Métodos para medir la riqueza léxica de los textos. Revisión y propuesta[1]

Methods for measuring the lexical richness of the texts. Review and proposal

Ramon Capsada Blanch y Joan Torruella Casañas
Institut de Sabadell
ICREA - Universitat Autònoma de Barcelona

Resumen. Con el presente trabajo se pretende hacer una revisión extensa de los diferentes métodos existentes para medir la riqueza léxica de textos y hacer una propuesta para su aplicación a un corpus textual. En primer lugar, se da una visión de los principales índices existentes para cuantificar la riqueza léxica, explicando cómo están definidos y evaluando sus fortalezas y sus debilidades a partir de actividades de carácter experimental. En segundo lugar, se plantea una propuesta de metodología de medición de la riqueza léxica para poderse utilizar en todo un corpus textual, de manera que se puedan establecer comparaciones entre textos y constituir una clasificación pautada del grado de riqueza léxica de cada uno de ellos dentro del conjunto del corpus.

Palabras clave: Riqueza léxica, Lingüística de corpus, Lingüística cuantitativa, Estadística léxica, Distribuciones de frecuencia de palabras, Estilometría.

Abstract. This paper aims to provide a comprehensive review of the different methods used to measure the lexical richness of texts and make a proposal for their application to text corpora. Firstly, it presents an overview of the main existing metrics to quantify lexical richness, explaining

Data de recepción: 18.02.2016 ▪ Data de aceptación: 18.10.2016.

[1] Esta investigación ha podido desarrollarse gracias a las ayudas de la DGICYT (FFI2014-51904-P) y del Comissionat per Universitats i Recerca de la Generalitat de Catalunya (SGR2014-1328).

how these are defined and evaluating their strengths and weaknesses by conducting experimental activities. Secondly, it proposes a methodology for measuring lexical richness that can be used across a complete text corpus so that we can both draw comparisons between texts and create a patterned rating of the degree of lexical richness for each of the texts within the whole corpus.

Keywords: Lexical richness, Corpus linguistics, Quantitative linguistics, Lexical statistics, Word frequency distributions, Stylometry.

1. INTRODUCCIÓN

En el campo de la lingüística cuantitativa, la medición de la riqueza léxica es un tema conocido y ampliamente tratado. A través de los años ha habido numerosas propuestas con el difícil objetivo de conseguir resumir la complejidad lingüística de un texto con la simplicidad de un número que permitiera comparar y ordenar los diversos textos según su riqueza de vocabulario. Con el presente trabajo hemos pretendido hacer una revisión extensa de los diferentes métodos existentes para medirla. Hemos querido explicar los diferentes números índice que cuantifican la riqueza léxica, entender como están definidos y cómo se calculan.

Hemos procurado que el trabajo tenga una intención didáctica y un carácter divulgativo, usando un lenguaje común y inteligible para cualquier lector interesado en el tema sin necesidad de tener otros conocimientos de matemáticas que no sean los elementales. No hemos evitado, sin embargo, en ciertos momentos, entrar en detalles concretos para dar al lector la posibilidad de apreciar el matiz de las partes y del procedimiento preciso de algunos argumentos.

Para poder poner a prueba los diferentes índices de medida de la riqueza léxica que se presentan, hemos llevado a cabo actividades de carácter experimental con los diferentes índices que hemos estudiado, haciendo análisis de carácter numérico para evaluar sus fortalezas y sus debilidades e indagar cuál podría ser la propuesta más sólida y recomendable.

Para poder llevar a cabo este trabajo de medida se han preparado varias aplicaciones informáticas programadas *ad hoc* que realizan los cálculos, los gráficos y los análisis que se presentan.

Los métodos estudiados se han aplicado en el análisis de la riqueza léxica de las diferentes obras que forman parte del *Corpus Informatitzat del Català Antic* (*CICA*). El *CICA* es un corpus informatizado, consultable a través de Internet (http://www.cica.cat), que actualmente está compuesto por 414 obras, desde el siglo XI al XVII, en lengua catalana y que contiene más de 9 000 000 de palabras.

Finalmente, también hemos tenido el atrevimiento de plantear nuestra propia propuesta de metodología de medición de la riqueza léxica de un texto, que se basa en la utilización conjunta de la información que proviene de los índices con mejor valoración, y en el uso de herramientas estadísticas elementales para conseguir un tratamiento global que supere el estudio aislado de un texto y se alcance un análisis de carácter relativo dentro del conjunto del corpus al que pertenece.

2. LA *TTR* (*TYPE-TOKEN RATIO*)

¿Cuáles son los métodos que permiten medir la mucha o poca profusión de vocabulario en un texto? La respuesta parece simple: el recuento de las palabras distintas que una obra contiene, de modo que, cuanto más elevado sea su número, más profusión léxica tendrá el texto. Aunque algunos investigadores en el campo del desarrollo del lenguaje infantil, como Miller (1991), han utilizado el *NDW* (*number of different words*) como medida de la variedad del vocabulario usado, a nadie se le escapa que esto no permite comparar la riqueza léxica de diferentes textos a menos que tengan el mismo tamaño, porque su valor depende precisamente de esta extensión. Por ello, será necesaria una cierta estandarización o normalización a partir de la longitud del texto. Esta no es una cuestión baladí, sino que ha sido históricamente tratada por muchos autores y que ha dado lugar a una multitud de propuestas procedimentales que se irán explicando a lo largo de este trabajo[2].

Como la longitud de un texto es el elemento determinante en la aparición de términos nuevos, parece natural que para conseguir un índice que permita hacer comparaciones sea necesario relacionar el número de palabras distintas (*type*) con el número total de palabras que el texto contiene (*token*); o sea, hacer la división del primer número entre el segundo, es como si se calculara la *riqueza media* de cada texto.

---

[2] En este punto, habría que especificar el concepto de *palabra*, ya que es un término que puede implicar muchas matizaciones. En este trabajo utilizaremos la acepción más amplia del término y que es la que se usa de una manera más habitual en los estudios basados en el tratamiento automático de la información. Entenderemos por palabra cualquier cadena de caracteres limitada por espacios en blanco o por cualquier signo de puntuación, es decir, cada unidad gráfica contará como una palabra. En sucesivos trabajos, sin embargo, será necesario especificar qué pasa con las variantes gráficas (*hixo*, *fixo*, *fijo*, *hijo*), aspecto muy relevante cuando se trata de textos antiguos, qué pasa con las palabras contractas (*al*, *del*) o con las palabras formadas por más de un elemento, ya sea a partir de un proceso morfológico de composición o no (*dormirse*, *lavavajillas*, *enhorabuena*). Asimismo, se deberá hablar de cómo se deben tratar las diferentes formas flexivas de un lema, qué hacer con los homógrafos y la polivalencia semántica y, en último término, decidir si para recuentos destinados a estudiar la riqueza léxica es necesario incluir las palabras gramaticales.

Esta es una de las medidas más antiguas (Templin, 1957) y más ampliamente utilizada para evaluar la riqueza léxica de un texto. Se llama *TTR*, que es la abreviación de la expresión inglesa *type-token ratio*. *Type*, traducida con la palabra española *tipo*, se refiere a las palabras distintas y *token* a cada una de las palabras (repetidas o no) que hay en el texto. La traducción literal de la palabra *token* es algo que sirve como una representación visible o tangible de una idea (símbolo, emblema, muestra), que tiene sentido si se piensa que los diferentes *tokens* repetidos son representaciones que aparecen en distintas partes del texto de un mismo *type*; la traducción más contextualizada sería: *palabra*, *ocurrencia*, *unidad gráfica*. Cada palabra *token* es una concreción particular de una palabra *type* general. Así la *TTR* es la razón o cociente que existe entre el número de tipos (*types*) y el número total de palabras (*tokens*), y su valor está comprendido entre 0 y 1.

El número total de palabras se suele representar con una *N* y el número de tipos con una *V*, ya que también corresponde al tamaño del vocabulario que utiliza el texto. Entonces se escribe:

$$TTR = \frac{V}{N}$$

Por ejemplo, en las dos muestras siguientes, se obtendría:

Verde que te quiero verde
verde viento verdes ramas
el barco sobre la mar
el caballo en la montaña.

Verde, que yo te quiero verde.

Con la sombra en la cintura
ella sueña en la baranda
verde carne, pelo verde
su cuerpo de fría plata.

(Federico García Lorca, *Romance sonámbulo*)

tipos: 28
palabras: 45
$TTR = \frac{28}{45} = 0{,}62$

Puedo escribir los versos más tristes esta noche.
Yo la quise, y a veces ella también me quiso.
En las noches como ésta la tuve entre mis brazos.
La besé tantas veces bajo el cielo infinito.

(Pablo Neruda, *Poema 20*)

tipos: 34
palabras: 36
$TTR = \frac{34}{36} = 0{,}94$

A lo largo de un texto, la *TTR* no es un valor constante, sino que va disminuyendo a medida que va aumentando la cantidad de *tokens* del texto que cogemos para su medida. Esta disminución de la *TTR* se explica a partir de la relación existente entre el número de tipos que van apareciendo con respecto al número de *tokens* del texto. En la Figura 1 se representa esta relación correspondiente en las primeras 400 palabras de la obra *Curial e Güelfa*[3].

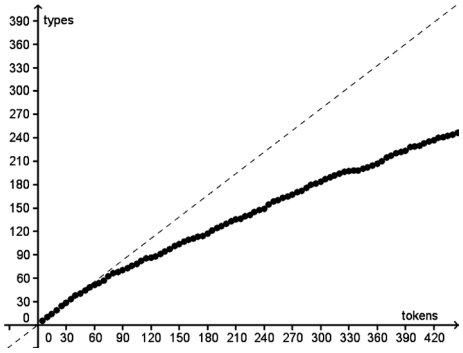


Figura 1

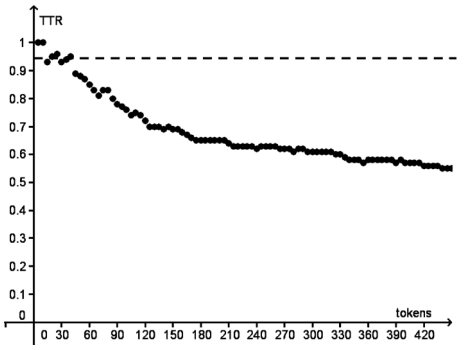


Figura 2

Los puntos del gráfico no siguen una línea recta sino que dibujan una curva que, a medida que avanza, va perdiendo pendiente separándose cada vez más de la recta de trazo discontinuo que se ha dibujado manteniendo la pendiente correspondiente a los primeros puntos de la curva. Esto sucede así porque la relación entre los *types* y los *tokens* de un texto no es una relación lineal (de proporcionalidad), en el sentido de que si de un texto se seleccionan el doble de *tokens* no habrá necesariamente el doble de tipos sino menos, ya que a medida que van apareciendo nuevos tipos cada vez quedarán menos por aparecer y, así, al ir aumentando el número de *tokens* los nuevos tipos irán apareciendo con más separación.

Como consecuencia, si vamos calculando la *TTR* a medida que vamos aumentando en número de *tokens*, como el número de tipos va aumentando más lentamente, el valor la división $\frac{V}{N}$ irá disminuyendo. Esto es, el valor de la *TTR* tenderá a disminuir a medida que la longitud del texto aumente.

En la Figura 2 se han representado los valores de la *TTR* correspondientes a los segmentos del mismo texto de *Curial e Güelfa*. Si la *TTR* fuera un valor constante,

[3] En la mayor parte de los ejemplos de este trabajo se ha utilizado la obra *Curial e Güelfa* en la edición de Antoni Ferrando (2007).

los puntos estarían sobre una recta horizontal como la dibujada con trazo discontinuo. Pero no es así, sino que se obtiene un gráfico manifiestamente decreciente.

En conclusión, se puede decir que la *TTR* es una medida que presenta limitaciones a la hora de estudiar la riqueza léxica, ya que su valor depende de la longitud del texto. Esto hace que este índice no sea útil para comparar la riqueza léxica de varios textos a menos que estos tengan la misma longitud.

3. CORRECCIONES DE LA FÓRMULA DE LA *TTR*

Ha habido numerosas propuestas para conseguir una medida más global que no dependiera de la longitud del texto, haciendo correcciones elementales a la fórmula inicial para calcular la *TTR*.

Si en vez de dividir el número de tipos (*types*) por el número total de palabras (*tokens*), se divide por la raíz cuadrada de este total, se obtiene el índice llamado *RTTR* (*root type-token ratio*) propuesto por Giraud (1960). Simbolizando el número de tipos por *V* y el número total de palabras por *N*, las fórmulas son:

$$TTR = \frac{V}{N} \qquad RTTR = \frac{V}{\sqrt{N}}$$

La idea está en que $\sqrt{N}$ crece más lentamente que *N* cuando ésta aumenta su valor, entonces para compensar la disminución de la *TTR* en crecer *N*, se divide por un número menor y, así, el resultado de la división es mayor.

En la Figura 3 se han representado la relación entre la longitud del segmento del texto y el índice *RTTR* correspondientes a la misma obra *Curial e Güelfa*. Se han tomado cuarenta puntos del texto uniformemente repartidos a través de toda su longitud que corresponden a muestras de texto cada vez más largas; en este caso concreto la primera muestra tiene 3500 palabras, la segunda 7000, la tercera 10 500 y así, sucesivamente, se va sumando 3500 palabras a la muestra anterior, hasta las 146 212 palabras totales que tiene la obra. De esta forma obtenemos una visión global de la evolución del índice.

En este gráfico, se observa que, a pesar de que su comportamiento es un poco mejor al de la *TTR*, tampoco se mantiene constante ya que al aumentar la longitud del texto el valor del índice también aumenta.

Siguiendo con la idea de corregir la fórmula para calcular la *TTR*, existe una familia de índices formada por los que utilizan el cálculo del logaritmo[4] de un

[4] El cálculo de logaritmos está muy relacionado con el cálculo de potencias; cuando una base dada la elevamos a un exponente se obtiene un resultado que se llama potencia. Cuando se calcula un

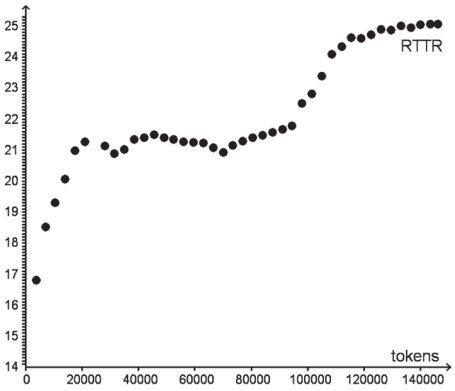


Figura 3

número. Existen múltiples fórmulas de este tipo que se han ido proponiendo a través de los años, de las que destacamos las cinco que se anotan a continuación, indicando también el investigador que la ha propuesto y el año de su formulación:

Herdan (1960): $C = \frac{log\ V}{log\ N}$

Somers (1966) : $S = \frac{\log(log\ V)}{\log(log\ N)}$

Maas (1972): $M = \frac{log\ N - log\ V}{log^2 V}$

Dugast (1978): $U = \frac{log^2 N}{log\ N - log\ V}$

Honoré: (1979) $H = 100\frac{log\ N}{1 - \frac{V_1}{V}}$

V_1 es el número de *hápax legomena* (palabras que aparecen una sola vez)

Diferentes autores (Baayen & Tweedie, 1998; Malvern *et alii*, 2004; McCarthy, 2005) han demostrado que la mayoría de estas medidas tampoco tienen un buen comportamiento respecto a la longitud del texto, ya que su valor aún depende marcadamente de esta longitud. Sin embargo, dos de ellas resultan bastante bien valoradas;

logaritmo se hace justo la operación inversa: se calcula a qué exponente hay que elevar la base para obtener la potencia. Así, por ejemplo el logaritmo (en base 10) de 1000 es 3 ya que 10^3 vale 1000 y el logaritmo de 1400 es 3,146 aproximadamente, ya que $10^{3,146}$ vale 1400. Los logaritmos se utilizan para trabajar con números grandes porque, una vez fijada la base, es suficiente trabajar con los exponentes que son números menores. Además, para multiplicar potencias con una misma base es suficiente sumar los exponentes; esto implica que al trabajar con logaritmos, las multiplicaciones se transforman en sumas y las funciones basadas en multiplicaciones en funciones lineales (de proporcionalidad). De ahí su utilización en la búsqueda de una función que relacione de forma lineal los *tokens* y los *types*.

son las medidas propuestas por Honoré y por Maas. Más adelante se analizará el índice de Honoré. En cuanto al índice de Maas se puede decir que presenta un buen comportamiento, ya que toma valores bastante constantes a lo largo de todo el texto.

En la Figura 4 se ha representado cómo evoluciona la medida de Maas cuando va aumentando la longitud del texto, en la misma obra *Curial e Güelfa*. Tal como se ha hecho en el anterior gráfico, en este también se han tomado cuarenta muestras uniformemente repartidas a través de toda la longitud de la obra. Se observa que este índice presenta poca variación; en concreto, varía desde el mínimo de 0,0191 hasta el máximo de 0,0204 en una gráfica con poca oscilación.

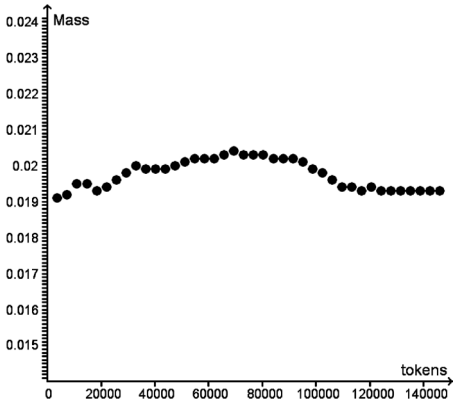


Figura 4

Por tanto, *M* puede considerarse una medida de la riqueza léxica con una aceptable estabilidad y el más recomendable de los índices que utilizan la corrección logarítmica. Aún así, hay que observar que tiene poca sensibilidad, es decir, cuando se mide en diferentes textos, el margen de variación de sus valores es bastante estrecho.

Una segunda observación sobre este índice es que tiene una variación inversa a la que tienen todos los demás índices; es decir, cuanto mayor es su valor, menor es la riqueza léxica del texto analizado, lo que hay que tener presente para evitar confusiones.

4. CÁLCULO DE LA RIQUEZA LÉXICA DIVIDIENDO EL TEXTO EN SEGMENTOS Y EN MUESTRAS

Hay otro conjunto de índices, cuyos creadores han querido aumentar su estabilidad basándose no en la modificación de la fórmula de cálculo de la *TTR*, sino en la

división del texto en partes más pequeñas y a partir de estas partes o segmentos del texto realizar el cálculo del índice de la riqueza léxica.

A continuación se presentan cinco índices que utilizan este tipo de procedimiento basado en el tratamiento de las partes: *MSTTR*, *MATTR*, *MTLD*, *D* y *HD-D*. La diferencia básica entre estos índices está en el método que utilizan para obtener las partes del texto a partir de los cuales efectuar su cálculo.

4.1 MSTTR: el valor medio de segmentos iguales de texto

El índice llamado *MSTTR* (*Mean segmental Type-Token Ratio*) fue propuesto por Johnson (1944). Para obtener este índice, el texto a analizar se divide en segmentos con la misma cantidad de palabras (normalmente 100 palabras por segmento). El residuo que queda al final, menor en número de palabras a las de un segmento, se desprecia. Para cada segmento completo se calcula su *TTR* y, al final, se hace la media aritmética de todas ellas. Esta media es el índice *MSTTR*.

