# EL ANÁLISIS LEXICOMÉTRICO DEL CRECIMIENTO DEL VOCABULARIO: ESTADO DE LA CUESTIÓN Y NUEVAS PERSPECTIVAS

*Vicente Sabido*

Departamento de Filología Española
Facultad de Filosofía y Letras
Universidad de Granada
18071 Granada (España)
E-MAIL: vsabido@ugr.es

**ABSTRACT**

*Un fenómeno frecuente en lexicometría y estilometría es el de la proliferación de estudios teóricos y metodológicos que, después, raramente se aplican a textos reales. El presente trabajo pretende ser una contribución puntual al aplicar a textos literarios hispánicos un modelo probabilístico del crecimiento léxico escasamente explotado hasta el momento.*

*Palabras-clave: lingüística computacional, lexicometría, estilometría.*

Dentro del vasto campo de la lexicometría, enriquecida en las últimas décadas por nuevos modelos estadísticos, uno de los asuntos tal vez menos estudiados sea el del crecimiento del vocabulario[1] en textos en lenguaje natural.

La noción de lo que denomino «crecimiento del vocabulario» es bastante simple. Cualquier discurso lingüístico consta de una secuencia de *tokens* que, de ordinario, ni son todos iguales ni son todos diferentes. El examen de un índice de un texto nos revela la existencia de una serie de *types* que aparecen una sóla vez en el discurso (los *hapax legomena*), otros que aparecen dos veces, tres, etc.

Ilustraré lo dicho con un breve texto tomado de la *Sonata de invierno* de Valle-Inclán:

> «Como soy muy viejo, he visto morir a todas **las** mujeres por quienes **en** otro tiempo suspiré **de** amor: **De una** cerré los ojos, **de** otra tuve **una** triste carta **de** despedida, y **las** demás murieron siendo abuelas, cuando ya me tenían **en** olvido.»

He señalado en negrita aquellos *types* ('las', 'en', 'de', 'una') que exhiben más de una ocurrencia en el fragmento. El resto son *hapax legomena*. A medida que el texto se alarga muchos *hapax* van dejando de serlo y pasan a engrosar la lista de *types* de frecuencia superior a 1, cosa que ocurre en el fragmento expuesto al llegar al *token* número 20 'De', que ya había aparecido en la posición 18 y que vuelve a ser usado dos veces más.

Si llamamos N al número de *tokens* del texto y V al número de *types*, en el fragmento anterior N=48 y V=33.

---

[1] Entiendo por *vocabulario* el conjunto de palabras que, únicas o repetidas, constituyen un texto. Por simplificar, asumo en adelante *palabra* como una tira de caracteres delimitada por espacios y signos de puntuación. Esta opción metodológica, lingüísticamente discutible, resulta para mi actual propósito suficientemente productiva. Utilizaré en lo sucesivo los anglicismos *type* y *token* para referirme, respectivamente, a cada palabra diferente que forma parte de un discurso y a las ocurrencias o realizaciones textuales de la misma.

Se trata, pues, de estudiar la forma en que el vocabulario (V) se constituye dinámicamente (de 1 a V) a lo largo del texto, desde el *token* primero hasta el último (de 1 a N).

Si encontramos (o construimos artificialmente) un texto de longitud suficiente en el cual todas las palabras que lo forman sean distintas, tendremos que:

$$f = N / V = 1, \text{ ya que } N=V.$$

Si en otro texto *sui generis* todas las palabras son iguales (imaginemos una «letanía emotiva» del tipo: 'oh, oh, oh, oh, oh, oh, oh, oh, oh...') resultará que:

$$f = N / V = N, \text{ ya que } V=1.$$

Entre ambos extremos ficticios se sitúan los discursos reales en lenguaje natural.

El problema del crecimiento del vocabulario es una cuestión de estructura, no de contenido. Lo que se trata de analizar es el grado cuantitativo de aportación de nuevos *types* a lo largo del discurso.

Partimos de la hipótesis, fácilmente intuible, de que al aumentar N aumenta igualmente V, pero no al mismo ritmo. Mientras que el crecimiento de N es constante, el de V disminuye según el texto avanza. El hablante, el escritor, repite forzosamente palabras ya utilizadas y, si el texto se extiende *ad infinitum*, llegará al momento en que las aportaciones léxicas se agoten. No obstante, en discursos reales, puede decirse que ningún hablante o escritor agota su léxico virtual.

