ANÁLISIS DE ORGANIZACIÓN Y CORRELACIONES EN TEXTOS ESCRITOS

Lev Guzmán Vargas
lguzmanv@ipn.mx,
Daniel Aguilar Velázquez,
Sepi-Upiita, Instituto Politécnico Nacional
Bibiana Obregón Quintana,
Facultad de Ciencias, Universidad Nacional Autónoma De México

Abstract

Se presentan resultados obtenidos de la caracterización de sistemas bajo el enfoque de transformaciones a redes complejas. En particular, sobre la organización temporal de las secuencias de palabras (textos escritos en idioma inglés), vistas como series de longitudes, y también como redes complejas. La presencia de escalamiento es una característica importante en sistemas complejos, ya que refiere la ausencia de escalas dominantes y, bajo ciertas circunstancias, ligada a estados críticos donde dominan propiedades emergentes y organizativas. Nuestro proceder se centró en evaluar las condiciones marcadas por la aparición de invariancia de escala, de los distintos sistemas estudiados. Se encuentra que las correlaciones entre longitudes de palabras se acentúan cuando la distancia entre ellas es grande en comparación a cuando están cercanas. Nuestros resultados fueron validados mediante tres técnicas de análisis no lineal.

I. Introducción

Muchos sistemas físicos, sociales y biológicos pueden describirse como sistemas complejos, compuestos de un número grande de componentes, que interactúan de manera no bien definida o intermitentemente. Cuando se observa la dinámica de un sistema complejo, a veces es difícil identificar o entender algunas características propias del sistema. Una de las propiedades de los sistemas complejos que ha llamado la atención de manera significativa es la capacidad de auto-organización y emergencia. En muchos sistemas complejos la característica común es que despliegan cierto grado de organización sin que esté actuando un principio externo de organización. Recientemente, el estudio de las redes complejas ha cobrado importancia, investigadores de distintas disciplinas científicas han descubierto que muchos sistemas físicos y biológicos pueden analizarse mediante modelos de redes complejas.

Diversos sistemas pueden ser concebidos como elementos acoplados, por ejemplo, redes neuronales, sistemas químicos, redes sociales o Internet. En particular, sistemas tecnológicos de información pueden naturalmente modelarse como redes, tal es el caso de Internet y la red de páginas Web, donde los nodos son ruteadores y páginas Web; y los enlaces son cables y URL's, respectivamente [1, 2]. Una red se concibe como un conjunto de nodos o vértices y un

conjunto de enlaces que unen a los nodos. Los enlaces pueden tener una dirección privilegiada, en cuyo caso se denominan dirigidos; y en algunas situaciones también representan cierta intensidad de relación, lo que conduce a redes con peso. La caracterización de una red compleja es fundamental para entender su estructura y funcionamiento. En este sentido, recientes estudios muestran que una red compleja puede caracterizarse mediante propiedades que indican el grado de organización.

Por otro lado, una característica importante en textos escritos es la aparición de correlaciones temporales creadas por la narración de ideas o historias. Sin embargo, la evaluación directa de estas correlaciones no es factible, porque las palabras se pueden utilizar en diferentes sentidos, haciendo un análisis cuantitativo difícil. En los últimos años, diversos métodos se han utilizado para explorar la presencia de correlaciones temporales en los textos [3, 4, 5], principalmente enfocados en la longitud o en la frecuencia de las palabras. Como se sabe, el lenguaje escrito es la conformación de las propiedades gramaticales y connotaciones semánticas con el propósito de expresar ideas o información [6, 7].

Se ha reportado que la organización temporal de las secuencias de longitud de palabra de los textos escritos se puede caracterizar por la presencia de ligeras correlaciones positivas [3, 4] y con variaciones locales de los exponentes de escala relacionados con la organización temporal de los textos. Aquí, usamos el método de visibilidad (VM) para transformar un texto escrito a una red compleja. Nuestros resultados muestran que la distribución de grado resultante sigue una ley de potencia, $P(k) \sim k^{\gamma}$, la cual exhibe dos regímenes diferentes de correlaciones en las distancias cortas y largas entre las palabras. Estos resultados se complementan con la aplicación del análisis de fluctuación sin tendencia (DFA) y los cálculos de los tiempos de recurrencia.

