DISTRIBUCIÓN DE FRECUENCIAS DE LA LONGITUD DE LAS PALABRAS EN ESPAÑOL: ASPECTOS DIACRÓNICOS Y DE ESTILOMETRÍA

Antonio Frías Delgado
Universidad de Cádiz

Abstract: En el presente trabajo analizamos la distribución de frecuencias de la longitud de las palabras en español comparando textos clásicos de finales del XVI y del XVII con otros del XIX y actuales. Mostramos que en español la longitud de las palabras (i) es un factor estable en la lengua, que (ii) podría ser usado para discriminar géneros, pero que (iii) no es suficiente para discriminar autorías.
Palabras clave: Diacronía, estilometría, lingüística computacional, lingüística de corpus.

I. INTRODUCCIÓN

Hemos abordado en este trabajo un estudio empírico de diversos textos con el fin de analizar cómo se distribuyen las frecuencias de las palabras del español cuando consideramos su longitud.

Diversos motivos podrían hacer que nos interesáramos por la longitud de las palabras que se emplean en un texto.

a) Desde el punto de vista de la estilometría, tal vez pueda resultar un rasgo significativo del estilo de un autor la preferencia por palabras más o menos largas de modo que, junto a otros rasgos, pueda añadir información suplementaria. De ser un factor relevante en la caracterización de un estilo, tendría importancia para resolver problemas de autoría, por ejemplo.

b) Desde un punto de vista diacrónico, sería interesante comprobar empíricamente si los textos mantienen o no alguna diferencia que pueda ser significativa en determinados momentos históricos. De existir, mostraría preferencias cambiantes en el tiempo.

c) Desde el punto de vista de la clasificación y categorización de textos resultaría significativo si pudiera comprobarse empíricamente que hay diferencias marcadas entre géneros.

Además del mero interés de estudio empírico, si la longitud de las palabras fuese un factor relevante tendría el valor computacional añadido de ser una característica fácil y rápidamente medible, sin necesidad de hacer consultas a diccionario ni efectuar análisis detallados y profundos que son computacionalmente muy costosos.

Nuestro interés en abordar este tipo de estudios estuvo motivado por los factores que acabamos de señalar. De un lado, la posible utilidad o no de la longitud de las palabras como factor relevante del estilo y susceptible de ser utilizado como criterio en problemas de atribución de autoría y categorización de textos. De otro, la sensación, como lector, de que los textos clásicos, digamos del siglo XVII, parecen tener un uso más abundante de palabras de longitud mayor que los textos del XIX o los actuales, por ejemplo.

II. METODOLOGÍA

Hemos efectuado una primera aproximación al tema considerando una treintena de textos que serían suficientes para ofrecer diferencias relevantes entre los textos españoles de distintas épocas, caso de existir, y mostrar patrones comunes entre los textos de un mismo autor, caso de que fuese un factor relevante para el estilo.

Los textos han sido obtenidos del *Proyecto Gutenberg*, de la *Biblioteca Virtual Cervantes* y de las ediciones digitales de diversos periódicos españoles (*El País, El Mundo, ABC, La Razón, La Vanguardia, Público*). No hemos incorporado textos de español de América. La Tabla 1 presenta los datos más generales de algunos de los textos usados y que son los que citaremos en este trabajo.

Texto	**Bytes**	**Palab**	**Ratio1**	**Lista**	**Ratio2**	**NSW**	**Ratio3**	**Ratio4**
Alemán	735.606	136.931	4,371	14.381	7,719	59.271	56,714	6,463
DQuijo	1966080	370.242	4,3	22.303	7,854	161.832	56,29	6,369
Avella	735.328	137.382	4,347	12.648	7,116	62.782	54,316	6,315
NovEj	757.636	142.772	4,306	13.122	7,628	61.779	56,728	6,435
Ercilla	725.690	128.868	4,633	13.774	7,641	65.893	48,867	6,564
Lope	66.369	12.167	4,454	3.339	6,674	6.229	48,804	6,209
Gongo	55.996	10.359	4,404	3293	6,419	5.374	48,122	6,062
Espronc	48.897	9.041	4,408	2.609	6,531	4.699	48,025	6,198
Galdós	2133324	394.685	4,04	29.371	8,172	176.647	55,243	6,504
Clarín	1678072	303.925	4,521	22.539	8,021	141.679	53,383	6,577
Pardo	970.752	173.512	4,578	22.843	8,02	83.143	52,082	6,734
Baroja	465.606	84.415	4,515	11.430	7,726	39.670	53,005	6,623
JRJ	84.960	16.457	4,162	3.356	6,698	7.536	54,207	6,01
Conde	152.118	27.781	4,475	5.558	7,508	12.826	53,831	6,619
Prensa	377.707	63.807	4,865	11.577	7,966	30.435	52,368	7,286

Tabla 1.

Leyenda de las columnas:

Texto: referencia al autor o la obra.
Bytes: tamaño del fichero de texto plano.
Palab: número de palabras del texto.
Ratio1: longitud media de las palabras del texto: cociente entre el número de caracteres, sin espacios blancos, y el número de palabras.
Lista: listado de las palabras diferentes que aparecen en el texto.
Ratio2: longitud media de las palabras del vocabulario: cociente entre el número de caracteres de Lista y el número de palabras.
NSW: número de palabras del texto al eliminar las *stopwords* o palabras vacías.
Ratio3: porcentaje de las palabras vacías respecto al total del texto.
Ratio4: longitud media de las palabras del texto sin ocurrencia de palabras vacías.

Textos:

Alemán: Segunda Parte del Guzmán de Alfarache. (1604).
DQuijo: Primera y segunda parte del Quijote, sin prólogos, ni dedicatorias. (1605,1615).
Avella: Quijote de Avellaneda. (1614)
NovEj: Novelas Ejemplares, Cervantes. (1613).
Ercilla: La Araucana (1590).
Lope: Sonetos, exclusivamente. (-1635).
Gongo: Sonetos, exclusivamente. (-1627).
Espronc: Espronceda, colección de poesías de distinta métrica. (-1842).
Galdós: Fortunata y Jacinta. (1886-1887).
Clarín: La Regenta. (1884-1885)
Pardo: Los Pazos de Ulloa y La Madre Naturaleza. (1886-1887).

Baroja: Las inquietudes de Shanti Andía. (1911).
JRJ: Juan Ramón Jiménez, colección de poesías de distinta métrica. (-1958).
Conde: Carmen Conde, Creció espesa la yerba. (1979).
Prensa: Recopilación de editoriales, artículos de opinión, deportes, economía, cultura, ciencia, etc. de distintos periódicos. Textos cortos de múltiples autores. Selección propia. (Enero, 2009)

Los textos han sido procesados para convertirlos en textos planos; se eliminaron los signos de puntuación al no ser de interés en este estudio. Obviamente, dados nuestros fines, no se aplicó ningún lematizador. Para la conversión en texto plano se utilizaron las herramientas suministradas en el NLTK (http://www.nltk.org) además de otras utilidades desarrolladas por nosotros expresamente y escritas en Python. Concretamente, el punto crucial de este proceso inicial, la segmentación en palabras de los textos, aunque no es problemática para el español, se efectuó con el método `wordpunct_tokenize` de NLTK. La corrección de sus resultados se comprobó manualmente en textos al azar.

De cada texto extrajimos su correspondiente listado de palabras, simplemente eliminando las repeticiones. Este listado no es exactamente el vocabulario que emplea el autor en el texto ya que mantiene las formas que sólo se diferencian en número, género, etc.

