UNIVERSIDAD DE CÓRDOBA

DEPARTAMENTO DE CIENCIAS DEL LENGUAJE

ÁREA DE LINGÜÍSTICA GENERAL

TESIS DOCTORAL

MÁS ALLÁ DEL CORPUS: *BIG DATA* EN LA INVESTIGACIÓN LINGÜÍSTICA. EVOLUCIÓN, ANÁLISIS Y PREDICCIÓN DEL USO DE LA LENGUA A TRAVÉS DE *TWITTER*

ADELA GONZÁLEZ FERNÁNDEZ

Bajo la dirección de la Profesora Dra. María Luisa Calero Vaquera

y de la Profesora Dra. Gloria Guerrero Ramos

MARZO 2016

TITULO: *Más allá del corpus: Big Data en la investigación lingüística. Evolución, análisis y predicción del uso de la lengua a través de Twitter*

AUTOR: *Adela González Fernández*

Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

www.uco.es/publicaciones
publicaciones@uco.es

TÍTULO DE LA TESIS:

Más allá del corpus: Big Data en la investigación lingüística. Evolución, análisis y predicción del uso de la lengua a través de *Twitter*

DOCTORANDA: Adela GONZÁLEZ FERNÁNDEZ

INFORME RAZONADO DE LAS DIRECTORAS DE LA TESIS

(se hará mención a la evolución y desarrollo de la tesis, así como a trabajos y publicaciones derivados de la misma).

La tesis doctoral que presenta Dª Adela González Fernández se inscribe en una innovadora línea de investigación interdisciplinar, donde, en el marco de las nuevas tecnologías, se interralacionan e interactúan la Lingüística de Corpus y la Terminótica, operando sobre los resultados ofrecidos por el *Big Data* (concepto cibernético caracterizado por los especialistas como el producto tangible y no estructurado de las interralaciones humanas a través de las nuevas tecnologías y las redes sociales). En el inicio de esta Tesis, y sustentando toda su estructura, se halla la siguiente hipótesis: la utilidad de la información disponible en las nuevas redes sociales y, en concreto, en el sistema de comunicación virtual denominado *Twitter*, para hacer un seguimiento tanto de la evolución como del estado inmediato del léxico español, en sus diversas variantes y zonas geográficas de uso, así como para realizar predicciones futuras sobre el comportamiento del mismo. Esta utilidad puede ser extrapolable en su uso a cualquier otro ámbito de estudio de la Lingüística (especialmente de la Lingüística aplicada), como se pretende demostrar aquí con un selecto muestrario de posibles aplicaciones.

Tras una primera parte teórica donde se traen a colación y explican conceptos previos como el de “Lingüística de corpus”, “The web as corpus”, “Big Data” y “Twitter”, la doctoranda, en la siguiente parte práctica y metodológica, muestra el funcionamiento y las utilidades que en el campo de la Lingüística presenta (o puede presentar) la herramienta informática por ella diseñada (supuesto que toda esta información se encuentra en soporte digital), cuyo objetivo final es demostrar la veracidad de la hipótesis de partida. Así, en esta segunda parte se muestran algunos resultados obtenidos de la aplicación, que se presentan bajo el formato de cuatro módulos: *Wordics Live* (módulo de información del uso de los ´terminos lingüísticos en tiempo real), *Wordics One* (análisis histórico de los términos utilizados en determinadas cuentas de *Twitter*), *Wordics Archive* (módulo de análisis histórico de la información) y *Wordics Data* (módulo de exportación de datos).

Los aspectos más novedosos de esta Tesis son los avances que presenta respecto a la metodología actualmente seguida en la Lingüística de Corpus, cuyas investigaciones clásicas ofrecen un evidente carácter estático y, al mismo tiempo, se encuentran encerradas en los límites temporales que vienen marcados por el momento histórico de la compilación del corpus. Por otra parte, la inmensa mayoría de los trabajos diseñados por la tradicional Lingüística de Corpus presentan una grave limitación en cuanto al tamño de los propios corpora así como al largo tiempo que debe emplearse en la elaboración de los mismos. Desde esta perspectiva, la presente

Tesis doctoral supone una notable contribución en el proceso de superación de dichas limitaciones prácticas y metodológicas, al tiempo que representa una innovadora incursión lingüística en la "minería" de datos ofreciods por las redes sociales, en este caso aplicada al análisis terminológico de la lengua española, a través de las muestras de lengua ofrecidas (también, si así se prefiere, en tiempo real) por la red social *Twitter*.

Los trabajos y publicaciones que, hasta la fecha, se han derivado de la investigación son los siguientes:

Artículos (en revistas indexadas):

-González Fernández, Adela. 2015. "Big Data as a Tool for Linguistic Research: Approaches to Trends in Bilingualism in Ten Latin-American Countries", *International Journal of Language and Applied Linguistics (IJLAL)*, special issue "Bilingual Education", 1: 1-12. Khate Sefid Press. ISSN: 2383-5014.

-González Fernández, Adela. 2016. "Análisis de las necesidades traductológicas en Europa a través de big data". *Skopos*, 7: 45-74. ISSN: 2255-3703 [en prensa].

Comunicaciones:

-*Innovación educativa de la titulación de Grado de Maestro de Educación Infantil*, en el 3rd. International Congress of Educational Sciences and Development. San Sebastián, 24-26 de junio de 2015.

-*La competencia bilingüe en el Grado de Educación Infantil de la Universidad de Córdoba*, en el I Congreso internacional sobre educación bilingüe. Córdoba, 17-20 de noviembre de 2015.

Pósteres:

-*Estudio de traducción y terminología contrastiva: el léxico del movimiento feminista en las lenguas inglesa, francesa y española*, en el I Congreso Científico de Investigadores en Formación en Agroalimentación de la eidA3el II Congreso científico de investigadores en formación de la UCO. Córdoba, 8 y 9 de mayo de 2012.

-*Student's assessment of the development of tutoring programs at University in the Faculty of Education*, en el 3rd. International Congress of Educational Sciences and Development. San Sebastián, 24-26 de junio de 2015.

Estancias en el extranjero:

-Luxemburgo: Prácticas en la *Dirección General de Traducción de la Comisión Europea*, del 5 al 30 de septiembre de 2011.

Por todo ello, se autoriza la presentación de la tesis doctoral.

Córdoba, 16 de marzo de 2016

Firma de las directoras

Fdo.:Dra. Mª Luisa Calero Vaquera

Fdo.: Dra. Gloria Guerrero Ramos

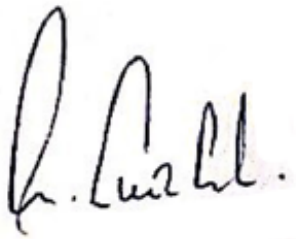
MÁS ALLÁ DEL CORPUS: BIG DATA EN LA INVESTIGACIÓN LINGÜÍSTICA. EVOLUCIÓN, ANÁLISIS Y PREDICCIÓN DEL USO DE LA LENGUA A TRAVÉS DE *TWITTER*

Tesis doctoral presentada por Adela González Fernández para optar al grado de Doctor

Córdoba, marzo de 2016

Fdo.: Adela González Fernández

La presente tesis doctoral ha sido dirigida por la Profesora Doctora María Luisa Calero Vaquera, Catedrática de Lingüística General de la Universidad de Córdoba, y codirigida por la Dra. Gloria Guerrero Ramos, Profesora Titular de la Universidad de Málaga. Esta Tesis cumple con los requisitos de la legislación vigente.



Fdo.: Profª Dra. María Luisa Calero Vaquera

A mis padres y a mi hermano.

Agradecimientos

Llegar al final de un trabajo de estas características no habría sido posible sin la implicación de muchas personas que, desde el cariño hacia la Lingüística y hacia la autora, han resultado de una ayuda inestimable.

En primer lugar, quiero expresar mi agradecimiento a la directora de este trabajo, la profesora María Luisa Calero Vaquera, quien ha sabido entenderme desde el primer momento y apoyarme en cualquier circunstancia y en cualquier idea, por arriesgada que pudiera parecer. Su aplomo y su sabiduría han estado presentes durante estos años de trabajo. Es también para mí un honor contar en esta tesis con la experiencia y el prestigio de la profesora Gloria Guerrero Ramos.

No pueden faltar palabras de agradecimiento en el ámbito académico para Alfonso Zamorano Aguilar, porque sin él nunca habría conocido el apasionante mundo de la Lingüística. También quiero agradecer su ayuda a Manuela Álvarez Jurado, su disponibilidad y su cercanía y, sobre todo, a Elena Gómez Parra, por prestarme su amistad y su confianza ciega desde el primer momento, sin darme siquiera oportunidad de devolvérselas.

En este trabajo y siempre, gracias a mi familia. A mi padre, por demostrarme que nunca se deja de aprender. A mi madre, por enseñarme el valor del trabajo y del esfuerzo, y el amor incondicional. A mi hermano, por ser mi mitad y alumbrar mis días. Ellos son mi felicidad y le dan sentido a mi vida.

Finalmente, aunque sean insuficientes, sirvan estas líneas para expresar mi profundo agradecimiento a Álvaro Maroto Conde, puntal imprescindible de mi vida, para quien nunca tendré palabras suficientes de gratitud. Este trabajo también es suyo y nunca habría sido posible sin su genialidad, su paciencia y su amor hacia mí. Mi eterno agradecimiento y devoción hacia él.

Índice

Índice de figuras

CAPÍTULO 1

CAPÍTULO 2

CAPÍTULO 3

CAPÍTULO 4

CAPÍTULO 5

CAPÍTULO 6

CAPÍTULO 7

CAPÍTULO 8

0 Introducción

0.1 Introducción y justificación

La mayor parte de los estudios en Lingüística requiere evidencias y ejemplos reales de uso de las lenguas. Esta necesidad de material auténtico ha sido suplida, desde las primeras recopilaciones de textos, por la utilización de corpus. La irrupción de los ordenadores en los trabajos manuales de acopio y gestión de material textual supuso un punto de inflexión en la historia de la Lingüística de Corpus, a partir del cual se removieron y se reestructuraron los cimientos de una ciencia todavía joven, que vio en la informática su principal punto de apoyo y el detonante de su propia revolución.

La relación entre estas dos disciplinas –la Lingüística y la Informática– ha pasado por distintas etapas desde sus inicios hasta la actualidad, y se ha vuelto tan estrecha que ya es difícil imaginarse el estudio sobre el lenguaje sin el soporte informático. Para Tognini-Bonelli (2001), la introducción de los ordenadores en el ámbito lingüístico y, más concretamente, en el trabajo con corpus, recorre fundamentalmente tres etapas. La primera de ellas consideraba a la Informática como una simple herramienta para el trabajo lingüístico –hasta el momento la mayor contribución a la Lingüística, según la autora–, gracias a la cual era posible gestionar y procesar la información de una manera más rápida y 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 removido cuestiones teóricas ya establecidas. Este hecho ha provocado que autores como Leech (1992), Halliday (1993) o la propia Tognini-Bonelli (2001) entre muchos otros, defiendan la posición de la

Lingüística de Corpus como ciencia, más allá del estatus metodológico que se le ha otorgado tradicionalmente.

Reconocen Renouf y Kehoe (2006) –y no les falta razón– 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.

Este trabajo surge como fruto de la investigación durante varios años en el ámbito de la Lingüística de Corpus y su vínculo con las nuevas tecnologías. La tesis doctoral que aquí presentamos supone, en consecuencia, un estudio interdisciplinar que aporta un enfoque novedoso para el estudio del lenguaje y de las lenguas, basado en *big data* y, en concreto, en la información que nos aportan la red de *microblogging Twitter*.

Entendemos por *big data*, *grosso modo*, los grandes conjuntos de información que por sus características no pueden ser obtenidos, gestionados ni procesados por herramientas tradicionales en un período de tiempo razonable. *Big data* se caracteriza por tres rasgos fundamentales que lo diferencian de la información tradicional: el enorme volumen de datos que lo componen, la velocidad con la que esos datos se generan y se transmiten y la variedad de formatos, temas, procedencias y tipos que lo forman. Las posibilidades que se le abren al investigador en cualquier campo, así como a las empresas e instituciones, gracias al análisis de esta información, son innumerables –trataremos de dar una muestra de ellas a lo largo del trabajo– porque gracias a *big data*

es posible realizar análisis más exhaustivos, basados en millones de datos y con muy poca inversión de tiempo, lo que repercute en mayor conocimiento y mayores beneficios, con mucho menor esfuerzo por parte del analista.

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

La elección de *Twitter* como fuente de información ha venido motivada, como decimos, por sus características textuales y por el gran alcance que tiene esta plataforma a nivel mundial. Los medios de comunicación, las empresas y los centros de investigación, tanto público como privados se están beneficiando de las ventajas que aporta el hecho de que millones de usuarios en todo el mundo lo utilicen para comunicarse. Además, tras la investigación previa, consideramos que la posibilidad de obtener en tiempo real la localización geográfica y temporal de los mensajes (tuits) publicados aporta unas ventajas inauditas para la investigación lingüística.

0.2 HIPÓTESIS DE PARTIDA

La hipótesis central y eje de este trabajo, se divide en dos ideas fundamentales:

1. La utilización de *big data* –y, en concreto, de la información contenida en *Twitter*– asistida por las tecnologías incorporadas a su manipulación, supone una mejora en la investigación lingüística, ya que enriquece las características de la metodología tradicional referidas al volumen de datos, al tiempo de recopilación y al tiempo de procesamiento de la información.

   Partimos de la idea de que, al igual que las grandes empresas, universidades y organismos oficiales están utilizando *big data* para avanzar en el ámbito de la ciencia y del conocimiento, el campo de la Lingüística no

puede ni debe quedarse atrás en este sentido, teniendo en cuenta los beneficios que el trabajo con *big data* está aportando en los casos mencionados. La Lingüística de Corpus puede verse enormemente beneficiada con esta nueva metodología, que permite obtener enormes cantidades de información textual como nunca antes se había hecho. Por otra parte, la mejora no solo se ve reflejada en el volumen de datos que podemos obtener para posteriores análisis, puesto que esto supondría un retroceso en el tiempo invertido en la investigación. Los avances tecnológicos que acompañan al progreso de la ciencia y sin los cuales no se podría concebir el trabajo con *big data* posibilitan, sin embargo, que el manejo de la información mejore en términos de recopilación, gestión y análisis de los datos. Gracias a ello, el gran volumen de información deja de ser un impedimento para convertirse en una ventaja a la hora de trabajar con corpus o con cualquier tipo de información textual con la intención de establecer conclusiones acerca de la lengua.

2. La metodología de trabajo lingüístico con *big data* incorpora nuevas dimensiones de análisis de textos, como son el etiquetado temporal y la geolocalización, lo que permite la realización de estudios sobre el lenguaje y las lenguas que no había sido posible llevar a cabo hasta el momento.

   Big data, asistido por las tecnologías oportunas, no solo ofrece información tradicional en grandes cantidades y en poco tiempo; otra de sus señas de identidad es la gran variedad de información que aporta en cualquier ámbito de conocimiento. Creemos que, en la investigación lingüística, los datos acerca del momento y del lugar en los que se produce un texto determinado son fundamentales para contextualizarlo y conocer más en profundidad las condiciones de producción lingüística que determinarán los análisis posteriores que puedan llevarse a cabo.

0.3 OBJETIVOS

Como no podría ser de otra manera, los objetivos que aquí nos planteamos se derivan de forma directa de estas dos ideas iniciales con las que trabajamos. Por lo

tanto, nuestro objetivo general consiste en demostrar la veracidad de la hipótesis y comprobar que no solo es posible utilizar *big data* para investigar el lenguaje y las lenguas, sino que, además, esta metodología supone una mejora con respecto al trabajo lingüístico tradicional porque aporta más información que hasta ahora, con una menor inversión de tiempo y de esfuerzo por parte del investigador.

Los objetivos específicos marcados para verificar la hipótesis se derivan, por tanto, de este objetivo general, y parten de la premisa de que no es posible llevar a cabo ningún estudio con *big data*, en cualquier ámbito y en el de la Lingüística, sin los medios técnicos necesarios para la adquisición, almacenamiento, gestión y análisis de la información. Podemos dividir en tres los objetivos específicos que nos planteamos cubrir con nuestra investigación:

-Crear una herramienta que permita la obtención de la información textual proveniente de *Twitter*, así como su almacenamiento y gestión, para un posterior análisis que tendrá que ser complementado, inevitablemente, por el lingüista.

-Desarrollar varios estudios que sirvan como muestra de las numerosas aplicaciones que se pueden derivar del trabajo con la herramienta y la metodología que presentamos y que demuestren la utilidad de *big data* en la investigación lingüística.

-Demostrar la conveniencia de la unión entre la Lingüística y la Informática y de la apuesta por la innovación en la investigación lingüística.

0.4 METODOLOGÍA

Con el fin de corroborar la hipótesis que acabamos de explicar, la metodología que hemos seguido en este trabajo consta de dos fases fundamentales, una de corte teórico y otra de aplicación práctica.

Durante la primera fase, hemos procurado profundizar en los aspectos teóricos que sustentan esta investigación y que la enmarcan en un entorno interdisciplinar. La segunda parte del trabajo tiene un carácter práctico y de aplicación, y se basa en la puesta en práctica de la herramienta que aquí presentamos. Puesto que la herramienta consta de cuatro módulos principales –tres de ellos diseñados para la recopilación y el análisis de la información–, presentamos distintos estudios realizados en cada uno de ellos que sirvan para ejemplificar algunas de las distintas aplicaciones que ofrece el sistema. Por este motivo, las distintas investigaciones que llevamos a cabo son breves e independientes entre sí y ninguna de ellas persigue profundizar en alguno de los temas

objeto de estudio, sino que tienen el objetivo de demostrar la utilidad de la herramienta para diferentes fines. Así, los estudios se clasifican en tres grandes bloques, que se corresponden con cada uno de los módulos de la herramienta; presentamos pues, para cada uno de ellos, la metodología específica, los resultados y las conclusiones propias.

0.5 ESTRUCTURA DEL TRABAJO

Esta tesis doctoral se encuentra dividida en dos partes fundamentales y diferenciadas entre sí. La primera de ellas constituye el *Marco Teórico* del trabajo, las bases sobre las que se asienta la investigación y que la encuadran en el marco de la Lingüística de Corpus y en su unión con *big data* y *Twitter*. Presentamos, por tanto, cuatro capítulos iniciales teóricos, dedicados a:

1. La Lingüística de Corpus
2. La web como corpus
3. *Big data*
4. *Twitter*

En el primero de ellos realizamos un recorrido histórico por la Lingüística de Corpus y aportamos algunos conceptos teóricos fundamentales en este ámbito. Después de justificar la necesidad de esta rama de la Lingüística en los estudios sobre el lenguaje, nos aproximamos al campo de la Lingüística Computacional, que tanta relación tiene con nuestro trabajo y tan unida se encuentra a la Lingüística de Corpus. En este breve acercamiento, explicaremos en qué consiste y cómo surgió la Lingüística Computacional y cuáles son sus áreas fundamentales de trabajo, para finalizar con un panorama de la situación actual en lo que a herramientas informatizadas para el análisis de corpus se refiere.

Las razones por las que existe un segundo capítulo dedicado exclusivamente a la web como corpus son varias. En primer lugar, la consideración de la *World Wide Web* como un corpus en sí misma, o como fuente para la elaboración de otros corpus, está tomando cada vez más fuerza dentro de la Lingüística de Corpus. Los científicos la consideran como la última etapa evolutiva de esta rama de la Lingüística y la mayoría de las líneas de investigación se están desarrollando en este sentido. Para nosotros, además, constituye el antecedente inmediato de nuestro trabajo –aunque su metodología

y objetivos sean radicalmente distintos–, puesto que nos proponemos utilizar la información textual contenida en *Twitter* para la elaboración de corpus de diversa índole que faciliten la investigación lingüística.

Los capítulos tercero y cuarto, dedicados a *big data* y a *Twitter*, respectivamente, se nos antojan estrictamente necesarios en nuestra investigación puesto que conforman la base sobre la que construimos nuestra hipótesis. Como cabe suponer, la perspectiva desde la que nos aproximamos a estos dos conceptos está influida por nuestra formación lingüística, aunque para una explicación más o menos completa son imprescindibles referencias al campo de la Informática, que se realizarán desde un lenguaje lo menos técnico posible.

La segunda parte de la tesis, denominada *Marco metodológico y de aplicación*, no tiene otro objetivo que dar cuenta de la metodología seguida, así como definir la estructura y el funcionamiento de la herramienta para presentar los estudios llevados a cabo para probar nuestra hipótesis en los distintos apartados que la conforman. Consta de dos partes diferenciadas en cuanto al contenido. Por un lado, el capítulo 5 (Metodología) actúa de puente entre la revisión teórica y la aplicación práctica de la herramienta por medio de distintos estudios. Por otro lado, los tres siguientes capítulos (6, 7 y 8) se centran en los estudios mencionados, dividiéndolos según el objetivo de cada uno de ellos. Esta segunda parte queda, por tanto, compuesta por los siguientes cuatro capítulos:

5. Metodología
6. Muestras de uso de *Wordics Live*
7. Muestras de uso de *Wordics One*
8. Muestras de uso de *Wordics Archive*

Así, en el capítulo 5 (Metodología), perteneciente al *Marco metodológico y de aplicación*, presentamos y explicamos la herramienta diseñada para esta investigación; mientras que en el resto de capítulos (6, 7 y 8, Muestras de uso de *Wordics Live*, Muestras de uso de *Wordics One* y Muestras de uso de *Wordics Archive*, respectivamente) aportamos una muestra de estudios lingüísticos –de los muchos posibles– llevados a cabo con esta metodología.