Es obvio, por otra parte, que para estudiar si ese crecimiento léxico obedece a algún tipo de ley o tendencia regular, es preciso el cotejo de dicho crecimiento en textos reales con um modelo matemático.

Ch. Muller ha propuesto un modelo teórico del crecimiento del vocabulario basado en la aplicación de la distribución binomial. Hay que decir que tal vez la más importante contribución de Muller al desarrollo de la lexicometría consiste precisamente en el hallazgo de la productividad de esa distribución estadística aplicada a las cuestiones de riqueza, crecimiento, conexión y especialización del vocabulario. Trataré de resumir su razonamiento:

Consideremos un texto cualquiera, cuyo desarrollo esquematizamos en una línea continua que va de 1 a N:

1_____A_____N

La probabilidad teórica -considerando una distribución aleatoria- de que un vocablo cualquiera de frecuencia f se encuentre exclusivamente entre A y N, es de:

$$\left(\frac{N-A}{N}\right)^f$$

Si f=1 la fórmula se comprenderá con más facilidad, ya que entonces su probabilidad de aparecer por vez primera después de A es de (N-A)/N: así, en un texto de N=1000 palabras, cualquier vocablo único, o sea de frecuencia 1, a partir de A=500, es decir, la mitad exacta del texto, tiene una probabilidad teórica de aparecer por vez primera de:

(1000-500)/1000=0,5

de modo que, si en dicho texto constatamos la existencia de 1000 vocablos de frecuencia 1, habría que esperar que la mitad de ellos aparecieran por vez primera en la primera mitad del texto y el resto en la segunda.

Para un vocablo de frecuencia 10 en el mismo texto, la probabilidad de aparecer por primera vez después de la palabra nº 500 sería de:

$$\left(\frac{1000-500}{1000}\right)^{10} = 0{,}00097$$

o sea, una probabilidad prácticamente nula. La contraria, o sea, la de aparecer antes de la palabra nº 500 sería igual a (1-0,00097)= 0,99903. Consideremos una palabra muy frecuente en casi cualquier texto castellano: la preposición 'de'. ¿Es imaginable en condiciones normales que en un discurso de cierta extensión dicha preposición tenga su primera ocurrencia en la página 148 de una novela de 150? Pero sí es imaginable que en esa misma novela cualquier *hapax legomena*, por el hecho de serlo, pueda aparecer con la misma probabilidad (0,5) en la primera o en la segunda mitad del volumen.

Hay que concluir por el momento que, para un *type* cualquiera, la probabilidad de aparecer por vez primera después de una posición A en un discurso de N *tokens* depende básicamente de su frecuencia en el mismo.

Y si agrupamos los vocablos por clases de frecuencias (f=1, f=2, f=3... f=n), el efectivo teórico de los vocablos que aparecerán por primera vez entre 1 y A será de:

$$V_f * \left(1-\left(\frac{N-A}{N}\right)^f\right)$$

Veamos un ejemplo. En el relato borgiano «La casa de Asterión» (N=862 ; V=389), deberían haber tenido su primera ocurrencia entre los *tokens* 1 y 300, según este modelo teórico binomial:

| f | $V_f$ | $V_f * \left(1-\left(\frac{N-A}{N}\right)^f\right) =$ | |
|---|---|---|---|
| 1 | 290 | $290 * (1-(862-300)/862)^1$ ) | 100,92 |
| 2 | 42 | $42 * (\text{«»})^2$ | 24,14 |
| 3 | 18 | $18 * (\text{«»})^3$ | 13,01 |
| 4 | 10 | $10 * (\text{«»})^4$ | 8,19 |
| ... | ..... | ..... | ..... |
| 37 | 1 | $1 * (\text{«»})^{37}$ | 0,99 |
| | | TOTAL = | 175 |

175 *types* o, en otras palabras, para A = 300 el crecimiento teórico de V será de 175 *types*. El hecho cierto es que el crecimiento real de V en el *token* 300 es de 173 *types*, sólo 2 menos que los que el modelo binomial predecía, lo cual implica que, en este caso, el ajuste entre realidad y modelo teórico es bastante preciso.

