UNIVERSIDAD DE MÁLAGA

PROGRAMA DE DOCTORADO EN MATEMÁTICAS

Sobre el comportamiento complejo de las palabras relevantes en textos: heterogeneidad espacial y correlaciones de largo alcance

CONCEPCIÓN CARRETERO CAMPOS

Tesis Doctoral

DIRECTORES
DR. PEDRO J. CARPENA SÁNCHEZ
DRA. ANA V. CORONADO JIMÉNEZ

Universidad de Málaga
2024

AUTORA: Concepción Carretero Campos

https://orcid.org/0000-0001-6501-6095

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña CONCEPCIÓN CARRETERO CAMPOS
Estudiante del programa de doctorado MATEMÁTICAS de la Universidad de Málaga, autor/a de la tesis, presentada para la obtención del título de doctor por la Universidad de Málaga, titulada: SOBRE EL COMPORTAMIENTO COMPLEJO DE LAS PALABRAS RELEVANTES EN TEXTOS: HETEROGENEIDAD ESPACIAL Y CORRELACIONES DE LARGO ALCANCE

Realizada bajo la tutorización de DR. CARLOS M. PARÉS MADROÑAL y dirección de DR. PEDRO J. CARPENA SÁNCHEZ Y DRA. ANA V. CORONADO JIMÉNEZ (si tuviera varios directores deberá hacer constar el nombre de todos)

DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a 30 de SEPTIEMBRE de 2024

Fdo.: CONCEPCIÓN CARRETERO CAMPOS Doctorando/a	Fdo.: CARLOS M. PARÉS MADROÑAL Tutor/a
Fdo.: PEDRO J. CARPENA SÁNCHEZ Y ANA V. CORONADO JIMÉNEZ Director/es de tesis	

Edificio Pabellón de Gobierno. Campus El Ejido. 29071
Tel.: 952 13 10 28 / 952 13 14 61 / 952 13 71 10
E-mail: doctorado@uma.es

Málaga, 30 de septiembre de 2024

Los abajo firmantes, **Pedro Juan Carpena Sánchez**, con D.N.I. *********, Doctor en Ciencias Físicas y Catedrático de Universidad del Departamento de Física Aplicada II de la Universidad de Málaga, y **Ana Victoria Coronado Jiménez**, con D.N.I. *********, Doctora en Ciencias Físicas y Profesora Titular de Universidad del Departamento de Física Aplicada II de la Universidad de Málaga, por la presente

CERTIFICAN:

Que Dª **Concepción Carretero Campos**, Licenciada en Matemáticas por la Universidad de Málaga, ha realizado en el Programa de Doctorado en Matemáticas bajo nuestra dirección el trabajo titulado:

Sobre el comportamiento complejo de las palabras relevantes en textos: heterogeneidad espacial y correlaciones de largo alcance

Revisado el mencionado trabajo, estimamos que puede ser defendido ante el Tribunal que ha de juzgarlo. Y para que conste donde sea necesario, autorizamos su presentación en la Universidad de Málaga

Málaga, 30 de septiembre de 2024

El abajo firmante, **Carlos María Parés Madroñal**, con D.N.I. *********, Doctor en Matemáticas y Catedrático de Universidad del Departamento de Análisis Matemático, Estadística e I.O y Matemática Aplicada de la Universidad de Málaga, por la presente

CERTIFICA:

Que Dª **Concepción Carretero Campos**, Licenciada en Matemáticas por la Universidad de Málaga, ha realizado en el Programa de Doctorado en Matemáticas bajo su tutorización el trabajo titulado:

Sobre el comportamiento complejo de las palabras relevantes en textos: heterogeneidad espacial y correlaciones de largo alcance

Revisado el mencionado trabajo, estima que puede ser defendido ante el Tribunal que ha de juzgarlo. Y para que conste donde sea necesario, autoriza su presentación en la Universidad de Málaga

A la memoria de mi madre.

Agradecimientos

Esta tesis doctoral se ha realizado en el marco de los proyectos P07-FQM-03163 de la Junta de Andalucía y FIS2012-36282 del Ministerio de Ciencia e Innovación, a los que agradezco su financiación.

Agradezco a mi director, el Dr. Pedro J. Carpena Sánchez, todo lo que me ha enseñado a lo largo de estos años. Gracias por compartir conmigo tus conocimientos y tu pasión por la investigación. A mi directora, la Dra. Ana V. Coronado Jiménez, gracias por haber querido incorporarte a la dirección de esta tesis y por tu disposición para lo que necesitase. A ambos, gracias por abrirme la puerta de vuestra casa y por toda vuestra paciencia.

Al Dr. Pedro A. Bernaola Galván y al Dr. Manuel Gómez Extremera, por sus contribuciones a mi aprendizaje y por interesarse por el progreso de este trabajo.

Al Dr. Carlos María Parés Madroñal, gracias por responder siempre amablemente a todos mis correos, por alegrarte de los avances y por acoger de buen grado los trámites administrativos.

Al Dr. Marcelo Montemurro, gracias por recibirme en la Universidad de Manchester y emplear tu tiempo en trabajar e intercambiar conocimientos conmigo.

A los que fueron mis compañeros del departamento de Física Aplicada II, por todos los ratos compartidos al inicio de este camino. Especialmente, Cristina, gracias por tantísimos años de amistad, de los que ya hasta perdemos la cuenta.

A mis compañeros del departamento de Matemática Aplicada, con los que he compartido paradas que me han llevado por otras sendas, con los que estoy compartiendo el final de este camino y con los que espero compartir el inicio de otros muchos, por todo el cariño que he recibido de ellos desde el primer día. A los que se han preocupado y ocupado, a los que me han acompañado y con los que he disfrutado de la maravillosa tarea de ser docente. Mención especial a mis chicas, Yolanda y Yadira, por darme el empujón que necesitaba y no dejarme sola en el camino. Y Alejandra, por llegar en el momento más oportuno y hacer que me sienta acompañada en las largas tardes de los últimos meses.

A Antonio, por conocer y sufrir esta memoria como si fuera la suya propia, y estar a mi lado en todo el proceso.

A mis alumnos, por hacer que disfrute tanto de este trabajo y desee que la universidad sea siempre mi lugar.

A mi familia, que es un tesoro. A mis padres, por la riqueza de la educación que me han proporcionado, tanto en el ámbito académico como en el personal: *"Mamá, aspiro a ser la docente y profesional que tú eras"*. A mis hermanos: *"Os quiero, es un regalo teneros"*. A mi tía, la Dra. Concha Campos, por ser una segunda madre para nosotros y el espejo en el que me miro en este camino académico. A mis primas, Alicia y Cristina, que son como hermanas, por su compañía en cada paso.

A mis amigas de siempre, que se emocionan por mis logros y que están en los momentos más importantes, ya sean malos o buenos, y por mucho que las complicaciones de la vida nos dificulten el vernos con asiduidad.

A Noemí, por tantas cosas valiosas que me ha enseñado en estos años y por ser pieza fundamental para haber podido llegar a este momento.

Los caminos muchas veces no son rectos, y nos encontramos con más obstáculos de los que preveíamos, o necesitamos hacer paradas que duran mucho más de lo esperado. En caminos así, acabas encontrando mucha gente que te acompaña, algunas con las que coincides en ciertas etapas y otras que están siempre ahí sin importarles si al final llegas a la meta porque, en ciertos momentos, hay cosas mucho más importantes que esa meta. Me considero afortunada por la gran valía humana de las personas que me rodean.

A todas ellas, GRACIAS.

Resumen

En las últimas décadas, el estudio del lenguaje y, más concretamente, de los textos escritos, ha sido objeto de investigaciones desde el punto de vista de la física estadística. Se conoce que las palabras relevantes, es decir, las que transmiten la información más importante sobre el contenido de un texto, presentan una distribución espacial heterogénea y se concentran en determinadas regiones del texto (en las que se desarrolla principalmente esa idea) formando agrupamientos o *clusters*. Cuanto más heterogénea sea la distribución espacial y más se aleje de la esperada por azar, mayor es la relevancia de la palabra. La conexión entre *clustering* y relevancia ha sido usada con éxito para detectar y extraer automáticamente las palabras relevantes (*keywords*) de un texto sin disponer de ninguna información previa sobre el mismo, y sin necesidad de un *corpus* de referencia. Medidas definidas para cuantificar *clustering* y otros tipos de aproximaciones funcionan adecuadamente cuando el texto es largo.

Con el objetivo de mejorar la detección en textos cortos, se han implementado algunas modificaciones a medidas de *clustering* ya definidas en la literatura, y se han realizado comparaciones con una medida basada en el cálculo de entropía por medio de la definición de métricas adecuadas para la evaluación y comparación de detectores de palabras clave. Posteriormente, se ha definido una nueva medida consecuencia de haber obtenido analíticamente la distribución exacta para las distancias entre apariciones sucesivas de una palabra esperada por azar. Esta distribución es válida independientemente de la frecuencia de aparición de la palabra considerada y de la longitud del texto bajo estudio. Sin embargo, en aproximaciones anteriores, se usaba como referencia para la distribución espacial esperada por azar la distribución geométrica, lo que es cierto sólo en el caso asintótico y, por tanto, afecta en mayor medida a palabras con baja frecuencia y en textos cortos. Se obtiene, entonces, que la nueva medida de *clustering* definida mejora la detección de palabras clave en textos cortos y, además, se estudian los valores de *clustering* extremos, lo que permite diferenciar entre palabras clave genéricas y específicas.

Además de la detección de palabras clave, otro foco de investigación son las correlaciones en textos. Se conoce que los textos tienen una distribución espacial compleja que da

lugar a la existencia de correlaciones de largo alcance, las cuales se han cuantificado para distintos autores y en diferentes idiomas. Es bien sabido que muchos sistemas físicos y biológicos (secuencias de ADN, dinámica cardíaca, música, señales sísmicas, etc.) presentan correlaciones de largo alcance, y que caracterizarlas y cuantificarlas permite comprender e interpretar la dinámica del sistema. La fuerte auto-atracción observada en la distribución espacial de palabras relevantes hace pensar que las correlaciones de largo alcance presentes en textos se deben a sus palabras clave. Dada una palabra en un texto concreto, se han cuantificado las correlaciones presentes en la secuencia binaria que represente sus apariciones, y se ha obtenido un vínculo claro entre las correlaciones de largo alcance y la relevancia de la palabra, pudiéndose usar el grado de correlaciones de largo alcance a escala grande para cada palabra como medida de relevancia. Las palabras comunes, que tienen una distribución homogénea, no contribuyen a las correlaciones presentes en el texto.

Las correlaciones de largo alcance mencionadas previamente sugieren modelar la manera en la que una palabra aparece en un texto mediante un proceso estocástico con correlaciones de largo alcance. Se realiza un estudio numérico sistemático de las distribuciones de los tiempos de paso por cero de procesos con correlaciones en ley de potencias, y se encuentran tres regímenes distintos entre los que se observa un régimen en el que la distribución sigue una *stretched exponential*, comportamiento que se había observado para la distribución de las distancias entre apariciones sucesivas de una palabra a escalas grandes. Se propone entonces un modelo capaz de reproducir la distribución espacial de una palabra en un texto, basado en las correlaciones de largo alcance observadas para la palabra e incorporando un factor de repulsión a escala corta. Se observa que, con un ajuste adecuado de los parámetros, el modelo reproduce no solo el grado de correlaciones de la secuencia binaria que representa sus apariciones, sino también la distribución espacial de la distancias entre apariciones de la palabra a todas las escalas y su grado de relevancia cuantificado mediante medidas de *clustering*. De este modo, se han conectado varios tópicos de investigación en textos (detección de palabras clave, cuantificación de correlaciones de largo alcance y estudio de la distribución de las distancias entre apariciones sucesivas de una palabra) y hemos encontrado una manera interesante de mostrar las dinámicas del lenguaje escrito: repulsión a corta escala y atracción contextual a escala grande para las palabras relevantes. Estas reglas conllevan que las palabras relevantes presenten una distribución espacial heterogénea que da lugar a *clusters* y correlaciones de largo alcance. Seguir trabajando en esta línea nos dará un conocimiento más profundo de las dinámicas subyacentes a la comunicación escrita.

Gran parte de los resultados obtenidos pueden aplicarse a otros tipos de secuencias simbólicas.

Índice general

Índice de tablas

Índice de figuras

Introducción

La física estadística de sistemas complejos es un área activa de investigación con múltiples aplicaciones en ámbitos tan diversos como son medicina, climatología, neurociencia y economía, entre otros. El uso de herramientas matemáticas combinado con el desarrollo de la capacidad computacional permite abordar problemas de gran complejidad y obtener conclusiones sobre la dinámica de sistemas complejos mediante desarrollos analíticos (cuando son posibles) y/o simulaciones numéricas. Entre esos objetos de estudio podemos destacar las dinámicas presentes en el lenguaje escrito, donde se puede enmarcar el desarrollo de esta investigación.

Esta tesis surge a partir de los trabajos de [Ortuño et al., 2002] y, posteriormente, [Carpena et al., 2009], en los que, siguiendo una aproximación similar a la que se empleaba en el estudio de las estadísticas de niveles del espectro de sistemas cuánticos desordenados, se analiza la distribución de palabras en un texto literario.

Se considera un texto como una secuencia simbólica, en la que cada una de las palabras distintas que lo componen es un símbolo diferente. Fijada una palabra concreta (símbolo concreto), las posiciones en las que aparece a lo largo del texto constituyen su 'espectro'. Se observa que el espectro de las palabras presenta un comportamiento diferente dependiendo del tipo de palabra. Para las palabras más informativas sobre el contenido del texto, que llamaremos palabras relevantes, la distribución espacial a lo largo de la secuencia presenta un comportamiento que denominaremos *clustering* o agrupamiento. Las posiciones en las que aparecen a lo largo de la secuencia no son aleatorias, tienen grandes fluctuaciones de frecuencia y tienden a agruparse formando *clusters*. Esto lleva a trabajar en las siguientes líneas:

I) Si la relevancia de una palabra está relacionada con su distribución espacial, esta puede ser un punto de partida para detectar las palabras más relevantes de un texto, comúnmente denominadas "palabras clave" (*keywords*). Como veremos posteriormente, la detección de palabras clave ha sido un campo amplio de investigación, en el que las primeras aproximaciones estaban basadas en la frecuencia de aparición de las palabras y solían necesitar de un conjunto de documentos de referencia (*cor-*

pus). La idea sería detectar las palabras más relevantes a partir de un único texto y sin tener ninguna información previa, solo por la manera en la que se distribuyen. Llamaremos medida de *clustering* a una medida que cuantifique cuánto se separa la distribución espacial de una palabra de la esperada si la palabra se distribuye al azar. La hipótesis es que a mayor *clustering*, mayor heterogeneidad de la distribución espacial y, en consecuencia, mayor relevancia de la palabra para el texto considerado.

II) El comportamiento de las palabras relevantes se puede interpretar también desde el punto de vista de un sistema complejo. Los elementos del sistema (las palabras) interaccionan entre sí a diferentes niveles dando lugar a una estructura compleja. Las fuertes interacciones de una palabra relevante consigo misma, nos llevan a considerar si una secuencia que represente sus apariciones a lo largo del texto presenta correlaciones de largo alcance. La hipótesis es que las correlaciones de largo alcance serán también una medida de relevancia y que, a partir de ellas, se puede plantear un modelo que reproduzca la manera en la que una palabra se "escribe" mediante procesos estocásticos con correlaciones de largo alcance.

En la primera línea de trabajo, basándonos en las medidas de *clustering* propuestas en [Ortuño et al., 2002; Carpena et al., 2009], tendremos como objetivo la mejora de la detección de las palabras clave en textos cortos, que es donde esas y otras aproximaciones proporcionan resultados menos precisos. Dado que queremos cuantificar cuánto se separa la distribución espacial de una palabra de lo esperado por azar, resulta fundamental qué distribución se va a considerar como referencia para una palabra distribuida al azar. A lo largo de la literatura se asumía como referencia la distribución geométrica, pero veremos que es solo correcta asintóticamente y obtendremos la distribución espacial exacta esperada por azar en el caso de un tamaño de texto y un número de apariciones de la palabra finitos (aplicable, en general, a cualquier secuencia simbólica). A partir de ahí podremos establecer una cota de *clustering* extremo, que permitirá clasificar las palabras relevantes en genéricas y específicas. Por otro lado, también plantearemos métricas adaptadas a la detección de palabras clave para poder comparar de la forma más objetiva posible distintas aproximaciones.

Y en la segunda línea, tendremos como objetivo comprobar si podemos establecer un vínculo entre la relevancia de una palabra y las correlaciones de largo alcance de una secuencia que represente sus apariciones. Para ello haremos una revisión de los métodos que se usan para cuantificar correlaciones de largo alcance y, especialmente, del *Detrended Fluctuation Analysis* y su aplicación a secuencias binarias. Y, finalmente, tendremos como objetivo el uso de procesos estocásticos con correlaciones de largo alcance para modelar la

forma en las que las palabras se distribuyen a lo largo del texto y reproducir sus principales propiedades.

Estructura de la tesis y publicaciones asociadas

La tesis consta de dos partes diferenciadas en las que se desarrollarán las dos líneas de trabajo mencionadas previamente.

La primera parte abarca los capítulos 1 al 3, organizados como sigue:

- En el **capítulo 1**, realizamos una revisión bibliográfica del problema de la detección de palabras clave en textos literarios y de las medidas de relevancia que se usarán en los capítulos posteriores.

- En el **capítulo 2**, nos centramos en implementar modificaciones que mejoren la detección de palabras clave en textos cortos, y en la comparación entre medidas de *clustering* y medidas basadas en entropía, por medio de la definición de métricas de evaluación adaptadas a este problema.

  El contenido de este capítulo corresponde a la publicación en revista indexada en JCR

  > **C. Carretero-Campos**, P. Bernaola-Galván, A. V. Coronado y P. Carpena,
  >
  > *Improving statistical keyword detection in short texts: entropic and clustering approaches.*
  >
  > Physica A 392, 1481 (2013).

- En el **capítulo 3**, obtenemos la distribución espacial esperada por azar para un símbolo concreto en una secuencia simbólica, analizamos sus propiedades y la aplicamos al problema de la detección de palabras clave.

  El contenido de este capítulo corresponde a la publicación en revista indexada en JCR

  > P. Carpena, P. A. Bernaola-Galván, **C. Carretero-Campos**, y A. V. Coronado,
  >
  > *Probability distribution of intersymbol distances in random symbolic sequences: Applications to improving detection of keywords in texts and of amino acid clustering in proteins.*
  >
  > Physical Review E 94, 052302 (2016).

La segunda parte abarca los capítulos 4 al 6, organizados como sigue:

- En el **capítulo 4** introducimos los conceptos teóricos asociados a procesos estocásticos con correlaciones de largo alcance que serán la base para los capítulos posteriores. Prestaremos especial atención a los métodos para generar (*Método de Filtrado de Fourier*) y cuantificar (*Detrended Fluctuacion Analysis*) tales correlaciones en secuencias reales y binarias.

- En el **capítulo 5** nos centramos en el estudio de tiempos de primer paso en procesos con correlaciones de largo alcance en ley de potencias. Sus propiedades estadísticas son fundamentales para describir la dinámica del sistema complejo subyacente. Estudiamos cómo dichas propiedades dependen de la fortaleza de las correlaciones del sistema.

  El contenido de este capítulo corresponde a la publicación en revista indexada en JCR

  > **C. Carretero-Campos**, P. Bernaola-Galván, P. Ch. Ivanov and P. Carpena,
  >
  > *Phase transitions in the first-passage time of scale-invariant correlated processes.*
  >
  > Physical Review E 85, 011139 (2012).

  con algunas secciones que provienen de la publicación

  > P. Carpena, A. V. Coronado, **C. Carretero-Campos**, P. Bernaola-Galván and P. Ch. Ivanov,
  >
  > *First-Passage Time Properties of Correlated Time Series with Scale-Invariant Behavior and with Crossovers in the Scaling.*
  >
  > Time Series Analysis and Forecasting, Springer International Publishing Switzerland (2016).

- En el **capítulo 6** establecemos un vínculo entre la detección de palabras clave, los estudios que reportan la presencia de correlaciones de largo alcance en textos literarios y los modelos que se plantean para reproducir la distribución espacial heterogénea observada para las palabras relevantes. Planteamos el uso del exponente de correlación a escalas grandes como medida de relevancia y la propuesta de un proceso con tales correlaciones como modelo de aparición de una palabra a lo largo del texto.

  El contenido de este capítulo corresponde a la publicación

C. Carretero-Campos, M. A. Montemurro, P. Bernaola-Galván, A. V. Coronado and P. Carpena,

Towards a deeper understanding of the complex behaviour observed in the distribution of words in written texts

en Proceedings of the European Conference on Complex Systems 2012. Springer International Publishing Switzerland (2013).

El trabajo en los contenidos de la tesis también ha dado lugar a las siguientes contribuciones a congresos:

- **C. Carretero Campos**, P. Bernaola-Galván, P. Ch. Ivanov y P. Carpena, *Propiedades del primer tiempo de paso en ruidos fractales correlacionados*. XXXII Bienal de Física (2009). Ciudad Real (España).
- **C. Carretero Campos**, P. Bernaola-Galván, A. V. Coronado y P. Carpena, *Correlaciones de largo alcance en secuencias binarias*. NoLineal 2010 (2010). Cartagena (España).
- **C. Carretero Campos**, P. Bernaola-Galván, A. V. Coronado y P. Carpena, *Detección automática de palabras clave en textos cortos*. XVII Congreso de Física Estadística (FisEs) (2011) Barcelona (España)
- **C. Carretero Campos**, M. A. Montemurro, P. Bernaola-Galván, A. V. Coronado y P. Carpena, *Correlaciones de largo alcance y palabras clave en textos*. NoLineal 2012 (2012). Zaragoza (España).
- **C. Carretero Campos**, M. A. Montemurro, P. Bernaola-Galván, A. V. Coronado and P. Carpena, *Towards a deeper understanding of the complex behaviour observed in the distribution of words in written texts*. The European Conference on Complex Systems 2012. Bruselas (Bélgica).
- P. Carpena, P. Bernaola-Galván, A.V. Coronado, **C. Carretero Campos**, M. Hackenberg and J.L. Oliver, *On the complex distribution of relevant words in the human genome*. The European Conference on Complex Systems 2012 (Satellite Meeting: Genomic Complexity) (2012). Bruselas (Bélgica).
- P. Carpena, A.V. Coronado, **C. Carretero Campos**, P. Bernaola-Galván and P. Ch. Ivanov, *First-passage time properties of scale-invariant correlated processes and of correlated processes with scaling crossovers*. NDES (2014), Albena (Bulgaria).

Parte I

Detección de palabras clave en textos literarios

Capítulo 1

Revisión bibliográfica

A lo largo de este capítulo realizaremos un acercamiento al problema de la detección automática de palabras clave en textos literarios. En primer lugar, consideraremos el interés general del problema y el objetivo con el que surge la idea de abordarlo, cuyo estudio constituirá la primera parte de este trabajo. A continuación, describiremos distintas aproximaciones a dicho problema presentes en la literatura.

Una vez realizada esta breve revisión, estaremos en disposición de adentrarnos en los dos siguientes capítulos. En el primero de ellos describiremos mejoras aplicadas a una medida bien establecida y referenciada en la bibliografía [Carpena et al., 2009] desarrollada para detectar palabras clave, así como métricas para evaluarla y compararla con otro tipo de aproximaciones. Y en el segundo mostraremos resultados analíticos que permitirán clasificar las palabras clave en dos grupos, genéricas y específicas.

1.1. Interés del problema y aplicaciones

En un mundo altamente digitalizado, la detección automática de palabras clave de un texto literario es un problema de relevancia y con múltiples aplicaciones. Por palabras clave entendemos aquellas que contienen la información más característica sobre el contenido del texto, y que nos transmiten las principales ideas sobre las que trata, las cuales permiten, por ejemplo, clasificar e indexar documentos y localizar información de nuestro interés. El objetivo fundamental de este trabajo será la detección de dichas palabras sin disponer de ninguna información previa acerca del contenido del texto (sin conocer el tipo de texto, su temática, etc.), únicamente mediante la forma en la que dichas palabras se distribuyen a lo largo del mismo, lo cuál está relacionado con la manera en la que se transmite la información. Cómo extraer esas palabras a partir de sus propiedades estadísticas sin recurrir a ninguna información externa ha sido un problema clave en las ciencias de la

información.

Este tipo de problema presenta dos claros focos de interés. Por un lado, el meramente teórico, en el cual se aborda la cuestión de pasar de elementos abstractos como las palabras clave, los textos y sus relaciones, a resultados cuantificables. Por otro lado, existe un interés evidente en las aplicaciones de este tipo de investigación: en la sociedad actual, de la era digital, puede ser una herramienta importante el análisis de textos masivos, por ejemplo, provenientes de internet, así como las aplicaciones a inteligencia artificial para reconocimiento de textos, bien sean científicos, artísticos, jurídicos, etc.

Podemos entender un texto de longitud N como un caso particular de una secuencia simbólica $\{S_1, S_2, ..., S_N\}$ de N símbolos (las palabras) que provienen de un alfabeto de m símbolos (número de palabras distintas: vocabulario), $S_i \in \{A_1, A_2, ..., A_m\}$. Veremos que una de las principales propiedades de un símbolo dado es su estructura espacial, es decir, cómo se distribuye a lo largo de la secuencia.

De esta forma, los resultados que se obtengan podrían ser aplicables a otras secuencias simbólicas como, por ejemplo, las secuencias de ADN, el texto biológico por excelencia. El ADN se puede considerar como un larguísimo texto sin comas ni espacios, del que se pretendería delimitar y localizar las palabras o motivos que lo componen (vocabulario) e investigar qué proporción de ellas pueden desempeñar algún papel en el control de la expresión génica y tienen función biológica (las cuales serían, por analogía, las palabras clave del ADN). Dicho objetivo no forma parte de esta memoria, pero explica el porqué de las características de la herramienta desarrollada (que no hace uso de información lingüística, no necesita un corpus de referencia y tiene la posibilidad de aplicarse a un texto sin espacios).

Cuando buscamos las palabras clave de un texto y desarrollamos algoritmos para detectarlas, tenemos la posibilidad de evaluar los resultados ya que podemos conocer a priori si una palabra es relevante o no y, en algunos casos, porque dispondremos además de un glosario que contiene las palabras más importantes del texto (y que nos permitirá hacer una evaluación más objetiva). Esto permite testear los algoritmos y conocer si están funcionando correctamente, lo cual justifica por qué para un objetivo como es la búsqueda de regiones con información biológica en la secuencia de ADN se plantean como punto de partida los textos literarios. La validez de la analogía lingüística mencionada y de la aplicabilidad de los métodos desarrollados en textos al ADN se puede ver en [Hackenberg et al., 2012].

Concluimos que, además de las claras aplicaciones que tiene para la recuperación de información el desarrollo de algoritmos que detecten de manera no supervisada las palabras clave de un texto, podemos encontrar otras aplicaciones menos inmediatas y más complejas que motivan dicho estudio.

1.2. Medidas de relevancia: estado del arte

En la literatura se han propuesto distintas aproximaciones al problema de la detección de palabras clave, con diferentes enfoques y estrategias. El objetivo es definir medidas que cuantifiquen el grado de relevancia de todas las palabras del texto y que, como consecuencia, nos permitan extraer las más relevantes o simplemente establecer un *ranking*.

Describiremos las primeras aproximaciones a este problema, que se basaban en un análisis de la frecuencia de ocurrencia de las palabras en el texto, para después centrarnos en dos estrategias diferentes desarrolladas en los últimos años para detectar palabras clave en un texto sin necesidad de usar un corpus externo, y que llamaremos aproximación basada en agrupamiento o *clustering*[1] y aproximación entrópica, respectivamente. Estas dos propuestas son las que hemos usado como referencias principales en los capítulos siguientes. Por último, mencionaremos otras propuestas presentes en la literatura.

1.2.1. Medidas basadas en la frecuencia de ocurrencia

El estudio de frecuencias de palabras en un texto comienza con el lingüista norteamericano George Kingsley Zipf, que formula la conocida ley de Zipf, ley empírica que establece que si ordenamos las palabras por su frecuencia[2] de aparición en el texto (de mayor a menor frecuencia), se tiene que [Zipf, 1949]:

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

con $\alpha \approx 1$, siendo $f(r)$ la frecuencia de aparición de la palabra y r su posición en el *ranking*. En la novela *Ulysses*, del escritor James Joyce, por ejemplo, la décima palabra más frecuente ($r = 10$) aparece 2653 veces ($f = 2653$), o la palabra en la posición 50 del *ranking* ($r = 50$) aparece 556 veces ($f = 556$).

Cuanto mayor sea la posición en el *ranking* de una palabra, menor es su frecuencia, ya que se han ordenado así. Necesariamente $f(r)$ es una función decreciente, pero además se observa empíricamente que se puede ajustar a una ley de potencias con exponente aproximadamente -1.

Se ha observado también esta relación no solo en textos literarios sino al estudiar otros sistemas como el tráfico en internet, la población de ciudades, en música, finanzas, fenómenos naturales,... [Li, 2002]. El rango de aplicabilidad de esta ley, su posible origen

[1] La RAE admite clúster como adaptación gráfica propuesta para la voz inglesa *cluster*, pero no contiene adaptación para el anglicismo *clustering*, que es el usado en la literatura para referirse a ese tipo de medidas.

[2] Salvo que se indique lo contrario, por frecuencia nos referiremos a la frecuencia absoluta: número de apariciones de la palabra en el texto.

y las desviaciones con respecto a la misma han sido estudiadas en la literatura [Moreno-Sánchez et al., 2016].

La primera propuesta para cuantificar la relevancia de una palabra en un texto se debe a [Luhn, 1958] que ya en 1958, teniendo como objetivo la creación automática de resúmenes, propone que la frecuencia de ocurrencia de una palabra en un texto proporciona una medida útil de su relevancia.

El objetivo del trabajo de Luhn es la creación de auto-resúmenes, principalmente de artículos de áreas de ciencia y tecnología. Estarán compuestos por frases del artículo seleccionadas según su contenido informativo. Se requiere por tanto una medida que permita comparar el contenido informativo de las frases, para determinar cuáles constituirán el auto-resumen. La medida se deriva de un análisis de las palabras que constituyen las frases, para lo cual se necesita previamente establecer un conjunto de las palabras significativas o relevantes para el texto.

Como hemos mencionado, Luhn propone usar la frecuencia de una palabra para medir su significación para el texto. La justificación de dicha propuesta la basa en el hecho de que normalmente un escritor repite ciertas palabras a medida que va desarrollando un tópico, lo cual sería un indicativo de su relevancia para el mismo. Todas las palabras del texto son ordenadas de mayor a menor frecuencia. Si representamos la frecuencia de cada palabra frente a su posición en el *ranking* se obtiene una curva como la que se muestra en la figura 1.1, como anticipaba la ley de Zipf. Las palabras con frecuencia muy alta, normalmente palabras comunes, se eliminan de la lista, así como las de frecuencia muy baja. El resto son consideradas palabras clave o significativas.

Una vez establecido el conjunto de palabras significativas, la relevancia de cada frase se mide a partir del número de palabras significativas que contiene y la proximidad entre ellas, cuantificando así la representatividad de la información contenida en la frase para el artículo. Aquellas con mayor puntuación constituyen el resumen.

Sin embargo, las limitaciones de la aproximación de Luhn son bien conocidas [Salton and McGill, 1986]. De hecho, aunque las aproximaciones basadas en un análisis de la frecuencia funcionan adecuadamente cuando se dispone de una colección de documentos como referencia (corpus) [Salton and McGill, 1986; Bookstein and Swanson, 1974, 1975; Harter, 1975a,b; Berger and Lafferty, 1999], no son suficientes para analizar un sólo texto sin información adicional sobre el mismo, que es nuestro objetivo fundamental. Si se dispone de un corpus, se puede comparar si una palabra que aparece con frecuencia alta en un texto, lo hace también en los demás, y distinguir palabras funcionales de palabras clave.

Cuando no disponemos de un corpus de referencia con el que poder comparar, la información proporcionada por la frecuencia no es muy útil. De hecho, si generamos

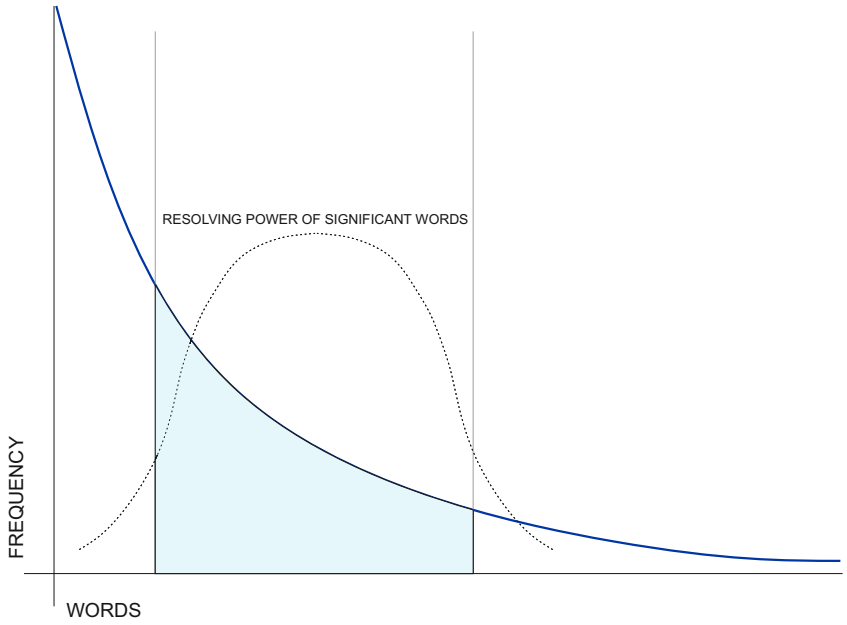

Figura 1.1: Frecuencia de cada palabra frente a su posición en el ranking. Basada en [Luhn, 1958].

textos aleatorios barajando las palabras del texto original, conservamos la frecuencia de ocurrencia de cada palabra (y se sigue cumpliendo la ley de Zipf), pero hemos destruido completamente la información. Ninguna palabra sería ahora relevante independientemente de su frecuencia. Obviamente, ello se debe a que la información que proporciona una palabra está controlada, no solo por su frecuencia, sino también por su estructura o distribución espacial a lo largo del texto correspondiente, como veremos a continuación.

1.2.2. Medidas basadas en *clustering*

Partiendo de la idea de que la información está contenida no sólo en las palabras en sí mismas, sino en el orden en el que están dispuestas a lo largo del texto, [Ortuño et al., 2002] mostraron que existe una relación crucial entre la relevancia de una palabra y su distribución espacial. Considerando el texto como una secuencia de N símbolos, entenderemos por distribución espacial de una palabra (símbolo concreto) con frecuencia n, la distribución de sus posiciones $j_1, j_2, ..., j_n$ ($j_i \in \{1, 2, ..., N\}$) a lo largo de la secuencia.

- Las palabras relevantes, es decir, aquellas que están más estrechamente relacionadas con los principales tópicos del texto, presentan una distribución muy inhomogénea. Normalmente se concentran en determinadas regiones del texto, presentando grandes fluctuaciones de frecuencia. Usando un lenguaje físico, podríamos decir que las

diferentes apariciones de una palabra relevante presentan un grado alto de auto-atracción, el cual da lugar a regiones con frecuencia alta de aparición (que denominaremos *clusters*) y regiones donde la palabra raramente aparece. El origen de esta auto-atracción está relacionado con la estructura de la información: un concepto importante se usa más a menudo en las regiones del texto donde se discute, y apenas aparece cuando se analiza otro concepto diferente.

- Por el contrario, las palabras comunes (artículos, preposiciones, etc.) presentan una distribución bastante homogénea a lo largo de todo el texto. Usando de nuevo un lenguaje físico, diríamos que las diferentes apariciones de una palabra no relevante no interactúan entre ellas, y por tanto aparecen distribuidas de forma prácticamente aleatoria.

Como ejemplo mostramos en la figura 1.2 dos palabras con frecuencia similar, el sustantivo 'Quijote' y la conjunción 'pero', del libro *Don Quijote* de Miguel de Cervantes. Observamos la distribución bastante homogénea de la palabra 'pero', palabra no informativa sobre el contenido del texto, frente a la distribución *clusterizada* de la palabra 'Quijote'. Observamos en este caso cómo, cuando se trabaja sin un corpus de referencia, la distribución espacial permite detectar relevancia, mientras que el valor de la frecuencia no. Ambas palabras tienen una frecuencia similar en el texto completo (alrededor de 2150) y también en la parte mostrada en la figura.

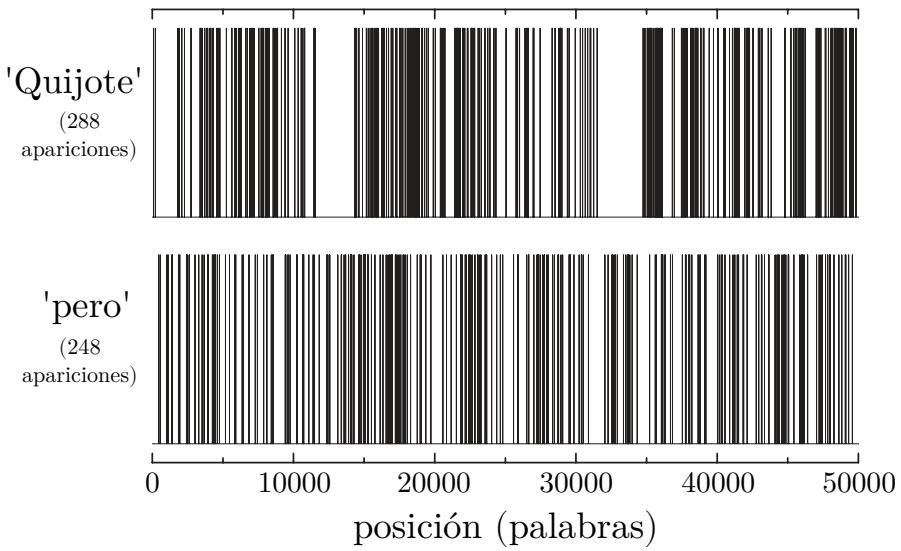

Figura 1.2: Posiciones del sustantivo 'Quijote' y la conjunción 'pero' a lo largo de las primeras 50000 palabras del libro *Don Quijote* de Miguel de Cervantes.

Estos conceptos físicos de auto-atracción o ausencia de interacción provienen del análi-

sis de niveles de energía de sistemas cuánticos desordenados, donde herramientas similares a la que presentaremos posteriormente fueron empleadas para analizar si los niveles de energía presentaban atracción, repulsión o ausencia de interacción (ver [Carpena et al., 2004] y las referencias que contiene).

Usando la conexión observada entre el *clustering* de una palabra y su relevancia, [Ortuño et al., 2002] definieron un método bastante efectivo para la detección automática de palabras clave basado en el análisis estadístico de las distribuciones de las distancias entre las sucesivas apariciones de una palabra. La idea es cuantificar cuánto se separa dicha distribución de lo esperado si la palabra se distribuye aleatoriamente, es decir, si las n posiciones en las que aparece $(j_1, j_2, ..., j_n)$, se escogen al azar entre las N posibles. La hipótesis es que a mayor separación con respecto al azar, mayor variabilidad de la distribución y, como consecuencia, mayor relevancia.

Partiendo de un texto de longitud N y de una palabra con frecuencia de ocurrencia n, denotaremos por $d_1, d_2, ..., d_{n-1}$ las distancias entre las sucesivas apariciones de la palabra a lo largo del texto, es decir, $d_i = j_{i+1} - j_i$, donde $i = 1, 2, ..., n-1$, siendo $j_1, j_2, ..., j_n$ las posiciones de la palabra. [Ortuño et al., 2002] propusieron usar el coeficiente de variación

$$\sigma = \frac{s}{\langle d \rangle} \tag{1.2}$$

como medida de la relevancia de la palabra, siendo $\langle d \rangle$ la distancia media entre sus apariciones y s la desviación estándar ($s = \sqrt{\langle d^2 \rangle - \langle d \rangle^2}$). Para cada palabra tendremos asociado un valor de σ. Se asume que el valor $\sigma = 1$ es el esperado para una palabra distribuida al azar [3]. La hipótesis es que a mayor valor de σ, más diferencia con respecto al azar, y mayor relevancia tendrá la palabra para el texto analizado. Se observa que, por ejemplo, en la versión inglesa de la Biblia, la palabra 'christ', que aparece $n = 571$ veces, tiene un valor de $\sigma = 18.42$, frente a la palabra 'king' para la que se obtiene un valor inferior de $\sigma = 8.15$, aunque su frecuencia es mayor ($n = 2542$). Se dirá que una palabra con $\sigma > 1$ presenta *clustering* y una palabra con $\sigma < 1$ repulsión (véase la figura 1.3).

Aunque este método resulta ser bastante eficiente en la detección de palabras clave, se ha comprobado que presenta algunas debilidades que conllevan identificaciones incorrectas [Zhou and Slater, 2003; Carpena et al., 2009]. Algunas han sido corregidas en [Carpena et al., 2009] definiendo una nueva medida denotada C que describiremos posteriormente. Tales debilidades pueden enumerarse como sigue:

1. No todas las palabras distribuidas al azar tienen el mismo nivel de *clustering* medido

[3] Por analogía con los espectros energéticos de sistemas desordenados. Puesto que la energía de esos espectros son aleatorias, la distribución normalizada a media 1 de separaciones entre energías consecutivas es la exponencial, $p(d) = \exp(-d)$, para la que $\sigma = 1$.

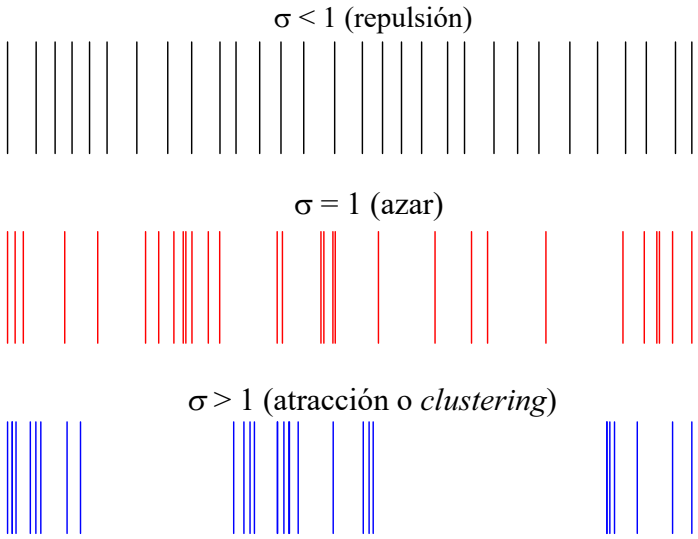

Figura 1.3: Repulsión, distribución aleatoria y *clustering* medido por σ.

por σ, ya que esta depende de su probabilidad $p = n/N$ de aparición en el texto.

Como se muestra en la figura 1.4, simulando la aparición de una palabra con probabilidad p en un texto aleatorio [4], y calculando el valor de σ en cada caso, se observa una dependencia del valor de p: en todos los casos σ crece con n y alcanza un valor asintótico para n grande, pero éste será diferente para cada p. El valor asintótico sólo será próximo a 1 cuando p tiende a 0. Por ejemplo, una palabra que aparece $n = 200$ veces con un valor de $\sigma = 0.96$ diríamos que presenta *clustering* si su probabilidad de aparición es $p = 0.1$ (porque el valor de σ es mayor al esperado por azar en ese caso), pero no así si $p = 0.05$, por ejemplo.

Para solucionarlo, en [Herrera and Pury, 2008; Carpena et al., 2009], se incorpora a la medida la dependencia de p, asumiendo que el valor esperado de σ para una palabra con probabilidad p distribuida aleatoriamente a lo largo de una secuencia de tamaño N es $\sigma = \sqrt{1-p}$. Este valor esperado se obtiene considerando la distribución geométrica como referencia para una palabra distribuida al azar[5]. Se define ahora la medida

$$\sigma_{nor} = \frac{\sigma}{\sqrt{1-p}} \tag{1.3}$$

En la figura 1.5 se observa cómo σ_{nor} no presenta dependencia de p ya que las curvas

[4]Se simulan textos aleatorios como secuencias binarias en las que el símbolo "1" aparece al azar con probabilidad p, representando la aparición de una palabra a lo largo del texto.