Conviene insistir en esta idea y recordar que no es nuestro afán cubrir por completo la enorme variedad y cantidad de estudios que nos posibilita la herramienta

que aquí presentamos, pues sería una tarea inabarcable que se apartaría del objetivo de nuestra investigación. Por ello, en los capítulos 6, 7 y 8 nos limitamos modestamente a ejemplificar con distintos estudios algunas de las posibles utilidades de la herramienta que nos permitan demostrar nuestra hipótesis de partida.

En último lugar, encontramos las conclusiones del trabajo, que formulamos desde una doble perspectiva. Por un lado, el trabajo nos conduce a conclusiones de índole teórica y disciplinar, puesto que podemos afirmar que la relación entre las dos disciplinas en torno a las cuales se construye esta tesis –la Lingüística y la Informática– es necesaria y, al mismo, tiempo, deseable para el avance y el progreso de la ciencia. En segundo lugar, también obtenemos conclusiones de carácter metodológico, al verificar la utilidad, la pertinencia y las ventajas de la utilización de *big data* y de *Twitter*, para ser más concretos, en la investigación lingüística, al mismo tiempo que enfatizamos la necesidad de disponer de las herramientas y soportes informáticos adecuados para que se puedan efectuar los distintos análisis lingüísticos.

Las páginas finales de esta tesis están dedicadas a la bibliografía utilizada. Por cuestiones prácticas y de comodidad para el lector, se incluyen los anexos en un CD adjunto al trabajo, para el que se aporta un índice al final de este.

PARTE I
MARCO TEÓRICO

1
Lingüística de Corpus

1.1 CONCEPTO DE *CORPUS*

A principios de los años 90 del siglo XX, Geoffrey Leech (1992: 106) se refirió a un corpus computerizado como "an unexciting phenomenon: a helluva lot of text, stored on a computer".

Naturalmente, cualquier otro intento de definición va más allá de esta afirmación, puesto que un corpus es algo más que un simple almacenamiento de textos en un ordenador. En este sentido, encontramos varias definiciones algo más específicas y parecidas entre sí, como la de Sinclair (1991: 171), que lo define como "a collection of natural-occurring language text, chosen to characterize a state of variety of a language" y la de Ezquerra *et al*. (1994: 10), quienes sostienen que "un corpus es un conjunto homogéneo de documentos lingüísticos de cualquier tipo (orales, escritos, literarios, coloquiales, etc.) que se toman como modelo de un estado o nivel de lengua predeterminado al cual representan o se pretende que representen".

Meyer (2002), en su defensa de la Lingüística de Corpus como metodología, explica que para entenderla así es necesario primero examinar su principal objeto de investigación: el corpus. La definición de corpus puede variar según lo general o lo concreto que se quiera ser. Su definición incluye el propósito lingüístico. Así, considera un corpus como una colección de textos o de partes de textos que pueden servir de base para un análisis lingüístico.

Por su parte, Aarts (1991: 45) profundiza un poco más y resalta una doble vertiente de corpus que no aparece en las definiciones anteriores:

> The corpus has a double function: on the one hand a corpus, especially in an enriched form, serves as a linguistic database for linguists studying the structure of the corpus language, and on the other the corpus in its raw form is for the

> corpus linguist the testbed for his hypotheses about the language, which he has expressed in a formal gramar.

Leech (1991: 8) explica que para los lingüistas estructuralistas americanos posteriores a Bloomfield, quienes estaban bajo la influencia del enfoque positivista y behaviorista de la ciencia y veían el corpus como la explicación primaria de la lingüística, corpus se definía como "a sufficiently large body of naturally occurring data of the language to be investigated". Pero señala una discontinuidad entre los lingüistas de corpus de entonces y los de ahora, y la sitúa, claro está, en Chomsky. Según Leech, treinta años después de la aparición del *Survey of English Usage* (SEU), la Lingüística de Corpus ha ampliado su alcance y su influencia hasta el punto en que se ha convertido en una corriente principal en sí misma, ese nuevo enfoque filosófico del que hablábamos antes.

También definen corpus Bowker and Pearson, en una primera aproximación, como "simply a body of text" (Bowker and Pearson, 2002: 9). Unas líneas más adelante, profundizando un poco más en el tema, indican que "a corpus can be described as a large collection of authentic texts that have been gathered in electronic forma according to a specific set of criteria". Señalan cuatro características que un corpus debe cumplir y que lo distinguen de otro tipo de recopilaciones de textos, a saber: "authentic, electronic, large and specific criteria".

Cuando estos autores hablan de "auténtico", quieren decir que esté construido a base de ejemplos reales del lenguaje y de una comunicación genuina entre personas que traten de temas cotidianos. En resumen, un corpus que se haya producido de manera natural y que no se haya creado expresamente para demostrar algo.

En segundo lugar, un corpus es "electrónico" si puede ser procesado por un ordenador. Según estos autores, el hecho de tener un corpus almacenado de forma electrónica, no solo contribuye al bien de la naturaleza, sino que nos permite el uso de programas de *software* (herramientas de análisis de corpus) que ayudan a manipular la información de muy distintas maneras. Además, ahorran tiempo y esfuerzo al lingüista, que será quien tenga después que analizar la información obtenida. Señalan también que, gracias a la aceleración del proceso a causa de la informática, los corpus[1]

[1] A lo largo del presente trabajo, utilizaremos la palabra *corpus* para referirnos al término tanto en singular como en plural, siguiendo las recomendaciones de la Real Academia Española y de la Fundéu BBVA. La palabra *corpora*, en plural, se utilizará cuando se esté escribiendo en inglés.

electrónicos son significativamente más grandes que los tradicionales, aunque la longitud exacta depende del propósito del estudio. A pesar de que esta no esté determinada, explican que, cuando especifican que un corpus debe ser grande, tienen en mente un número de textos mucho mayor de los que sería posible obtener de manera manual.

En cuarto y último lugar, Bowker and Pearson, en el mismo texto, opinan que un corpus no es una colección de textos al azar, sino que deben ser seleccionados con unos criterios específicos para que así sean representativos de un lenguaje particular. Estos criterios dependerán de la naturaleza y del objetivo del proyecto concreto que se lleve a cabo en cada momento.

Por otro lado, Parodi (2010: 25) elabora su propia definición de corpus y propone que "un corpus es una colección o conjunto de textos que está formado por al menos dos o más textos" y añade que "un corpus debe contener un número importante de textos que comparten ciertos rasgos definitorios, limitado solo por características inherentes a la naturaleza de los mismos".

Así pues, tras un breve repaso por algunas de las definiciones existentes, Corpas Pastor (2001) y el propio Parodi (2010) determinan que la definición más aceptada hasta el momento en el entorno de los lingüistas de corpus es la propuesta por el EAGLES (*Expert Advisory Group on Language Engineering Standards*) (1996: 4), un proyecto de la Unión Europea para la estandarización y la coordinación de los trabajos realizados con las diferentes lenguas de Europa, según el cual un corpus se trata de "a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language". Además, no solo EAGLES (1996), sino también Parodi (2010) y Biber *et al*. (2001) proponen una serie de recomendaciones o características representativas y relevantes a la hora de considerar un corpus, que incluyen, entre otras, referencias al tamaño, al tipo de muestras, a la representatividad, al procesamiento semiautomático, al formato o a la procedencia.

En este sentido, uno de los principales criterios que se tienen en cuenta a la hora de diseñar los corpus es el del tamaño. A pesar de que un corpus suele definirse como "una colección grande de textos", el adjetivo calificativo queda poco preciso. No hay normas establecidas que indiquen el tamaño ideal de un corpus; según Bowker and Pearson (2002), suele ser una decisión basada en las necesidades del trabajo, la disponibilidad de información y la cantidad de tiempo que tiene el investigador. No obstante, se suele asumir que para estudiar un lenguaje específico no es necesariamente

válida la premisa de que mejor será el corpus mientras más grande sea, simplemente por el hecho de que puede darse el caso de que, al tratarse de una lengua de especialidad, no aparezca el término concreto en corpus del lenguaje general o que no estén adscritos a ese ámbito específico. Además de estas razones, es más difícil y requiere más tiempo encontrar textos especializados del lenguaje específico y una muestra no excesivamente amplia puede ser lo suficientemente representativa.

Por este motivo, los corpus destinados a fines específicos suelen ser de menor tamaño que aquellos que se refieren al lenguaje general.

Sin embargo, los corpus pequeños tienen asociadas una serie de desventajas en comparación con los de mayor tamaño, entre ellas, la probabilidad de que no contengan todos los conceptos, los términos o los patrones importantes para la investigación o que no sea posible hacer generalizaciones por no disponer de información suficiente.

Otro de los factores que hay que tener en cuenta en el diseño de un corpus es si este va a ser cerrado o abierto. Los corpus cerrados contienen una información limitada que se enmarca dentro de un período de tiempo determinado. Por el contrario, los corpus abiertos disponen de una mayor flexibilidad que permite añadir, eliminar y actualizar textos, de manera que reflejan la evolución y el estado actual del lenguaje. Se tratan, los segundos, de corpus dinámicos capaces de reflejar los cambios naturales del lenguaje, que se encuentra en constante cambio y evolución. Estos corpus se conocen como *monitor corpora*.

La cuestión del tamaño de los textos se relaciona con la conveniencia de incluir los textos completos o fragmentos. De nuevo, el propósito de la investigación decidirá en este punto, teniendo en cuenta que, en un análisis del lenguaje general, fragmentos textuales pueden ser lo suficientemente representativos para la consecución de nuestros objetivos, mientras que el tamaño de un texto individual sí puede ser relevante en estudios sobre el lenguaje específico porque la información que nos interese puede aparecer en cualquier parte del documento y, además, el lugar donde esté utilizado puede ser altamente revelador.

También es interesante distinguir entre el tamaño del corpus y el número de textos que lo componen; no es lo mismo disponer, por ejemplo, de un corpus de 100.000 palabras que provengan de un número escaso de autores o de textos, que un corpus del mismo tamaño, en lo que a número de palabras se refiere, constituido por muchos textos de distinta procedencia.

El medio en el que el texto se encuentre es otro de los factores que considerar a la hora del diseño del corpus. Es obvio que la recopilación de corpus orales o semiorales entraña mayor dificultad que la de los textos escritos, entre otras razones porque la tarea de transcribir el material oral es mucho más laboriosa y requiere más tiempo que la recopilación de los textos escritos. Además, la estudiada actitud de los hablantes cuando son grabados, la posible intimidación o la falta de naturalidad que pueden adquirir son inconvenientes que también aparecen en estos casos.

Otros factores, como el tema del corpus, el tipo de texto (entre experto-experto, entre experto-no experto, no experto-no experto, etc.), la autoría del documento y de la edición de este, el idioma (material auténtico o traducido) o la fecha de publicación deben ser tenidos en cuenta en el diseño de los corpus.

1.2 ORÍGENES Y EVOLUCIÓN DE LA LINGÜÍSTICA DE CORPUS

La historia de la Lingüística de Corpus (LC), considerando a esta tal y como la entendemos hoy en día, es relativamente corta.

Hasta el siglo XIX, la tradición de trabajos lingüísticos basados en corpus se caracterizaba por una recopilación de textos con la finalidad de tener el único acercamiento posible al estudio de lenguas muertas. Más adelante, hasta mediados del siglo XX, los corpus se empezaron a utilizar para investigar acerca de la adquisición del lenguaje infantil, para el establecimiento de normas ortográficas y estudios de lenguas extranjeras, fundamentalmente –glosarios, estudios comparativos, gramáticas, etc. También Kruisinga (Van Essen, 1983) y Poutsma (Svartvik, 1992) recopilaron ejemplos para el estudio de la gramática. No podemos olvidarnos tampoco de ejemplos como Käding, que recopiló unos 11 millones de palabras del alemán para comprobar la frecuencia de distribuciones y secuencias de letras. Este corpus, simplemente por su tamaño y por la época en la que se elaboró –finales del siglo XIX- es comparable a los corpus actuales (McEnery y Wilson, 2003). O Thorndike, que también recopiló, en 1921, un corpus millonario, de unos 4,5 millones de las palabras más frecuentes del inglés (Rojo, 2008). En 1944 trabajó conjuntamente con Lorge en un corpus de 18 millones de formas para elaborar una lista de las 30.000 palabras más frecuentes del inglés (McEnery y Wilson, 2003; Rojo, 2008).

Durante estos años, encontramos corpus lexicográficos, como el que utilizó James Murray para la elaboración del *Oxford English Dictionary*. La primera edición

del diccionario se completó en 1928, con más de cuatro millones de fichas de citas aportadas por voluntarios. Murray suplió la falta de herramientas informáticas con el trabajo y la ayuda de su familia, que le ayudó a poner en orden las fichas (Murray, 1977).

También Alexander Ellis necesitó la ayuda de unas ochocientas personas para la elaboración de *The Existing Phonology of English Dialects* (Svartvik, 2007). Se trataba de trabajos colosales realizados de forma manual en los que se analizaban cantidades masivas de ejemplos reales de la lengua. Otto Jespersen explica en su autobiografía el proceso de la siguiente manera:

> I am above all an observer I quite simple cannot help making linguistic observations. In conversations at home and abroad, in railway compartments, when passing people in streets and on roads, I am constantly noticing oddities of pronunciation, forms and sentence constructions … For these notes I have found it practical to use small slips of paper … It is imposible for me top ut even a remotely accurate number on the quantity of slips I have had or still have: a lot of them have been printed in my books, particularly the four volumes of Modern English Grammar, but at least just as many were scrapped when the books were being drafted, and I still have a considerable number of drawers filled with unused material. I think a total of 3-400,000 will hardly be an exaggeration. Darwin (*sans comparaison !*) says: "I am a complete millionaire in odd and curious facts". But he also says: "My mind seems to have become a kind of machine for grinding general laws out of large collections of facts. (Jespersen, 1838: 213-215, *apud* Svartvik, 1992: 7)

En esta primera mitad del siglo XX y en el marco del estructuralismo americano, la LC se asentó como metodología empírica y como única vía para el estudio de las lenguas. Gries (2009) no quiere olvidarse de esta rama americana de los primeros lingüistas, que también contribuyeron a los orígenes de la LC, como los trabajos de Sapir, de Bloomfield o de Harris. También Fries, en *The Structure of English*, recopiló evidencias lingüísticas de conversaciones reales para estudiar la estructura del inglés oral (Fries, 1973).

Estos corpus, a diferencia de los anteriores, incluían datos orales y también tenían como finalidad el estudio de lenguas vivas y de su fonética (Villayandre, 2010). Sin embargo, puesto que la recopilación y el análisis se efectuaban de forma manual, la

representatividad era muy escasa, debido a la imposibilidad del manejo de datos suficientes.

Por tanto, la extracción manual de información textual ha sido la forma tradicional de recopilar información para la descripción lingüística. Sin embargo, en su forma moderna, lo que últimamente se ha venido llamando *Lingüística de Corpus*, se remonta a la década de los sesenta. El inicio de este nuevo enfoque –grandes recopilaciones de textos procesables por una máquina– estuvo marcado por la aparición de los ordenadores y de la posibilidad de introducir textos. Es en esta etapa cuando comienza la llamada "segunda generación de Lingüística de Corpus". No obstante, ya en los años 40, Roberto Busa había establecido el vínculo entre los corpus y los ordenadores (McEnery y Hardie, 2012), demostrando que las concordancias se podían aplicar de forma efectiva a los textos electrónicos y dando paso, así, a lo que conocemos como la primera generación de los programas de concordancia.

Estas recopilaciones presentaban características muy evolucionadas en comparación con sus predecesoras. El punto de inflexión, como decimos, fue la incursión de la informática, pero esto tuvo consecuencias directas, como el aumento del número de palabras (un millón), el carácter representativo –Alphonse Juilland desarrolló diccionarios basados en corpus con un grado similar de representatividad en varios idiomas (McEnery y Hardie, 2012)– de los datos y el predominio de textos escritos debido a las dificultades técnicas que encontraban los ordenadores para el trabajo con documentos orales.

El primero de los corpus[2] pertenecientes a esta época empezó a gestarse en Inglaterra en 1959, con Randolph Quirk a la cabeza. El creador del *Survey of English Usage Corpus* (SEU)[3] se mudó de Durham al *London University College* para emprender el primer gran proyecto de recolección de palabras del inglés escrito y hablado. Muchos lingüistas que posteriormente han sido grandes referencias en el campo de la LC participaron en este proyecto que, a pesar de no estar informatizado en un primer momento (en la actualidad sí lo está), reunía todas las características para situarlo en la segunda generación de la LC. El corpus, una vez terminado, está

[2] Svartvik (2007) explica que todavía en los primeros momentos del *Survey of English Usage*, el término *corpus* no era muy común y que Randolph Quirk todavía usaba en los primeros días del *Survey of English Usage* las variantes "*descriptive register*", "*primary material*" o "*texts*". Tampoco estaban seguros de la formación del plural en la palabra *corpus*; Svartvik cuenta cómo una mañana, tomando café, dudaban entre si la forma correcta del plural debía ser *corpuses* o *corpora*, hasta que alguien entre los asistentes dio por concluida la discusión diciendo: "Yo creo que es *corpi*".

[3] http://www.ucl.ac.uk/english-usage

compuesto por un millón de palabras, pertenecientes a 200 textos de 5000 formas cada uno de inglés británico oral y escrito, desde 1955 hasta 1985 (UCL, 2011). De este modo, introduce novedades con los corpus existentes hasta el momento porque no solo le presta la misma atención a la lengua oral que a la escrita, sino que lleva a cabo un registro completo de los fenómenos que contienen los textos estudiados (Rojo, 2008). El SEU, debido a que su material se ha recopilado durante treinta años, se puede considerar como el primer intento de aportar una fuente de información para el estudio diacrónico del inglés británico (McEnery y Hardie, 2012).

Uno de los lingüistas que ayudó a la elaboración del corpus que sentaría las bases de la LC posterior fue Jan Svartvik, quien se trasladó desde la Universidad de Uppsala a la Universidad de Durham con una beca –bajo la dirección de Quirk– para emprender la recopilación de textos orales recogidos del *Noveno Congreso Internacional de Lingüistas* en Cambridge, Massachusetts. Junto a él, David Crystal, Sidney Greenbaum o Geoffrey Leech se encontraban, entre otros, dentro del prestigioso elenco de lingüistas que participaron en este proyecto y del que formaron parte para publicar dos grandes obras de referencia que tuvieron de base el SEU. Estas obras son *A Grammar of Contemporary English* (Quirk *et al., 1972*) y su edición ampliada, *A Comprehensive Grammar of the English Language* (Quirk *et al*., 1985), elaboradas por el llamado "gang of four" –Quirk, Svartvik, Greenbaum y Leech.

Svartvik recuerda con cariño esta época y el momento en el que W. Nelson Francis, de la Universidad de Brown (EE.UU.), llegó al despacho de Randolph Quirk y soltó encima de su mesa una de las antiguas cintas de ordenador a la vez que pronunciaba las palabras: "*Habeas corpus*" (Svartvik, 2007: 14). Se trataba, naturalmente, del *Brown University Standard Corpus of Present-Day American English*[4], conocido como *Brown Corpus* y elaborado por W. Nelson Francis y Henry Kučera. Este fue el primer corpus que se construyó pensando en ser gestionado por un ordenador y, por ende, por programas informáticos, y el que verdaderamente marcó la transición a la nueva etapa de la LC. Este segundo proyecto, en el que Quirk trabajaba de asesor, está constituido también por un millón de palabras, procedentes de 500 muestras, con 2000 palabras cada una. El tamaño relativamente amplio de muestras se debe a que, debido a las limitaciones de los ordenadores de la época, la

[4] http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM

representatividad se conseguía con un conjunto amplio de muestras de tamaño reducido (Rojo, 2008).

Sin embargo, la irrupción de la figura de Chomsky en el panorama lingüístico a mediados del siglo XX y la expansión del generativismo, que coincidieron con la aparición del *Brown Corpus*, supusieron un obstáculo importante para el desarrollo de la Lingüística de Corpus y para todo aquello que tuviera que ver con frecuencias, estadísticas, hechos lingüísticos concretos y reales, etc., a lo que por aquel entonces se oponían frontalmente. El cambio radical del enfoque de los estudios lingüísticos y la imposición de la filosofía racionalista como única vía para el estudio del lenguaje ganaron terreno frente a la metodología empírica sobre la que se sustentaba el trabajo de la LC.

A pesar de que Svartvik (2007) asegura que su círculo de trabajo no se vio demasiado afectado por la corriente generativista dominante –aunque admite que, a veces, llamarse lingüista de corpus era algo parecido a que su nombre apareciera en las listas de pasajeros del *Titanic*– (Svartvik, 2007: 15), lo cierto es que este fue el motivo por el que se rompió la continuidad entre la LC de los estructuralistas americanos y la LC que conocemos actualmente.

Las críticas vertidas por Chomsky al empirismo de los corpus son obvias. Partiendo de la intuición como único recurso válido para la investigación lingüística y de la idea de que la labor del lingüista es reflejar la competencia, que será quien determine la gramaticalidad o agramaticalidad de los textos, la actuación y los corpus no son elementos válidos para el estudio del lenguaje. Además, argumentaba que la naturaleza cerrada de los corpus de la época no podía reflejar el carácter infinito de las lenguas que, a pesar de tener un inventario finito de signos, ofrecían infinitas posibilidades de combinaciones entre ellos. Por otra parte, la introspección, como único criterio válido para el estudio de las lenguas, ahorra tiempo al lingüista empírico, que tiene que invertir horas y esfuerzo en la elaboración de los corpus que no dan cuenta del estado de una lengua.

> Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite.

> The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list. (Chomsky, 1958: 159)[5]

Abercrombie (1965) se sumó a las críticas realizadas por Chomsky con una vertiente más práctica, ya que argumentaba que la LC del momento requería una serie de técnicas y herramientas informáticas sin las cuales la investigación no solo era cara, sino costosa y poco fiable.

La lingüística generativa, por tanto, dejaba poco espacio para lo que no se ajustara a lo que ellos consideraban una práctica lingüística aceptable, por lo que el primer pilar que Francis y Kučera sentaran para la Lingüística de Corpus no tuvo una gran acogida por muchos lingüistas en aquel primer momento. En palabras de Francis (1992: 28, *apud* Meyer, 2002), la corriente lingüística predominante describió la creación del *Brown Corpus* como "a useless and foolhardy enterprise" porque la intuición del hablante nativo era "the only legitimate source of gramatical knowledge", y eso no se podía obtener con un corpus. El mismo Francis fue protagonista de una conversación en 1962 en la que Robert J. Lees, discípulo de Chomsky y, claro está, partidario de la lingüística generativa, le preguntó al creador del *Brown Corpus*, el día que se conocieron en una conferencia, a qué se dedicaba; cuando Francis le contó lo que traía entre manos (a cargo de una beca de la Oficina Estadounidense de Educación), Lees le contestó que era una auténtica pérdida de tiempo y del dinero del gobierno, y que, como hablante nativo que era Francis, él solo era capaz de producir en diez minutos más ejemplos de la gramática inglesa que los que pudiera encontrar en millones de textos elegidos al azar (Francis, 1982).

No es esta la única anécdota al respecto que cuenta Francis, quien empezó una conferencia en 1984 advirtiendo a su público de que llevaba puesto un pisacorbatas con la forma de una llave inglesa que le habían regalado sus alumnos después de constituir el "linguistic plumber's union", cuyo emblema era esa herramienta. Esto se produjo como respuesta irónica a un colega suyo de literatura irlandesa, llamado David O'Kraus, después de que este le dijera que cualquiera que utilizara un ordenador para trabajar con literatura de calidad no era más que un fontanero. Un tiempo más tarde, volvieron a coincidir y O'Kraus le comentó a Francis, haciendo referencia a su

[5] Palabras pronunciadas en la *3rd Texas Conference on Problems of Linguistic Analysis*, citado en Leech (1991: 8), Rojo (2008) y Svartvik (2007: 15).

encuentro anterior: "Ah, Nelson, me bhoy, it was not you I was after callin' a plumber; it was them other fellows like Henry Kučera" (Francis, 1985: 5[6]).

Charles J. Fillmore ilustra estas dos visiones oponiendo, por un lado, a los lingüistas de sillón y, por otro, a los lingüistas de corpus, y los define de la siguiente manera:

> Armchair linguistics does not have a good name in some linguistics circles. A caricature of the armchair linguist is something like this. He sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abroptly shouting, "Wow, what a neat fact!", grabs his pencil, and writes something down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like. (Fillmore, 1992: 35)

Por el contrario:

> Corpus linguistics does not have a good name in some linguistics circles. A caricature of the corpus linguist is something like this. He has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he es busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus as the second word of a sentence. (Fillmore, 1992: 35)

Concluye el lingüista americano, que se reconoce a sí mismo como un lingüista de sillón, haciendo un llamamiento a la unidad y reconociendo la necesidad de la unión de ambas corrientes para un adecuado análisis de las lenguas:

> I have two observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicón and gramar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way. My conclusion is that the two kinds of linguists

[6] Discurso pronunciado en la cena de la *5ª Conferencia del ICAME* en Windmere, Inglaterra, el 21 de mayo de 1985. Disponible en *ICAME News*, 10:5-7.

> need each other. Or better, that the two kinds of linguists, wherever possible, should exist in the same body. (Fillmore, 1992: 35)

El tiempo ha terminado demostrando que las afirmaciones de Chomsky acerca del carácter sesgado de los corpus y de la inutilidad de las estadísticas estaban equivocadas, dejando claro que el uso de los corpus era, y sigue siendo, fundamental en numerosas parcelas del estudio del lenguaje. Sin embargo, nos recuerda Rojo (2010: 1153) que no podemos olvidar el contexto en el que estas críticas fueron realizadas y que, de algún modo, eran ciertas porque la concepción de corpus que el padre de la lingüística generativa tenía en mente es distinta de la realidad actual. En parte, la superación de la falsa suposición de que toda la Lingüística de Corpus es exclusivamente descriptiva y de que a los gramáticos generativistas no les interesa la información sobre la que se basan sus teorías, junto con la de que el análisis de los corpus no contribuye a la teoría lingüística, han facilitado la reconciliación de posturas. Según Leech (1992, *apud* Meyer, 2002), el principal argumento utilizado por parte de los generativistas en contra de la Lingüística de Corpus es que la información está más centrada en la actuación que en la competencia y, por tanto, es más descriptiva que teórica. Sin embargo, continúa Leech, esta distinción está exagerada y no es tan grande como se suele creer, ya que la actuación es producto de la competencia.

Volviendo a los corpus de la segunda generación, parece comúnmente aceptado, pues, que la historia de la Lingüística de Corpus viene ligada a la existencia de los ordenadores, lo que puede ser la explicación a que no se haya profundizado demasiado en sus antecedentes, aunque, si citamos al pionero y creador del primer corpus informatizado, Francis, no está muy de acuerdo con esta afirmación y habla de los antecedentes del *Brown Corpus* como "language corpora B.C." –i.e. *before the use of computers*. (Francis, 1992: 17).

En un alarde de humildad –y refiriéndose solo a los corpus del mundo anglosajón– explica Francis que, antes del *Brown Corpus*, ha habido muchos otros e importantes (Francis, 1992). Sin embargo, para hablar de ellos es fundamental tener en cuenta los propósitos lingüísticos que caracterizan a los corpus y que los distinguen de otras recopilaciones de textos, como puede ser el *Oxford Book of English Verse*, que tiene un propósito literario, o el *Corpus Juris Civilis* de Justiniano. Precisamente por este motivo, realiza especial hincapié en la última frase de su propia definición de corpus: "a collection of texts assumed to be representative of a given language, dialect,

or other subset of a language, to be used for linguistic analysis" (Francis, 1992: 17). Señala también aquí este autor como una aproximación algo más cercana a lo que hoy en día podemos considerar como corpus el *Latin Corpus Glossary*, del siglo XVIII, y lo menciona como el posible primer ejemplo de diccionario bilingüe latín-inglés.

En definitiva, Francis se centra en tres grandes líneas de corpus: la lexicografía, la dialectología y la gramática. En la primera, menciona los tres grandes proyectos lexicográficos en lengua inglesa, a saber: el diccionario de Johnson, en el siglo dieciocho; el trabajo de Murray para el *Oxford English Dictionary* (OED), en el siglo diecinueve y el de los editores Merriam-Webster, en el veinte. Entre el segundo tipo de corpus, aquellos con el propósito de elaborar atlas dialectológicos, destacan los de Ellis y Wright en el siglo diecinueve, y los de Kurath, Lowman y McDavid, por un lado, y Orton y Dieth, por el otro, en el siglo veinte. Por último, entre los corpus gramaticales se centra en el antecedente inmediato del suyo propio, el *Survey of English Usage Corpus* (SEU), del que ya hemos hablado.

Desde el punto de vista de Leech (1991: 9), este último pertenece a la primera era de los corpus, mientras que el de Francis se encuentra en la segunda (aunque, posteriormente, por el hecho de estar informatizado el SEU, se ha venido incluyendo también en la segunda etapa). La principal desventaja de los corpus informatizados, desde el punto de vista de Leech, es la tendencia a dejar a un lado la información oral (por su dificultad a la hora de registrarla) –esta objeción queda ahora obsoleta– para favorecer los ejemplos escritos; el SEU, por tanto, se libraba de este inconveniente y mantenía aproximadamente un cincuenta por ciento de datos orales y otro tanto de escritos.

Los dos corpus que inauguraron la época dorada de la Lingüística de Corpus fueron pronto sucedidos por otros, entre los que destaca el *Lancaster-Oslo/Bergen* (LOB) *Corpus*[7], bajo la dirección de Geoffrey N. Leech (Universidad de Lancaster), en colaboración con Stig Johansson –en su primera fase, en Lancaster, puesto que en la segunda etapa, en Oslo, fue Johansson quien tomó el mando de la dirección– y el *Norwegian Computing Center for the Humanities*. Este corpus se ideó como el equivalente británico al *Brown Corpus* e, igual que este, está compuesto de 500 fragmentos, con 2000 palabras cada uno y se elaboró siguiendo los mismos principios que su homólogo americano.

[7] http://clu.uni.no/icame/manuals/LOB/INDEX.HTM

En 1975, tras la vuelta de Svartvik a Suecia en la Universidad de Lund, el que fuera el joven ayudante de Quirk en los inicios del SEU, emprendió, con la ayuda de una serie de alumnos de postgrado, la informatización de los textos orales del SEU, proyecto que se conoció como el *Survey of Spoken English* (SSE). La unión entre el SEU –cuyo director era Greenbaum desde que sucediera a Quirk en 1983– y del SSE dio lugar al *London-Lund Corpus of Spoken English* (LLC)[8].

En este punto, la historia de la Lingüística de Corpus del inglés (ECL –English Corpus Linguistics) avanzaba más en la investigación del inglés británico que en la del inglés americano. A principios de los años 90, los investigadores tenían la posibilidad de acudir a compilaciones escritas de inglés británico que se elaboraron después del SEU, como el *British National Corpus* (BNC) o el *Bank o English*, así como de estudiar la lengua oral con el *London-Lund Corpus*. Sin embargo, el *Brown Corpus* seguía siendo el mayor corpus público disponible de inglés americano y los corpus orales eran inexistentes (McEnery y Hardie, 2012).

Unos años más tarde, el equipo de *University College London*, con Sidney Greenbaum a la cabeza, emprendió el que hasta el momento se considera el mayor corpus para el estudio comparativo de variedades del inglés, el llamado *International Corpus of English* (ICE)[9]. La idea original era la compilación de corpus de inglés oral tanto británico como americano para establecer comparaciones entre las dos variedades, pero el LLC no podía servir como el componente británico porque los textos para comparar debían ser coetáneos, y los del LLC se compilaron en los treinta años que siguieron a 1955 (Greenbaum, 1991, 1992). El lingüista reconoce que, a pesar de la existencia de corpus de distintas variedades del inglés computerizados en India (*Kolhapur Corpus*) o en Australia (*Macquarie Corpus*) que seguían el modelo del *Brown* y del LOB, estos carecían de parte oral, como ocurría en Estados Unidos. Este hecho le hizo pensar que el corpus tendría mucho más valor si incluía otras variedades del inglés, y ese fue el resultado final. El inicio del proyecto de un corpus internacional se sitúa en enero de 1988 cuando Charles Meyer aceptó la compilación de un corpus oral del inglés americano, como parte del esfuerzo internacional (Greenbaum, 1991). El corpus final cuenta con un nutrido número de variedades del inglés, entre las que se encuentran las de Canadá, Hong Kong, Nigeria, India o Irlanda[10]. Greenbaum (1991)

[8] http://clu.uni.no/icame/manuals/LONDLUND/INDEX.HTM
[9] http://www.ucl.ac.uk/english-usage/projects/ice.htm
[10] Consultar http://ice-corpora.net/ICE/INDEX.HTM para más información.

admite que, tanto la fecha de recopilación (textos pertenecientes al período comprendido entre 1990 y 1993) y el número de palabras (un millón, que representan formas orales y escritas de la lengua) son cifras arbitrarias –aunque el número de palabras siguió la práctica comenzada por el SEU, el *Brown* y el LOB–, pero lo suficientemente representativas para que el corpus fuera adecuado.

Los esfuerzos en los estudios de ECL se unificaron en una organización internacional que se creó en febrero de 1977 en Olso (Leech y Johansson, 2009), llamada ICAME (*International Computer Archive of Moder English*). ICAME es el punto de encuentro para los investigadores en ECL, quienes fomentan el desarrollo científico a través de conferencias anuales y la publicación de una revista científica de difusión pionera en este ámbito, denominada *ICAME Journal*.

Esta visión de la historia de la LC, centrada en la ECL, no queda completa para los intereses de esta investigación sin aludir a otras tradiciones científicas, fundamentalmente la hispánica.

En este sentido, la Real Academia Española publicó entre 1726 y 1739 el *Diccionario de autoridades*[11] o el *Diccionario de construcción y Régimen de la lengua castellana*, que comenzó Cuervo en 1872, con una selección equilibrada de obras y de ejemplos (Montes, 1998; Rojo, 2008). Tampoco podemos olvidar el proyecto de frecuencias de palabras de Juilland y Chang en 1964, como menciona Rojo (2008) o el *Diccionario fraseológico del Siglo de Oro*, de Julio Cejador. Rojo (2008), en este repaso por la tradición hispánica de corpus, no deja atrás tampoco obras como las de Keniston acerca de la prosa y la sintaxis españolas o la de Salvador Fernández Ramírez con su *Gramática española* de 1962.

También a la Real Academia Española pertenece el *Corpus diacrónico del español*[12] (CORDE). Este corpus cuenta con doscientos cincuenta millones de registros escritos pertenecientes a distintos géneros (narrativo, lírico, dramático, científico-técnico, histórico, jurídico, religioso, periodístico, etc.), de todas las épocas (Edad Media, Siglos de Oro y Época Contemporánea) y lugares en los que se habló el español desde sus inicios hasta 1974. Por su perspectiva diacrónica, otorga un 74% de su contenido para el español peninsular y un 26% para Latinoamérica (Pitkowsky y Vásquez, 2009). Se trara, por tanto, de la fuente de consulta de referencia para un

[11] http://www.rae.es/recursos/diccionarios/diccionarios-anteriores-1726-1996/diccionario-de-autoridades
[12] http://www.rae.es/recursos/banco-de-datos/corde
Se puede consultar el corpus en: http://corpus.rae.es/cordenet.html

estudio diacrónico de la lengua española; además, sirvió como base para la elaboración del *Nuevo diccionario histórico del español*.

Más recientemente, y coincidiendo con el final del CORDE, se creó en 1975, en el seno también de la Real Academia Española, el *Corpus de referencia del español actual*[13] (CREA). Este corpus está compuesto por un conjunto de documentos informatizados de diversa índole y procedencia. La primera versión del CREA incluye textos fechados desde 1975 hasta 2004, con textos escritos pertenecientes a libros, revistas y periódicos, así como textos orales procedentes de grabaciones de radio o televisión. Estos textos tienen su origen, aproximadamente en la misma proporción, en España y de Latinoamérica, aunque el porcentaje no se encuentra tan igualado cuando se trata de material procedente de la lengua oral (presente en un 10%) y de la lengua escrita (que constituye el 90%). Desde su última versión, de junio de 2008, dispone de algo más de ciento sesenta millones de formas procedentes de textos que se han agregado al bloque de la prensa americana y de los libros, correspondientes al período de tiempo comprendido entre el año 2000 y el 2004. Además, se añaden frecuencias absolutas y normalizadas de formas ortográficas, a la vez que se elimina la diferenciación entre mayúsculas y minúsculas, y se suprimen las cifras (Pitkowsky y Vásquez, 2009).

No podemos cerrar el recorrido por los corpus más representativos del español sin mencionar los creados en el marco del Instituto Universitario de Lingüística Aplicada (IULA) de la Univerdad Pompeu Fabra, entre los que destaca el proyecto *Corpus Tècnic*[14] y todas las herramientas construidas en torno a este dedicadas a su explotación lingüística.

Estos dos últimos pertenecen a la última generación de corpus, de millones de palabras e informatizados, a la que también pertenecen corpus como el *British National Corpus*[15] (BNC), el *Bank of English*[16] (BoE), que hemos mencionado anteriormente, o el *Corpus of Contemporary American English*[17] (COCA). El primero de ellos cuenta con cien millones de palabras de inglés británico oral y escrito. Se construyó entre 1991 y 1994 y fue elaborado por la editorial *Oxford University Press*, junto con otras editoriales, entidades y universidades, entre las que destacan la Universidad de

[13] http://www.rae.es/recursos/banco-de-datos/crea
Se puede consultar el corpus en: http://corpus.rae.es/creanet.html
[14] https://www.iula.upf.edu/recurs01es.htm
[15] http://www.natcorp.ox.ac.uk/
[16] http://corpus.byu.edu/coca/compare-boe.asp
[17] http://corpus.byu.edu/coca/

Lancaster y la Universidad de Oxford. El BoE (que forma parte del proyecto COBUILD) comenzó a compilarse en 1991 en la Universidad de Birmingham bajo la dirección de John Sinclair y contiene más de quinientos millones de palabras. Por su parte, el COCA almacena más de cuatrocientos cincuenta millones de palabras recopiladas desde 1990 hasta 2012. Se trata del mayor corpus de inglés americano, con más de ciento sesenta mil textos orales y escritos procedentes de diversos géneros.

1.3 LINGÜÍSTICA DE CORPUS: APROXIMACIÓN AL CONCEPTO

> "Corpus-linguistics is normal linguistics, and any other kind of linguistics is something odd or special that needs to be justified"
>
> Christian Mair (1992: 98)

Hace algo más de medio siglo, como ya hemos apuntado, la Lingüística de Corpus tal y como se conoce hoy en día, se inició como un campo complementario a la corriente de la Lingüística general. Sin embargo, a pesar de su corta historia, ha sufrido una evolución vertiginosa y no solo ha sabido superar las dificultades que encontró en sus inicios, sino que se ha convertido en una herramienta fundamental en prácticamente cualquier investigación lingüística. Así lo explicaba Wallace Chafe (1992: 96) en el primer *Simposio de lingüística de corpus* celebrado en Estocolmo y frente a un surtido número de sus colegas más representativos en este campo:

> What, then, is a "corpus linguist"? I would like to think that it is a liguist who tries to understand language, and behind language the mind, by carefully observing extensive natural samples of it and then, with insight and imagination, constructing plausible understandings that encompases and explain those observations. Anyone who is not a corpus linguist in this sense is, in my opinion, missing much that is relevant to the linguistic enterprise.

La expansión de los estudios de Lingüística de Corpus ha traído consigo una revolución en los estudios sobre el lenguaje que ha quedado patente en todos los ámbitos, como los estudios traductológicos, los estudios del lenguaje para fines específicos, la investigación del procesamiento del lenguaje natural, la lingüística computacional o la enseñanza de idiomas, entre otros.

No obstante, y a pesar de opiniones bastante extendidas similares a la de Halliday (1993) –quien afirma que la Lingüística de Corpus ha provocado un salto cualitativo en nuestra concepción sobre el lenguaje– su definición, como ocurre con muchas otras áreas científicas, no está exenta de polémica ni de opiniones contradictorias. El principal obstáculo a la hora de llegar a un consenso estriba en la consideración de la Lingüística de Corpus bien como metodología o ben como teoría con estatus propio. La literatura a favor de ambas consideraciones abunda, aunque otros expertos, como Gries (2009), consideran que las consecuencias prácticas de una y otra postura no son tan distintas.

Autores como McEnery y Hardie (2012) son partidarios de una visión de la Lingüística de Corpus no como una disciplina que se ocupa del estudio de un aspecto concreto del lenguaje, sino como un conjunto de procedimientos o métodos para su estudio. Meyer (2002) o Bowker y Pearson (2002: 9) tienen una opinión parecida, cuando estos últimos definen lingüística de corpus como "an approach or a methodology for studying language use. It is an empirical approach that involves studying examples of what people have actually said, rather than hypothesizing about what they might or should say". Afirman también que la Lingüística de Corpus hace un gran uso de la informática, lo que permite que la información pueda ser manipulada de formas en las que resultaría imposible con material impreso, aunque autores como De Kock (2001) ven su relación la informática como un impedimento para su estatus metodológico. A estas ideas de corte metodológico se opondrán otros lingüistas que no están de acuerdo con la caracterización de la LC como metodología, como los teóricos llamados Neo-Firthianos (McEnery and Hardie, 2012), entre los que destacan J. Sinclair (1991), M. Stubbs (2007), Teubert (2005), Chafe (1992) o Tognini-Bonelli (2001). Estos autores, relacionados de un modo u otro con la Universidad de Birmingham, comparten con J. R. Firth la idea de que el corpus en sí mismo conforma una teoría propia acerca del lenguaje.

En este sentido, Parodi (2010: 15) explica que:

> La LC no se entiende como una rama o un área de la lingüística tal como son la fonología, la semántica, la sintaxis, sino que como un método de investigación que puede ser empleado en todas las ramas o áreas de la lingüística, en todos los niveles de la lengua y desde enfoques teóricos diferentes. Sus aplicaciones son múltiples y no limitan las posibilidades de indagación.

En su opinión, a pesar de que le otorga cierto carácter metodológico, apunta que no puede considerarse una metodología estricta porque permite diversas opciones a la hora del estudio del lenguaje y tiene sus propios principios reguladores. Para este investigador de la Escuela Lingüística de Valparaíso, la suya es una visión interdisciplinar que, aunque deja abierta la puerta a una visión metodológica del concepto, se aleja de posturas mucho más radicales, como la de Teubert (2005) quien afirma que la LC constituye un enfoque teórico para el estudio del lenguaje o la de Mahlberg (2005), que sostiene que la LC es en sí misma un marco teórico propio e independiente.