De esta manera se consigue que toda la información que hay en el texto intervenga de manera homogénea, ya que cada segmento tiene el mismo peso en el cálculo de la media final. Esto hace que la longitud del texto deje de ser determinante, ya que en los textos largos se formarán más cantidad de segmentos y esta cantidad no tendrá importancia porque su longitud es la misma y se calcula la media de todos ellos. En la evaluación de este índice (Malvern *et alii*, 2004; Bowker & Pearson, 2002), se ha observado que cuando se trabaja con textos largos (más de 1000 palabras) presenta una buena estabilidad; en cambio, en textos más cortos no es tan buena. Una de las causas de esta baja estabilidad en textos cortos proviene de la necesidad de despreciar la parte sobrante de palabras al segmentar el texto; entonces, al ser el texto corto, la longitud del residuo obtenido puede tener relevancia en el cálculo. Si, con la intención de mejorar la estabilidad, se disminuye la longitud de los segmentos, entonces, al tener estos pocas palabras y, en consecuencia, pocos tipos distintos, las *TTR* que se obtienen de cada segmento son siempre muy cercanas a 1, lo que provoca que las *MSTTR* de diferentes textos sean muy parecidas. Por tanto, el hecho de querer conseguir una mayor estabilidad nos hará perder sensibilidad en la comparación entre diferentes textos.

En la Figura 5 se observa su buen comportamiento en textos largos. El índice se ha calculado, como siempre, seleccionando cuarenta muestras uniformemente repartidas a través de toda la longitud de la misma obra *Curial e Güelfa*. Su valor se mantiene, para números grandes de *tokens* (a partir de 50 000), muy cercano a 0,715. Para los números de *tokens* menores del inicio, varía más, y toma los valores extremos de 0,713 y de 0,727.

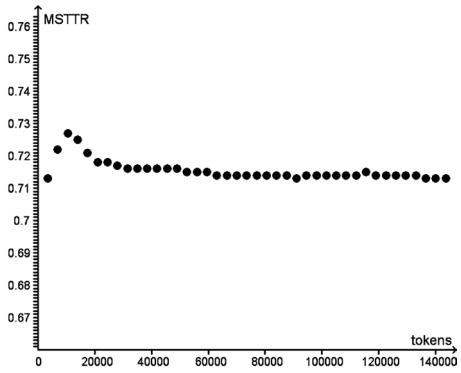


Figura 5

El hecho de que el cálculo de la *MSTTR* esté basado en los cortos segmentos de 100 palabras es la causa de su buen comportamiento; pero este mismo hecho es lo que provoca sus defectos: presenta una baja sensibilidad ya que una partición con tantísimos segmentos pequeños, al calcular los valores medios, provoca un resumen demasiado simple de la información global que el texto contiene.

4.2 MATTR: el valor medio de tantos segmentos como palabras

Michael A. Covington y Joe D. McFall (2010) definieron el índice llamado *MATTR* (*Moving-Average Type-Token Ratio*) de forma similar a *MSTTR* pero con la intención de mejorarlo para superar sus limitaciones. Ambos índices calculan la media de las *TTR* de segmentos del texto de igual longitud, pero el nuevo índice hace intervenir una cantidad mucho mayor de información ya que selecciona más segmentos y con más palabras cada uno: selecciona un segmento de 500 palabras para cada una de las palabras del texto.

El cálculo se hace a partir de una *ventana* de 500 palabras que se va *moviendo suavemente* a través de todo el texto avanzando palabra a palabra. Se comienza calculando la *TTR* para las palabras 1-500, luego por las 2-501, luego 3-502 y así sucesivamente hasta terminar el texto. Finalmente, se calcula la media de todas las *TTR*. El hecho de utilizar mucha más información y de solaparse los diferentes segmentos lo máximo posible produce que no se pierda tanta información global, ya que el valor del índice no se ve afectado por las interacciones accidentales entre los límites de los segmentos y los límites de las unidades naturales del texto (párrafos, capítulos, etc.).

Otro aspecto interesante del cálculo del índice *MATTR* es que se pueden utilizar las *TTR*s individuales de cada segmento para poder observar conjuntamente los

numerosos valores obtenidos con el fin de analizar cómo va evolucionando la riqueza léxica, palabra a palabra, a través de todo el texto. Esto se hace a través de un gráfico como el de la Figura 6. Este gráfico corresponde al texto *Llibre de Job* de Jeroni Cuencas[5]. Por cada una de las 23 181 palabras de este texto, exceptuando las 499 primeras, se calcula la *TTR* de un segmento de texto formado por 500 palabras (las 499 que la preceden y ella misma). Por lo tanto en este gráfico aparecen 22 681 puntos, los cuales, al estar tan apretados, dan ese aspecto de línea quebrada. Esta gráfica es una descripción detallada de cómo va variando la riqueza léxica del texto en cada una de sus partes.

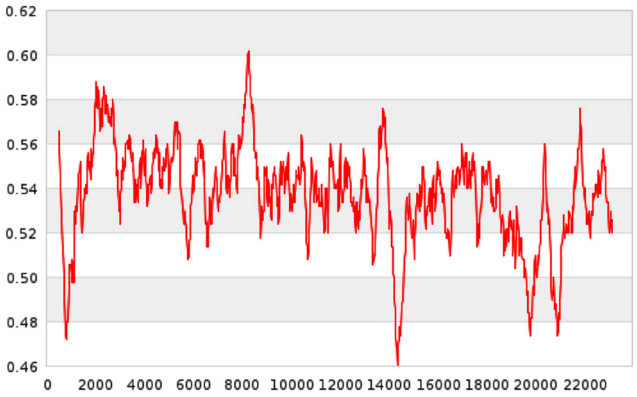


Figura 6

Para observar cómo se comporta el índice *MATTR* respecto a la longitud del texto, se puede establecer un gráfico (Figura 7) como los que ya se han establecido

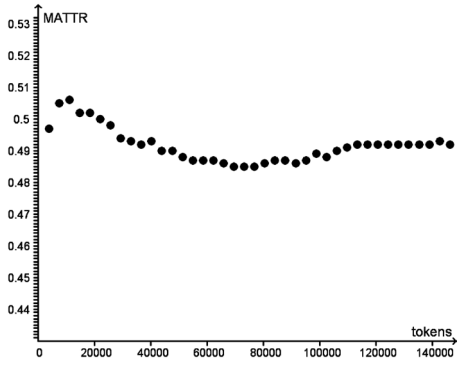


Figura 7

[5] Obra del siglo XVI, editada por Jaume Riera i Sans (1976).

para los índices anteriores, seleccionando cuarenta muestras uniforme-mente repartidas a través de toda la longitud de la misma obra *Curial e Güelfa*. De esta manera, se obtiene una distribución de los puntos parecida a la del gráfico del *MSTTR*, pero en este caso los valores son menores (ya que se trabaja con muestras de 500 *tokens* en vez de 100).

El peor comportamiento también se da al principio en las muestras de menos *tokens* donde los valores oscilan más. A partir de los 25 000 *tokens* el gráfico se va estabilizando en torno al valor 0,490. El valor varía más, sin embargo, que en el caso de la *MSTTR*, el cual se mantenía más constante (para poder comparar diferentes índices se han tomado en todos los casos tres cifras de significación). El valor final del *MATRR* que corresponde a todo el texto es de 0,492. Una explicación de esta mayor variabilidad está en el hecho de que el *MATRR* trabaja con mucha más información. Al disponer de más información este índice es más *sensible*.

4.3 MTLD: Segmentos de texto léxicamente saturados

El índice llamado *MTLD* (*Measure of Textual Lexical Diversity*) fue propuesto por McCarthy (2005). El punto de partida también es similar a los anteriores *MSTTR* y *MATTR*, ya que divide el texto en segmentos y calcula la *TTR* de cada uno de ellos, pero, en este caso, la longitud de cada segmento es variable y depende precisamente del valor que vaya tomando la *TTR* a medida que se va aumentando la longitud del segmento de texto.

El proceso para obtener la *MTLD* de un texto es el siguiente: de forma secuencial, comenzando desde el principio del texto, se va creando el primer segmento aumentándolo palabra a palabra y calculando para cada adición el valor que toma la *TTR* del segmento. Como el valor va disminuyendo a medida que se alarga el texto, cuando este valor se hace inferior a un determinado umbral previamente fijado (normalmente 0,72) se considera que el segmento está completo y se empieza a crear uno de nuevo siguiendo el mismo sistema y repitiendo la operación hasta el final del texto. Al finalizar este proceso, la *MTLD* se calcula dividiendo el número total de palabras del texto entre el número total de segmentos que se han podido formar.

Por ejemplo (entre paréntesis figura el valor de la *TTR*):

...quien (1,00) *ríe* (1,00) *primero (1,00)*, *es* (1,00) *quien* (0,80) *ríe* (0.667) *mejor (1,00)*...

TTR es menor a 0,72:
- Empieza un nuevo segmento.
- Se incrementa el número de segmentos.

Al llegar al final del texto se calcula $MTLD = \frac{N}{n}$, *donde N* es el número de palabras del texto y *n* es el número de segmentos.

Para que no queden residuos sin procesar, si al final del texto queda un segmento sin alcanzar la *TTR* umbral este segmento no se desprecia sino que se obtiene un número residual menor que uno (calculado proporcionalmente a la cantidad que le falta a la *TTR* de este segmento para llegar a uno) que se suma al número de segmentos completos.

Con la finalidad de suavizar los efectos del azar en la longitud de los segmentos del texto, el valor definitivo de *MTLD* se calcula haciendo la media de los dos valores obtenidos al repetir el proceso de cálculo dos veces, uno en sentido directo siguiendo el orden de lectura y otro en sentido inverso. El valor de *MTLD* obtenido de este modo representa, por término medio, el número de palabras que tienen los segmentos, es decir, su longitud media. Cuanto más largos sean los segmentos, menos veces se habrá alcanzado el umbral inferior de la *TTR* (0,72) y mayor será la riqueza léxica del texto.

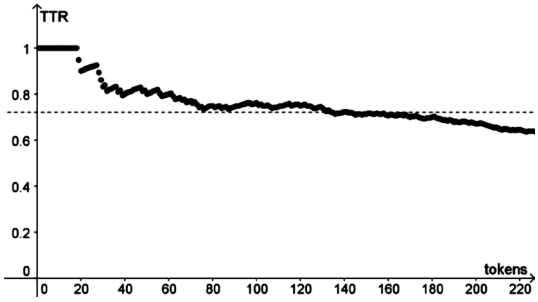
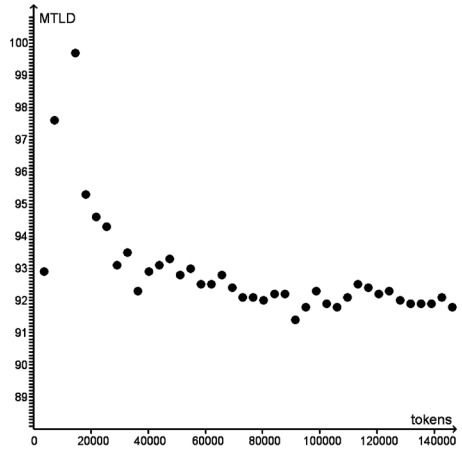


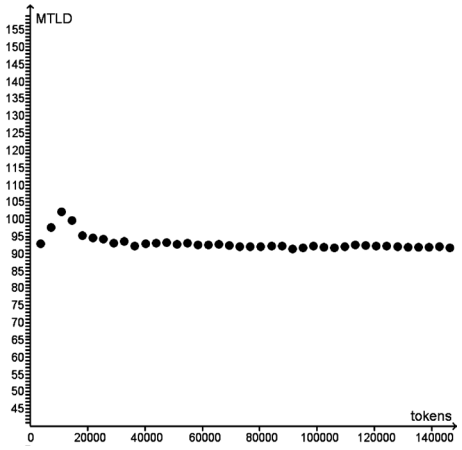
Figura 8

El valor de la *TTR* umbral de 0,72 se ha establecido a partir de la observación práctica. En la Figura 8 se presenta cómo va variando la *TTR* en función de la longitud que tiene el texto (tal como ya se ha visto en la Figura 2). Se ha escogido arbitrariamente una muestra de 220 *tokens* del texto. Se ha dibujado una recta horizontal con trazo discontinuo que indica el valor 0,72. Se observa que para longitudes pequeñas del texto, el valor de la *TTR* tiene marcadas oscilaciones. Esto se debe a que con pocas palabras aparecen tipos nuevos con mucha frecuencia, haciendo variar la *TTR* de forma notable. Pero a medida que la longitud del texto va aumentando, la línea que describe la relación se va estabilizando, porque los tipos nuevos cada vez aparecen con menos frecuencia. Como consecuencia, a partir de un cierto punto,

ir añadiendo palabras tiene muy poca influencia en la modificación de la variación léxica; por tanto, procesar estas palabras resulta *poco rentable*. Es por este motivo que en el cálculo del *MTLD* se corta el segmento de texto en un punto determinado: cuando se ha llegado a la saturación léxica del segmento. Esta es la gran ventaja del índice *MTLD*, la cual le da más eficacia que el *MSTTR*. En el caso actual los segmentos están saturados de información léxica, en el caso anterior los segmentos tenían siempre una longitud arbitraria de 100 palabras, independientemente de la información léxica que el segmento tuviera. A través de las diversas pruebas que se han hecho (McCarthy & Jarvis, 2010), se ha determinado que el mejor umbral de corte es precisamente el valor de 0,72.

Para analizar la variabilidad de este índice respecto a la longitud del texto, volvemos a utilizar el gráfico que trabaja con cuarenta muestras uniformemente repartidas a través de toda la obra *Curial e Güelfa*. En la Figura 9 se observa que este gráfico presenta más dispersión que en el caso del *MSTTR* y del *MATTR*, tanto en los valores iniciales donde la variación sigue siendo más notable como en los valores más avanzados donde se mantiene mayor variación que en los índices anteriores.





Figuras 9 y 10

Hay que indicar, sin embargo, que los gráficos son poco precisos para estudiar la estabilidad de un índice ya que el aspecto de un gráfico depende mucho del valor de la unidad con la que se gradúan los ejes. Así, en las dos Figuras 9 y 10 se representa la misma relación, pero el aspecto es muy diferente: parece que el comportamiento del índice en la segunda figura sea mucho más estable que en la primera, pero realmente la relación es la misma, lo único que ha cambiado es la forma de representarla: en

la Figura 9, cada unidad mínima del eje vertical representa la cantidad 0,1, mientras que en la Figura 10 cada unidad mínima del eje vertical representa 1. Es como si en la primera figura se le hubiera aplicado una ampliación vertical de 10 veces, con respecto a la segunda. ¿Cuál es la escala más conveniente en la representación para estudiar la estabilidad de este índice? La respuesta a esta pregunta no es inmediata. Más adelante, en el apartado 11, volveremos sobre esta cuestión para intentar ir más allá del método gráfico y encontrar un método numérico que nos permita medir con más rigor el grado de variación que los diferentes índices presentan.

Con todo, el *MTLD* parece ser que es uno de los mejores índices para medir la riqueza léxica. El propio autor, Philip McCarthy (2010), junto con Scott Jarvis hicieron un completo trabajo de evaluación, comprobando su validez y comparándolo con los demás índices, obteniendo unos buenos resultados. Comprobaron que el *MTLD* es una medida estable ya que depende poco de la longitud del texto y que sirve para trabajar tanto con textos cortos como con textos largos. También evidenciaron que tiene una buena sensibilidad y que, en cuanto a la comparación con los otros índices, presenta una alta correlación con los que mejor se comportan y una correlación baja con los que se comportan peor.

4.4 Parámetro D: muestras aleatorias de palabras

El índice llamado parámetro *D*, propuesto por Malvern (1989), tiene una larga historia y ha sido estudiado y utilizado por un número considerable de investigadores que han propuesto varias mejoras y evoluciones. Entre estos estudiosos cabe citar Sichel (1986), Malvern (1989), Malvern & Richards (1997), McKee & Malvern & Richards (2000), Jarvis (2002), Harris Wright & Silverman & Newhoff (2003) y McCarthy & Jarvis (2007).

Uno de los aspectos más innovadores de este índice respecto de los demás hasta ahora presentados es el hecho de utilizar muestras aleatorias de palabras del texto en vez de los segmentos secuenciales utilizados en el cálculo de los *MSTTR*, *MATTR* y *MTLD*.

Los segmentos secuenciales son las cadenas de palabras del texto en el orden que estas se presentan cuando se lee; en cambio, las muestras aleatorias se obtienen seleccionando las palabras al azar a partir de todas las palabras del texto.

Cuando se trabaja con muestras aleatorias, en vez de utilizar un solo segmento para cada longitud (como sucedía en el caso secuencial), se seleccionan muchas muestras aleatorias para cada una de las longitudes y se calcula el valor medio de sus *TTR*. Al final, para cada longitud determinada, se tendrá un valor de la *TTR*

correspondiente a una *parte general* del texto de esta longitud determinada. Decimos *parte general* porque no se refiere a una única muestra concreta sino a una multitud de muestras (en teoría, todas las posibles, aunque a la práctica se utilizan 100) de una misma longitud y a su valor medio de la *TTR*.

La otra innovación que presenta este índice *D* es el hecho de que para definirlo no se basa en el cálculo de una única *TTR* (que era el caso de los índices estudiados anteriormente) sino en el estudio del conjunto de las *TTR* obtenidas en las diferentes *partes generales* de texto y, en particular, en valorar la distribución que tienen unas respecto a las otras, para comprobar cómo van variando, a partir del gráfico de puntos en el que cada uno representa la longitud de *una parte general* del texto y su *TTR*.

En la Figura 11 se observa esta representación gráfica de las *TTR* para muestras aleatorias menores de 1000 palabras de la obra *Tirant lo Blanch*[6]. La diferencia entre esta gráfica y las presentadas anteriormente es que en la gráfica actual los valores de las *TTR* son unos valores medios obtenidos a partir de 100 muestras aleatorias que representan una *parte general* del texto de un determinado tamaño; en cambio, en las otras gráficas cada punto representa la *TTR* (u otro índice) de un único trozo de texto secuencial. Los valores medios del gráfico de la Figura 11, precisamente por ser valores medios, hacen que el gráfico tenga un aspecto más *suave* que en el caso de los otros gráficos (Figuras 2 y 8) que se presentaban más quebrados.

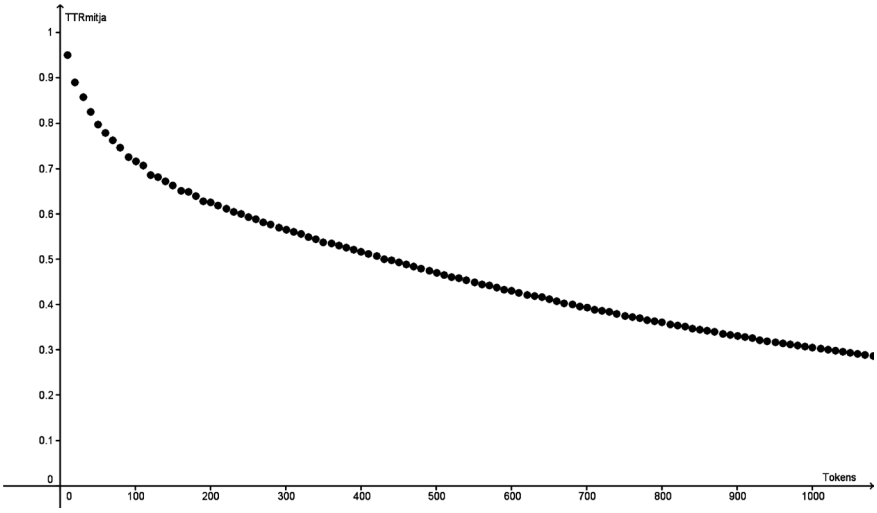


Figura 11

[6] Joanot Martorell, *Tirant lo Blanch*, edición de Albert Hauf (2005).

La idea clave para establecer el índice *D* está en el hecho de encontrar una familia de funciones matemáticas con una fórmula relativamente simple y que se ajuste a la línea de puntos obtenida en el gráfico de la longitud de los fragmentos generales de texto y su *TTR*, como es el caso de la Figura 11.

