ISSN: 2255-3703

Big data para el análisis de las necesidades traductológicas en cinco capitales de Europa

Adela González Fernández
Universidad de Córdoba
l52gofea@uco.es

Fecha de recepción: 12.12.2015
Fecha de aceptación: 30.01.2016

Resumen: La cantidad de información digital disponible está aumentando de forma masiva. Las grandes empresas como *Google*, *Facebook*, *Microsoft*…, así como los correos electrónicos, las descargas de música o cualquier operación realizada a través de internet genera cantidades gigantescas de información no estructurada. El término *big data* se refiere a estos enormes conjuntos de información que no pueden ser gestionados ni analizados mediante herramientas tradicionales en una cantidad de tiempo aceptable. Esta investigación pretende demostrar la utilidad de *big data* como fuente de información para el estudio en el campo de la lingüística, lo que nos permitirá llevar a cabo estudios que no podrían realizarse con los métodos tradicionales. En este trabajo, utilizamos la información disponible en el servicio de microblogging *Twitter* para analizar los idiomas más hablados en cinco capitales europeas y conocer así las necesidades traductológicas que presentan. De esta forma, intentamos demostrar la utilidad de *big data* como herramienta para la investigación lingüística. Hemos analizado los tuits generados en las ciudades de Berlín, Bruselas, París, Madrid y Londres en el período de tiempo comprendido entre el 21 de agosto y el 21 de septiembre de 2015. Para ello, hemos desarrollado una herramienta de autor que nos permite obtener la información, almacenarla, gestionarla y analizarla. Pretendemos así obtener resultados basados en millones de datos y analizados en tiempo real, lo que nos ahorra costes y tiempo a la hora de llevar a cabo la investigación y nos permite conocer de manera inmediata los idiomas que se utilizan en cualquier momento y en cualquier lugar.

Palabras clave: *big data*, análisis lingüístico, *Twitter*, necesidades traductológicas, Europa.

Big data for the analysis of translation demand in five European capitals

Abstract: The amount of digital global information available is exploding. Huge quantities of unstructured information is being generated by big companies, such as *Google*, *Facebook*, *Microsoft*… The term *big data* refers to these huge datasets that cannot be managed and analysed by traditional tools and software in a tolerable

time. This research aims to show the utility of big data as a source of information for linguistic research, which enables us to carry out investigations that could not be done in a traditional way. In this paper, we use the information available in the microblogging service *Twitter* to analyse the most spoken languages in five European capitals, in order to know the demands for translation jobs. Thus, we prove the utility of big data as a tool for linguistic research. We have studied the total number of tweets generated in Berlin, Brussels, Paris, Madrid and London, during 21 August 2015 and 21 September 2015. For this purpose, we have developed an authoring tool through which we obtain, stored, processed and analysed the information. Our objective is to obtain results based on millions of data in real time, which saves not only time, but also costs in the research process and allows us to know the languages used anywhere and anytime.

Key words: big data, linguistic analysis, *Twitter*, translation demand, Europe.

Sumario: Introducción. 1. *Twitter* como fuente de información dentro de *big data*. 2. Funcionamiento de *Twitter*. 3. Metodología

Introducción

La cantidad de información que se genera en el mundo crece a pasos agigantados. Sin embargo, su naturaleza es muy diferente a la de la información en el pasado. La proliferación masiva de aparatos electrónicos y el frenético desarrollo de la informática han supuesto una revolución sin precedentes en el mundo de las comunicaciones. La cantidad y la variedad de datos disponibles, así como la velocidad a la que se generan, se almacenan y se transmiten, crece y evoluciona como nunca antes lo había hecho. Debido a la combinación de los dispositivos móviles, de internet y de la computación en la nube, los datos que se generan a partir de cámaras, micrófonos, sensores, etc. conformarán, dentro de poco, la mayor parte de la información disponible.

Según un informe de *International Data Corporation* (IDC), uno de los líderes mundiales en análisis de información masiva, durante la próxima década, el universo digital crecerá un 40% anual e incluirá no sólo el número creciente de personas y empresas que utilizan internet, sino también los pequeños aparatos conectados (Gantz y Reinzel, 2012). Este universo digital se compone de toda la información digital creada, replicada y consumida en un año, proveniente de las acciones más diversas que se puedan realizar en el día a día, como escribir o subir fotos o vídeos de los teléfonos móviles a las redes sociales, realizar una compra por internet, sacar dinero del cajero automático, realizar llamadas de teléfono, utilizar sistemas de control automáticos, etc.

IDC estimó en un estudio realizado en 2012 que en 2005 se crearon y replicaron 130 exabytes –un exabyte equivale a mil millones de gigabytes- y en 2013 ya habría 4,4 zettabytes (4,4 billones de gigabytes). Sus analistas calculan que, para el año 2020, la cantidad de información generada alcanzará 44 ZB, prácticamente tantos bits como estrellas hay en el universo. Esto significa que, desde ahora hasta 2020, la cifra crecerá anualmente más del doble hasta alcanzar más de 5.200 gigabytes por cada habitante de la Tierra.

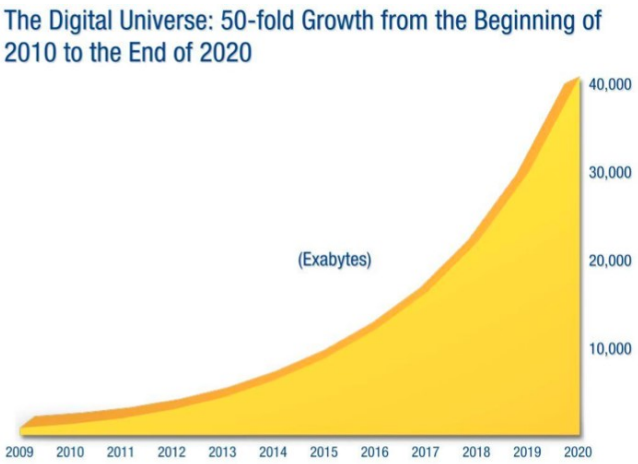


Figura 1: Crecimiento del universo digital. Fuente: IDC's Digital Universe Study, sponsored by EMC, 2012

En este escenario de explosión de la información global, el término *big data* se utiliza, a grandes rasgos, para describir los enormes conjuntos de información que se generan. Además, esta información se nos presenta, en la mayoría de los casos, de forma no estructurada, es decir, no está ordenada y lista para procesar, lo que implica la necesidad de nuevos sistemas de análisis de información más potentes y más rápidos, de manera que puedan analizarla en tiempo real. *Big data* presenta nuevas oportunidades, desafíos y problemas a los que las empresas, las organizaciones y los investigadores debemos enfrentarnos para conseguir comprender la información y obtener el máximo beneficio. Empresas de todo el mundo, gobiernos, universidades y organismos oficiales están cada vez más interesados en el gran potencial que ofrece *big data* y están empezando a invertir para acelerar su investigación y aplicación.

Son muchas las empresas, organizaciones e investigadores que han definido el término *big data*. Por ejemplo, Manyika *et al*. (2011:1) lo definen

para McKinsey –una consultora global del campo de la administración estratégica- como "conjuntos de información cuyo tamaño excede la capacidad de las herramientas de software de bases de datos convencionales para obtener, almacenar, procesar y analizar la información". De forma parecida, IBM –el gigante azul- se refiere a *big data* como "la información que no puede ser procesada ni analizada mediante procesos o herramientas convencionales" (Zikopoulos *et al*., 2012: 3). Los analistas de IBM, sin embargo, alertan de que el nombre puede llevar a error porque puede implicar que la información de la que se disponía antes no era cuantiosa o que la única característica de *big data* es el enorme volumen de información, cuando las dos afirmaciones, según ellos, son falsas.

