Universidad Nacional Autónoma de México

Facultad de Ciencias

EVOLUCIÓN DEL LENGUAJE EN EL TIEMPO

QUE PARA OBTENER EL TÍTULO DE:
Físico

PRESENTA:

Sergio Ańgel Sánchez Chávez

DIRECTOR DE TESIS:
DR. CARLOS FRANCISCO PINEDA ZORRILLA

2016

Ciudad Universitaria, D. F.

1. Datos del alumno
Sergio Ángel
Sánchez Chávez
55 23 32 11 97
Universidad Nacional Autónoma de México
Facultad de Ciencias
Física
407026472

2. Datos del tutor
Dr.
Carlos Francisco
Pineda Zorrilla

3. Datos del sinodal 1
Dr.
Adonis Germinal
Cocho Gil

4. Datos del sinodal 2
Dr.
David Phillip
Sanders

5. Datos del sinodal 3
Dr.
Carlos
Gershenson García

6. Datos del sinodal 4
Dr.
Ruben Yvan Maarten
Fossion

7. Datos del trabajo escrito
Evolución del lenguaje en el tiempo
66 p.
2016

Agradecimientos

Durante el efímero tránsito que tenemos en este mundo, establecemos distintos lazos afectivos con las personas que conocemos. Algunos de estos lazos se rompen en poco tiempo mientras que otros perduran hasta el final, pero (casi) todos ellos dejan una huella en nosotros.

En primer lugar agradezco a mis padres, Lucina y Ángel, por tanto que me han dado. A mi hermana Sonia, que bien o mal siempre hemos estado juntos. Y a mi segunda madre, mi madrina Mercedes.

Especialmente quiero agradecer a los doctores Carlos Pineda, Carlos Gershenson, Germinal Cocho y Jorge Flores. Por lo que me han apoyado y lo mucho que he aprendido con ellos. Gracias ha ustedes he podido desarrollar este tema.

Por último pero no menos importante me gustaría recordar a mis amigos, desde los que conservo de la infancia hasta quienes sigo viendo con frecuencia. Disculpen si no menciono a cada uno por su nombre, ustedes saben quienes son. Gracias, por que han hecho más divertido este camino.

Gracias UNAM.

Índice general

Resumen

El estudio sobre lenguaje es bastante amplio y diverso, desde los procesos neurofisiológicos que están involucrados en su adquisición hasta la manera de interactuar de distintos idiomas en una comunidad. Nuestro análisis se limita a la parte escrita del lenguaje, involucra la frecuencia de las palabras usadas en los libros en un determinado año y como cambia su frecuencia de uso a lo largo del tiempo.
Hemos propuesto una nueva medida a la que llamamos *diversidad de rango*, que logra reflejar estos cambios de frecuencia de uso. Lo hemos comprobado en 6 idiomas diferentes como lo son: español, inglés, francés, italiano, alemán y ruso. Además proponemos un modelo, basado en caminatas aleatorias gaussianas que reproduce el cambio de uso de las palabras a través del tiempo. Y que tiene una *diversidad de rango* similar a la de los idiomas.

Capítulo 1

Distribuciones de frecuencias en palabras

Los primeros estudios textos con el uso de herramientas estadísticas fueron publicados en 1932 por George Zipf, donde analizó la frecuencia de aparición de palabras en un texto [**?**]. Utilizó como texto de análisis *Ulysses* de James Joyce[1]. Los resultados que obtuvo le fascinaron porque él tenía cierta obsesión con los patrones matemáticos [**?**]. Encontró que la primeras tres palabras más repetidas del inglés eran *the*, *of* y *and*, además que la frecuencia de palabras seguía una relación $y \sim \frac{1}{x^a}$, donde y es el número de veces que aparece una palabras en el texto mientras que x es la posición que ocupa una palabra en el texto, con un exponente a cercano a uno. Cabe señalar que Zipf, previamente había realizado un análisis de algunos extractos de textos chinos; haciendo el conteo sobre la repetición de las sílabas en esos textos [**?**]. Con una muestra de cerca de 20,000; y observó una relación lineal en escala log-log:

$$P_k \sim \frac{1}{k^a}. \tag{1.0.1}$$

Después que Zipf mostrara sus resultados, se han encontrado otros fenómenos, no relacionados con la lingüística, que pueden ser razonablemente aproximados por una distribución del tipo potencial. Entre estos fenómenos podemos nombrar: población de las ciudades, número de llamadas telefónicas, distribución de las notas en una melodía, tamaño de los archivos de computadora, el número de visitas a las página web, el número de citas a artículos científicos, etcétera [**?**].

[1] El conteo de las palabras lo delegó a sus estudiantes. Recordemos que Rutherford hizo lo mismo en el conteo de partículas α.

Hoy en día con una computadora es fácil obtener la distribución de las frecuencias por palabras en un texto, hagámoslo con cuatro grandes obras de la literatura como son *Don Quijote de la Mancha, Los miserables, La divina comedia* y *Ulises*[2]; y obtengamos el **ranking** de palabras, donde la primera posición, o **rango**, la ocupa la palabras que más repite en el texto, la siguiente palabra con mayor repetición ocupa la posición número dos y así sucesivamente. El resultado se muestra en la siguiente gráfica 1.1, que en escala log-log sigue una tendencia lineal lo que nos indica una ley de potencias.

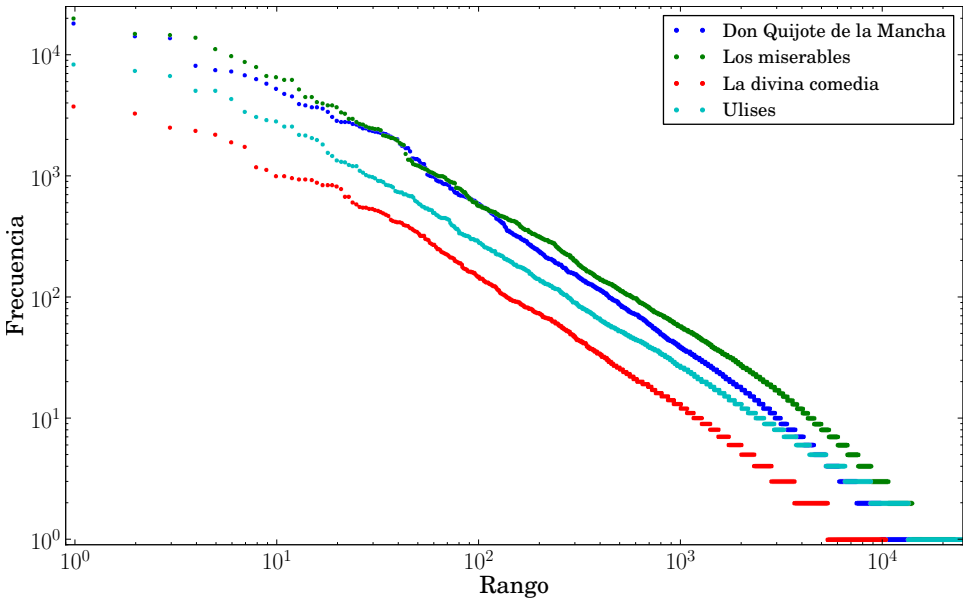

Figura 1.1: Ranking de la frecuencia de la palabras en 4 obras de la literatura universal. Cada una de las diferentes palabras en cada libro esta representada por un punto. La frecuencia de cada una, i.e., el número de veces que se repite cada palabra, esta sobre el eje y. En el eje x el rango, o la posición en el ranking, que ocupa cada palabra. Ambos ejes están en escala logarítmica.

[2]Lo más conveniente es descargarlo de *Project Gutenberg*, una página de acceso abierto que ha digitalizado más de 40,000 libros (principalmente en inglés).

1.1. La base de datos, Google Books

Desde que Zipf publicó sus resultados hasta nuestra fecha, la tecnología ha avanzado a tal grado que contar las palabras de un libro sea tan sencillo como descargar el archivo en texto plano y programar unas cuantas líneas de código. En caso de que el texto no se encuentre digitalmente, sólo hay que escanearlo y con un programa de reconocimiento de caracteres extraer el texto. Ignoraremos los errores que aparecen durante este procedimiento, pues nuestro trabajo se basa en estadística sobre un número grande de datos y con el desarrollo diario de la tecnología estos errores de escaneo disminuyen.

Del escaneo de libros se ha encargado Google, específicamente la división de Google Books bajo el proyecto llamado: N-gram Viewer [**?**]. Google ha escaneado el 4 % de todos los libros publicados hasta el 2009, en varios idiomas: español, francés, alemán, ruso e inglés. Este último idioma también lo podemos encontrar separado en inglés británico y americano. En el 2012 la base se actualizó y se agregaron nuevos idiomas como el chino, italiano y hebreo.

El corpus o total de libros escaneados es 5,195,769 obtenidos de 40 universidades alrededor del mundo más la contribución de algunas editoriales. El conjunto total de datos esta divido por n-gramas, donde un **n-grama** es un secuencia continua de palabras o signos. Google tiene disponible al público las bases de datos de 1 a 5 n-gramas.[3]

Para este trabajo nos dimos a la tarea de descargar, ordenar y limpiar de erratas las bases de datos de inglés, español, francés, alemán, ruso e italiano. Seleccionamos estos idiomas por ser los que tenían las bases de datos más grandes y también porque conocemos lo suficiente los idiomas para entender mínimamente las bases de datos. En cada base de datos tratamos de remover la mayor cantidad de palabras que no pertenecen a cada idioma. La forma en que limpiamos la base elimina totalmente las palabras que contienen caracteres que no pertenecen al alfabeto del idioma en cuestión. Para dejarlo en claro pongamos un ejemplo: en el idioma inglés no existe el carácter *ñ* por lo tanto la palabra *baño* fue eliminada. Algunas palabras erróneas puedes seguir apareciendo en el corpus. Esto puede ser debido a errores en el reconocimiento de caracteres (OCR), errores al clasificar a que idioma corresponden los libros, o simplemente la inclusión de una palabra foránea en el texto.

El proceso de depuración anterior se realizó en cada uno de los seis idiomas en consideración. La lista de letras permitidas se puede ver en el apéndice.

[3] `http://storage.googleapis.com/books/ngrams/books/datasetsv2.html`

Los datos que nos sirven para trabajar son la palabra y el número de veces que se repite en todo el conjunto de libros escaneados y clasificados para cada idioma. La figura 1.2 muestra el número total de palabras por año para cada lenguaje. Como se ve, el número de palabras aumenta al transcurrir el tiempo, a pesar de que existen fluctuaciones año con año.

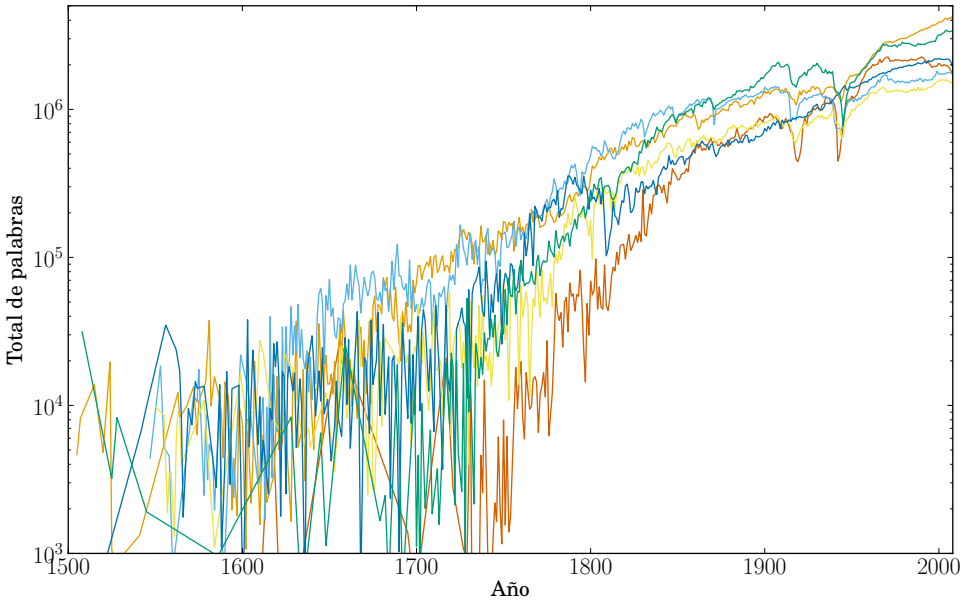

Figura 1.2: Cantidad de palabras a través de los años en cada una de las distintas bases de datos que Google para cada idioma. El código de colores es el siguiente: inglés (■), francés (■), alemán (■), italiano (■), español (■) y ruso (■). Este código de colores se usará para los idiomas a lo largo de todo el trabajo.

De esta gráfica podemos extraer cierta información y relacionarla con eventos históricos. La base de datos más robusta es la de inglés. El tamaño del corpus de inglés y del francés al inicio es similar y del orden de 10^8 1-gramas. Esto es atribuible al dominio académico que Francia tenía en Europa durante el siglo XIX y que después Estados Unidos ha tomado. Las demás bases, si bien comienzan con diferentes tamaños, al paso de los años tienden ser similares y con un tamaño del orden de 10^9 palabras. Notorio también son los dos picos invertidos entre 1900-1950, que corresponden con las dos guerras mundiales, muy pronunciados en el caso del idioma ruso. En este idioma también es notorio el incremento en el número de palabras entre la segunda guerra mundial y la caída de la URSS. Por lo menos en esta gráfica para

el español no notamos los efectos de las guerras, o fue mínimo el efecto de éstas, pues carece de los picos pronunciados como los demás idiomas.

1.2. Ley de Zipf

Aunque es común referirse a (1.0.1) como una "ley", en realidad estamos tratando con un modelo basado en una distribución de probabilidad. En este caso tenemos que puede ser una distribución de probabilidad continua o discreta. Técnicamente es más fácil trabajar con distribuciones de probabilidad continuas. Aunque los rankings de origen son discretos se suele utilizar distribuciones continuas cuando la cantidad de datos es significativa [**?**]. En nuestro caso trataremos los datos como una distribución de probabilidad continua.
Se define la distribución de probabilidad tipo Zipf como:

$$p(k) = Ck^a, \tag{1.2.1}$$

donde $p(k)$ es un función de densidad de probabilidad, con C como la constante de normalización, válida para $k \geq k_{min}$, ya que diverge en $k = 0$. Esta constante de normalización está dada por, $C = (a-1)k_{min}^{a-1}$.
La distribución (1.2.1), se le puede sacar otras propiedades como momentos sin embargo estas no serán necesarias para el análisis en este trabajo.
La distribución (1.2.1), al igual que la distribución normal, también posee momentos bien definidos que dependen de a. También es común oír que una distribución tipo Zipf es una distribución libre de escalas porque no importa a que escala nos fijemos el comportamiento es el mismo. Los datos pueden ser continuos o discretos y en ambos casos (1.2.1) se pueden manejar funciones de densidad de probabilidad continua o discreta. Técnicamente es más fácil trabajar la función de densidad de probabilidad para una variable continua. Si se tiene que los datos son discretos pero k es muy grande, como en nuestro caso, podemos usar la densidad de probabilidad continua. Un desarrollo más amplio puede encontrarse en el trabajo de M.E.J Newman [**?**].
Al ajustar la distribución Zipf (sin normalizar) al idioma inglés para el año 2000, obtenemos la gráfica 1.3, que a simple vista vemos que es una tosca aproximación para la precisión estadística de la base de datos. Para justificar ésta discrepancia, principalmente en las colas, se han hecho diversas suposiciones. Una de ellas es que la cantidad de datos es insuficiente, otra que las distribuciones a usar deben ser finitas (tener una k_{max}) ya que las palabras usadas no son infinitas [**?**].

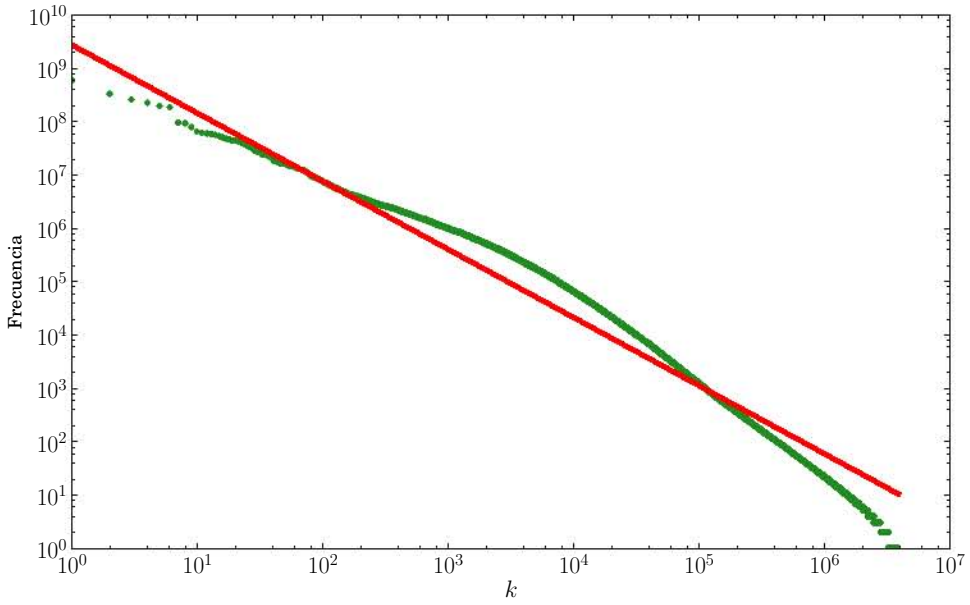

Figura 1.3: Frecuencia de palabras del idioma inglés (■) para el año 2000 y ajuste de la distribución Zipf (■).

1.3. Otras distribuciones

En la figura 1.3 notamos que la frecuencia de palabras no sigue exactamente una línea recta en escala logarítmica. La discrepancia es visible entre el rango 10^3 y 10^4. La distribución (1.3.1a) (línea roja) atraviesa los datos (línea verde) en dos puntos.

Frecuentemente se dice que la "cola"(rangos altos) de la distribución es pesada cuando en los rangos altos la frecuencia de los últimos elementos disminuye más rápidamente. Otros tipos de distribuciones logran ajustar mejor la cola al incorporar el término e^{-bk}.

Comparamos la ley de Zipf con otras cuatro distribuciones, todas ellas con

el término común $1/k^a$. Las distribuciones son las siguientes:

$$m_1(k) = \mathcal{N}_1 \frac{1}{k^a}, \tag{1.3.1a}$$

$$m_2(k) = \mathcal{N}_2 \frac{e^{-bk}}{k^a}, \tag{1.3.1b}$$

$$m_3(k) = \mathcal{N}_3 \frac{(n-k)^d}{k^a}, \tag{1.3.1c}$$

$$m_4(k) = \mathcal{N}_4 \frac{(n-k)^d e^{-bk}}{k^a}, \tag{1.3.1d}$$

$$m_5(k) = \mathcal{N}_5 \begin{cases} \frac{1}{k^a} & k \leq k_c \\ \frac{k_c^{a'-a}}{k^{a'}} & k > k_c \end{cases}. \tag{1.3.1e}$$

Las distribuciones (1.3.1a), (1.3.1b),(1.3.1c) y (1.3.1d) se les puede asociar un origen común. Es más, los distribución (1.3.1b) y (1.3.1c) son casos particulares de (1.3.1d). A estos últimos también se les conoce como distribución tipo γ y distribución tipo β, respectivamente. Álvarez-Martínez et. al desarrollan distribuciones [**?**] que da origen a las ecuaciones (1.3.1b),(1.3.1c) y (1.3.1d).