Para cada texto también construimos una variante sin ocurrencia de palabras vacías.

No existe un único conjunto de palabras vacías (*stopwords*) para el español. La lista de la Real Academia de las palabras más frecuentes, así como algunas listas de palabras vacías de otros proyectos, incluye palabras con significación que no nos parece que debieran formar parte de la misma. A partir de diversas fuentes hemos elaborado personalmente una lista con 382 palabras que incluye las formas de los verbos ser, haber y estar.

Para un análisis de la distribución de frecuencias según el número de sílabas, hemos construido personalmente un contador silábico; dado que lo que nos interesa conocer es sólo el número de sílabas de la palabra, no es necesario agrupar los caracteres en sílabas. Sin necesidad de contar las sílabas, una buena aproximación se obtiene simplemente agrupando longitudes: 1-3:1; 4-5: 2; 6-7:3… Evidentemente, este último procedimiento es inexacto, pero sirve para tener una idea general si no queremos contar expresamente las sílabas.

No hemos eliminado los nombres propios. Hemos supuesto que su efecto sería similar en todos los textos por lo que no afectaría a las comparaciones entre ellos.

La Tabla 2 presenta el porcentaje de palabras hasta una longitud de 14 para una selección de textos.

Ercilla	7,63	20,67	14,68	8,20	12,90	10,89	9,15	6,83	4,35	2,71	1,32	0,52	0,12	0,03
Avllan	7,41	22,81	16,89	10,36	12,60	9,89	7,78	5,26	3,06	2,07	0,93	0,48	0,26	0,13
DQuij	7,71	23,41	17,43	9,68	12,04	9,66	7,84	5,22	3,43	1,93	0,84	0,46	0,22	0,09
Epronc	6,80	25,00	13,78	8,14	12,29	12,08	8,98	6,89	3,36	1,84	0,56	0,20	0,07	0,01
Galdós	6,27	25,38	15,37	11,14	11,78	8,69	7,56	5,22	3,76	2,22	1,37	0,66	0,32	0,15
Almán	7,21	22,85	16,80	10,29	12,62	9,77	7,82	5,34	3,28	2,02	1,03	0,50	0,28	0,11
Gongo	4,99	23,59	15,64	10,29	13,79	12,10	8,20	5,65	3,14	1,50	0,72	0,29	0,09	0,02
JRJ	5,40	27,45	16,01	11,49	12,63	9,84	7,32	4,44	2,75	1,65	0,64	0,25	0,08	0,03
Lope	5,35	23,28	15,78	9,85	12,32	12,16	9,07	5,86	3,36	1,63	0,86	0,30	0,12	0,03
NoEje	7,90	23,44	17,14	9,81	12,02	9,30	7,65	5,59	3,44	2,02	0,83	0,45	0,22	0,11
Bazán	6,55	24,79	14,62	8,97	11,43	9,02	8,23	6,21	4,30	2,88	1,44	0,83	0,43	0,18
Clarín	5,77	24,63	15,37	9,27	12,30	9,37	8,09	6,17	4,24	2,48	1,19	0,62	0,27	0,13
Baroja	6,62	25,00	14,92	8,48	11,79	9,74	8,24	6,04	3,92	2,60	1,34	0,61	0,40	0,20
Conde	6,71	25,69	13,26	10,46	12,00	9,84	7,18	6,10	3,62	2,33	1,30	0,76	0,39	0,18
Prensa	5,43	25,06	15,24	7,51	9,45	8,19	8,26	6,90	5,25	3,74	2,20	1,23	0,71	0,47
AmLib	7,73	23,78	17,48	9,71	11,72	8,92	8,64	5,35	3,12	1,96	0,80	0,39	0,18	0,18
CsEng	7,39	23,23	16,69	11,11	12,50	8,94	8,10	4,79	3,51	2,11	0,78	0,33	0,37	0,06
Celoso	7,97	23,57	16,92	10,13	13,20	9,96	7,19	4,75	2,72	2,03	0,75	0,38	0,26	0,10
LicVid	8,59	22,24	17,94	8,61	11,42	9,80	7,34	5,76	3,91	2,55	0,90	0,50	0,23	0,15
EspIng	8,20	22,47	16,81	9,11	12,06	8,32	8,48	6,93	3,48	2,31	0,90	0,51	0,23	0,09

FuSan	7,57	23,90	17,18	9,45	12,09	8,99	7,66	5,55	3,59	2,05	1,01	0,48	0,30	0,10
Gitan	7,67	23,58	16,90	10,61	11,72	9,96	7,17	5,79	3,23	1,76	0,83	0,44	0,21	0,09
Freg	7,83	23,78	17,22	10,14	12,19	9,50	7,09	5,45	3,45	1,91	0,72	0,37	0,19	0,08
RinCor	8,16	23,48	16,48	10,15	12,26	9,32	6,92	4,95	4,44	2,17	0,76	0,60	0,19	0,08

Tabla 2.

Novelas Ejemplares: El amante liberal, El casamiento engañoso, El celoso extremeño, El licenciado Vidriera, La española inglesa, La fuerza de la sangre, La gitanilla, La ilustre fregona, Rinconete y Cortadillo.

III. ANÁLISIS DE LA DISTRIBUCIÓN DE LAS FRECUENCIAS

III.1. Variabilidad entre lenguas

Un primer dato que hay que tener en cuenta sobre la distribución de la longitud de las palabras es que presenta un rasgo característico importante: varía de lengua a lengua.

