ISSN: 2255-3703

Big data y corpus lingüísticos para el estudio de la densidad léxica

Adela González Fernández
Universidad de Córdoba
adela.gonzalez@uco.es

Fecha de recepción: 20.7.2018
Fecha de aceptación: 15.09.2018

Resumen: La unión entre la Informática y de la Lingüística es cada vez más frecuente en las investigaciones en el campo del lenguaje y de las lenguas. La Lingüística de corpus, en especial, se está viendo beneficiada por este emparejamiento, gracias a los avances a la hora de gestionar y procesar los corpora. En este trabajo damos un paso más y proponemos el trabajo en Lingüística de corpus a través de *big data*, en general, y de *Twitter*, en particular. Gracias a la creación de una herramienta informática diseñada específicamente para el trabajo lingüístico en *big data*, obtendremos una inmensa cantidad de información textual que nos servirá para la compilación de *corpora* mediante los que estudiaremos la diversidad léxica en el lenguaje de cuatro escritores españoles. Para ello, extraeremos los tuits publicados por ellos en sus cuentas de *Twitter* y los procesaremos a través de nuestra herramienta para obtener la información deseada. Intentaremos demostrar, también, la mejora que esta nueva metodología supone en este tipo de estudios.

Palabras clave: lingüística de corpus; densidad léxica; diversidad léxica; *big data*; *Twitter*.

Big data and linguistic corpora as a tool for the study of lexical density

Abstract: The merger of Computer Sciences and Linguistics is increasingly common in researches in the field of languages. Corpus linguistics, specially, is benefiting from this matching, due to the improvements in management and processing of corpora. In this work we go one step further and suggest working in Corpus linguistics with big data, in general, and *Twitter*, in particular. Thanks to the development of a software specifically designed for linguistic work with big data, we will obtain a vast amount of information which will be used to compile linguistic corpora through which we will study the lexical density in four Spanish writers. In order to do that, we will obtain the tweets published by them in their *Twitter* accounts and we will process them with our software tool. We also aim to prove the benefits that this methodology implies in this kind of research.

Key words: corpus linguistics; lexical density; lexical diversity; big data; *Twitter.*

Sumario: Introducción. 1. Marco teórico. 2. Metodología. 3. Resultados. Conclusiones.

Introducción

La relación entre la Lingüística y la Informática ha pasado por distintas etapas desde sus inicios hasta la actualidad, y se ha vuelto tan estrecha que ya es difícil imaginarse el estudio sobre el lenguaje sin el soporte informático. Para Tognini-Bonelli (2001), la introducción de los ordenadores en el ámbito lingüístico y, más concretamente, en el trabajo con corpus, recorre fundamentalmente tres etapas. La primera de ellas consideraba a la Informática como una simple herramienta para el trabajo lingüístico –hasta el momento la mayor contribución a la Lingüística, según la autora–, gracias a la cual era posible gestionar y procesar la información de una manera más rápida y cómoda. La siguiente fase se caracterizó no solo por la mayor abundancia de ejemplos reales de información, sino por la propia naturaleza de los ordenadores, que afectó al marco metodológico de la investigación gracias a una mayor velocidad, sistematización y volumen de los datos. La década de los noventa fue testigo de la tercera etapa que describe Tognini-Bonelli, gracias al increíble aumento de la información procesable con la ayuda del ordenador, que contribuyó no solo a la mejora cualitativa sino también a la cuantitativa y, con ellas, a la revolución que ha aportado nuevos enfoques y removido cuestiones teóricas ya establecidas. Este hecho ha provocado que autores como Leech (1992), Halliday (1993) o la propia Tognini-Bonelli (2001) entre muchos otros, defiendan la posición de la Lingüística de Corpus como ciencia, más allá del estatus metodológico que se le ha otorgado tradicionalmente.

Reconocen Renouf y Kehoe (2006) que, tras más de veinte años, la Lingüística de Corpus acoge una mayor variedad de actividades, relacionadas con la elaboración de corpus de pequeño, mediano o gran tamaño, así como con la construcción de corpus multidimensionales. Además, también está relacionada con el análisis de estos corpus, su evaluación y la revisión de teorías existentes. Insisten los autores en el prólogo de su libro *The Changing Face of Corpus Linguistics* (Renouf y Kehoe, 2006) en que no son estos los únicos aspectos en los que la Lingüística de Corpus está sufriendo cambios, y nos recuerdan que la lengua es un fenómeno cambiante y que el concepto de corpus se está viendo modificado a partir de la disponibilidad de textos accesibles desde la *World Wide Web*. Leech (2007: 133) refuerza esta idea cuando afirma que:

> in one sense corpus linguistcs appear to inhabit an expanding universe. The Internet provides a virtually boundless resource for the methods of corpus linguistics. In addition, there is continuing growth in the number

> and extent of text archives and other text resources.... This is greatly to be welcomed, obviously.