En general, es comúnmente aceptada la idea de que *big data* posee tres características fundamentales (a partir de las cuales aparecen otras). Estas tres cualidades son: volumen, velocidad y variedad, y dan lugar a lo que se conoce como el modelo de las 3 V. El primero en introducir este concepto fue Loug Laney, en 2011, aunque no se refería a *big data*, sino a la gestión de la información en general. Fueron empresas como IBM o Microsoft las que lo aplicaron a *big data*. En este modelo, volumen se refiere a las grandes cantidades de información que se generan diariamente por individuos y por empresas y que crece a pasos agigantados. La velocidad significa la rapidez con la que la información se genera y se mueve. Zikopoulos *et al*. (2012) hacen hincapié en este aspecto de información en movimiento, yendo así más allá que otras definiciones que limitan la velocidad a esa rapidez que acabamos de explicar con la que la información se genera, se almacena y se recupera. Por último, la variedad hace alusión a los diferentes tipos de información compleja disponible, como puede ser la información tradicional y estructurada, la semiestructurada o la no estructurada, procedente de páginas web, de blogs, de correos electrónicos, de documentos, de redes sociales o de cualquier otra fuente de información digital.

Por otra parte, empresas como IDC, definen el concepto de la siguiente forma: "las tecnologías de *big data* describen una nueva generación de tecnologías y arquitecturas diseñadas para extraer valor económico de grandes cantidades y de una amplia variedad de información, gracias a la captura de datos, descubrimiento y/o análisis a alta velocidad" (Carter, 2011: 1). Esta definición fue el comienzo de una nueva visión de *big data* que incluía una cuarta V: el valor. También IBM, por su parte, añadió una cuarta V, distinta de esta última, para veracidad. Se referían con ella a la confianza que nos ofrece la información, ya que, en este nuevo panorama, aumentan las dificultades a la hora de controlar la calidad y la

exactitud de la información. Además de éstas, se han añadido otras características V al concepto, entre las que se encuentran variabilidad o visualización.

Otras empresas y organizaciones, como NIST, se centran en el aspecto tecnológico de *big data*, que la definen como "donde el volumen, la velocidad de adquisición o la representación de la información limitan la capacidad de llevar a cabo análisis efectivos con enfoques relacionales tradicionales o requieren el uso del escalado horizontal para un procesamiento eficaz" (Cooper y Mel, 2012: 15). En la misma línea, Dumbill (2012: 1) se refiere al concepto como "información que excede la capacidad de procesamiento de los sistemas de bases de datos habituales. La información es demasiado grande, se mueve demasiado rápido o no se ajusta a las estructuras de las bases de datos. Para obtener valor de esta información, es necesario escoger un sistema alternativo para procesarla".

En términos generales, podemos definir *big data* como conjuntos de información que no pueden ser comprendidos, obtenidos y procesados por herramientas de software y de hardware tradicionales en un período de tiempo tolerable (Chen, Mao y Liu, 2014).

En los últimos años, además, los costes de almacenamiento de un gigabyte han caído drásticamente, con una estimación de 2,00 dólares a 0,20 dólares desde 2012 a 2020 (IDC, 2012). Esto nos permite almacenar estas enormes cantidades de información a un precio muy bajo, de manera que la infraestructura del universo digital crecerá un 40% durante este período.

Gracias a la ubicuidad de los teléfonos móviles y a la generalizada disponibilidad de aplicaciones y de internet, es cada vez más frecuente el uso de las redes sociales, como *Twitter*, *Facebook* o *Instagram*, las descargas de música, la elaboración y publicación de documentos, la utilización de GPS, la visualización de vídeos de *Youtube*, etc. Un análisis adecuado de toda la información que se desprende *de* aquí permite a las empresas hacer perfiles personalizados de cada usuario y ofrecer productos diseñados con unas características muy específicas.

Por otro lado, Gens (2015), en otro informe para IDC, predice que, en este año, el mercado global de *big data* y los análisis de éste alcanzarán los 125 mil millones de dólares en todo el mundo en software, hardware y servicios. Además, el gasto en análisis de los medios se triplicará en el mismo año. Las estrategias de *big data* empleadas en los últimos años se harán mucho más complejas gracias a un aumento en las fuentes de información, en los usuarios y en las aplicaciones, lo que aumentará la ratio

de servicios profesionales de la tecnología en un 25% en los próximos cinco años.

1. *Twitter* como fuente de información dentro de *big data*

Twitter es una red social de microblogging con una marca de "¿qué está pasando?" (Yoon, Elhadad y Bakken, 2013) que permite a los usuarios registrarse gratis y publicar posts cortos de 140 caracteres llamados *tuits*. Estos usuarios utilizan la plataforma para compartir sus sentimientos, sus preocupaciones, sus gustos, sus inquietudes, sus actividades diarias y cualquier tipo de comentarios relacionados con su vida cotidiana. Los tuits pueden ser publicados y también seguidos por otros usuarios llamados seguidores, ya sea mediante la aplicación móvil, el cliente de escritorio o nativo, desde la página web Twitter.com, o con el servicio de mensaje corto (SMS). Aunque el servicio es gratis, el envío de tuits a través de SMS comporta las tarifas propias de cada operadora telefónica.

Yoon *et al.* (2013) afirman que los tuits son una fuente de información en tiempo real de lo que pasa en el mundo. El contenido de los tuits no depende de un estímulo intermitente específico, sino que representa una información más naturalista y tiene la ventaja adicional de estar disponibles en grandes cantidades Los usuarios de *Twitter* siguen a otros o bien tienen seguidores. A diferencia de la mayoría de las redes sociales, como *Facebook* o *MySpace*, estas relaciones de seguidores no requieren reciprocidad. Un usuario puede seguir a otro y que éste no lo siga a él. Ser seguidor en *Twitter* significa que el usuario recibe todos los mensajes (tuits) de aquellas personas a las que se sigue.

Desde que Jack Dorsey, uno de los fundadores de la compañía, escribiera su primer tuit el 21 de marzo de 2006, la plataforma ha experimentado un enorme crecimiento que en solo seis años le hizo llegar a 140 millones de usuarios activos que publicaban 340 millones de tuits diarios (Twitter, 2012). En 2015, *Twitter* ya cuenta con más de 500 millones de usuarios, de los cuales 302 participan de forma activa en la plataforma (Smith, 2015), lo que se traduce en 500 millones de tuits cada día (Twitter, 2015).

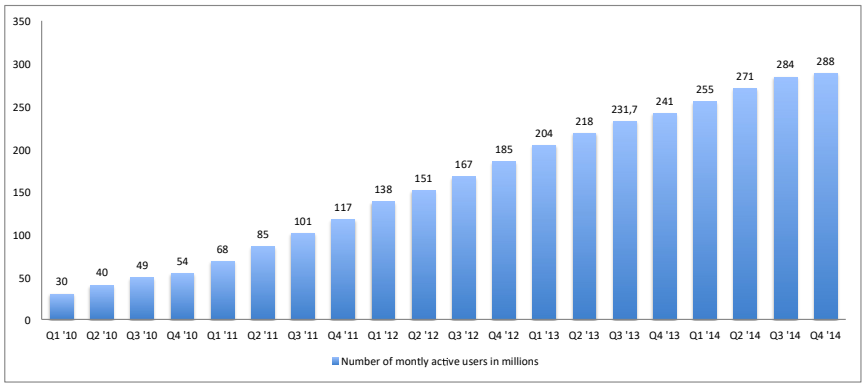


Fig. 2. Usuarios activos de *Twitter* por mes desde 2010 hasta 2014. Fuente: Statista, 2015

Con este panorama, las empresas de todo el mundo están empezando a invertir grandes sumas económicas que colaboran en la investigación sobre *big data* y redes sociales para sacar el máximo provecho de su potencial y de este incontrolable aumento de información.