[5]Nótese que en [Carpena et al., 2016a] se deriva la distribución exacta para n apariciones de una palabra distribuidas al azar en un texto de longitud N y su correspondiente coeficiente de variación. Comentaremos las ventajas de usar dicha distribución exacta para textos cortos en lugar de la geométrica en el Capítulo 3.

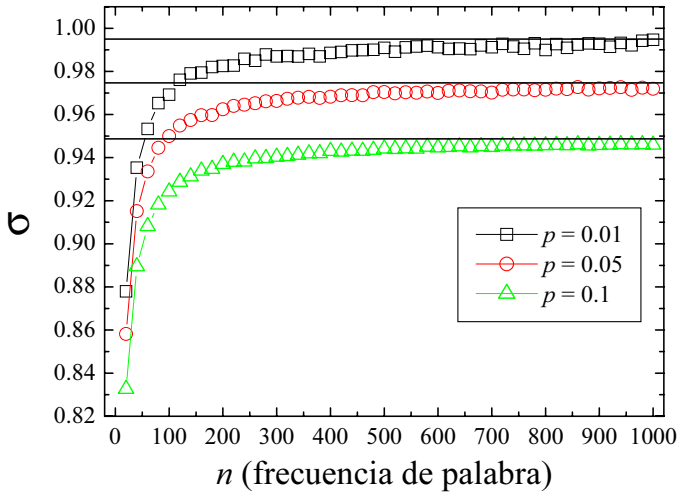

Figura 5: Comportamiento del valor medio de σ en textos aleatorios simulados para una palabra que aparece con distintos valores de probabilidad p en función de la frecuencia de aparición n de la palabra en el texto. Cada punto representa el valor medio obtenido para $10^{10}/n$ simulaciones. Las lineas horizontales corresponden al valor teórico $\sigma_{\text{geo}} = \sqrt{1-p}$.

problema, proponemos en este proyecto una primera mejora a la medida de clustering σ, consistente en normalizar σ por el valor esperado en una secuencia aleatoria, $\sqrt{1-p}$, definiendo así la medida σ_{nor}:

$$\sigma_{\text{nor}} = \frac{\sigma}{\sqrt{1-p}} \tag{8}$$

Esta medida corrige completamente los problemas que acabamos de mencionar: por un lado hace que cualquier palabra aleatoria, independientemente de su p, tenga el mismo valor de σ_{nor} para el mismo número de apariciones de la palabra, tal y como podemos comprobar en la figura 3, donde hemos considerado las mismas simulaciones mostradas en la figura 2 pero representando ahora σ_{nor}. Nótese como las curvas colapsan perfectamente, aún correspondiendo a palabras con distinto valor de p. Por otro lado, esta medida hace que para textos largos, se verifique que $\sigma_{\text{nor}} = 1$ corresponda a una palabra aleatoria, $\sigma_{\text{nor}} > 1$ indique una palabra con clustering y $\sigma_{\text{nor}} < 1$ signifique repulsión.

$$\langle \sigma_{ser} \rangle = \frac{2n-1}{2n+2} \qquad sd(\sigma_{ser}) = \frac{1}{\sqrt{n}\left(1+2.8n^{-0.85}\right)} \tag{1.4}$$

Teniendo en cuenta las debilidades mencionadas y las posibilidades de mejora, los autores definen la medida C de la siguiente forma:

$$C\left(\sigma_{nor}, n\right) = \frac{\sigma_{nor} - \langle \sigma_{nor} \rangle (n)}{sd(\sigma_{nor})(n)} \tag{1.5}$$

De esta manera C combina la información proporcionada por el *clustering* σ_{nor} y por la frecuencia n. Ahora, $C = 0$ indica que la palabra se distribuye al azar y, a mayor valor de C, mayor es el *clustering* y, por tanto, la relevancia. Valores negativos de C indican

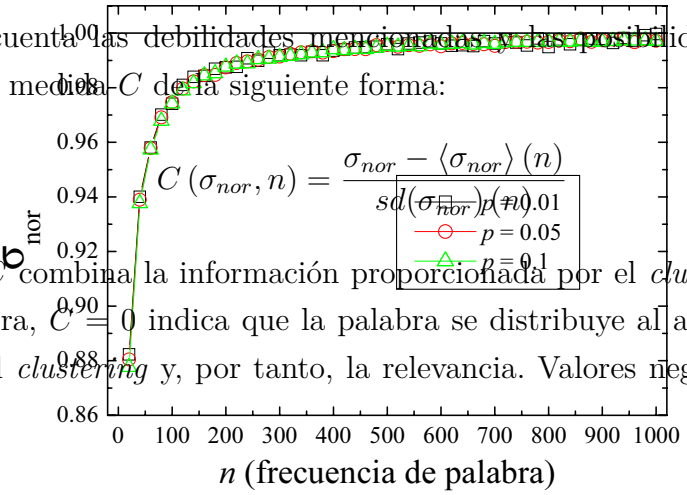

Figura 6: Comportamiento del valor medio de σ_{nor} en textos aleatorios simulados para una palabra que aparece con distintos valores de probabilidad p en función de la frecuencia de aparición n de la palabra en el texto. Cada punto representa el valor medio obtenido para $10^{10}/n$ simulaciones. El solapamiento

número de apariciones de la palabra, tal y como podemos comprobar en la figura 3, donde hemos considerado las mismas simulaciones mostradas en la figura 2 pero representando ahora σ_{nor}. Nótese como las curvas colapsan perfectamente, aun correspondiendo a palabras con distinto valor de p. Por otro lado, esta medida hace que para textos largos, se verifique que $\sigma_{\text{nor}} = 1$ corresponda a una palabra aleatoria, $\sigma_{\text{nor}} > 1$ indique una palabra con clustering y $\sigma_{\text{nor}} < 1$ signifique repulsión.

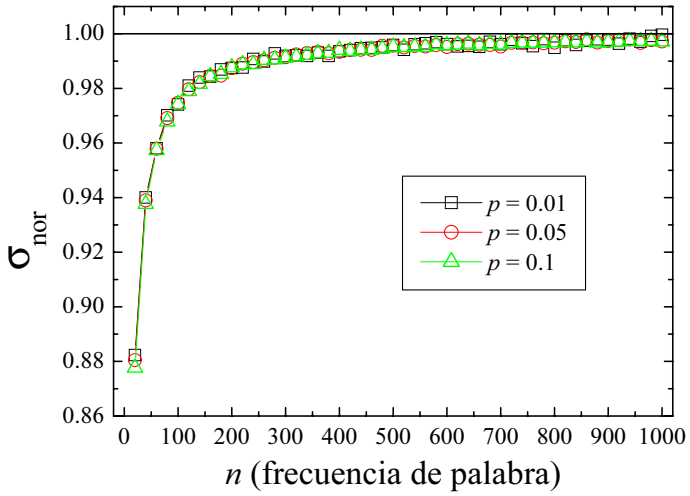

Figura 6: Comportamiento del valor medio de σ_{nor} en textos aleatorios simulados para una palabra que aparece con distintos valores de probabilidad p en función de la frecuencia de aparición n de la palabra en el texto. Cada punto representa el valor medio obtenido para $10^{10}/n$ simulaciones. El solapamiento de las curvas muestra como σ_{nor} elimina el efecto de p, y además como en el caso de textos largos, una palabra aleatoria alcanza el valor $\sigma_{\text{nor}} = 1$.

Figura 1.3: Comportamiento del valor medio de σ_{nor} en textos aleatorios simulados para una palabra que aparece con distintos valores de probabilidad p en función de la frecuencia de aparición n de la palabra en el texto. Cada punto representa el valor medio obtenido para $10^{10}/n$ simulaciones. El solapamiento de las curvas muestra como σ_{nor} elimina el efecto de p y además como, en el caso de textos largos, una palabra aleatoria alcanza el valor $\sigma_{\text{nor}} = 1$.

repulsión.

Siguiendo también la hipótesis de [Ortuño et al., 2002], en [Zhou and Slater, 2003] incorporan información secuencial, dando lugar a una medida que detecta el incremento de *clustering* y no se ve afectada por una única aparición inusual de una palabra en el texto.

Nótese, como comentamos previamente, que la relación entre una distribución *clusterizada* y la relevancia, no es sólo cierta para palabras en textos, sino también para cadenas cortas de nucleótidos en secuencias de ADN, donde el *clustering* está relacionado con funciones biológicas [Hackenberg et al., 2010, 2011, 2012].

1.2.3. Medidas basadas en la entropía

Otro método diferente para detectar palabras clave, basado en cuantificar el contenido de información de la secuencia de ocurrencias de cada palabra a lo largo del texto, y que usa la entropía de Shannon [Shannon and Weaver, 1949] para su definición, fue propuesto por [Herrera and Pury, 2008].

Se considera una partición de un texto de longitud N en P partes. Para cada palabra w se calcula la medida $E_{nor}(w)$, que cuantificará su relevancia para el texto, mediante el

siguiente procedimiento:

Una medida de probabilidad sobre la partición $\{p_i(w)\}$ se puede definir como sigue

$$p_i(w) = \frac{f_i(w)}{\sum_{j=1}^{P} f_j(w)} \quad (i = 1, ...P), \tag{1.6}$$

donde $f_i(w)$ es la frecuencia relativa de ocurrencia de la palabra w en la i-ésima parte. La entropía de Shannon de esta distribución viene dada por la expresión

$$S(w) = -\frac{1}{\ln(P)} \sum_{i=1}^{P} p_i(w) \ln(p_i(w)). \tag{1.7}$$

Para eliminar la dependencia de $S(w)$ de la frecuencia, teniendo en cuenta que para una palabra que aparece n veces en un texto aleatorio $S_{ran} \approx 1 - \frac{P-1}{2n \ln(P)}$, se define la medida normalizada

$$E_{nor}(w) = \frac{1 - S(w)}{1 - S_{ran}(w)}. \tag{1.8}$$

De este modo, las palabras distribuidas aleatoriamente tendrán valores de E_{nor} cercanos a 1. Por el contrario, cuanto mayor sea el valor de E_{nor}, mayor será la heterogeneidad de la distribución de la palabra a lo largo del texto y por tanto su relevancia.

Nótese que el número de partes P en las que se divide el texto, y la elección de cómo realizar dicha división, influye completamente en los resultados de E_{nor}, como veremos en el Capítulo 3. Ello hace que la medida E_{nor} conlleve cierto grado de arbitrariedad en la elección de P o un conocimiento previo de la estructura del texto.

Una vez definida la medida E_{nor}, en [Herrera and Pury, 2008] se realiza una comparación de dicho método con las aproximaciones basadas en el *clustering* definidas en [Ortuño et al., 2002] y [Zhou and Slater, 2003]. Para ello, se analiza el texto *The Origin of Species* de Charles Darwin, usando sus capítulos como partición natural en el cálculo de la medida entrópica E_{nor}. En dicho artículo se muestra que la medida entrópica proporciona buenos resultados como detector de palabras relevantes y que su comportamiento es tan bueno o mejor que las basadas en el *clustering*. En dicha comparación no se usa la medida C definida por [Carpena et al., 2009], que hemos descrito en la sección anterior, ya que ésta fue publicada con posterioridad. En el Capítulo 2 compararemos los comportamientos de E_{nor} y C, especialmente en textos cortos.

1.2.4. Otras propuestas

Además de la medida entrópica y las técnicas de *clustering* descritas en la sección anterior, que usaremos como referencia en capítulos posteriores, en la literatura relativa

a este campo de investigación pueden encontrarse diversas propuestas que tratan aproximaciones relacionadas con, entre otros, métodos basados en grafos, redes, o distintas variantes de medidas de entropía.

Algunas referencias al respecto que pueden resultar de interés son [Montemurro and Zanette, 2002; Mehri and Darooneh, 2011; Mehri et al., 2019; Mihalcea and Tarau, 2004; Rose et al., 2010; Tohalino et al., 2023].

Capítulo 2

Detección de palabras clave en textos cortos. Métricas de evaluación

La medidas para la detección de palabras clave basadas en la cuantificación de *clustering*, C, y en el uso de la entropía de Shannon, E_{nor}, descritas en el capítulo anterior, proporcionan resultados satisfactorios cuando se aplican en textos largos. Sin embargo, algunas cuestiones importantes quedan aún sin resolver. Por un lado, el planteamiento de modificaciones enfocadas a mejorar la detección en textos cortos y, por otro, el uso de métricas apropiadas para cuantificar el comportamiento de las medidas.

El problema de la detección automática de palabras clave en textos cortos es especialmente importante desde el punto de vista práctico (artículos científicos, páginas web, etc). Y, debido al tamaño pequeño de la muestra, se esperan resultados peores en ambas medidas.

Con respecto a la cuantificación del comportamiento de los detectores de palabras clave, ha habido intentos de ir más allá de los resultados cualitativos. En [Herrera and Pury, 2008] usan una versión del glosario del libro *The Origin of Species* para identificar de una forma lo más objetiva posible las palabras relevantes para el texto. Esto permite comparar el comportamiento de diferentes medidas de relevancia mediante una adaptación de los conceptos de precisión y exhaustividad (*precision and recall*), conceptos que se usan habitualmente en el ámbito de recuperación de información.

En este capítulo presentamos algunas modificaciones para la medida de *clustering* C [Carretero-Campos et al., 2013] al objeto de mejorar resultados anteriores, enfocándonos principalmente en el ámbito de aplicación a textos cortos. Se lleva a cabo un estudio comparativo con la medida entrópica E_{nor} a cuyo fin se introducen nuevas métricas para la evaluación del comportamiento de detectores de palabras clave basadas en las necesidades de un usuario típico.

2.1. Proponiendo mejoras a una medida de *clustering*

En esta sección, en primer lugar, dado el interés de la detección correcta de palabras clave en textos cortos, analizaremos el comportamiento de C para palabras con frecuencia pequeña, que será la situación típica en un texto corto. A continuación, después de demostrar que la medida C, tal como está definida, no detecta una palabra relevante cuyas apariciones estén concentradas en una única región, nos plantearemos la incorporación de condiciones de contorno que posibiliten su detección.

2.1.1. Cómputo para palabras poco frecuentes

En el capítulo anterior vimos que la medida C se define con el objetivo de corregir la fuerte dependencia que presentaba la medida σ_{nor} de la frecuencia de ocurrencia de la palabra:

$$C\left(\sigma_{nor}, n\right) = \frac{\sigma_{nor} - \langle \sigma_{nor} \rangle \left(n\right)}{sd(\sigma_{nor})\left(n\right)} \tag{2.1}$$

Si una palabra aparece n veces en un texto, nos interesaba cuantificar cuánto se distancia su valor de σ_{nor} de lo esperado por azar. Para ello se utilizaban los ajustes

$$\langle \sigma_{nor}(n) \rangle = \frac{2n-1}{2n+2},\ sd(\sigma_{nor}(n)) = \frac{1}{\sqrt{n}\left(1 + 2.8n^{-0.865}\right)} \tag{2.2}$$

En la figura 2.1 podemos observar que, aunque en general el ajuste a dichas funciones es bastante bueno, pierde precisión para valores pequeños de n. Para $n \leq 30$ hay desviaciones con respecto al ajuste que sería interesante corregir para mejorar la detección en palabras que aparezcan poco, que es lo habitual en un texto corto.

Como solución proponemos emplear en el rango de frecuencia $n \leq 30$ el valor obtenido mediante simulación de textos aleatorios para cada valor individual de n, en lugar de los ajustes presentados en las ecuaciones 2.2.

Esta modificación en el cálculo de C no implica ningún coste computacional significativo. Las simulaciones de textos aleatorios necesarias para computar los valores de $\langle \sigma_{nor} \rangle$ y $sd(\sigma_{nor})$ para $n \leq 30$ se realizan sólo una vez y los valores resultantes se guardan. Cuando tengamos que analizar una palabra que aparece con una frecuencia $n \leq 30$ en cualquier texto, simplemente tendremos que mirar en la lista precalculada.

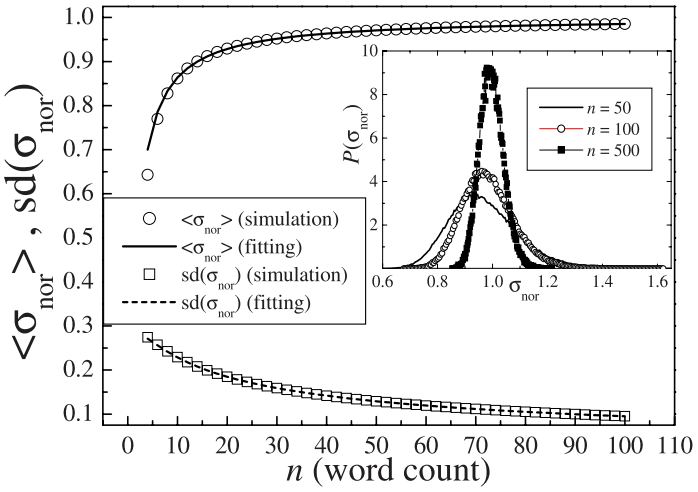

Figura 2.1: Simulación y ajustes de $\langle\sigma_{nor}\rangle$ y $sd(\sigma_{nor})$ para distintos valores de n

2.1.2. Condiciones de contorno

Como hemos comentado anteriormente, si una palabra aparece en un texto concentrada en una única región del mismo (formando un único *cluster*) es muy probable que la medida C no la detecte como una palabra relevante. Esto se debe a que el conjunto de distancias entre apariciones sucesivas de esta palabra estará formado por un conjunto homogéneo de distancias cortas (*intra-cluster*). Sin embargo, es inmediato pensar que palabras relevantes para un texto pueden comportarse de esa manera, como sería el caso de un tópico del que se hable mucho en una región del texto de la cual es un tema principal, pero que luego no se mencione más. Como ejemplo mostramos en la figura 2.2 las apariciones de la palabra 'wax'(cera) en el libro *The Origin of Species*. Esta palabra aparece únicamente en el capítulo 7, donde se estudia el comportamiento de las abejas.

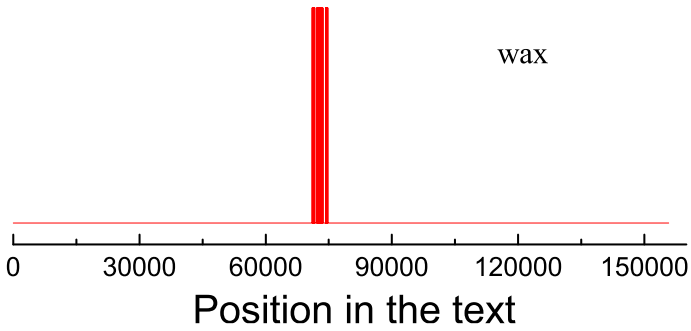

Figura 2.2: Posiciones de la palabra 'wax' ($n = 39$) en el libro *The Origin of Species*

[Zhou and Slater, 2003] ya observaron esta debilidad en la medida σ (que es previa a C) y propusieron como solución el uso de condiciones de contorno. Su propuesta consistía en tener en cuenta no sólo la región del texto en la que aparece la palabra, sino la longitud

total del mismo, añadiendo dos distancias artificiales: una que fuese desde el principio del texto hasta la primera vez que aparece la palabra, y otra entre la última ocurrencia de la palabra y el final del texto.

Siguiendo esta propuesta calculamos C de la manera descrita en las secciones anteriores, pero con la salvedad de que para cada palabra con frecuencia n ahora consideraremos un conjunto de $n+1$ distancias. Suponiendo que el texto tiene longitud total N, establecemos las posiciones 0 y $N+1$ como fronteras (ver figura 2.3, parte a)), lo cual en la práctica es equivalente a considerar para cada palabra analizada dos apariciones artificiales en las posiciones 0 y $N+1$. Tendremos ahora el conjunto de distancias $\{d_0, d_1, d_2, ..., d_{n-1}, d_n\}$, en el que hemos incluido las dos distancias nuevas d_0 y d_n, siendo $d_0 = p_i$ y $d_n = N+1-p_f$ con p_i y p_f la primera y última posición real de la palabra a lo largo del texto. A la medida C calculada mediante este procedimiento la denotaremos C_0.

Si una palabra se distribuye de manera homogénea a lo largo del texto, tal y como ocurre con las palabras que no son relevantes, las distancias d_0 y d_n serán típicamente del orden del resto de distancias del conjunto. Como consecuencia, el valor de *clustering* no se modificará de forma significativa, es decir, $C_0 \simeq C$. Sin embargo, para una palabra localizada en un único *cluster*, que (como hemos discutido más arriba) podría ser relevante, habrá una probabilidad alta de que d_0, d_n o ambas sean mayores que el resto de distancias del conjunto, lo cual incrementará el valor de *clustering* que teníamos para esa palabra, es decir, $C_0 > C$.

El uso de las condiciones de contorno que hemos descrito nos permite incluir en el conjunto de distancias información acerca de las regiones del texto en las que no aparece la palabra analizada. Sin embargo, dichas condiciones de contorno no son las únicas posibles. Tal y como es habitual cuando se estudian las propiedades de muchos sistemas físicos, nos plantemos también el efecto que tendría considerar condiciones de contorno periódicas. Al igual que antes, tendremos en cuenta la longitud total del texto, pero ahora conectamos el final del texto con el principio, es decir, consideramos un texto circular (ver Fig. 2.3, parte b)). En este caso, sólo incluimos una distancia adicional, que iría desde la última aparición de la palabra hasta la primera. Para una palabra con frecuencia n, consideraremos entonces el conjunto de distancias $\{d_1, d_2, ..., d_{n-1}, d_*\}$, en el que incluimos la distancia $d_* = N - p_f + p_i$, siendo de nuevo p_i y p_f la posición inicial y final de la palabra en el texto, y N la longitud total del mismo. A la medida C calculada mediante este procedimiento la denotaremos C_1.

También en este caso, como las palabras no relevantes se distribuyen de manera muy homogénea a lo largo del texto, la nueva distancia d_* será del orden de las demás distancias del conjunto, y el valor de clustering apenas se modifica, es decir, $C_1 \simeq C$. Sin embargo, para una palabra relevante localizada en un único *cluster*, la nueva distancia d_* será

mayor que cualquier otra del conjunto incrementando notablemente el valor de *clustering*, es decir, $C_1 > C$, lo que permite la detección de dicha palabra como relevante.

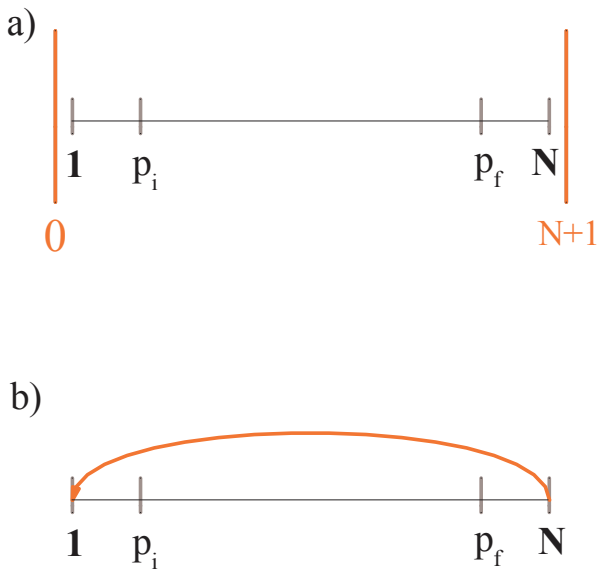

Figura 2.3: Esquema de los dos tipos de condiciones de contorno.

2.2. Reducción del glosario: preprocesamiento

Como ya hemos comentado, y debido a la existencia de un glosario para cuantificar la bondad de los resultados, usaremos para el análisis el libro 'The Origin of Species by means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life' de Charles Darwin (6th Edition) [1]. Siguiendo el procedimiento de [Herrera and Pury, 2008], eliminamos los comentarios sobre las diferentes ediciones, la tabla de contenidos, el glosario y el índice, para evitar la introducción de apariciones de palabras que no están estrictamente en el contenido del texto que queremos analizar. Como resultado, obtenemos finalmente un texto con longitud $N = 193786$ palabras que contiene un vocabulario formado por 8186 palabras diferentes.

Si computamos E_{nor} (definida en (1.8)) y C para esas 8186 palabras, obtendremos dos *rankings* diferentes de relevancia. El problema aquí es que el concepto de relevancia es subjetivo ya que una palabra no es relevante o irrelevante en sí misma, sino que su contenido de información puede ser significativo para un texto determinado y no tener ninguna importancia en otro. Por este motivo, [Herrera and Pury, 2008] sugirieron el uso de un glosario como punto de referencia. Éste permitiría la evaluación y comparación

[1] Descargado de la web del Proyecto Gutenberg

de la bondad de los resultados obtenidos con diferentes detectores de palabras clave que apliquemos al libro de Darwin. Dicho glosario se creó a mano [Herrera and Pury, 2008] partiendo del glosario original y del índice terminológico, y contiene 283 palabras que aparecen con frecuencia al menos 9 en el libro completo.

Sin embargo, observamos que hay palabras que aparecen en el glosario sólo en singular o sólo en plural ('varieties', 'groups', 'character',...), mientras que otras aparecen tanto en singular como en plural ('condition','conditions'; 'habit', 'habits'; 'stage', 'stages';...). Como el glosario se emplea para identificar las palabras relevantes para el libro, esto tiene una consecuencia bastante importante: palabras como 'variety', 'group' y 'characters' no se están considerando relevantes; mientras que 'varieties', 'groups' y 'character', sí.

Hasta ahora estábamos considerando cada cadena de caracteres entre dos espacios en blanco como una palabra diferente. Sin embargo, para corregir la incongruencia mencionada en el glosario, proponemos identificar el singular y plural de cada palabra como la misma palabra (por ejemplo, 'descendant' y 'descendants'). Con este objetivo, consideramos las reglas de formación de plurales para los sustantivos en inglés y las implementamos en nuestro algoritmo. Notemos que, como consecuencia, también identificaremos el infinitivo y la tercera persona del singular del presente de los verbos (por ejemplo, 'descend' and 'descends'). Después de llevar a cabo dicha identificación entre singular y plural, obtenemos un vocabulario de 7104 palabras, en lugar de 8186. Con respecto al glosario, éste se ve reducido de 283 a 249 palabras. La diferencia crucial es que ahora tanto el singular como el plural de cada una de las palabras del glosario son considerados relevantes para el texto.

Por otro lado, la identificación entre singular y plural que aplicamos para corregir las incongruencias del glosario podría considerarse también como el caso más simple de preprocesamiento lingüístico o, en un lenguaje físico, como la corrección de primer orden a los resultados obtenidos sin ningún procesamiento. Un preprocesamiento lingüístico del texto, como por ejemplo la reducción de todas las palabras a sus raíces comunes, o la identificación de formas sustantivas y adjetivas, podría mejorar el comportamiento de los detectores de palabras clave. Pero el número de reglas (sintácticas y gramaticales) necesarias para ese preprocesamiento lingüístico extensivo sería grande y los costes computacionales derivados serían altos, por lo que estaría fuera del objetivo de este trabajo. Sin embargo, la identificación entre singular y plural tiene pocas reglas y puede ser implementada en tiempo real en el algoritmo mientras lee el texto de entrada para identificar el vocabulario sin ningún coste computacional discernible. Por medio de esta implementación, obtenemos un glosario más compacto y un vocabulario de palabras realmente diferentes; a la vez que, como veremos posteriormente, comprobamos que mejora los resultados de los detectores de palabras clave por medio de las métricas de evaluación que definiremos en la sección

posterior.

2.3. Métricas de evaluación

Ante la existencia de diferentes tipos de medidas de relevancia es inmediato plantearse si se pueden proponer métricas que cuantifiquen el comportamiento de dichos detectores y que nos permitan evaluar los resultados obtenidos con cada una de las aproximaciones.

2.3.1. Precisión y exhaustividad

En el ámbito de los Sistemas de Recuperación de Información, con el objetivo de evaluar el comportamiento de un algoritmo de búsqueda, se han usado tradicionalmente dos métricas llamadas *precision* y *recall* (precisión y exhaustividad) [Hand et al., 2001]. Dada una búsqueda concreta dentro de un conjunto de N documentos, estos deben ser previamente clasificados como relevantes o no para dicha búsqueda. Suponiendo que el algoritmo que queremos evaluar nos proporciona n documentos potencialmente relevantes, *precision* y *recall* se definen como sigue [Hand et al., 2001]:

- *precision* es la fracción de los n documentos recuperados que son realmente relevantes, es decir,
$$\text{precision} = \frac{\#\text{relevantes} \cap \#\text{recuperados}}{\#\text{recuperados}}$$
- *recall* es la proporción de documentos relevantes que el algoritmo ha recuperado, es decir,
$$\text{recall} = \frac{\#\text{relevantes} \cap \#\text{recuperados}}{\#\text{relevantes}}$$

A medida que aumentemos el número n de documentos recuperados por el algoritmo, el valor de *recall* irá aumentando, ya que iremos recuperando más documentos relevantes. Pero a su vez es usual que el valor de *precision* vaya disminuyendo, ya que es difícil que al aumentar n obtengamos sólo documentos relevantes.

2.3.2. Adaptación al problema de la detección de palabras clave

[Herrera and Pury, 2008] propusieron una adaptación de estas dos métricas al problema de la evaluación de detectores de palabras clave usando el libro *The Origin of Species*. Como ya hemos mencionado, consideran como palabras relevantes aquellas palabras del vocabulario del libro que pertenezcan al glosario que hemos definido en la sección anterior, de modo que tendríamos un conjunto de 283 palabras relevantes. De esta manera todas

las palabras del vocabulario son preclasificadas como relevantes o no para el libro. Una vez preclasificadas las palabras por medio del glosario, Herrera y Pury definen en el contexto de detección de palabras clave las siguientes métricas de evaluación:

- *precision of a keyword extractor* (pr_{ke})

$$\mathrm{pr}_{ke} = \frac{283}{LP},$$

  donde LP denota la posición en el *ranking* de la última palabra relevante recuperada.

- *recall of an index of relevance* (r_{ir})

$$\mathrm{r}_{ir} = \frac{NG}{283},$$

  donde NG denota el número de palabras relevantes recuperadas hasta la posición 283 del *ranking*.

Cada detector proporciona un *ranking* de todas las palabras del vocabulario del libro en orden decreciente de relevancia. Herrera y Pury obtuvieron que los valores de pr_{ke} y r_{ir} para E_{nor} eran mejores que para otras medidas descritas en la literatura (la medida C fue definida con posterioridad a [Herrera and Pury, 2008]).

Sin embargo, si nos basamos en las necesidades de un usuario típico, las métricas pr_{ke} y r_{ir} no parecen ser adecuadas para evaluar la calidad de un detector de palabras clave. El número de palabras clave que normalmente se emplea para sintetizar el tópico de cualquier texto escrito es pequeño y, como consecuencia, el usuario de un detector de palabras clave sólo requiere un número reducido de palabras. De modo que la información sobre que método recupera antes todas las palabras del glosario (en este caso 283) sería secundaria, siendo más importante la precisión de las palabras situadas en las primeras posiciones del *ranking*, por ser las únicas que tendría en cuenta un usuario típico.

Podemos deducir de la definición que el valor de pr_{ke} sólo depende de la posición del *ranking* en la que se detecta la última palabra del glosario (LP), independientemente de cuando hemos detectado el resto de palabras relevantes, incluso aunque éstas se encontrasen en las primeras 282 posiciones. Esta dependencia de la detección de una única palabra (la última), que además probablemente no es importante para el usuario, que sólo tiene en cuenta las primeras palabras del *ranking*, hace que el resultado proporcionado por pr_{ke} no resulte útil en este contexto. De hecho, no podemos asegurar que el glosario usado es perfecto, y esta métrica se vería totalmente afectada por la posibilidad de que sólo una de entre todas las 283 palabras del glosario estuviese preclasificada erróneamente como relevante.

Por otro lado, el valor de r_{ir} se obtiene a partir del número de palabras relevantes recuperadas dentro de las primeras 283 entradas del ranking (NG). Y, como ya hemos argumentado, en una situación práctica un usuario sólo tendrá en cuenta las palabras de las primeras posiciones del *ranking*, y no miraría 283 entradas. Como consecuencia el uso de r_{ir} tampoco nos daría información útil en este contexto.

2.3.3. Métricas basadas en las necesidades del usuario

Por los motivos que acabamos de exponer, proponemos [Carretero-Campos et al., 2013] dos métricas alternativas para evaluar detectores de palabras clave basándonos en las necesidades del usuario. Por un lado, sugerimos que el objetivo principal para evaluar el comportamiento de un detector de palabras clave debería ser cuantificar si las primeras palabras del *ranking* proporcionadas por el detector son relevantes para el texto o no. Y, por otro lado, pensamos que también sería interesante incluir información acerca de la completitud del detector, es decir, acerca de las posiciones en las que recupera cada una de las palabras relevantes, pero dando más peso a aquellas detectadas antes, ya que serían las que usualmente consideraría un usuario.

Precision at n

En primer lugar, definiremos *precision at n*, $\mathrm{pr}(n)$, como la fracción de las primeras n palabras del *ranking* que son relevantes. Si denotamos por $\mathrm{key}(n)$ el número de palabras relevantes dentro de las n primeras palabras del *ranking*, tenemos que

$$\mathrm{pr}(n) = \mathrm{key}(n)/n, \quad (0 \leq \mathrm{pr}(n) \leq 1) \tag{2.3}$$

Si evaluamos $\mathrm{pr}(n)$ para valores pequeños de n ($n \leq 50$) tendremos información acerca de la relevancia de las palabras situadas en las primeras posiciones del *ranking*. Al igual que hicieron en [Herrera and Pury, 2008], consideraremos que una palabra es relevante si y sólo si pertenece al glosario.

En la figura 2.4 mostramos los resultados de esta métrica de evaluación aplicada a las medidas C, C_0, C_1 y E_{nor}. La partición usada para calcular E_{nor} son los 16 capítulos del libro, como en [Herrera and Pury, 2008]. Podemos observar que C_0 se comporta como un detector perfecto hasta $n = 17$ ($\mathrm{pr}(n) = 1$). Esto significa que las primeras 17 palabras del *ranking* de relevancia obtenido por medio de C_0 pertenecen al glosario. Si aceptamos el glosario como punto de referencia acerca de cuáles son las palabras verdaderamente relevantes para el libro, el resultado obtenido para C_0 es excelente: las primeras 17 palabras del *ranking* han sido correctamente identificadas como relevantes. También obtenemos

buenos resultados para C, C_1 y E_{nor} ya que en todos los casos se verifica que hasta $n = 50$, $\mathrm{pr}(n) \geq 0.6$: al menos el 60 % de las palabras han sido identificadas correctamente como relevantes.

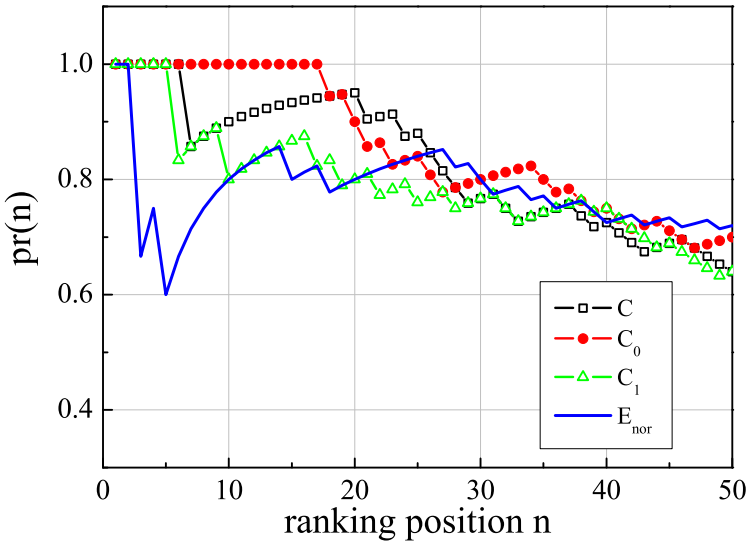

Figura 2.4: Comportamiento de $\mathrm{pr}(n)$ para $n \leq 50$ obtenido con C (cuadrados negros), C_0 (círculos rojos), C_1 (triángulos verdes) y E_{nor} (línea azul) para el libro *The Origin of Species* usando el glosario preparado en [Herrera and Pury, 2008].

El valor de precisión decrece cada vez que la palabra recuperada no pertenece al glosario, de modo que la decisión de cuáles son las palabras seleccionadas en el glosario tiene un papel decisivo. Merece la pena notar que algunas palabras que no están incluidas en el glosario, y por tanto que han sido preclasificadas como no relevantes, aparecen entre las primeras palabras de los cuatro *rankings* de relevancia y de hecho tienen un contenido semántico relevante para el texto. Quizá se debe a que dichas palabras son en algún sentido 'palabras comunes' y el autor no consideró necesario incluirlas explícitamente en el glosario. Ese es el caso, por ejemplo, de la palabra 'island' (posición 18 para C_0 y 17 para C_1) e 'islands' (posición 7 para C, 20 para C_0 y 5 para E_{nor}).

Si identificamos previamente el singular y plural de cada palabra, tal y como explicamos en la sección anterior, obtenemos los resultados que mostramos en la figura 2.5. De aquí en adelante cada vez que realicemos dicha identificación denotaremos a las medidas como C^*, C_0^*, C_1^* y E_{nor}^*. Aunque ya no observamos un rango tan amplio de máxima precisión como el obtenido anteriormente para C_0, observamos que tanto C^*, como C_0^* y C_1^* tienen valores de *precision* mayores que 0.8 (80 %) hasta $n = 50$, mejorando los resultados obtenidos sin la identificación entre singular y plural. En el rango de n pequeño ($n < 10$), observamos que las medidas de *clustering* funcionan mejor que la entrópica.

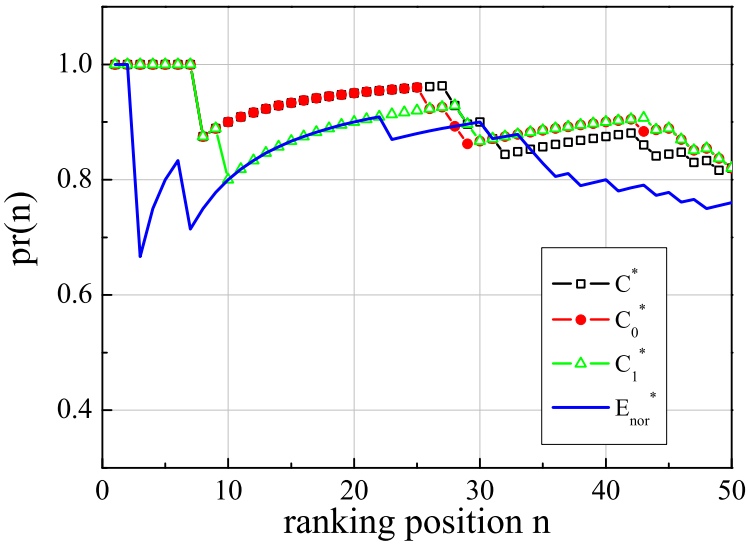

Figura 2.5: Análogamente a la figura 2.4, pero incluyendo la identificación previa del singular y plural de cada palabra. Las medidas calculadas con dicha identificación se denotan con *.

Average precision

En segundo lugar, para obtener información acerca de la completitud del detector, adaptaremos la métrica denominada *average precision* (AP) al problema de detección de palabras clave. Es una métrica de evaluación comúnmente empleada en el ámbito de Recuperación de Información que proporciona información acerca del comportamiento global de un algoritmo [Zhu, 2004; Aslam et al., 2005; Yilmaz and Aslam, 2006; Robertson et al., 2010]. En el contexto de detección de palabras clave definimos AP como sigue:

$$\mathrm{AP} = \frac{1}{R}\sum_{n=1}^{L} \mathrm{pr}(n) \times \mathrm{rel}(n), \tag{2.4}$$

donde $\mathrm{pr}(n)$ fue definida en (2.3) y $\mathrm{rel}(n)$ es igual a 1 si la palabra en la posición n del ranking es relevante y 0 en otro caso; L es el número total de palabras en el ranking (vocabulario) y R el número total de palabras relevantes (glosario). De modo que AP se define como el promedio de las precisiones en las posiciones del ranking en las que se recupera cada una de las palabras relevantes. Como consecuencia, tiene en cuenta todas las palabras relevantes, pero da más peso a aquellas detectadas antes. En el caso de un detector perfecto, las primeras R palabras del ranking serían relevantes y, por tanto, $\mathrm{AP} = 1$. Los resultados obtenidos para $C, C_0, C_1, E_{nor}, C^*, C_0^*, C_1^*$ y E_{nor}^* se muestran en la figura 2.6 y confirman las conclusiones extraídas previamente. Aquí también observamos que los resultados obtenidos usando la identificación singular-plural son todos mejores que los obtenidos sin ella. Podemos estimar una mejora promedio de alrededor del 10 % en los

valores de AP para todas las medidas. En ambos casos, la medida C con condiciones de contorno (C_0 y C_0^*) presenta el mejor valor de precision promedio y la medida entrópica E_{nor} el peor, aunque sin diferencias extremas.

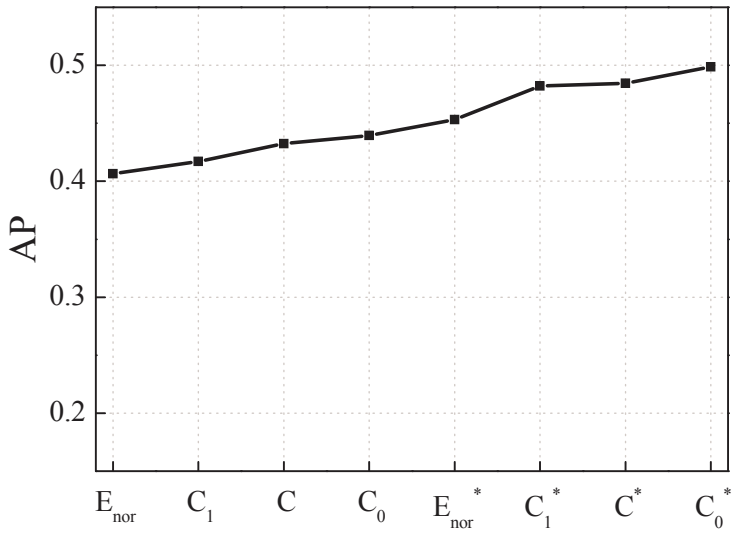

Figura 2.6: *Average precision* AP de los detectores $C, C_0, C_1, E_{nor}, C^*, C_0^*, C_1^*$ y E_{nor}^* aplicados al libro *The Origin of Species*.

Resumiendo esta sección, proponemos dos métricas (pr(n) y AP) para evaluar el comportamiento de detectores de palabras clave. Por medio de dichas métricas concluimos que: i) La identificación singular-plural, además de proporcionarnos un glosario más compacto, mejora satisfactoriamente el comportamiento de los detectores de palabras clave, y ii) el comportamiento de las medidas basadas en *clustering* C, C_0 y C_1 es igual o mejor que el de la medida entrópica E_{nor}.

2.4. Dependencia de E_{nor} con respecto a la partición

En el capítulo 1 se mostraba como la definición de E_{nor} (ecuación 1.8), que usa la entropía de Shannon, requería de una partición P del texto que queramos analizar. Los resultados mostrados hasta ahora para la medida E_{nor} han sido obtenidos a partir de una partición natural del libro *The Origin of Species* en sus 16 capítulos. A priori se puede pensar que lo más adecuado es realizar una partición del texto teniendo en cuenta divisiones naturales, ya que se pretende detectar las palabras más significativas comparando su distribución de frecuencias entre las partes en las que se ha dividido el texto. En lugar de los capítulos, si queremos considerar otra división natural, podríamos pensar en los párrafos.