Siguiendo esta línea, hemos utilizado para este trabajo una metodología novedosa que supera la etapa de la *Web como corpus* (Kilgarriff 2001; Kilgarriff y Grefenstette 2003) puesto que aúna el trabajo con corpus y con *big data*.

Entendemos por *big data*, *grosso modo*, los grandes conjuntos de información que por sus características no pueden ser obtenidos, gestionados ni procesados por herramientas tradicionales en un período de tiempo razonable. *Big data* se caracteriza por tres rasgos fundamentales que lo diferencian de la información tradicional: el enorme volumen de datos que lo componen, la velocidad con la que esos datos se generan y se transmiten y la variedad de formatos, temas, procedencias y tipos que lo forman. Las posibilidades que se le abren al investigador lingüístico en este sentido son muy numerosas, puesto que gracias a *big data* es posible realizar análisis más exhaustivos, basados en millones de datos y con muy poca inversión de tiempo, lo que repercute en mayor conocimiento y mayores beneficios, con mucho menor esfuerzo por parte del analista.

Dentro de *big data*, hemos seleccionado el servicio de *microblogging* más utilizado mundialmente, *Twitter*, como base de datos para llevar a cabo análisis sobre el lenguaje debido a su naturaleza, más centrada en elementos textuales que gráficos o audiovisuales, así como su consideración por parte de los usuarios como plataforma para comunicarse con los demás, expresar opiniones y sentimientos o para transmitir información. Como Gantz y Reinsel (2011) afirman, los medios sociales, como *Twitter*, son las nuevas fuentes de información porque han construido sistemas en los que los consumidores, de manera consciente o no, generan flujos de información continuos que tienen la capacidad de expandirse rápidamente gracias a las características de Internet.

Son numerosas las investigaciones acerca de la densidad lingüística, como el *Proyecto Aracne* de la Fundéu BBVA (2015), Ávila (1991), Gómez Molina (2004), Rodríguez-Tapia (2016), etc., aunque ninguna de ellas trabaja con las técnicas de *big data*. Algunos estudios lingüísticos de otra índole que han seguido esta técnica son, por ejemplo, el de Grieve, Nini, Guo y Kasakoff (2015) Huang, Guo, Kasakoff y Grieve (2016), donde estudian la variedad y el cambio lingüístico a través de *Twitter*; o Renouff (2015), quien destaca las ventajas de la utilización de *big data* para la Lingüística de Corpus.

Esta investigación pretende demostrar la utilidad de trabajar con esta metodología para la obtención de información que nos permita determinar la

densidad léxica de determinados tipos de textos. En este caso concreto, hemos querido seleccionar para el estudio, y para evidenciar la posibilidad de realizar este tipo de estudios a través de una herramienta informática sobre el corpus de *Twitter*, el lenguaje de cuatro escritores españoles.

1. Marco teórico

Antes de comenzar, es importante tener en cuenta algunas consideraciones teóricas acerca de la cuestión de la densidad léxica.

Johansson (2008) define *diversidad léxica* como una "medida que da cuenta del número de palabras diferentes que hay en un texto", mientras que explica que la *densidad léxica* "mide la proporción de elementos léxicos (sustantivos, verbos, adjetivos y algunos adverbios) en el texto" (Johansson 2008: 63). Explica la autora, además, que es teoréticamente posible que un texto tenga una alta diversidad léxica –que contenga muchos tipos distintos de palabras– y una baja densidad léxica –porque aparezcan más pronombres y verbos auxiliares que sustantivos o verbos léxicos– o viceversa.