La familia de funciones, propuesta por Malvern, que cumple bien estas condiciones es la siguiente:

$$y = \frac{d \cdot \sqrt{1 + \frac{2 \cdot x}{d}} - 1}{x}$$

Donde la *x* representa los diferentes valores de la longitud (número de palabras) de cada parte general de texto y la *y* es la *TTR* correspondiente. La *d* es un parámetro, es decir, un valor fijo para cada uno de los casos de función que conforman esta familia. En la gráfica de la Figura 12, están las funciones que corresponden a los valores $d = 10$, $d = 40$, $d = 100$.

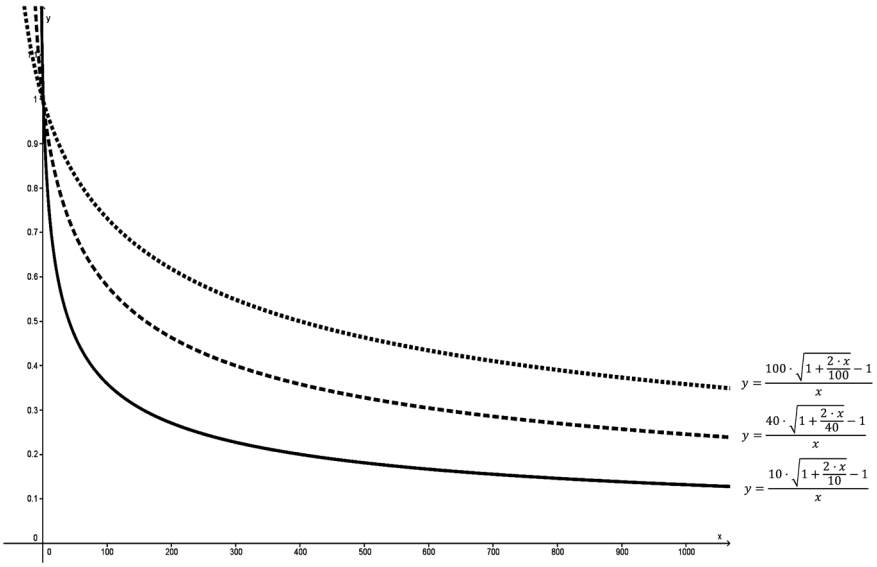


Figura 12

El planteamiento se hace de la siguiente manera: de las muchas funciones que se disponen dentro de esta familia, se trata de escoger aquella función que *se ajuste más* a la distribución de puntos que se obtienen al hacer el gráfico que de forma

empírica previamente se ha calculado. Cuando se dice que la función se ajuste a la distribución de puntos se quiere decir que las distancias de cada uno de los puntos empíricos a la línea que representa la función sean las mínimas posibles, entre los diferentes casos de funciones disponibles en la familia.

Sólo se trabaja con 16 puntos diferentes (muestras de 35, 36, ..., 50 palabras cada una), ya que, después de las pruebas correspondientes, se ha observado que trabajando con estos puntos era cuando se alcanzaba el mejor ajuste a la distribución de los valores empíricos y que si se aumentaba el número de puntos tampoco se producía una mejora significativa.

De este modo, partiendo de las 16 parejas de valores observados, compuestas por la longitud de las muestras de texto y su *TTR*, se deduce la función de la familia descrita que más se ajusta al gráfico de los 16 puntos. El valor que tiene el parámetro *d* en esta función es precisamente el índice *D* que sirve para medir la riqueza léxica del texto que se está analizando.

Sobre la valoración de la validez de este índice, históricamente se ha ido comprobando que presenta una buena estabilidad y sensibilidad, aunque también hay investigadores que ven problemas en su comportamiento, especialmente en textos largos, ya que presenta dependencia respecto a la longitud del texto. También ha recibido críticas el hecho de utilizar muestras aleatorias, por la razón que se destruye la integridad del texto, desapareciendo su estructura textual, influenciando en la medida de su riqueza léxica.

4.5 Índice HD-D. La teoría de las probabilidades mejora la precisión de la medida

Además de las criticas indicadas anteriormente, el revés más fuerte que ha recibido este índice *D* ha sido por parte de McCarthy & Jarvis (2007), al darse cuenta de que este índice no es más que una aproximación de una bien conocida función de probabilidad: la *distribución hipergeométrica*. Estos autores empezaron comprobando, utilizando diversos corpus, que los valores obtenidos empíricamente y los que se pueden calcular directamente a partir de la distribución hipergeométrica tenían altísimas correlaciones (entre 0,971 y 0,913, sobre un máximo de 1).

El valor de *D* y el obtenido de la distribución hipergeométrica son diferentes ya que trabajan en *escalas* de valores distintas, pero su significado es el mismo. McCarthy & Jarvis nombraron este nuevo índice como índice *HD-D* (el *D* de la distribución hipergeométrica).

Dentro del ámbito de la estadística y teoría de la probabilidad, la distribución hipergeométrica es bien conocida y sirve para calcular la probabilidad de obtener

un determinado número de elementos de una misma clase cuando se selecciona una muestra (sin reposición) de un conjunto más amplio. Este es precisamente el proceso de muestreo que, según se ha explicado, se utiliza para calcular el índice *D*.

Si se tiene un texto con un total de *N* palabras y con *d* palabras de un *tipo* determinado, cuando se extrae una muestra de *n* palabras, la probabilidad de que aparezcan en la muestra *x* palabras del *tipo* que nos referimos se puede calcular con la fórmula de la distribución hipergeométrica. Por ello, se puede escribir:

$$P\ (x, n, d, N) = \text{resultado de una fórmula.}$$

Obviamos explicitar esta fórmula por su carácter técnico y por no ser necesaria en la práctica ya que hoy en día no se hacen los cálculos *manualmente*, sino que cualquier hoja de cálculo (tipo Excel) o lenguaje de programación tiene la fórmula de la distribución hipergeométrica implementada, cosa que permite hacer este cálculo automáticamente.

Cuando determinábamos el índice *D*, lo hacíamos a partir del cálculo de la media de la *TTR* de 100 muestras de un tamaño determinado. Estos tamaños eran de 35, 36..., 50 palabras. Pero McCarthy & Jarvis argumentan que para el cálculo del índice *HD-D* es suficiente hacerlo para el único caso de las muestras que tienen 42 palabras y se puede escribir $MTTR_{42}$ para representar su *TTR* media.

Por lo tanto se trata de calcular el valor de $MTTR_{42}$, no a partir del laborioso proceso de analizar 100 muestras de forma empírica (contando experimentalmente los *types* y los *tokens* observando directamente las muestras obtenidas al azar), sino aplicando directamente la fórmula de la distribución hipergeométrica. No se contarán frecuencias experimentales sino que se calcularán probabilidades. Haciéndolo así se obtienen ventajas significativas: se gana mucha simplicidad en el proceso de cálculo —ya que sólo hay que aplicar la fórmula— y, lo que es más importante, la precisión obtenida es óptima y totalmente inigualable por el método empírico, ya que en este caso sólo se trabajaba con 100 muestras y, en cambio, con el cálculo probabilístico, el valor de la probabilidad calculada coincide con el valor que se obtendría al analizar todas las muestras posibles (de 40 palabras, en este caso), que son cantidades astronómicas; por ejemplo, un texto de 1000 palabras tiene $5{,}56 \times 10^{71}$ (un número de 72 cifras!) muestras posibles de 40 palabras.

Concretamos más y suponemos que el texto que se está estudiando tiene 11 tipos de palabras diferentes (*types*) que se representan con las letras A, B, C, D, E, F, G, H, I, J, K. La probabilidad de obtener una palabra del tipo A una o más veces, cuando se extrae una muestra, se representa por P_A. Y lo mismo para las probabilidades de los otros tipos: P_B, P_C, P_D …, P_K.

Se demuestra que el valor de $MTTR_{42}$ se puede calcular a partir de la fórmula siguiente, donde las diferentes probabilidades se pueden calcular a partir de la distribución hipergeométrica:

$$mTTR_{42} = \frac{P_A + P_B + P_D + \ldots + P_K}{40}$$

Obviamos la demostración de la fórmula anterior por no recargar en exceso la exposición.

Siguiendo el esquema de cálculo del índice *D*, además de la media $MTTR_{42}$, también se deberían calcular las otras quince medias $MTTR_{15}$, $MTTR_{16}$, $MTTR_{17}$ ..., $MTTR_{50}$, para luego hacer el gráfico y encontrar la función que más se ajustase al conjunto de los puntos, y el valor del parámetro de esta función sería el índice *D*. Pero McCarthy & Jarvis argumentan que no es necesario hacer todo este proceso, sino que basta con calcular una de las medias y esta media será precisamente el nuevo índice *HD-D*. Estos autores lo demuestran calculando la correlación entre cada media y el índice *D*, y obtienen valores extremadamente altos (del orden de 0,97) en todos los casos, por lo que concluyen que cualquiera de estas medias nos servirá para definir el nuevo índice *HD-D*.

Entonces sólo se trata de elegir, de forma convencional, la *TTR* media de una de estas muestras para definirla como índice *HD-D*. McCarthy & Jarvis proponen que sea $MTTR_{42}$, ya que 42 es el valor que queda a medio camino entre 35 y 50, que es el intervalo que se utiliza en el cálculo de *D*.

Por lo tanto se puede escribir:

$$HDD = mTTR_{42} = \frac{P_A + P_B + P_D + \ldots + P_K}{42}$$

Así pues, a partir de datos globales de un texto que son su número total de palabras y el número de palabras de cada tipo, se calcula la $MTTR_{42}$ de forma exhaustiva utilizando la distribución hipergeométrica. Este valor $MTTR_{42}$ representa la riqueza léxica *media*, en términos de *tipos/palabras*, de todas las muestras de 42 palabras ($2{,}97 \times 10^{74}$ muestras, en un texto de 1000 palabras). Esta clase de muestra de 42 palabras se elige como representante de todo el texto, de tal manera que la riqueza léxica media de estos millones y millones de muestras se asigna como la medida de la riqueza léxica del texto. Este es el índice *HD-D*.

Como se puede observar, a partir de este nuevo tratamiento del tema, el histórico y tan valorado índice *D* no queda en muy buena posición: está calculado de forma empírica midiendo (aunque se utilizan 100 muestras) una parte ínfima de las

muestras teóricamente posibles con la imprecisión que ello conlleva y, además, se deduce a partir de resultados redundantes (al utilizar 16 tipos diferentes de muestras) que nada aportan a la mejora de su cálculo.

En las siguientes tablas se presentan los cálculos de las *TTR* para las muestras de los tamaños propuestas en el cálculo del índice *D* (para la obra *Curiel e Güelfa*). En una columna están los valores calculados según el sistema de obtener y medir las muestras de forma empírica de acuerdo con el método del índice *D* y, en la otra columna, están los valores calculados a partir de la distribución hipergeométrica según el método del índice *HD-D*. En dichas tablas se observa que hay ligeras diferencias entre los valores de las dos columnas; también se comprueba que en los dos casos los valores van decreciendo ligeramente, siguiendo el perfil decreciente de una de las funciones de la Figura 12. Se ha marcado en negrita el valor correspondiente a la muestra de 42 palabras ya que este es el valor del índice *HD-D*.

	Método D	**Método H-D**
TTR_{35}	0.843	0.852
TTR_{36}	0.841	0.850
TTR_{37}	0.840	0.848
TTR_{38}	0.829	0.847
TTR_{39}	0.834	0.845
TTR_{40}	0.829	0.843
TTR_{41}	0.834	0.841
TTR_{42}	0.825	**0.839**

	Método D	**Método H-D**
TTR_{43}	0.827	0.837
TTR_{44}	0.819	0.835
TTR_{45}	0.813	0.834
TTR_{46}	0.816	0.832
TTR_{47}	0.809	0.830
TTR_{48}	0.809	0.828
TTR_{49}	0.809	0.827
TTR_{50}	0.809	0.825

Valor del índice HD-D = TTR42 = 0,839

El índice *HD-D* es una importante mejora respecto a su predecesor *D* ya que su cálculo es mucho más simple y exacto. Aún así sigue teniendo la limitación de que su valor aún depende de la longitud del texto. Este problema perdura, ya que la distribución hipergeométrica depende del tamaño de la población donde se aplica. Así, según antes se ha indicado, cuando se utiliza su fórmula para el cálculo de la probabilidad, uno de los cuatro valores necesarios es precisamente el número total de palabras que el texto tiene. Pero esta variación no es grande.

De este modo, si se hace el gráfico habitual (Figura 13), seleccionando cuarenta muestras uniformemente repartidas a través de toda la longitud de la misma obra *Curial e Güelfa*, se comprueba que su comportamiento en textos largos es bueno, ya que su variación respecto a la longitud del texto es pequeña, similar a como se

comportan los últimos índices estudiados (*MSTTR*, *MATTR* y *MTLD*). Su valor se mantiene, para números de *tokens* grandes, muy cercano a 0,860. Para los números de *tokens* menores, varía más, destacando su valor inicial de 0,845 que es el mínimo.

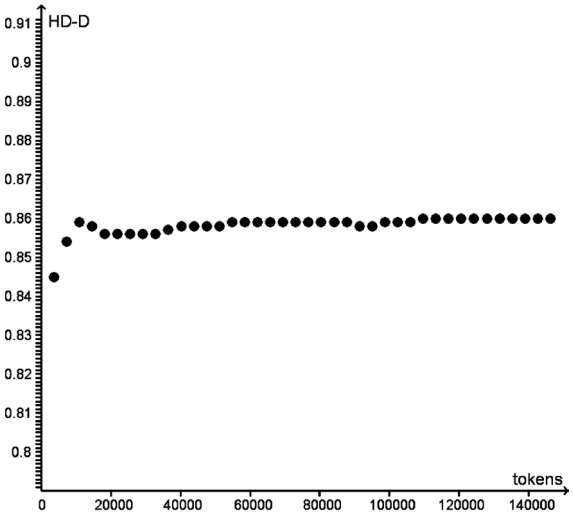


Figura 13

5. DISTRIBUCIONES DE FRECUENCIAS DE PALABRAS

Cuando se ha presentado el índice de Honoré, en el apartado 3, lo hemos definido a partir del número de *hapax legomena* que un texto contiene. Un *hapax legomena* (de la lengua griega *hapax*, 'única', y *legomenon*, 'lectura') es cualquier palabra que sólo aparece en el texto una sola vez.

Para poder saber cuántos hápax tiene un texto o cuántas palabras se repiten dos veces o bien las veces que nos puedan interesar, se debe repasar el texto palabra por palabra (*tokens*) y anotar para cada uno de los *types* que van apareciendo cuántas veces se repite en todo el texto; es decir, es necesario calcular la frecuencia que tiene cada *type*. De esta manera se construye la llamada *Lista de frecuencias de palabras*.

A continuación se presenta una parte de la Lista de frecuencias de palabras de la obra ya citada *Llibre de Job*. Sólo se han incluido los primeros 50 tipos, ordenados según su frecuencia. Se ha simbolizado, como suele ser habitual, los *types* con el símbolo *t* y su frecuencia con el símbolo *m*:

t	*m*
ý	1664
de	880
la	746
que	667
el	547
en	455
per	454
a	447
no	364
los	331
les	299
com	235
ab	208
déu	199
del	186
si	158
me	153

t	*m*
és	151
al	130
se	130
aquell	121
qui	114
què	96
job	95
des	94
ni	92
quant	88
qual	85
tu	83
coses	81
mi	80
· l	80
ha	79
· s	75

t	*m*
mia	74
home	73
jo	73
ell	71
· m	69
aquells	68
més	68
sa	66
són	66
terra	64
sens	60
aquella	59
o	58
vida	58
a	58
perquè	57
son	57

La información que ofrece la *Lista de frecuencias de palabras* puede ser tratada, a su vez, estadísticamente contando cuántas veces se repite cada frecuencia, obteniéndose así la frecuencia de cada frecuencia. Así, se construye la llamada *Distribución de frecuencias (de las frecuencias de palabras)*:

m	V_m
1	3226
2	624
3	267
4	163
5	91
6	61
7	49
8	32
9	23
10	19
11	32
12	19
13	17
14	10
15	14
16	11
17	7
18	6
19	6
20	2
21	7
22	9
23	6

m	V_m
24	6
25	5
26	8
27	5
29	3
30	5
31	4
32	1
33	2
34	4
35	2
36	2
38	1
39	1
40	3
41	2
42	1
43	2
44	3
45	2
46	2
47	2
49	1

m	V_m
50	2
55	1
56	2
57	2
58	3
59	1
60	1
64	1
66	2
68	2
69	1
71	1
73	2
74	1
75	1
79	1
80	2
81	1
83	1
85	1
88	1
92	1
94	1

m	V_m
95	1
96	1
114	1
121	1
130	2
151	1
153	1
158	1
186	1
199	1
208	1
235	1
299	1
331	1
364	1
447	1
454	1
455	1
547	1
667	1
746	1
880	1
1664	1

V = 4824 N = 23181

Para ello es necesario tomar la *m* (que es el conjunto de las frecuencias de los *types*) de la Lista de frecuencias de palabras y tratarla como una variable estadística calculando la frecuencia de cada uno de sus valores; es decir, calcular cuántos *types* existen que se repitan *m* veces, para cada uno de los valores de *m*: 1, 2, 3... hasta el máximo número de repeticiones, que en el caso de la obra *Llibre de Job* es de 1664 veces, que sólo ocurre con el *type ý*. De esta manera se obtiene una nueva tabla que se presenta ordenada de mayor a menor frecuencia; esta frecuencia se representa con el símbolo V_m. Como ya sabemos, para representar el total de *types* se utiliza la letra *V*.

Se cumple $V = V_1 + V_2 + V_3 + \ldots + V_{1664}$. Esto también se puede expresar[7]: $V = \sum V_i$ que significa lo mismo. La lectura de esta tabla nos dice que existen 3226 *types* que se repiten una sola vez, 624 que se repiten dos veces, 267 que se repiten tres veces, y así sucesivamente hasta llegar al último caso: solamente un *type* que se repite 1664 veces.

En esta *Distribución de frecuencias* están representadas de forma resumida y compacta, pero completa, todas las 23 181 palabras de la obra. Esta es la ventaja de este tipo de tabla: nos da una visión global, resumida y estructurada del vocabulario que se utiliza en el texto.

También se puede representar el gráfico de esta distribución (Figura 14), el cual nos permite tener una visión todavía más sintética de las palabras que componen esta obra.

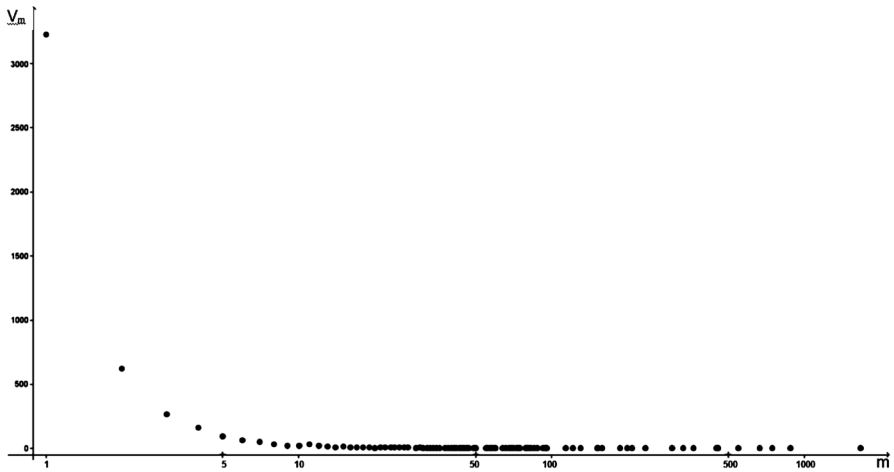


Figura 14

7 Como ya es sabido el símbolo Σ es utilizado en estadística para indicar la suma de un conjunto de términos.

Teniendo en cuenta que los valores que representa la *m* en el eje horizontal van desde 1 hasta 1664 y que nos interesa poder observar en detalle los puntos con los valores de la *m* pequeños, se ha utilizado, en este eje horizontal, *la escala logarítmica*. Esto quiere decir que las divisiones igualmente espaciadas en este eje se han marcado con 1, 10, 100, 1000... en vez de 0, 1, 2, 3... De esta manera se consigue una compresión progresiva para los valores grandes de este eje horizontal, permitiendo la representación de todo su rango en un gráfico bastante corto.