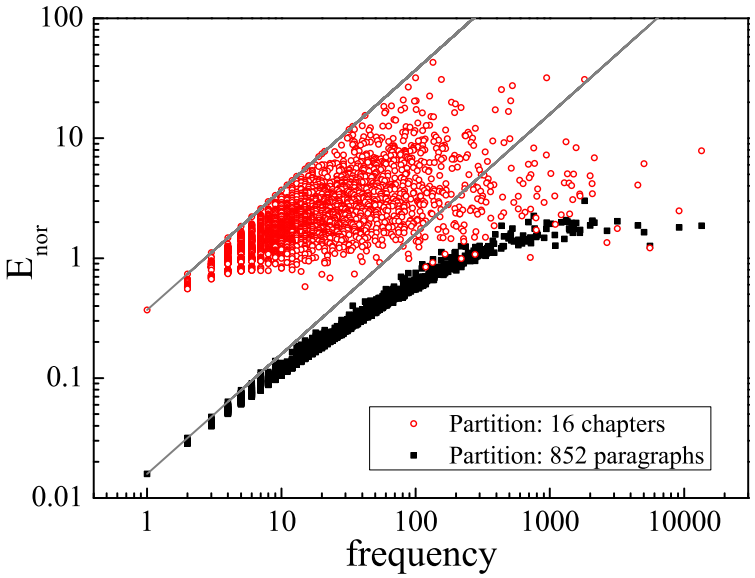

Figura 2.7: Los valores de E_{nor} para todas las palabras de *The Origin of Species* obtenidas usando la partición en capítulos (círculos rojos) y en párrafos (cuadrados negros) como función de la frecuencia de la palabra. Las líneas muestran el máximo valor posible de E_{nor} para cada valor de frecuencia en ambos casos.

Dividimos entonces el texto en los 852 párrafos que contiene y calculamos E_{nor} usando dicha partición. En la figura 2.7 mostramos los valores de E_{nor} obtenidos para las 8186 palabras del libro como función de su frecuencia, tanto para el caso de la partición en capítulos, como para el caso de la partición en párrafos. Observando dicha figura se puede deducir que, si consideramos la partición en párrafos, la medida entrópica E_{nor} no resultará un detector de palabras clave fiable, ya que no distingue entre diferentes grados de relevancia para palabras con la misma frecuencia (mientras que sí lo hace la obtenida en el caso de partición en capítulos). De hecho, solo 2 de las 20 primeras palabras del ranking obtenido con dicha medida pertenecen al glosario ('species' y 'varieties'). Las 18 palabras restantes son claramente no relevantes ya que no tienen contenido significativo para el texto, como es el caso de 'will', 'we', 'have', 'a', 'from' y 'be'. Si previamente llevamos a cabo la identificación entre singular y plural descrita en la sección anterior, los resultados de la medida entrópica obtenidos a partir de una división en párrafos siguen siendo realmente pobres (ver línea discontinua en la figura 2.8). Observamos por tanto que la elección de la partición determina completamente los resultados obtenidos con E_{nor} y que, aunque los párrafos constituyen también una división natural, en este caso el cálculo de E_{nor} a partir de dicha partición no distingue las palabras relevantes del texto.

De modo que, al usar la medida entrópica E_{nor}, el problema es qué criterio seguir para elegir una partición adecuada del texto. La dependencia de los resultados con respecto

a la partición elegida fue mencionada también en [Herrera and Pury, 2008], pero sin llevar acabo un estudio sistemático. Hay muchos textos (especialmente los textos cortos) cuya única división natural existente son los párrafos, para los que, en el ejemplo que acabamos de ver, los resultados proporcionados por E_{nor} no son muy precisos. Como consecuencia, podríamos pensar que para ese tipo de textos (que no tienen secciones ni capítulos) no tenemos una partición natural que asegure a priori la bondad de los resultados obtenidos al calcular E_{nor}. Por ese motivo, nos planteamos estudiar cómo la medida E_{nor} se comporta en divisiones artificiales del texto y, en concreto, analizaremos los resultados obtenidos a partir de varias divisiones del texto en partes iguales. Para cuantificar y comparar los resultados obtenidos con distintas particiones usaremos de nuevo las métricas de evaluación pr(n) and AP definidas en la sección anterior.

En las figuras 2.8 y 2.9 comparamos los resultados obtenidos calculando E^*_{nor} a partir de divisiones del texto en $45, 250, 500$ y 800 partes iguales con los resultados obtenidos en divisiones naturales. Notemos que estos resultados incluyen la identificación entre singular y plural porque hemos comprobado que mejora la precisión de los detectores de palabras clave. Los resultados para pr(n) ($n \leq 50$) y para AP, que tienen en cuenta la información acerca de la precisión de las primeras palabras del *ranking*, así como el 'comportamiento global' del detector, están de nuevo en acuerdo. A pesar de usar divisiones artificiales, observamos que en general para un número de partes iguales P del orden del número de capítulos, como $P = 45$, la precisión de la medida entrópica es bastante buena. Sin embargo, los resultados empeoran cuando aumentamos el número de partes iguales P en las que dividimos el texto. Para $P = 800$ los resultados son casi tan malos como usando la división en párrafos (figura 2.8).

Por tanto, podemos concluir que el factor importante que controla el comportamiento de E_{nor} es el número mínimo de palabras incluido en cada una de las partes en las que se divide el texto, independientemente de si la división es natural o artificial. Por un lado, cada parte tiene que ser suficientemente grande como para formar una unidad con contenido significativo que permita detectar relevancia [Montemurro and Zanette, 2010]. Por otro lado, debemos tener un número mínimo de partes para poder realizar una comparación estadísticamente significativa entre las frecuencias de la palabras en cada una de las partes.

Para finalizar esta sección, notar que los resultados obtenidos con las medidas C^* (C^*_0, C^*_1) son igual o mejores que los obtenidos con las mejores elecciones de partición para la medida entrópica E^*_{nor}, pero con la ventaja de no tener que realizar ninguna elección. Esta es una ventaja importante, sobre todo a la hora de analizar textos cortos, tal como discutiremos en la siguiente sección.

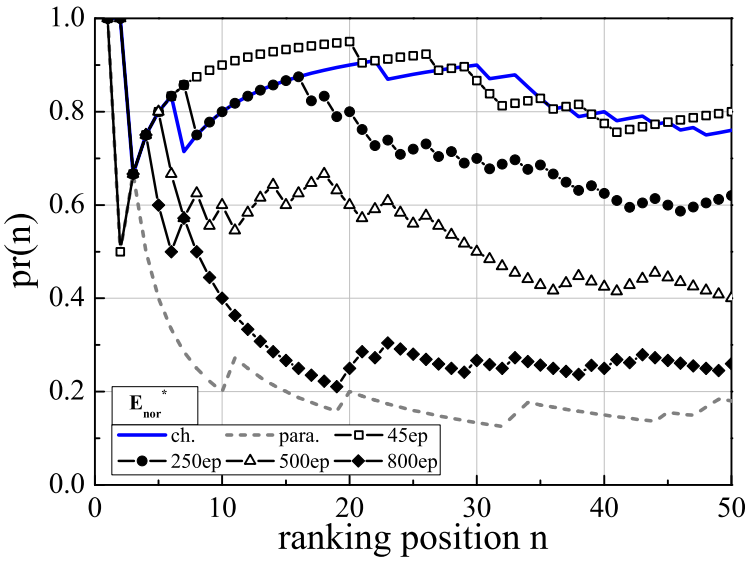

Figura 2.8: Comportamiento de pr(n) para $n \leq 50$ obtenido para la medida entrópica E^*_{nor} usando diferentes tipos de particiones del libro *The Origin of Species*: particiones naturales (capítulos (línea azul) y párrafos (línea discontinua gris)) y particiones artificiales (45 (cuadrados negros abiertos), 250 (círculos negros sólidos), 500 (triángulos negros abiertos) y 800 (rombos negros sólidos) partes iguales).

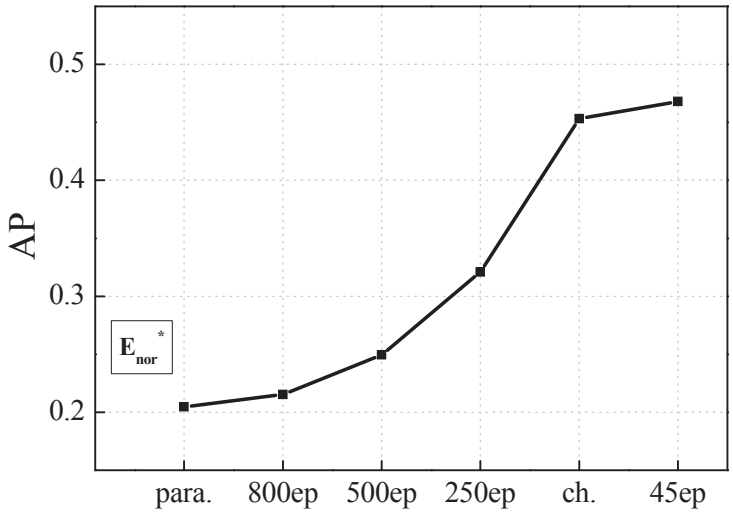

Figura 2.9: Average precision AP de la medida entrópica E^*_{nor} usando diferentes tipos de particiones del libro *The Origin of Species*: particiones naturales (capítulos y párrafos) y artificiales (45, 250, 500 y 800 partes iguales).

2.5. Resultados en textos cortos

Una de las principales características deseable en un detector de palabras clave es que sea capaz de funcionar correctamente cuando el texto es corto. En general es difícil de conseguir porque la frecuencia de la mayoría de palabras va a ser pequeña, lo cual complica el análisis estadístico. Como consecuencia, es común que se identifiquen erróneamente las palabras más informativas del texto. Sin embargo, la detección y extracción de palabras clave de textos cortos es de claro interés práctico ya que puede ser aplicado a artículos científicos, páginas web, etc. Además, si un detector de palabras clave es capaz de detectar las palabras relevantes de un texto corto, será capaz de hacerlo en uno largo también, pero lo contrario no es necesariamente verdad.

Por ello, en esta sección nos centraremos en evaluar el comportamiento de E_{nor} y C en textos cortos. Concretamente, vamos a considerar dos tipos de textos cortos: i) Un texto corto del que disponemos de un glosario para cuantificar el comportamiento de los detectores de palabras clave, y ii) Textos cortos genéricos, para los que presentamos resultados cualitativos.

2.5.1. Textos cortos con glosario

En este caso seleccionamos como texto corto de referencia el capítulo más corto del libro *The Origin of Species*. De este modo tendremos un glosario que nos permita aplicar las métricas de evaluación que hemos definido previamente.

Elegimos por tanto el capítulo III del libro, titulado *Struggle for existence*, el cual tiene una longitud de 6349 palabras, que es aproximadamente un 3 % de la longitud total del libro. El glosario que usamos para este capítulo consiste en todas las palabras del glosario original que aparecen en el capítulo con frecuencia al menos 3. Una vez llevada a cabo la identificación entre singular y plural, el capítulo contiene un vocabulario de 1236 palabras y el glosario contiene 50. Todos los resultados que mostraremos más abajo incluyen esta identificación.

Para calcular E_{nor} necesitamos realizar una partición del capítulo. Las divisiones naturales que contiene este capítulo son secciones (6) y párrafos (28). Calcularemos E_{nor} considerando ambas divisiones y, tal y como hicimos con el libro completo, también lo calcularemos usando divisiones artificiales del capítulo en partes iguales. Como *a priori* no tenemos idea sobre cual va a ser la mejor partición, mostraremos los resultados para una partición en 14 partes iguales como ejemplo. Presentamos en las figuras 2.10 y 2.11 los resultados obtenidos para las métricas pr(n) y AP calculadas a partir de los rankings de relevancia de C^*, C_0^*, C_1^*, y E_{nor}^* en divisiones naturales y artificiales.

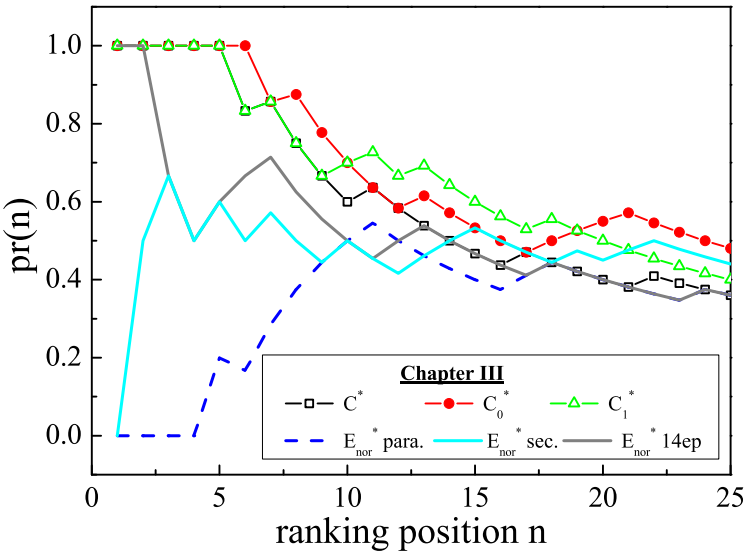

Figura 2.10: Comportamiento de pr(n) para $n \leq 25$ obtenido con C^* (cuadrados negros), C_0^* (círculos rojos), C_1^* (triángulos verdes) y E_{nor}^* con divisiones naturales en párrafos (línea azul discontinua) y secciones (línea celeste), y división artificial en 14 partes iguales (línea gris) para el capítulo III del libro *The Origin of Species*.

Con respecto a la métrica pr(n), observamos que para n pequeño ($n < 10$), que es un número típico de palabras clave que puede requerir un hipotético usuario, las medidas C^*, C_0^*, C_1^* funcionan mejor que los tres casos de E_{nor}^*. En particular, para $n < 6$ las medidas C^*, C_0^* y C_1^* funcionan como detectores perfectos (pr = 1), mientras que E_{nor}^* calculada a partir de una partición en párrafos representa el caso opuesto, ya que las 4 primeras palabras del ranking obtenido en ese caso son erróneas ('or', 'been', 'we', 'the'). En cuanto a E_{nor}^* calculada en una partición en secciones y en 14 partes iguales, el comportamiento para n pequeño es intermedio entre los dos casos anteriores. Cuando consideramos más palabras del ranking (mostramos hasta $n = 25$ en la figura 2.10) las diferencias entre los resultados obtenidos con las diferentes medidas se van reduciendo, aunque incluso en este rango, probablemente más allá de los requerimientos de un usuario de un detector de palabras clave, C^*, C_0^* y C_1^* se comportan igual o mejor que los tres casos considerados para E_{nor}^*.

Merece la pena notar que en el caso de las medidas C^*, C_0^* y C_1^*, entre las palabras que se consideran erróneamente detectadas como relevantes (es decir, aquellas que no pertenecen al glosario y que por tanto disminuyen pr(n)) incluidas entre las 25 primeras del ranking, podemos encontrar palabras como 'advantage', 'competition', y 'relations'. Tales palabras son objetivamente relevantes para el capítulo considerado (*Struggle for existence*), de modo que los valores de pr(n) podrían ser claramente mejores para esas medidas. Sin embargo, entre las palabras consideradas erróneas en los casos de E_{nor}^*, sólo

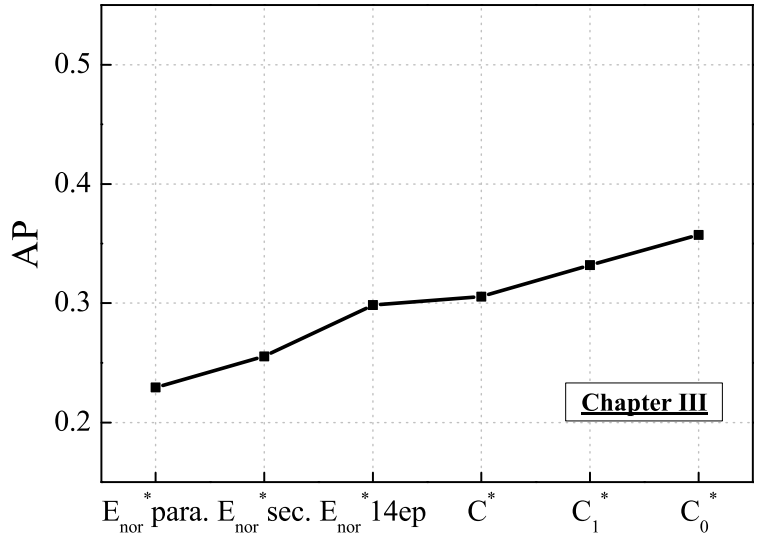

Figura 2.11: Average precision AP de los detectores C^*, C_0^*, C_1^* y E_{nor}^* con divisiones naturales en párrafos y secciones, y división artificial en 14 partes iguales obtenida en el capítulo III del libro *The Origin of Species*.

encontramos ejemplos como 'been', 'we', 'had', 'said' o 'was'. Todas ellas son claramente no relevantes para el capítulo considerado, de modo que los valores de E_{nor}^* no mejorarían incluso en el caso en el que se consideraran más palabras como relevantes además de las pertenecientes al glosario.

En cuanto al comportamiento global, cuantificado por AP, mostramos en la figura 2.11 los resultados para C^*, C_0^*, C_1^*, y para E_{nor}^* en tres particiones (párrafos, secciones y 14 partes iguales). Dichos resultados corresponden también al capítulo más corto (capítulo III) del libro *The Origin of Species*, y complementan a los resultados mostrados en la figura 2.10. Observamos que, de manera similar a los resultados obtenidos para el libro completo, la medida C_0^* es la que presenta el mejor valor de AP, y que el peor caso corresponde a la medida entrópica E_{nor}^* obtenida a partir de la división en párrafos del capítulo. Además, C^*, C_0^* y C_1^* se comportan igual o mejor que los tres casos de E_{nor}^*.

Debido a que los resultados presentados para pr(n) y AP dependen completamente del glosario que se ha seleccionado, mostramos en la figura 2.12 los valores de C^* y E_{nor}^* (calculada a partir de la división en párrafos) obtenidos para las 1236 palabras del capítulo en función de su frecuencia n. Estos valores son independientes del glosario y muestran que C^* distingue mucho mejor diferentes grados de relevancia para palabras con la misma frecuencia. Por el contrario, en el caso de E_{nor}^* a partir de la división en párrafos, los valores para palabras con la misma frecuencia y diferente grado de relevancia son muy similares, lo que produce una detección incorrecta de las palabras relevantes del capítulo. En particular, las palabras responsables de producir valores bajos de precisión

están realmente mal identificadas ('will', 'would', 'from', 'been',...). Los valores de precisión decrecen cuando se selecciona una palabra que no está en el glosario. Sin embargo, esto puede ocurrir por dos razones: la palabra es relevante, pero no estaba incluida en el glosario (ya que el glosario no es perfecto), o la palabra no es relevante y no está incluida en el glosario. En el caso que estamos estudiando, las palabras responsables del mal resultado de E^*_{nor} son del segundo tipo.

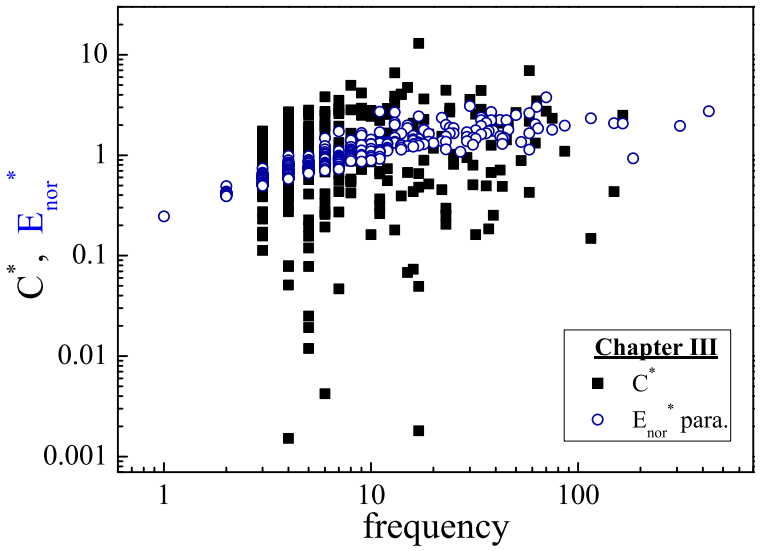

Figura 2.12: Comparación entre los valores de C^* (cuadrados negros) y E^*_{nor} calculada con división en párrafos (círculos azules) frente a la frecuencia de ocurrencia de todas las palabras del capítulo III de *The Origin of Species*.

2.5.2. Textos cortos genéricos

Hasta ahora hemos usado un libro y uno de sus capítulos como texto largo y corto respectivamente, motivados por la disponibilidad de un glosario para cuantificar la bondad de los resultados. Como ya hemos comentado más arriba, y debido a la naturaleza estadística de los detectores de palabras clave, los resultados peores se esperan en general en textos cortos debido a la baja frecuencia de las palabras. Sin embargo, el análisis de textos cortos es probablemente el más importante desde el punto de vista práctico. Por esa razón, presentamos en esta subsección los resultados obtenidos para otros textos cortos genéricos. Hemos seleccionado varias entradas de Wikipedia [2], y en particular conceptos científicos con longitudes es el rango 500-3000, típicamente más pequeños que artículos científicos estándar. Hemos elegido dichas entradas siguiendo dos criterios: sus longitudes,

[2]Descargamos las entradas de Wikipedia en formato xml, las limpiamos y producimos un texto bruto que es posteriormente analizado con ambos métodos.

y la facilidad para evaluar cualitativamente los resultados obtenidos para los conceptos seleccionados, ya que no dispondremos de un glosario.

En la tabla 2.1 mostramos las 10 primeras palabras clave extraídas usando C y E_{nor} para tres de esas entradas. En el caso de E_{nor}, en el que como ya hemos visto necesitamos realizar una partición del texto, hemos usado la división natural en secciones de cada una de las entradas. Observamos que en general ambos métodos recuperan palabras relevantes, pero la medida de *clustering* C funciona mejor que la medida entrópica E_{nor}, ya que el número de palabras realmente relevantes entre las 10 primeras es mayor que en el caso de E_{nor}. Notemos también que, tal como era de esperar, ambos métodos funcionan mejor en la entrada más larga ('statistics'), pero incluso en este caso el número de palabras realmente relevantes entre las 10 primeras es mayor para C (9) que para E_{nor} (6-7).

Teniendo en cuenta los resultados que hemos presentado en esta sección, concluimos que las medidas derivadas de una aproximación basada en *clustering* proporcionan resultados fiables cuando se aplican a textos cortos (iguales o mejores que usando E_{nor}), sin necesitar ninguna información a priori, mientras que E_{nor} necesita una elección previa sobre qué tipo de partición natural o artificial del texto debe ser considerada. La detección correcta o incorrecta de las palabras clave dependerá de dicha elección.

term	speed (476 words)		sound (1306 words)		statistics (3903 words)	
measure / rank	C	E_{nor}	C	E_{nor}	C	E_{nor}
1	per	is	speed	the	statistics	value
2	time	time	pressure	waves	the	the
3	hour	per	waves	a	population	population
4	h	symbol	noise	sound	hypothesis	hypothesis
5	distance	or	an	medium	measurements	interval
6	interval	the	level	in	experimental	measurements
7	or	an	intensity	and	models	can
8	units	s	energy	is	sample	is
9	an	hour	pa	pressure	probability	sample
10	km	a	hz	speed	significance	true

Tabla 2.1: *Ranking* de las 10 primeras palabras clave extraídas usando C y E_{nor} (en secciones) de las entradas en Wikipedia 'speed', 'sound' y 'statistics'.

2.6. Conclusión

En este capítulo proponemos dos métricas de evaluación, pr(n) y AP, especialmente indicadas para cuantificar el comportamiento de detectores de palabras clave y basadas en las necesidades de un usuario típico. Estas dos métricas nos permiten evaluar y comparar dos diferentes aproximaciones al problema de la detección de palabras clave, la entrópica y la basada en *clustering*, para la que además presentamos algunas mejoras. Comparamos el comportamiento de los dos detectores de palabras clave en textos cortos y largos. Como

texto largo, elegimos el libro *The Origin of Species* debido a la disponibilidad de un glosario que se usará como referencia. En el caso de textos cortos, presentamos por un lado resultados cuantitativos obtenidos en el capítulo más corto de *The Origin of Species* y testados con el glosario, y, por otro lado, resultados cualitativos obtenidos de entradas de Wikipedia, los cuales consideramos como nuestros textos cortos genéricos.

Aunque ambos métodos funcionan razonablemente bien en textos largos, concluimos que el comportamiento de la medida entrópica presenta una fuerte dependencia de la partición del texto seleccionado, dando resultados fiables solo para elecciones adecuadas, lo cual no es trivial ya que no las conocemos a priori. Sin embargo, las medidas basadas en *clustering* parecen dar resultados exitosos tanto en textos largos como cortos sin ninguna información previa del texto. Incluso en los casos donde la partición del texto lleva a resultados precisos usando la aproximación entrópica, las medidas basadas en *clustering* proporcionan resultados iguales o mejores.

Capítulo 3

Distribución de las distancias entre símbolos en secuencias aleatorias. Aplicación a la detección de *clustering*

En capítulos anteriores hemos visto cómo la distribución espacial de una palabra a lo largo de un texto está relacionada con su relevancia, siendo esta mayor cuanto más nos separemos de la distribución esperada por azar. Medidas que cuantifican dicha desviación han resultado útiles para detectar las palabras clave de un texto, sin disponer de información previa sobre el contenido del mismo.

Recordemos que se caracterizaba la distribución espacial de una palabra a lo largo del texto usando la distribución $p(d)$ de las distancias entre apariciones sucesivas de la palabra. También vimos que la distribución que se asumía como referencia para una palabra distribuida al azar [Ortuño et al., 2002; Herrera and Pury, 2008; Carpena et al., 2009; Carretero-Campos et al., 2013; Oliver et al., 2002; Hackenberg et al., 2012, 2010; Altmann et al., 2009] es la distribución geométrica. Una vez que la palabra aparece, la distancia hasta la próxima aparición se interpreta como el número de ensayos necesarios para tener un primer éxito, entendiendo por éxito que la palabra vuelva a aparecer. Si denotamos por $p \equiv n/N$ la probabilidad de encontrar la palabra en el texto, entonces

$$p_{\text{geo}}(d) = p(1-p)^{d-1}. \tag{3.1}$$

Sin embargo, la distribución $p_{\text{geo}}(d)$ es la que se obtiene asintóticamente en el límite para N y n tendiendo a infinito, manteniendo constante el ratio n/N. Y, como conse-

cuencia, puede llevar a resultados incorrectos si la palabra tiene una frecuencia baja y el texto es corto (es decir, para N y n pequeños). El objetivo de este capítulo es entonces obtener la distribución exacta $p_{N,n}(d)$ esperada por azar para N y n finitos, y usarla en la detección de palabras clave. De hecho, veremos en este capítulo que su uso permite mejorar la detección en textos cortos, y también distinguir entre palabras clave genéricas o de uso más general frente a palabras clave muy específicas.

3.1. Distribución entre apariciones sucesivas de una palabra esperada por azar

Consideramos un texto de longitud N, que podemos identificar como el intervalo $[1, N]$, en el que las posiciones de una palabra serán números enteros en dicho intervalo. Consideramos una palabra que aparece al azar n veces en el texto ($n \leq N$), situada en las posiciones j_i, para $i = 1, 2, \ldots, n$, con $0 < j_1 < j_2 < \ldots < j_n < N + 1$. El conjunto de $n - 1$ distancias entre apariciones sucesivas de la palabra viene dado por $d_i = j_{i+1} - j_i$ con $i = 1, 2, \ldots, n-1$. Queremos obtener la probabilidad de encontrar dos apariciones de la palabra a distancia d.

Incluimos en el conjunto dos distancias adicionales, $d_0 = j_1 - 0 = j_1$ y $d_n = N + 1 - j_n$, equivalentes a considerar dos apariciones "artificiales" de la palabra en las posiciones 0 y $N + 1$, que podemos interpretar como condiciones de contorno. Nótese que, entre los dos tipos de condiciones de contorno consideradas en el capítulo anterior, estas son las que mejores resultados nos dieron. Tenemos entonces un conjunto de $n+1$ distancias que toman valores en $\{1, 2, \ldots, N + 1 - n\}$. Veremos cómo la inclusión de esas dos distancias simplifica el cálculo de $p_{N,n}(d)$ y no modifica el resultado final.

Para calcular $p_{N,n}(d)$ contamos el número de formas en las que se puede obtener una distancia d, analizando todas las posibles maneras de distribuir n apariciones de la palabra en un texto de longitud N. Tenemos en cuenta tres situaciones distintas:

i) La primera aparición de la palabra se sitúa en la posición $j_1 = d$, lo que es equivalente a obtener una distancia d al principio del intervalo $[0, N + 1]$: $d_0 = d$. Esto deja $N - d$ posiciones posibles para las $n - 1$ apariciones restantes de la palabra, es decir, hay $\binom{N-d}{n-1}$ configuraciones diferentes con una distancia d al principio del intervalo $[0, N + 1]$.

ii) La última aparición de la palabra se sitúa en la posición $j_n = N + 1 - d$, lo que es equivalente a obtener una distancia d al final del intervalo $[0, N + 1]$: $d_n = d$. Por simetría, nos encontramos en la misma situación que en i): esto deja $N - d$

posiciones posibles para las $n-1$ apariciones restantes de la palabra, es decir, hay $\binom{N-d}{n-1}$ configuraciones diferentes con una distancia d al final del intervalo $[0, N+1]$.

iii) d es una distancia "real" entre dos apariciones de la palabra: $d_i = d$ para algún $i \in \{1, 2, \ldots, n-1\}$. Esto ocurrirá si tenemos dos apariciones consecutivas de la palabra en las posiciones $\{1, 1+d\}$, $\{2, 2+d\}$, $\{3, 3+d\}$,..., ó $\{N-d, N\}$. Como consecuencia hay $N-d$ situaciones diferentes en las que podemos encontrar dos apariciones consecutivas de una palabra a distancia d. Cada una de esas situaciones deja $N-d-1$ posiciones posibles para las $n-2$ apariciones restantes. Tenemos entonces $(N-d)\binom{N-d-1}{n-2}$ configuraciones diferentes con una distancia d en el interior de $[0, N+1]$.

Teniendo en cuenta los tres casos, el número total de formas en las que se puede obtener una distancia d es:

$$2\binom{N-d}{n-1} + (N-d)\binom{N-d-1}{n-2} = (n+1)\binom{N-d}{n-1} \tag{3.2}$$

Podemos obtener entonces $p_{N,n}(d)$ dividiendo (3.2) por el número total de formas de obtener cualquier distancia del conjunto $\{1, 2, ..., N-n+1\}$, esto es:

$$p_{N,n}(d) = \frac{(n+1)\binom{N-d}{n-1}}{\sum_{k=1}^{N-n+1}(n+1)\binom{N-k}{n-1}} = \frac{\binom{N-d}{n-1}}{\binom{N}{n}} \tag{3.3}$$

con $d = 1, 2, \ldots, N-n+1$.

Observemos que, si N y n son pequeños, podemos obtener $p_{N,n}(d)$ numéricamente comprobando todas las posibles configuraciones de las n apariciones de la palabra. Pero, en general, no se puede hacer en un tiempo razonable, debido a que el número de configuraciones sería muy grande. De ahí la utilidad del resultado obtenido en (3.3).

En la figura 3.1 se muestran varios ejemplos de $p_{N,n}(d)$ para $N = 200$ y diferentes números de apariciones n de la palabra. Para $n = 40$ el número de configuraciones posibles es $\binom{200}{40} \simeq 2 \times 10^{42}$, que no se pueden comprobar en un tiempo razonable. Para ese caso, además del resultado exacto de la ecuación (3.3), mostramos el resultado numérico obtenido generando 10^8 configuraciones aleatorias.

Antes de analizar las principales características de la distribución $p_{N,n}(d)$ obtenida en (3.3) observamos que:

i) La probabilidad de la distancia más pequeña posible, $d = 1$, es

$$p_{N,n}(1) = \frac{n}{N}, \tag{3.4}$$

que coincide con la probabilidad de encontrar la palabra en una posición determinada. Y la probabilidad de la distancia más grande posible, $d = N + 1 - n$, es

$$p_{N,n}(N + 1 - n) = \frac{1}{\binom{N}{n}}. \tag{3.5}$$

ii) Para $n = 1$, es decir, si la palabra aparece solo una vez, todas las distancias tienen la misma probabilidad,

$$p_{N,1}(d) = \frac{1}{N}. \tag{3.6}$$

Para $n = 2$, la probabilidad decrece linealmente con la distancia d,

$$p_{N,2}(d) = \frac{2(N - d)}{N(N - 1)}. \tag{3.7}$$

Y, en general, para una palabra que aparece n veces, $p_{N,n}(d)$ es un polinomio de grado $(n - 1)$ en d.

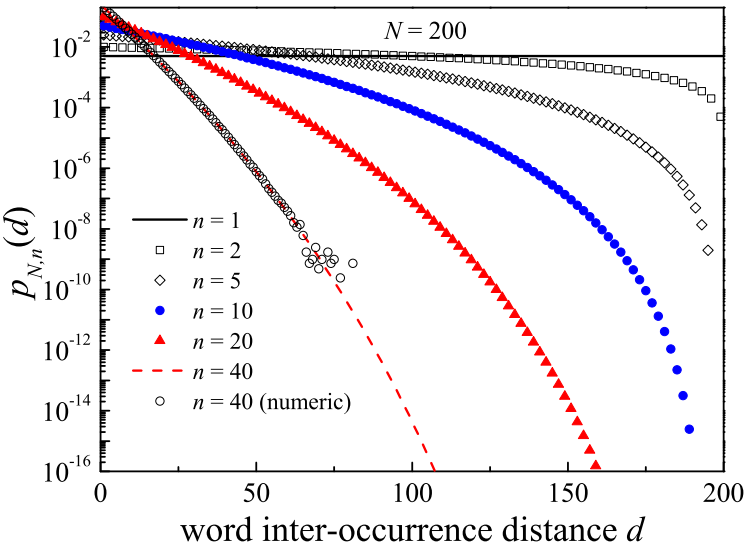

Figura 3.1: Ejemplos de distribuciones $p_{N,n}(d)$ obtenidas de la ecuación (3.3) para diferentes números de apariciones n y para $N = 200$. En el caso $n = 40$ también se muestra (círculos) la distribución $p_{N,n}(d)$ obtenida numéricamente generando 10^8 configuraciones aleatorias.

3.2. Algunas propiedades de la distribución

Hemos obtenido analíticamente que la distribución exacta esperada por azar para N y n finitos es

$$p_{N,n}(d) = \frac{\binom{N-d}{n-1}}{\binom{N}{n}} \ , \quad d = 1, 2, \ldots, N-n+1.$$

Por construcción, y por ser además una función de masa de probabilidad, es claro que

$$\sum_{d=1}^{N+1-n} p_{N,n}(d) = \sum_{d=1}^{N+1-n} \frac{\binom{N-d}{n-1}}{\binom{N}{n}} = \frac{\binom{N}{n}}{\binom{N}{n}} = 1 \tag{3.8}$$

La inclusión de las condiciones de contorno $d_0 = j_1$ y $d_n = N + 1 - j_n$ implica que $\sum_{i=0}^{n} d_i = N + 1$. Así que la distancia media es $\langle d \rangle = (N+1)/(n+1)$ para cualquier configuración de n apariciones de la palabra. El mismo resultado se obtiene de $p_{N,n}(d)$, entendiendo esta como una función de masa de probabilidad:

$$\langle d \rangle = \sum_{d=1}^{N+1-n} d\, p_{N,n}(d) = \sum_{d=1}^{N+1-n} d \frac{\binom{N-d}{n-1}}{\binom{N}{n}} = \frac{N+1}{n+1} \tag{3.9}$$

Sin incluir las condiciones de contorno, la media esperada es la misma porque la distribución es también $p_{N,n}(d)$ [1]. Sin embargo, medias obtenidas de diferentes configuraciones de las n posiciones de la palabra serán en general distintas a (3.9), lo que se evita usando condiciones de contorno.

Para el momento de segundo orden tenemos

$$\langle d^2 \rangle = \sum_{d=1}^{N+1-n} d^2 \frac{\binom{N-d}{n-1}}{\binom{N}{n}} = \frac{(N+1)(2N-n+2)}{(n+1)(n+2)} \tag{3.10}$$

y el de tercer orden viene dado por

$$\langle d^3 \rangle = \frac{(N+1)(12N - 7n - 6Nn + 6N^2 + n^2 + 6)}{(n+1)(n+2)(n+3)} \tag{3.11}$$

[1] Se tendrá una distancia d si hay dos apariciones consecutivas de la palabra en alguna de las posiciones $\{1, 1+d\}, \{2, 2+d\}, \{3, 3+d\} \ldots, \{N-d, N\}$. Existen, por tanto, $(N-d)\binom{N-d-1}{n-2}$ formas diferentes de obtener una distancia d en el rango $(1, N+1-n)$, luego

$$p_{N,n}(d) = \frac{(N-d)\binom{N-d-1}{n-2}}{\sum_{j=1}^{N+1-n}(N-j)\binom{N-j-1}{n-2}} = \frac{\binom{N-d}{n-1}}{\binom{N}{n}}.$$

Es posible calcular momentos de orden superior a partir de la distribución $p_{N,n}(d)$, aunque conllevaría una mayor complejidad en sus expresiones explícitas.

De la expresión de los dos primeros momentos, se llega a que la varianza viene dada por

$$\sigma^2 = \langle d^2 \rangle - \langle d \rangle^2 = \frac{n(N+1)(N-n)}{(n+1)^2(n+2)} \tag{3.12}$$

Observemos que, para N fijo, σ^2 alcanza su máximo en $n = 1$ y es decreciente en n, tendiendo a 0 cuando n tiende a N. Por otro lado, fijando n, $\sigma^2 \sim N^2/n^2$ para N grande.

En lo que respecta a la distribución acumulada $P_{N,n}(k) \equiv \text{Prob}\{d \leq k\}$, su expresión explícita viene dada por

$$P_{N,n}(k) = \sum_{d=1}^{k} \frac{\binom{N-d}{n-1}}{\binom{N}{n}} = 1 - \left(\frac{N-k}{n}\right) \frac{\binom{N-(k+1)}{n-1}}{\binom{N}{n}} \tag{3.13}$$

y puede ser expresada en términos de la función de masa de probabilidad (3.3), es decir,

$$P_{N,n}(k) = 1 - \left(\frac{N-k}{n}\right) p_{N,n}(k+1) \tag{3.14}$$

o, en una forma más compacta, como

$$P_{N,n}(k) = 1 - \frac{\binom{N-k}{n}}{\binom{N}{n}}. \tag{3.15}$$

Las ecuaciones (3.14) y (3.15) nos permiten obtener la función de distribución acumulada complementaria $Q_{N,n}(k) \equiv \text{Prob}\{d > k\} = 1 - P_{N,n}(k)$:

$$Q_{N,n}(k) = \left(\frac{N-k}{n}\right) p_{N,n}(k+1) = \frac{\binom{N-k}{n}}{\binom{N}{n}}. \tag{3.16}$$

3.2.1. Propiedades asintóticas

Estudiamos ahora la distribución $p_{N,n}(d)$ para N grande, la cual, por definición de número combinatorio, vendría dada por

$$p_{N,n}(d) = \frac{(N-d)!n!(N-n)!}{(n-1)!(N-d-(n-1))!N!}. \tag{3.17}$$

De este modo, empleando las aproximaciones para N grande

$$\frac{(N-d)!}{N!} \sim \frac{1}{N^d} \tag{3.18}$$

$$\frac{(N-n)!}{(N-n-(d-1))!} \sim (N-n)^{d-1} \tag{3.19}$$

obtenemos que

$$\hat{p}_{N,n}(d) = \frac{n}{N^d}\,(N-n)^{d-1} = \frac{n}{N}\left(1-\frac{n}{N}\right)^{d-1} \tag{3.20}$$

donde denotamos por $\hat{p}_{N,n}(d)$ a la distribución asintótica para N grande.

Observamos que $\hat{p}_{N,n}(d)$ no es más que la distribución geométrica $p_{\text{geo}}(d)$ (3.1) con $p \equiv n/N$. Hemos comprobado que asumir la distribución geométrica como la referencia para una palabra distribuida al azar en un texto es solo correcto en el límite para N y n grandes. Así que puede llevar a resultados incorrectos si se usa en un texto corto, como veremos posteriormente.

3.2.2. Variabilidad máxima

Estudiamos ahora la variabilidad de la distribución $p_{N,n}(d)$ mediante el coeficiente de variación (c_v)

$$c_v \equiv \frac{\sqrt{\sigma^2}}{\langle d \rangle} = \frac{\sigma}{\langle d \rangle} \tag{3.21}$$

usado habitualmente para caracterizar la desviación del azar de series temporales en muchos campos científicos como inmunología [Reed et al., 2002], dinámicas humanas [Goh and Barabási, 2008] o sistemas complejos [Guo et al., 2017]. Recordemos que dicho coeficiente era el empleado como medida de relevancia en [Ortuño et al., 2002], donde se denotaba σ a la desviación estándar del conjunto de distancias normalizadas.

Usando (3.9) y (3.12) obtenemos para $p_{N,n}(d)$ que

$$c_v(N,n) = \sqrt{\frac{n(N-n)}{(N+1)(n+2)}} \tag{3.22}$$

Observamos que el coeficiente c_v crece con N y tiende a 1 cuando $N \to \infty$. Por otro lado, c_v alcanza un valor máximo como función de n. En la figura 3.2 se muestra el comportamiento de c_v en función de n para varios valores de N y se observa claramente el máximo de c_v. Si consideramos n como una variable continua, de $\partial c_v/\partial n = 0$, llegamos a que el máximo para c_v se obtiene para

$$n_{\text{máx}} = \sqrt{2N+4} - 2 \tag{3.23}$$

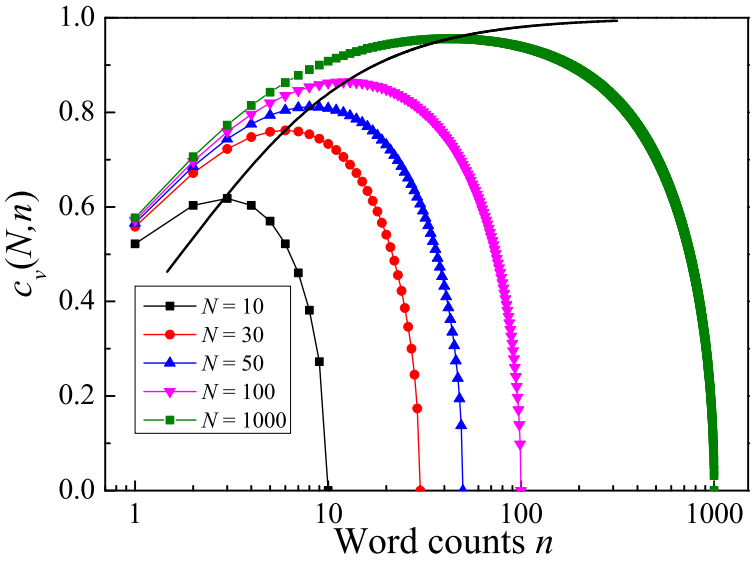

Figura 3.2: Coeficiente de variación $c_v(N, n)$ como función de n para varios valores de longitud del texto N. Se observa que, en cada caso, hay un valor de c_v máximo $c_{v,\text{máx}}$ que se alcanza para un n particular, $n_{\text{máx}}$. La línea continua une los valores de $c_{v,\text{máx}}$ como función de $n_{\text{máx}}$.

Incluimos también en la figura 3.2 la línea del máximo de c_v ($c_{v,\text{máx}}$) obtenido de $c_{v,\text{máx}} = c_v(N, n_{\text{máx}})$ y dibujado como función de $n_{\text{máx}}$.

A partir de (3.23), y teniendo en cuenta que la longitud típica N de los textos es al menos de varios cientos, podemos concluir que la distribución presenta máxima diversidad (o mínima homogeneidad) para $n \simeq \sqrt{2N}$.

Nótese que la distribución de las distancias entre apariciones sucesivas $p_{N,n}(d)$ se obtiene considerando una palabra con n apariciones distribuidas al azar en un texto de longitud N, lo que normalmente es una referencia para homogeneidad. Sin embargo, dependiendo de n, cambia la homogeneidad de la distribución e incluso presenta un mínimo.

3.3. Cuantificando *clustering*

Siguiendo con la idea, vista en capítulos anteriores, de que las palabras relevantes (palabras clave) en un texto tienden a formar *clusters* (se usan más frecuentemente en unas partes del texto que en otras, dando lugar a una distribución espacial con muchas fluctuaciones), veamos si los resultados obtenidos ahora nos permiten plantear una medida de *clustering* que capture con más precisión las desviaciones de la distribución espacial de una palabra de lo esperado si se distribuye al azar.