Gelbukh, Sidorov y Chanona (2002: 3) explican que una de las desventajas de los corpus tradicionales es que presentan pocas o ninguna ocurrencia de muchas palabras, mientras que otras aparecen muy repetidas, debido al fenómeno conocido como la ley de Zipf (Zipf 1965). También Rayson, Walkerdine, Fletcher y Kilgarriff (2006) achacan a esta ley el hecho de que la mitad de las palabras que hay en los corpus aparecen una sola vez, por lo que los grandes corpus son necesarios para asegurar la inclusión de palabras y frases fundamentales y para aumentar las posibilidades de aparición

Algunos autores, como Daller, van Hout y Treffers-Daller (2003) intercambian indistintamente el uso del término *diversidad léxica* con el de *riqueza léxica*, mientras que otros, como Malvern *et al*. (2004), en Johansonn (2008), indican que la *diversidad léxica* es solo una parte del concepto multidimensional de *riqueza léxica*. En esta línea, Laufer y Nation (1995: 309) enumeran varios elementos para medir la riqueza léxica, que son: originalidad léxica, densidad léxica, sofisticación léxica y variación léxica. En este trabajo, siguiendo a Daller, van Hout y Treffers-Daller (2003), así como a Gregori Signes y Clavel Arroitia (2015) y al *Proyecto Aracne* de la Fundéu BBVA (2015), utilizaremos los términos *densidad* léxica, *riqueza léxica* y *diversidad léxica* como sinónimos.

Tradicionalmente, la riqueza léxica de los textos se ha medido a través de la denominada TTR (*Type-Token Ratio*), donde *token* se refiere al número total de palabras que un texto contiene (*casos*, en español) y *type* al

repertorio de palabras distintas (*tipos*, en español) (Bergman y Paavola 2003). Cuanto más variado sea el vocabulario de un texto, mayor diversidad léxica tendrá, lo que significa que estará compuesto por un alto número de palabras distintas y poco repetidas. La relación tipo-caso se obtiene de dividir los primeros entre los segundos:

TTR=tipo/caso

La siguiente oración, por ejemplo, tiene una TTR de 1, puesto que tiene el mismo número de tipos que de casos, porque no se repite ninguna palabra:

El viaje consistió únicamente en cinco días cabalgando y descansando, sin conversaciones.

La oración consta de doce palabras y ninguna de ellas se repite, por lo que 12/12=1. Por el contrario, la siguiente oración:

Él se detuvo unos pasos por delante y se dio la vuelta.

Tiene una ratio de 0,91, puesto que, de 12 palabras que la componen, hay una ("se") que se repite dos veces. Por lo tanto: 11/12=0,91.

La relación entre los tipos y los casos se mueve inevitablemente entre los valores 0 y 1, puesto que, siendo los tipos de un texto *n* –dependiendo de su longitud–, los casos, como mínimo, darán un valor de 1 y, como máximo, de *n*. Esto quiere decir que los valores posibles de TTR oscilarán entre 1/n –cuando haya solo una palabra distinta–y n/n (que es igual a 1) –cuando no se repita ninguna palabra.

En este trabajo, mostramos los resultados de densidad lingüística en términos de porcentajes, como es habitual en los estudios lingüísticos, para facilitar la comprensión. Estos porcentajes resultan de multiplicar por cien el resultado de la división entre tipos y casos.

A pesar del uso generalizado de este procedimiento, existen algunos problemas y limitaciones que es necesario tener en cuenta a la hora de utilizarlo para medir la densidad léxica de un texto.

En primer lugar, la relación entre los tipos y los casos surgió como recurso para analizar la lengua inglesa, cuya variación morfológica es mucho menor que la del español. Por ello, la ratio TTR considera distintas las palabras que comparten un mismo lema, algo que puede resultar útil para lenguas poco flexivas. Sin embargo, esto implica que un mismo lema en un idioma flexivo como el español es contabilizado como varias palabras distintas, según vaya variando en género y en número (o persona y tiempo, en el caso de los verbos). De esta forma, los artículos determinados *el*, *la*,

los, *las*, son contemplados como cuatro palabras distintas, mientras que en inglés es solo una: *the*.

Por otro lado, es obvio que esta relación se ve muy condicionada por la longitud de los textos, ya que, cuanto mayor sea su extensión, más probabilidades hay de que las palabras aparezcan en más de una ocasión.

2. Metodología

Para poder llevar a cabo un estudio de estas características, en el que la fuente de información y el material para la elaboración del corpus es *big data*, hemos realizado el diseño y la implementación de una herramienta informática que permite extraer de *Twitter* la información lingüística que se le solicite, así como almacenarla, gestionarla y analizarla (González-Fernández 2016). En este caso, hemos confeccionado un *corpus* con el material textual de las cuentas individuales de esta plataforma de cuatro escritores españoles de reconocido prestigio y con una participación activa en ella. Los autores son: Mónica Carrillo (@MonicaCarrillo), Arturo Pérez Reverte (@perezreverte), Daniel Sánchez Arévalo (@sanchezarevalo) y Lucía Etxebarria (@LaEtxebarria). La elección de estos escritores y no otros viene motivada por su alta participación en la red social y, por consiguiente, las posibilidades que ofrecían para la realización de un estudio significativo.