También Gatto (2008: X) resalta la conveniencia de trabajar con la LC como enfoque filosófico para comprender la relación entre las "implicaciones teóricas de la lingüística de corpus" y el "surgimiento de nuevos métodos que extraigan el máximo beneficio del inmenso potencial de la web como corpus", aspecto que trataremos más adelante. Desde esta perspectiva, Stubbs (2007: 127) afirma:

> Studies of large corpora provide two main contributions to linguistics. First, they provide many new and surprising facts about language use. This is an important test for an approach to language study: it can help us to learn new things. Second, by looking at language from a new point of view, corpus studies can help to solve paradoxes which have plagued linguistics for at least a hundred years.

La visión de Gatto de la LC como un "nuevo enfoque filosófico" es compartida por lingüistas como Leech (1992) o como Tognini-Bonelli (2001), quienes opinan que se ha convertido en una nueva forma de entender la investigación lingüística y en un área de conocimiento capaz de definir sus propias reglas. No obstante, estas palabras de Leech no deben llevarnos a engaño en su aparente defensa de la LC como marco teórico propio porque, como aclaramos unas líneas más abajo, él es uno de los defensores de su carácter metodológico e instrumental.

La propia Tognini-Bonelli (2001) introdujo las expresiones *corpus-driven linguistics* y *corpus-based linguistics* para explicar estos dos puntos de vista y sus diferencias a la hora de definir el concepto de lingüística de corpus.

Para ella, el término *corpus-based linguistics* se refiere a una metodología cuya finalidad es explicar, probar o ejemplificar teorías o descripciones que fueron formuladas antes de la aparición de los grandes corpus. Explica que, tradicionalmente, las teorías lingüísticas son el resultado de reflexiones obtenidas a partir de una gran experiencia y un estudio exhaustivo, sumados a la competencia o intuición del hablante nativo. La naturaleza de esa competencia no se puede determinar porque no es parte de una metodología de investigación, sino del propio investigador. Desde su punto de vista, los lingüistas que trabajan de esta forma se posicionan en un punto en el que trabajan con datos y teorías que consideran adecuados, de modo que "sesgan la información". El motivo por el que el corpus se considera útil en estos casos es porque, a veces, señala pequeños errores que aparecen en el modelo adoptado y, además, porque son una "fuente valiosa de información cuantitativa" (Tognini-Bonelli, 2001: 66).

Sin embargo, como herramienta para estudiar el lenguaje, difiere de otras metodologías en la falta de sistematicidad procedimental y metodológica. Si bien es cierto que hay muchas generalizaciones que se pueden hacer respecto a la Lingüística de Corpus y que algunas de sus ideas más importantes están bastante consensuadas, la realidad es que se trata de un área heterogénea.

Por el contrario, continúa Tognini-Bonelli (2001), en el enfoque denominado *corpus-driven*, el corpus vas más allá de un mero depósito de ejemplos que respalden teorías ya elaboradas, los enfoques teóricos reflejan de una forma directa las evidencias aportadas por el corpus; es decir, los ejemplos se toman de forma literal y no se ajustan para que encajen en las categorías predeterminadas del lingüista. Los patrones, las listas de frecuencias y los datos contenidos en el corpus conforman la base y las evidencias para la elaboración de la teoría, del mismo modo que la ausencia de elementos también es altamente significativa. Para ella, la teoría no existe de manera independiente a los ejemplos reales, y los estudios teóricos y descriptivos "reflejan la evidencia" (Sinclair, 1991: 4, *apud* Tognini-Bonelli, 2001).

En cualquier caso, parece comúnmente aceptado que, a pesar de que los procedimientos sigan desarrollándose y renovándose (como es nuestra intención demostrar en este trabajo), hay algunos recursos, tales como la frecuencia o la concordancia, que se erigen como el aspecto central y el punto fuerte de esta rama de la lingüística. Independientemente de la visión que compartan los lingüistas de corpus, las definiciones que aportan suelen ser neutras y válidas en cualquier contexto, como las de Aijmer y Altenberg (1991: 1), cuando afirman que "corpus linguistics can be described

as the study of language on the basis of text corpora"; Abaitua (2002b: 62) que la entiende como el "área de la lingüística especializada en el aprovechamiento de los corpus" o Svartvik (1992: 7), quien la define como "the use of large collections of text available in machine-readable form".

Generalizando, por tanto, podemos considerar la Lingüística de Corpus como aquella rama de la Lingüística que trabaja con un número relativamente grande (no es fácil ser precisos en este aspecto, ni en la cantidad, ni en la conveniencia) de textos almacenados en formato electrónico en un espacio de tiempo razonable y que constituyen una base adecuada para un estudio específico del lenguaje. El uso de los ordenadores alivia en enorme medida el trabajo manual, el denominado por Svartvik "*donkey work*" (Leech, 1991: 25), tanto por el arduo trabajo que suponía introducir los textos en el corpus, como por el tiempo que requería y la posibilidad de error que tenía. Esta es, en consecuencia, otras de las características comunes a los corpus actuales: las herramientas informáticas que facilitan el trabajo y que nos ofrecen, hasta el momento, recursos como frecuencias de uso y concordancias –fundamentalmente–, es decir, información para el análisis cuantitativo y cualitativo de los textos (McEnery and Hardie, 2012).

Precisamente este es el motivo por el que Leech (1992: 106) argumenta que la Lingüística de Corpus debería llamarse realmente "*computer corpus linguistics* (CCL)". A este respecto, indica que la LC no conforma un ámbito de estudio concreto, sino que consiste en una base metodológica que sirve de soporte para la realización de la investigación lingüística. Y así, como metodología, se combina fácilmente con otras ramas de la lingüística, que van desde la fonética hasta la sociolingüística. Explica Leech que la única otra rama de la lingüística a la que le ocurre lo mismo es a la Lingüística Computacional y que el solapamiento creciente que se está produciendo entre estas dos disciplinas implica que ya no se pueda hablar de corpus sin ordenadores. Sin embargo, desde su punto de vista, CCL no es simplemente una nueva metodología para el estudio del lenguaje, sino un nuevo ámbito de investigación y un nuevo enfoque filosófico para el estudio, como hemos señalado más arriba. Sitúa al ordenador como elemento central y única arma poderosa que hace posible este nuevo tipo de lingüística y explica que aquí la tecnología, como ha pasado durante siglos en las ciencias naturales, ha adquirido un papel más importante que simplemente el de apoyar y facilitar la investigación. En sus palabras: "I see it as the essential means to a new kind

of knowledge, and as an "open sesam" to a new way of thinking about language" (Leech, 1992: 106).

A pesar de ser partidario de la LC como disciplina científica y no como metodología, John M. Sinclair (1992: 379) coincide con Leech al afirmar que la calidad de muchas disciplinas científicas se ha visto mejorada gracias a la aparición de los ordenadores, pero que pocas han experimentado un cambio tan profundo como lo ha hecho la Lingüística, donde las nuevas metodologías y la relación entre especulaciones y hechos se van a ver extremadamente alteradas.

La CCL de Leech tiene cuatro características fundamentales que, desde su punto de vista, la distinguen: a) le da más importancia a la actuación que a la competencia; b) se centra en la descripción lingüística y no en los universales lingüísticos; c) trabaja tanto con modelos cuantitativos como cualitativos del lenguaje y d) se basa en una visión empírica y no racionalista de la investigación científica.

En la misma línea, Aijmer y Altenberg (1991: 3) opinan que la creciente dependencia del material procesable por ordenador y de este último como herramienta para la investigación ha llevado a la Lingüística tradicional a una forzosa cooperación entre los lingüistas tradicionales y los informáticos, de manera que la Lingüística de Corpus se está convirtiendo en un área interdisciplinar en la que el trabajo en equipo y la unión de diferentes enfoques están prosperando.

Es importante señalar que, cuando hablamos de textos en los corpus, no siempre nos referimos a textos completos, sino a fragmentos o a archivos (orales, escritos o semiorales) de información textual susceptibles de ser procesados por ordenador. Esta es, precisamente, la característica que supuso el punto de inflexión en los corpus hace aproximadamente medio siglo. Los corpus anteriores, como el de Fries, el del Murray o el de Johnnson, como ya hemos explicado, fueron recopilados a mano. La mejora, por tanto, no solo fue a nivel de reducción de tiempo y de esfuerzo o al menor margen de error, sino en cuanto al tamaño de los corpus, que, en su mayoría, se vio aumentado notablemente gracias a la capacidad de almacenaje y procesamiento de los ordenadores.

1.4 POR QUÉ LA LINGÜÍSTICA DE CORPUS

En una conferencia en julio de 1964 en la Sociedad Lingüística del *American Summer Institute*, Chomsky ironizó con la siguiente afirmación: "*I live in New York* is more frequent than *I Live in Dayton Ohio*", puesto que el número de habitantes en

Nueva York es mucho mayor que el de Ohio. Svartvik replicó que la frecuencia lingüística no debe trivializarse de esa manera y demostró que hay ciertos patrones que no pueden reducirse a aspectos como la población de las ciudades o las preferencias de la gente a la hora de poner nombre a sus hijos (*apud* Halliday, 1991: 30).

Que no hace falta defender ni demostrar la importancia de los corpus como fuente de información y de investigación parece obvio para Halliday, afirmación que ha quedado validada por muchos otros autores, como Svartvik, pero lo que no resulta tan obvio para él es el estatus teórico de las frecuencias de uso. En cualquier caso, esto no quiere decir que, en su manera de concebir la investigación, el estudio de corpus no ocupe un lugar central en las investigaciones teóricas sobre el lenguaje. De hecho, dos son para él las principales razones que justifican la utilidad de la LC en los estudios sobre el lenguaje. En primer lugar, a raíz de sus estudios de cantonés y mandarino, resalta la importancia de la estadística para estudiar la gramática de forma cuantitativa, ya que es la mejor manera de llevar estos estudios a cabo (Halliday, 1992: 61). Por otro lado, el potencial semántico del sistema solo emerge en su totalidad cuando se trata de la lengua oral y, más concretamente, de la interacción espontánea y natural (Halliday, 1992: 62).

Chafe (1992) menciona que la observación sirve para unir los productos de la imaginación con la realidad y que los métodos que se han usado para analizar el lenguaje y la mente pueden ordenarse en dos dimensiones: una de ellas opone comportamientos públicos con la introspección. La del comportamiento, como él la denomina, aporta una información física y pública, mientras que la introspección trabaja con información disponible solamente para el observador. La otra dimensión contrasta la manipulación artificial pero intencionada de la realidad con la observación de fenómenos que ocurren de manera natural. De esta forma, enfrenta, por un lado, información conductual o de comportamiento con la observación obtenida de la introspección; y, por otro lado, opone la observación artificial con la natural. Cada una de ellas ofrece sus ventajas y sus limitaciones. Por ejemplo, la observación de comportamiento, según la nomenclatura de Chafe, está disponible para la verificación pública, y este aspecto resulta positivo si tenemos en cuenta que uno de los objetivos de la ciencia es producir conocimiento para que sea compartido y verificado por un gran número de personas. El inconveniente, desde su perspectiva, es que en los últimos tiempos la ciencia se ha interesado poco en los procesos mentales y, para él, en última instancia, es ahí donde debemos prestar todo nuestro interés. Afirma que no hay nada de

malo en estar interesado en el comportamiento, pero que no es posible alcanzar un entendimiento completo de este sin llegar a comprender la mente, ya que aquí es donde reside ese comportamiento. En resumen, la información basada en el comportamiento es verificable, pero indirecta, mientras que la información basada en la introspección es directa, pero difícil de verificar. Es decir, el aspecto positivo de los fenómenos naturales es que son más cercanos a la realidad; el negativo es el hecho de que puede darse el caso de que estos fenómenos se den muy pocas veces y haya que esperar mucho o realizar muchas observaciones para satisfacer la curiosidad. Sin embargo, este inconveniente puede subsanarse si tenemos en cuenta que la baja ocurrencia de algún fenómeno puede ser una observación en sí misma. Por tanto, esto deja de ser una desventaja para convertirse en una fuente de conocimiento.

Y define así el término *corpus* como "compilaciones más o menos cuantiosas de lenguaje que se dan de forma natural" (Chafe, 1992: 88). Les otorga las ventajas mencionadas anteriormente: están basados en comportamientos públicos y están disponibles para todo el que quiera examinarlos, de manera que satisfacen el criterio de verificabilidad. Además, explica, que mientras que la información basada en el comportamiento siempre es ajena al proceso mental, al lenguaje no le ocurre esto porque lo considera una ventana a la mente, compleja, imperceptible e imperfecta, pero la mejor de las aproximaciones que tenemos hacia ella. Al ser naturales, más que manipulados, están más cerca de la realidad y, aunque esto implique la incapacidad para identificar fenómenos hasta que no ocurran, el hecho de la no ocurrencia, como hemos mencionado anteriormente, es un fenómeno en sí mismo que necesita una explicación.

En resumen, Chafe (1992) concluye que hay dos buenas razones para usar la información disponible en los corpus en lugar de –o a la vez que– la introspección y la elicitación. Una de ellas se basa en el enfoque cuantitativo de los corpus, que se hace atractivo y accesible gracias a la disponibilidad de información procesable por el ordenador y a los *software* para la extracción y el procesamiento estadístico de la esta; la segunda se centra en el enfoque cualitativo. Independientemente del tamaño, los corpus son útiles porque la información contenida en ellos suele ser de mejor calidad que la derivada de la instrospección y la elicitación; además de que, para ciertos problemas, son la única información disponible. Se trata de ejemplos reales de uso que se enmarcan en contextos concretos, lo que hace posible el estudio sistemático de las interrelaciones entre la forma lingüística y la función comunicativa (Mair, 1992).

Podemos reunir en cuatro principios el análisis de corpus cuantitativos que realiza Chafe (Mair, 1992). Para empezar, el propósito de un corpus no es limitar la información a una supuesta muestra representativa, sino aportar un marco de trabajo para conocer aspectos sobre el lenguaje general. Es decir, el objeto de la LC no es la explicación de lo que hay presente en el corpus, sino el entendimiento del lenguaje. En segundo lugar, la LC persigue explicaciones acerca de la formas sintácticas funcionales y basadas en el discurso, más que buscarlas al considerar la naturaleza de la sintaxis como un algoritmo formal autónomo. En tercer lugar, aunque es cierto que los corpus son útiles en cualquier análisis del lenguaje, es en el estudio de situaciones espontáneas y no planificadas donde se vuelven fundamentales, así como en la conversación natural.

Por su parte, para Svartvik (1992: 8-9) existe también un buen número de razones para la utilización de los corpus, que pueden resumirse de la siguiente forma:

-En el campo de la Lingüística, el uso de corpus permite realizar afirmaciones más objetivas que la mera introspección. Los hablantes nativos pueden hablar bien, pero no saben qué o cómo lo han hecho.

-La verificabilidad es un requisito habitual en la investigación científica, por lo tanto, la Lingüística, que se define como el estudio científico del lenguaje, no debería quedar exenta de este procedimiento.

-Para los estudiantes de lenguas, los diferentes ejemplos de uso del lenguaje son fundamentales. Para los estudios diacrónicos de la lengua, es el único recurso disponible.

-Al igual que los corpus son necesarios para describir los distintos usos del lenguaje, también lo son para conocer la frecuencia de ocurrencia de los elementos lingüísticos en las diferentes variedades del lenguaje. Hay, por tanto, una correlación entre la frecuencia relativa y el registro.

-Para muchos lingüistas, los corpus no son solo simples ejemplos del lenguaje, sino una fuente teórica.

-En muchos campos de la Lingüística aplicada, como la enseñanza de lenguas, la traducción automática, el reconocimiento del habla, etc., los corpus se están convirtiendo en una fuente fundamental de información.

-Los corpus ofrecen la posibilidad de la llamada *total accountability* de las características lingüísticas. Los antiguos trocitos de papel donde se anotaba la información podían llegar a ser útiles, pero no alcanzaban la *total accountability*.

-Otro inconveniente del método tradicional es el uso individual surgido de cada recopilación de ejemplos. El acceso a los corpus estandarizados e informatizados permite compartir la información y no considerar la estadísticas como una suma de estudios basados en corpus, sino como una investigación que aporta resultados acumulativos.

-Los corpus libres aportan información lingüística al investigador que de otra manera habría sido difícil o imposible de obtener.

-Por último, los corpus se presentan como una herramienta fundamental para aquellos estudiantes de lenguas que no son nativos, puesto que, para ellos, la introspección y la intuición no sirven.

Podemos observar la similitud de opiniones y las coincidencias entre los argumentos de Svartvik y de otros autores, como Chafe (1992) o Mair (1992). Además, Svartvik deja abierta la puerta para aquellos investigadores que trabajen con la LC como rama teórica de la Lingüística e introduce su utilidad no solo para hablantes nativos y los investigadores del lenguaje, sino también para los estudiantes de lenguas extranjeras.

Tras enumerar las virtudes de los corpus y las ventajas de su uso, Svartvik también menciona algunos aspectos negativos que, a su juicio, hay que tener presentes a la hora de trabajar con ellos. Por un lado, advierte del peligro que la automatización de los corpus supone si llega a eliminar por completo el trabajo manual, pues hay ciertas áreas en Lingüística que lo necesitan. Además, recuerda que no siempre el tamaño es garantía de éxito en el corpus, es importante también tener en cuenta la idoneidad. Deja en último lugar el riesgo más importante para él: la distancia que se crea entre el usuario final del corpus y el material textual primario, bien porque la información no haya sido correctamente tratada o bien porque no se haya consultado de la forma adecuada.

Gracias a los corpus, por lo tanto, se han podido investigar cuestiones relativas a la estructura del lenguaje, a la variación y a la historia que antes era imposible responder e incluso imaginar desde el punto de vista de la introspección (Biber y Finegan, 1992).

Las opiniones de Svartvik (1992) acerca de la utilidad de los corpus no estaban muy alejadas de la realidad, de la misma manera que no se equivocó al prever el camino que tomaría la LC en los últimos años del siglo XX y los albores del XXI. Unos pronósticos que, aun hoy, siguen estando vigentes, porque, aunque se están cumpliendo, todavía tienen mucho camino que recorrer.

El primer gran cambio que se avecinaba era, naturalmente, la extensión de los corpus futuros, que serían considerablemente mayores que sus predecesores. En el momento de escribir el artículo, el tamaño estándar de un millón de palabras estaba empezando a ser reemplazado por el de cien millones, como el BNC.

Este aumento en tamaño vendría propiciado por los avances en *hardware* y en la producción (bienvenida) de *software* más fáciles de usar, y también gracias a nuevas formas de considerar las funcione de un corpus. Aquí encuentran su origen conceptos como el de *monitor corpus*, que se refiere a corpus de tamaños no finitos, pero con el lenguaje en movimiento, analizado a través de filtros en tiempo real.

Además, la dirección que estaba tomando el uso de los corpus se relacioanaba con un aumento en su actividad y utilización, enmarcadas en un clima más liberal dentro de la Lingüística y con un acceso más fácil, más eficiente y más barato. Otros avances tecnológicos también ejercerían una fuerte influencia en el futuro de la Lingüística de Corpus, como el análisis de grandes cantidades de textos y de la comunicación internacional para el intercambio de información, que, según él, se verían también facilitados por el acceso *online* a fuentes textuales que se encontraban –y todavía lo están– en constante crecimiento.

Sin embargo, es importante, advierte Svartvik (1992: 11-12), que no olvidemos que nuestro campo de estudio es el lenguaje humano y que toda esta tecnología debe ser administrada por recursos humanos, lingüistas que combinen la herramientas facilitadas por la informática con el trabajo humano.

1.5 LINGÜÍSTICA COMPUTACIONAL

1.5.1 Introducción

La afirmación de que el estudio del lenguaje, tanto desde el punto de vista empírico, como desde el intuitivo, ha sido una de las actividades más antiguas de la civilización no alberga ningún tipo de duda ni de discusión al respecto. A lo largo de la historia hemos explorado la naturaleza del lenguaje para comprender el papel que este juega en la mente y en la comunicación humanas. Con el paso de los siglos, esta rama de conocimiento ha estado unida a otras áreas con las que ha establecido lazos conceptuales y procedimentales y ha sido en los últimos tiempos cuando la aparición de la Informática ha hecho que la investigación sobre el lenguaje evolucione y explore

nuevas metodologías que han añadido otra dimensión a los estudios lingüísticos. De hecho, la mayoría de los autores citados en este capítulo, como hemos visto, no conciben la LC sin la intrínseca unión con los ordenadores. Esto ha sido posible, entre otras razones, gracias al apoyo de técnicas y herramientas que permiten almacenar ejemplos reales de usos lingüísticos y analizar la información desde nuevas perspectivas. Esta relación entre los lingüistas y los informáticos, condenados a entenderse, es lo que Henry Kučera denomia "the odd couple" (Kučera, 1991: 401).

La introducción de este nuevo enfoque ha contribuido de dos formas básicas en el campo de la Lingüística (Sekhar, 2008):

a) Ha permitido verificar viejas teorías sobre el lenguaje y la conveniencia o no de mantenerlas.

b) Ha ampliado el alcance del uso directo de evidencias e información lingüísticas en los trabajos lingüísticos y en las tecnologías del lenguaje.