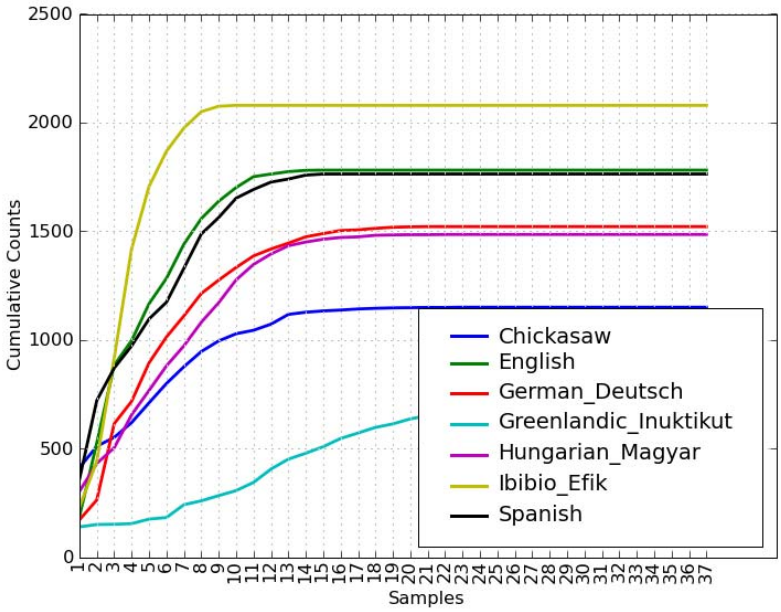

Figura 1A: Frecuencias acumuladas

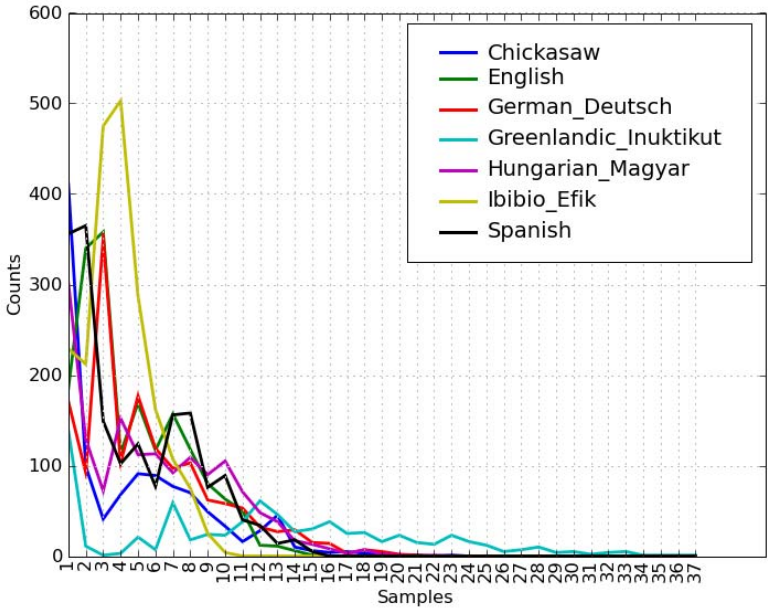

Figura 1B: Frecuencias sin acumulación

Aunque la distribución de las figuras 1A y 1B está creada a partir de un texto corto, como la Declaración Universal de los Derechos Humanos, puede darnos una primera aproximación al tema, a la vez que muestra las diferencias entre lenguas.

La explicación para el comportamiento general que evidencia la figura 1 parece evidente: la distribución va a depender en gran medida de la longitud de las palabras más frecuentes en cada lengua. Para textos españoles la longitud 2 es la más frecuente, en tanto que para textos en inglés, por ejemplo, es la longitud 3 la más frecuente.

La longitud de la palabra escrita, podría pensarse, es una característica bastante arbitraria. Una medida más razonable de la complejidad de las palabras de una lengua podría ser la sílaba. Seguramente este debe ser el caso para objetivos más relacionados con cuestiones psicolingüísticas o más en el espíritu y la línea de los originales trabajos de Zipf. A nuestros efectos de posible medida directa, rápida e inmediata del texto es suficientemente razonable.

III.2. Distribuciones de diccionarios y listados del español

Si consideramos las palabras de un diccionario del español, vemos que muestra la siguiente distribución de frecuencias (Figura 2).

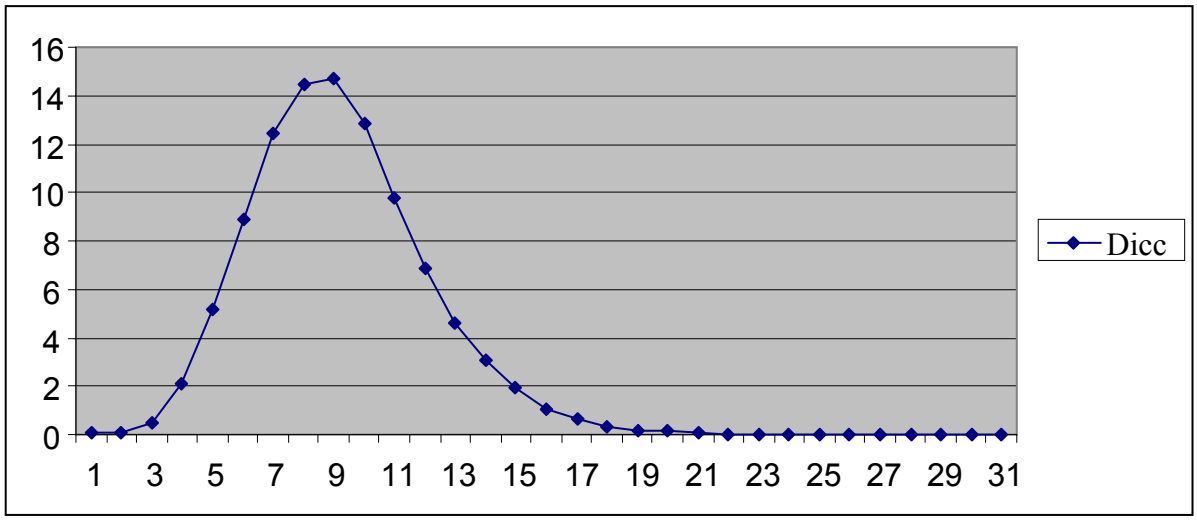

Figura 2.

Hemos utilizado como fuente el diccionario español de FreeLing 2.1 elaborado por la Universidad de Barcelona y la Politécnica de Cataluña. La versión que hemos utilizado consta de 69.994 palabras. Las palabras de longitud mayor de 20 son prácticamente todas términos científicos. En este caso son 107, un porcentaje realmente mínimo. Para un español no científico, podemos asumir razonablemente que 20 es una longitud máxima. Si aceptamos este supuesto, llama la atención el que la distribución tiene el aspecto de una distribución normal, con media de 9,158, desviación típica de 2,812 y coeficiente de variación de Pearson de 0,307.

A partir de las palabras del diccionario construimos una bolsa o listado de palabras, que incluye formas flexivas y derivadas. El listado de palabras de FreeLing 2.1 que hemos usado consta de 556.012 palabras. Su distribución aparece en la figura 3.

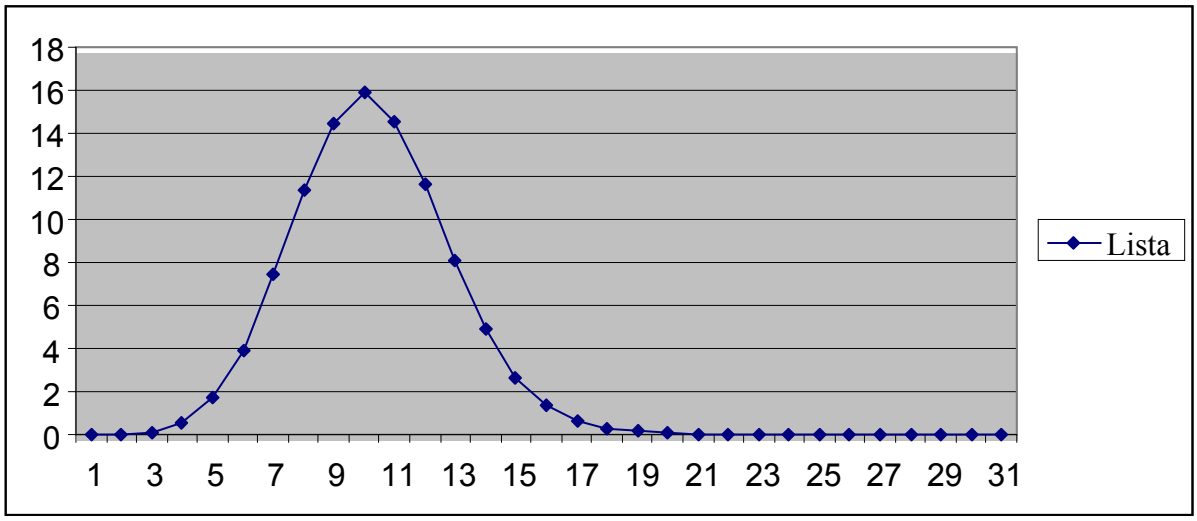

Figura 3.

La figura 4 muestra ambas distribuciones superpuestas.