La herramienta que hemos mencionado contiene distintas funcionalidades que permiten al usuario trabajar con la información de diversas maneras y obtener resultados desde perspectivas diferentes, según las necesidades de la investigación. Los datos son extraídos a partir de la API oficial de *Twitter*: *TwitterAPI*. Entendemos por API (*Application Programming Interface*):

> Un conjunto de reglas (código) y especificaciones que las aplicaciones pueden seguir para comunicarse entre ellas: sirviendo de interfaz entre programas diferentes de la misma manera en que la interfaz de usuario facilita la interacción humano-software (Merino, 2014).

La herramienta obtiene los tuits a partir de esta interfaz, que facilita el intercambio de datos, los procesa y los transforma en lenguaje JSON. JSON (*JavaScript Object Notation*) es un formato ligero de intercambio de datos, independiente del lenguaje y basado en el texto que permite la transmisión de información estructurada entre todos los lenguajes de programación. A partir de este momento, se realizan, mediante toda una serie de técnicas informáticas, se realizan procesos de análisis de los datos desde una perspectiva puramente lingüística. Puesto que nuestro interés, en este caso, ha sido realizar un análisis de la densidad léxica en el corpus extraído,

hemos realizado la búsqueda en las cuentas de los escritores mencionados y hemos aplicado la función de análisis de densidad léxica.

El corpus obtenido, que por motivos de espacio no podemos mostrar aquí, contiene una media aproximada de 2300 tuits por cada escritor; es decir, unos 9200 tuits en total. Las fechas de publicación de estos tuits varían, puesto que, aunque el número de tuits de cada uno es similar –en la Figura 5 se muestran los datos–, no lo es la frecuencia con la que escriben o no todos han escrito exactamente en las mismas fechas. Por lo tanto, se tratan de los últimos tuits obtenidos con fecha de marzo de 2016.

3. Resultados

A continuación, mostramos los resultados obtenidos tras el análisis de todos estos tuits, de forma individual para cada escritor. Como podemos ver en las distintas figuras (Figura 1, Figura 2, Figura 3 y Figura 4), la información devuelta por la herramienta hace referencia a los conceptos de tipos, casos y densidad de la lengua, explicados anteriormente.

Mónica Carrillo:

Descripción	Valor
Tipos (Type)	6925
Casos (Token)	24678
Densidad de la lengua	28.06%

Figura 1. Relación tipo/caso en la cuenta de Mónica Carillo

Arturo Pérez-Reverte:

Descripción	Valor
Tipos (Type)	6054
Casos (Token)	21768
Densidad de la lengua	27.81%

Figura 2. Relación tipo/caso en la cuenta de Arturo Pérez-Reverte

Daniel Sánchez Arévalo:

Descripción	Valor
Tipos (Type)	9119
Casos (Token)	37768
Densidad de la lengua	24.14%

Figura 3. Relación tipo/caso en la cuenta de Daniel Sánchez Arévalo

Lucía Extebarria:

Descripción	Valor
Tipos (Type)	8860
Casos (Token)	31583
Densidad de la lengua	28.05%

Figura 4. Relación tipo/caso en la cuenta de Lucía Etxebarria

La información que nos aportan estas tablas consiste en un recuento del número de casos de los tuits recopilados, por un lado (fila 2); el número total de tipos (fila 1), por otro; y, por último, la relación entre ambos valores, representada en términos de porcentaje (fila 3), como hemos visto anteriormente.

Si comparamos los datos pertenecientes a los cuatro escritores, podemos ver cómo Mónica Carillo y Lucía Etxebarria obtienen prácticamente el mismo nivel de densidad léxica –28,06% y 28,05%, respectivamente–. Arturo Pérez-Reverte está muy próximo a ellas, con una diferencia de un 0,24%, mientras que Daniel Sánchez Arévalo se queda más atrás con 24,14%.

Para que el análisis resulte más sencillo, presentamos a continuación los datos obtenidos en una tabla comparativa:

Datos	Carrillo	Pérez-Reverte	Sánchez Arévalo	Etxebarria
Número de tuits	2509	1629	2820	2201
Tipos	6925	6054	9119	8860
Casos	24678	21768	37768	31583
Densidad	**28,06%**	**27,81%**	**24,14%**	**28,05%**

Figura 5. Comparación densidad léxica entre los autores analizados

Como se puede observar, en la tabla se ha incluido un dato más, relativo al número de tuits que la herramienta devuelve de cada autor. Este dato es importante porque, recordemos, la longitud del texto determina el cálculo de su densidad léxica, como ya hemos explicado.

