UNIVERSIDAD AUTÓNOMA DEL
ESTADO DE MORELOS

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS
INSTITUTO DE INVESTIGACIÓN EN CIENCIAS BÁSICAS Y APLICADAS
Centro de Investigación en Ciencias

Análisis estadístico de textos

TESIS

QUE PARA OBTENER EL GRADO DE
DOCTOR EN CIENCIAS

PRESENTA

Diego Leonardo Espitia Cabrejo

DIRECTOR DE TESIS
DR. HERNÁN LARRALDE RIDAURA

Cuernavaca, Morelos Noviembre 2019

Et Earello Endorenna utúlien. Sinome maruvan ar
Híldinyar tenn'Ambarmetta.

– Elessar Telcontar

Resumen

En el presente trabajo presentamos los resultados del análisis estadístico de alrededor de 120 textos escritos en 14 idiomas distintos: Español, Inglés, Alemán, Francés, Turco, Ruso, Islandés, Checo, Danés, Finlandés, Hebreo, Húngaro, Italiano y Latín; al igual que de secuencias aleatorias de caracteres que usamos como hipótesis nula.

Para dicho análisis usamos diferentes herramientas. La primera de ellas es la teoría de redes. Específicamente construimos redes de co ocurrencia de palabras y calculamos, entre otras propiedades de la red, el *Clustering Coefficient*. Este nos permite proponer un método para la identificación automática de lenguas, a la vez que permite establecer una medida que puede ayudar definir una "distancia" entre idiomas.

Por otro lado, encontrando la posición de las palabras en el texto y haciendo estadística sobre la distancia (dada en número de palabras) entre dos repeticiones de un término, proponemos un método que permite encontrar de manera automática las palabras clave del texto. Si usamos dicho método en un mismo documento, pero escrito en diferentes idiomas, podemos correlacionar de manera precisa las mismas palabras claves en los diferentes lenguajes, estableciendo así que las palabras claves identificadas por éste método, parecen ser invariantes ante la traducción.

Tabla de Contenidos

Figuras

Tablas

Introducción 1

En las últimas décadas los límites de lo que tradicionalmente ha estudiado la física se han vuelto cada vez más difusos. En los años 50 o 60 era impensable que un físico se aventurara a indagar cuestiones propias de las humanidades o de las ciencias sociales y abandonara sus esfuerzos por entender a los átomos o a las estrellas.

A partir de los años 70, empezaron a surgir modelos tales como el modelo de Schelling sobre segregación [1], o el modelo de Axelrod, sobre la diseminación cultural [2], por mencionar algunos, que empezaron a difuminar el límite entre distintas disciplinas y a construir puentes entre las ciencias sociales y las ciencias naturales.

Tambíen desde la biología empezaron a surgir preguntas muy interesantes, cuya respuesta requiere un enfoque que va más allá del hecho de clasificar, describir e identificar a los seres vivos. Cuestiones tales como el rol que cumple la quimiotáxis en la fecundación de mamíferos o peces; o la relación entre la información y el ADN, entre muchas otras, necesitan de un nuevo enfoque que requiere el concurso de varias disciplinas científicas.

Es en este contexto en donde surge un nuevo paradigma en el quehacer científico: Los sistemas complejos [3].

Los sistemas complejos investigan la relación que existe entre las distintas partes que forman a un sistema, y la relación de este con su entorno. Dichas relaciones pueden dar lugar a comportamientos colectivos no triviales [3]. A diferencia de los sistemas simples, como podrían ser por ejemplo un péndulo, un oscilador; los sistemas complejos presentan caracteristicas, tales como la emergencia, [1] transiciones de fase, [2] son adaptativos [3] y presentan correlaciones a escalas grandes.

Al estudiar sistemas complejos, a menudo se encuentra que aunque los sistemas se pueden ver diferentes a escala microscópica, para escalas muy grandes, la descripción matemática se vuelve igual para muchos de ellos, y por lo tanto se dice que los sistemas complejos tienen propiedades universales, siendo éste el origen de la multidisciplinariedad que parece ser el sello de los sistemas complejos [3]

Ahora bien, existen dos herramientas que son muy importantes para el estudio de los sistemas complejos, y que juegan y jugarán, un papel fundamental en el desarrollo de esta nueva ciencia, el llamado *Big Data* y la teoría de redes [4].

Gracias al advenimiento de las computadoras, y de la internet, los datos, y sobre todo la información que se pueda extraer de ellos, son parte importante del análisis de sistemas complejos. Técnicas tales como el *Machine Learning* y la minería de datos son cada vez más populares para encontrar patrones estadísticos que lleven a inferir correlaciones. Y he aquí un debate interesante: Estamos en la era del

1: Que es cuando el sistema tiene propiedades que sus partes no tienen por sí mismas

2: Dado que los elementos que componen a un sistema complejo, interaccionan de manera tal que el comportamiento del sistema no puede predecirse del comportamiento de los elementos que lo forman, los sistemas complejos pueden cambiar su comportamiento repentinamente.

3: Adaptatividad hace referencia al hecho que los sistemas complejos responden a interacciones externas o internas, de manera que son capaces de autoorganizarse.

fin de las teorías; ahora la correlación reemplaza la causalidad, y la ciencia puede avanzar incluso sin modelos que ayuden a teorizar sobre el comportamiento de los sistemas bajo estudio. [5].

Desde ese punto de vista, este trabajo quizá sea anticuado, ya que no se hará uso de *Big Data*, justamente lo contrario, se usará una cantidad reducida de datos, y trataremos de entender los mecanismos que subyacen en la interacción de las partes del sistema que bajo estudio.

Es así, que en este trabajo, se hará un uso extensivo de la teoría de redes la cual permite estudiar de manera sistemática a los datos que estamos interesados en entender. Pero, ¿Qué datos se estudiarán en este trabajo?, los estamos viendo, las palabras son nuestros datos.

1.1 La Física y el Lenguaje

¿Por qué un físico podría estudiar el lenguaje? La respuesta trivial a ésta pregunta es: porque puede; simplemente la física es aquello que un físico hace.[3] Sin embargo, esta respuesta resulta ser muy simplista. Como se mencionó anteriormente, la universalidad permite que los sistemas complejos sean estudiados desde distintos puntos de vista, y los lenguajes al ser sistemas complejos, admiten una descripción estadística, que se ve muy favorecida por un enfoque multidisciplinario.

3: Esta frase ha sido mencionada por Sir Sam Edwards y por David Gross, entre otros.

En particular, los lenguajes humanos son sistemas complejos interesantes, ya que como portadores de información de alta complejidad, deben permitir una alta tasa de transmisión de información, y a la vez, deben ser robustos ante posibles errores de comunicación. Estas dos características, han hecho que los lenguajes hayan evolucionado de forma tal que tengan estructura estadística muy particular, en un punto de equilibrio entre el orden y el desorden. [6].

Resulta pues interesante caracterizar dicha estructura estadística, ya que se puede obtener de allí información muy importante, no solo sobre el lenguaje en sí, con miras a mejorar los algoritmos de traducción o corrección de textos, sino en relación con campos que no son estrictamente lingüísticos o computacionales, y que permiten investigar sobre el origen de los idiomas [7], textos encriptados [6, 8], variaciones del lenguaje, y como éstas pueden relacionarse con diferencias genéticas [9], o como la teoría de la información se puede usar para caracterizar idiomas [10], por mencionar algunos ejemplos.

Dichos ejemplos muestran la riqueza y diversidad de enfoques con los cuales se puede estudiar a los lenguajes, así como también la diversidad de herramientas que pueden usarse.

Ahora bien, en este trabajo haremos uso de la lingüística cuantitativa y de la teoría de redes para hacer el análisis estadístico de textos.

Lingüística cuantitativa

La lingüística cuantitativa es una subdisciplina de la lingüística que se ocupa de estudiar propiedades estructurales de los lenguajes con el fin de descubrir las leyes matemáticas subyacentes en los fenómenos lingüísticos, y así poder estudiar diversos aspectos de la lengua como la dinámica de su evolución y sus propiedades emergentes [11].

Aunque existen diversas leyes matemáticas asociadas a los lenguajes, quizá las que más han llamado la atención a físicos, matemáticos y linguistas son la Ley de Zipf [12], y la ley de Herdan-Heaps [13].

La ley de Zipf, es una ley que de manera empírica afirma que la frecuencia de cada palabra en un texto, está cerca de ser inversamente proporcional a su rango, es decir a su posición en una tabla de frecuencias

$$f(n) \propto \frac{1}{n^{\alpha}}, \tag{1.1}$$

donde α es el exponente que caracteriza a la distribución, generalmente $\alpha \approx 1$, y n es el rango de una palabra dada. [14–17]. Es decir la frecuencia de las palabras en un texto, sigue una ley de potencias. La ley de Zipf, aunque encontrada en el contexto de la lingüística, aparece de manera natural en muchos otras disciplinas, como son las poblaciones de ciudades [18], redes neuronales [19], etc.

Figura 1.1: Palabras más frecuentes en el Retrato de Dorian Gray en Español. El tamaño representa la frecuencia de las palabras, siendo las 5 más frecuentes de color anaranjado.

La ley de Herdan-Heaps, es también una ley empírica que relaciona longitud N de un texto, con $V(N)$, el número de palabras distintas que aparecen en él, está dada por [13]:

$$V(n) \propto N^{\beta}, \tag{1.2}$$

donde $0 < \beta < 1$, es decir el vocabulario en un texto crece sublinealmente con la longitud del documento. Se ha propuesto que las leyes de Zipf y Herdan-Heaps, se encuentran relacionadas entre sí.

En particular, resulta interesante ver que se puede derivar la ecuación (1.2) a partir de (1.1) [20]. Además se ha mostrado empíricamente que $\alpha \approx 1/\beta$, [13], estableciendo así un tipo de propiedad posiblemente universal entre la frecuencia de las palabras que forman un documento, con el vocabulario y la longitud del mismo.

Dichas leyes son las más usadas para modelar estadísticamente a los idiomas y son muy estudiadas en el contexto de búsqueda y recuperación de información [13], comunicación animal [21], e incluso, busqueda de vida inteligente extraterrestre [22].

Teoría de la información y Complejidad de Kolmogorov

La teoría de la información es una disciplina que estudia el intercambio y procesamiento de la información contenida en una secuencia simbólica s, a través de canales de comunicación imperfectos. Desde este punto de vista, la cantidad clave es la entropía de Shanon $H(s)$ que es una medida de la incertidumbre promedio de s. Es decir, es el número más pequeño de unidades de información (típicamente la información se mide en bits) necesario para, en promedio, describir la salida de la secuencia aleatoria s [23].

Por otra parte, la complejidad de Kolmogorov, es una medida que estima la cantidad de información de una secuencia de caracteres [23]. Sea s, una secuencia simbólica, entonces la complejidad de dicha secuencia $K(s)$, se define como la longitud (en bits) del programa más pequeño que produce como salida a la cadena. [24]. Dicho de otra manera, un texto tiene complejidad máxima si el programa mínimo usado para producirlo, es de longitud mayor o igual a s.

Aunque conceptualmente diferentes, la entropía de Shannon y la complejidad de Kolmogorov, son medidas que estan relacionadas, ya que la complejidad de Kolmogorov converge, a medida que la longitud de s tiende a infinito, a la entropía $H(s)$ [25].

Si se pudiese encontrar el programa más pequeño que describe a una secuencia simbólica, entonces podría usarse dicho programa para estimar la entropía de dicha secuencia. Aunque determinar el programa más pequeño que produce la cadena es un problema complicado, se ha propuesto que los algoritmos de compresión (por ejemplo el algoritmo de Lempel–Ziv–Welch, usado en los programas *.zip*. *.tar*, [26]), pueden usarse para estimar una aproximación a la complejidad de Kolmogorov, y por lo tanto usar dichos algortimos para estimar una cota superior de la entropía de un idioma [6, 10, 25].

Ahora bien, usando por ejemplo, el algoritmo de compresión *LZ77* [26], se puede comprimir un texto de tal manera que su complejidad estaría dada por

$$K(s) \approx \text{tamaño del archivo comprimido} + \text{tamaño del descompresor del algoritmo } LZ77. \quad (1.3)$$

Dado que el tamaño del descompresor será siempre el mismo, es decir constante para cualquier secuencia s, de longitud L; podemos estimar la complejidad de Kolmogorov por unidad de texto como

$$\frac{K(s)}{L} \approx CR^{-1}, \quad (1.4)$$

donde CR se conoce como la razón de compresión de datos. La razón de compresión es una forma apropiada para medir qué tan bien un algoritmo de compresión comprime un determinado conjunto de datos, al mirar la relación que existe entre el número de bits requeridos para representar los datos antes de la compresión y el número de bits necesarios para representar los datos después de la compresión [27]. Esta relación está dada por

$$CR = \frac{\text{Tamaño del texto sin comprimir}}{\text{Tamaño del texto comprimido}}. \quad (1.5)$$

Por lo tanto a partir de la definición (1.5), se puede por ejemplo calcular la razón de compresión para *El Conde de Montecristo* de Alexandre Dumas (padre).

	Alemán	**Francés**	**Inglés**	**Español**	**Turco**
Texto	2.1664	2.1433	2.0974	2.0986	2.1686
R1	1.9521	1.8771	1.8656	1.8789	1.9256
R2	1.8283	1.7811	1.7492	1.7722	1.8232
R3	1.6067	1.5789	1.5588	1.5819	1.5962
R4	1.2988	1.3097	1.2915	1.3044	1.3014

Tabla 1.1: Razón de compresión para *El Conde de Montecristo* en varios idiomas. Los textos $R1$ a $R4$ son textos aleatorizados de distintas maneras, pero construidos a partir de *El Conde de Montecristo*.

En la Tabla 1.1, se puede ver la razón de compresión para *El Conde de Montecristo*, y diferentes textos construidos a partir de él. Por ejemplo, $R1$ se construyó a partir de intercambiar aleatoriamente las palabras del texto. $R2$ se obtuvo al quitar todos los espacios en el documento original, y reintroducirlos de manera aleatoria (más detalles en el capítulo 2)[4] . $R3$ es igual a $R2$, pero barajando las "palabras" aleatoriamente; y $R4$ se obtuvó al intercambiar aleatoriamente todas las posiciones de las letras del documento original. De esta manera tenemos diferentes secuencias simbólicas, que van desde las mas complejas ($R4$), a las menos complejas (texto original).

4: Por ejemplo, las primeras palabras del texto son:
`el conde de montecristo alexandre dumas el castillo de if...`
y serían reemplazadas por:
`elc ond ed em o n te cristoal e xa ndr ed u maselc asti llod e ifc...`

Tal y como se ve en la tabla, no se observan grandes desviaciones de la razón de compresión para distintos idiomas. Esto implicaría que la estimación para la entropía de los textos comprimidos, es similar para los distintos lenguajes de la tabla.

Esto está en concordancia con lo hallado en [28], donde se muestra que un documento escrito en diferentes idiomas, se comprime aproximadamente en la misma proporción, y por lo tanto, de acuerdo con [25], la

entropía sería aproximadamente igual. En otras palabras, la entropía por símbolo es una propiedad universal de los lenguajes naturales.

1.2 Teoría de Redes

En esta sección se dará un breve resumen sobre la teoría de redes. A menos que se mencione lo contrario, toda la información de esta sección, proviene del libro *Network Science* de Albert-László Barabási, [29].

Para entender un sistema complejo, es clave entender cómo sus componentes interaccionan entre sí. Se dice entonces que una red es un catálogo de los componentes de un sistema, llamados nodos, y las interacciones que hay entre ellos, son representadas a través de aristas o enlaces.

Definición de una red

Formalmente, una red se define a través de un grafo; esto es, un par ordenado $G(V, E)$ tal que:
$\cdot V$ es un conjunto de nodos;
$\cdot E \subseteq x, y | (x, y) \in V^2 \wedge x \neq y$,
es un conjunto de pares ordenados (desordenados), llamados aristas o enlaces.

Los enlaces en una red pueden ser dirigidos o no dirigidos, dependiendo si los enlaces son pares ordenados o desordenados. En este trabajo nos centraremos en estudiar redes no dirigidas en donde los nodos serán las palabras, y los enlaces indicarán adyacencia de palabras en el texto.

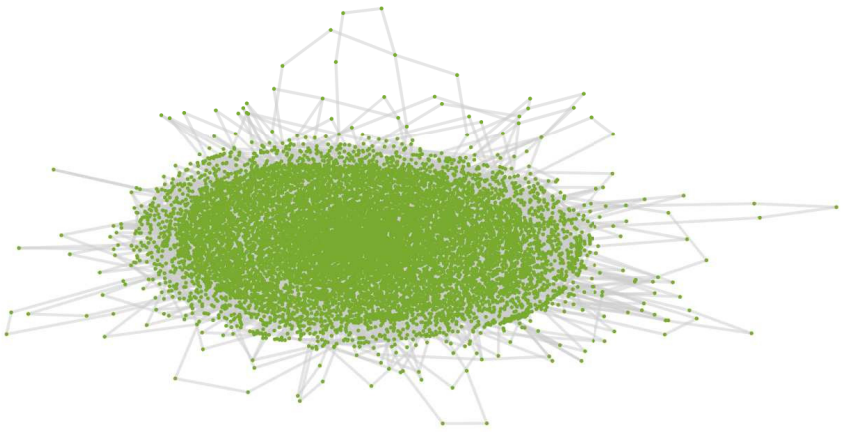


Figura 1.2: El retrato (del grafo) de Dorian Gray

En la Figura 1.2, se ve la representación de la red de co-ocurrencia de palabras en el libro *El retrato de Dorian Gray*. Claramente se observa que la red es demasiado compleja para comprender sus propiedades a través de una inspección visual, por lo tanto, es necesario recurrir a las herramientas de la teoría de redes para caracterizarla.

Matriz de Adyacencia

Para la descripción completa de una red, se requiere hacer una lista de los enlaces entre los nodos. Esto se logra a través de la matriz de adyacencia A de una red. La matriz de adyacencia de una red no dirigida de nodos tiene N filas y N columnas, siendo sus elementos:

- $A_{ij} = 1$ si hay un enlace que une al nodo i con el nodo j
- $A_{ij} = 0$ si los nodos i y j no están conectados entre sí

Al no tener una dirección de preferencia, la matriz de adyacencia de una red no dirigida es simétrica, es decir: $A_{ij} = A_{ji}$.

El grado de un nodo, es decir, el número de enlaces que tiene, se puede calcular directamente de la matriz de adyacencia como la suma de las filas de la matriz A:

$$k_i = \sum_{j=1}^{N} A_{i,j}. \tag{1.6}$$

Esta cantidad es muy importante en la teoría de redes, ya que la topología de una red, depende directamente de su distribución de grados.

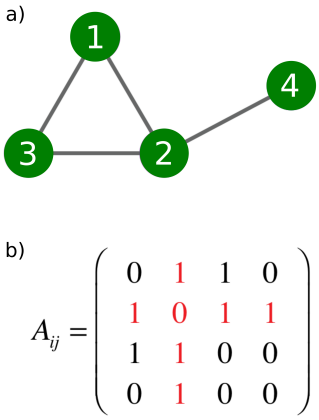


Figura 1.3: a) Ejemplo de un grafo. b) Matriz de Adyacencia.

Distribución de grados $p(k)$

La distribución de grados de una red $p(k)$, proporciona la probabilidad de que un nodo seleccionado al azar en la red tenga un grado k. Dado que $p(k)$ es una probabilidad, debe cumplirse que:

$$\sum_{k=1}^{\infty} p(k) = 1. \tag{1.7}$$

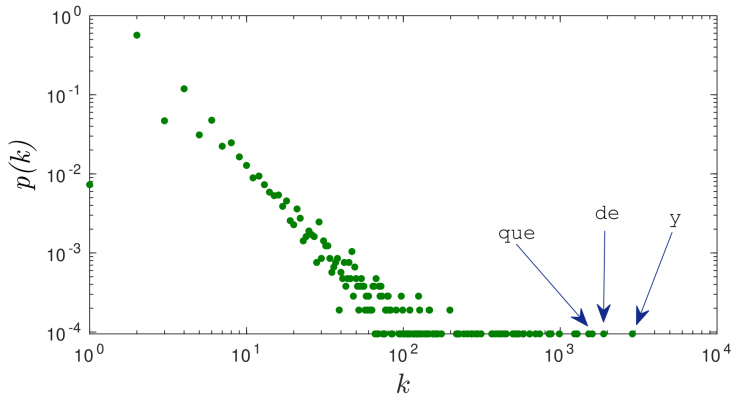


Figura 1.4: Distribución de grados para *el retrato Dorian Gray* en Español. Las palabras *y, de, que* tienen los valores de k más grandes y se denominan *hubs* de la red

En la Figura 1.4 se ve la distribución de grados de la red de co-ocurrencia para el *Retrato de Dorian Gray*. Nóstese que los *hubs*, es decir los nodos que tienen los valores de k más grandes, corresponden a aquellas palabras que son más frecuentes en el texto, ver Figura 1.1.

La existencia de estos nodos, densamente conectados, tiene como consecuencia que $p(k)$ tenga una distribución de ley de potencias[5] , es decir, la distribución de grados de la red de co-ocurrencia de palabras esta caracterizada por una distribución de la forma:

$$p(k) \propto k^{-\alpha}. \tag{1.8}$$

5: Nótese que la ley de potencias se satisface si ignoramos los nodos con grado $k = 1, 3, 5, 7$. Estos nodos son poco comunes en el texto, y vienen de casos muy puntuales, como cuando la palabra se encuentra precedida y sucedida por una palabra idéntica (*y fue, y dijo*), para $k = 1$ por ejemplo

En la Figura 1.4, se ve que en una gráfica log – log, $p(k)$ depende linealmente de k, siendo α la pendiente de dicha línea. Ahora bien, toda red que tenga una distribución de grados dada por una ley de potencias de la forma (1.8), se le llama **red libre de escala**. Muchas de las redes que aparecen al estudiar sistemas complejos, tienen esta característica, y por lo tanto, esta universalidad es una propiedad omnipresente en muchas disciplinas. Para determinar si una red es libre de escala, debe usarse un método estadístico apropiado que permita determinar el valor de α, algo que, cómo veremos más adelante no es trivial.

Existen también redes aleatorias, cuya distribución de grados sigue una distribución de Poisson. Estas redes se caracterizan por no tener *hubs*, por lo tanto los enlaces se distribuyen uniformemente entre los nodos.

Muchas propiedades de las redes libres de escala dependen del valor del exponente α en la ecuación (1.8). Por ejemplo, el tamaño del *hub* más grande esta dado por

$$k_{max} = k_{min} N^{\frac{1}{\alpha-1}}, \tag{1.9}$$

donde N es el tamaño de la red. En general se pueden encontrar 3 regímenes dependiendo del valor de α:

- **Régimen anómalo** ($\alpha \leq 2$): En este caso, el número de enlaces del *hub* crece más rápido que el tamaño de la red, tal como se ve de la ecuación (1.9), en donde el exponente es mayor a 1. Esto significa que para valores de N suficientemente grandes, la red no puede existir, ya que habrían más enlaces que nodos.
- **Régimen libre de escala** ($2 < \alpha < 3$): De la ecuación (1.9) se puede ver que el exponente en este caso es menor que 1, y por lo tanto el *hub* más grande crece por debajo del valor de N. Esto tiene como consecuencia importante que las redes en este caso serán de mundo ultra-pequeño.[6] Las redes estudiadas en este trabajo están en este régimen.
- **Régimen de redes aleatorias** ($\alpha > 3$): En este caso, es difícil distinguir a una red libre de escala de una red aleatoria del mismo tamaño, ya que los *hubs* crecen lentamente y para fines prácticos en estos nodos, el valor del grado k_{max} es comparable al k promedio de la red.

6: Se dice que una red es de mundo pequeño si la distancia promedio d que separa dos nodos elegidos al azar depende logarítmicamente del tamaño del sistema.
En el caso de redes libres de escala, dichas distancias son más pequeñas que en una red aleatoria de tamaño igual (del orden de $d \sim \ln \ln N$), de ahí que se denominen ultra-pequeñas.

Clustering Coefficient

El grado k_i, da información sobre cuántos vecinos tiene el i-ésimo nodo, pero no contiene información sobre la relación que puede haber entre los vecinos del nodo. El *Clustering Coefficient* mide la densidad de enlaces que tienen los vecinos del i-ésimo nodo; por lo tanto $C_i = 0$ significa que no existen enlaces entre los vecinos del nodo i; y $C_i = 1$ implica que cada uno de los vecinos de i se vinculan entre sí.

El *Clustering Coefficient* se define como

$$C_i(k_i) = \frac{2L_i}{k_i(k_i - 1)}, \tag{1.10}$$

donde, L_i representa el número de enlaces entre los k_i vecinos del nodo i.

1.3 Estimación del exponente de $p(k)$

Calcular correctamente el valor de α en la ecuación (1.8) requiere el uso de un método estadístico apropiado. Una simple regresión lineal no aplica en este caso, ya que existen fluctuaciones muy grandes en la cola de la distribución (ver Figura 1.4), y los resultados obtenidos pueden ser muy inexactos [30].

Para encontrar el valor de α, se debe seguir un método estadístico que resumiremos a continuación. Supongamos que un conjunto discreto de datos $\{x_1, x_2, \ldots, x_N\}$, se grafica en escala log – log, tal como en la Figura 1.4. Para determinar el valor de la pendiente en esta gráfica, se usa un *Maximum Likelihood Estimator* para la distribución de probabilidad de la ecuación (1.8), dado por:

$$\hat{\alpha} = 1 + N \left[\sum_{i=1}^{N} \ln \frac{x_i}{x_{min} - \frac{1}{2}} \right]^{-1}, \tag{1.11}$$

donde x_{min} es el dato para el cual se empieza a observar el comportamiento de ley de potencias. En nuestro caso, podemos determinar x_{min} mediante una inspección visual; de la Figura 1.4 es fácil ver que la ley de potencias se empieza a observar para aquellos nodos en donde $k > 7$. Sin embargo este valor se debe determinar de manera sistemática. Para ello, consideraremos la distribución cumulativa de $P(x)$ para el caso discreto

$$P(x) = 1 - \frac{\zeta(\hat{\alpha}, x)}{\zeta(\hat{\alpha}, x_{min})}, \tag{1.12}$$

donde $\zeta(\hat{\alpha}, x) = \sum_{n=0}^{\infty} (n + x)^{-\hat{\alpha}}$ es la función zeta generalizada de Hurwitz. Usando la ecuación (1.12) como distribución de referencia, es posible utilizar el test de Kolmogorov-Smirnov, el cuál es una prueba estadística que cuantifica la distancia entre las distribuciones cumulativas de referencia y de una muestra de datos; permitiendo aceptar o rechazar la hipótesis que los datos provienen o no de la distribución de referencia. Es decir, se puede comparar $P(x)$ y la distribución cumulativa de los datos $S(x)$ tal que

$$D = \max_{x \geq x_{min}} = |S(x) - P(x)|. \tag{1.13}$$

Por lo tanto, se debe calcular D para cada uno de los valores del conjunto de datos $\{x\}$ y se escoge el x_{min} en aquel valor que minimiza a D.

