ELUA, 31, 2017, págs. 171-186 DOI: 10.14198/ELUA2017.31.09

ESTUDIO DE NEOLOGISMOS A TRAVÉS DE *BIG DATA* EN UN CORPUS TEXTUAL EXTRAÍDO DE TWITTER

STUDY OF NEOLOGISMS THROUGH *BIG DATA* IN A TEXTUAL CORPUS OBTAINED FROM TWITTER

Adela González Fernández
Universidad de Córdoba
adela.gonzalez@uco

Recibido: 01/09/2017
Aceptado: 12/10/2017

Resumen

El uso de la informática es cada vez más habitual en el trabajo lingüístico. 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 proponemos la utilización de *big data* y de *Twitter* para comprobar su utilidad a la hora de estudiar la formación y aparición de neologismos. Gracias a la creación de una herramienta informática diseñada específicamente para el trabajo lingüístico en *big data*, obtenemos una inmensa cantidad de datos textuales que nos servirán para la compilación de corpora. Estos datos no solo nos ofrecen información lingüística, sino también temporal y espacial. Mediante la selección de unos parámetros específicos, estudiaremos distintos términos con diferentes patrones de formación para comprobar cómo se forman, dónde y cuándo se introducen en la lengua y cómo y cuánto se utilizan.

PALABRAS CLAVE: neologismos; lingüística de corpus; *big data*; *Twitter*.

Abstract

The use of computers is more and more usual in linguistic research. Corpus linguistics, in particular, is benefiting from this matching, due to the improvements in management and processing of corpora. In this work we propose to test the usefulness of big data and Twitter as a resource to study the formation and appearance of neologisms in language. 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. This data not only provides us linguistic information, but also facts related to temporal and spatial information. Through the selection of some specific settings, we will study various terms with different formation patterns, in order to know how they are formed, when and where they are introduced in the language and how they are used.

KEYWORDS: neologisms; corpus linguistics; big data; *Twitter*.

Para citar este artículo / To cite this article: González Fernández, Adela (2017). Estudio de neologismos a través de big data en un corpus textual extraído de Twitter. *ELUA*, 31: 171-186. doi: 10.14198/ELUA2017.31.09

Enlace / Link: http://dx.doi.org/10.14198/ELUA2017.31.09

1. INTRODUCCIÓN

Parece difícil pensar, actualmente, en la investigación lingüística aplicada sin el apoyo de la informática en cualquiera de sus vertientes. Sin embargo, la unión entre ambas disciplinas es relativamente reciente. Son tres las etapas por las que, según Tognini-Bonelli (2001), han pasado desde sus inicios: 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 más 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 ha 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.

También Renouf y Kehoe (2006) afirman 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 la línea de estos autores, proponemos una metodología novedosa que va más allá 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 (González Fernández 2016: 92). *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 un mayor conocimiento y en 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 por 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.

En esta investigación nos proponemos utilizar tanto las técnicas de análisis de *big data* como la información textual que nos ofrece *Twitter* para comprobar la utilidad de ambos en el estudio de la lengua, en concreto, de los neologismos. Para ello, se ha utilizado un software informático diseñado y desarrollado específicamente para este tipo de investigación[1].

Uno de los objetivos de esta herramienta es poder crear uno o varios corpora con todos estos datos que sirvan de base para la investigación lingüística y que aporten la mayor cantidad y variedad de información posible para poder sacar el máximo partido al concepto de A3 –*anytime, anywhere, anybody*. Para la construcción de los corpus, obtenemos la información directamente de la API[2] de *Twitter*, que sirve de interfaz entre este y la herramienta, y facilita la interacción humano-software. Además, dispone de distintos módulos que nos ofrecen información en tiempo real, desde un punto de vista histórico o a través de cuentas específicas de usuarios.

Esta investigación pretende comprobar la utilidad de *Twitter* para la investigación lingüística y, más concretamente, para conocer el comportamiento y la evolución de los neologismos en un lugar determinado y en un tiempo (incluso real) que se establezca. Para este propósito, aportaremos solo algunos ejemplos de este fenómeno en la lengua española que siguen distintos patrones de formación.

Conviene insistir, por tanto, en la idea de que no es nuestra intención elaborar un estudio detallado acerca de los neologismos en español, el modo, lugar o tiempo de penetración en nuestra lengua o su desarrollo, sino aportar una serie de ejemplos que validen nuestra hipótesis. Por este motivo, no consideramos oportuno profundizar en consideraciones teóricas acerca de la neología y de los neologismos[3]. Nos limitaremos, por tanto, a hacer referencia al estado de la cuestión en la investigación neológica y a ofrecer algunas ejemplos que demuestren la conveniencia de la utilización de *big data* para el conocimiento lingüístico.

Consideramos que entre las virtudes de este recurso se encuentra la de servir como apoyo a las últimas corrientes en la investigación neológica en español, liderada por la Red de Observatorios de Neología del Castellano (NEOROC)[4] y cuyo interés principal se

1 La información completa acerca de este *software* se puede encontrar en González Fernández (2016).

2 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).