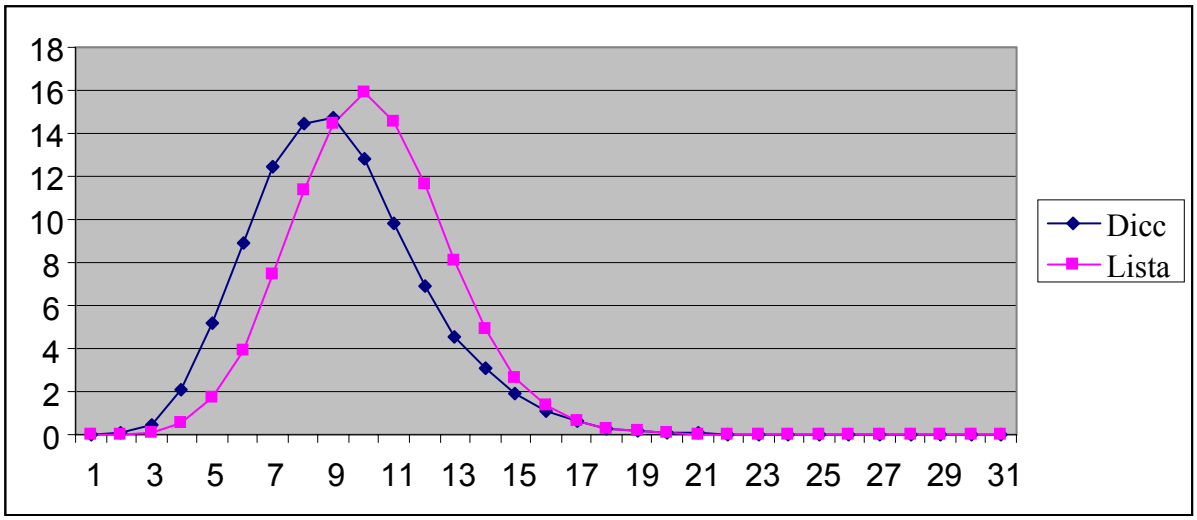

Figura 4.

Como es de esperar, al introducir flexión y derivación la curva se desplaza a la derecha. Llama la atención que el listado de palabras del español, hasta un número razonable de longitud en torno a 20, tiene el aspecto de una distribución normal, con media de 10,233, desviación típica de 2,548 y coeficiente de variación de Pearson de 0,249.

III.3. Distribuciones de textos: aspectos de diacronía

Como es de esperar, el aspecto de la distribución de las frecuencias de la longitud de las palabras de que consta un diccionario o un gran listado, que se aproxima a una distribución normal, varía cuando consideramos los textos completos. La ley de Zipf va a causar un importante cambio.

Los estudios cuantitativos sobre textos en español son menos numerosos y detallados que los que existen para el inglés. Sabemos (Woods, 2001) que las palabras más frecuentes del español del Siglo de Oro no coinciden siempre con las del español actual. "De" no era en el Siglo de Oro la palabra más frecuente en todos los textos con gran diferencia, como ocurre hoy. En *El Quijote*, por ejemplo las palabras más frecuentes son "que" -20.090- e "y" -17.704-. Pero, ¿afectan estos cambios también a la distribución de las frecuencias de la longitud en general?

Consideremos un texto clásico, *Don Quijote,* cuya distribución de frecuencias acumuladas de las longitudes de las palabras está representada en la figura 5.

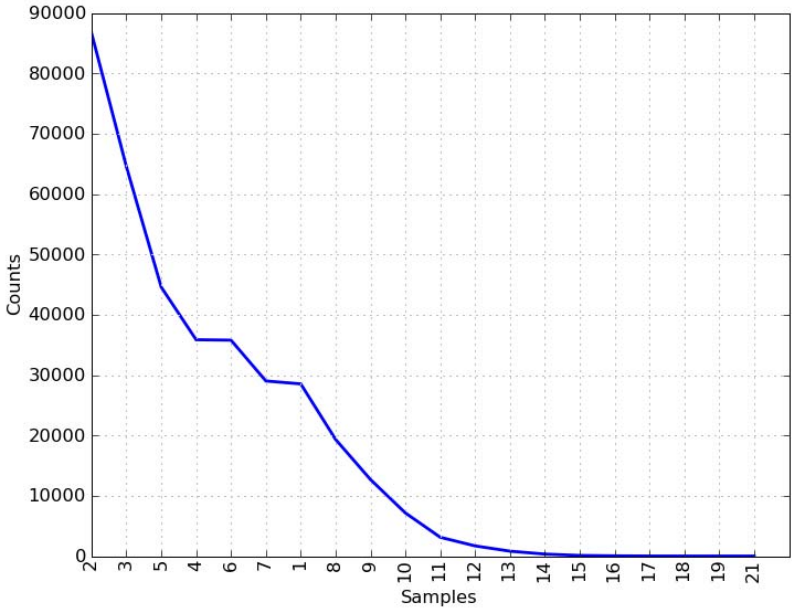

Figura 5.

Como referencia comparativa de otra lengua, la figura 6 muestra también la distribución para la traducción inglesa del *Quijote* del Proyecto Gutenberg.

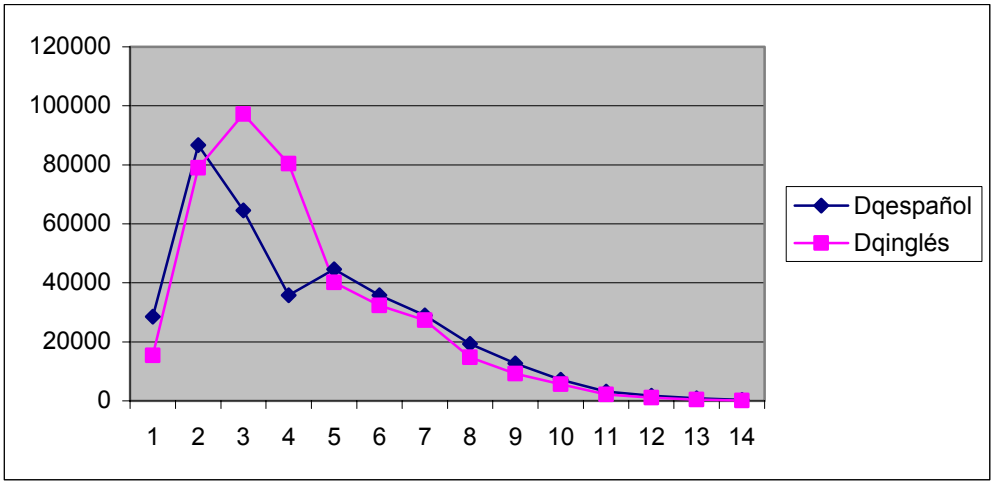

Figura 6.

Reparemos no sólo en la diferencia de la curva, sino en que los valores más frecuentes en cada una de las lenguas son distintos.

Nos haremos una mejor idea de la distribución si en vez de considerar frecuencias acumuladas, representamos los valores que toma cada longitud (figura 7).