Según esto, la baja densidad de Daniel Sánchez Arévalo puede venir explicada por el mayor número de publicaciones –y, por ende, de casos– con respecto a sus compañeros. Por otro lado, Lucía Etxebarria, de quien se han obtenido menos tuits que de su compañera, Mónica Carrillo, tiene un número mayor de casos que esta, lo que indica que ha introducido más variedad en los 140 caracteres disponibles para cada *post*. Esto quiere decir que el alto valor de densidad léxica de Etxebarria se ve reforzado, teniendo en cuenta que, aunque haya publicado menos tuits, ha producido más casos que los dos anteriores.

Puesto que la longitud del texto y el número de casos son factores limitantes a la hora de establecer una densidad léxica real, estos resultados no pueden considerarse como valores absolutos. Las comparaciones, por tanto, no son estrictamente fieles a la realidad ya que sería necesario analizar el mismo número de casos en los distintos sujetos de estudio, para poder determinar con fiabilidad cuáles son las densidades en cada uno de ellos.

Para ello, la herramienta que se ha diseñado propone una solución, que consiste en una gráfica segmentada que permite seleccionar un número concreto de casos en un momento determinado y en todas las cuentas. Esto quiere decir que es posible saber, dado un número determinado de palabras producidas, cuál es la relación entre tipos y casos de ese usuario y, por lo tanto, obtener una comparación real.

En el momento de utilizar esta función de la herramienta, hemos seleccionado el punto de la gráfica coincidente con 20.000 palabras en los

cuatro casos. El motivo de esta cifra es la búsqueda de un valor alto, pero que todos los escritores compartieran, para que pudiera hacerse la comparación. Como Pérez-Reverte es el que menos casos tiene, se ha escogido una cifra próxima a su máximo de palabras. A continuación mostramos las gráficas individuales, donde la línea roja diagonal representa el 100% de los casos y que la línea indica el número de tipos en relación a los casos concretos, según vayan aumentando a lo largo del tiempo. Como se puede observar, aparece, en verde, el número de casos seleccionado –20000– y, en color negro, los tipos contabilizados dentro de ese número de casos. El porcentaje de la densidad es el resultado de la división del número de tipos entre el número de casos y su multiplicación por 100.

Mónica Carrillo:

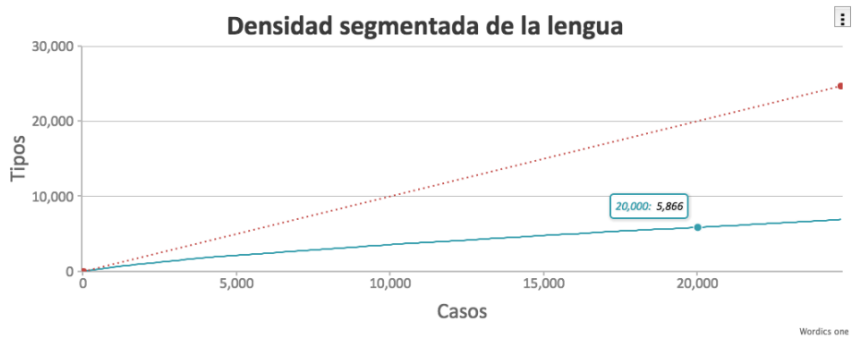

Figura 6. Densidad léxica segmentada de Mónica Carrillo

Arturo Pérez-Reverte:

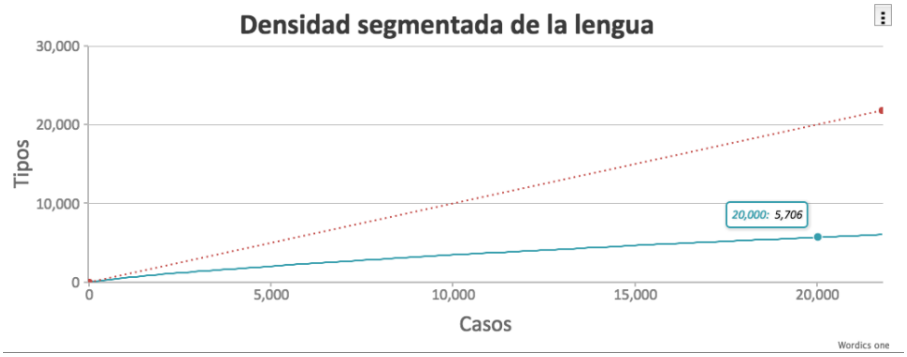

Figura 7. Densidad léxica segmentada de Arturo Pérez-Reverte

Daniel Sánchez Arévalo:

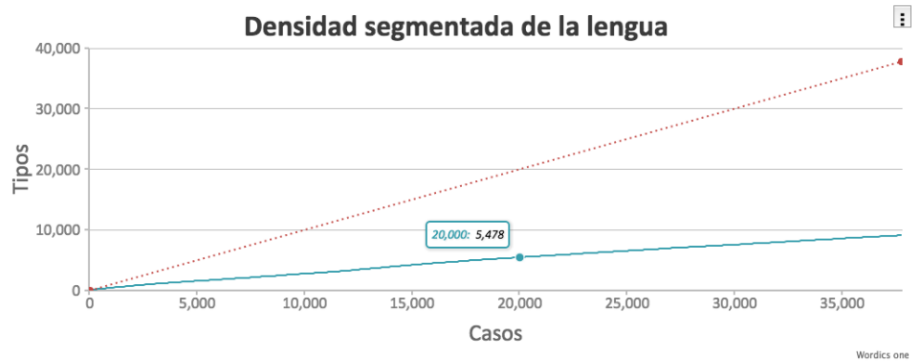

Figura 8. Densidad léxica segmentada de Daniel Sánchez Arévalo

Lucía Etxebarria:

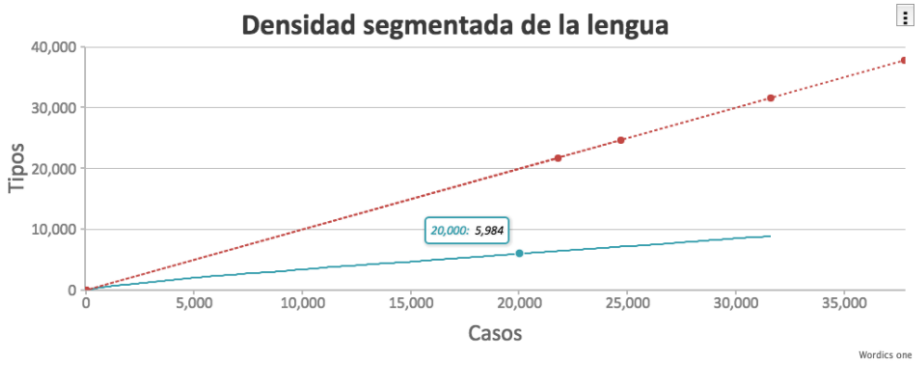

Figura 9. Densidad léxica segmentada de Lucía Etxebarria

En la figura siguiente, aparece una tabla a modo de resumen con los resultados obtenidos:

Datos	Carrillo	Pérez-Reverte	Sánchez Arévalo	Etxebarria
Casos	20.000	20.000	20.000	20.000
Tipos	5866	5706	5478	5984
Densidad	**29,3%**	**28,5%**	**27,39%**	**29,9%**

Figura 10. Comparación densidad léxica segmentada entre los autores analizados

Como era de esperar, los valores de la densidad léxica no solo varían con respecto a los valores totales, sino que todos ellos aumentan y, además, se acercan entre sí; lo que demuestra que, efectivamente, cuanto mayor sea longitud del texto que analicemos, menor será la variedad de palabras que éste contenga. Sin embargo, a pesar de la variación de porcentajes, el orden de los autores que mayor o menor densidad presentan no se ve alterado prácticamente, con la excepción de Lucía Etxebarria, que supera a Mónica Carrillo en seis décimas. Una vez más, por tanto, comprobamos que el hecho de tener más casos le supone, a la postre, que su valor de densidad se iguale al de Carrillo, a pesar de los casi 7000 casos de diferencia entre ambas autoras.

Conclusiones

Tras la realización de este estudio, nos preguntamos si la densidad léxica está relacionada con un lenguaje pobre, inexacto o incorrecto. Aunque, en principio, no tenemos la convicción de que exista necesariamente una relación directa entre estos dos aspectos, es posible que un lenguaje cuidado y sobre el que se ha reflexionado conduzca, casi de manera natural, a un aumento en la variedad de las palabras que lo componen. Esto redunda en una reflexión más profunda acerca de la lengua que tiene como resultado un uso más correcto de sus reglas.