Por este motivo, una de las redes sociales que más se prestan a su estudio y que más estudios está produciendo es *Twitter*, gracias a su naturaleza abierta, a su alto alcance y a su continua actualización en tiempo real. El trabajo de Salathé, Vu, Khandelwal y Hunter (2013) es un ejemplo de la gran variedad de estudios que se están llevando a cabo en este sentido y que demuestran la eficacia de *Twitter* como fuente de información. Estos autores realizaron un estudio en el que explicaba el poder de las comunidades online para extender el sentimiento negativo ante las vacunas, como también hizo UNICEF (2013), analizando las redes sociales en Europa. Para ello, se analizó la información existente en *Facebook*, *Twitter* y también algunos blogs.

Pero no sólo en el campo de la salud, también en el de la economía, el desarrollo o la sociología, se utilizan las redes sociales para aprovechar la información que contienen. Por citar algunos, Asur y Huberman (2010) demuestran que el análisis de estos medios, si es lo suficientemente amplio y está adecuadamente diseñado, suele ser más exacto que otras técnicas para la extracción de información, como los sondeos o las encuestas de opinión. En su estudio, los autores analizan las opiniones sobre películas en *Twitter* para predecir los ingresos en taquilla.

En 2011, *United Nations Global Pulse*, la iniciativa de Naciones Unidas para la utilización de *big data* para el desarrollo y la acción humanitaria, llevó a cabo una investigación en la que se puso de manifiesto la relación entre las leyes sociales y económicas. En el proyecto, se llevó a

cabo un análisis de tuits escritos en inglés, japonés e indonesio relacionados con diferentes temas (préstamos, deudas, vivienda y comida). Los resultados confirmaron que los precios oficiales de la comida en Indonesia coincidían con el número de tuits referidos al precio del arroz.

Otro ejemplo de la utilidad de la información disponible en *Twitter* es el de la previsión de la gripe (Broniatowsky, Dredze y Paul, 2014). Estos autores llevaron a cabo un estudio en el que afirman que el análisis de la información disponible en los tuits mejora la predicción de la prevalencia de la gripe y ayuda a detectar los índices de la enfermedad en tiempo real. También demostraron que los modelos que usan información de *Twitter* pueden reducir los errores de previsión de un 17% a un 30% y que son mejores indicadores que los de *Google Flu Trends* (GFT) –la herramienta oficial de *Google* basada en el análisis de ciertos términos de búsqueda para analizar y predecir la evolución de la gripe. Según Lazer, Kennedy, King y Vespignani (2014), sin embargo, el análisis de *big data* todavía tiene algunos problemas; el principal de ellos es que presenta ciertas limitaciones, que se pueden superar con otros sistemas de *big data*, para lo cual *Twitter* se presenta como uno de los más apropiados por cuestiones de granularidad, sobreajuste, replicabilidad, etc. El estudio de Broniatowsky *et al*. (2014) concluye que el número de tuits indicativos de la enfermedad real mantienen una fuerte coincidencia con los índices de los organismos oficiales de Estados Unidos para el control de las enfermedades -*Centers for Disease Control and Prevention* (CDC)-, así como el *Departamento de Salud e Higiene de Nueva York*, mientras que el de Lazer *et al*. demuestra que los análisis de GFT tienen un margen de error considerable con respecto a los CDC.

Otros autores, como Sakaki, Okazaki y Matsue (2010), han demostrado la utilidad de *Twitter* en la prevención de desastres naturales, como terremotos o tifones. También hay estudios que enumeran las ventajas de la utilización de *Twitter* en la educación como herramienta para el fomento del aprendizaje activo de idiomas (Borau, Ullrich, Feng & Shen, 2009) y como plataforma para el aprendizaje orientado a los procesos en educación superior (Ebner, Lienhardt, Rohs y Meyer, 2009).

Para Ritter, Clark, Mausam & Etzioni (2011: 1524) “los tuis son una fuente única de información, más actualizada e inclusive que los artículos de noticias, gracias a la facilidad para tuitear y la proliferación de aparatos móviles”.

Obviamente, aprovechar todo este potencial requiere esfuerzos en innovación en cuanto a operaciones y procesos, como señalan Manyika *et al.* (2011).

Sin embargo, a pesar de que algunos trabajos, como el de Kwak, Lee, Park y Moon (2010) resaltan que *Twitter*, gracias a su API abierta, a su peculiaridad de relaciones unilaterales entre usuarios y a los mecanismos de retuiteo, ofrece una oportunidad sin precedentes para investigadores de muy variados ámbitos -entre los que señala a los lingüistas-, no existen hasta el momento investigaciones que utilicen la información que nos brinda esta plataforma para un análisis lingüístico profundo a través de las herramientas informáticas adecuadas.

La utilización de *big data* en la investigación lingüística, como en el resto de las áreas que hemos analizado, nos permite obtener información precisa y objetiva acerca de diferentes aspectos de la lengua que sería extremadamente difícil extraer de otra forma. Llevar a cabo una investigación lingüística basada en los millones de datos que aporta *Twitter* usando métodos tradicionales supondría una tarea casi imposible si tenemos en cuenta las cuatro premisas que Bowker y Pearson (2003) señalan a la hora de elaborar un corpus. Para ellos, todo corpus lingüístico debe ser auténtico, electrónico, amplio y ordenado con unos criterios específicos, lo que se llevaría un gasto extraordinario de tiempo y de recursos humanos para poder igualar un corpus como el nos ofrece *Twitter*.

Gracias a la plataforma, la información es pública y está disponible, y se almacena de manera automática en los propios servidores de *Twitter*. Esto posibilita llevar a cabo búsquedas de millones de tuits en tan sólo un par de segundos.

Otro aspecto importante que hay que tener en cuenta es el carácter cambiante de las lenguas. Si en última instancia consiguiéramos elaborar un corpus fiable y completo de esas características, para cuando estuviera terminado, probablemente ya se hubiera quedado obsoleto. La vertiginosa velocidad a la que cambia el lenguaje y las circunstancias de los hablantes no pueden permitirse esperar el tiempo que necesitaría el corpus para su creación por medios tradicionales porque su evolución seguiría un ritmo más rápido que el de la propia investigación lingüística. Pero *Twitter* no sólo nos permite analizar información en tiempo real, también nos ofrece la posibilidad de acceder a información histórica, algo que puede resultar muy útil en cualquier tipo de investigación lingüística.

El objetivo principal de esta investigación es demostrar que, a través de *big data* en general y de la red social *Twitter* en particular, tenemos la posibilidad de obtener información precisa acerca de la evolución del lenguaje y de obtener los datos suficientes para llevar a cabo análisis apropiados de su estado actual, así como predicciones certeras acerca de su comportamiento en el futuro.

En palabras de IBM (2015): "*Twitter* es distinta a cualquier otra fuente de información del mundo. Es una plataforma mundial de información en tiempo real, pública y conversacional, en la que voces de todo el mundo hablan sobre cualquier tema imaginable".

2. Funcionamiento de *Twitter*

La característica fundamental que distingue a *Twitter* de otras redes sociales, como Facebook o Instagram, es que es abierta al público. Esto implica que cualquier persona con conexión a internet puede acceder a ella, aunque no tenga registrada una cuenta de usuario. Además, los ajustes predeterminados también son públicos, lo que significa que cualquier usuario puede seguir a cualquier otro en *Twitter* público sin que este último de su aprobación o sin la necesidad de un contacto directo entre usuarios.