3 Algunas obras de referencia en este ámbito son Bastuji (1974), Guilbert (1975), Dubois (1979), Rondeau (1984), Cabré (1993), Alvar Ezquerra (1993), Guerrero Ramos (1995) o Díaz Hormigo (2004a, 2004b, 2008, 2010, 2015).

4 NEOROC está coordinada por el Observatori de Neologia del Institut de Linguistica Aplicada (IULA) de la Universitat Pompeu Fabra (http://www.iula.upf.edu/rec/neoroc) y la conforman una serie de nodos, entre los que se encuentran el Grupo de Investigación en Neología de la Universidad de Cádiz (NEOUCA) y las universidades de Málaga, Valencia, País Vasco, Salamanca (http://neousal.usal.es/), Murcia y Alicante, además del propio IULA (Díaz Hormigo, 2015).

centra en la detección, la selección, el análisis, el almacenamiento, la difusión y el estudio contrastivo de la neología léxica en las distintas variedades del español de la Península, como explica Díaz Hormigo (2015). Nuestro software nos permite elaborar un corpus (más o menos amplio, a voluntad) de neologismos que aporte información, una vez más, acerca de dónde, cuándo y cómo se utilizan las palabras de nueva formación que, además, puede ser complementado con otros dos módulos que presenta la herramienta: el del estudio en tiempo real y el de usuarios concretos de *Twitter*. En el primer caso, porque, como explica Rondeau: "el concepto de neología es esencialmente diacrónico, porque está ligado al dinamismo de las lenguas vivas, en constante evolución a pesar de la impresión de estabilidad que tienen de ella los sujetos hablantes" (Rondeau 1983: 121, *apud* Fuentes *et allii*. 2009: 106). Y, en el segundo caso, para poder determinar el ámbito en el que se crean estas palabras (por ejemplo, NEOROC estudia la producción de neologismos en los medios de comunicación).

2. METODOLOGÍA

El criterio fundamental que se ha seguido para decidir el carácter neológico de una palabra determinada ha consistido en comprobar si la unidad en cuestión analizada se encuentra recogida en el diccionario de referencia de la Real Academia Española (DRAE, 23a ed.). Además, puesto que el criterio lexicográfico ya no se considera el único para reconocer una unidad léxica como neológica (Estornell, 2009), hemos tenido también en cuenta su uso y su comportamiento en el corpus. Nos proponemos aportar algunos ejemplos de neología léxica, siguiendo la clasificación de Guerrero Ramos (1995).

Como hemos apuntado en el apartado anterior, la herramienta, diseñada específicamente para el análisis lingüístico de la información textual que contiene *Twitter*, ofrece al investigador la posibilidad de extraer material lingüístico para su posterior análisis. En este sentido, son varias sus funcionalidades, como, por ejemplo, realizar consultas por idiomas, por regiones geográficas, por fechas, por unidades léxicas y por expresiones regulares. Esto hace posible que el investigador pueda introducir sus criterios de búsqueda e incluso, en la investigación sobre neologismos, lanzar búsquedas según el proceso de formación de palabras.

Puesto que tratamos de estudiar solo algunos casos de palabras nuevas, no hemos llevado a cabo un proceso de vaciado de estas, según las recomendaciones de Cabré *et allii*. (2004). La metodología seguida ha consistido en seleccionar distintas opciones de análisis (simple o comparado) –dependiendo de las necesidades en cada caso– que facilita la herramienta a la hora de introducir palabras en el buscador. En la opción de análisis simple, la herramienta lanza una sola búsqueda sobre la unidad que se quiere consultar y devuelve los resultados correspondientes a dicha búsqueda. Por el contrario, el análisis comparado permite realizar dos o más búsquedas simultáneas, de manera que los resultados de las distintas consultas aparecen de forma conjunta para que se puedan establecer relaciones y comparaciones entre ellos.

En nuestro caso, a pesar de que la herramienta lo permite, no ha sido necesario utilizar filtrado por coordenadas, puesto que nuestro objetivo consiste en comprobar la utilización de determinadas palabras en la lengua española a nivel mundial. Por este motivo, sí hemos hecho uso de la selección de idioma.

Dado que nuestro objetivo es seguir la evolución de un término concreto a lo largo del tiempo, la función utilizada ha sido la de búsqueda histórica en *Twitter*, puesto que el análisis en tiempo real no nos aportaría la información deseada. El rango de fechas que se ha establecido para la búsqueda de la información ha comprendido desde el 15 de septiembre de 2015 hasta el 15 de marzo de 2016.

A pesar de que la herramienta ofrece la opción de obtener, para cada palabra, la información numérica en cuanto al índice de frecuencias –detallada por idiomas–, el corpus completo de tuits, las KWIC –*Key Word in Context*– (palabras clave en contexto) y las colocaciones de cada palabra, nos hemos limitado a extraer la gráfica con la frecuencia de uso y algunos los ejemplos reales de utilización del término en cuestión, ya que sería inabarcable aportar todos los ejemplos extraídos, teniendo en cuenta que estamos trabajando con cantidades masivas de datos.