En los escritores estudiados se cumple esta premisa con el autor que presenta menor densidad léxica, Sánchez Arévalo. Realizando un análisis más exhaustivo de sus producciones textuales, comprobamos que comete algunos usos no normativos, relacionados con el queísmo o algunas expresiones como *en base a*. Además, utiliza numerosos rasgos característicos del lenguaje coloquial, la expresión *en plan,* o representaciones de la risa (*jajajaja)*:

7 nov 2015, 13:39 h.	@faustianovich Te he tocado la patata? ;) Me alegro. (Me acabo de dar cuenta que este tweet nunca se

	envió. Qué cosas)
25 may 2014, 22:18 h.	Las radios del Spotify siempre empiezan bien, con buen criterio en base a tu selección, pero luego se les empieza a ir la olla cosa fina.
01 mar 2016, 21:09 h.	@Lucia_Alvarez_ Jajaja, qué maja! Llévatela a la isla a jugar con Noesunponi!
15 feb 2016, 21:24 h.	@YUSAN_5 Jajajaja. A mí también me falta lamentablemente mi talento
20 mar 2013, 20:57 h.	@SallyBurton @ariadnunii @quimyo A nosotros aunque vayamos de guays y tal en plan estrellitas del celuloide, también nos animais los días ;)
30 jun 2014, 23:37 h.	@iguardans Pues sí, deberían estar en plan SGAE, que te pillan siempre en cualquier lugar remoto del planeta donde pongas un chunda chunda

Como decimos, el uso de expresiones de tipo informal, de emoticonos, sufijos, anglicismos y expresiones coloquiales acercan la lengua de Sánchez Arévalo en *Twitter* a la oralidad mucho más que la del resto de escritores, en los que apenas encontramos este tipo de ejemplos. Por ejemplo, este autor utiliza la expresión “en plan” en 9 ocasiones, mientras que de los otros, solo Pérez Reverte lo utiliza, en 2 ocasiones y Etxebarria, 1. En cuanto a la expresión onomatopéyica de la risa, aparecen 53 ocurrencias en Sánchez Arévalo y 1 en Etxebarria.

Es posible que aquí resida la explicación a la menor variedad léxica del escritor, ya que los porcentajes de densidad varían según se trate de material oral o escrito. Ure (1971) y Halliday (1985) afirman que los textos con menor densidad son más sencillos de entender y, además, suelen pertenecer a la lengua oral, mientras que aquellos con mayor variedad son característicos de la lengua escrita. Sin embargo, ya hemos mencionado que el lenguaje en *Twitter* es un híbrido, que se encuentra a medio camino entre la oralidad y la escritura, con lo que sería difícil determinar con exactitud cuáles serían los valores estándar dentro de los cuales debería oscilar el porcentaje.

Consideramos fundamental, también, concluir la conveniencia de la utilización de *big data*, y más en concreto, de *Twitter*, para la investigación lingüística. Como podemos observar en este estudio, centrado en la densidad léxica, la cantidad de información que compone los corpora es considerablemente grande, mientras que el tiempo invertido para la compilación del corpus ha sido incomparablemente menor que si se hubiera

realizado de una forma más tradicional, ya que la herramienta ha generado de forma automática las listas de tuits que componen los corpora.

Referencias bibliográficas

ÁVILA, R. (1991): "Densidad léxica y adquisición del vocabulario: niños y adultos". En HERNÁNDEZ, C., *et al.* (Eds.), *El español de América. Actas del III congreso internacional del español en América, Valladolid, 3 a 9 de julio de 1989,* vol. 2, Castilla y León: Consejería de Cultura y Turismo, 621-630.

BERGMAN, M. & PAAVOLA, S. (Eds.). (2003): *The commens dictionary of Peirce's terms. Peirce's terminology in his own words* [en línea]. [ref. de 15 de diciembre de 2016]. Disponible en web: <http://www.commens.org/dictionary/term/token>.

DALLER, H., VAN HOUT, R. y TREFFERS-DALLER, J. (2003): "Lexical richness in the spontaneous speech of bilinguals". *Applied Linguistics*, 24/2, 197-222.

GANTZ, J. y REINSEL, D. (2011): Extracting Value from Chaos. *IDC iView*, 1-12 [en línea]. [ref. de 20 de diciembre de 2016]. Disponible en web: https://www.emc.com/collateral/analyst-reports/idc-extracting- value-from-chaos-ar.pdf.