II. Métodos y resultados

Para transformar una serie de tiempo en una red compleja, recurrimos al método de visibilidad (MV) propuesto por [8]. Este se introdujo básicamente para transformar las series de tiempo irregulares en redes, y con ello extraer información complementaria sobre la organización temporal de la secuencia original. De acuerdo con el MV, para construir la red de visibilidad correspondiente de una serie de tiempo, para cualquier par de valores de datos, $\{t_a, y_a\}$ y $\{t_b, y_b\}$, definimos un vínculo entre ellos, si no hay ningún otro elemento, $\{t_c, y_c\}$, colocado en el medio que intercepta la línea que conecta los dos valores, es decir, $\{t_c, y_c\}$ satisface, $y_c < y_b + (y_a - y_b)\,(t_b - t_c) / (t_b - t_a)$.

Por lo tanto, podemos definir un nodo para cada elemento de la serie, así la red resultante es siempre conexa, sin dirección, y es invariante bajo transformaciones de la serie [8]. Como se muestra por Lacasa et al. [9}, una serie de tiempo estocástica se puede caracterizar a través de sus distribuciones de grado, P(k), en particular por sus exponentes, $\propto$, ya que en varios casos

se presenta un comportamiento de ley de potencia, $P(k) \sim k^{\alpha}$ donde se tiene la relación, $\propto = 4 - \beta$, siendo beta el exponente espectral para el caso del movimiento browniano fraccional [9]. Para realizar nuestro análisis, se recopilaron 30 textos escritos provenientes del sitio Gutenberg Project http://www.gutenberg.org, los cuales consisten de palabras en formato de texto plano (el listado de los nombres de los textos se presenta en la Figura 3). Con esta información se generó una secuencia (serie) conformada por palabras para caracterizar los distintos textos mediante estadística. Enseguida, se determinó la longitud de todas las palabras en términos del número de letras que conforman cada una (Ver Fig. 1a).

Una vez hecha la transformación, para cada secuencia (libro) se construyeron las distribuciones de grado, las cuales describen el nivel de conectividades alcanzado por cada elemento (palabra). El objetivo de esta transformación es, determinar el nivel de correlación o interdependencia entre las longitudes de palabras ubicadas en distintas partes del texto, esto es, determinar la memoria existente entre la aparición de una palabra de longitud dada, y la aparición posterior de otra con longitud igual. Se determinó que las distribuciones de grado siguen una forma funcional del tipo ley de potencia, $P(k) \sim k^{\gamma}$, donde γ es un exponente que caracteriza las conexiones y que sigue dos regímenes (Ver Fig. 1b) [10]. Por un lado, para escalas cortas, el valor del exponente es muy cercano al observado para secuencias sin memoria, mientras que, para escalas largas, el valor del exponente revela presencia de correlaciones de largo plazo (Ver Fig. 2a). Adicionalmente, se realizaron testes estadísticos para determinar si es posible mejorar nuestros resultados juntando los datos de todas las muestras (libros). Se encontró que al juntar los datos la estadística mejoró y con ello se determinaron exponentes más estables. También, se generaron versiones aleatorias (barajadas) de los textos para destruir las correlaciones, encontrando que el exponente que caracteriza estas versiones es cercano al valor $\gamma_r \approx 2$.

En seguida, se aplicó el método análisis de fluctuación sin tendencia (DFA por sus siglas en inglés) para confirmar la presencia de correlaciones en las secuencias originales de longitud de palabras. Los valores obtenidos para los exponentes invariantes de escala son concordantes con los reportados en otros trabajos, y están en acuerdo con los obtenidos mediante la transformación de visibilidad, confirmando que, para escalas grandes, la memoria de las secuencias de longitud es del tipo de largo plazo (ver Figura 2). En la misma dirección, se evaluó el tiempo de recurrencia para distintos valores de longitudes de palabras, esto es, se contabilizó el número de palabras de longitud menor a un valor l_0 que transcurren antes de una palabra exceda este valor. El resultado de esta estadística es descrito en la Figura 3, donde se observa que los datos son bien descritos por una distribución del tipo exponencial estirada, $G(t) \sim e^{-a\tau^{b}}$), con a y b dos parámetros de ajuste. Se observa que conforme el umbral l_0 crece el valor de b decrece, indicando una presencia de memoria conforme esto ocurre.