Las palabras neológicas estudiadas se han seleccionado con el objetivo de ejemplificar diversos tipos de neologismos creados según los criterios de la profesora Guerrero Ramos (1995) y que presentan una cierta frecuencia de uso en español. Los neologismos estudiados son:

(1) Poliamor
(2) Beticismo, sevillismo y madridismo
(3) Preferentista
(4) Trolear y troleo
(5) Posturear y postureo
(6) Veroño
(7) Juernes
(8) Brexit
(9) Googlear
(10) Spoiler
(11) Selfi y selfie

A continuación, mostramos los gráficos resultantes de la búsqueda para cada palabra y los contextos de uso de aquellos términos que hemos considerado relevantes o con un significado todavía poco asentado.

3. RESULTADOS

Dentro de los denominados "neologismos de forma", Guerrero Ramos (1995) incluye la creación de palabras por combinación de elementos léxicos existentes o, lo que es lo mismo, los procesos tradicionalmente conocidos como composición y derivación. Sin entrar en cuestiones polémicas acerca de la consideración de los prefijos como medios para obtener una palabra derivada o compuesta[5], y teniendo en cuenta que la prefijación es uno de los procedimientos más frecuentes en la formación de palabras, mostramos como caso particular el uso del prefijo *poli* para la formación de la palabra *poliamor* (Figura 1):

5 Se puede consultar Alvar Ezquerra (1993), Lang (1992), Varela Ortega (2005) o Díaz Hormigo (2015), para ahondar más en esta cuestión.

Figura 1. Gráfica de la frecuencia de uso de *poliamor*.

Este término, referido a una relación sentimental entre más de dos personas, sigue el proceso de creación morfológica propio del idioma español y el modelo de palabras como *politraumatismo* o *politeísta*. Encontramos 26 apariciones de la palabra en total, en el tiempo preestablecido. A continuación podemos observar un ejemplo (ejemplo 1):

(1) a. 8 oct 2015, 19:20:41: No puedo dar mi opinión sobre el poliamor porque aún no sé a fondo en qué consiste, pero no, NO ES MALO.
b. 8 oct 2015, 19:21:22: Por lo que he leído, el poliamor consiste en tener relaciones con más personas en vez de tener solo una.

De la misma manera que sucede con la prefijación, la sufijación es otro de los recursos más importantes en español a la hora de creación de nuevas palabras. En el ejemplo que mostramos a continuación (Figura 2), añadiendo un sufijo se han formado palabras nuevas que mantienen la misma categoría gramatical que la palabra de la que proceden. Concretamente, en este caso se han formado sustantivos a partir de otros sustantivos, a los que se ha añadido el sufijo *–ismo*. Puesto que hemos estudiado tres términos de similar formación, hemos utilizado la opción de análisis comparado que nos proporciona la herramienta para introducir en el buscador, de forma simultánea, las palabras *beticismo, madridismo* y *sevillismo*, referidas a las aficiones de los equipos de fútbol del Real Betis Balompié, Real Madrid y Sevilla Fútbol Club, respectivamente. Así, hemos obtenido una gráfica con la frecuencia de aparición de los tres términos, en la que la línea gris lisa representa a *beticismo*; la rayada, a *madridismo*; y la negra, a *sevillismo*, como se puede observar en la leyenda:

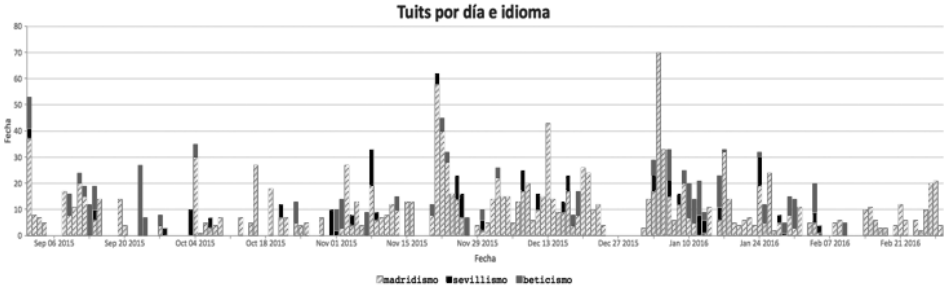

Figura 2. Gráfica de la frecuencia de uso de *madridismo*, *sevillismo* y *beticismo*.

Los resultados numéricos en español son:

– *Beticismo*: 288 apariciones.
– *Madridismo*: 1269 apariciones.
– *Sevillismo*: 163 apariciones.