GELBUKH, A. SIDOROV, G. y CHANONA, L. (2002): "Compilation of a Spanish Representative Corpus". En *International Conference on Computational Linguistics and Intelligent Text Processing CICLing02, Lecture Notes in Computer Science 2276*, Springer, 285-288.

GÓMEZ MOLINA, J. R. (2004): "Los contenidos léxico-semánticos". En SÁNCHEZ LOBATO, J., y SANTOS GARGALLO, I. *Vademécum para la formación de profesores. Enseñar español como segunda lengua (L2) / lengua extranjera (LE)*, Madrid: SGEL, 789-810.

GONZÁLEZ-FERNÁNDEZ, A. (2016): *Más allá del corpus: Big data en la investigación lingüística. Evolución, análisis y predicción del uso de la lengua a través de Twitter.* Tesis doctoral, Universidad de Córdoba, Córdoba, España.

GREGORY SIGNES, C. y CLAVEL ARROITIA, B. (2015): "Analysing lexical density and lexical diversity in university students' written discourse". En *Procedia: Social and Behavioral Sciences, Current Work in Corpus Linguistics: Working with Traditionally- conceived Corpora and Beyond. Selected Papers from the 7th International Conference on Corpus Linguistics (CILC2015)*, 198: 546-446.

HALLIDAY, M. A. K. (1985): *Spoken and written language*. Geelong Vict.: Deakin University.

HALLIDAY, M. A. K. (1993): "Quantitative studies and probabilities in grammar". En HOEY, M. (Ed.), *Data, description, discourse,* London: HarperCollins, 1-25.

JOHANSSON, V. (2008): "Lexical diversity and lexical density in speech and writing: a developmental perspective". *Working Papers*, 53, Lund University, Department of Linguistics and Phonetics, 61-79.

KILGARRIFF, A. (2001): "Web as corpus". En *Proceedings of the Corpus Linguistics Conference (CL 2001)*. *University Centre for Computer Research on Language Technical Paper* Vol. 13, Special Issue, Lancaster University, 342-344.

KILGARRIFF, A. y GREFENSTETTE, G. (2003): Introduction to the Special Issue on the Web as as Corpus. *Computational linguistics*, 29(3), 333-347.

LAUFER, B. y NATION, P. (1995): "Vocabulary Size and Use: Lexical Richness in L2 Written Production". *Applied Linguistics,* 16(3), 307-322.

LEECH, G. (2007): "New resources, oro just better old ones? The Holy Grail of representativeness". En HUNDT, M., NESSELHAUF, N., y BIEWER, C. (Eds.) *Corpus linguistics and the web,* Amsterdam: Rodopi, 132-150.

LEECH, G. y FALLON, R. (1992): "Computer Corpora. What do they tell us about Culture?". *ICAME Journal*, 16, 29-50.

MALVERN, D., RICHARDS, B., CHIPERE, N. y Duran, P. (2004): *Lexical diversity and language development: quantification and assessment*. New York: Palgrave Macmillan.

MERINO, M. (2014): ¿Qué es una API y para qué sirve?, *Ticbeat*. [en línea]. [ref. de 30 de enero de 2017]. Disponible en web: <http://www.ticbeat.com/tecnologias/que-es-una-api-para-que-sirve/>.

FUNDÉU BBVA. *Proyecto Aracne*. [en línea]. [ref. de 25 de febrero de 2017]. Disponible en web: <http://www.fundeu.es/aracne/>.

RAYSON, P., WALKERDINE, J., FLETCHER, W. H. y KILGARRIFF, A. (2006): Annotated web as corpus. En *Proceedings of the 2nd International Workshop on Web as Corpus,* Stroudsburg, PA: ACM, 27-33.

RENOUF, A. y KEHOE, A. (Eds.). (2006): *The changing face of corpus linguistics*. Amsterdam: Rodopi.

RODRÍGUEZ-TAPIA, S. (2016): "Clasificación cuantitativa de los textos según su grado de especialidad: parámetros para la elaboración de los índices de densidad terminológica y de reformulación de un corpus sobre insuficiencia cardíaca". *Anuario de Estudios Filológicos*, 39, 227-250.

TOGNINI-BONELLI, E. (2001): *Corpus linguistics at work*. Amsterdam: J. Benjamins.

URE, J. (1971): Lexical density and register differentiation. En Perren, J. E., y Trim, J. L. M. (eds.), *Applications of linguistics,* Cambridge: Cambridge University Press, 443-452.

ZIPF, G. (1965): *The Psycho-Biology Of Language*. Cambridge: MIT Press.