Proponemos ahora la medida de *clustering* de una palabra definida como

$$K(N,n) = \frac{c_{v,\text{obs}}(N,n)}{c_{v,\text{exp}}(N,n)}, \tag{3.24}$$

donde N es la longitud del texto, n el número de apariciones de la palabra en dicho texto y $c_{v,\text{obs}}(N,n)$ ($c_{v,\text{exp}}(N,n)$) el coeficiente de variación observado (esperado si las apariciones fueran al azar).

La diferencia con la medida σ_{nor} (y, en consecuencia, con C) radica en la forma en la que se calculan $c_{v,\text{exp}}(N,n)$ y $c_{v,\text{obs}}(N,n)$. En particular, $c_{v,\text{exp}}(N,n)$ se calculaba usando la distribución geométrica (3.1) en lugar de la obtenida para n y N finitos (3.3), para la que c_v viene dado por la ecuación (3.22).

Algunos autores definen una medida de *clustering* solo mediante el numerador de la ecuación (3.24) ($c_{v,\text{obs}}(N,n)$) con el nombre de *intermittency index* [Altmann et al., 2009; Amancio, 2015; Amancio et al., 2013]. En este caso usan como referencia para una palabra distribuida al azar la distribución exponencial, para la que $c_{v,\text{exp}}(N,n) = 1$. Sin embargo, como hemos observado previamente, $c_{v,\text{exp}}$ no es constante (figura 3.2), depende de N y n, y toma valores alejados de la unidad.

Con la definición de la ecuación (3.24), los resultados de $K(N,n)$ para una palabra concreta se interpretan de la siguiente forma:

i) $K(N,n) > 1$ implica que las fluctuaciones en la distribución de distancias observada son mayores que las esperadas por azar, lo que conlleva que la palabra se atrae a sí misma y forma *clusters*.

ii) $K(N,n) \simeq 1$ indica que las fluctuaciones de las distancias son esencialmente las esperadas si la palabra se distribuye aleatoriamente a lo largo del texto.

iii) $K(N,n) < 1$ sugiere pocas fluctuaciones de distancias, implicando la existencia de repulsión.

Analizamos ahora los cambios mencionados en la forma de obtener $c_{v,\text{obs}}(N,n)$ y $c_{v,\text{exp}}(N,n)$, y sus implicaciones.

3.3.1. $c_{v,\text{exp}}(N,n)$: distribución geométrica vs. exacta

Vimos anteriormente que la distribución geométrica

$$p_{\text{geo}}(d) = p(1-p)^{d-1} \tag{3.25}$$

se asumía como la esperada para una palabra con n ocurrencias distribuidas aleatoriamente en un texto de longitud N (siendo $p \equiv n/N$). Sin embargo, vimos también (sección 3.2.1) que esto es sólo correcto asintóticamente, para N y n tendiendo a infinito.

Por otro lado, habíamos obtenido la distribución exacta para N y n finitos, $p_{N,n}(d)$ (véase (3.3)) como:

$$p_{N,n}(d) = \frac{\binom{N-d}{n-1}}{\binom{N}{n}}. \tag{3.26}$$

Si pensamos en el caso de libros largos, tendremos N del orden de 10^5–10^6 palabras y las palabras relevantes con una frecuencia n en el rango de varias decenas a unos pocos cientos. Con estos números, las diferencias entre el caso asintótico y la distribución exacta son pequeñas y será suficiente con la aproximación geométrica (ver más abajo). Por ello los detectores de palabras clave que usan medidas de *clustering* similares a (3.24) con la distribución geométrica como la esperada por azar funcionan bastante bien en libros suficientemente largos [Ortuño et al., 2002; Carpena et al., 2009; Herrera and Pury, 2008; Carretero-Campos et al., 2013; Mehri et al., 2015].

Sin embargo, para textos cortos o moderadamente largos (como artículos científicos, informes, páginas web, etc.), $N \sim 10^3$–10^4 y n es como mucho del orden de unas pocas decenas. En este caso, si se asume $p_{\text{geo}}(d)$ como la distribución de distancias esperada por azar en lugar de $p_{N,n}(d)$, se pueden introducir errores importantes cuando estimamos la probabilidad de una distancia dada.

En la figura 3.3, ilustramos las diferencias entre la distribución geométrica y el resultado exacto para N y n finito en varios casos:

- La Figura 3.3a muestra la distribución exacta (símbolos) y la correspondiente distribución geométrica (líneas continuas) para diferentes valores de longitud del texto N y de frecuencia de la palabra n, en función de la variable espacial natural, la distancia normalizada $\bar{d} = d/\langle d \rangle$. Para cada combinación de N y n, la probabilidad p de la correspondiente distribución geométrica se obtiene como $p = n/N$. Para valores grandes de n (ver casos $n = 200$ en la figura), que son los esperados en un libro largo, los resultados para ambas distribuciones son muy similares independientemente de N. Sin embargo, para valores pequeños de n (ver casos $n = 5$ en la figura), se observa que los resultados son similares solo en el rango $\bar{d} \leq 2$ mientras que, para $\bar{d} > 2$, los resultados para $p_{N,n}(d)$ son órdenes de magnitud más pequeños que los de $p_{\text{geo}}(d)$. Así que, si se usa la distribución geométrica para estimar la probabilidad de una distancia en ese rango, dicha probabilidad estaría severamente sobreestimada.

- La figura 3.3b muestra el ratio $r(d) \equiv p_{N,n}(d)/p_{\text{geo}}(d)$, también como función de la

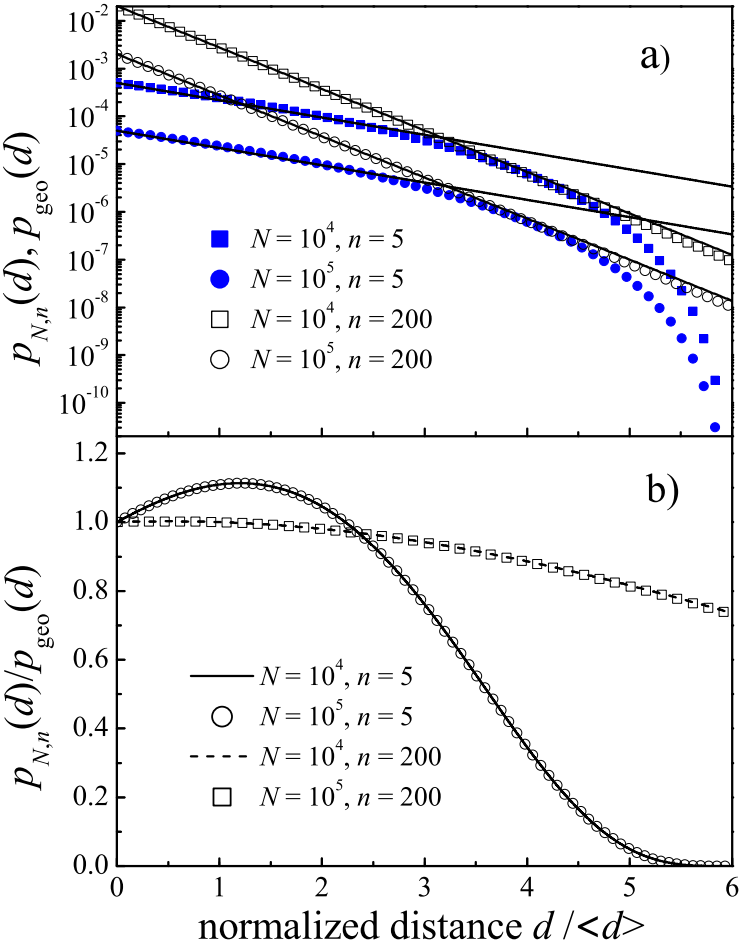

Figura 3.3: a) Distribución exacta de las distancias entre apariciones $p_{N,n}(d)$ (símbolos) y la correspondiente distribución geométrica $p_{\text{geo}}(d)$ (líneas continuas) como función de la distancia normalizada $d/\langle d\rangle$ para diferentes combinaciones de valores de N y n. Para las distribuciones geométricas, $p = n/N$. b) El ratio $p_{N,n}(d)/p_{\text{geo}}(d)$ para los 4 casos mostrados en a). Nótese que las curvas colapsan para palabras con la misma frecuencia n, indicando que n es la variable natural para medir la desviación del caso asintótico.

distancia normalizada $\bar{d}$, para comparar mejor ambos casos para distintos valores de n. En primer lugar, observamos que las curvas con el mismo valor de n colapsan, independientemente de N. Esto sugiere que n es la variable natural para medir la desviación del caso asintótico. En segundo lugar, observamos que, como esperábamos, para n grande, $r(d) \sim 1$ en todo el rango estudiado de $\bar{d}$. Sin embargo, para n pequeño, se presentan dos regímenes:

1. En el rango $0 < \bar{d} \leq 2$ el ratio $r(d)$ es mayor que uno, luego la probabilidad de una distancia en ese rango es subestimada cuando usamos $p_{\text{geo}}(d)$. En dicho intervalo se observa un máximo para $r(d)$. Usando (3.25) and (3.26) se puede probar que para N grande el máximo se alcanza en $d_{\max} = N/n$ y su valor es $r(d_{\max}) \simeq 1 + 1/(2n)$.

2. En el rango $\bar{d} > 2$, $r(d)$ decrece abruptamente a medida que crece $\bar{d}$, conllevando una gran sobreestimación al usar $p_{\text{geo}}(d)$.

Estas diferencias entre las distribuciones exacta y geométrica se verán también reflejadas en sus respectivos coeficientes de variación. Recordemos que en el caso geométrico

$$c_{v,\text{geo}}(N,n) = \sqrt{1 - n/N} = \sqrt{1-p}, \tag{3.27}$$

y en el caso de $p_{N,n}(d)$ (ver ecuación (3.22))

$$c_{v,\text{exact}}(N,n) = \sqrt{\frac{n(N-n)}{(N+1)(n+2)}}. \tag{3.28}$$

En la figura 3.4 se muestran $c_{v,\text{geo}}$ y $c_{v,\text{exact}}$ en función de n, y consideramos dos casos. En el primero fijamos la longitud del texto N (línea discontinua y círculos) y en el segundo fijamos p de modo que $N = n/p$ (línea continua y cuadrados). Tal y como esperábamos, a medida que aumenta n, $c_{v,\text{exact}}$ tiende asintóticamente a $c_{v,\text{geo}}$ en ambos casos. Así que, para valores grandes de n, como ocurre en libros largos, no habría diferencia en usar $c_{v,\text{exact}}$ o $c_{v,\text{geo}}$ en la medida de relevancia (3.24). Sin embargo, para n pequeño, que es la situación normal en textos cortos, las diferencias son sustanciales. De hecho se observa que $c_{v,\text{geo}}/c_{v,\text{exact}} > 1$ y, cuanto más pequeño sea n, mayor es el ratio, indicando que el uso de $c_{v,\text{exact}}$ en lugar de $c_{v,\text{geo}}$ en (3.24) es más sensible a la detección de *clustering* en palabras de frecuencia baja. Veremos más ejemplos en secciones posteriores.

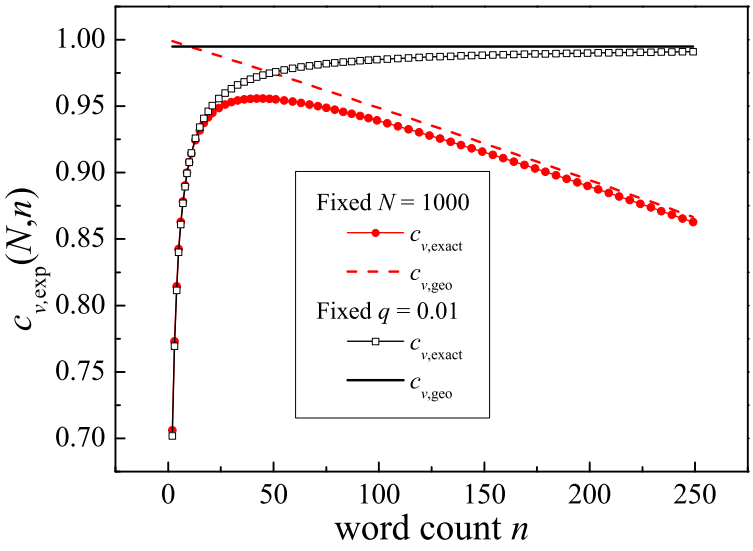

Figura 3.4: Coeficientes de variación esperados obtenidos de la distribución geométrica ($c_{v,\mathrm{geo}}$) y del resultado exacto ($c_{v,\mathrm{exact}}$) como función de n. Distinguimos entre un texto con una longitud fija de $N = 1000$ palabras y una palabra con una probabilidad fija $p = 0.01$ (y, por tanto, $N = n/p$).

3.3.2. $c_{v,\mathrm{obs}}(N,n)$: la necesidad de condiciones de contorno

Consideremos una palabra que aparece n veces en un texto de longitud N en las posiciones $j_1, j_2, \ldots, j_n$. Si no consideramos condiciones de contorno, tendremos el conjunto de $n-1$ distancias entre apariciones obtenido de $d_i = j_{i+1} - j_i$. En este caso, la media $\langle d_{\mathrm{obs}} \rangle$ difiere en general de la esperada por azar $\langle d \rangle = (N+1)/(n+1)$ ya que, dependiendo de la distribución espacial particular de la palabra dada, $\sum_{i=1}^{n-1} d_i$ varía en el rango $n - 1 \leq \sum_{i=1}^{n-1} d_i \leq N - 1$. Recordemos que $\langle d \rangle$ es la misma independientemente de si aplicamos condiciones de contorno o no (3.9). Sin embargo, con condiciones de contorno, la media observada coincide siempre con la esperada ya que en este caso tenemos $n+1$ distancias con $\sum_{i=1}^{n+1} d_i = N + 1$.

Obviamente $\langle d_{\mathrm{obs}}^2 \rangle$ será también muy diferente si la evaluamos a partir de las $n-1$ distancias o considerando condiciones de contorno con $n+1$ distancias. Como consecuencia, el coeficiente de variación observado $c_{v,\mathrm{obs}}(N,n)$ cambiará sustancialmente dependiendo de si se imponen condiciones de contorno o no.

Mostramos ahora un ejemplo de *clustering* extremo que sería interpretado como una fuerte auto-repulsión si no se usan condiciones de contorno. En efecto, supongamos que tenemos un texto de longitud N y una palabra que aparece n veces, lo más *clusterizada* posible, con todas sus apariciones separadas por distancia uno desde la posición m, es decir, en $m, m+1, \ldots, m+n-1$ donde $m \in [1, N+1-n]$. Obtenemos entonces que $\langle d_{\mathrm{obs}} \rangle = \langle d_{\mathrm{obs}}^2 \rangle = 1$, de donde $c_{v,\mathrm{obs}}(N,n) = 0$. Esto lleva (3.24) a $K(N,n) = 0$ que,

como vimos previamente, indica repulsión extrema. Sin embargo, si usamos condiciones de contorno, tendremos dos distancias adicionales $d_0 = m$ y $d_n = N+1-(m+n-1)$, y esto nos llevará a obtener el valor de *clustering* más alto posible, como veremos posteriormente. Este ejemplo muestra la necesidad del uso de condiciones de contorno.

3.3.3. Valores de *clustering* extremos

Una vez que hemos analizado cómo calcular adecuadamente $c_{v,\text{exp}}$ y $c_{v,\text{obs}}$ para usarlos en la medida de *clustering* $K(N,n)$ (3.24), estudiaremos ahora cuáles son los valores extremos de $K(N,n)$.

Consideramos un texto de longitud N y una palabra que aparece n veces en el texto. Fijados N y n, el valor de $c_{v,\text{exp}}(N,n)$ viene dado por la ecuación (3.28). Por otro lado, $c_{v,\text{obs}}$ será máximo cuando σ^2_{obs} alcance también su valor máximo. Asumiendo condiciones de contorno, esto ocurrirá cuando todas las n apariciones de la palabra estén concentradas al principio, al final, o divididas entre el principio y el final del texto. En todos estos casos, el conjunto de $n+1$ distancias consistirá en n distancias de valor 1 y una distancia de valor $N+1-n$. Por tanto, como $\langle d_{\text{obs}} \rangle = (N+1)/(n+1)$, tenemos que:

$$\sigma^2_{\text{obs,max}} = \frac{n + (N+1-n)^2}{n+1} - \left(\frac{N+1}{n+1}\right)^2. \tag{3.29}$$

Usando esta expresión en (3.21) podemos obtener el valor máximo de $c_{v,\text{obs}}$ y, mediante (3.24), el valor máximo de *clustering*

$$K_{\max}(N,n) = \sqrt{\frac{(N-n)(n+2)}{N+1}} \simeq \sqrt{n+2}, \tag{3.30}$$

donde la aproximación funciona porque típicamente $n \ll N$.

Una situación de *clustering* extremo cercana al valor máximo aparece cuando hay un único cluster que no está ni al principio ni al final del texto. Consideremos un único cluster que empieza en la posición j, ocupando entonces las posiciones $j, j+1, \ldots, j+n-1$. El conjunto de distancias está formado por $n-1$ distancias de valor 1, y por las dos distancias adicionales de valores j y $N+1-(j+n-1)$. Procediendo como antes, podemos obtener σ^2_{obs} y su correspondiente $K(N,n;j)$, que en este caso depende también de j. Obviamente, para $j=1$ y $j=N-n+1$, $K(N,n;j) = K_{\max}(N,n)$ ya que el cluster está localizado al principio o al final del texto, respectivamente. Para j en el rango $1 < j < N-n+1$,

$K(N,n;j)$ tiene un mínimo en $j_{\min} = (N-n+2)/2$, donde el valor de *clustering* es

$$\begin{aligned} K(N,n;j_{\min}) &= \sqrt{\frac{(N-n)(n+2)(n-1)}{2n(N+1)}} \\ &= \sqrt{\frac{n-1}{2n}}\, K_{\max}(N,n) \end{aligned} \tag{3.31}$$

Definimos $K_{\rm b}(N,n) \equiv \sqrt{(n-1)/2n}\, K_{\max}(N,n)$ y lo consideraremos como la cota inferior para *clustering* extremo.

Valores de *clustering* extremo similares se obtienen para situaciones en las que una palabra está localizada casi en su totalidad en un único *cluster* con unas pocas apariciones aisladas fuera del *cluster*, o una palabra localizada en dos *clusters* muy distantes.

Este análisis nos lleva a concluir que las palabras con *clustering* extremo se pueden identificar cuando el valor de *clustering* correspondiente está en el rango $K_{\rm b}(N,n) \leq K(N,n) \leq K_{\max}(N,n)$. Usaremos esta propiedad en la siguiente sección.

Notemos que los valores de *clustering* extremo mostrados en las ecuaciones (3.30) y (3.31) se han obtenido bajo la hipótesis de tener *clusters* con muchas distancias de valor 1. Obviamente esto no ocurre en un texto real (aunque sí puede ocurrir en otras secuencias simbólicas), donde las distancias típicas entre palabras son siempre mayores que 1 incluso dentro de un *cluster*. Sin embargo, dado que dentro del *cluster* las distancias típicas son mucho más pequeñas que la media esperada (3.9), y que el coeficiente de variación mide las fluctuaciones de las distancias comparadas con la media, los valores de *clustering* extremo de las ecuaciones (3.30) y (3.31) son también aplicables en este caso.

3.4. Aplicaciones

Los valores de la medida de *clustering* (3.24) pueden cambiar sustancialmente si se calcula como hemos propuesto en las secciones anteriores, usando que $c_{v,\exp}(N,n) = c_{v,\text{exact}}(N,n)$ y aplicando condiciones de contorno para obtener $c_{v,\text{obs}}(N,n)$, o si se calcula usando que $c_{v,\exp}(N,n) = c_{v,\text{geo}}(N,n)$ y sin aplicar condiciones de contorno para obtener $c_{v,\text{obs}}(N,n)$. Para distinguir ambos casos, las denotaremos $K(N,n)$ y $\hat{K}(N,n)$, respectivamente. Nótese que $\hat{K}(N,n)$ no es más que la medida σ_{nor} descrita en el capítulo 1.

Como consecuencia de lo analizado en las secciones anteriores, la hipótesis es que usar $K(N,n)$ en lugar de of $\hat{K}(N,n)$ debería mejorar los resultados en dos aspectos:

i) $K(N,n)$ debería detectar mejor el *clustering* asociado a las palabras relevantes, especialmente en el caso de textos cortos y palabras de frecuencia baja. La razón es

el uso de $c_{v,\text{exact}}(N,n)$ en $K(N,n)$ en lugar del asintótico $c_{v,\text{geo}}(N,n)$ en $\hat{K}(N,n)$, el cual sobreestima el *clustering* esperado por azar para palabras de frecuencia baja. Hasta donde nosotros sabemos, es la primera medida de relevancia que incorpora la información procedente de la distribución exacta esperada por azar en lugar de la asintótica.

ii) Solo cuando se consideran condiciones de contorno los diferentes regímenes de distribución espacial de las palabras pueden ser adecuadamente asociados a valores de *clustering*: $K > 1$ implica *clustering*, $K \simeq 1$ indica distribución al azar y $K < 1$ auto-repulsión. Esto no es siempre cierto cuando se usa $\hat{K}$, como se ha mostrado anteriormente con un ejemplo de *clustering* extremo interpretado como repulsión.

Como vimos en el capítulo anterior, el libro *The Origin of Species* se ha usado como la referencia estándar para muchos algoritmos de detección de palabras clave [Herrera and Pury, 2008; Carretero-Campos et al., 2013; Mehri et al., 2015], en parte porque contiene su propio glosario que permitía decidir la relevancia de una palabra de forma menos subjetiva que en otros casos. Al analizar el libro completo, todas las palabras relevantes son relativamente frecuentes (n alrededor de varias decenas), por lo que $c_{v,\text{exact}}$ y $c_{v,\text{geo}}$ son casi idénticos (ver figura 3.4). De hecho, otros detectores de palabras clave funcionan bien en este caso. Mostramos en la tabla 3.1 el *ranking* de las 10 palabras más relevantes obtenidas usando K y $\hat{K}$ y, por comparar, el obtenido con la medida E_{nor} [Herrera and Pury, 2008] usando la división en capítulos como partición. Los tres *rankings* son bastante similares y las palabras clave extraídas reflejan muy bien los contenidos del libro. Nótese que la palabra 'wax', de la que ya hablamos en la sección 2.1.2, aparece en el octavo lugar en el *ranking* de K, pero en el de $\hat{K}$ ocupa la posición 1406. Aquí se refleja el efecto de usar condiciones de contorno ya que la frecuencia de la palabra es relativamente alta y, por tanto, tiene poca influencia.

Sin embargo, si se considera un texto corto, las diferencias entre K y $\hat{K}$ se acentuarán debido al efecto de la frecuencia baja de las palabras (que también puede afectar considerablemente a otras medidas de relevancia). Mostramos, como hicimos en el capítulo 2 de esta memoria, los resultados de analizar el capítulo más corto del libro, titulado *Struggle for existence* (Capítulo III). En la tabla 3.2 se muestran resultados para K, $\hat{K}$ y E_{nor} (usando partición en párrafos). Observamos, como esperábamos, que K presenta el mejor comportamiento ya que tanto $\hat{K}$ como E_{nor} incluyen entre las primeras palabras de sus *rankings* palabras no relevantes. En el caso de $\hat{K}$ aparecen 'had' y 'said'. Para entender mejor las diferencias entre K y $\hat{K}$ en textos cortos, mostramos en la tabla 3.3 algunas palabras clave del capítulo analizado, junto a sus frecuencias y sus posiciones en ambos *rankings*.

K ranking	$\hat{K}$ ranking	E_{nor} ranking
formations	formations	hybrids
sterility	cells	sterility
hybrids	sterility	i
bees	hybrids	species
instincts	bees	islands
instinct	instincts	forms
cells	workers	instincts
wax	slaves	varieties
fertility	instinct	breeds
slaves	diagram	fertility

Tabla 3.1: *Ranking* de las 10 palabras más relevantes extraídas del libro *The Origin of Species*, de Charles Darwin. Las palabras están ordenadas en orden decreciente de K (primera columna), $\hat{K}$ (segunda columna), y la medida entrópica E_{nor} [Herrera and Pury, 2008] (tercera columna).

K ranking	$\hat{K}$ ranking	E_{nor} ranking
varieties	varieties	or
selection	had	been
bees	cattle	we
advantage	trees	the
heath	said	heath
individual	climate	i
competition	species	bees

Tabla 3.2: *Ranking* de las 7 palabras más relevantes extraídas del capítulo III del libro *The Origin of Species*, de Charles Darwin. Las palabras están ordenadas en orden decreciente de K (primera columna), $\hat{K}$ (segunda columna) and E_{nor} [Herrera and Pury, 2008] (tercera columna).

Se observa que, en general, para palabras con frecuencia relativamente grande, como 'varieties', las dos medidas son similares. De hecho, es la primera palabra de ambos *rankings* de relevancia. Sin embargo, para palabras relevantes con frecuencia pequeña ($n < 10$), observamos que K mejora considerablemente los resultados. Las palabras más relevantes para el capítulo ('selection', 'advantage', 'individual', 'competition', etc) están en la parte superior de su *ranking*, no así en el de $\hat{K}$, en el que ocupan posiciones bastante más lejanas.

word	word count n	rank K	rank $\hat{K}$
varieties	16	1	1
selection	6	2	36
advantage	5	4	197
individual	6	6	219
competition	7	7	11
natural	7	16	207

Tabla 3.3: *Ranking* de palabras relevantes extraídas del capítulo más corto (Capítulo III) del libro *The Origin of Species*, de Charles Darwin. Incluimos la frecuencia de la palabra en el capítulo (segunda columna), y la posición de la palabra en el *ranking* de relevancia obtenido usando K (tercera columna) o $\hat{K}$ (cuarta columna).

3.4.1. Palabras clave genéricas vs. específicas

Los resultados de *clustering* extremo de la sección 3.3.3 nos permiten sugerir una forma de clasificar las palabras clave (palabras con valores altos de K) en dos clases: genéricas y específicas.

Las palabras clave genéricas son palabras con valores altos de K, pero que se usan a lo largo de todo el texto. Se pueden identificar atendiendo a dos características principales: deben ser relativamente frecuentes y su valor de K, aunque alto, debe ser más pequeño que la cota de *clustering* extremo $K_{\mathrm{b}}(N, n)$ determinada en la sección 3.3.3.

Por otro lado, las palabras clave específicas están descritas por las siguientes particularidades: su frecuencia no puede ser grande y su valor de K debería estar próximo o superar la cota de *clustering* extremo. Notemos que esto solo puede ocurrir cuando la palabra está concentrada en un único cluster o en una situación similar, implicando que la palabra se usa solo en un contexto muy específico del texto.

En la figura 3.5 mostramos los valores de $K(N, n)$ para todas las palabras del vocabulario de *The Origin of Species* como función del número de apariciones de la palabra n. También incluimos dos líneas que corresponden al valor máximo de *clustering* $K_{\max}(N, n)$ y a la cota para *clustering* extremo $K_{\mathrm{b}}(N, n)$.

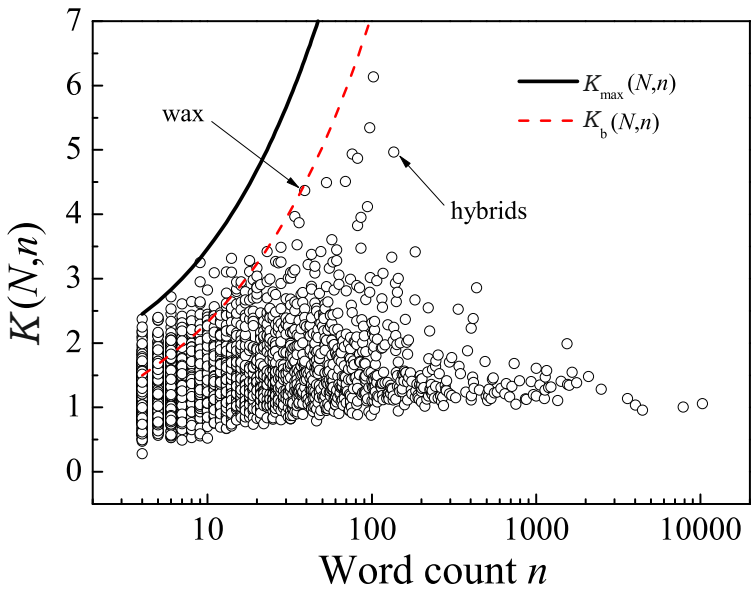

Figura 3.5: Valores de *clustering* $K(N,n)$ para las palabras del vocabulario del libro *The Origin of Species* como función de la frecuencia n. Incluimos solo palabras con $n > 3$. Las líneas corresponden al valor máximo de *clustering* $K_{\max}(N,n)$ y a la cota inferior de *clustering* extremo, $K_{\rm b}(N,n)$.

Indicamos en la figura dos ejemplos típicos correspondientes a palabras clave genéricas y específicas.

El caso de 'wax', ya comentado cuando discutíamos los resultados de la tabla 3.1, corresponde a una palabra clave específica ya que su valor de K está situado prácticamente sobre la línea de *clustering* extremo. Esta palabra sólo se detecta como palabra clave si se usa K (en lugar de $\hat{K}$), como consecuencia de aplicar condiciones de contorno.

La especificidad de 'wax' se puede apreciar mejor en la figura 3.6 (panel superior) donde las posiciones de las 39 apariciones de la palabra en el texto se indican con líneas verticales. De hecho, 'wax' aparece solo en el intervalo 71066-74741 (como ya comentamos en el capítulo 2 de esta memoria), incluido dentro del capítulo VII del libro, en el que se estudia el comportamiento de las abejas.

En cuanto a 'hybrids', es una palabra de mayor frecuencia ($n = 136$) y también mayor valor de K que 'wax'. Sin embargo, su valor de K está situado bastante por debajo de la línea de *clustering* extremo (ver la figura 3.5) y esta situación corresponde a una palabra clave genérica: altamente *clusterizada* pero que no se usa en un único contexto. De hecho esto se confirma en la figura 3.6 (panel inferior), donde mostramos las posiciones de las 136 apariciones de 'hybrids' a lo largo de todo el texto, y donde se muestra cómo la palabra está *clusterizada* pero se usa en diferentes contextos.

Se puede consultar la aplicación de estos resultados a otros textos en el Apéndice A.

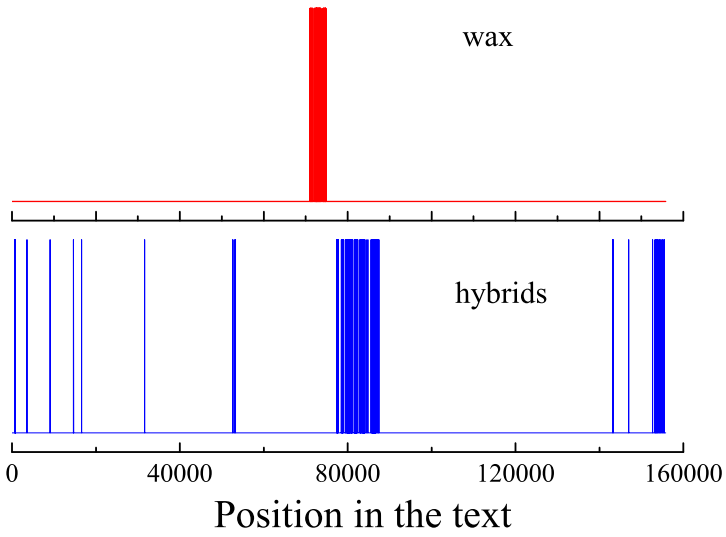

Figura 3.6: Posiciones de las palabras 'wax' ($n = 39$, panel superior) y 'hybrids' ($n = 136$, panel inferior) en el libro *The Origin of Species*

3.5. Conclusión

En este capítulo, se ha obtenido analíticamente la distribución exacta de las distancias entre apariciones sucesivas de una palabra en un texto (o, en general, de un símbolo en una secuencia simbólica) asumiendo que las posiciones en las que aparece se eligen al azar. Tradicionalmente se consideraba la distribución geométrica como la referencia en ese caso, pero comprobamos que solo se verifica asintóticamente. Se analizan las propiedades de la distribución y, en particular, se muestra que existe un valor de frecuencia para el que la variabilidad de la distribución es máxima. El conocimiento de la distribución exacta mencionada, junto a la aplicación de condiciones de contorno, permite mejorar la detección de *clustering* en secuencias simbólicas, especialmente si el símbolo aparece pocas veces. En textos, el análisis de valores extremos de *clustering* permite clasificar las palabras clave en genéricas y específicas.

Parte II

Correlaciones de largo alcance asociadas a palabras clave

Capítulo 4

Correlaciones de largo alcance. Generalidades

En la primera parte de esta memoria hemos visto que existe una relación clara entre la relevancia de una palabra y su distribución espacial a lo largo del texto, y que una medida adecuada de *clustering* nos permite detectar exitosamente las palabras clave de un texto, sin necesidad de disponer de información previa.

El *clustering* de una palabra relevante conlleva la presencia de grandes fluctuaciones de su frecuencia de aparición. Es intuitivo plantearse si la secuencia que representa las apariciones de dicha palabra a lo largo del texto tiene correlaciones de largo alcance. Si el *clustering* y las correlaciones están relacionados, una medida que cuantifique la existencia de correlaciones de largo alcance también nos serviría para detectar relevancia.

De hecho veremos en el capítulo 6 cómo la distribución espacial tan heterogénea de las palabras relevantes se debe a fuertes interacciones entre ellas, que se manifiestan como correlaciones de largo alcance en ley de potencias que se extienden a escalas mucho mayores de las que cabría esperar por las reglas sintácticas de escritura en el lenguaje humano. Y, tal como intuíamos, comprobaremos que el grado de correlaciones de largo alcance de una palabra va a ser una buena medida de su relevancia para el texto.

Antes de abordar dicho estudio, emplearemos este capítulo para definir formalmente el concepto de correlación de largo alcance, y realizar una breve revisión de métodos usados en la literatura para caracterizarlas y cuantificarlas. En concreto profundizaremos en uno de los métodos, denominado *Detrended Fluctuation Analysis* (DFA), que será el que emplearemos para cuantificar las correlaciones en textos. Analizaremos su comportamiento aplicándolo a secuencias sintéticas generadas con correlaciones controladas mediante el Método de Filtrado de Fourier. Por último, haremos un breve inciso sobre la generación y cuantificación de correlaciones en secuencias binarias, que nos será de utilidad para el

estudio de las estadísticas de tiempos de paso que realizaremos capítulos posteriores.

4.1. Caracterización

Hay evidencia en la literatura de que muchos sistemas físicos y biológicos poseen propiedades invariantes de escala caracterizadas por correlaciones de largo alcance: secuencias de ADN [Peng et al., 1992; Buldyrev, 2006], dinámica cardíaca [Ivanov et al., 1996, 1999; Kantelhardt et al., 2003; Ashkenazy et al., 2001], fluctuaciones electroencefalográficas humanas [Robinson and Harsch, 2002; Sapir et al., 2003], actividad motora humana [Hu et al., 2004], el modo de andar [Hausdorff et al., 2001; Ashkenazy et al., 2002], etc. También en meteorología [Koscielny-Bunde et al., 1998; Talkner and Weber, 2000; Govindan et al., 2001], señales sísmicas [Varotsos et al., 2003b], economía [Ivanov et al., 2004; Lee, 2009], música [Jennings et al., 2004], etc. Para comprender e interpretar la dinámica de tales sistemas surge la necesidad de entender, caracterizar y cuantificar las correlaciones de largo alcance.

Nótese que este comportamiento fue observado por primera vez por el hidrologista británico Harold Edwin Hurst cuando estudiaba cómo regular el flujo del río Nilo, el cual era conocido por el hecho de que largos periodos de sequía eran seguidos por largos periodos de inundaciones. En la figura 4.1 podemos observar los niveles anuales mínimos correspondientes a los años 622 - 1281.

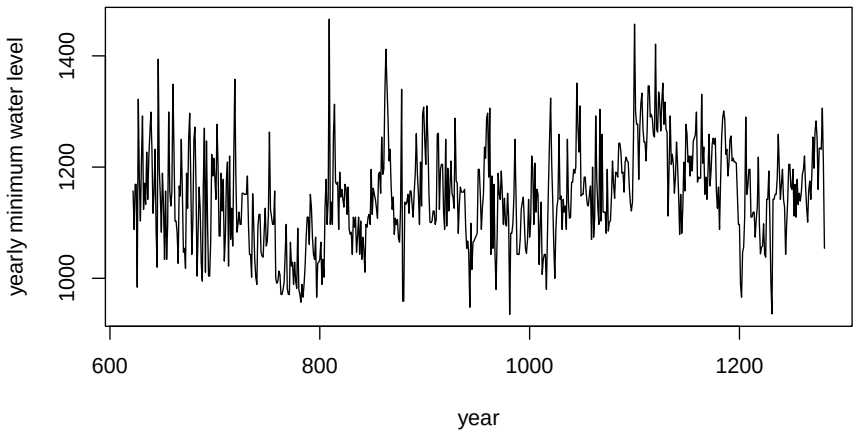

Figura 4.1: Niveles mínimos de agua anuales del río Nilo durante los años 622-1281 (obtenida de [Beran, 1994])

4.1.1. Definiciones previas

Para entender formalmente el concepto de correlaciones de largo alcance, necesitamos algunas definiciones previas [Beran, 1994]:

Sea $\{X_i\}_{i\in\mathbb{Z}}$ un proceso estocástico tal que $\mu_i = E(X_i)$ y $\sigma_i^2 = var(X_i)$ existen y son finitas. Definiremos la **autocovarianza** entre X_i y X_j, $\gamma(i,j)$, mediante la fórmula:

$$\gamma(i,j) = E\left[(X_i - \mu_i)\,(X_j - \mu_j)\right] \tag{4.1}$$

Diremos que el proceso $\{X_i\}_{i\in\mathbb{Z}}$ es **estacionario** (en sentido amplio: *wide sense stationary* **WSS**) si se cumplen las siguientes condiciones:

- La media $\mu_i = E(X_i)$ es constante, independiente de i: $\mu_i = \mu = cte$.
- La autocovarianza entre X_i y X_j sólo depende del intervalo transcurrido entre i y j: $\gamma(i,j) = \gamma(i-j) = \gamma(j-i)$. Si $k = |i-j|$, abusaremos de notación y escribiremos $\gamma(k)$.

Nótese que, como consecuencia, la varianza $\sigma_i^2 = var(X_i)$ también es constante, independiente de i, ya que: $\sigma_i^2 = var(X_i) = \gamma(i,i) = \gamma(0) = \sigma^2 = cte$.

La **autocorrelación** entre X_i y X_j viene definida de la siguiente forma:

$$\rho(i,j) = \frac{\gamma(i,j)}{\sigma_i \sigma_j} \tag{4.2}$$

Entonces, si el proceso es estacionario, tendremos que:

$$\rho(i,j) = \rho(i-j) = \rho(j-i) = \rho(k) = \frac{\gamma(k)}{\sigma^2} \tag{4.3}$$

Obtenemos así una medida de la dependencia entre eventos que están separados una distancia k.

Una vez definidos estos conceptos ya tenemos las herramientas necesarias para explicar qué es un proceso estocástico estacionario con correlaciones de largo alcance [Beran, 1994].

4.1.2. Caracterización

Caracterización I (mediante la función de autocorrelación).

Sea $\{X_i\}_{i\in\mathbb{Z}}$ un proceso estocástico estacionario. Diremos que es un proceso con **correlaciones de largo alcance o dependencia fuerte** si existen un número real $\delta \in (0,1)$

y una constante $c_\rho > 0$ verificando que:

$$\lim_{k\to\infty} \frac{\rho(k)}{c_\rho k^{-\delta}} = 1 \tag{4.4}$$

Diremos también que se trata de un proceso con **memoria larga** (*long-memory processes*), ya que la dependencia entre eventos disminuye muy lentamente al incrementar la distancia entre ellos. De hecho se tiene que las correlaciones decaen a cero tan lentamente que no son sumables:

$$\sum_{k=-\infty}^{\infty} \rho(k) = \infty \tag{4.5}$$

Este comportamiento en ley de potencias, $\rho(k) \approx c_\rho k^{-\delta}$ cuando $k \to \infty$, indica la carencia de cualquier escala característica y, por tanto, la invariancia de escala o comportamiento fractal de la secuencia (ver sección 4.1.4). Por el contrario, los procesos con correlaciones sumables

$$\sum_{k=-\infty}^{\infty} \rho(k) < \infty \tag{4.6}$$

serán llamados procesos con **correlaciones de corto alcance o dependencia débil o memoria corta**.

Caracterización II (mediante la función de densidad espectral).

También podemos caracterizar los procesos con correlaciones de largo alcance por medio de la función de **densidad espectral**. El teorema de Wiener-Kintchine establece que la densidad espectral de un proceso estocástico WSS es la transformada de Fourier de la correspondiente función de autocorrelación. Entonces, dado un proceso estocástico $\{X_i\}_{i\in\mathbb{Z}}$ estacionario con función de autocorrelación $\rho(k)$, definimos la densidad espectral en la forma [Priestley, 1981]:

$$f(\lambda) = \frac{\sigma^2}{2\pi} \sum_{k=-\infty}^{\infty} \rho(k) \exp(ik\lambda) \tag{4.7}$$

Diremos ahora que el proceso tiene correlaciones de largo alcance o dependencia fuerte o memoria larga si existen un número real $\beta \in (0, 1)$ y una constante $c_f > 0$ tales que:

$$\lim_{\lambda\to 0} \frac{f(\lambda)}{c_f\, |\lambda|^{-\beta}} = 1 \tag{4.8}$$

Observemos que las dos definiciones son equivalentes, lo cual nos permite relacionar

los exponentes δ y β:

$$-\beta = \delta - 1 \Rightarrow \beta = 1 - \delta \tag{4.9}$$

Todo lo anterior se extiende convenientemente al caso de procesos en tiempo continuo (se omiten los detalles).

4.1.3. Estimadores

Las dos caracterizaciones descritas en la sección anterior son asintóticas, y de ahí la dificultad que presenta el estudio de las correlaciones de largo alcance en secuencias finitas.

Función de autocorrelación estimada.

En la práctica, dada una secuencia estocástica $\{X_i\}_{i=1,\dots,N}$, la función de autocorrelación a distancia k se estima mediante la función $C(k)$ definida de la siguiente forma:

$$C(k) = \frac{\frac{\sum_{i=1}^{N-k} X_i X_{i+k}}{N-k} - \frac{\sum_{i=1}^{N-k} X_i}{N-k}\frac{\sum_{i=1}^{N-k} X_{i+k}}{N-k}}{\frac{\sum_{i=1}^{N}(X_i - \langle X\rangle)^2}{N}}, \text{ donde } \langle X\rangle = \frac{\sum_{i=1}^{N} X_i}{N} \tag{4.10}$$

De aquí en adelante nos referiremos a la función $C(k)$ como función de autocorrelación, sobreentendiendo que se trata de la función de autocorrelación estimada.

Por tanto, en la práctica diremos que tenemos correlaciones de largo alcance cuando la función $C(k)$ se comporta como una ley de potencias en la forma:

$$C(k) \propto \frac{1}{k^{\delta}} \tag{4.11}$$

Diremos también que las secuencias con correlaciones de corto alcance tendrán una función de autocorrelación $C(k)$ que decrece de manera suficientemente rápida como para que sea sumable. Para una secuencia aleatoria pura, es esperado que los coeficientes $C(k)$ sean próximos a 0 para $k \neq 0$.

Espectro de potencias.