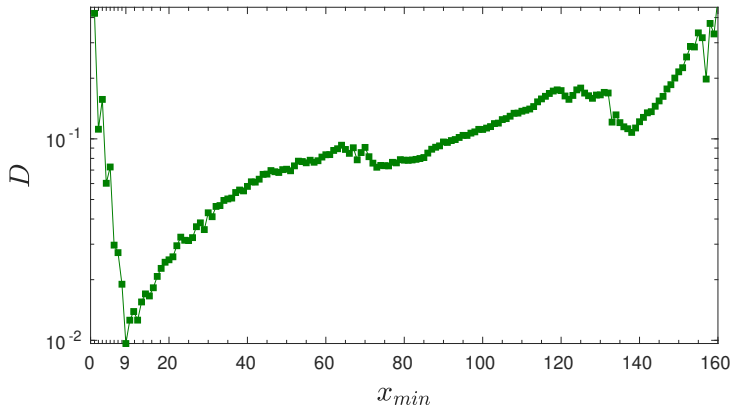


Figura 1.5: Test de Kolmogorov-Smirnov, donde se observa que el valor de $x_{min} = 9$ para $P(k)$ obtenido de los datos de la red de co-ocurrencia del *El retrato de Dorian Gray*

En la Figura 1.5, se ve que el valor de $x_{min} = 9$ minimiza a D, lo cual concuerda satisfactoriamente con la inspección visual. Determinado este valor, podemos usar la ecuación (1.11) para estimar $\hat{\alpha}$. En nuestro ejemplo $\hat{\alpha} = 2.1500$

Ahora bien, dado el conjunto de datos $\{x\}$ y la estimación para $\hat{\alpha}$ que los ajusta, es deseable saber si la hipótesis que los datos $\{x\}$ se distribuyen o no como una ley de potencias, es correcta. Para ello se usa una prueba de bondad de ajuste que genere un *p-value*, el cuál permite cuantificar dicha hipótesis. Para ello se debe generar datos sintéticos a partir de los valores estimados de $\hat{\alpha}$ y x_{min}.

Los datos sintéticos se pueden generar de la siguiente manera: con una probabilidad $P_z = n_z/N$, generamos $N - x_{min}$ números aleatorios a partir de la distribución $p(x)$ obtenida con los parámetros $\hat{\alpha}$ y x_{min}. Con probabilidad $P = 1 - P_z$, seleccionamos al azar un elemento del conjunto de datos que no sigue el comportamiento de ley de potencias, es decir seleccionamos elementos del subconjunto $\{x < x_{min}\}$. De esta manera generamos un conjunto de datos sintéticos que sigue a la ley de potencias para valores mayores a x_{min}, pero que para valores menores sigue la misma distribución (diferente a una ley de potencias) que el conjunto de datos.

En nuestro ejemplo N sería el número de datos (10560) y n_z sería el número de datos que siguen el comportamiento de una ley de potencias (10551).

Para obtener el *p-value* generamos 1000 conjuntos de datos sintéticos, a partir de las estimaciones de $\hat{\alpha}$ y x_{min}; luego calculamos nuevamente el test de Kolmogorov-Smirnov para cada uno de los datos sintéticos en relación con la ley de potencia que mejor ajusta al conjunto de datos.

$$D^* = \max_{x \geq x_{min}} = |DS(x) - P(x)|, \tag{1.14}$$

donde $DS(x)$ serían los datos sintéticos, y $P(x)$ estaría dado por la ecuación (1.12), con $x_{min} = 9$ y $\alpha = 2.1500$

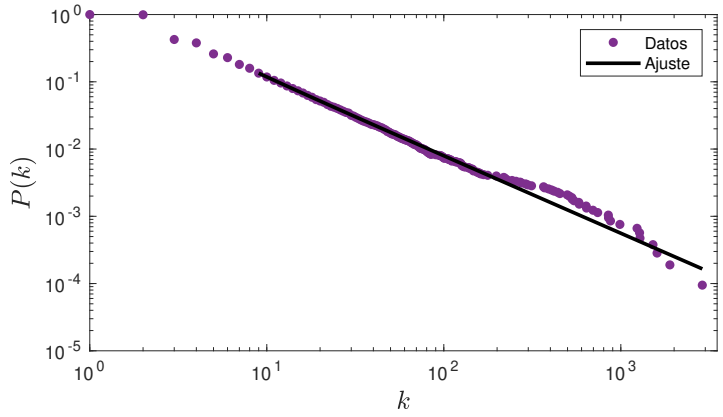


Figura 1.6: Distribución cumulativa de $P(k)$ para *El Retrato de Dorian Gray* y su respetivo ajuste

Finalmente, contamos la fracción de veces en que la estadística resultante es mayor que el valor de los datos empíricos. Esta fracción es nuestro *p-value*, es decir:

$$p\text{-}value = \sum_{i=1}^{1000} \frac{D_i^* \geq D}{N}. \tag{1.15}$$

Entonces si *p-value* es muy pequeño, se dice que no es plausible que el modelo ajuste a los datos. De acuerdo con Clauset et al, [30], si *p-value* ≤ 0.1 la hipótesis de la ley de potencias se debe descartar. En nuestro caso *p-value* $= 0.9110$

Se puede también estimar el error standard de α usando

$$\sigma = \left(N \left[\frac{\zeta''(\hat{\alpha}, x_{min})}{\zeta(\hat{\alpha}, x_{min})} - \left(\frac{\zeta'(\hat{\alpha}, x_{min})}{\zeta(\hat{\alpha}, x_{min})} \right)^2 \right] \right)^{-1/2}, \tag{1.16}$$

donde la prima denota la derivada respecto de α. Para el ejemplo que hemos venido trabajando $\sigma = 0.0305$

1.4 Sobre el presente trabajo

En este trabajo, presentamos un estudio que involucra el análisis de la ley de Zipf, la ley de Herdan-Heaps y de la distribución de grados de redes de co-ocurrencia en 14 idiomas diferentes: Español, Inglés, Francés, Alemán, Turco, Ruso, Islandés, Checo, Danés, Finlandés, Hebreo, Húngaro, Italiano y Latín; además de analizar textos aleatorios, generados a partir de éstos lenguajes naturales.

Los textos aquí analizados se componen principalmente de obras literarias, muchas de ellas clásicos de la literatura universal, tales como *Don Quijote* o *El retrato de Dorian Gray*. En el apéndice (B), se enlistan los textos usados, que son alrededor de 120.

El análisis que hemos hecho de estos textos, permite afirmar que las leyes de Zipf y Herdan-Heaps son leyes universales, no solo de los idiomas, sino de secuencias simbólicas de texto aleatorias, generadas a

partir de lenguajes naturales. Nuestros resultados apoyan la idea que dichas leyes no son realmente leyes exclusivas de la lingüística y que debe hacerse mucho trabajo aún para entender los comportamientos relativos a fenómenos de escala [15].

Sin embargo, el *Clustering Coefficient* de las redes de co-ocurrencia, depende del idioma, y dicha dependencia es diferente en secuencias aleatorias y en lenguajes naturales, lo cual permite proponer una medida para distinguir entre idiomas. Esto podría ser importante para ayudar a resolver el problema de la clasificación de lenguajes, que en la actualidad depende en gran medida del uso de *Machine Learning*. Ya que nuestro método no usa éstos enfoques, puede ser de utilidad para estudiar la relación genealógica entre idiomas.

Por otra parte, mucho del análisis estadístico de textos que se ha hecho en la actualidad, se ha enfocado en la construcción de traductores automáticos. Dichos traductores, como por ejemplo el Google Translator, usan extensivamente técnicas de *Machine Learning* y *Big Data*. Como lo mencionamos anteriormente, estos enfoques dejan a un lado el uso de modelos, para centrarse en la búsqueda de correlaciones, que aunque son muy importantes, no necesariamente son útiles para entender los fenómenos subyacentes en los lenguajes.

En este trabajo se estudia el problema de la traducción de máquina, a través de la identificación de las palabras claves de un mismo texto en diferentes idiomas. Así como lo hicimos en el caso anterior con las leyes de Zipf y Herdan-Heaps, también haremos una revisión de algunos de los métodos para la detección automática de palabras clave, y se propondrá un nuevo método que parece ser exitoso para encontrar palabras claves en textos escritos en diferentes idiomas.

Estos estudios estadísticos pueden usarse también para determinar si textos encriptados satisfacen leyes lingüísticas, y por lo tanto ayudar a su desciframiento. Esta razón es la que motivó el estudio del manuscrito de Voynich, el cual se ha considerado como el santo grial de la criptografía [31]. Los resultados en este apartado apoyan la idea que el manuscrito de Voynich, no es un engaño, en el sentido que parece estar escrito en un lenguaje natural y no es una secuencia extraña de símbolos. Los métodos aquí desarrollados permiten proponer un posible idioma de escritura, y además se poden identificar algunas "palabras clave", que podrían ser de ayuda para su desciframiento.

La distribución del presente escrito es como sigue: en el capítulo 2 analizaremos las propiedades universales y no universales de los textos, y mostraremos como el *Clustering Coefficient* puede usarse para medir una distancia entre idiomas.

En el capítulo 3 se explorarán los métodos para detectar automaticamente palabras claves en textos, y se mostrará las ideas detrás del método que proponemos para encontrar dichas palabras.

En el capítulo 4 se aplicarán los métodos desarrollados al manuscrito de Voynich.

Finalmente se mostrarán conclusiones y perspectivas de este trabajo.

Clustering Coefficient y distancia entre lenguas

2

En este capítulo se analizarán las propiedades estadísticas de 91 textos en 7 idiomas diferentes (Español, Inglés, Francés, Alemán, Turco, Ruso e Islandés), así como de textos con espacios insertados aleatoriamente (Ver Apendice B para consultar una lista de los textos utilizados). Veremos que con estos textos, relativamente pequeños, (alrededor de 11260 palabras diferentes), las leyes estadísticas universales, tales como, la ley de Zipf, la ley de Herdan-Heap, coinciden bastante bien con los resultados obtenidos en otros lugares.

También se construirá la red de co-ocurrencia de palabras de cada texto. Si bien, los resultados sugieren que la distribución de grados es nuevamente universal, se observa que el *Clustering Coefficient* (Coeficiente de agrupamiento), que dependen en gran medida de la estructura local de la red, puede usarse para diferenciar entre idiomas, así como para distinguir idiomas naturales de textos aleatorios.

2.1 Motivación

El estudio estadístico de los lenguajes naturales ha sido un campo que se ha venido desarrollando en las últimas décadas [7, 10, 12, 14, 32, 33]. Probablemente el resultado más sorprendente en éste campo es la ley de Zipf, que establece que si se ordenan las palabras de un texto en función de su frecuencia (de la más usada, hasta las que aparecen una sola vez), una gráfica de frecuencia como función del rango, en una escala log-log, sigue aproximadamente una ley de potencias, para todos los idiomas [12, 16].

Este tipo de resultados universales han despertado durante mucho tiempo el interés de físicos y matemáticos, así como lingüistas [17, 34, 35]. De hecho, se ha dedicado una gran cantidad de esfuerzo para tratar de comprender el origen de la ley de Zipf, en algunos casos se argumenta que surge del hecho de que los textos llevan información [36], en otros casos se dice que es el resultado de la simple casualidad [15, 37].

Otra interesante ley empírica que se obtiene de estudiar los textos es la ley de Heaps-Herdan, que describe cómo el vocabulario, es decir, el conjunto de palabras diferentes, crece con el tamaño de un texto, [38, 39]. Vale la pena señalar que se ha argumentado que esta ley es una consecuencia de la ley de Zipf. [13, 20, 40]

Una herramienta diferente utilizada para caracterizar textos es la red de adyacencia (o co-ocurrencia) [41–44]. Los nodos en esta red representan las palabras en el documento, y dichos nodos se conectan entre sí, si las palabras son adyacentes en el texto. Dichos enlaces pueden ser dirigidos según el orden en que las palabras aparecen, o no dirigidos.

Aquí se estudiarán las propiedades de la red de adyacencia de textos en varios idiomas, utilizando enlaces no dirigidos.

Al representar al texto como una red, será posible describir sus propiedades utilizando las herramientas de la teoría de redes. [29].

Quizá la manera más simple de caracterizar a una red es su distribución de grados, es decir, la fracción de nodos con un número dado de enlaces. Como se mencionó en la introducción, la distribución de grados de una red, puede seguir una ley de potencias, que parece ser universal para todos los idiomas. Como discutiremos más adelante, dicha universalidad puede ser entendida como una consecuencia de la ley de Zipf.

Un aspecto interesante del estudio estadístico de idiomas, es la identificación de idiomas, que está intimamente relacionado con el problema de medir la distancia[8] entre lenguas [45]. A su vez este problema se encuentra en el centro del debate sobre la existencia de un idioma proto-humano, el cuál, de acuerdo con algunas teorías (muy controvertidas y lejos de ser un problema resuelto) sería el origen de todas la lenguas existentes, ver por ejemplo [7] y las referencias allí mencionadas.

8: La "distancia" entre dos lenguas, es un concepto que no se ha logrado definir satisfactoriamente, ya que las diferencias lingüísticas entre idiomas son diversas y difíciles de cuantificar. Sin embargo muchos trabajos lingüísticos, se basan en el hecho que dicha distancia existe. Poder definir y medir esta distancia es uno de los problemas abiertos más interesantes en el *NLP*.

Ahora bien, desde un punto de vista más práctico, la variedad de idiomas que pueden encontrarse en la internet, hace que sea útil desarrollar métodos para la identificación automática de lenguajes. La solución más común a éste problema se ha enmarcado en lo que se conoce como *Natural Language Processing* (NLP), y el uso de *Machine Learning* [46].

El problema con este enfoque es que deja fuera a una gran cantidad de lenguas, que por su número de hablantes no tienen un interés comercial para las empresas, pero que desde el punto de vista lingüístico e histórico son muy importantes.

Por esto es importante plantear métodos estadísticos para distinguir textos e idiomas y más aun, poder proponer métricas que puedan agrupar idiomas en diferentes familias [45, 47, 48].

En este capítulo se usará el *Clustering Coefficient* [29] para mostrar que, aunque la distribución de grados de la red de co-ocurrencia es común a todos los idiomas, la estadística del *Clustering Coefficient*, parece ser distinto para diferentes idiomas.

Es importante notar que el *Clustering Coefficient* generalmente disminuye en función de tamaño de la red [49], por lo tanto, debemos comparar textos con redes de co-ocurrencia de tamaños similares.

En la siguiente sección presenta la distribución de rango vs frecuencia para los textos estudiados; también medimos cómo aumenta el vocabulario con el tamaño del texto, así como las respectivas distribuciones de grado de las redes correspondientes a cada texto.

Estos resultados los comparamos con la hipótesis nula "aleatoria".

Esta hipótesis nula consiste en un conjunto de textos construidos de la siguiente manera:

1. Primero se eliminan todos los espacios del texto original, obteniendo una secuencia simbólica.
2. Tomando la primera letra de la secuencia, y con probabilidad de 1/2, se agrega la siguiente letra, o un espacio y la siguiente letra de la secuencia.
3. Se repite el paso 2, letra por letra, hasta terminar con la secuencia.

La razón por la que la hipótesis nula se construye de esta manera en lugar de aleatorizar las letras de manera independiente tal y como se hace usualmente [15, 50], es que las letras consecutivas no son independientes, por el contrario, están correlacionadas para garantizar la pronunciabilidad de las palabras, o según reglas ortográficas.

El método aquí propuesto para construir los textos aleatorios, conserva la mayoría de las correlaciones entre letras consecutivas y sus frecuencias, pero destruye la gramática del idioma original. Los textos aleatorios del presente capítulo se construyeron a partir del libro *El Retrato de Dorian Gray*. En total se hicieron 12 realizaciones del proceso anteriormente descrito, dos veces para cada idioma (a excepción del Islandés.)

Todos los textos fueron intervenidos para eliminar los signos de puntuación, números, paréntesis y otros símbolos poco comunes, y todas las letras se convirtieron en minúsculas.Además, los textos, se analizan en su alfabeto original (alfabeto cirílico para textos Rusos o los caracteres especiales en Islandés) usando la codificación UTF-8.

También se calcula el *Clustering Coefficient* de la red de co-ocurrencia para cada texto. Dado que el *Clustering Coefficient* depende de manera no trivial del tamaño de las redes, cortamos los textos para que todos tengan esencialmente el mismo tamaño de vocabulario ($\simeq$ 11260).

Las distribuciones de dichos coeficientes son más o menos similares para textos del mismo idioma, sin embargo se observa que las distribuciones son diferentes para distintos idiomas y para la hipótesis nula. Los datos muestran que las mayores diferencias entre los distintos idiomas provienen de aquellos nodos cuyo *Clustering Coefficient* toma valores de 0 o 1.

Con esto en mente, se puede hacer un gráfico de la dispersión de estos nodos para cada idioma. Aunque existe traslape entre algunos idiomas, otras lenguas son claramente diferenciadas entre sí. Esto nos permite ajustar una distribución Gaussiana bi-variada a los datos de cada lengua, y por lo tanto estimar una densidad de probabilidad que indica el idioma en el cual esta escrito un texto.

2.2 Textos y Leyes Universales

En la Figura 2.1 se muestra la ley de Zipf para algunos de los textos, incluidos los textos aleatorios construidos como se describió anteriormente. Está claro que todos los textos reproducen convincentemente la

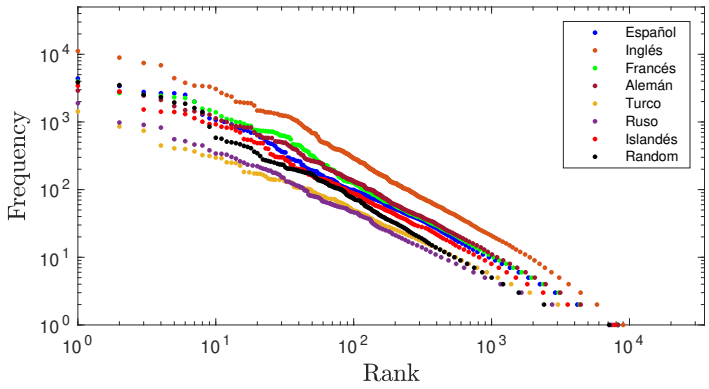


Figura 2.1: Ley de Zipf $f(n) = \frac{c_1}{n^\alpha}$ para textos escritos en Inglés, Ruso, Turco, Francés, Alemán, Español e Islandés (LogBinned). Los puntos negros representan los textos aleatorios construidos como se describe en el texto.

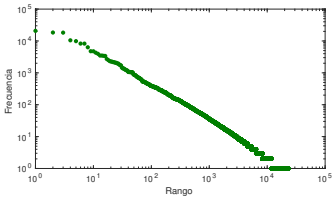


Figura 2.2: Ley de Zipf para *Don Quijote* en Español. Nótese que las palabras con $f = 1$ y 2 representan aproximadamente la mitad del vocabulario.

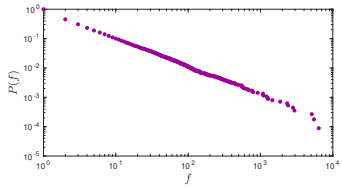


Figura 2.3: Distribución Cumulativa de frecuencias para Don Quijote, $P(f) = \frac{c_2}{f^{\alpha_z}}$

Derivando 2.1 con respecto a n tenemos:

$$\frac{1}{N} = -\frac{c_2}{\left(\frac{c_1}{n^\alpha}\right)^{\alpha_z}} \frac{\alpha c_1}{n^{\alpha+1}}.$$

Usando $n = (c_1/f)^{1/\alpha}$ y $c_2 = f^{\alpha_z} P(f)$, tenemos

$$P(f) = -\frac{c_1^{(1+1/\alpha)}}{\alpha N c_1} \frac{1}{f^{(1+1/\alpha)}}.$$

De aquí se puede ver que si la Ley de Zipf tiene una pendiente de $\alpha \simeq 1$, entonces $P(f) \simeq 1/f^2$

ley de Zipf: $f(n) \sim 1/n^\alpha$ donde $n = 1, 2, ...N_{tot}$ es el rango de palabras, N_{tot} es el tamaño del vocabulario y f es la frecuencia de las palabras.

El hecho que los textos aleatorios sigan también a la ley de Zipf, está en contraste con trabajos anteriores en los cuales se argumenta que hay diferencias entre la ley de Zipf de textos y secuencias aleatorias [51]. Ésto podría deberse al hecho de que nuestra construcción de texto aleatorio conserva las correlaciones entre las letras, mientras que las letras en [51] se colocaron de forma independiente. También observamos que en todos los casos, palabras que aparecen solo una o dos veces (llamadas *hapax legomena* y *dix legomena* respectivamente) representan aproximadamente la mitad del vocabulario en cada texto. Ver figura Figura 2.2.

Si usamos el método del *Maximun Likelihood* para discernir y cuantificar cuando datos empíricos siguen una ley de potencias, [30], encontramos que $P(f)$, para los textos aquí analizados tiene una distribución de ley de potencias. Ver Figura 2.3, Tabla 2.1 y Apéndice (A).

La Figura 2.1 es la típica gráfica de rango vs frecuencia, de la cual se deriva, a través de un ajuste de mínimos cuadrados que $\alpha \simeq 1$. Ahora bien, si n/N_{tot} es la fracción de palabras con frecuencias mayores o iguales a $f(n)$, entonces

$$\frac{n}{N_{tot}} \sim \int_{f(n)}^{\infty} P(f)\, df, \tag{2.1}$$

donde $P(f) \simeq 1/f^{\alpha_z}$ es la distribución de frecuencia del vocabulario, siendo $\alpha_Z \simeq 2$, por lo tanto si $f(n) \sim 1/n^\alpha$, luego $P(f) \sim 1/f^{1+1/\alpha}$, que está en buen acuerdo con lo que observamos. Ver tablas en el Apéndice (A)

La Figura 2.4 muestra el tamaño del vocabulario $V(N)$, en función de la longitud N del texto considerado. Una vez más, todos los textos, incluidos los textos aleatorios, siguen la ley de Heaps-Herdan $V(N) \sim kN^\beta$ razonablemente bien. Nuevamente, los parámetros que describen los diversos textos se dan en el Apéndice (A)

Continuando con las leyes universales que describen textos, en la Figura 2.5 mostramos un ejemplo de la distribución de grados para

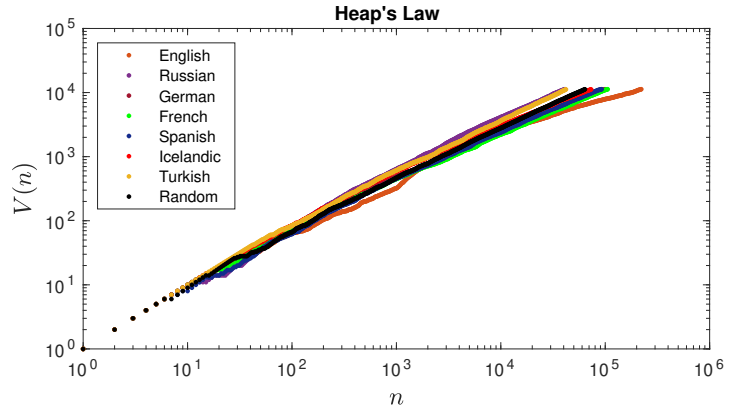


Figura 2.4: La ley de Herdan-Heap para varios idiomas (Inglés, Ruso, Francés, Alemán, Español, Islandés y Turco). Los puntos negros representan textos al azar.

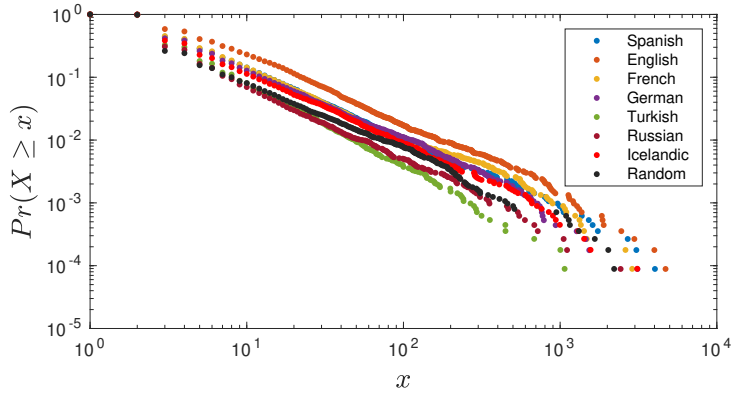


Figura 2.5: Distribución cumulativa de los grados para varios idiomas (Español, Inglés, Francés, Alemán, Turco, Ruso e Islandés). Los puntos negros representan a los textos aleatorios.

la red de adyacencia de los textos estudiados en este trabajo. Está claro que a excepción de los grados impares bajos ($k = 1, 3, 5, 7$), la distribución parece seguir una ley de potencias. Ver Figura 2.6 Los parámetros correspondientes a los textos se dan en el Apéndice (A).

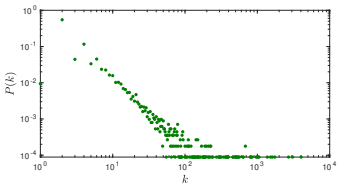


Figura 2.6: Distribución de grados para Don Quijote en Español

Como se mencionó anteriormente, este comportamiento asintótico es una consecuencia de la ley de Zipf. Si se supone que cada vez que aparece una palabra, el grado de entrada k_{in} (alternativamente, el grado de salida k_{out}) del nodo correspondiente aumenta aproximadamente en uno, entonces puede esperarse que el grado de entrada crezca proporcionalmente a la frecuencia de cada palabra. Además, se puede esperar que el grado total de un nodo sea $k \approx k_{in} + k_{out} \approx 2k_{in}$ (claramente esto no siempre es cierto, aunque no ocurre muy amenudo: por ejemplo, una palabra puede aparecer dos veces, separada solamente por otra, y seguida (precedida) a su vez de palabras diferentes, lo que lleva a un nodo con grado $k = 3$; ver Figura 2.7).