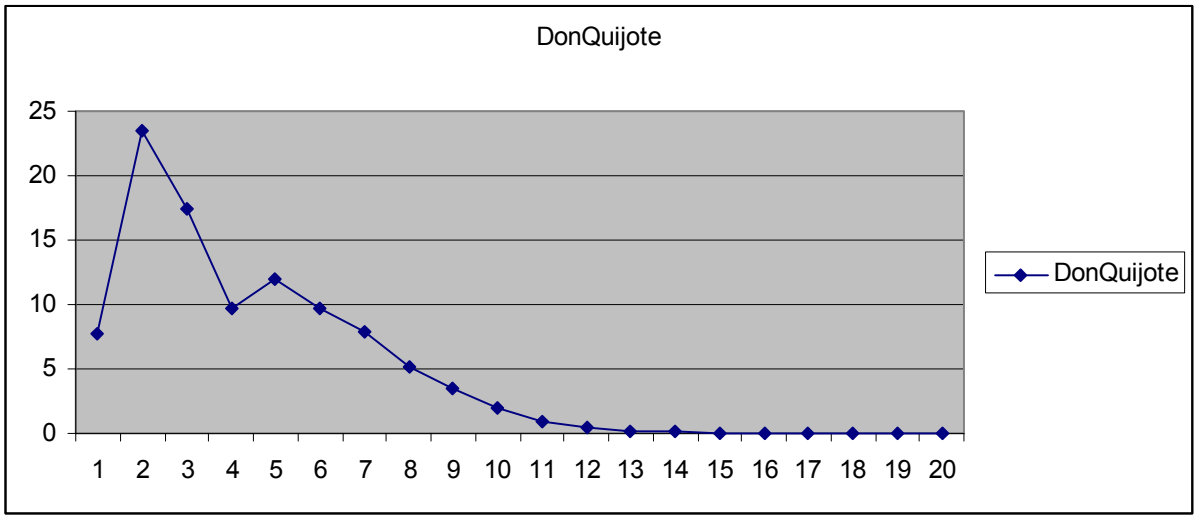

Figura 7.

Dado que los textos tienen extensiones muy variables, en este trabajo, excepto en la figura 6, siempre hemos representado los porcentajes en vez de los valores absolutos para permitir las comparaciones entre textos.

Aunque hay una palabra de 21 caracteres, "bienintencionadamente", de hecho el porcentaje de palabras con una longitud mayor de 12 es muy pequeño. Si efectuásemos una agrupación razonable de caracteres, por ejemplo, considerando las palabras según su número se sílabas en vez de su longitud, el gráfico recuerda a los típicos de la ley de Zipf, aunque con una caída más suave, como muestra la figura 8.

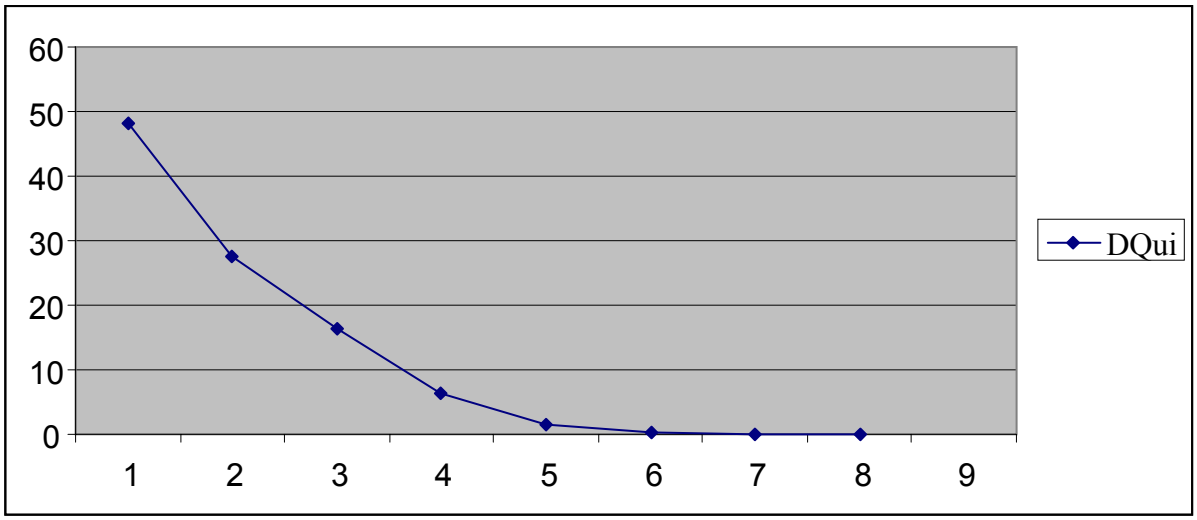

Figura 8.

Éste es un dato que hemos encontrado bastante constante en todos los textos que hemos analizado. En torno al 47% de las palabras de un texto constan de 1 sílaba; en torno a un 27%, de 2 sílabas; en torno a un 17% de 3 sílabas; la variabilidad es mayor a partir de las palabras de 4 sílabas: sobre un 6-7% para 4 sílabas y 1,5% para 5 sílabas; el total de las palabras de más de 5 sílabas supone apenas un 0,2%.

Una comparación de la distribución del *Quijote* de la figura 7 con las distribuciones de obras similares, las de Mateo Alemán y Avellaneda, se presenta en la figura 9.

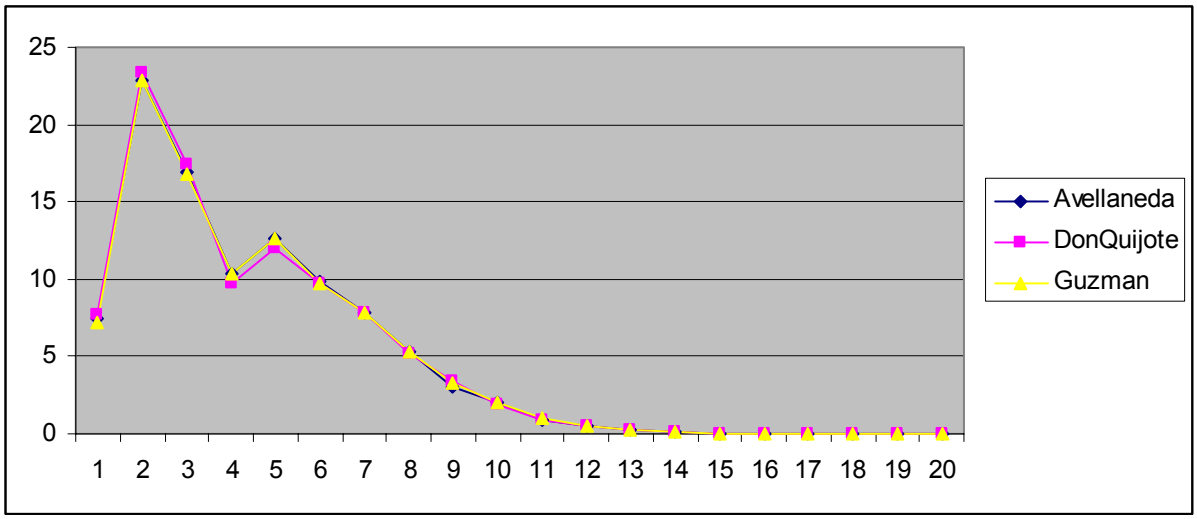

Figura 9.

Como podemos comprobar, el español de los grandes textos clásicos ofrece una distribución muy similar, atendiendo a la longitud de las palabras.

Podríamos pensar que las características propias de la poesía, sus reglas métricas, tienen importantes efectos diferenciadores respecto a la novela. La figura 10 muestra la distribución de *La Araucana* y colecciones de sonetos de Lope de Vega y de Góngora.