Estas dos ideas, aunque quizá algo básicas y sencillas, hablan de una nueva dimensión añadida al campo de la Lingüística, gracias a la aparición y al avance de la tecnología informática, lo que ha dado como resultado el surgimiento de la llamada Lingüística Computacional. Esta Lingüística Computacional se entiende como una de las áreas de la Inteligencia Artificial, cuyo objetivo es el tratamiento del lenguaje como el instrumento fundamental de la comunicación humana, directamente ligado a la cognición.

La Lingüística de Corpus está, por tanto, íntimamente ligada a la Lingüística Computacional, pues aporta grandes cantidades de ejemplos reales del lenguaje almacenados en bases de datos en forma de corpus de manera sistemática. Por otra parte, la Lingüística Computacional también nos ofrece también herramientas sofisticadas que analizan estos corpus para extraer la información lingüística.

La fuerte motivación lingüística y cognitiva que nos ha llevado siempre a investigar la forma de comunicarnos a través del tiempo y del espacio, junto con la motivación técnica que dirige la construcción de sistemas informáticos inteligentes capaces de realizar una interacción lingüística eficiente con los humanos, se han unido para desarrollar sistemas como traducción automática, extracción de información, generación y comprensión del lenguaje, etc. Pero para el diseño y el desarrollo de estos

sistemas, necesitamos entender de forma empírica el lenguaje natural y sus características. Aquí es donde los corpus se convierten en indispensables.

Por tanto, a pesar de que los trabajos pioneros en Lingüística de Corpus de hace cincuenta años tuvieron unos inicios difíciles y encontraron muchas trabas para hacerse un hueco en la predominante gramática generativa, poco a poco, esta orientación ha ido ganando adeptos gracias a lo que se presenta como una nueva metodología y un nuevo enfoque ayudados por los ordenadores. Sin olvidar, claro está, que es al lingüista a quien corresponde el último análisis de la información aportada por las máquinas.

Debido a la corta historia de la Lingüística Computacional, no es de extrañar la variedad terminológica a la hora de nombrarla y también la diversidad de definiciones. La literatura al respecto abunda (Meya y Huber, 1986; Moreno, 1990; Gómez Guinovart, 1998; Johnson y Johnson, 1998; Crystal, 2000; Martí y Castelló, 2000; Domínguez, 2002, Pérez y Moreno, 2009, etc.). Algunos de los otros términos que le han puesto nombre a esta disciplina son: lingüística informática (Moreno, 1990), procesamiento del lenguaje natural (Meya y Huber, 1986), procesamiento de datos lingüísticos (Meya y Huber, 1986) o ingeniería lingüística (Meya y Huber, 1986; Gómez Guinovart, 1998; Pérez y Moreno, 2009).

En líneas generales, nos parece acertada la definición que aportan Pérez y Moreno (2009: 68) y que dice así:

> La Lingüística Computacional constituye un campo científico de carácter interdisciplinar, vinculado a la lingüística y a la informática, cuyo fin fundamental es la elaboración de modelos computacionales que reproduzcan distintos aspectos del lenguaje humano y que faciliten el tratamiento informatizado de las lenguas.

Gómez Guinovart (1998), además, la considera una subdisciplina de la Inteligencia Artificial que, a su vez, es una parte de la Informática. El objetivo de la Inteligencia Artificial es la construcción de ordenadores que simulen un comportamiento inteligente. Minsky (1968: 2) la define como "the science of making machines to do things that would require intelligence if done by humans" y, en una línea similar, Kurzweil (1990 *apud* Russell y Norvig, 1995: 3), como: "the art of creating machines that perform functions that require intelligence when performed by people".

Guillermo Rojo (2006) determina que la interacción entre la Informática y la Lingüística se efectúa en tres niveles distintos:

a) El primero es aquel en el que los ordenadores se utilizan como herramienta que contribuye al desarrollo del trabajo lingüístico –que, hasta la aparición de estos, se llevaba a cabo de forma manual. Es decir, la función de los ordenadores se limita a facilitar y aligerar la tarea del lingüista, ahorrándole tiempo y esfuerzo. Aquí encontramos, por ejemplo, los procesadores de textos.

b) En el siguiente nivel que determina Rojo, los ordenadores ayudan al lingüista a manejar grandes cantidades de datos y a sistematizar el trabajo. La construcción y el manejo de los grandes corpus informatizados se encuentran en este nivel.

c) Por último, el autor señala un tercer nivel, el más interesante para él, en el que los ordenadores interactúan con el ser humano, comprenden las lenguas naturales y son capaces de reproducirlas. En este último nivel es donde tiene cabida la Lingüística Computacional.

1.5.2 Recorrido histórico

La Lingüística Computacional encuentra sus orígenes a mediados del siglo XX. La mayoría de los autores coinciden en situar su nacimiento en los últimos años de la Segunda Guerra Mundial (Domínguez, 2002; Pérez y Moreno, 2009; Villayandre, 2010), cuando Estados Unidos y la Unión Soviética comenzaron a trabajar en proyectos cuyo objetivo era la elaboración de programas de traducción entre el inglés y el ruso. Los organismos oficiales, entre los que se encontraban los servicios de inteligencia y las fuerzas armadas, fueron los impulsores de estos proyectos y los principales inversores.

Muy pocos años después, los científicos Alan Turing, a quien se le considera uno de los padres de la informática –y quien fue contratado por el gobierno británico para construir máquinas capaces de descifrar mensajes clave– y Claude Shannon fueron determinantes para la evolución de la Inteligencia Artificial y de la Lingüística Computacional. El primero de ellos (que en su conocido artículo "Intelligent Machinery" (Turing, 1996) ya había comenzado a contemplar la posibilidad de construir máquinas capaces de pensar) fue quien desarrolló la *Teoría de los Autómatas*. Shannon, por su parte, también contribuyó a la construcción de autómatas aplicando una teoría de probabilidad basada en el modelo de Markov en 1948 (Shannon y Weaver, 1998), un modelo estocástico de probabilidad en el que la probabilidad de que ocurra un

evento depende del evento anterior. Estos trabajos sentaron las bases de la líneas científicas que surgirían en los años siguientes, donde son protagonistas Chomsky –centrado en el análisis sintáctico–, por un lado, y Minsky, Shannon y Weaver[18], por el otro, más volcados en el ámbito de la inteligencia artificial. Estos dos últimos, matemáticos ambos, acabarían formulando en 1949 la *Teoría de la Información y de la Comunicación*, tan influyente, todavía hoy, en el campo de la Lingüística y también en el de la Informática. (Shannon y Weaver, 1998). Según esta teoría, para que la información que parte de una fuente de comunicación llegue al receptor, este tiene que descodificarla, y esta descodificación se basa en un proceso probabilístico que determina en parte la probabilidad de la siguiente descodificación (Cerny, 2010). Como apunta Villayandre (2010), la información codificada en forma de símbolos puede ser procesada tanto por los ordenadores –en el caso de que sean símbolos numéricos–, como por la mente humana –en el caso de que no sean numéricos. Fue Shannon quien dio un paso más allá y contempló la posibilidad de que los ordenadores también fueran capaces de descifrar símbolos no numéricos en su artículo "A Chess Playing Machine" (1950), en el que sugería que la flexibilidad de las máquinas las posibilitaba para entender el lenguaje (Villayandre, 2010).

Estos años fueron bastante prolíficos, debido al interés que la Inteligencia Artificial comenzó a despertar en los ámbitos científicos y a los avances que vieron la luz durante este tiempo, como el método Bayes de reconocimiento de caracteres, para determinar la autoría de los documentos o las técnicas de reconocimiento de voz, entre otros. A partir de aquí, dada la posibilidad de que los ordenadores simularan el pensamiento humano, surgió el Procesamiento del Lenguaje Natural (PNL), denominación que tardaría tiempo en consensuarse y asentarse, pero que atrajo la atención de las investigaciones y cuya aplicación más importante fue la traducción automática.

El auge que experimentara durante estos años la Inteligencia Artificial y, más concretamente, la traducción automática, sufrió un duro revés en la década de los 60, cuando la escasez de resultados, debida a la complejidad del lenguaje –y, según Villayandre (2010), a la ausencia y la falta de adecuación de las teorías lingüísticas– se apoderó de la situación y dio lugar a un informe negativo de la *National Academy of Science* (Domínguez, 2002). El conocido informe ALPAC (*Automatic Language*

[18] Warren Weaver planteó la posibilidad de aprovechar la velocidad, la capacidad y la flexibilidad lógica que habían adquirido los ordenadores para utilizarlos en la traducción 1949 (Shannon y Weaver, 1998).

Processing Advisory Committe, 1964, *apud* Hutchins, 2003) admitía la falta de logros y daba comienzo a la drástica reducción de la financiación y al cambio de rumbo en las investigaciones en el área de la Inteligencia Artificial.

Mientras tanto, la irrupción de Noam Chomsky en este panorama con *Syntactic Structures* (1957) y *Aspects of a Theory of Language* (1965) resultó, una vez más, determinante. Su naturaleza y su tradición lingüísticas no impidieron que su gramática generativa y su teoría de los lenguajes formales fueran unas de las más influyentes en el campo de la Informática y de la Lingüística Computacional.

En los años que siguieron al informe ALPAC, la Inteligencia Artificial se centró en otras áreas del procesamiento automático del lenguaje, especialmente en la elaboración de corpus, coincidiendo con la aparición del *Brown Corpus*. Se desarrollaron programas informáticos como *ELIZA* [19], capaces de mantener conversaciones con los usuarios y que demostraron las teorías de Turing (Villayandre, 2010) y también técnicas de recuperación de información. Durante los años 70, empresas como IBM y AT&T Lab Research investigaron sobre las técnicas de tratamiento y reconocimiento del habla que siguieron consolidándose en la década siguiente, donde apareció, entre otros avances, *PROLOG*, un lenguaje de programación que utilizaba la lógica como base creado por Colmerauer en 1970 (Llorens y Castel, 1996-2001). Los esfuerzos durante esta época se centraron en la comprensión y en la simulación de los procesos que subyacen al lenguaje (Meya y Huber, 1986). En la década de los 80, la traducción automática vuelve a resurgir con fuerza, al igual que también se recupera la Lingüística de Corpus, y siguen apareciendo nuevos trabajos sobre generación del lenguaje.

El momento más decisivo, en cambio, y más influyente para la Lingüística Computacional –y nos atrevemos a decir que para todas las áreas del conocimiento– llegó en los años 90 con la aparición de la *World Wide Web*. La introducción masiva de la Informática y de Internet en el mundo trajo consigo cambios paradigmáticos en la ciencia y una enorme ampliación de sus posibilidades. A partir de este momento, la expansión del conocimiento y la velocidad a la que la investigación ha avanzado han dado impulso a logros científicos sin precedentes en todos los ámbitos del saber, y la Lingüística Computacional, naturalmente, también se ha beneficiado de ellos. Las

[19] *ELIZA* fue desarrollado por Joseph Weizenbaum en el MIT, entre 1964 y 1966. En el programa, *ELIZA* emula ser un psicoterapeuta y los usuarios interactúan con él. En el siguiente enlace, se puede dialogar con *ELIZA*: http://www-ai.ijs.si/eliza/eliza.html

traducciones automáticas, los sistemas de diálogo, la recuperación de información, y el almacenamiento y análisis de cantidades masivas de datos han experimentado transformaciones que han tomado nuevas direcciones para las que ya no hay marcha atrás. Incluso han aparecido materias nuevas, fruto de la combinación de ámbitos más tradicionales con las nuevas tecnologías, como la *terminótica* (Cabré, 1993: 359), que se encarga "en general, de las relaciones entre la informática y la terminología; y, en particular, que trata de la aplicación de la informática al trabajo terminológico".

Desde el punto de vista de esta autora, la relación entre la Informática y la Lingüística ha traído consigo nuevas aplicaciones "que se pueden clasificar según el grado de complejidad creciente del tratamiento informático que requieren sus objetivos" (Cabré, 1993: 356). Así, distingue cuatro niveles, que abarcan desde las aplicaciones que no manipulan ni analizan los datos (como los sistemas de tratamiento de textos) hasta los sistemas inteligentes que pretenden realizar tareas propias del ser humano (como pueden ser los sistemas de vaciado automático de términos, los sistemas de traducción automática o los generadores de texto). En la zona intermedia entre estos dos extremos sitúa Cabré las herramientas lingüísticas automatizadas, como los diccionarios automatizados, y en un nivel informático mayor, los sistemas automáticos que manipulan los datos, entre los que se encuentran los analizadores, lematizadores, programas de tratamiento estadístico, etc.

1.5.3 Áreas de trabajo

Cualquier aspecto del lenguaje humano susceptible de ser trabajado con los ordenadores puede despertar el interés de la Lingüística Computacional. La creación de programas informáticos que emulen de la forma más fiel posible el comportamiento humano es la última meta de esta disciplina.

Domínguez (2002) enumera en seis las áreas de trabajo de la Lingüística Computacional:

1. *Tagging* o etiquetamiento morfológico: se trata del etiquetamiento de las palabras de forma aislada que nos da información acerca de su morfología.
2. *Parsing* o análisis sintáctico: consiste en el proceso de análisis de oraciones según su sintaxis. Dentro del campo de la Inteligencia Artificial, se incluyen procedimientos de interpretación semántica. Los algoritmos necesarios para esta tarea siguen dos tipos de procedimientos (Meya y Huber 1986,

Grishman, 1991, Domínguez, 2002; Villayandre, 2010): *bottom-up*, si se parte de símbolos para llegar a estructuras más complejas; y *top-down*, que parte oraciones dadas para hacer hipótesis sobre cómo están constituidas. Domínguez hace también alusión al llamado análisis superficial o *shallow parsing*, que analiza algunos componentes de la oración sin llegar a ser exhaustivo.

3. Técnicas de reconocimiento de voz y conversión de texto a voz: las técnicas de reconocimiento de voz se hacen a través de sistemas automáticos (*automatic speech recognition*) que transcriben la voz humana en datos procesables por un ordenador. Mediante un sistema de probabilidades y basándose en la teoría de la comunicación de Shannon, identifican las oraciones con más probabilidades de ser la que parte del transmisor de la señal acústica para decodificarla. El objetivo de la conversión de texto a voz, por otro lado, es generar automáticamente los sonidos que un ser humano produciría al leer un texto.
4. Recuperación inteligente de información o *information retrieval*: este campo incluye todos los sistemas automáticos de obtención y análisis de información para su utilización posterior por los usuarios. Estos sistemas son los que se utilizan en los corpus actuales y en los buscadores de Internet.
5. Sistemas de diálogo y sistemas expertos: consisten en la transmisión de información entre los usuarios y el ordenador, gracias al almacenamiento digital previo del conocimiento de expertos en un área determinada.
6. Traducción automática: no cabe duda de que estos sistemas no han alcanzado la perfección y sus limitaciones provocan que no sean capaces de sustituir a los profesionales, pero han experimentado grandes avances en tres líneas distintas: la traducción palabra por palabra, la traducción por transferencia y la traducción por medio de una interlingua (generalmente, a través del inglés o del esperanto).

Prácticamente las mismas áreas de trabajo son las que menciona Gómez Guinovart (1998), aunque con una clasificación un poco más particular. Él divide en tres los campos fundamentales en los que la Lingüística Computacional encuentra su

aplicación y los ordena partiendo del más ligado a la Lingüística para finalizar con el más relacionado con la Informática. Así, establece las siguiente líneas:

1. La informática aplicada a la investigación lingüística: aquí incluye los etiquetados morfológicos y sintácticos.
2. La implementación de teorías lingüísticas: según el autor, esta línea posee tres objetivos: a) la elaboración de teorías o modelos lingüísticos, b) descripción de fenómenos lingüísticos concretos enmarcados en estas teorías y c) comprobación de una forma automatizada de la consistencia de una teoría lingüística o de sus predicciones. En este grupo incluye los sistemas de planificación lingüística o formalismos lingüísticos, diseñados para representar conocimientos lingüísticos, entendidos por los ordenadores y que sirven para la comprobación de las teorías. Para ellos, entre otras, se usa el programa *PROLOG*, que hemos mencionado anteriormente.
3. Las aplicaciones lingüísticas de la informática: esta línea está centrada en el PNL y la comprensión y creación de lenguajes naturales. Aquí tienen cabida por tanto, las tecnologías del habla (reconocimiento del habla y síntesis), la traducción automática (Gómez distingue entre traducción totalmente automática y traducción asistida por ordenador) y la extracción de información.

Villayandre (2010), por último, en su trabajo de tesis sobre Lingüística Computacional, clasifica en tres grandes grupos las aplicaciones que considera más importantes:

1. Traducción automática.
2. Interacción en lenguaje natural, donde incluye las interfaces y los sistemas de diálogo.
3. Recuperación y extracción de información.

Aparte de estas tres, también nombra las herramientas de ayuda a la escritura (como los correctores ortográficos o sintácticos), la creación automática de resúmenes, la extracción de terminología, la indexación automática, la síntesis y el reconocimiento del habla y *data mining* (o minería de datos). En la figura que aparece a continuación, podemos ver la clasificación que la autora realiza para todas ellas, distinguiendo entre aquellas aplicaciones que están basadas en el texto y las que están basadas en el habla:

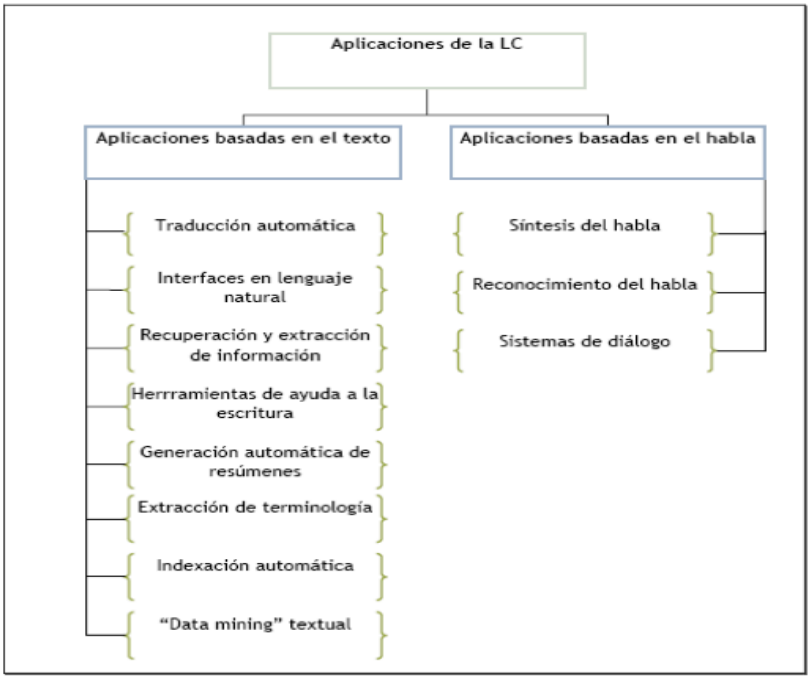


Figura 1.1. Clasificación de las aplicaciones de la Lingüística Computacional según Villayandre
Fuente: Villayandre (2010)

1.5.4 Herramientas para el análisis de corpus

Las herramientas actuales con las que trabaja la Lingüística de Corpus se mueven en los ámbitos de la anotación textual y el etiquetado. La Lingüística Computacional ha desarrollado programas de fácil manejo para los investigadores, diseñados para extraer información de los corpus de una forma mucho más rápida y segura que la manual.

En este sentido, existen numerosas herramientas de *software* que permiten descifrar los corpus en términos de información morfológica (*tagging*), sintáctica (*parsing*) y semántica. Sin embargo, el elevado número de programas disponibles no nos garantiza variedad en los sistemas de análisis, pues las funciones que cumplen unos y otros son casi idénticas.

Entre estas funciones, se encuentran el conocido *part of speech tagging*, o POS, que realiza funciones de etiquetado morfológico, con distintos tipos de estrategias. Por ejemplo, TAGGIT fue el primer etiquetador automático que se aplicó a corpus de gran tamaño, entre ellos, el *Brown Corpus*. El programa se basaba en reglas adquiridas de forma semiautomática a partir de un diccionario de tres mil entradas más una lista de sufijos (Dipper, 2008).

En los años 80, la Universidad de Lancaster desarrolló otro tipo de anotación, basado en TAGGIT, conocido como CLAWS (*the Constituent Likelihood Automatic Word-tagging System*) (Garside, 1987). Este programa heredó de TAGGIT aspectos como las entradas del diccionario, los sufijos o las reglas entre palabras. Sin embargo, introdujo un programa de frecuencias estadísticas como novedad que permitió más precisión a la hora de desambiguar aquellas palabras que lo necesitaran (aunque es importante señalar que la desambiguación, todavía hoy, no está del todo conseguida). Corpus como el BNC han sido anotados con este sistema (Leech, Garside y Bryant, 1994).

También existe la posibilidad de lematizar las palabras del corpus, esto es, asignarles su lema para que todas aquellas palabras que compartan el mismo lema puedan ser incluidas en una sola búsqueda.