Veamos algunos ejemplos de uso (ejemplo 2):

(2) a. 1 sept 2015, 00:47:43: Que grande!!! Gran ilusión para todo el beticismo! #LaMejorAficion https://t.co/wy5wZjDVgc
b. 1 sept 2015, 22:16:21: Yo no sé qué pasará en el campo, pero sí sé que Joaquín ha devuelto mucha ilusión al beticismo. Y eso, amigos, es otra cosa. @elpelotazocsr
c. 25 ene 2016, 00:05:26: Bonito día para el beticismo. Ya era hora! https://t.co/sa5Xr7Cvko
d. 1 sept 2015, 20:23:36: @alexcibernetica La prensa me temo que no está con él y hay una parte muy importante del madridismo muy sensible a lo que la prensa dice.
e. 9 feb 2016, 00:37:06: que gozada ver el madridismo patas arriba #ChiringuitoMadrid
f. 3 ene 2016, 22:24:27: El aficionado del fútbol ha vivido un partidazo pero el madridismo ha vivido un atraco en los que no se han pitsdo 3 penaltis claros a favor
g. 8 nov 2015, 22:17:54: Gran Sevilla ante un Madrid mediocre. Las críticas del sevillismo a su equipo son desmesuradas, un año más.
h. 4 feb 2016, 22:55:49: Sevillismo puro! La giralda presume orgullosa...? https://t.co/hLx2pSzVAl
i. 4 feb 2016, 22:55:49: Sevillismo puro! La giralda presume orgullosa...? https://t.co/hLx2pSzVAl

Un nuevo ejemplo de sufijación nominal pero, en este caso, con cambio de categoría gramatical, se trata del término *preferentista*, referido a aquellas personas en posesión de participaciones preferentes en una entidad bancaria, del que encontramos cinco apariciones el día 6 de octubre de 2015 (ejemplo 3):

(3) a. 6 oct 2015, 10:22:50: En la España de las raíces vigorosas, puedes morir como un preferentista arruinado o ser recibido por el ministro del Interior por "amenazas
b. 6 oct 2015, 13:10:04: Acabo de escuchar a un preferentista estafado decir: "Ya que son mayoría en el gobierno lo van a ser tb en Soto del Real". #CárcelPaRato
c. 6 oct 2015, 14:18:49: Preferentista q le han estafado 58.000 €, ha pedido verse con Pedro Sánchez #PSOE y no la recibe. Como cordero que quiere verse con el lobo.

No faltan tampoco otros tipos de sufijación, como la verbal, con sufijos del tipo *–ear*; es el caso, por ejemplo, del término *trolear*, que procede de la palabra *trol* y que significa, en el ámbito de Internet:

> acción y al efecto de intervenir en un foro digital con el objetivo de generar polémica, ofender y provocar de modo malintencionado a los demás usuarios, a menudo enviando multitud de mensajes que pretenden captar la atención e impedir el intercambio o desarrollo habitual de dicho foro. (Fundéu BBVA, s.f.)

Otros significados aportados también por la Fundéu, de carácter más general, se refieren a "intervenir con ánimo de hacer fracasar algo", como concepto de *boicotear* o *provocar*, y "tomar el pelo, vacilar o gastar una broma, por lo general pesada". También procedente de la misma palabra ha aparecido la forma sufijada *troleo*, para crear nuevo un sustantivo con el mismo lexema. Mostramos a continuación (Figura 3) un ejemplo comparado de ambos términos en el que se pueden observar sus frecuencias de uso (*trolear* aparece en color negro y *troleo* está representado con la barra a rayas):

Figura 3. Gráfica de la frecuencia de uso de *troleo* y *trolear*.

En total, el verbo (*trolear*) se utiliza en 21 ocasiones en el período de tiempo indicado (ejemplo 4), mientras que el sustantivo (*troleo*) lo hace en 57 (ejemplo 5). Observemos algunos de los contextos en los que se utilizan ambos términos:

(4) a. 23 sept 2015, 23:19:23: He sido formado por mi tío el calvo en el arte de trolear y vacilar
b. 13 nov 2015, 23:41:30: Esa cuenta es falsa, y precisamente lo que buscan es trolear. Yo en tu lugar no los mencionaria ;) No valen la pena @Pienso1ro
c. 9 dic 2015, 22:20:27: Mola esto de trolear de vez en cuando a los fanáticos PPSOE y alguno de Cuñadanos. Risas aseguradas. Sonríe #SiSePuede

(5) a. 9 sept 2015, 22:30:29: Se esta quedando con todos vosotros, vaya troleo jajajajajajaja #CantizanoEH @ElHormigueroMx
b. 8 dic 2015, 16:18:10: La estrategia de troleo de @ahorapodemos...o si no convencenos manipulamos. Vieja tactica de los años 30 https://t.co/ioNHiso4UM
c. 8 dic 2015, 14:23:48: @subversivos_ como se nota el troleo, pensar que el lector conservador del ABC voto por el koletas, roza el delirio

Exactamente los mismos patrones de comportamiento que *trolear* y *troleo* siguen los neologismos *posturear* (ejemplo 6) y *postureo* (ejemplo 7), creados a partir de sufijación verbal (*–ear*) y nominal (*–eo*), obteniendo, en este último caso, un nuevo sustantivo sobre la base del sustantivo inicial *postura*. No obstante, la utilización de ambos términos es enormemente dispareja, puesto que *posturear* únicamente se utiliza 5 veces en la franja de fechas establecida, frente a las 10.178 de *postureo* (Figura 4):

Figura 4. Gráfica de la frecuencia de uso de *postureo* y *posturear*.