Los distribución (1.3.1c) y (1.3.1d) tienen una restricción que depende del número de elementos de la distribución. Es decir, se pide conocer el número total n de elementos que forman la distribución. Esto se refleja en el factor $(n-k)^d$.

Altmann et.al [**?**] utilizaron una variación a la ley de Zipf, la distribución (1.3.1e) con $a = 1$. Esta distribución les permite dividir las palabras en dos secciones, la primera ($k \leq k_c$) la relacionan con el núcleo de palabras más usadas y la segunda parte ($k > k_c$) con las palabras menos usadas de cada idioma.

El ajuste de las distintas distribuciones (1.3.1a)-(1.3.1e), se realizó por medio de mínimos cuadrados en el espacio log-log. Es posible ver reflejado en los parámetros de los distribución, ciertas cosas como la falta de datos en algunos idiomas, especialmente para años anteriores a 1800. Esto es posible verlo en las gráficas ya que antes de esa fecha casi todos los parámetros varían enormemente. Ésta es la principal razón de utilizar los datos de 1800 en adelante, para los subsecuentes análisis.

En algunos casos se puede asociar los picos o anomalías a cierto hechos históricos, como la primera y segunda guerra mundial, que se refleja en el parámetro a, independientemente del distribuciones. El inglés y el francés se ven como los idiomas que presentan menos variaciones en los parámetros.

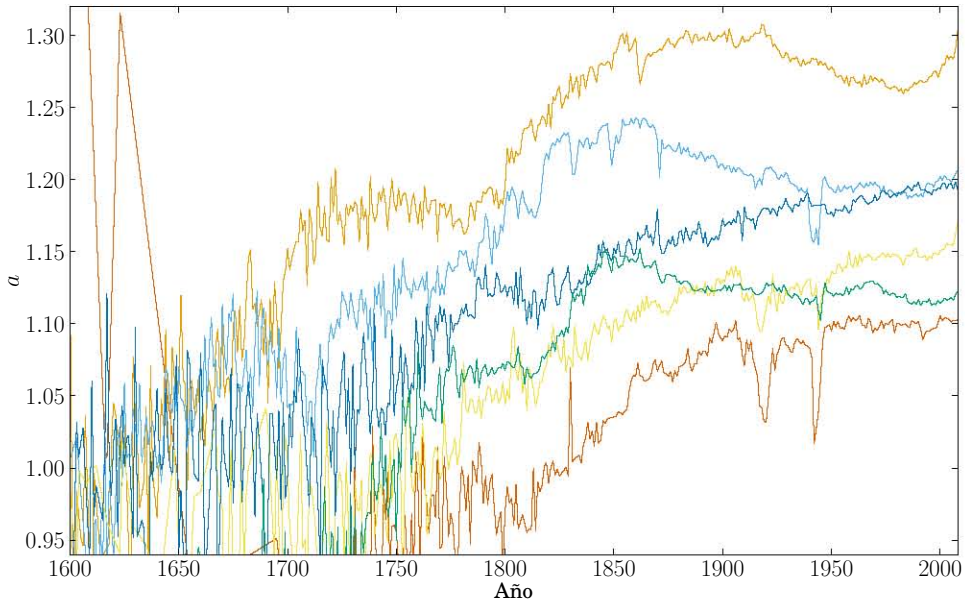

Figura 1.4: Parámetros de las distribuciones (1.3.1a), a través del tiempo. El código de colores de esta figura y las cuatro siguientes es el mismo que el de la fig. 1.2

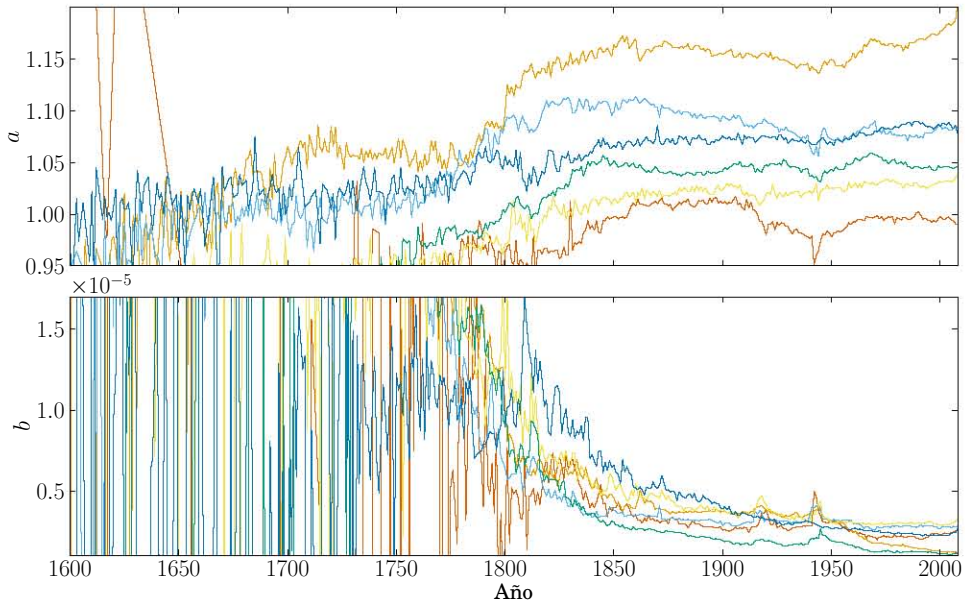

Figura 1.5: Parámetros de las distribuciones (1.3.1b), a través del tiempo.

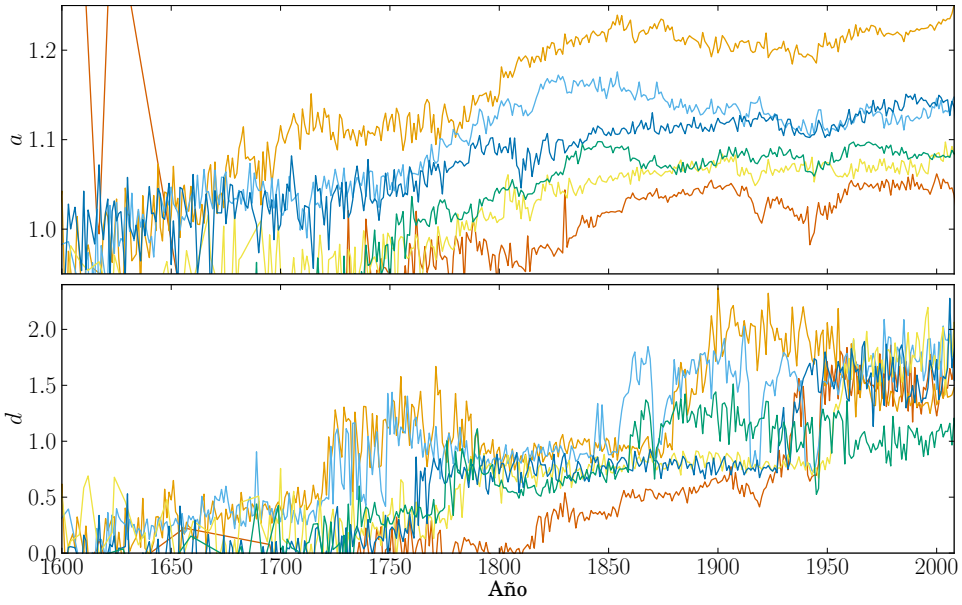

Figura 1.6: Parámetros del distribuciones (1.3.1c), a través del tiempo.

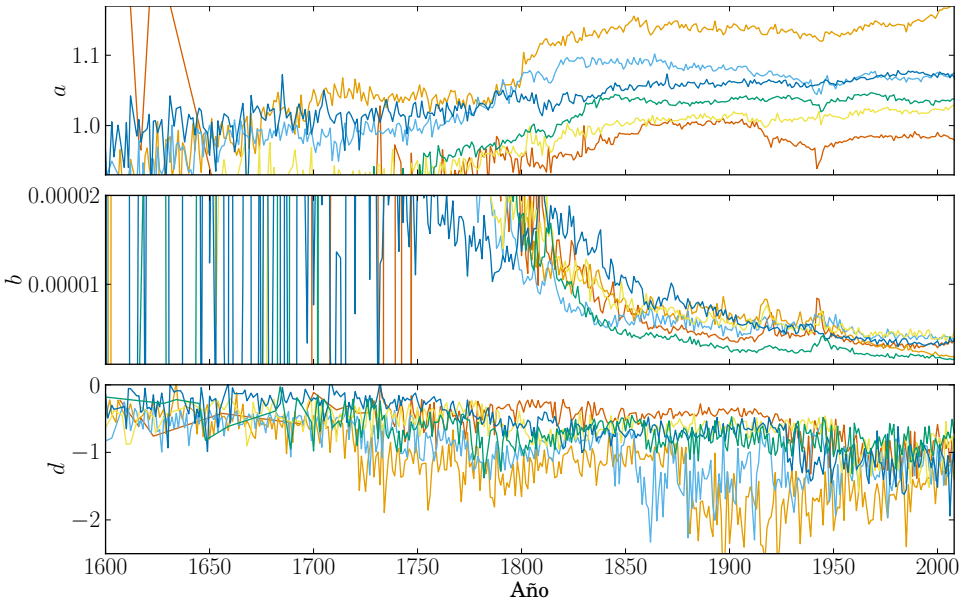

Figura 1.7: Parámetros de las distribuciones (1.3.1d), a través del tiempo.

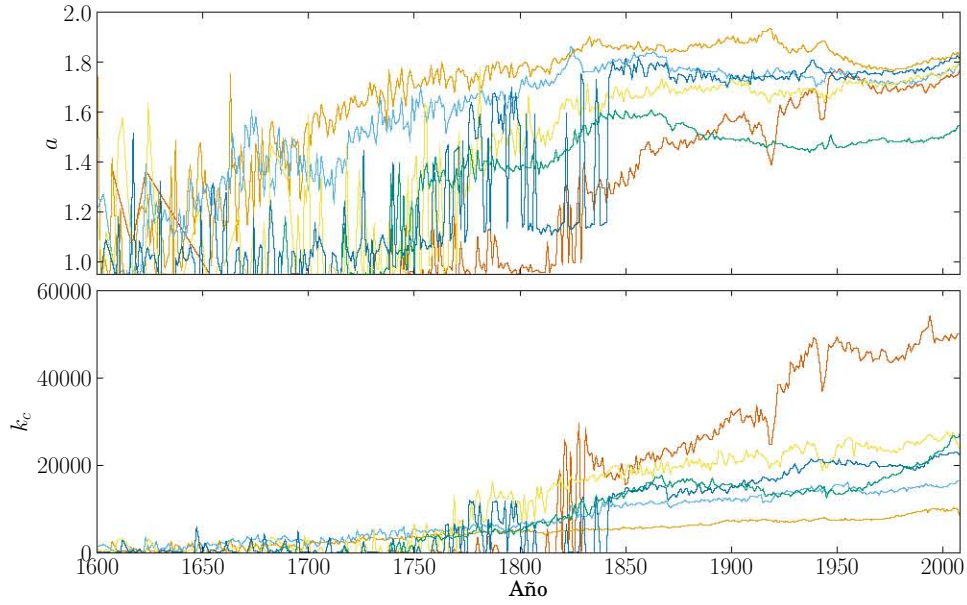

Figura 1.8: Parámetros de las distribuciones (1.3.1e), a través del tiempo.

1.4. Bondad de los ajustes

La bondad del ajuste cuantifica que tan bien una distribución propuesta ajusta datos experimentales. En nuestro caso usamos la prueba de χ^2 para ver que tanto se ajustan las diferentes distribuciones a los datos de palabras en cada lenguaje y tener algún indicio de cual puede ser mejor. El motivo para usar esta prueba, es que es comúnmente usada en la literatura relacionada con este proyecto [**?**].

Cabe señalar que el resultado de la prueba no es un criterio absoluto para tomar partido sobre una distribución respecto a otras, ya que al aumentar el número de parámetros la diferencia entre la distribución y los datos tiende a disminuir. Sin embargo es deseable que la distribución más apropiada se capaz de describir el fenómeno con la menor cantidad de parámetros.

Veamos los diferencias normalizadas, $\epsilon(k)$, entre las distribuciones (1.3.1) y el ranking de palabras para el año 2000 del idioma inglés. La fig. 1.9 la obtuvimos al calcular,

$$\epsilon_i(k) = \frac{m_i(k) - x_k}{x_k}, \tag{1.4.1}$$

donde x_k es la frecuencia de la palabra en el rango k-ésimo, $m_i(k)$ es la frecuencia predicha por la distribución i en el rango k-ésimo, e i puede ser

cualquiera de las 5 distribuciones en (1.3.1).

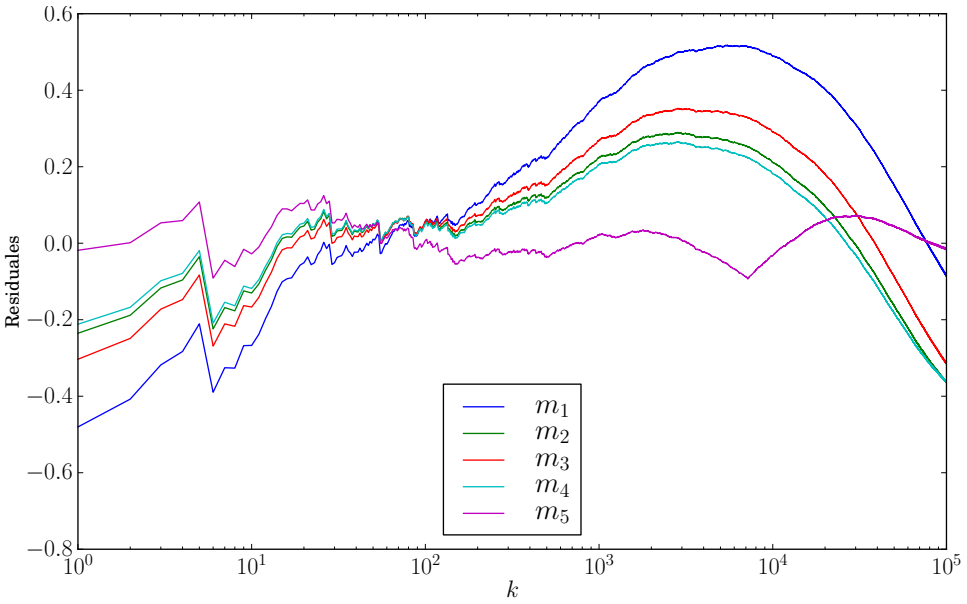

Figura 1.9: Diferencia entre la distribución de palabras en el idioma inglés para el año 2000 y cada uno de los diferentes distribuciones de (1.3.1). Ver eq. (1.4.1).

Si $\epsilon_i(k) = 0$ para toda k y alguna i, significaría que la distribución i describe perfectamente los datos. De la gráfica anterior, podemos ver que la distribución cuya diferencia es la menor de las 5 usadas es (1.3.1e), mientras que el peor ajuste es (1.3.1a). La forma de la gráfica es muy simular para las diferencias de (1.3.1a)-(1.3.1d), a excepción de la diferencia de la distribución (1.3.1e). Al parecer todas las distribuciones son muy cercanas a un punto que se encuentra alrededor de $k = 70$. También es llamativo el hecho de que los cinco distribución coincidan casi totalmente entre los rangos $40 < k < 50$.
Lo más deseado es que la desviación de los datos con la distribución propuesta sea $O_i - E_i$ sea pequeña en caso contrario ni de coña estaríamos pensando que se trata de la distribución pensada. Sin embargo el problema es cuantificar que tan "pequeño" es lo pequeño. Lo que se espera es que el valor de O_i caiga alrededor de E_i con una desviación estándar del orden de $\sqrt{E_i}$. Bajo las consideraciones anteriores tenemos un estimado de las desviaciones de la distribución con respecto a los datos que para cada punto sería $O_i - E_i \sim \sqrt{E_i}.$, es decir tanto la diferencia como la fluctuación deben ser

similares y su razón del orden de 1. Como esto se hace para cada punto, la suma sobre los errores los n puntos de la distribución puede presentar sesgos, por lo tanto es mejor tomar el cuadrado tanto de la diferencia como de la fluctuación. Así, definimos

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_i - E_i)^2}{E_i}. \tag{1.4.2}$$

Si suponemos que los datos observados corresponde a una distribución propuesta, como lo hemos hecho hasta ahora, la diferencia y la fluctuación de cada punto deben de ser $\frac{O_i - E_i}{\sqrt{E_i}} \sim 1$, y por lo tanto la suma total para todos los puntos en cuestión será del orden de n. Entre mejor sea el ajuste $\chi^2 \lesssim n$, de lo contrario, si $\chi^2 \gg n$ podemos suponer que la distribución propuesta no es la adecuada. Cabe señalar que entre más grande sea el conjunto de datos, la comparación punto a punto puede inducir sesgos; para evitarlo se agrupan los datos en clases;[4] intervalos consecutivos y que además no se traslapan unos con otros.

Una forma más precisa, y más usada, de comparar la discrepancia de los datos con las distribuciones, (1.3.1), es obtener el valor p de cada una de ellas. Lo primero es proponer una hipótesis que deseamos probar a la que llamaremos hipótesis nula, H_0. A las demás se les llamará hipótesis alternativa H_1. Esta hipótesis debe ser acerca de los parámetros de la distribución. Entonces el valor p se define como la probabilidad de obtener cierto valor de la distribución de probabilidad normalizada de χ^2, suponiendo la hipótesis nula como cierta. Este valor p se compara con un valor de significancia α dado, para poder rechazar o aceptar la hipótesis nula. Lo común suele ser tomar $\alpha = 0.05$ o $\alpha = 0.01$. Si el valor p es mucho más chico que α se dice que tenemos evidencia significativa contra la hipótesis nula, H_0, y por lo tanto debemos inclinarnos por la hipótesis alternativa, H_1. Por el contrario, si el valor p es mucho mayor que α decimos que no tenemos evidencia significativa contra la hipótesis nula y podemos inclinarnos por ella. Podemos escoger mal la hipótesis y caer en alguno de los siguientes errores que se muestran en la tabla 1.4.1. Así es que debemos interpretar bien el resultado.

[4]Esta es la traducción, más común, de lo que inglés se llaman *bins*.

Decisión	H_0 es verdadera	H_1 es falsa
No rechazar H_0	–	Error tipo II
Rechazar H_0	Error tipo I	–

Tabla 1.4.1: Tipos de errores que se comenten al aceptar o rechazar la hipótesis nula.

