Se estudian seis idiomas con herramientas estadísticas para encontrar similitudes

Por Felipe Jiménez Rodríguez

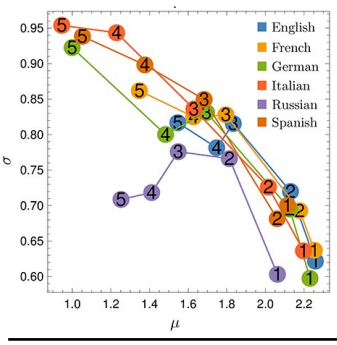
3 de agosto de 2018

"El comportamiento del lenguaje es robusto y adaptativo" dijo en entrevista Carlos Gershenson, miembro del Centro de Ciencias de la Complejidad (C3) e investigador del Instituto de Matemáticas Aplicadas y Sistemas de la UNAM. Ambas, robustez y adaptabilidad, son también propiedades que poseen los sistemas complejos y los seres vivos.

El lenguaje es adaptativo porque permite la incorporación de nuevas palabras de acuerdo al contexto —por ejemplo, con el uso masivo de las redes sociales y nuevas tecnologías aparecieron palabras como tuitear, hackear o link— o deja en desuso otras. Sin embargo, esos cambios no afectan la compresión entre personas a través de generaciones. El lenguaje sigue funcionando a pesar de los cambios, una propiedad conocida como robustez.

En un estudio publicado por la revista Frontiers in Physics en mayo de 2018, del que Gershenson es autor, investigadores de México, Estados Unidos, Finlandia y Rusia, usaron la base de datos de *N*-gramas de Google Books —una colección de palabras que pueden encontrarse en textos impresos entre 1500 y 2009—. Un *N*-grama es una combinación de n palabras: por ejemplo, pares de palabras son 2-gramas. La colección de *N*-gramas de Google Books contiene textos en 8 idiomas diferentes. Aplicando diferentes medidas y métodos a la base de datos, los investigadores encontraron ajustes de datos similares para seis diferentes lenguajes y *N*-gramas de cinco palabras.

Los científicos explican en el artículo que este tipo de análisis puede contrastarse con teorías existentes del lenguaje, invitando a los lingüistas a la colaboración transdisciplinaria. Asimismo, se espera que esta investigación ayude a mejorar los algoritmos de predicción de texto, como los usados en los dispositivos inteligentes.



RANK DYNAMICS OF WORD USAGE AT MULTIPLE SCALES

José A. Morales, Ewan Colman, Sergio Sánchez, Fernanda Sánchez-Puig, Carlos Pineda, Gerardo Iñiguez, Germinal Cocho, Jorge Flores y Carlos Gershenson.

Front. Phys. (2018) May 22, 6:46 doi. org/10.3389/fphy.2018.00045

https://www.frontiersin.org/articles/10.3389/fphy.2018.00045/full

Unidad de Comunicación y Diseño

T. (+52) 55 5622 6730 Ext. 2017 y 2018
E. comunicacion@c3.unam.mx
disenio@c3.unam.mx

Centro de Ciencias de la Complejidad (C3)

Circuito Centro Cultural s/n /frente a Universum), Cd. Universitaria, Coyoacán 04510, Cidad de México

www.c3.unam.mx
@C3UNAM
Centro de Ciencias de la Complejiadd C3

Un equipo internacional de investigadores de la Facultad de Ciencias, Instituto de Física, Centro de Ciencias de la Complejidad (C3) e Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas UNAM, México; la *University of Vienna*, Austria; *Next Games* y la *Aalto University*, Finlandia; el *Massachusetts Institute of Technology*, Cambridge, Estados Unidos; y la *ITMO University*, Rusia estudiaron los N-gramas de seis lenguas: inglés, español, francés, ruso, alemán e italiano.

La investigación busca extender un estudio de palabras individuales (1-gramas) previamente realizado en 2015. El nuevo estudio analiza ahora también pares de palabras (2-gramas), hasta frases de cinco palabras (5-gramas) para identificar qué diferencias se observan entre los cinco tipos de *N*-gramas cuando se aplican diferentes medidas y métodos estadísticos.

El periodo de tiempo en el que se enfocó el estudio fue entre 1855 y 2009, esto debido a que para algunos idiomas no había información suficiente para años anteriores a 1855. Se analizó la frecuencia de uso de N-gramas, obteniendo los rangos de uso en cada año para el intervalo de años estudiado, desde los más usados hasta los menos usados.

A los rangos de uso de *N*-gramas anuales en los seis idiomas se les aplicó diferentes medidas como la diversidad de rango, la probabilidad de cambio, la entropía de rango y la complejidad de rango, todas propuestas por los autores.

En general, el comportamiento de los datos obtenidos puede ser aproximado por una curva sigmoide (en forma de S). "Lo que nos dice esa curva es que las palabras (N-gramas) más usadas cambian menos, lo que permite al lenguaje tener robustez. Por el contrario, las palabras menos usadas varían más, lo que se traduce en adaptabilidad", explicó Gershenson.

¿Es necesario estudiar frases de n palabras? ¿No se pueden deducir propiedades generales de todos los *N*-gramas a partir de sólo una palabra (1-grama)? Para responder esto, los investigadores usaron como modelo nulo un conjunto de pares de palabras (2-gramas) sin gramática y lo compararon con datos reales.

Encontraron que el modelo nulo variaba más que los datos de los 2-gramas reales, lo que los lleva a conlcuir que se deben estudiar los diferentes tipos de N-gramas simultáneamente para comprender la estructura y el uso del lenguaje de una manera más integral a escalas múltiples.

Los investigadores destacan los patrones en común para los seis lenguajes y los cinco tipos de *N*-gramas estudiados, por ejemplo, el ajuste de los datos a la curva sigmoide. Esto es, los seis lenguajes se comportan de manera similar. ¿Qué significa esto? Aquí falta el trabajo y retroalimentación de lingüistas, filósofos del lenguaje y otros para dar un contexto a lo que dicen los datos.

Queda también la interrogante de si estos comportamientos son válidos para todos los lenguajes, cuestión que puede ser contestada a través de mayor disponibilidad de bases de datos.

Revista *Frontiers in Physics* : https://www.frontiersin.org/articles/10.3389/fphy.2018.00045/full