Universitat Oberta de Catalunya

artnodes
REVISTA DE ARTE, CIENCIA Y TECNOLOGÍA

http://artnodes.uoc.edu

ARTÍCULO

NODO «HUMANIDADES DIGITALES: SOCIEDADES, POLÍTICAS, SABERES»

Aracne: estudio de la variación lingüística en la prensa española entre 1914 y 2014

Elena Álvarez Mellado
Fundéu BBVA - UNED

Leticia Martín-Fuertes Moreno
Fundéu BBVA

Fecha de presentación: abril de 2018
Fecha de aceptación: septiembre de 2018
Fecha de publicación: noviembre de 2018

Cita recomendada

Álvarez Mellado, Elena; Martín-Fuertes Moreno, Leticia. 2018. «Aracne: estudio de la variación lingüística en la prensa española entre 1914 y 2014». En: Nuria Rodríguez-Ortega (coord.). «Humanidades digitales: sociedades, políticas, saberes». *Artnodes*. N.º 22: 7-12. UOC. [Fecha de consulta: dd/mm/aa] http://dx.doi.org/10.7238/a.v0i22.3200

Resumen

Aracne es un proyecto de lingüística de corpus que tiene como propósito medir cuantitativamente cómo ha evolucionado el lenguaje de la prensa española entre 1914 y 2014, con especial atención a los rasgos de riqueza léxica. El proyecto ha consistido en la creación de un corpus de prensa de dos millones de palabras, confeccionado a partir de noticias extraídas de las hemerotecas de cuatro periódicos centenarios (*El Norte de Castilla, El Diario de Mallorca, El Heraldo de Aragón* y *La Vanguardia*). A partir de este corpus, se han obtenido distintas medidas de riqueza por décadas (variación, densidad léxica y complejidad de los artículos) y se han comparado los valores obtenidos en las distintas épocas. Los resultados muestran que la riqueza y la complejidad de los textos periodísticos se han mantenido notablemente estables en los últimos cien años.

Palabras clave
riqueza lingüística, riqueza léxica, complejidad léxica, corpus de prensa

Aracne: a study of linguistic variation in the Spanish press between 1914 and 2014

Abstract
*Aracne is a corpus linguistic project aimed at quantitatively measuring how the language of the Spanish press evolved between 1914 and 2014, paying special attention to the features of lexical richness. The project has involved the creation of a two-million word corpus compiled on the basis of news items extracted from the newspaper archives of four centenarian newspapers (*El Norte de Castilla*,* El Diario de Mallorca*,* El Heraldo de Aragón *and* La Vanguardia*). On the basis of this corpus, different measures of richness have been obtained by decade (variation, lexical density and complexity of the articles) and the values obtained for the different periods have been compared. The results evince that, over the last century, the richness and complexity of journalistic texts has remained substantially stable.*

Keywords
linguistic richness, lexical richness, lexical complexity, newspaper corpus

1. Introducción y objetivos

¿Cómo ha cambiado la lengua de la prensa en los últimos cien años? ¿Es más rico el lenguaje periodístico de hoy que el de hace cien años? ¿Qué palabras eran más frecuentes antes que ahora? ¿Qué nos dice la variación de vocabulario de los cambios históricos y políticos que ha experimentado la sociedad española a lo largo del último siglo? Aracne es un proyecto de lingüística de corpus que tiene como propósito medir cuantitativamente cómo ha evolucionado el lenguaje de la prensa española entre 1914 y 2014, con especial atención a los rasgos de riqueza léxica. En este artículo, presentamos las conclusiones y cuestiones metodológicas relativas al proyecto.

2. Material y método

El proyecto ha consistido en la selección, procesamiento y análisis de un corpus de dos millones de palabras procedentes de cuatro diarios españoles que han venido publicándose de forma ininterrumpida desde 1914 al 2014 (*La Vanguardia*, *El Norte de Castilla*, *Heraldo de Aragón*, *Diario de Mallorca*). Para las mediciones de riqueza y frecuencia, se ha considerado la década como la unidad temporal mínima sobre la que trabajar. La selección de ejemplares para confeccionar el corpus se ha hecho a través de un muestreo aleatorio, para garantizar una representación no sesgada de las distintas décadas, meses y días de la semana.

