	*J<o>an*	35	*J<u>an*	134	169	20.7
9	*vecino (sus.)*	*ve<z>ino*	116	*ve<c>ino*	49	165	70.3
10	*mil (num.)*	*mi<ll>*	95	*mi<l>*	35	130	73.1
11	*cual (relat.)*	*<q>ual*	111	*<c>ual*	17	128	86.7
12	*mi (poses.)*	*m<j>*	26	*m<i>*	102	128	20.3
13	*saber (vb.)*	*sa<u>er*	43	*sa<b>er*	73	116	37.1
14	*escribano (sus.)*	*escri<u>ano*	92	*escri<b>ano*	19	111	82.9
15	*bien (adv.)*	*<v>ien*	30	*<b>ien*	79	109	27.5
16	*merced (sus.)*	*mer<z>ed*	30	*mer<c>ed*	75	105	28.6
17	*ver (vb.)*	*<b>er*	40	*<v>er*	61	101	39.6

F.a.: Frecuencia de forma antigua / F.m.: Frecuencia de forma moderna
RV: ratio de variación

Fig. 3.1. Lemas variantes más frecuentes

Nos interesa seguir la historia de estos lemas en todos los documentos reunidos y transcritos por el equipo del Proyecto.

3.2. Frecuencia normalizada

Nuestro propósito actual es buscar las diferencias entre las formas paleográficas y las modernas, numéricamente más significativas. Para el mismo propósito, creemos conveniente tomar en consideración la frecuencia media en distintas franjas cronológicas, divididas en 50 años. Naturalmente cuanto más se presenten estas diferencias, más significativas serán estadísticamente. En este sentido, la primera diferencia más numérica, con la frecuencia total de 483, entre el antiguo signo tironiano en forma parecida a la letra griega tau (τ) transcrito en «&» y la forma moderna «y» es mucho más importante que, por ejemplo, la forma antigua *sáuado* correspondiente a *sábado*, cuya frecuencia es 1.

Antes de buscar directamente las diferencias que más frecuencia tienen, hay que cambiar las frecuencias absolutas en las frecuencias normalizadas por el número total de formas en cada franja cronológica, puesto que la frecuencia total varía de manera pronunciada, como vemos en la siguiente tabla[9]:

Año	1400	1450	1500	1550	1600	1650	1700	1750	1800	1850
F. total	857	1 484	6 876	6 798	1 379	9 719	1 783	8 597	3 223	2 026

Fig. 3.2.1. Frecuencia total de cada franja cronológica

La frecuencia normalizada (FN) se calcula de la siguiente manera:

FN = FA / Frecuencia total * 1000

En este caso conviene utilizar el multiplicador 1000, en lugar de 100 (porcentaje), 10 000, 100 000, etc. puesto que 1000 se aproxima al mínimo de la frecuencia total (857). La siguiente tabla ofrece las frecuencias absolutas de las formas de la conjunción «y»:

Año	1400	1450	1500	1550	1600	1650	1700	1750	1800	1850
<y>			155	281	111	851	104	439	171	70
<&>	18	129	335	1						
<e>	66	9	52	164	7	4		1	3	

Fig. 3.2.2. Formas de la conjunción «y». Frecuencia absoluta

[9] Utilizamos la franja cronológica con intervalo de 50 años. La cantidad redonda de 50 años corresponde al promedio de la edad de persona a lo largo de historia. Conviene utilizar el número redondo, por ejemplo, 10, 20, 25, 50, 100, 200,... El intervalo de 25 no conviene por ofrecer las frecuencias bastante reducidas y el de 100 parece ser demasiado lato para detectar los cambios lingüísticos. El número del año corresponde al inicio de la franja, de modo que 1400, por ejemplo, comprende los años de 1400 a 1449.

La transformamos en la tabla de frecuencia normalizada por 1000 palabras:

Año.50	1400	1450	1500	1550	1600	1650	1700	1750	1800	1850
año (sus.)			0.6	3.5	4.4	4.2	6.2	3.4	4.7	3.5
anno (sus.)	2.3	8.1	2.5	0.1				0.6		
bien (adv.)	8.2	2.0	2.5	3.1	3.6	1.5	0.6	0.6	0.6	
vien (adv.)				0.9		1.0		0.2		
cual (relat.)				0.1		0.8	0.6			3.5
qual (relat.)	5.8	3.4	3.6	4.3	2.9	2.6	1.7	1.3	0.9	
escribano (sus.)				0.1	0.7	0.6		0.9	0.6	0.5
escriuano (sus.)	1.2	1.3	2.5	2.8	2.2	1.7		0.1		
haber (vb.)	4.7		0.4	0.3		1.0	1.1	5.1	5.6	5.4
_aber (vb.)	1.2	3.4	3.3	4.6	2.9	3.3	6.2	1.6		
hacer (vb.)			1.5	1.0	2.2	0.8	2.2	5.1	4.0	6.9
facer (vb.)	7.0	3.4	3.8	0.7		0.7		0.8		
Juan (n. prop.)	3.5	6.7	4.2	6.2		2.6	1.1	1.0	3.7	0.5
Joan (n. prop.)				1.0	6.5	1.5				
merced (sus.)			0.1	2.6		1.9	0.6	3.8	1.2	
merzed (sus.)						1.4	0.6	0.5		
mi (poses.)			0.1	2.8	25.4	1.1	1.7	2.9	1.6	1.5
mj (poses.)	5.8		1.0	1.2						

mil (num.)						0.8	2.8	0.8	2.8	3.0
mill (num.)	1.2	10.8	1.7	3.8	2.2	2.8	2.2	0.6		
saber (vb.)	1.2	0.7	1.0	1.3	0.7	2.5	1.7	2.4	1.6	0.5
sauer (vb.)			0.1	2.5	0.7	0.5	0.6	1.2		
señor (sus.)			0.6	1.5	7.3	2.6	7.3	5.0	4.0	6.9
sennor (sus.)	4.7	8.1	6.0					0.9		
un (art.)				1.9	0.7	7.0	5.6	8.0	6.8	4.9
vn (art.)	2.3	2.0	4.9	18.5	3.6	4.3	4.5	1.2		
vecino (sus.)			1.0	0.7	0.7	1.5		1.3	2.8	0.5
vezino (sus.)		2.7	4.7	2.2		1.1	1.1	2.9		
ver (vb.)	2.3	0.7	1.2	1.3	2.2	0.5	0.6	2.9	1.9	0.5
ber (vb.)			0.3	0.3		1.3		1.2	0.3	
villa (sus.)	1.2	7.4	10.3	4.9	2.2	3.4	0.6	7.2	7.1	3.0
uilla (sus.)				0.1	2.2	4.4	3.4	0.3		
y (conj.)			22.5	41.3	80.5	87.6	58.3	51.1	53.1	34.6
& (conj.)	21.0	86.9	48.7	0.1						
e (conj.)	77.0	6.1	7.6	24.1	5.1	0.4		0.1	0.9	

Fig. 3.2.3. Formas variantes frecuentes.
Frecuencia normalizada por 1000 palabras

3.3. Frecuencia media

Para medir la importancia numérica de la variación, la frecuencia total (FT) sirve como indicador de la misma, cuya fórmula es la siguiente:

$FT = \Sigma F(i)$

donde F(i) es cada frecuencia, por ejemplo, en el conjunto de datos (63. 0, 88.3, 108.5), su frecuencia total (FT) es 63.0+88.3+108.5 = 259.3.

De aquí en adelante, en lugar de la frecuencia total (FT), utilizamos la frecuencia media, que sirve como un valor representativo del conjunto, junto con la desviación media, que explicaremos en la sección inmediatamente posterior. La frecuencia media se calcula con la siguiente fórmula:

$FM = \Sigma F(i) / N$

donde F(i) es frecuencia de alguna forma y N es el número de franjas (=10). Por ejemplo, veamos el siguiente conjunto de datos, que representa la frecuencia del lema «el» (artículo), normalizada por 1000 palabras de cada franja cronológica con intervalo de 50 años:

«el» (art.)	1400	1450	1500	1550	1600	1650	1700	1750	1800	1850
FN.1000	63.0	88.3	108.5	80.3	91.4	93.7	112.2	100.3	91.2	124.4

Fig. 3.3.1. «el» (artículo). Frecuencia normalizada

Su frecuencia media (FM) es:

FM = (63.0+88.3+108.5+80.3+91.4+93.7+112.2+100.3+91.2+124.4) / 10 = 95.3

En la siguiente tabla parcial correspondiente al lema «y», hemos agregado la frecuencia media (FM):

N.FN. 1000	1400	1450	1500	1550	1600	1650	1700	1750	1800	1850	**FM**
<y> (conj.)			22.5	41.3	80.5	87.6	58.3	51.1	53.1	34.6	**42.9**
<&> (conj.)	21.0	86.9	48.7	0.1							**15.7**
<e> (conj.)	77.0	6.1	7.6	24.1	5.1	0.4		0.1	0.9		**12.1**

FM: Frecuencia media

Fig. 3.3.2. Formas del lema «y».
Frecuencia normalizada por 100 palabras

3.4. Desviación media

Por otra parte también es importante observar la desviación que representa el grado de cambios numéricos. Por ejemplo, la temperatura media de 20 grados con desviación media de 15 grados, cuya vacilación media

está entre 20 - 15 = 5 grados y 20 + 15 = 35 grados, no es la misma que la idéntica temperatura media, 20 grados, con desviación media de 5 grados (20 - 5 = 15 grados, 20 + 5 = 25 grados). El primer caso es mucho más variable que el segundo. Y ahora nos interesa observar el grado de variabilidad en los cambios lingüísticos.

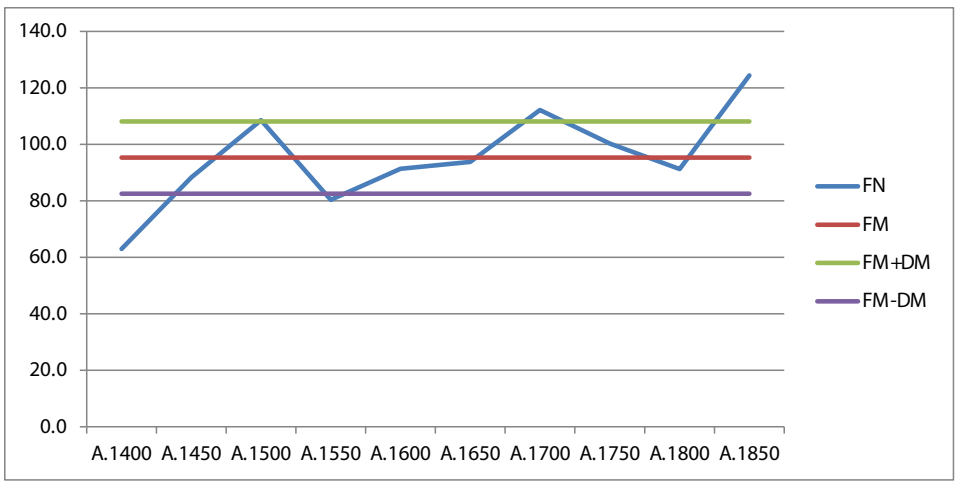
Existen varias maneras de medir el grado de variabilidad. Una de ellas es la desviación media (DM), cuya fórmula es:

$$DM = \Sigma|F(i) - FM| / N$$

donde |F(i) - FM| representa el valor absoluto de la diferencia entre cada frecuencia F(i) y la frecuencia media (FM). La desviación media (DM) del mismo dato es:

$$DM = (|63.0-95.3|+|88.3-95.3|+|108.5-95.3|+|80.3-95.3|+|91.4-95.3|+|93.7-95.3|+|112.2-95.3|+|100.3-95.3|+|91.2-95.3|+|124.4-95.3|) / 10 = 12.8$$

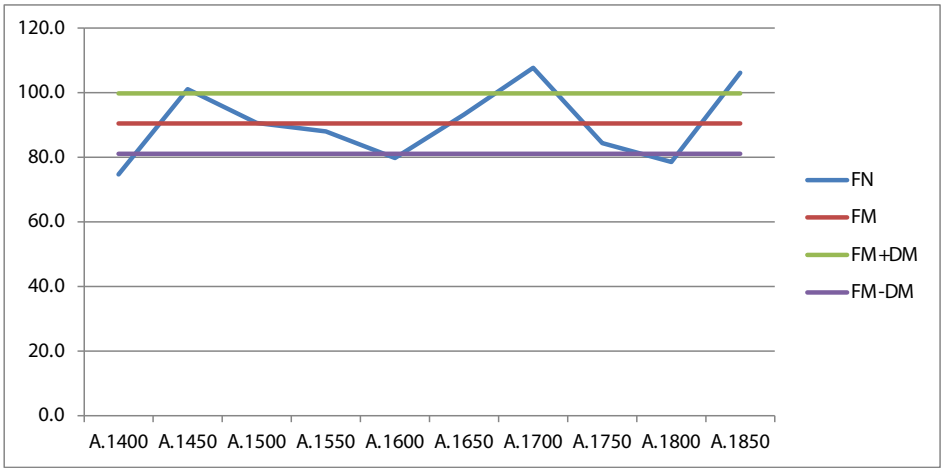
De esta manera, comprendemos que la desviación media significa el promedio de la desviación con respecto a la frecuencia media. Es decir, las diez frecuencias en promedio vacilan entre el dintel (la frecuencia media + la desviación media por arriba) y el umbral (frecuencia media - desviación media), lo que representa el siguiente gráfico:



FN: frecuencia normalizada por 1000 palabras
FM: frecuencia media, DM: desviación media

Fig. 3.4.1. Lema «el» (art.).

Lo mismo que el caso del lema del artículo definido «el», la preposición más frecuente, «de», también es relativamente constante en su forma gráfica:



FN: frecuencia normalizada por 1000 palabras
FM: frecuencia media, DM: desviación media

Fig. 3.4.2. Lema «de» (prep.)