Dada una secuencia estocástica $\{X_i\}_{i=1,\dots,N}$, llamaremos espectro de potencias y denotaremos por $S(f)$ (f denota frecuencia), al módulo al cuadrado de la transformada de Fourier de la secuencia. Tendremos entonces correlaciones de largo alcance si:

$$S(f) \propto \frac{1}{f^{\beta}} \tag{4.12}$$

Al caracterizar los procesos estacionarios con correlaciones de largo alcance mediante las ecuaciones (4.4) y (4.8), los exponentes δ y β estaban comprendidos entre 0 y 1, y relacionados en la forma $\beta = 1-\delta$. Sin embargo, en general, cuando calculemos el espectro de potencias de secuencias obtenidas de procesos reales, se pueden encontrar comportamientos en la forma (4.12) para valores de β fuera de ese rango, tanto menores que 0 como mayores que 1 (ya que muchas señales físicas y biológicas exhiben no estacionariedad). Diremos que:

i) Si $\beta = 0$, la secuencia no tiene correlaciones. En este caso el espectro es constante y por tanto contiene todas las frecuencias con el mismo peso. Por analogía a lo que ocurre con la luz blanca a este tipo de secuencias se les llama "ruido blanco" (*white noise*).

ii) Si $\beta > 0$, las correlaciones son positivas. Como casos particulares tenemos $\beta = 2$, llamado "ruido rojo" o "ruido browniano" en honor a Robert Brown. Este tipo de espectro es el correspondiente al movimiento browniano (ver sección 4.1.4), que no es estacionario. Y el caso $\beta = 1$, llamado "ruido $1/f$" o "ruido rosa", por estar comprendido entre el blanco y el rojo.

iii) Si $\beta < 0$, tenemos anticorrelaciones o correlaciones negativas.

En la figura 4.2 podemos observar gráficamente la diferencia entre anticorrelaciones o correlaciones negativas ($\beta < 0$), ausencia de correlaciones ($\beta = 0$) y correlaciones positivas ($\beta > 0$). Para ello hemos generado mediante el Método de Filtrado de Fourier (que será descrito en 4.2) cuatro secuencias todas de tamaño $N = 2^9$, y caracterizadas por un espectro en ley de potencias que sigue la ecuación (4.12) con $\beta = -1$, $\beta = 0$, $\beta = 1$ y $\beta = 2$, respectivamente.

Notemos que la función $C(k)$ sólo puede aplicarse a secuencias estacionarias y cuando $\beta \geq 1$, δ satura en 0 [Coronado and Carpena, 2005]:

$$\delta = \begin{cases} 1-\beta & \text{si } \beta < 1 \\ 0 & \text{si } \beta \geq 1 \end{cases} \tag{4.13}$$

4.1.4. Ruido gaussiano fraccionario y movimiento browniano fraccionario

Hemos visto anteriormente cómo las secuencias estocásticas con correlaciones de largo alcance son caracterizadas por estadísticas en ley de potencias. Esto indica la carencia de una escala característica y, por tanto, la invariancia de escala o comportamiento fractal de la secuencia. Por eso también son denominadas señales fractales o ruidos fractales.

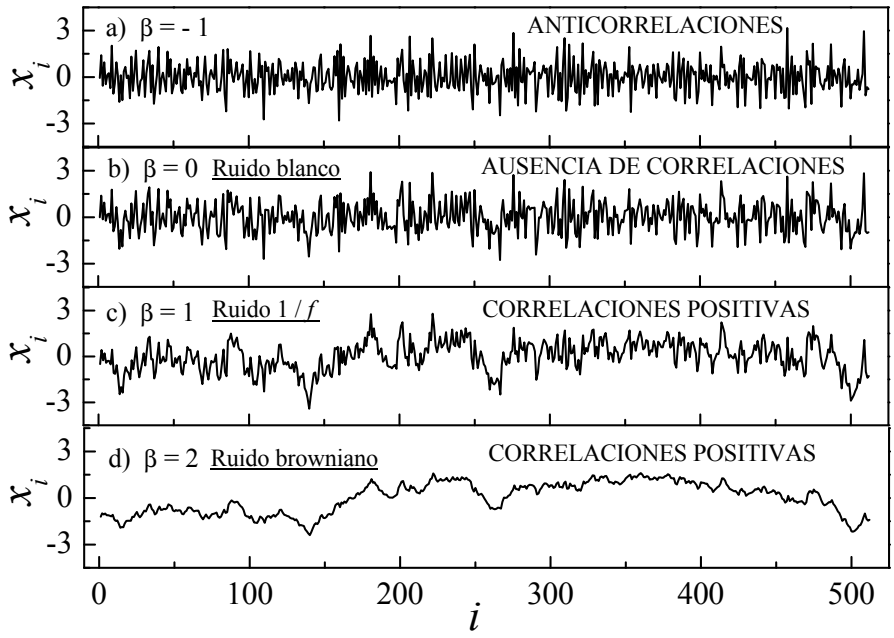

Figura 4.2: Comportamiento de las correlaciones de una secuencia estocástica en función del exponente de su espectro de potencias β: $\beta < 0$ indica anticorrelaciones o correlaciones negativas; $\beta = 0$ indica ausencia de correlaciones y $\beta > 0$ indica correlaciones positivas

En sentido determinístico hablaríamos de que la misma estructura geométrica es observada independientemente de la distancia a la cual uno la mira, pero en el contexto de procesos estocásticos [Beran, 1994] la fractalidad o autosimilitud es considerada en sentido estadístico: el proceso reescalado convenientemente en tiempo y estado es igual en distribución al proceso original. Hablamos entonces de obtener una réplica del total desde el punto de vista de las distribuciones. Se han observado propiedades fractales en muy diversos tipos de series temporales, presentando propiedades estadísticas similares independientemente de la escala de observación.

Los procesos autosimilares fueron introducidos por Kolmogorov en 1941, pero su relevancia estadística no se vio hasta que Mandelbrot los introdujo en el ámbito estadístico en 1968 [Mandelbrot and Van Ness, 1968]. En esta sección profundizaremos un poco más en los procesos autosimilares, concretamente en los que tienen incrementos estacionarios, lo cual nos llevará a definir dos tipos de procesos estocásticos de gran relevancia denominados movimiento browniano fraccionario y ruido gaussiano fraccionario.

Proceso autosimilar.

Un proceso $\{Y_t\}$ con parámetro continuo t es autosimilar con coeficiente de autosimilitud H [Beran, 1994], si para cualquier factor positivo c, el proceso reescalado con escala

ct, $\{c^{-H}Y_{ct}\}$ es igual en distribución al proceso original $\{Y_t\}$. Tendrá incrementos estacionarios si $\{X_t^T\}$ definido como $X_t^T = Y_t - Y_{t-T}$ es estacionario, para cualquier $T > 0$. El parámetro H caracterizará tanto al proceso autosimilar como a sus incrementos y variará en el rango $(0, 1)$.

Por simplicidad de notación, llamaremos $\{X_t\}$ al proceso de incrementos para $T = 1$.

Ruido gaussiano fraccionario y movimiento browniano fraccionario.

Supongamos ahora que el proceso de los incrementos $\{X_t\}$ es gaussiano y $E(X_t) = 0$. Asumiendo también que $E(Y_t) = 0$ y $Y_0 = 0$, la función de autocorrelación del proceso de los incrementos (que es estacionario) viene dada por la siguiente expresión [Beran, 1994]:

$$\rho(k) = \frac{1}{2}\left\{|k+1|^{2H} - 2\,|k|^{2H} + |k-1|^{2H}\right\} \tag{4.14}$$

Al ser gaussiano, mediante la media y la función de autocorrelación tenemos el proceso $\{X_t\}$ totalmente determinado. Por tanto para cada $H \in (0, 1)$ hay exactamente un proceso gaussiano que es el incremento estacionario de un proceso autosimilar con parámetro H, aquel cuya función de autocorrelación viene dada por la ecuación (4.14). Este proceso es llamado ruido gaussiano fraccionario (fGn, del inglés *fractional Gaussian noise*) y el correspondiente proceso autosimilar movimiento browniano fraccionario (fBm, del inglés *fractional Brownian motion*). Denotaremos a dicho proceso autosimilar por $B_H(t)$.

Los fBm's son una familia de procesos definidos por [Mandelbrot and Van Ness, 1968] como una generalización del movimiento browniano ordinario. Son de gran interés práctico, ya que proporcionan modelos útiles para una gran cantidad de series temporales naturales.

Movimiento browniano.

El movimiento browniano, $B(t)$, recibe su nombre del botánico Robert Brown [Brown, 1828] que estudió el movimiento de granos de polen en el agua. Inicialmente surge como modelo para el movimiento de partículas suspendidas en un líquido o un gas, pero posteriormente ha sido utilizado en la representación de procesos aleatorios en diversos campos aplicados (finanzas, biología, etc.). Definimos el movimiento browniano [Beran, 1994] como un proceso estocástico $B(t)$ gaussiano con incrementos $B(t_1) - B(t_2)$ independientes de media cero y varianza $|t_1 - t_2|$. De modo que la desviación estándar del incremento $B(t+T) - B(t)$ con $T > 0$ es $T^{1/2}$, lo cual se conoce como "ley $T^{1/2}$" [Mandelbrot and Van Ness, 1968]. Autosimilitud con $H = 1/2$ o, lo que es lo mismo, que $c^{-1/2}B(ct)$ tenga la misma distribución que $B(t)$, se sigue fácilmente de esta definición . Luego podemos

decir que $B(t) = B_H(t)$ para $H = 1/2$.

Para $H \neq 1/2$ los fBm's, $B_H(t)$, serán también procesos gaussianos de media 0 pero cuyos incrementos $B_H(t_1) - B_H(t_2)$ están correlacionados. La varianza será ahora proporcional a $|t_1 - t_2|^{2H}$, satisfaciendo la "ley T^H" para la desviación estándar [Mandelbrot and Van Ness, 1968]. Veremos ahora cómo los fBm's se van a dividir en tres familias según el valor del parámetro de autosimilitud H.

Consideremos primero un fGn caracterizado por el parámetro H. Se trata entonces de un proceso gaussiano con función de autocorrelación dada por la ecuación (4.14). Cuando $k \to \infty$ se tiene entonces que [Beran, 1994]:

$$\rho(k) \approx H(2H-1)k^{2H-2} \tag{4.15}$$

Como consecuencia,

1. Si $1/2 < H < 1$, las correlaciones decaen a 0 tan lentamente que no son sumables. Decimos entonces que el fGn tiene correlaciones de largo alcance o memoria larga.

2. Si $H = 1/2$, todas las correlaciones en intervalos no nulos son 0, es decir, no hay correlaciones. El fGn es un ruido blanco (la definición de ruido blanco fue dada en 4.1.3).

3. Si $0 < H < 1/2$, el fGn tiene correlaciones negativas o anticorrelaciones.

Si traducimos esto en términos de los fBm's diríamos [Delignieres et al., 2006]:

1. Si $1/2 < H < 1$, $B_H(t)$ es persistente ya que presenta una correlación positiva entre sus incrementos.

2. Si $H = 1/2$, los incrementos son independientes (ruido blanco). Sabemos que se trata del movimiento browniano ordinario $B(t)$.

3. Si $0 < H < 1/2$, $B_H(t)$ es antipersistente. Presenta una correlación negativa entre sus incrementos.

Los fGn's y fBm's son procesos interconvertibles (dado un fBm, sus incrementos constituyen un fGn y la integral de un fGn es un fBm) caracterizados por el mismo parámetro H. Se van a poder diferenciar mediante el exponente de la densidad espectral:

- La densidad espectral del proceso de los incrementos, el fGn, se comporta cerca del origen en la forma [Beran, 1994]:

$$f(\lambda) \approx c_f\, |\lambda|^{1-2H} \tag{4.16}$$

Esto nos permite relacionar H con β para un fGn:

$$\beta = 2H - 1 \tag{4.17}$$

Observemos que para un ruido blanco, como $H = 1/2$ tendremos $\beta = 0$, tal como vimos en 4.1.3 (el ruido blanco correspondía a un espectro constante).

- Para un fBm la no estacionariedad conlleva que no posee espectro convencional. Esto se soluciona mediante el espectro de Wigner-Ville (ver [Lowen and Teich, 2005]) el cual decae como $f^{-(2H+1)}$ (f representa frecuencia). Entonces podemos deducir también la relación entre H y β para un fBm:

$$\beta = 2H + 1 \tag{4.18}$$

Observemos ahora que para el movimiento browniano ordinario $H = 1/2$ implica $\beta = 2$. Por eso en 4.1.3 decíamos que el espectro con exponente $\beta = 2$ correspondía al movimiento browniano.

Como $0 < H < 1$, tenemos entonces que:

- ▷ Para un fGn: $-1 < \beta < 1$
- ▷ Para un fBm: $1 < \beta < 3$

Finalicemos esta sección ilustrando gráficamente fBm's y sus correspondientes incrementos fGn's para distintos valores de H, indicando también el correspondiente valor para β. Pondremos un ejemplo de cada una de las tres familias en la figura 4.3.

4.2. Generación: Método de Filtrado de Fourier

Dedicaremos esta sección a describir el método que emplearemos en nuestras simulaciones para generar secuencias estocásticas sintéticas con correlaciones de largo alcance: el Método de Filtrado de Fourier (FFM, del inglés *Fourier Filtering Method*). La generación de dichas secuencias con correlaciones controladas será de utilidad para ilustrar el funcionamiento del método que emplearemos para cuantificar las correlaciones y que será descrito en la sección posterior. Además, necesitaremos también de dicha generación para el estudio de las estadísticas de tiempos de paso de procesos con correlaciones que realizaremos en el capítulo 5, así como para crear modelos que posean correlaciones de largo alcance que, como veremos, reproducirán el modo en el que las palabras relevantes se distribuyen a lo largo de un texto (capítulo 6).

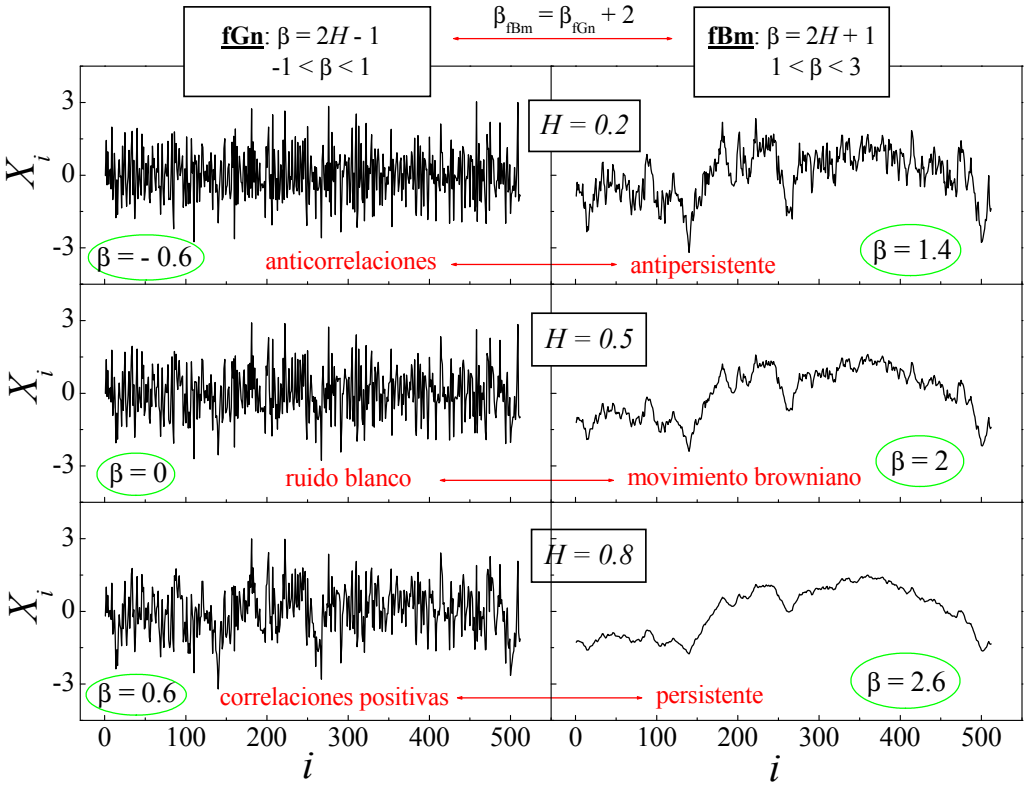

Figura 4.3: Ejemplos de fBm's y sus correspondientes incrementos fGn's para distintos valores de H, un caso de cada una de las tres familias que hemos diferenciado: $0 < H < 1/2$, $H = 1/2$ y $1/2 < H < 1$. Indicamos también el valor de β en cada caso. Todas las secuencias de tamaño $N = 2^9$

4.2.1. Transformada discreta de Fourier. Algoritmo FFT

La herramienta principal que utiliza este método es la transformada discreta de Fourier (DFT), computada mediante el algoritmo FFT. La DFT permite pasar la secuencia al dominio de las frecuencias f. Lo que ocurra a pequeñas frecuencias nos dará idea del comportamiento de la secuencia a distancias grandes (si la secuencia es espacial) o tiempos grandes (si la secuencia es temporal), y lo que ocurra a frecuencias grandes nos informará del comportamiento a distancias o tiempos pequeños.

Supongamos que partimos de una secuencia $\{h_i\}$ de tamaño N, de la que queremos calcular la transformada de Fourier. Supongamos que N es par (más adelante veremos que de hecho va a ser siempre potencia entera de 2). La transformada discreta de Fourier de $\{h_i\}$ será una secuencia de N números complejos $\{H_n\}$ tales que:

$$H_n = \sum_{k=0}^{N-1} h_k \exp(2\pi i k n/N) \tag{4.19}$$

Observemos que nuestros datos son discretos y tomados a distancia unidad. Por tanto, el rango de frecuencias en el que podemos estimar la transformada será $-f_c \leq f \leq f_c$, con $f_c = 1/2$ la llamada frecuencia crítica de Nyquist ($f_c = 1/2\Delta$, con Δ el intervalo de muestreo). Debido a las propiedades de periodicidad (H periódica en n con periodo N) y de simetría ($H(-n) = H(N-n)$) de la DFT, podemos tomar n en la ecuación (4.19) variando desde 0 a $N-1$ con las siguientes correspondencias:

$n = 0 \longrightarrow$ frecuencia 0

$1 \leq n \leq N/2 - 1 \longrightarrow$ frecuencias positivas $0 < f < 1/2$

$N/2 + 1 \leq n \leq N - 1 \longrightarrow$ frecuencias negativas $-1/2 < f < 0$

$n = N/2 \longrightarrow f = 1/2$ y $f = -1/2$

La computación de la DFT (ver ecuación(4.19)) es un proceso $O(N^2)$, siendo N el tamaño de la secuencia. Para reducir este coste computacional surge el algoritmo FFT. FFT son las siglas de *Fast Fourier Transform* o Transformada Rápida de Fourier. Generalmente se conoce su existencia desde los trabajos de Cooley y Tukey (1965). No fue una idea original, ya que el paso crítico de factorización había sido descrito ya mucho antes por Karl Friedrich Gauss. Este algoritmo nos permite reducir la computación a $O(N \log_2 N)$ operaciones. La diferencia entre N^2 y $N \log_2 N$ es inmensa para N grande.

El algoritmo se basa en el uso recursivo del lema de Danielson-Lanczos. En 1942, Danielson y Lanczos probaron que una DFT de longitud N puede ser reescrita como la suma de dos DFT, cada una de longitud $N/2$. De este modo, partiendo de un N potencia entera de 2, llegaremos recursivamente hasta transformadas de longitud 1. El algoritmo explota las propiedades de la exponencial compleja.

Observemos ahora que las secuencias con las que trabajamos tienen valores reales, no complejos. Para calcular su transformada podríamos hacer uso del algoritmo FFT considerando todas las partes imaginarias nulas. Pero resultaría ineficiente, ya que haríamos operaciones innecesarias. Lo que se hace en este caso es guardar los datos adecuadamente en un array complejo de longitud la mitad ($N/2$), del que calcularemos su transformada mediante la FFT. Luego, haciendo uso de las propiedades de simetría de la transformada de Fourier de datos reales, obtendremos el resultado buscado. Esto se hace por medio de la rutina REALFT [Press et al., 1992].

Una vez que sabemos cómo calcular eficientemente y con el menor coste computacional la transformada de Fourier discreta de una secuencia de números reales, nuestro interés será el comportamiento del espectro de potencias $S(f)$ (definido en 4.1.3). Sabemos que tendremos correlaciones de largo alcance si:

$$S(f) \propto f^{-\beta}$$

4.2.2. Descripción del Método de Filtrado de Fourier

La DFT, como ya mencionamos, forma parte esencial del Método de Filtrado de Fourier (FFM), el cual en esencia consiste en lo siguiente: partiendo de una secuencia de números aleatorios sin correlaciones introduciremos correlaciones entre las variables mediante un filtro en ley de potencias. Ilustraremos ahora cómo actúa el método paso por paso.

Supongamos que queremos generar una secuencia $\{X_i\}$ de tamaño N (potencia entera de 2) con correlaciones de largo alcance caracterizadas mediante el exponente de correlación β en la ecuación (4.12). El FFM consiste en lo siguiente:

i) Generamos una secuencia $\{\eta_i\}$, $i = 1, \ldots, N$ (estacionaria) de números aleatorios no correlacionados con distribución gaussiana de media cero y varianza unidad ($N(0, 1)$). En la figura 4.4 lo mostramos para $N = 2^{12}$

ii) Pasamos al dominio de frecuencias mediante la transformada de Fourier discreta computada usando REALFT (véase 4.2.1), y calculamos el espectro de potencias $S(f)$ (sin constante de normalización ya que lo que nos importa es la pendiente). Como partíamos de una secuencia sin correlaciones, sabemos que su espectro de potencias será plano (ruido blanco) como podemos observar en la figura 4.5. Ahí mostramos el espectro de potencias de la secuencia de la figura 4.4.

iii) Para tener correlaciones de largo alcance modificamos el espectro multiplicando por una ley de potencias, $f^{-\beta}$ (filtro), donde β es el exponente de correlación deseado. Por tanto el espectro se aproximará en doble escala logarítmica a una recta de pendiente $-\beta$ (aunque conservará una componente estocástica). En la figura 4.6 se muestra en doble escala logarítmica el espectro de nuestra secuencia de tamaño $N = 2^{12}$ modificado para $\beta = 1.6$.

iv) Calcularemos ahora la transformada de Fourier inversa para obtener una secuencia de números aleatorios que, por construcción, va a tener exponente de correlación β. Observemos que queremos obtener una secuencia de números reales. La rutina REALFT nos permite también calcular la transformada de Fourier inversa de un array de datos complejos de modo que obtengamos datos reales. De este modo llegaremos, tal como era nuestro objetivo, a una secuencia $\{X_i\}$ de tamaño N (potencia entera de 2) con correlaciones de largo alcance caracterizadas mediante el exponente de correlación β en la ecuación (4.12). En la figura 4.6 mostramos tal secuencia para $N = 2^{12}$ y $\beta = 1.6$.

De modo que, eligiendo el exponente β, controlamos las correlaciones de largo alcance que queremos introducir en la secuencia. Ya sabemos (véase 4.1.3) que $\beta < 0$ indica anticorrelaciones o correlaciones negativas; $\beta = 0$ indica ausencia de correlaciones y $\beta > 0$ indica correlaciones positivas que aumentan con β. También sabemos (véase ahora 4.1.4) que si $-1 < \beta < 1$ estamos generando fGn's y si $1 < \beta < 3$ fBm's.

Por último notar que existe una mejora del Método de Filtrado de Fourier propuesta por [Makse et al., 1996]. El objetivo de esta mejora es evitar ciertas limitaciones que presenta el FFM en el estudio de sistemas grandes.

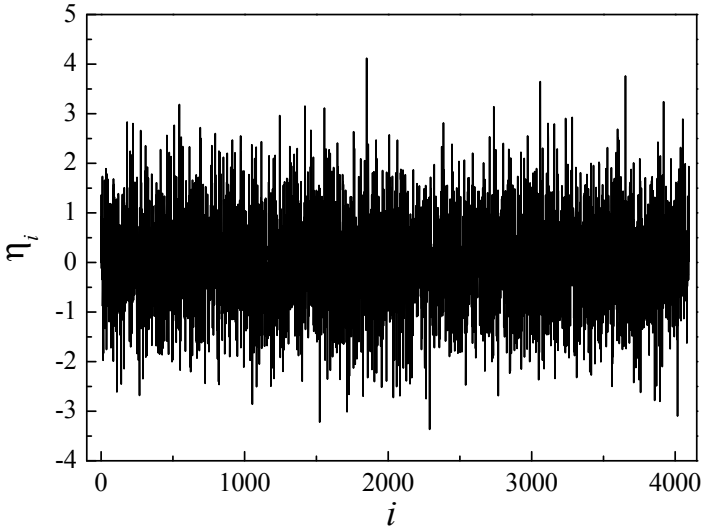

Figura 4.4: Secuencia de números aleatorios no correlacionados generada con tamaño $N = 2^{12}$

4.3. Cuantificación de correlaciones: DFA

Como ya comentamos anteriormente, el método que usaremos para cuantificar correlaciones es el *Detrended Fluctuation Analysis* (DFA). Este método fue introducido por [Peng et al., 1994] para el estudio de las correlaciones presentes en secuencias de ADN. No sólo se ha empleado en este ámbito, sino que es comúnmente usado para medir correlaciones en muchos campos de investigación. Podemos encontrar en la literatura la aplicación del *Detrended Fluctuation Analysis* en campos tan diversos como la climatología [Orun and Koçak, 2009], en economía [Yuan et al., 2009], en el comportamiento del sistema de control postural humano [Blázquez et al., 2009], en datos de seísmos [Kalimeri et al., 2008], en datos de tráfico [Shang et al., 2008], etc.

Como veremos en la descripción del método, el DFA proporciona un nuevo exponente

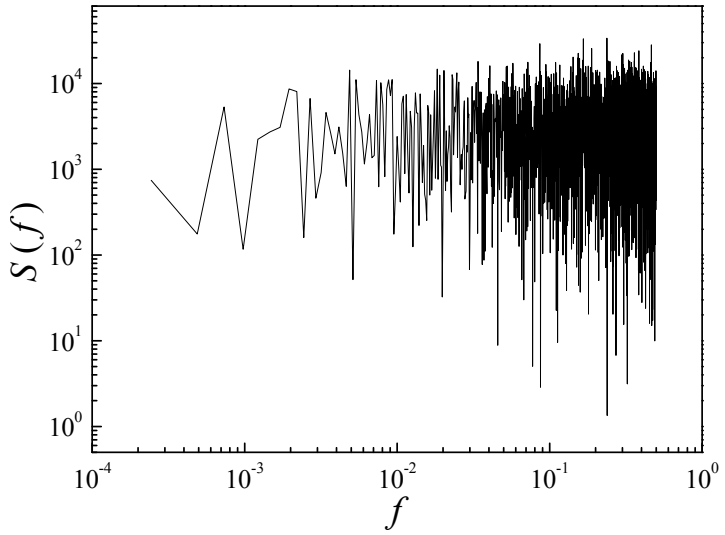

Figura 4.5: Espectro de potencias de la secuencia de la figura 4.4 en doble escala logarítmica

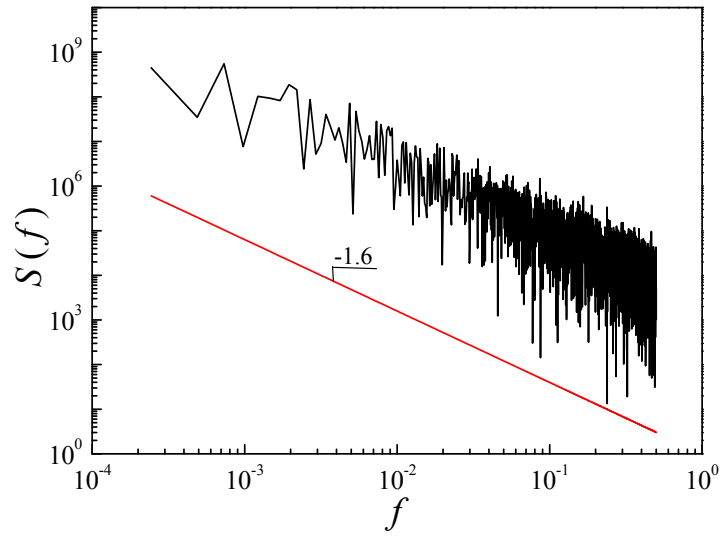

Figura 4.6: Espectro de potencias modificado para $\beta = 1.6$ y mostrado en doble escala logarítmica

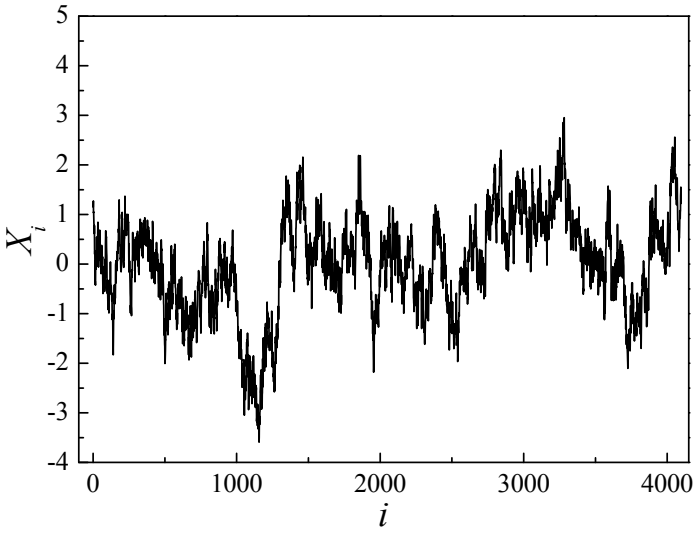

Figura 4.7: Secuencia $\{X_i\}$ con correlaciones de largo alcance generada con exponente de correlación $\beta = 1.6$ y tamaño $N = 2^{12}$

de escala (o exponente de correlación) α para representar las propiedades de correlación de la secuencia. Empezaremos dando una descripción detallada que nos muestre cómo actúa el método paso por paso.

4.3.1. Descripción del método

DFA es un método de análisis de escala para estimar correlaciones de largo alcance en ley de potencias. Nos da como resultado un parámetro α que cuantifica las correlaciones presentes en la secuencia. Notemos que los exponentes α y β se relacionan de la siguiente manera [Buldyrev, 2006; Kantelhardt et al., 2001]:

$$\alpha = \frac{\beta + 1}{2} \tag{4.20}$$

Para ilustrarlo partiremos de una secuencia sintética que generamos mediante FFM (4.2) con $\beta = 0.4$ y tamaño $N = 2^{12}$ (ver figura 4.8), a la cual le aplicaremos el DFA. Por la ecuación (4.20) esperamos obtener $\alpha = 0.7$. El DFA consta de los siguientes pasos:

Sea $\{X_i\}(i = 1, \ldots, N)$ la secuencia de la cual queremos medir las correlaciones. Como primer paso la integraremos (restándole la media), obteniendo así una nueva secuencia que denotaremos $\{Y_i\}(i = 1, \ldots, N)$:

$$Y(i) = \sum_{k=1}^{i} \left(X_k - \langle X \rangle\right), \text{ donde } \langle X \rangle = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{4.21}$$

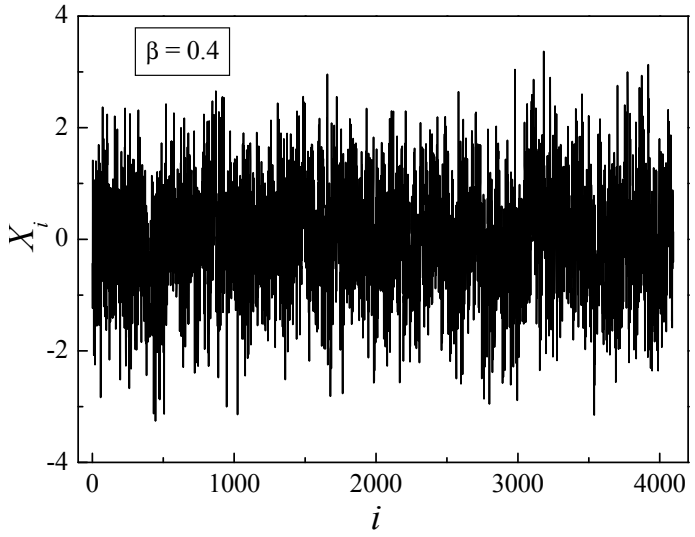

Figura 4.8: Secuencia $\{X_i\}$ con correlaciones de largo alcance generada con exponente de correlación $\beta = 0.4$ y tamaño $N = 2^{12}$

Ahora dividimos la secuencia integrada $Y(i)$ en cajas de longitud l, que podrán ser solapantes o no solapantes. En el caso de cajas no solapantes, tendríamos $n = N/l$ cajas. En cada caja, calculamos el ajuste lineal por mínimos cuadrados, $Y_{fit}(i)$, que llamamos "tendencia local".

Nótese aquí que existen variantes del DFA según el orden m del ajuste polinómico que se realice en cada caja y se les denomina DFA-m. Las diferentes formulaciones del DFA están enfocadas a eliminar tendencias en la secuencia estudiada: DFA de orden m elimina tendencias (polinómicas) de orden $m-1$ en la secuencia.

Nótese también que en nuestras simulaciones numéricas habrá que tener en cuenta que como la longitud l de caja es un número entero (ya que el índice i es entero, puesto que nuestra secuencia es discreta), será habitual encontrarnos con casos en los que queden al final de la secuencia un número de puntos inferior a l, insuficientes para añadir una última caja. Un caso sería el de cajas no solapantes y tamaño de secuencia no divisible por la longitud de caja considerada (es decir, N/l no es entero, ver figura 4.9). Existen distintas formas de solucionarlo, escogemos añadir una última caja que cubra los últimos l puntos, sin importarnos el solapamiento que se produzca con la anterior (ver figura 4.10).

Una vez realizados los ajustes lineales, restamos la tendencia local en cada caja. Se obtiene así la "fluctuación sin tendencia" ("*detrended fluctuation*"):

$$f_j(i) = Y(i) - Y_{fit}(i), \text{ para la } j\text{-ésima caja} \tag{4.22}$$

Calculamos la media de los cuadrados de la fluctuación en cada caja y promediamos

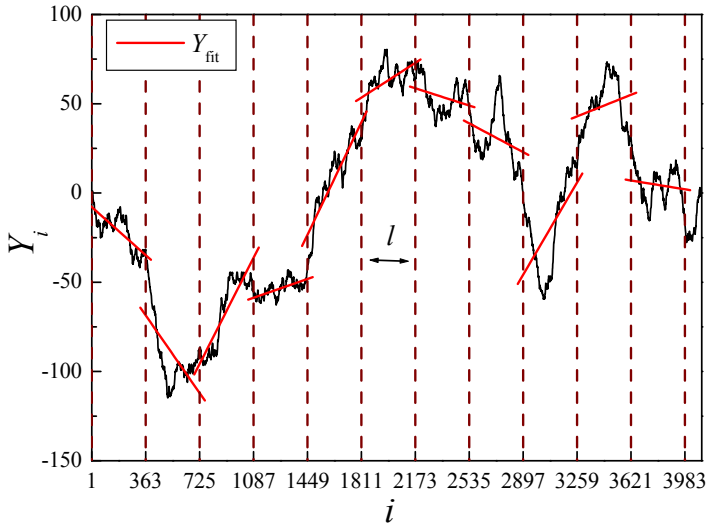

Figura 4.9: Secuencia de la figura 4.8 integrada y dividida en cajas no solapantes de longitud $l = 362$. En cada caja se realiza un ajuste lineal que denominamos Y_{fit}. Observemos que se quedan sin cubrir los puntos de la secuencia desde $i = 3983$ hasta $i = 4096$

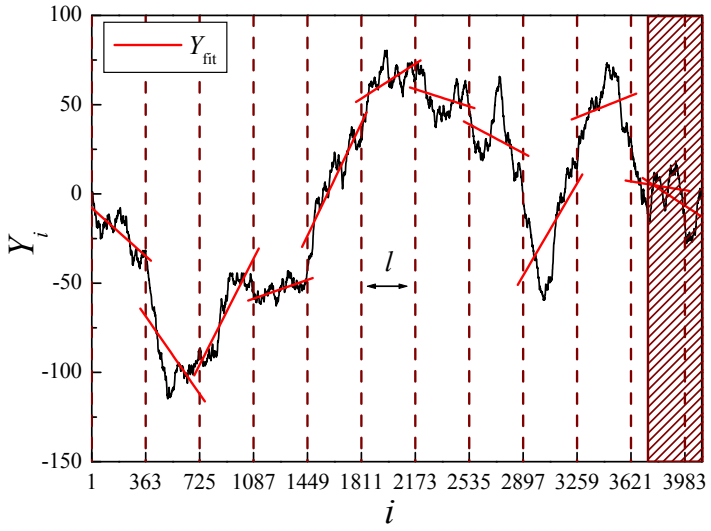

Figura 4.10: Ídem figura 4.9, añadiendo una última caja de longitud $l = 362$

(suponiendo que tenemos n cajas de longitud l). Finalmente, hacemos la raíz cuadrada:

$$F(l) = \left(\frac{1}{n} \sum_{j=1}^{n} \left[\frac{1}{l} \sum_{i \in j-box} [f_j(i)]^2 \right] \right)^{1/2} \tag{4.23}$$

Notemos que si se consideran cajas no solapantes y $n = N/l$, y entonces $F(l)$ es la raíz media cuadrática de las fluctuaciones:

$$F(l) = \left(\frac{1}{N} \sum_{i=1}^{N} [Y(i) - Y_{fit}(i)]^2 \right)^{1/2} \tag{4.24}$$

La computación se repite sobre diferentes valores de l para obtener una relación entre l y $F(l)$. Una relación en ley de potencias,

$$F(l) \propto l^{\alpha} \tag{4.25}$$

indica la presencia de *scaling*. El exponente α vendrá dado entonces por la pendiente de un ajuste lineal por mínimos cuadrados de $\log F(l)$ frente a $\log l$ y representa las propiedades de correlación de la secuencia. Se le llama exponente de escala o exponente de correlación: si $\alpha = 0.5$ no hay correlación (ruido blanco), si $\alpha < 0.5$ o $\alpha > 0.5$ hay correlaciones negativas o positivas que aumentan con α, respectivamente ($\alpha = 1.5$ corresponde al movimiento browniano).

En las figuras 4.9 y 4.10 mostramos la división en cajas de longitud l no solapantes (para que visualmente se observara mejor). Por tanto, con excepción de la primera caja y en algunos casos también de la última (si la añadimos como en figura 4.10), cada caja tendrá como inicio el punto obtenido al sumarle l al inicio de la caja anterior. Sin embargo, con el objetivo de disponer de más estadística (promediando mayor número de cajas) en nuestras simulaciones numéricas consideraremos cajas solapantes. Se pueden obtener cajas solapantes de muy diversas formas, nosotros estableceremos que cada caja tenga como inicio el punto obtenido al sumarle $\sqrt{l}$ al inicio de la caja anterior. El máximo solapamiento se obtendría si los puntos de inicio de cajas sucesivas estuviesen separados sólo por un punto.

Por otro lado es importante el rango en el que dejamos variar l. Tomaremos longitudes que variarán entre 8 y $N/10$ (siendo N el tamaño de la secuencia). Este es el rango de longitudes en el cual DFA proporciona resultados precisos [Hu et al., 2001; Carpena et al., 2022]. Parece conveniente tomar dichas longitudes equiespaciadas logarítmicamente ($l_1 = 8, l_2 = 8a, l_3 = 8a^2, ...$) ya que el parámetro α se obtendrá ajustando linealmente $\log(F(l))$ frente a $\log l$. Concretamente nosotros lo haremos para $a = \sqrt{2}$, es decir, tomaremos las

longitudes potencias de raíz de 2 (truncadas para obtener enteros) desde 8 a la última menor o igual a $N/10$.

Observemos en la figura 4.11 la gráfica de $F(l)$ frente a l en doble escala logarítmica para la secuencia de la figura 4.8, habiendo considerado cajas con el solapamiento mencionado anteriormente y l variando en el rango que hemos descrito. Ciertamente tenemos un comportamiento en ley de potencias ya que en doble escala logarítmica observamos una recta. Ajustando la pendiente obtendremos α. Tal como esperábamos $\alpha \approx 0.7$, concretamente $\alpha = 0.708 \pm 0.003$.

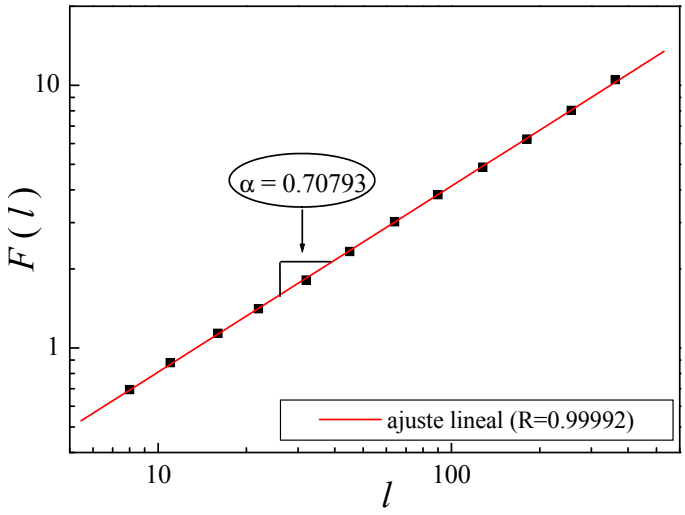

Figura 4.11: $F(l)$ frente a l en doble escala logarítmica para la secuencia de la figura 4.8 y el correspondiente ajuste para obtener α

4.3.2. Ventajas respecto a otros métodos

Una vez que sabemos cómo funciona el DFA, discutiremos el porqué de su elección para el estudio de las correlaciones que realizaremos en capítulos posteriores. El DFA, tal como hemos visto, nos proporciona un exponente de correlación α que caracterizará las correlaciones presentes en dichas secuencias: $\alpha = 0.5$ indica ausencia de correlaciones, valores de $\alpha < 0.5$ muestran anticorrelaciones y valores de $\alpha > 0.5$ corresponden a correlaciones positivas que aumentan con α.

Sin embargo, previamente habíamos descrito otros exponentes. Recordemos que decíamos que una secuencia $\{X_i\}$ tiene correlaciones de largo alcance cuando:

- La función de autocorrelación se comporta como una ley de potencias en la forma $C(k) \propto 1/k^{\delta}$

- El espectro de potencias sigue una ecuación en la forma $S(f) \propto 1/f^{\beta}$

Decíamos también que los exponentes δ y β se relacionaban siguiendo la ecuación (4.13). Veíamos que en el rango $(0, 1)$ se tenía $\delta = 1 - \beta$. Sin embargo, β puede tomar valores fuera de ese rango ($\beta < 0$ indicaba anticorrelaciones o correlaciones negativas; $\beta = 0$ ausencia de correlaciones y $\beta > 0$ correlaciones positivas) y ese tipo de comportamiento no era detectado por la función de autocorrelación. De hecho, para secuencias con $\beta \geq 1$, δ satura en 0 y no las distingue unas de otras. La función de autocorrelación sólo se puede aplicar a secuencias estacionarias (fGn's). Además se sabe que no es un estimador suficientemente preciso de las correlaciones ya que da resultados muy ruidosos y δ no es un buen exponente para caracterizar las correlaciones a menos que las secuencias sean muy largas y con correlaciones de largo alcance extremas [Coronado and Carpena, 2005].

Descartando el uso de la función de autocorrelación como herramienta para medir las correlaciones, podíamos pensar en analizar el espectro de potencias de las secuencias que obtengamos. Sabemos además que el exponente β nos permite distinguir entre fGn ($-1 < \beta < 1$) y fBm ($1 < \beta < 3$). Lo que ocurre es que el cálculo de FFT's es ruidoso y el ajuste que tendríamos que realizar para obtener β (ajuste lineal de $\log S(f)$ frente a $\log f$) es menos limpio que el que se hace para obtener α (comparar figuras 4.6 y 4.11).

El DFA, sin embargo, nos proporciona un exponente de correlación α que permite analizar secuencias estacionarias (fGn's) y no estacionarias (fBm's) y distinguir entre ellas; y a su vez es obtenido mediante un ajuste lineal poco ruidoso. Mediante la ecuación (4.20) que relaciona α y β, podemos concluir que fGn's corresponden al rango $0 < \alpha < 1$ y fBm's a $1 < \alpha < 2$ [Delignieres et al., 2006]. Observamos cómo recorre de manera continua fGn-fBm, conservando el sentido intuitivo derivada-integral: $\alpha_{fGn} = \alpha_{fBm} - 1$. En la figura 4.12 mostramos de nuevo la figura 4.3 pero completada con los valores de α. Notemos también que α se relaciona con el exponente de autosimilitud H en la forma $\alpha = H$ para los fGn's y $\alpha = H + 1$ para los fBm's.