La manera de proceder para llevar a cabo el estudio ha sido:

1. Obtención de los ejemplares seleccionados en formato imagen digital de las hemerotecas de los medios colaboradores.
2. Procesamiento mediante tecnología OCR para extracción del texto del ejemplar.
3. Supervisión humana (asistida por ordenador mediante el uso de macros, corrección semiautomática y expresiones regulares) de los textos producidos por el OCR, para garantizar una calidad aceptable de los textos y separación del texto continuo del ejemplar producido por el OCR en artículos.
4. Lematización y procesamiento de los textos mediante tecnología de procesamiento de lenguaje natural (PLN).
5. Medición de las variables (variación léxica, densidad y complejidad de los artículos).
6. Cálculo de medias y agrupación de los datos en intervalos de tiempo.
7. Visualización de los datos y análisis.

3. Medición de los datos

¿Cómo se mide la riqueza lingüística? ¿Es posible establecer de manera objetiva y numérica las diferencias de complejidad lingüística entre dos textos? Para observar la variación de la riqueza de los artículos del corpus a lo largo del tiempo, se han medido y comparado tres

rasgos que se consideran indicadores de la complejidad lingüística de un texto (Johansson, 2008): la variación léxica (representada por la relación entre palabras distintas y palabras totales o types/tokens ratio, TTR), la densidad léxica (número de palabras semánticamente plenas en relación con las palabras totales del texto) y el nivel de sofisticación lingüística (complejidad léxica y gramatical de los textos).

1. Para la variación léxica se han considerado dos índices, ambos inspirados en la variable tradicional TTR (*types/tokens ratio*, es decir, palabras distintas entre palabras totales), pero en una propuesta ligeramente modificada para adaptarlos mejor al procesamiento de textos en español y al propósito del proyecto. Puesto que la variable TTR no permite comparar la riqueza de textos de distinta longitud, las mediciones y comparaciones de nuestros índices TTR (o sus modificaciones) se han hecho sobre artículos de extensión semejante. Por otro lado, la variable TTR tiene limitaciones muy evidentes al aplicarla a lenguas flexivas como el español, ya que la medición tradicional del TTR considera como tipos distintos (*types*) formas diferentes de una misma palabra (plurales, femeninos, conjugaciones). El cálculo del TTR en el proyecto, por lo tanto, se ha hecho considerando las formas lematizadas de las palabras, no la diferenciación tradicional de *types* y *tokens*. Consideramos que esta aproximación es más apropiada que la convencional, pues recoge de una manera más fiel el funcionamiento morfológico del español y lo que se considera variación léxica.
2. El segundo rasgo que se ha tenido en cuenta ha sido la densidad léxica, medida como la relación entre el número de palabras con categoría semántica (nombres, adjetivos, verbos, adverbios acabados en -mente) y palabras totales del texto. Los valores de la densidad léxica oscilan entre 0 y 1, tendiendo a 1 los textos muy densos (es decir, con una alta proporción de palabras semánticamente plenas) y a 0 aquellos en los que aparecen muchas palabras más gramaticales o estructurales.
3. Por último, se ha observado la complejidad lingüística de los textos, medida como el grado de elaboración léxica y gramatical de los artículos procesados (longitud de las frases, frecuencia y nivel de complejidad de las palabras, uso de tiempos verbales, elaboración sintáctica).

4. Resultados y conclusiones

A tenor de los resultados obtenidos, se puede observar que los tres rasgos medidos para observar la variación de la riqueza lingüística a lo largo del corpus se mantienen llamativamente estables en los cien años que comprende el estudio (véanse las figuras 1, 2, 3 y 4).[1]

Si bien el objetivo fundamental de proyecto ha sido medir la evolución de la riqueza lingüística, el análisis del corpus permite observar otros rasgos relacionados con los cambios históricos en el lenguaje de la prensa, como la evolución que han sufrido los adjetivos en grado superlativo (-ísimo), la evolución del modo subjuntivo, el espectacular desplome de los tratamientos de persona «don» y «señor» o la frecuencia relativa de algunas palabras cultural e históricamente relevantes como «guerra», «peseta», «fanega», «comunismo», «alemán», «nuclear» o «europeo».[2]