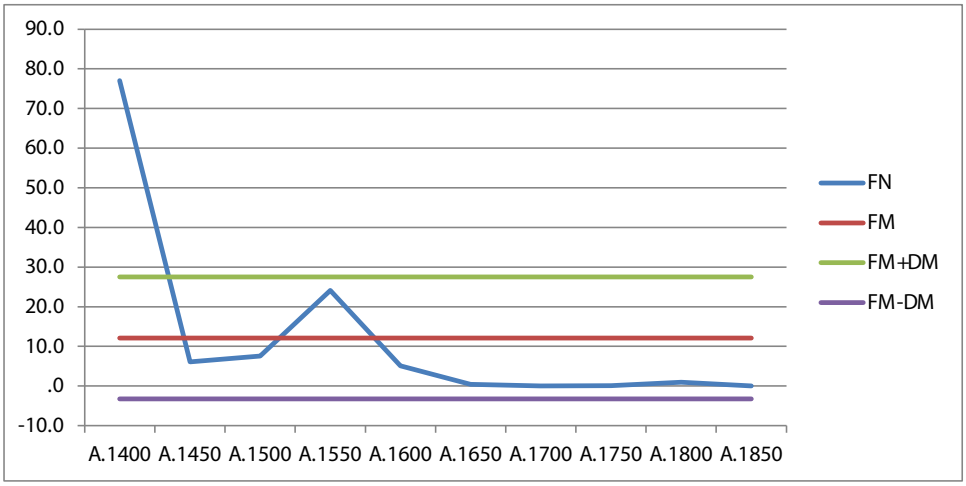
Pero ahora, antes de proceder a analizar el grado de variabilidad, conviene reconsiderar el uso de la desviación media (DM) para estos casos de las formas variantes, puesto que se presenta un problema técnico estadístico ineludible. Tras averiguar su característica, intentaremos solucionarlo inmediatamente.

Explicamos el problema con un ejemplo de <&> («et»), que corresponde a la actual conjunción copulativa «y». Su frecuencia media (FM) es 12.1 y la desviación media (DM) es: 15.4. La siguiente tabla presenta la frecuencia normalizada por 1000 palabras (FN), frecuencia media y dos líneas, superior e inferior con respecto a la desviación media: FM+DM y FM-DM:

&(conj.)	1400	1450	1500	1550	1600	1650	1700	1750	1800	1850
FN	77.0	6.1	7.6	24.1	5.1	.4	.0	.1	.9	.0
FM	12.1	12.1	12.1	12.1	12.1	12.1	12.1	12.1	12.1	12.1
FM+DM	27.5	27.5	27.5	27.5	27.5	27.5	27.5	27.5	27.5	27.5
FM-DM	-3.2	-3.2	-3.2	-3.2	-3.2	-3.2	-3.2	-3.2	-3.2	-3.2

FN: frecuencia normalizada por 1000 palabras
FM: frecuencia media / DM: desviación media

Fig. 3.4.3. Forma «e» (conj.)



FN: frecuencia normalizada por 1000 palabras
FM: frecuencia media / DM: desviación media

Fig. 3.4.4. Forma «e» (conj.)

Nos damos cuenta de que la desviación media (DM = 15.4) excede la frecuencia media (FM = 12.1), y por lo tanto la línea inferior de la desviación media (FM-DM) presenta un valor negativo (12.1 - 15.4 = -3.2)[10]. Tenemos entendido que la desviación media (DM) es un valor medio de desviación con respecto a la frecuencia media (FM). Esto quiere decir que cada frecuencia oscila por promedio entre la (FM + DM) y (FM - DM). Sin embargo, ahora ocurre que el límite inferior FM - FD está por bajo cero (-3.2). Naturalmente esto no es concebible, puesto que la frecuencia por su naturaleza nunca puede ser un valor negativo (bajo cero), a diferencia de, por ejemplo, la temperatura.

La razón por la que se ha presentado el límite bajo negativo en FM - DM (12.1 - 15.4 = -3.2) se debe a que estamos tratando un dato sumamente variado, de modo que la desviación media (DM) se presenta bastante grande por el valor que hay de suma frecuencia, concretamente 77.0, que empuja el valor medio de desviación hacia arriba. Al aplicarlo a la parte inferior, resulta que sobrepasa el nivel cero[11].

[10] Por valores inferiores que hay en el segundo decimal, puede haber la diferencia de 0.1 en el cálculo de sustracción.

[11] Lo mismo que la desviación media, la desviación típica (DT), o «desviación estándar», también presenta el problema del valor negativo en caso del dato extremamente desigual. Hay que considerar este inconveniente a la hora de aplicar la DT junto con la frecuencia media (FM), en forma de FM + DT y FM - DT, puesto que la segunda,

Para solucionar el inconveniente del valor negativo en la parte inferior de la diferencia de frecuencia media y la desviación media, proponemos utilizar las fórmulas de la «desviación media superior» (DMS) y la «desviación media inferior» (DMI) de manera separada. En el conjunto de datos, por ejemplo, {77.0, 6.1, 7.6, 24.1, 5.1, 0.4, 0.0, 0.1, 0.9, 0.0}, las frecuencias superiores a la frecuencia media (12.1) son: {77.0, 24.1} y las frecuencias inferiores a la frecuencia media son {6.1, 7.6, 5.1, 0.4, 0.0, 0.1, 0.9, 0.0}.

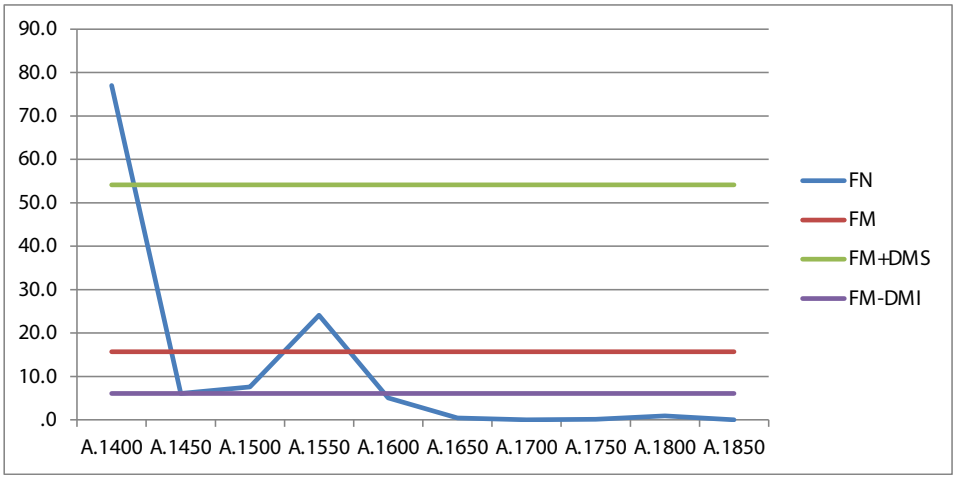
La desviación media superior (DMS) es:

DMS = [(77.0 - 12.1) + (24.1 - 12.1)] / 2 = 38.5

Y la desviación media inferior (DMI) es:

DMI = ((12.1-6.1) + (12.1-7.6) + (12.1-5.1) + (12.1-0.4) + (12.1-0.0) + (12.1-0.1) + (12.1-0.9) + (12.1-0.0)) / 8 = 9.6

De esta manera, ahora la problemática parte inferior no presenta el valor negativo: 12.1 - 9.6 = 2.5.



FN: frecuencia normalizada por 100 palabras, FM: frecuencia media
DMS: desviación media superior, DMI: desviación media inferior

Fig. 3.4.5. Forma «e» (conj.).

FM - DT, puede presentar el valor negativo. La presentación de FM + DT y FM - DT es una práctica usual en el tratamiento estadístico descriptivo.

Para saber el grado medio de desviación, proponemos utilizar la «desviación media dual» (DMD)[12], que es la suma de la desviación media superior (DMS) + la desviación media inferior (DMI) dividida por el rango (el valor máximo: MAX - el valor mínimo: MIN).

DMD = (DMS + DMI) / (MAX - MIN) = (38.5 + 9.6) / (77.0 - 0.0) = 0.625

Ahora bien, estamos en condición de medir el grado de variación en forma de la desviación media dual (DMD), junto con la frecuencia media (FM):

Forma variante	FM	DM	DMS	DMI	DMD
año (sus.)	3.0	1.7	1.2	2.8	.657
anno (sus.)	1.4	1.8	2.9	1.3	.519
bien (adv.)	2.3	1.7	2.1	1.4	.422
vien (adv.)	.2	.3	.5	.2	.695
cual (relat.)	.5	.7	1.1	.5	.461
qual (relat.)	2.6	1.4	1.4	1.4	.464
escribano (sus.)	.4	.3	.3	.3	.696
escriuano (sus.)	1.2	.9	.9	.9	.662
haber (vb.)	2.4	2.3	2.8	1.9	.845
_aber (vb.)	2.6	1.6	1.3	1.9	.525
hacer (vb.)	2.4	1.8	3.0	1.3	.615
facer (vb.)	1.6	1.8	3.1	1.3	.627
Juan (n. prop.)	3.0	1.9	1.9	1.9	.568
Joan (n. prop.)	.9	1.3	2.1	.9	.465
merced (sus.)	1.0	1.1	1.4	.9	.593
merzed (sus.)	.2	.3	.6	.2	.571
mi (poses.)	3.7	4.3	21.7	2.4	.949
mj (poses.)	.8	1.1	1.9	.8	.459
mil (num.)	1.0	1.1	1.8	.8	.884
mill (num.)	2.5	2.0	3.3	1.4	.433
saber (vb.)	1.4	.5	.7	.5	.575
sauer (vb.)	.6	.5	.9	.4	.515

[12] Los tres términos, «desviación media superior» (DMS), «desviación media inferior» (DMI) y «desviación media dual» (DMD), son de nuestras propuestas.

señor (sus.)	3.5	2.6	2.6	2.6	.709
sennor (sus.)	2.0	2.6	4.3	1.8	.755
un (art.)	3.5	3.0	3.0	3.0	.741
vn (art.)	4.1	3.1	3.9	2.6	.353
vecino (sus.)	.9	.6	.8	.5	.477
vezino (sus.)	1.5	1.3	1.6	1.1	.589
ver (vb.)	1.4	.7	.9	.6	.634
ber (vb.)	.3	.4	.9	.2	.851
villa (sus.)	4.7	2.7	2.7	2.7	.546
uilla (sus.)	1.0	1.4	2.3	1.0	.735
y (conj.)	42.9	23.2	23.2	23.2	.530
& (conj.)	15.7	21.9	36.5	15.7	.600
e (conj.)	12.1	15.4	38.4	9.6	.624

FM: frecuencia media, DM: desviación media,
DMS: desviación media superior,
DMI: desviación media inferior,
DMD: desviación media dual

Fig. 3.4.6. Grados de variación

Las diez formas variantes con la frecuencia media (FM) más elevada son:

N	Forma variante	**FM**	DM	DMS	DMI	DMD
1	*y (conj.)*	**42.9**	23.2	23.2	23.2	.530
2	*& (conj.)*	**15.7**	21.9	36.5	15.7	.600
3	*e (conj.)*	**12.1**	15.4	38.4	9.6	.624
4	*villa (sus.)*	**4.7**	2.7	2.7	2.7	.546
5	*vn (art.)*	**4.1**	3.1	3.9	2.6	.353
6	*mi (poses.)*	**3.7**	4.3	21.7	2.4	.949
7	*señor (sus.)*	**3.5**	2.6	2.6	2.6	.709
8	*un (art.)*	**3.5**	3.0	3.0	3.0	.741
9	*año (sus.)*	**3.0**	1.7	1.2	2.8	.657
10	*Juan (n. prop.)*	**3.0**	1.9	1.9	1.9	.568

Fig. 3.4.7. Primeras diez formas variantes más frecuentes

Y las diez formas variantes con la desviación media (DM) más elevada son:

N	Forma variante	FM	DM	DMS	DMI	**DMD**
1	*mi (poses.)*	3.7	4.3	21.7	2.4	**.949**
2	*mil (num.)*	1.0	1.1	1.8	.8	**.884**
3	*ber (vb.)*	.3	.4	.9	.2	**.851**
4	*haber (vb.)*	2.4	2.3	2.8	1.9	**.845**
5	*sennor (sus.)*	2.0	2.6	4.3	1.8	**.755**
6	*un (art.)*	3.5	3.0	3.0	3.0	**.741**
7	*uilla (sus.)*	1.0	1.4	2.3	1.0	**.735**
8	*señor (sus.)*	3.5	2.6	2.6	2.6	**.709**
9	*escribano (sus.)*	.4	.3	.3	.3	**.696**
10	*vien (adv.)*	.2	.3	.5	.2	**.695**

Fig. 3.4.8. Primeras diez formas variantes más variadas

Vamos a prestar atención a estas formas, puesto que las primeras son las más frecuentes y las segundas son las más variables.

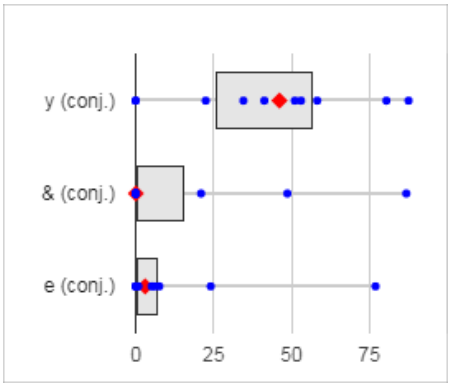
4. VISUALIZACIÓN DE DISTRIBUCIÓN

Nuestro objetivo del actual estudio no es solamente preparar las tablas numéricas, sino también, o más bien, describir los cambios históricos generales. Para ello, intentaremos elaborar gráficos que ofrezcan las imágenes ilustrativas que faciliten nuestras propias interpretaciones de las grandes tendencias cronológicas. Vamos a comparar los métodos tradicionales de visualización con los nuevos propuestos por nosotros.

4.1. Visualización de variación

Tradicionalmente, como medio visualizador de variación de frecuencias se ha venido utilizando el diagrama de caja, o de «caja y bigote», inventado por el matemático estadounidense, John W. Tukey en 1969. Consiste en formar la caja en los dos límites del primer cuartil (25 % de número de datos menores) y del tercer cuartil (75 %). Dentro de ella se encuentra el punto del segundo cuartil o «mediana» (50 %). En la línea se colocan los puntos de todos los datos. La caja así formada representa el ámbito del 50 % (75 % - 25 %). También, en lugar del punto medio, la mediana (50 %),

se puede utilizar la frecuencia media (FM) y para los límites, desviación típica (DT).