Observamos que, como las medidas que indican correlaciones de largo alcance están relacionadas, el conocimiento de una estadística nos lleva a poder deducir la otra. El comportamiento en ley de potencias se conserva generalmente a lo largo de estas medidas y sabemos relacionar los exponentes. Sin embargo, el cálculo de cada una de las medidas está sujeta a diferentes limitaciones matemáticas, que hemos visto que hacen más propio el uso de unos métodos que de otros.

Sin embargo, en los últimos años se han puesto en evidencia algunas limitaciones del DFA [Carpena et al., 2017, 2022], intrínsecas al propio método, que hay que tener en cuenta para no obtener conclusiones erróneas al aplicarlo.

Por un lado, para escalas pequeñas (del orden de $l \in [3, 12]$), la función $F(l)$ no se

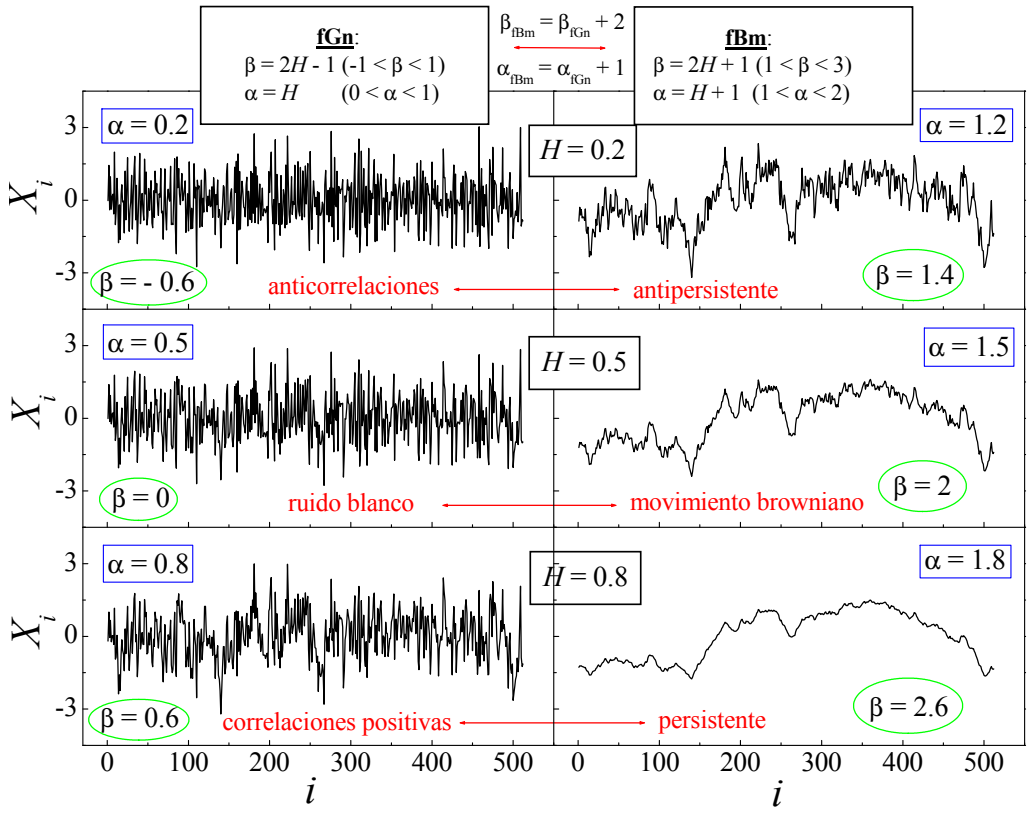

Figura 4.12: Completamos la figura 4.3 añadiendo los valores de α en cada caso.

comporta como una ley de potencias [Carpena et al., 2022]. Ello conlleva una sobreestimación del exponente de correlación a esas escalas, que no es un efecto de tamaño finito, sino una limitación intrínseca del DFA. Nótese que en el capítulo 6, en el que haremos ajustes del exponente de escala del DFA distinguiendo dos regiones para el ajuste (escalas pequeñas y grandes) estaremos ya fuera de la región en la que se observan esos problemas.

Por otro lado, en la sección 4.4, en la que nos centraremos en el uso del DFA en secuencias binarias, comentaremos que el DFA puede llevar a interpretaciones erróneas cuando se aplica a la secuencia binaria obtenida al considerar el signo del proceso continuo subyacente [Carpena et al., 2017].

4.3.3. Efectos de tamaño finito

Por último, es importante que el método que utilicemos se vea afectado lo menos posible por el hecho de trabajar con secuencias finitas. Por ello presentaremos en esta sección un estudio sistemático de los efectos de tamaño en el DFA. Recalcularemos el exponente de correlación α de secuencias sintéticas generadas con correlaciones controladas para comprobar la consistencia del método, y testear si recuperamos el exponente impuesto en

la generación.

Dicho estudio fue realizado en [Coronado and Carpena, 2005] para secuencias generadas con exponentes de correlación en el rango $0 < \beta < 2$, concluyendo que los resultados obtenidos aplicando el DFA son prácticamente los mismos una vez que tenemos un tamaño $N \geq 2^{10}$ y que en comparación con otros métodos es el que menos se ve afectado por efectos de tamaño. Aquí ampliaremos el rango de ese resultado, analizando los efectos de tamaño del DFA en secuencias sintéticas generadas con $-1 < \beta < 3$ y para tamaños $N = 2^{10}, 2^{12}, 2^{14}, 2^{16}, 2^{18}, 2^{20}, 2^{22}$. Fijando un β de ese rango en nuestra generación, variamos el tamaño N de la secuencia y vemos cómo afecta eso al resultado mostrado por el DFA.

Tal como se hace en [Coronado and Carpena, 2005], generaremos para cada β y N, $2^{25}/N$ secuencias mediante el FFM (excepto en el caso $N = 2^{22}$ que serán $2^{27}/N$ para tener suficiente estadística) y obtendremos un α promedio y su desviación. Si el DFA produce resultados correctos, habría que obtener $\alpha = (\beta + 1)/2$. De este modo, sabremos en que rangos de correlación y longitudes de secuencia el DFA proporciona buenos resultados.

Debido a que conocemos la relación entre α y β, fijar un β en la generación equivale a fijar un α. Hablaremos a partir de ahora en términos de α. Por tanto, generamos en el rango $0 < \alpha < 2$ y para los tamaños mencionados. A cada una de las secuencias le aplicaremos el DFA, el cual nos proporciona el correspondiente exponente de correlación. Calcularemos el promedio y lo denotaremos α_{DFA} (para diferenciar el impuesto en la generación del calculado mediante el DFA) y su desviación (que mostraremos mediante flechas verticales en las figuras).

Presentamos los resultados en la figura 4.13. En ella mostramos la relación entre el exponente alfa promedio que proporciona el DFA, α_{DFA}, en función del que imponemos en la generación, α. Obviamente tendríamos que obtener todos los puntos sobre la recta $y = x$. Observamos que prácticamente para todo el rango de tamaños los puntos solapan sobre dicha recta. Las flechas verticales muestran la desviación. Sólo para correlaciones cercanas a $\alpha = 2$ (a partir de $\alpha = 1.8$ aproximadamente) y para correlaciones cercanas a $\alpha = 0$ (a partir de $\alpha = 0.4$ aproximadamente) observamos separaciones de la recta que se van corrigiendo a medida que aumenta N. Para α cercano a 0, sobreestimamos ligeramente las correlaciones y para α cercano a 2, las subestimamos.

Se sabe que es difícil estimar anticorrelaciones (α por debajo de 0.5) [Hu et al., 2001] por lo que usualmente el proceso que se sigue es integrar la secuencia (acumularla) y a la secuencia acumulada calcularle el DFA. Si es que tenemos anticorrelaciones, la secuencia acumulada será antipersistente (ver figura 4.12). Al exponente que obtengamos, habría que restarle 1, para obtener el de la secuencia de partida. Implementamos este proceso para el caso de las anticorrelaciones para ver si así mejoramos los resultados anteriores y

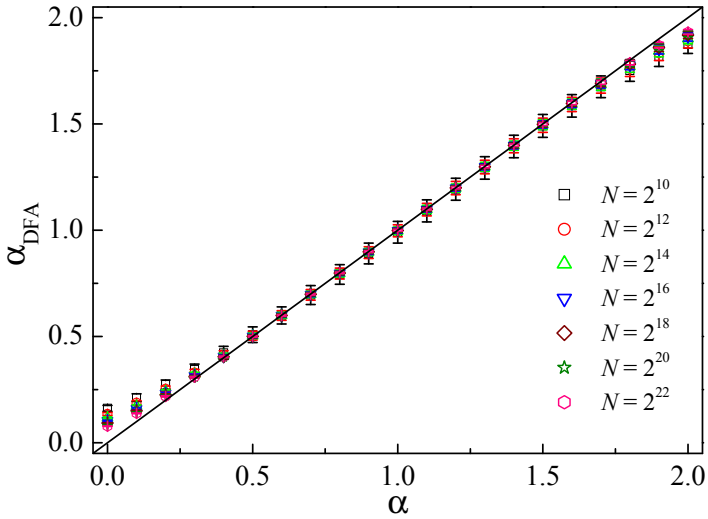

Figura 4.13: Relación entre el exponente alfa promedio que proporciona el DFA, α_{DFA}, en función del que imponemos en la generación, α. Observamos el resultado obtenido variando el tamaño de las secuencias. La linea muestra la recta $y = x$ y las flechas verticales las desviaciones respecto a los valores medios calculados.

se alejan menos de la recta.

En la figura 4.14 comparamos, en el caso de las anticorrelaciones y para tamaño $N = 2^{22}$, el α_{DFA} calculado directamente con el calculado mediante la acumulada. Para obtener este último, para cada α impuesto en la generación de nuevo hemos realizado $2^{27}/N$ iteraciones, de las cuales hemos obtenido el promedio y la desviación. Observamos cómo claramente los resultados son mejores con el cálculo realizado mediante la acumulada y solapan completamente en la recta $y = x$. De este modo se solucionan los problemas de sobreestimación en ese rango.

Con respecto a las correlaciones cercanas a $\alpha = 2$ la subestimación se debe a que estamos ya con correlaciones de largo alcance grandes y necesitaríamos mayor tamaño de la secuencia para conseguir el solapamiento con la recta $y = x$. Es un efecto del tamaño finito de las secuencias.

Por último, haremos referencia al hecho de que el método implementado es el DFA de orden 1. Hemos visto cómo en el rango $0 < \alpha < 2$ proporciona resultados precisos. De hecho, hemos recalculado el α de secuencias sintéticas de las que ya conocíamos su exponente de correlación (impuesto en la generación), para comprobar la consistencia del método. Sin embargo, es importante destacar, como mostraremos en la figura 4.15, que si permitimos correlaciones superiores a $\alpha = 2$ el DFA-1 satura. De hecho, para una tendencia lineal o para una tendencia cuadrática se puede realizar el cálculo analítico y se obtiene $\alpha = 2$. Por muchas correlaciones que haya en la secuencia, el DFA-1 no es capaz

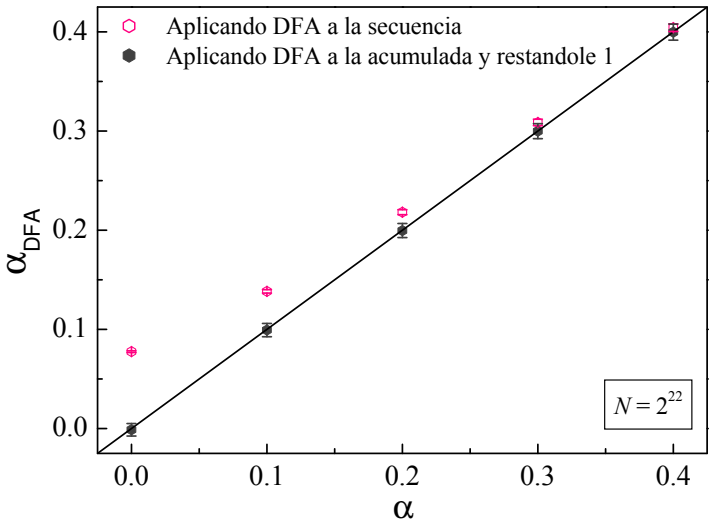

Figura 4.14: Comparación, en el caso de anticorrelaciones ($\alpha < 0.5$) y para tamaño $N = 2^{22}$, del α_{DFA} calculado directamente con el calculado restándole 1 al obtenido aplicando el DFA a la secuencia acumulada. De nuevo la línea muestra la recta $y = x$ y las flechas verticales las desviaciones respecto a los valores medios calculados.

de medir más de $\alpha = 2$.

La figura 4.15 muestra α_{DFA} para secuencias de tamaño $N = 2^{22}$ en el rango $0 < \alpha < 3$: claramente se observa que a partir de $\alpha = 2$, α_{DFA} deja de aumentar y satura en 2 (las desviaciones respecto a los valores medios son pequeñas y las flechas verticales quedan ocultas por el propio símbolo). Aunque sabemos, porque está impuesto en la generación, que tienen un exponente de correlación mayor, el DFA-1 no es capaz de medir más de 2. Notemos que para los valores de $\alpha < 0.5$ en la figura 4.15 el correspondiente α_{DFA} ha sido obtenido aplicando el DFA a la acumulada y restando 1, tal y como explicamos anteriormente. Notemos también que la saturación de α_{DFA} en 2 se soluciona usando un DFA de orden superior. No es necesario en nuestro caso ya que nos centraremos en el rango de correlación $0 < \alpha < 2$, que es el rango que presentan las correlaciones de largo alcance presentes en la mayoría de los sistemas naturales estudiados en la literatura.

4.4. Correlaciones de largo alcance en secuencias binarias

En las secciones anteriores hemos descrito el método que usaremos para generar secuencias reales con correlaciones de largo alcance (4.2: FFM), así como el que emplearemos para cuantificarlas (4.3: DFA).

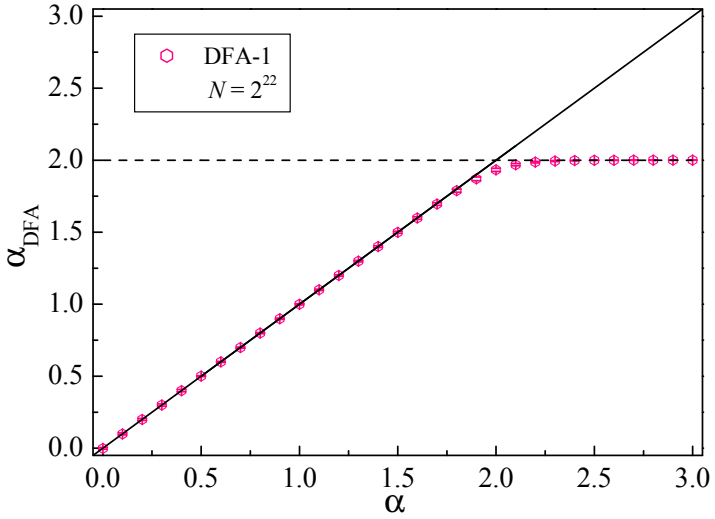

Figura 4.15: α_{DFA} para secuencias de tamaño $N = 2^{22}$ en el rango $0 < \alpha < 3$: DFA-1 satura en $\alpha_{DFA} = 2$

Para finalizar este capítulo nos centraremos en el caso particular de las secuencias binarias, cómo generar secuencias binarias con correlaciones, y qué resultados se obtienen al cuantificar las correlaciones de la secuencia binaria mediante el DFA. Estos resultados tendrán aplicaciones en capítulos posteriores, en especial cuando analicemos las apariciones de una palabra en un texto como una secuencia binaria.

4.4.1. Generación

Una técnica estándar para generar secuencias binarias con correlaciones consiste en lo siguiente:

Dado un valor de α y un tamaño N (potencia entera de 2 si queremos hacer uso del algoritmo FFT) generaremos, tal y como hemos visto en secciones previas, una secuencia real X_i con correlaciones de largo alcance caracterizadas por dicho exponente: consideramos $\beta = 2\alpha - 1$ y aplicamos el Método de Filtrado de Fourier (4.2.2). Supongamos que X_i es una secuencia de media 0 generada mediante tal procedimiento. A partir de dicha secuencia existen distintas posibilidades de obtener una secuencia binaria. Podemos simplemente considerar el signo del proceso continuo subyacente o, lo que es lo mismo, considerar como umbral el paso por 0 y mapear a 1 o -1 según los valores de la secuencia fuesen mayores o menores que 0, respectivamente. Para $i = 1, ..., N$

$$X_{bin}(i) = \text{sign}(X(i)) = \begin{cases} 1 & \text{si } X(i) > 0 \\ -1 & \text{en otro caso} \end{cases} \tag{4.26}$$

De esta forma obtendremos una secuencia binaria X_{bin} con probabilidad $p = 1/2$ de aparición de 1'os. La información de la secuencia real que pasa a la binaria es la contenida en el cambio de signo. Este tipo de secuencias binarias ocurre en sistemas de diversa naturaleza. El tamaño de los segmentos de signo constante o, lo que es lo mismo, las distancias entre dos cruces consecutivos de la secuencia X_i por 0, constituyen lo que se denomina tiempos de retorno o, en algunos contextos, tiempos de primer paso (FPT, del inglés *first passage time*). Estudiaremos sus propiedades estadísticas (relevantes para describir la dinámica del sistema complejo subyacente) en función del exponente de correlación de la secuencia real X_i de partida, en el capítulo 5.

Este proceso puede ser generalizado al paso por cualquier umbral. O, dicho de otra forma, dada cualquier probabilidad $p \in (0, 1)$ podemos obtener, a partir de nuestra secuencia real, una secuencia binaria con dicha probabilidad de aparición de 1'os, buscando el umbral apropiado para el mapeo. Usaremos este procedimiento para generar las secuencias binarias que modelarán las ocurrencias de las palabras a lo largo de un texto en el capítulo 6.

4.4.2. Cuantificación

Una vez generadas las secuencias binarias por el procedimiento descrito, es inmediato plantearse si conservarán el exponente de correlación de la secuencia real de partida.

Para cada valor del exponente de correlación α (que consideraremos en el rango $0 < \alpha < 3$), generaremos la secuencia real con dicho exponente de correlación, a partir de ella obtendremos la secuencia binaria X_{bin} correspondiente considerando como umbral el paso por 0, aplicaremos el DFA y observaremos qué resultados se obtienen. Analizaremos también cómo va a depender el resultado obtenido del tamaño de secuencia considerado.

Al igual que hicimos en la sección anterior trabajaremos con tamaños de secuencia $N = 2^{10}, 2^{12}, 2^{14}, 2^{16}, 2^{18}, 2^{20}, 2^{22}$. Para cada α de entrada y cada tamaño de secuencia N generamos $2^{25}/N$ secuencias binarias (excepto $2^{27}/N$ en el caso $N = 2^{22}$). Aplicamos DFA a cada una de esas secuencias binarias y al promedio obtenido de todas ellas lo denominaremos α_{bin}. Calcularemos también la desviación respecto de dicha media. Hay, excepto para $\alpha = 0.5$ de partida, cierta dependencia del tamaño N de la secuencia que desaparece prácticamente una vez que llegamos a $N = 2^{18}$ aproximadamente. Por brevedad, en la figura 4.16, mostraremos el resultado obtenido para secuencias de tamaño $N = 2^{22}$, donde ya sabemos que los efectos de tamaño finito son muy pequeños.

En principio distinguimos 4 regímenes diferentes en el exponente de correlación de las secuencias binarias α_{bin}, dependiendo del exponente α impuesto de entrada en la generación de la secuencia real:

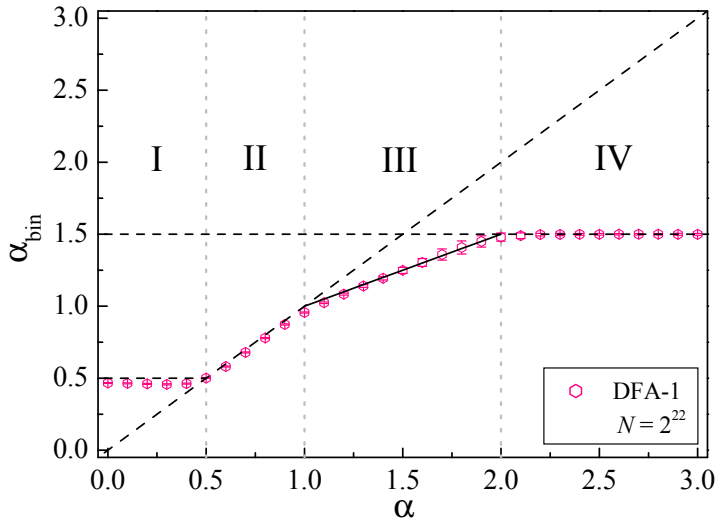

Figura 4.16: α_{bin} para secuencias de tamaño $N = 2^{22}$ en el rango $0 < \alpha < 3$

I. Si $\boldsymbol{\alpha < 0.5}$: El exponente de correlación de la secuencia binaria satura en $\alpha_{bin} = 0.5$. Esto sorprendentemente nos diría que para una secuencia real con anticorrelaciones el cambio de signo es aleatorio. Interpretaríamos que no encontramos en la secuencia binaria, como podríamos esperar, más probabilidad de cambiar de signo, que de no cambiar. Sin embargo, como comentaremos más adelante, se trata de un resultado erróneo del DFA.

II. Si $\boldsymbol{0.5 < \alpha < 1}$: El exponente de correlación α_{bin} de la secuencia binaria coincide con el de la secuencia real de partida α. Esto nos diría que toda la información de la secuencia real de partida queda reflejada en su cambio de signo.

III. Si $\boldsymbol{1 < \alpha < 2}$: El exponente de correlación α_{bin} de la secuencia binaria es sistemáticamente menor que el de la secuencia real α de partida, aunque aumenta linealmente con dicho exponente: $\alpha_{bin} = (\alpha + 1)/2$. La fuerza de las correlaciones positivas de la secuencia real se traduce en la existencia de regiones grandes en las que la secuencia no cambia de signo, lo que implica pérdida de información al pasar a binaria.

IV. Si $\boldsymbol{\alpha > 2}$: Observamos que, al igual que ocurría con el exponente recalculado de la secuencia real, el exponente de correlación obtenido al aplicarle a las secuencias binarias generadas el DFA, también satura cuando el α impuesto en la generación de las secuencia real es mayor que 2. Pero con las secuencias binarias hay una diferencia importante: la saturación se produce en $\alpha_{bin} = 1.5$. Sabemos que DFA-1 puede medir correlaciones por encima de 1.5 (porque lo hace con las reales), así que dicha saturación no se debe a las limitaciones del DFA de orden 1, si no que se tiene

que deber al hecho de que la secuencia es binaria. De hecho, en este régimen, la secuencia binaria va a consistir en pocos segmentos de gran longitud. Este hecho es cierto incluso en el límite de secuencias grandes, donde el número de cambios de signo es siempre finito (como veremos en el próximo capítulo). Se puede comprobar cómo en ese caso el DFA siempre proporciona un exponente igual a 1.5.

Sin embargo, en [Carpena et al., 2017] mostramos evidencia analítica y numérica de que cuando la secuencia binaria que contiene la información del signo de la secuencia real de partida presenta anticorrelaciones en ley de potencias, el DFA interpreta erróneamente ausencia de correlaciones. Cuando X_i es estacionaria, se conoce una relación analítica entre la función de autocorrelación $C(k)$ y la función de fluctuación del DFA $F(\ell)$ [Höll and Kantz, 2015; Talkner and Weber, 2000]. A partir de esta relación, y usando que también se conoce la relación entre la función de autocorrelación de la secuencia real y la del signo [Apostolov et al., 2008], llegamos a

$$F_{s,DFA}(\ell) = \sqrt{\left(1 - \frac{2}{\pi}\right)\frac{\ell^2 - 4}{15\ell} + \frac{2}{\pi}F^2_{DFA}(\ell)}. \tag{4.27}$$

Esta ecuación relaciona la función de fluctuación del DFA de la secuencia original, $F_{DFA}(\ell)$, con la del signo $F_{s,DFA}(\ell)$. Se observa que, si $\alpha < 0.5$, para ℓ grande, domina el primer término de la raíz (y es proporcional a ℓ) implicando que $F_{s,DFA}(\ell) \sim \ell^{1/2}$ de donde $\alpha_{bin} = 0.5$. Esto explica el resultado numérico mostrado anteriormente en el caso $\alpha < 0.5$, en el que obteníamos $\alpha_{bin} = 0.5$. Para $0.5 < \alpha < 1$, es el segundo término de la ecuación 4.27 el que domina para ℓ grande, de lo que se deduce $F_{s,DFA}(\ell) \sim \ell^{\alpha}$ y, por tanto, $\alpha_{bin} = \alpha$.

Analíticamente se obtienen los mismos resultados que numéricamente, pero son erróneos para $\alpha < 0.5$. De la relación entre la función de autocorrelación de la secuencia real $C(k)$ y la del signo $C_s(k)$ [Apostolov et al., 2008] se llega a

$$C_s(k) \simeq \frac{2}{\pi}C(k). \tag{4.28}$$

De aquí se deduce que si la función de autocorrelación de X_i tiene un comportamiento en ley de potencias, también lo tendrá la del signo. Es más, el exponente de la ley de potencias coincide. El resultado proporcionado por el DFA, provocado por el primer sumando en la ecuación 4.27 (y, por tanto, intrínseco al propio método), lleva a conclusiones erróneas.

Concluimos entonces que los regímenes I y II de la gráfica anterior son en realidad un único régimen en el que se conservan las correlaciones de la secuencia de partida. En el próximo capítulo veremos también cómo estos regímenes están relacionados con las

propiedades estadísticas de las distribuciones de las longitudes entre pasos consecutivos por 0 de X_i o, lo que es lo mismo, de los segmentos de signo constante de $X_{bin}(i)$.

Capítulo 5

Tiempos de primer paso en procesos con correlaciones de largo alcance

Como ya hemos comentado previamente, veremos en el capítulo 6 que la distribución espacial heterogénea de las palabras relevantes se manifiesta como correlaciones de largo alcance en ley de potencias. En dicho capítulo, propondremos también modelar las ocurrencias de la palabra en el texto a partir de los pasos por umbral de un proceso con correlaciones de largo alcance, del que la palabra modelada heredará las correlaciones.

Aquí vamos a analizar sistemáticamente las propiedades estadísticas de los pasos por umbral (pasos por cero, en este caso) de señales con correlaciones de largo alcance, y cómo dependen dichas propiedades de las correlaciones de la señal subyacente. Así, podremos sentar las bases que nos permitirán proponer el modelo del capítulo 6.

Una de las formas en la que se investiga la dinámica de sistemas complejos es mapeándolos en *random walks* generalizados unidimensionales. Las características fundamentales de dichos procesos vienen dadas por las propiedades estadísticas de los tiempos de primer paso [Condamin et al., 2007] (FPT [1]): la forma funcional de su distribución de probabilidad y su longitud promedio.

Se han observado diferentes formas funcionales para la distribución de probabilidad de los FPT en estudios empíricos:

i) Exponencial pura en procesos sin correlaciones [Bunde and Havlin, 1995].

ii) *Stretched exponential*[2] en varios sistemas complejos naturales y sociales: desde disparo neuronal [Schindler et al., 2004], fluctuaciones climáticas [Bunde et al., 2005]

[1] Nótese aquí que se usará el término FPT para referirnos a las longitudes ℓ entre dos pasos consecutivos por 0 del proceso (*zero-level crossings*), para seguir la terminología empleada en [Carretero-Campos et al., 2012].

[2] Función de distribución acumulada complementaria de la distribución de Weibull. Usaremos el término *stretched exponential* por ser el empleado en la literatura en este contexto.

o dinámicas del corazón [Reyes-Ramírez and Guzmán-Vargas, 2010], al tráfico de internet [Leland et al., 1994; Cai et al., 2009] y la actividad bursátil [Ivanov et al., 2004; Wang et al., 2009].

iii) Ley de potencias para ciertos procesos de intermitencia on-off relacionados con circuitos electrónicos no lineales [Ding and Yang, 1995] y difusión anómala [Shlesinger et al., 1993; Rangarajan and Ding, 2000; Khoury et al., 2011; Eliazar and Klafter, 2009].

La diferencia en la forma de la distribución se suele atribuir a las especificidades del sistema individual. Identificar factores comunes que puedan ser responsables de comportamientos similares de la distribución de los tiempos de primer paso en sistemas diferentes no ha sido objeto de investigaciones. De hecho, estos sistemas exhiben diferentes comportamientos invariantes de escala con correlaciones de largo alcance y no se conoce cómo el grado de las correlaciones del sistema se relaciona con las propiedades estadísticas del FPT.

En el desarrollo de este capítulo nuestra hipótesis es que las correlaciones son el factor común en procesos complejos de diferente naturaleza que presentan propiedades estadísticas similares para los tiempos de primer paso. Y, a la inversa, que sistemas con las mismas propiedades en sus tiempos de primer paso tienen un grado similar de correlaciones.

5.1. Metodología

Nuestro objetivo en este capítulo es investigar cómo el grado de correlaciones del proceso afecta a las propiedades fundamentales de los tiempos de primer paso: la forma funcional de su distribución de probabilidad y la longitud promedio.

En general, vamos a realizar un estudio numérico de las propiedades de los tiempos de primer paso tanto en función de las correlaciones de la señal como de su tamaño. Para ello, usamos el Método de Filtrado de Fourier [Makse et al., 1996] (descrito en 4.2) para generar señales fractales con media cero, desviación estándar uno y un grado de correlaciones de largo alcance en ley de potencias fijado a priori mediante el exponente α del DFA. Para cada combinación de N y α, los resultados que mostramos se obtienen mediante simulaciones de Monte Carlo con $2^{32}/N$ realizaciones.

Recordemos que el algoritmo genera una secuencia de números aleatorios no correlacionados (señal aleatoria) siguiendo una $N(0,1)$ en el espacio real, y la pasa al dominio de las frecuencias (f) mediante la transformada de Fourier obteniendo un ruido blanco. Multiplicando ahora por una ley de potencias de la forma $f^{-(2\alpha-1)/2}$, y volviendo al espacio

real mediante la transformada inversa de Fourier, obtenemos una secuencia con correlaciones cuantificadas por el exponente α que, por construcción, corresponde al exponente de escala del *Detrended Fluctuation Analysis* (DFA) [Peng et al., 1994] (ver sección 4.3.1). Como consecuencia, el espectro de potencias de la secuencia resultante será una ley de potencias de la forma $S(f) \sim 1/f^{\beta}$, con $\beta = 2\alpha - 1$.

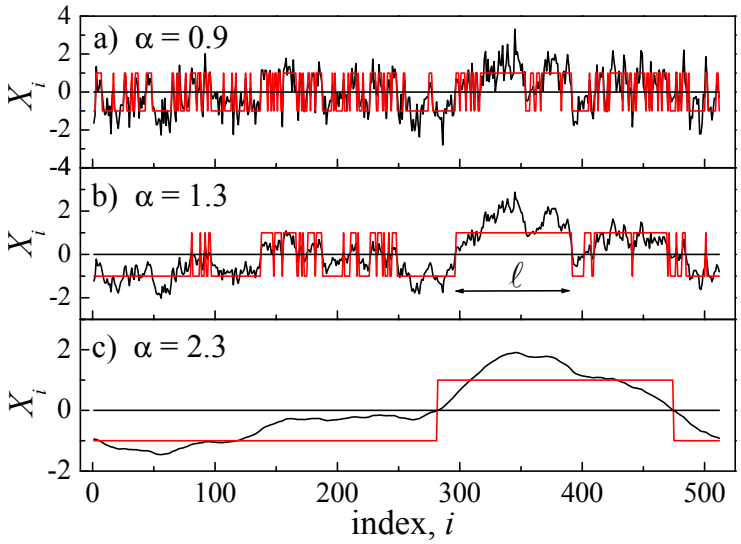

Figura 5.1: Ejemplos de tres procesos invariantes de escala (línea negra continua), cada uno de tamaño $N = 2^9$, y con diferentes grados de correlaciones cuantificadas por el exponente de escala α obtenido usando el método DFA [Peng et al., 1994]. Valores crecientes de α indican un grado más alto de correlaciones. El tiempo de primer paso (FPT), definido como el intervalo ℓ entre dos pasos por cero consecutivos del proceso, se representa mediante segmentos de signo constante $+1$ o -1 (línea roja). Obsérvese el cambio en el perfil del proceso al incrementar las correlaciones, lo que lleva a valores más grandes de ℓ y al correspondiente cambio en las estadísticas del FPT.

Dada la relación biunívoca entre ambos exponentes, podemos usar cualquiera de los dos como referencia. En las próximas secciones usaremos como referencia el exponente α del DFA, por ser el método estándar en el estudio de series temporales con correlaciones de largo alcance [Hu et al., 2001; Coronado and Carpena, 2005; Carpena et al., 2007; Blázquez et al., 2009; Xu et al., 2011; Ma et al., 2010], y también poder aplicarse a señales no estacionarias del mundo real.

Recordemos que, i) si $\alpha = 0.5$, la secuencia no tiene correlaciones (ruido blanco), ii) si $\alpha < 0.5$ las correlaciones son negativas (anticorrelaciones) y iii) $\alpha > 0.5$ indica correlaciones positivas. Procesos con $0 < \alpha < 1$ son *fractional Gaussian noises* (fGns) y procesos con $1 < \alpha < 2$ son *fractional Brownian motions* (fBms). En particular, $\alpha = 1.5$ corresponde al *classical random walk*.

Consideraremos procesos con α en el rango $0 < \alpha < 3$ y, para todos ellos, la longitud ℓ del FPT se define como la distancia entre dos cruces por cero consecutivos del proceso (véase figura 5.1). Aunque, estrictamente hablando, la terminología de FPT se reserva para fBms ($1 < \alpha < 2$), la usaremos en el rango completo de α ($0 < \alpha < 3$) por simplicidad.

5.2. Distribuciones de los tiempos de primer paso

Los resultados de las simulaciones descritas en la sección 5.1 muestran la existencia de tres regímenes diferentes (véase figura 5.2a) para la densidad de probabilidad $p(\ell)$[3] de las longitudes de los tiempos de primer paso ℓ en función del grado de correlaciones α del proceso (*stretched exponential*, cola en ley de potencias y saturación), separados por dos puntos de transición. Describimos estos tres regímenes a continuación.

5.2.1. Régimen en *stretched exponential*

Para $\alpha < 1$, obtenemos que la densidad de probabilidad $p(\ell)$ se comporta como una *stretched exponential*, es decir,

$$p(\ell) \sim \exp\left[-\left(\ell/\ell_0\right)^{\varepsilon}\right]. \tag{5.1}$$

Nótese que en la forma funcional de $p(\ell)$ tendríamos también un factor multiplicativo $(\ell/\ell_0)^{\varepsilon-1}$, con ℓ_0 el parámetro de escala. Sin embargo, en los ajustes realizados en las simulaciones numéricas no se llega a apreciar la ley de potencias ya que solo afecta a distancias ℓ pequeñas.

También observamos que el parámetro ε (*stretching parameter* o parámetro de forma) depende del exponente de correlación α del proceso como sigue:

- Para $\alpha = 0.5$, se tiene que $\varepsilon = 1$, que corresponde a una exponencial pura, coherente con el hecho de que es un ruido blanco.
- Para $\alpha < 0.5$, $\varepsilon > 1$, y crece a medida que α decrece. En este caso, $p(\ell)$ decae más rápido que la exponencial.
- Para $\alpha > 0.5$, $\varepsilon < 1$, y decrece con α. En este caso, $p(\ell)$ es una *stretched exponential* real (exponencial estirada) y la cola de $p(\ell)$ se vuelve más pesada a medida que crece α.

[3] Nótese que ℓ puede alcanzar valores muy grandes ($\ell \gg 1$), lo que justifica el considerarla una variable continua.

Estos resultados coinciden con la observación experimental para una gran variedad de fenómenos [Reyes-Ramírez and Guzmán-Vargas, 2010; Ivanov et al., 2004; Wang et al., 2009] y con trabajos previos en los que se simulan y estudian procesos en este rango de correlaciones [Bunde et al., 2005]. En cuanto a derivaciones analíticas, en [Newell and Rosenblatt, 1962] obtienen que la forma en *stretched exponential* es una cota superior para los *zero-level crossings* (o los FPTs, como los denominamos aquí) en fGns, es decir, en el rango $0 < \alpha < 1$.

5.2.2. Régimen de cola en ley de potencias

Para $1 < \alpha < 2$, el modelo (5.1) no es válido, y encontramos empíricamente que $p(\ell)$ se comporta de la forma

$$p(\ell) \sim \frac{f(\ell)}{\ell^{\delta}}, \tag{5.2}$$

donde la función $f(\ell)$ solo afecta a corta escala (valores pequeños de ℓ) y tiende a una constante a medida que crece ℓ.

La función $f(\ell)$ es la responsable de la curvatura de $p(\ell)$ que se observa a escalas muy pequeñas, curvatura que podemos apreciar en la figura 5.2b. Por otro lado, $f(\ell)$ evita la divergencia de (5.2) en el límite para ℓ pequeño. Sin embargo, la cola de la distribución se comporta como una ley de potencias de exponente δ.

Obtenemos numéricamente que el exponente δ y el exponente de correlación α se relacionan siguiendo la igualdad $\delta = 3 - \alpha$. Estos resultados concuerdan con conclusiones anteriores para la distribución de FPT en este régimen: argumentos de *scaling* presentados en [Ding and Yang, 1995] y una derivación heurística mostrada en [Rangarajan and Ding, 2000] basada en resultados sobre el valor máximo de un fBm [Molchan, 1999], conducen a un comportamiento de cola como el de (5.2).

Nótese que la forma funcional de $p(\ell)$ y la relación entre δ y α encontradas para el rango $1 < \alpha < 2$ generalizan el resultado conocido para la distribución de los FPT de un *random walk* [Bunde and Havlin, 1995] ($\alpha = 1.5$) donde

$$p(\ell) \sim \frac{e^{-a/\ell}}{\ell^{3/2}}. \tag{5.3}$$

Para $\alpha = 1$, correspondiente al ruido $1/f$, encontramos una transición entre los dos regímenes. En este caso, $p(\ell)$ presenta un comportamiento intermedio y decae más lento que (5.1) pero más rápido que (5.2), como se muestra en la figura 5.2(b).

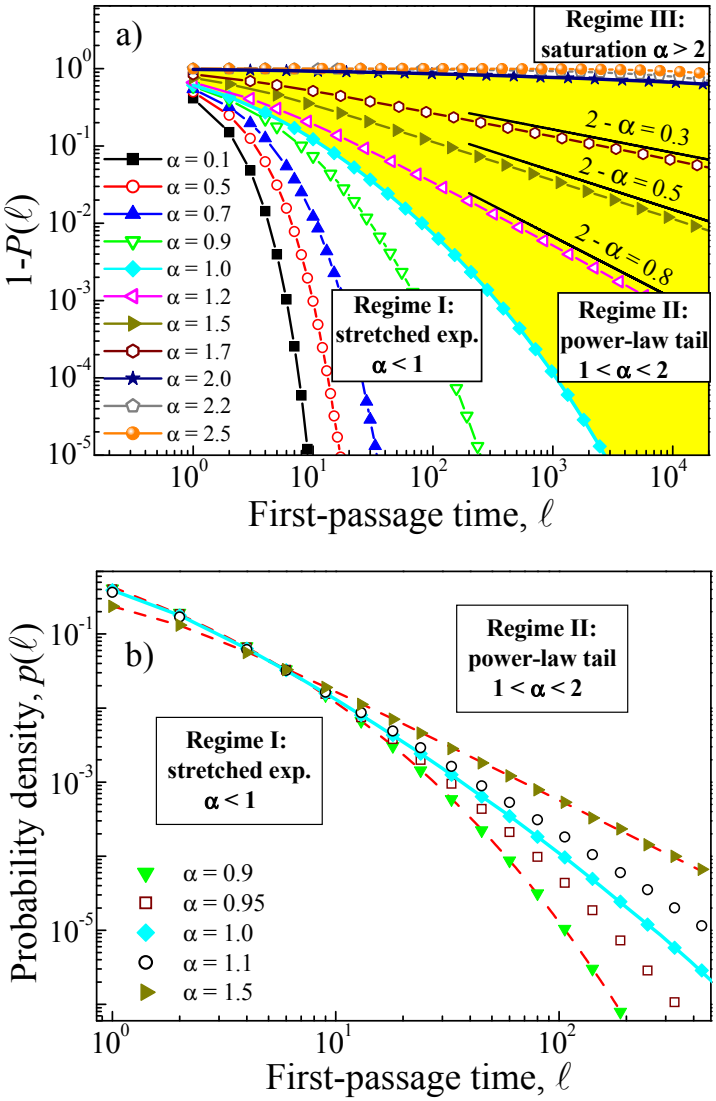

Figura 5.2: a) Distribución de probabilidad acumulada complementaria $1 - P(\ell)$ de los intervalos ℓ entre pasos consecutivos por cero para procesos invariantes de escala de tamaño $N = 2^{24}$ y grados diferentes de correlaciones cuantificadas por el exponente de escala α. b) Densidad de probabilidad $p(\ell)$ para valores pequeños de ℓ para procesos cercanos al punto de transición $\alpha = 1$. Las líneas discontinuas corresponden a ajustes con el modelo (5.1) para $\alpha = 0.9$, y con el modelo (5.3) para $\alpha = 1.5$.

5.2.3. Régimen de saturación

Para $\alpha > 2$, obtendríamos $\delta = 3-\alpha < 1$ y en esta situación $p(\ell)$ no se puede normalizar en el límite de un tamaño de sistema grande N.

En este régimen, observamos que $p(\ell)$ se aplana a medida que crece α (véase figura 5.3) y tiende a $p(\ell) = 1/N$, la cual se muestra mediante un rectángulo sombreado en la figura 5.3. Sin embargo, los efectos de tamaño finito son muy importantes, y se aprecia un pico en $\ell = N/2$, más pronunciado a medida que crece α.

En la práctica, muchos de los FPTs son del orden del tamaño del sistema y, en consecuencia, la probabilidad acumulada complementaria $1 - P(\ell)$ es esencialmente plana independientemente de α (véase 5.2a)).

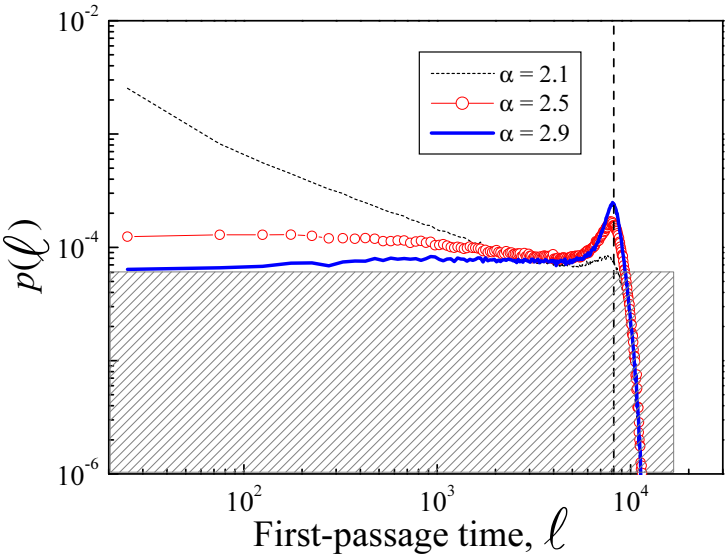

Figura 5.3: Densidad de probabilidad $p(\ell)$ para procesos con diferentes valores de α en el régimen de saturación. Los resultados corresponden a un tamaño de sistema $N = 2^{14}$ y se han obtenido con 10^5 realizaciones para cada valor de α. El rectángulo sombreado corresponde a la distribución uniforme $p(\ell) = 1/N$.