En el gráfico y en la tabla se observa que los valores V_m, que son las frecuencias de *m*, presentan una gran variabilidad y disminuyen con mucha rapidez: así va desde 3226 de los hápax hasta 1 de los *types* que se repiten muchas veces; sólo hace falta llegar a m=5 para que V_5 ya baje por debajo de 100, y a partir de m=59 los valores V_m ya toman el valor 1 o muy cercanos. Todo esto nos indica que el fenómeno más frecuente, y con diferencia, es que los *types* generen muy pocos *tokens*; esto es, que la mayoría de palabras se repitan muy pocas veces, destacando notablemente el caso de una sola repetición (hápax). En este texto, el 67% de los *types* se repiten una sola vez y el 90% de los *types* se repiten como máximo cinco veces. Por tanto, los *types* muy repetidos son una minoría; ahora bien, son pocos pero tienen mucha potencia de replicación, ya que generan un elevadísimo número de palabras. Para ilustrar la magnitud de esta alta generación de *tokens* a partir de escasos *types*, se puede observar que del total de los 4824 *types* que el texto contiene sólo hacen falta seleccionar los 68 *types* más replicados para alcanzar el 50% del total de las 23 181 palabras del texto. Los 4756 *types* restantes se repiten en muy pocos *tokens*: de una a 42 veces.

6. LA LEY DE ZIPF Y EL PARÁMETRO Z

Este comportamiento descrito anteriormente es general en todos los textos. Es un esquema comentado muchas veces por diferentes autores como, por ejemplo, Guillermo Rojo (2002: 9): «Esta clara característica de los textos, que he sintetizado en diversas ocasiones diciendo que están siempre formados por unas pocas formas que aparecen mucho y muchas formas que aparecen poco o muy poco».

George Kingsley Zipf (1949) hizo una formulación cuantitativa de este comportamiento estableciendo la llamada ley de Zipf. Esta ley enuncia una relación numérica muy simple: la palabra más frecuente tiene una frecuencia igual al doble de la frecuencia de la segunda palabra más frecuente, el triple de la tercera, el cuádruple de la cuarta y así sucesivamente. Para poder cuantificar esta relación se debe construir una nueva tabla que se denomina *Tabla de rangos y frecuencias* (ver la tabla siguiente).

rango	frec.	type(s)	rango	frec.	type(s)	rango	frec.	type(s)	rango	frec.	type(s)
1	1664	ý	25	94	dels	57	49	ma	129	23	amichs, ...
2	880	de	26	92	ni	59	47	meu, mon	138	22	aygües, ...
3	746	la	27	88	quant	61	46	hòmens, te	145	21	emperò, ...
4	667	que	28	85	qual	63	45	als, has	147	20	està, ...
5	547	lo	29	83	tu	66	44	d', quals, sua	153	19	altres, ...
6	455	en	30	81	coses	68	43	contra, ...	159	18	part, ...
7	454	per	32	80	mi, ·l	69	42	tots	166	17	algun, ...
8	447	a	33	79	ha	71	41	capítol, hon	177	16	mira, ...
9	364	no	34	75	·s	74	40	cosa, hé, ...	191	15	carn, ...
10	331	los	35	74	mia	75	39	just,	201	14	altre, ...
11	299	les	37	73	home, yo	76	38	sobre	218	13	mans, ...
12	235	com	38	71	ell	78	36	lloch, sos	237	12	cel, ...
13	208	ab	39	69	·m	80	35	paraules, ...	269	11	altra, ...
14	199	déu	41	68	aquells, més	84	34	ans, ...	288	10	orella, ...
15	186	del	43	66	sa, són	86	33	sia, ·ls	311	9	consell, ...
16	158	si	44	64	terra	87	32	meus	343	8	bona, ...
17	153	me	45	60	sens	91	31	castich, ...	392	7	llança, ...
18	151	és	46	59	aquella	96	30	dix, ...	453	6	arbre, ...
20	130	al, se	49	58	o, vida, a	99	29	bé, ...	544	5	avant, ...
21	121	aquell	51	57	perquè, son	104	27	ells, ...	707	4	valor, ...
22	114	qui	53	56	el, té	112	26	boca, ...	974	3	alegre, ...
23	96	què	54	55	·t	117	25	fins, ...	1598	2	noble, ...
24	95	job	56	50	li, ser	123	24	entre, ...	4824	1	capa, ...

Tabla de rangos y frecuencias del *Llibre de Job*

Esta tabla es similar a la de la *Lista de frecuencias de palabras* hecha al principio de este apartado, ordenando los diferentes *types* según su frecuencia de forma decreciente. La diferencia está en que se añade el número de orden que corresponde a cada *type* según esta ordenación. Este número de orden se denomina rango. Así en la tabla correspondiente a la obra *Llibre de Job*, la palabra *ý* tiene rango 1, la palabra *de* rango 2, la palabra *la* rango 3 y así sucesivamente. Cuanto hay varios *types* con la misma frecuencia se ordenan de forma aleatoria y se les da rangos consecutivos; en este caso, en la tabla, sólo se ha representado el último rango que es el mayor. Así, por ejemplo, hay 6 *types* cuya frecuencia es 24 y se les ha asignado los rangos 118, 119, 120, 121, 122 y 123; en la tabla sólo aparece el rango 123; uno de estos 6 *types* es la palabra *entre*.

Una vez se ha construido esta tabla, la ley de Zipf tiene una formulación muy simple: *rango* × *frecuencia* = *constante*. La *constante* es característica de cada texto.

Esta ley sólo se cumple de forma aproximada. En los *types* extremos, tanto en la zona inferior como en la superior, se desvía notablemente del cálculo teórico. Para los valores intermedios el comportamiento es mejor; en estos casos, y para esta obra en particular, el valor del producto se puede redondear (con una cifra significativa) en 3000, tal como se muestra en la siguiente tabla, donde para ilustrar estas observaciones, se ha realizado el cálculo *rango* × *frecuencia* para dos series de rangos, los 10 primeros y los que van del 39 al 54:

rango	*frecuencia*	*producto*
1	1664	1664
2	880	1760
3	746	2238
4	667	2668
5	547	2735
6	455	2730
7	454	3178
8	447	3576
9	364	3276
10	331	3310

rango	*frecuencia*	*producto*
39	69	2691
41	68	2788
43	66	2838
44	64	2816
45	60	2700
46	59	2714
49	58	2842
51	57	2907
53	56	2968
54	55	2970

Era de esperar que los complejos sistemas que regulan los aspectos sintácticos, léxicos y semánticos del lenguaje no se pudieran sintetizar en un modelo teórico tan simple como es esta ley de Zipf.

A partir de su formulación ha habido una inmensa cantidad de autores que han basado sus teorías de lingüística cuantitativa en el desarrollo y mejora de esta ley. Una de las derivaciones más destacables de esta ley es la llamada *Distribución generalizada de Zipf* formulada por Orlov & Chitashvili (1983). A partir de ella, R. Harald Baayen, en varios de sus trabajos (1998, 2001) formula una función que permite calcular el valor del número de *types* a partir del número total de *tokens* que el texto tiene:

$$V(N) = \frac{Z}{\log(p \cdot Z)} \cdot \frac{N}{N - Z} \log\left(\frac{N}{Z}\right)$$

Para cada valor de *N* (la longitud, en número de *tokens*, de diferentes muestras de un texto) se puede calcular el valor de *V (N)* (el número de *types*). A demás de estas dos variables, en la fórmula aparecen dos símbolos más: el parámetro *p* y el parámetro *Z*. El término parámetro se refiere a que son valores característicos de cada texto, es decir, que se mantienen constantes para cualquier muestra de un texto

determinado. El parámetro *p* se puede calcular haciendo la división del número de hápax entre el total de los *tokens* del texto.

Pero el parámetro que tiene interés es el *Z*, que se interpreta como el tamaño mínimo de la muestra del texto donde aún se mantiene la ley de Zipf, porque para muestras de longitud pequeña la ley ya no se cumple, tal como se ha indicado anteriormente. La importancia de *Z* está en que se puede utilizar como medida de la riqueza léxica del texto y que, teóricamente, se mantendrá constante independientemente de la longitud de la muestra. La razón por la que *Z* se puede interpretar como una medida de la riqueza léxica es porque si se analiza la fórmula de la función anterior se observa que un aumento de *Z* conlleva un aumento de *V(N)*.

Por tanto, se trata de utilizar esta fórmula en sentido inverso: se partirá de unos valores fijos de *N* y de *V (N)* como dos características conocidas del texto y se deducirá el valor de Z, obteniéndose una medida de la riqueza léxica del texto.

Sólo queda el problema técnico de resolver la ecuación, que no es trivial ya que la incógnita *Z* aparece tres veces implicada en operaciones diferentes y dos de ellas bajo el efecto del logaritmo. La forma de conseguirlo es utilizando uno de los métodos de resolución de tipo iterativo.

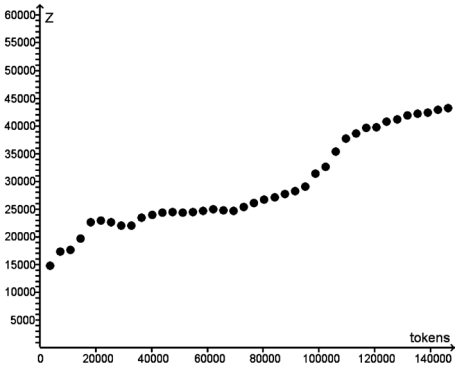


Figura 15

En nuestro aplicativo informático[8], hemos utilizado el llamado método de Newton y lo hemos aplicado en las muestras habituales de la obra *Curial e Güelfa*, obteniéndose el gráfico de la Figura 15. En este gráfico, se observa que los valores obtenidos de *Z* presentan una remarcable variabilidad, ya que van desde el mínimo de 14 792 hasta el máximo de 43 185.

[8] Hemos programado dos aplicaciones informáticas especialmente diseñadas para hacer los cálculos y el tratamiento de la información necesarios para llevar a cabo el estudio actual. La primera aplicación, llamada *DiLe*, partiendo de un archivo en formato *txt,* donde está la obra, construye su *Distribución de frecuencias* (de las frecuencias de palabras) y calcula todos los índices que aquí se presentan con sus gráficos, tanto de los valores empíricos como de los aleatorios, con un número de puntos configurable; también permite hacer gráficos conjuntos de dos o más textos para poderlos comparar. La segunda aplicación, denominada *DiLeCo*, confecciona el análisis numérico de todos los textos de un corpus determinado (en este caso se ha aplicado al *CICA*). El análisis que elabora es el que se explica en el apartado 11 de este trabajo. Esta segunda aplicación también calcula el *NOMC* de cada texto según el proceso que se indica en el apartado 12 y los califica según se explica en el apartado 13.

7. LA TASA MEDIA DE CRECIMIENTO Y EL ÍNDICE DE HONORÉ

También vinculada a esta Ley de Zipf, existen otras propiedades, una de ellas es la que hace referencia al comportamiento de los hápax (Rojo, 2002, p. 11; 2008, p. 19 y ss.): la proporción entre el número de hápax y el total de los *types* en una muestra del texto de tamaño cualquiera mantiene un valor aproximadamente constante que está alrededor del 50%. En nuestro caso, como ya se ha visto anteriormente, da un valor un poco mayor:

$$\frac{\text{Núm.}\, hapax}{\text{Núm.}\, types} = \frac{V_1}{V} = \frac{3226}{4824} = 0{,}67$$

Esta proporción ya había aparecido en la definición del índice de Honoré (en el apartado 3), tiene significado lingüístico y mantiene una estrecha relación con la llamada *Tasa de crecimiento del vocabulario* (*TC*), que mide a qué *velocidad* crece el vocabulario a medida que aumenta el tamaño del texto. Esta tasa se puede estimar (hacer un cálculo aproximado) con la proporción entre el número de hápax y el total de los *tokens* del texto:

$$TC = \frac{V_1}{N}$$

En vez de dividir entre el total de types, se hace entre el total de palabras. Este cociente ya ha aparecido justo en el apartado anterior cuando se calculaba el parámetro p. En este sentido, la TC también se puede interpretar como una probabilidad, la probabilidad de que después de haber leído N tokens de un texto se obtenga un type nuevo que aún no haya aparecido en el total de los N tokens precedentes (Baayen, 2001). Estas consideraciones también se pueden hacer respecto al índice $\frac{V_1}{V}$ haciendo referencia al total del vocabulario en vez del total de palabras.

Si estas tasas fueran constantes podrían ser utilizadas como índices de medida de la riqueza léxica del texto. La tasa de crecimiento ya está claro que no es constante, porque si lo fuera, la curva de crecimiento de los *types* respecto del total de palabras debería seguir una línea recta y eso ya se ha explicado que no se cumple.

Aunque Rojo (2002) defiende que el índice $\frac{V_1}{V}$ se mantiene constante con independencia del tamaño del corpus y que caracteriza el corpus ya que para cada uno se obtiene un valor diferente del índice, cuando se sale del ámbito de corpus de gran

tamaño[9] y se aplica a textos de un tamaño más moderado, este índice no se mantiene estable. De ahí la necesidad que tuvo Honoré (1979) de integrarlo en la fórmula logarítmica para conseguir una mejor estabilidad en el índice que él definió:

$$H = 100 \cdot \frac{logN}{1 - \frac{V_1}{V}}$$

Todavía existe otro índice que definió Sichel (1975) muy relacionado con estos, ya que utiliza los *dis legonema* (tipos con 2 repeticiones) en vez de los hápax para calcular su proporción respecto al total de tipos:

$$S = \frac{V_2}{V}$$

La estabilidad de este índice *S* es mejor que la de $\frac{V_1}{V}$.

En las Figuras 16 y 17, aparecen los gráficos que muestran el comportamiento de los índices *H* y $\frac{V_1}{V}$ en la obra *Curial e Güelfa*, seleccionando los habituales 40 puntos uniformemente repartidos. Hemos seguido representando las unidades gráficas verticales con tres cifras significativas. Haciéndolo así se observa que el índice H presenta un comportamiento aceptable ya que se mantiene bastante constante respecto a la longitud del texto.

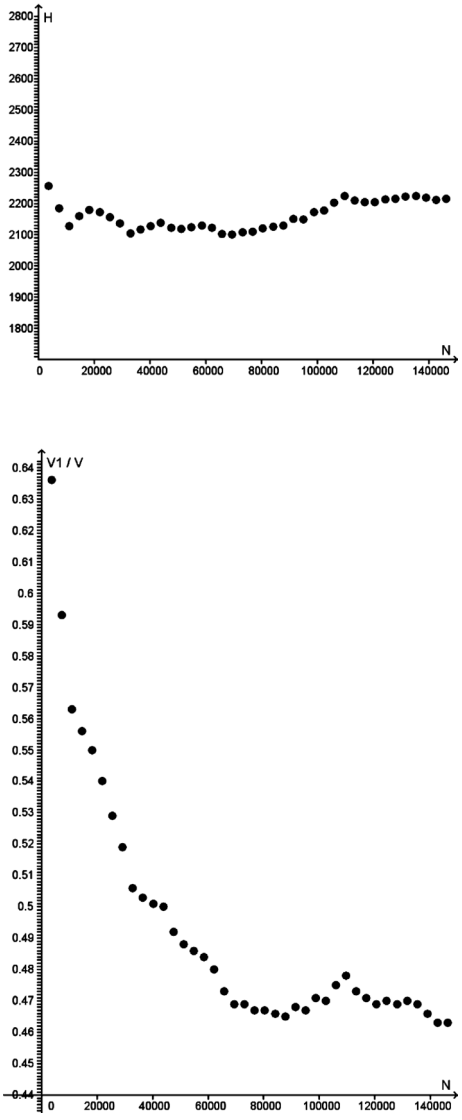


Figura 16-17

[9] Guillermo Rojo ha analizado el *Corpus de Referencia del Español Actual* (*CREA*), construido por la Real Academia Española, que tiene un tamaño inmenso, el cual está formado por 117 millones de *types*. En este análisis los resultados sobre la estabilidad del índice V_1/V son muy buenos.

Contrariamente, el índice V_1/V presenta una variación mucho más marcada, especialmente en los valores menores de N.

8. LA MEDIA Y LA DESVIACIÓN TÍPICA. EL ÍNDICE K DE YULE-HERDAN

En el cálculo de los índices anteriores hemos partido de la *Distribución de frecuencias*, la tabla de las frecuencias de *m* que hemos presentado en el apartado 5. No obstante, a partir de esta tabla también se pueden calcular los dos parámetros más habituales en el cálculo estadístico: el primero es el valor medio de *m*, es decir, la frecuencia media de palabras, y el segundo es la desviación típica de *m*.

La *frecuencia media* de palabras se representa $\overline{m}$ por y se calcula:

$$\overline{m} = \frac{\sum(m_i \cdot V_i)}{\sum V_i} = \frac{N}{V}$$

donde m_i representa los diferentes valores que toma *m* (número de repeticiones de los *types* o, lo que es lo mismo, las diferentes frecuencias) y V_i el número de veces que se presenta este valor m_i. *N* es el total de *tokens* y *V* el total *types*.

El nombre *frecuencia media* de palabras tiene sentido ya que indica, por término medio, cuántos *tokens* corresponden para cada *type*, es decir, cuál es la repetición (frecuencia) media de cada *type*. En realidad este parámetro es justamente el valor inverso a nuestra antigua conocida *TTR*:

$$\overline{m} = \frac{N}{V} = \frac{1}{\frac{V}{N}} = \frac{1}{TTR}$$

Por lo tanto, la *frecuencia media* no nos aporta información nueva, sólo nos la presenta de forma simétrica. Como consecuencia, $\overline{m}$ tampoco será un índice constante que nos sirva para medir la riqueza léxica de un texto ya que, al igual que la *TTR*, su valor dependerá de forma muy notable de *N*.

La *desviación típica* de *m* se representa por σ_m y se define según la forma usual:

$$\sigma_m = \sqrt{\frac{\sum V_i \cdot (m_i - \overline{m})^2}{V}}$$

donde todos los símbolos que aparecen ya son conocidos.

Este parámetro mide la dispersión de los diferentes valores de la variable *m*, es decir, si son muy diferentes o muy iguales entre sí, y lo hace calculando cuál es por término medio la desviación de cada valor m_i de la variable *m* respecto a su valor medio $\overline{m}$. Por eso se multiplican las diferencias $m_i - \overline{m}$ por el número de veces que se presenta este valor m_i (es decir por V_i). Se deben elevar al cuadrado las desviaciones $m_i - \overline{m}$ por motivos técnicos (ya que la definición $\overline{m}$ de hace que estas diferencias den, dos a dos, resultados iguales pero con signos opuestos, lo que provocaría una suma total igual a cero, y en elevar al cuadrado se convierten todas las cantidades en positivas, evitando así la suma nula). Para calcular el valor medio se dividirá la suma total de estas diferencias (al cuadrado) entre el total de *types* (V). Por último, se debe efectuar la raíz cuadrada del cociente para deshacer el aumento de dimensión provocado al elevar al cuadrado las diferencias calculadas inicialmente.