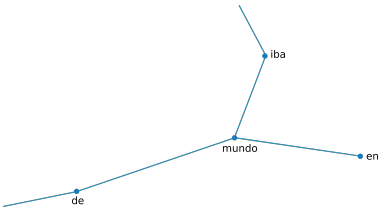


Figura 2.7: La frase *de mundo en mundo iba*, es un ejemplo de un nodo con $k = 3$

Un razonamiento similar al de la ecuación 2.1 para $P(k)$, nos lleva a concluir que si $f(n) \sim 1/n^{\alpha}$, entonces $P(k) \sim 1/k^{1+1/\alpha}$, que nuevamente está de acuerdo con lo que observamos. En la tabla Tabla 2.1 se puede ver esto claramente.

Por ejemplo, la pendiente de la gráfica rango vs frecuencia es ($\alpha \approx 1$) [51]. Usando la ecuación (2.1) se ve que dado el valor de α, los valores de la pendiente para $P(f)$ y $P(k)$ deben ser aproximadamente iguales a 2, lo cual se corrobora al usar el *Maximun Likelihood* y ver que α_Z y α_k se acercan a dicho valor.

Tabla 2.1: Resultados para Español de los textos analizados

Libro	Longitud	Vocabulario	α_k	σ_k	α_Z	σ_Z	β	σ_h
El Conde de Montecristo	92378	11275	2.15	0.03	1.87	0.01	0.781	0.002
Don Quijote	113068	11277	2.11	0.03	1.84	0.01	0.810	0.002
Los tres mosqueteros	106869	11242	2.10	0.03	1.86	0.01	0.746	0.002
Unamuno	104769	11219	2.05	0.03	1.89	0.01	0.765	0.002
Valle-Inclan	76657	11252	2.24	0.03	2.04	0.02	0.780	0.002
Concha Espina	60356	11226	2.33	0.04	2.12	0.03	0.814	0.001
Angelina	71434	11281	2.23	0.03	2.02	0.02	0.810	0.002
Ilíada	91203	11275	2.19	0.03	1.96	0.02	0.799	0.002
Odisea	92381	11290	2.18	0.02	1.96	0.02	0.797	0.002
Pio Baroja	85227	11273	2.21	0.03	2.03	0.03	0.787	0.001
La compañia blanca	76186	11232	2.18	0.03	1.97	0.02	0.786	0.002
Moby Dick	69986	11230	2.15	0.03	2.00	0.01	0.795	0.002
Veinte Mil Leguas...	76443	11214	2.22	0.03	2.01	0.02	0.788	0.001

Es importante notar que los valores del error estandar σ_k, σ_Z y σ_h son tales que no permiten distinguir estadísticamente entre los textos estudiados, incluso entre disitintos idiomas o textos aleatorios.

2.3 Clustering coefficient

Hasta ahora, los resultados confirman que todos los textos exhiben las estadísticas universales observadas en los lenguajes naturales. En realidad, se podría argumentar que las leyes de Zipf, Herdan-Heaps y la distribución de grados, pueden ser "demasiado universales", no pudiendo distinguir claramente los textos escritos en idiomas reales y textos aleatorios.

Además, todas estas leyes parecen ser consecuencia de la ley de Zipf, y esta ley refleja solo la frecuencia de las palabras, no su orden. Por lo tanto, las tres leyes siguen vigentes si las palabras de los textos se barajaran aleatoriamente. Claramente, barajar las palabras destruye cualquier relación que pueda existir entre palabras sucesivas en un texto. Dicha relación es lo que transmite significado. Por lo tanto, esperamos que el *Clustering Coefficient* [29] de la matriz de adyacencia, que depende en gran medida de la estructura local de las redes, distinga entre textos aleatorios y textos reales, e incluso entre textos en diferentes idiomas.

El *Clustering Coefficient* $C_i(k_i)$ del nodo i con grado k_i es definido como la proporción del número de enlaces entre los vecinos del nodo i sobre el número total de enlaces que serían posibles para este nodo $k_i(k_i - 1)/2$. Por lo tanto, claramente, $0 \leq C_i(k_i) \leq 1$. Los Hapax legomena, por ejemplo, corresponden principalmente a nodos con grado $k = 2$, por lo tanto su *Clustering Coefficient* solo puede tomar los valores de 0 o 1 (grado $k = 1$ es posible si el hapax aparece seguido y precedido por la misma palabra, pero estos son casos raros).

En términos generales, los valores del *Clustering Coefficient* varían en función del tamaño de la red [49], por lo tanto, para comparar los *Clustering Coefficients* de las redes correspondientes a diferentes textos, se han recortado los textos para que tengan aproximadamente el mismo tamaño de vocabulario ($\simeq 11260$).

En la Figura 2.8 se muestra un ejemplo del *Clustering Coefficient* como función del grado k. Es importante notar que hay muchos nodos que tienen el mismo grado k cuyos valores para $C(k)$ pueden ser iguales. En

la figura, Los puntos rojos representan el *Clustering Coefficient* promedio para cada k, y la línea negra es el *Log Binning* este promedio [29].

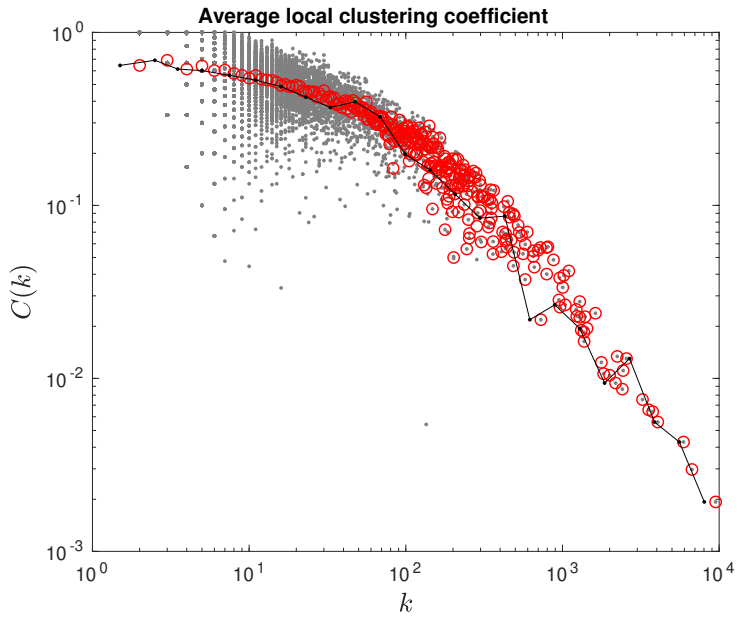


Figura 2.8: Clustering coefficient como función del grado de *Don Quijote* en Español. Los puntos rojos representan el $C(k)$ promedio para cada k, y la línea negra es el *Log Binning* dicho promedio

2.4 Diferenciación del lenguaje

Para cuantificar las diferencias entre idiomas, se define la cantidad $N(C)$ como:

$$N(C) = \frac{\text{Número de nodos cuyo valor de } C(k) \text{ es el mismo}}{\text{Vocabulario}} \tag{2.2}$$

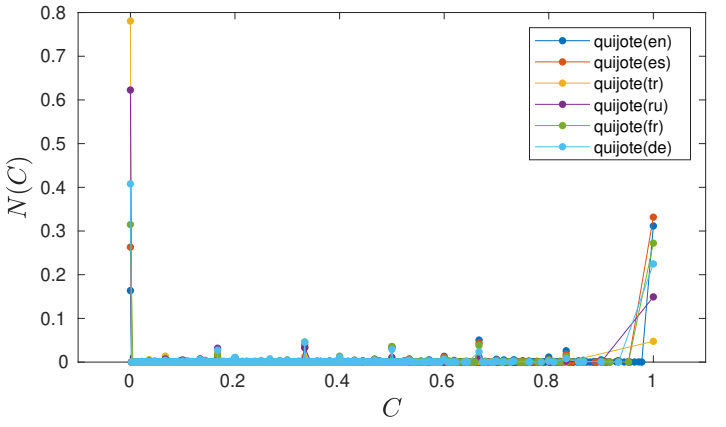


Figura 2.9: Fracción de nodos con el mismo *Clustering coefficient* para Don Quijote en Inglés, Español, Turco, Ruso, Francés y Alemán.

En la Figura 2.9 mostramos $N(C)$ vs C para Don Quijote en seis idiomas diferentes. Del gráfico queda claro que $N(0)$ y $N(1)$ muestran el mayor grado de variación entre los distintos idiomas, por lo tanto, proponemos centrarnos en estos dos números para caracterizar a los lenguajes. En la Figura 2.10 se muestra un diagrama de dispersión

de $N(1)$ vs $N(0)$ para los textos en cada idioma. Usando estimadores de máxima verosimilitud, podemos ajustar los datos a distribuciones gaussianas bi-variadas de la forma

$$f(x, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi\sqrt{\det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(x-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu})\right), \tag{2.3}$$

donde $\boldsymbol{\mu}$ es un vector de promedios y $\boldsymbol{\Sigma}$ es la matriz de co-varianza. Estos valores se pueden estimar de la siguiente forma

$$\begin{aligned} \hat{\mu}_n &= \frac{1}{n}\sum_{j=0}^{n} x_j \\ \hat{\Sigma}_n &= \frac{1}{n}\sum_{j=0}^{n}(x_j - \hat{\mu}_n)(x_j - \hat{\mu}_n)^T. \end{aligned} \tag{2.4}$$

Usando entonces $N(0)$ y $N(1)$ se puede graficar el contorno para cada uno de los ajustes, ver Figura 2.10.

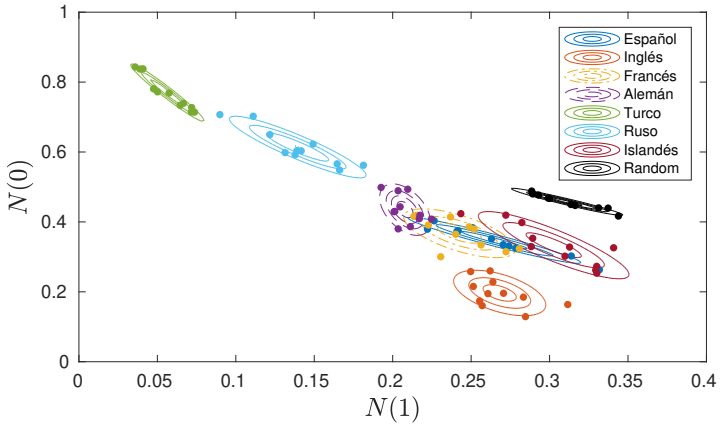


Figura 2.10: Distribución normal bivariada de $N(0)$ y $N(1)$ para los diferentes textos y secuencias aleatorias. Los puntos representan los datos, los contornos son los ajustes.

En la figura Figura 2.10 se ve que hay una clara distinción entre idiomas y textos aleatorios. Además, se puede obervar que los idiomas tienden a agruparse de manera coherente con las relaciones conocidas entre ellos. Por ejemplo, se observa que los contornos correspondientes a Francés y Español muestran una fuerte superposición, lo que es de esperarse ya que son idiomas que comparten un mismo origen [52].

Por otro lado, el Ruso está lejos del Francés o el Español. Esto sugiere que éstos ajustes se pueden utilizar cuantitativamente para la clasificación de idiomas a través de su "distancia". Por ejemplo, el Francés y el Español que son lenguas romances, parecen estar más cerca entre sí que el Ruso y el Turco, que tienen orígenes distintos.

Para probar la validez de estos resultados, se puede calcular $N(0)$ y $N(1)$ para otro conjunto de libros y se averiguará en qué idiomas están escritos.

Para ello, se usarán los valores de μ y Σ estimados con las ecuaciones (2.4) para cada uno de los idiomas estudiados. Entonces, reemplazando $N(0)$ y $N(1)$ en la ecuación (2.3), se podrá asignar la densidad de probabilidad que un texto esté escrito en uno u otro idioma.

Tabla 2.2: Densidad de probabilidad para los textos aleatorios y para los textos escritos en los idiomas estudiados. Se desprecian aquellos valores menores a 1×10^{-8}.

Books	Español	Inglés	Francés	Alemán	Turco	Ruso	Islandés	Random
MobyDick(es)	10.572	0.00014223	125.25	9.1582	0	0.19125	4.0324	0
TwentyThousand...(es)	250.58	0.0033275	182.17	34.068	0	0.019141	0.45346	0
TwentyYearsLater(en)	0	230.94	0.013046	0	0	0	1.179e-07	0
BramStoker(en)	0	65.208	0.036547	0	0	0	0.00013916	0
Voltaire(fr)	266.07	0.003546	196.17	11.899	0	0.022266	0.91734	0
Miserables(fr)	23.99	0.00077475	127.1	0.0030604	0	0.02086	21.05	0
MobyDick(de)	0.026313	7.5812e-07	31.707	325.5	0	2.6555	0.8309	0
Dostoevsky(de)	0.0023808	0.00076044	4.8089	6.3034	0	2.4945e-07	1.9212e-06	0
MobyDick(tr)	0	0	0	0	977.45	2.4039	0	0
JulesVerne(tr)	0	0	0	0	25.009	0.77189	0	0
AroundWorld...(ru)	0	0	0	0	0.0098057	53.27	0	0
MysteriousIsland(ru)	0	0	0	0	5.7008	13.406	0	0
Smásögur I(is)	0	0	0.022572	0	0	0	15.199	0
Smásögur II(is)	0	0	0.58821	0	0	4.4908e-05	23.281	0
RandomTextA	0	0	0	0	0	0	4.7682e-07	5.8013
RandomTextB	0	0	0	0	0	0	0	0.045203

En la Tabla 2.2, se puede ver, por ejemplo, que es más probable que el texto *Smásögur I* (Cuentos en Islandés) esté escrito en Islandés que en cualquiera de los otros idiomas analizados, o que sea un texto aleatorio.

No es sorprendente que sea difícil saber si Voltaire en Francés está realmente escrito en Francés o en Español, así como tampoco es fácil saber si *Moby Dick* en Español está escrito en Español o Francés realmente, y en ambos casos la predicción de máxima probabilidad falla. Sin embargo, está claro que estos libros no están escritos en ninguno de los otros idiomas aquí estudiados, ni corresponden a un texto aleatorio. Por otro lado, *Veinte Mil leguas de viaje submarino* en español y *Les Miserables* en Francés, están correctamente identificados, así como todos los demás textos analizados, incluidos los textos aleatorios.

A diferencia de [45], en donde se proponen metodologías para asignar distancias entre lenguas, basadas en el uso de corpus, nuestro trabajo utiliza una cantidad relativamente pequeña de textos. Además, como se puede ver en las tablas presentadas en el Apéndice (A), la longitud de los textos usados no es necesariamente la longitud del texto completo. Los textos se cortaron con la longitud adecuada para que todos tuvieran aproximadamente el mismo vocabulario ($\simeq$ 11260). Por lo tanto, las longitudes reales oscilaron entre 368076 palabras para los libros de *Jane Austen* en Inglés, hasta 26347 palabras para el texto que llamamos *Turco I*. Esto es importante no solo por razones computacionales, sino que también puede ser importante para los estudios que tratan de esclarecer la relación entre idiomas para los cuales no existen grandes corpus, algo muy común por ejemplo en estudios lingüísticos de lenguas indígenas. El método propuesto en este trabajo puede ser útil en tales casos, ya que el único requisito que se necesita es el de recortar los textos para satisfacer un tamaño de vocabulario apropiado.

Detección automática de palabras clave

3

En este capítulo se describirá el método que desarrollamos para encontrar palabras clave en un texto. Este método se basa en la idea que una palabra clave en un documento es aquella en donde la distribución de las posiciones a lo largo del texto, difiere de una distribución aleatoria.

Compararemos a este método, que llamamos *Term Frequency - Random Sampling Distribution* (TF-RSD) con otros métodos que aparecen en la literatura, y se mostrará como el TF-RSD, puede usarse para encontrar palabras claves en textos escritos en diferentes idiomas.

3.1 Word Burstiness: ¿La clave para encontrar palabras clave?

Una forma que resulta conveniente para caracterizar estadísticamente a las palabras en un texto, es medir la distancia (dada en número de palabras) entre dos repeticiones de un mismo término. Esta distancia varía en función del tipo de palabras. Por ejemplo, las palabras como *de, la, y, que*, que son las más frecuentes, aparecerán distribuidas a lo largo de todo el texto, mientras que otras palabras, pueden aparacer únicamente en ciertas partes del libro.

Esta heterogeneidad en la distribución de ocurrencias de palabras en un texto se conoce como *Word Burstiness* [14], y se ha propuesto como una forma de detectar automáticamente palabras claves en un documento [53].

Sea x_i la distancia entre dos ocurrencias de una palabra, entonces se define a la distancia normalizada como:

$$u_i = \frac{x_i}{\langle x_i \rangle}, \qquad \text{donde} \qquad \langle x_i \rangle = \frac{N}{f_i}, \tag{3.1}$$

siendo f_i la frecuencia de la palabra, y N el número de palabras en el documento.

En la Figura 3.1 se puede ver la distribución cumulativa de u_i para las palabras *Quijote, Sancho, la, y* en el *Don Quijote* en Español. De esta figura se observa que la distribución de u_i para las palabras funcionales[10] *la, y* se aproxima a una distribución de Poisson. De igual manera, aquellas palabras, que suponemos importantes como *Quijote* o *Sancho*, se alejan de dicha distribución

Esta diferencia entre las palabras funcionales, y las palabras importantes, se puede entender de la siguiente manera. Si la probabilidad de encontrar a una palabra en un lugar dado en el texto fuese la misma

10: Se les dice palabras funcionales (llamdas también *stop words*) a aquellas que no tienen ningún contenido semántico y su rol es puramente gramatical [54]

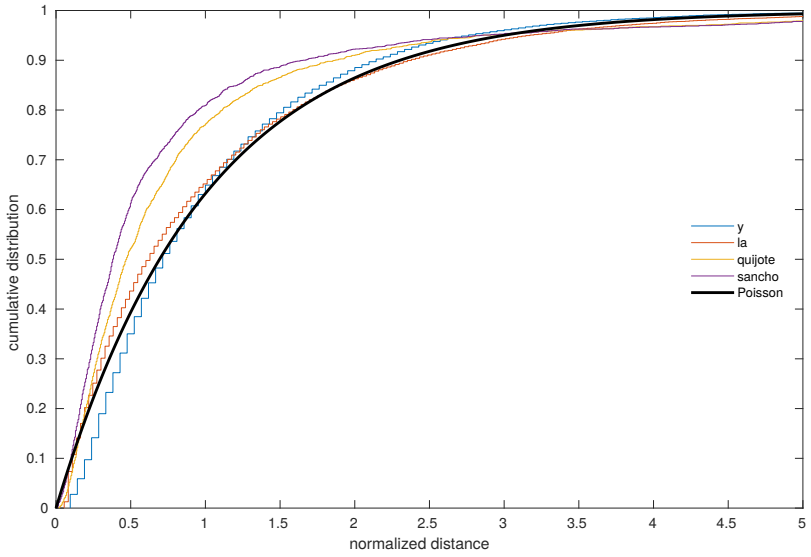


Figura 3.1: Distribución cumulativa de u_i para algunas palabras de Don Quijote en español. La línea negra es la distribución de Poisson.

en todo el documento, es decir, si no hubiese correlaciones entre las repeticiones sucesivas de la palabra en cuestión, la probabilidad de que la distancia entre términos sea u_i estaría dada por una distribución (cumulativa) de Poisson de la forma

$$P(u_i) = 1 - \exp(-u_i). \tag{3.2}$$

Por otra parte, si hay correlaciones en las repeticiones de una palabra, la distribución de u_i se desviará de una distribución de Poisson. Por ejemplo, la palabra *Quijote*, en varios idiomas tiene una distribución tipo LogLogistic

$$P(u_i) = \frac{1}{1 + \left(\frac{u_i}{\alpha}\right)^{-\beta}} \tag{3.3}$$

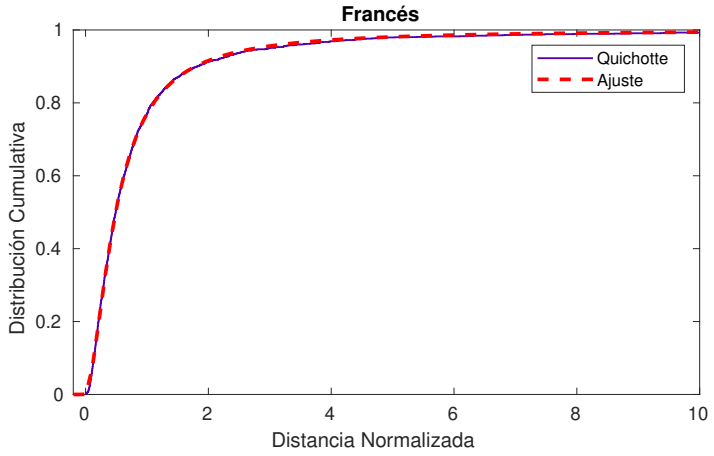


Figura 3.2: Distribución Cumulativa de u_i para la palabra *Quijote* en el *Don Quijote* en Francés. El ajuste (línea roja) corresponde a una distribución tipo LogLogistic.

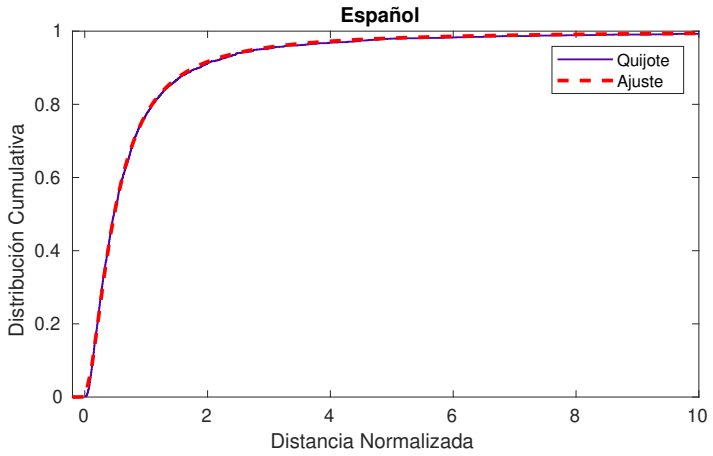


Figura 3.3: Distribución Cumulativa de u_i para la palabra *Quijote* en el *Don Quijote* en Español. El ajuste (línea roja) corresponde a una distribución tipo LogLogistic.

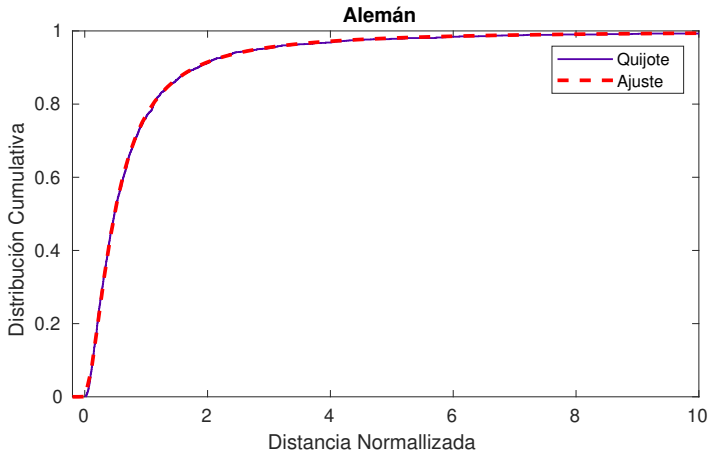


Figura 3.4: Distribución Cumulativa de u_i para la palabra *Quijote* en el *Don Quijote* en Alemán. El ajuste (línea roja) corresponde a una distribución tipo LogLogistic.

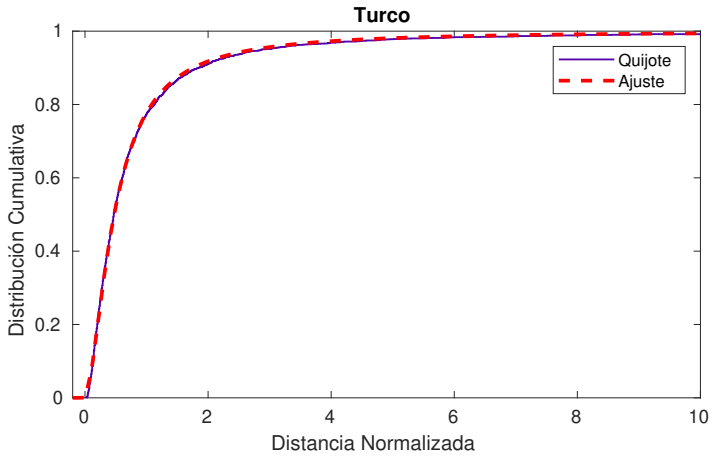


Figura 3.5: Distribución Cumulativa de u_i para la palabra *Quijote* en el *Don Quijote* en Turco. El ajuste (línea roja) corresponde a una distribución tipo LogLogistic.

Por otra parte, un análisis detallado de la distribución de u_i para la palabra y muestra que dicho término tiene en lugar de una distribución de Poisson, una de tipo Birbaum-Saunders de la forma

$$P(u_i) = \varphi\left(\frac{\sqrt{u_i} - \sqrt{\frac{1}{u_i}}}{\gamma}\right), \tag{3.4}$$

donde γ es el *parámetro de forma* y φ es la función cumulativa de la distribución normal.

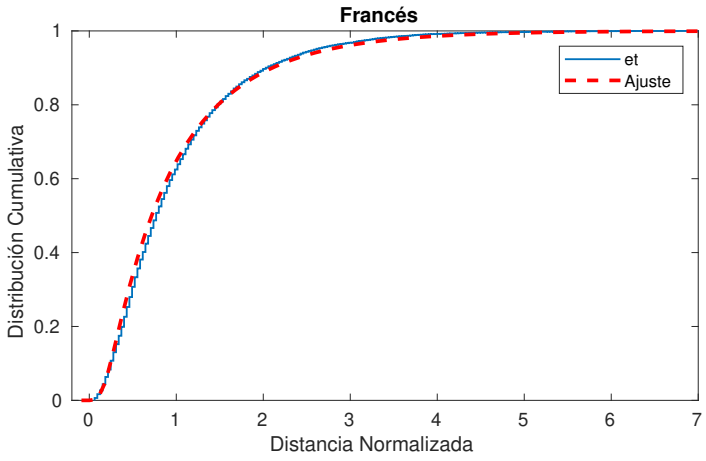


Figura 3.6: Distribución Cumulativa de u_i para la palabra *y* en el *Don Quijote* en Francés. El ajuste (línea roja) corresponde a una distribución tipo Birbaum-Saunders.