En nuestro caso la hipótesis nula es: que los valores de los parámetros calculados son los correctos (para cada año y para cada modelo) y tanto la distribución de los datos corresponden a ese modelo. Así que para obtener el valor p, tenemos dos opciones: miramos las tablas de la distribución χ^2 para n, donde es el número de grados de libertad que tomamos en cuenta, y el valor obtenido de χ^2 o lo calculamos directamente con una función de un numpy, Mathematica, Maple, etc. En todos los casos el valor p que obtuvimos fue del orden de 10^{-20}, es decir, prácticamente cero y mucho menor que un nivel de significancia, $\alpha = 0.01$. Por lo tanto tenemos evidencia contra la hipótesis nula y por lo tanto no tenemos fuertes elementos para decir que los datos corresponden a alguna de las distribuciones mencionadas en (1.3.1).

Capítulo 2

Diversidad de rango

En la capítulo anterior comprobamos que ninguna de las distribuciones propuestas (1.3.1), logran ajustarse a los datos, de manera totalmente convincente. La que mejor lo hace, doble Zipf, propone un núcleo de palabras para cada lenguaje que va desde 7,000 hasta 60,000 [**?**]. Esto no está en concordancia con lo que los lingüistas han propuesto, por lo menos para el caso del idioma inglés y español [**?**, **?**, **?**, **?**]. Su núcleo es del orden de 2,000 palabras. En este capítulo se propondrá y explicará una nueva medida que está más apegada a los números obtenidos por los lingüistas.

2.1. Clasificación de las palabras

Uno de los muchos procedimientos que se siguen al abordar un problema, es dividirlo en unidades más pequeñas para su estudio, tratar de ubicar las partes fundamentales, entenderlas y en base a ello analizar el problema original. Si esto no da resultado se pueden agrupar esta partes fundamentales por ciertas características o estructuras, encontrar nuevas propiedades y abordar el problema nuevamente. Así podemos ir escalando en los distintos niveles, hasta encontrar la solución al problema.
Hasta ahora hemos tratado con conjuntos enormes de palabras que están agrupados en idiomas. Esta separación no la hemos hecho nosotros, aunque sí la hemos dado por buena. Dentro de estos conjuntos, la unidad básica es la palabra. Así desprovista de contexto, como es nuestro caso, podemos buscar características por las que podamos agruparlas en subconjuntos más pequeños que el idioma; pero más grandes que sólo palabras. Las características que busquemos para formar subconjuntos pueden ser muchas y no todas nos serán de ayuda.

Una manera sencilla de clasificarlas puede ser por la letra con la que inician o por el número de letras que contienen, otra por el número de silabas. Estas clasificaciones, aunque son sencillas, no nos son útiles. Agruparlas por temas también es posible: palabras médicas, palabras matemáticas, palabras religiosas, etcétera. También podríamos clasificar las palabras por artículos, adjetivos, adverbios, etc. Mientras que las primeras clasificaciones son fáciles de algoritmizar y programar en una computadora, las segundas son mucho más difíciles de implementar en un computadora y las terceras están a medio camino entre las primeras y las segundas.
La clasificación que emplearemos fue propuesta por C. C. Fries [**?**]. Las palabras se pueden dividir en 2 tipos: palabras que expresan una idea o contenido en sí misma; y palabras que son auxiliares en la construcción del mensaje que se quiera dar [**?**]. A las primeras se les llama palabras de contenido y a las segundas palabras funcionales. En el diccionario Merriam-Webster [**?**] las podemos encontrar definidas como las palabras que son usadas principalmente para mostrar relaciones gramaticales entre otras palabras.
La lista de lo que se entiende por palabras funcionales, es para fines prácticos, pequeña, considerando el número de palabras que aparecen en cualquier diccionario. A continuación presentaremos una lista de lo que se consideran palabras funcionales.

- Artículos
- Pronombres
- Conjunciones
- Verbos auxiliares
- Interjecciones
- Partículas
- Preposiciones

Puede presentarse que alguna palabra sea usada como funcional en un contexto y en otro contexto como palabra de contenido. Por ejemplo, tenemos las conjugaciones del verbo *haber* en español, suelen usarse como verbos auxiliares de los tiempos compuestos; pero también se puede usar como verbo propio. En la base de Google, al ser 1-gramas no tenemos manera alguna de diferenciar qué porcentaje de es debido al uso funcional o al de contenido. Necesitamos el contexto en el que se usa la palabra para poder discriminar qué tipo de palabra es.

De la lista anterior vemos que el universo de palabras funcionales es realmente pequeño. Un cálculo rápido nos permite contar alrededor de 200 palabras. Tomando en cuenta esto, no es para nada descabellado decir que si tomamos un diccionario y seleccionamos una palabra al azar (ignorando el texto de la definición de cada palabra), la palabra elegida seguramente será de contenido. Ésto no se debe confundir con elegir una palabra al azar de un texto cualquiera. En este caso la probabilidad de elegir una palabra funcional no es baja, ya que, al ser un texto estructurado, se apoya en pronombres, conjunciones, artículos, etc. Este hecho se ve reflejado cuando tomamos las primeras 20 palabras más usadas en cualquier idioma (de los 6 que seleccionamos para este trabajo). La gran mayoría de ellas es funcional, incluso en el idioma ruso que carece de artículos.
En las siguientes dos tablas colocamos, como ejemplo, las primeras 20 palabras del inglés, tabla 2.1.1, y del español, tabla 2.1.2, en el año 2000. Como puede verse la gran mayoría son palabras de contenido, tanto en inglés como en español.

the (1)	a (6)	it (11)	be (16)
of (2)	is (7)	with (12)	by (17)
and (3)	that (8)	was (13)	i (18)
to (4)	for (9)	on (14)	are (19)
in (5)	as (10)	not (15)	this (20)

Tabla 2.1.1: Las 20 palabras más usadas del idioma inglés en el año 2000. Entre paréntesis está indicada la posición en el ranking.

de (1)	que (6)	las (11)	no (16)
la (2)	a (7)	por (12)	para (17)
en (3)	los (8)	un (13)	su (18)
y (4)	del (9)	con (14)	es (19)
el (5)	se (10)	una (15)	al (20)

Tabla 2.1.2: Las 20 palabras más usadas del idioma español en el año 2000. Entre paréntesis está indicada la posición en el ranking.

Para este trabajo utilizamos el inglés como idioma base. Elaboramos una lista con todas las palabras funcionales en inglés. Con esta lista podemos separar a las palabras funcionales de las de contenido. En el caso de pal-

abras que tienen más de un significado con uno de ellos funcional y otro de contenido, elegimos la palabra como funcional. También advertimos que no aplicamos la lematización sobre el conjunto de palabras, es decir, las inflexiones de una palabra son contabilizadas como palabras diferentes. Por ejemplo, "volaba", "volaste" y "volar" son contabilizadas como palabras distintas.

En la gráfica 2.1 comparamos las primeras 20 palabras, tanto de contenido como funcionales, en inglés con las respectivas palabras para los demás idiomas. Usamos Google-Translate para traducir cada palabra en inglés a cada uno de los 5 idiomas restantes. Formamos una lista de las palabras funcionales en inglés y con esta construimos las otras cinco listas de palabras funcionales para los demás idiomas. Ciertamente aparecieron algunos errores de traducción, los cuales fueron subsanados manualmente al revisar cada una de las listas. Las líneas continuas corresponden al grado de correspondencia entre las primeras 20 palabras funcionales del inglés respecto las 20 primeras palabras funcionales de los otros idiomas. Las coincidencias son remarcables, por arriba del 0.8, es decir 80 %, con excepción del lenguaje ruso, el cual, como ya lo mencionamos anteriormente, no tiene artículos. Repetimos el mismo procedimiento con las 20 primeras palabras de contenido. Aquí vemos que las coincidencias van en un rango de 0.4 a 0.6, al final reduciéndose de 0.5 a 0.6, lo cual puede reflejar el crecimiento de la base de datos, al incrementarse la heterogeneidad de cada base. Para suavizar la gráfica, cada punto es el promedio de 10 años.

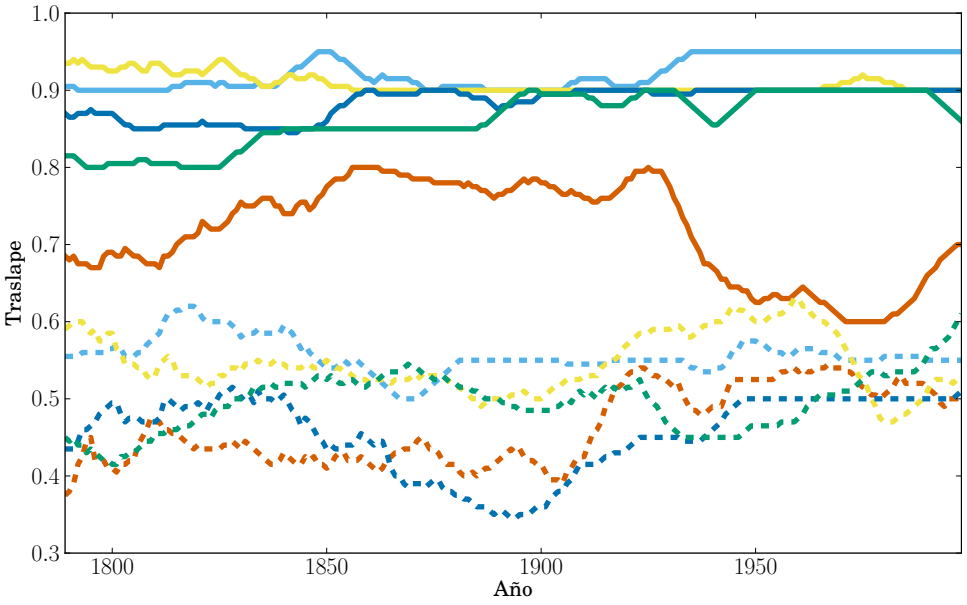

Figura 2.1: Traslape de las palabras en inglés de contenido y funcionales en los otros idiomas. La gráfica compara el grado de similitud de las palabras de contenido y funcionales en inglés con los demás idiomas. Las líneas continuas son las palabras funcionales, mientras que las líneas puntuadas son la de contenido. El código de colores es el mismo que el de la fig. 1.2. Cada punto es el promedio de los 10 años anteriores.

2.2. Evolución temporal de palabras

Conviene preguntarnos por las **trayectorias**, i.e., el recorrido de rangos de las palabras en el tiempo. No tomaremos en cuenta la separación funcional/contenido, queremos ver el comportamiento de cualquier palabra. Para ello veamos las trayectorias de varias palabras al azar para el idioma inglés. Tomamos como base el idioma inglés, ya que tiene la base de datos más grande, de los seis lenguajes. Tomemos como año inicial 1800, dado que es el año donde el número de datos ya es significativo en las distintas bases de datos de los idiomas. Elegiremos palabras entre los rangos 1–20, 400–800 y 2000–5000, ver figs. 2.2, 2.3 y 2.4.

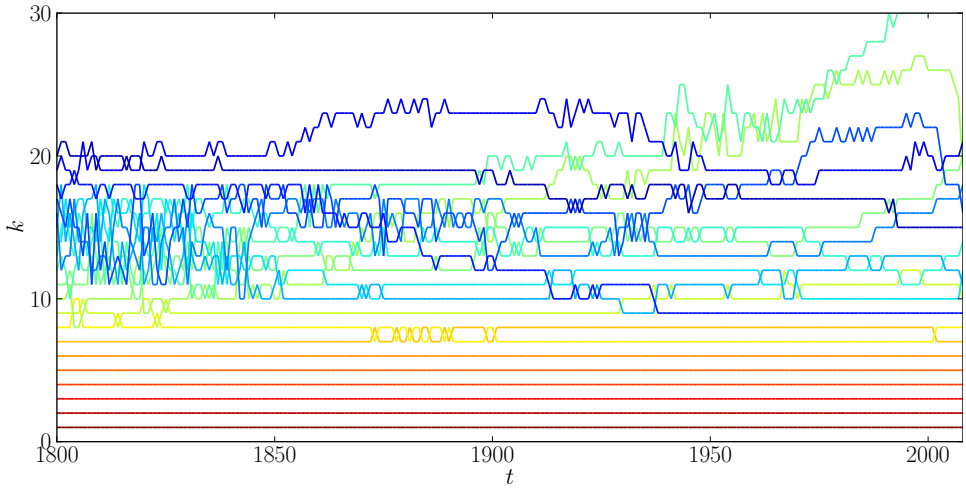

Figura 2.2: Trayectoria de las primeras 20 palabras del idioma inglés en el año 1800. La figura muestra el cambio de rango experimentado, cada año, por estas palabras hasta el año 2008.

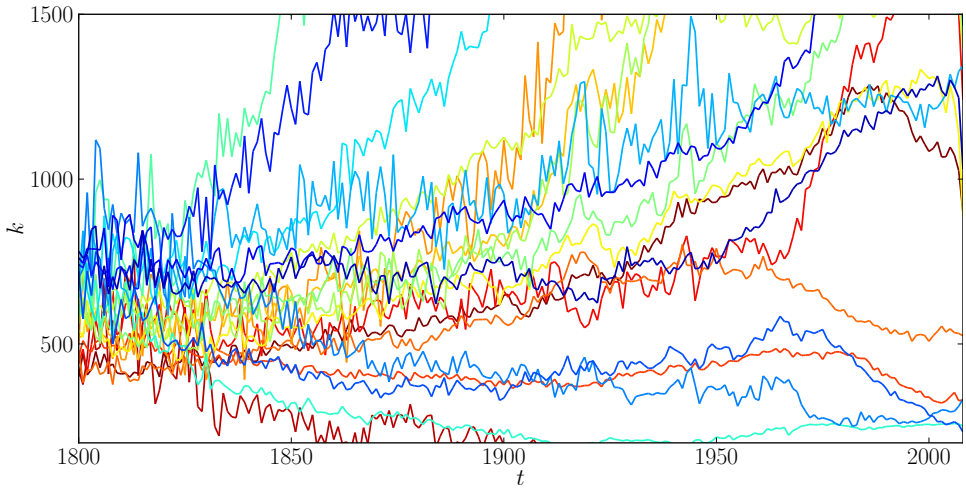

Figura 2.3: Trayectoria de 20 palabras, entre los rangos 400–800 del idioma inglés en el año 1800. La figura muestra el cambio de rango experimentado, cada año, por estas palabras hasta el año 2008.

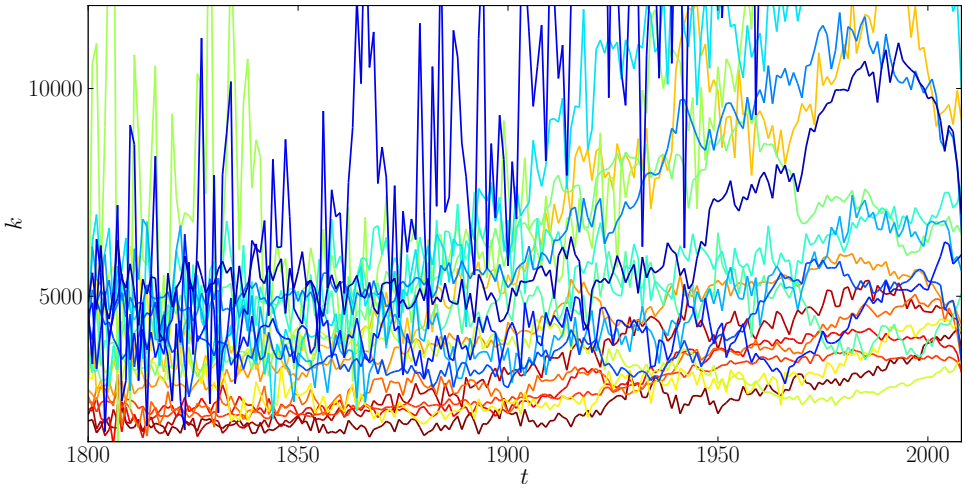

Figura 2.4: Trayectoria de 20 palabras, entre los rangos 2000–5000, del idioma inglés en el año 1800. La figura muestra el cambio de rango experimentado, cada año, por estas palabras hasta el año 2008.

Comentemos las tres figuras anteriores. Algo que esperábamos, o por lo menos ya intuíamos, es que las primeras palabras casi no cambian de rango a través del tiempo. Estas deben ser palabras funcionales, por ejemplo en inglés "the" y "of" son las palabras más usadas y no cambian su posición a lo largo de los años. Conforme el rango inicial es más alto, la posición que empieza a ocupar la palabra a través de los años tiene un comportamiento aleatorio y al parecer el **salto** es más grande entre mayor sea el rango inicial.

2.3. Diversidad de rango

Hasta el momento, la mayoría de los análisis sobre leyes de potencias se han enfocado en las propiedades que presentan las distribuciones en un momento dado. En nuestro caso, cada distribución de palabras corresponde a un año específico y en el capítulo anterior nos fijamos en la variación de los parámetros de los modelos, (1.3.1), a lo largo del tiempo. Sin embargo, hemos visto en las secciones anteriores de este capítulo el comportamiento de los elementos (palabras) de las distribuciones a lo largo del tiempo. En base a ello, podemos proponer una característica que sea común en los seis diferentes idiomas.

Esta nueva característica la hemos llamado **diversidad de rango**; se construye de la siguiente manera:

1. Se fija el año inicial y el número de años , Δt, a evaluar.

2. Se toma el primer rango de todos los años a evaluar y contamos el número de palabras distintas que aparecen en ese rango, con ello ya tenemos el primer punto de la gráfica.

3. Se sigue con el segundo rango y se aplica el procedimiento anterior: se cuenta cuántas palabras diferentes aparecen durante el periodo de tiempo en el rango dos y así tenemos nuestro segundo punto de la gráfica.

4. Este procedimiento se lleva a cabo para el número de rangos que uno desee o que permita la base de datos.

5. Para normalizar los datos, dividimos los resultados entre el número años Δt y así obtenemos $d(k)$.

De esta manera llegamos a una gráfica tipo sigmoide. Proponemos una forma funcional, que sea la cumulativa de una curva gaussiana centrada en μ y de ancho σ:

$$\Phi_{\mu,\sigma}(\log_{10}(k)) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\log_{10}(k)} \mathrm{e}^{-\frac{(\mathrm{y}-\mu)^2}{2\sigma^2}} \mathrm{dy}. \tag{2.3.1}$$

Esta curva se puede caracterizar con los dos parámetros, μ y σ.

Ahora aplicaremos esta nueva medida en las bases de datos de los 6 idiomas. Los resultados son presentados en la figura 2.5. El número de años $\Delta t = 208$ y los rangos analizados son de 1 hasta 10000. Para poder ajustar la sigmoide, se utilizó un ventaneo (*windowing*) logarítmico, sobre el eje x. Con los valores de μ y σ para los seis idiomas, obtuvimos su promedio $\mu = 2.29$ y un ancho de $\sigma = 0.55$.