5.3. Comportamiento del valor medio

Una propiedad interesante es el comportamiento de $\langle \ell \rangle$ en función del tamaño del sistema N, que es diferente en los tres regímenes mencionados en la sección anterior (véase figura 5.4).

i) En el rango $\alpha < 1$ (régimen en *stretched exponential*), $\langle \ell \rangle$ tiende asintóticamente a un valor constante en el límite de un tamaño de sistema N grande, como podemos

observar en la figura 5.4a. En este régimen, el comportamiento de $\langle \ell \rangle$ como función de N se ajusta a un modelo del tipo:

$$\langle \ell \rangle = \langle \ell \rangle_\infty \left(1 - \frac{1}{cN^b} \right), \tag{5.4}$$

donde b y c son constantes positivas, y $\langle \ell \rangle_\infty$ representa el valor asintótico.

Nótese que, a medida que aumenta α, la convergencia al valor asintótico $\langle \ell \rangle_\infty$ es más lenta, y que los valores de $\langle \ell \rangle_\infty$ también crecen con α.

ii) En el rango $1 < \alpha < 2$ (régimen de cola en ley de potencias), sin embargo, encontramos que $\langle \ell \rangle$ diverge con el tamaño del sistema N siguiendo una ley de potencias (véase figura 5.4b):

$$\langle \ell \rangle \sim N^\gamma \tag{5.5}$$

Esto concuerda con el hecho de que, en este rango de α, vimos que la cola de la función de densidad $p(\ell)$ seguía una ley de potencias (5.2). De hecho, si se verifica la ecuación 5.2, entonces

$$\langle \ell \rangle = \int_1^N \ell \, p(\ell) \, d\ell \sim \int_1^N \ell \, \ell^{-\delta} \, d\ell \sim N^{2-\delta}. \tag{5.6}$$

Por tanto, $\gamma = 2 - \delta$, y como $\delta = 3 - \alpha$ obtenemos $\gamma = \alpha - 1$. Como se puede ver en la figura 5.4b, nuestros ajustes numéricos a leyes de potencias corresponden a valores de γ que verifican esa relación.

iii) En el rango $\alpha > 2$ (régimen de saturación), $\langle \ell \rangle$ también diverge con el tamaño del sistema N siguiendo una ley de potencias, pero con exponente constante $\gamma = 1$ para todos los valores de α (véase figura 5.4b), es decir, $\langle \ell \rangle \sim N$.

Nótese que $\langle \ell \rangle$ no puede crecer más rápido que el tamaño del sistema N, excluyendo valores $\gamma > 1$.

En cuanto a las transiciones entre regímenes distintos, para $\alpha = 1$ se observa una transición de fase de un comportamiento convergente en el valor medio de los tiempos de primer paso a un comportamiento divergente (figura 5.4). En este punto de transición, $\langle \ell \rangle$ no converge a un valor finito, como en el régimen en *stretched exponential*, ni diverge con N como una ley de potencias, como en el régimen de cola en ley de potencias. Encontramos que $\langle \ell \rangle$ diverge logarítmicamente en el límite termodinámico $N \to \infty$: $\langle \ell \rangle \sim \log N$.

En el punto de transición entre el régimen de cola en ley de potencias y el de saturación, $\alpha = 2$, encontramos que $\langle \ell \rangle \sim N / \log N$. Este comportamiento es intermedio entre los dos

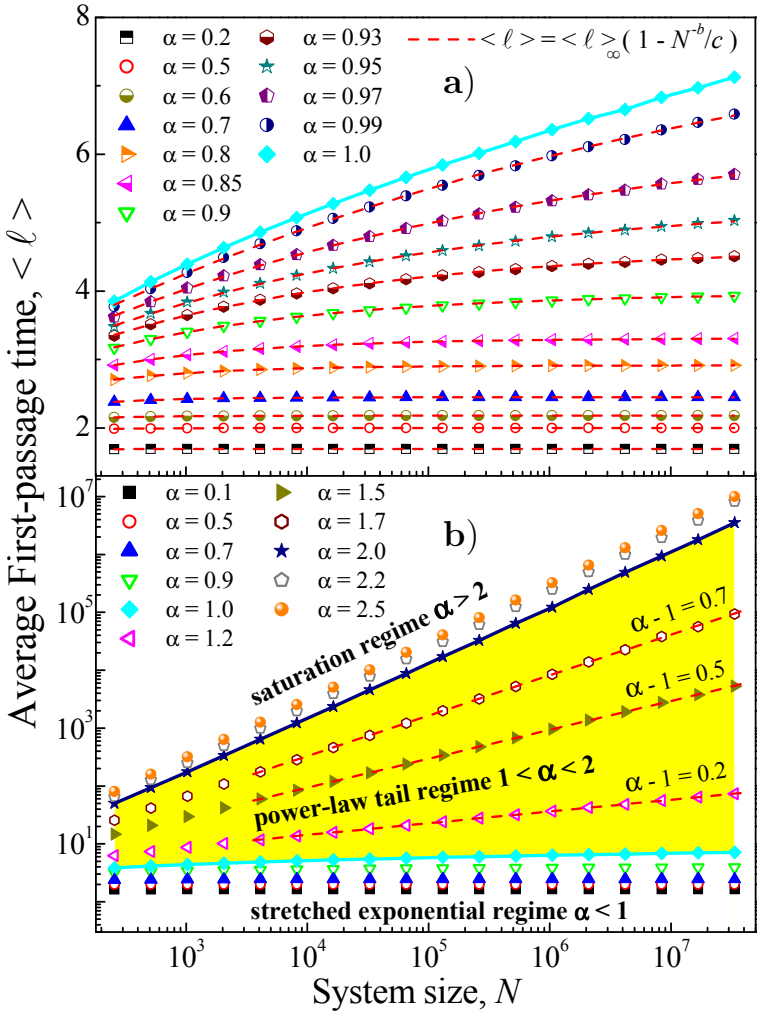

Figura 5.4: a) Comportamiento convergente de $\langle \ell \rangle$ como función del tamaño del sistema N en el régimen en *stretched exponential*($\alpha < 1$). Las líneas discontinuas representan los ajustes con (5.4).b) Dependencia de $\langle \ell \rangle$ con N para procesos invariantes de escala con correlaciones diferentes para los tres regímenes que identificamos en la figura 5.2. Obsérvese que el panel a) es una ampliación de la parte inferior del panel b). Las líneas discontinuas en el régimen de cola en ley de potencias corresponden a los ajustes $\langle \ell \rangle \sim N^{\gamma}$, with $\gamma = \alpha - 1$.

regímenes: $\langle \ell \rangle$ crece más rápido que cualquier ley de potencias con $\gamma < 1$, pero más lento que una ley de potencias con $\gamma = 1$ (figura 5.4b).

El comportamiento de $\langle \ell \rangle$ en el límite termodinámico se puede resumir en un diagrama de fases que mostramos en la figura 5.5. En el régimen en *stretched exponential* (panel izquierdo en la figura 5.5), donde $\langle \ell \rangle$ converge en el límite termodinámico, la elección natural del parámetro de orden es el valor asintótico, $\langle \ell \rangle_\infty$, que crece con α y diverge cuando $\alpha \to 1^-$.

En los otros dos regímenes (panel derecho en la figura 5.5), como $\langle \ell \rangle$ diverge con N de la forma $\langle \ell \rangle \sim N^\gamma$, un parámetro de orden conveniente para describir el comportamiento de $\langle \ell \rangle$ es el exponente γ que tiende a cero cuando $\alpha \to 1^+$, y converge a $\gamma = 1$ cuando $\alpha \to 2^-$. En el régimen de saturación $\alpha > 2$, el parámetro de orden permanece constante: $\gamma = 1$.

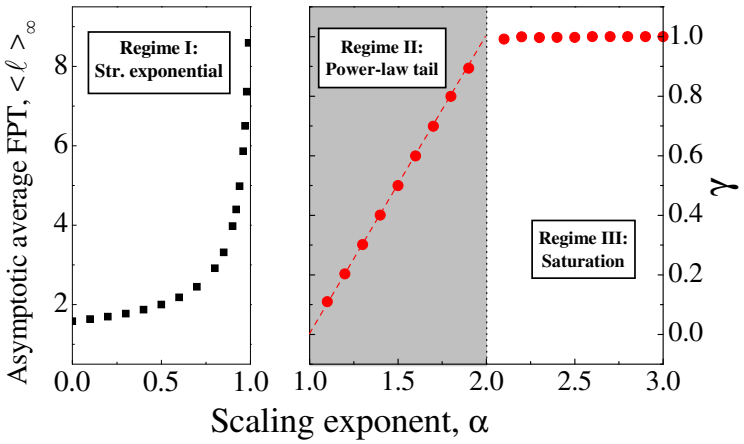

Figura 5.5: Diagrama de fase de las transiciones de régimen en *stretched exponential* a cola en ley de potencias y a saturación. Los símbolos corresponden a resultados numéricos, y la línea discontinua a la curva $\gamma = \alpha - 1$. Para $\alpha < 1$ (panel izquierdo) el parámetro de orden es el valor asintótico $\langle \ell \rangle_\infty$ (figura 5.4a), mientras que para $\alpha > 1$ (panel derecho) el parámetro de orden es el exponente γ de la ecuación (5.5).

Las principales propiedades de la longitud promedio $\langle \ell \rangle$ de FPT y la densidad de probabilidad $p(\ell)$ en los tres regímenes están también resumidas en la Tabla 5.1.

Los resultados que hemos obtenido para el comportamiento de $\langle \ell \rangle$ en los tres regímenes también se pueden entender en términos de los efectos de tamaño finito de la distribución $P(\ell)$. En la figura 5.6 observamos que:

i) Para procesos con $\alpha < 1$, $P(\ell)$ es esencialmente independiente del tamaño del sistema N. De este modo, el FPT medio $\langle \ell \rangle$ está bien definido y para N suficientemente grande no hay efectos de tamaño apreciables, dando lugar a un valor asintótico finito $\langle \ell \rangle_\infty$ (véase figura 5.4(a)).

α	$p(\ell)$	$\langle \ell \rangle$
Régimen I (*stretched exp.*) $0 < \alpha < 1$	$\sim \exp\left[-\left(\ell/\ell_0\right)^{2-2\alpha}\right]$	$\lim_{N\to\infty}\langle \ell \rangle = \langle \ell \rangle_\infty$, constante fijado α $\langle \ell \rangle_\infty$ crece con α
Régimen II (*power-law tail*) $1 < \alpha < 2$	$\sim 1/\ell^{3-\alpha}$	$\sim N^{\alpha-1}$
Régimen III (saturación) $2 < \alpha < 3$	se aplana con α efectos de tamaño importantes en $\ell = N/2$	$\sim N$

Tabla 5.1: Propiedades de la densidad de probabilidad $p(\ell)$ y la longitud de FPT media $\langle \ell \rangle$ en los tres regímenes distintos como función del exponente de correlación α del DFA.

ii) En el punto de transición $\alpha = 1$, donde $\langle \ell \rangle$ diverge logarítmicamente (figura 5.4(a)), los efectos de tamaño del sistema en $P(\ell)$ son más pronunciados (véase figura 5.6, panel central).

iii) Por encima del punto de transición $\alpha > 1$, para cualquier realización finita hay un corte en la cola de la ley de potencias de $P(\ell)$ que escala con el tamaño N del sistema (véase figura 5.6, panel inferior), asegurando la cola en ley de potencias de la distribución incluso en el límite termodinámico $N \to \infty$, y por tanto $\langle \ell \rangle$ diverge siguiendo una ley de potencias en N (figura 5.4(b)).

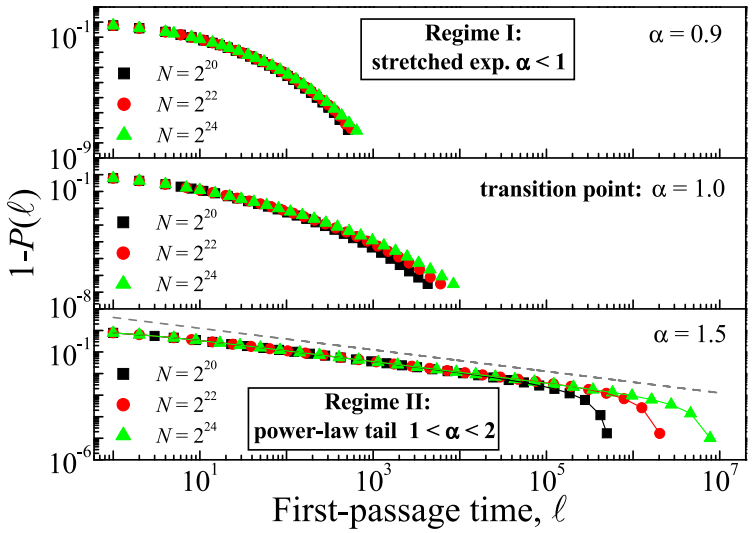

Figura 5.6: Dependencia de la distribución acumulada complementaria $1-P(\ell)$ del tamaño del sistema N. La transición del régimen en *stretched exponential* al de cola en ley de potencias es estable e independiente de N. Las distribuciones que se muestran en todos los paneles se obtienen mediante simulaciones de Monte Carlo con $2^{32}/N$ realizaciones.

Otra cantidad importante relacionada con la dependencia de las estadísticas del FPT con el tamaño del sistema N es el número medio de segmentos de FPT, $\langle n \rangle$. Se definen los segmentos como partes continuas del proceso con signo constante, cuyas fronteras son los cruces por cero (véase figura 5.1). Encontramos que el comportamiento de $\langle n \rangle$ en los tres regímenes es esencialmente el inverso de $\langle \ell \rangle$, a saber:

i) En el régimen en *stretched exponential*, $\langle n \rangle$ diverge de la forma $\langle n \rangle \sim N$ independientemente de α.

ii) En el régimen de cola en ley de potencias, $\langle n \rangle$ diverge más lentamente, $\langle n \rangle \sim N^{\lambda}$, donde el exponente $\lambda = 2 - \alpha$ decrece cuando $\alpha \to 2^{-}$.

iii) En el régimen de saturación ($\alpha > 2$), $\langle n \rangle$ converge con $N \to \infty$ a un valor asintótico constante $\langle n \rangle_\infty$, que decrece a medida que crece α.

5.4. Relación con secuencias binarias

Los resultados obtenidos en este capítulo están estrechamente relacionados con el comportamiento de secuencias binarias.

Como vimos en el capítulo anterior, una técnica estándar para generar procesos fractales binarios con correlaciones es simplemente considerar el signo del proceso fractal continuo subyacente. De esta forma, las secuencias binarias obtenidas están compuestas de segmentos de solo dos valores posibles, $+1$ o -1. El tamaño de esos segmentos son los FPTs de la señal original (figura 5.1). Este tipo de secuencias binarias se dan en sistemas de diversa naturaleza: señales sísmicas [Varotsos et al., 2003b], transporte por membranas [Varotsos et al., 2003a], cadenas de ADN [Carpena et al., 2007, 2011] y sólidos binarios desordenados [Usatenko et al., 2008; Carpena et al., 2002].

5.5. Procesos con *crossovers*

El análisis hecho a lo largo de este capítulo se puede extender a procesos con correlaciones de largo alcance que, en lugar de un único exponente de escala, presenten dos regímenes diferentes caracterizados por distintos exponentes de correlación a escalas pequeñas y grandes con un *crossover* separando ambas regiones [Carpena et al., 2016b]. Esto ocurre en sistemas complejos regulados por mecanismos que compiten actuando a diferentes escalas temporales.

Para poder hacer un estudio sistemático de las propiedades estadísticas de los tiempos de primer paso en estos procesos se usa una versión modificada del Método de Filtrado de Fourier. En lugar de multiplicar en el dominio de las frecuencias (f) un ruido blanco por una ley de potencias de la forma $f^{-(2\alpha-1)/2}$, se multiplica por

$$Q(f) = \begin{cases} f^{-(2\alpha_l-1)/2} & \text{si} \quad f \leq f_c \\ f_c^{(\alpha_s-\alpha_l)}\, f^{-(2\alpha_s-1)/2} & \text{si} \quad f > f_c \end{cases} \tag{5.7}$$

y, volviendo al espacio temporal mediante la transformada inversa de Fourier, obtenemos una secuencia con un *crossover* en t_c ($f_c \equiv t_c^{-1}$) y con exponentes de escala α_s y α_l del DFA, a escalas temporales pequeñas y grandes, respectivamente.

Los resultados muestran que, en general, la forma funcional de la función de densidad $p(\ell)$ presenta un comportamiento mixto: a escalas cortas se comporta como la esperada

(según lo mostrado a lo largo de este capítulo) para un proceso con un único exponente de escala α_s y, a escalas grandes, como la esperada para un proceso con un único exponente de escala α_l,

$$p(\ell) = \begin{cases} p_{\alpha_s}(\ell) & \text{si} \quad \ell < g_{t_c} \\ p_{\alpha_l}(\ell) & \text{si} \quad l > g_{t_c} \end{cases} , \tag{5.8}$$

con $g(t_c)$ una función monótona en t_c.

Los detalles de estos resultados y su comparación con observaciones experimentales se pueden consultar en [Carpena et al., 2016b].

5.6. Conclusión

En este capítulo hemos visto que las correlaciones se pueden considerar el factor unificador que controla las propiedades estadísticas de los FPTs en una amplía clase de procesos fractales, independientemente de las especificidades del sistema dinámico particular considerado.

Cuando las correlaciones están en el rango $\alpha < 1$ (tales como registros climáticos [Bunde et al., 2005] o en actividad bursátil [Ivanov et al., 2004; Wang et al., 2009]), la función de densidad de probabilidad de los FPTs $p(\ell)$ se comporta como *stretched exponential*, con un valor medio finito incluso para tamaños de sistema divergentes. Puesto que este régimen corresponde a las correlaciones que se observarán en las palabras, no es sorprendente que las distancias entre apariciones consecutivas de una palabra se comporten como una *stretched* (a escala larga), como veremos en el próximo capítulo, y que dará pie al modelo que describiremos.

Por el contrario, cuando las correlaciones están en el rango $1 < \alpha < 2$, como por ejemplo en procesos de difusión anómalos [Shlesinger et al., 1993; Rangarajan and Ding, 2000; Khoury et al., 2011; Eliazar and Klafter, 2009], la cola de $p(\ell)$ sigue una ley de potencias, $p(\ell) \sim \ell^{-\delta}$ con $\delta = 3 - \alpha$, generalizando los resultados conocidos para el *classical random walk* ($\alpha = 3/2$), para el que $\delta = 3/2$. En este caso, el valor medio de FPT crece como ley de potencias con el tamaño del sistema con un exponente menor que uno.

Para el caso de procesos con correlaciones extremas ($2 < \alpha < 3$), que se pueden ver como integraciones de fBm's, las densidades de probabilidad $p(\ell)$ son esencialmente planas, y el valor medio de FPT diverge con el tamaño del sistema.

Estos resultados se pueden extender a procesos que están controlados por un exponente de correlación a escala pequeña y otro a escala grande, concluyendo que si las correlaciones y el *crossover* es similar, lo serán también las propiedades estadísticas de los tiempos de

primer paso, independientemente de las características concretas del sistema dinámico bajo estudio.

Capítulo 6

Correlaciones de largo alcance en palabras clave: un modelo que las reproduce

En las últimas décadas ha habido un interés creciente por el estudio del lenguaje humano en el contexto de los sistemas complejos. Los textos escritos son buenos candidatos para tal estudio, ya que están compuestos por elementos individuales (palabras) que interactúan entre ellos en formas complejas y a diferentes niveles, controlados por las reglas gramaticales del idioma particular, el género literario, el estilo del escritor y la información que el texto quiere transmitir.

Las aproximaciones se han centrado en tres temas principales, a saber:

i) La detección automática de palabras clave.

   Como vimos en la primera parte de esta tesis, la idea es detectar las palabras clave de un texto (es decir, las palabras relacionadas con los tópicos principales) sin el uso de información externa. Recordemos que una estrategia exitosa para abordar este problema [Ortuño et al., 2002; Carpena et al., 2009; Carretero-Campos et al., 2013; Carpena et al., 2016a] se basaba el hecho de que las palabras relevantes se atraen entre sí, y se concentran en determinadas regiones del texto formando *clusters* y dando lugar a grandes fluctuaciones de frecuencia. Sin embargo, las palabras comunes se distribuyen de manera más homogénea. Por tanto, cuanto mayor sea el *clustering*, mayor será la relevancia. Se obtenía así un *ranking* de relevancia cuantificando adecuadamente el *clustering* de cada palabra.

ii) Correlaciones de largo alcance en textos.

   Las interacciones complejas entre palabras que se producen a muchos niveles, que

comentamos previamente, dan lugar a una estructura espacial compleja no trivial en textos que se puede cuantificar por medio del análisis de las correlaciones de largo alcance. Los resultados de esos análisis concluyen que los textos escritos presentan estructuras de largo alcance que dan lugar a correlaciones de largo alcance que han sido cuantificadas para diferentes textos e idiomas [Montemurro and Pury, 2002; Bhan et al., 2006; Alvarez-Lacalle et al., 2006; Şahin et al., 2009; Altmann et al., 2012; Ogura et al., 2019].

iii) Modelos para reproducir la distribución espacial de las palabras.

Se intenta modelar el patrón de apariciones de una palabra en un texto y, concretamente, reproducir la distribución $p(d)$ de las distancias entre apariciones sucesivas de una palabra dada (también denominados tiempos de recurrencia). Algunos resultados muestran [Altmann et al., 2009; Tanaka-Ishii and Bunde, 2016] que, en general, para cualquier palabra $p(d)$ se comporta como una *stretched-exponential*[1]. Aunque este resultado concuerda bastante bien con observaciones experimentales para d grande, falla a distancias cortas, donde $p(d)$ se sobreestima sistemáticamente.

En este capítulo veremos que estas tres cuestiones no son independientes, sino que están profundamente relacionadas y no son más que diferentes caras del mismo problema. En primer lugar, mostramos que las correlaciones de largo alcance observadas en textos se deben a sus palabras clave: cuantificamos las correlaciones de largo alcance de cada palabra del texto, y mostramos que las palabras relevantes presentan fuertes correlaciones de largo alcance y son las responsables de las correlaciones observadas en el texto completo, mientras que las palabras comunes no presentan correlaciones y no contribuyen a las correlaciones del texto. De hecho, presentamos resultados que indican que el grado de correlaciones de una palabra es también una buena medida de su relevancia para el texto. En segundo lugar, presentamos un modelo capaz de reproducir la distribución espacial de una palabra en un texto. Partiendo de las correlaciones de la palabra y de su frecuencia en el texto, el modelo predice la posición de las apariciones consecutivas de la palabra y reproduce todas sus propiedades interesantes: sus correlaciones de largo alcance, su estructura espacial caracterizada por la distribución de distancias $p(d)$ en el rango completo de d (resolviendo el problema de resultados previos) y también el grado de relevancia de la palabra cuantificado por la medida de *clustering* mencionada anteriormente.

[1] Obsérvese que este es el comportamiento que obtuvimos en el capítulo anterior para la distribución de las longitudes ℓ de los FPT's en el rango $\alpha < 1$.

6.1. Metodología

6.1.1. Cuantificación de correlaciones de largo alcance en palabras

Representamos las ocurrencias de cada palabra del texto mediante una secuencia binaria de la siguiente forma:

Dado un texto de longitud N, para cada palabra distinta w del texto generamos una secuencia binaria $x_w(i)$ $(i = 1, 2, \ldots, N)$ asignando el valor 1 en todas las posiciones i del texto donde la palabra aparece y 0 en el resto. De este modo, tenemos una secuencia binaria distinta para cada una de las palabras que conforman el vocabulario del texto. Elegimos un mapeo del texto en una secuencia binaria para evitar la introducción de correlaciones no reales debidas a la asignación numérica [Voss, 1992].

Cuantificaremos las correlaciones de dichas secuencias binarias por medio del exponente de escala α del DFA (véase sección 4.4). Recordemos que el DFA mide el promedio de las fluctuaciones $F(\ell)$ de la secuencia a diferentes escalas ℓ y que una relación en ley de potencias,

$$F(\ell) \propto \ell^{\alpha} \tag{6.1}$$

indica la presencia de *scaling*. Si $\alpha = 0.5$ la secuencia no tiene correlaciones, mientras que si $\alpha < 0.5$ o $\alpha > 0.5$ presenta correlaciones negativas o positivas, respectivamente.

6.1.2. Modelo para reproducir la distribución espacial

El objetivo del modelo que queremos desarrollar es que reproduzca la distribución de una palabra a lo largo del texto, así como sus correlaciones de largo alcance. Proponemos generar una secuencia binaria con correlaciones de largo alcance que represente las ocurrencias de la palabra a lo largo del texto de la siguiente forma:

i) En un primer paso, mediante el Método de Filtrado de Fourier (FFM) [Makse et al., 1996], generamos una secuencia de números reales aleatorios $x(i)$ siguiendo una N(0,1) con correlaciones de largo alcance en ley de potencias caracterizadas por el exponente de escala α.

ii) En un segundo paso, asumimos que la palabra aparece a lo largo del texto en las posiciones i en las que la secuencia $x(i)$ supera un umbral (r), de manera que las ocurrencias de la palabra se pueden representar por medio de la secuencia binaria

$x_{bin}(i)$ obtenida de

$$x_{bin}(i) = \begin{cases} 1 & \text{si} \quad x(i) \geq r \\ 0 & \text{en otro caso} \end{cases} \tag{6.2}$$

El umbral r lo obtenemos numéricamente imponiendo que la probabilidad de tener valores de la secuencia por encima del umbral coincida con la frecuencia de la palabra que queremos modelar.

Denotamos ahora por $p_0(d)$ la distribución de las distancias entre apariciones de la palabra que estamos modelando. En el capítulo anterior, hemos obtenido por medio de simulaciones numéricas que la forma funcional de la distribución acumulada complementaria $Q_0(d)$ (definida como $Q_0(d) \equiv \int_d^\infty p_0(x)\,dx$) es *stretched exponential* [Carretero-Campos et al., 2012] , es decir,

$$Q_0(d) \sim e^{-(d/c)^\beta} \tag{6.3}$$

donde las constantes β y c dependen de la palabra considerada.

Como comentamos previamente, esta forma funcional (línea discontinua en la figura 6.5) fue propuesta por [Altmann et al., 2009] para caracterizar la distribución de las distancias entre apariciones sucesivas de la misma palabra. Dicha expresión concuerda bastante bien con las observaciones experimentales para d grande, pero sobreestima sistemáticamente la distribución a distancias cortas d (véase la figura 6.5).

Sin embargo, observamos que las palabras reales tienen repulsión a distancias cortas, debido a las restricciones impuestas por la gramática que no permite la aparición de la misma palabra a distancias muy cortas. Para incorporar este fenómeno, proponemos un factor de repulsión $f(d)$ que modifica la distribución de distancias entre apariciones obtenida mediante el procedimiento descrito anteriormente. Aceptamos una distancia d (y, por tanto, un '1' a una distancia d del anterior en la secuencia binaria) con probabilidad $f(d)$ dada por

$$f(d) = \begin{cases} 1 - e^{-\frac{(d-(d_{min}-1))}{a}} & \text{si} \quad d \geq d_{min} \\ 0 & si \quad d < d_{min} \end{cases} \tag{6.4}$$

donde d_{min} es la distancia entre ocurrencias mínima observada para la palabra real a lo largo del texto. El parámetro a da información sobre la escala característica de la repulsión y depende de la palabra considerada. Con este factor de repulsión, tenemos que

$$p(d) \sim f(d)\, p_0(d)\ . \tag{6.5}$$

Obsérvese que la repulsión $f(d)$ modifica a $p_0(d)$ disminuyendo su resultado solo a distancias cortas. Para distancias d suficientemente grandes, $f(d) \simeq 1$ y, en consecuencia, $p(d) \simeq p_0(d)$. De esta forma, $p(d)$ presenta el comportamiento correcto a todas las escalas.

Los tres parámetros del modelo (el exponente de correlación α, el umbral r y el parámetro de escala de la repulsión a) pueden ser estimados inicialmente del texto que queremos analizar: α se calcula usando el DFA en la secuencia binaria obtenida del texto para la palabra considerada, r se obtiene a partir de la frecuencia de la palabra en el texto, y a como la distancia d a la que la distribución de distancias de la palabra real se separa de la *stretched exponential* (véase la figura 6.5). Este procedimiento puede ser difícil de implementar debido a la dependencia entre los parámetros.

Nótese que aquí cabría plantearse de nuevo si el exponente de correlación α_{bin} de la secuencia binaria obtenida a partir de un umbral r va a coincidir con el de la secuencia real α de partida. En el capítulo 4 (sección 4.4.2) vimos que, si $0 < \alpha < 1$, se conservan las correlaciones si se consideran los pasos por 0. En trabajos recientes [Kalra and Santhanam, 2021] en los que se estudia si se pueden inferir las correlaciones de largo alcance de una serie temporal a partir de sus eventos extremos, resultados numéricos muestran que para un umbral $r = \mu+\sigma$ (siendo μ y σ la media y desviación estándar de la señal, respectivamente) las correlaciones son similares. Sin embargo, se observa que, a medida que se aumente el umbral, parece que empiezan a disminuir. Para un umbral alto (como ocurrirá con las palabras) es entonces esperable que se pierdan correlaciones.

6.2. Vínculo entre relevancia y correlaciones de largo alcance

Mostramos aquí, al igual que en la primera parte de la tesis, resultados obtenidos usando el libro *The Origin of Species* de Charles Darwin (6th Edition), que tiene una longitud de $N = 193786$ palabras y un vocabulario de 8186 palabras distintas.

Al tratarse de un texto largo, no hay grandes diferencias entre usar la medida de *clustering* C [Carpena et al., 2009] ó $K(N, n)$ [Carpena et al., 2016a], diferencias que habría si se tratase de un texto corto. En primer lugar, calculamos la medida de relevancia C para las 8186 palabras distintas para obtener un *ranking* de relevancia (ordenado en orden decreciente de C). Además, para cada palabra con frecuencia al menos 100 (un total de 252 palabras), consideramos la secuencia binaria que representa sus apariciones a lo largo del texto (véase metodología) y cuantificamos las correlaciones de dicha secuencia por medio del DFA. Consideramos un valor mínimo para la frecuencia con el objetivo de evitar resultados afectados por una baja estadística.

En la figura 6.1 mostramos los resultados del DFA aplicado a una palabra relevante, 'species' (posición 4 en el *ranking*), y a una palabra no relevante, 'but' (posición 8185 en el *ranking*). Observamos que a distancias cortas se obtiene para ambas palabras un

exponente de correlación α próximo a 0.5. Sin embargo, a una distancia que está más allá de los efectos de las reglas gramaticales, observamos un *crossover* en el exponente de escala de la palabra relevante hacia un valor de α significativamente mayor que 0.5, lo que indica fuertes correlaciones de largo alcance.

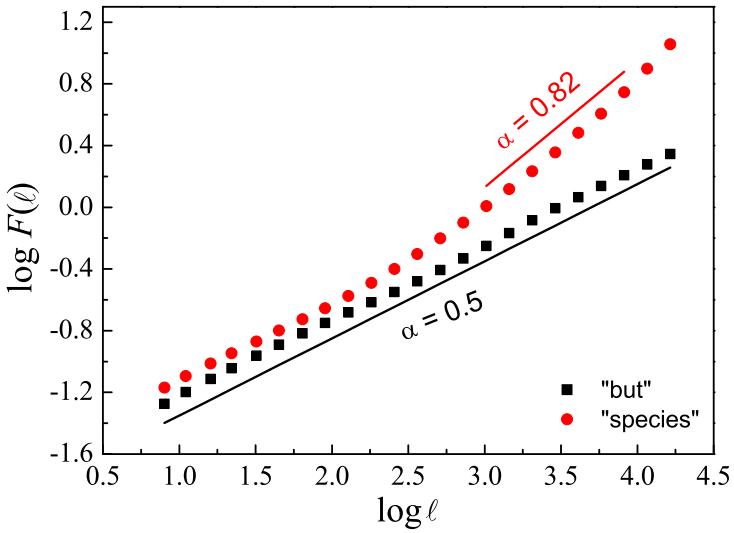

Figura 6.1: Método DFA aplicado a las secuencias binarias que representan las apariciones de las palabras 'species' (relevante) y 'but' (no relevante) a lo largo del libro *The Origin of Species* de Charles Darwin

Nótese que el exponente 0.5 obtenido a escalas pequeñas se extiende en un rango que supera las escalas a las que la función $F(\ell)$ no se comporta como una ley de potencias [Carpena et al., 2022]. Por otro lado, vimos también que [Carpena et al., 2017] (sección 4.4.2), cuando la secuencia es estacionaria, el DFA interpreta erróneamente ausencia de correlaciones en la secuencia del signo en lugar de anticorrelaciones. Así que, en trabajos futuros, podríamos plantearnos la posibilidad de que, a escalas pequeñas, el exponente 0.5 que observamos se corresponda a anticorrelaciones debido a la repulsión que imponen las reglas gramaticales. La existencia de un *crossover* a escalas intermedias en las correlaciones aparece en general solo para palabras relevantes (palabras con valores altos de C), mientras que palabras no relevantes tienen un exponente de correlación próximo a 0.5 en todas las escalas y no presentan un *crossover* a un segundo régimen. Para ilustrar este comportamiento calculamos un ajuste lineal por mínimos cuadrados de $\log F(\ell)$ versus $\log \ell$ en un rango de distancias ℓ desde 1000 a 10000 para obtener un valor para el exponente de correlación a escalas grandes, que denotaremos α_2.

En la figura 6.2 mostramos la distribución de probabilidad de α_2 para las 50 palabras más informativas y para las 50 menos informativas del *ranking* de relevancia. La separación entre las dos distribuciones sugiere que α_2 se puede usar para distinguir las palabras

relevantes de un texto: cuanto mayor sea el valor de α_2, mayor será la relevancia, como podemos ver en la tabla 6.1. Obsérvese que las distribuciones de probabilidad de α_2 para las 50 palabras más frecuentes y para las 50 menos frecuentes casi no difieren entre ellas (véase el recuadro en la figura 6.2), de modo que estos resultados son independientes de la frecuencia de aparición.

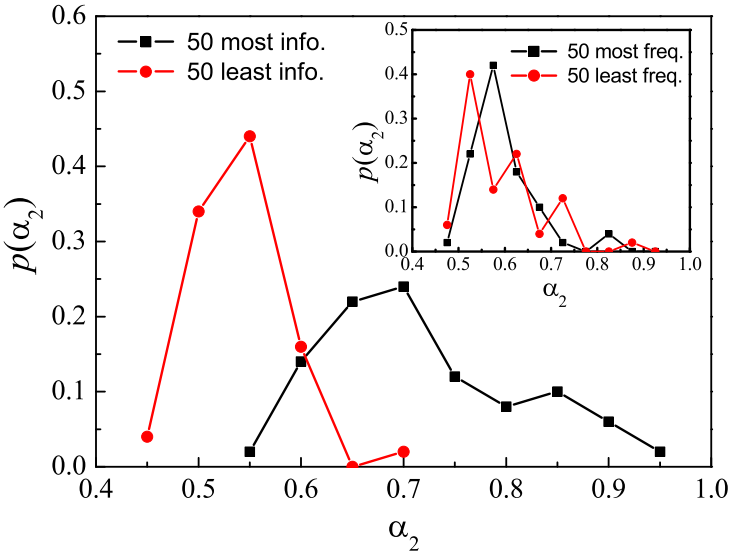

Figura 6.2: Distribución de probabilidad del exponente de escala α_2 obtenido mediante un ajuste lineal por mínimos cuadrados de $\log F(\ell)$ versus $\log \ell$ en un rango de distancias ℓ desde 1000 a 10000. Para las 50 palabras más informativas versus las 50 menos informativas del libro *The Origin of Species* de Charles Darwin. Inset: lo mismo para las 50 palabras más frecuentes versus las 50 menos frecuentes.

El vínculo entre correlaciones de largo alcance y relevancia también se ha observado para otros libros y en otros idiomas, lo que sugiere que se trata de una característica universal (véase Apéndice A). Otros trabajos que usan secuencias binarias en las que la escala temporal son las frases, y la función de autocorrelación en lugar del DFA, llegan a conclusiones similares [Ogura et al., 2019].

6.3. Propiedades que reproduce el modelo

Presentamos algunos resultados del modelo que hemos propuesto para reproducir las propiedades de correlación de las palabras. En la figura 6.3 se muestra el proceso de generación de una palabra artificial que modele la distribución de la palabra 'parts' a lo largo del libro *The Origin of Species* (para una elección adecuada de los parámetros, véase metodología). En este caso, los parámetros son $\alpha = 0.88$, $r = 2.96989$ y $a = 10$. Nótese que se necesita partir de un exponente de correlación α superior al de la palabra 'parts'

word	α_2	word	α_2
seeds	0.945	case	0.500
islands	0.905	also	0.499
young	0.901	whether	0.498
water	0.889	hence	0.495
flowers	0.859	so	0.493
forms	0.849	either	0.491
varieties	0.843	both	0.484
organs	0.836	give	0.484
breeds	0.833	then	0.471
species	0.816	us	0.464

Tabla 6.1: Las primeras 10 palabras y las últimas 10 palabras extraídas del libro *The Origin of Species* mediante el exponente de correlación a escalas grandes α_2

(por ello no está incluida en la tabla 6.1), ya que el valor del umbral es alto, a lo que hay que sumar el efecto de la repulsión. Mostramos aquí los resultados para 'parts' como palabra representativa promedio (su frecuencia en el texto es 233), pero el modelo se puede aplicar a cualquier palabra. Como podemos ver en la figura 6.3b y 6.3c, la distribución espacial de la palabra modelada concuerda con la de la palabra real, dando lugar a una estructura de *clusters* similar.

Si aplicamos el DFA a las secuencias binarias generadas mediante 256 iteraciones del modelo y promediamos las curvas, obtenemos que el modelo conserva con precisión la estructura de correlaciones de la palabra (véase la figura 6.4). La curva promedio del modelo proporciona un ajuste muy bueno de las correlaciones de la palabra 'parts' a todas las escalas. Observamos que el modelo parece reproducir la distribución de una palabra a lo largo del texto, así como sus correlaciones de largo alcance.

Nos centramos ahora en la distribución de probabilidad de las distancias entre apariciones de la palabra 'parts', para ver si el modelo también es capaz de reproducirla. Como comentamos previamente, para ajustar esta distribución se ha usado en la literatura la *stretched exponential* [Altmann et al., 2009]. En la figura 6.5 representamos $\log(-\log(Q(d)))$ versus $\log d$, donde $Q(d)$ es la distribución acumulada complementaria de las distancias entre apariciones de la palabra. En esta escala, la *stretched exponential* (véase la ecuación 6.3) corresponde a una línea recta de pendiente β, pero ese comportamiento solo se observa para distancias grandes. A distancias cortas, que se ven afectadas por las reglas gramaticales, la distribución real está sobreestimada. Sin embargo, si representamos la distribución de distancias proporcionada por nuestro modelo (mostrada por la línea roja de la figura 6.5), vemos que el modelo es capaz de reproducir fielmente el comportamiento en todo el rango de distancias ya que hemos tenido en cuenta la repulsión

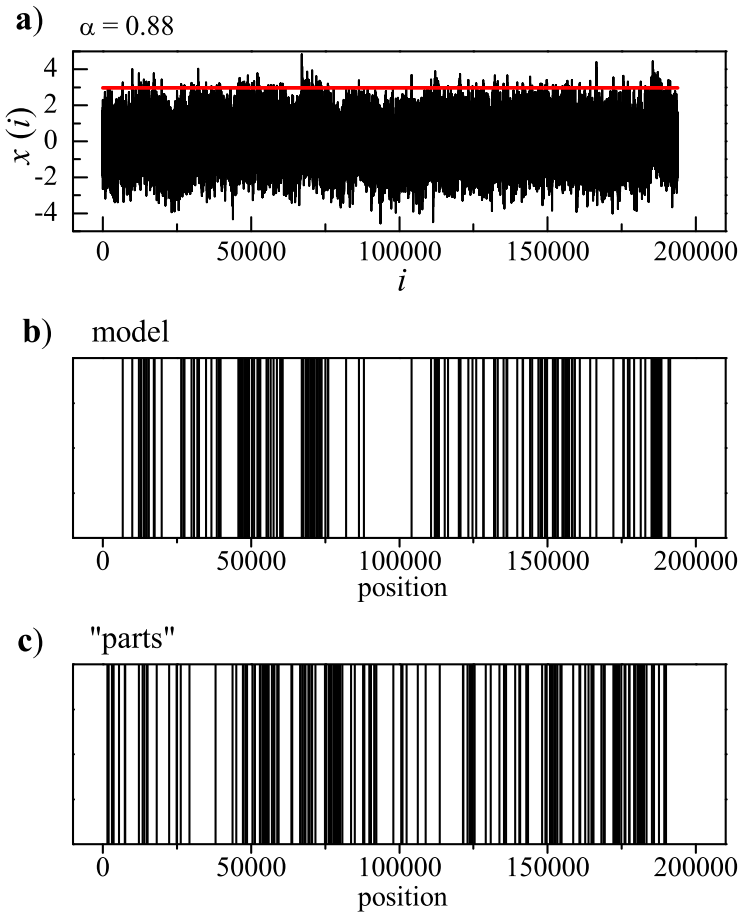

Figura 6.3: Proceso de generación de una palabra artificial que modele la distribución de la palabra 'parts' a lo largo del libro *The Origin of Species*: a) secuencia de números aleatorios $x(i)$ con exponente de correlación α (*línea negra*) y un umbral r (*línea roja*) usado para convertir la secuencia real en una binaria, b) apariciones de la palabra modelada a lo largo del texto, c) apariciones de la palabra real a lo largo del texto

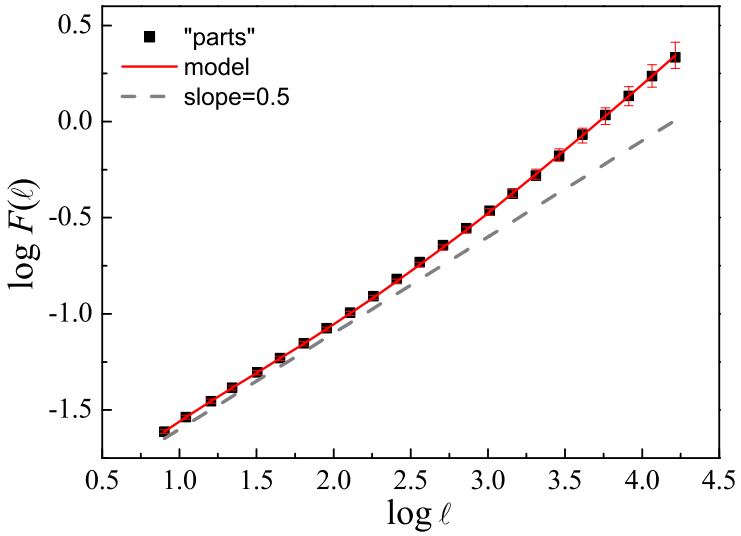

Figura 6.4: *Detrended Fluctuation Analysis* aplicado a la secuencia binaria que representa las apariciones de la palabra 'parts' (*cuadrados negros*) a lo largo del libro *The Origin of Species* de Charles Darwin, y el ajuste obtenido por medio de la palabra modelada (*línea roja continua*). Se dibuja una recta con pendiente 0.5 (*línea gris discontinua*) a efectos comparativos.

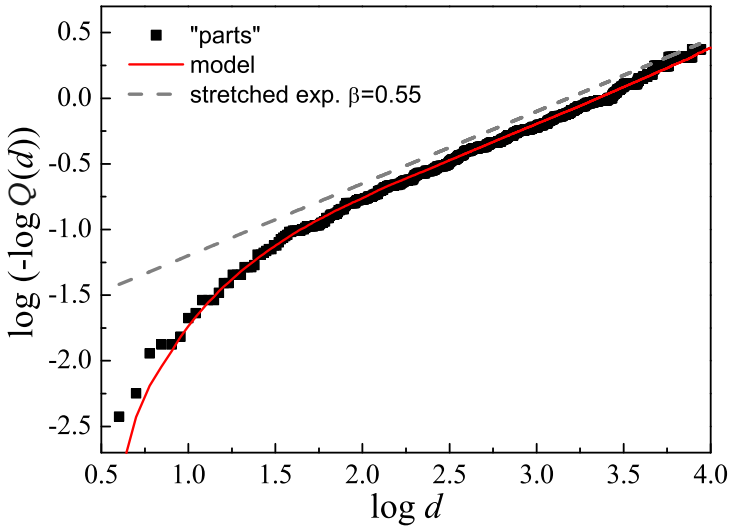

Figura 6.5: Distribución acumulada complementaria $Q(d)$ de las distancias entre apariciones de la palabra 'parts' (*cuadrados negros*) en el libro *The Origin of Species* (representada en una escala en la que la *stretched exponential* es una línea recta, véase la ecuación 6.3), y el ajuste obtenido por medio de la palabra modelada (*línea roja continua*). Se dibuja una recta con pendiente $\beta = 0.55$ (*línea gris discontinua*) a efectos comparativos.

que presentan las palabras reales a distancias cortas.