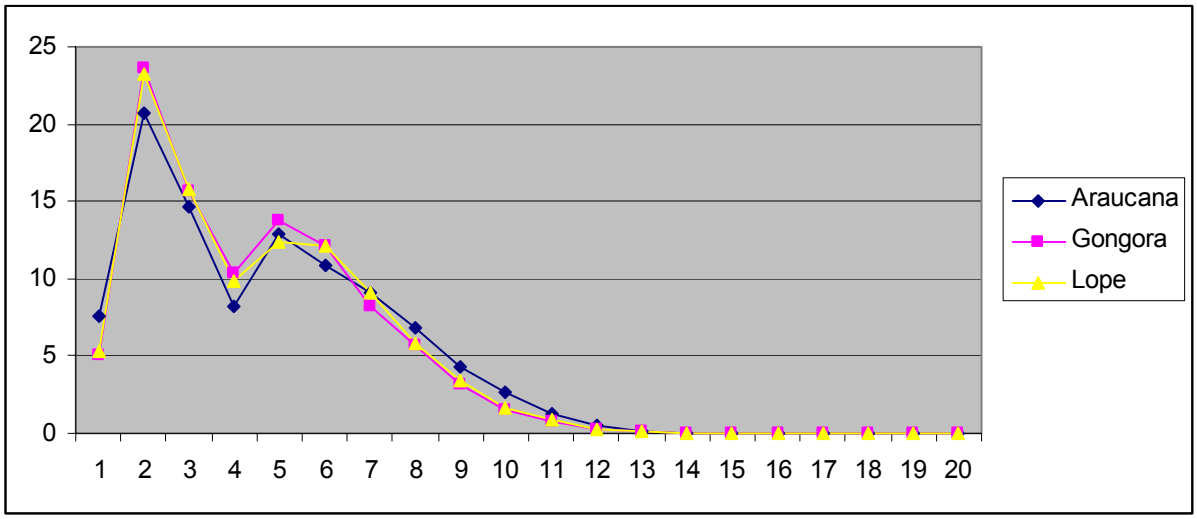

Figura 10.

Puede observarse, sin embargo, que la distribución es muy similar.

Comparemos el texto de *Don Quijote* con este otro de finales del XIX, *Fortunata y Jacinta* (figura 11).

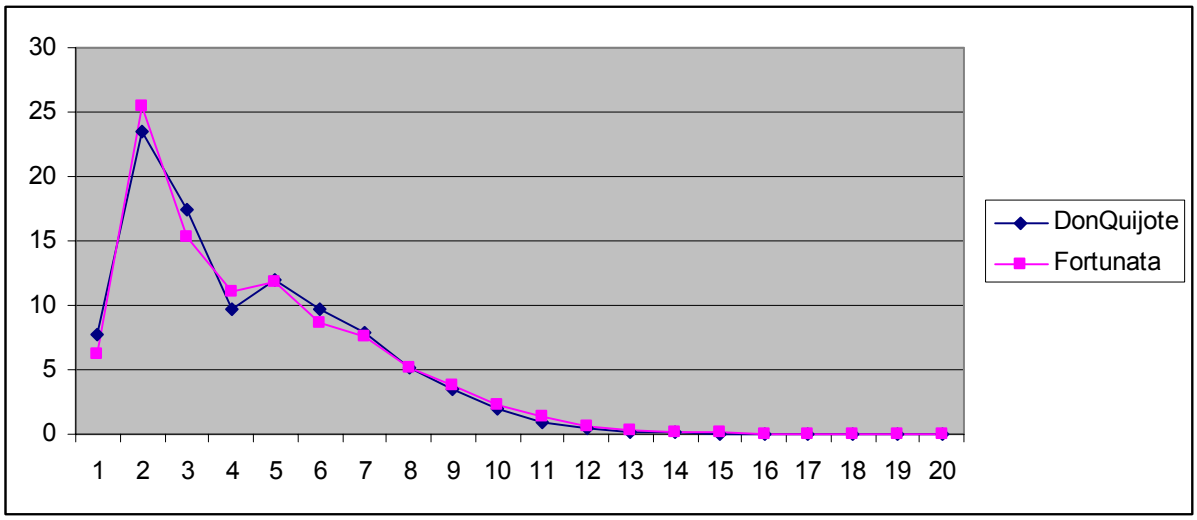

Figura 11.

Desde luego que no hay diferencias tan marcadas como para que a simple vista pueda considerarse que ha habido un fuerte cambio en la lengua en casi tres siglos. La situación es similar cuando comparamos la obra de Cervantes con las de Baroja, Clarín y Pardo Bazán, como muestra la figura 12.

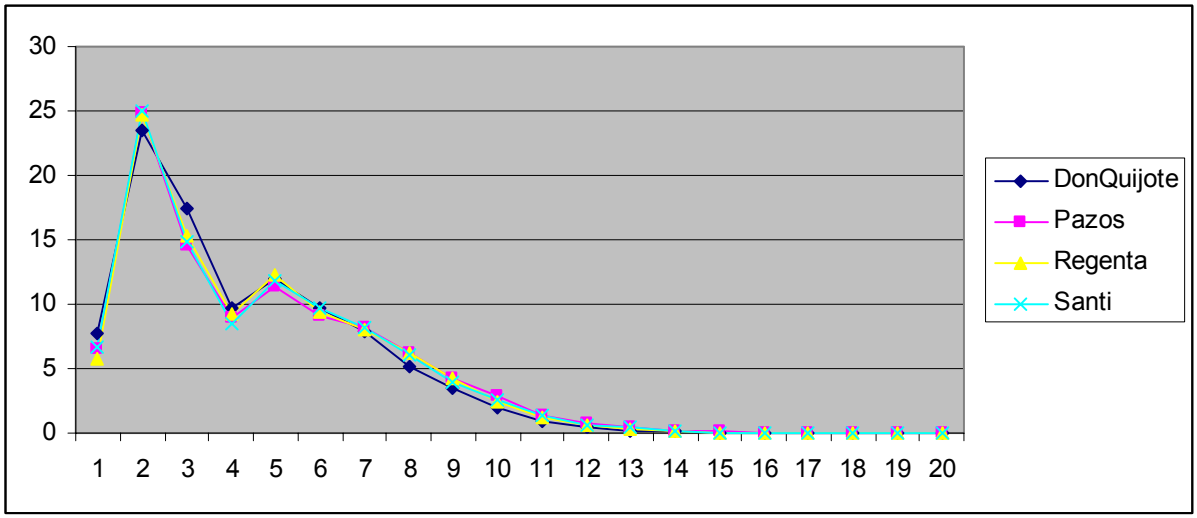

Figura 12.

Si analizamos la distribución de la figura 13 en la que se compara un texto clásico, otro actual y una colección de textos cortos de múltiples autores y géneros, veremos que las diferencias no pueden achacarse a un cambio fuerte en las preferencias de los autores de distintas épocas, sino más bien al factor género.

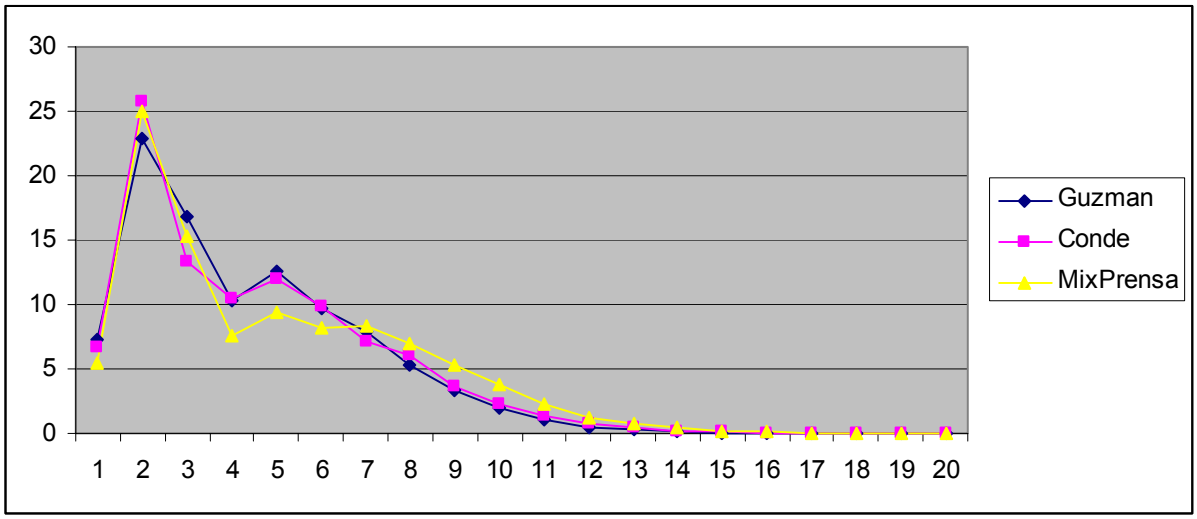

Figura 13

La figura 13 presenta al menos dos rasgos dignos de consideración.