Estas gráficas de evolución de la frecuencia de palabras relacionadas con los avatares culturales e históricos de los últimos cien años

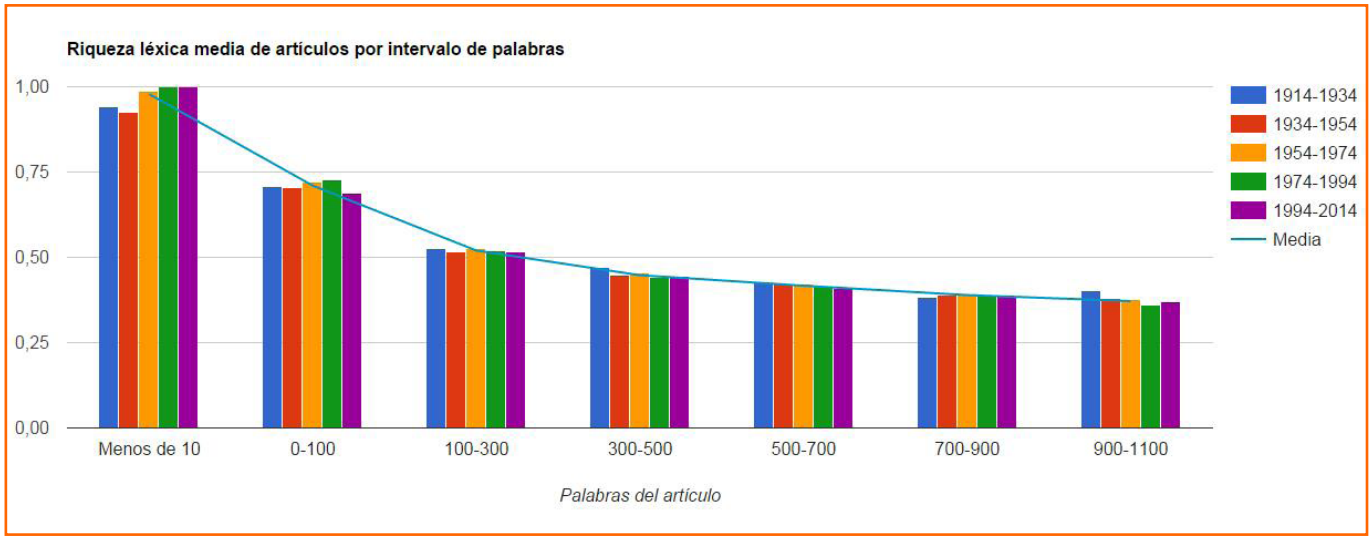

Figura 1. Riqueza léxica media de artículos por intervalo de palabras

1. Gráficas y visualizaciones disponibles en http://www.fundeu.es/aracne/riqueza.html.
2. Gráficas y visualizaciones disponibles en http://www.fundeu.es/aracne/gram.html y http://www.fundeu.es/aracne/lexico.html.

