ESTADÍSTICA DEL SIGNIFICADO:
DETECCIÓN DEL CONTENIDO SEMÁNTICO DE UN TEXTO ESCRITO

Damián H. Zanette

El lenguaje escrito codifica la información que contiene en la forma de secuencias de palabras. A medida que escribimos un texto, la construcción de su contenido semántico –es decir, de su significado– impone condiciones sobre qué palabras aparecen, con qué frecuencia se usan, y cómo se distribuyen. ¿Sería posible, estudiando la distribución de palabras en un texto, extraer información sobre su contenido?

En primer lugar, comprobamos un hecho empírico interesante: las palabras que están más relacionadas con la temática de un texto aparecen con frecuencia relativamente alta, pero su distribución es muy heterogénea, acumulándose en partes específicas del texto. En cambio, las palabras funcionales tales como conjunciones, preposiciones y artículos, tienen altas frecuencias pero su distribución es más uniforme. La Fig. 1 ilustra este hecho para tres palabras de *On the Origin of Species*, de Ch. Darwin: *islands* e *instinct* están concentradas en ciertos capítulos, mientras que *more* se distribuye de modo más parejo.

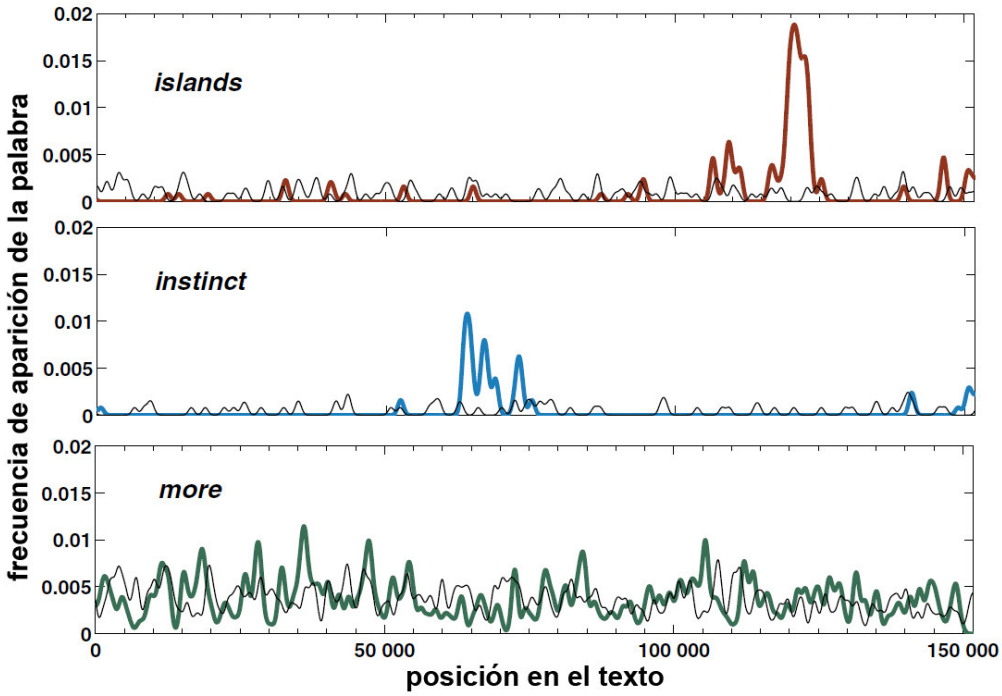

Fig. 1. Frecuencia de aparición de las palabras *islands*, *instinct* y *more* a lo largo de *On the Origin of Species*.

Para cuantificar la heterogeneidad en la distribución de una palabra calculamos cuánta información contiene –o cuál es su déficit de entropía– respecto de una distribución aleatoria. Para realizar el cálculo, dividimos el texto en partes (los "casilleros") de igual longitud. La información de la distribución ordenada respecto de la aleatoria depende

del tamaño de los casilleros (la "escala" o *graining*). Existe una escala en la que la información necesaria para ordenar las apariciones alcanza un valor máximo. Las palabras con distribución más heterogénea –que, como vimos, son generalmente las más significativas del texto– tienen mayor información en el máximo.

On the Origin of Species	*Analysis of the Mind*	*Moby Dick*
on	image	I
species	memory	whale
varieties	images	you
hybrids	word	Ahab
forms	belief	is
islands	words	ye
of	desire	Queequeg
will	sensations	thou
selection	object	me
genera	you	of
plants	past	he
seeds	knowledge	captain
sterility	box	boat
fertility	content	the
characters	consciousness	Stubb
breeds	appearances	his
groups	movements	Jonah
water	mnemic	was
the	feeling	whales
formations	proposition	my

Tabla 1. Listas de palabras de tres textos clásicos, ordenadas de acuerdo a la información de su distribución a lo largo del texto.