Existe un tercer parámetro estadístico muy relacionado con estos dos, el llamado *coeficiente de variación*, que se representa por v_m y se define como el cociente de σ_m entre $\overline{m}$:

$$v_m = \frac{\sigma_m}{\overline{m}}$$

El coeficiente de variación viene a ser como una desviación típica estandarizada para conseguir una validez más universal. Al dividir entre la media conseguimos eliminar el efecto del tamaño de los datos, pudiendo comparar valores de dispersión que provienen de conjuntos cuyos datos tienen tamaños diferentes. Se puede decir que esta división entre la media reduce la medida de la dispersión a una escala común.

El índice K de Yule-Herdan

Los valores de estos tres parámetros estadísticos $\overline{m}$, σ_m y v_m dependen de la longitud del texto, pero G. Herdan observó (en 1955) que al dividir el coeficiente de variación v_m entre $\sqrt{V}$ se conseguía un nuevo índice que presentaba un comportamiento mucho más estable respecto a la longitud del texto, tomando valores casi constantes. Además, este nuevo índice estaba muy relacionado con el llamado *Characteristic K* que había definido G. U. Yule varios años antes (1944) partiendo de unos presupuestos diferentes.

Se puede simbolizar por K_H el índice definido por Herdan. Este autor lo definió de la siguiente manera:

$$K_H = \left(\frac{v_m}{\sqrt{V}}\right)^2 = \frac{\frac{{\sigma_m}^2}{\overline{m}^2}}{V} = \frac{{\sigma_m}^2}{\overline{m}^2 \cdot V}$$

Sustituyendo en esta fórmula las expresiones que hemos deducido para $\overline{m}$ y σ_m se obtiene la fórmula explícita que nos permitirá calcular K_H directamente a partir de la *Distribución de frecuencias*:

$$K_H = \frac{\sum V_i \cdot {m_i}^2}{N^2} - \frac{1}{V}$$

Cuando Yule había definido, unos años antes, su índice *Characteristic K*, lo había hecho partiendo de supuestos de carácter probabilístico, suponiendo que la aparición de las diferentes palabras en un texto se rige por el modelo de bolsa llena de bolas de diferentes colores, el llamado modelo probabilístico de Poisson: el uso de las palabras es similar a la selección aleatoria de las bolas de diferentes colores depositadas dentro de una bolsa; cada bola corresponde a un *token* y cada color diferente corresponde a un *type*; todas las bolas tienen la misma probabilidad de salir y cada extracción no está influenciada por las anteriores (se hace volviendo a poner en la bolsa las bolas que se sacan). Estos supuestos de aleatoriedad e independencia en el caso de los textos sólo se cumplen parcialmente, ya que las palabras no están utilizadas al azar sino siguiendo las propias leyes del lenguaje que reflejan la cohesión lexical tanto a nivel de frase como a nivel del discurso más general de la obra.

La fórmula que dedujo Yule es la siguiente:

$$K = 10\,000 \cdot \left(\frac{\sum V_i \cdot {m_i}^2}{N^2} - \frac{1}{N} \right)$$

Las fórmulas para los dos índices K_H y K casi coinciden: se diferencian en usar $\frac{1}{V}$ en vez de $\frac{1}{N}$, que para textos largos toman valores casi insignificantes, y en el factor 10 000 que Yule utiliza sólo para agrandar los ínfimos valores que se obtienen y facilitar así su lectura.

Por tanto, en la práctica, el significado y uso de estos dos índices son equivalentes; sólo varía su deducción. La ventaja de K_H es que no hace falta la suposición del modelo probabilístico de Poisson que, como ya hemos indicado, presenta limitaciones. K_H se deduce directamente del cálculo de los parámetros estadísticos de la *Distribución de frecuencias* (media, desviación típica, coeficiente de variación), lo que, además de superar estas limitaciones teóricas, también permite hacer una interpretación más directa de su significado a partir de la información que nos dan de estos parámetros estadísticos.

En este sentido parece que, de entrada, K_H no tiene por qué medir la riqueza léxica del texto ya que, según su definición inicial $K_H = \frac{v_m^2}{V}$, proviene de la medida del coeficiente de variación v_m (nuevamente *normalizado* al volverlo a dividir entre V), el cual, a su vez, proviene de la desviación típica; por tanto K_H mide la dispersión de las diferentes frecuencias m_i en las que aparecen los *types* respecto a la frecuencia media $\overline{m}$.

Pero si nos fijamos bien, vemos que las frecuencias m_i indican la repetición de las palabras, por lo que K_H es una medida de la *dispersión de la repetición de las palabras*. Cuando la dispersión sea baja (y también el valor de K_H) es porque la repetición está concentrada en pocas palabras, es decir, habrá pocas palabras que se repitan y esto implicará una riqueza léxica alta. De forma simétrica justificaríamos que un valor de K_H alto implicaría una riqueza léxica baja. Todo esto nos indica que el índice K_H mide la repetición léxica y, por tanto, también mide la riqueza léxica, pero variando de forma inversa a la mayoría de los otros índices: valores altos del índice implican unas riquezas léxicas bajas y, viceversa, valores bajos del índice implican riquezas léxicas altas.

En la Figura 18, se ha medido el índice *K* en los habituales 40 puntos uniformemente repartidos y en sus correspondientes muestras de la obra *Curial e Güelfa*. En el gráfico se observa que el comportamiento de este índice es bueno ya que tiene poca variación respecto a la longitud del texto. Al principio del gráfico es donde está la mayor variación, después casi se estabiliza con valores cercanos a 98.

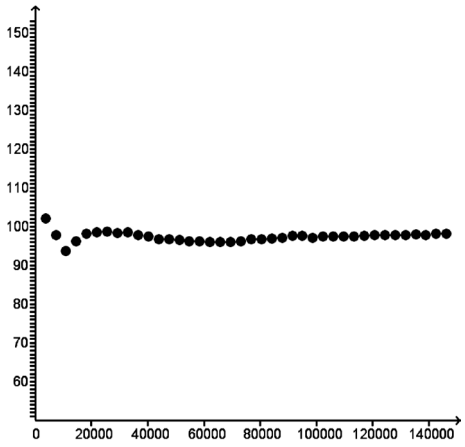


Figura 18

9. LA ALEATORIZACIÓN DE LAS MUESTRAS DA UNA INTERPRETACIÓN MÁS COMPLETA DE LOS DIFERENTES ÍNDICES

Aquí termina todo este carrusel de los índices! En total se ha visto 15 índices diferentes. Ahora trataremos de analizarlos conjuntamente para compararlos y ver cuáles tienen un mejor comportamiento.

En los apartados 4.4 y 4.5 al trabajar con los índices *D* y *HD-D* se han utilizado las muestras aleatorias de palabras; en cambio, en el cálculo de los otros índices se han utilizado muestras obtenidas de forma secuencial tal como se presentan en el texto real.

Como ya se ha explicado en el apartado 4.4, la diferencia entre una muestra aleatoria y una muestra secuencial está en que en la primera cada uno de los *tokens* se obtiene eligiendo al azar una palabra del texto, mientras que en la segunda, los *tokens* siguen el orden real establecido en el texto. El ya explicado modelo probabilístico de Poisson, que se puede concretar con la bolsa llena de bolas de diferentes colores (cada bola corresponde a un *token* y cada color diferente, a un *type*, presentando todas las bolas la misma probabilidad de salir y no estando cada extracción influenciada por las anteriores), sólo se ajusta de forma completa en el caso de las muestras aleatorias. Se debe tener en cuenta que un fragmento del texto real no se puede considerar originado sólo por causas aleatorias ya que la propia estructura del texto condiciona la elección de las palabras cuando el autor lo crea, tanto en el nivel más particular de la estructura sintáctica correspondiente a cada frase como también en el nivel de la estructura de la organización global del texto.

Por ello, para mejorar la comprobación de si los diferentes índices son constantes al variar la longitud del texto, se harán dos procesos diferentes de cálculo: además del que ya se ha realizado sobre muestras secuenciales del texto real, se hará un segundo cálculo sobre un gran número de muestras aleatorias. Si los resultados son buenos en este segundo cálculo, se habrá comprobado la constancia del índice en el texto de forma *general* o *teórica*, en el sentido que se mantiene constante el valor medio del índice del gran número de muestras aleatorias del texto. Si, además, los resultados también son buenos en el caso del cálculo secuencial, se habrá comprobado la constancia del índice en una muestra en particular, aquella que sigue el orden del texto real.

En la Figura 19 se muestran los gráficos que se obtienen según este doble procedimiento para diferentes índices llevados a cabo en la obra *Llibre de Job*. De forma similar a como ya se ha hecho con los gráficos anteriores, se han tomado veinte muestras uniformemente repartidas a través de toda la longitud de las 23 181 palabras

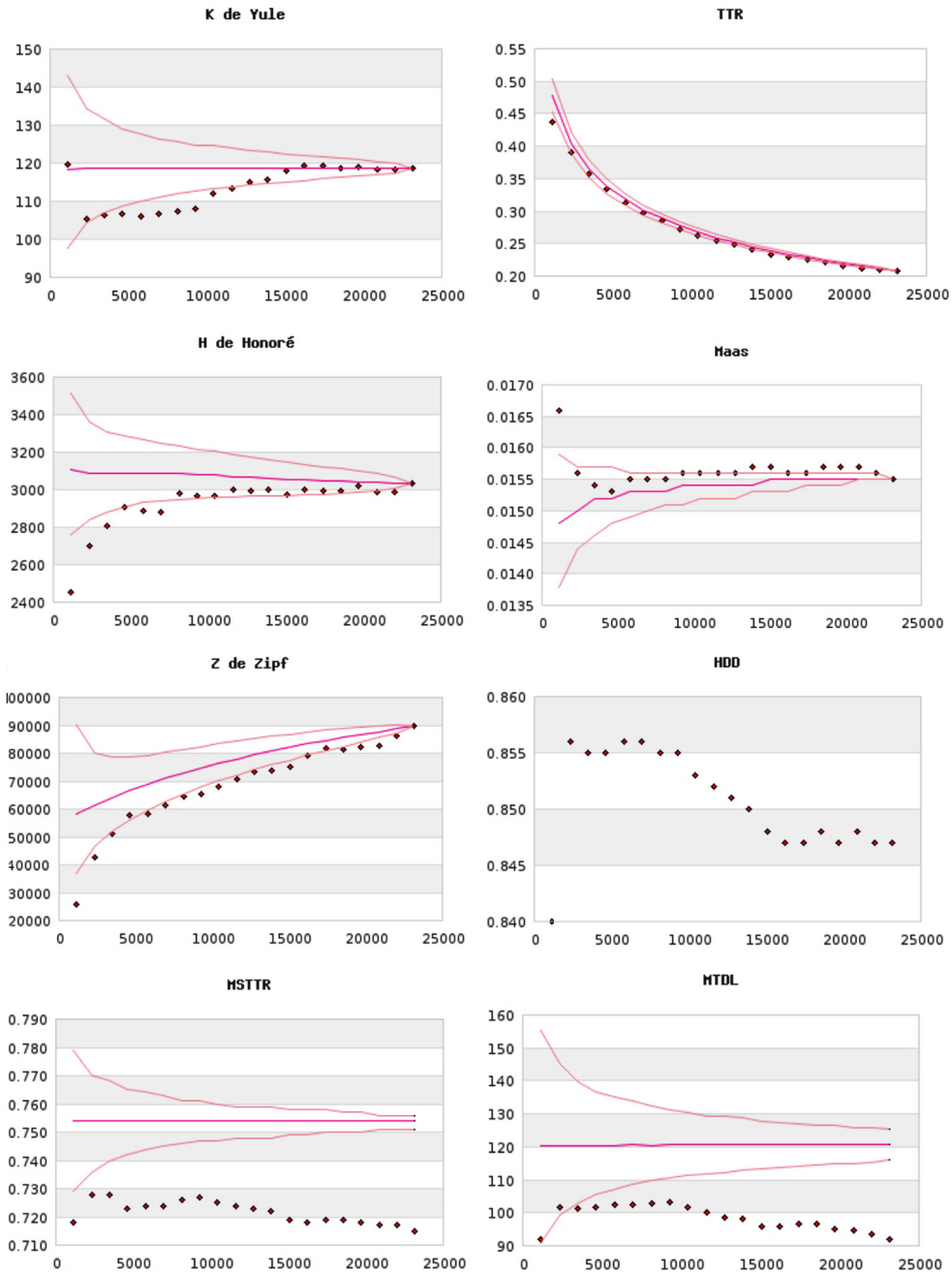


Figura 19

de la obra, que corresponderán a muestras que se van incrementando cada una con 1159 palabras. La línea discontinua formada por los puntos corresponde a los resultados de la muestra secuencial. Las tres líneas continuas corresponden a los resultados de las muestras aleatorias y se han obtenido procediendo de la manera que se explica a continuación. Para cada una de las veinte longitudes se han tomado 5000 muestras al azar (usando nuestro aplicativo informático, mediante sorteo aleatorio entre todas las palabras del texto) con las que se ha calculado el valor del índice correspondiente, obteniéndose por cada uno de los 20 puntos de medición 5000 valores diferentes del índice. Con estos 5000 valores se han calculado dos cosas: primero su media, la línea más intensa y centrada es el gráfico de estas medias; y segundo, el intervalo de confianza del 95%, el cual se hace ordenando los 5000 valores, descartando el 2,5% de los valores menores (125 valores) y los 2,5% de los valores mayores y representando el máximo y el mínimo de los valores restantes; las dos líneas menos intensas y situadas a ambos lados de la línea de las medias corresponden al gráfico de estos intervalos de confianza. Este procedimiento lo propone R. Harald Baayen en trabajos ya citados (1998 y 2001) y es una aplicación del conocido *Método de Montecarlo* para calcular intervalos de confianza.

La línea de las medias representa el comportamiento *teórico* de cada índice en el texto. Las líneas de los intervalos de confianza nos delimitan la zona donde deben caer los puntos de la observación secuencial para que su comportamiento, con sus variaciones, sea fruto del azar y que no exista ninguna razón significativa que explique este comportamiento; cuando los puntos de la observación empírica salgan de estos límites de los intervalos de confianza, indicará que la variación del índice no se puede considerar provocada por azar sino por las características del propio texto ordenado que dan unos valores al índice que se alejan de los que habría que esperar. Que esto sea así hay una probabilidad del 95%.

Si se revisan estos gráficos (Figura 19) se observa que, para la obra analizada, los cuatro índices *K*, *H*, *MSTTR* y *MTLD*, presentan la línea correspondiente a las medias totalmente horizontal; esto significa que desde el punto de vista teórico tienen muy buen comportamiento ya que mantienen siempre un valor constante al variar la longitud del texto. El índice de *Maas* ya presenta más variación teórica aunque tiende a la estabilización, y el índice *Z* tiene unos valores teóricos claramente ascendentes. El índice *HD-D* sólo tiene representados la serie de puntos porque ya inicialmente se calcula con todas las muestras aleatorias posibles del texto, y estos puntos ya representan los valores teóricos y no se pueden calcular los valores secuenciales siguiendo el orden real del texto; con todo, estos valores teóricos de la *HD-D* varían bastante.

Finalmente, el gráfico de la *TTR*, como era de esperar, sigue esa línea claramente descendente que ya habíamos visto anteriormente.

Si nos fijamos ahora en la sucesión de puntos que, para cada índice, representan las medidas correspondientes a las muestras que siguen el orden real de la obra, se observa que estas siguen distribuciones similares a las que ya se han ido viendo en los apartados anteriores en los gráficos correspondientes a la explicación de cada índice, aunque en aquel caso eran de la obra *Curial e Güelfa* y cada texto presenta sus distribuciones específicas. En el caso de los actuales gráficos de la Figura 19, se comprueba que el comportamiento de estas medidas correspondientes al texto real es notablemente menos estable que el que presentaban las medidas aleatorias y la propiedad de ser constantes es más discutible en todos los índices representados. Esta pérdida de constancia se justifica por el hecho de que las medidas se basan en un solo caso (el del texto real), mientras que la línea de las medias se calcula a partir de 5000 muestras de texto aleatorio. En la mayoría de los índices también se observa que algunos puntos están fuera del intervalo de confianza. Por ejemplo, en el índice *K* los puntos correspondientes desde la tercera hasta la décima muestra, están por debajo de la línea inferior del intervalo: esto significa que en esta zona del texto existe una razón lingüística que provoca una disminución significativa de la repetición léxica (que es lo que realmente mide este índice); en el índice *H*, para las seis primeras muestras, se observa una disminución significativa de la riqueza léxica; en el índice *Z*, la mayoría de puntos se encuentran por debajo del intervalo y, contrariamente, en el índice de *Maas* los puntos correspondientes a la segunda mitad están por encima del intervalo.

Los gráficos de los seis índices hasta ahora comentados (*K*, *TTR*, *H*, *Maas*, *Z* y *HD-D*) siguen todos un mismo patrón en el sentido de que las cuatro líneas del gráfico, a medida que aumenta la longitud de la muestra, van convergiendo y al final coinciden todas en el mismo punto que corresponde al valor del índice para el texto completo. Cuando trabajamos con el texto completo, las 5000 muestras formadas por la totalidad de las 23 181 palabras de la obra del *Llibre de Job* tienen todas ellas las mismas palabras (todas), la única diferencia está en el orden, pero en el cálculo de estos índices el orden que ocupan las palabras no tiene importancia, sino sólo el hecho de que las palabras sean diferentes. Por eso el valor del índice será el mismo para las 5000 muestras y así los cuatro valores correspondientes al texto real, a la media y los dos extremos del intervalo coincidirán. Las muestras correspondientes al penúltimo punto tienen 23 181-1159 = 22 022 palabras; esto quiere decir que todas las muestras deben ser *bastante iguales* ya que sólo tendrán, como máximo, 1159 palabras diferentes, por lo que los valores de los índices de las muestras serán muy

similares y, por tanto, también lo serán los cuatro valores (real, media y extremos). Estos razonamientos justifican la convergencia de las cuatro líneas del gráfico ya que cuanto más cerca esté el punto de medición del final del texto, más se parecerán las 5000 muestras y también los cuatro valores mencionados.

Pero este patrón de convergencia no lo siguen los índices *MSTTR* y *MTLD* ya que, como se ve en los gráficos de la Figura 19, aunque las líneas que representan los extremos del intervalo tienden hacia la línea de las medias, no llegan a juntarse. Esto se debe a que el cálculo de estos dos índices se basa en dividir la muestra en un gran número de segmentos (de 100 palabras para el *MSTTR* y de longitud variable para el *MTLD*), lo que provocará que, por ejemplo, en dos muestras con todas las 23 181 palabras de la obra, debido a su formación aleatoria, el orden de las palabras será diferente y, por tanto, los segmentos en que estarán divididas cada una de las dos muestras también; por este motivo, se obtendrán valores diferentes en el cálculo del índice y no habrá convergencia completa de los cuatro valores (real, media y extremos).

En estos dos gráficos también se comprueba que la riqueza léxica calculada secuencialmente en el texto real es bastante inferior al que da la media de los valores aleatorios. Esto es así ya que, en los segmentos formados en las muestras aleatorias, el hecho de estar seleccionados al azar hace que tengan más variación, en cambio la estructura y significado lingüísticos de las muestras del texto real condicionan la elección de las palabras proporcionando una variación menor.

10. COMPARACIÓN GRÁFICA DE DIFERENTES TEXTOS

Según se ha visto, el análisis de la riqueza léxica presenta dos aspectos comprometidos que nos obligan a estar atentos en su tratamiento: primero, la constancia de los índices en aumentar la longitud del texto no siempre se mantiene y, segundo, en las representaciones gráficas, el orden de magnitud que se decide utilizar en los números (hasta las unidades, o hasta las décimas, o hasta ...) de la escala gráfica hará que el gráfico quede más ampliado o menos ampliado, condicionando la observación de esta constancia de los índices. Cuando se quiere hacer una lectura del gráfico para obtener valores *absolutos* de la riqueza léxica (referidos a un único texto) nos faltan referencias para saber con qué orden de magnitud se deben representar los valores. Tanto la constancia de los índices como el orden de magnitud en su representación pierden importancia cuando se utilizan los gráficos para hacer una lectura *relativa*, esto es, cuando se comparan los gráficos de dos textos para saber cuál tiene una mayor riqueza léxica.