Por último, calculamos el valor promedio de la medida de relevancia C sobre todas las iteraciones del modelo, y la comparamos con el valor obtenido para la palabra real. Para la palabra 'parts', para la que $C = 9.71$, obtenemos un promedio $< C >= 12.55$ con desviación estándar $sd_C = 3.9$, y por tanto el valor real C está dentro del intervalo de confianza.

Los resultados del modelo son generales: hemos visto que es posible encontrar parámetros adecuados que proporcionan ajustes similares para otras palabras de este libro, para otros libros y en diferentes idiomas (resultados pendientes de sistematizar y detallar en trabajos futuros en los que se automatice la obtención de los parámetros).

6.4. Conclusión

Los resultados mostrados a lo largo de este capítulo nos llevan a concluir que: i) El grado de correlaciones de largo alcance de una palabra a escalas grandes está directamente relacionado con su relevancia y ii) Las correlaciones de largo alcance que se han observado en textos se deben a la distribución compleja de las palabras relevantes a lo largo del texto. Las palabras comunes, que tienen una distribución homogénea, no contribuyen a la existencia de correlaciones de largo alcance.

Por otro lado, hemos propuesto un modelo que es capaz de conservar todas las propiedades de interés de las palabras en textos escritos: reproduce el comportamiento complejo caracterizado por la presencia de correlaciones a escalas grandes, la estructura espacial caracterizada por la distribución de distancias, y el grado de relevancia cuantificado por medidas de relevancia basadas en *clustering*.

Conclusiones

Esta investigación estudia las secuencias simbólicas formadas por las palabras que constituyen un texto, y la información que se puede obtener a partir de la manera en la que se distribuye una palabra concreta a lo largo de la secuencia. Esto nos lleva a estudiar propiedades que también serán aplicables a otros tipos de secuencias simbólicas.

En la primera línea de trabajo, teniendo como objetivo la mejora en la detección de palabras clave en textos, especialmente cortos, y en la evaluación de los resultados obtenidos podemos destacar que:

I. Hemos realizado una adaptación de métricas para evaluar el comportamiento de distintos detectores de palabras clave, por medio de las cuales hemos comparado dos aproximaciones distintas, basadas en cuantificación de *clustering* y en cálculo de entropía, respectivamente. Hemos implementado mejoras en la medida de *clustering* y hemos estudiado la dependencia de la medida entrópica respecto de la partición del texto necesaria para su cómputo. Hemos comparado el comportamiento de ambas aproximaciones en textos de distinto tamaño y hemos obtenido que la medida entrópica proporciona resultados precisos solo para elecciones adecuadas de la partición del texto considerado, mientras que la medida de *clustering* proporciona resultados iguales o mejores, sin necesidad de realizar ninguna partición previa.

II. Hemos obtenido la expresión analítica de la distribución de las distancias entre ocurrencias sucesivas de un símbolo en una secuencia de tamaño N, suponiendo que sus n apariciones se sitúan al azar. Al estudiar las propiedades de dicha distribución, hemos encontrado un valor de frecuencia para el que el coeficiente de variación es máximo y hemos comprobado que en el caso asintótico, es decir, en el límite para n y N grandes, se obtiene la distribución geométrica. El conocimiento de la distribución exacta, junto al uso de condiciones de contorno, nos ha permitido definir una medida de *clustering* $K(N,n)$ más sensible a la detección de palabras de frecuencia baja (que serán comunes en un texto corto) y obtener su valor máximo. Hemos encontrado que estaremos en una situación de *clustering* extremo si $K_{\rm b}(N,n) \leq K(N,n) \leq K_{\rm max}(N,n)$, siendo $K_{\rm b}(N,n) \equiv \sqrt{(n-1)/2n}\, K_{\rm max}(N,n)$.

Estos resultados nos han permitido clasificar las palabras clave en dos clases: genéricas (relativamente frecuentes y con valor de K inferior a la cota $K_b(N,n)$) y específicas (su frecuencia no es grande y su valor de K supera la cota de *clustering* extremo).

En la segunda línea, en la que teníamos como objetivo si se podía establecer un vínculo entre la relevancia de una palabra y las correlaciones de largo alcance, y establecer un modelo que reproduzca su distribución espacial a partir de las correlaciones, podemos destacar que:

I. Hemos realizado un estudio numérico sistemático de las distribuciones de las longitudes entre pasos consecutivos por 0 para procesos invariantes de escala con correlaciones en ley de potencias. Al realizar dicho estudio hemos encontrado tres regímenes (*stretched exponential*, cola en ley de potencias y saturación) con comportamiento distinto de la distribución, en función de las correlaciones del proceso. Hemos obtenido conclusiones similares en procesos que presentan un *crossover* de un exponente de correlación a escala pequeña a otro a escala grande. Con este estudio hemos establecido la base para el modelo que propondremos para reproducir las correlaciones de largo alcance en palabras.

II. Hemos representado las ocurrencias de una palabra en un texto mediante una secuencia binaria y hemos analizado sus correlaciones. Hemos obtenido que el grado de correlaciones de una palabra a escala grande está directamente relacionado con su relevancia, pudiendo usarse para detectar palabras clave, que hemos concluido que son las responsables de las correlaciones de largo alcance presentes en textos.

III. Hemos planteado un modelo que, basado en las correlaciones a larga escala de una palabra, en su frecuencia de aparición y con el uso de un factor de repulsión reproduce el comportamiento de las palabras relevantes: su patrón de apariciones formando *clusters*, la forma funcional de su distribución espacial y el *crossover* a un exponente de correlación mayor a larga escala en la secuencia binaria que representa sus apariciones.

Aplicaciones y líneas futuras

La medida de *clustering* $K(N,n)$ que hemos definido se puede aplicar a otro ejemplo paradigmático de secuencias simbólicas, las proteínas, para detectar el *clustering* de aminoácidos. Una proteína es una secuencia de 20 aminoácidos diferentes que se puede ver

como un texto en el que cada palabra del vocabulario es un aminoácido. El conocimiento de la distribución exacta obtenida es importante en este caso debido a que la longitud promedio de las proteínas es relativamente corta y también lo son las frecuencias de aminoácidos individuales. Para más información y detalles al respecto, se puede consultar [Carpena et al., 2016a].

Por otro lado, el análisis numérico sistemático de las distribuciones de las longitudes entre pasos consecutivos por 0 para procesos invariantes de escala con correlaciones en ley de potencias, tiene implicaciones en el estudio de sistemas correlacionados-desordenados. Para más detalles al respecto, se puede consultar [Carretero-Campos et al., 2012].

Como líneas futuras de continuación de esta investigación podemos mencionar:

I. Explorar distintas posibilidades para la automatización de la estimación de los parámetros del modelo planteado en el capítulo 6.

II. Estudiar la distribución de los 'segundos vecinos', es decir, de las distancias entre la aparición de una palabra y la segunda vez que aparece a partir de ese momento. Este análisis permitiría determinar si la interacción entre apariciones de una misma palabra ocurre sólo a 'primeros vecinos', es decir, si el autor del texto 'olvida' más allá de la anterior aparición de una palabra dada.

III. Investigar si los textos generados por inteligencia artificial presentan las mismas propiedades estadísticas que los generados por humanos. Sería interesante, por ejemplo, comparar la distribución de valores de la medida de *clustering* K para todas las palabras del vocabulario de un texto real escrito por una persona y de un texto generado con IA de la misma longitud, o mejor todavía, de un conjunto de textos de ambos tipos. La distribución de valores de K recoge tanto la frecuencia como la estructura de las palabras en el texto, y puede servir para encontrar similitudes y diferencias entre ambos tipos de texto.

IV. Plantear un modelo físico de partículas en un medio lineal discreto que interaccionan con las mismas reglas que las palabras (relevantes): fuerzas repulsivas a distancias cortas y atractivas a distancias largas. Estudiar qué fuerzas habría que considerar para conseguir que la distribución de las partículas en el modelo unidimensional se parezca a la observada en las palabras. Esto permitiría modelar un texto como un auténtico modelo físico en el que establecer analogías interesantes.

Apéndice A

A lo largo de esta memoria los métodos planteados para detectar palabras relevantes se han aplicado fundamentalmente al texto literario *The Origin of Species*, por ser un texto de referencia en este contexto, que se usa para evaluar y comparar distintas aproximaciones. Para mostrar que en realidad los resultados obtenidos tienen carácter general, en este apéndice detallamos la aplicación de dichos métodos a otros dos textos: por un lado, *A Brief History of Time*, de Stephen Hawking y, por otro, esta tesis doctoral. Nótese que se trata de un texto en inglés y otro en español, con el objetivo de mostrar cómo las técnicas que proponemos son efectivas independientemente del idioma considerado.

Nos centraremos en aplicar la medida de *clustering* $K(N, n)$ desarrollada en la primera parte de esta memoria que nos permitía, no sólo obtener las palabras más relevantes de un texto, sino también hacer una distinción entre palabras clave genéricas y específicas (véase capítulo 3). Por otro lado, también mostraremos los resultados obtenidos si queremos clasificar las palabras a partir del exponente de correlación a escala grande α_2 de la secuencia binaria que representa las apariciones de cada palabra, como hicimos en la segunda parte de esta memoria (véase capítulo 6).

A Brief History of Time

Una vez eliminado prólogo, glosario y agradecimientos, el libro *A Brief History of Time*, de Stephen Hawking, tiene una longitud de $N = 61016$ palabras y un vocabulario formado por 4262 palabras distintas. Si computamos la medida de *clustering* $K(N, n)$ (definida en la sección 3.3) obtenemos el *ranking* de relevancia que se muestra en la tabla A.1. Observamos cómo las 10 primeras palabras nos dan una idea clara de los contenidos de los que trata el texto. Cabe mencionar que en el glosario que aparece al final del libro encontramos, entre otros, los términos 'black hole', 'imaginary time', 'string theory' y 'elementary particles' en los que aparecen las 5 primeras palabras del *ranking* obtenido.

Recordemos que en el capítulo 3 (sección 3.3.3) se estudiaron los valores extremos de $K(N, n)$ y se estableció el valor máximo de *clustering*, $K_{\max}(N, n)$, y la cota inferior de

word	K
black	4.676
imaginary	4.101
string	3.998
hole	3.961
particles	3.850
thermodynamic	3.788
disorder	3.616
newton	3.533
inflationary	3.396
holes	3.371

Tabla A.1: *Ranking* de las 10 palabras más relevantes extraídas del libro *A Brief History of Time*, de Stephen Hawking, mediante la medida de *clustering* $K(N, n)$.

clustering extremo, $K_{\mathrm{b}}(N, n)$. Aplicando estos resultados podemos observar (véase figura A.1) que, aunque las palabras 'particles' (posición 5 en el *ranking* de K) y 'thermodynamic' (posición 6 en el *ranking* de K) tienen valores muy próximos de K, sólo una de ellas supera la cota de *clustering* extremo. En el caso de 'thermodynamic', su valor de K es superior a $K_{\mathrm{b}}(N, n)$, lo que nos permite clasificarla como palabra clave específica, ya que sólo se supera la cota cuando la palabra se usa en un contexto específico del texto. Podemos comprobar que esta es la situación de 'thermodynamic' observando sus posiciones a lo largo del texto representadas en la figura A.2. Todas sus apariciones ($n = 22$) están contenidas en el capítulo 9, titulado *The Arrow of Time*, haciendo referencia a 'the thermodynamic arrow of time'. Sin embargo, como vemos en las figuras A.1 y A.2, la palabra 'particles' está situada bastante por debajo de la cota de *clustering* extremo y, aunque está *clusterizada*, se usa en diferentes contextos. 'Particles' tiene, por tanto, las características de una palabra clave genérica. Estos dos ejemplos muestran la validez de los umbrales determinados en el capítulo 3 para distinguir entre palabras clave genéricas y específicas.

En cuanto a la conexión entre las correlaciones de largo alcance de una palabra y su relevancia en un texto vamos a considerar ahora, como hicimos en el capítulo 6, la secuencia binaria que representa las apariciones a lo largo del texto de cada palabra con frecuencia mayor o igual a 100 (para evitar resultados afectados por una baja estadística). Recordemos que llamábamos α_2 al exponente de correlación a escalas grandes que se obtenía al aplicar el método DFA a dichas secuencias. En la tabla A.2 observamos en *A Brief History of Time* el mismo vínculo entre correlaciones de largo alcance y relevancia que ya mostramos en el capítulo 6 para *The Origin of Species* (tabla 6.1). Observamos que las primeras palabras aparecen en términos contenidos en el glosario del libro tales como 'strong force', 'neutron star', 'conservation of energy', 'light cone' y 'space-time',

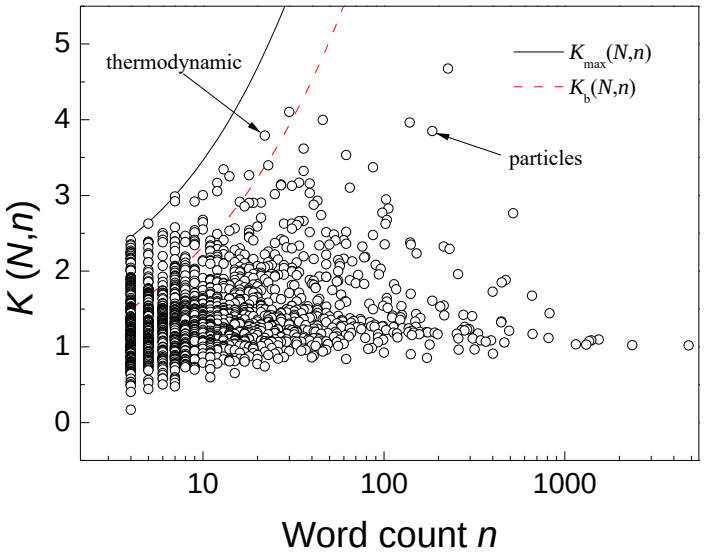

Figura A.1: Valores de *clustering* $K(N,n)$ para las palabras del vocabulario del libro *A Brief History of Time* como función de la frecuencia n. Incluimos solo palabras con $n > 3$. Las líneas corresponden al valor máximo de *clustering* $K_{\max}(N,n)$ y a la cota inferior de *clustering* extremo, $K_{\rm b}(N,n)$.

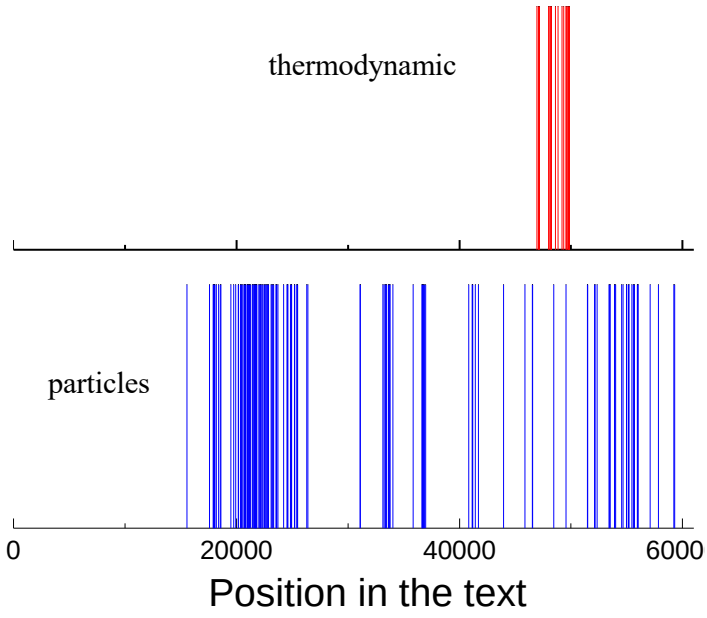

Figura A.2: Posiciones de las palabras 'thermodynamic' ($n = 22$, panel superior) y 'particles' ($n = 185$, panel inferior) en el libro *A Brief History of Time* en el que $N = 61016$.

a diferencia de las últimas palabras, claramente no informativas sobre el contenido del texto (como, por ejemplo, 'so', 'but' y 'this'). Estos resultados siguen sugiriendo, como ya comentamos, que el grado de correlaciones de largo alcance de una palabra a escalas grandes está directamente relacionado con su relevancia.

word	α_2	word	α_2
force	0.902	as	0.501
stars	0.840	very	0.501
energy	0.839	some	0.497
light	0.812	all	0.491
star	0.804	because	0.486
i	0.799	so	0.484
universe	0.791	other	0.481
space	0.787	however	0.476
particle	0.776	but	0.471
particles	0.749	this	0.465

Tabla A.2: Las primeras 10 palabras y las últimas 10 palabras extraídas del libro *A Brief History of Time* mediante el exponente de correlación a escalas grandes α_2.

Sobre el comportamiento complejo de las palabras relevantes en textos: heterogeneidad espacial y correlaciones de largo alcance

Realizamos ahora un análisis similar de esta tesis doctoral, concretamente el texto formado por el resumen, introducción, capítulos 1 al 6 y conclusiones (eliminamos índices y bibliografía). Dicho texto tiene una longitud de $N = 28710$ palabras y un vocabulario formado por 2866 palabras distintas.

En la tabla A.3 mostramos las 10 primeras palabras del *ranking* asociado a la medida de *clustering* $K(N, n)$. El lector podrá juzgar que se trata de palabras clave en el contenido desarrollado en esta memoria tales como 'correlaciones', 'alcance', 'palabra', 'secuencia' y 'binaria'. Podemos de nuevo fijarnos en un par de palabras con valores altos de K que podamos distinguir como palabras clave específicas o genéricas, según superen o no la cota de *clustering* extremo $K_{\mathrm{b}}(N, n)$. Observamos que la palabra 'régimen' está por encima de dicha cota y se trataría de una palabra clave específica ya que sus apariciones están contenidas, casi en su totalidad, en el capítulo 5 de esta memoria (véase figura A.4). En dicho capítulo se habla de la distinción entre 'régimen en *stretched exponential*', 'régimen de cola en ley de potencias' y 'régimen de saturación' en las distribuciones de los tiempos de primer paso. Sin embargo, la palabra 'secuencia', por debajo de la cota de *clustering* extremo (figura A.3), aparece en distintos contextos de esta memoria (como se observa en la figura A.4) y diremos que es una palabra clave genérica de esta tesis doctoral.

word	K
correlaciones	6.545
dfa	4.410
secuencia	4.348
exponente	4.245
alcance	3.958
binaria	3.858
régimen	3.614
glosario	3.613
escala	3.583
palabra	3.534

Tabla A.3: *Ranking* de las 10 palabras más relevantes extraídas de esta tesis doctoral (resumen, introducción, capítulos 1 al 6 y conclusiones) mediante la medida de *clustering* $K(N, n)$.

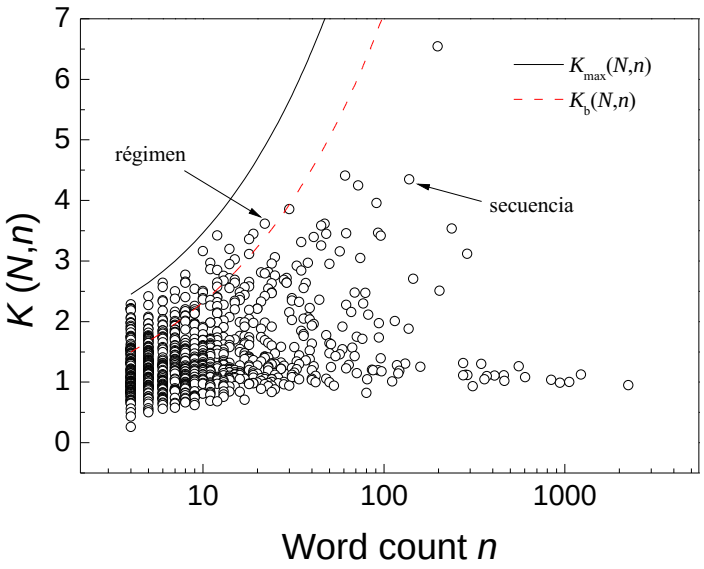

Figura A.3: Valores de *clustering* $K(N, n)$ para las palabras del vocabulario de esta tesis doctoral (resumen, introducción, capítulos 1 al 6 y conclusiones) como función de la frecuencia n. Incluimos solo palabras con $n > 3$. Las líneas corresponden al valor máximo de *clustering* $K_{\max}(N, n)$ y a la cota inferior de *clustering* extremo, $K_{\mathrm{b}}(N, n)$.

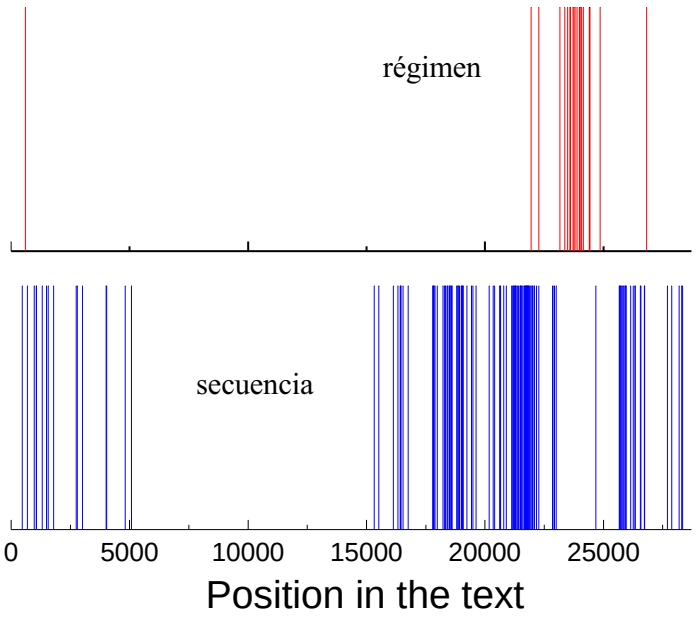

Figura A.4: Posiciones de las palabras 'régimen' ($n = 22$, panel superior) y 'secuencia' ($n = 138$, panel inferior) en esta tesis doctoral (resumen, introducción, capítulos 1 al 6 y conclusiones) en la que $N = 28710$.

word	α_2	word	α_2
correlaciones	0.721	que	0.513
secuencia	0.696	con	0.508
palabra	0.675	el	0.502
distribución	0.667	a	0.502
palabras	0.606	como	0.486

Tabla A.4: Las primeras 5 palabras y las últimas 5 palabras extraídas de esta tesis doctoral (resumen, introducción, capítulos 1 al 6 y conclusiones) mediante el exponente de correlación a escalas grandes α_2.

Por último, mostramos en la tabla A.4 las primeras y últimas cinco palabras ordenadas por el exponente de correlación a escala grande α_2. Podemos observar el vínculo entre correlaciones de largo alcance y relevancia, estando de nuevo 'correlaciones', 'secuencia' y 'palabra' entre las primeras palabras del *ranking*. Nótese que, al ser este texto más corto, el número de palabras con frecuencia al menos 100 es bastante más reducido.

Consideramos que los resultados mostrados en este Apéndice, correspondientes a dos textos tan distintos como *A Brief History of Time* y esta memoria, ilustran que las técnicas desarrolladas en la misma funcionan adecuadamente en textos con diferentes estilos y en distintos idiomas, y tienen por tanto validez general.

Bibliografía

Altmann, E. G., Cristadoro, G., and Esposti, M. D. (2012). On the origin of long-range correlations in texts. *Proceedings of the National Academy of Sciences*, 109(29):11582–11587.

Altmann, E. G., Pierrehumbert, J. B., and Motter, A. E. (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLOS ONE*, 4(11):1–7.

Alvarez-Lacalle, E., Dorow, B., Eckmann, J.-P., and Moses, E. (2006). Hierarchical structures induce long-range dynamical correlations in written texts. *Proceedings of the National Academy of Sciences*, 103(21):7956–7961.

Amancio, D. R. (2015). Authorship recognition via fluctuation analysis of network topology and word intermittency. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(3):P03005.

Amancio, D. R., Altmann, E. G., Rybski, D., Oliveira, Jr, O. N., and Costa, L. d. F. (2013). Probing the statistical properties of unknown texts: Application to the voynich manuscript. *PLOS ONE*, 8(7):1–10.

Apostolov, S. S., Izrailev, F. M., Makarov, N. M., Mayzelis, Z. A., Melnyk, S. S., and Usatenko, O. V. (2008). The signum function method for the generation of correlated dichotomic chains. *Journal of Physics A: Mathematical and Theoretical*, 41(17):175101.

Ashkenazy, Y., Ivanov, P. C., Havlin, S., Peng, C.-K., Goldberger, A. L., and Stanley, H. E. (2001). Magnitude and sign correlations in heartbeat fluctuations. *Phys. Rev. Lett.*, 86:1900–1903.

Ashkenazy, Y., M. Hausdorff, J., Ch. Ivanov, P., and Eugene Stanley, H. (2002). A stochastic model of human gait dynamics. *Physica A: Statistical Mechanics and its Applications*, 316(1):662–670.

Aslam, J. A., Yilmaz, E., and Pavlu, V. (2005). The maximum entropy method for analyzing retrieval measures. In *Proceedings of the 28th Annual International ACM*

SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05, page 27–34, New York, NY, USA. Association for Computing Machinery.

Beran, J. (1994). *Statistics for Long-Memory Processes*. Chapman & Hall/CRC.

Berger, A. and Lafferty, J. (1999). Information retrieval as statistical translation. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 222–229, New York, NY, USA. Association for Computing Machinery.

Bhan, J., Kim, S., Kim, J., Kwon, Y., il Yang, S., and Lee, K. (2006). Long-range correlations in korean literary corpora. *Chaos, Solitons & Fractals*, 29(1):69–81.

Blázquez, M., Anguiano, M., de Saavedra, F. A., Lallena, A. M., and Carpena, P. (2009). Study of the human postural control system during quiet standing using detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications*, 388(9):1857–1866.

Bookstein, A. and Swanson, D. R. (1974). Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 25(5):312–316.

Bookstein, A. and Swanson, D. R. (1975). A decision theoretic foundation for indexing. *Journal of the American Society for Information Science*, 26(1):45–50.

Brown, R. (1828). Xxvii. a brief account of microscopical observations made in the months of june, july and august 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *The Philosophical Magazine*, 4(21):161–173.

Buldyrev, S. V. (2006). *Power Law Correlations in DNA Sequences*, pages 123–164. Springer US, Boston, MA.

Bunde, A., Eichner, J. F., Kantelhardt, J. W., and Havlin, S. (2005). Long-term memory: A natural mechanism for the clustering of extreme events and anomalous residual times in climate records. *Phys. Rev. Lett.*, 94:048701.

Bunde, A. and Havlin, S., editors (1995). *Fractals in Science*. Springer-Verlag.

Cai, S.-M., Fu, Z.-Q., Zhou, T., Gu, J., and Zhou, P.-L. (2009). Scaling and memory in recurrence intervals of internet traffic. *Europhysics Letters*, 87(6):68001.

Carpena, P., Bernaola-Galván, P., Coronado, A. V., Hackenberg, M., and Oliver, J. L. (2007). Identifying characteristic scales in the human genome. *Phys. Rev. E*, 75:032903.

Carpena, P., Bernaola-Galván, P., Hackenberg, M., Coronado, A. V., and Oliver, J. L. (2009). Level statistics of words: Finding keywords in literary texts and symbolic sequences. *Phys. Rev. E*, 79:035102.

Carpena, P., Bernaola-Galván, P., and Ivanov, P. C. (2004). New class of level statistics in correlated disordered chains. *Phys. Rev. Lett.*, 93:176804.

Carpena, P., Bernaola-Galván, P., Ivanov, P. C., and Stanley, H. E. (2002). Metal–insulator transition in chains with correlated disorder. *Nature*, 418(6901):955–959.

Carpena, P., Bernaola-Galván, P. A., Carretero-Campos, C., and Coronado, A. V. (2016a). Probability distribution of intersymbol distances in random symbolic sequences: Applications to improving detection of keywords in texts and of amino acid clustering in proteins. *Phys. Rev. E*, 94:052302.

Carpena, P., Coronado, A. V., Carretero-Campos, C., Bernaola-Galván, P., and Ivanov, P. C. (2016b). First-passage time properties of correlated time series with scale-invariant behavior and with crossovers in the scaling. In Rojas, I. and Pomares, H., editors, *Time Series Analysis and Forecasting*, pages 89–102. Springer International Publishing.

Carpena, P., Gómez-Extremera, M., and Bernaola-Galván, P. A. (2022). On the validity of detrended fluctuation analysis at short scales. *Entropy*, 24(1).

Carpena, P., Gómez-Extremera, M., Carretero-Campos, C., Bernaola-Galván, P., and Coronado, A. V. (2017). Spurious results of fluctuation analysis techniques in magnitude and sign correlations. *Entropy*, 19(6).

Carpena, P., Oliver, J. L., Hackenberg, M., Coronado, A. V., Barturen, G., and Bernaola-Galván, P. (2011). High-level organization of isochores into gigantic superstructures in the human genome. *Phys. Rev. E*, 83:031908.

Carretero-Campos, C., Bernaola-Galván, P., Ivanov, P. C., and Carpena, P. (2012). Phase transitions in the first-passage time of scale-invariant correlated processes. *Phys. Rev. E*, 85:011139.

Carretero-Campos, C., Bernaola-Galván, P., Coronado, A., and Carpena, P. (2013). Improving statistical keyword detection in short texts: Entropic and clustering approaches. *Physica A: Statistical Mechanics and its Applications*, 392(6):1481–1492.

Condamin, S., Bénichou, O., Tejedor, V., Voituriez, R., and Klafter, J. (2007). First-passage times in complex scale-invariant media. *Nature*, 450(7166):77–80.

Coronado, A. V. and Carpena, P. (2005). Size effects on correlation measures. *Journal of Biological Physics*, 31(1):121–133.

Delignieres, D., Ramdani, S., Lemoine, L., Torre, K., Fortes, M., and Ninot, G. (2006). Fractal analyses for 'short' time series: A re-assessment of classical methods. *Journal of Mathematical Psychology*, 50(6):525–544.

Ding, M. and Yang, W. (1995). Distribution of the first return time in fractional brownian motion and its application to the study of on-off intermittency. *Phys. Rev. E*, 52:207–213.

Eliazar, I. and Klafter, J. (2009). Statistical resilience of random populations to random perturbations. *Phys. Rev. E*, 79:011103.

Goh, K.-I. and Barabási, A.-L. (2008). Burstiness and memory in complex systems. *Europhysics Letters*, 81(4):48002.

Govindan, R., Vjushin, D., Brenner, S., Bunde, A., Havlin, S., and Schellnhuber, H.-J. (2001). Long-range correlations and trends in global climate models: Comparison with real data. *Physica A: Statistical Mechanics and its Applications*, 294(1):239–248.

Guo, F., Yang, D., Yang, Z., Zhao, Z.-D., and Zhou, T. (2017). Bounds of memory strength for power-law series. *Phys. Rev. E*, 95:052314.

Hackenberg, M., Barturen, G., Carpena, P., Luque-Escamilla, P. L., Previti, C., and Oliver, J. L. (2010). Prediction of cpg-island function: Cpg clustering vs. sliding-window methods. *BMC Genomics*, 11(1):327.

Hackenberg, M., Carpena, P., Bernaola-Galván, P., Barturen, G., Alganza, Á. M., and Oliver, J. L. (2011). Wordcluster: detecting clusters of DNA words and genomic elements. *Algorithms for Molecular Biology*, 6(1):2.

Hackenberg, M., Rueda, A., Carpena, P., Bernaola-Galván, P., Barturen, G., and Oliver, J. L. (2012). Clustering of DNA words and biological function: A proof of principle. *Journal of Theoretical Biology*, 297:127–136.

Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. MIT Press.

Harter, S. P. (1975a). A probabilistic approach to automatic keyword indexing. part i. on the distribution of specialty words in a technical literature. *Journal of the American Society for Information Science*, 26(4):197–206.

Harter, S. P. (1975b). A probabilistic approach to automatic keyword indexing. part ii: An algorithm for probabilistic indexing. *Journal of the American Society for Informaiton Science*, 26(4):280–289.

Hausdorff, J. M., Ashkenazy, Y., Peng, C.-K., Ivanov, P. C., Stanley, H., and Goldberger, A. L. (2001). When human walking becomes random walking: fractal analysis and modeling of gait rhythm fluctuations. *Physica A: Statistical Mechanics and its Applications*, 302(1):138–147. Proc. Int. Workshop on Frontiers in the Physics of Complex Systems.

Herrera, J. P. and Pury, P. A. (2008). Statistical keyword detection in literary corpora. *The European Physical Journal B*, 63(1):135–146.

Höll, M. and Kantz, H. (2015). The relationship between the detrendend fluctuation analysis and the autocorrelation function of a signal. *The European Physical Journal B*, 88(12):327.

Hu, K., Ivanov, P. C., Chen, Z., Carpena, P., and Eugene Stanley, H. (2001). Effect of trends on detrended fluctuation analysis. *Phys. Rev. E*, 64:011114.

Hu, K., Ivanov, P. C., Chen, Z., Hilton, M. F., Stanley, H., and Shea, S. A. (2004). Non-random fluctuations and multi-scale dynamics regulation of human activity. *Physica A: Statistical Mechanics and its Applications*, 337(1):307–318.

Ivanov, P. C., Amaral, L. A. N., Goldberger, A. L., Havlin, S., Rosenblum, M. G., Struzik, Z. R., and Stanley, H. E. (1999). Multifractality in human heartbeat dynamics. *Nature*, 399(6735):461–465.

Ivanov, P. C., Rosenblum, M. G., Peng, C.-K., Mietus, J., Havlin, S., Stanley, H. E., and Goldberger, A. L. (1996). Scaling behaviour of heartbeat intervals obtained by wavelet-based time-series analysis. *Nature*, 383(6598):323–327.

Ivanov, P. C., Yuen, A., Podobnik, B., and Lee, Y. (2004). Common scaling patterns in intertrade times of u. s. stocks. *Phys. Rev. E*, 69:056107.

Jennings, H. D., Ivanov, P. C., M. Martins, A. d., da Silva, P., and Viswanathan, G. (2004). Variance fluctuations in nonstationary time series: a comparative study of music genres. *Physica A: Statistical Mechanics and its Applications*, 336(3):585–594.

Kalimeri, M., Papadimitriou, C., Balasis, G., and Eftaxias, K. (2008). Dynamical complexity detection in pre-seismic emissions using nonadditive Tsallis entropy. *Physica A: Statistical Mechanics and its Applications*, 387(5):1161–1172.

Kalra, D. S. and Santhanam, M. S. (2021). Inferring long memory using extreme events. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(11):113131.

Kantelhardt, J. W., Havlin, S., and Ivanov, P. C. (2003). Modeling transient correlations in heartbeat dynamics during sleep. *Europhysics Letters*, 62(2):147.

Kantelhardt, J. W., Koscielny-Bunde, E., Rego, H. H., Havlin, S., and Bunde, A. (2001). Detecting long-range correlations with detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications*, 295(3):441–454.

Khoury, M., Lacasta, A. M., Sancho, J. M., and Lindenberg, K. (2011). Weak disorder: Anomalous transport and diffusion are normal yet again. *Phys. Rev. Lett.*, 106:090602.

Koscielny-Bunde, E., Bunde, A., Havlin, S., Roman, H. E., Goldreich, Y., and Schellnhuber, H.-J. (1998). Indication of a universal persistence law governing atmospheric variability. *Phys. Rev. Lett.*, 81:729–732.

Lee, C.-Y. (2009). Characteristics of the volatility in the korea composite stock price index. *Physica A: Statistical Mechanics and its Applications*, 388(18):3837–3850.

Leland, W., Taqqu, M., Willinger, W., and Wilson, D. (1994). On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15.

Li, W. (2002). Zipf's law everywhere. *Glottometrics*, 5:14–21.

Lowen and Teich (2005). *Fractal-Based Point Processes*. Wiley.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Ma, Q. D. Y., Bartsch, R. P., Bernaola-Galván, P., Yoneyama, M., and Ivanov, P. C. (2010). Effect of extreme data loss on long-range correlated and anticorrelated signals quantified by detrended fluctuation analysis. *Phys. Rev. E*, 81:031101.

Makse, H. A., Havlin, S., Schwartz, M., and Stanley, H. E. (1996). Method for generating long-range correlations for large systems. *Phys. Rev. E*, 53:5445–5449.

Mandelbrot, B. B. and Van Ness, J. W. (1968). Fractional brownian motions, fractional noises and applications. *SIAM Review*, 10(4):422–437.

Mehri, A., Agahi, H., and Mehri-Dehnavi, H. (2019). A novel word ranking method based on distorted entropy. *Physica A: Statistical Mechanics and its Applications*, 521:484–492.

Mehri, A. and Darooneh, A. H. (2011). The role of entropy in word ranking. *Physica A: Statistical Mechanics and its Applications*, 390(18):3157–3163.

Mehri, A., Jamaati, M., and Mehri, H. (2015). Word ranking in a single document by Jensen–Shannon divergence. *Physics Letters A*, 379(28):1627–1632.

Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing order into text. In Lin, D. and Wu, D., editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Molchan, G. M. (1999). Maximum of a fractional brownian motion: Probabilities of small values. *Communications in Mathematical Physics*, 205(1):97–111.

Montemurro, M. A. and Pury, P. A. (2002). Long-range fractal correlations in literary corpora. *Fractals*, 10(04):451–461.

Montemurro, M. A. and Zanette, D. H. (2002). Entropic analysis of the role of words in literary texts. *Advances in Complex Systems*, 05(01):7–17.

Montemurro, M. A. and Zanette, D. H. (2010). Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems*, 13(02):135–153.

Moreno-Sánchez, I., Font-Clos, F., and Corral, Á. (2016). Large-scale analysis of zipf's law in english texts. *PLOS ONE*, 11(1):1–19.

Newell, G. F. and Rosenblatt, M. (1962). Zero Crossing Probabilities for Gaussian Stationary Processes. *The Annals of Mathematical Statistics*, 33(4):1306 – 1313.

Ogura, H., Amano, H., and Kondo, M. (2019). Measuring dynamic correlations of words in written texts with an autocorrelation function. *Journal of Data Analysis and Information Processing*, 7(2):46–73.

Oliver, J. L., Carpena, P., Román-Roldán, R., Mata-Balaguer, T., Mejıías-Romero, A., Hackenberg, M., and Bernaola-Galván, P. (2002). Isochore chromosome maps of the human genome. *Gene*, 300(1):117–127. Natural Selection and the Neutral Theory.

Ortuño, M., Carpena, P., Bernaola-Galván, P., Muñoz, E., and Somoza, A. M. (2002). Keyword detection in natural languages and DNA. *Europhysics Letters*, 57(5):759.

Orun, M. and Koçak, K. (2009). Applicatıon of detrended fluctuation analysis to temperature data from turkey. *International Journal of Climatology*, 29(14):2130–2136.

Peng, C.-K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M., and Stanley, H. E. (1992). Long-range correlations in nucleotide sequences. *Nature*, 356(6365):168–170.

Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., and Goldberger, A. L. (1994). Mosaic organization of DNA nucleotides. *Phys. Rev. E*, 49:1685–1689.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C*. Cambridge University Press, Cambridge, USA, second edition.

Priestley, M. B. (1981). *Spectral analysis of time series*. Academic Press, London.

Rangarajan, G. and Ding, M. (2000). First passage time distribution for anomalous diffusion. *Physics Letters A*, 273(5):322–330.

Reed, G. F., Lynn, F., and Meade, B. D. (2002). Use of coefficient of variation in assessing variability of quantitative assays. *Clinical and Vaccine Immunology*, 9(6):1235–1239.

Reyes-Ramírez, I. and Guzmán-Vargas, L. (2010). Scaling properties of excursions in heartbeat dynamics. *Europhysics Letters*, 89(3):38008.

Robertson, S. E., Kanoulas, E., and Yilmaz, E. (2010). Extending average precision to graded relevance judgments. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, page 603–610, New York, NY, USA. Association for Computing Machinery.

Robinson, H. P. C. and Harsch, A. (2002). Stages of spike time variability during neuronal responses to transient inputs. *Phys. Rev. E*, 66:061902.

Rose, S., Engel, D., Cramer, N., and Cowley, W. (2010). *Automatic Keyword Extraction from Individual Documents*, chapter 1, pages 1–20. John Wiley & Sons, Ltd.

Şahin, G., Erentürk, M., and Hacinliyan, A. (2009). Detrended fluctuation analysis in natural languages using non-corpus parametrization. *Chaos, Solitons & Fractals*, 41(1):198–205.

Salton, G. and McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., USA.

Sapir, N., Karasik, R., Havlin, S., Simon, E., and Hausdorff, J. M. (2003). Detecting scaling in the period dynamics of multimodal signals: Application to parkinsonian tremor. *Phys. Rev. E*, 67:031903.

Schindler, M., Talkner, P., and Hänggi, P. (2004). Firing time statistics for driven neuron models: Analytic expressions versus numerics. *Phys. Rev. Lett.*, 93:048102.

Shang, P., Lu, Y., and Kamae, S. (2008). Detecting long-range correlations of traffic time series with multifractal detrended fluctuation analysis. *Chaos, Solitons & Fractals*, 36(1):82–90.

Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL.

Shlesinger, M. F., Zaslavsky, G. M., and Klafter, J. (1993). Strange kinetics. *Nature*, 363(6424):31–37.

Talkner, P. and Weber, R. O. (2000). Power spectrum and detrended fluctuation analysis: Application to daily temperatures. *Phys. Rev. E*, 62:150–160.

Tanaka-Ishii, K. and Bunde, A. (2016). Long-range memory in literary texts: On the universal clustering of the rare words. *PLOS ONE*, 11(11):1–14.

Tohalino, J. A. V., Silva, T. C., and Amancio, D. R. (2023). Using citation networks to evaluate the impact of text length on keyword extraction. *PLOS ONE*, 18(11):1–17.

Usatenko, O., Melnik, S., Kroon, L., Johansson, M., Riklund, R., and Apostolov, S. (2008). Spectral analysis and synthesis of 1d dichotomous long-range correlated systems: From diffraction gratings to quantum wires. *Physica A: Statistical Mechanics and its Applications*, 387(19):4733–4739.

Varotsos, P. A., Sarlis, N. V., and Skordas, E. S. (2003a). Attempt to distinguish electric signals of a dichotomous nature. *Phys. Rev. E*, 68:031106.

Varotsos, P. A., Sarlis, N. V., and Skordas, E. S. (2003b). Long-range correlations in the electric signals that precede rupture: Further investigations. *Phys. Rev. E*, 67:021109.

Voss, R. F. (1992). Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys. Rev. Lett.*, 68:3805–3808.

Wang, F., Yamasaki, K., Havlin, S., and Stanley, H. E. (2009). Multifactor analysis of multiscaling in volatility return intervals. *Phys. Rev. E*, 79:016103.

Xu, Y., Ma, Q. D., Schmitt, D. T., Bernaola-Galván, P., and Ivanov, P. C. (2011). Effects of coarse-graining on the scaling behavior of long-range correlated and anti-correlated signals. *Physica A: Statistical Mechanics and its Applications*, 390(23):4057–4072.

Yilmaz, E. and Aslam, J. A. (2006). Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, page 102–111, New York, NY, USA. Association for Computing Machinery.

Yuan, Y., tian Zhuang, X., and Jin, X. (2009). Measuring multifractality of stock price fluctuation using multifractal detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications*, 388(11):2189–2197.

Zhou, H. and Slater, G. W. (2003). A metric to search for relevant words. *Physica A: Statistical Mechanics and its Applications*, 329(1):309–327.

Zhu, M. (2004). Recall, precision and average precision. Technical report, Department of Statistics and Actuarial Science, University of Waterloo.

Zipf, G. K. (1949). *Human Behavior and the Principle Least Effort: An Introduction to Human Ecology*. Addison-Wesley, Cambridge, MA.