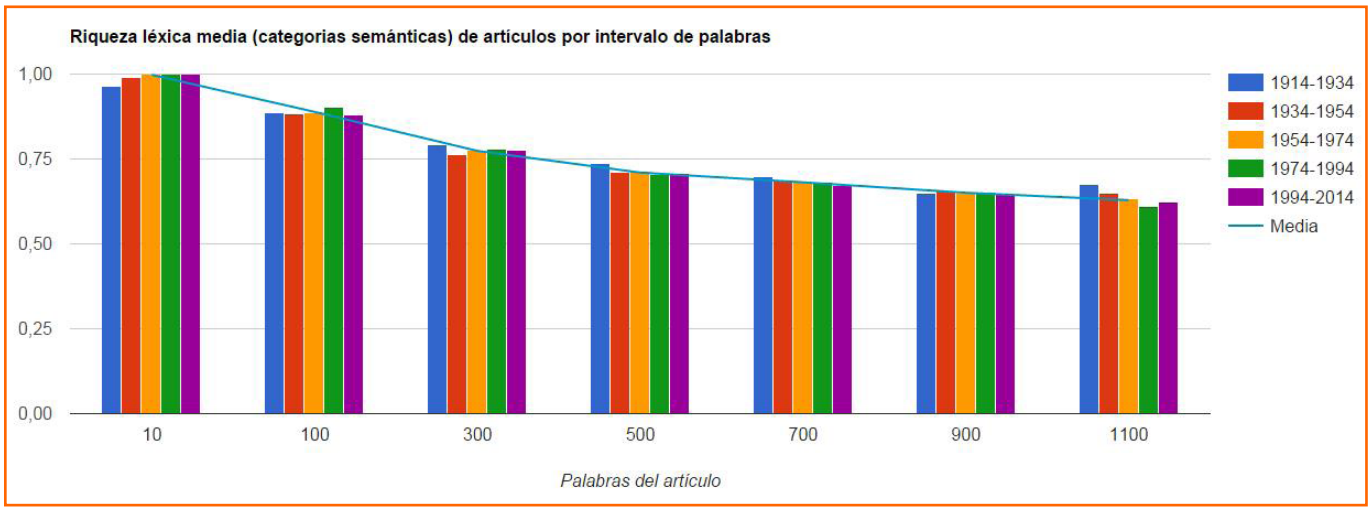

Figura 2. Riqueza léxica media (categorías semánticas) de artículos por intervalo de palabras