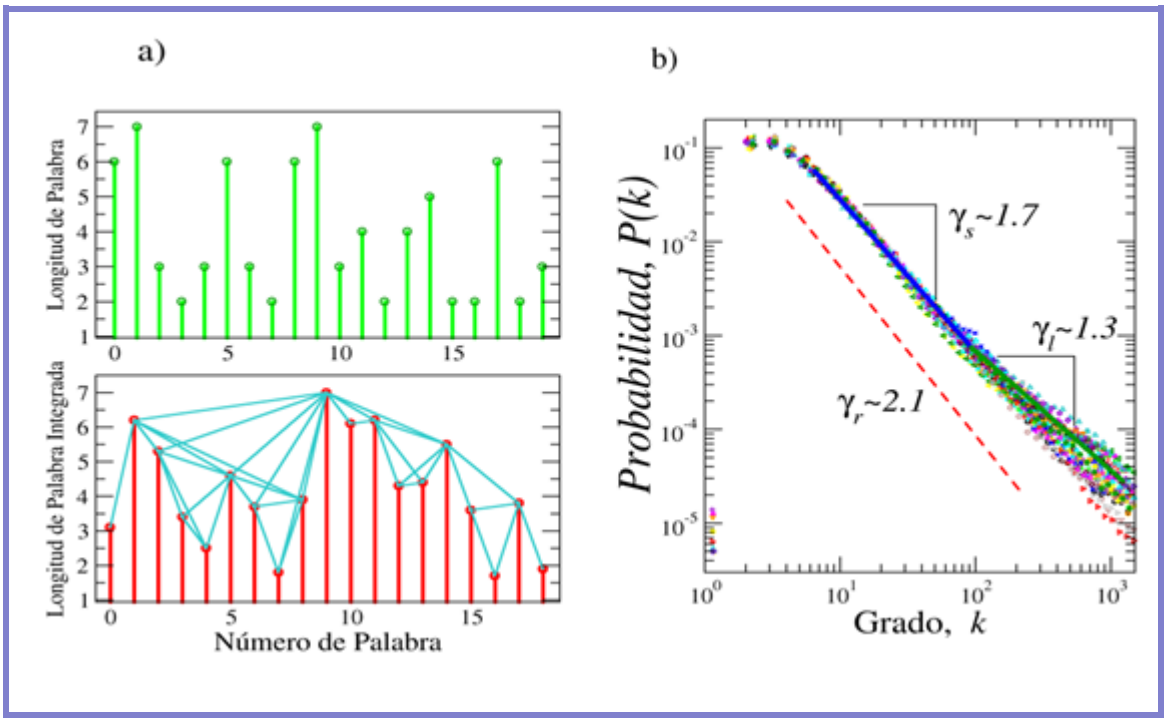
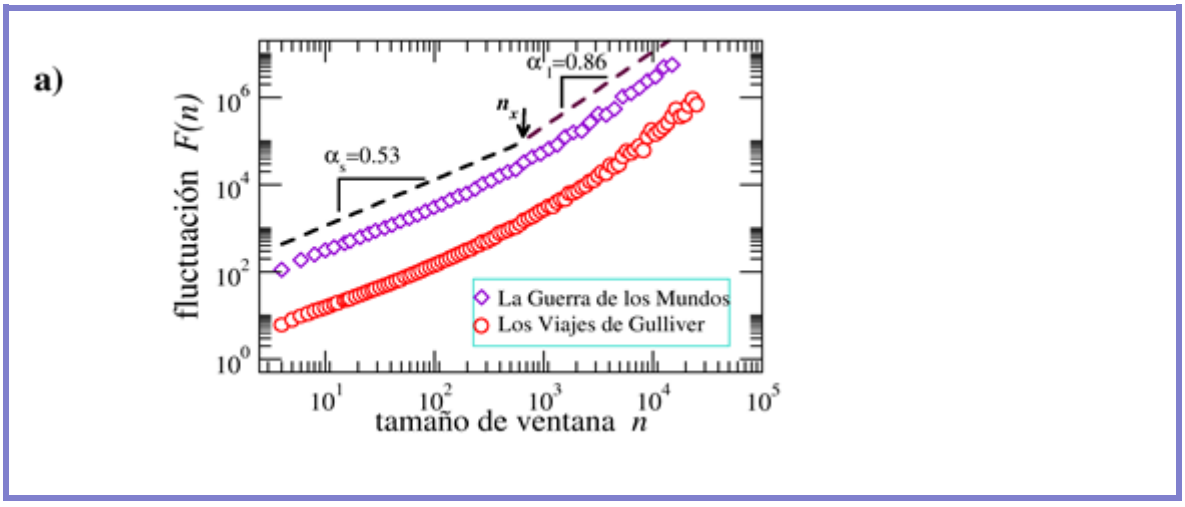