Pero la cuestión de la anotación tampoco se escapa de la polémica en cuanto a su utilidad a la hora de analizar la información textual contenida en los corpus. El argumento más utilizado en su contra es que la anotación "contamina" la información original y dificulta la visualización de los patrones lingüísticos. Esta visión es apoyada fundamentalmente por los investigadores que defienden el enfoque de *corpus-driven linguistics*, puesto que ven el corpus como el punto de partida para el análisis. Puesto que su objetivo es observar los patrones lingüísticos contenidos en el corpus, la anotación no aporta ninguna ventaja a la investigación y puede generar "ruido". Así lo expresa Sinclair, para quien la anotación pudo haber sido útil hace cincuenta años, cuando los primeros sistemas operativos y *software* no eran capaces de procesar texto:

> The interspersing of tags in a language texts is a perilous activity, because the text thereby loses its integrity, and no matter how careful one is the original text cannot be reliably retrieved... In corpus-driven linguistics you do not use pre-tagged text, but you process the raw text directly and then the patterns of this uncontaminated text are able to be observed. (Sinclair, 2004: 191)

Desde su punto de vista, mientras sigamos confiando en las etiquetas, estaremos centrando nuestra atención en los modelos antiguos de investigación que se basaban en corpus pequeños. Sin embargo, los corpus anotados no se ajustan a las necesidades de la sociedad de la información porque no son lo suficientemente "sensibles". Se ha

demostrado que no son adecuados para lo que él denomina "textos abiertos" (*open texts*)[20], que constituyen una parte esencial para entender el lenguaje humano.

Por otro lado, los defensores de la anotación la consideran un paso necesario para probar una teoría lingüística concreta (Anthony, 2013).

En cualquier caso, como hemos señalado unas líneas más arriba, existen diversas herramientas utilizadas para estos propósitos, muy parecidas entre sí y que resultan útiles para ciertas tareas del lingüista.

Para elaborar concordancias y listas de frecuencias con el sistema de KWIC (*Key Word In Context*), algunas de las más conocidas son: *WordSmith Tools* (Scott, 2012), *MonoConc Pro* (*Barlow, 2000*), *AntConc* (Anthony, 2012) o *The Sketch Engine*, desarrollada por Kilgarriff y su equipo en 2004 (Kilgarriff *et al*. 2004)

Por otro lado, en cuanto a los programas más utilizados para la anotación morfológica y sintáctica, también existe una amplia oferta de *software* que ofrecen a los lingüísticas la posibilidad de llevar a cabo estas tareas de forma semiautomática. Por nombrar algunos de ellos, encontramos *SALTO*, *UAM CorpusTool* o *WordStat*, para la anotación morfológica; o *CLaRK* o *NooJ*, para la sintáctica. No obstante, casi todos llevan a cabo diversas funciones de anotación y etiquetado, aunque la mayoría de ellos solo trabajan con el inglés.

[20] "By 'open text' I mean unrestricted text, any text that you find that is a reasonable sample of a particular language in use; indeed in the way in which language research has developed there is clear evidence that the study of open text is frequently avoided" (Sinclair, 2004: 186).

2

La web como corpus (*The web as corpus)*

2.1 CONCEPTOS PREVIOS

La *World Wide Web* se ha convertido en el principal punto mundial de encuentro de información, comunicación, cultura y comercio. Su inmensidad, gratuidad y fácil accesibilidad hacen de ella un recurso de un valor incalculable para la investigación y la expansión del conocimiento. Usamos la web prácticamente para cualquier propósito, ya sea reservar un viaje, realizar una operación bancaria o consultar un tutorial acerca de cualquier ámbito. Naturalmente, detrás de estas posibilidades que nos ofrece Internet, subyacen otras formas de explotar su potencial relacionadas con las investigaciones científicas. En el ámbito de la Lingüística, en particular, la ventaja es obvia: la gigantesca cantidad de material textual que hace posible, por primera vez en la historia, estudiar innumerables ejemplos reales de utilización de las lenguas, producidos por distintos individuos en situaciones totalmente diferentes unas de otras (Baroni y Bernardini, 2006), y que también posibilita el acceso a fuentes de información secundarias, a material bibliográfico, etc.

Así pues, la Lingüística, como el resto de la ciencias, también se beneficia de la web y de los miles de millones de palabras y textos que contiene. Pero no solo los lingüistas tradicionales pueden sacar partido de las ventajas y novedades que nos ofrece; dada su naturaleza de textos procesables por ordenador, el tamaño, la accesibilidad y su constante actualización, la Lingüística de Corpus y las áreas relacionadas con ella –fundamentalmente la Lingüística Computacional– han visto ampliadas sus posibilidades de manera exponencial. Dentro de la Lingüística Computacional, el Procesamiento del Lenguaje Natural (PNL), la recuperación de la información, la minería de textos y las tecnologías del lenguaje en general son los campos que más están avanzando en esta línea. Los usos, por lo tanto, que se le pueden dar a la web en el área de los estudios

lingüísticos son muy variados, y van desde algo tan simple como comprobar la ortografía de una palabra o su frecuencia de apariciones hasta la construcción de corpus.

Teniendo en cuenta que la finalidad última de la Lingüística de Corpus es la observación directa del lenguaje en contextos naturales y auténticos por hablantes cuyo objetivo sea el establecimiento de la comunicación y no demostrar la competencia lingüística, la *World Wide Web* se presenta como un nuevo horizonte lleno de posibilidades para la investigación y la fuente más rica y accesible de material disponible.

Por otra parte, la vertiginosa velocidad a la que se producen los cambios y a la que aumenta la cantidad de información disponible plantea también la necesidad de la creación de herramientas y sistemas de trabajo que se ajusten a la nueva realidad.

Actualmente, parece haber poca discusión acerca de la idoneidad de la expresión *The web as corpus* –la web como corpus– y de su consideración como tal. La expresión inglesa fue introducida por primera vez en 2001 por Adam Kilgarriff y, dos años más tarde, desarrollada en el conocido e influyente artículo que lleva por título esta misma frase y en el que Kilgarriff y Grefenstette (2003) dan argumentos a favor del estatus de corpus de la web. A partir de una comparación con la definición de corpus de McEnery y Wilson (2003) y de una reflexión filosófica en torno a las preguntas: *¿qué se considera un corpus para una tarea determinada?* y *¿qué es un corpus?*, los creadores de *The web as corpus* concluyen con un rotundo sí a la cuestión de si la web puede considerarse un corpus.

Son muchas y obvias las ventajas del *web corpus*, aunque no podemos olvidar que también presenta algunas limitaciones. Fletcher (2012) enumera el tamaño, el amplio espectro que cubre, la constante actualización y la multimodalidad (audio, vídeo y texto) como sus principales puntos fuertes. Dentro de sus limitaciones, la autoría de las páginas web, la intención con la que se publican, el público al que se dirigen, la atención o el cuidado con los que se han escrito o la representatividad y precisión que presentan son los principales retos a los que nos enfrentamos.

Aun así, las investigaciones siguen avanzando en este campo, donde se están desarrollando estudios lingüísticos de todo tipo en los que se relacionan de forma casi indisoluble la Lingüística de Corpus y la Lingüística Computacional. En esta línea, y con la consideración de la web como corpus, aparecen los trabajos presentado en las conferencias anuales de la Asociación de Lingüística Computacional que comenzaran en 1999 (Kilgarriff y Grefenstatte, 2003). Entre muchos otros, destaca, por ejemplo, el

trabajo de Resnik (1999), quien elaboró corpus paralelos de inglés y francés extraídos de la web con técnicas como la identificación automática del lenguaje, entre otras. Mihalcea y Moldovan (1999) desarrollaron un método de desambiguación de los sustantivos, verbos, adjetivos y adverbios de un texto a partir de las estadísticas obtenidas de la web. Jones y Ghani (2000) demostraron que se podían realizar búsquedas automáticas basadas en la probabilidad de aparición de una palabra en un texto de una lengua minoritaria que produjeran un corpus con más ejemplos de esas mismas palabras. También Fujii e Ishikawa (2000) utilizaron los recursos disponibles en la web para extraer definiciones de textos técnicos a modo de enciclopedia. Un par de años más tarde, Keller, Lapata y Ourioupina (2002) demostraron también que la web puede utilizarse para obtener frecuencias de pares de palabras (adjetivo-sustantivo, sustantivo-sustantivo, etc.) que pasan inadvertidos en un corpus determinado; este trabajo fue ampliado tres años más tarde por Lapata y Keller (2005).

Kilgarriff y Grefenstette (2003) mencionan a otros autores que utilizan la web para la desambiguación de significados, como Rigau *et al*. (2002) para *The Meaning Project*, para obtener estadísticas léxicas para frases preposicionales (Volk, 2001), para la creación de web corpora *ad hoc* (Fletcher, 2004) o para la creación de modelos estadísticos del lenguaje con el objetivo de crear corpus equilibrados (Villaseñor Pineda *et al*., 2003). Además, se han desarrollado sistemas de pregunta-respuesta utilizando como fuente la redundancia presente en los grandes corpus de la web (Greenwood *et al*., 2002; Dumais *et al*., 2002) en la Universidad de Sheffield y en Microsoft, respectivamente, y también basados en la recuperación de información de la web (*AnswerBus*) para realizar esas preguntas y respuestas en cinco idiomas –inglés, francés, español, italiano, alemán y portugués– (Zheng, 2002). Agirre *et al*. (2000), Varantola (2002) y Fletcher (2004d), por otra parte, han utilizado las aplicaciones que brinda la web para otras áreas de la lingüística, como la relación entre conceptos y temas, la traducción o la enseñanza de idiomas.

Otros proyectos, como el *WaCky Project*[21] (*Web as Corpus kool ynitiative*) dan muestra de la creciente tendencia en Lingüística de utilizar la web para la investigación. El primer trabajo que se llevó a cabo en el marco de este proyecto fue emprendido por Baroni, Bernardini, Ferraresi y Zanchetta (2009) y en él se elaboraron tres grandes corpus de inglés, alemán e italiano con los que demuestran la utilidad y la necesidad

[21] http://wacky.sslmit.unibo.it/doku.php?id=start

urgente de una interfaz libre basada en la web que permita un acceso sencillo para aquellos lingüistas que no estén muy versados en la informática y que les ayude a realizar una investigación extensa con corpus desde un punto de vista cualitativo y cuantitativo.

Por otro lado, la Universidad de Oslo (Guevara, 2010) desarrolló un web corpus del noruego gracias a un *web-crawler* que analizaba los documentos obtenidos de Internet a través de los buscadores comerciales de Google y Yahoo! Precisamente en el desarrollo de esta herramienta de *web-crawling* para la compilación automática de corpus especializados de la web se basan los trabajos de de Groc (2011), de Suchomel y Pomikálek (2012) o de Schäfer *et al*. (2014). Naturalmente, quedan atrás numerosísimos estudios relacionados con la web como corpus que tienen su punto de encuentro en las conferencias de *Web as Corpus* auspiciadas anualmente por la Asociación de Lingüística Computacional. En ellas, siguen participando pioneros de esta línea de investigación, como Kilgarriff o Fletcher, así como investigadores de distintas universidades europeas que se centran en los problemas específicos relacionados con la recopilación de información y su normalización, y en los procesos de construcción de web corpus específicos.

Antoinette Renouf (2007: 28) considera la irrupción de los web corpus como la última etapa dentro de la evolución del estudio de corpus:

> -A partir de la década de los 60 del siglo pasado: existencia de pequeños corpus de un millón de palabras (o menos). Estos corpus son estándares, generales o especializados, multimodales y multidimensionales.
>
> -A partir de los 80: grandes corpus de varios millones de palabras. Tienen las mismas características que los de la generación anterior, a excepción del tamaño.
>
> -A partir de los 90: "Modern Diachronic" corpus, dinámicos y abiertos.
>
> -A partir del año 1998: la *Web as Corpus*. Textos procedentes de la web como fuente de información lingüística.
>
> -A partir del año 2005: computación distribuida (*The Grid*) y consolidación de las tipologías de corpus existentes.

No obstante, como ocurre en cualquier ámbito del saber, también hay voces críticas, recelosas del rumbo que está empezando a tomar la Lingüística de Corpus, que reclaman la validez de los corpus tradicionales y su mayor adecuación para el estudio del lenguaje. No podemos olvidar, por ejemplo, la rotundidad con la que Sinclair, uno de los principales lingüistas especializados en el trabajo de corpus, expresa esta idea:

> The World Wide Web is not a corpus, because its dimensions are unknown and constantly changing, and because it has not been designed from a linguistic perspective. At present it is quite mysterious, because the search engines, through which the retrieval programs operate, are all different, none of them are comprehensive, and it is not at all clear what population is being sampled. Nevertheless, the WWW is a remarkable new resource for any worker in language (…) and we will come to understand how to make best use of it. (Sinclair, 2005: 15)

No hay duda de que las oposiciones de Sinclair aquí mostradas y los argumentos en contra de la web como corpus son hoy en día absolutamente rebatibles, como iremos demostrando a lo largo del presente trabajo. No obstante, y a pesar de su contundencia, en este artículo Sinclair deja una puerta abierta al futuro de los web corpus como una valiosa fuente para cualquier investigador que será mejorada en el futuro. Diez años después de esta afirmación, ha quedado demostrado que no se equivocaba.

En respuesta a estas ideas ancladas en el pasado, Hundt, Nesselhauf y Biewer (2007) hacen un ejercicio de empatía con aquellos estudiosos en los que esta nueva visión de corpus pueda generar desconfianza:

> The standard size of modern corpora is no longer 1 million but rather 100 million words. Why, then, should anyone want to use any material other than carefully compiled corpora? Why take the risk of using databases that are unlikely to meet the requirement of representativeness? (Hundt *et al*., 2007: 1)

Y admiten que, aunque es comprensible el temor y las reservas que algunos puedan tener, existe una serie de motivos por los que la idea de utilizar la web como corpus resulta enormemente ventajosa.

Para empezar, argumentan que en algunas áreas de la Lingüística, entre las que se encuentra el trabajo lexicográfico, siguen resultando insuficientes los grandes corpus de un millón de palabras, como el BNC (*British National Corpus*). Esto se debe a que el estudio de las innovaciones léxicas, de la morfología o incluso de algunos aspectos básicos de la gramática, necesitan más material que el disponible en estos corpus. En este sentido, Guillermo Rojo (2008: 15) afirma que:

> La utilización de la web como un gran corpus es una posibilidad que está recibiendo notable atención en los últimos años y tiene partidarios decididos. Resulta discutible, por tanto, la conveniencia de construir corpus generales, que nunca van a alcanzar el tamaño que tiene la red.

Por otra parte, continúan (Hund *et al*., 2007) una inmensa mayoría de variedades [del inglés] no encuentran representación en los grandes corpus, ni siquiera en el ICE (*International Corpus of English*), que se centra fundamentalmente en el inglés británico y en el americano.

En tercer lugar, destaca el hecho de que los nuevos avances tecnológicos, como el correo electrónico, los blogs, los foros, etc. han dado lugar a nuevas tipologías textuales que conforman un objeto de estudio en sí mismas y a las que los creadores de los megacorpus no se enfrentaban. Se trata, también, de textos a medio camino entre la escritura y la oralidad, más cercanos a los patrones de esta última, pero con formato escrito. Tipologías, estas, cada vez más asimiladas y estudiadas, como lo demuestran autores como Gómez Torrego (2001), Yus (2001), López Quero (2003), Araujo y Melo (2003), Galán (2007) o Calero Vaquera (2014). Además, Hund *et al*. (2007) destacan la nueva e "interesante dimensión" (Hund *et al*., 2007: 2) que las tipologías presentes en la web añaden al estudio de los fenómenos sociopragmáticos y que se producen con la utilización de lengua privada dentro del dominio público.

Los argumentos de tipo económico también tienen peso en su defensa de la web como corpus si tenemos en cuenta la enorme inversión necesaria para la creación de un corpus de referencia general y el alto número de probabilidades de que haya quedado obsoleto para cuando esté terminado. También aquí, Rojo apoya la cuestión económica preguntándose: "Dada la ingente cantidad de textos existentes en la parte pública de Internet, ¿tiene sentido mantener las más que considerables inversiones necesarias para mantener los corpus existentes, ampliarlos y, en su caso, crear otros nuevos?" (Rojo, 2008: 22). Renouf (2003) incide también en este aspecto, a la vez que nos presenta *WebCorp* como herramienta para solventar la naturaleza no lingüística de los buscadores comerciales, como Google.

Por último, nos recuerdan las autoras que la lengua utilizada en la web es en sí misma una de las mayores fuentes de influencia del cambio lingüístico. Afirman que

para evaluar correctamente el impacto que las conversaciones *online* ("*weblish*" o "*netspeak*"[22]) tienen en la lengua es fundamental conocer el fenómeno en sí mismo.

De todos estos aspectos, el tamaño es el problema que más afecta a los corpus tradicionales y una de sus principales desventajas con respecto a los corpus obtenidos de la web. Muchos de los corpus equilibrados de referencia, elaborados cuidadosa y minuciosamente para el estudio lingüístico, pueden ofrecer información limitada o incluso no ofrecer evidencias de ciertos aspectos. Rojo (2008) enumera este como el principal punto fuerte de la web, junto con la constante actualización y creación de nuevos contenidos, que hacen que los resultados obtenidos puedan llegar a ser más interesantes que los que se puedan obtener de cualquier otro corpus. Por otra parte, este autor enumera tres posibles problemas derivados de la web como corpus: a) la dependencia de buscadores comerciales concebidos con fines distintos a las consultas lingüísticas, b) la gran cantidad de textos que no pueden ser descargados, a pesar de que los materiales de Internet son más numerosos y ricos que los de un corpus tradicional y c) el carácter dinámico de Internet, que constituye al mismo tiempo una ventaja y un inconveniente, porque esta inestabilidad hace que los resultados estén cambiando constantemente.

A nuestro modo de ver, sin embargo, ninguna de las tres desventajas que menciona Rojo lo es. Como respuesta a la primera de ellas, aunque es absolutamente cierto que los buscadores de Internet no nacieron como herramienta lingüística y, por lo tanto, no están pensados inicialmente para realizar análisis y estudios sobre la lengua, este problema ha sido superado con la creación de herramientas específicas de análisis que trabajan sobre los buscadores para extraer la información que sea del interés del lingüista, como *WebCorp* (Kehoe y Renouf, 2002) o *KwicFinder* (Fletcher, 2001), entre otras; sin olvidar, por supuesto, la herramienta que en este trabajo presentamos.

En segundo lugar, la privacidad de algunos de sus contenidos no es un aspecto característico ni exclusivo de los web corpus, puesto que los corpus tradicionales también encuentran limitaciones con muchas de sus fuentes. La ventaja que en este sentido sí que ofrecen los corpus construidos a partir de la web es que la no aparición de textos o de fragmentos de textos privados se puede ver compensada con el millonario

[22] David Crystal (2006: 17) desarrolla estos conceptos y añade otras variantes, entre las que se encuentran: "*netlish*", "*Internet language*", "*cyberspeak*", "*electronic discourse*", "*electronic language*", "*interactive written discourse*" o "*computer-mediated communication*" (CMC). Cada una de ellas, explica, tiene diferentes implicaciones. Por ejemplo, *netlish* deriva directamente del inglés y está perdiendo actualidad porque en la web está dejando de ser poco a poco el idioma más utilizado en la red conforme avanza el plurilingüismo.

número de ejemplos que obtenemos de la parte libre, mientras que el número de evidencias que podemos extraer de un corpus tradicional es muy inferior.

Por último, y en respuesta a la tercera desventaja que Guillermo Rojo atribuye a los web corpus, consideramos que el hecho de que los resultados sean susceptibles de rápidos y continuos cambios, aunque es cierto que puede desembocar en la imposibilidad de reproducirlos, no es sino un reflejo de la naturaleza dinámica del lenguaje. Si la realidad cambia, los resultados de los trabajos deben reflejar esta evolución, lo que constituirá no un error en las conclusiones de trabajos anteriores, sino nuevos resultados. La lengua se encuentra, desde el principio de los tiempos, en continua evolución, sean cuales sean los tipos de corpus con los que nos aventuremos a analizarla. Negar esta evidencia supondría cerrar los ojos ante la realidad y un retraso en las investigaciones lingüísticas. No obstante, nos parece importante admitir el problema de la irreplicabilidad de los resultados, aspecto necesario a la hora de verificar o de refutar los resultados de cualquier investigación (Lüdeling, Evert y Baroni, 2007). Sin embargo, las nuevas herramientas diseñadas específicamente para trabajar con información lingüística son capaces de solventar también este problema porque tienen la capacidad de almacenar la información en la base de datos y de poder recuperarla en cualquier momento. Las posibilidades de poder retomar datos para un nuevo estudio dependen, no obstante, de la forma concreta en la que se materialice la utilización de la web como corpus, que, como veremos a continuación, es un concepto genérico que engloba distintos enfoques.

También Baroni y Ueyama (2006) realizan una férrea defensa de la conveniencia de utilizar la web como corpus y de las ventajas que esta presenta. Las resumen en tres: a) las derivadas del gran tamaño, b) las relacionadas con la posibilidad de construir corpus de forma rápida y económica en lenguas para las que no existen corpus de referencia y c) las que tienen que ver con la gran cantidad de géneros presentes en la web y que no podemos encontrar en las fuentes tradicionales escritas, como los blogs y toda la comunicación interactiva. A pesar de ello, también podemos encontrar algunos problemas en la utilización de la web que tienen que ver con el "ruido" que se genera debido al material no lingüístico, con la posible falta de control del investigador acerca del contenido del corpus provocada por la utilización de métodos automáticos de minería de textos y, por último, con los asuntos relacionados con el *copyright* de algunos de los documentos que se utilicen para la construcción del corpus. Inconvenientes, estos, que para los autores no son lo suficientemente relevantes

como para desechar la posibilidad que nos brinda la web en la investigación lingüística, ya que muchos de ellos no son problemas exclusivos de este tipo de corpus.