A la hora de publicar un post y enlazarlo con un tema de actualidad o sobre el que estén hablando otros usuarios, basta con añadir una etiqueta o *hashtag* al post con el símbolo almohadilla -#- delante de la palabra clave o etiqueta (*#hashtag*). *Twitter* agrupa los posts por *hashtags* y los clasifica en un índice en el que se muestran los temas más populares, llamados *trending topics*, esto es, los temas de los que está hablando la gente agrupados en tiempo real. De esta forma, cualquiera que introduzca una palabra clave o un *hashtag* en el buscador puede ver qué está pasando en el mundo y de qué está hablando la gente. Los *trending topics* se pueden clasificar también por país, ciudad o estado.

Puesto que los tuits se publican en la página web de *Twitter*, no desaparecen al cerrar la aplicación, de manera que se almacenan y se tener acceso a ellos en cualquier momento, incluso sin ser miembros registrados.

Registrarse, sin embargo, es un proceso sencillo que consiste simplemente en crear una cuenta y escoger un nombre de usuario, que vendrá precedido por el símbolo @ (@nombredeusuario). En el perfil de cada usuario aparece una lista denominada "siguiendo" los (cuentas que se siguen) y otra de "seguidores" (cuentas que siguen al propio usuario). La relación entre un usuario y sus seguidores no es necesariamente bidireccional, con lo que un usuario puede seguir a otro que no lo siga y viceversa.

Además de texto, los tuits pueden incluir hipervínculos, pero debido a las limitaciones de espacio, generalmente se utiliza un servicio de acortamiento de URL, como *t.co* (el servicio propio de *Twitter*), *tinyurl.com* o *bit.ly*. El funcionamiento de estos servicios consiste en introducir la URL extendida y la herramienta la acorta de forma instantánea y automática.

Los tuits pueden ser contestados, retuiteados (RT + @nombredeusuario + contenido del tuit) o también marcados como favoritos con un icono de una pequeña estrella. Si se cambia el contenido de un tuit al hacer retuit, se convierte en un tuit modificado (MT). También se puede mencionar a otro usuario añadiendo @nombredeusuario en el contenido del tuit (para ello no es necesario que exista una relación bidireccional de seguidor/seguido) e incluso enviar mensajes privados denominados mensajes directos. En este caso, el destinatario del mensaje sí debe ser un seguidor.

En la figura 3 podemos ver las partes que componen un tuit y que acabamos de explicar.

Fig. 3. Anatomía de un tuit: *Twitter*

3. Metodología

Ya hemos comentado que las redes sociales y, más concretamente, *Twitter*, se están utilizando con cada vez más frecuencia como medios de comunicación a través de los cuales la gente expresa sus opiniones, preocupaciones, etc. La libertad que otorga el hecho de que no existan restricciones de temas, idiomas, horarios o geografía, hace de él un medio inigualable como plataforma para la expresión personal y medio de interacción entre las personas.

Esto resulta de gran utilidad para la investigación lingüística, ya que tenemos la posibilidad de acceder a las producciones lingüísticas de los usuarios sean cuales sean su idioma, su sexo, su edad o su situación social. Además, el enorme número de usuarios que deja su producción lingüística y nos la brinda para la investigación hace que los resultados extraídos estén basados en millones de muestras, lo que aporta veracidad a la investigación.

En este estudio, hemos seleccionado cinco capitales europeas para ejemplificar uno de los posibles usos de *big data* en la investigación lingüística. Las capitales seleccionadas han sido Berlín, Bruselas, París, Madrid y Londres. Para el estudio, hemos establecido el período de tiempo

de un mes, comprendido entre el 21 de agosto y el 21 de septiembre de 2015.

El principal objetivo de esta investigación es demostrar la utilidad de *big data*, y en concreto de *Twitter*, como fuente para la investigación lingüística. Para llevar a cabo el análisis de los resultados, hemos desarrollado una herramienta de autor que registra todos los tuits publicados durante ese período de tiempo en toda Europa. A la hora de realizar del estudio, se ha llevado a cabo una selección que ha concluido con las cinco capitales presentes en este trabajo. De este modo, la herramienta ha almacenado la identificación única de cada tuit, el texto, el usuario, las coordenadas geográficas (latitud y longitud) y la fecha y hora de publicación. En este caso, para nuestro estudio, no ha sido relevante analizar el contenido del texto y sólo hemos extraído el idioma del mismo, utilizando la norma ISO 639-1. El filtro de las cinco capitales se ha realizado mediante una *bounding box*, es decir, un recuadro que engloba geográficamente una zona determinada mediante dos pares de coordenadas (vértice noreste y vértice suroeste).

La ventaja de este tipo de análisis estriba en la posibilidad de llevar a cabo un estudio inmediato del estado lingüístico de cualquier zona del mundo de cualquier dimensión o característica, en tiempo real y con una muestra de millones de usuarios. En nuestro caso, hemos realizado el estudio sobre 58.475.875 tuits.

Una vez obtenidos los datos, se ha desarrollado un visualizador que extrae la información obtenida de la base de datos y que muestra una representación gráfica en un mapa para el segmento temporal y espacial seleccionado, con los distintos idiomas clasificados por colores. Así, podemos comprobar a simple vista la distribución geográfica de las lenguas utilizadas en las capitales. Además, la herramienta nos permite obtener las estadísticas porcentuales de la utilización de los idiomas en las áreas seleccionadas. A continuación, analizamos los resultados obtenidos de las cinco capitales europeas.

A continuación, presentamos el análisis de los datos obtenidos con una ficha para cada una de las capitales. En ella, indicamos las coordenadas utilizadas para el establecimiento de la *bounding box*, así como la superficie analizada y el número de habitantes de cada capital. En cuanto a la información relativa a los tuits analizados, incluimos, en primer lugar, un gráfico con los porcentajes de los cuatro idiomas más hablados en cada ciudad y el porcentaje del resto de idiomas utilizados, además de una tabla con el número exacto de tuits publicados en cada idioma. Por último, añadimos seis gráficos más, todos con el mapa de la zona geográfica

delimitada por las coordenadas. El primero de ellos se trata de ese mapa sin información sobre los tuits, en el segundo aparecen todos los tuits de los diferentes idiomas superpuestos, cada uno con un color distinto. Los colores se indican en los cuatro siguientes mapas, donde vienen desglosados los tuits por idioma.

Berlín

Coordenadas seleccionadas

NE 52.667511, 13.72616

SW 52.330269, 13.05355

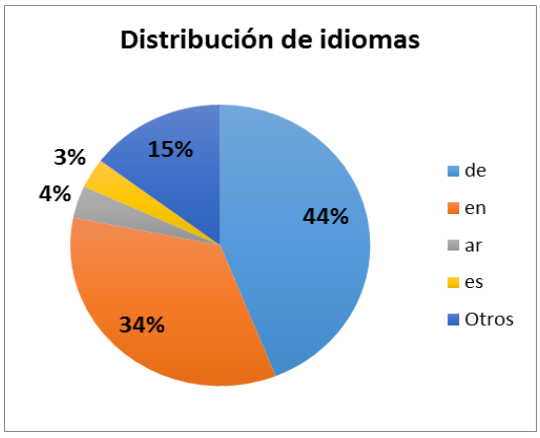
Población

3,502 millones (ONU, 2012)

Superficie

891,8 km^2

Análisis

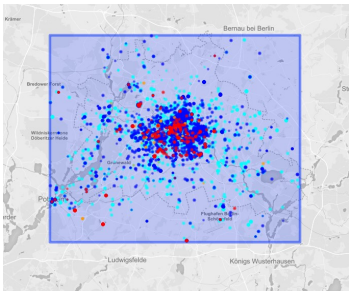


Número de tuits
21/8 – 21/9

de	113394
en	88182
ar	9529
es	8811
otros	37937

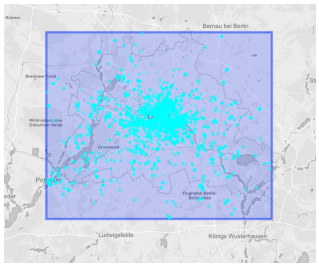
Región seleccionada y distribución geográfica de idiomas