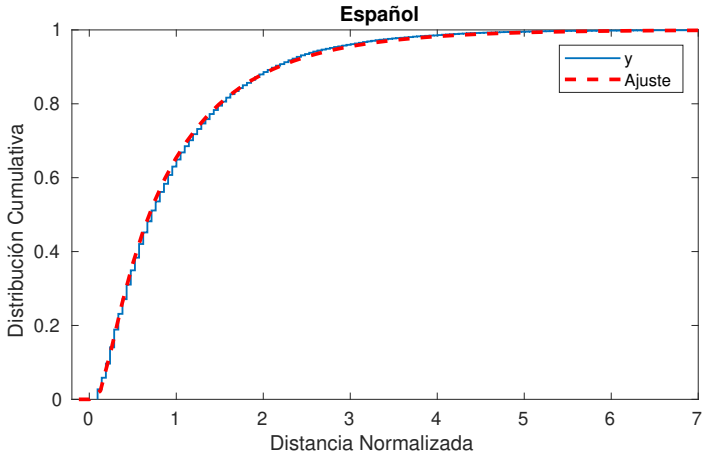


Figura 3.7: Distribución Cumulativa de u_i para la palabra *y* en el *Don Quijote* en Español. El ajuste (línea roja) corresponde a una distribución tipo Birbaum-Saunders.

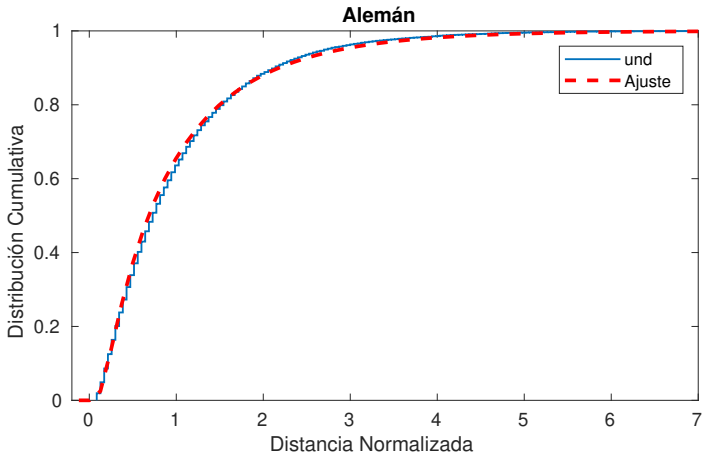


Figura 3.8: Distribución Cumulativa de u_i para la palabra *y* en el *Don Quijote* en Alemán. El ajuste (línea roja) corresponde a una distribución tipo Birbaum-Saunders.

Resultados similares a la Figura 3.3 se encuentran para la palabra *Sancho*. De la misma manera en la palabra funcional *la*, las distancias tienen distribuciones cumulativas del tipo Birbaum-Saunders. Otras palabras que a priori podrían ser importantes para este libro, como *Rocinante* o *Dulcinea*, tienen distribuciones cumulativas de u_i que pueden ser del tipo Exponenciales, Gamma, Burr, etc.

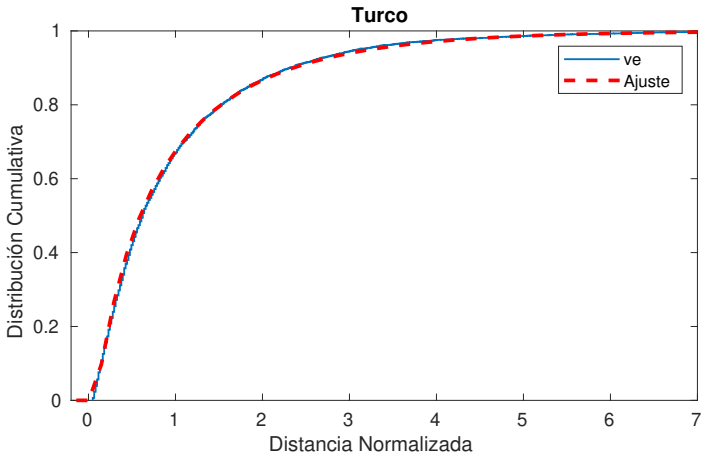


Figura 3.9: Distribución Cumulativa de u_i para la palabra y en el *Don Quijote* en Turco. El ajuste (línea roja) corresponde a una distribución tipo Birbaum-Saunders.

Ahora bien, lo que sí se puede concluir de las figuras anteriores, es que existe una clara diferencia entre la distribución de distancias de palabras funcionales y palabras consideradas importantes.

Esta observación llevó a *Ortuño et al* [53], a proponer la desviación estándar de u_i como una medida de la importancia de una palabra en un texto.

$$\sigma_i = \langle (u_i - \langle u_i \rangle)^2 \rangle^{1/2}, \tag{3.5}$$

es decir a mayor valor de σ_i, más importante es la palabra.

De la Figura 3.10 es claro que la palabra más importante de la biblia sería *Jesus*, algo que tiene mucho sentido.

M. Ortuño *et al.*: Keyword detection in natural languages and DNA **761**

Table I – *Words with larger σ in The Bible.*

Word	Frequency	σ	Word	Frequency	σ
jesus	983	24.18	david	1064	8.86
christ	571	18.42	king	2542	8.15
paul	162	11.56	pharisees	87	8.06
disciples	244	10.88	jeremiah	148	8.00
peter	164	10.17	gospel	104	7.91
joab	145	10.03	solomon	305	7.67
faith	247	9.34	mordecai	60	7.45
saul	420	9.17	esther	57	7.43
absalom	108	9.12	joshua	217	7.42
john	137	9.03	elisha	58	7.39

Figura 3.10: Imagen tomada de [53], donde se muestran las palabras claves de la Biblia en Inglés.

Al aplicar este método al *Don Quijote* en Español, se obtiene un conjunto de palabras clave, ver Tabla 3.1, que son difíciles de reconocer como términos importantes para este libro.

Tabla 3.1: *Word Burstiness* para *Don Quijote* en Español

	Palabra	σ	Frecuencia		Palabra	σ	Frecuencia
1	camila	8.7433	148	12	zoraida	5.7286	78
2	lotario	8.5605	142	13	ricote	5.0703	35
3	duquesa	7.565	192	14	renegado	4.9673	60
4	vuesa	7.3367	203	15	basilio	4.842	53
5	anselmo	7.1175	138	16	leonela	4.7589	44
6	antonio	6.753	65	17	grisóstomo	4.7452	29
7	fernando	6.5889	135	18	mono	4.7424	44
8	dorotea	6.3675	112	19	gobernador	4.6908	174
9	altisidora	6.302	64	20	rodríguez	4.6718	45
10	cardenio	5.8497	101	21	marcela	4.6587	28
11	luscinda	5.836	99	22	tosilos	4.4788	28
38	sancho	3.9325	2241	59	quijote	3.2745	2241
90	dulcinea	2.8065	286	93	rocinante	2.7656	206

En este caso, la palabra más importante de *Don Quijote* parece ser *Camila*[11] , la cual es difícil asociar al *Don Quijote*. Para entender la razón por la cual este método le da una importancia preponderante a la palabra *Camila*, en lugar de por ejemplo *Quijote*, o *Rocinante*; se puede graficar la posición de dicha palabra a lo largo del documento.

11: Camila, junto a Lotario y Anselmo, son personajes de una de las novelas cortas, llamada *el curioso impertinente*, que aparece intercalada en el *Don Quijote*.

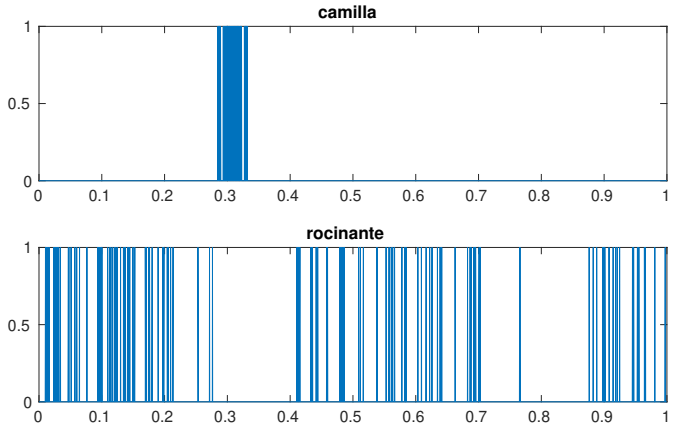


Figura 3.11: Posición de la palabra *Camila* y *Rocinante* en el *Don Quijote*

En la Figura 3.11, que para abreviar la llamaremos *Código de Barras*, se ha graficado la posición de cada palabra en el texto. Dado que el número de palabras de un mismo libro, va a depender del idioma en el cuál esta escrito, es conveniente representar la posición de la palabra en el libro, por un número en el intervalo $[0, 1]$, esto con el fin de poder comparar libros en distintos lenguajes.

Al comparar el *código de barras* de *Rocinante* con *Camila*, es claro que la palabra *Rocinante*, a diferencia de *Camila*, se encuentra distribuida a lo largo del texto.

Ahora bien, dado que σ_i es una medida de la dispersión de la distribución de distancias u_i, de acuerdo con este método, una palabra será importante cuanto más agrupada se encuentre. Dicha definición de "clave" falla al ignorar el hecho que pueden existir palabras importantes que se encuentran distribuidas a lo largo de todo el texto, como es el caso de *Rocinante* (cualquier lector pondría al caballo de *Don Quijote* como un personaje más importante que *Camila*).

En el caso de la biblia el método parece funcionar ya que *Jesús* es uno de los personajes principales del nuevo testamento, y esta palabra está confinada (al igual que *Camila*) a una zona muy específica del libro. Fuera de los Evangelios, la palabra *Jesús* prácticamente no aparece

en la biblia, al igual que *Camila* sólo aparece en algunos capítulos del Quijote.

3.2 Term Frequency - Inverse Document Frequency (TF-IDF)

El TF-IDF es quizá el método más usado para encontrar las palabras claves en un corpus. Por definición, los corpus estan compuestos de una colección, a menudo bastante grande, del orden de millones de textos. El TF-IDF está definido por [46]

$$\text{TF-IDF} = -f_t \times \log \frac{n_t}{N_d}, \tag{3.6}$$

donde f_t es la frecuencia de la palabra en cuestión, n_t es el número de documentos en donde la palabra aparece y N_d es el número total de documentos en el corpus. Al término $\log n_t/N_d$ se le conoce como *Inverse Document Frequency*.

Es importante resaltar que existen diferentes esquemas para "pesar" a cada uno de los términos de la ecuación (3.6). Por ejemplo, f_t puede ser binario, la palabra aparece o no aparece en el corpus, puede ser log-normalizado, se reemplaza f_t por $\log(1 + f_t)$. A su vez el término IDF también puede definirse de distintas maneras, por ejemplo reemplazando $\log \frac{n_t}{N_d}$ por $\log \frac{N_d - n_t}{n_t}$, entre otros, cuyo proposito es realizar cálculos específicos para diferentes tipos de corpus [55].

Para entender como funciona el TF-IDF, vamos a suponer que tenemos un corpus compuesto de 3 documentos.

Documento 1 Yo amo a mi gato
Documento 2 Yo odio a mi gato y al mezcal
Documento 3 El mezcal es mi afición y mi pasión

Con estos documentos podemos construir la Tabla 3.2, donde las columnas contienen al vocabulario del corpus, y las filas representan a cada uno de los documentos. Los números en la tabla son la frecuencia de cada palabra. Esta manera de organizar a las palabras de un corpus se le conoce como bolsa de palabras (*bag of words*) y es la base de partida para todo lo relativo a minería de textos y recuperación de información. [46]

Tabla 3.2: Ejemplo de la construcción de una bolsa de palabras

	yo	amo	a	mi	gato	odio	y	al	mezcal	el	es	afición	pasión
Doc 1	1	1	1	1	1	0	0	0	0	0	0	0	0
Doc 2	1	0	1	1	1	1	1	1	0	1	0	0	0
Doc 3	0	0	0	2	0	0	1	0	1	1	1	1	1

Usando la frecuencia inversa del documento, ecuación (3.6) podemos calcular el TF-IDF para cada una de las palabra del vocabulario del corpus

Tabla 3.3: Ejemplo de la construcción de TF-IDF

	yo	amo	a	mi	gato	odio	y	al	mezcal	el	es	afición	pasión
Doc 1	0.40	1.09	0.40	0	0.40	0	0	0	0	0	0	0	0
Doc 2	0.40	0	0.40	0	0.40	1.09	0.40	1.09	0.40	0	0	0	0
Doc 3	0	0	0	0	0	0	0.40	0	0.40	1.09	1.09	1.09	1.09

De la Tabla 3.3 vemos que la importancia de las palabras se puede medir dependiendo del valor de TF-IDF que cada una tenga. Es claro por ejemplo que la palabra *mi* es la menos importante del corpus, ya que aparece en todos los documentos, luego $\log 1 = 0$ y sin importar su frecuencia su valor de TF-IDF sera nulo. Esto ocurre típicamente con las palabras funcionales. Por lo tanto, las palabras claves del corpus serán aquellas que aparezcan un determinado número de veces (el factor f_t), pero que no se presenten en la mayoría de documentos. En este ejemplo palabras como *odio, amo, afición, pasión* vienen a ser las palabras claves del conjunto de textos.

Del ejemplo anterior se puede ver que el TF-IDF se convierte en una competencia de qué término domina más, si f_t o $\log n_t/N_d$. Para ver esto de manera más clara, podemos tomar como corpus a cada uno de los versículos del nuevo testamento de la biblia en Inglés. [12]

12: Este es un ejemplo tomado de *Text Mining with MATLAB* [46]

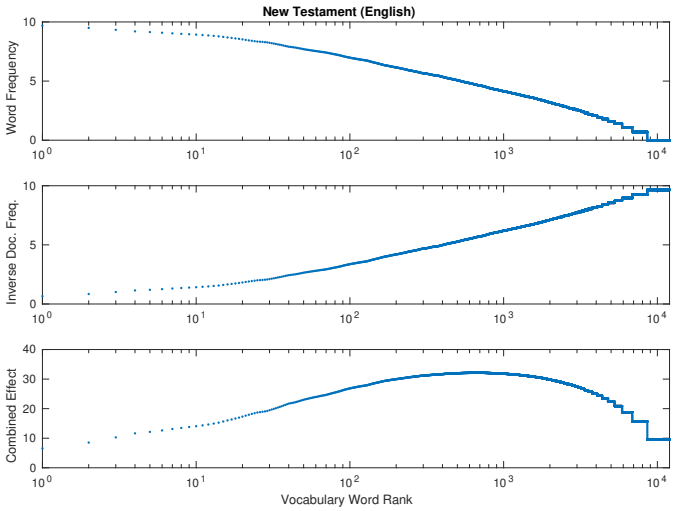


Figura 3.12: a) Rango del vocabulario como función de la frecuencia. b) Rango del vocabulario como función del la frecuencia inversa del documento. c) Efecto combinado de TF-IDF

En la Figura 3.12 se observa cómo es el efecto de cada uno de los términos de la ecuación (3.6). En el primer recuadro se ve la ley de Zipf. En el segundo recuadro, se observa cómo se comporta el IDF, como función del vocabulario ordenado de mayor a menor frecuencia. En el recuadro final, vemos cómo aquellas palabras que tienen una frecuencia alta (típicamente palabras funcionales), tienen un TF-IDF bajo, ya que estas tienden a aparecer en todos los documentos; por el contrario, aquellas palabras que aparecen solamente en un documento, tienen también un TF-IDF bajo, ya que dichas palabras son importantes solo para el documento en cuestión, pero no importantes desde el punto de vista de todo el corpus. Las palabras que son relevantes para este caso, son aquellas que se encuentran a medio camino entre uno y otro, es decir, tiene frecuencias relativamente altas, pero están

limitadas a aparecer en solo un cierto número de versículos.

Como se ha visto, el TF-IDF es un método apropiado para encontrar palabras claves, pero su eficacia está ligada intimamente con la existencia (y el tamaño) del corpus. Como se mencionó en el capítulo anterior, para ciertos campos, como el estudio de lenguas indígenas, la existencia del corpus no está garantizada, y por lo tanto, no es posible usar este método.

Como se vio también en la primera sección de este capítulo, el *word burstiness* falla al no considerar que pueden existir palabras relevantes para el texto que se distribuyen a lo largo del texto.

Sin embargo las ideas detrás de estos dos métodos son útiles para proponer uno que soluciona los problemas antes mencionados.

3.3 Term Frequency - Random Sampling Distribution (TF-RSD)

Hasta ahora dos cosas son claras; primero, la distribución de distancias entre ocurrecias de una palabra, parece depender de si esta es funcional o no. Segundo, las palabras funcionales generalmente aparecen en todos los documentos de un corpus, por eso en el TF-IDF tienen valores cercanos a cero.

Como se mencionó previamente, nuestro interés está en estudiar textos, no colecciones de ellos. Es así que el primer método presentado es apropiado para encontrar palabras claves en un libro; pero falla en su definición de qué es clave en un texto.

El segundo método es efectivo para hallar palabras clave, pero es apropiado con corpus de textos, y si lo quisiésemos aplicar al Quijote por ejemplo, no sería claro como dividir al texto; ya que fraccionarlo en capítulos, páginas, párrafos o un cierto número de partes, es en general arbitrario.

Vamos a suponer sin embargo que, dado un texto, lo dividimos en k partes o cajas iguales. Luego, la probabilidad conjunta de que una palabra dada, aparezca l veces en cada caja; si las posiciones del término están uniformemente distribuidas, viene dada por

$$P(l_1, \dots l_k) = \frac{\prod_{i=1}^{k} \binom{\nu}{l_i}}{\binom{\nu k}{f_t}}, \tag{3.7}$$

donde ν es el número de palabras en cada una de las partes en las que dividimos el texto, l_i es la frecuencia de la palabra en cada una de las cajas,[13] y f_t es la frecuencia total de la palabra.

13: Es importante notar que l_i está sujeta a la condición que $\sum_{i=1}^{k} l_i = f_t$;

De la ecuación (3.7) es posible determinar el número de cajas en las cuales la palabra dada aparece (Ver apéndice C para más detalles)

$$m_t(k) = k\left[1 - \left(1 - \frac{1}{k}\right)^{f_t}\right]. \tag{3.8}$$

Ahora bien, inspirados en la ecuación (3.6) proponemos entonces que la importancia de una palabra en un texto puede describirse a través de

$$\text{TF-RSD} = -f_t \times \log \frac{n_t}{m_t(k)}, \tag{3.9}$$

donde, de forma similar a como se definió en el TF-IDF, n_t es el número de cajas en las cuales la palabra aparece.

Para un valor de k dado, el término $\log \frac{n_t}{m_t(k)}$ representa la competencia entre dos distribuciones estadísticas, n_t siendo la distribución real de una palabra; y $m_t(k)$ la distribución uniforme aleatoria de la palabra, en cada una de las k cajas. Es por esta razón que a este término lo llamamos *Random Sampling Distribution*.

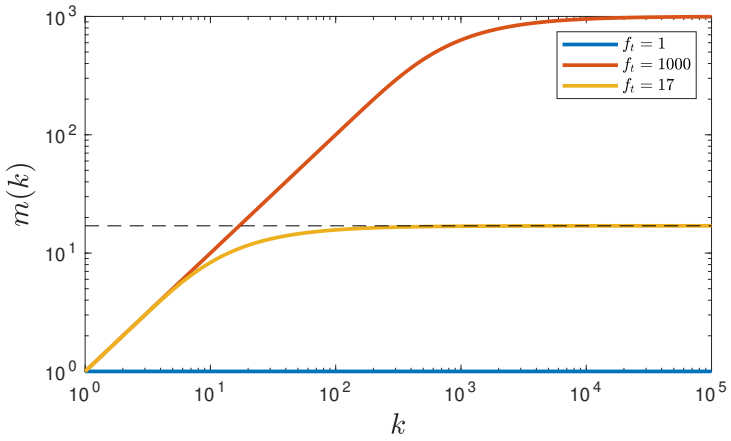


Figura 3.13: $m(k)$ como función de k para tres valores distintos de la frecuencia f.

En la Figura 3.13 se observa como varía $m(k)$, como función de k, para tres distintos valores de la frecuencia. Si $f_t = 1$, se ve que la palabra es un *Hápax Legómenon*. En este caso, $m(k) = n_t = 1$ y de manera trivial se cumple que TF-RSD= 0. Es decir, la ecuación (3.9) asigna un valor nulo a todas aquellas palabras que solo aparecen una vez en todo el documento.

Cuando $f_t > 1$, $m(k)$ toma distintos valores como función del número de cajas, pudiéndose distinguir dos casos. El primero, cuando $k \to N$ (o de manera equivalente cuando $f_t \gg 1$), se puede obervar que $m(k) \approx f_t$. En este límite, en el que hay tantas cajas como número de palabras, es fácil ver que $n_t \sim f_t$ y por lo tanto, $\log \frac{n_t}{m_t(k)} \to 0$.

El segundo caso sería cuando cuando los valores de k son intermedios, digamos entre 2 y 1000 en la Figura 3.13. En este caso se puede ver que $m(k) \approx k$ y el valor de RSD dependerá de si la distribución de la palabra en las k cajas es o no aleatoria.

Si la distribución del término en cuestión es aleatoria, o dicho de otro modo, si no existen correlaciones en las repeticiones de la palabra, entonces $n_t \approx k$ y nuevamente TF-RSD$\rightarrow 0$.

En caso contrario, cuando $n_t \neq k$, que es lo mismo que decir que existen correlaciones en la distribución de la palabra, el término $\log \frac{n_t}{m(k)}$ tendrá valores mucho mayores que 0, dependiendo de que tanto difiere la distribución del término de la aleatoriedad, y de cuantas veces aparece la palabra en el documento.

Entonces, de acuerdo con este análisis, se puede afirmar que una palabra será importante si, aparece un determinado número de veces, y si adicionalmente su distribución de apariciones es diferente de la aleatoriedad.

De la discusión anterior, se puede concluir que para encontrar las palabras clave en un texto, es importante dividir el documento en k partes, de tal manera que exista un muestreo apropiado de las palabras. Para ello vamos a encontrar el valor de k que maximize a

$$\text{TF-RSD}_{total} = -\sum_i f_i \log \frac{n_t}{m_i(k)}, \tag{3.10}$$

donde la suma se realiza sobre todo el vocabulario del texto.

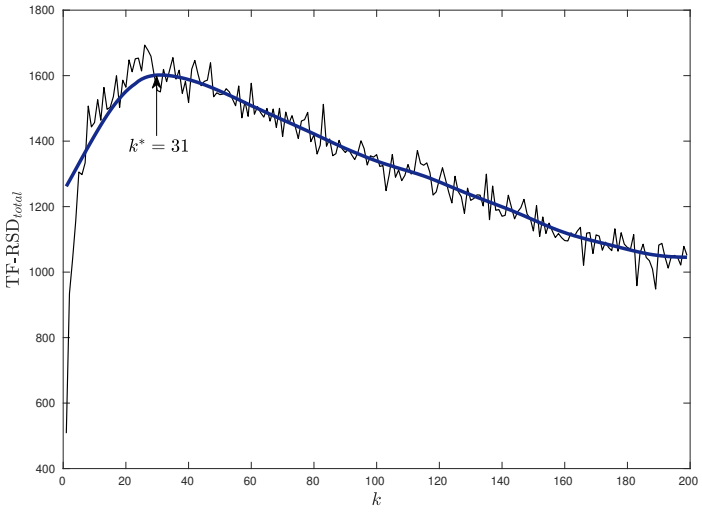


Figura 3.14: TF-RSD$_{total}$ como función de k para *EL Principito* en Español. La línea negra se obtiene de la ecuación (3.10), la línea azul se obtiene de usar un filtro para suavizar los datos.

Si se divide sistemáticamente el texto en $k = 2, 3, 4, 5, 6, 7...$ partes iguales, se puede hacer una gráfica TF-RSD$_{total}$ como función de k. Por ejemplo, para el libro *El Principito*, (Figura 3.14), la división óptima sería para $k^* = 31$. Es importante notar que se usó un filtro de promedio móvil que suaviza los datos, reemplazando cada punto del conjunto de datos con el promedio gaussiano de los datos vecinos definidos dentro de una ventana [56].

Con este resultado se puede usar la ecuación (3.9) para determinar las palabra importantes en *El Principito*. Ver Tabla 3.4

Tabla 3.4: TF-RSD para *El Principito* en Español

	Palabra	TF-RSD	Frecuencia		Palabra	TF-RSD	Frecuencia
1	zorro	36.768	32	12	tres	15.691	5
2	tierra	31.835	113	13	serio	14.246	17
3	estrellas	27.005	32	14	planeta	13.218	22
4	veces	26.460	56	15	había	13.047	26
5	cordero	25.053	15	16	tiempo	12.748	17
6	flor	22.091	41	17	días	12.655	10
7	dijo	21.542	33	18	triste	10.400	4
8	volvió	21.386	9	19	mil	9.2708	15
9	vienes	18.459	17	20	tarde	9.1301	19
10	tenía	16.062	15	21	hombres	8.2007	2
11	dibujo	15.961	46	22	principito	7.3667	15

En esta tabla podemos ver las 22 palabras que de acuerdo con nuestro método, serían las más importantes para *El Principito*.

No es difícil ver que a pesar de ser un texto relativamente corto, con un vocabulario de 2667 palabras, encontramos términos que ciertamente se pueden relacionar con esta obra. Si se aplica el método descrito a *El Principito* en Inglés, se obtiene que, en este caso $k^* = 46$, y las palabras claves serían

Tabla 3.5: TF-RSD para *El Principito* en Inglés

	Palabra	TF-RSD	Frecuencia		Palabra	TF-RSD	Frecuencia
1	nothing	34.116	32	12	until	13.167	7
2	sheep	32.295	40	13	prince	12.748	179
3	never	27.410	41	14	day	11.935	32
4	flower	25.316	54	15	beautiful	10.751	13
5	fox	25.246	35	16	thousand	10.751	13
6	drawing	23.782	24	17	world	10.751	13
7	years	20.085	12	18	told	10.485	8
8	said	19.897	196	19	asked	10.413	27
9	planet	17.157	69	20	king	9.9954	32
10	three	14.814	17	21	tell	9.9453	11
11	say	13.840	23	22	true	9.9170	14

Comparando la Tabla 3.4 con la Tabla 3.5, es posible identificar algunas palabras claves en Español y sus respectivas traducciones en Inglés, tales como *Zorro-Fox*, *Cordero-Sheep*, *Flor-Flower*.