a) Por una parte, una mayor frecuencia de palabras de gran longitud. Cuando se inspeccionan manualmente las palabras de mayor longitud se comprueba que casi siempre se debe a la incorporación de términos científicos y técnicos, así como a palabras compuestas que no existían en el español clásico ("audiovisuales", "socialdemocracia", etc.).

b) Por otra parte, la menor frecuencia de palabras de longitud 4 y 5. Al examinar estas palabras no aparece ninguna anomalía que implique cambio en las preferencias de la lengua de una época histórica a otra. Se trata más bien de características propias del género periodístico. Comparando la distribución de frecuencias de las sílabas se observa mejor la diferencia, como aparece en la figura 14.

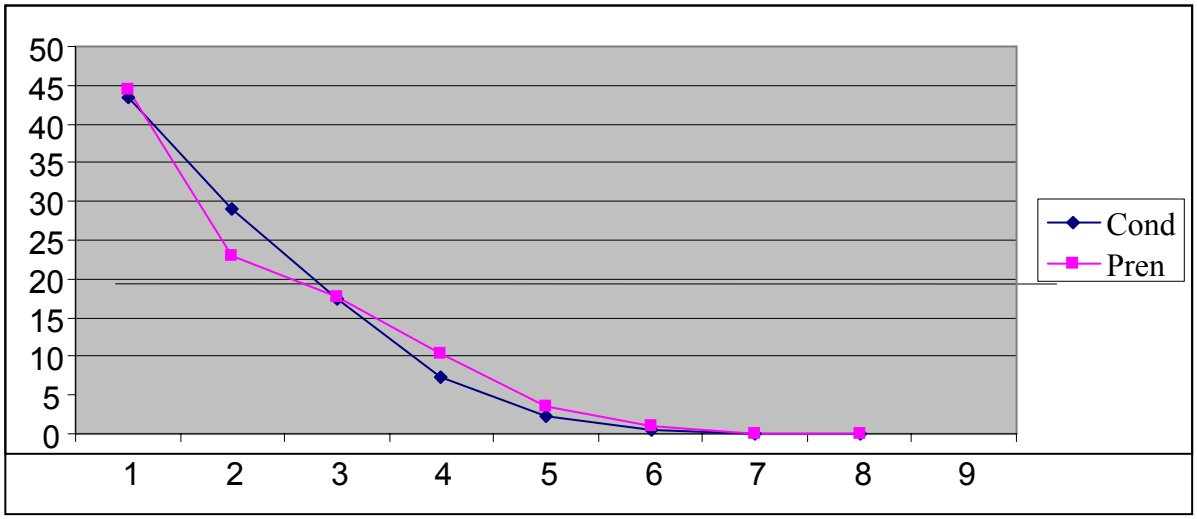

Figura 14.

Puede observarse que el efecto se centra en una disminución de las palabras de 2 sílabas. En general, las palabras de dos sílabas juegan un papel importante en las diferencias de distribución entre textos en español. Para el caso que nos ocupa, una posible hipótesis explicativa podría ser la de que (i) la concisión a que está obligado el texto periodístico corto aumenta la frecuencia de términos técnicos, más informativos, de longitud larga; (ii) la ley de Zipf explica la constancia de la frecuencia de palabras de una sílaba; (iii) el uso culto no especializado general de la lengua española, que podemos suponer es el de la novela, tiene un alto porcentaje de palabras bisílabas. Una exageración, efectista para titular de prensa, sería: el español coloquial es bisílabo.

Volviendo al trabajo de Woods (2001) citado antes, creemos que habría que analizar en más profundidad si algunas de las diferencias observables se deben a un cambio histórico en el español o más bien a características del género periodístico.

Como era esperable, si analizamos la distribución de frecuencias en los textos a los que hemos eliminado las palabras vacías, nos da una gráfica más semejante a la de una distribución normal, como si fuese la de un mero listado. (Figura 15)

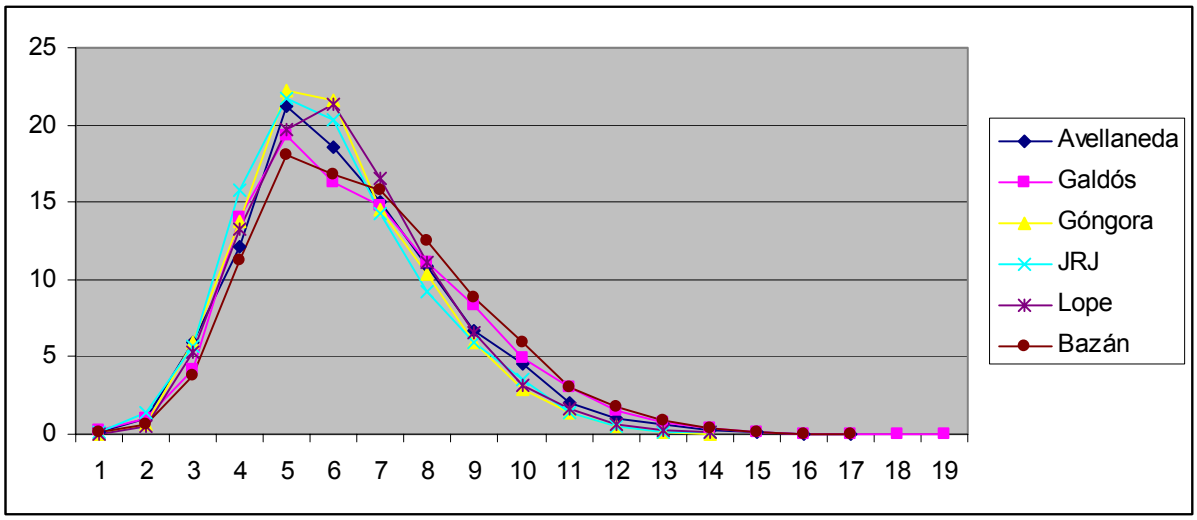

Figura 15.

Dijimos antes que la longitud media de las palabras de un diccionario español puede aproximarse a 9. Viendo los datos de la Tabla 1, observamos que la longitud media de las palabras para textos suficientemente largos es bastante estable, en torno a 4,4; sea prosa o poesía. También es bastante estable la longitud media de las palabras cuando eliminamos de los textos las palabras vacías, excepto en el caso de los textos de prensa; en torno al 6.2. Los textos poéticos, como era esperable, tienen un porcentaje menor de palabras vacías. Juan Ramón Jiménez es una excepción para la que no encontramos hipótesis explicativa. El menor porcentaje de palabras vacías no se debe a la longitud de los textos, como lo prueba el caso de *La Araucana*. Puede observarse también que hay diferencias entre poesía y prosa en la longitud media de las palabras descontando las repeticiones.