Región geográfica

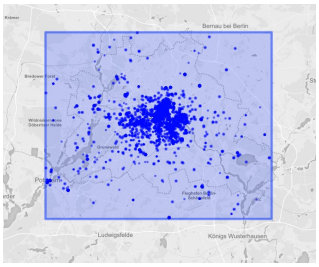


Distribución de idiomas

Distribución detallada por idioma

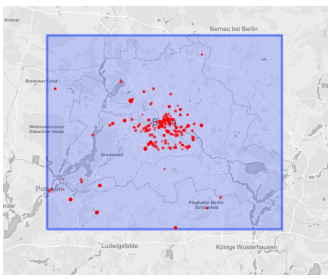


Distribución - Alemán



Distribución – Inglés

Distribución - Árabe



Distribución – Español

Comentarios

El alemán, como lengua oficial, es el idioma más hablado, que se concentra en el centro de la ciudad y se va repartiendo radialmente hacia

las afueras de manera homogénea. Podemos ver cómo el inglés, como segunda lengua más utilizada, se habla más o menos en los mismos lugares que el idioma oficial, aunque en menor cantidad. Con respecto al tercer y cuarto idiomas, resulta curioso observar cómo el árabe, que se habla más que el español, tiene una zona de influencia mucho más restringida que este último.

Así, la utilización del árabe se limita fundamentalmente a la parte más céntrica de la ciudad y sólo en algunos puntos concretos, aunque también se utiliza en algún otro punto de la región, sobre todo al sur. Por el contrario, el español, aunque menos hablado, tiene una distribución más amplia hacia el norte y el oeste. Además, en la parte céntrica está mucho más repartido que el árabe, lo que significa que, a pesar de que haya menos hablantes, su presencia está más homogéneamente distribuida que la de los arábigo parlantes.

Bruselas

Coordenadas seleccionadas

NE 50.913971, 4.43709

SW 50.79628, 4.31393

Población

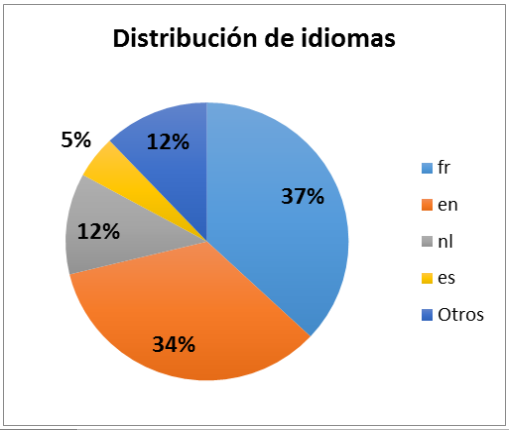
177.307 (ONU, 2012)

Superficie

32,61 km^2

Análisis

Número de tuits

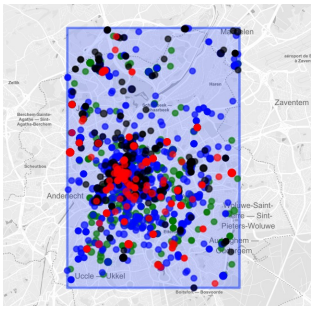


21/8 – 21/9

fr	39763
en	36830
nl	12654
Es	5331
otros	1330

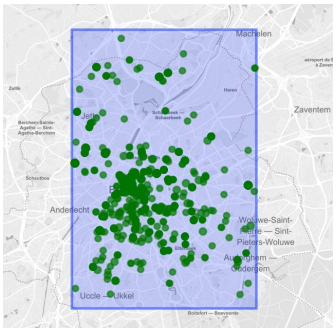
Región seleccionada y distribución geográfica de idiomas

Región geográfica



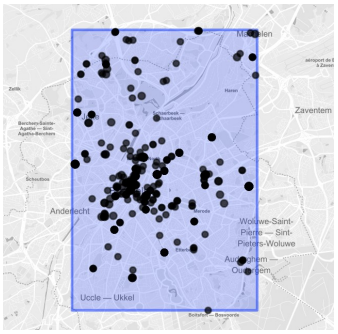
Distribución de idiomas

Distribución detallada por idioma

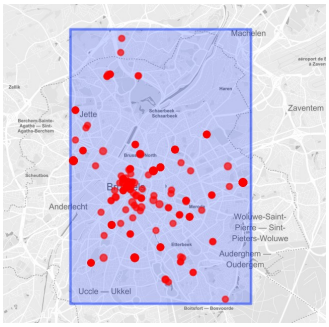


Distribución - Francés

Distribución – Inglés



Distribución - Neerlandés



Distribución – Español

Comentarios

En Bruselas resulta significativa la diferencia entre el número de hablantes de las distintas lenguas. Mientras que entre la primera y la segunda lengua –francés e inglés- la diferencia de uso es relativamente pequeña, entre el inglés y el neerlandés (dentro de cuyo registro se contabilizan los tuits escritos en flamenco) es mucho mayor, al igual que con el español, que se trata del cuarto idioma más utilizado.

Los mapas reflejan esta situación y muestran que tanto el francés como el inglés se utilizan más o menos por las mismas zonas de la región, sobre todo en el centro, aunque también observamos cómo el inglés se extiende por algunas zonas del norte y del noreste. El rastro que dejan los tuits escritos en holandés mantiene muchas semejanzas con el que dejan los del inglés por la zona centro y norte, aunque es cierto que, en la parte sur, este idioma se utiliza mucho menos que los dos primeros. Por su parte, el español es el idioma menos utilizado en la lista de los cuatro primeros. Su presencia predomina, como en el resto de los casos, en la zona centro de la ciudad y se extiende de forma radial hacia el este y hacia el sur. En la parte norte, sólo en el oeste es donde encontramos evidencias de la utilización del español. Por tanto, si comparamos el holandés y el español, vemos cómo el primero se extiende más hacia el noreste, donde el español tiene poca presencia, y éste último se encuentra más repartido por el sur que el holandés.

París
Coordenadas seleccionadas

49.04694, 2.63791

48.658291, 2.08679

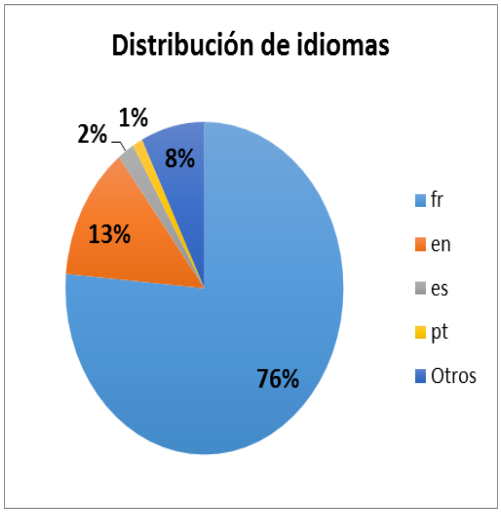
Población

2,244 millones (ONU, 2012)

Superficie

105,4 km^2

Análisis

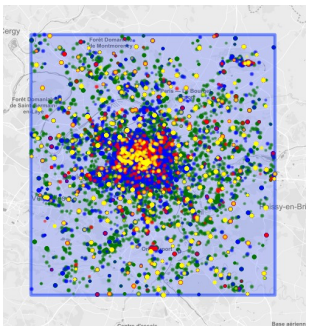


Número de tuits
21/8 – 21/9

fr	1472037
en	252066
es	37757
pt	22696
otros	142681

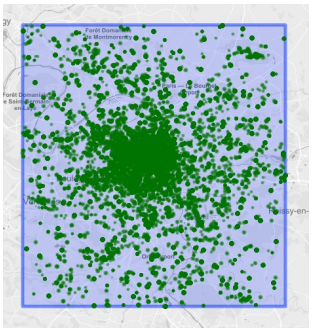
Región seleccionada y distribución geográfica de idiomas