Ahora bien, si se aplica el TF-RSD al *Don Quijote* en Español, se obtiene que el valor óptimo en el cual se debe dividir el texto es $k^* = 247$, ver Figura 3.15.

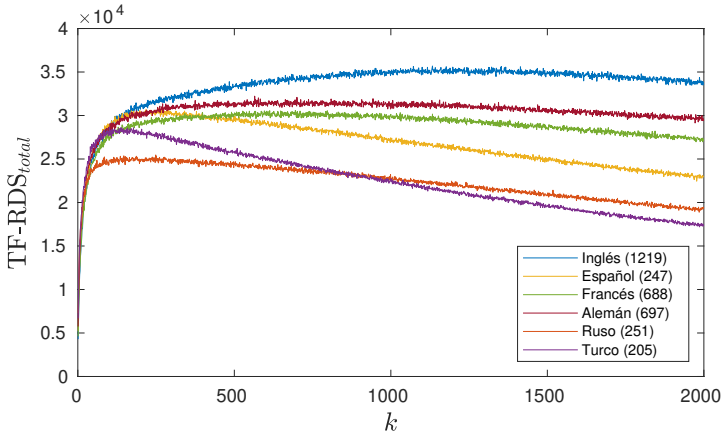


Figura 3.15: TF-RSD como función de k para Don Quijote en varios idiomas. Nóstese que el valor óptimo de k depende del idioma.

	Palabra	TF-RSD	Frecuencia		Palabra	TF-RSD	Frecuencia
1	sancho	492.26	2143	12	fernando	182.13	259
2	quijote	333.56	2158	13	vuestra	174.04	135
3	tanta	319.76	148	14	estos	170.56	846
4	puso	313.79	142	15	os	170.36	199
5	cura	285.06	313	16	gobernador	166.73	458
6	anselmo	271.19	138	17	señora	161.83	174
7	te	266.16	717	18	caballero	155.32	515
8	don	245.38	2629	19	panza	151.13	657
9	merced	204.26	895	20	pudo	146.1	325
10	duquesa	198.24	189	21	dulcinea	145.19	99
11	padre	182.13	259	22	desde	141.51	277

Tabla 3.6: TF-RSD para Don Quijote en Español

Aplicando la ecuación (3.9), se puede obtener la Tabla 3.6, donde se enlistan las palabras clave para el *Don Quijote* en Español. Aquí se puede ver que, tal y como se supuso en un principio, palabras como *Sancho*, o *Quijote* tienen valores altos de TF-RDS. También aparecen palabras tales como *caballero* o *cura*, que son relevantes para este libro. A este conjunto de palabras, también pertenecen algunos términos que, de acuerdo a la definición dada anteriormente, son palabras funcionales; tal es el caso de *te*, *estos*, *os* y *desde*. Esto significa que dichas palabras no se encuentran distribuidas aleatoriamente a lo largo del documento.

Esto es mucho más evidente en el caso del *Quijote* en Inglés. Para $k^* = 1219$, se ve que las palabras claves serían

	Palabra	TF-RSD	Frecuencia		Palabra	TF-RSD	Frecuencia
1	her	1303.1	2110	12	don	525.09	2587
2	you	985.11	2291	13	your	460.2	1338
3	i	847.83	6470	14	they	437.16	2662
4	thou	787.28	1214	15	quixote	433.45	2011
5	sancho	713.95	2064	16	was	430.34	3377
6	she	674.21	1211	17	thee	429.54	750
7	him	656.91	3418	18	is	426.85	3584
8	he	647.48	5839	19	we	352.4	939
9	his	620.56	4347	20	had	326.5	2141
10	my	611.96	2785	21	said	314.97	2613
11	me	555.45	2485	22	will	303.57	1629

Tabla 3.7: TF-RSD para Don Quijote en Inglés

En la Tabla 3.7 se observa que salvo las palabras *quijote*, *sancho* y *don*, todas las palabras son funcionales. Recordemos que para obtener la ecuación (3.9), se está suponiendo que las palabras funcionales se encuentran distribuidas aleatoriamente, cosa que como se vio no es cierta. (Ver Figura 3.6). Por ejemplo, la distribución de u_i para la palabra *I* (*Yo* en Inglés) se ve en la Figura 3.16

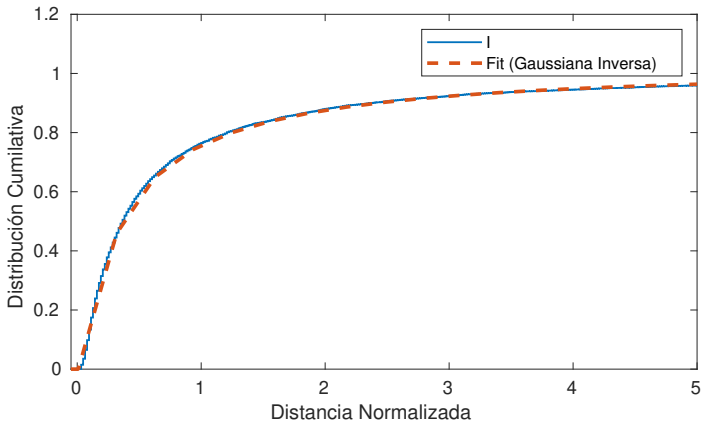


Figura 3.16: Distribución de u_i para la palabra *I* en el *Don Quijote* en Inglés. La línea azul es la distribución cumulativa de u_i para la palabra *I*. La línea roja es el ajuste. En este caso, la distribución que ajusta los datos es una Gaussiana Inversa.

En este caso, la distribución que ajusta a los datos es una Gaussiana Inversa, de la forma

$$P(u_i) = \frac{1}{2}\left\{1 + erf\left[\sqrt{\frac{\lambda}{2u_i}}\left(\frac{u_i}{\mu} - 1\right)\right]\right\} + \frac{\exp(2\lambda)/\mu}{2}\left\{1 + erf\left[\sqrt{\frac{\lambda}{2u_i}}\left(\frac{u_i}{\mu} - 1\right)\right]\right\}, \tag{3.11}$$

donde $\mu > 0$ es el promedio, $\lambda > 0$ es el parámetro de escala y erf es la función error[14] .

14: La función error está dada por: $erf(x) = \frac{1}{\sqrt{\pi}}\int_{-x}^{x} \exp(-t^2)dt$

Las palabras equivalentes *moi, yo, ich, ben* en Francés, Español, Alemán y Turco, tienen distribuciones tipo Gamma, Birbaum-Saunders, Gaussiana Inversa y Birbaum-Saunders respectivamente.

El problema de la detección automática de palabras funcionales es una cuestión abierta en el campo del procesamiento de lenguaje natural (NLP). Existen diferentes propuestas para automatizar la detección de *Stop Words* [57, 58], sin embargo hasta el momento, la solución más sencilla es la de conocer de antemano listas de palabras funcionales y removerlas en el pre-procesamiento de los textos [46].

Si se remueven las palabras funcionales de las listas obtenidas al aplicar la ecuación (3.9), se obtendrán un conjunto de palabras clave, que sin lugar a dudas, son muy relevantes para el *Don Quijote*

Tabla 3.8: Palabras clave, de mayor a menor valor de TF-RSD para *Don Quijote* en Inglés, Francés y Alemán. Para configurar esta lista, hemos removido las palabras funcionales de cada idioma

Inglés	Traducción	Francés	Traducción	Alemán	Traducción
sancho	sancho	sancho	sancho	sancho	sancho
don	don	don	don	quijote	quijote
quixote	quijote	quichotte	quijote	don	don
knight	caballero	histoire	historia	ritter	caballero
took	tomó	chevalier	caballero	pfarrer	sacerdote
worship	culto	épée	espada	lotario	lotario
lothario	lothario	visage	cara	gnaden	gracia
camilla	camilla	duc	duque	sag	decir
love	amor	père	padre	wirt	anfitrión
tears	lágrimas	grâce	gracia	wahrheit	verdad
duchess	duquesa	seigneur	señor	wert	valor
just	justo	répliqua	respondió	wort	palabra
master	maestro	maître	maestro	vater	padre
say	decir	fort	fuerte	herzog	duque
man	hombre	gouverneur	gobernador	essen	comer
art	arte	curé	sacerdote	herr	señor
father	padre	écuyer	escudero	schenke	taberna
lady	dama	chevaliers	caballeros	stelle	lugar
let	dejar	zoraïde	zoraïde	pansa	pansa
good	bien	âne	culo	geschichte	historia
know	saber	cheval	caballo	sage	saga
barber	barbero	femme	mujer	riesen	gigante

Los resultados de la Tabla 3.8 indican que al aplicar el *Term Frequency-Ransom Sampling Distribution*, combinado con un conocimiento previo de palabras funcionales, es posible encontrar palabras clave en el *Quijote* que al compararlas con las palabras clave obtenidas en la Tabla 3.1 son más relevantes para el libro en cuestión.

Corpus paralelos

Ahora que se tiene un método que permite encontrar palabras clave, es interesante ver si dichas palabras son las mismas al hacer una traducción del texto. Para ello será necesario el uso de los llamados corpus paralelos.

Se llama corpus paralelo a la colección de textos en donde se encuentra algún documento en un lenguaje L_1 y sus traducciones a un conjunto de idiomas $L_2 ... L_n$.

En el presente trabajo se elegirán textos de la literatura universal, que por su caracter de clásicos, han sido traducidos y revisados extensamente por parte de la comunidad literaria, y que se pueden encontrar en diferentes idiomas.

Si se supone que las palabras claves de un texto deben aparecer en aproximadamente la misma posición, independientemente del idioma en el que se encuentre escrito, sería posible entonces correlacionar dos palabras en un corpus paralelo (ver Figura 3.17).

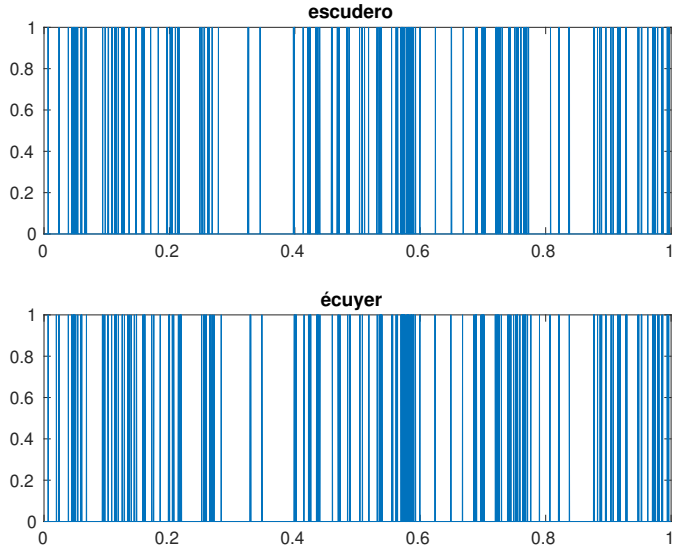


Figura 3.17: *Código de barras* para la palabra *escudero* en el *Don Quijote* en Español y Francés

Por el contrario, dos palabras, (como Rocinante y escudero), que no guardan ninguna relación entre sí, no tienen ningún tipo de correspondencia en su posición, y por lo tanto en un corpus paralelo, estas dos palabras no deberían estar correlacionadas.

Para comprobar que dicha hipótesis es cierta, se puede pensar en los *códigos de barras* como "señales". Esto es, alrededor de cada una de las posiciones vamos a suponer que en lugar de haber una línea, (como en la Figura 3.11 y en la Figura 3.17) hay una gaussiana de la forma

$$\Phi(x) = \sum_{i=1}^{f} \left(\frac{1}{2\pi\sigma^2} \right)^{(1/4)} e^{\frac{(x-\alpha_i)^2}{4\sigma^2}}, \tag{3.12}$$

donde f sería la frecuencia de la palabra, α_i sería la i-ésima posición de la palabra en el texto y σ sería la anchura de la gaussiana.

En la figura (Figura 3.18), se puede ver esquemáticamente que si la palabra se encuentra aproximadamente en la misma posición en el corpus paralelo, entonces su significado probablemente esté relacionado. En caso contrario, dos palabras cuya posición en el texto no coincide, posiblemente no guarden ninguna relación entre sí.

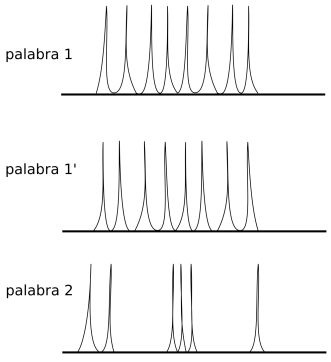


Figura 3.18: Esquema de la representación de las palabras a través de Gaussianas al rededor de su posición

Usando la ecuación (3.12), se puede calcular la superposición de estas dos "señales" como

$$O(\alpha_i, a_j) = \int_{-\infty}^{\infty} \Phi_1(x)\Phi_2(x)dx = \sum_{i=1}^{f_\alpha}\sum_{j=1}^{f_a} e^{-\frac{1}{8\sigma^2}(\alpha_i - a_j)^2}, \tag{3.13}$$

donde f_α, f_a serían las frecuencias de las palabras y α_i, a_j serían las posiciones de las palabras en cada uno de los textos del corpus paralelo.

De la ecuación (3.13) se puede calcular el coeficiente de correlación de Pearson de la siguiente manera:

$$\rho = \frac{\int_{-\infty}^{\infty} \Phi_1(x)\Phi_2(x)dx - (\int_{-\infty}^{\infty} \Phi_1(x)dx)(\int_{-\infty}^{\infty} \Phi_2(x)dx)}{\sqrt{\int_{-\infty}^{\infty} \Phi_1^2(x)dx \int_{-\infty}^{\infty} \Phi_2^2(x)dx}}. \tag{3.14}$$

El coeficiente de correlación de Pearson, permitirá entonces encontrar la relación que existe entre dos palabras del corpus paralelo. Solo restaría encontrar cuál es la forma apropiada de calcular σ, el ancho de las Gaussianas.

Dado que los textos que aquí se manejan son de tamaños que van desde *el principito* (con cerca de 10 mil palabras), hasta textos mucho más largos, como *Don Quijote*, (con alrededor de 400 mil palabras), debemos encontrar una manera de obtener σ que funcione para todos los documentos. En general la guía que usará en este trabajo es la siguiente:

$$\sigma = \frac{1}{\max(L_1 L_2)} 300, \tag{3.15}$$

donde $\max(L_1 L_2)$ es el máximo entre la longitud del texto 1 y la longitud del texto 2. Dicho de otro modo estamos suponiendo que hay 150 palabras a lado y lado de la posición real de la palabra.

Usando la ecuación (3.14) a las primeras 100 palabras ordenadas de acuerdo al valor que se obtiene con el TF-RSD (ecuación (3.9)) para *el Quijote* en Español y Francés, se obtiene la correlación entre palabras clave para estos dos idiomas.

En la Tabla 3.9, se pueden observar varias palabras clave para el *Don Quijote* en Español y Francés. Se puede ver que las palabras *Sancho, Quijote, Don* son identificadas correctamente en los dos idiomas.

Se identifican también las palabras *zoraida y teresa*, que son respectivamente las esposas de Ruy Pérez de Viedma[15] y Sancho Panza.

Nótese también que, a diferencia del *Word Burstiness* que parece asignar importancia solo a palabras que hacen referencia a personajes, el método aquí propuesto encuentra como palabras importantes a algunos términos que sin duda son relevantes para un libro de caballería, por ejemplo, *espada, caballero, escudero*; así como también palabras que, para el *Quijote* toman especial relevancia, tales como

15: Este personaje, también conocido como el cautivo, es el protagonista de la historia de un soldado que, al igual que Cervantes, lucha en la batalla de Lepanto, es tomado prisionero y con la ayuda de Zoraida (hija de su captor) logra escapar de su encierro y regresar a tierras cristianas.

	Palabra	Palabra	ρ
1	teresa	thérèse	0.96429
2	don	don	0.91498
3	gobernador	gouverneur	0.91445
4	zoraida	zoraïde	0.90398
5	sancho	sancho	0.90239
6	quijote	quichotte	0.89707
7	señor	seigneur	0.83647
8	venta	hôtellerie	0.83182
9	caballero	chevalier	0.82824
10	historia	histoire	0.81172
11	escudero	écuyer	0.80831
12	cura	curé	0.78845
13	don	quichotte	0.78301
14	espada	épée	0.73282
15	mujer	femme	0.70841
16	padre	père	0.68064
17	ventero	hôtellerie	0.65784
18	duquesa	duc	0.63029

Tabla 3.9: Correlación de las palabras clave para Don Quijote en Español y Francés

padre, cura, venta y ventero, estos últimos haciendo referencia al lugar para el hospedaje de viajeros, que es dónde transcurren muchas de las historias de este libro.

Como ejemplo final, en la Tabla 3.10 se observan los resultados de aplicar el método para *Veinte Mil Leguas de Viaje Submarino* de Jules Verne, en Inglés y Español

	Palabra	Palabra	ρ
1	ned	ned	0.935297163
2	red	rojo	0.9238526258
3	captain	capitán	0.9202891835
4	meters	metros	0.9197023735
5	lounge	salón	0.9171271175
6	nautilus	nautilus	0.9154708244
7	nemo	nemo	0.904683302
8	conseil	conseil	0.8795681811
9	air	aire	0.8451853084
10	nemo	capitán	0.8423326706
11	you	usted	0.8379069442
12	land	land	0.8378178748
13	captain	nemo	0.8373246312
14	abraham	fragata	0.8303461806
15	surface	superficie	0.8287551219
16	seas	mares	0.8275242012
17	professor	profesor	0.8007180364

Tabla 3.10: Correlación de las palabras clave para *Veinte mil leguas de viaje submarino* en Español e Inglés

Nuevamente es posible ver que las palabras encontradas por el procedimiento aquí descrito, son relevantes para la trama del libro. Se encuentran correlaciones altas para el término *nautilus*, se puede reconocer a personajes importantes como *nemo, ned, profesor, conseil*, y palabras que se refieren a cuestiones marinas *metros, mares, superficie*.

Al igual que en *el Quijote*, en este libro se correlacionan palabras que casi siempre aparecen una al lado de la otra, como *don* y *quijote*, o *capitán* y *nemo*, o *fragata* y *abraham*.[16]

16: La fragata Abraham Lincoln es el nombre de la embarcación, capitaneada por el cazador de ballenas Ned Land, en la cual viaja el Profesor Pierre Aronnax. Este barco zarpa con el objetivo de dar caza a un monstruo marino y termina encontrando al Capitán Nemo y al Nautilus.

Aplicación: El manuscrito de Voynich

4

En este capítulo se aplicarán los métodos desarrollados anteriormente, a un manuscrito que es considerado como uno de los enigmas más interesantes en la historia de la criptografía, el manuscrito de Voynich [31].

Dado su carácter misterioso, el manuscrito de Voynich se ha prestado para el desarrollo de mucha pseudo ciencia, pero también ha llamado la atención de científicos de diversas disciplinas, y varias investigaciones serias se han hecho al respecto. Los estudios recientes sobre el manuscrito han derivado en dos hipótesis: una de ellas afirma que las correlaciones estadísticas encontradas en el texto son compatibles con aquellas encontradas en lenguajes reales [59–61]. Otros investigadores sostienen que el manuscrito no es más que un elaborado engaño, y que las correlaciones estadísticas que se han encontrado en él, pueden explicarse con otros métodos [62, 63].

Los resultados obtenidos del análisis hecho en este trabajo, apoyan la idea que el manuscrito no es un engaño, y que por el contrario, las correlaciones estadísticas halladas en él, son similares a las encontradas en lenguajes naturales.

4.1 ¿Qué dice el Manuscrito de Voynich?

Nadie sabe. El manuscrito de Voynich toma su nombre del coleccionista y anticuario Wilfrid Voynich (1865-1930). Es un texto escrito en un lenguaje y alfabeto desconocido; y que hasta la fecha no ha podido ser descifrado. Lo que parece ser claro es que el manuscrito de Voynich data de finales de la baja edad media. Análisis de Carbono 14 indican que el pergamino en el cuál esta escrito data de entre 1404 y 1435. Los estudios realizados en las tintas usadas, indican que los pigmentos utilizados contienen varios tipos de minerales molidos de uso común durante toda la edad media. No se han encontrado rastros de componentes que sean inconsistentes con la datación por radiocarbono del pergamino. Los pigmentos se consideran en su mayoría de bajo costo; así como el trabajo hecho en el manuscrito se considera promedio, comparado con documentos similares de la misma época [64, 65].

Nada se sabe del manuscrito antes del año 1608. El ex libris [17] indica que fue propiedad de Jacobus Horčický (1565-1622), un farmacéuta, que fue nombrado como químico imperial en 1607 en la corte del Sacro Emperador Romano-Germánico Rodolfo II de Habsburgo (1552-1612), famoso por su patronazgo a la astronomía, (fue benefactor de Tycho Brahe y de Johannes Kepler) astrología, a las ciencias ocultas y al arte.

17: Se llama Ex Libris a la marca de propiedad que suele colocarse en un libro, y que contiene el nombre del dueño del ejemplar o de la biblioteca propietaria.

Figura 4.1: Rodolfo II del Sacro Imperio Romano Germánico.
https://upload.wikimedia.org/wikipedia/commons/6/64/Joseph_Heintz_d._Ä._002.jpg

Horčický curó al emperador de una enfermedad grave, razón por la cual fue ascendido a la clase nobiliaria con el titulo de Tepenec en el año de 1608. Él escribió su nombre en el margen inferior del primer folio del manuscrito usando su título nobiliario (Jacobus de Tepenec), siendo ésta la primera referencia historicamente comprobable que se tiene del manuscrito [66].

No es claro como el libro quedó en posesión de Georgius Barschius (1585-1662). Este oficial del tribunal de justicia de Praga, legó el manuscrito a Johannes Marcus Marci (1595-1667), rector de la universidad de Praga. Marci en 1666 decide enviar el manuscrito a Roma con su amigo y mentor Athanasius Kircher (1601-1680), profesor de la Pontificia Universidad Gregoriana, (antiguamente Colegio Romano), uno de los más grandes eruditos de su tiempo. Kircher trabajó en el desciframiento del manuscrito, pero los resultados de sus investigaciones son desconocidos [31].

El libro permaneció oculto por cerca de 250 años, hasta que Wilfrid Voynich, lo halló en la Villa Mondragone en Italia en 1912. En los primeros años de estudio, Voynich contactó a numerosos eruditos, uno de ellos el filósofo William Romaine Newbold (1865-1926), quien se adjudicó el desciframiento del texto, basándose en copias de algunas páginas del manuscrito. Su propuesta de desciframiento, sin embargo, rápidamente fue considerada como errónea [31].

Tras la muerte de Voynich, su esposa Ethel Lilian Voynich, facilitó al sacerdote Theodore C. Petersen (1883-1966), profesor de la Universidad Católica de América, una fotocopia completa del manuscrito. Él a su vez suministró copias del libro a varios expertos criptólogos, entre ellos a William Frederick Friedman (1891-1969), quien fue jefe de la división de criptoanálisis de la *National Security Agency* (NSA) y pionero en el uso de computadoras para analizar de manera sistemática el textos encriptados. Friedman, cuyos trabajos fueron de importancia para descifrar el código *PURPLE*, usado por el imperio del Japón durante la segunda guerra mundial, comienza con el análisis del Manuscrito de Voynich. Para ello reúne a un grupo de expertos en diversos campos (criptólogos, egiptólogos, matemáticos, lingüistas). Este grupo trabajó de 1944 a 1946 y realizó los primeros análisis de frecuencia de palabras y letras usando computadoras; así como una transcripción del texto que fuese posible leer en una computadora. En 1951, Friedman logra entusiasmar al criptólogo británico John Hessell Tiltman (1894-1982), quien desarrolló métodos de criptoanálisis que fueron de importancia para el desciframiento del código *Lorentz*, usado por la *Wehrmacht* en la segunda guerra mundial. Los resultados de sus trabajo se pueden encontrar en *The Voynich Manuscript, an elegant enigma* [31], el cuál es una de las principales referencias en el estudio del manuscrito de Voynich.

De esta primera época de estudio, los resultados más importantes son las conclusiones hechas por Friedman y Tiltman quienes afirmaron que el manuscrito no es un simple código de sustitución, y que además debe estar escrito en alguna lengua no natural. También es muy impor-

tante destacar el trabajo de Prescott Currier, quien descubrió que en el manuscrito existen dos lenguajes estadísticamente diferentes [61].

Tras la muerte de Ethel Lilian Voynich, el libro fue heredado a una de sus amigas Anne Nill. Ella vendió el libro al coleccionista Hans P. Kraus. Incapaz de encontrar un comprador, Kraus donó el manuscrito a la Universidad de Yale, que lo mantiene en la Biblioteca Beinecke de libros raros. En el 2016, la Universidad de Yale publicó un facsímil del manuscrito de Voynich, que contiene además ensayos de varios expertos, en donde se resume todo lo conocido sobre el mismo [64].

Estudios estadísticos en años recientes, han usado la información mutua de las palabras del manuscrito para encontrar aquellos términos que serían más informativos [59]. También se han usado métodos como el *Word Burstiness* y la teoría de redes para tratar de determinar las palabras clave y el lenguaje en el que el Voynich podría estar escrito, aunque sin resultados concluyentes para esta última parte [60]. Es importante notar que de estos trabajos se puede concluir básicamente que el texto escrito en el manuscrito parece ser compatible con lenguajes naturales.

Por otra parte, las hipótesis que apoyan la idea que el manuscrito es un engaño, se basan en el hecho que, a pesar del trabajo de reconocidos criptoanalístas como lo fueron Friedman y Tiltman, y del uso de computadores, no se ha podido extraer aún una explicación satisfactoria para un texto que fue escrito en la baja edad media, cuando los sistemas de encriptación no eran muy sofisticados, y que en principio debería ser fácil de descifrar.

También ha llamado mucho la atención la historia alrededor del origen del manuscrito. Está bien documentado que el Emperador Rodolfo II era muy devoto del ocultismo, llegando a tener una colección de curiosidades llena de libros raros de alquímia y magia, por los cuales pagaba muy bien, y que atrajeron a la corte a toda laya de alquimístas, charlatanes, científicos y embusteros, con ánimos de obtener dinero fácil. En particular los nombres de John Dee (1527-1608), matemático y ocultista y Edward Kelley (1555-1597), alquimista y ocultista, ambos de origen inglés han sido relacionados de manera circustancial con el Voynich. Se sabe bien que estos personajes eran considerados charlatanes (aún en su tiempo) y que perfectamente habrían podido llevar a cabo un engaño para estafar a Rodolfo II [31, 66].