Así una manera simple y eficaz para saber si un texto tiene mayor riqueza léxica que otro es hacer en un mismo gráfico la representación de los valores de los dos índices según la longitud del texto.

En la Figura 20 se ha hecho para dos textos: *Llibre de Job* y *Il·lustracions dels comtats de Rosselló*, *Cerdanya y Conflent*[10]. Se ha seguido representando los valores correspondientes a las diferentes muestras que siguen el orden secuencial del texto real y también los valores aleatorios (tanto los valores medios como los dos extremos de los intervalos de confianza al 95%) correspondientes a 5000 muestras obtenidas en el azar.

La conclusión es clara e inequívoca: la obra *Llibre de Job* tiene mayor riqueza léxica que la obra *Il·lustracions*. Los ocho gráficos muestran esta misma conclusión. Cabe recordar que los índices *K* y *Mass* miden la inversa de la riqueza léxica (la repetición), por eso el orden de los valores para los dos textos en estos gráficos se presentan invertidos.

En esta comparación, la constancia de los índices no tiene una importancia decisiva sino que lo que realmente importa es la situación relativa entre las dos distribuciones de los puntos correspondientes a cada texto. En este sentido, incluso se puede utilizar la *TTR*, índice que se caracteriza por su manifiesta no-constancia, ya que, aunque sus valores disminuyen de forma muy notable, cada valor de la *TTR* del *Llibre de Job* siempre se mantiene inequívocamente por encima del correspondiente valor de la *TTR* de *Il·lustracions*.

La comparación de la riqueza léxica de diferentes textos a través de los gráficos también se puede hacer para más de dos textos. Por motivos de claridad gráfica, sólo se representan los puntos que corresponden a la medida secuencial del texto real y ya no aparecen las medidas aleatorias.

El la Figura 21 se presentan los gráficos que comparan cinco textos diferentes. Por razones de coherencia entre los diferentes índices, es de esperar que el orden en que aparecen las líneas de cada texto sea el mismo en todos los gráficos, aunque hay que recordar, una vez más, que en los casos de los índices *K* y *Maas* el orden es el inverso. Este hecho se cumple en todos los casos, excepto en uno: en el gráfico del índice *K*. En este gráfico, el orden con el que aparece el *texto 1* respecto al *texto 4* es el contrario a lo que aparece en todos los demás gráficos. Este comportamiento del índice *K* de no preservar el orden se ha observado también en otras comparaciones.

[10] Francesc Comte, *Il·lustracions dels comtats de Rosselló*, *Cerdanya y Conflent*, ed. Joan Tres (1995).

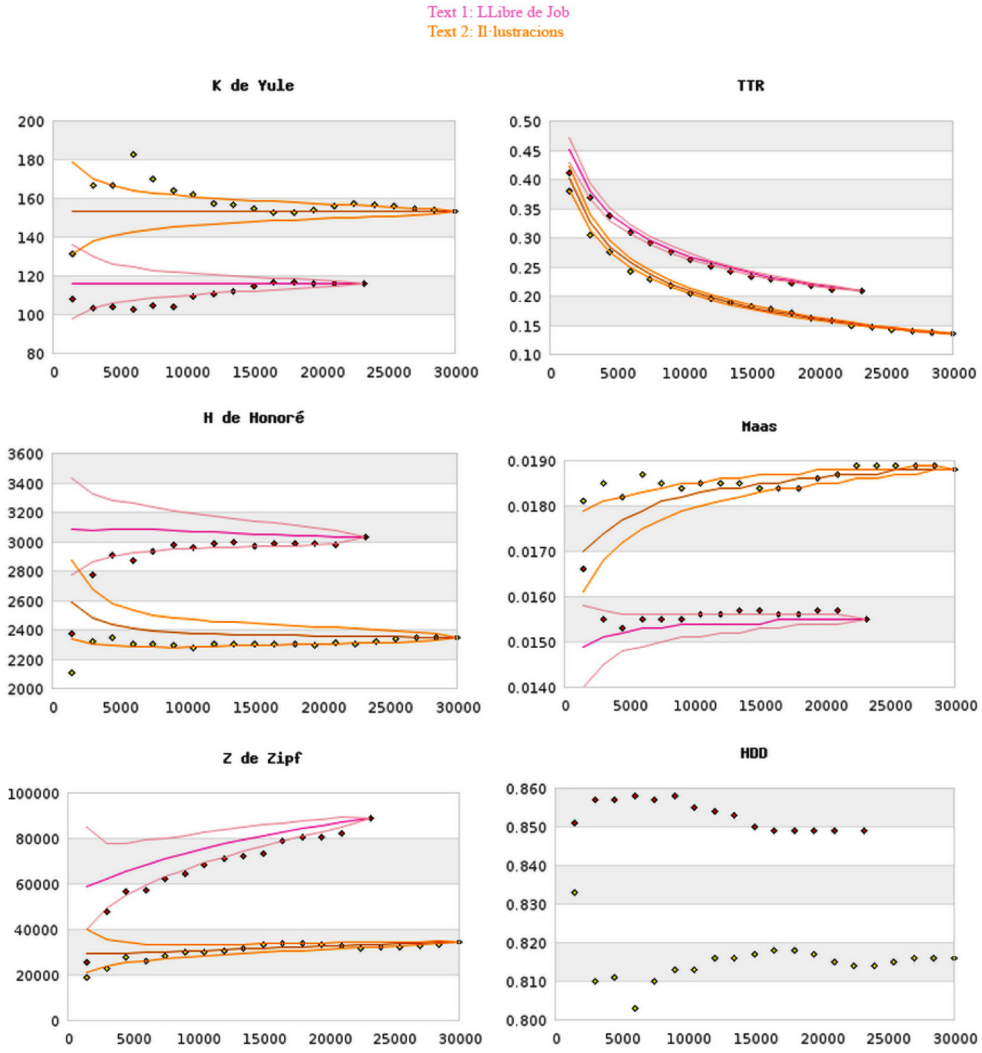

Text 1: LLibre de Job
Text 2: Il·lustracions
K de Yule
TTR
H de Honoré
Maas
Z de Zipf
HDD

Figura 20

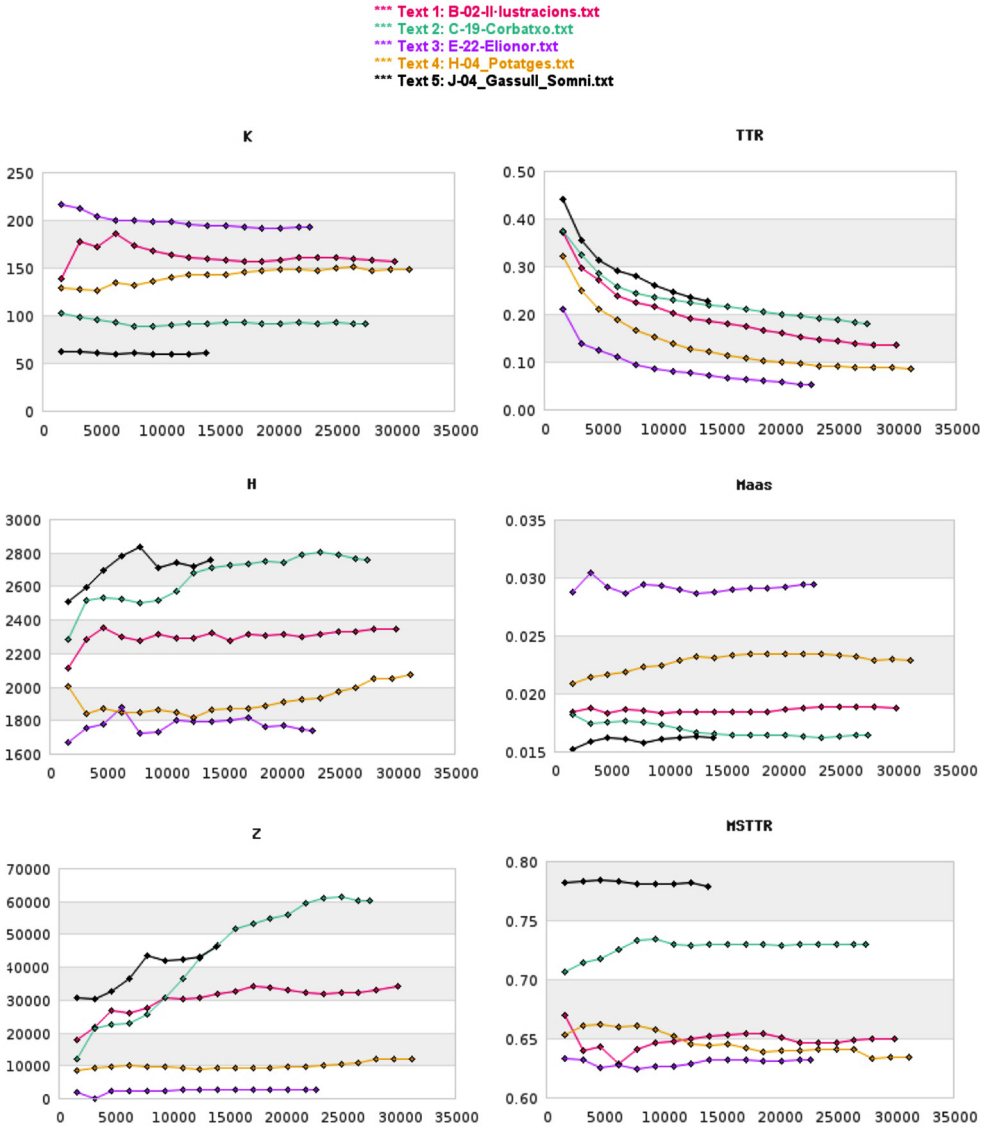

*** Text 1: B-02-Il·lustracions.txt
*** Text 2: C-19-Corbatxo.txt
*** Text 3: E-22-Elionor.txt
*** Text 4: H-04_Potatges.txt
*** Text 5: J-04_Gassull_Somni.txt
K
TTR
H
Maas
Z
MSTTR

Figura 21

11. CUALIDADES DESEABLES QUE DEBE TENER UN ÍNDICE DE MEDIDA DE LA RIQUEZA LÉXICA

Los gráficos anteriores también presentan algunas limitaciones. Por ejemplo, aparecen líneas de diferentes índices que se cruzan o tienen tendencia a cruzarse; así las líneas del *texto 5* y del *texto 2*, en la mayoría de los gráficos (índices *Z*, *H*, *TTR* y *Maas*), se cruzan o se cruzarían si el *texto 5* fuera más largo. ¿Cuál de los dos textos tiene mayor riqueza léxica?

A estas limitaciones resultantes de la comparación entre textos, también se deben añadir las ya mencionadas en apartados anteriores, como la no constancia de los índices y la dificultad de elegir un orden de magnitud adecuado.

En este apartado se analizarán estas limitaciones con herramientas numéricas. Para ello se examinarán las cualidades deseables que un buen índice debe tener para medir la riqueza léxica de forma satisfactoria y eficaz. A nuestro juicio, un *buen* índice debe tener tres cualidades:

1. Debe ser estable. Para un texto determinado, su valor tiene que mantenerse constante independientemente del tamaño de la muestra.
2. Debe ser sensible. Debe poder tomar una gama con una gradación de valores suficientemente amplia que permita diferenciar todos los textos, también los que tengan riqueza similar.
3. Debe ser coherente con los demás índices. Los valores de un *buen* índice calculados en conjuntos amplios de textos deben estar fuertemente relacionados con los valores obtenidos con otros índices *buenos*, y no lo deben estar con los valores de índices *no buenos*.

Los índices que se analizarán para saber si sus comportamientos se pueden considerar como *buenos* serán los siguientes: *K* de Yule, *H* de Honoré, *Maas*, *Z* de Zipf, *MSTTR*, *MTLD* y *HD-D*. A partir de lo expuesto en los apartados anteriores, se han escogido estos índices por ser los mejor considerados en la literatura especializada y también porque son los que mejores expectativas nos han ofrecido en las pruebas previas que nosotros hemos realizado. También se analizará la *TTR* como caso de índice *no bueno*, como elemento de contraste.

El análisis se hará utilizando un número elevado de textos con la intención de que los resultados obtenidos tengan valor estadístico en el sentido de que superen el comportamiento de casos particulares y representen patrones de comportamiento de carácter general. En concreto, se aplicarán los cálculos a la totalidad de los 414 textos que conforman el *Corpus Informatitzat del Català Antic* (*CICA*).

11.1 Estudio de la estabilidad

Para estudiar la estabilidad de los diferentes índices, para cada texto, se marcan 50 puntos de medición uniformemente distribuidos, de tal manera que se obtienen 50 muestras cada vez más largas. Si se representa por *N*, el número total de palabras del texto, la primera muestra tiene *N*/50 palabras, la segunda 2·*N*/50, la tercera 3·*N*/50 y así sucesivamente. Se hacen dos tipos de análisis: el primero es el secuencial y en él las muestras siguen el orden real del texto; el segundo es el aleatorio, donde se miden 500 muestras aleatorias para cada longitud. Es el mismo procedimiento que se ha utilizado para hacer los gráficos de los apartados anteriores, pero esta vez con 50 puntos de medición. Para cada tipo de análisis se calcula el valor del índice para cada uno de los 50 tamaños de las muestras, después se calcula la media de los 50 valores obtenidos y finalmente se observa cuánto se desvía, por término medio, cada valor respecto a la media; esto se hace calculando el coeficiente de variación[11]. De esta manera, se obtiene un coeficiente de variación para cada texto. Para terminar se calcula la media de todos ellos y también su coeficiente de variación. Analizando los 414 textos del *CICA*, se han obtenido los siguientes resultados:

Análisis secuencial	*TTR*	*Z*	*K*	*MTLD*	*H*	*MSTTR*	*Maas*	*HD-D*
Media de los coeficientes de variación	0,38	0,26	0,06	0,05	0,05	0,03	0,03	0,008
Coeficiente de Variación de los coeficientes de variación	0,20	0,34	0,70	0,64	0,40	0,69	0,53	0,59

En esta tabla se han ordenado los índices de forma decreciente según el valor medio de los coeficientes de variación (*cv*). El valor mayor, como era de esperar, corresponde a la *TTR* ya que es el índice genuinamente no constante. Destaca el valor correspondiente a *Z* puesto que también es alto, de un orden similar al que corresponde a la *TTR*; esto indica que *Z* no puede considerarse constante ya que está a un nivel cercano a la *TTR* y muy alejado de los resultados obtenidos para los otros índices. Estos otros índices obtienen una media de los *cv* baja, lo que indica que, en términos generales, se puede considerar que tienen una constancia aceptable. *K*, *MTLD* y *H* obtienen una media de los *cv* cercana a 0,05; *MSTTR* y *Maas* están en un segundo grupo con una constancia algo superior y, finalmente, *HD-D* se encuentra en una posición destacada con un valor bajo de 0,008, lo que hace que sea, y con

[11] Según se ha explicado en el apartado 12, el coeficiente de variación se calcula dividiendo la desviación típica entre la media.

diferencia, el índice más constante, consecuencia de su cálculo partiendo de muestras aleatorias.

En la segunda fila de la tabla aparece, para cada índice, lo que se ha denominado *Coeficiente de Variación de los coeficientes de variación* (*CVcv*), que es el resultado que se obtiene al calcular el coeficiente de variación de la colección formada por cada uno de los coeficientes de variación de todos los textos. Los valores menores del *CVcv* corresponden a la *TTR* y *Z*, lo que indica que existe poca variación en los *cv* obtenidos para los diferentes textos; la mayoría tienen unos valores de un mismo orden que, como ya se ha indicado, son valores altos, lo que indica que estos dos índices se comportan mal en la gran mayoría de los textos. Por el contrario, los *CVcv* de los otros índices son mayores, destacando el 0,70 de *K*; esto indica que existen valores de los *cv* de los textos que son notablemente diferentes, es decir, no todos presentan valores bajos, sino que habrá algunos textos con un *cv* alto y que, por tanto, tendrán el índice mucho menos constante de lo que indica el promedio general de todo el corpus.

Análisis aleatorio	*TTR*	*Z*	*H*	*MSTTR*	*Maas*	*HD-D*	*MTLD*	*K*
Media de los coeficientes variación	0,42	0,09	0,03	0,03	0,03	0,008	0,006	0,004
Coeficiente de Variación de los coeficientes variación	0,17	0,49	0,70	0,75	0,37	0,59	0,46	0,48

Estos son los resultados de las medidas aleatorias de las 500 muestras obtenidas al azar. En este caso, los resultados se extreman: la *TTR* muestra más variabilidad (pasa de 0,38 a 0,42) y los otros índices mejoran en constancia, destacando el índice *Z* que pasa de 0,26 a 0,09, lo que indica que, trabajando con muestras aleatorias, *Z* obtiene una estabilidad aceptable, propiedad que se pierde cuando se trabaja con muestras que siguen el orden secuencial del texto. *MSTTR* y *Maas* se mantienen en valores similares a los resultados empíricos y *MTLD* y *K* mejoran significativamente, lo que implica que si se utilizaran estos índices con muestras aleatorias, que no es el caso habitual, se conseguirían unas altas garantías de estabilidad.

Estos resultados confirman la mayoría de las observaciones realizadas a partir de los gráficos de los apartados anteriores y en particular los de la de la Figura 19.

Es importante observar que los valores altos de la segunda fila correspondientes al *CVcv* nos indican, de forma similar al caso del análisis empírico, una alta variabilidad en los valores de los *cv* obtenidos para los diferentes textos y esto significa, una vez más, que las conclusiones sobre la estabilidad de los índices obtenidas anteriormente son precisamente de carácter general, pero que existen casos particulares que

son textos que no se comportan bien, presentando unos índices menos constantes, así como la existencia de algunos textos con unos índices excelentemente constantes.

Precisamente, para poder determinar cuál es la importancia de estos textos que presentan índices con comportamiento no constante, se ha querido realizar un segundo análisis de la estabilidad de los índices. Para ello se ha recurrido a herramientas propias de la inferencia estadística (aquella que permite obtener conclusiones generales para toda la población a partir del estudio de una muestra), utilizando un procedimiento basado en la técnica estadística llamada contraste de hipótesis.

Se parte de la idea de que si el índice se mantiene constante entonces el gráfico formado por los puntos, cuyas coordenadas son la longitud de la muestra y el valor del índice, correspondientes a todas las muestras posibles del texto debe ser una recta horizontal; es decir, la pendiente p de la recta debe ser cero. Para comprobarlo se parte de un conjunto de 50 muestras del texto cada vez más largas de la misma manera como se ha hecho en los análisis anteriores. Se calcula la pendiente p' de la recta de regresión (la que se ajusta más) de estos 50 puntos. La técnica de contraste de hipótesis permite definir un test estadístico para contrastar si a partir del valor obtenido de p' a partir de los 50 puntos, se puede deducir que el valor de p (correspondiente a todas las muestras posibles) es cero. Si esto es cierto se dice que se cumple la Hipótesis nula (*Ho*). Para comprobarlo, pues, se parte de los valores de las coordenadas de los 50 puntos, que representamos por (x_i, y_i), y se debe calcular el valor de p' así como el valor de un parámetro estadístico que se representa con la letra t. Esto se hace a partir de las fórmulas siguientes:

$$p' = \frac{50 \cdot \sum_{i=1}^{50} x_i \cdot y_i - \left(\sum_{i=1}^{50} x_i\right) \cdot \left(\sum_{i=1}^{50} y_i\right)}{50 \cdot \sum_{i=1}^{50} {x_i}^2 - \left(\sum_{i=1}^{50} x_i\right)^2}$$

$$t = \frac{p'}{\sqrt{\frac{\sum_{i=1}^{50} {y_i}^2 - \frac{\left(\sum_{i=1}^{50} y_i\right)^2}{50} - p'^2 \cdot \left(\sum_{i=1}^{50} {x_i}^2 - \frac{\left(\sum_{i=1}^{50} x_i\right)^2}{50}\right)}{48 \cdot \sum_{i=1}^{50} (x_i - m)^2}}}$$

El valor de t depende de la cantidad de puntos que se utilizan para su cálculo, que en nuestro caso ha sido de 50. La teoría estadística matemática dice que si se cumple la hipótesis de que el valor de p sea cero, los diferentes valores de t se ajustarán a una distribución de probabilidad determinada que se llama *t de Student*. La probabilidad de equivocarnos decidiendo que la hipótesis no se cumple aun siendo cierta se llama el nivel de significación del test. Con un nivel de significación del 0,05 y con un total

de 50 puntos, el valor que tiene esta distribución *t de Student* es de 2,0106. De esta manera se obtiene el criterio para aceptar o rechazar la hipótesis de que el valor de *p* sea cero: se acepta si el valor de *t* es menor de 2,0106, de lo contrario se rechaza.