Podemos observar algunos ejemplos de uso a continuación:

(6) a. 11 nov 2015, 13:25:30: os quejáis de que sólo se habla de Francia y vosotros estáis hablando hoy más que nunca de Siria sólo por posturear
b. 14 nov 2015, 15:04:02: Es muy gracioso de verdad, ¿no sería más fácil que hicieseis algo por ayudar en vez de posturear con hastags?
c. 14 nov 2015, 16:07:11: Yo Postureo del verbo posturear #YoPostureo #MiguelitoStyle #DeMayorQuieresSerComoYo #Malaga #Style #Sabado https://t.co/Tz8PEJo4qB

(7) a. 1 sept 2015, 13:38:34: Con la edad que tienes y te comportas como una cría de 15 años. Que te gusta la tele y el postureo @shailagal
b. 1 sept 2015, 18:14:59: Ahora mismo esto llega a niveles de postureo que ni la Nasa conoce... http://t.co/9TnnbXoDsp
c. 5 sept 2015, 00:45:42: Me da coraje toda esa gente que se compra cámaras reflex por postureo

Los mecanismos de formación de palabras nuevas a partir de la acronimia, es decir, a través del truncamiento de las voces que forman un término (Guerrero Ramos 1995: 35) también son abundantes en nuestro idioma. Tenemos, por ejemplo, los casos de *veroño* o *juernes*. El primero de ellos –*veroño*– surge de la unión de *verano*+*otoño*; *juernes*, de la misma manera, es la suma de las palabras *jueves*+*viernes*. La figura siguiente (Figura 5) muestra un total de 152 apariciones de *veroño*:

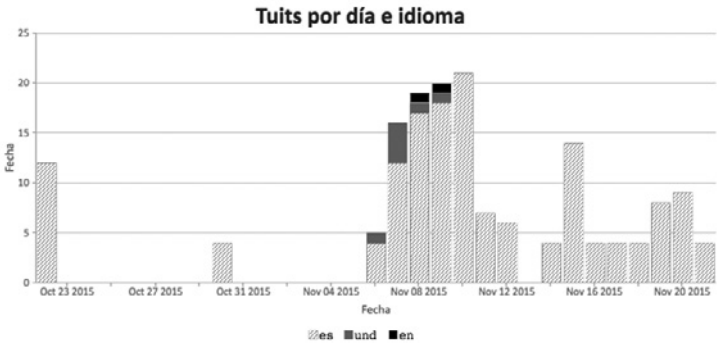

Figura 5. Gráfica de la frecuencia de uso de *veroño*.

Estos son algunos ejemplos reales de uso de este término (ejemplo 8):

(8) a. 30 oct 2015, 14:55:35: A partir de ahora tenemos cinco estaciones. Queda inaugurado el Veroño! ¡Ozú Qué calor más grande! @jotaerrepp_70 @lalibretacolora
b. 30 oct 2015, 12:13:50: Hemos creado para una clienta una camiseta de gasa y punto, perfecta para disfrutar del #veroño!!! 🌻🔆🍂🍁 https://t.co/waDVdkgY9R
c. 9 nov 2015, 19:58:26: Quien dice Noviembre? En Málaga seguimos con el veroño ..y parece que se quiere quedar #DíasCalurosos

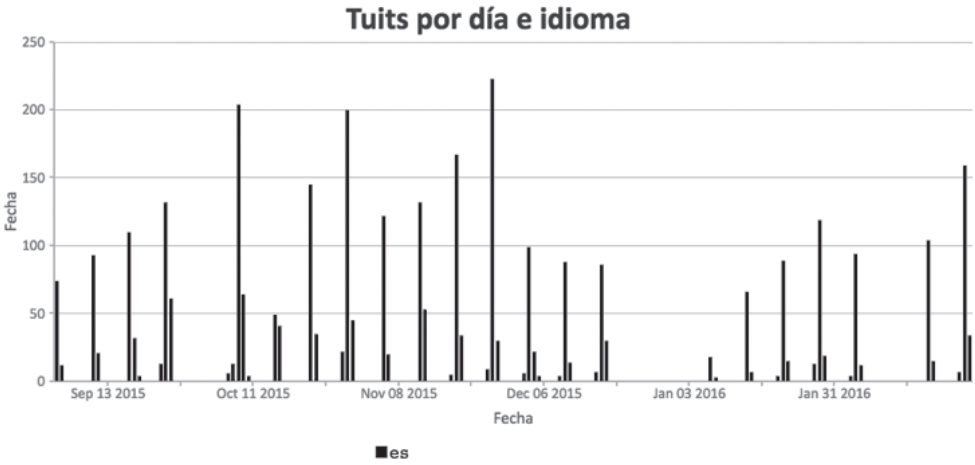

Figura 6. Gráfica de la frecuencia de uso de *juernes*.

Resulta muy llamativa la gráfica de *juernes* (Figura 6) si tenemos en cuenta que todos los máximos relativos que aparecen en ella coinciden con los distintos jueves del año. Además, podemos observar cómo la utilización del término se reduce considerablemente durante las vacaciones navideñas. Este hecho nos da una idea del tipo de usuarios que suele utilizar esta palabra en *Twitter*: estudiantes universitarios que salen a divertirse los jueves como si de viernes se tratara, puesto que no suelen tener clase al día siguiente, y que durante el período vacacional no acusan la diferencia entre un día y otro porque pueden salir a divertirse cualquier día.