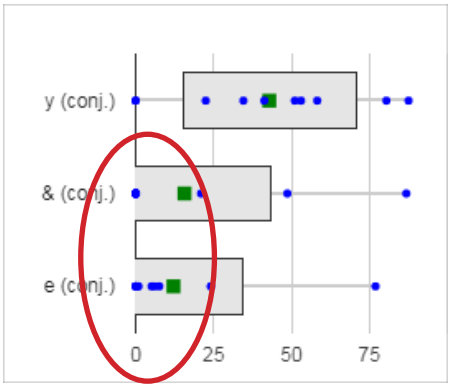


Fig. 4.1.1. Mediana ± cuartil / Fig. 4.1.2. Frecuencia media ± desviación típica

El diagrama de caja es útil para conocer la tendencia de distribución de frecuencia y la amplitud de la caja sirve como indicador de la variación. Sin embargo, a nuestro modo de ver, la versión original del diagrama de caja con mediana y cuartiles peca de la simplificación excesiva. Se trata de una división cuatripartita equitativa sin considerar las situaciones precisas de cada dato. Por ejemplo, el rango intercuartílico (RIC), indicador de la variación, no sirve como indicador preciso de variación, puesto que, por ejemplo, el dato de (0, 2, 5, 8, 10) y el de (0, 2, 3, 8, 50) presentan el mismo RIC, 8 - 2 = 6. El diagrama de caja con mediana y cuartiles tiene el mérito de ser de fácil interpretación y, lo que es importante para nosotros, nunca da valores negativos en los datos de frecuencia.

El diagrama de caja con frecuencia media y desviación típica tiene el mérito de ser más preciso en medir la variación en forma de caja, puesto que la variación se calcula por medio de la desviación típica, más sofisticada que el cuartil. Sin embargo, según el gráfico de arriba se presentan valores negativos en la parte inferior de la caja, lo que es difícil de interpretar, como hemos visto en la sección 3.3. Otro defecto sería la dificultad de entender el concepto de desviación típica (DT) o «desviación estándar» para el público no versado en la estadística:

$$DT = \sqrt{\Sigma\,(F(i) - FM)^2 / N}$$

En cambio, el concepto de desviación media es simple y fácil de entender:

DM = Σ |F(i) - FM| / N

Se trata del promedio de todas las diferencias entre cada frecuencia F(i) y la frecuencia media (FM). Hemos elaborado el programa para instalar en nuestro sistema LYNEAL, presentado en la Introducción (sec. 1). Sin embargo, como hemos visto en la sección 3.3 y el siguiente gráfico izquierdo, presenta el mismo problema de valor negativo, lo mismo que la caja de desviación típica, en la parte inferior:

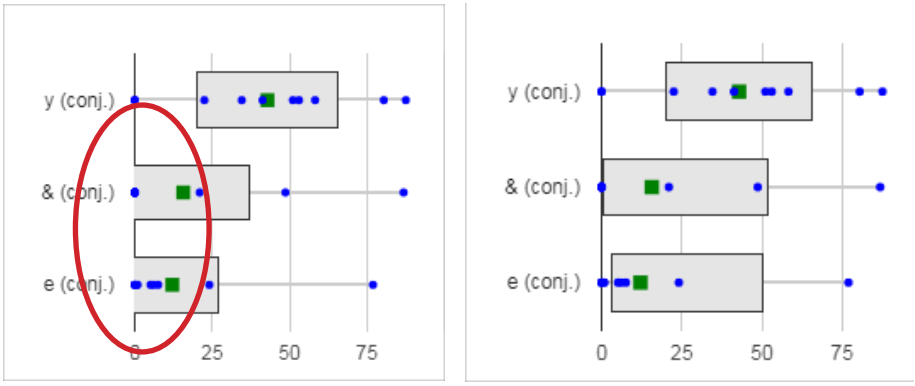


Fig. 4.1.3. Media ± Desviación media / Fig. 4.1.4. Media ± DM superior e inferior

Hemos solucionado el mismo problema que causa el diagrama de caja con desviación media con la introducción de la desviación media superior y la inferior (sec. 3.3), como muestra el gráfico de arriba derecho y el programa en JavaScript que ofrecemos a continuación (sec. 3.4.):

```
function MeanDev(Xn, sel){ // Desviación media
//Xn: vector vertical, sel=0: desviación media (DM),
// 1: DM superior, 2: DM inferior, 3: DM dual
 var n, av, s1=0, s2=0, c1=0, c2=0;
 n = NR(Xn); av = AmA(Xn); //cuento; promedio
 for(var i = 1; i <= n; i++) {
 if (sel == 0) {s1 += Math.abs(Xn[i][1] - av); ++c1} //DM
 if ((sel == 1 || sel == 3) && Xn[i][1] > av) {s1 += Xn[i][1] - av; ++c1;}
 //DM superior
 if ((sel == 2 || sel == 3) && Xn[i][1] < av) {s2 += av - Xn[i][1]; ++c2;}
 //DM inferior
 }
 if (sel == 0) return s1 / c1;
 if (sel == 1) return s1 / c1;
 if (sel == 2) return s2 / c2;
 if (sel == 3) return ((s1 / c1) + (s2 / c2)) / RgV(Xn); //RgV: rango del vector
}
```

4.2. Visualización de frecuencias

Nuestro mayor interés está en la observación de vicisitudes de cambios lingüísticos históricos en forma de frecuencias normalizadas. Cuando se trata de los datos secuenciales, como el caso de cambios cronológicos, generalmente se utiliza el gráfico de línea como el siguiente:

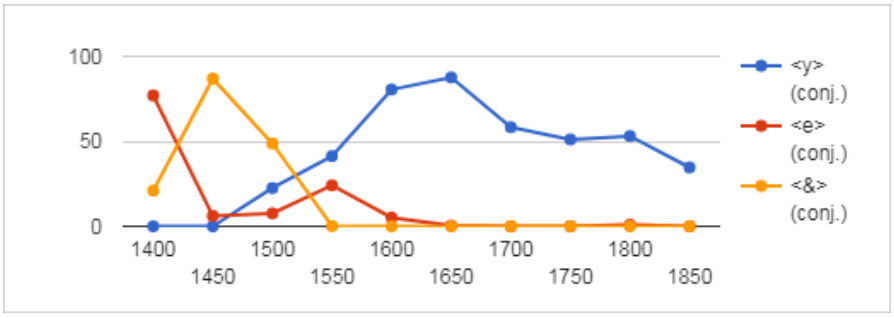


Fig. 4.2.1. Formas del lema «y» en gráfico de línea. Frecuencia normalizada

De esta manera podemos presenciar los altibajos de frecuencias en forma de líneas. Por ejemplo, la historia de la grafía de la conjunción «y» empieza con el signo tironiano (&), que desaparece en 1550. En cambio se utiliza la letra <e> desde 1400 hasta 1600, que cae en desuso en 1650. La forma actual <y> aparece en 1500 y continúa hasta la actualidad. Por estas líneas sabemos que el orden relativo es: <e> → <&> → <y>. Sin embargo, los cambios lingüísticos siempre presentan épocas de transición, donde se cruzan distintas formas en el mismo período. A partir del gráfico anterior, podemos formular la distribución cronológica en el siguiente cuadro esquemático:

Año	1400	1450	1500	1550	1600	1650-1850
<y>	-	-	+	++	++	++
<&>	+	++	++	-	-	-
<e>	++	(+)	(+)	(+)	(+)	-

Fig. 4.2.2. Esquema de las formas del lema «y»: <y>, <&>, <e>

Este esquema nos ayuda a comprender la realidad histórica de cambios. En el gráfico de línea anterior es difícil esquematizarla. La verdad es que hacer el esquema anterior nos ha costado un poco de trabajo.

Otro gráfico usual en la observación de frecuencias es el de barra:

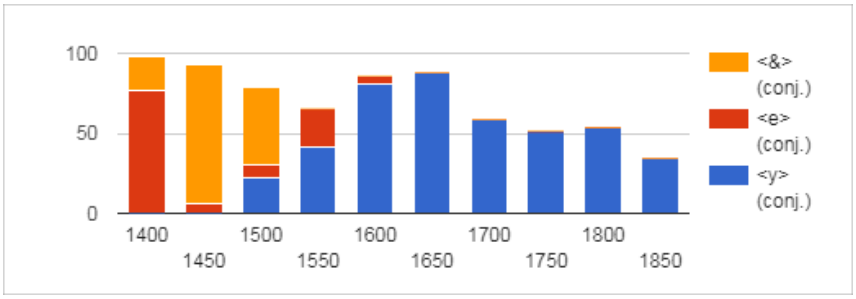


Fig. 4.2.3. Formas del lema «y» en gráfico de barra. Frecuencia normalizada

La barra es adecuada para ver la cantidad y proporción de cada caso, pero ahí es muy difícil dibujar el esquema de tendencias.

Para dar solución a la dificultad de esquematización en el gráfico de línea y en el de barra, hemos programado la tabla de color en la siguiente forma, que presenta inmediatamente el esquema cuantitativo de cambios:

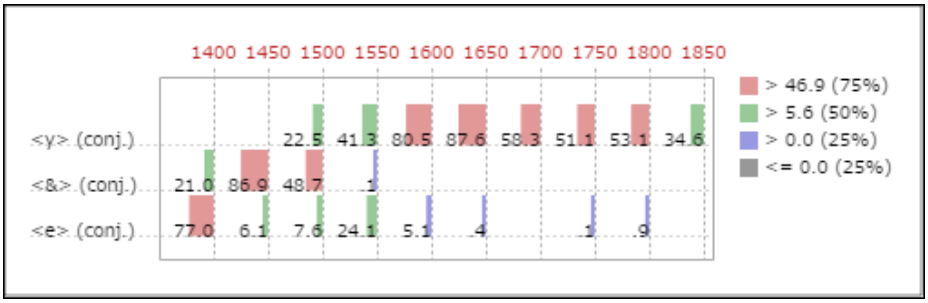


Fig. 4.2.4. Lema: «y» (conjunción) / formas: *<y>, <&>, <e>*.

Hemos utilizados cuatro colores: rojo, verde, azul y gris para distinguir los cuatro grupos: el grupo superior al tercer cuartil (> 75 %), el grupo entre el tercer y el segundo cuartil (> 50 %), el grupo superior al primer cuartil (> 25 %) y el resto, menos o igual que el primer cuartil (<=25 %). A pesar de que hemos rechazado el uso de cuartiles para la medición del grado de variación (sec. 4.1.), hemos admitido su utilidad en dividir los datos en cuatro grupos igualados de número de miembros.

A continuación, veamos la distribución de las formas actuales y las antiguas más frecuentes. Por cuestión de espacio utilizable, nos limitaremos a mostrar los gráficos de distribución y no comentaremos estos casos interesantes e importantes:

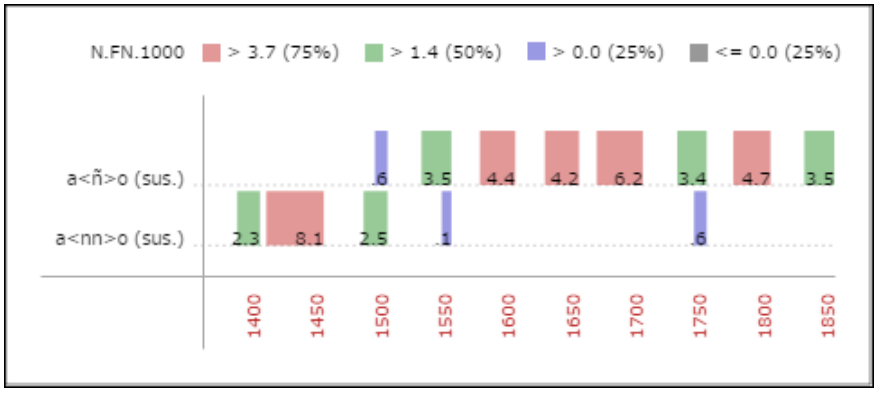


Fig. 4.2.5. Lema: «año» (sustantivo): formas: *a<ñ>o, a<nn>o*.

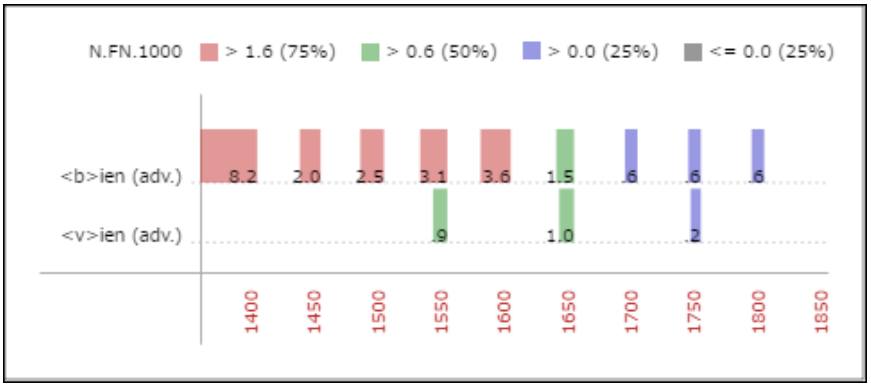


Fig. 4.2.6. Lema: «bien» (adverbio): formas: *<b>ien, <v>ien*.

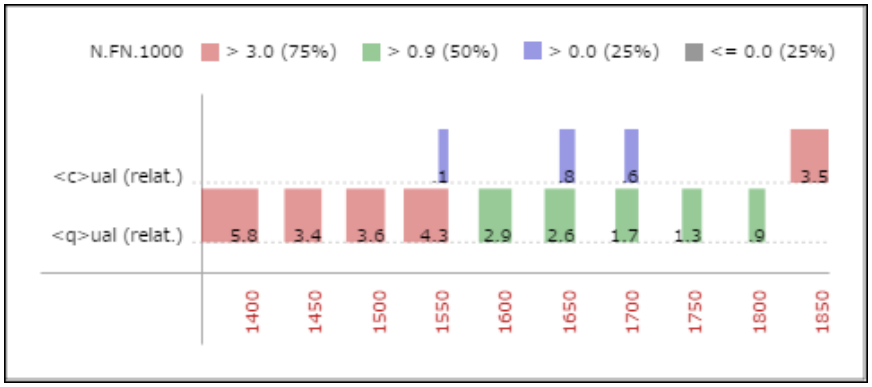


Fig. 4.2.7. Lema: «cual» (relativo): formas: *<c>ual, <q>ual*.