Exactamente lo mismo le ocurre al pionero en la consideración de la web como corpus, Adam Kilgarriff. En su artículo titulado *The Web as corpus*, no deja de reconocer los puntos débiles de este nuevo concepto con el que, asume, tenemos que trabajar todos porque "está con nosotros" (Kilgarriff, 2001: 1). En una comparación entre la web y el BNC, denomina a este último como "an English country garden", y admite que "whatever perversities the BNC has, the web has in spades":

> First, not all documents contain text, and many of those that do are not only text. Second, it changes all the time. Third, like Borges's Library of Babel, it contains duplicates, near duplicates, documents pointing to duplicates that may not be there, and documents that claim to be duplicates but are not. Next, the language has to be identified (and documents may contain mixes of language). Then comes the question of text type: to gain any perspective on the language we have at our disposal in the web, we must classify some of the millions of web pages, and we shall never do so manually, so corpus linguists, and also web search engines, need ways of telling what sort of text a document contains: chat or hate-mail; learned article or bus timetable. (Kilgarriff, 2001: 1)

Sin embargo, aunque a primera vista pueda parecer que la intención de Kilgarriff es disuadirnos en la novedosa tarea que constituye hacer de la web un corpus, no se trata sino de argumentos en los que subyacen los retos a los que debemos enfrentarnos y que impulsan esta nueva práctica:

> These may sound like arguments for *not* studying the web: for scientific progress, we need to fix certain parameters so we can isolate the features we want to look at, and the web is not a good environment for that. This is true. For the web to be useful for language study, we must address its anarchy. If the web is a torrent and nothing more, it is not useful; for it to be useful, we must channel off manageable quantities to irrigate the pastures of scientific and technological progress.

2.2 LA WEB COMO CORPUS VS. LA WEB PARA LOS CORPUS (THE WEB AS/FOR CORPUS)

Los distintos usos que se le han dado y se le siguen dando a la web como corpus, así como las diferentes perspectivas desde las que se pueda explotar su potencial para la investigación dentro de la Lingüística de Corpus, han resultado en la distinción entre dos expresiones diferenciadas: *web as corpus* y *web for corpus* (De Schryver, 2002 y Fletcher, 2004, 2007 y 2012).

La primera de ellas, *the web as corpus* –la web como corpus–, considera la web como una fuente de información lingüística que puede ser utilizada directamente como un corpus en sí mismo. Por el contrario, *the web for corpus* o, lo que es lo mismo, la web para la construcción de corpus, puede ser utilizada como fuente de información adecuada y útil para la elaboración de corpus *offline*.

En un análisis de estas dos perspectivas, Hundt *et al*. (2007) consideran que los principales problemas derivados del primer enfoque se resumen en que no sabemos con exactitud el tamaño del corpus, la tipología de textos que contiene o la calidad del material que aporta. Además, los resultados no se pueden repetir debido a la naturaleza efímera de la web y algunas páginas se muestran invisibles a los buscadores (inconvenientes, estos, similares a los que argumentaba Rojo y explicados unas líneas más arriba). Por otra parte, Hundt *et al*. (2007) reconocen la utilidad de la utilización de la web como corpus para el estudio de algunos fenómenos, como la creación de neologismos, para lo que admiten que la web es una de las mejores fuentes de información existentes; también para encontrar información anecdótica del tipo de si un determinado adjetivo era en el pasado utilizado por los hablantes nativos de inglés o no, información imposible de encontrar en los corpus tradicionales, incluso en los más grandes.

El uso de la web para construir corpus, por el contrario, permite adquirir archivos, textos y fuentes de información que los lingüistas están comenzando a utilizar para la compilación de corpus. Las autoras expresan su convicción de que, en el futuro, esta será la única forma de obtener cantidades razonables de información, refiriéndose a algunas variedades del inglés. Son tres las ventajas que le atribuyen a este enfoque de web corpus:

1. Control de las páginas o del material que utilizamos como fuente de información.

2. Accesibilidad gracias a las herramientas de análisis de corpus, que nos permiten realizar consultas que no se pueden utilizar con tanta facilidad en la información sin procesar de la web.
3. Mayor nivel de análisis, ya que se pueden utilizar estas herramientas.

Sin embargo, las mayores limitaciones de este uso de la web se siguen centrando en la falta de herramientas eficientes para la extracción de la información y creación de los corpus.

Ahondando un poco en este tema, Baroni y Bernardini (2006) identifican cuatro sentidos distintos que se le pueden atribuir a la expresión *web as corpus*, teniendo en cuenta que no existe un punto de vista unitario y consensuado acerca de la mejor forma de explotar la web para la investigación lingüística:

1. La web como sustituta del corpus: los investigadores utilizan la web para resolver cuestiones que podrían solucionar con un corpus, pero de manera mucho más rápida, accesible e inmediata. Esto ocurre por varias razones, a saber: los corpus que hay disponibles son demasiado pequeños o no existen para tal propósito, los interesados no tienen acceso a un corpus o incluso puede darse la posibilidad de que ni siquiera sepan qué es un corpus. Los buscadores comerciales se utilizan con propósitos lingüísticos y "oportunistas" (Baroni y Bernardini, 2006: 10), como puede ser una traducción determinada en un momento dado, o la comprobación de la ortografía o del uso de una palabra confiando en los resultados de Google, como hicieron Chklovsky y Pantel (2004) en su estudio sobre los verbos. Otros autores, algunos de ellos ya mencionados, también utilizaron los resultados del buscador para sus propósitos lingüísticos, como Grefenstette (1999) para identificar de posibles traducciones, Turney (2001) para el estudio de los sinónimos, Keller y Lapata (2003) para obtener frecuencias de pares de palabras o Nakov y Hearst (2005) Herramientas como *WebCorp* o *KwicFinder* se utilizan para estos propósitos.
2. La web como tienda de corpus: los lingüistas que hacen un uso de la web en este sentido, aprovechan las posibilidades que esta ofrece para recopilar textos obtenidos de Internet a través de los buscadores para

construir un corpus (en el sentido tradicional del término) que esté siempre disponible. Para afinar la búsqueda, se delimitan los parámetros ofrecidos por el buscador, como el idioma, la procedencia del texto, la URL, etc. Lüdeling *et al*. (2007) también aportan pautas para la extracción de información ya sea a través del buscador o de la obtención de páginas de Internet, ya sea al azar, de forma controlada, automática o manualmente. Para este propósito, la herramienta *BootCat* puede resultar útil.

3. La web como corpus en sí mismo: mientras que para los dos usos anteriores –que tienen que ver con cuestiones "oportunistas" –, los corpus en papel podrían haber sido también válidos si no fuera por el hecho de que no están digitalizados, el concepto de la web como corpus es radicalmente nuevo. Principalmente porque supone investigar la naturaleza de la web. En concreto, los autores se centran, en este apartado, en la web como corpus que representa a la lengua inglesa.
4. Mega-corpus/mini-web: se trata de la postura más radical en cuanto a la utilización de la web como corpus y se refieren a esta como un intento de crear un nuevo corpus en forma de mini-web o de mega-corpus que se adapte a la investigación lingüística. Este nuevo concepto heredaría características de sus dos fuentes de inspiración, es decir, de la web, por un lado, y del corpus, por otro. A la primera se parecería en el gran tamaño, en la constante actualización, en el material procedente de las páginas web y en una rápida interfaz basada en la web para acceder a la información. Con el corpus compartiría características como la anotación, la posibilidad de consultas sofisticadas o su carácter relativamente estable. De este nuevo concepto podrían beneficiarse tanto investigadores interesados en aspectos del lenguaje a través de la web como investigadores cuyo punto de interés sea el conocimiento de la web a través del lenguaje.

Este modo de ver las cosas evidencia la riqueza de la relación entre la web y los corpus, así como y las insondables posibilidades que se abren entre estos dos ámbitos. Evidentemente, las primeras razones que apoyan la colaboración y la interconexión

entre ambos son de carácter práctico, como hemos visto: tamaño, rapidez, accesibilidad, economía, etc. Pero, yendo un poco más allá y en la línea de las ideas de Baroni y Bernardini que acabamos de explicar, en esta relación subyacen razones más potentes para su buen funcionamiento, de tipo cuantitativo y cualitativo; además, este enfoque se aborda desde un punto de vista no solo lingüístico y tecnológico, sino también social (Crystal, 2006):

> Writers on the Internet struggle to find ways of expressing its unprecedented impact… Language being such a sensitive index of social change, it would be surprising indeed if such a radically innovative phenomenon did not have a corresponding impact on the way we communicate… The Internet is not just a technological fact; it is a social fact. (Crystal, 2006: 237)

2.3 ASPECTOS FUNDAMENTALES

El impacto que la llegada de la *World Wide Web* ha tenido en la Lingüística de Corpus y en su propia concepción ha removido los cimientos de esta disciplina y modificado o, al menos, cuestionado algunos de sus principios. Este es uno de los motivos por los que Renouf y Kehoe hablan de "the changing face of corpus linguistics" (2006: 3). La web como corpus abre las puertas para el estudio y la revisión profunda de algunos aspectos teóricos y metodológicos sobre los que se asienta el trabajo con corpus y que explicaremos a continuación, siguiendo la clasificación de Gatto (2014a).

2.3.1 Autenticidad

Si hay algo en lo que los lingüistas están de acuerdo con respecto a los corpus es que los ejemplos de la lengua contenidos en estos deben ser reales, naturales y auténticos. La mayoría de los autores reflejan este aspecto en sus definiciones de corpus. Recordemos, por ejemplo, la definición de Chafe, cuando afirmaba que le gustaría pensar que un lingüista de corpus es alguien que intenta comprender la lengua y la mente observando detenidamente grandes ejemplos naturales de la primera[23]; o

[23] "What, then, is a "corpus linguist"? I would like to think that it is a linguist who tries to understand language, and behind language the mind, by carefully observing extensive natural samples of it and then, with insight and imagination, constructing plausible understandings that encompass and explain those observations." Chafe (1992: 96)

Bowker y Pearson (2002), que defendían que la Lingüística de Corpus trabaja con ejemplos de lo que la gente ha dicho realmente, más que especular con lo que podrían haber dicho[24]. De la misma forma, Sinclair (1991), Leech (1992), McEnery y Wilson (2003), Baroni y Bernardini (2006) o McEnery y Hardy (2012), entre muchos otros, resaltan la autenticidad como una de las características clave para los corpus.

Gatto (2008) afirma que el motivo por el que la web ha adquirido el estatus de corpus no es la digitalización de los textos y su disponibilidad o su fácil acceso. La auténtica razón para que esto haya ocurrido es que los textos que la componen son textos reales, resultado de situaciones comunicativas genuinas de personas que utilizan la lengua en sus rutinas habituales.

Surge de nuevo aquí la siempre polémica contraposición entre la introspección y el empirismo de la actuación, cuya balanza se ha ido inclinando hacia uno u otro lado según el momento histórico que atravesara. En la actualidad, afortunadamente para los lingüistas de corpus, o quizá como fruto de los buenos resultados de sus investigaciones, los estudios sobre el lenguaje basados en datos empíricos y la consideración de la web como corpus le han ganado el pulso a la introspección que Chomsky defendiera en su día con tanto ahínco.

Teubert (2005: 5) parece tenerlo bastante claro cuando afirma que: "los conceptos y categorías derivados del estudio introspectivo del lenguaje o de modelos provenientes de otras disciplinas (por ejemplo, computación) pueden no ser apropiados para la descripción de la información lingüística auténtica".

También Sinclair opina lo mismo cuando habla acerca del "growing respect for real examples" (Sinclair, 1991: 5) y cuando afirma que está de moda mirar a la sociedad más que a la mente para encontrar ejemplos reales, al contrario de lo que pasaba hace treinta años, cuando "starved of adequate data, linguistics languished" (Sinclair, 1991: 1).

En este contexto de revitalización de la Lingüística de Corpus, la revolución que ha supuesto la irrupción de la *World Wide Web* no ha hecho sino contribuir al afianzamiento de esta disciplina que años atrás se había visto menospreciada. Gatto (2008) no solo está de acuerdo con esta afirmación, sino que responsabiliza en parte a la web, a la revolución digital y a los avances tecnológicos de la creciente popularidad

[24] "Simply speaking, corpus linguistics is an approach or a methodology for studying language use. It is an empirical approach that involves studying examples of what people have actually said, rather than hypothesizing about what they might or should say." (Bowker y Pearson, 2002: 9)

vivida por la Lingüística de Corpus. La enorme cantidad de ejemplos reales, electrónicos, disponibles y accesibles han contribuido a la preferencia científica por el empirismo. Esta es la razón por la que la web constituye "a fabulous linguist's playground" (Kilgarriff y Grefenstette, 2003: 345).

2.3.2 Representatividad

Íntimamente ligado a la autenticidad se encuentra el concepto de representatividad, una cuestión tan polémica como estudiada y que Leech define en una primera aproximación como "the degree to which a corpus is representative" (Leech, 2007: 133). Es difícil, por tanto, no encontrar referencias a este aspecto en una gran parte de las definiciones de corpus que aportan los diferentes autores. Biber, por ejemplo, explica que:

> A corpus is not simply a collection of texts. Rather, a corpus seeks to represent a language or some part of a language. The appropriate design for a corpus therefore depends upon what it is meant to represent. The representativenes of the corpus, in turn, determines the kinds of research questions that can be addressed and the generalizability of the results of the research. (Biber, 1998: 246)

McEnery y Wilson (2003: 32), por su parte afirman: "a corpus in modern linguistics, in contrast to being simply any body of text, might more accurately be described as a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration".

También Parodi incluye la representatividad dentro de las ocho características que, a su parecer, debe reunir un corpus a la hora de su construcción; los siete aspectos restantes que habría que tener en cuenta para esta tarea son: extensión, formato, diversificación, marcado o etiquetado, procedencia, tamaño de las muestras y clasificación y adscripciones de tipos disciplinar, temático, etc. (Parodi, 2010: 23).

Como es natural, también Chomsky tiene aquí mucho que decir puesto que, una vez más, estamos ante una perspectiva del lenguaje orientada hacia la actuación –o la *parole* de Saussure (Saussure, 2002). Según el padre de la gramática generativa, existe una dicotomía entre "*externalized language* o *E-language*" e "*internalized language* o *I-*

language" o, lo que es lo mismo, lengua externa y lengua interna, que surgen como consecuencia de la competencia y la actuación, respectivamente. En esta distinción, la Lingüística, por supuesto, debe centrarse en la lengua interna, por lo que un corpus es del todo inútil: "Linguistics should be concerned with I-language and knowledge of I-language, that is with truth about the mind/brain, putting aside the irrelevant concept of E-language, however construed." (Chomsky, 1987: 45, *apud* Leech, 2007: 135)

Los ejemplos reales de la lengua en el nivel de la actuación lingüística individual son, por tanto, fundamentales para la generalización de las afirmaciones sobre la lengua interna de Chomsky o la *langue* de Saussure. Aquí reside la verdadera importancia de la representatividad, porque es la garantía donde se asienta la validez de las afirmaciones acerca del lenguaje que se formulen a raíz de un corpus. Sin embargo, la famosa frase de Chosmky citada en el capítulo anterior, en la que asegura que todo corpus natural se encuentra sesgado (Chomsky, 1958, *apud* Leech 1991), se ha interpretado en ocasiones fuera de contexto y quizá no sea aplicable a los corpus actuales. En este punto, autores como Leech (2007) o Halliday (1991) asumen que lo que le queda al lingüista de corpus es llegar a comprender mejor la lengua-I a través del estudio de la lengua-E, puesto que son dominios totalmente independientes y la primera es la manifestación de la segunda.

Es, como decimos, un asunto polémico el de la representatividad. Tognini-Bonelli así lo reconoce cuando lo caracteriza como "a vexed question" (2001: 57). A este respecto, Kilgarriff y Grefenstette se cuestionan el alcance de la palabra y se preguntan: "¿representativo de qué?" (2003: 8) y admiten que, dejando al lado los corpus muy especializados, no es fácil determinar los límites de la representatividad de un corpus general.

Leech, sin embargo, concreta mucho más la cuestión y acota la definición de representatividad de la siguiente forma: "In practical terms, a corpus is representative to the extent that findings based on its contents can be generalized to a larger hypothetical corpus" (Leech, 1991: 27). Idea que sigue manteniendo años más tarde cuando le dedica varias páginas a este aspecto (Leech, 2007). También Váradi, a pesar de diferir en las ideas de Leech acerca de la Lingüística de Corpus, tiene un concepto similar de la representatividad y expresa que para él no significa otra cosa que diseñar un corpus que sirva como modelo de la totalidad del uso de la lengua de una comunidad (Leech, 1991: 587).

Las visiones más críticas con la Lingüística de Corpus desechan por completo la idea de la representatividad y argumentan que, mientras los defensores de los corpus

resaltan las ventajas de trabajar con una base empírica para formular generalizaciones, estudiar las variaciones y probar las teorías lingüísticas, estas cuestiones son inaceptables sin representatividad. Por tanto, si no se puede asegurar la representatividad de un corpus, cualquier verdad que se obtenga de él, es solo cierta para ese corpus y no puede generalizarse a nada más. Váradi (2001), con un gran escepticismo hacia la utilidad de los corpus, vierte duras acusaciones contra estos y subraya los problemas e inconsistencias metodológicas empleados en la práctica de corpus actual que, desde su punto de vista, desembocan, aunque de forma no intencionada, en una falta de rigor científico.

Según Biber (Biber, 1992: 174), la representatividad está condicionada por "the kind of texts included, the number of texts, from within texts, and the length of text samples". Una tarea difícil de acometer como demuestran las sugerentes expresiones con las que ha sido etiquetada, como "el Santo Grial", por Leech (2007: 134) o "la caja de Pandora", por Kilgarriff (2003: 333). También el profesor de la Universidad de Lancaster habló de ella años atrás como "un acto de fe" (Leech, 1991: 27). Sin embargo, Leech compensa el pesimismo derivado de la imposibilidad de alcanzar la representatividad con la idea de que se pueden dar pasos que nos acerquen a ella.

El inmenso tamaño de la web parece mitigar de algún modo el problema de la representatividad debido a que el número de ejemplos reales de usos de la lengua se multiplica exponencialmente con respecto a los corpus tradicionales. Siguiendo con el mismo autor, "the web as corpus makes the notion of a representative corpus redundant" (Leech, 2007: 144) porque, si toda la web puede ser explorada con un motor de búsqueda, no es necesario un corpus representativo porque tenemos todo el universo textual a nuestra disposicón. No obstante, resalta la idea de que aunque pueda parecer que la web convierte a los corpus en una herramienta prescindible y sustituible por los buscadores comerciales, esto se desvanece cuando tenemos en cuenta que no están diseñados para realizar búsquedas lingüísticas y, por lo tanto, adolecen de gran parte de las ventajas que aportan los corpus. Gatto (2008) opina que, a pesar de ello, en la web reside todo el potencial para obtener la representatividad porque los textos que contiene son el producto de interacciones humanas y reflejan a la comunidad internacional en tiempo real. Manning y Schütze (1999: 119) afirman que: "A sample is representative if what we find for the sample also holds for the general population". Por consiguiente, el alcance, la variedad y el tamaño de la web son capaces de compensar los límites de la representatividad. En palabras de Kilgarriff y Greffenstette (2003: 343), "the web is not

representative of anything else. But nor are other corpora, in any well- understood sense".

2.3.3 Tamaño

Mucho se ha debatido acerca del tamaño de los corpus y su influencia en la representatividad y la utilidad de estos. Parece innegable que el resurgir de la lingüística de corpus vino de la mano de las mejoras técnicas de los ordenadores y del aumento de la información lingüística disponible para su análisis. Sin embargo, como es habitual en casi todos los aspectos relativos a la lingüística de corpus, la idoneidad o la conveniencia de grandes corpus también han sido extensamente debatidas. Argumentos a favor y en contra de la utilización de grandes corpus abundan, al igual que también existen autores que no entran en polémica y no aportan especificaciones acerca del tamaño ideal de los corpus. Así lo expresan, por ejemplo, Bowker y Pearson (2002: 45):

> Unfortunately, there are no hard and fast rules that can be followed to determine the ideal size of a corpus. Instead, you will have to make this decision based on factors such as the needs of your project, the availability of data and the amount of time that you have.

A pesar de ello, aconsejan no admitir el gran tamaño como algo siempre deseable:

> It is very important, however, not to asume that bigger is always better. You may find that you can get more useful information from a corpus that is small but well designed than from one that is larger but is not customized to meet your needs. (Bowker y Pearson, 2002: 45, 46)

Atkins, Clear y Ostler (1992), en su guía para la elaboración y el diseño de corpus, tampoco precisan el tamaño más conveniente para estos y puntualizan que se trata de una cuestión que todavía está esperando a ser resuelta.

Fletcher (2012), por el contrario, no duda al afirmar que uno de los factores responsables de que la web haya alcanzado tal popularidad es su propio tamaño. Paradójicamente, el continuo crecimiento de la web y su gran tamaño son, al mismo

tiempo, algunas de sus mayores virtudes y de sus más importantes limitaciones como objeto de investigación lingüística. Esto entra en contraposición con definiciones de corpus como la de McEnery y Wilson (2003: 30), en la que destacan el carácter finito como una de las características de este: "a body of text of a finite size". Sin embargo, los autores reconocen la existencia y las virtudes de corpus que no se ajustan a esta definición. Se refieren, claro está, a los *monitor corpus*, como el que construyera Sinclair en el proyecto Cobuild. El mismo Sinclair define *monitor corpus* como un corpus "which has no final extent because, like language itself, it keeps on developing" (Sinclair, 1991: 25). Podríamos considerar a la web, por tanto, la última versión en la evolución de los *monitor corpus*.