Apliquemos el procedimiento al relato completo, analizando la relación entre el crecimiento teórico y el crecimiento real del vocabulario y aplicando el test estadístico del Chi[2]:

---

[2] De *El Aleph* (1949),

| N | Incr_Real | Incr_Teór. | Dif. | $Chi^2$ |
|---|---|---|---|---|
| 100 | 71 | 74 | -3 | 0,1216 |
| 200 | 49 | 54 | -5 | 0,4629 |
| 300 | 53 | 47 | 6 | 0,7659 |
| 400 | 50 | 43 | 7 | 1,1395 |
| 500 | 44 | 41 | 3 | 0,2195 |
| 600 | 36 | 38 | -2 | 0,1052 |
| 700 | 28 | 36 | -8 | 1,7777 |
| 800 | 39 | 35 | 4 | 0,4571 |
| 862 | 19 | 21 | -2 | 0,1904 |

TOTAL $Chi^2$= 5,2402

G.D.L.= 8

Probabilidad de rechazar H0 = 26,83 %, o sea, no significativa. Vemos que, en este cuento de Borges, el crecimiento real del vocabulario coincide netamente con el calculado según el modelo teórico.

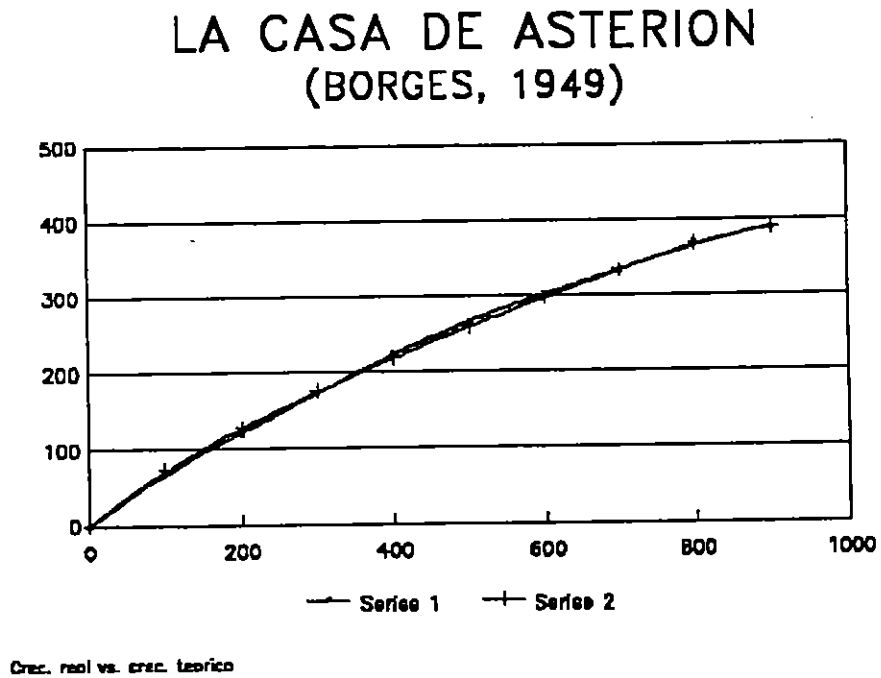


Fig. nº 1

Un análisis similar de los 17 restantes cuentos del libro *El Aleph*, al que pertenece «La casa de Asterión», evidencia que dichos textos se ajustan con precisión al modelo teórico binomial del crecimiento léxico desarrollado por Ch. Muller y apenas puesto en práctica hasta el momento.

Otro experimento posterior, el de analizar el crecimiento de *types* desde el relato «El inmortal» hasta «La intrusa», primero y último de *El Aleph*, revela un desajuste notable entre el modelo teórico y la realidad. El test del $Chi^2$ nos lleva a rechazar la hipótesis nula con un nivel de significación superior

al 99 %. La razón del desajuste -y una lectura no muy profunda del libro así lo corrobora- es fácil de entender: el volumen se compone de 18 piezas muy distintas entre sí tanto temática como estilísticamente.