Región geográfica

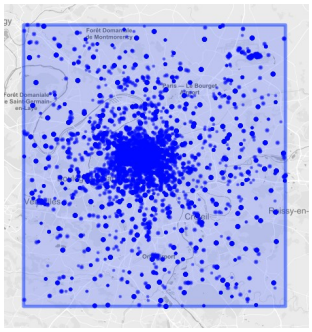


Distribución de idiomas

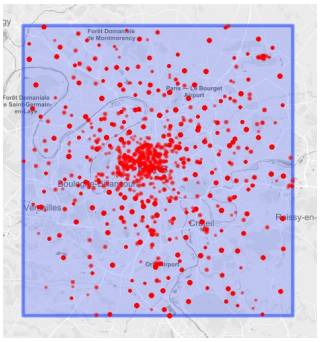
Distribución detallada por idioma



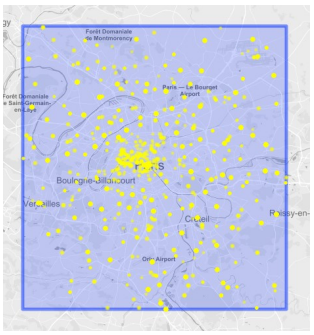
Distribución - Francés



Distribución – Inglés



Distribución - Español



Distribución – Portugués

Comentarios

En el caso de la capital francesa, la lengua oficial presenta un claro predominio sobre las restantes, encabezadas por el inglés, que sólo representa algo más de un octavo del porcentaje de tuits escritos en francés. En una situación similar se encuentra el español con respecto a su antecedente más inmediato, con sólo un 2% del total de los tuits analizados. Por su parte, el portugués, a pesar de ser el cuarto idioma más utilizado en esta ciudad, sólo representa un 1% del total.

Tanto el francés como el inglés se utilizan prácticamente en las mismas zonas geográficas: se concentran en el centro de la ciudad y se distribuyen radialmente hacia las afueras, perdiendo densidad conforme se van alejando. La gran diferencia entre ambos, sin embargo, radica en el número de hablantes de cada idioma, siendo, como acabamos de comentar, mucho mayor el del idioma galo. El mismo patrón de distribución geográfica sigue el español, aunque con un número mucho menor de tuits. Podemos ver cómo se concentran, no sólo el español, sino también el resto de los idiomas, en un punto al noreste y en otro al sur de la región; estos puntos coinciden con los aeropuertos de Charles De Gaulle y de Orly, y también en Versailles. El portugués se encuentra también repartido de manera muy homognea aunque, al igual que todos los demás, más concentrado en el centro.

Madrid

Coordenadas seleccionadas

NE 40.520081, -3.5349

SW 40.325939, -3.79887

Población

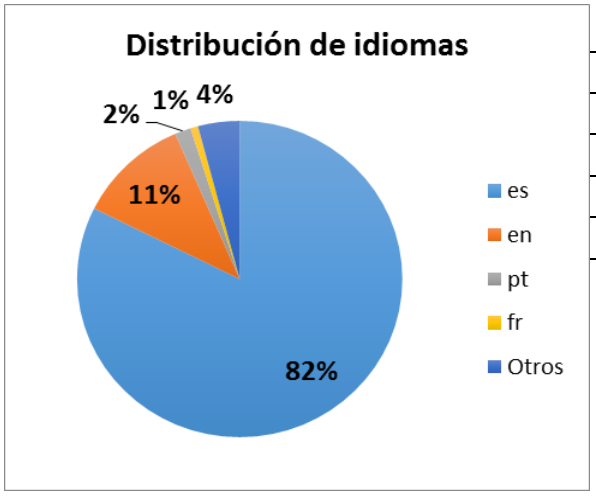
3,165 millones (ONU, 2014)

Superficie

605,8 km^2

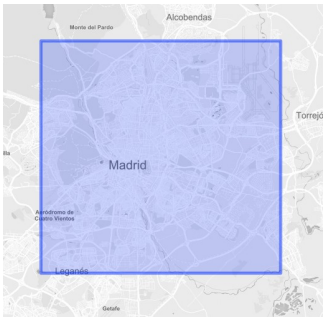
Análisis
Número de tuits

21/8 – 21/9

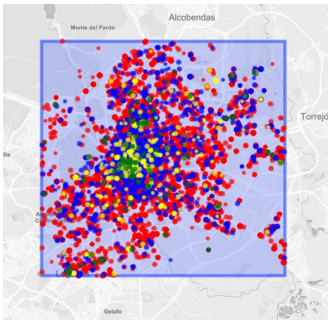


es	482812
en	65389
pt	9316
fr	4444
otros	24248

Región seleccionada y distribución geográfica de idiomas

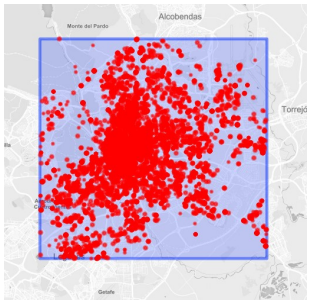


Región geográfica

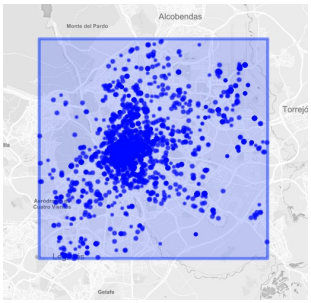


Distribución de idiomas

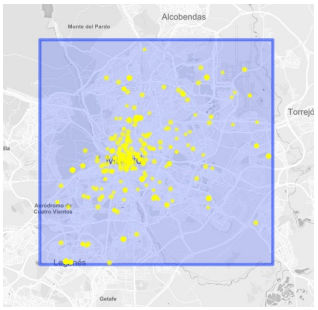
Distribución detallada por idioma



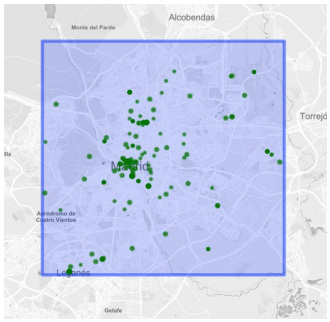
Distribución - Español



Distribución – Inglés



Distribución - Portugués



Distribución – Francés

Comentarios

Los cuatro idiomas más utilizados en Madrid coinciden con los de París, aunque con distinto orden. Naturalmente, el español es la lengua más usada entre los usuarios de *Twitter*. Existe aquí una diferencia aún mayor entre esta primera lengua y la segunda más usada –el inglés-, que sólo se utiliza en un 11% del total. Mucho menor todavía es el número de usuarios que utilizan el portugués o el francés para comunicarse.

A diferencia de otras ciudades, aunque se sigue manteniendo la premisa de que el mayor uso de todos los idiomas se da en el centro de las ciudades, esta concentración en Madrid adquiere más tamaño cuando se trata del español que del inglés (y, por supuesto, también del portugués y del francés). También es distinta la distribución a la de la capital francesa porque, en este caso, los idiomas no se distribuyen de manera radial, sino que crecen fundamentalmente hacia el norte y hacia el suroeste, fundamentalmente el español. El crecimiento hacia el noreste también es significativo, pero la densidad de tuits es menor que en los casos anteriores. El uso del francés, mucho menor que el del portugués, se circunscribe, como hemos apuntado, más al centro y apenas se extiende hacia el norte; también encontramos presencia de este idioma en la zona próxima a Leganés, en la parte sur.