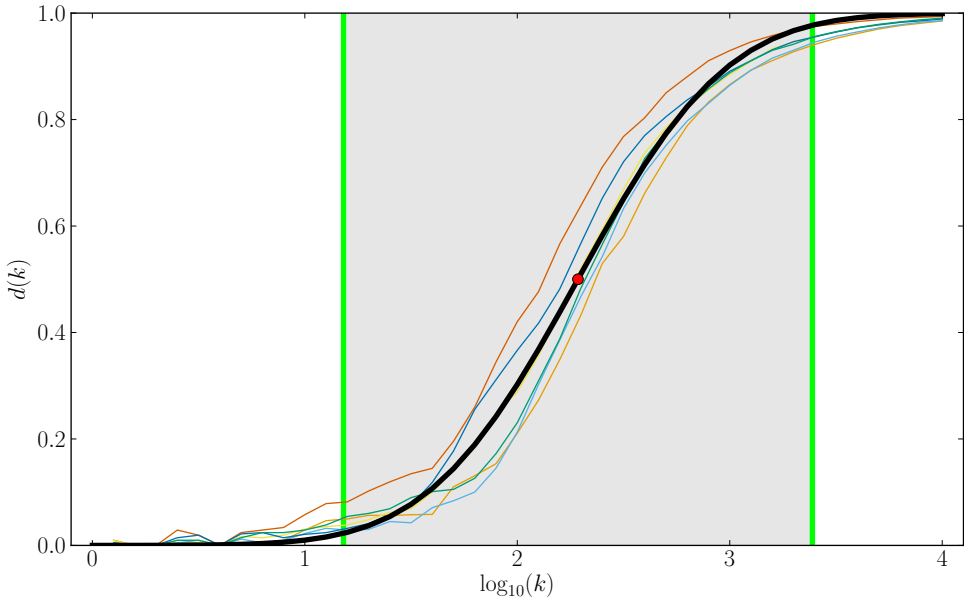

Figura 2.5: Diversidad de rango para distintos idiomas. La diversidad de rango calculada entre los años 1800-2008, para cada uno de los 6 idiomas, respetando el mismo código de colores de la fig. 1.2. La diversidad promedio de los seis idiomas anteriores se muestra en (■). El punto ● indica el valor promedio de las $\hat{\mu}$ para los seis idiomas, mientras que las líneas verticales (■) indican la región comprendida entre $\mu \pm 2\sigma$.

Para darnos una idea de la robustez de esta medida, calculemos, para inglés, la diversidad de rango con diferentes Δt desde 10 hasta 100 años. Como año inicial tomemos el año 1900 y al igual que las anteriores gráficas, el rango máximo a evaluar es el 10000. El resultado es la gráfica 2.6.

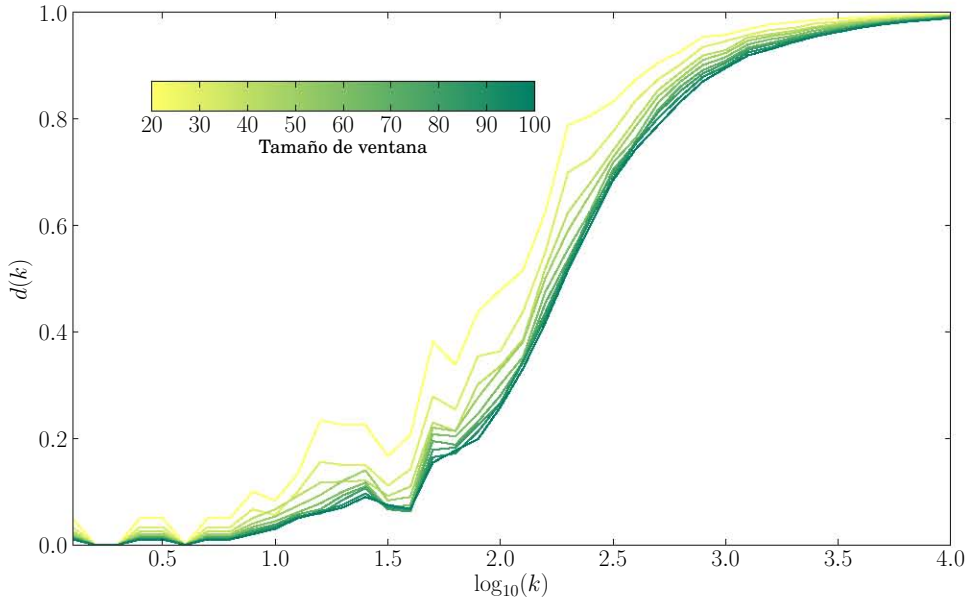

Figura 2.6: Diversidad de rango con diferentes tamaños de ventana, Δt. El año inicial es 1900 en todos los casos y el cálculo de la diversidad es sobre los primeros 10000 rangos. El idioma analizado es inglés.

El resultado es alentador: conforme es mayor el Δt la forma de la curva se mantiene y va acercándose, a lo que parece ser, un límite.

2.4. Cabeza, cuerpo y cola de los lenguajes

En la distribución gaussiana, es común dar especial importancia al intervalo $\mu \pm 2\sigma$, que es donde cae el 95 % de los elementos de la distribución. Estos puntos, delimitan tres regiones; ver fig. 2.5. La primera región, de 0 a $\mu - 2\sigma$, a la que hemos denominado **cabeza**, corresponde a palabras que a lo largo del tiempo no cambian o cambian muy poco de rango. Ya que su diversidad de rango es baja, lo cual queda se refleja en la trayectorias, fig. 2.2. La segunda región comprendida entre $\mu - 2\sigma$ y $\mu + 2\sigma$, el **cuerpo**, contiene palabras que cambian de rango entre ellas y a medida que nos alejamos de $\mu - 2\sigma$, la posibilidad que regresen o vuelvan a ocupar su rango disminuye. Podríamos decir que las palabras en esta región son las mismas a lo largo de los años, sólo hacen cambios de rango entre ellas. La última comprende de $\mu + 2\sigma$ hacia ∞, la **cola**; en esta región es menos probable que las palabras permanezcan sólo en un rango y salten al cuerpo, ya ni pensemos que a la

cabeza, por lo que permanecen en esta región. Creemos que son palabras usadas en ambientes más específicos (médico, científico, militar, etc.).
En el párrafo anterior hablamos en términos de μ y σ, lo cual es conveniente para realizar los cálculos. Sin embargo es más conveniente decir que el punto $\mu - 2\sigma$ corresponde al rango $k_- = 15$; mientras que $\mu + 2\sigma$ corresponde al rango $k_+ = 2448$. Este último número es ligeramente mayor que el número que señalan lingüistas para el mínimo de palabras para poder expresarse en un idioma. Que sea mayor es probable que se deba a que no aplicamos la lematización, como lo explicamos en la sección 2.1.
En la fig. 2.7 aparecen; el centro de la sigmoide μ y los puntos $\mu \pm 2\sigma$ para todos los idiomas. Aquí queremos resaltar que estos puntos en el año inicial tienen un tendencia de crecimiento no despreciable, pero al correr del tiempo llega a estabilizarse; esto debe ser por el crecimiento de las bases de datos. Un caso especial es el idioma ruso, que en 1918 tuvo una reforma ortográfica, lo cual notamos en la abrupta caída de k_- a partir de ese año. También de la misma figura notamos que la mayor influencia de estos cambios es en k_-, que tiene que ver con las palabras funcionales.

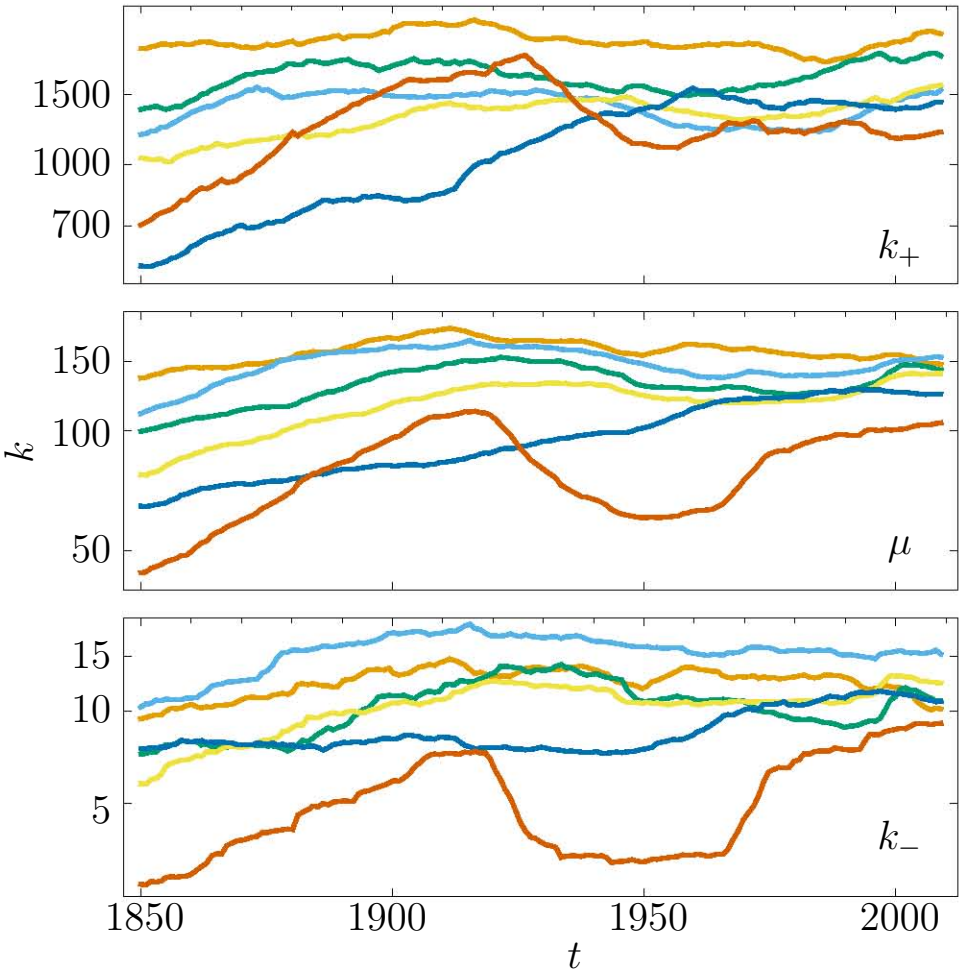

Figura 2.7: Evolución en el tiempo del centro de la sigmoide μ, el fin de la cabeza k_- y el inicio de la cola k_+ para los diferentes idiomas. El intervalo de tiempo utilizado es $\Delta t = 50$ y t es el punto final del intervalo. El código de colores es el mismo que el de la fig. 1.2.

Capítulo 3

Modelo de trayectorias

En este capítulo proponemos y estudiamos un modelo que reproduce las trayectorias de las palabras a través del tiempo y la diversidad de rango.

3.1. Motivación del modelo

Al ver las trayectorias de las palabras, figs. 3.5-3.7, aventuramos la siguiente hipótesis. Las palabras al ocupar la siguiente posición en el ranking del siguiente año saltan a otro rango aleatoriamente. Sin embargo, este salto es proporcional al rango donde actualmente se encuentran. De tal manera debemos encontrar un proceso o algoritmo que pueda reproducir los resultados encontrados en el capítulo anterior.
Como hemos encontrado, las palabras funcionales (artículos, pronombres, etc.), tienen un rango bajo y sus trayectorias son prácticamente rectas horizontales. Podemos definir una variable, para cualquier palabra, que nos será de utilidad en el análisis siguiente. La llamaremos **diferencia de paso**, y es el valor de la diferencia entre el rango en el año actual y el rango en el año siguiente, $k_{t+1} - k_t$. La diferencia de pasos de las primeras diez palabras, ver fig. 2.2, es cero o uno. Por otro lado, tenemos que rangos más altos tienen saltos más grandes y variados, aumentando el valor del salto conforme subimos de rango.
Seleccionamos tres palabras en diferentes posiciones en el ranking y nos fijamos en sus histogramas de diferencias de paso. Estos parecen ajustarse a una distribución normal, ver fig. 3.1 [izquierda]. Si estudiamos la diferencia de pasos relativa (es decir dividir cada $[k_{t+1} - k_t]$ entre k_t), podemos ver más claramente que parecen gaussianas, y con un ancho similar, ver fig. 3.1 [derecha]. De hecho, en estas tres palabras se ve que entre mayor es el rango

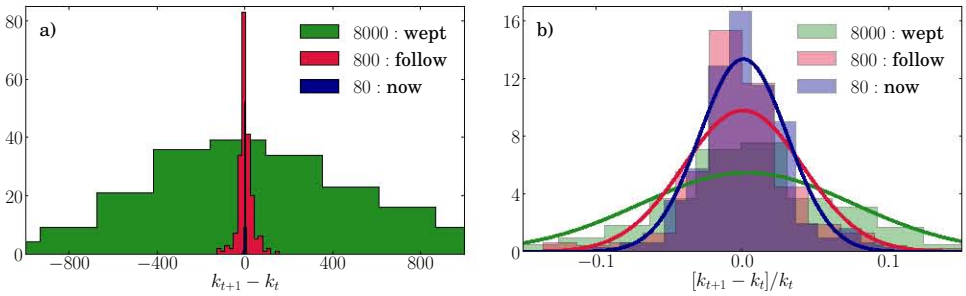

Figura 3.1: [izquierda] Cada histograma corresponde a lo valores de los saltos de rango de un palabra a través del tiempo. El número que antecede a cada palabra es su posición en el ranking de 1800. En este caso las tres palabras son del idioma inglés y las trayectorias van desde 1800 al 2008. [derecha] Cada histogramas corresponde a alguna de las tres palabras usadas en la fig. 3.1. El número que antecede a cada palabra es su posición en el ranking de 1800. La única modificación consiste normalizar el valor de los saltos, dividiendo entre el rango de origen del salto, k_t.

la desviación estándar crece. Quiere decir que hay cierta relación simple entre el rango y el tamaño de salto. Habrá excepciones; pero la gran mayoría sigue este comportamiento, como veremos más adelante.
Si el análisis anterior lo aplicamos para las primeras 10000 palabras del inglés del año 1800, en vez de sólo 3 palabras, obtenemos el promedio, $\hat{\mu}$, y la desviación estándar, $\hat{\sigma}$, de la diferencia de pasos asociada a cada rango, fig. 3.2. De esta misma figura es posible ver que existe una relación entre $\log(\hat{\sigma})$ y $\log(k)$. En particular, proponemos que la relación que gobierna es del tipo lineal,

$$Y = mX + b, \tag{3.1.1}$$

donde $Y = \log_{10}(\hat{\sigma})$ y $X = \log_{10}(k)$, de tal forma que:

$$\hat{\sigma} = \alpha k^m. \tag{3.1.2}$$

Veamos qué pasa con el histograma de las diferencia de pasos normalizados y tratemos de obtener más información. El primer hecho que podemos rescatar de la figura 3.3 es que el promedio de los saltos, es prácticamente cero o fluctúa alrededor de él, en la mayoría de los casos. En cambio vemos que la desviación estándar crece marginalmente. Si tomamos, nuevamente, un histograma de todos los valores de la desviación, el valor que mayormente se repite es 0.07.

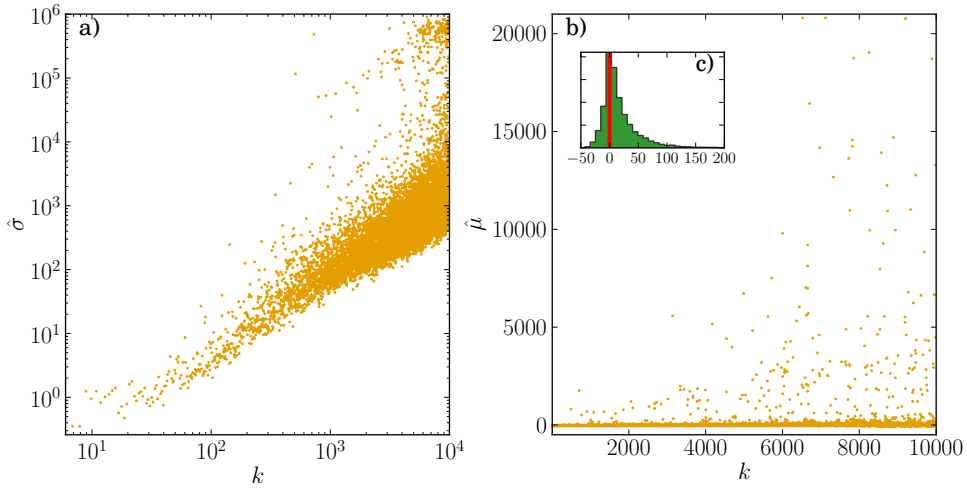

Figura 3.2: El procedimiento realizado para tres palabras del idioma inglés en la fig. 3.1 lo aplicamos para las trayectorias de las primeras 10000 palabras del año 1800. En a) se graficó la desviación estándar, $\hat{\sigma}$, y en b) el promedio, $\hat{\mu}$. Finalmente c) es el histograma de todos las $\hat{\mu}$ de b); la línea roja indica el salto de tamaño cero, es decir ningún salto.

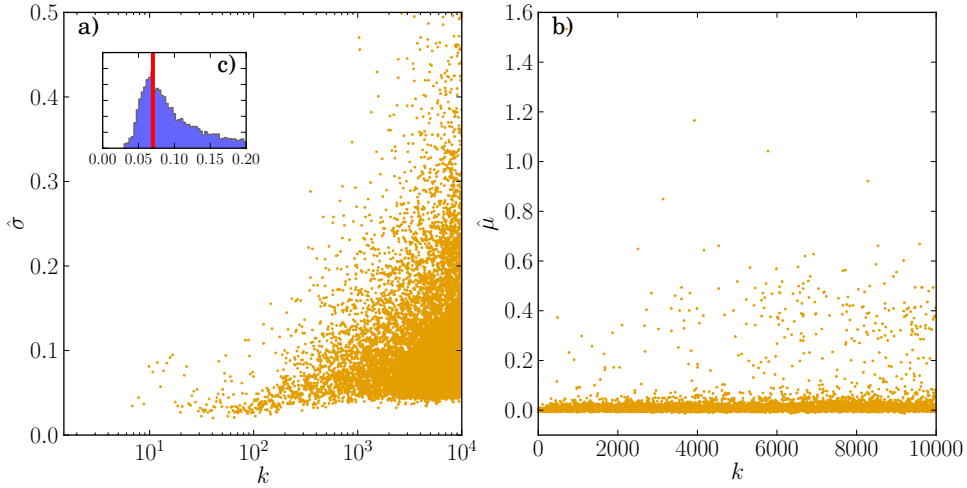

Figura 3.3: Gráfica similar a la fig. 3.2, pero usando saltos relativos, $(k_{t+1} - k_t)/k_t$.