Daré otro ejemplo literario bastante alejado en el tiempo: la traducción en octava rima de fray Luis de León del bíblico *Cantar de Cantares* atribuido a Salomón.

| N | V | INCR. |
|---|---|---|
| 350 | 201 | 201 |
| 700 | 324 | 123 |
| 1050 | 455 | 131 |
| 1400 | 564 | 109 |
| 1750 | 677 | 113 |
| 2100 | 773 | 96 |
| 2450 | 853 | 80 |
| 2800 | 941 | 88 |
| 3150 | 1018 | 77 |
| 3500 | 1100 | 82 |
| 3531 | 1106 | 6 |

| N | Incr_Real | Incr_Teor. | Dif. | Chi2 |
|---|---|---|---|---|
| 350 | 201 | 210 | -9 | 0,3857 |
| 700 | 123 | 145 | -22 | 3,3379 |
| 1050 | 131 | 123 | 8 | 0,5203 |
| 1400 | 109 | 110 | -1 | 9,09E-03 |
| 1750 | 113 | 99 | 14 | 1,9797 |
| 2100 | 96 | 93 | 3 | 9,67E-02 |
| 2450 | 80 | 87 | -7 | 0,5632 |
| 2800 | 88 | 82 | 6 | 0,4390 |
| 3150 | 77 | 77 | 0 | 0 |
| 3500 | 82 | 74 | 8 | 0,6486 |
| 3531 | 6 | 6 | 0 | 0 |

total chi2= 8,9674

G.D.L.= 10

Probabilidad de rechazar H0 = 39,03 % (no significativa)

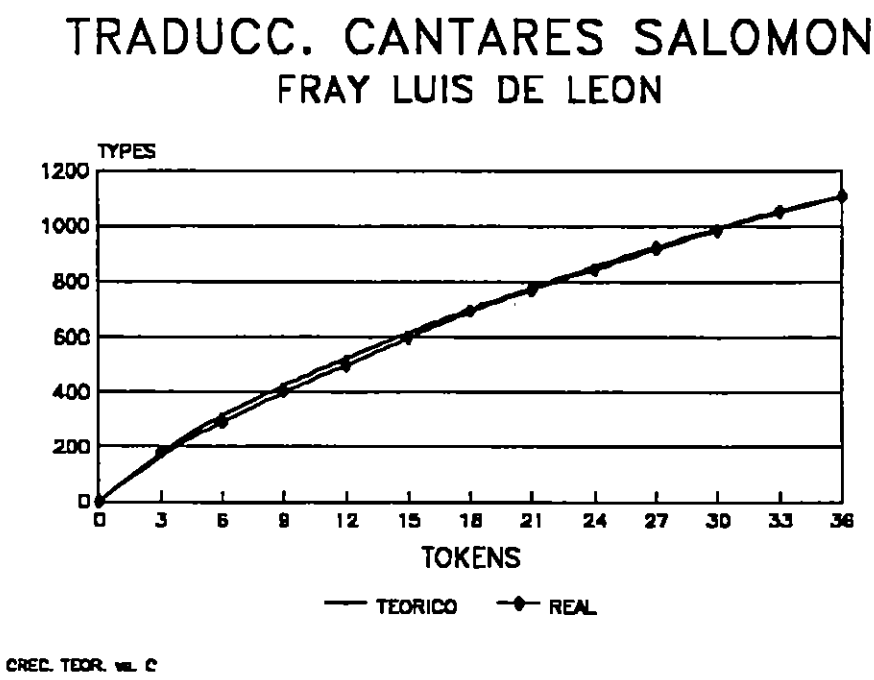


Fig. nº 2

No puede, por tanto, rechazarse la hipótesis nula. Parece, pues, que la voluntad del autor ha dado lugar a un texto de crecimiento léxico regular, próximo al del del modelo binomial mülleriano. Tal vez el tema unitario del poema, un lírico diálogo amoroso, y la contención clásica del estilo de Fray Luis expliquen el fenómeno.

El desajuste entre el crecimiento real del vocabulario en un discurso determinado y su crecimiento teórico según el modelo binomial es -básicamente- un *indicador* de fluctuaciones que llaman la atención al investigador sobre un fenómeno textual, cuya explicación debe buscarse en la reiteración o innovación temático-estilística y que, por tanto, remite a un análisis detallado de la obra estudiada. El ordenador es, obviamente, herramienta indispensable para la tarea preliminar, lexicométrica. Para finalizar, sólo dejar constancia que la riqueza léxica de un texto, la regularidad o la ausencia de regularidad del crecimiento de su vocabulario, no implican necesariamente mayor expresividad, exactitud o valor estético. Grandes obras de la literatura universal utilizan un vocabulario relativamente reducido y son más frecuentes los experimentos frustrados que exhiben una espectacular abundancia de vocabulario. En el lenguaje natural, si consideramos como la más relevante la función comunicativa, hay que saber decir lo que se pretende decir, y esto es, en principio, no guarda una relación de necesidad con el mayor o menor número de vocablos que se empleen para ello[3].