Algo parecido ocurre con *veroño*, que se utiliza mucho más en la época de la estación de otoño en la que se presuponen temperaturas más bajas que las que se están produciendo en ese momento. En el período analizado registramos un total de 3317 apariciones de *juernes*. A continuación vemos algunos ejemplos concretos (ejemplo 9):

(9) a. 3 sept 2015, 20:17:50: Alguien sale de juernes
b. 4 sept 2015, 13:20:27: Juernes con mis amores 💅❤ @ Bora Bora Polinesian Bar https://t.co/l3ulh1IGHt
c. 10 sept 2015, 09:29:01: Menudo juernes nos espera

Mediante el mismo procedimiento de formación neológica denominado acronimia se están introduciendo en nuestro idioma nuevas palabras, pero ya procedentes de idiomas

extranjeros –fundamentalmente el inglés. El préstamo es, precisamente, "uno de los medios fundamentales de cualquier lengua para su enriquecimiento neológico", como afirma Guerrero Ramos (1995: 36). El idioma español, como cualquier otra lengua, está acogiendo constantemente términos procedentes de otros idiomas, dentro de los cuales predomina manifiestamente el inglés.

Estos prestamos se comportan de distintas maneras y también tienen orígenes diversos, aunque la mayoría de ellos también son neologismos en su lengua de origen formados por mecanismos similares a los nuestros. Es, por ejemplo, el caso de *brexit*: el término utilizado desde hace unos meses para nombrar la salida de Gran Bretaña de la Unión Europea y formado a partir de las palabras *Britain+exit*. Resulta llamativo cómo aspectos relevantes de la actualidad condicionan el uso de palabras de este estilo, que surgen por la necesidad de nombrar una nueva realidad de gran relevancia en el panorama político internacional y comienzan a usarse con la naturalidad y frecuencia que exija la realidad que les ha dado origen.

Desde el 1 de septiembre hasta el 19 de febrero, la herramienta registró 15.988 ocurrencias de *brexit* en todo el mundo, de las cuales 252 fueron en español (Figura 7):

Figura 7. Gráfica de la frecuencia de uso de *brexit* en español.

El mismo mecanismo de formación ha sufrido la palabra *nomophobia*, que ha sido adaptada a la ortografía española con la forma gráfica *nomofobia*. Este término, con el significado de "pánico a encontrarse sin teléfono móvil" surgió, en inglés, de la unión de las formas *no* + *mobile phone* + *phobia*, de manera que formó un acrónimo que ha sido trasladado y adaptado enseguida al español.

Otros términos de creación neológica han aparecido en español mediante sufijación aplicada a palabras de origen anglosajón, como es el caso de los verbos *whatsappear* (*WhatsApp* + *-ear*), con sus múltiples variantes –*wasapear*, *whasapear*, *wasear, guasapear, guasear*, etc.– y derivados –*whatsapeo*, *whasapeo*, *wasapeo*, etc.–, o *googlear* (*guglear*). Para el caso de *WhatsApp* y todas sus variantes, hemos utilizado la opción de "elegir palabra" que aparece en la herramienta, es decir, hemos llevado a cabo una búsqueda mediante expresiones regulares. Para ello, se ha introducido la secuencia "w*s*p*" para determinar los criterios restrictivos de constitución de palabras. Esto quiere decir que la herramienta ha buscado

todas las palabras del corpus que contengan las letras que aparecen en la secuencia más una o varias letras en los lugares donde hay asteriscos. El resultado ha sido 1634 palabras que cumplen esta característica, de las cuales 35 están relacionadas con *WhatsApp*, lo que demuestra la enorme variedad formal que presenta todavía este término debido a su reciente introducción en español. *Googlear*, por el contrario, parece estar más asentada; mostramos a continuación (Figura 8) la evolución de esta palabra:

Figura 8. Gráfica de la frecuencia de uso de *googlear*.

Este término aparece con una frecuencia de 27 apariciones en español, en el período de tiempo analizado. Algunos de sus contextos de uso son los siguientes (ejemplo 10):

(10) a. 24 feb 2016, 14:37:23: Está materia me gusta tanto que no voy a llegar a rendir porque me detengo a googlear todo lo que me resulta interesante.
b. 25 feb 2016, 22:30:55: Tuve que googlear quien/que es Silvia Peyrou porque me daba cosa confundirla con la que trabaja en la mercería de mi barrio.
c. 29 feb 2016, 18:51:48: Voy a googlear a Axel para aprenderme una canción asi la puedo tararear mañana a la mañana.

Otros términos, también procedentes del inglés, se están introduciendo en español sin cambios ortográficos ni fonológicos –en términos generales y en la medida de lo posible–, como ocurre con *spoiler* o *bullying*, que se trata de extranjerismos no adaptados.

De *spoiler*, de hecho, encontramos 4686 apariciones en español, como podemos comprobar en la Figura 9:

Figura 9. Gráfica de la frecuencia de uso de *spoiler* en español.