3.2. Modelo de caminatas aleatorias para la diversidad de rango

Con lo que hemos visto hasta este momento sobre la diversidad de rango, las trayectorias de las palabras e histogramas estamos en condiciones de elaborar un modelo que nos permita reproducir las trayectorias y, por ende, la diversidad de rango. El procedimiento es el siguiente:

1. Se crea una lista con un determinado número N de **palabras artificiales**. En realidad estas palabras artificiales son etiquetas o mejor dicho objetos, que no cambian durante todo el tiempo analizado. Por ejemplo, se pueden nombrar *palabra-1, palabra-2, . . . , palabra-N*. Esta lista representa el ranking de N palabras en el año inicial.

2. El siguiente paso es obtener el rango que ocuparán las palabras al año siguiente. El rango de cada palabra se puede ver como un número entero asociado a ella. Por ejemplo, tomemos la palabra que ocupa el rango k, *palabra-k*. El salto que dará al siguiente año depende del rango k y, como vimos en la sección anterior, supondremos que el tamaño del salto está dado por un número aleatorio gaussiano de promedio cero y desviación estándar proporcional al rango k, $G(0, \alpha k)$. Por lo tanto el nuevo número, s, asignado para el siguiente año a *palabra-k* es,

   $$s_{t+1} = k_t + G(0, \alpha k_t).$$

   Esto se hace para todas las palabras del ranking y cada palabra artificial queda asociada a un número s_{t+1}. Nótese que s_{t+1} no es un número entero, ni es el rango correspondiente al año siguiente, $t + 1$. Es una variable auxiliar.

3. Las palabras artificiales se reordenan en base a los números s_{t+1}. A la palabra con el número s_{t+1} más pequeño le corresponde el rango 1, y así sucesivamente, hasta que la palabra con el s_{t+1} más grande le corresponde el rango N. Lo que nos da el ranking del año siguiente.

4. Los pasos anteriores se repiten el número de veces que deseemos y que corresponderá al número de años.

Falta por determinar los parámetros que usemos para la simulación, los dos primeros son claros; el número de años y el número de palabras sintéticas. El último parámetro, pero no menos importantes, es α, estos los obtuvimos de la gráfica 3.3. Hicimos el ajuste de estos datos con la ecuación (3.1.2), de la

misma forma que se hizo para obtener la diversidad de rango; agrupando en casillas logarítmicas y sacando un promedio sobre cada casilla. A esta curva promedio fue la que se ajusto a una recta. De tal manera que obtuvimos una $m \approx 1$ y una $\alpha \approx 0.1$.

El procedimiento anterior es una primera aproximación al problema. El siguiente listado tiene algunos puntos que se pueden considerar importantes y que nuestro modelo no toma en cuenta.

- El nacimiento y muerte de palabras. Hay palabras que caen en desuso total y desaparecen del vocabulario común, otras que sufren alteraciones en su forma escrita y algunas más son prestamos lingüísticos que terminan por incorporarse al vocabulario. Por ejemplo "computadora". En nuestro modelo son las mismas palabras en todos los años.

- El cálculo de α, parámetro del modelo. Para su cálculo sólo utilizamos las palabras del año inicial, y no todas las palabras que pueden aparecer a lo largo de los años. En la fig. 3.4 se puede ver el porcentaje de las 10000 primeras palabras de 1800 que sobreviven año con año hasta el 2008. Podemos ver este porcentaje disminuye pero siempre queda por arriba de 60 %.

- La interacción entre palabras. No hemos tomado en cuenta la sintaxis del idioma, como el modelo del gas ideal donde no se toman en cuenta las interacciones entre partículas. Así en este modelo no se toma en cuenta la interacción entre palabras. Esto es una aproximación ya cada idioma posee una estructura para armar oraciones.

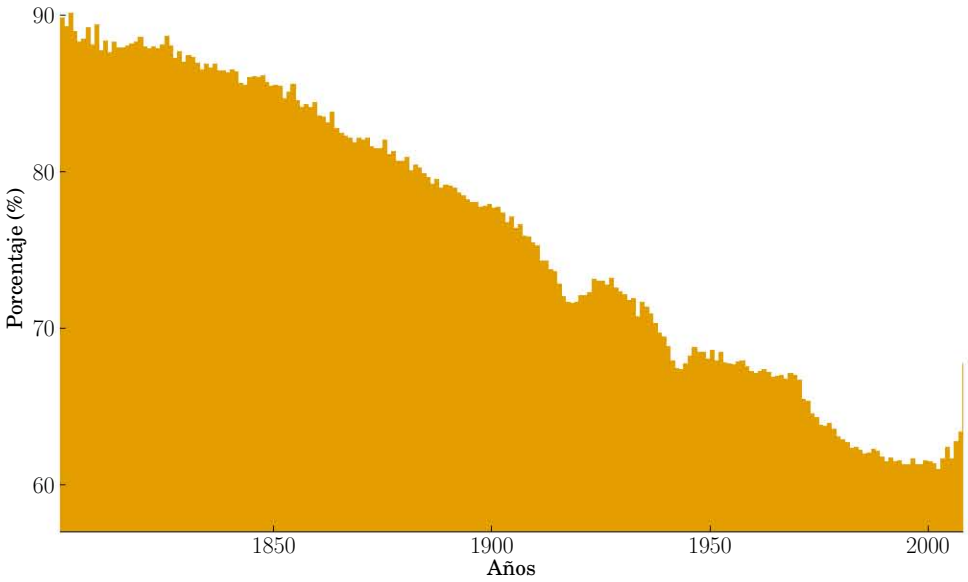

Figura 3.4: Porcentaje de la primeras 10000 palabras del año 1800 que comparten las primeras 10000 palabras para todos los demás años.

3.3. Comparación entre simulaciones y datos experimentales

Ya estamos en condiciones de contrastar nuestro modelo con los datos. Para comparar con los datos reales, alimentamos el modelo con un número de palabras igual a 20000. La diversidad de rango experimental la cortamos en 10000 palabras; para calcular la diversidad de rango sintética igualmente lo haremos a 10000; sin embargo las otras 10000 palabras actúan como reservorio. El número de años simulados será 209, mismo periodo de tiempo que para las palabras reales. Tomaremos $\alpha = 0.066$.

Los resultados de las trayectorias simuladas están en las gráficas 3.5, 3.6 y 3.7 y corresponden a la región de la cabeza, cuerpo y cola, respectivamente. La diversidad de rango simulada se compara con la diversidad de rango de los idiomas estudiados en la fig. 3.8.

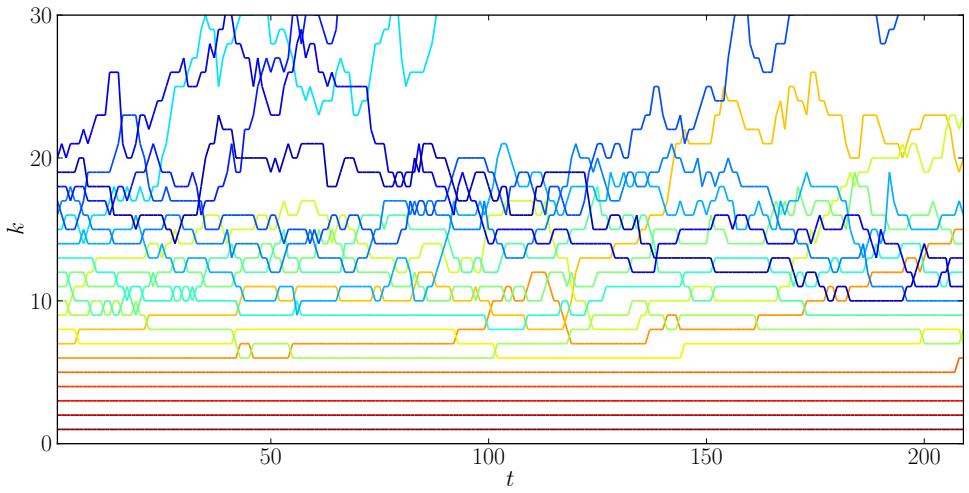

Figura 3.5: Trayectorias de las primeras 20 palabras artificiales en la zona de la cabeza, a lo largo de 208 años.

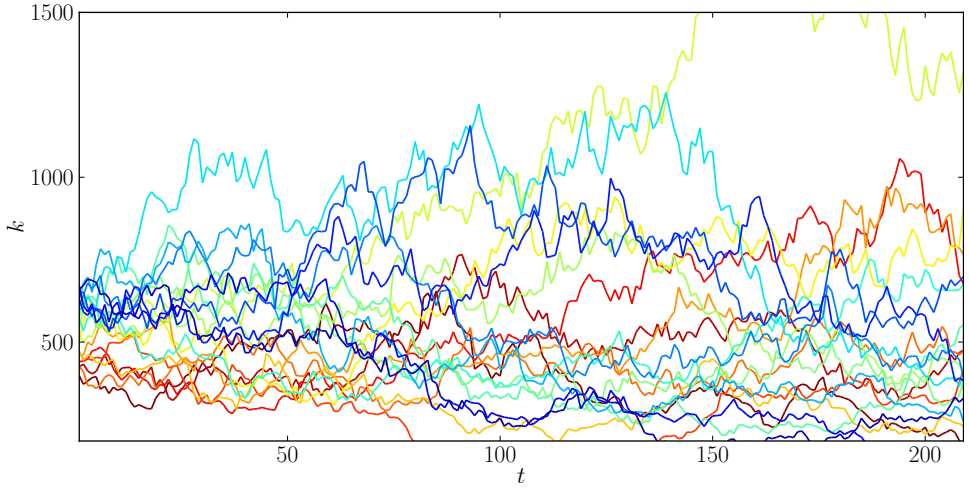

Figura 3.6: Trayectorias 20 de palabras artificiales ubicadas en el cuerpo, a lo largo de 208 años.

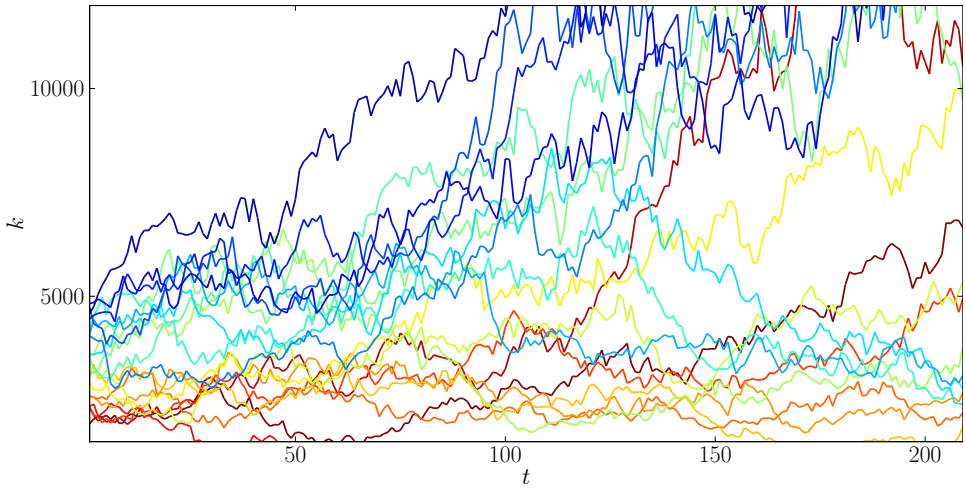

Figura 3.7: Trayectorias 20 de palabras artificiales ubicadas en la cola, a lo largo de 208 años.

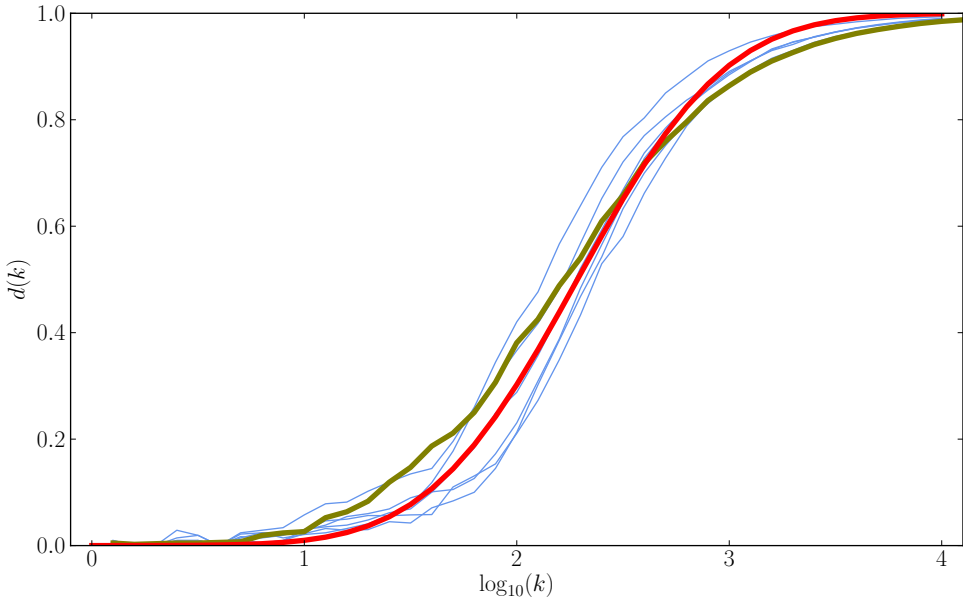

Figura 3.8: Diversidad de rango sintética vs. diversidad de rango de cada idioma. La diversidad de rango obtenida a lo largo de 208 años (1800-2008) para cada uno de los 6 idiomas (■), nuestro ansatz eq. (2.3.1) (■) y la diversidad de rango simulada (■).

En la sección 3.1 ajustamos los histogramas, correspondientes a los saltos de las palabras, con una distribución gaussiana. Sin embargo, se podría objetar que los histogramas de pasos pueden ajustarse a distribuciones de similar forma que la distribución gaussiana, por ejemplo, la distribución lorentziana. Hay dos razones por las que se descartan otras distribuciones diferentes a la guassiana. En primera, la sencillez. Siempre se trata de buscar el modelo más sencillo que funcione, y en segunda, y más importante, ahora con el modelo se puede ver que si en vez de tomar el número aleatorio $G(0, \hat{\sigma}))$ usamos un número aleatorio lorentziano $L(0, \hat{\sigma})$, las trayectorias seguidas no se parecen en nada a la trayectorias de la palabras ni la diversidad de rango corresponde a la obtenida de los idiomas. Esto inmediatamente descarta un comportamiento exclusivamente lorentziano, fig. 3.9.

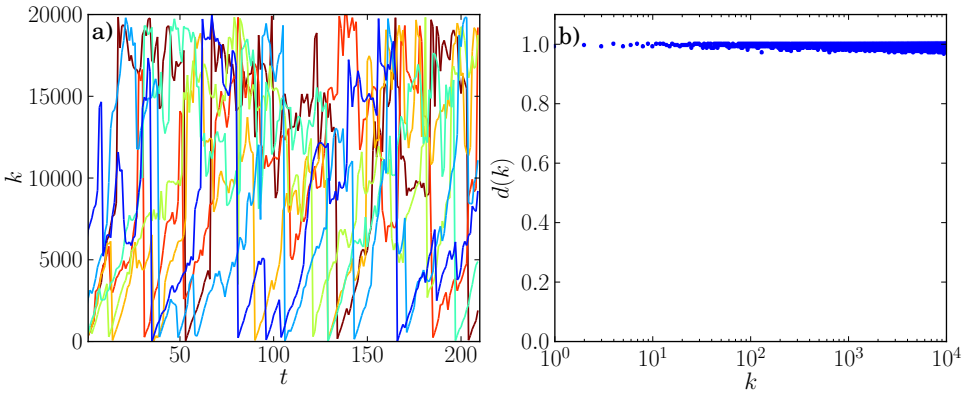

Figura 3.9: En a), hay graficadas algunas trayectorias de palabras artificiales generadas con números aleatorios lorentzianos. En b) esta la diversidad de rango simulada, debida a todas la trayectorias de palabras artificiales obtenidas de la simulación.

3.4. Exploración del modelo

Exploramos más a fondo el modelo y encontramos un rango para α en el cual el comportamiento de las trayectorias de las palabras y la diversidad de rango es similar al observado en los 6 idiomas analizados. Encontramos que con 20000 palabras y 210 años, el rango $0.01 \leq \alpha \leq 0.25$ es el adecuado.
En la fig. 3.10 vemos que mientras más pequeño sea α, μ aumenta, es decir que al graficar la diversidad de rango, el centro del sigmoide se desplaza a la derecha. Esto sucede porque hay mayor número de palabras que no cambian de rango, las de rangos bajos. Podemos decir que existe un k_{max} donde las

palabras con rango menor son trayectorias rectas, ver la primera columna de la fig. 3.11. Esto se traduce como $d(k) = 0$ para todas las $k < k_{max}$, y como consecuencia el crecimiento de μ, fig. 3.12. En cambio si α se incrementa, la probabilidad de que las palabras de rango bajo cambien de rango al siguiente paso, aumenta. Esto trae como consecuencia que la diversidad de rango para palabras de rango bajo aumente, o que k_{max} disminuya. Las trayectorias producidas en esta situación están en la última columna de la fig. 3.11. La diversidad de rango llega que desaparecer ya que para todos los rangos $d(k) \sim 1$ y tendríamos un caso parecido a lo que sucede con las trayectorias lorentzianas. De la fig. 3.10 vemos que esto sucede alrededor de $\alpha^{-1} \sim 4$, o, $\alpha \sim 0.25$. Mención a parte merece σ por que fluctúa alrededor de 0.7, para $1/200 < \alpha \leq 0.25$. En este régimen, σ no varia apreciablemente.

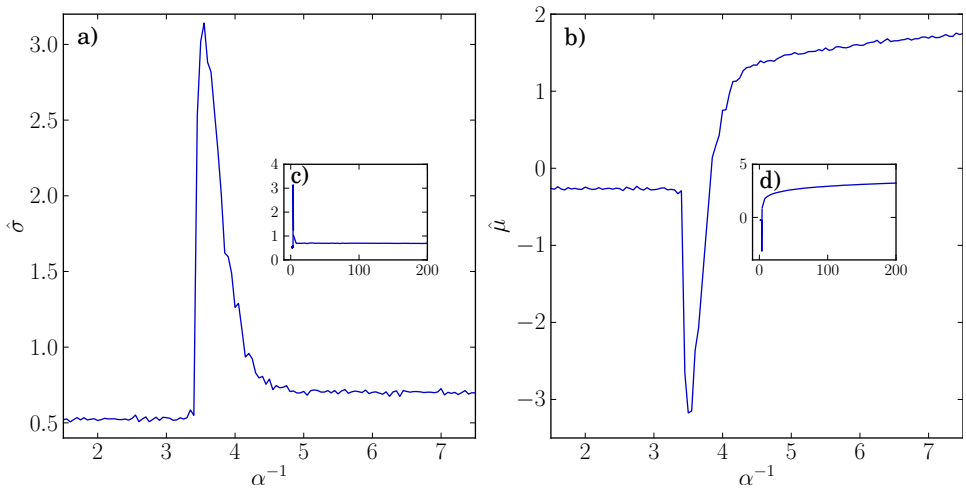

Figura 3.10: Simulación del modelo de caminatas aleatorias para un rango de $2 \leq \alpha^{-1} \leq 200$, 20000 palabras artificiales y 200 años. En a) y c) se grafica σ como función de α^{-1}, mientras que b) y d) se grafica μ como función de α^{-1}.