Londres

Coordenadas seleccionadas

NE 51.692322, 0.33403

SW 51.286839, -0.51035

Población

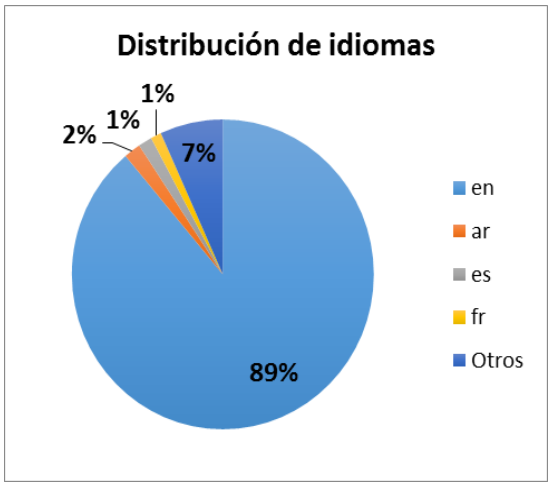
8,539 millones (ONU,2014)

Superficie

1572,8 km^2

Análisis
Número de tuits

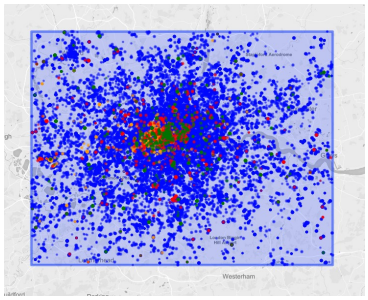
21/8 – 21/9



en	2416695
ar	49125
es	40073
fr	30953
otros	183223

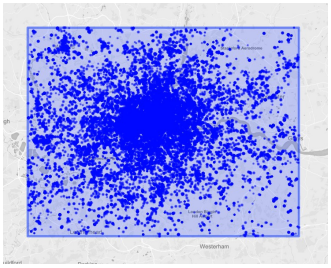
Región seleccionada y distribución geográfica de idiomas

Región geográfica

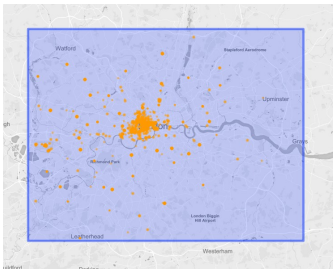


Distribución de idioma

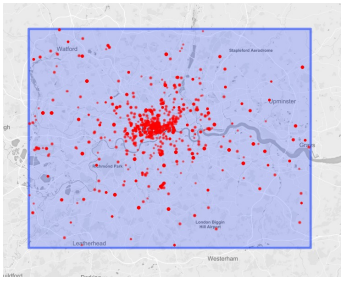
Distribución detallada por idioma



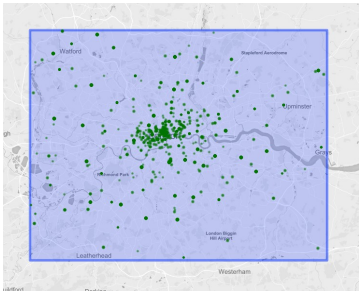
Distribución - Inglés



Distribución – Árabe



Distribución - Español



Distribución – Francés

Comentarios

El caso de Londres es el más llamativo de los cinco ejemplos presentados en esta investigación porque se trata de la capital en la que la diferencia entre la lengua oficial y el resto es más abrupta, a pesar de ser la ciudad que mayor número de tuits registra. En este sentido, el 89% de los tuits estudiados son en inglés, mientras que los tres idiomas siguientes –árabe, español y francés- sólo representan un 2% (árabe) y un 1% (español y francés). Sólo en el 7% restante incluimos los demás idiomas utilizados en la ciudad, entre los que se encuentran el portugués, el turco, el italiano o el japonés, entre muchos otros.

En cuanto a la organización geográfica, aquí sí podemos ver una distribución radial, partiendo del centro de la ciudad. Sin embargo, la diferencia entre el inglés y el resto de los idiomas es muy significativa, puesto que el inglés se utiliza con muchísima densidad en prácticamente la totalidad de la región estudiada, mientras que los otros tres se concentran

en una zona céntrica mucho más restringida. Las diferencias entre la utilización del árabe y del español radican en que el primero de ellos se utiliza menos en la zona este que el segundo, aunque también podemos observar cómo ambos experimentan una crecida en un punto del oeste. Este punto coincide con el aeropuerto de Heathrow, al igual que ocurría en París. El uso del francés se extiende también de manera relativamente homogénea a partir del centro de la ciudad, aunque, como podemos ver, con un índice mucho más pequeño. Aunque el francés y el árabe se circunscriben a la zona centro y se hablan en un radio del mismo tamaño, aproximadamente se observa un agrupamiento distinto de los tuits, lo que indica que en algunas zonas o barrios de la ciudad se habla más un idioma que el otro.

Conclusiones

Como ya hemos apuntado anteriormente, el objetivo fundamental de este estudio es demostrar que las enormes cantidades de información disponibles en la web, correctamente administradas y analizadas, constituyen una valiosa fuente de información para la investigación en el ámbito de la lingüística. Gracias a *big data*, tenemos la posibilidad de trabajar con millones de datos, pero es importante procesarlos adecuadamente y analizarlos de manera que se puedan convertir en información útil. En esta investigación concreta, gracias a la herramienta de autor que hemos explicado, hemos registrado los tuits publicados en Europa durante un mes y los hemos seleccionado por idiomas y por localización geográfica para comprender en tiempo real cuáles son los idiomas que se hablan en un lugar determinado en una fecha concreta.

Aunque la herramienta devolvía todos los tuits publicados en cada una de las *bounding boxes* descritas, hemos considerado oportuno mencionar los cuatro idiomas más utilizados por los usuarios de *Twitter* para comunicarse y englobar el resto en una sola categoría.

Desde un punto de vista profesional traductológico vemos, por ejemplo, que el inglés predomina como segunda lengua más usada en todas las capitales en las que no funciona como idioma oficial. Londres, lógicamente, es la excepción. Además, es llamativa la diferencia en esta ciudad entre el uso del inglés y el del resto de los idiomas. Parece claro que el nivel de plurilingüismo en una capital cuyo idioma oficial es el inglés es mucho menor que en el resto de ciudades, en las que esta lengua figura siempre en segundo lugar. Los datos dejan pocas dudas acerca del carácter vehicular y universal de este idioma. Por otro el lado, es curioso también el hecho de que, en las capitales estudiadas, el alemán sólo se habla de forma

significativa en Berlín, mientras que otros idiomas, como el español, el árabe, el francés o el portugués, tienen más presencia en los países europeos. De hecho, estos cuatro, junto con el inglés y el alemán son los únicos seis idiomas registrados como los más utilizados en estas cinco ciudades. Mención especial merece el uso del neerlandés, ya que se trata de uno de los idiomas oficiales de Bruselas. Idiomas, sin embargo, que podrían presentar oportunidades para el mundo de la traducción, ya sea por su proximidad con los países estudiados o por el gran número de hablantes que tienen (como podrían ser el italiano, el chino o el turco), no dejan rastros relevantes en estas ciudades.

Observamos que, mediante la aplicación de *big data* al campo de la lingüística, podemos obtener una visión global y veraz de las necesidades traductológicas de un lugar concreto en una fecha determinada. Huelga decir que a ningún profesional del sector se le escapa cuáles son los idiomas más utilizados en las distintas capitales. Sin embargo, puesto que los idiomas son un ente vivo y se encuentran en constante evolución, su situación es susceptible de cambio conforme pasa el tiempo o cambian los factores sociales, políticos o económicos. Una investigación de este tipo no sólo ahorra tiempo y costes con respecto a metodología tradicional, sino que nos brinda la posibilidad de obtener un fotograma del estado de cualquier idioma en el momento que deseemos e incluso, también, en tiempo real.