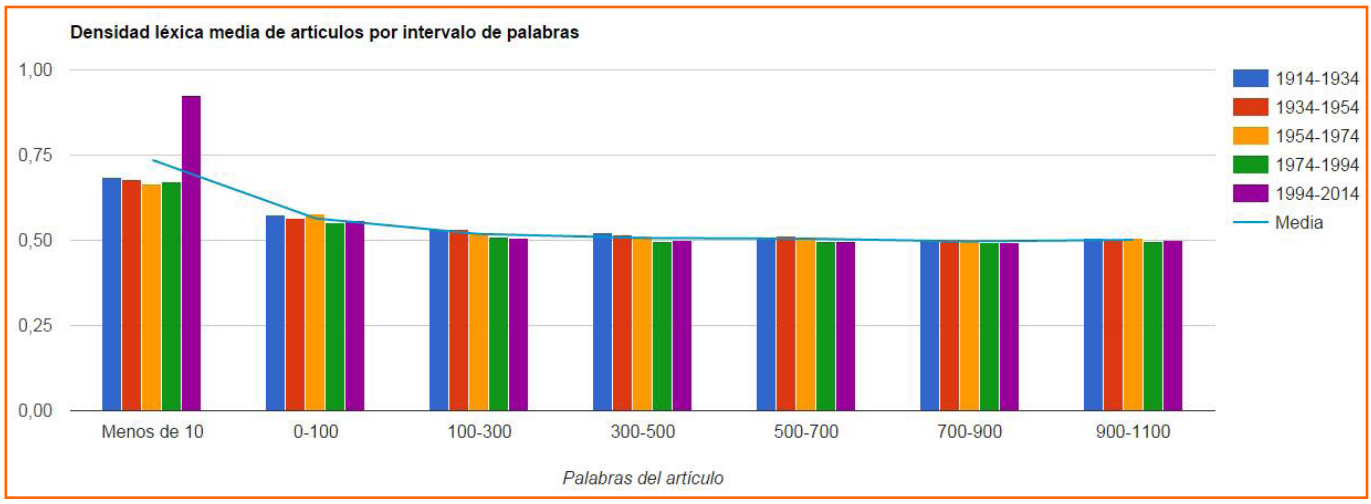

Figura 3. Densidad léxica media de artículos por intervalo de palabras

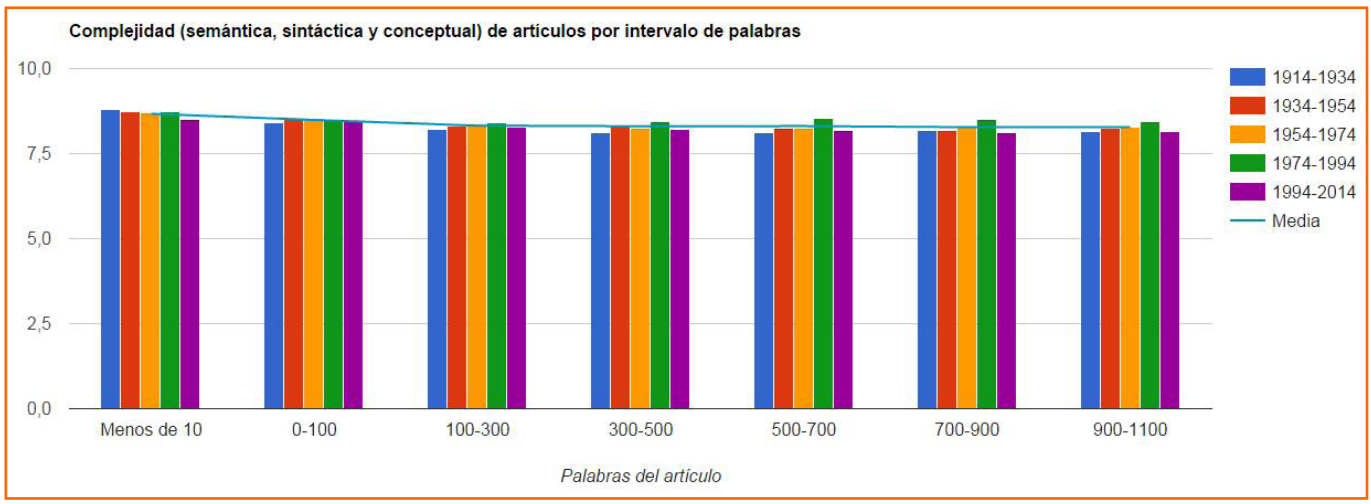

Figura 4. Complejidad (semántica, sintáctica y conceptual) de artículos por intervalo de palabras

permiten asomarse al léxico que caracterizó cada década. El perfil de picos y valles que dibuja la frecuencia léxica constituye una radiografía de las hemerotecas españolas, es un reflejo de la propia historia de España y Europa a lo largo de los últimos cien años e invita a atisbar cómo era la sociedad que producía y redactaba estas noticias.

5. Agradecimientos

Proyecto Aracne es un proyecto impulsado por la Fundación del Español Urgente (Fundéu) y financiado por el banco BBVA.

6. Bibliografía

Johansson, V. 2008. «Lexical diversity and lexical density in speech and writing: a developmental perspective». *Working papers in Linguistics Lund University*, vol. 53: 61-79.

CV

Elena Álvarez Mellado
Fundéu BBVA y Universidad de Educación a Distancia (UNED)
ealvarezmellado@gmail.com

c/ Juan del Rosal, 16
28040 Madrid

Elena Álvarez Mellado es licenciada en Lingüística por la Universidad Complutense de Madrid. En 2009, recibió el primer premio del certamen universitario Arquímedes de introducción a la investigación científica en el área de ciencias sociales y humanidades. Entre 2010 y 2016, trabajó como lingüista computacional en la empresa de tecnología lingüística Molino de Ideas, y durante 2015 dirigió el proyecto de investigación Aracne para la Fundación del Español Urgente (Fundéu BBVA). Ha sido personal investigador en el Laboratorio de Innovación en Humanidades Digitales de la ETSI de Informática de la UNED, como parte del proyecto POSTDATA. Actualmente es estudiante de posgrado en la Universidad de Brandeis (Massachusetts) con una beca de la Fundación LaCaixa. Compagina su labor como lingüista con la faceta de divulgadora: en 2016 publicó el libro de divulgación lingüística *Anatomía de la lengua*, y desde 2017 escribe una columna sobre lengua en eldiario.es, columna por la que recibió en 2018 el Premio Nacional de Periodismo Miguel Delibes.

CV

Leticia Martín-Fuertes
Fundéu BBVA
leticia.mfm@gmail.com

Av. de Burgos, 8 - 28036 Madrid

Leticia Martín-Fuertes Moreno es licenciada en Filología Clásica por la Universidad Autónoma de Madrid (UAM). Fue colaboradora habitual de la revista universitaria *Hermes*, dedicada a la actualidad del mundo clásico, y editora en LID Editorial Empresarial. En 2015 trabajó como lingüista computacional en Molino de Ideas, participando en el proyecto de investigación Aracne para la Fundación del Español Urgente (Fundéu BBVA). Desde entonces ha trabajado como lingüista computacional en Bitext y, actualmente, en Google, contratada por Adecco. También es colaboradora del máster de Humanidades Digitales de la Universidad Nacional de Educación a Distancia (UNED).