Estudio de la aplicabilidad de la ley de Zipf y de la ley de Heaps en los corpus de aprendientes de inglés

Nicolas Ballier, Paula Lissón

Estudio de la aplicabilidad de la ley de Zipf y de la ley de Heaps en los corpus de aprendientes de inglés.

Nicolas Ballier[1] Paula Lissón[1]
(1) Université Paris-Diderot, EA 3967 – CLILLAC-ARP, France
nicolas.ballier@univ-paris-diderot.fr, paula.lisson@etu.univ-paris-diderot.fr

Keywords: *corpus de aprendientes, complejidad léxica, Zipf-Mandelbrot, crecimiento del vocabulario, hápax legomena.*

Este trabajo se centra en la aplicabilidad de la ley de Zipf-Mandelbrot (Zipf, 1949; Mandelbrot, 1953) y de la ley de Heaps (1978) en los corpus de aprendientes. Para ello, realizaremos una comparación entre las curvas de crecimiento del vocabulario en textos escritos por nativos ingleses y en textos escritos por aprendientes de inglés.

La ley de Zipf-Mandelbrot establece que, en un texto dado, la distribución de las palabras está relacionada con su frecuencia. Esto se traduce en que el texto estará compuesto por pocas palabras con mucha frecuencia, y por muchas palabras con poca frecuencia. En un estudio reciente, Bentz y Buttery (2014) muestran que a) la ley de Zipf-Mandelbrot puede ser utilizada como medida de estudio de la diversidad léxica y, b) no todas las lenguas siguen de la misma forma la ley de Zipf-Mandelbort. Nuestra hipótesis es que los aprendientes de inglés no siguen exactamente la ley de Zipf-Mandelbort, y que su curva de crecimiento del vocabulario es diferente con respecto a la curva de los nativos, lo que podría ayudarnos a clasificar a los aprendientes en diferentes niveles.

La ley de Heaps (1978), complementaria a la ley de Zipf, establece que el crecimiento del vocabulario de un texto dado es una función del tamaño de dicho texto. Si aumentáramos el tamaño del texto, aunque el crecimiento del vocabulario seguiría siendo ascendiente, dejaría de ser linear, ya que a medida que se incrementa el número de palabras, la posibilidad de que aparezcan palabras nuevas se ve reducida. Nuestra hipótesis es que los aprendientes presentan un crecimiento del vocabulario más limitado, por lo que la producción de hápax legomena sería inferior a la predicción propuesta por la ley de Heaps (aproximadamente la raíz cuadrada del número total de *tokens*).

Para probar nuestra hipótesis, estudiaremos la aplicabilidad de la ley de Zipf-Mandelbrot y de la ley de Heaps en un corpus escrito de estudiantes hispanófonos de inglés, **NOCE** (Díaz-Negrillo, 2007), y compararemos los resultados con los de un corpus de producciones escritas de nativos ingleses, **LOCNESS** (Paquot, 2015). De esta forma, analizaremos la valencia de las leyes aquí propuestas, mostrando así las variaciones entre los nativos y los no nativos.

A partir del número de *tokens* y de hápax legomena de nuestro corpus de aprendientes, generaremos los espectros de frecuencia que nos permitirán crear las curvas de crecimiento del vocabulario. Para ello, emplearemos el paquete {zipfR} (Evert & Baroni, 2006), implementado en el programa R (R Core Team, 2016). Siguiendo los pasos de Ballier y Gaillat (2016), utilizaremos la función "compare.richness.fnc" implementada en {languageR} (Baayen, 2007) para comparar el crecimiento del vocabulario entre las producciones de nativos y no nativos.

A continuación, desarrollaremos la extrapolación de las curvas de crecimiento de vocabulario (ver figura 1) según los tres modelos de *Large Number of Rare Events* (LNRE) incluidos en

{zipfr}: "Generalized Inverse Gauss-Poisson" (R Harald Baayen, 2001, 2008), "Zipf-Mandelbrot" y "Finite Zipf-Mandelbrot" (Evert, 2004). Finalmente, comparemos los resultados de los tres modelos para identificar cuál de ellos es más adecuado en el análisis de los corpus de aprendientes.

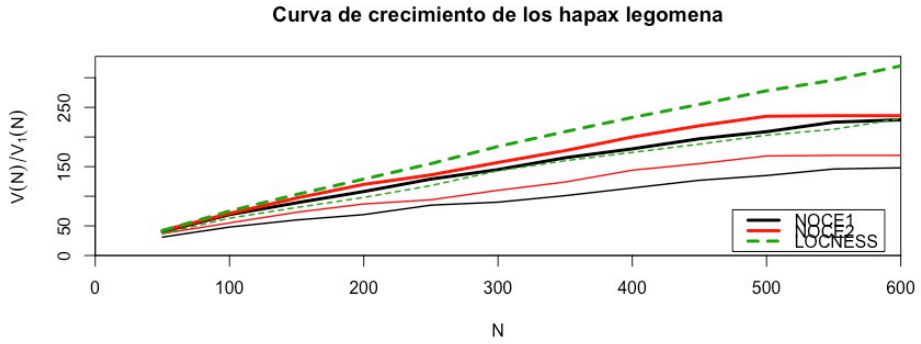

Figura 1: curva de crecimiento de los hápax legomena en producciones escritas de dos aprendientes (NOCE) y de un nativo (LOCNESS)

Referencias bibliográficas

Baayen, R. H. (2001). *Word frequency distributions* (Vol. 18). Springer Science & Business Media.

Baayen, R. H. (2007). languageR: data sets and functions for «Analyzing Linguistic Data» (Versión 1.0). Recuperado a partir de http://reaktanz.de/?c=hacking&s=koRpus

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Ballier, N., & Gaillat, T. (2016). Classification d'apprenants francophones de l'anglais sur la base des métriques de complexité lexicale et syntaxique (Vol. 9, pp. 1-14). Presentado en JEP-TALN-RECITAL 2016.

Bentz, C., & Buttery, P. (2014). Towards a computational model of grammaticalization and lexical diversity (pp. 38-42). Presentado en Proc. of 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)@ EACL.

Díaz-Negrillo, A. (2007). *A Fine-grained Error Tagger for Learner Corpora* (PhD Thesis). University of Jaen, Jaen.

Evert, S. (2004). A simple LNRE model for random character sequences (Vol. 2004). Presentado en Proceedings of JADT.

Evert, S., & Baroni, M. (2006). The zipfR library: Words and other rare events in R. *Relazione presentata all'useR*.

Heaps, H. S. (1978). *Information retrieval: Computational and theoretical aspects*. Academic Press, Inc.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. *Communication theory*, *84*, 486-502.

Paquot, M. (2015). LOCNESS.

R Core Team. (2016). R: A language and environment for statistical computing. (Versión 3.3.1 (2016-06-21)). Vienna, Austria.: R Foundation for Statistical Computing. Recuperado a partir de https://www.R-project.org/

Zipf, G. K. (1949). Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Addison-Wisley. *Cambridge, MA*.