Este término (*spoiler*) se usa para referirse a la revelación de contenido sustancial de una trama de una novela, serie o película y que puede acabar con el interés de quien lo sigue (ejemplo 11):

(11) a. 1 sept 2015, 00:16:04: Cuando tu madre te hace el spoiler del último capítulo de la mejor serie del mundo... 😭👋🔫 #breakingbad
b. 2 sept 2015, 03:43:51: Quiero un spoiler sobre lo que va a pasar con el mundo
c. 2 sept 2015, 14:37:22: Creo que Aomine va a explica ahora algo de la Ultimate Zone. Lo intuyo porque ya me hicieron el spoiler xd.

A pesar de que este estudio ha consistido en una pequeña muestra del uso en *Twitter* de algunos recientes neologismos del español, no podemos concluir sin referirnos al término de procedencia inglesa que, a pesar de no aparecer aún recogido en el DRAE (23a ed.), fue nombrado la palabra del año 2014 por la Fundéu BBVA: *selfie* y su adaptación a la ortografía española *selfi*. Comprobamos, mediante el análisis comparado de la herramienta, cómo todavía sigue predominando en español la forma gráfica inglesa, con 11.554 apariciones (Figura 10), frente a las 111 de la forma adaptada al español, *selfi* (Figura 11):

Figura 10. Gráfica de la frecuencia de uso de *selfie* en español.

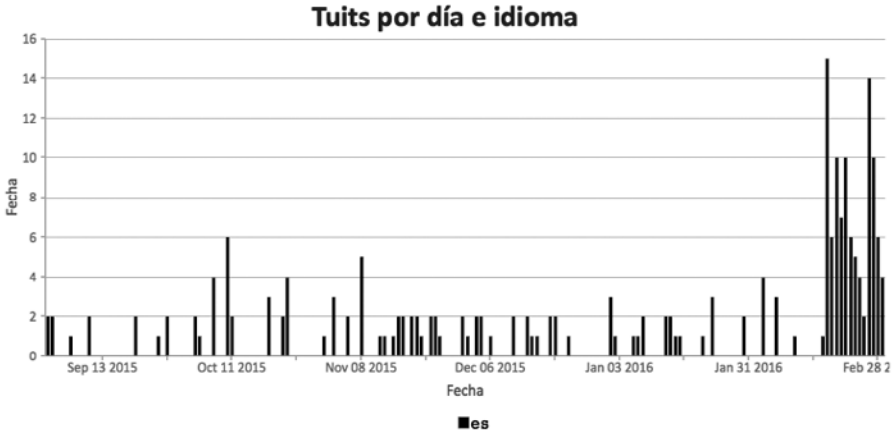

Figura 11. Gráfica de la frecuencia de uso de *selfi* en español.

Pese a la enorme diferencia registrada en cuanto al número de apariciones del *selfie* y *selfi*, podemos observar que la línea de tendencia que sigue la gráfica es similar en ambos casos. El notable incremento de tuits generados se explica porque a partir de mediados de febrero de 2016 el sistema se escaló para poder procesar los tuits publicados en todo el mundo, mientras que, hasta ese momento, solo analizaba los generados en Europa.

4. CONCLUSIONES

En general, el trabajo con *big data* está a la vanguardia de la mayoría de los ámbitos de investigación. Esto se debe, fundamentalmente, a las ventajas que este nuevo concepto nos aporta y a los avances con respecto a la metodología tradicional de investigación. En primer lugar, las cantidades masivas de datos, a las que antes no se tenía acceso, suponen una fuente de información de un enorme valor; pero, además, el tipo de información al que accedemos es de una naturaleza distinta y posee propiedades de las que hasta el momento no podíamos obtener rendimiento. En el trabajo lingüístico, en concreto, y gracias la herramienta a la que hemos hecho alusión, podemos obtener datos como las coordenadas espaciales y temporales exactas de emisión de un enunciado lingüístico en una gran cantidad de idiomas o la visualización en tiempo real de una producción concreta. Claro está, que intentar abordar un trabajo y analizar información proveniente de la totalidad de la información que circula en Internet es descabellado y poco productivo, debido no solo al volumen de la información, sino a su estructura y a su procedencia. Este es uno de los motivos por los que se ha seleccionado *Twitter* como fuente de información. El otro motivo principal es, lógicamente, el carácter eminentemente lingüístico de esta plataforma de comunicación social, que hace de ella una fuente inagotable de información, además de su utilización por gran parte de la población mundial como herramienta de comunicación y de información.

Sin embargo, como ya hemos apuntado, no se puede acometer una investigación de estas características con las técnicas de trabajo tradicionales. Por ello hemos desarrollado esta herramienta informática que es capaz de extraer, almacenar, gestionar y analizar la información obtenida.