Figura 1. a) Método de Visibilidad. b) Distribución de grado



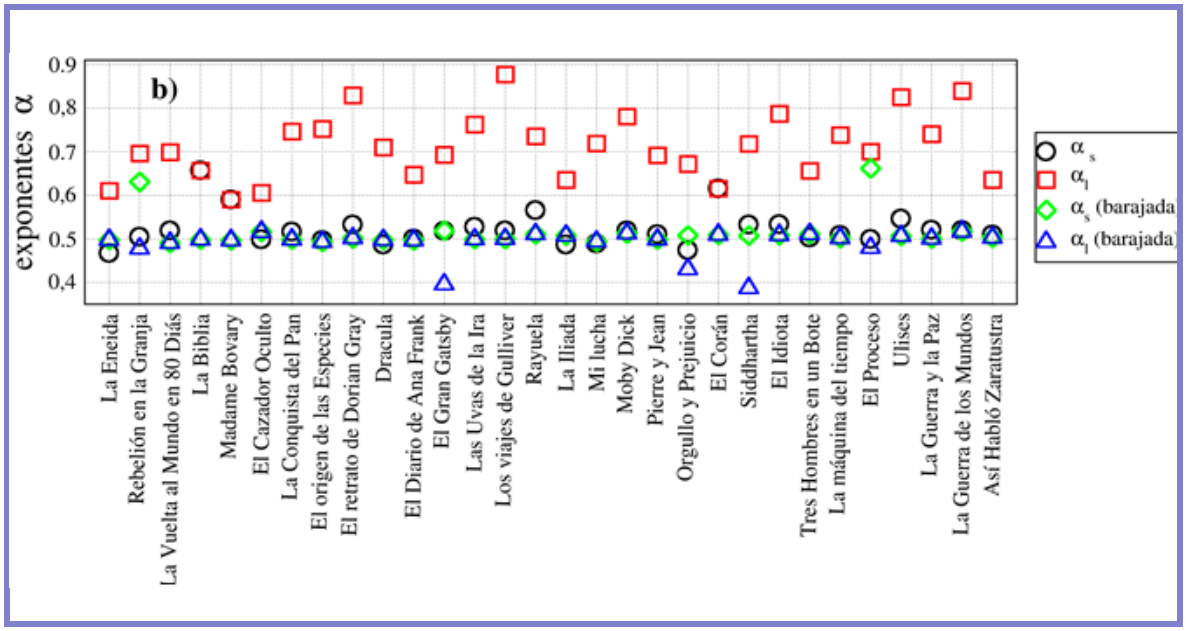


Figura 2. Análisis de fluctuación sin tendencia a) Dos casos representativos: La guerra de los mundos y Los viajes de Gulliver. b) Exponentes de correlación para los 30 libros analizados