Hay evidencia histórica que John Dee, estuvo en contacto con el erudito italiano Gerolamo Cardano (1501-1576), quien es conocido por ser el primer matemático en usar números negativos para resolver ecuaciones algebraicas. Entre sus muchos aportes, Cardano es el inventor de un instrumento llamado rejilla de Cardano.

Este dispositivo consiste generalmente en una pieza hecha de metal con perforaciones, que al ponerse de manera apropiada sobre un manuscrito revela un mensaje oculto en un texto ordinario. Ver Figura 4.2

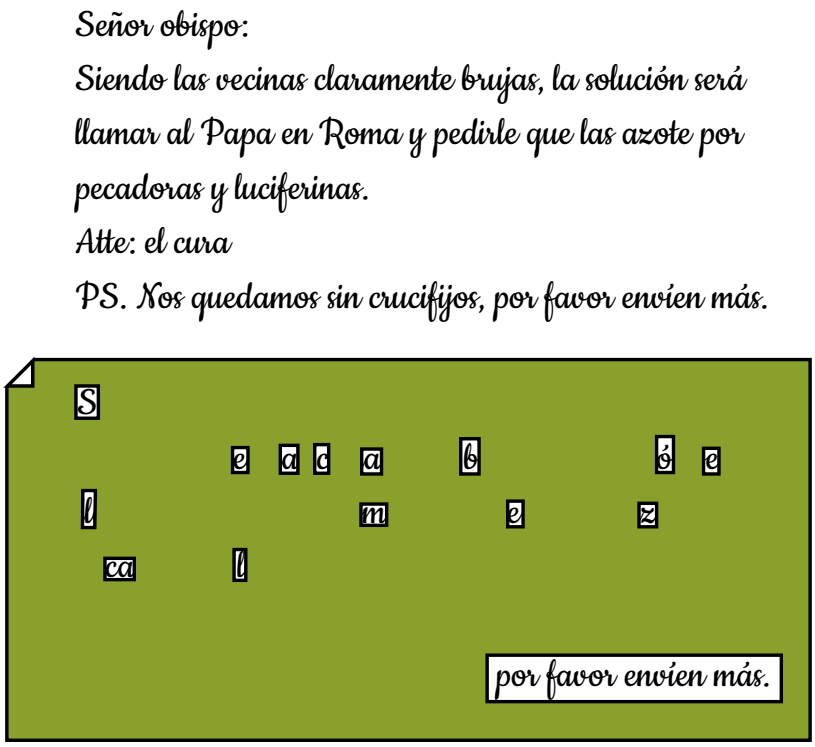


Figura 4.2: Ejemplo de como se utiliza una rejilla de Cardano.

En [62] se propone que la rejilla de Cardano pudo haberse usado para producir el manuscrito de Voynich. La idea aquí es la de usar un texto base, dividido en sílabas y escrito en papel en una lengua natural. Luego, una rejilla de Cardano, podría usarse para seleccionar de manera más o menos aleatoria, las sílabas que compondrían el texto del manuscrito. Algo parecido al método que usamos en el capítulo 2 para producir texto aleatorio.

Aunque esta porpuesta no es más que una especulación, ha llamado la atención de las personas interesadas en el Voynich, pues propone una forma (junto con la invención del alfabeto) eficiente para generar un texto, y que podría explicar algunas de las correlaciones estadísticas que se han encontrado en el análisis del texto escrito en el Voynich.

En [63] se propone usar caminatas aleatorias para determinar la existencia del correlaciones de largo alcance en textos escritos en lenguajes naturales. La idea de este método se basa en la propuesta de *Kokol et al* [67]. Si se asigna un código binario a cada letra de un escrito (Por ejemplo, $A = 00000$, $B = 00001$, etc), entonces se pueden usar los números 0 o 1 como pasos de una caminata aleatoria. Kokol muestra que la desviación estandar de los pasos es una ley de potencias de la forma $F(l) \propto l^{\alpha}$, con $\alpha \approx 0.5$ para lenguajes naturales.

Los resultados para el Voynich usando las caminatas aleatorias, muestran que el valor de $\alpha \approx 0.8$, siendo esto inconsistente con un texto escrito en algún idioma, pero consistente con un proceso estocástico como el descrito anteriormente.

4.2 Descripción del Manuscrito de Voynich

El manuscrito de Voynich es un códice de pergamino, que mide 225 x 160 mm y tiene unos 5 cm de grosor. Consta de 102 folios (se supone que originalmente fueron 116 de los cuales faltan 14). El manuscrito está escrito de manera elegante, pero con un alfabeto totalmente desconocido.

Casi todas sus páginas contienen ilustraciones de hierbas, constelaciones o sistemas de tubos de origen incierto, que transportan líquidos y están poblados por pequeñas figuras femeninas. Aunque se han identificado varias similitudes con las ilustraciones de otros manuscritos, las ilustraciones del Voynich son en gran medida exclusivas de este manuscrito. Para identificar a las páginas del manuscrito se usa la notación *f* (para folio) seguido del número de folio, seguido de *r* (para recto - el frente) o *v* (para verso - el reverso) de la página. Un ejemplo sería f17v, para el reverso de la página 17 del manuscrito.

Existe debate si el orden en el que se encuentra el manuscrito en la actualidad es el correcto, pues hay evidencia que fue reencuadernado. Esto hace que sea difícil saber cómo clasificarlo. A pesar de ello, y de acuerdo con las ilustraciones del manuscrito, este se ha dividido en 6 secciones: [31]

- **herbal**, con dibujos de hierbas y plantas que no son fácilmente identificables.
- **astronómica**, con ilustraciones del sol, la luna, estrellas y símbolos zodiacales.
- **cosmológica**, con dibujos, típicamente circulares de estrellas y constelaciones.
- **biológica**, la cual contiene dibujos *anatómicos* de pequeñas figuras femeninas
- **farmacéutica**, llamada así ya que hay dibujos de contenedores y de partes de plantas (hojas y raíces)
- **recetas**, la cual consiste en párrafos pequeños cada uno acompañado por el dibujo de una estrella en el margen.

El análisis de las ilustraciones del manuscrito es el principal medio para tratar de identificar el origen del Voynich. En particular llama la atención el castillo que aparece en una de las páginas (ver Figura 4.3) que es de estilo gibelino, típico del norte de Italia.

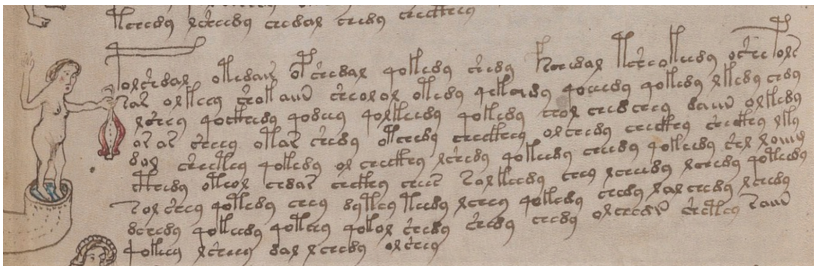
Figura 4.3: Imagen de un castillo de estilo gibelino (cuyas almenas tienen forma de V) y que aparece en folio 85v del manuscrito de Voynich.

Otra imagen que ha llamado la atención, es la ilustración de la constelación de Sagitario que aparece en el manuscrito (Ver Figura 4.5). En ella se representa a esta constelación con el dibujo de un hombre ataviado con un gorro del tipo cola de zorro y que en lugar de tener un arco, tiene una ballesta. Esta representación es poco común, pero se ha encontrado en manuscritos del sur de Alemania, escritos entre los años 1400 y el 1500 depués de cristo [68].

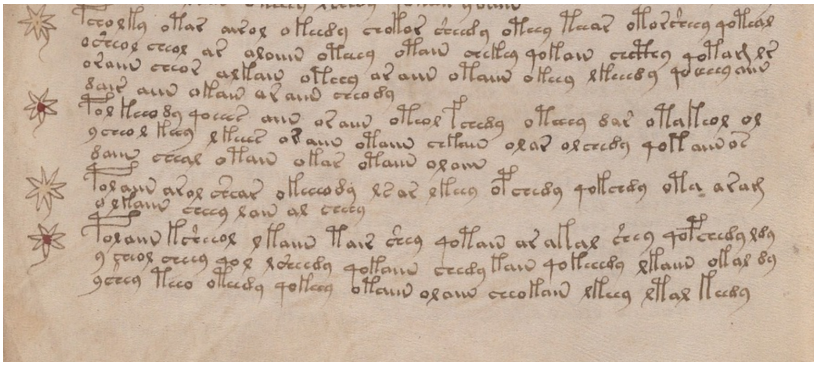
El análisis de estas ilustraciones ha llevado a especular que el o los autores del manuscrito de Voynich provienen de Europa central, y parece plausible que en esta región fue producido el manuscrito.

Herbal

Biológica



Recetas



Farmacéutica

Figura 4.4: Secciones herbal, biológica, recetas y farmacéutica del manuscrito de Voynich

Cosmológica

Astronómica

Figura 4.5: Secciones cosmológica y astronómica del manuscrito de Voynich

Análisis del texto

Ahora bien, en cuanto al análisis del texto, es necesario representar los glifos del manuscrito de una manera que permita ser leído por una computadora. Para ello se han propuesto varias transliteraciones del alfabeto. En este trabajo usaremos la transcripción denominada *European Voynich Alphabet* (EVA). En esta transcripción se utilizan las letras del alfabeto latino para representar los símbolos hallados en el manuscrito [69]. En la Figura 4.6 se muestra la equivalencia entre los caracteres latinos y los glifos del Voynich.

BASIC EVA CHARACTERS

	EVA	Capitalised EVA
'		
a		
b		
c		
d		
e		
f		
g		
h		
i		
j		
k		
l		
m		
n		
o		
p		
q		
r		
s		
t		
u		
v		
x		
y		
z		

PUNCTUATION CHARACTERS

	EVA	
*		unreadable
,		possibly a space
-		drawing intruding into text
.		space
=		end of paragraph
?		missing word
???		missing words
!		interlinear non-coding spacer
%		interlinear coding spacer

"UNOFFICIAL EVA"

"		plume on top of connector
+		plume intruding in connector

META CODES

#		line comment
{ }		in-line comment
< >		folio/locus indicator
[|]		alternative readings
\		line split (not in original)
$		weirdo code header
&		extended-eva header
;		end of extended-eva or weirdo code
( )		ligature notation

Figura 4.6: Transliteración del Manuscrito de Voynich, tomada de [69]

Esta transliteración es popular debido a que permite "pronunciar" las palabras encontradas en el manuscrito. Por ejemplo, las últimas tres líneas de la receta mostrada en la Figura 4.4 se leerían así:

```
polaiin.ksheeol.lkaiin.tair.shey.qotai!n.ar.akal.shey.qopchedy.ldy-
ycheol.cheey.qol.lsheedy.qokaiin.chedy.kai!n.qokeeedy.lkaiin.okal.dy-
yshey.teeo.oteedy.qokeey.otaiin.olaiin.cheokai!n.lkeey.ltal.keedy=
```

4.3 Resultados

En este apartado, se mostrarán los resultados para el manuscrito de Voynich que encontramos usando las herramientas desarrolladas en este trabajo.

Probablemente la primera observación que se hizo del Voynich y que aquí corroboramos, es que el texto del manuscrito satisface la ley de Zipf.

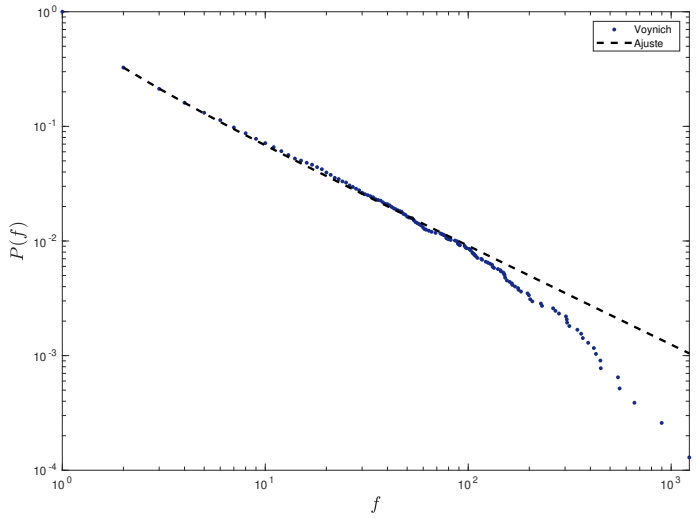


Figura 4.7: Distribución Cumulativa de frecuencias para el Manuscrito de Voynich. La línea negra es el ajuste hecho con el *Maximum Likelihood Estimator* para leyes de potencia.

En la Figura 4.7 se oberva la distribución cumulativa de la frecuencia de palabras en el Voynich. En este caso se puede estimar que la pendiente de la ley de potencias $P(f) \sim f^{-\alpha_Z}$ es $\alpha_Z = 1.8600 \pm 0.0171$, y el *p-value* = 0.1490

En la Figura 4.8 se ve el comportamiento de la ley de Heaps para el Voynich. De manera similar, en la Figura 4.10 se puede observar la distribución cumulativa para la distribución de grados de la red de co ocurrencia de palabras en el manuscrito.

De estas figuras es claro que el comportamiento estadístico del texto extraído del manuscrito de Voynich, es similar al comportamiento encontrado previamente para los idiomas estudiados, y no es posible ver ninguna diferencia significativa entre el "Voychinés" y los lenguajes naturales que aquí analizamos.

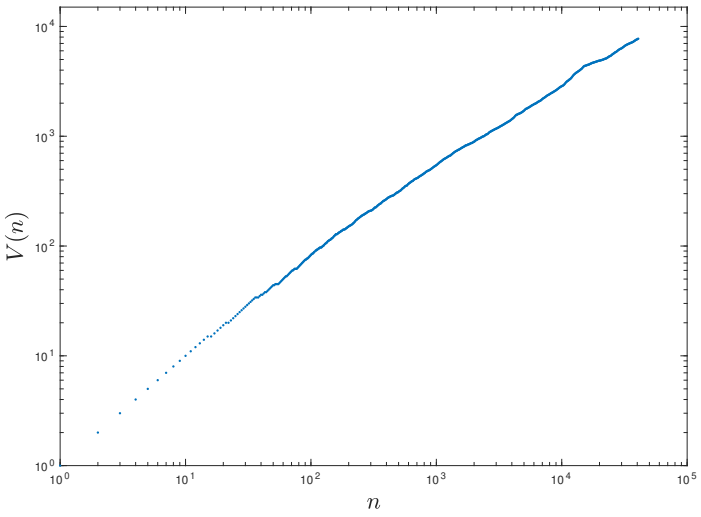


Figura 4.8: Ley de Heaps para el Manuscrito de Voynich. Para este caso $\beta = 0.7854 \pm 0.0018$

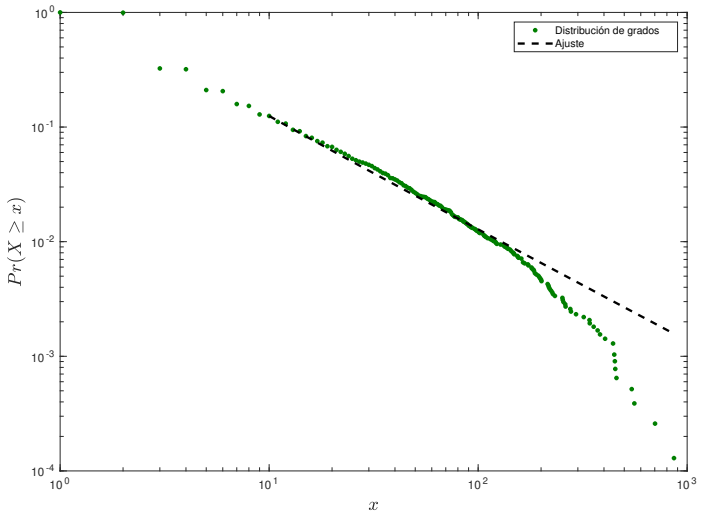


Figura 4.9: Cumulativa de la distribución de grados de la red de co-ocurrencia del Manuscrito de Voynich. Para este caso tenemos $\alpha_k = 1.9700 \pm 0.0312$ y *p-value*= 0.3194.

Los resultados más interesantes aparecen cuando se aplica el método de la distribución bi variada de los valores de $N(0)$ y $N(1)$ para la red de co-ocurrencia de las palabras del manuscrito. En este caso se hizo un proceso idéntico al realizado en el capítulo 2, teniendo en cuenta que el vocabulario del Voynich es de 7732 palabras, y además explorando idiomas que no se analizaron previamente.

La Tabla 4.1 resume los resultados encontrados. En este caso se puede ver que, de acuerdo con el método, el idioma en el cual está escrito el manuscrito de Voynich es cercano a la colección de textos en idioma Checo, que elegimos para estimar las distribuciones bi-variadas de los nodos cuyo *Clustering Coefficient* son 0 o 1.

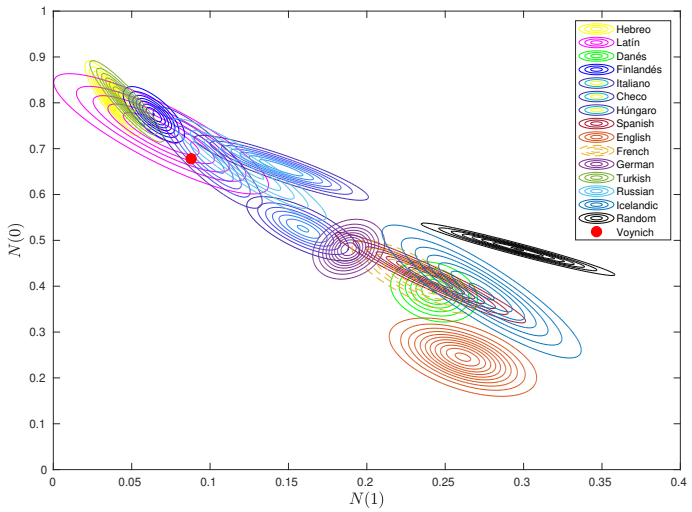


Figura 4.10: Distribución normal bivariada de $N(0)$ y $N(1)$ para el manuscrito de Voynich.

Tabla 4.1: Likelihood para el manuscrito de Voynich, los valores menores a 1×10^{-8} fueron redondeados a 0.

Idioma	Likelihood
Español	0.00025171
Inglés	0
Francés	0
Alemán	0
Turco	0.82594
Ruso	5.026
Islandés	3.803e-07
Checo	140.0559
Danés	0
Finlandés	4.2889
Hebreo	0
Húngaro	0.25293
Italiano	0.015372
Latín	88.1825
Random	0

Hasta donde sabemos, este es el primer resultado que postula un idioma para el manuscrito de Voynich extraído de un análisis estadístico del texto. Como se mencionó anteriormente, el análisis de algunas de las ilustraciones del manuscrito postulan que el origen de este es de Europa central. El análisis estadístico aquí realizado es consistente con dicha hipótesis y dada la estrecha relación que existe entre el Voynich y la ciudad de Praga, resultaría interesante realizar un estudio mucho más detallado sobre la posibilidad que el Checo sea el idioma en el cual se encuentra encriptado el manuscrito del Voynich.

En cuanto a las palabras clave, si se hace una gráfica de TF-RSD$_{total}$ como función de k es posible ver que por la división óptima sería para $k^* = 116$, tal y como se muestra en la Figura 4.12.

Utilizando la ecuación (3.9), se obtiene la Tabla 4.2, en donde se enlistan las palabras que de acuerdo con el método desarrolado en el capítulo 3, serían claves en el manuscrito de Voynich.

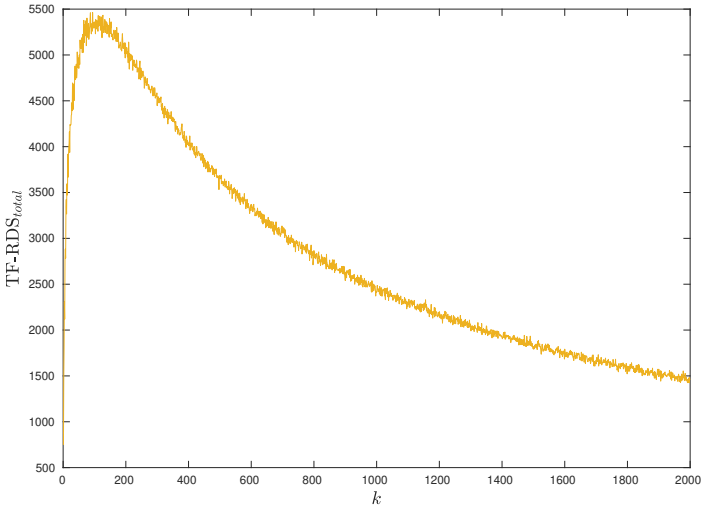


Figura 4.11: TF-RSD como función de k para manuscrito de Voynich. Se puede obervar que $k^* = 116$.

Tabla 4.2: TF-RSD para le manuscrito de Voynich.

	Palabra	TF-RSD	Frecuencia		Palabra	TF-RSD	Frecuencia
1	qokain	172.14	279	12	dy	67.723	273
2	qokeedy	163.62	305	13	dair	64.252	106
3	shedy	158.28	426	14	lchedy	62.826	119
4	chedy	148.84	500	15	al	61.366	254
5	qokedy	137.41	270	16	okeey	58.689	176
6	qol	105.77	151	17	otedy	55.235	155
7	qokaiin	92.722	262	18	sho	52.361	129
8	qokeey	91.263	306	19	shey	52.229	282
9	chor	85.661	216	20	ar	49.494	352
10	chol	73.545	393	21	okain	47.713	143
11	qokal	73.025	191	22	or	46.784	369

Si se buscan estas palabras en el documento, se puede ver que los términos *qokain* y *qokeedy* aparecen en 57 y 63 de las 116 particiones en las que dividimos el texto, respectivamente.

Llama la atención que dichos términos aparecen con mucha frecuencia en el folio *f75r*, 13 y 11 veces respectivamente (ver Figura 4.12). Es en esta página que con diferencia, se concentra el mayor número de repeticiones dichos términos. Ciertamente este folio podría ser punto de partida interesante para empezar un análisis del manuscrito de Voynich.

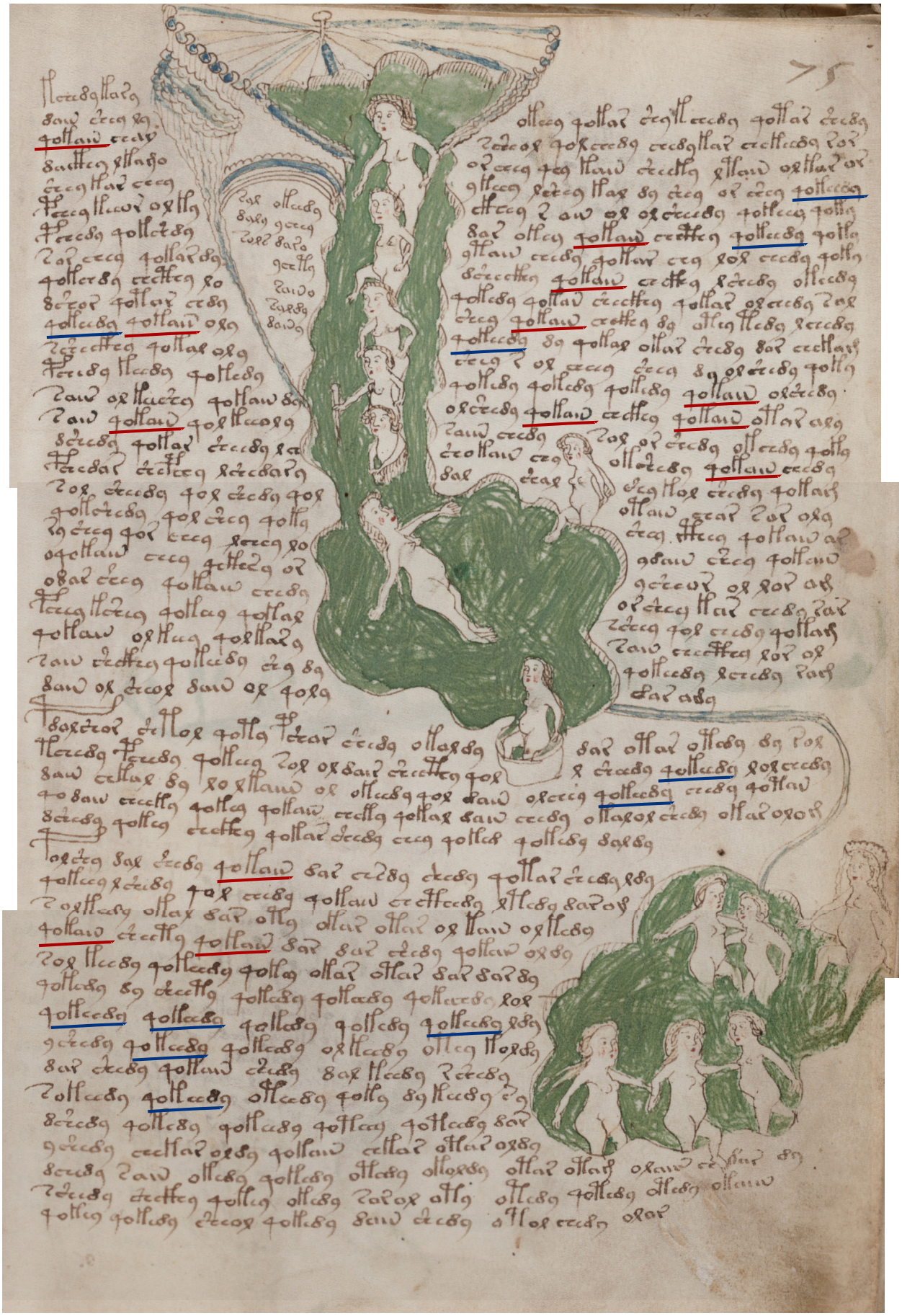


Figura 4.12: Folio 75r del manuscrito de Voynich, donde podemos encontrar la palabras *qokain* (líneas rojas) y *qokeedy* (líneas azules) varias veces.

5 Conclusiones

En este trabajo se analizaron estadísticamente alrededor de 120 textos en catorce idiomas diferentes, así como textos aleatorios construidos aleatorizando los espacios entre palabras sin alterar el orden de las letras en el texto. La primera conclusión a la que podemos llegar es que los textos escritos en lenguajes naturales y también los textos aleatorios, presentan comportamientos indistinguibles de las leyes de Zipf y Herdan-Heaps.