Se ha aplicado este criterio de contraste de hipótesis para cada uno de los índices en cada uno de los 414 textos del corpus y después se han calculado dos valores medios que representan el comportamiento global de cada índice en todo el corpus: la media de las *p'* y el porcentaje de textos que cumplen la hipótesis nula (*Ho*), es decir, que superan el test y que permiten afirmar que *p* vale cero (a un nivel de significación del 0,05). Se han obtenido los resultados que se muestran en la tabla que aparece a continuación:

K		*TTR*		*H*		*MSTTR*		*MTLD*		*Z*		*Maas*		*HD-D*	
p'	*Ho*	*p'*	*Ho*	*p'*	*Ho*	*p'*	*Ho*	*p'*	*Ho*	*p'*	*Ho*	*p'*	*Ho*	*p'*	*Ho*
0,006	49	0,007	0	0,0009	46	0,0001	65	0,0008	52	0,000021	44	0,004	44	0,0009	43

Una vez más se confirman los malos resultados de la *TTR* con un 0% de textos que superen el test; así sigue sirviendo como contraste para los otros índices. El índice que obtiene mejores resultados es el *MSTTR* con un 65% de textos con pendiente cero y siendo la pendiente media significativamente menor (de 0,0001). Los otros índices también presentan valores de la pendiente media muy bajos y un porcentaje de superación del test de pendiente nula alrededor de 50%. No son resultados excelentes, pero, teniendo en cuenta que el test es muy estricto, son suficientes para concluir que estos índices presentan una estabilidad aceptable, siempre hablando en términos generales y teniendo en cuenta que habrá algunos textos en los que los índices presentarán una estabilidad insuficiente.

11.2 Estudio de la sensibilidad

Para estudiar la sensibilidad de los diferentes índices se toma, para cada uno de ellos, el valor máximo I_{max} y el valor mínimo I_{min} del conjunto de los valores obtenidos en todos los textos. La diferencia $I_{max} - I_{min}$ nos da la amplitud de la variación de los valores que el índice toma en todo el corpus, el cual se denominará *rango*. Al dividir esta diferencia entre el valor mínimo se obtiene el *rango relativo*:

$$rango\ relativo = \frac{I_{max} - I_{min}}{I_{min}}$$

Al dividir entre I_{min} se consigue equiparar las diferentes *escalas* en las que se expresa cada índice. El *rango relativo* expresará cuántas veces hace falta el valor mínimo del índice para abarcar sus posibles medidas diferentes. Cuanto mayor sea el rango relativo, mayor será la *sensibilidad* del índice.

Pongamos el ejemplo del índice *K*. Analizando el corpus se han obtenido los siguientes valores:

$$I_{max} = 816;\ I_{min} = 41$$

$$rango\ relativo = \frac{I_{max} - I_{min}}{I_{min}} = \frac{816 - 41}{41} = \frac{775}{41} = 19$$

Esto quiere decir que cuando se mide el índice *K* de todos los textos del corpus se obtienen valores comprendidos en una amplitud igual a 19 veces el valor I_{min} (41).

Al hacer este cálculo para los diferentes índices de todos los textos del corpus, se obtienen los siguientes resultados:

	MSTTR	*HD-D*	*Maas*	*H*	*K*	*TTR*	*MTLD*	*Z*
Rango relativo	2	9	12	15	19	21	37	172

Esta tabla indica que el índice *MSTTR* tiene muy baja sensibilidad, los índices *HD-D*, *Maas* y *H* tienen una sensibilidad intermedia, *K* y *MTLD* tienen una sensibilidad alta y Z la tiene muy alta.

Esta tabla es el resultado del análisis secuencial. El análisis aleatorio da los mismos resultados para todos los índices, exceptuando el *MSTTR* y el *MTLD*. Esto es así porque, como ya se había visto en los gráficos de la Figura 19, los valores de los índices cuando se miden en la totalidad del texto coinciden en los dos análisis, exceptuando los dos índices mencionados, ya que su cálculo se basa en dividir la muestra en un gran número de segmentos que, en el caso de ser aleatoria dará valores diferentes. Los resultados obtenidos son:

	MSTTR	*MTLD*
Rango relativo	2	73

Sin embargo, los rangos del índice *MSTTR* coinciden en los dos análisis (en realidad difieren en dos décimas). En el caso del *MTLD*, el rango aumenta considerablemente en el análisis aleatorio. Tal como se vio en el estudio de la estabilidad,

aquí también se observa que si se utilizara el índice *MTLD* con muestras aleatorias mejoraría su eficacia en aumentar significativamente su sensibilidad.

En resumen, si se interpretan conjuntamente los resultados presentados en el estudio de la estabilidad y de la sensibilidad de los índices se llega a las siguientes conclusiones:

— El índice *Z* no se puede considerar un buen indicador de la riqueza léxica ya que no se mantiene constante respecto a la longitud del texto, como indica la elevada media de los coeficientes de variación (0,26) obtenida en el análisis secuencial.

— El índice *MSTTR* tampoco puede considerarse un buen indicador de la riqueza léxica ya que tiene un rango relativo de valor muy pequeño (2), lo que indica que tiene muy poca sensibilidad.

— En general, los índices *K*, *MTLD*, *H*, *Maas* y *HD-D* se pueden considerar buenos indicadores de la riqueza léxica, ya que presentan una estabilidad aceptable (su media de los *cv* es baja) y también una buena sensibilidad (su rango es grande). Particularmente, sin embargo, habrá textos que por sus características algunos de estos índices no tendrán un buen comportamiento, en especial en cuanto a su constancia (su Coeficiente Variación de la Media de los *cv* es grande).

— Cuando se aplican estos índices utilizando muestras aleatorias su comportamiento mejora significativamente, ya que aumentan tanto su constancia como su sensibilidad. *K* y *MTLD* son los que destacan más en esta mejora en la aleatorización. Sin embargo, ni en uno ni en otro, los autores que los definieron propusieron su uso aleatorio por razones de mantener la coherencia textual al efectuar la medida. Contrariamente, el índice *HD-D* en su definición inicial sus autores ya la hicieron con la aplicación con muestras aleatorias y siempre se calcula con esta aleatorización, y es por eso que ha dado tan buenos resultados cuando se ha valorado su constancia, aunque en la valoración de su sensibilidad sus resultados no han sido tan buenos.

11.3 Estudio de la coherencia

Para estudiar la coherencia de los índices se han calculado los *coeficientes de correlación de Pearson* entre las diferentes parejas que se pueden formar con los 8 índices analizados. El *coeficiente de correlación* sirve para medir el grado de relación que existe entre dos variables. Se simboliza con una *r*. Su valor es un número comprendido entre -1 y 1. Si *r* vale 1, entonces existe una dependencia total (y lineal)

entre las dos variables; si *r* vale 0, no existe relación (lineal) entre las dos variables. Los valores negativos de *r* también indican relación, pero inversa, al aumentar uno, el otro disminuye, siendo el valor -1 lo que corresponde a la relación inversa total.

La coherencia se debería manifestar con la existencia de fuertes correlaciones entre los valores de los índices considerados buenos indicadores de la riqueza léxica, y también con la no existencia de correlación entre los índices buenos y los que no son considerados buenos.

Se han efectuado los cálculos de *r* en los valores de los índices calculados para todos los textos del corpus, tanto en las medidas empíricas como en las medidas aleatorias, y se han obtenido los siguientes resultados:

	H	MSTTR	MTLD	Mass	HD-D	Z	TTR
K	-0,503	-0,722	-0,703	0,645	-0,961	-0,454	-0,379
H		0,628	0,563	-0,814	0,609	0,752	0,254
MSTTR			0,885	-0,795	0,829	0,586	0,557
MTLD				-0,675	0,793	0,733	0,534
Mass					-0,757	-0,693	-0,555
HD-D						0,562	0,428
Z							0,205

Análisis secuencial. Valores de *r*

	H	MSTTR	MTLD	Mass	HD-D	Z	TTR
K	-0,503	-0,890	-0,812	0,645	-0,961	-0,454	-0,379
H		0,685	0,620	-0,814	0,609	0,752	0,254
MSTTR			0,899	-0,837	0,976	0,644	0,403
MTLD				-0,717	0,892	0,788	0,445
Mass					-0,757	-0,693	-0,555
HD-D						0,562	0,428
Z							0,205

Análisis aleatorio. Valores de *r*

Como era de esperar, muchos valores coinciden en las dos tablas ya que, tal como ya se ha indicado, los valores de los índices cuando se calculan en el texto completo son iguales tanto en el tratamiento empírico como en el tratamiento aleatorio, exceptuando en los índices *MSTTR* y *MTLD*. Por tanto, sólo los coeficientes *r* donde intervienen estos dos índices serán diferentes en las dos tablas.

Estos resultados confirman la idea inicialmente formulada de que entre los índices que son buenos indicadores de la riqueza léxica debe haber correlaciones altas y también la idea complementaria de que entre los que son buenos indicadores y los que no lo son, debe haber correlaciones bajas. Es el caso de la *TTR*, la cual presenta los valores más bajos de correlación con todos los demás índices.

Los resultados con muestras aleatorias, una vez más, son más consistentes que los secuenciales, ya que presentan valores menores para la *TTR* y mayores para los otros índices.

Destaca las altísimas correlaciones de *HD-D* con los demás índices, especialmente en los resultados aleatorios. Algunos de ellos llegan casi al 100% (0,976 entre *HD-D* y *MSTTR* así como 0,961 entre *HD-D* y *K*). El índice *HD-D* sólo presenta dos valores *r* que no están en este tan alto nivel, los de las correlaciones con *Z* y con *H* (exceptuando la *TTR*, como índice *malo* que es). También cabe destacar los altos valores de *r* que presentan *MTLD* y *MSTTR*, tanto entre ellos como en la relación entre estos dos índices con los de *K* y *Maas*, además del ya indicado *HD-D*.

Un último comentario sobre los valores negativos de *r* para los índices *K* y *Maas*, los cuales están de acuerdo con la ya indicada relación inversa de estos dos índices respecto a los otros.

12. EL ECLECTICISMO COMO UNA FORMA DE SÍNTESIS. CÁLCULO DEL NÚMERO DE ORDEN MEDIO EN EL CORPUS

A partir de los análisis realizados en el apartado anterior sobre las cualidades que debe tener un buen índice de la riqueza léxica, hemos considerado que hay cinco de estos índices que cumplen aceptablemente estas cualidades deseables: *MTLD*, *K*, *HD-D*, *H* y *Maas*.

También hemos visto que si se estudia la correlación (estudio de la coherencia) existente entre estos índices, aunque en algunos casos es muy alta, aparecen valores intermedios, lo que ratifica que estos índices miden aspectos diferentes.

La propuesta que ahora haremos nosotros es la de no escoger un único índice como el paladín de la riqueza léxica, sino, por el contrario, tratar de utilizar la información que nos pueden aportar todos estos cinco índices que han demostrado que presentan un buen comportamiento en sus medidas. De esta manera trabajaremos con más información y, además, el posible mal comportamiento de los índices en algunos textos, como ya hemos indicado que sucede, podrá quedar compensado.

Además de esta cuestión sobre si debe elegirse un único índice o bien un conjunto de los mejores, también queremos incidir en una segunda idea: cómo superar

la dificultad de dar un valor relativo a las medidas que se obtienen cuando se calcula el índice. Cuando decimos, por ejemplo, que la obra *Tirant lo Blanc* tiene un valor de $K = 110$, esto ¿qué significado relativo tiene? ¿Es una obra con alta, baja o intermedia riqueza léxica? Hace falta definir una escala de referencia que indique cuál es el valor relativo de los resultados que se obtienen.

Para conseguir estos dos objetivos, proponemos calcular para todos los textos del corpus (en nuestro caso, el *Corpus Informatitzat del Catalá Antic*) un *nuevo índice* que se obtenga a partir de la intervención de los cinco índices seleccionados y que también nos dé idea de cómo está situado el texto respecto a la totalidad de los textos del corpus.

Para calcular este *nuevo índice* se propone proceder de la siguiente manera:

1º. Para cada uno de los cinco índices se hace una ordenación de todos los textos del corpus, de menor a mayor según el valor que toma el índice, obteniéndose cinco listas de los textos ordenadas.
2º. Como consecuencia, a cada texto se le asocian cinco números de orden según las ordenaciones obtenidas.
3º. Para cada texto, se calcula la media aritmética de estos cinco números de orden. Esta media la llamamos *Número de Orden Medio en el Corpus* (*NOMC*).
4º. Se realiza una nueva ordenación de todos los textos del corpus, de menor a mayor según el *NOMC*. El número de orden que corresponde a cada texto según esta nueva ordenación lo llamamos *Número de Orden Medio en el Corpus, Relativo* (*NOMCr*).
5º. Para estandarizar este nuevo índice y conseguir una referencia convencional que permita observar qué posición relativa tiene cada texto respecto a la totalidad del corpus, se convierte el valor del *NOMCr* de cada texto a percentiles según el cálculo que se indica a continuación. Lo llamamos *Número de Orden Medio en el Corpus Percentil* (*NOMC%*).

$$NOMC\% = \frac{NOMCr}{N.\,total\ textos} \cdot 100$$

Los percentiles se obtienen cuando se transforma todo el rango de valores en un nuevo rango comprendido entre 0 y 100. El *NOMC%* indica el lugar que ocuparía el índice del texto una vez se hubiera efectuado la ordenación indicada, y recalculando los valores que corresponderían si el corpus tuviera un total de justamente 100 textos.

Todo este procedimiento queda ilustrado en las tablas que se exponen a continuación. La primera corresponde a las cinco ordenaciones obtenidas según el valor que

toma cada índice. Solo se presentan las diez primeras y las diez últimas posiciones. Para cada texto, consta el código de la obra[12] y el valor del índice correspondiente. La segunda tabla corresponde a la ordenación según el *NOMC*, con sus diferentes versiones. Aparecen los diez primeros y los diez últimos textos de la ordenación que se obtiene con los cuatrocientos catorce textos que contiene el corpus *CICA*.

El *NOMC%* cumple las dos condiciones que se habían propuesto: en su definición intervienen los cinco índices seleccionados y tiene un valor relativo respecto a la totalidad de los textos del corpus. Estas son las dos cualidades de este índice, es un índice *complejo* ya que se ha construido utilizando información que proviene de cinco fuentes diferentes, y es un índice *global* ya que se ha elaborado con información no sólo procedente de un texto aislado, sino del conjunto del corpus al que pertenece.

	K		H		MTLD		Mass		HD-D	
1	F-40	816.3	F-40	486	F-40	7.0	F-40	0,0889	F-40	0,119
2	E-45	754.0	J-14	821	E-35	11.9	J-14	0,0462	E-37	0,357
3	E-105	645.5	E-42	860	E-45	14.4	E-35	0,0375	E-65	0,357
4	E-35	538.9	E-32	887	E-105	14.9	E-32	0,0372	J-14	0,579
5	E-37	453.5	E-38	1212	E-37	15.8	E-37	0,0363	E-35	0,593
6	E-93	409.6	F-10	1221	L-02	16.2	E-93	0,0345	E-31	0,595
7	E-48	366.7	E-86	1246	E-32	16.3	E-39	0,0333	E-45	0,609
8	E-52	348.6	E-39	1248	E-38	22.1	F-37	0,0319	E-93	0,614
9	J-14	344.4	E-30	1250	E-93	22.1	E-45	0,0308	E-105	0,619
10	E-64	339.5	E-36	1277	F-10	24.7	D-15	0,0305	E-42	0,619
					…					
405	J-21	65.6	J-02	3004	J-13	155.8	J-01	0,0141	J-27	0,894
406	A-12	64.2	C-07	3033	J-03	157.3	E-55	0,0138	J-19	0,895
407	J-23	63.5	E-62	3038	J-01	165.7	J-20	0,0133	A-12	0,897
408	J-12	62.6	J-10	3053	J-10	167.1	L-01	0,0133	J-12	0,898
409	J-05	62.3	E-48	3181	A-13	168.5	J-12	0,0132	J-11	0,898
410	J-15	62.1	E-64	3225	J-09	181.5	L-02	0,0131	J-05	0,899
411	J-04	60,7	J-20	3278	A-12	196.1	J-10	0,013	J-09	0,901
412	J-11	59.3	J-19	3488	J-19	208	J-02	0,0124	J-15	0,902
413	J-09	58.7	J-15	4040	J-15	210,3	J-19	0,0118	J-04	0,902
414	J-02	40,9	L-01	4184	J-02	265.4	J-15	0,0112	J-02	0,931

Las 5 ordenaciones de los textos según los índices (de menor a mayor riqueza léxica)

[12] La letra inicial de código representa la tipología y su significado es el siguiente: A: prosa de ficción, B: crónicas y obras historiográficas, C: obras religiosas y morales, D: prosa cancilleresca, E: textos administrativos, F: textos jurídicos, G: libros de corte, H: textos científicos y técnicos, I: epistolarios y dietarios, J: poesía, L: obras gramaticales y lexicográficas. El número del código corresponde al orden de entrada de la obra al corpus.

Obra	NOMC	NOMCr	NOMC%
F-40	1	1/414	0,24
E-35	5.6	2/414	0,48
J-14	6.4	3/414	0,72
E-32	9	4/414	0,97
E-37	10	5/414	1,21
E-39	11.4	6/414	1,45
E-38	14.8	7/414	1,69
F-10	16	8/414	1,93
E-30	19.6	9/414	2,17
E-42	20,4	10/414	2,42
...			
J-17	398.4	405/414	97,83
J-04	398.6	406/414	98,07
J-09	399.2	407/414	98,31
J-07	400	408/414	98,55
J-01	400,6	409/414	98,79
J-12	405	410/414	99,03
J-10	405.4	411/414	99,28
J-19	408.8	412/414	99,52
J-02	411.8	413/414	99,76
J-15	412.4	414/414	100

Cálculo del Número de Orden Medio en el Corpus:
NOMC (de menor a mayor riqueza léxica)

13. CALIFICACIÓN DE LAS OBRAS DEL CORPUS

Aunque los percentiles nos dan una posición relativa del valor obtenido del índice para un texto determinado, no informan con bastante precisión de cuál es el nivel de calidad de la riqueza léxica del texto. Esto sí que ocurriría si los diferentes valores del índice aparecieran todos por igual, con la misma frecuencia, pero normalmente este no es el caso, ya que hay valores que son más frecuentes que otros. Por ello, para saber en qué nivel de calidad se encuentra el valor obtenido del índice para un texto determinado, hay que tener en cuenta la distribución de las frecuencias del conjunto formado por todos los resultados. En este caso, es mejor trabajar con el conjunto inicial de resultados (*NOMC*) antes de transformarlos en percentiles.