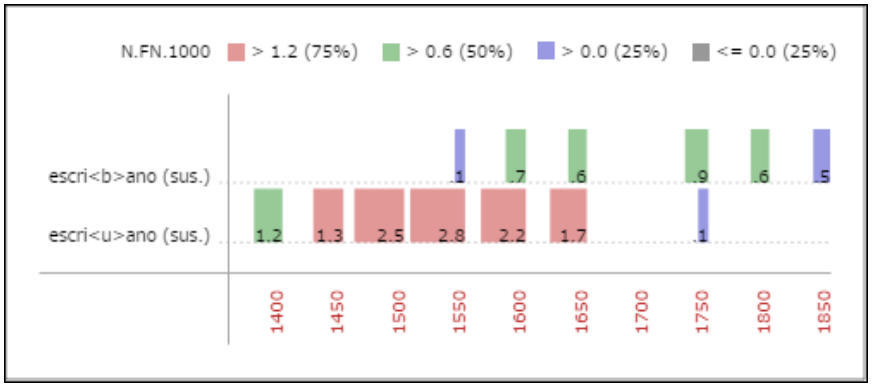


Fig. 4.2.8. Lema: «escribano» (sustantivo): formas: *escri<b>ano, escri<u>ano*.

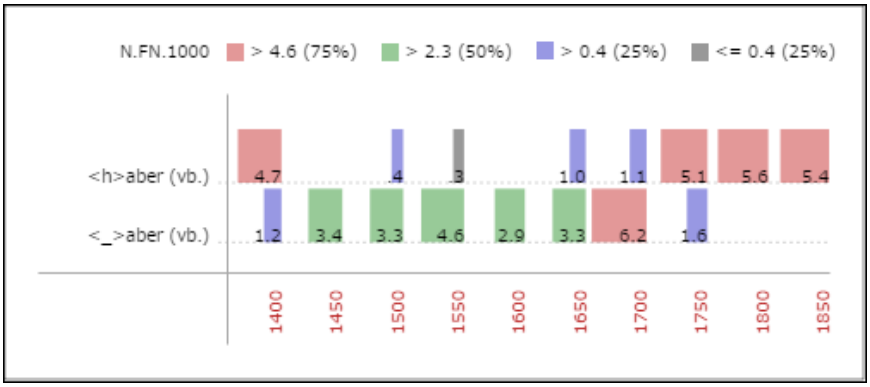


Fig. 4.2.9. Lema: «haber» (verbo): formas: *<h>aber, <_>aber*.

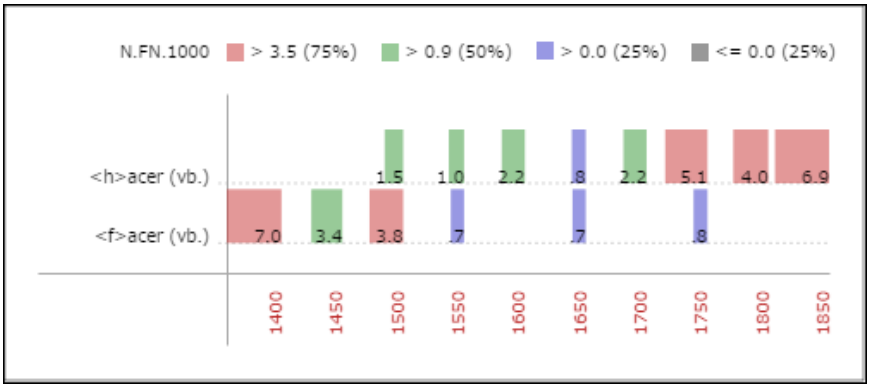


Fig. 4.2.10. Lema: «hacer» (verbo): formas: *<h>acer, <f>acer*.

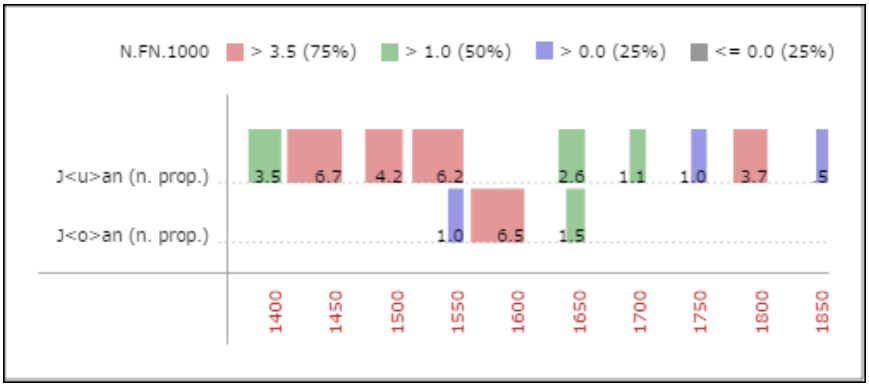


Fig. 4.2.11. Lema: «Juan» (nombre propio): formas: *J<u>an, J<o>an*.

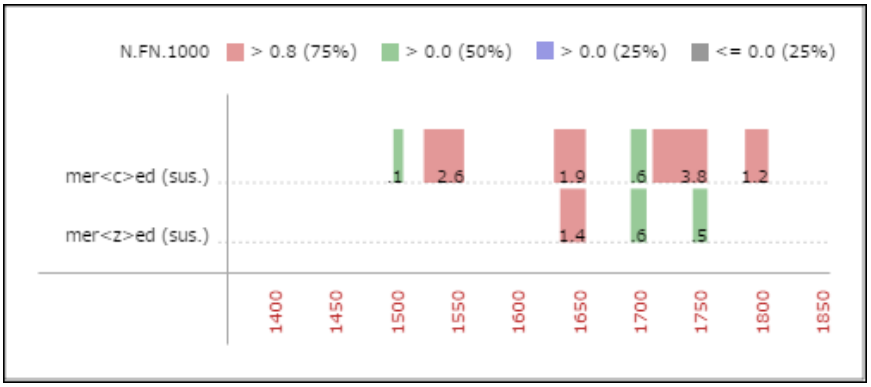


Fig. 4.2.12.Lema: «merced» (sustantivo): formas: *mer<c>ed, mer<z>ed*.

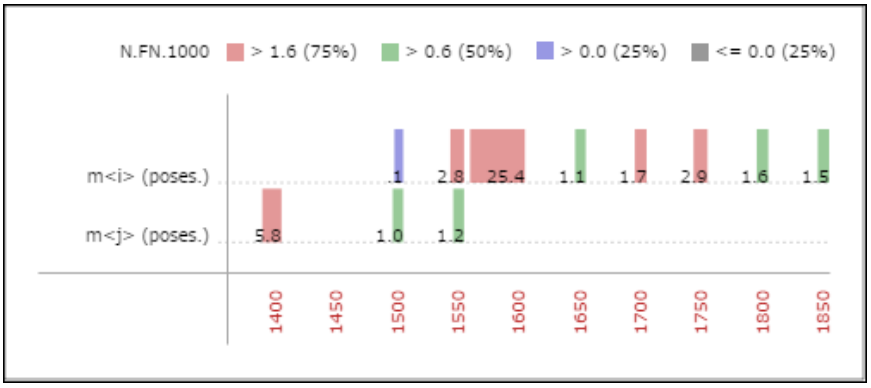


Fig. 4.2.13. Lema: «mi» (posesivo): formas: *m<i>, m<j>*.

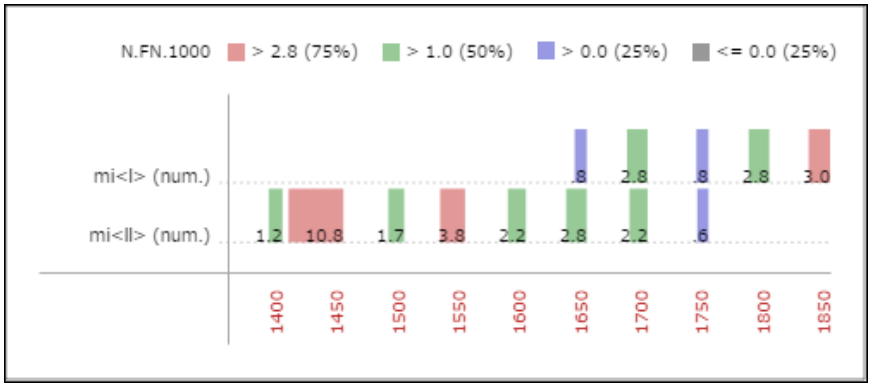


Fig. 4.2.14. Lema: «mil» (numeral): formas: *mi<I>, mi<II>*.

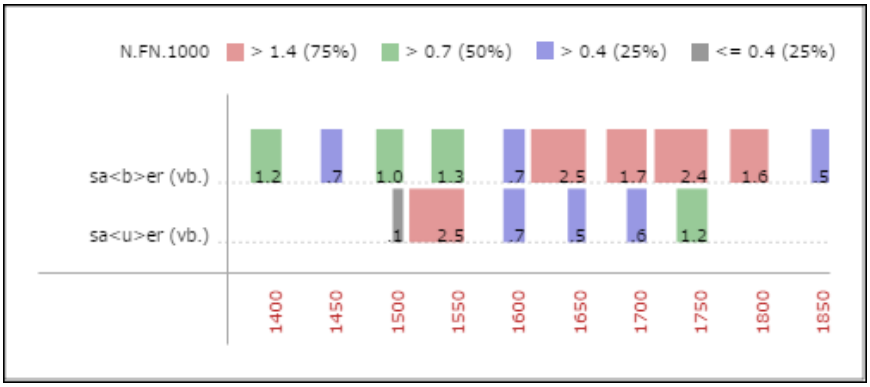


Fig. 4.2.15. Lema: «saber» (verbo): formas: *sa<b>er, sa<u>er*.

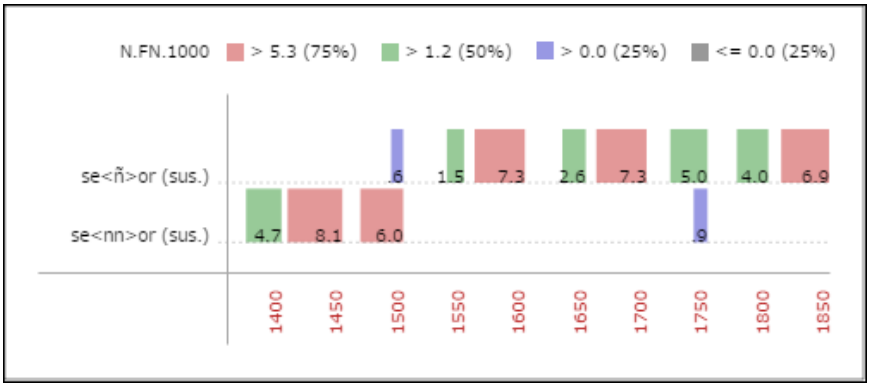


Fig. 4.2.16. Lema: «señor» (sustantivo): formas: *se<ñ>or, se<nn>or*.

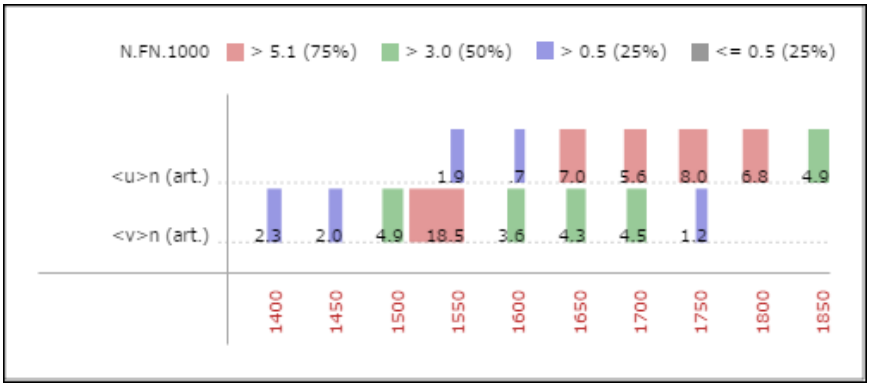


Fig. 4.2.17. Lema: «un» (artículo): formas: *<u>n, <v>n*.

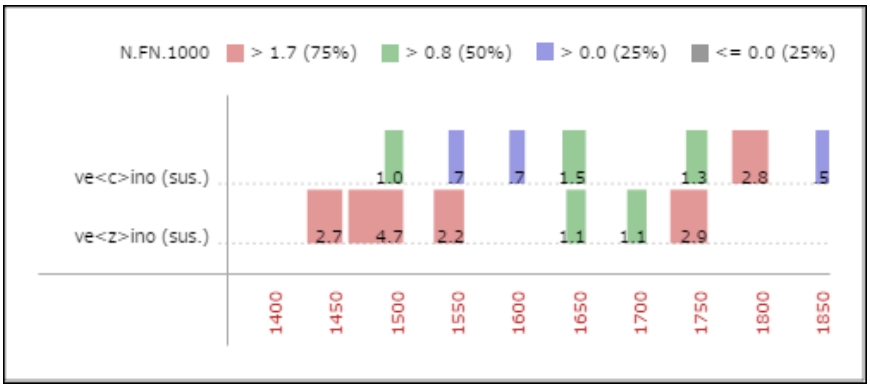


Fig. 4.2.18. Lema: «vecino» (sustantivo): formas: *ve<c>ino, ve<z>ino*.

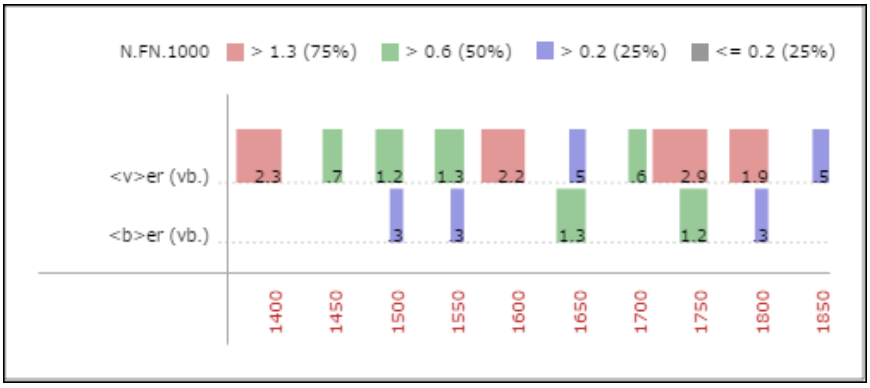


Fig. 4.2.19. Lema: «ver» (verbo): formas: *<v>er, <b>er*.