De acuerdo con los resultados encontrados en este trabajo, los textos aleatorios satisfacen las mismas leyes que los textos escritos en lenguajes naturales. Existe aún mucha controversia respecto a si los textos aleatorios satisfacen o no la ley de Zipf [15, 37, 51]. Aquí, nosotros ofrecemos dos argumentos que refuerzan la idea que la ley de Zipf, no es estrictamente una ley lingüística, sino probablemente algo más general.

El primer argumento se basa en el hecho que, mediante el uso del método estadístico desarrollado para analizar si datos empiricos siguen una ley de potencias [30], los textos aleatorios que aquí se analizaron satisfacen la ley de Zipf, con una pendiente estadísticamente igual a la de los textos escritos en lenguajes naturales.

En segundo lugar, para todos los textos usados en la construcción de la red de co-ocurrencia de palabras, se puede argumentar que la distribución de grados de los nodos de la red, satisface una ley de potencias de la forma

$$P(k) \propto \frac{1}{k^{1+1/\alpha}}, \tag{5.1}$$

donde α es el valor de la pendiente de la distribución de rango frecuencia de palabras de un texto, es decir la pendiente de la ley de Zipf. Dado que $\alpha \simeq 1$ en todos los casos, entonces de la ecuación (5.1) se sigue que $P(k) \sim k^{-2}$. Esto último se se encuentra en buen acuerdo con los resultados obtenidos del análisis estadístico de $P(k)$ para las redes de co-ocurrencia, incluyendo a aquellas construidas con textos aleatorios, en donde encontramos que la pendiente de la disitribución de grados es $\alpha_k \approx 2$ (ver Apendice A).

Ahora bien, dado que la ley de Herdan-Heaps, se puede derivar de la ley de Zipf [13], entonces estas tres leyes (ley de Zipf, ley de Herdan-Heaps y la distribución de $P(k)$) se encuentran relacionadas entre sí y, de acuerdo con nuestros resultados, tienen un caracter universal.

Sin embargo, la distribución del *Clustering Coefficient*, captura de manera efectiva la estructura local de la red de co-ocurrencia de palabras.

Se puede observar que en dicha estructura, aquellos nodos cuyo $C(k)$ toma valores de 0 o 1, son los más abundantes en la red. Adicionalmente, la estadística hecha con estos nodos parece ser sensible al idioma, lo cual nos permite proponer un método que claramente diferencia entre textos aleatorios y textos escritos en lenguajes naturales; y que además permite establecer diferencias entre idiomas, y proponer una forma de estimar "distancias" entre lenguajes, o al menos entre familias de lenguajes.

Esto podría ser de utilidad para la clasificación de lenguajes dentro de familias lingüísticas. A diferencia de otros métodos, nuestra propuesta no requiere de la existencia de corpus de textos de gran tamaño, lo cual puede ser de mucha utilidad para el estudio y clasificación de aquellas lenguas (por ejemplo las lenguas indígenas), en donde la existencia de corpus lingüísticos no está garantizada.

Por otro lado, estudiamos algunos métodos para encontrar palabras claves en un texto. Dichos métodos presentan problemas a la hora de ser aplicados a libros, pues o bien no parecen reconocer el hecho que pueden existir palabras claves que se distribuyen a lo largo del texto (*Word Burstiness*), o bien dependen de corpus lingüísticos (*Term Frequency - Inverse Document Frequency*).

Sin embargo, tomando elementos claves de éstos dos métodos, podemos proponer un nuevo procedimiento que puede encontrar palabras clave en en documentos relativamente cortos, y que además reconoce a las mismas palabras clave de un texto escrito en distintos idiomas.

Esto podría ser de utilidad para mejorar los algoritmos de traducción automática de textos, que parecen depender exclusivamente de enfoques basados en el *Machine Learning*.

Es interesante notar también que la distribución de distancias de ciertas palabras funcionales parece ser común a varios lenguajes, tal y como pasa con las palabras *y, la*. Otras palabras como *yo* y sus equivalentes en otras lenguas, presentan otro tipo de distribuciones de distancias. Sería entonces muy interesante hacer un estudio sistemático de como es la distribución de distancias de las palabras funcionales en diferentes idiomas.

Adicionalmente, siguiendo la propuesta de [53], en la cual, se puede aplicar su método al estudio de secuencias de ADN, nuestro método perfectamente podría adaptarse también a estas secuencias (y en general a cualquier secuencia simbólica) y ser de utilidad en campos fuera de la lingüística.

Como un ejemplo de esto último, aplicamos los métodos desarrollados en este trabajo para estudiar el texto del manuscrito de Voynich. De este análisis encontramos resultados que son interesantes, ya que fue posible proponer la posibilidad que el documento este escrito en Checo, lo cuál es consecuente con algunas propuestas del origen del Voynich, y además propusimos una lista de palabras clave.

En particular la palabra *qokain*, que de acuerdo con nuestro análisis resulto ser la palabra más importante, aparece con mucha frecuencia en el *folio 75r*, que pertenece a la sección biológica.

Este podría ser un interesante punto de partida para un nuevo análisis del manuscrito de Voynich.

En los trabajos [59] y [60], se encontró que, de acuerdo con sus métodos, palabras tales como *qokain, qokeedy, shedy, qokain, chor, qokal, lchedy, al, dy, okeey* son importantes para el manuscrito de Voynich. Los resultados hallados en este trabajo, se encuentran en concordancia con los de ellos, ya que dichos términos fueron determinados también por nuestra metodología.

Apéndices

A

Tablas y Resultados

En este apéndice presentamos tablas de resultados para los datos analizados en este trabajo. α_k y σ_k representan el parámetro y el error estándar de la ley de potencia para la distribución de grados de las redes de co-ocurrencia $P(k) \propto k^{\alpha_k}$. Igualmente, α_Z y σ_z representan el parámetro y el error estándar de la distribución de frecuencias $P(f) \propto f^{-\alpha_Z}$. Los valores de la ley de Heaps $V(n) \propto N^{\beta}$ y σ_h se obtuvieron mediante el ajuste de mínimos cuadrados.

Español

Tabla A.1: Resultados para Español.

Book Name	Length	Vocabulary	α_k	σ_k	k_{min}	*p-value*	α_Z	σ_Z	f_{min}	*p-value*	β	σ_h
The Count of Montecristo	92378	11275	2.15	0.03	9	0.439	1.87	0.01	1	0.640	0.781	0.002
Don Quixote	113068	11277	2.11	0.03	10	0.206	1.84	0.01	1	0.119	0.810	0.002
The Three Musketeers	106869	11242	2.10	0.03	11	0.669	1.86	0.01	1	0.203	0.746	0.002
Unamuno	104769	11219	2.05	0.03	10	0.602	1.89	0.01	1	0.107	0.765	0.002
Valle-Inclan	76657	11252	2.24	0.03	8	0.532	2.04	0.02	5	0.331	0.780	0.002
Concha Espina	60356	11226	2.33	0.04	9	0.190	2.12	0.03	4	0.445	0.814	0.001
Angelina	71434	11281	2.23	0.03	8	0.180	2.02	0.02	3	0.583	0.810	0.002
Iliad	91203	11275	2.19	0.03	8	0.658	1.96	0.02	4	0.419	0.799	0.002
Odyssey	92381	11290	2.18	0.02	6	0.289	1.96	0.02	4	0.510	0.797	0.002
Pio Baroja	85227	11273	2.21	0.03	8	0.601	2.03	0.03	7	0.362	0.787	0.001
The White Company	76186	11232	2.18	0.03	9	0.126	1.97	0.02	3	0.510	0.786	0.002
Moby Dick	69986	11230	2.15	0.03	9	0.249	2.00	0.01	2	0.533	0.795	0.002
TwentyThousand	76443	11214	2.22	0.03	9	0.105	2.01	0.02	3	0.860	0.788	0.001

Inglés

Tabla A.2: Resultados para Inglés.

BookName	Length	Vocabulary	α_k	σ_k	k_{min}	*p-value*	α_Z	σ_Z	f_{min}	*p-value*	β	σ_h
Don Quixote	221474	11278	2.16	0.03	17	0.417	1.90	0.02	15	0.823	0.731	0.003
The Count of Montecristo	178516	11261	2.17	0.03	19	0.244	1.97	0.03	16	0.934	0.703	0.002
The Three Musketeers	233220	11266	2.14	0.03	24	0.972	1.91	0.03	33	0.579	0.704	0.003
Jane Austen	368076	11270	2.16	0.03	34	0.805	1.93	0.04	84	0.702	0.660	0.003
Celebrated Crimes	156044	11274	2.20	0.03	17	0.569	2.05	0.04	28	0.505	0.726	0.002
Les Miserables	131649	11254	2.17	0.04	19	0.198	1.97	0.03	10	0.867	0.736	0.002
Anna Karenina	259749	11268	2.13	0.03	20	0.965	1.86	0.02	11	0.105	0.687	0.002
War And Peace	201580	11223	2.17	0.04	33	0.612	1.94	0.03	25	0.590	0.699	0.002
Brothers Karamazov	291642	11212	2.12	0.03	27	0.647	1.85	0.02	22	0.600	0.686	0.003
Oscar Wilde	174912	11262	2.15	0.04	28	0.865	1.92	0.03	24	0.774	0.716	0.002
Charles Dickens	183844	11266	2.12	0.03	20	0.738	1.89	0.02	9	0.992	0.714	0.002
Twenty Years Later	231543	11257	2.12	0.04	29	0.854	1.92	0.04	44	0.718	0.701	0.003
Bram Stoker	221752	11265	2.13	0.03	23	0.182	1.88	0.03	20	0.804	0.691	0.002

Francés

Tabla A.3: Resultados para Francés.

Book Name	Length	Vocabulary	α_k	σ_k	k_{min}	*p-value*	α_Z	σ_Z	f_{min}	*p-value*	β	σ_h
The Count of Montecristo	105525	11271	2.10	0.03	9	0.378	1.89	0.02	3	0.681	0.745	0.002
Don Quixote	111728	11237	2.10	0.02	8	0.495	1.89	0.02	5	0.628	0.746	0.002
The Three Musketeers	111274	11268	2.07	0.03	11	0.520	1.85	0.01	1	0.326	0.768	0.002
Oscar Wilde	85015	11206	2.15	0.03	8	0.422	1.92	0.01	1	0.538	0.783	0.002
Madame Bobary	72966	11292	2.22	0.03	8	0.001	2.00	0.01	2	0.940	0.782	0.002
Honoré de Balzac	78495	11264	2.17	0.03	9	0.062	1.98	0.02	3	0.160	0.799	0.002
Homero	149951	11236	2.11	0.03	11	0.682	1.86	0.02	7	0.212	0.722	0.002
Notre Dame	69988	11282	2.18	0.03	9	0.012	1.98	0.01	1	0.965	0.784	0.001
Lesuieur	85886	11250	2.17	0.03	8	0.122	1.97	0.02	4	0.729	0.778	0.002
Guy de Maupassant	74709	11257	2.16	0.03	9	0.068	1.93	0.01	1	0.499	0.795	0.002
Twenty Thousand Leagues Under the Sea	74369	11272	2.23	0.03	8	0.001	2.00	0.02	3	0.895	0.781	0.002
Voltaire	81450	11267	2.15	0.03	9	0.002	1.95	0.01	2	0.160	0.772	0.001
Les Miserables	79011	11275	2.14	0.03	9	0.238	1.95	0.01	1	0.769	0.784	0.001

Alemán

Tabla A.4: Resultados para Alemán.

Book Name	Length	Vocabulary	α_k	σ_k	k_{min}	*p-value*	α_Z	σ_Z	f_{min}	*p-value*	β	σ_h
The Count of Montecristo	99693	11263	2.06	0.02	8	0.669	1.82	0.01	1	0.156	0.738	0.002
Don Quixote	81741	11323	2.07	0.03	10	0.716	1.92	0.01	1	0.385	0.921	0.006
The Three Musketeers	107870	11271	2.04	0.03	13	0.623	1.82	0.01	1	0.629	0.743	0.002
Honoré de Balzac	75986	11287	2.05	0.03	11	0.414	1.93	0.01	1	0.772	0.783	0.002
Rudolf Hans Bartsch	58874	11288	2.07	0.04	18	0.496	1.94	0.03	5	0.705	0.805	0.002
Felix Dahn I	67330	11268	2.17	0.05	23	0.616	1.96	0.01	1	0.404	0.785	0.002
Felix Dahn II	75792	11257	2.09	0.02	8	0.658	1.91	0.01	1	0.248	0.781	0.002
Charles Dickens I	82374	11274	2.06	0.02	8	0.128	1.90	0.01	1	0.853	0.779	0.002
Cahrles Dickens II	81893	11285	2.00	0.03	9	0.256	1.92	0.01	1	0.536	0.822	0.003
Alfred Döblin	56757	11240	2.12	0.03	10	0.595	2.02	0.02	2	0.526	0.787	0.002
Gustave Falke	62815	11202	2.07	0.03	10	0.939	1.96	0.02	3	0.225	0.788	0.002
MobyDick	72414	11215	2.08	0.03	12	0.676	1.94	0.01	2	0.329	0.779	0.002
Crime and Punishment	96492	11260	2.19	0.05	28	0.366	1.80	0.01	1	0.879	0.756	0.002

Turco

Tabla A.5: Resultados para Turco.

BookName	Length	Vocabulary	α_k	σ_k	k_{min}	*p-value*	α_Z	σ_Z	f_{min}	*p-value*	β	σ_h
The Count of Montecristo	42040	11198	2.26	0.06	19	0.524	2.07	0.02	2	0.455	0.822	0.001
Don Quixote	35207	11241	2.27	0.05	12	0.162	2.18	0.01	1	0.310	0.881	0.002
The Three Musketeers	40731	11280	2.22	0.04	10	0.145	2.07	0.02	2	0.317	0.857	0.002
Tale of Two Cities	37838	11292	2.26	0.04	11	0.371	2.14	0.01	1	0.113	0.855	0.002
Oscar Wilde	35065	11205	2.29	0.05	14	0.182	2.13	0.02	2	0.367	0.866	0.002
Jules Verne I	35595	11264	2.29	0.05	12	0.713	2.11	0.02	2	0.592	0.845	0.001
David Copperfield	39672	11213	2.23	0.04	10	0.711	2.09	0.02	2	0.595	0.854	0.002
Crime and Punishment	39716	11279	2.25	0.04	12	0.197	2.10	0.01	1	0.756	0.855	0.002
Turkish I	26347	11240	2.50	0.08	16	0.944	2.34	0.01	1	0.357	0.913	0.001
Turkish II	26765	11244	2.43	0.07	14	0.934	2.31	0.02	2	0.204	0.904	0.002
Turkish III	27564	11288	2.34	0.06	12	0.487	2.21	0.04	4	0.336	0.883	0.001
MobyDick	33500	11224	2.31	0.06	16	0.352	2.19	0.01	1	0.455	0.881	0.001
Jules Verne II	40060	11225	2.25	0.04	10	0.386	2.06	0.01	1	0.189	0.863	0.002

Ruso

Tabla A.6: Resultados para Ruso

Book Name	Length	Vocabulary	α_k	σ_k	k_{min}	*p-value*	α_Z	σ_Z	f_{min}	*p-value*	β	σ_h
Don Quixote	41169	11277	2.16	0.04	10	0.799	2.21	0.01	1	0.485	0.864	0.002
The Count of Montecristo	47282	11234	2.16	0.04	10	0.615	2.11	0.01	1	0.367	0.802	0.002
The Three Musketeers	51306	11277	2.14	0.03	10	0.869	2.06	0.01	1	0.196	0.818	0.002
Anna Karenina	53333	11242	2.12	0.04	11	0.625	2.02	0.02	2	0.175	0.823	0.002
War And Peace	45596	11321	2.14	0.03	9	0.019	2.09	0.01	1	0.232	0.821	0.002
Brothers Karamazov	47083	11293	2.11	0.05	16	0.861	2.16	0.01	1	0.785	0.835	0.002
Twenty Thousand Leagues Under the Sea	35961	11297	2.29	0.05	10	0.766	2.21	0.01	1	0.108	0.865	0.002
Anton Chekhov	45423	11282	2.18	0.04	11	0.713	2.13	0.01	1	0.714	0.869	0.002
Oscar Wilde	43504	11321	2.10	0.04	12	0.792	2.00	0.04	7	0.624	0.823	0.002
Honoré de Balzac	35407	11280	2.15	0.05	12	0.886	2.05	0.04	5	0.429	0.881	0.002
Twenty Years Later	48539	11250	2.10	0.04	11	0.636	1.99	0.03	4	0.801	0.823	0.002
Moby Dick	34748	11234	2.16	0.05	11	0.578	2.07	0.03	4	0.856	0.857	0.002
Crime and Punishment	40035	11217	2.19	0.05	16	0.724	2.15	0.01	1	0.678	0.835	0.001

Islandés

Tabla A.7: Resultados para Islandés.

BookName	Length	Vocabulary	α_k	σ_k	k_{min}	*p-value*	α_Z	σ_Z	f_{min}	*p-value*	β	σ_h
TorfhildiHólm	73242	11202	2.18	0.06	29	0.838	1.97	0.01	1	0.156	0.773	0.002
SagaI	99051	11184	2.03	0.02	8	0.148	1.88	0.01	1	0.569	0.753	0.001
SagaII	141436	11248	1.95	0.02	6	0.501	1.76	0.01	1	0.551	0.714	0.002
SagaIII	103020	11270	2.00	0.02	8	0.964	1.84	0.01	2	0.640	0.734	0.001
SagaIV	116521	11235	1.99	0.02	6	0.256	1.81	0.01	1	0.102	0.735	0.002
SagaV	106061	11290	1.98	0.02	6	0.465	1.84	0.01	1	0.659	0.729	0.001
SagaVI	116956	11296	2.21	0.06	50	0.634	1.83	0.01	1	0.118	0.734	0.002
SagaVII	119928	11287	2.20	0.06	49	0.794	1.81	0.01	1	0.216	0.742	0.001
JónTrausti	66577	11238	2.05	0.03	9	0.278	1.94	0.02	3	0.553	0.785	0.001
JónThoroddsen	89739	11249	2.02	0.03	8	0.148	1.85	0.02	4	0.273	0.757	0.001
ÞorgilsGjallanda	65357	11285	2.10	0.03	8	0.227	1.97	0.02	2	0.295	0.786	0.001
SmásögurI	58932	11287	2.10	0.03	9	0.717	1.98	0.02	4	0.811	0.803	0.001
SmásögurII	61272	11226	2.10	0.04	12	0.301	1.99	0.02	2	0.126	0.803	0.001

Random

Tabla A.8: Resultados para los textos aleatorios.

Book Name	Length	Vocabulary	α_k	σ_k	k_{min}	*p-value*	α_Z	σ_Z	f_{min}	*p-value*	β	σ_h
Random I	63904	11258	2.05	0.03	10	0.901	1.88	0.02	3	0.952	0.805	0.001
Random II	62391	11251	2.00	0.04	12	0.335	1.88	0.02	3	0.678	0.788	0.001
Random III	62619	11286	2.02	0.03	9	0.522	1.90	0.02	3	0.445	0.802	0.001
Random IV	61148	11208	1.99	0.03	11	0.256	1.91	0.02	3	0.856	0.808	0.001
Random V	63181	11291	2.04	0.03	8	0.407	1.93	0.02	2	0.225	0.791	0.001
Random VI	62430	11302	2.00	0.04	14	0.294	1.87	0.03	5	0.247	0.796	0.001
Random VII	66740	11224	2.10	0.06	29	0.588	1.88	0.04	10	0.704	0.804	0.001
Random VIII	65939	11251	1.98	0.03	10	0.008	1.86	0.02	4	0.478	0.812	0.002
Random IX	62318	11247	2.03	0.03	9	0.258	1.90	0.02	3	0.151	0.810	0.001
Random X	61574	11239	1.98	0.03	11	0.395	1.92	0.02	2	0.102	0.814	0.001
Random A	66795	11277	2.01	0.03	9	0.812	1.87	0.03	5	0.523	0.797	0.001
Random B	65996	11262	2.01	0.04	11	0.895	1.88	0.03	4	0.755	0.797	0.001

B

Textos usados

Aquí presentamos listas de los textos que se utilizaron en este trabajo. La gran mayoría de los documentos se obtuvieron del proyecto Gutemberg (https://www.gutenberg.org/), excepto los textos en Ruso, Turco, Islandés, Hebreo y Checo, que se obtuvieron de otras fuentes.

Tabla B.1: Documentos en Español e Inglés. (Fuente: Proyecto Gutemberg).

Español	
Alexandre Dumas	The Count of Montecristo The Three Musketeers
Miguel de Cervantes	Don Quixote
Miguel de Unamuno	Niebla Una Historia De Pasión
Ramón del Valle-Inclan	Memorias Del Marqués De Bradomin: Sonata De Otoño Sonata De Verano Sonata De Primavera Sonata De Invierno
Concha Espina	Agua De Nieve La Esfinge Maragata Dulce Nombre
Rafael Delgado	Angelina
Homer	Iliad Odyssey
Pío Baroja	Memorias De Un Hombre De Acción: El Aprendiz De Conspirador Los Caminos Del Mundo
Arthur Conan Doyle	The White Company
Herman Melville	Moby Dick
Jules Verne	Twenty Thousand Leagues Under the Sea

Inglés	
Miguel de Cervantes	Don Quixote
Alexandre Dumas	The Count of Montecristo The Three Musketeers Celebrated Crimes Twenty Years Later
Jane Austen	Mansfield Park Northanger Abbey Persuasion Sense and Sensibility
Victor Hugo	Les Miserables
Leon Tolstói	Anna Karenina War and Peace
Fyodor Dostoevsky	Brothers Karamazov
Oscar Wilde	The Picture of Dorian Gray The Happy Prince and Other Tales De Profundis A House Of Pomegranates The Canterville Ghost Selected Prose Of Oscar Wilde
Charles Dickens	Oliver Twist A Tale Of Two Cities
Bram Stoker	Dracula The Jewel of Seven Stars

Tabla B.2: Documentos en Turco y Ruso. Fuente Turco: www.ekitapcilar.com
Turkish I, II and III se obtuvieron University of Oxford Text Archive http://ota.ox.ac.uk/desc/0387
Fuente Ruso: https://www.e-reading.club

Turco	
Alexandre Dumas	The Count of Montecristo The Three Musketeers
Miguel de Cervantes	Don Quixote
Charles Dickens	A Tale of Two Cities David Copperfield
Turkish I Turkish II Turkish III	Modern prose: samples from literary texts and newspapers
Jules Verne	Twenty Thousand Leagues Under the Sea From the Earth to the Moon Around the World in 80 Days
Herman Melville	Moby Dick
Fyodor Dostoevsky	Crime and Punishment

Ruso	
Alexandre Dumas	The Count of Montecristo The Three Musketeers Twenty Years Later
Miguel de Cervantes	Don Quixote
Oscar Wilde	The Portrait of Dorian Gray De Profundis
Honoré de Balzac	Fater Goriot A Woman of Thirty
Jules Verne	Twenty Thousand Leagues Under the Sea Mysterious Island Around the World in 80 Days
Anton Chekhov	Short Stories Compilation
Fyodor Dostoevsky	Brothers Karamazov
Leo Tolstoy	Anna Karenina War And Peace

Tabla B.3: Fuente: Gutemberg Project

Francés	
Miguel de Cervantes	Don Quixote
Alexandre Dumas	The Count of Montecristo The Three Musketeers
Victor Hugo	The Hunchback of Notre-Dame Les Miserables
Jules Verne	Twenty Thousand Leagues Under the Sea
Guy de Maupassant	Ball of Fat Moonlight Contes de la Bécasse
Oscar Wilde	The Portrait of Dorian Gray Intentions
Gustave Flaubert	Madame Bovary
Honoré de Balzac	The Human Comedy. Scenes from private life: At the Sign of the Cat and Racket The Ball at Sceaux The Purse The Vendetta Madame Firmiani A Second Home Domestic Bliss The Imaginary Mistress Study of a Woman Albert Savarus
Homer	Iliad
Daniel Lesueur (Jeanne Lapauze)	Amour D'Aujourd'Hui
Voltaire	Candide

Alemán	
Alexandre Dumas	The Count of Montecristo The Three Musketeers
Miguel de Cervantes	Don Quixote
Honoré de Balzac	Grosse Und Kleine Welt (Short Stories) A Woman of Thirty
Rudolf Hans Bartsch	Grenzen der Menschheit Vom sterbenden Rokoko
Felix Dahn	Ein Kampf um Rom I Ein Kampf um Rom II
Charles Dickens	Oliver Twist A Tale of Two Cities
Alfred Döblin	Die Lobensteiner reisen nach Böhmen
Gustav Falke	Der Mann im Nebel
Herman Melville	Moby Dick
Fyodor Dostoevsky	Crime and Punishment

Tabla B.4: Documentos en Islandés. Las sagas se obtuvieron https://sagadb.org/. Los otros textos se obtuvieron de https://www.snerpa.is/net/index.html

	Islandés
Torfhildi Hólm	Brynjólfur Biskup Sveinsson
Sagas I	Bandamanna Saga Bardar Saga Bjarnar Saga Droplaugarsona Saga Gisla Saga Hrafnkels Saga Eiríks Saga Eyrbyggja Saga
Sagas II	Brennu-Njáls Saga Laxdæla Saga
Sagas III	Egils Saga Grettis Saga
Sagas IV	Finnboga Saga Fljótsdæla Saga Flóamanna Saga Fóstbræðra Saga Grænlendinga Saga Gull-Þóris Saga
Sagas V	Gunnars Saga Gunnlaugs Saga Hænsna-Þóris Saga Hallfreðar Saga Harðar Saga Hávarðar Saga Heiðarvíga Saga Hrana Saga
Sagas VI	Kjalnesinga Saga Kormáks Saga Króka-Refs Saga Ljósvetninga Saga Reykdæla Saga Svarfdæla Saga Þórðar Saga
Sagas VII	Þorsteins Saga Hvíta Þorsteins Saga Síðu-Hallssonar Valla-Ljóts Saga Vatnsdæla Saga Víga-Glúms Saga Víglundar Saga Vopnfirðinga Saga Færeyinga Saga Ölkofra Saga Laxdæla Saga
Jón Trausti	Anna Frá Stóruborg Borgir
Jón Thoroddsen	Maður Og Kona
Þorgils Gjallanda	Upp Við Fossa Gamalt Og Nýtt
Smásögur I	Brúðardraugurinn Írafells - Móri Sagan Af Heljarslóðarorrustu Ferðasaga Þórðar Saga Geirmundarsonar Grímur Kaupmaður Deyr Hans Vöggur
Smásögur II	Kærleiksheimilið Brennivínshatturinn Gulrætur Í vinnunni Einræða Vordraumur Kvöld, nótt, morgunn

Tabla B.5: Documentos en Checo y Danés.
Fuente Checo https://www.databazeknih.cz/
Fuente Danés: Proyecto Gutemberg

Checo	
Fyodor Dostoevsky	Nétochka Nezvánova The House of the Dead
František Omelka	Blesky nad Beskydami Pověst Z Východu Vlci Proti
Karel Čapek	R.U.R Bílá nemoc
Michal Viewegh	Báječná léta pod psa
Mikhail Bulgakov	The Master and Margarita
Josef K. Šlejhar	Kuře Melancholik
Romain Rolland	Pierre et Luce
Henryk Sienkiewicz	Quo Vadis?