III.4. Distribuciones de textos: cuestiones de estilometría

En la atribución de autoría por procedimientos cuantitativos hay dos procesos básicos: (a) el conjunto de factores o rasgos que pensamos identifica el estilo del autor y (b) los algoritmos para su tratamiento. Como rasgos de estilo se han propuesto decenas de características. La longitud de las palabras figura entre las primeras que aparecen, ya en el siglo XIX. Aunque

los estudios empíricos sobre textos ingleses no avalan que la longitud de las palabras sea un rasgo discriminante (Grieve, 2007), hemos querido comprobar empíricamente lo que ocurre con textos en español.

Las distribuciones que hemos venido considerando, como puede comprobarse a partir de los datos reflejados en la Tabla 2, muestran que la longitud de las palabras es un factor con escaso poder discriminatorio entre autores. No puede demostrarse que, respecto a este factor, haya una consistencia entre diversos textos de un mismo autor que los diferencien marcadamente de los de otro autor. La figura 16 muestra la distribución de algunas novelas ejemplares.

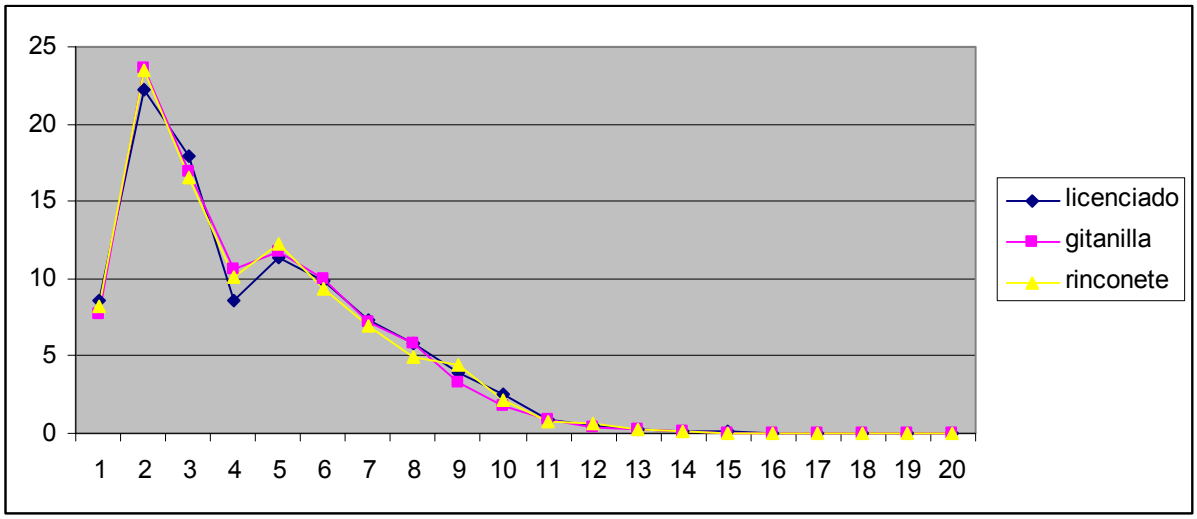

Figura 16.

Si usamos técnicas analíticas más sofisticadas, el resultado no mejora. Cualquier métrica de distancia o de diferencia que usemos mostrará importantes diferencias entre diversos textos de un mismo autor a la vez que similitudes entre textos de diversos autores. Por ejemplo, cualquier métrica razonable muestra una diferencia mayor entre *Don Quijote* y *El licenciado Vidriera*, que entre *Don Quijote* y *Guzmán de Alfarache*.

La figura 17 muestra la distribución comparada de *El Licenciado Vidriera* y el texto que Adolfo de Castro pretendió hacer pasar por cervantino, *El Buscapié*. Como puede observarse, la diferencia es mínima.

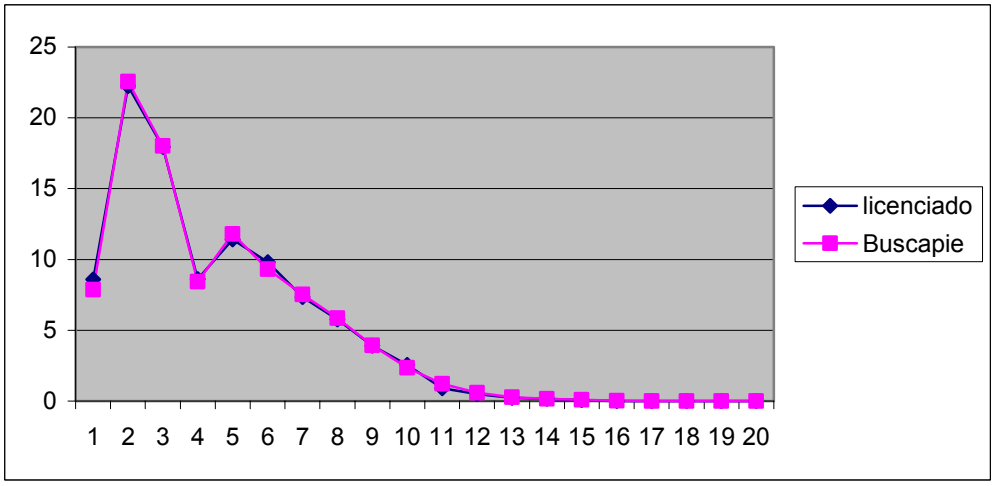

Figura 17.

IV. CONCLUSIONES

Volviendo a nuestras hipótesis de partida, nuestro análisis parece avalar:

a) Que la lengua española se ha mantenido bastante estable desde el siglo XVI hasta ahora en lo que se refiere a la frecuencia de distribución de las longitudes de las palabras.

Si un lector de textos clásicos al encontrar palabras como "dábanoslo" se siente tentado de conjeturar que el español clásico tenía una mayor frecuencia de palabras largas, no parece ser éste el caso. Al contrario: en el español actual hay un gran número de palabras técnicas y compuestas que han alargado el idioma. La sensación del lector de textos clásicos puede deberse a un mayor coste de procesamiento ante palabras más infrecuentes; "socialdemocracia" es casi el doble de larga y la procesamos con menor coste.

b) Que la longitud de las palabras usadas en un texto en español no constituye un factor estable como para discriminar problemas de atribución de autoría.

c) Que la frecuencia de distribución de la longitud de las palabras podría ser usada como factor para categorizar textos por género.

d) Que las palabras bisílabas juegan un papel importante en la lengua española como factor discriminatorio, por razones que desconocemos y de forma que debe ser analizada en más detalle.

V. REFERENCIAS BIBLIOGRÁFICAS

Bird, S., Klein, E., and Loper, E. (2008). *Natural Language Processing*. http://www.nltk.org/book

Grieve, J. (2007). Quantitative authorship attribution: an evaluation of techniques. *Literay and Linguistic Computing*, 22, 251-270.

Holmes, D. (1994). Authorship attribution. *Computers and the Humanities,* 28, 87-106.

Holmes, D. (1998). The Evolution of Stylometry in Humanities Scholarship, *Literary and Linguistic Computing*, 13, 111-117.

Koppel, M. And Scheler, J. (2009). Computational methods in authorship attribution, *Journal of the American Society for Information Science and Technology*, 60, 9-26.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval,* 3, 233-334.

Luyckx, K. and Daelemans, W. (2008). Authorship Attribution and Verification with Many Authors and Limited Data. In *Proceedings of the 20th Belgian-Netherlands Conference on Artificial Intelligence (BNAIC 2008)*, 335-336.

Woods, M.J. (2001). Spanish word frequency: a historical surprise. *Computers and the Humanities*, 35, 231-236.