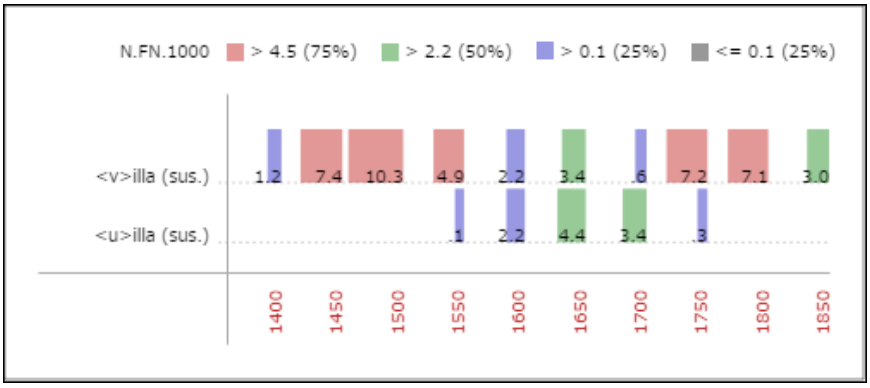


Fig. 4.2.20. Lema: «villa» (sustantivo): formas: *<v>illa, <u>illa*.

Otro problema que causa el gráfico de línea es su limitación numérica de casos. Por ejemplo, nos interesan no solamente las tres formas pertenecientes al lema «y», sino un conjunto de numerosos casos. A la hora de preparar el gráfico de línea con el conjunto grande, las líneas y la leyenda de líneas resultan casi ilegibles, como muestra el siguiente gráfico:

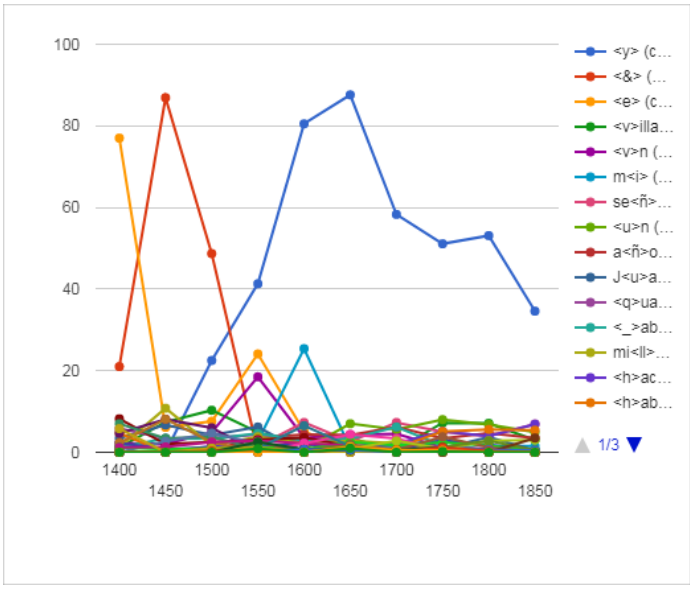


Fig. 4.2.21. Gráfico de línea: 35 casos de las formas variables

Lo mismo pasa también en el gráfico de barra:

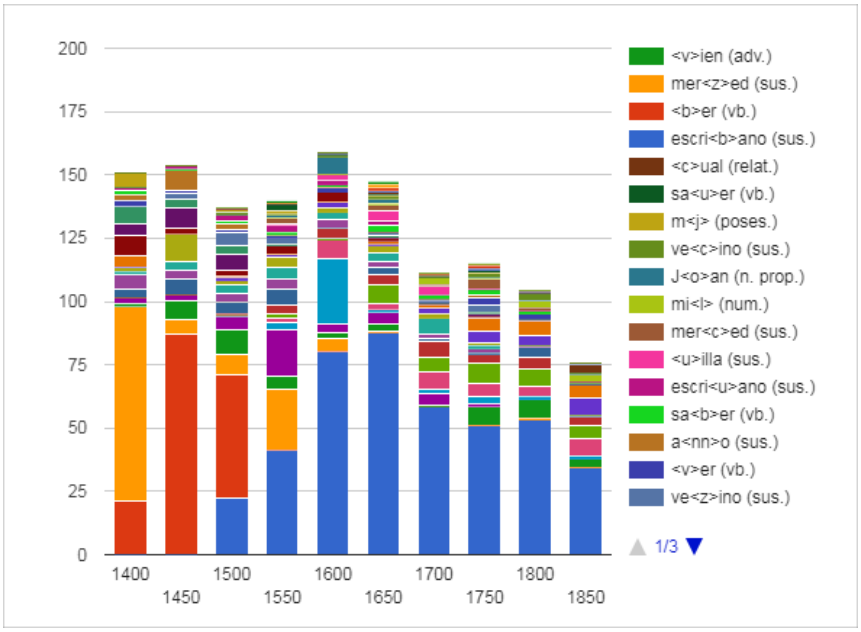


Fig. 4.2.22. Gráfico de barra: 35 casos de las formas variables

En cambio, la tabla de color, propuesta por nosotros, puede visualizarse sin problema[13]:

[13] El Microsoft Excel posee la función parecida, que se realiza directamente en las celdas. Nuestro método difiere de ella por realizarse en la pantalla de canvas de html en la página web, con ejes vertical y horizontal graduables como veremos más adelante (4.3)

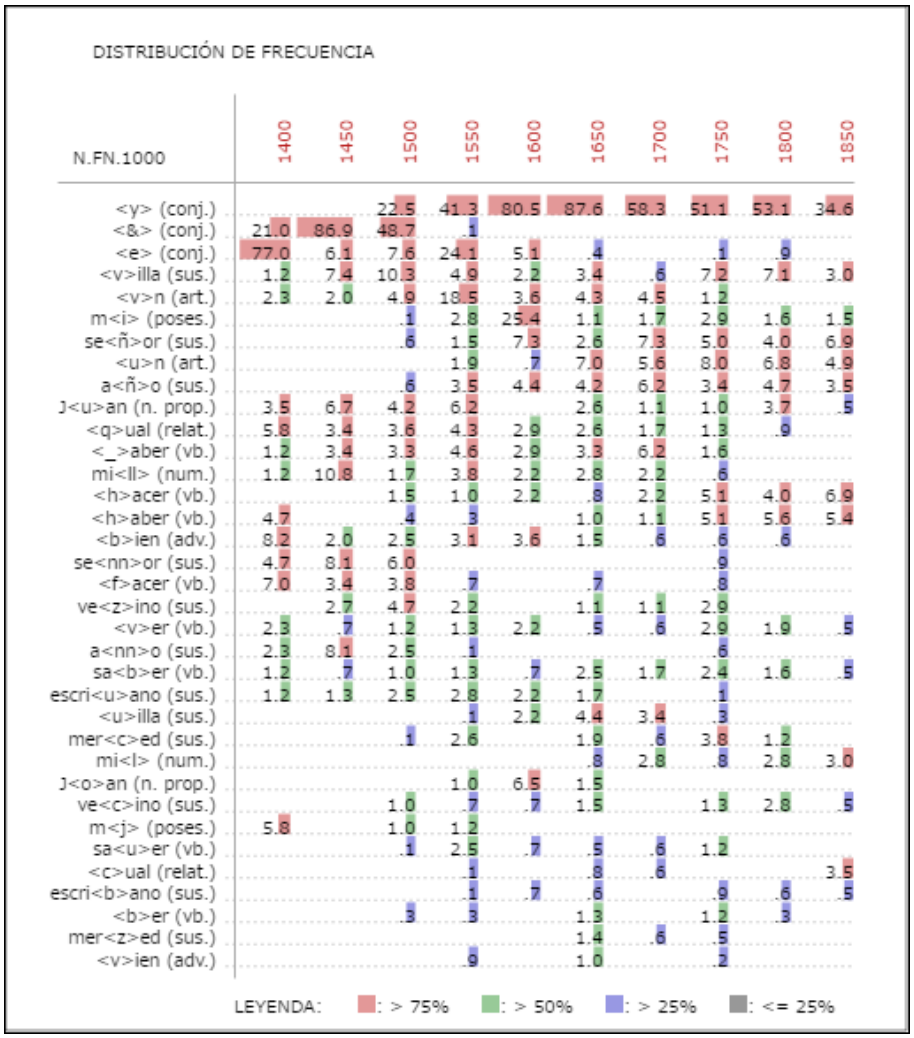


Fig. 4.2.23. Tabla de color: 35 casos de las formas variables.

El siguiente gráfico de círculo en color representa las frecuencias de cada caso en cada periodo de años por el color y el tamaño de círculos:

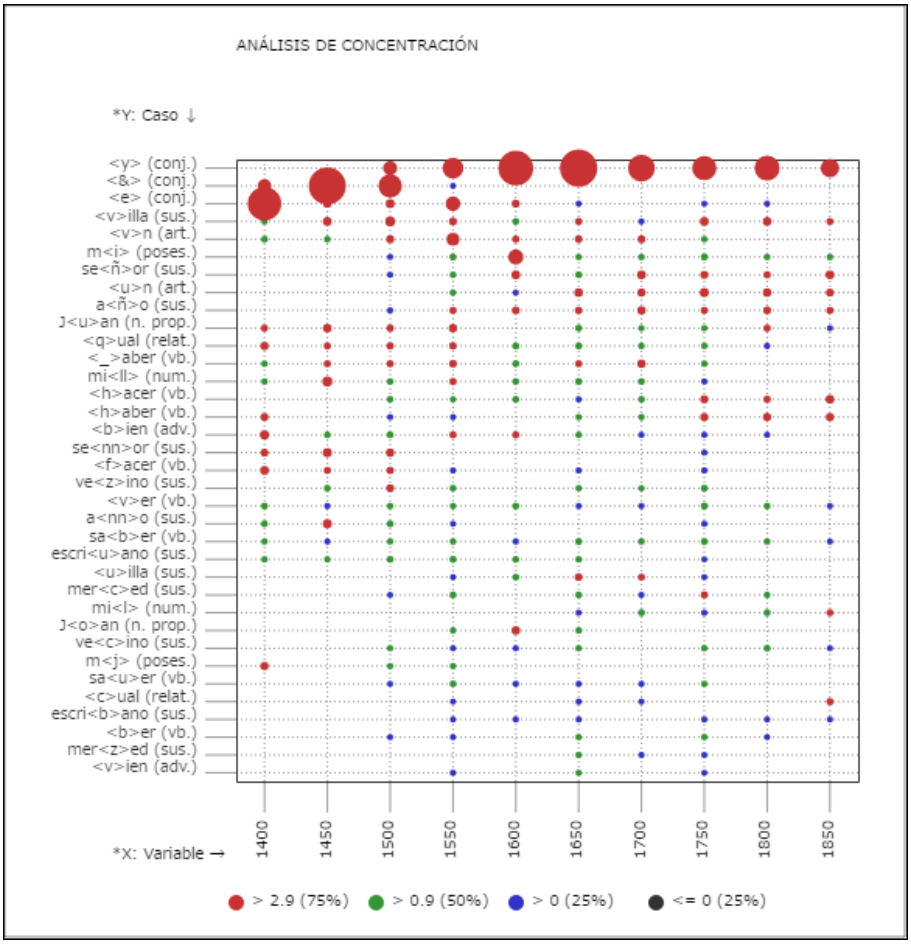


Fig. 4.2.24. Gráfico de círculo en color: 35 casos de las formas variables

Por estos gráficos podemos observar fácilmente los cambios de cada forma en relación con otras en el mismo plano bidimensional.

4.3. Visualización concentrada

En la sección anterior (sec. 4.2.), hemos visto que las formas variantes pertenecientes al mismo lema están distribuidas no al azar, sino siguiendo un patrón. Cuando se trata de tres formas, es fácil de hallar el patrón de distribución, como el caso de <&>, <e> e <y>. La patronización de dos variantes es fácil y se realiza automáticamente, puesto que siempre el orden es el mismo: forma antigua > forma actual.

Cuando se trata de 35 casos, pertenecientes a los 17 lemas variantes más frecuentes, la patronización con cambios del orden de casos (forma lingüística) y variables (franja cronológica) se convierte en una tarea difícil por no decir imposible de realizar[14]. El método de distancia conjunta, que hemos denominado «Análisis de Concentración», consiste en buscar el patrón concentrado en la línea diagonal de parte inferior izquierda a la parte superior derecha, lo que está realizado en el gráfico siguiente (análisis bilateral de concentración):

[14] Bertin (1977, 1988) ha propuesto un método manual, mientras que Benzecri («Análisis de correspondencia») y Hayashi («Cuantificación tipo III») han encontrado método de cálculo matricial independientemente uno de otro. Para nuestra aplicación del método de Hayashi, véase Moreno Fernández (1990: 152-158). Por nuestra parte, hemos propuesto un método basado en el cálculo de distancia conjunta de frecuencias horizontales y verticales (Ueda, 1993; Moreno Fernández, 1999: 364-368).