Podemos decir, por tanto, que *big data* es una fuente de información de gran valor para la investigación lingüística que nos permite obtener resultados fiables basados en millones de datos y de forma inmediata, algo inviable con métodos tradicionales. Las puertas que se nos abren con *big data* requieren, no obstante, inversiones y esfuerzos en innovación, no sólo en la adquisición de nuevas metodologías, sino también en el desarrollo de nuevas herramientas que sean capaz de cubrir las necesidades de este nuevo horizonte. A la hora de llevar a cabo estudios lingüísticos, *Twitter* presenta ventajas adicionales porque se trata de un sistema de comunicación en continua actualización y se presenta como un reflejo del uso real del lenguaje que hacen los usuarios. Además, nos permite establecer criterios de búsqueda específicos que se ajusten al perfil y a los objetivos de cada investigación. Por ejemplo, para este estudio hemos delimitado la búsqueda por fecha, idioma y localización geográfica.

Confiamos en que este trabajo abra el camino para futuras investigaciones lingüísticas en la misma línea y siente las bases de nuevas metodologías que favorezcan el desarrollo de esta disciplina, basadas en la técnica y en el progreso científico.

Referencias bibliográficas

ASUR, S., & HUBERMAN, B. A. (2010). Predicting the future with social media. Web intelligence and Intelligent Agent Technology. (Wi-IAT), 2010 IEEE/WIC/ACM International Converence on IEEE.

BORAU, K., ULLRICH, C. FENG, J. y SHEN, R. (2009). Microblogging for language learning: using Twitter to train communicative and cultural competence. *Proceedings of the International Conference on Advances in Web Based Learning*.

BOWKER, L. y PEARSON, J. (2002). *Working with Specialized Language. A practical guide to using corpora*. London: Routledge.

CARTER, P. (2011). Big data analytics: Future Arquitectures, Skills and Roadmaps for the CIO. Recuperado de http://www.sas.com/resources/asset/BigDataAnalytics-FutureArchitectures-Skills-RoadmapsfortheCIO.pdf

CHEN, M., MAO, S. y LIU, Y. (2014). Big data: a survey. *Mobile Networks and Applications*, 19. 171-209. doi: 10.1007/s11036-013-0489-0

COGAN, P., ANDREW, M., Bradonjic, M., Tucci., Kennedy, W. S. y Sala, A. (2012). Reconstruction and analysis of Twitter conversation graphs. Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research. (25-31). New York: ACM. doi: 10.1145/2392622.2392626

COOPER, M. y MELL, P. Tackling big data. NIST Information Technology Laboratory. Computer Security Division. US Department of Commerce. Recuperado de http://csrc.nist.gov/groups/SMA/forum/documents/june2012presentations/fcsm_june2012_cooper_mell.pdf

CUKIER, K. (2010). Data, data everywhere: a special report on managing information. *The Economist Newspaper*.

CUNHA, E., MAGNO, G., COMARELA, G., ALMEIDA, V., GONÇALVES, M. A. y BENEVENUTO. (2011). Analyzing the dynamic evolution of hashtags on Twitter: a language-based approach. *Proceedings of the Workshop on Language in Social Media*. (58-65). Stroudsburg, PA: Association for Computational Linguistics.

DUMBILL, E. (2012). What is big data? An introduction to the big data landscape. *O'Reilly*. Recuperado de https://beta.oreilly.com/ideas/what-is-big-data

EBNER, M., LIENHARDT, C. ROSH, M. y MEYER, I. (2010). Microblogs on Higher Education – A chance to facilitate informal and process-oriented learning? Computers and Education, 55, pp. 92-100.

GANTZ, J. y REINSEL, D. (2011). Extracting value from chaos. *IDC iView*, pp. 1-12. Retrieved from http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

GENS, F. (2015). IDC Predictions 2015: Accelerating Innovation — and Growth — on the 3rd Platform. Recuperado de http://www.sap.com/bin/sapcom/en_us/downloadasset.2014-12-dec-19-22.idc-predictions-2015-accelerating-innovation--and-growth--on-the-3rd-platform-pdf.bypassReg.html

HONG, L., Convertino, G. y CHI, E. H. (2011). Language matters in Twitter: a large scale study. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.* (518-521). Menlo Park, California: The AAAI Press.

IBM (2015). Retrieved from http://www.ibm.com/big-data/us/en/big-data-and-analytics/ibmandTwitter.html.

KWAK, Lee, PARK y MOON (2010). What is Twitter, a Social Network or a News Media? (26-30). *WWW '10. Proceedings of the 19th International Conference on Wolrd Wide Web.* (591-600). New York: ACM. doi: 10.1145/1772690.1772751.

LANEY, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. *Meta Group Research Note*, 6 February. Recuperado de http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

LAZER, D., KENNEDY, R., KING, G. & VESPIGNANI, A. The parabale of Google Flu: traps in big data. *Science*, 343 (6176), pp. 1203-1205. doi: 10.1126/science.1248506

MANYIKA, J., CHUI, M., BROWN, B., BUGHIN, J., DOBBS, R. ROXBURGH, C. y BYERS, A. H. (2013). *Big data: the next frontier for innovation, competition and productivity*. McKinsey Globan Institute.

NAAMAN, M., BOASE, J. y LAI, C.H. (2010). Is it really about me? Message content in social awareness streams. *Proceedings of the 2010 ACM conference on computer supported cooperative work*. (189-192). New York: ACM. doi: 10.1145/1718918.1718953

PAUL, M. J., DREDZE, M. y BRONIATOWSKY, D. (2014). Twitter improves Influenza forescasting. PLOS Current Outbreaks, 1. doi: 10.1371/current.outbreaks.90b9ed0f59bae4ccaa683a39865d9117

POKORNOWSKI, M. (2015). The fourth V, as in evolution: How evolutionary linguistics can contribute to data science. Theoria et Historia Scientiarum, 11, pp. 45-61. doi: http://dx.doi.org/10.12775/ths-2014-003

RITTER, A., CLARK, S., MAUAM y ETZIONI, O. (2011). Named entity recognition in Tweets: an experimental study. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. (1524-1534). Stroudsburg, PA: Association for Computational Linguistics.

SALATÉ, M., VU, D.Q., KHANDELWAL, S. y HUNTER, D.R. (2013). The Dynamics of Health Behaviour Sentiments on a Large Online Social

Network. EPJ Data Science (2:4). Springer. Recuperado de http://www.epjdatascience.com/content/pdf/epjds16.pdf

SAKAKI, T., OKAZAKI, M. Y MATSUO, Y. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors. En WWW'10

Smith, C. (2015, June 5). By the numbers: 150+ amazing Twitter statistics. DMR. Retrieved from http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-Twitter-stats/

The Economist (2011). Drowning in numbers. Retrieved from http://www.economist.com/blogs/dailychart/2011/11/big-data-0

Twitter (2012, March 21). Twitter turns six. Twitter. Retrieved from https://blog.Twitter.com/2012/Twitter-turns-six

United Nations Global Pulse. Mining Indonesian tweets to understand food price crisis, 2014. Retrieved from http://www.unglobalpulse.org/sites/default/files/Global-Pulse-Mining-Indonesian-Tweets-Food-Price-Crises%20copy.pdf

Uso de Twitter/Datos de la empresa. (2015). Retrieved from https://about.Twitter.com/es/company

YOON, S., ELHADAD, N. & BAKKEN, S. (2013). A Practical Approach for Content Mining of Tweets. *American Journal of Preventive Medicine*, 45 (1), pp. 122-129.

ZIKOPOULOS, P.C., EATON, C., DEROOS, D., DEUTSCH, T. & LAPIS, G. (2012). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. United States of America: McGraw Hill.