Danés	
Arne Magnussen	Den vidtudraabte Besættelse udi Thisted
Herman Bang	Det graa hus Enkens Søn Franz Pander Haabløse Slægter Fædra Hendes Højhed Stuk Det Hvide Hus Min Gamle Kammerat Tine Ved Vejen Ludvigsbakke
Sophus Bauditz	Absalons Brønd
Ingvor Bondesen	Skovstrup-Folk
Gudmund Nyeland Brandt	Stauder
Guy Boothby	Doktor Nikola
Sergius Stepniak	En Nihilist
Sophie Breum	Hyld Og Humle

Tabla B.6: Documentos en Finlandés y Húngaro. Fuente: Proyecto Gutemberg

Finlandés	
Jon Olof Åberg	Aina Adlercreutzin Sanansaattaja Erkki Ollikainen Kaarle Xii Vanginvartijana Karhu-Antin Anni Ja Spof'In Pistooli Rajalahden Torppa Sandelsin Urhea Joukko
Ari Aalto	Keltakukkia
Elise Polko	Soitannollisia satuja ja jutelmia
Elias Lönnrot	Elämä-kerrallisia piirteitä
Juhani Aho	Ensimmäiset Novellit Aatteiden Mies Helsinkiin Hellmannin Herra Esimerkin Vuoksi; Maailman Murjoma
Friedrich von Schiller	Kavaluus Ja Rakkau
August Ahlqvist	Muistelmia Matkoilta Venäjällä Vuosina Uusi Suomalainen Lukemisto

Húngaro	
Abonyi Árpád	A vörös Regina
Ambrus Zoltán Munkái	Álomvilág A Berzsenyi-Leányok Giroflé És Girofla
Anonimus	Vasárnapi Könyv
Mihály Babits	Timár Virgil Fia
István Bársony	Magyar Élet
Lenke Beniczkyné Bajza	Végzetes Tévedés
Elek Benedek	Édes Anyaföldem!

Tabla B.7: Documentos en Italiano y Latín. Fuente: Proyecto Gutemberg

Italiano	
Miguel de Cervantes	Don Quixote
Oscar Wilde	The Portrait of Dorian Gray
Jules Verne	TwentyThousand Leagues Under the Sea
Edoardo Gibbon	Storia Della Decadenza E Rovina Dell'Impero Romano
E. Masi *et al*	La Vita Italiana Nel Rinascimento
Carlo Botta	Storia D'Italia Dal 1789 Al 1814
Francesco Crispi	Politica Estera
Pietro Fanfani	Racconto Storico Del Secolo XIV
Lord Macaulay	Storia D'Inghilterra
Federico De Roberto	La Messa de Nozze

Latín	
Arcadius Avellanus	Mysterium Arcae Boulé Pericla Navarchi Magonis
Reginald Bainbridge Appleton	Puer Romanus Fabulae
Aurelius Augustinus Hipponensis	Confessiones
Daniel Bernoulli	Dissertatio inauguralis physico-medica de respiratione
Charles François Lhomond	Urbis Romae Viri Inlustres
Jacobus Berzelius	Nova Analysis Aquarum Medeviensium
Robert J. Breckinridge	Secreta Monita Societatis Jesu
Herman Boerhaave	De Usu Ratiocinii Mechanici in Medicina

Tabla B.8: Documentos en Hebreo. Fuente: https://benyehuda.org/

Hebreo	
Carl Ewald	Tales
Fyodor Dostoevsky	Crime and Punishment
Chaim M. Horowitz	Halachic Scriptures Of Geonim
Sam Vaknin	New Hebrew Short Fiction
Leon Tolstoy	The Cossacks
Nahum Sokolow	A Journey to Poland
Lytton Strachey	Queen Victoria
Jane Austen	Love and Pride
Anton Chekhov	Short Stories
Zevi Scharfstein	History of Chanukah in Israel in recent generations

C

Derivación de $m(k)$

En esta sección se va a derivar la expresión para $m(k)$, ecuación (3.8). Esta ecuación da cuenta del número de cajas en las cuales una palabra aparece, si dicha palabra se encuentra uniformemente disitribuida a lo largo de un documento.

Dado un documento de longitud N, lo dividimos en k partes, o cajas, cada una con un tamaño (número de palabras) ν. Entonces, la probabilidad conjunta de que una palabra aparezca $l_1, \ldots l_k$ veces está dada por

$$P(l_1, \ldots l_k) = \frac{\phi(l_1, \ldots l_k)}{\binom{\nu k}{f}}, \tag{C.1}$$

donde f es la frecuencia de la palabra, y $\phi(l_1, \ldots l_k)$ es la probabilidad que la palabra aparezca en una caja y no en ninguna otra. Esta probabilidad se puede escribir como

$$\phi(l_1, \ldots l_k) = \prod_{i=1}^{k} \binom{\nu}{l_i}, \tag{C.2}$$

donde l_i es la frecuencia de la palabra en cada una de las cajas. Nóstese que $\sum_i l_i = f$. Ahora bien, dada una palabra, se puede sumar sobre todas las maneras de escoger una caja de las k que tenemos en donde el término aparezca (es decir que $l > 0$ en esta caja y $l = 0$ en las demás), esto implica sumar $l = 1 \ldots \infty$

$$\phi(m|f) = \binom{k}{m} \sum_{l_1=1}^{\infty} \cdots \sum_{l_m=1}^{\infty} \prod_{i=1}^{m} \binom{\nu}{l_i} \delta\left(f - \sum_i l_i\right) \tag{C.3}$$

dado que esta suma es dificil de hacer, podemos usar la función generadora

$$\phi(m|\mu) = \sum_{f=0}^{\infty} e^{\mu f} \phi(m|f) = \binom{k}{m} \sum_{l_1=1}^{\infty} \cdots \sum_{l_m=1}^{\infty} \prod_{i=1}^{m} \binom{\nu}{l_i} e^{\mu l_i} \tag{C.4}$$

Usando el teorema del binomio se tiene que

$$\phi(m|\mu) = \binom{k}{m} \prod_{i=1}^{m} [(1 + e^{\mu}) - 1]^{\nu} = \binom{k}{m} z^m \tag{C.5}$$

donde $z = [(1 + e^{\mu}) - 1]^{\nu}]$. Entonces, el promedio $\langle m|f \rangle$, que sería el número promedio de cajas sería

$$\sum_{m=0}^{k} m\phi(m|\mu) = \sum_{m=0}^{k} m \binom{k}{m} z^m = z \frac{d}{dz} (1 + z)^k \tag{C.6}$$

$$\sum_{m=0}^{k} m\phi(m|\mu) = \sum_{m=0}^{k} m\binom{k}{m} z^m = zk(1+z)^{k-1}. \tag{C.7}$$

Usando $\nu k = N$ se tiene que

$$\sum_{m=0}^{k} m\phi(m|\mu) = k\left[(1+e^{\mu})^{N} - (1+e^{\mu})^{N-\nu}\right] \tag{C.8}$$

$$\sum_{m=0}^{k} m\phi(m|\mu) = \sum_{f=0}^{\infty} e^{\mu f}\left(\sum_{m} m\phi(m|f)\right) = \sum_{f=0}^{\infty} k\left[\binom{N}{f} - \binom{N-\nu}{f}\right] e^{\mu f} \tag{C.9}$$

De la anterior ecuación podemos reconocer los coeficientes como

$$m(k) = \sum_{m=0}^{k} m\phi(m|\mu) = k\left[\binom{N}{f} - \binom{N-\nu}{f}\right] \tag{C.10}$$

Dividiendo por $\binom{N}{f}$ tenemos

$$m(k) = k\left[1 - \frac{\binom{N-\nu}{f}}{\binom{N}{f}}\right] \tag{C.11}$$

Usando la definición del binomial tenemos que

$$m(k) = k\left[1 - \frac{(N-\nu)!(N-f)!}{N!(N-\nu-f)!}\right] \tag{C.12}$$

Usando $\gamma(n) = (n-1)!$ y la aproximación

$$\frac{\Gamma(n+a)}{\Gamma(n+b)} \sim n^{a-b} \tag{C.13}$$

Se puede ver que

$$m(k) \simeq k\left[1 - \left(\frac{N-\nu}{N}\right)^{f}\right] \tag{C.14}$$

de esta ecuación se sigue que

$$m(k) \simeq k\left[1 - \left(1 - \frac{1}{k}\right)^{f}\right] \tag{C.15}$$

Bibliografía

[1] Thomas C. Schelling. 'Models of Segregation'. In: *The American Economic Review* 59.2 (1969), pp. 488–493 (cited on page 1).

[2] Robert Axelrod. 'The Dissemination of Culture: A Model with Local Convergence and Global Polarization'. In: *Journal of Conflict Resolution* 41.2 (1997), pp. 203–226 (cited on page 1).

[3] Yaneer Bar-Yam. *Dynamics of Complex Systems*. 1st. Perseus Books, 1997 (cited on page 1).

[4] Scholtes Ingo. 'Understanding complex systems: When Big Data meets network science'. In: *it – Information Technology. Methods and Applications of Informatics and Information Technology* 57.4 (2015), pp. 252–256. DOI: https://doi.org/10.1515/itit-2015-0012 (cited on page 1).

[5] Chris Anderson. *The End Of Theory: The Data Deluge Makes The Scientific Method Obsolete*. Available at https://www.wired.com/2008/06/pb-theory/ (2019/09/08) (cited on page 2).

[6] Zanette D.H Montemurro M.A. *Complexity and Universality in the Long-Range Order of Words. In: Degli Esposti M., Altmann E., Pachet F. (eds) Creativity and Universality in Language. Lecture Notes in Morphogenesis*. 1st. Springer, Cham, 2016 (cited on pages 2, 4).

[7] Murray Gell-Mann and Merritt Ruhlen. 'The origin and evolution of word order'. In: *Proceedings of the National Academy of Sciences* 108.42 (2011), pp. 17290–17295. DOI: 10.1073/pnas.1113716108 (cited on pages 2, 13, 14).

[8] Kevin Knight, Beáta Megyesi, and Christiane Schaefer. 'The Secrets of the Copiale Cipher'. In: *Journal for Research into Freemasonry and Fraternalism* (May 2012). DOI: 10.1558/jrff.v2i2.314 (cited on page 2).

[9] Krishna R. Veeramah et al. 'Little genetic differentiation as assessed by uniparental markers in the presence of substantial language variation in peoples of the Cross River region of Nigeria'. In: *BMC Evolutionary Biology* 10.1 (2010), p. 92. DOI: 10.1186/1471-2148-10-92 (cited on page 2).

[10] I. Kontoyiannis et al. 'Nonparametric entropy estimation for stationary processes and random fields, with applications to English text'. In: *IEEE Transactions on Information Theory* 44.3 (May 1998), pp. 1319–1327. DOI: 10.1109/18.669425 (cited on pages 2, 4, 13).

[11] Haitao Liu and Wei Huang. 'Quantitative linguistics: state of the art, theories and methods'. In: *J Zhejiang Univ (Humanit Social Sci)* 43 (Jan. 2012), pp. 178–192 (cited on page 3).

[12] George Zipf. *Human behavior and the principle of least effort: an introduction to human ecology*. 1st Edition. Addison-Wesley Press, Cambridge, MA, 1949 (cited on pages 3, 13).

[13] Ricardo Baeza–Yates and Gonzalo Navarro. 'Block addressing indices for approximate text retrieval'. In: *Journal of the American Society for Information Science* 51.1 (2000), pp. 69–82. DOI: 10.1002/(SICI)1097-4571(2000)51:1<69::AID-ASI10>3.0.CO;2-C (cited on pages 3, 4, 13, 55).

[14] D. H. Zanette. 'Statistical Patterns in Written Language'. In: *arXiv:1412.3336* (2014) (cited on pages 3, 13, 23).

[15] W. Li. 'Random texts exhibit Zipf's-law-like word frequency distribution'. In: *IEEE Transactions on Information Theory* 38.6 (Nov. 1992), pp. 1842–1845. DOI: 10.1109/18.165464 (cited on pages 3, 12, 13, 15, 55).

[16] Steven T. Piantadosi. 'Zipf's word frequency law in natural language: A critical review and future directions'. In: *Psychonomic Bulletin & Review* 21.5 (Oct. 2014), pp. 1112–1130. DOI: 10.3758/s13423-014-0585-6 (cited on pages 3, 13).

[17] Isabel Moreno-Sánchez, Francesc Font-Clos, and Álvaro Corral. 'Large-Scale Analysis of Zipf's Law in English Texts'. In: *PLOS ONE* 11.1 (Jan. 2016), pp. 1–19. DOI: 10.1371/journal.pone.0147073 (cited on pages 3, 13).

[18] Xavier Gabaix. 'Zipf's Law for Cities: An Explanation'. In: *The Quarterly Journal of Economics* 114.3 (1999), pp. 739–767 (cited on page 3).

[19] David J. Schwab, Ilya Nemenman, and Pankaj Mehta. 'Zipf's Law and Criticality in Multivariate Data without Fine-Tuning'. In: *Phys. Rev. Lett.* 113 (6 Aug. 2014), p. 068102. DOI: 10.1103/PhysRevLett.113.068102 (cited on page 3).

[20] Leonid Boytsov. 'A Simple Derivation of the Heap's Law from the Generalized Zipf's Law'. In: *arXiv:1711.03066* abs/1711.03066 (2017). DOI: http://arxiv.org/abs/1711.03066 (cited on pages 4, 13).

[21] Ana María Palmero et al. 'Information theory reveals that individual birds do not alter song complexity when varying song length'. In: *Animal Behaviour* 87 (2014), pp. 153–163. DOI: https://doi.org/10.1016/j.anbehav.2013.10.026 (cited on page 4).

[22] Laurance R. Doyle et al. 'Information theory, animal communication, and the search for extraterrestrial intelligence'. In: *Acta Astronautica* 68.3 (2011). SETI Special Edition, pp. 406–417. DOI: https://doi.org/10.1016/j.actaastro.2009.11.018 (cited on page 4).

[23] Andreia Teixeira et al. 'Entropy Measures vs. Kolmogorov Complexity'. In: *Entropy* 13.3 (Mar. 2011), pp. 595–611. DOI: 10.3390/e13030595 (cited on page 4).

[24] Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto. 'Language Trees and Zipping'. In: *Phys. Rev. Lett.* 88 (4 Jan. 2002), p. 048702. DOI: 10.1103/PhysRevLett.88.048702 (cited on page 4).

[25] A. Kaitchenko. 'Algorithms for estimating information distance with application to bioinformatics and linguistics'. In: *Canadian Conference on Electrical and Computer Engineering 2004 (IEEE Cat. No.04CH37513)*. Vol. 4. May 2004, 2255–2258 Vol.4. DOI: 10.1109/CCECE.2004.1347695 (cited on pages 4, 5).

[26] J. Ziv and A. Lempel. 'Compression of individual sequences via variable-rate coding'. In: *IEEE Transactions on Information Theory* 24.5 (Sept. 1978), pp. 530–536. DOI: 10.1109/TIT.1978.1055934 (cited on page 4).

[27] Khalid Sayood. *Introduction to Data Compression*. Third Edition. Elsevier Inc, 2006 (cited on page 5).

[28] F. Behr et al. 'Estimating and comparing entropies across written natural languages using PPM compression'. In: *Data Compression Conference, 2003. Proceedings. DCC 2003.* Mar. 2003, pp. 416–. DOI: 10.1109/DCC.2003.1194035 (cited on page 5).

[29] Albert-László Barabási. *Network Science*. 1st. Cambridge University Press, 2016 (cited on pages 6, 14, 18, 19).

[30] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. 'Power-Law Distributions in Empirical Data'. In: *SIAM Review* 51.4 (2009), pp. 661–703. DOI: 10.1137/070710111 (cited on pages 9, 11, 16, 55).

[31] M D'Imperio. *The Voynich Manuscript: An Elegant Enigma*. 1st Edition. Laguna Hills, CA: Aegean Park Press, 1981 (cited on pages 12, 41–43, 45).

[32] Reinhard Köhler. 'Syntactic Structures: Properties and Interrelations'. In: *Journal of Quantitative Linguistics* 6.1 (1999), pp. 46–57. DOI: 10.1076/jqul.6.1.46.4137 (cited on page 13).

[33] Eduardo Altmann and Martin Gerlach. 'Statistical laws in linguistics'. In: *arXiv: 1502.03296* (Feb. 2015) (cited on page 13).

[34] Jake Ryland Williams et al. 'Zipf's law holds for phrases, not words'. In: *Scientific Reports* 5 (Aug. 2015). Article, p. 12209 (cited on page 13).

[35] Mark Newman. 'The power of design'. In: *Nature* 405.6785 (2000), pp. 412–413. DOI: 10.1038/35013189 (cited on page 13).

[36] R. Ferrer i Cancho. 'The variation of Zipf's law in human language'. In: *The European Physical Journal B - Condensed Matter and Complex Systems* 44.2 (Mar. 2005), pp. 249–257. DOI: 10.1140/epjb/e2005-00121-8 (cited on page 13).

[37] Peter Zörnig. 'Zipf's law for randomly generated frequencies: explicit tests for the goodness-of-fit'. In: *Journal of Statistical Computation and Simulation* 85.11 (2015), pp. 2202–2213. DOI: 10.1080/00949655.2014.925113 (cited on pages 13, 55).

[38] Gustav Herdan. *Type-token mathematics*. The Hague: Mouton, 1960 (cited on page 13).

[39] Harold Stanley Heaps. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, 1978 (cited on page 13).

[40] Leo Egghe. 'Untangling Herdans law and Heaps law: Mathematical and informetric arguments'. In: *Journal of the American Society for Information Science and Technology* 58.5 (2007), pp. 702–709. DOI: 10.1002/asi.20524 (cited on page 13).

[41] Ramon Ferrer i Cancho, Ricard V. Solé, and Reinhard Köhler. 'Patterns in syntactic dependency networks'. In: *Phys. Rev. E* 69 (5 May 2004), p. 051915. DOI: 10.1103/PhysRevE.69.051915 (cited on page 13).

[42] Camilo Akimushkin, Diego Raphael Amancio, and Osvaldo Novais Oliveira Jr. 'Text Authorship Identified Using the Dynamics of Word Co-Occurrence Networks'. In: *PLOS ONE* 12.1 (Jan. 2017), pp. 1–15. DOI: 10.1371/journal.pone.0170527 (cited on page 13).

[43] HaiTao Liu and Jin Cong. 'Language clustering with word co-occurrence networks based on parallel texts'. In: *Chinese Science Bulletin* 58.10 (Apr. 2013), pp. 1139–1144. DOI: 10.1007/s11434-013-5711-8 (cited on page 13).

[44] Y. Matsuo and M. Ishizuka. 'Keyword Extraction From A Single Document Using Word Co-Occurrence Statistical Information'. In: *International Journal on Artificial Intelligence Tools* 13.01 (2004), pp. 157–169. DOI: 10.1142/S0218213004001466 (cited on page 13).

[45] Pablo Gamallo, José Ramom Pichel, and Iñaki Alegria. 'From language identification to language distance'. In: *Physica A: Statistical Mechanics and its Applications* 484 (2017), pp. 152–162. DOI: https://doi.org/10.1016/j.physa.2017.05.011 (cited on pages 14, 21).

[46] Rafael E. Banchs. *Text Mining with MATLAB*. 1st Edition. Springer, 2013 (cited on pages 14, 29, 30, 36).

[47] Barry R. Chiswick and Paul W. Miller. 'Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages'. In: *Journal of Multilingual and Multicultural Development* 26.1 (2005), pp. 1–11. DOI: 10.1080/14790710508668395 (cited on page 14).

[48] Filippo Petroni and Maurizio Serva. 'Measures of lexical distance between languages'. In: *Physica A: Statistical Mechanics and its Applications* 389.11 (2010), pp. 2280–2283. DOI: https://doi.org/10.1016/j.physa.2010.02.004 (cited on page 14).

[49] Agata Fronczak, Piotr Fronczak, and Janusz A. Hołyst. 'Mean-field theory for clustering coefficients in Barabási-Albert networks'. In: *Phys. Rev. E* 68 (4 Oct. 2003), p. 046126. DOI: 10.1103/PhysRevE.68.046126 (cited on pages 14, 18).

[50] George A. Miller. 'Some Effects of Intermittent Silence'. In: *The American Journal of Psychology* 70.2 (1957), pp. 311–314 (cited on page 15).

[51] Ramon Ferrer-i-Cancho and Brita Elvevåg. 'Random Texts Do Not Exhibit the Real Zipf's Law-Like Rank Distribution'. In: *PLOS ONE* 5 (Mar. 2010), pp. 1–10. DOI: 10.1371/journal.pone.0009411 (cited on pages 16, 17, 55).

[52] Haspelmath Martin Hammarström Harald Forkel Robert. *Glottolog 4.0.* Available at https://glottolog.org/resource/languoid/id/roma1334 (2019/08/15) (cited on page 20).

[53] M Ortuño et al. 'Keyword detection in natural languages and DNA'. In: *EPL (Europhysics Letters)* 57 (Jan. 2007), p. 759. DOI: 10.1209/epl/i2002-00528-3 (cited on pages 23, 27, 56).

[54] T.P. Klammer, M. Schulz, and A.D. Volpe. *Analyzing English Grammar*. Pearson/Longman, 2006 (cited on page 23).

[55] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 'Scoring, term weighting, and the vector space model'. In: *Introduction to Information Retrieval*. Cambridge University Press, 2008, pp. 100–123. DOI: 10.1017/CBO9780511809071.007 (cited on page 29).

[56] MathWorks®. *Matlab Reference Documentation*. Available at https://www.mathworks.com/help/matlab/ref/smoothdata.html (2019/10/03) (cited on page 33).

[57] W. John Wilbur and Karl Sirotkin. 'The automatic identification of stop words'. In: *Journal of Information Science* 18.1 (1992), pp. 45–55. DOI: 10.1177/016555159201800106 (cited on page 36).

[58] Edgar Roman-Rangel and Stephane Marchand-Maillet. 'Stopwords Detection in Bag-of-Visual-Words: The Case of Retrieving Maya Hieroglyphs'. In: *New Trends in Image Analysis and Processing – ICIAP 2013*. Ed. by Alfredo Petrosino, Lucia Maddalena, and Pietro Pala. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 159–168 (cited on page 36).

[59] Marcelo A. Montemurro and Damián H. Zanette. 'Keywords and Co-Occurrence Patterns in the Voynich Manuscript: An Information-Theoretic Analysis'. In: *PLOS ONE* 8.6 (June 2013), pp. 1–9. DOI: 10.1371/journal.pone.0066344 (cited on pages 41, 43, 57).

[60] Diego R. Amancio et al. 'Probing the Statistical Properties of Unknown Texts: Application to the Voynich Manuscript'. In: *PLOS ONE* 8.7 (July 2013), pp. 1–10. DOI: 10.1371/journal.pone.0067310 (cited on pages 41, 43, 57).

[61] Kevin Knight Sravana Reddy. 'What We Know About The Voynich Manuscript'. In: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (2011), pp. 78–86 (cited on pages 41, 43).

[62] Gordon Rugg. 'An Elegant Hoax? A Possible Solution To The Voynich Manuscript'. In: *Cryptologia* 28.1 (2004), pp. 31–46. DOI: 10.1080/0161-110491892755 (cited on pages 41, 44).

[63] Andreas Schinner. 'The Voynich Manuscript: Evidence of the Hoax Hypothesis'. In: *Cryptologia* 31.2 (2007), pp. 95–107. DOI: 10.1080/01611190601133539 (cited on pages 41, 44).

[64] Deborah E. Harkness Raymond Clemens. *The Voynich Manuscript*. 1st Edition. Yale University Press, 2016 (cited on pages 41, 43).

[65] René Zandbergen. *The Radio-Carbon Dating of the Voynich Manuscript*. Available at http://www.voynich.nu/extra/carbon.html (2019/19/08) (cited on page 41).

[66] René Zandbergen. *History of the Voynich Manuscript*. Available at http://www.voynich.nu/history.html (2019/19/08) (cited on pages 42, 43).

[67] Peter Kokol et al. 'Computer and Natural Language Texts - A Comparison Based on Long-Range Correlations'. In: *JASIS* 50 (1999), pp. 1295–1303 (cited on page 44).

[68] René Zandbergen. *Analysis of the illustrations*. Available at http://www.voynich.nu/illustr.html (2019/19/08) (cited on page 45).

[69] René Zandbergen. *Text Analysis - Transliteration of the Text*. Available at http://www.voynich.nu/transcr.html (2019/19/08) (cited on page 48).

DR. VICTOR BARBA LÓPEZ
COORDINADOR DEL POSGRADO EN CIENCIAS
PRESENTE

Atendiendo a la solicitud para emitir DICTAMEN sobre la revisión de la TESIS titulada **"Análisis estadístico de textos"** que presenta el alumno **Diego Leonardo Espitia Cabrejo (5920150108)** para obtener el título de **Doctor en Ciencias.**

Nos permitimos informarle que nuestro voto es:

NOMBRE	DICTAMEN	FIRMA
Dr. Mariano López de Haro ICF-UNAM	Aprobado	
Dr. Markus Franziskus Muller CINC-UAEM	Aprobado	
Dr. Gustavo Martínez Mekler ICF-UNAM	Aprobado	
Dr. François Leyvraz ICF-UNAM	Aprobado	
Dr. Maximino Aldana ICF-UNAM	Aprobado	
Dr. Raúl Salgado García CINC-UAEM	Aprobado	
Dr. Hernán Larralde Ridaura ICF-UNAM	Aprobado	