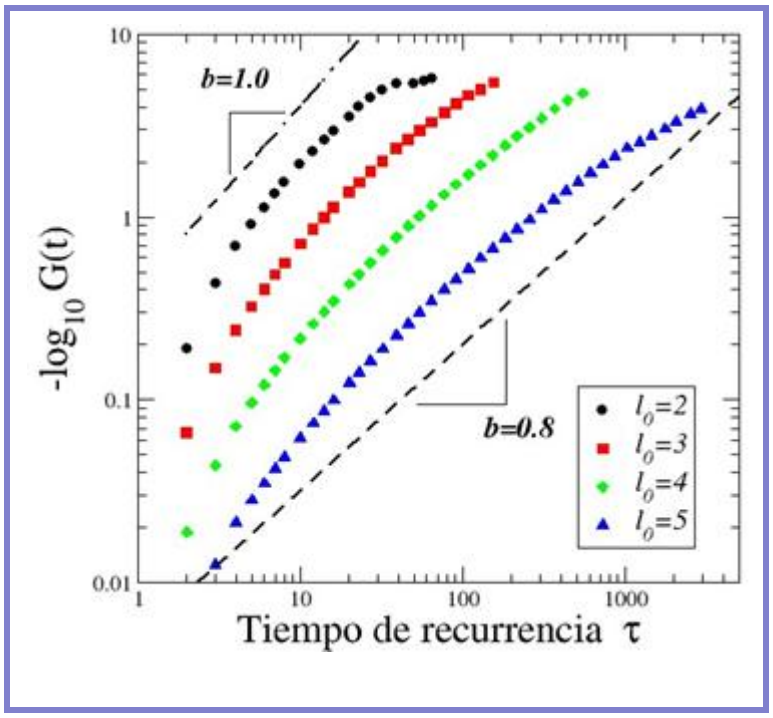


Figura 3. Distribuciones de tiempo de recurrencia en plano log-log

III. Conclusiones

Las propiedades de las distribuciones de frecuencia de palabras han sido muy estudiadas desde hace varias décadas [11], pero no se sabe mucho acerca de la correlación entre las palabras y las longitudes de palabra. Aquí, se han estudiado las propiedades de correlación de secuencias de longitud de palabra de grandes textos literarios. Nuestros resultados, basados en el método de visibilidad (VM), revelan que la distribución de grado de las redes de visibilidad integradas de longitud de palabra, puede ser descrita por una función de ley de potencia con aproximadamente dos regímenes, en cambio la distribución correspondiente de los datos de reacomodo se caracteriza por el valor del exponente γ = 2.1, como se esperaba para los datos no correlacionados [9]. Estos resultados han sido corroborados por los resultados del análisis de fluctuación sin tendencia (DFA), donde se encontró que, para las escalas pequeñas, las secuencias poseen $\propto_s \approx 0.5$, indicando que no hay correlación, mientras que, para grandes escalas, las secuencias de longitud de palabra tienen correlaciones positivas, según lo expresado por el exponente ∝_l≈0.7. Además, las distribuciones del tiempo de recurrencia también presentan una desviación con respecto al comportamiento exponencial puro como el parámetro aumenta el umbral, es decir, para pequeños valores de l_0, las longitudes de las palabras cortas dominan la dinámica sin memoria, mientras que para las grandes l_0, las altas longitudes de palabra se distribuyen con tiempos de recurrencia, que exhiben memoria (b < 1). Por lo tanto, los tres métodos encuentran que la falta de correlación en longitudes de palabra a pequeñas escalas se sustituye por una correlación positiva en longitudes de palabra con separaciones más largas entre palabras. Las estructuras lingüísticas a gran escala en estos libros son diferentes, de manera importante, en comparación con las estructuras de pequeña escala. Esperamos que estos nuevos hallazgos sirvan como base para una interpretación lingüística o semántica sobre el lenguaje o, más concretamente, sobre el lenguaje escrito.

IV. Referencias

[1] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.-U., 2006. Complex networks: Structure and dynamics. Physics Reports 424, 175-308. DOI:10.1016/j.physrep.2005.10.009

[2] Newman, M. E. J., 2003. The structure and function of complex networks. SIAM Review 45, 167-256. DOI:10.1137/S003614450342480 [3] Kosmidis, K., Kalampokis, A., Argyrakis, P., 2006. Language time series analysis. Physica A 370, 808–816. http://dx.doi.org/10.1016/j.physa.2006.02.042

[4] Ausloos, M., 2012. Generalized Hurst exponent and multifractal function of original and translated texts mapped into frequency and length time series. Physical Review E 86, 031108. DOI:10.1103/PhysRevE.86.031108.

[5] Rodriguez, E., Aguilar-Cornejo, M., Femat, R., Alvarez-Ramirez, J., 2014. Scale and time dependence of serial correlation in word-length time series of written texts. Physica A 414, 378–386. http://dx.doi.org/10.1016/j.physa.2014.07.063.

[6] Nowak, M.A., Plotkin, J.B., Jansen, V.A.A., 2000. The evolution of syntactic communication. Nature 404, 495-498. DOI:10.1038/35006635

[7] Hauser, M.D., Chomsky, N., Fitch, W.T., 2002. The faculty of language: What is it, who has it and how did it evolve?. Science 298, 1569-1579. DOI:10.1126/science.298.5598.1569.

[8] Lacasa, L.; Luque, B.; Ballesteros, F.; Luque, J.; Nuño, J.C., 2008. From time series to complex networks: The visibility graph. Proc. Natl. Acad. Sci. USA, 105, 4972–4975. DOI:10.1073/pnas.0709247105.

[9] Lacasa, L.; Luque, B.; Luque, J.; Nuño, J.C., 2009. The visibility graph: A new method for estimating the Hurst exponent of fractional Brownian motion. Europhys Letters 86, 30001. DOI:10.1209/0295-5075/86/30001.

[10] Guzmán-Vargas, L., Obregón-Quintana, B., Aguilar-Velázquez, D., Hernández-Pérez, R., Liebovitch, L., 2015. Word-length correlations and memory in large texts: a visibility network analysis. Entropy 17, 7798-7810. DOI:10.3390/e17117798. [11] Zipf, G.K., 1935. The Psycho-Biology of Language: An Introduction to Dynamic Philology; M.I.T. Press: Cambridge, MA, USA.