La Tabla 1 muestra algunas palabras ordenadas por su información máxima para *On the Origin of Species*, *Analysis of the Mind*, de B. Russell, y *Moby Dick*, de H. Melville. En los tres casos vemos que, en su gran mayoría, las palabras con mucha información son muy relevantes al contenido semántico de los respectivos textos.

Una vez calculada la información de la distribución de cada palabra como función de la escala, es posible evaluar la información total de la distribución de todas las palabras del texto. La Fig. 2 muestra el resultado para los tres libros mencionados más arriba. Para que nuestras conclusiones tengan valor estadístico, aplicamos el mismo análisis a un cuerpo de unos 5000 libros del Proyecto Gutenberg (www.gutenberg.org). En este análisis masivo, encontramos que la escala a la cual la información alcanza su máximo es muy parecida para todos los textos, situándose alrededor de unos pocos miles de palabras. Interpretamos esta escala característica notando que coincide con la longitud típica necesaria para desarrollar el núcleo de cualquier línea argumental. Por lo que sabemos, esta es la primera vez en que se detecta cuantitativamente este tipo de escala, asociada al contenido semántico de un texto e intermedia entre las escalas relevantes a la gramática y la longitud total de un libro.

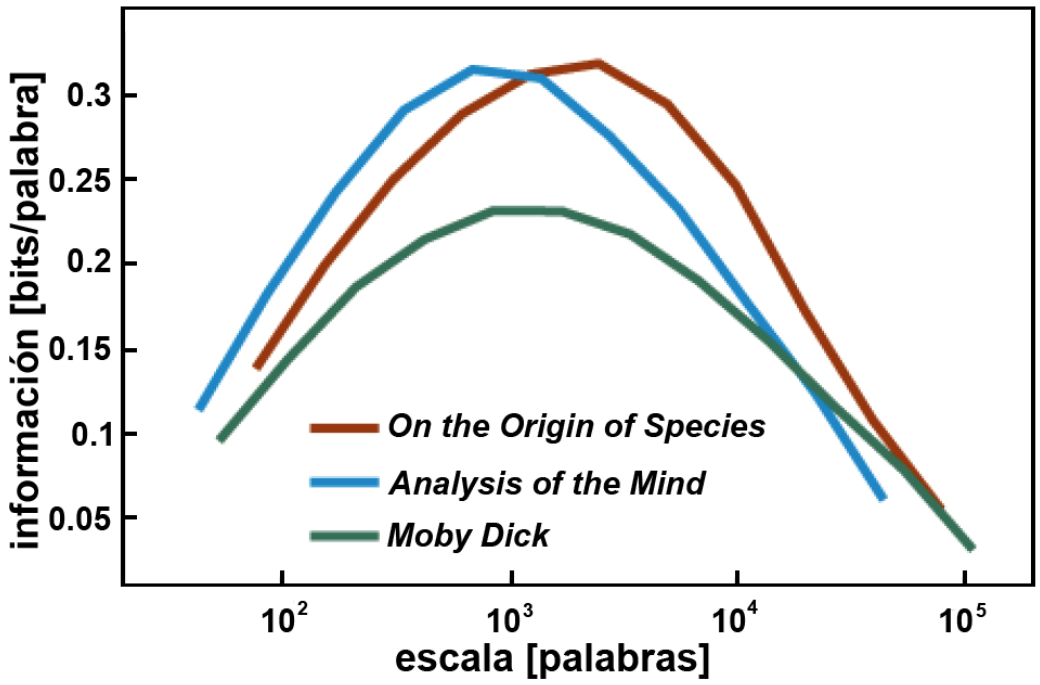

Fig. 2. Información como función de la escala para las tres obras consideradas en la Tabla 1.

La detección de contenido semántico es un ingrediente esencial en el procesamiento automático de la información bibliográfica. El método que hemos presentado tiene la ventaja de no requerir conocer *a priori* el idioma en que está escrito el texto; basta con poder individualizar sus palabras.

Más información:

Montemurro M. A. and Zanette D. H., *The statistics of meaning: Darwin, Gibbon and Moby Dick*, Significance, Dec. 2009, 166-169.

Montemurro M. A. and Zanette D. H., *Towards the quantification of the semantic information encoded in written language*, http://arxiv.org/abs/0907.1558, to appear in Adv. Compl. Sys. (2010).