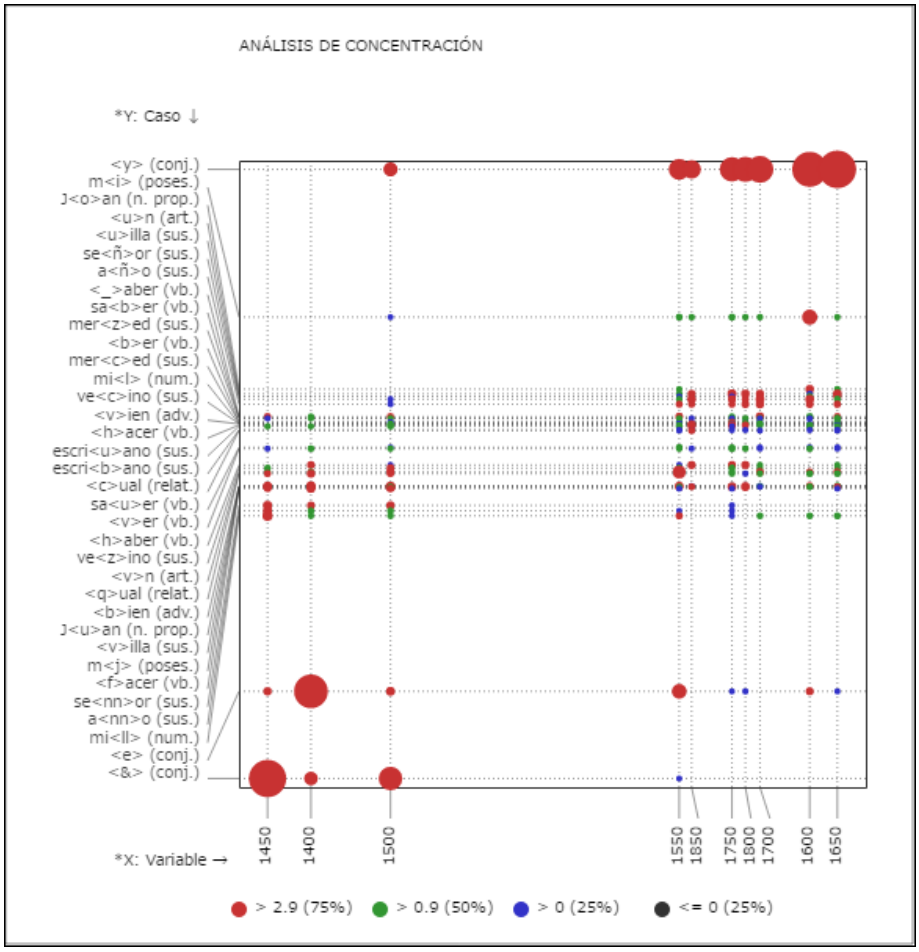


Fig. 4.3.1. Concentración bilateral

Este gráfico no es conveniente, puesto que el orden de las franjas cronológicas, 1450, 1400, 1500, 1550,1850, ..., 1650, conseguido por buscar la patronización diagonal más concentrada, no es convincente. El orden cronológico debe ser constante manteniendo la secuencia de 1400, 1450, ..., 1850. Para el análisis cronológico de forma lingüísticas, únicamente el eje de formas es sujeto al cambio de orden. El siguiente gráfico es el resultado del análisis unilateral de concentración, producido por el cambio de orden de las formas:

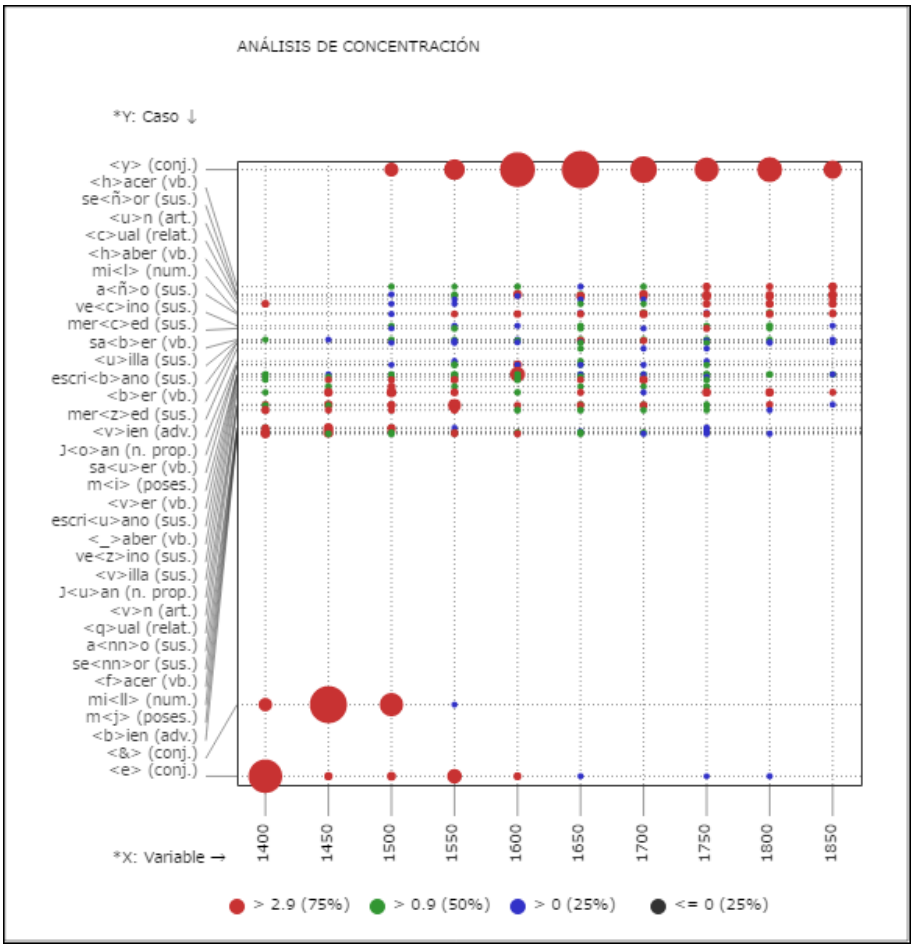


Fig. 4.3.2. Concentración unilateral

Por este gráfico ahora sí que podemos comprobar la concentración de <&> y <e> en los años relativamente tempranos, 1400, 1450, 1500, 1550. En cambio la forma <y> es moderna por excelencia: 1600 en adelante.

Se efectúa la periodización automática por medio de las áreas concentradas. En primer lugar, el programa busca el punto de división óptima calculando la densidad de frecuencias en las cuatro divisiones: A, B, C, D, de las cuales, las divisiones A y D deberían presentar la máxima densidad de frecuencias. En cambio, cuanta menos densidad de frecuencia se presenten en las zonas B y C, mejor. El programa busca todos los puntos de intersección de la línea divisoria vertical y la horizontal y devuelve el punto donde se ha calculado el máximo valor de la siguiente fórmula:

[(A + D) - (B + C)] / [(A + D) + (B + C)]

donde A, por ejemplo, representa la densidad de frecuencia, que es la frecuencia media de la zona A. El resultado se presenta en el gráfico siguiente, donde «A: 1 +» significa que la división A es positiva en cuanto a la línea divisoria 1. En cambio la división B es negativa:

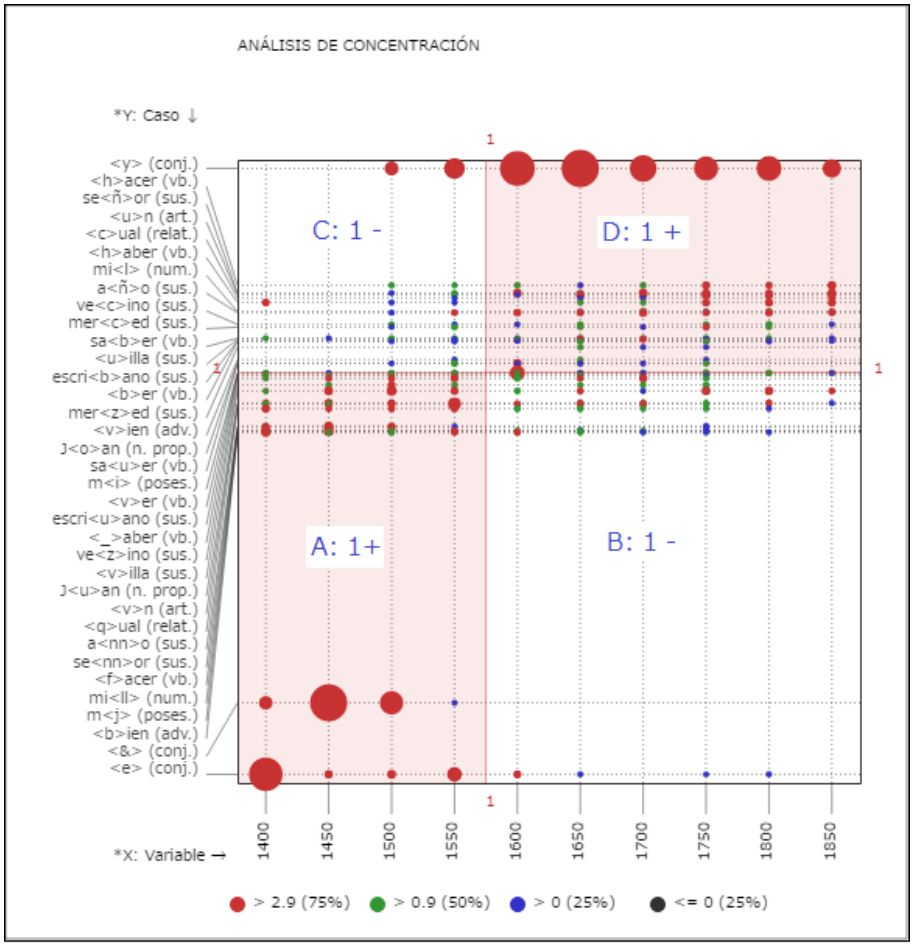


Fig. 4.3.3. División 1.

De esta manera la primera división de la franja cronológica se encuentra entre 1550 y 1600.

El programa busca el segundo punto de intersección entre 1550 y 1600.

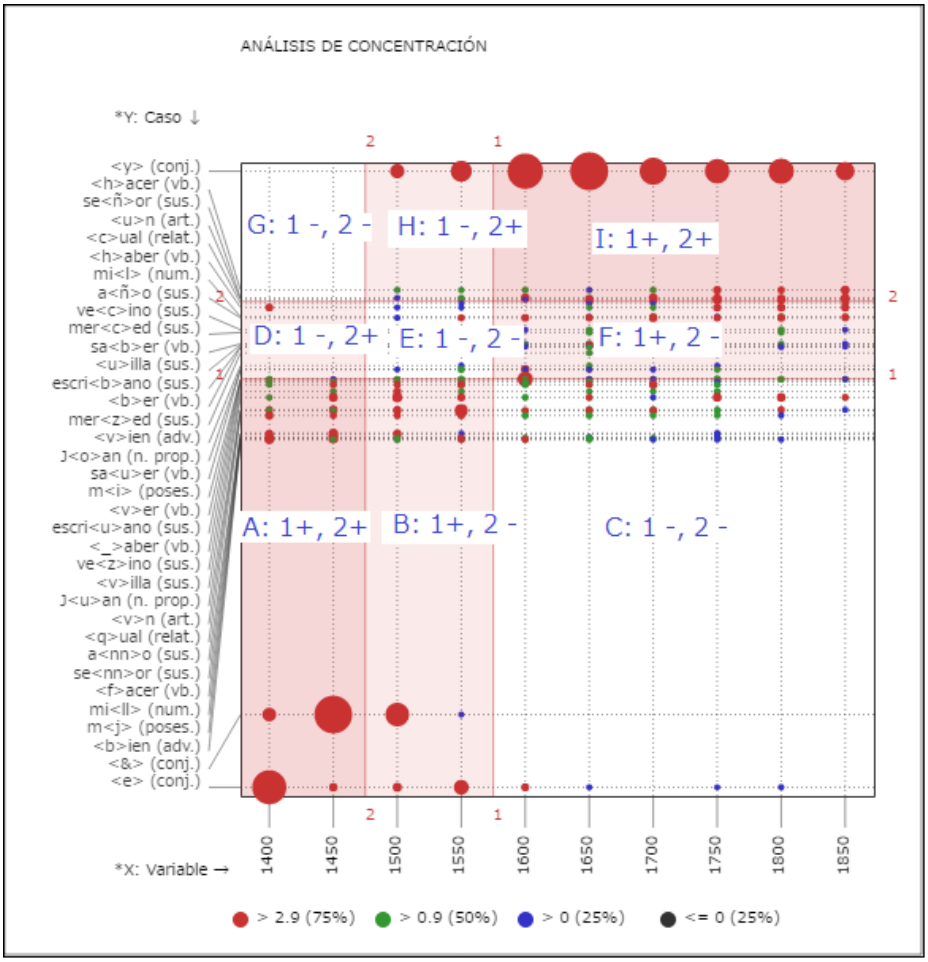


Fig. 4.3.4. División 1, 2

Ahora tenemos una periodización automática de franjas: 1400-1450, 1500-1550 y 1600-1850. La división A es importante en el sentido en que las dos líneas 1 y 2 coinciden en reconocer como la parte positiva: A: 1+, 2+. Lo mismo puede decirse de la división I: 1+, 2+. Las divisiones B, D, F, H son menos importantes, puesto que solo las apoyan una de las dos líneas. En cambio, ninguna línea divisoria apoya las divisiones C, E, G. Cada división contiene sus respectivas formas lingüísticas.

El programa sigue buscando las posibles líneas divisorias. Las terceras líneas divisorias se dibujan entre 1500 y 1550, lo que justifica la periodización cuatripartita: 1400-1450, 1500, 1550, 1600-1850.

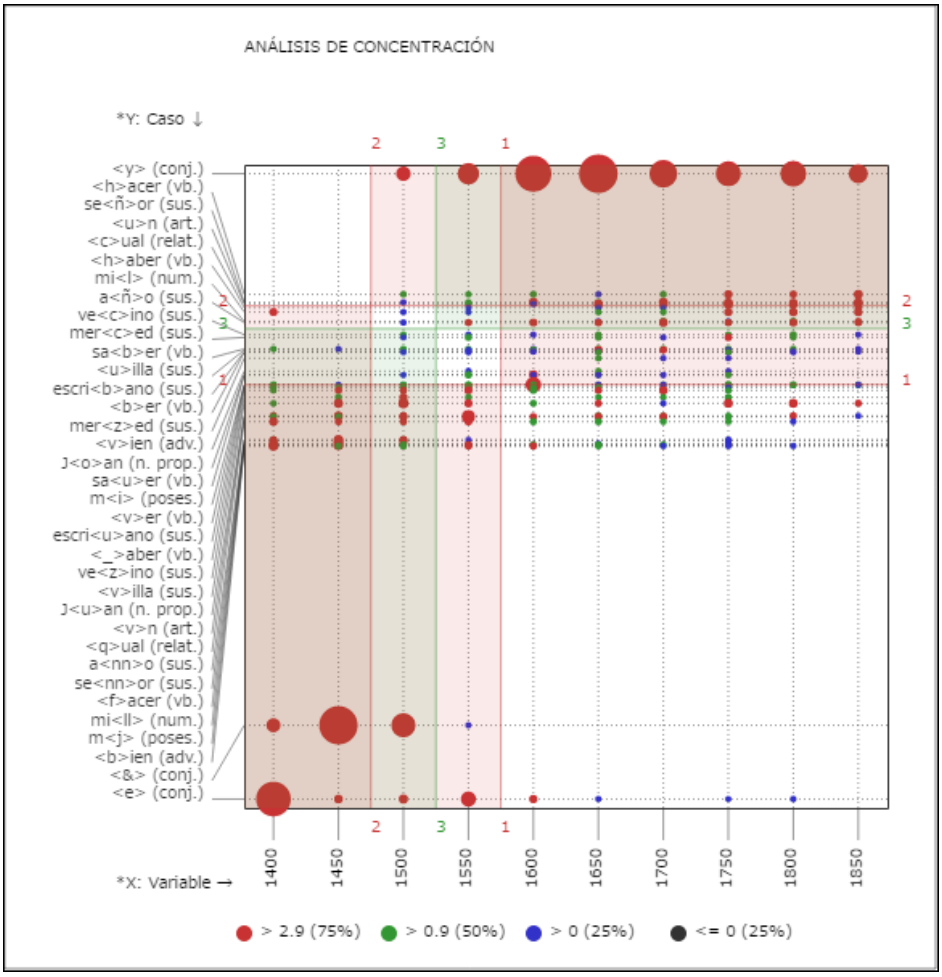


Fig. 4.3.5. División 1, 2, 3

Por las formas variantes más frecuentes, observamos que los mayores cambios se han presentado en 1450, 1500 y 1550. A partir de 1600, la lengua ha continuado con una elevada homogeneidad relativa.