En la Figura 22 se presenta el diagrama de barras de esta distribución en el caso de que se agrupen los valores en 15 clases (las barras de color más claro). La media de esta distribución es de m = 207,5 y su desviación típica es σ = 104,2. En tono más oscuro, también se presentan las frecuencias que se obtendrían si el conjunto de

resultados se ajustaran exactamente a una *distribución Normal* con la misma media y con la misma desviación típica que tiene el conjunto de valores medidos. En este doble diagrama se puede observar que los datos experimentales se ajustan bastante a la *distribución Normal*. De hecho si se hace un test *chi cuadrado* (χ^2) de bondad del ajuste de la distribución experimental a la *Normal* (de media 207,5 y de desviación típica 104,2) lo supera, pero de una manera no demasiado holgada: el valor del *estadístico de contraste* es 13,7 y (tal como debe ser) es menor que el valor 22,4 de la distribución *chi cuadrado* correspondiente.

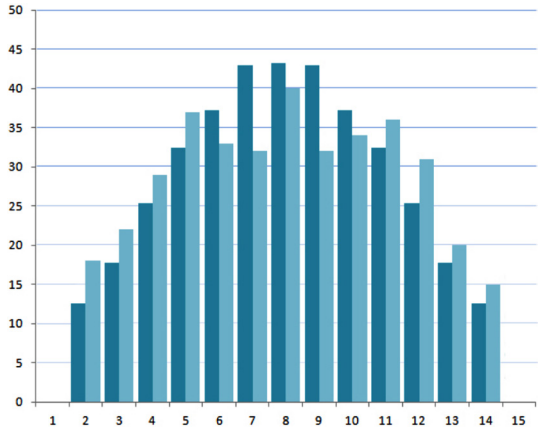
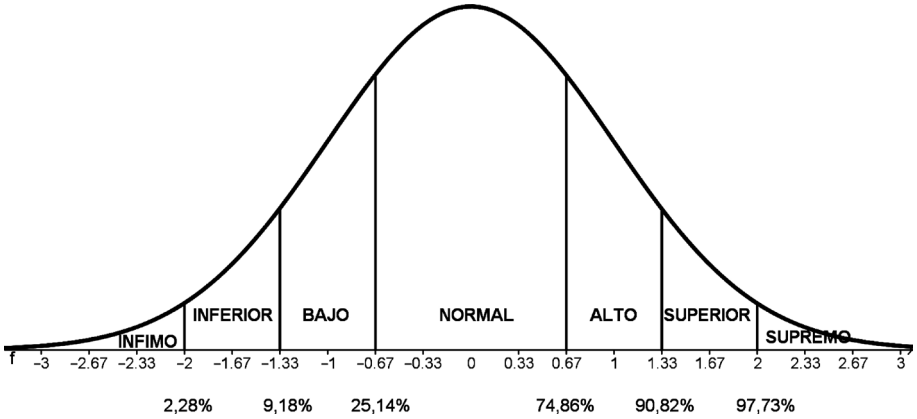


Figura 22

El nivel de calidad intermedio corresponde a los valores más frecuentes que son aquellos que están cercanos a la media y a medida que se separan de esta media la frecuencia de los valores va disminuyendo y, por tanto, su nivel de calidad se hace más extremo, ya sea en positivo o en negativo. Como la desviación típica nos indica lo alejados que se encuentran los diferentes valores respecto de la media, será precisamente esta desviación típica la que dará los criterios para saber el nivel de calidad de cada valor.

En este tipo de clasificación habitualmente se definen 7 clases: una clase central y, simétricamente, tres por encima y tres por debajo. Los puntos centrales de estas clases son *m* para la clase central, $m + \sigma$ y $m - \sigma$ para las primeras clases superior e inferior. Entonces, los límites de las 7 clases quedarán como indican la tabla y el gráfico que se presentan a continuación.

Límites	Nivel de calidad	Percentil dist. Normal	Percentil del NOMC
por debajo de $m-2\sigma$	Ínfimo	por debajo de 2,28	no existe
de $m-2\sigma$ hasta $m-1{,}33\cdot\sigma$	Inferior	de 2,28 hasta 9,18	de 0 hasta 10,63
de $m-1{,}33\cdot\sigma$ hasta $m-0{,}67\cdot\sigma$	Bajo	de 9,18 hasta 25,14	de 10,63 hasta 29,47
de $m-0{,}67\cdot\sigma$ hasta $m+0{,}67\cdot\sigma$	Normal	de 25,14 hasta 74,86	de 29,47 hasta 71,01
de $m+0{,}67\cdot\sigma$ hasta $m+1{,}33\cdot\sigma$	Alto	de 74,86 hasta 90,82	de 71,01 hasta 90,10
de $m+1{,}33\cdot\sigma$ hasta $m+2\sigma$	Superior	de 90,82 hasta 97,73	de 90,10 hasta 100
por encima de $m+2\sigma$	Supremo	por encima de 97,73	no existe



Distribución Normal con estas 7 zonas, el número de σ en el eje horizontal y los percentiles correspondientes

Este tipo de clasificación la han propuesto diferentes autores para calificar en diversos grados los valores de variables estadísticas que siguen una distribución normal. Entre estos autores destacan algunos investigadores en el campo de la psicología (Wechsler, McCall, Standford, por ejemplo), con la intención de desarrollar las escalas para medir la inteligencia.

Como la distribución de los valores del *NOMC* no se ajusta exactamente a la distribución Normal, los percentiles no coinciden del todo. Así, tal como se ve en la tabla, por ejemplo, en la primera y la última clase no hay ningún texto y la clase central presenta una longitud menor.

De esta forma, las 414 obras del corpus *CICA* quedan clasificadas según el nivel de riqueza léxica que presentan. A modo de ejemplo podemos citar algunas obras del

corpus indicando a que clase pertenecen. Para cada obra, consta el código (según los criterios indicados anteriormente en la nota 12) y el percentil escrito entre paréntesis:

Obras con una riqueza léxica *inferior*: F-37: *Capbreu de la Vall de Vives*, anónimo, s. XIII (4,83); I-45: *Lletres en català als bisbes d'Urgell 1*, anónimo, s. XIII (5,8); H-20: *Pràtica mercantívol*, Joan Ventallol, s. XVI (10,39).

Obras con una riqueza léxica *baja*: H-25: *Taula general*, Ramon Llull, s. XIII (13,29); H-13: *Lo sisè seny*, Ramon Llull, s. XIV (15,7); F-19: *Speculum prioris*, anónimo, s. XV (28,74).

Obras con una riqueza léxica *normal*: C-06: *Libre de Sancta Maria*, de Ramon LLull, s. XIV (33,09); B-06: *Llibre dels fets del rei en Jaume*, Jaume I, s. XIV (44,44); *Cartes d'una catalana del segle XIV al seu marit 2*, I-02: Sereneta de Tous, segle XIV (62,32).

Obras con una riqueza léxica *alta*: B-10: *Cròniques d'Espanya*, Pere Miquel Carbonell, s. XVI (74,15); A-02: *Tirant lo blanc*, Joanot Martorell, s. XV (84,06); A-04: *Curial e Güelfa*, anònim, s. XV (84,78).

Obras con una riqueza léxica *superior*: C-07: *Llibre de Job*, Jeroni Conques, s.XVI (91,55); J- 26: *Poesies*, Ausiàs March, s. XV (93,24); J-02: *Spill*, Jaume Roig s. XV (99,76).

14. CONCLUSIONES

Sin ánimo de ser exhaustivos, en este trabajo se han analizado 15 índices diferentes para medir la riqueza léxica de un texto (*TTR*, *RTTR*, Herdan, Somers, Maas, Dugast, Honoré, *MSTTR*, *MATTR*, *MTLD*, parámetro *D*, *HD-D*, *Z* de Zipf, Sichel y *K* de Yule). Este elevado número de índices que hemos hallado estudiando la literatura respectiva demuestra que es un tema tratado por muchos autores y de forma extensa. Nuestra primera intención ha sido la de presentar una perspectiva bastante amplia y completa que permitiera una visión global del tema, partiendo de la definición elemental de la *TTR*, de sus correcciones simples con radicales o logaritmos, presentando también procesos de cálculo más complejos basados en el uso masivo de muestras, en la ley de Zipf, y en la Distribución de frecuencias (de las frecuencias de palabras). La realización de gráficos de estos índices con 40 puntos uniformemente distribuidos en la longitud del texto que se corresponden a muestras de tamaño creciente, así como tener en consideración estudios ya realizados por otros autores, han sido los dos puntales donde hemos basado una primera evaluación, la cual nos ha permitido hacer una primera selección de los siete índices más recomendables: *MTTR*, *MTLD*, *HD-D*, Honoré, Mass, *Z* de Zipf y *K* de Yule.

A los índices de este primer grupo selecto se les ha aplicado un segundo examen gráfico, pero esta vez con muestras aleatorias. No sólo se han hecho los cálculos sobre el texto real sino también sobre 5000 muestras elegidas al azar para cada uno de los tamaños marcados por los puntos del gráfico. Esto ha permitido diferenciar el comportamiento teórico del índice y de su comportamiento correspondiente al orden real del texto y observar si este comportamiento se encontraba dentro del margen de error tolerable por el azar. La estabilidad teórica de estos índices es, en general, buena; la estabilidad empírica no es tan buena, aunque es aceptable, exceptuando el índice *Z* de Zipf que es el que presenta más variabilidad.

La eficacia de estas representaciones gráficas y el comportamiento no excelentemente estable de los índices nos han hecho ver que un método simple y eficiente para comparar la riqueza léxica de dos o más textos es precisamente el gráfico conjunto de sus valores. En este caso, la constancia de los índices no tiene una importancia decisiva, lo que realmente importa es la situación relativa de las dos distribuciones de los puntos correspondientes a cada texto.

Para profundizar en la valoración de estos índices se ha llevado a cabo un análisis de tipo estadístico en el conjunto de las obras del *Corpus Informatizat del Català Antic* sobre las tres cualidades que consideramos que debe tener un buen índice de medida de la riqueza léxica: la estabilidad, la sensibilidad y la coherencia.

Las dos pruebas realizadas para valorar la estabilidad han sido, por un lado, el cálculo de la media de los coeficientes de variación de los valores que toma el índice en 50 muestras de longitud creciente en cada texto y, por el otro lado, el contraste de la hipótesis de que la pendiente de la recta de regresión de los puntos, cuyas coordenadas son la longitud de la muestra y el valor del índice, sea cero. En este caso, el único índice que resultó manifiestamente no estable fue el *Z*; los otros índices dieron resultados medianamente buenos, lo que nos ha hecho concluir que en términos generales los otros seis índices son estables, pero no en todos los textos, puesto que existe un número de casos particulares donde el índice no es estable. *HD-D*, *MTLD* y *K* (estos dos últimos cuando se utilizaban muestras aleatorias) han sido los que han obtenido los mejores resultados de estabilidad.

Para valorar la sensibilidad se ha calculado el rango relativo de cada índice, dividiendo la diferencia del valor máximo y el mínimo que toma en todo el corpus entre este valor mínimo, obteniéndose cuántas veces el valor mínimo del índice hacen falta para abarcar sus posibles medidas diferentes. En este caso, ha sido el índice *MSTTR* el que no superaba la prueba, presentando una sensibilidad muy pequeña. Los otros seis índices presentaban una sensibilidad buena.

El cálculo del coeficiente de correlación de Pearson entre las diferentes parejas que se pueden formar con los 8 índices analizados (los siete en cuestión más la *TTR* como contraste) nos ha servido para evaluar la coherencia, obteniéndose los resultados previstos: entre los índices que son buenos indicadores de la riqueza léxica se presentan correlaciones altas y entre los índices que son buenos indicadores y los que no lo son se presentan correlaciones bajas.

La interpretación conjunta de los resultados obtenidos con esta triple valoración de la estabilidad, de la sensibilidad y de la coherencia, nos ha hecho concluir que existen cinco índices medidores de la riqueza léxica que se pueden considerar buenos (aunque no excelentes): *MTLD*, *HD-D*, *K* de Yule, Honoré y Mass.

Precisamente nuestra propuesta de definir el *Número de Orden Medio en el Corpus* (*NOMC*) parte del uso conjunto de estos cinco índices con la doble intención de utilizar el máximo de información disponible y la de posibilitar la compensación del posible mal comportamiento de alguno de los índices en algún texto. Por este motivo *NOMC* es un índice *complejo*. También es un índice *global* ya que se elabora con información no sólo procedente de un texto aislado, sino del conjunto del corpus al que pertenece. Asimismo, es un índice con *valor relativo* ya que se expresa en forma de percentiles, obteniéndose la posición relativa de cada texto respecto al conjunto total.

Por último, también hemos propuesto utilizar la distribución de frecuencias de los valores del *NOMC* en el corpus para dar una calificación del nivel de calidad de la riqueza léxica de las obras que componen el *CICA*. Como la distribución de frecuencias se ajusta a una distribución *Normal*, nos hemos inspirado en la escala de Wechsler (que utiliza el número de desviaciones típicas que el valor del índice está alejado de la media) para clasificar las obras en las siete categorías establecidas: *Ínfimo*, *Inferior*, *Bajo*, *Normal*, *Alto*, *Superior* y *Supremo*, aunque en el caso del *CICA* no existen obras en las dos categorías extremas.

BIBLIOGRAFÍA

Baayen, R. H. & Tweedie, F. J. (1998): "How variables may a constant be? Measures in lexical richness in perspective", *Computers and the Humanities* 32, pp. 323-352. https://doi.org/10.1023/A:1001749303137

Baayen, R. H. (2001): *Word frequency distributions*. Dordrecht: Kluver Academic Publishers. https://doi.org/10.1007/978-94-010-0844-0

Baayen, R. H. (2008): *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511801686

Bowker, L & Pearson, J. (2002): *Working with specialized language. A practical guide to using corpora*. London / New York: Routledge. https://doi.org/10.4324/9780203469255

Carroll, J. B. (1964): *Language and Thought*. New Jersey: Prentice-Hall, Englewood Cliffs.

CICA = Torruella, J & Pérez Saldanya, M. & Martines, J. (dirs.) (2013): *Corpus Informatitzat del Català Antic*. http://www.cica.cat [última consulta: 30/01/2016].

Covington, M. A. & McFall, J. D. (2010): "Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR)", *Journal of Quantitative Linguistics* 17/2, pp. 94-100. https://doi.org/10.1080/09296171003643098

Dugast, D. (1978): "Sur quoi se fonde la notion d'etendue theoretique du vocabulaire?", *Le Francais Moderne* 46, pp. 25-32.

Ferrando, A. (ed.) (2007): *Curial e Güelfa. Edició a cura de* –. Tolouse: Anacharsis.

Guiraud, P. (1960): *Problèmes et Méthodes de la Statistique Linguistique*. Dordrecht: D. Reidel.

Harris Wright, H. & Silverman, S. W. & Newhoff, M. (2003): "Measures of lexical diversity in aphasia", *Aphasiology* 17, pp. 443-452. https://doi.org/10.1080/02687030344000166

Hauf, A. (ed.) (2005): *Joanot Martorell. Tirant lo Blanch*. València: Tirant lo Blanch.

Herdan, G. (1955): "A new derivation and interpretation of Yule's 'Characteristic' K", *Zeitschrift für angewandte Mathematik und Physik* 6/4, pp. 332-339. https://doi.org/10.1007/BF01587632

Herdan, G. (1960): *Quantitative Linguistics*. London: Butterworth.

Honoré, A. (1979): "Some *Measures of Richness of Vocabulary*", *ALLC Bulletin* 7/2, pp. 172-177.

Jarvis, S. (2002): "Short texts, best-fitting curves and new measures of lexical diversity", *Language Testing* 19, pp. 57-84. https://doi.org/10.1191/0265532202lt220oa

Johnson, W. (1944): "Studies in language behavior: I. A program of research", *Psychological Monographs* 56, pp. 1-15. https://doi.org/10.1037/h0093508

Maas, H. D. (1972): "Zusammenhang zwischen Wortschatzumfang und L'ange eines Textes", *Zeitschrift für Literaturwissenschaft und Linguistik* 8, pp. 73-79.

McKee, G. & Malvern, D. & Richards, B. (2000): "Measuring vocabulary diversity using dedicated software", *Literary and Linguistic Computing* 15/3, pp. 323-337. https://doi.org/10.1093/llc/15.3.323

Malvern, D. D. (1989): *Thetype-token characteristic - an empirical investigation of a mathematical model for thetype-token ratio*. Reading: University of Reading, Faculty of Education and Community Studies. Unpublished working paper.

Malvern, D. D. & Richards, B. J. (1997): "A new measure of lexical diversity", in A. Ryan & A. Wray (eds.): *Evolving Models of Language*. Clevedon: Multilingual Matters, pp. 58-71.

Malvern, D. *et alii* (2004): *Lexical Diversity and Language Development. Quantification and Assessment*. New York: Palgrave Macmillan.

McCarthy, P. M. (2005): "An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD)", *Dissertation Abstracts International* 66/12, UMI No. 3199485.

McCarthy, P. M. & Jarvis, S. (2007): "Vocd: A theoretical and empirical evaluation", *Language Testing* 24/4, pp. 459-488. https://doi.org/10.1177/0265532207080767

McCarthy, P. M. & Jarvis, S. (2010): "MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment", *Behavior Research Methods* 42, pp. 381-392. https://doi.org/10.3758/BRM.42.2.381

Miller, J. F. (1991): "Quantifying productive Language disorders", in J. F. Miller (ed.): *Research on child language disorders: a decade of progress*. Austin: Pro-Ed, pp. 211-220.

Muller, C. (1968): *Initiation à la statistique linguistique*. Paris: Librairie Larousse.

Orlov, J. & Chitashvili, R. (1983): "Generalized Z-distribution generating the well-known rank-distributions", *Bulletin of the Academy of Sciences* 110, pp. 269-272.

Riera i Sans, J. (ed.) (1976): *Llibre de Job. Versió del segle XVI. Edició a cura de –*. Barcelona: Insititut d'Estudis Catalans.

Rojo, G. (2002): "Sobre la Lingüística basada en el análisis de corpus". Ponencia plenaria en las Jornadas sobre corpus lingüísticos (organizadas por Uzei, San Sebastián, octubre de 2002).

Rojo, G. (2008): "Lingüística de corpus y lingüística del español". Ponencia plenaria en el XV Congreso de la Asociación de Lingüística y Filología de América Latina (Montevideo, 18-21 de agosto de 2008). Disponible en: http://gramatica.usc.es/~grojo/Publicaciones/ Lgca_corpus_lgca_espanol.pdf.

Sichel, H. S. (1975): "On a distribution law for word frequencies", *Journal of the American Statististical Association* 70/351, pp. 542-547. https://doi.org/10.1080/01621459.1975.10482469 https://doi.org/10.2307/2285930

Sichel, H. S. (1986): "Word frequency distributions and type-token characteristics", *Mathematical Scientist* 11, pp. 45-72.

Somers, H. H. (1966): "Statistical methods in literary analysis", in J. Leeds. (ed.): *The computer and literary style*. Kent: Kent State University, pp. 128-140.

Templin, M. C. (1957): *Certain languages kills in children: Their development and interrelation ships*. Westport: Greenwood.

Tres, J. (ed.) (1995): *Francesc Comte. Il·lustracions dels comtats de Rosselló, Cerdanya y Conflent*. Barcelona: Curial.

Van Gijsel, S. & Speelman, D. & Geeraerts, D. (2005): "A Variationist, Corpus Linguistic Analysis of Lexical Richness", in *Proceedings from the Corpus Linguistics Conference Series*, vol. 1/1, pp. 1-16. http://www.birmingham.ac.uk/research/activity/corpus/ publications/conference-archives/2005-conf-e-journal.aspx.

Yule, G. U. (1944): *The Statistical Study of Literary Vocabulary*. London/Cambridge: Cambridge University Press.

Zipf, G. K. (1949): *Human Behavior and the Principle of Least Effort*. Cambridge/Massachusetts: Addison-Wesley.