Introducción al uso del paquete {koRpus} de R para el tratamiento de la complejidad léxica y la legibilidad en corpus lingüísticos.

Nicolas Ballier[1], Paula Lissón[1], Verónica C. Trujillo-González[2]
(1) Université Paris-Diderot, EA 3967 – CLILLAC-ARP, Francia
(2) IATEXT (Instituto de Análisis y Aplicaciones Textuales), ULPGC, España

nicolas.ballier@univ-paris-diderot.fr
paula.lisson@etu.univ-paris-diderot.fr
veronica.trujillo@ulpgc.es

La lingüística de corpus, un ámbito cada vez más interdisciplinar, reúne técnicas provenientes de disciplinas, a priori, tan dispares como el Tratamiento Automático del Lenguaje (TAL), la aplicación de modelos estadísticos o, incluso, la extracción de datos mediante el uso de programas informáticos utilizados principalmente en la minería de datos. Así, el uso de lenguajes de programación como Python o R (R Core Team, 2016) está cada vez más extendido en la lingüística de corpus, gracias a la creación de librerías o paquetes especializados en el tratamiento de corpus.

En la actualidad, existen más de 50[1] paquetes diseñados para la investigación de corpus lingüísticos, concebidos desde enfoques distintos, que responden a diferentes objetivos (exploración de corpus orales/escritos, lexicometría y textometría, entre otros). Nuestra propuesta se basa en la explotación del paquete {koRpus} (Michalke, 2016) , que contiene una implementación de 15 métricas de diversidad léxica y 34 métricas de legibilidad. Estas métricas nos permiten evaluar cuantitativamente la riqueza y la variedad del léxico y de las estructuras sintácticas empleadas en un corpus dado (cf. Lissón, 2017).

Mediante el uso de un anotador automático de partes del discurso (Treetagger; Schmid, 1995) y sus versiones en múltiples idiomas, las métricas pueden aplicarse a corpus de diversas lenguas, permitiendo así el desarrollo de estudios comparativos multilingües. Además, el formato en el que {koRpus} genera los resultados (conocido como *data frame* en R, cf. Ballier, 2016) facilita su posterior tratamiento estadístico. Así, por ejemplo, técnicas como el Análisis de Componentes Principales (ACP), el Análisis Factorial (AF), el Análisis Discriminante Lineal (ADL) o la creación de *random forests* (Tagliamonte y Baayen, 2012) pueden llevarse a cabo fácilmente con otros paquetes de R.

Nuestra contribución describe y analiza la cadena de tratamiento que debe llevarse a cabo para la utilización de las métricas implementadas en {koRpus}, desde la importación del corpus en formato *txt*, pasando por la "tokenización" con TreeTagger, la extracción y visualización de los resultados, hasta la modelización estadística con técnicas de reducción de la dimensionalidad. Se trata de mostrar cómo el paquete {koRpus} puede ser utilizado para la investigación de la variedad léxica y la legibilidad. Estos dos conceptos son clave en la investigación de la adquisición y el aprendizaje de lenguas extranjeras, pero también en otras aplicaciones como la autoría de textos y la lingüística forense.

La expansión de las tecnologías computacionales en la lingüística de corpus abre un nuevo paradigma de investigación para los lingüistas. La creación y difusión de corpus cada

[1] https://cran.r-project.org/web/views/NaturalLanguageProcessing.html

vez más extensos obliga, en cierta manera, al investigador a familiarizarse con programas complejos que faciliten la interoperabilidad y el paso del estudio cualitativo al cuantitativo, y viceversa. En ese sentido, nuestra contribución a HDH 2017 consiste en presentar la metodología necesaria para llevar a cabo estudios sobre el léxico y la legibilidad utilizando el lenguaje de programación R, mostrando el espectro de posibilidades que el tratamiento automático de datos proporcionado por el paquete {koRpus} ofrece.

Referencias

Ballier, N. (2016). R, pour un écosystème du traitement des données? L'exemple de la linguistique. Presentado en Données, métadonnées des corpus et catalogage des objets en sciences humaines et sociales.

Lissón, P. (2017). *A corpus-based study of lexical diversity and readability in EFL learners' productions: towards a critical statistical approach* (Unpublished MA thesis). Université Paris Diderot, Paris, France.

Michalke, M. (2016). koRpus: An R Package for Text Analysis (Versión 0.06-5). Recuperado a partir de http://reaktanz.de/?c=hacking&s=koRpus

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/

Schmid, H. (1995). Treetagger| a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, *43*, 28.

Tagliamonte, S. A., & Baayen, R. H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, *24*(2), 135-178. https://doi.org/10.1017/S0954394512000129