---

[3] Los ejemplos desarrollados en esta comunicación han sido posibles gracias a la aplicacón LEXI. LEXI es una herramienta para el análisis del crecimiento del vocabulario en textos escritos en caracteres ASCII (o sea, los caracteres de las lenguas occidentales o no cuya transcripción se ajuste al código ASCII extendido).

Los textos sometidos a análisis deben estar grabados encódigo ASCII puro (es decir, el programa ignora signos depuntuación y ciertos caracteres de control que la mayoría de losprocesadores de texto utilizan para ciertas tareas). Normalmentelos procesadores de textos más usuales poseen una opción degrabar texto en modo ASCII puro (también llamado «modo no-documento», «modo DOS», etc.).

LEXI convierte todas las palabras a minúsculas. En efecto,no parece tener objeto que la aplicación que ofrecemos distingacomo palabras diferentes 'Agua' / 'AGUA' / 'agua; 'La' / 'LA' /'la'; etc.

## BIBLIOGRAFIA

GUIRAUD, P., *Les Caractères Statistiques du Vocabulaire*, Paris, P.U.F., 1954.

GUIRAUD, P., *Problèmes et Méthodes de la Statistique Linguistique*, Dordrecht-Holland, D.Reidel Publishing Company, 1959.

HERDAN, G., *The Calculus of Linguistics Observations*, The Hague, Mouton & Co,1962.

HERDAN, G., *Quantitative Linguistics*, London,Butterworths, 1964.

MULLER, CH., *Essai de Statistique Lexicale, l'Illusion comique de P. Corneille*, Paris, Klincksieck, 1964.

HERDAN, G., *The Advanced Theory of Language as Choice and Chance*,

BERLIN-HEILDELBERG-NEW YORK, Springer, 1966.

MULLER, CH., *Etude de Statistique Lexicale, le Vocabulaire du Théâtre de Corneille*, Paris, Larousse, 1967.

MULLER, CH., *Estadística lingüística*, Madrid, Gredos, 1973.

KOCK, J. de., *Introducción a la lingüística automática de las lenguas románicas*, Madrid, Gredos, 1974.

MULLER, CH., *Principes et Méthodes de la Statistique Lexicale*, Paris, Classiques Hachette,1977.

DUGAST, D., *La Statistique Lexicale*, Genève, Slaktine, 1980.

---

El programa ofrece la posibilidad de crear hasta tres archivos de resultados distintos: el primero, con el nombre del fichero original y la extensión .CNT, muestra una serie puramente numérica de crecimiento léxico. El segundo, con extensión .CRC,da un listado de todas las palabras distintas por orden deaparición en el discurso escrito, acompañadas de su frecuenciade utilización. El tercero, con la extensión .ORD, hace una tareasemejante, sólo que en este caso las palabras, con su frecuenciaabsoluta, se exhiben en orden alfabético. Hay que hacer notar queeste 'orden alfabético' no corresponde al normativo de la RealAcademia de la Lengua Española (ya que el programa ha sido diseñado para casi cualquier otra lengua), sino al que exige la propia estructura del código ASCII extendido (por lo cual, laspalabras que en español comiencen por 'ñ' serán colocadas después de las que lo hagan por 'z'; lo mismo podría decirse de las vocales acentuadas o con diéresis).

La pregunta «FRECUENCIA DE MUESTREO» solicita una cifra (=>1) según la cual LEXI exhibirá el número de palabras nuevas del discurso escrito adaptándose al número introducido por el usuario.

Esta aplicación requiere un ordenador IBM PC o compatiblecon un mánimo de 512 K de RAM y una unidad de diskette. Si el texto a analizar es muy extenso, puede ser necesario un disco duro.

LEXI ha sido desarrollado por V. Sabido y J.M. Guirao en la Universidad de Granada. El programa, cuya propiedad intelectual pertenece a los autores y a dicha Universidad, puede ser sin embargo distribuido con finalidad investigadora o educativa, con las únicas reservas de que a) no se modifiquen los archivos LEXI.EXE, LEAME.COM y README.COM y b) no se cobre ningún tipo de tarifa, a excepción de los mínimos gastos de reproducción y/o envío.