Tras analizar los resultados, creemos que es posible utilizar *Twitter* para poder obtener información acerca de los comportamientos que en cuestión neológica adopta una lengua determinada, que en este caso ha sido el español. Además, consideramos que, gracias a investigaciones de este tipo, es posible ayudar a establecer las tendencias predominantes en los procesos de formación de palabras nuevas y conocer sus mecanismos de formación, además de contribuir a la actualización de las teorías actuales sobre la formación de neologismos y aportar evidencias para la actualización de obras lexicográficas. Tareas, todas ellas, comprendidas dentro de los objetivos de las investigaciones desarrolladas en el marco del NEOUCA enumerados por Díaz Hormigo (2015).

Referencias bibliográficas

Alvar Ezquerra, M. (1993). *La formación de palabras en español*. Madrid: Arco/Libros.

Bastuji, J. (1974). "Aspects de la néologie sémantique". En Guilbert, L. *et al*. (eds.). *La néologie lexicale. Langages*, 36, pp. 6-19.

Cabré, M. T. (1993). *La terminología. Teoría, metodología, aplicaciones*, traducción de Carles Tebé. Barcelona: Antártida/Empuries.

Cabré, M. T. *et allii*. (2004). *Metodología del trabajo en neología: criterios, materiales y procesos*. Barcelona: Universitat Pompeu Fabra. Papers de l'IULA. Sèrie Monografies, 9: http://www.iula.upf.edu/04mon009.htm (30-01-17)

Díaz Hormigo, M. T. (2004a). "Restricciones del sistema y restricciones de la norma en la formación de palabras", *Lingüística en la Red II*: http://www.linred.com (6-1-17)

Díaz Hormigo, M. T. (2004b). "Neología y tecnología: a propósito de los programas de detección automática de neologismos", *Español Actual*, 82, pp. 116-119.

Díaz Hormigo, M. T. (2008). "La investigación lingüística de la neología léxica en España. Estado de la cuestión", *LynX. Panorámica de estudios lingüísticos*, 7, pp. 5-60.

Díaz Hormigo, M. T. (2010). "Revisión historiográfica de los conceptos "neología" y "neologismo"". En Assunção, C., G. Fernandes y M. Loureiro, (eds.). *Ideias Linguísticas na Península Ibérica (séc. XIV a séc. XIX)*. Münster:Nodus Publikationen, I, pp. 167-176.

Díaz Hormigo, M. T. (2015). "Neología aplicada y lexicografía para la (necesaria) actualización de las entradas de los elementos de formación de palabras en diccionarios generales", *Revista de lingüística y lenguas aplicadas*, 10, pp 12-20.

Diccionario de la Real Academia Española (DRAE), 23a Edición. (2014). Real Academia española: http://www.rae.es/ (30-09-15)

Dubois, J. (Coaut.). (1979). *Diccionario de lingüística.* Madrid: Alianza.

Estornell Pons, M. (2009). *Neologismos en la prensa: criterios para reconocer y caracterizar las unidades neológicas*. Valencia: Universitat de València.

Fuentes, M., S. Gerding, S. Constanza, A. Pecchi, G. Kot. y P. Cañete. (2009). "Neología léxica: reflejo de la vitalidad del español de Chile", *RLA, Revista de lingüística teórica y aplicada*, 47(1), pp. 103-124.

Fundación del Español Urgente. (2016). *Fundéu BBVA*: http://www.fundeu.es/ (30-09-15)

Gantz, J. y D. Reinsel. (2011). "Extracting Value from Chaos", *IDC iView*, pp. 1-12: https://www.emc.com/collateral/analyst-reports/idc-extracting- value-from-chaos-ar.pdf (15-08-15)

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.

Guerrero Ramos, G. (1995). *Neologismos en el español actual*. Madrid: Arco/Libros.

Guilbert, L. (1975). *La créativité lexicale*. Paris: Larousse.

Halliday, M. A. K. (1993). "Quantitative studies and probabilities in grammar". En Hoey, M. (ed.), *Data, description, discourse*. London: HarperCollins, pp. 1-25.

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, pp. 342-344: http://ucrel.lancs.ac.uk/publications/CL2003/CL2001%20conference/papers/kilg arri.pdf (30-09-15)

Kilgarriff, A. & Grefenstette, G. (2003). "Introduction to the Special Issue on the Web as as Corpus". *Computational linguistics*, 29(3), pp. 333-347.

Lang, M. F. (1992). *Formación de palabras en español. Morfología derivativa productiva en el léxico moderno*. Madrid: Cátedra.

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

Leech, G. (2007). "New resources, oro just better old ones? The Holy Grail of representativeness". En Hundt, M., N. Nesselhauf y C. Biewer, (eds.) *Corpus linguistics and the web.* Amsterdan: Rodopi, pp. 132-150.

Merino, M. (2014). "¿Qué es una API y para qué sirve?" *Ticbeat*: http://www.ticbeat.com/tecnologias/que-es-una-api-para-que-sirve/ (30-01-16)

Renouf, A. y A. Kehoe (eds.). (2006). *The changing face of corpus linguistics*. Amsterdam: Rodopi.
Rondeau, G. (1984). *Introduction à la terminologie*. Chicoutimi (Québec): Gaëtan Morin.
Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: J. Benjamins.
Varela Ortega, S. (2005). *Morfología lexica: la formación de palabras*. Madrid: Gredos.