En el análisis de concentración, los ejes reordenados se sitúan según su similitud mutua de distribución. Lo observamos en el gráfico anterior. Con las tres excepciones, <y>, <&> y <e>, la mayoría se sitúan en lugares cercanos. Aunque esta situación refleja bien las relaciones entre formas lingüísticas, no es conveniente para saber los sitios de las líneas divisorias horizontales. El sistema LYNEAL posee la selección de «Ejes equidistantes», que facilita la colocación de todas las formas en el eje con distancias iguales:

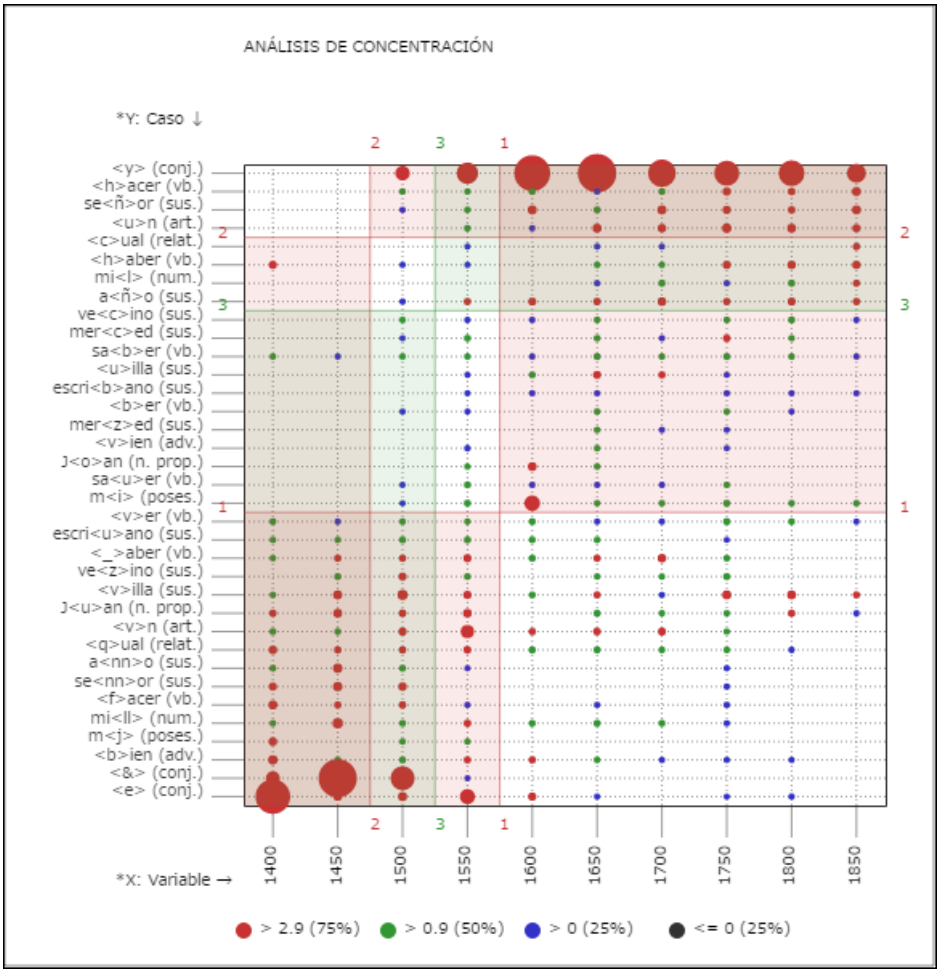


Fig. 4.3.6. División 1, 2, 3. Distancias iguales

De esta manera, ahora podemos ver dónde se trazan las tres líneas divisorias horizontales. El análisis de concentración aclara que las dos formas, *<y>, <h>acer, se<ñ>or* y *<u>n*, que están en la parte superior a la línea 2, pertenecen a los periodos relativamente tardíos, mientras que las formas que se encuentran debajo de la línea 1 aparecen en las franjas tempranas: 1400 y 1450. Entre los dos grupos extremos se encuentran las formas que se sitúan en períodos intermedios.

Estas observaciones están basadas en la distribución de frecuencias de datos, en la que correlacionan la cronología y las formas lingüísticas. La misma correlación no existía en el estado inicial de patronización, sino que se ha conseguido por los cálculos de distancia conjunta de los puntos

reactivos con distintas magnitudes de frecuencia.

Volvemos a observar la situación patronizada en el gráfico de distribución de frecuencia. Estamos ante los cambios cronológicos adecuadamente ordenados:

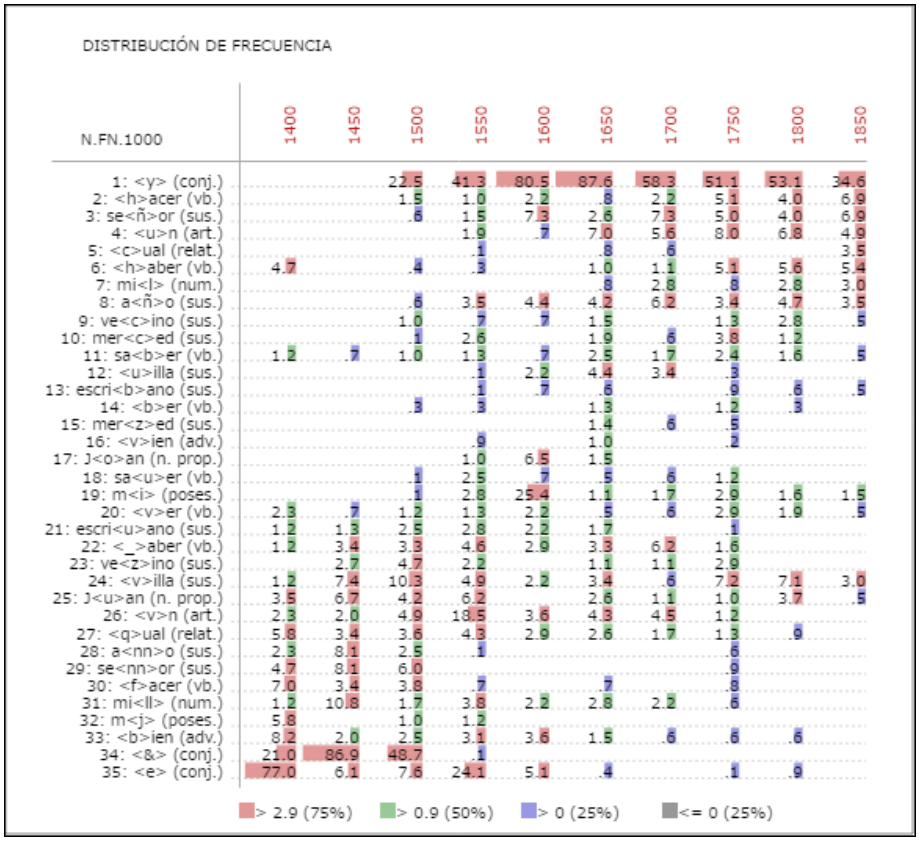


Fig. 4.3.7. Distribución concentrada de frecuencia

5. CONSIDERACIONES FINALES

Apreciamos los estudios históricos anteriores de la lengua española que ofrecen capítulos dedicados a los períodos divididos por sus características lingüísticas y extralingüísticas. Unos utilizan los términos generales de la historia política, social y/o cultural. Otros recurren a la cronología de reinados, siglos o sitios representantes de difusión (Burgos, Toledo, Sevilla). A nuestro modo de ver, son periodizaciones basadas en el «criterio exterior», por tratar cada época en relación con las divisiones de denominación no lingüística.

Por otra parte, pensamos que también es posible llevar a cabo una periodización con el «criterio interno» basada en las distribuciones de los rasgos lingüísticos. El mérito de la periodización interna estriba en que cada periodo caracteriza los rasgos lingüísticamente sin depender de las divisiones preexistentes. Naturalmente en la periodización externa se caracteriza cada época por rasgos lingüísticos, pero la clasificación sigue siendo preestablecida. En su estudio, se nota la dirección que parte de la época para llegar a los fenómenos lingüísticos. Es bien sabido que un rasgo existente en un determinado reinado o siglo no garantiza necesariamente la desaparición del mismo en el siguiente.

Como una posible alternativa, proponemos cambiar de dirección: partir del mundo lingüístico para llegar a los períodos, puesto que nuestro mayor interés está en los cambios o constancias de los aspectos lingüísticos. Actualmente, gracias a la colección enriquecida de datos lingüísticos, que necesita trabajo, tiempo y dedicación, estamos en condición de intentar tomar otro camino.

Hemos observado que los cambios de estado de la lengua no ofrecen divisiones tajantes, puesto que se presentan transiciones y apariciones precursores puntuales y seguidores esporádicos, como ocurre en el sol naciente con luces anteriores de la madrugada o en la puesta del sol con crepúsculo de cierta prolongación. Pero todo esto no niega que haya cambios históricos de estado lingüístico suficientemente definido. Nuestro método de patronización ofrece una solución al problema de la continuidad de cambios lingüísticos. La transición de una forma a otra no se realiza de manera tan nítida como A → B, sino más bien como A →A:B→B, donde se observa la época de coexistencia de la forma antigua y la nueva: A:B. En realidad, el cambio lingüístico tampoco se realiza de manera tan simple como A →A:B→B, sino que siempre presenta frecuencias sumamente variadas. De ahí que venga la necesidad del análisis estadístico, como ocurre en todas las ciencias sociales y naturales y, actualmente, humanas también.

En nuestra experiencia de docentes universitarios, hemos notado que la disciplina estadística construye un punto débil o una barrera difíciles de salvar para los estudiantes de ciencias humanas. Se debe a que en nuestra carrera de filología y literatura casi nunca se ha exigido la asignatura de matemática, estadística ni informática. Por esta razón, hemos pensado que en esta situación es necesario proponer nuestro propio método de cuantificación, fácil de comprender, sin depender del conocimiento previo de matemáticas avanzadas, a diferencia de los métodos aplicados a las ciencias exactas. El cálculo de desviación media dual y el de patronización, tratados

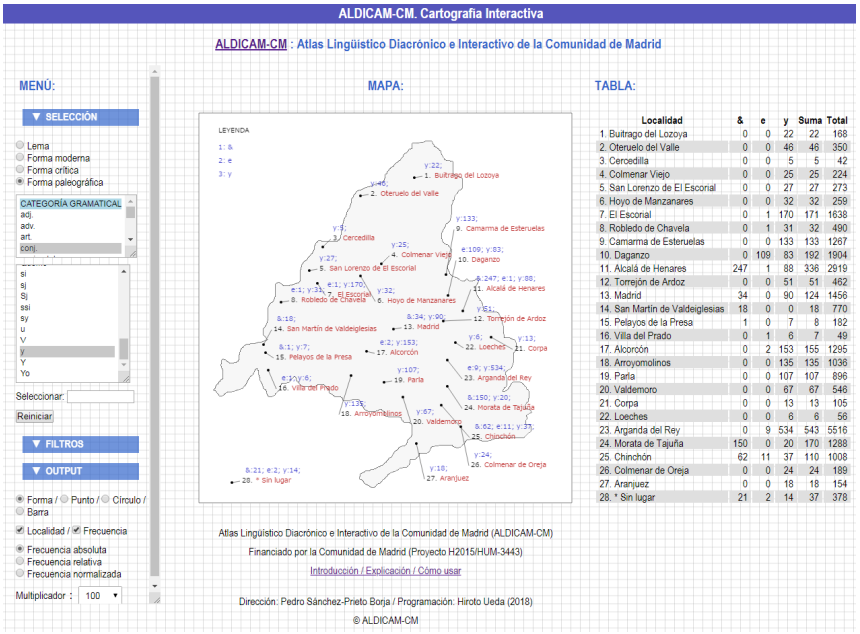
en el estudio actual, son dos ejemplos de ellos (Ueda / Moreno Sandoval, 2017).

El estudio de Sánchez-Prieto Borja y Vázquez Balonga (2018a) comenta que muchas de las opiniones de los autores clásicos tienen un carácter impresionista y otras no han sido valoradas por la crítica en toda su complejidad por no haber recurrido a fuentes documentales. Somos conscientes de que nuestro estudio actual tampoco ha indagado toda la complejidad de la variación cronológica, ni el número actual de documentos es suficiente para abordar un vaciado exhaustivo de todas las formas y todos los lemas existentes en ellos. En este sentido nuestro estudio no puede ser más que un informe intermedio de investigación sobre la historia de grafías en documentos madrileños.

Según Sánchez-Prieto (comunicación personal), dentro de poco la cantidad de documentos tratados en el Proyecto de ALDICAM va a ser de 700 con extensión a períodos anteriores, lo que garantizará la fiabilidad estadística en los posteriores análisis. En el momento en que se hayan reunido estos documentos, intentaremos de nuevo analizar los grandes rasgos de cambios gráficos generales, posiblemente con más minuciosas divisiones cronológicas con intervalo de 25 o 20 años en lugar de 50. Suponemos que no va a haber grandes diferencias de conclusión puesto que este estudio se ha enfocado en los lemas más frecuentes, que se suelen repetir de manera similar en distintos textos en general. Pero ahora con la ampliación de materiales la aproximación a la realidad histórica va a ser más precisa y fiable.

Para dar colofón al presente trabajo, ofrecemos la última noticia del proyecto en colaboración. Estamos elaborando una plataforma de cartografía lingüística, automática e interactiva. Las imágenes que presentamos seguidamente muestran algunas fotos instantáneas de variación geocronológica del lema «y», con formas de <&>, <e>, <y>. Con estas fotos observamos una realidad histórica en distintas épocas divididas por la periodización interna y ahora en la extensión geográfica. Por ellas podemos confirmar el nacimiento de la nueva forma <y> en lugares centrales puntuales en su crepúsculo matutino de la historia y persistencia de las antiguas, <&>, <e>, en las localidades periféricas esporádicas en el vespertino.