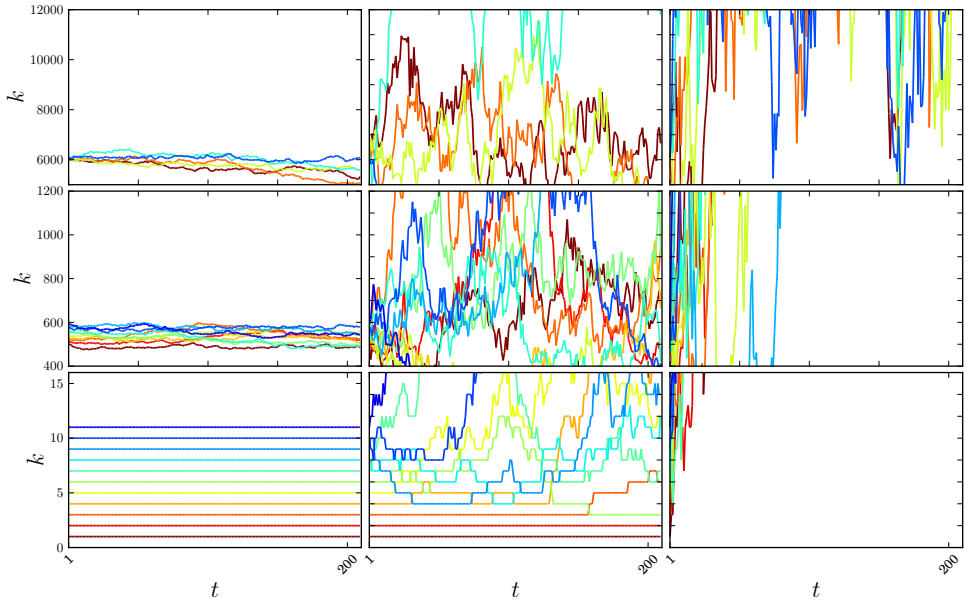

Figura 3.11: Trayectorias de palabras artificiales en las tres diferentes regiones con tres distintos valores de α. Las trayectorias de la primer columna (de izquierda a derecha) son debidas a la simulación $\alpha = 0.005$, la segunda columna con $\alpha = 0.079$ y la tercera con $\alpha = 0.251$. La primera fila (de arriba hacia abajo) las trayectorias corresponde a la región de la cola, la segunda fila a región del cuerpo y la tercera a la región de la cabeza. El $\Delta t = 210$ y el número de total de palabras es 20000.

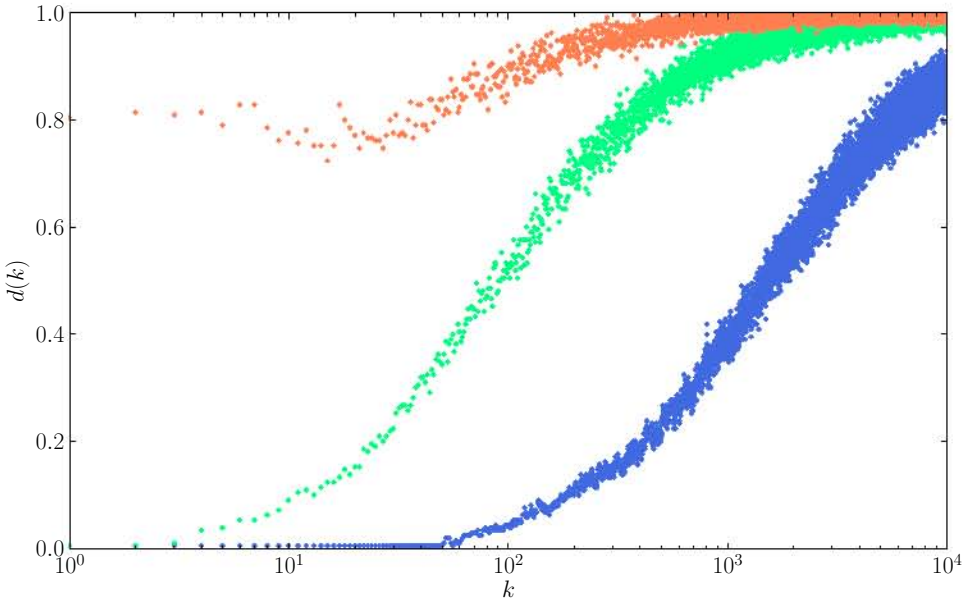

Figura 3.12: La diversidad de rango de las trayectorias simuladas, para $\alpha = 0.005$ (■), $\alpha = 0.079$ (■) y $\alpha = 0.251$ (■). El Δt es 210 y el número de total de palabras es 20000.

Capítulo 4

Conclusiones

En está tesis hemos visto la distribución de las frecuencias de las palabras para 6 distintos idiomas, tomando como base el idioma inglés. Tratamos de ajustar cinco distintas distribuciones y vimos cómo varían sus parámetros año con año. Estos últimos están correlacionados con el número de libros escaneados por Google cada año y con eventos históricos. Sin embargo, ninguno de los modelos presentados es satisfactorio para describir los datos.

En la búsqueda de encontrar un modelo, una regla o una ley común a los datos de los idiomas, hemos incorporado el factor del tiempo en este estudio. Analizando las trayectorias seguidas por las palabras a través del tiempo, caracterizándolas por cómo cambian el rango en cada paso de tiempo. Esto nos llevó a dos conclusiones importantes, que son el aporte de nuestro trabajo: la **diversidad de rango** y el modelo de caminatas aleatorias gaussianas invariantes de escala.

La diversidad de rango es una nueva caracterización que toma en cuenta la variación de las palabras en cada rango en un tiempo dado. Esta resulta ser muy similar para los 6 idiomas. Además, nos permite dividir las palabras en tres regiones: cabeza, cuerpo y cola.

Fuimos capaces de elaborar un modelo de caminatas aleatorias que reproduce aceptablemente los resultados de las trayectorias de las palabras y su diversidad de rango. Este modelo sólo tiene un parámetro, no toma en cuenta la muerte de palabras y el nacimiento de nuevas palabras, ni algún otro factor externo que pudiese modificar las trayectorias, aún así es capaz de reproducir aceptablemente, entre otros, la diversidad.

Observamos que nuestro estudio está hecho a un nivel estadístico. No estamos enfocados en aspectos sintácticos, semánticos o gramaticales del lenguaje humano. Sin embargo hay preguntas que quedan por responder ¿Por qué la

diversidad de rangos se aproxima a una distribución lognormal? ¿Cuáles son los procesos y mecanismos que son requeridos para esto? ¿Para qué otro tipo de sistemas es válida? ¿Existe un Δt característico para cada sistema o base de datos?

Para responder algunas de las preguntas basta con aplicar la diversidad de rango a diferentes sistemas rankeados; deportes, economía, redes sociales, etcétera. En algunos de ellos la forma del ranking (en log-log) es diferente a la que tenemos en palabras. Por ejemplo, se puede ver en la figura 4.1, la distribución del número de llamadas de celular hechas en una hora determinada en Senegal, es muy diferente a la distribución de rango para idiomas. En la figura 4.2 se puede ver la diversidad de rango para esa misma hora pero durante todo un año. Es bastante alentador ver que a pesar de ser datos tan diferentes, la diversidad de rango tiene una forma similar a la que obtuvimos para palabras.

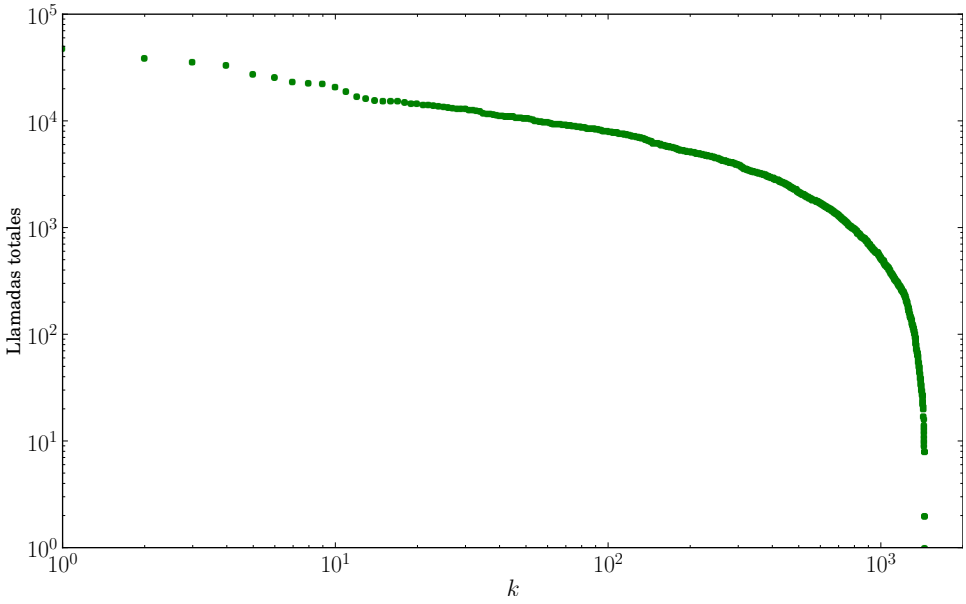

Figura 4.1: Distribución de rango-frecuencia para el número de llamadas hechas por celulares en Senegal, entre las 15 hrs. y 16 hrs. el 10 de Marzo de 2013. Cada elemento del ranking corresponde a una antena de recepción-emisión, se ordenan de acuerdo al número total de llamadas salientes en la antena.

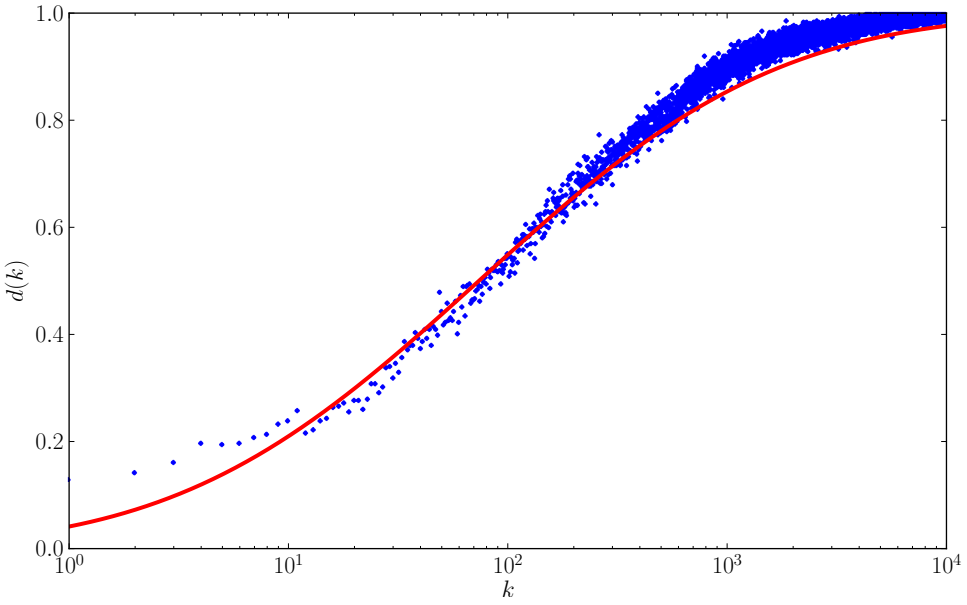

Figura 4.2: Diversidad de rango del número de llamadas realizadas entre las 15 hrs. y 16 hrs. durante el año 2013. Los valores de los parámetros de la sigmoide son $\mu = 1.87$ y $\sigma = 1.08$.

Apéndice A

Artículo en PLoS ONE

PLOS ONE

RESEARCH ARTICLE

Rank Diversity of Languages: Generic Behavior in Computational Linguistics

Germinal Cocho[1,2], Jorge Flores[1], Carlos Gershenson[3,2]*, Carlos Pineda[1], Sergio Sánchez[4]

1 Instituto de Física, Universidad Nacional Autónoma de México, Mexico City, Mexico, **2** Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City, Mexico, **3** Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Mexico City, Mexico, **4** Facultad de Ciencias, Universidad Nacional Autónoma de México, Mexico City, Mexico

* cgg@unam.mx

Citation: Cocho G, Flores J, Gershenson C, Pineda C, Sánchez S (2015) Rank Diversity of Languages: Generic Behavior in Computational Linguistics. PLoS ONE 10(4): e0121898. doi:10.1371/journal.pone.0121898

Academic Editor: Eduardo G. Altmann, Max Planck Institute for the Physics of Complex Systems, GERMANY

Received: October 14, 2014

Accepted: February 5, 2015

Published: April 7, 2015

Data Availability Statement: Data available from the Google Ngrams dataset at https://books.google.com/ngrams/datasets.

Funding: GC received support from project IN107414 from the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica of the Universidad Nacional Autonoma de Mexico. CG was supported by SNI membership 47907 of Consejo Nacional de Ciencia y Tecnologia, Mexico. CP received support from the projects 153190 from Consejo Nacional de Ciencia y Tecnologia and IA101713 from the Programa de Apoyo a Proyectos

Abstract

Statistical studies of languages have focused on the rank-frequency distribution of words. Instead, we introduce here a measure of how word ranks change in time and call this distribution *rank diversity*. We calculate this diversity for books published in six European languages since 1800, and find that it follows a universal lognormal distribution. Based on the mean and standard deviation associated with the lognormal distribution, we define three different word regimes of languages: "heads" consist of words which almost do not change their rank in time, "bodies" are words of general use, while "tails" are comprised by context-specific words and vary their rank considerably in time. The heads and bodies reflect the size of language cores identified by linguists for basic communication. We propose a Gaussian random walk model which reproduces the rank variation of words in time and thus the diversity. Rank diversity of words can be understood as the result of random variations in rank, where the size of the variation depends on the rank itself. We find that the core size is similar for all languages studied.

Introduction

Statistical studies of languages have become popular since the work of George Zipf [1] and have been refined with the availability of large data sets and the introduction of novel analytical models [2–7]. Zipf found that when words of large corpora are ranked according to their frequency, there seems to be a universal tendency across texts and languages. He proposed that ranked words follow a power law $f \sim 1/k$, where k is the rank of the word—the higher ranks corresponding to the least frequent words—and f is the relative frequency of each word [8, 9]. This regularity of languages and other social and physical phenomena had been noticed beforehand, at least by Jean-Baptiste Estoup [10] and Felix Auerbach [11], but it is now known as Zipf's law.

Zipf's law is a rough approximation of the precise statistics of rank-frequency distributions of languages. As a consequence, several variations have been proposed [12–15]. We compared

de Investigación e Innovación Tecnológica of the Universidad Nacional Autonoma de Mexico. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Zipf's law with four other models, all of them behaving as $1/k^a$ for a small k, with $a \approx 1$, as detailed in the SI. We found that all models have systematic errors so it was difficult to choose one over the other.

Studies based on rank-frequency distributions of languages have proposed two word regimes [15, 16]: a "core" where the most common words occur, which behaves as $1/k^a$ for small k, and another region for large k, which is identified by a change of exponent a in the distribution fit. Unfortunately, the point where exponent a changes varies widely across texts and languages, from 5000 [16] to 62,000 [15]. A recent study [17] measures the number of most frequent words which account for 75% of the Google books corpus. Differences of an order of magnitude across languages were obtained, from 2365 to 21077 words (including inflections of the same stems). This illustrates the variability of rank-frequency distributions. The core of human languages can be considered to be between 1500 and 3000 words (not counting different inflections of the same stems), based on basic vocabularies for foreigners [18], creole [19], and pidgin languages [20]. For example, Voice of America's Special English [21] and Wikipedia in Simple English use about 1500 and 2000 words, respectively (not counting inflections). The Oxford Advanced Learner's Dictionary lists 3000 priority lexical entries [22]. This suggests that the change of exponent a or another arbitrary cutoff in rank-frequency distributions does not reflect the size of the core of languages.

In view of these problems with rank-frequency distributions, we propose a novel measure to characterize statistical properties of languages. We have called this measure *rank diversity* and it tells us how words change their rank in time. With rank diversity, three regimes of words are identified: "heads", "bodies" and "tails". This measure of rank diversity follows the same simple functional law with similar parameters for all data analyzed. In particular, this is so for the six European languages studied here using a large data set of more than 6.4×10^{11} words from Google Books [23], which contains about 4% of all books written until 2008. It should be noted that this data set includes all different inflected forms (such as plural, different tense/aspect forms, etc.) found in the book corpus. Data sets such as this have allowed the study of "culturomics": how cultural traits such as language have changed in time [24–30].

The rank diversity follows a scale-invariant behavior regarding its fluctuations, which inspires a model based on random walks, with scale-invariant random steps. This model reproduces the behavior of diversity and thus captures the essence of the evolution of word rank across different languages.

Rank diversity of words

In what follows we shall consider six European languages from the Indo-European family. They are English and German; Spanish, French and Italian; and Russian. They belong to different linguistic branches: Germanic, Romance, and Slavic, respectively. The native speakers of these languages account for approximately 17% of the world population.

We shall start by taking into account the 20, say, most used words in the six languages, that is, the lowest-ranked words. Using, for the sake of uniformity, the first sense or first meaning given by Google Translate, once these words are translated into English, the coincidences in all six languages are remarkable (see Table S1 in S1 Text). This could have been foreseen, since most of the lowest-ranked words are articles, prepositions or conjunctions, *i.e.* what is called function words. A different matter, as we shall see, would result if we had considered only nouns, verbs, adverbs or adjectives, known as content words.

In order to quantify this fact, we present in Fig. 1 the time evolution of the overlap of the first 20 lowest-ranked words in the five languages with respect to the corresponding list of English. From the upper part of this figure we can see that along two centuries this overlap

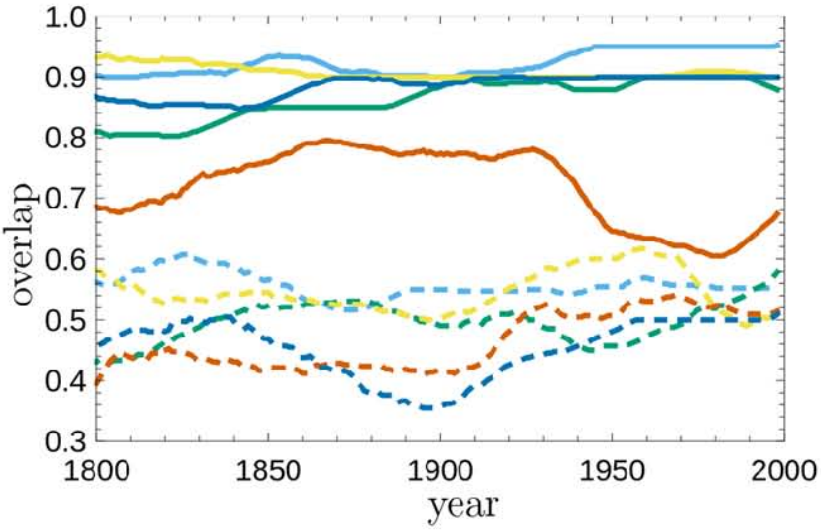

Fig 1. Overlap of the 20 most frequent words (continuous lines), and of the 20 most frequent *content* words (dashed lines) across languages, with respect to English, as a function of time. When words have more than one meaning, the first sense, according to Google Translate, was used. The color code for languages is as follows: light blue for French, green for German, yellow for Italian, dark blue for Spanish, and dark orange for Russian. Additionally, light orange will be used for English when required (see also Fig. 2). The same color coding for languages will be used throughout the rest of the article.

doi:10.1371/journal.pone.0121898.g001

fluctuates around 0.9, a rather large number, except for Russian, since this language does not have articles. These data reveal that these Indo-European languages have shared structural properties, notwithstanding that they belong to distinct linguistic branches.