Nuestro objetivo final es preparar un inventario léxico exhaustivo de formas variantes y constantes, con índices de frecuencia y desviación, que facilitará estudios históricos seguros con evidencias documentales.



Localidad	&	e	y	Suma	Total
1. Buitrago del Lozoya	0	0	22	22	168
2. Oteruelo del Valle	0	0	46	46	350
3. Cercedilla	0	0	5	5	42
4. Colmenar Viejo	0	0	25	25	224
5. San Lorenzo de El Escorial	0	0	27	27	273
6. Hoyo de Manzanares	0	0	32	32	259
7. El Escorial	0	1	170	171	1638
8. Robledo de Chavela	0	1	31	32	490
9. Camarma de Esteruelas	0	0	133	133	1267
10. Daganzo	0	109	83	192	1904
11. Alcalá de Henares	247	1	88	336	2919
12. Torrejón de Ardoz	0	0	51	51	462
13. Madrid	34	0	90	124	1456
14. San Martín de Valdeiglesias	18	0	0	18	770
15. Pelayos de la Presa	1	0	7	8	182
16. Villa del Prado	0	1	6	7	49
17. Alcorcón	0	2	153	155	1295
18. Arroyomolinos	0	0	135	135	1036
19. Parla	0	0	107	107	896
20. Valdemoro	0	0	67	67	546
21. Corpa	0	0	13	13	105
22. Loeches	0	0	6	6	56
23. Arganda del Rey	0	9	534	543	5516
24. Morata de Tajuña	150	0	20	170	1288
25. Chinchón	62	11	37	110	1008
26. Colmenar de Oreja	0	0	24	24	189
27. Aranjuez	0	0	18	18	154
28. * Sin lugar	21	2	14	37	378

Fig. 5.1. ALDICAM. Cartografía lingüística

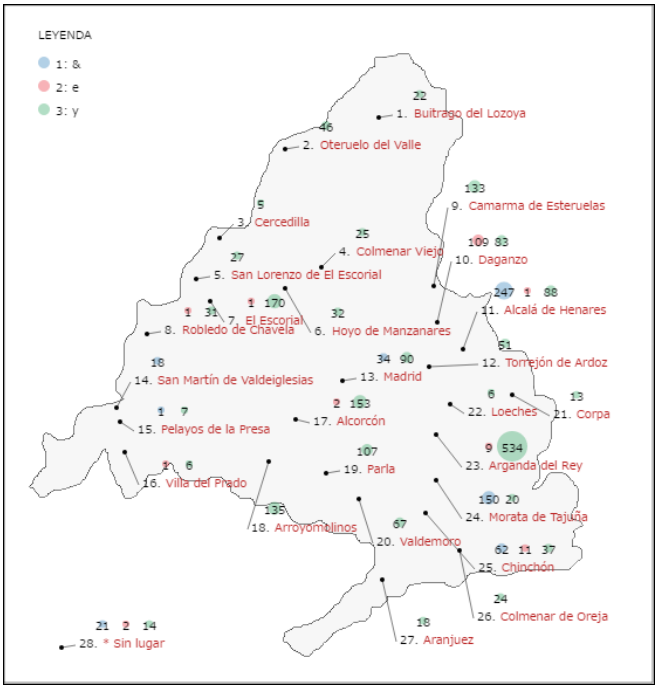


Fig. 5.2. Frecuencia absoluta. Año 1400-1850: 1. <&> / 2. <e> / 3. <y>

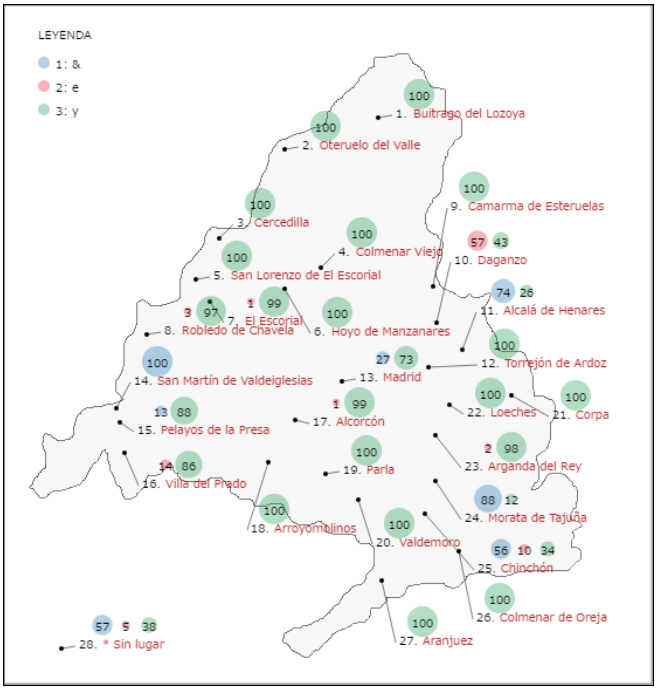


Fig. 5.3. Frecuencia relativa. Año 1400-1850: 1. <&> / 2. <e> / 3. <y>

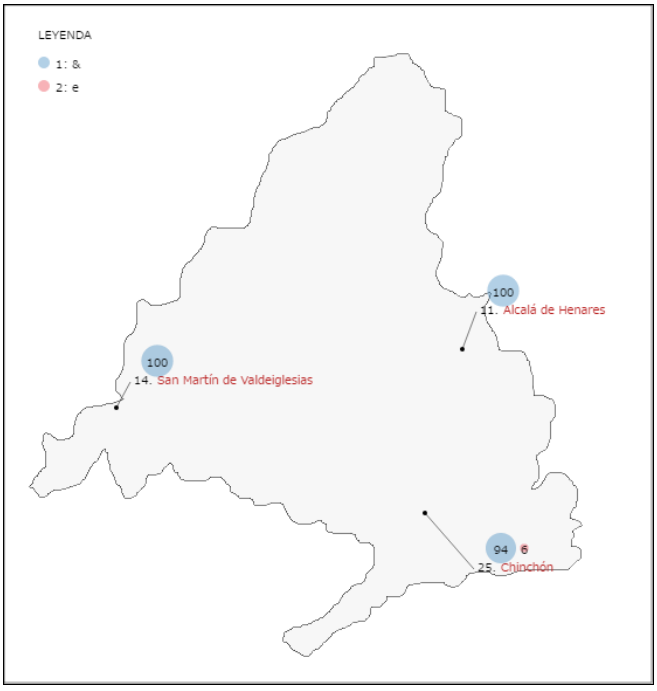


Fig. 5.4. Frecuencia relativa. Año1400-1450: 1. <&> / 2. <e>

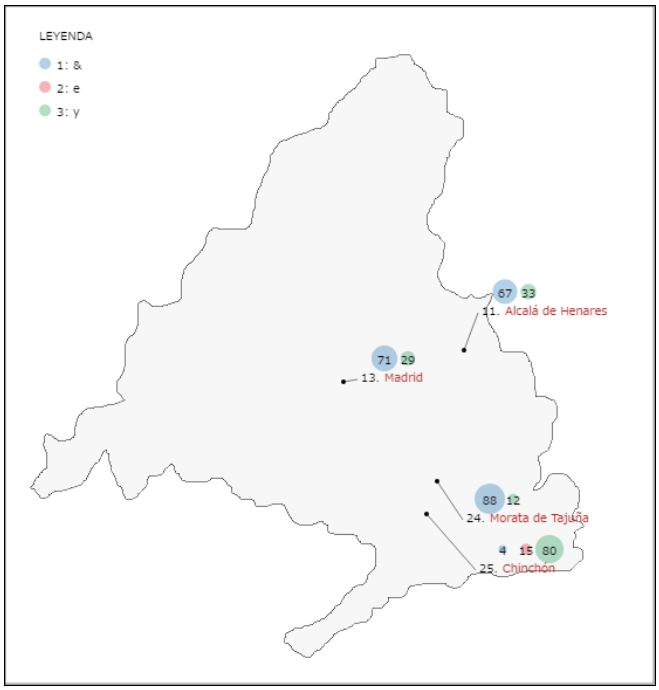


Fig. 5.5 Frecuencia relativa. Año 1500: 1. <&> / 2. <e> / 3. <y>

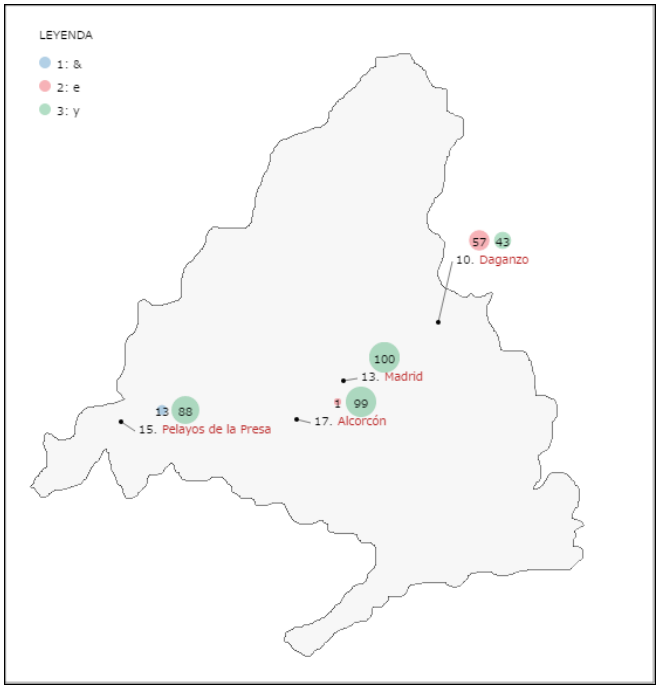


Fig. 5.6. Frecuencia relativa. Año 1550: 1. <&> / 2. <e> / 3. <y>

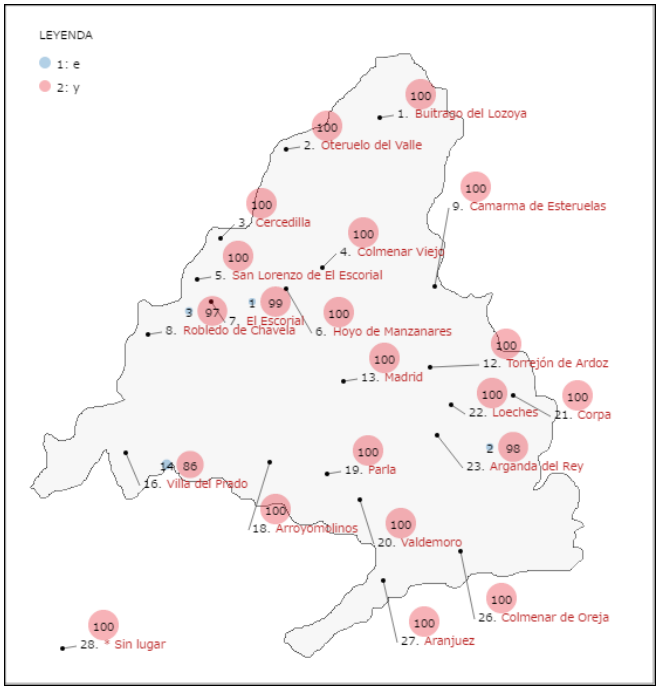


Fig. 5.7. Frecuencia relativa. Año 1600-1850: 1. <e> / 2. <y>

BIBLIOGRAFÍA

Almeida Cabrejas, B. (2014): «Scriptores con bajo y medio nivel socioeducacional en documentos del siglo XIX del Archivo Municipal de Alcalá de Henares: acercamiento a sus usos gráfico» en Díaz Moreno, R. y Belén Almeida (eds.): *Estudios sobre la historia de los usos gráficos en español*. Lugo: Axac, 167-210.

Bertin, Jacques. 1977, (trad.) 1988. *La gráfica y el tratamiento gráfico de la información*. Introducción y versión castellana de Antonio Muñoz Carrión. Madrid. Taurus.

Martín Aizpuru, Leyre / Ueda, Hiroto Ueda. 2018. «Fonema oclusivo velar sordo y sus grafías correspondientes en español. Variaciones y cambios de <k>, <c>, <ch>, <qu> en espacio, tiempo y tipos textuales», comunicación presentada en el *XI Congreso Internacional de Historia de la Lengua Española,* Pontificia Universidad Católica del Perú, 6-10 de agosto, 2018.

Moreno Fernández, Francisco. 1990. *Metodología sociolingüística*. Madrid. Gredos.

_____. 1999. «Análisis cuantitativo de campos léxicos», en José Mauel Blecua / Gloria Clavería / Carlos Sánchez / Joan Torruella (eds.) *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Seminario de Filología e Informática, Universidad Autónoma de Barcelona.

Sánchez-Prieto Borja, Pedro / Vázquez Balonga, Delfina (2016): «¿Seseo en el centro peninsular?», Revista de Historia de la Lengua Española, 10, 201-207.

Sánchez-Prieto Borja, Pedro / Vázquez Balonga, Delfina (2018a): «Toledo frente a Madrid en la conformación del español moderno: el sistema pronominal átono», en Revista de Filología Española, XCVIII, 1º enero-junio 2018, pp. 157-187.

Sánchez-Prieto Borja, Pedro / Vázquez Balonga, Delfina (2018b): «El léxico en los documentos de la Comunidad de Madrid (ss. XVI- XIX)», en Dolores Corbella, Alejandro Fajardo y Jutta Langenbacher (eds.), Historia del léxico español y Humanidades Digitales. Berna: Peter Lang, pp. 343-379.

Ueda, Hiroto. 1993. «División dialectal de Andalucía. Análisis computacional», *Actas del Tercer Congreso de Hispanistas de Asia, Asociación Asiática de Hispanistas,* Tokio.

https://lecture.ecc.u-tokyo.ac.jp/~cueda/kenkyu/chiri/andaluz/andaluz.pdf

Ueda, Hiroto. 2017. «Las grafías <u>, <v> y <b> a lo largo de la historia del español. Análisis separado de frecuencias y análisis conjunto multivariante», presentada en el *V Congreso Internacional de la Red CHARTA* (14 de junio de 2017, Universidad de Lausana).

Ueda, Hiroto / Moreno Sandoval, Anonio. 2017. *Análisis de datos cuantitativos para estudios lingüísticos.*

https://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo/4-numeros/doc/numeros-es.pdf

Vázquez Balonga (2014): *Textos para la Historia del Español* VIII. Archivo Municipal de Arganda del Rey. Alcalá de Henares: Servicio de publicaciones de la Universidad de Alcalá.