The lowest-ranked words used to construct the upper part of Fig. 1 are essentially the same along centuries (See Figs S3-S8 in S1 Text). But this is not the case for content words, as can be seen in Table S2 in S1 Text. First, and as also shown by the dashed curves in Fig. 1, the overlap of these words with respect to English for the other five languages (including Russian) is of the order of 0.5. These values are much lower than the overlap of function words. Second, the most common nouns vary considerably with time. On the one hand, nouns like *time*, *man*, *life* and their translation to the other languages are present independently of the century. On the other hand, words like *god* and *king* have a low rank in the eighteenth century but have a larger rank in the last century. The rank change in time of these nouns reflect cultural facts.

What is discussed in the previous paragraph is an example of what could be called rank diversity $d(k)$. This is, in the present study, the number of different words occurring at a specific rank k over a given period of time Δt. We found that the resulting rank diversity curves for the six languages studied between 1800 and 2008 are similar to each other, as shown in Figs. 2 and 6. Low ranks have a very low diversity, as few words appear in the same ranks for the years we have studied.

As shown by the continuous lines in Fig. 2, the sigmoid curve fits very well $d(k)$ for all languages considered, except for low k where the statistical fluctuations are larger due to the small sample size. The sigmoid is the cumulative of a Gaussian distribution, i.e.

$$\Phi_{\mu,\sigma}(\log_{10} k) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\log_{10} k} \mathrm{e}^{-\frac{(y-\mu)^2}{2\sigma^2}} \mathrm{d}y, \tag{1}$$

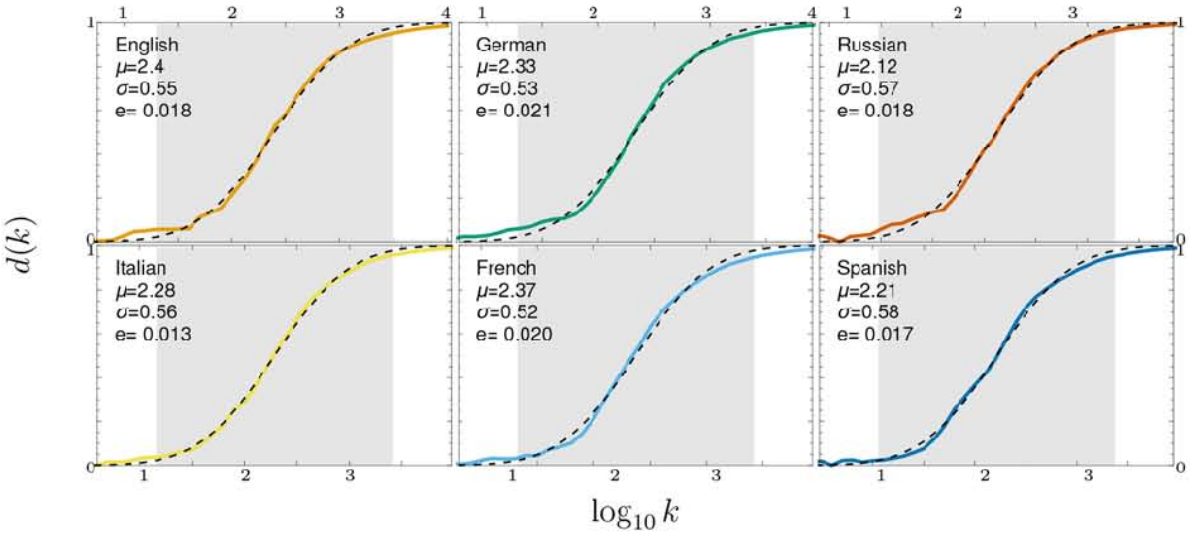

Fig 2. Rank diversity. Diversity d as a function of the rank k for different languages from 1800 to 2008, where $d(k)$ measures how many different words appear for a given rank k during the time considered ($\Delta t = 208$). For example, for English, $d(1) = 1/208$, as the word 'the' appears in the first rank for all years considered. Although we have analyzed up to $k \approx 10^6$, rank diversity for $k > 10^4$ is not shown as $d(k) \approx 1$, *i.e.*, a different word appears in each rank every year. Data are windowed over time, with a slot of size $\delta \log_{10} k = 0.1$, for the sake of clearness. Additionally, the sigmoid defined in Equation 1 is shown as a black dashed curve, with the best fit parameters, also reported in each subfigure. The mean square error e between the data and the fit, is also given. The shaded region corresponds to the average "body" of all languages.

doi:10.1371/journal.pone.0121898.g002

and is given as a function of $\log k$. The values of μ and σ reported in Fig. 2 were obtained adjusting Equation 1 to the rank diversity calculated for each individual language. The mean value μ identifies the point where $d(k) \approx 0.5$, while the width σ gives the scale in which $d(k)$ gets close to its extremal values. When $\log k$ is much larger than $\mu+\sigma$, $\Phi_{\mu,\sigma}(\log k)$ gets exponentially close to one, whereas when $\log k$ is much smaller than $\mu-\sigma$ it gets exponentially close to zero. It is customary in statistics to define a bulk of the Gaussian between $\mu \pm 2\sigma$, where 95% of the population lies. Along the same lines, we define three regions, marked by

$$\log_{10} k_{\pm} = \mu \pm 2\sigma. \tag{2}$$

First, we find what we shall call the head of the language, distributed with ranks between 1 and k_-; a second region, identified as the body of the language, lies between k_- and k_+; and finally the tail, beyond k_+. From the values reported in Fig. 2, we see that $9 < k_- < 22$, while k_+ lies between 1832 and 3099. As shown in Fig. 3, these regions are robust to changes in the historical period considered and to the data set size (larger for recent years).

The bodies of languages consist of words that have limited change in time. Based on the size of basic vocabularies, it can be argued that the "core" of English is between 1500 and 3000 words, as mentioned in the introduction, which is consistent with our results. If we agree that the rank diversity identifies the core (head and body) of English, then it can be argued that the size of the core of the other five languages studied is similar [31], which is also supported by the high similarity across languages in Fig 2.

The tails of languages are formed by words which vary their rank considerably in time. This implies that they are more dependent on the text and its domain than words from the core. It can be assumed that words belonging to the head and body of languages have a high

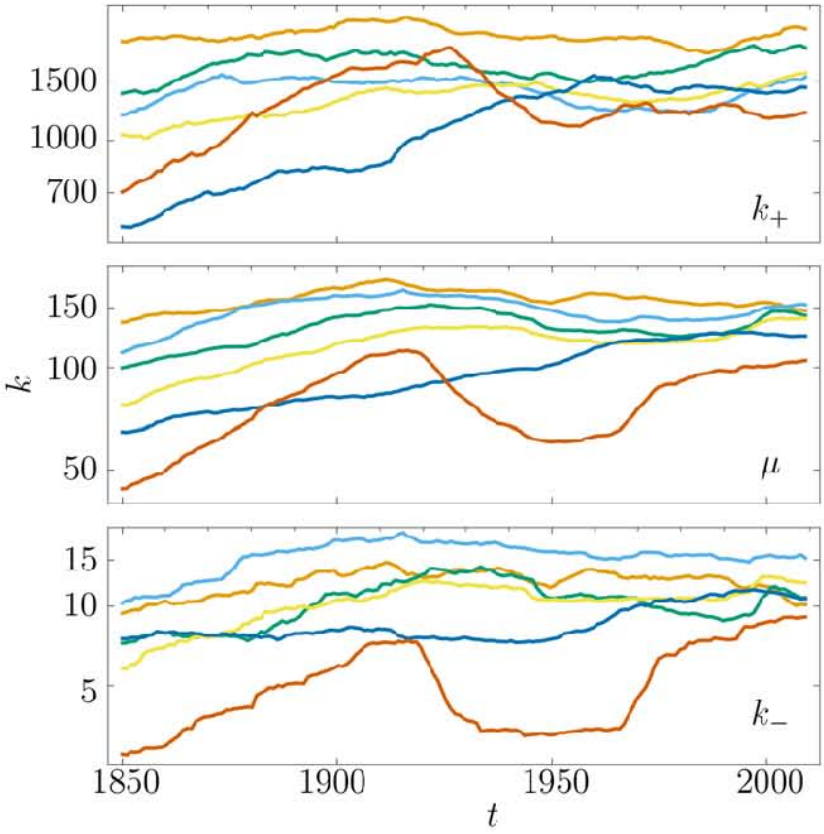

Fig 3. Evolution in time of the center of the sigmoid (middle panel), and the borders of the head and body (bottom panel) and body and tail (top panel) for the different languages along time for intervals of fifty years, *i.e.* $\Delta t = 50$. Head words have $k \leq k_-$, body words have $k_- < k \leq k_+$, and tail words have $k_+ < k$. See Fig. 2 for color coding.

doi:10.1371/journal.pone.0121898.g003

probability of being used in any text, while words from the tail would appear only in specific texts and domains.

Note that we obtain language cores slightly larger than those proposed by linguists. This is to be expected, as the Google Books data set treats words forms inflected for different persons, tenses, genders, numbers, cases, and so forth, as distinct items, while dictionaries count only stems (presented as citation forms, i.e. the basic form that users are most likely to look up). For example, the core for English obtained using rank diversity consists of 2448 words, but within these there are only 1760 different stems in the year 2008. Moreover, the studied data set contains several proper names which are not included in basic vocabulary lists. For English, 55 out of 2448 are proper names in 2008.

The rank evolution of particular words in time, belonging to the head, body, and tail of English is shown in Fig. 4a. This ratifies the results shown in Fig. 2, where low-ranked words exhibit little variation in time and this variation increases with the rank. More trajectories are presented in the SI. As mentioned above, words from the head vary little over time. However, the way in which words from the body or tail vary their rank in time appears to be similar, although at a different scale. This similarity leads us to propose a model of rank diversity where the amount of rank variation depends only on the rank.

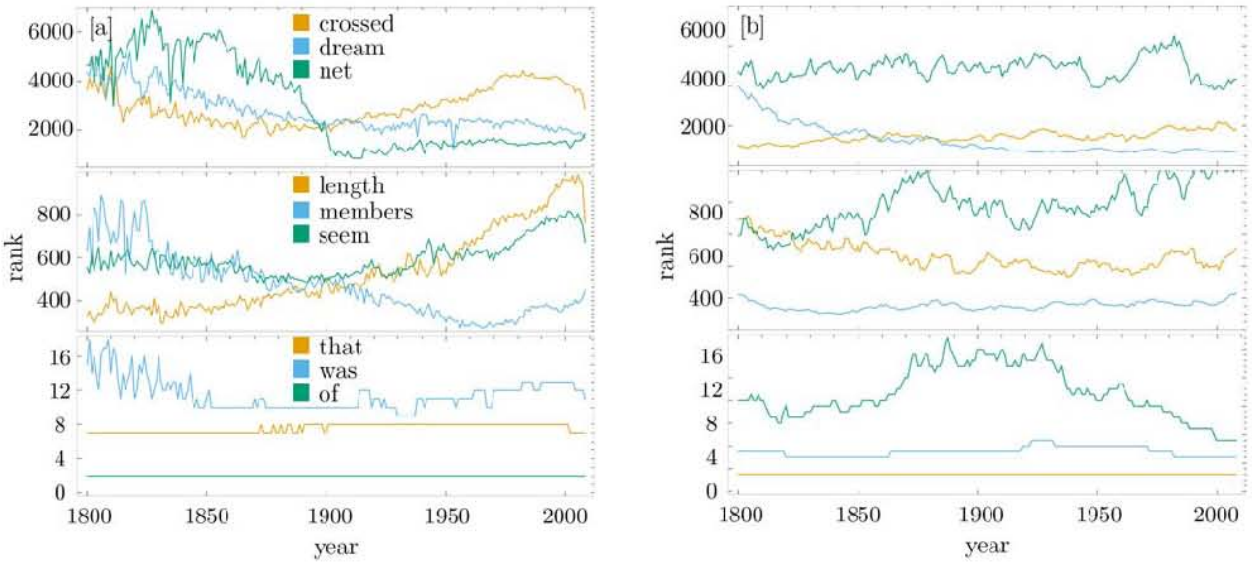

Fig 4. Rank evolution. [a]: Evolution of the rank for several particular, but random words in different regimes in the English language. From bottom to top we show words with initial ranks of order 1 (head), 100 (body) and 1000 (tail). [b]: Evolution of the rank for several particular, but random words in different regimes, for our scale-free Gaussian walker, i.e. the simulated language we have generated.

doi:10.1371/journal.pone.0121898.g004

A random walk model for rank diversity

We consider the relative size of frequency changes, or flights as they are sometimes called in statistical physics, defined as $(k_{t+1}-k_t)/k_t$ where k_t is the rank at discrete time t of a given element. We present in Fig. 5 the distribution of these frequency changes for English, our largest data set, and in Fig. S10 in S1 Text for all languages. Notice that, on average, the relative jumps seem to be largely independent of the value of the rank. We propose, based on this fact, a simple model to understand the evolution of rank diversity of words.

We shall call this model a scale-invariant random Gaussian walk, since a word with rank k_t, is converted to rank k_{t+1} according to the following procedure: One defines an auxiliary variable s_{t+1} at time t+1 by the relation

$$s_{t+1} = k_t + G(0, k_t\hat{\sigma}), \tag{3}$$

where $G(0, \tilde{\sigma})$ is a Gaussian random number generator of width $\tilde{\sigma}$ and mean 0. This means that the random variable s_{t+1} has a width distribution proportional to k_t. Words with very low ranks will change very slowly or not at all, while those with higher k have a larger rank variation in time, as reflected by $d(k)$. Once the values of s_{t+1} for all words are obtained, they are ordered according to their magnitude. This new order gives new rankings, *i.e.* the k values at time t+1. There is a small correlation of the jumps between different times in this model. This is consistent with the observed behavior of the six languages dealt with here, as can be seen in Fig. S11 in S1 Text. The only parameter in the model is the width $\hat{\sigma}$, which is the most common standard deviation of the relative frequency changes of each data set.

A word of caution must be said. In Fig. 5, two curves are plotted. In green, a Lorentzian distribution, and in red a Gaussian distribution, both centered at zero, and with a width obtained by best fit to the data presented here. Although the Lorentzian fits these data somewhat better than the Gaussian, we use the latter in our model, since the long tails of the Lorentzian would

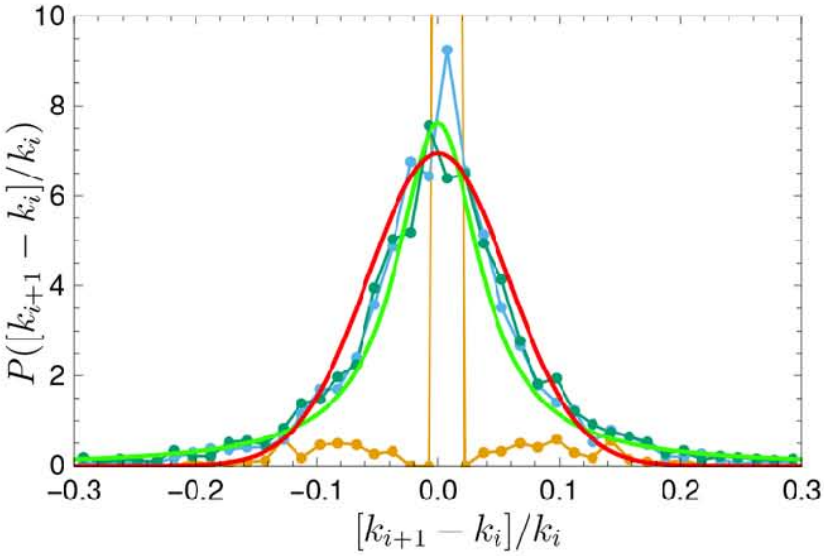

Fig 5. Distribution of relative size of frequency changes $[k_{t+1}-k_t]/k_t$ in the case of English for words in the head (gold) (that start with rank between 1 and 10), the body (blue)(rank between 200 and 210), and the tail (green) (rank between 5000 and 5010). Notice that for words in the head, the granularity of the model (Equation 3) shows up as large deviations from the Gaussian. For the body and tail, the relative jumps are similar independently of the initial rank of the word. We also show, as a thick green curve, the Lorentzian distribution which best fits the average of the curves for the body and tail. A Gaussian, with zero mean and the most common standard deviation $\hat{\sigma} = 0.0575$, is also shown in red for comparison (see text for details). The corresponding plot for other languages is shown in the supplementary information.

doi:10.1371/journal.pone.0121898.g005

yield long flights in words (not observed in the historical data) and a very different function $d(k)$. One should recall that the Lorentzian does not have a finite second moment, so this might be the reason for this distribution to be inadequate. It is probable that a truncated Lorentzian could be a better choice, but we leave this detail open as a possible refinement to our model.

With this model we have produced the evolution of a random simulated language; see [32] for other approaches. Fig. 4b shows examples of rank trajectories at different scales, exhibiting similarities with those of actual words shown in Fig. 4a. Moreover, if its diversity $d(k)$ is calculated with the $\hat{\sigma}$ corresponding to the most popular width of the distribution of relative size of flights for all words in the English language from 1800 to 2008, the results coincide with the sigmoid obtained for all six languages analyzed, as shown in Fig. 6.

Discussion

Within statistical linguistics, the frequency-rank distributions of several languages of European origin have been analyzed for many years now. However, no simple model can reproduce the detailed properties of this distribution (see SI). In particular, there has been the proposal that there exist two different regimes for ranks, but these regimes have not been satisfactorily validated in the empirical data. Due to these difficulties we have been led to introduce a statistical measure, which we have called rank diversity, to describe the statistical properties of natural languages. A simulated random language was generated which reproduces the observed features quite well.

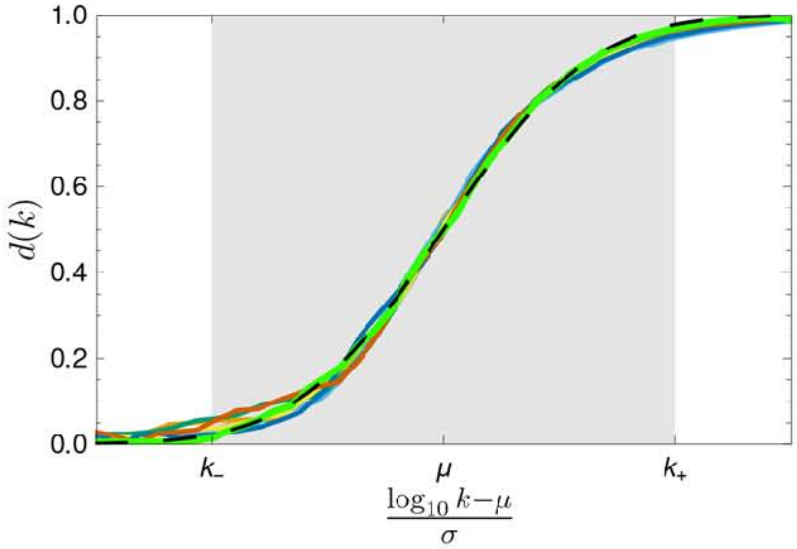

Fig 6. Rank diversity for the simulated language. The green curve represents the diversity corresponding to the language dynamics of a single realization of the Gaussian random walk model. We also include data for all languages studied, but normalized so that $k_{\pm}$ coincide. The ansatz for the rank diversity is plotted as a parameter-free cumulative of a Gaussian with zero mean and unit variance as a dashed black curve.

doi:10.1371/journal.pone.0121898.g006

Our random walk model mimics the evolution of languages to produce a simulated rank diversity which closely matches that of historical data. We consider that statistical similarities across languages and the simplicity of the model to reproduce them sufficient evidence to claim that rank diversity of words is universal. This does not imply that all languages have the same rank diversity curves, but that the rank diversity distribution of all the languages studied here can be fitted properly with Equation 1. Certainly, different languages have different curves that fit them better, just as different exponents fit better a Zipf distribution of different languages. For the languages studied, $1.6 \leq \mu \leq 2.1$ and $0.4 \leq \sigma \leq 0.6$.

This universality could be used to favor nativist explanations of human language [33, 34], where language is claimed to be determined by innate constraints. However, the high-ranked diversity of language tails could be used in favor of adaptationist explanations as well [35], as the precise rank of tail words is highly contingent. In recent years, explanations of human language relating biological evolution (genetically encoded innate properties) and learning (epigenetical adaptation) with culture have gained strength [36–38]. Even so, few assumptions are necessary to explain some general aspects of the evolution of human languages [39]. The present work shows that the evolution of word frequency can be explained with Gaussian random walks, where the size of the change in word frequency is proportional to its rank, *i.e.* frequent words change less than infrequent words. This explanation does not require innate properties, adaptive advantages, nor culture. This does not imply that the latter are irrelevant for other aspects of language evolution. Note that our study is carried out at a statistical level. We do not address syntactic, semantic, and grammatical aspects of human language [40–43], which are certainly important.

Why does the rank diversity approach a lognormal distribution? Which processes and mechanisms are required for this? There is one condition for a variable to have a lognormal distribution. This condition is that the variable should be the result of a high number of

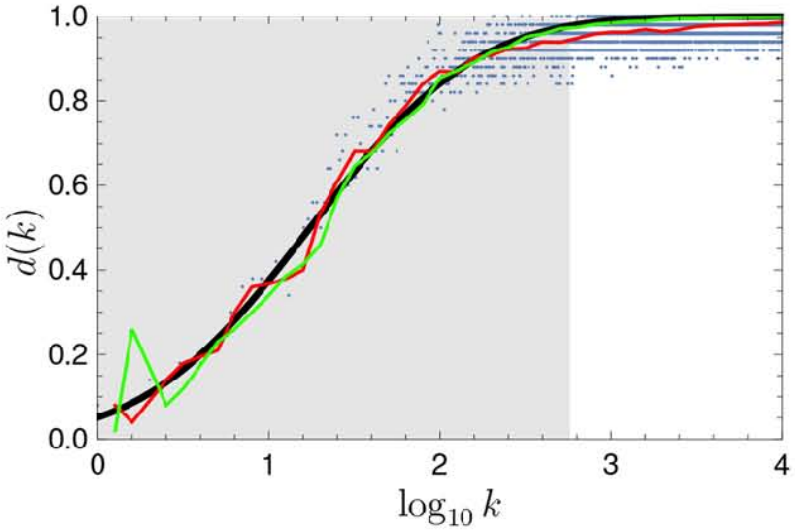

Fig 7. Rank diversity of male chess players obtained from the trimestral FIDE rankings from April, 2001 to May, 2012 ($\Delta t = 50$), considering the first 10,000 ranks. Blue dots show rank diversity, windowed in the red line. The black line shows the sigmoid fit with $\mu = 1.24$ and $\sigma = 0.76$. The green line shows a simulation with $\hat{\sigma} = 0.18$. Notice that there is no head as $\mu - 2\sigma < 0$. This is to be expected, as many players enter and leave the ranking during the years considered.

doi:10.1371/journal.pone.0121898.g007

different and independent causes which produce positive effects composed multiplicatively. Thus, each cause has a negligible effect on the global result [44]. Our Gaussian random walk model supports this as a suitable explanation: the statistical distribution of d is always lognormal, there is a high number of components (words), each word has a negligible effect compared to the language properties, *i.e.* large changes in word frequency (ranking) do not cause large changes in the statistical properties of each language, and the rank of each word is partially a cumulative product of its rank in previous times, as expressed in Equation 3. Languages statistically comply with these dynamics, and that serves as an explanation for their evolution and structure.

In future work, it will be relevant to study the rank diversity of n-grams with $n > 1$ [45], other linguistic corpora and phenomena with dynamic rank distributions [27, 46–48] and more generally with temporal networks [49–52]. A specific example would be the ranking of chess players, given by the World Chess Federation (Fédération Internationale des Échecs). The rank diversity in this case is provided in Fig. 7, which shows that the sigmoid is appropriate also for this case.

Supporting Information

S1 Text. Figure S1. Rank distributions of words according to frequency. [a]: Normalized word frequency f_R as a function of the rank k for several languages for books published in the year 2000. The color code for languages is as follows: (light blue) for French, (green) for German, (yellow) for Italian, (orange) for English, (dark blue) for Spanish, and (red) for Russian. [b]: Word frequency f_R as a function of the rank k for English and several years, normalized so that the most frequent element has relative frequency one. In the inset, the unnormalized frequency f is shown.

Figure S2. Comparison between the different models, Equations S1–S5, and the frequency of rank distribution. We use the data for the year 2000 and all languages under consideration. The logarithm base 10 of the ratio of the observed values and the model is plotted. It can be appreciated that different models fit better in different regions. However there is no model that fits all languages and all regions much better than the others.
Figure S3. Rank variations in time of twenty words from three different scales for English.
Figure S4. Rank variations in time of twenty words from three different scales for German.
Figure S5. Rank variations in time of twenty words from three different scales for French.
Figure S6. Rank variations in time of twenty words from three different scales for Italian.
Figure S7. Rank variations in time of twenty words from three different scales for Spanish.
Figure S8. Rank variations in time of twenty words from three different scales for Russian.
Figure S9. Rank variations in time of twenty words from three different scales for our simulated language.
Figure S10. Distribution of relative flights for all languages studied. A similar plot as the one presented in Fig. 5 is shown for other languages. The same color coding and details are used.
Figure S11. Correlations for relative frequency changes for different languages. Black line shows correlations for simulated language.
(PDF)

Acknowledgments

We are grateful to the editor and the two anonymous referees for their useful comments.

Author Contributions

Conceived and designed the experiments: GC JF CG CP SS. Performed the experiments: GC JF CG CP SS. Analyzed the data: GC JF CG CP SS. Contributed reagents/materials/analysis tools: GC JF CG CP SS. Wrote the paper: GC JF CG CP SS.

References

1. Zipf GK (1932) Selective Studies and the Principle of Relative Frequency in Language. Cambridge, MA, USA: Harvard University Press.
2. Mandelbrot B (1953) An informational theory of the statistical structure of language. In: Jackson, W, editor, Communication Theory, the Second London Symposium, London: Betterworth, chapter 36. pp. 486–502. URL http://www.uvm.edu/~pdodds/files/papers/others/1953/mandelbrot1953a.pdf.
3. Hawkins JA, Gell-Mann M, editors (1992) The Evolution of Human Languages: Proceedings of the Workshop on the Evolution of Human Languages, Held August, 1989 in Santa Fe, New Mexico. Perseus Books.
4. Ferrer i Cancho R, Solé RV (2002) Zipf's law and random texts. Advances in Complex Systems 5: 1–6. doi: 10.1142/S0219525902000468
5. Baek SK, Bernhardsson S, Minnhagen P (2011) Zipf's law unzipped. New Journal of Physics 13: 043004. doi: 10.1088/1367-2630/13/4/043004
6. Corominas-Murtra B, Fortuny J, Solé RV (2011) Emergence of Zipf's law in the evolution of communication. Phys Rev E 83: 036115. doi: 10.1103/PhysRevE.83.036115
7. Perc M (2012) Evolution of the most common English words and phrases over the centuries. Journal of The Royal Society Interface 9: 3323–3328. doi: 10.1098/rsif.2012.0491
8. Newman ME (2005) Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46: 323–351. doi: 10.1080/00107510500052444
9. Clauset A, Shalizi CR, Newman ME (2009) Power-law distributions in empirical data. SIAM Review 51: 661–703. doi: 10.1137/070710111
10. Petruszewycz M (1973) L'histoire de la loi d'Estoup-Zipf: documents. Mathématiques et Sciences Humaines 44: 41–56.

11. Auerbach F (1913) Das gesetz der bevölkerungskonzentration. Petermanns Geographische Mitteilungen 59: 74–76.
12. Booth AD (1967) A "law" of occurrences for words of low frequency. Information and Control 10: 386–393. doi: 10.1016/S0019-9958(67)90201-X
13. Montemurro MA (2001) Beyond the Zipf–Mandelbrot law in quantitative linguistics. Physica A: Statistical Mechanics and its Applications 300: 567–578. doi: 10.1016/S0378-4371(01)00355-7
14. Font-Clos F, Boleda G, Corral A (2013) A scaling law beyond Zipf's law and its relation to Heaps' law. New Journal of Physics 15: 093033. doi: 10.1088/1367-2630/15/9/093033
15. Gerlach M, Altmann EG (2013) Stochastic model for the vocabulary growth in natural languages. Phys Rev X 3: 021006.
16. Ferrer i Cancho R, Solé RV (2001) Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. Journal of Quantitative Linguistics 8: 165–173. doi: 10.1076/jqul.8.3.165.4101
17. Bochkarev V, Solovyev V, Wichmann S (2014) Universals versus historical contingencies in lexical evolution. Journal of The Royal Society Interface 11: 20140841. doi: 10.1098/rsif.2014.0841
18. Takala S (1985) Estimating students' vocabulary sizes in foreign language teaching. In: Practice and Problems in Language Testing, Afinla, volume 8. pp. 157–165. URL https://www.jyu.fi/ hum/laitokset/solki/afinla/julkaisut/arkisto/40/takala.
19. Hall RA (1953) Haitian Creole: Grammar, Texts, Vocabulary. Philadelphia: American Folklore Society.
20. Romaine S (1988) Pidgin and Creole Languages. London: Longman.
21. Beare K (2014) Voice of America Special English Dictionary. About.com English as 2nd Language. URL http://esl.about.com/cs/reference/a/aavoa.htm.
22. Hornby AS (2005) Oxford Advanced Learner's Dictionary. Oxford, UK: Oxford University Press. URL http://www.oxfordlearnersdictionaries.com/wordlist/english/oxford3000/ox3k_A-B/.
23. Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, et al. (2011) Quantitative analysis of culture using millions of digitized books. Science 331: 176–182. doi: 10.1126/science.1199644 PMID: 21163965
24. Wijaya DT, Yeniterzi R (2011) Understanding semantic change of words over centuries. In: Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web. ACM, pp. 35–40.
25. Serrà J, Corral Á, Boguñá M, Haro M, Arcos JL (2012) Measuring the evolution of contemporary western popular music. Scientific Reports 2: 521. doi: 10.1038/srep00521 PMID: 22837813
26. Petersen AM, Tenenbaum J, Havlin S, Stanley HE (2012) Statistical laws governing fluctuations in word use from word birth to word death. Scientific Reports 2: 313. doi: 10.1038/srep00313 PMID: 22423321
27. Blumm N, Ghoshal G, Forró Z, Schich M, Bianconi G, et al. (2012) Dynamics of ranking processes in complex systems. Physical Review Letters 109: 128701. doi: 10.1103/PhysRevLett.109.128701 PMID: 23005999
28. Acerbi A, Lampos V, Garnett P, Bentley RA (2013) The expression of emotions in 20th century books. PLoS ONE 8: e59030. doi: 10.1371/journal.pone.0059030 PMID: 23527080
29. Perc M (2013) Self-organization of progress across the century of physics. Scientific Reports 3: 1720. doi: 10.1038/srep01720
30. Febres G, Jaffe K, Gershenson C (2014) Complexity measurement of natural and artificial languages. Complexity Early View.
31. Hernández H (1988) Hacia un modelo de diccionario monolingüe del espaáol para usuarios extranjeros. In: Actas del Primer Congreso Nacional de ASELE. pp. 159–166. URL http://cvc.cervantes. es/ensenanza/biblioteca_ele/asele/pdf/01/01_0307.pdf.
32. Steels L (1997) The synthetic modeling of language origins. Evolution of Communication 1: 1–34. doi: 10.1075/eoc.1.1.02ste
33. Chomsky N (1965) Aspects of the Theory of Syntax. Massachusetts Institute of Technology. M.I.T. Press. URL http://books.google.com.mx/books?id = u0ksbFqagU8C.
34. Hauser M, Chomsky N, Fitch W (2002) The faculty of language: What is it, who has it, and how did it evolve? Science 298: 1569. doi: 10.1126/science.298.5598.1569 PMID: 12446899
35. Pinker S, Bloom P (1990) Natural language and natural selection. Behavioral and Brain Sciences 13: 707–727. doi: 10.1017/S0140525X00081061
36. Kirby S (1999) Function, Selection, and Innateness: The Emergence of Language Universals. Oxford University Press.

37. Kirby S, Dowman M, Griffiths TL (2007) Innateness and culture in the evolution of language. Proceedings of the National Academy of Sciences 104: 5241–5245.
38. Chater N, Reali F, Christiansen MH (2009) Restrictions on biological adaptation in language evolution. Proceedings of the National Academy of Sciences 106: 1015–1020.
39. Nowak MA, Krakauer DC (1999) The evolution of language. Proceedings of the National Academy of Sciences 96: 8028–8033.
40. Steels L (1995) A self-organizing spatial vocabulary. Artificial Life 2: 319–332. doi: 10.1162/artl.1995.2.3.319 PMID: 8925502
41. Sandler W, Meir I, Padden C, Aronoff M (2005) The emergence of grammar: Systematic structure in a new language. Proceedings of the National Academy of Sciences of the United States of America 102: 2661–2665.
42. Gell-Mann M, Ruhlen M (2011) The origin and evolution of word order. Proceedings of the National Academy of Sciences 108: 17290–17295.
43. Beuls K, Steels L (2013) Agent-Based Models of Strategies for the Emergence and Evolution of Grammatical Agreement. PLoS ONE 8: e58960+. doi: 10.1371/journal.pone.0058960 PMID: 23527055
44. Brockmann D, Helbing D (2013) The hidden geometry of complex, network-driven contagion phenomena. Science 342: 1337–1342. doi: 10.1126/science.1245200 PMID: 24337289
45. Ha LQ, Sicilia-Garcia EI, Ming J, Smith FJ (2002) Extension of Zipf's law to words and phrases. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING. pp. 315–320.
46. Batty M (2006) Rank clocks. Nature 444: 592–596. doi: 10.1038/nature05302 PMID: 17136088
47. Braha D, Bar-Yam Y (2006) From centrality to temporary fame: Dynamic centrality in complex networks. Complexity 12: 59–63. doi: 10.1002/cplx.20156
48. Hausmann R, Hidalgo CA, Bustos S, Coscia M, Simoes A, et al. (2014) The Atlas of Economic Complexity: Mapping Paths to Prosperity. MIT Press.
49. Gross T, Sayama H, editors (2009) Adaptive networks: Theory, Models and Applications. Understanding Complex Systems. Berlin Heidelberg: Springer. URL http://dx.doi.org/10.1007/978-3-642-01284-6.
50. Gautreau A, Barrat A, Barthélemy M (2009) Microdynamics in stationary complex networks. Proceedings of the National Academy of Sciences 106: 8847–8852.
51. Perra N, Gonçalves B, Pastor-Satorras R, Vespignani A (2012) Activity driven modeling of time varying networks. Scientific Reports 2: 469. doi: 10.1038/srep00469 PMID: 22741058
52. Holme P, Saramäki J (2012) Temporal networks. Physics Reports 519: 97–125. doi: 10.1016/j.physrep.2012.03.001

Bibliografía

[1] George Kingsley Zipf. *Selected studies of the principle of relative frequency in language*. Harvard Univ. Press, 1932.

[2] Trevor J Barnes and Matthew W Wilson. Big data, social physics, and spatial analysis: The early years. *Big Data & Society*, 1(1):2053951714535365, 2014.

[3] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[4] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[5] Mark EJ Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[6] Martin Gerlach and Eduardo G. Altmann. Stochastic model for the vocabulary growth in natural languages. *Phys. Rev. X*, 3:021006, May 2013.

[7] R Alvarez-Martínez, G Cocho, RF Rodríguez, and G Martínez-Mekler. Birth and death master equation for the evolution of complex networks. *Physica A: Statistical Mechanics and its Applications*, 402(C):198–208, 2014.

[8] Albert Sydney Hornby. *Oxford Advanced Learner's Dictionary*. Oxford University Press, Oxford, UK, 2005.

[9] Sauli Takala. Estimating students' vocabulary sizes in foreign language teaching. In *Practice and Problems in Language Testing*, volume 8, pages 157–165. Afinla, 1985.

[10] Kenneth Beare. *Voice of America Special English Dictionary*. About.com English as 2nd Language, 2014.

[11] Humberto Hernández. Hacia un modelo de diccionario monolingüe del español para usuarios extranjeros. In *Actas del Primer Congreso Nacional de ASELE*, pages 159–166, 1988.

[12] Charles Carpenter Fries. The structure of english: An introduction to english sentences, 1952.

[13] Kenneth Beare. Content and function words. `http://esl.about.com/od/learningtechniques/a/Content-And-Function-Words.htm`, 2014. [Online; accesado 2014-12-22].

[14] Merriam-Webster. Function word. `http://www.merriam-webster.com/dictionary/function+word`, 2014. [Online; accesado 2014-12-22].