Esta naturaleza infinita de continuo crecimiento se presenta como uno de los problemas que más preocupan a algunos de los investigadores en el campo.

(Gelbukh, Sidorov y Chanona, 2002: 3) explica que una de las desventajas de los corpus tradicionales es que presentan pocas o ninguna ocurrencia de muchas palabras, mientras que otras aparecen muy repetidas, debido al fenómeno conocido como la ley de Zipf[25] (Zipf, 1965). También Rayson, Walkerdine, Fletcher y Kilgarriff (2006) achacan a esta ley el hecho de que la mitad de las palabras que hay en los corpus aparecen una sola vez, por lo que los grandes corpus son necesarios para asegurar la inclusión de palabras y frases fundamentales y para aumentar las posibilidades de aparición.

Baroni y Ueyama (2006) también defienden el concepto de grandes corpus para el procesamiento del lenguaje natural, basándose en artículos como el de Banko y Brill (2001), que demuestran que incluso los algoritmos sencillos de desambiguación funcionan mejor en el seno de grandes cantidades de información, Mair (2006) o Turney (2001). Afirman que incluso para los lenguajes muy especializados, la inmensidad de la web puede ser útil también para la construcción de pequeños corpus, puesto que solo una base de datos tan grande como la web puede contener la información suficiente para la construcción de ese corpus.

Guillermo Rojo (2008, *online*) tampoco duda en afirmar que: "con toda claridad, es necesario seguir construyendo corpus y su tamaño debe ser lo más grande que podamos conseguir".

[25] En 1935, George Kingsley Zipf, profesor de Lingüística de la Universidad de Harvard, estableció que el número de apariciones de una palabra es inversamente proporcional a su número de orden.

Por otro lado, Halliday (2007: 298) distingue entre el corpus como objeto y el corpus como instrumento. En el segundo de los casos, su importancia reside en que se constituye como una ventana para el estudio de la lengua y, por lo tanto, el tamaño es muy importante porque mientras mayor sea el corpus, más información aportará sobre el sistema.

Para el enfoque explicado en el primer capítulo de este trabajo, denominado *corpus-driven linguistics* y defendido por Tognini-Bonelli (2001), Teubert (2005), Sinclair (2004) o Gatto (2008, 2014a), entre otros, el tamaño es una cuestión fundamental puesto que el corpus no se concibe para comprobar o refutar una teoría, sino como base para la elaboración de una nueva. Como resultado, un corpus pequeño solo aporta evidencias del fenómeno del lenguaje objeto de la investigación y, como consecuencia de ello, se corresponde con un pequeño fotograma de la complejidad del lenguaje. Por el contrario, un corpus de gran tamaño es capaz de ofrecer una visión más amplia y completa. Sinclair (2004: 189) defiende abiertamente esta postura:

> There is no virtue in being small. Small is not beautiful; it is simply a limitation. If within the dimensions of a small corpus, using corpus techniques, you can get results that you wish to get, then your methodology is above reproach –but the results will be extremely limited, and also the range of features that you can observe. The main virtue of being large in a corpus is that the underlying regularities have a better chance of showing through the superficial variations, and there's a lot of variation in the surface realization of linguistic units in a corpus.

En este punto, y haciendo referencia a la ley de Zipf, argumenta que en un corpus de gran tamaño existen más posibilidades de encontrar las palabras con bajo índice de frecuencia.

En lo que se refiere a la web, calcular su tamaño no solo es una tarea inabarcable, sino que, además, las vagas estimaciones al respecto conducen a unos resultados efímeros, debido al dinamismo que la caracteriza. Eric Schmidt, director ejecutivo de Google hasta 2011, estima que el tamaño del la *World Wide Web* ronda los cinco millones de terabytes de información, de los cuales la empresa solo tiene indizados 200 terabytes. Esto supone el 0,004% del total (Domínguez, 2015). Por otro

lado, *Internet Live Stats* (2015) calcula que el número de páginas web en Internet rondaba, en 2015, los mil millones.

Sea cual fuere el tamaño exacto, es indiscutible que la web proporciona a los lingüistas una colección de textos infinitamente mayor que cualquier otro corpus existente. En los primeros años de la Lingüística de Corpus, conseguir un corpus de un tamaño suficiente como para obtener una evidencia significativa suponía un problema. Actualmente, la situación se ha dado la vuelta por completo y el reto está en conseguir sacar el máximo provecho de los grandes corpus grandes sin que el científico se vea sobrepasado, como advierte Hunston: "the sheer quantity of linguistic information can be overwhelming for the observer" (2002: 25).

Sin embargo, estudios como Baroni y Kilgarriff (2006), Banko y Brill (2001) o Keller y Lapata (2003) demuestran que las grandes cantidades de información –incluso aquellas que aportan "ruido" – son más convenientes que las bases de datos pequeñas, especialmente cuando se trata de procesamiento del lenguaje natural (Gatto, 2014a).

2.3.4 Contenido

La principal consecuencia lógica del crecimiento desmesurado del tamaño de la *World Wide Web* es el aumento del contenido. Sin duda, otro de los aspectos fundamentales a la hora de valorar las cualidades de un corpus, que tiene mayores repercusiones aun cuando se trata de la web como corpus. Además, el incremento de contenido viene también de la mano de la representatividad, ya que, conforme aumenta el número de páginas web disponibles sin restricciones en cuanto al tema, a la lengua, al tipo de texto, al formato, etc., crecen las posibilidades de que todo ese contenido sea representativo.

Hay que tener en cuenta que las limitaciones prácticas a la hora de seleccionar el contenido para la construcción de un corpus han sido, a menudo, determinantes en el producto final. Así lo cree Hunston (2002: 27) al afirmar que aspectos como el copyright, el formato o la disponibilidad indudablemente influyen en el diseño de los corpus. Y, al fin y al cabo, el contenido de los corpus es precisamente lo que determina el alcance de las generalizaciones que se pueden extraer de los mismos (Gatto, 2008).

La diversidad de contenidos presentes en la web no deben, en ningún caso, disuadirnos en su utilización para propósitos lingüísticos. Como afirmaba Kilgarriff, es necesario enfrentarse a la "anarquía" (2001: 1) para poder sacar provecho de ella. Esta

anarquía viene provocada por la posibilidad de que cualquier usuario pueda generar contenido en tiempo real, con formato electrónico, de cualquier tipo o género y en cualquier lengua. Pero aquí reside la verdadera riqueza de la web y es de donde parece fundamental tratar de obtener el máximo provecho.

Siguiendo la clasificación de Gatto (2014a), analizaremos cuatro aspectos importantes dentro del contenido de la web.

2.3.4.1 Lengua

Si hay algún punto en el que el contenido de los corpus de la web se ha beneficiado con respecto al de los corpus tradicionales, ese es su carácter plurilingüe. La práctica totalidad de los idiomas existentes en el planeta están contenidos en la web al alcance de un botón, incluso los idiomas más minoritarios, que hasta ahora presentaban graves dificultades a la hora de entrar en contacto con ellos. Así lo expresaba ya David Crystal cuando todavía la web no había alcanzado ni el tamaño ni la variedad a los que nos tiene acostumbrados hoy en día:

> The Web is an eclectic medium, and this is seen also in its multilinguistic inclusiveness. Not only does it ofer a home to all linguistics styles within a language; it offers a home to all language –one thier communities have a functioning computer technology. (Crystal, 2006: 216)

La riquísima variedad de idiomas que habitan en la red ha dado lugar a que uno de los ámbitos que, en el campo de la Lingüística, más presencia tiene en los trabajos derivados de la web como corpus sea la enseñanza de lenguas extranjeras o el análisis de aspectos gramaticales o léxicos de idiomas concretos.

Grefenstette y Nioche estimaron en el año 2000, en un estudio sobre los idiomas europeos de la web, que a pesar de que el inglés era el idioma más utilizado, el uso del resto de idiomas crecía a mayor velocidad que el inglés. En noviembre de 2015, la tendencia a nivel mundial sigue siendo la misma: el inglés es el idioma más hablado en la red (casi 900 millones de usuarios), mientras que el resto de idiomas continúa creciendo y fluctuando entre los demás puestos. En el siguiente gráfico, podemos observar la distribución actual de las distintas lenguas en la red:

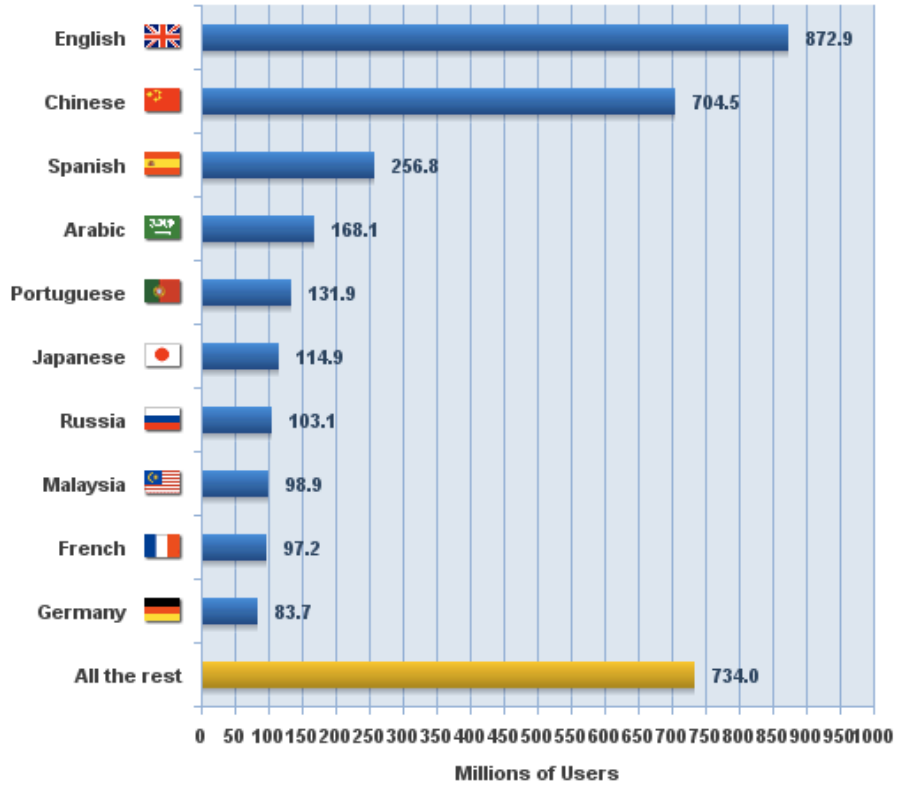


Figura 2.1. Los diez idiomas más utilizados en Internet en noviembre de 2015

Fuente: Internet World Stats, 2016

Si comparamos los datos de esta tabla con los idiomas más hablados en el mundo según Lewis, Paul, Simons y Fennig (2015), vemos que los tres primeros puestos de la clasificación los siguen ostentando los mismos idiomas, con la diferencia de que el inglés, que es el primer idioma más utilizado en la web, es el tercer idioma más hablado en el mundo, precedido, en orden, por el chino y el español. Para poder contextualizar mejor los datos de la tabla anterior, a continuación los acompañamos de los porcentajes relacionados con el número de personas que tienen acceso a Internet, en la tercera columna; el aumento de usuarios de Internet, en la cuarta; el porcentaje de usuarios de Internet de ese idioma a nivel mundial, en la siguiente columna; y la estimación del número de hablantes de cada lengua, en la última columna.

Top Ten Languages Used in the Web - November 30, 2015 (Number of Internet Users by Language)					
TOP TEN LANGUAGES IN THE INTERNET	Internet Users by Language	Internet Penetration (% Population)	Users Growth in Internet (2000 - 2015)	Internet Users % of World Total (Participation)	World Population for this Language (2015 Estimate)
English	872,950,266	62.4 %	520.2 %	25.9 %	1,398,283,969
Chinese	704,484,396	50.4 %	2,080.9 %	20.9 %	1,398,335,970
Spanish	256,787,878	58.2 %	1,312.4 %	7.6 %	441,052,395
Arabic	168,176,008	44.8 %	6,592.5 %	5.0 %	375,241,253
Portuguese	131,903,391	50.1 %	1,641.1 %	3.9 %	263,260,385
Japanese	114,963,827	90.6 %	144.2 %	3.4 %	126,919,659
Russian	103,147,691	70.5 %	3,227.3 %	3.1 %	146,267,288
Malay	98,915,747	34.5 %	1,626.3 %	2.9 %	286,937,168
French	97,180,032	25.2 %	709.9 %	2.9 %	385,389,434
German	83,738,911	87.8 %	204.3 %	2.5 %	95,324,471
TOP 10 LANGUAGES	2,632,248,147	53.5 %	787.0 %	78.2 %	4,917,011,992
Rest of the Languages	734,013,009	31.3 %	1,042.9 %	21.8 %	2,342,890,251
WORLD TOTAL	3,366,261,156	46.4 %	832.5 %	100.0 %	7,259,902,243

Figura 2.2. Los diez idiomas más hablados en la web y número millones de usuarios de Internet de cada idioma.

Fuente: Internet World Stats, 2016

Que estas cifras resulten más o menos sorprendentes se trata de una cuestión subjetiva y que queda sujeta al juicio del lector. Sin embargo, en una comparativa de este estudio con su homólogo cinco años atrás, el análisis puede resultar más relevante y clarificador:

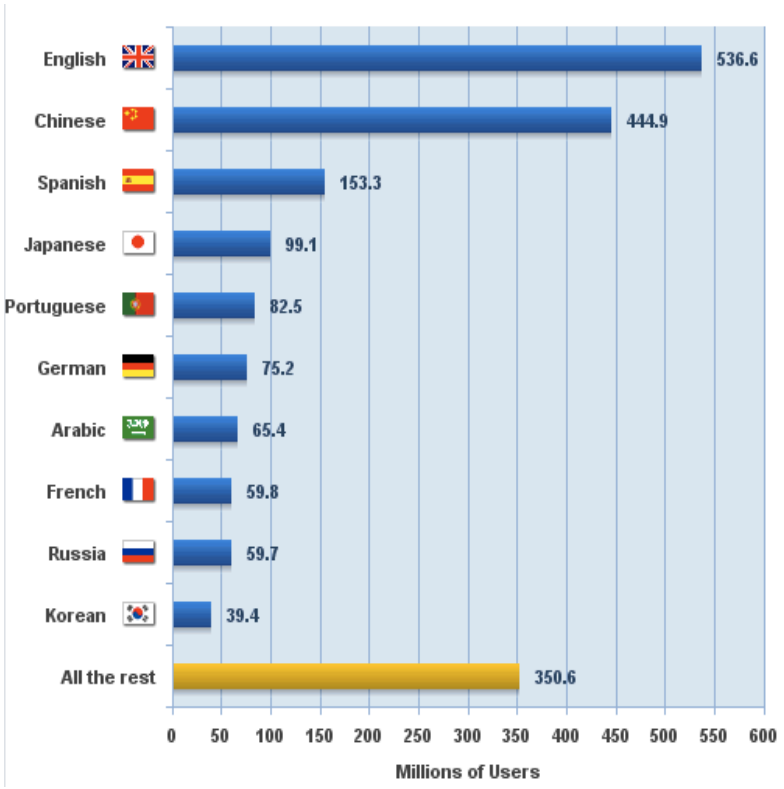


Figura 2.3. Los diez idiomas más utilizados en Internet en noviembre de 2010

Fuente: Internet World Stats, 2016

Queda claro que el número de usuarios de Internet en las distintas lenguas ha crecido desde 2010. Por ejemplo, el inglés ha aumentado en 336 millones, mientras que el chino lo ha hecho en casi 260. El porcentaje de español ha sumado, en estos últimos cinco años, más del 100%.

Los idiomas más utilizados también han variado a partir del cuarto puesto. Mientras que en 2010, el japonés, el portugués, el alemán y el árabe ocupaban cuarto, quinto, sexto y séptimo lugar, respectivamente, en 2015, estos puestos están tomados, en orden, por el árabe, el portugués, el japonés y el ruso. Resulta curioso el movimiento que realizan los distintos idiomas y cómo idiomas como por ejemplo, el alemán, quedan relegados en este último año a la novena posición. Sin embargo, encontramos la explicación de estos cambios en la cada vez mayor conectividad a Internet de países con gran número de habitantes en los que el desarrollo tecnológico ha tardado más en llegar. Por el contrario, países como Alemania, siempre a la vanguardia en medios tecnológicos, han sufrido muchos menos cambios. Es, por tanto, la consecuencia del fenómeno conocido como "brecha digital".

2.3.4.2 Temas

La *World Wide Web* se ha convertido en un medio de comunicación tan extendido y generalizado que es difícil encontrar alguna actividad humana que no haya sido alcanzada por ella. El advenimiento de la web 2.0 acentuó aún más el perfil multitemático de la web y gracias a ella se nos permitió a los usuarios no solo consumir información, sino también generarla e incluso compartirla.

Como es lógico, la costumbre de clasificar los textos en tipologías ha tratado de buscarse un hueco en Internet y se han llevado a cabo numerosos intentos para organizar el contenido y sacarle así el máximo provecho. Chakrabarti (2003: 7) nos recuerda que: "organizing knowledge into ontologies is an ancient art, descended from philosophy and epostemology". Sin embargo, las pretensiones de clasificar de forma tradicional algo tan radicalmente distinto a lo que respondía perfectamente a los criterios de clasificación han quedado en gran medida frustradas porque no pueden adaptarse a la propia naturaleza inclasificable de la web.

Los primeros intentos fueron llevados a cabo en los años 90 por Jerry Yang y David Filo, en la Universidad de Stanford (Gatto, 2014a). Estos dos estudiantes de doctorado crearon el directorio de Yahoo! para ayudar a sus amigos a encontrar las páginas web que fueran de su interés. Hacia la misma época surgió el *Open Directory Project*, también conocido como *DMoz* y el único que perdura en la actualidad. La forma de proceder de estos directorios se basa –o se basaba– en el trabajo de editores humanos (voluntarios, en el caso de *DMoz*), quienes se encargan de introducir a mano en el directorio las páginas web que lo soliciten o que cumplan los requisitos establecidos.

Los beneficios que los directorios ofrecen se reducen principalmente a dos. Por un lado, la catalogación del contenido web hace la búsqueda más fácil y rápida y, una vez que el usuario ha entrado en la primera categoría, la búsqueda se redirige de manera que se le ofrecen las páginas relacionadas que puedan ser de su interés. Por otro lado, las páginas a las que el usuario es conducido han pasado por el filtro de los editores antes de llegar a estar presentes en el directorio, lo que da ciertas garantías de calidad y las distingue de otras páginas que no hayan sido capaces de formar parte de él.

Lógicamente, conforme ha ido creciendo el número de páginas web disponibles, la tarea se ha vuelto inabarcable y esto ha dado lugar al cierre de la mayoría de los directorios, excepto *DMoz*, que a pesar de estar desactualizado y prácticamente en desuso, se sigue manteniendo a flote debido a su nulo coste de mantenimiento.

Figura 2.4. Directorio de *Open Directory Project* (*DMoz*)

Fuente: Dmoz.org, consultado 27 de agosto de 2015

En cualquier caso, y a pesar de las facilidades que los directorios puedan brindarnos como usuarios regulares de Internet, los beneficios que para la Lingüística de Corpus se puedan extraer no son tan numerosos o, cuando menos, hay que tomarlos con mucha prudencia. La principal razón es que una etiqueta que se refiera de forma general a un tema concreto no es suficiente para que un lingüista pueda discriminar el contenido web; para lo único que sirven estos directorios es para que tareas de búsqueda de información resulten algo más efectivas (Gatto, 2014a).

2.3.4.3.Registros, géneros y tipologías textuales

El desarrollo tecnológico también ha traído como consecuencia la aparición de nuevas tipologías textuales, como los chats, los blogs, los SMS, etc. que se encuentran a medio camino entre el registro oral y el escrito y que han suscitado numerosos estudios

que prestan atención a las diferencias en el uso del lenguaje entre unos y otros, como Piñol (1999), Gómez Torrego (2001), Crystal (2006), Calvo Revilla (2002) o López Quero, Calero Vaquera y Zamorano Aguilar (2004), entre muchísimos otros, ya citados más arriba.

Las hasta ahora claras líneas divisorias entre la oralidad y la escritura, y el registro formal o informal han empezado a desdibujarse hasta el punto que nuevos registros, géneros y tipologías que se ajusten a las características de la lengua de Internet están empezando a surgir.

Sharoff (2007) admite esta idea y pone en duda que las categorías existentes hasta el momento sean útiles para los textos procedentes de la web:

> Texts in representative corpora are typically classified into their domain and genre. However, it is not clear if existing domain and genre typologies can be applied at all to unlabeled data collected from the Web, for instance, to results of crawling. (Sharoff, 2007: 83)

Además, hay que tener en cuenta que Internet no está solo compuesto de nuevos géneros o tipologías textuales: no podemos olvidar que los textos tradicionales también están presentes en la web y que han sido convertidos a formato electrónico sin perder sus características originales, lo que aumenta aún más la variedad disponible. Marina Santini (2007) sugiere analizar esta problemática desde dos perspectivas distintas: desde el "hibridismo" y desde la "individualización". El primero de ellos tiene que ver con la "variación multigénero" presente en las páginas web individuales, mientras que la "individualización" se refiere a la ausencia de un género conocido dentro de una página web.

Por lo tanto, la solución que autores como Santini (2007) o Mehler, Sharoff y Santini (2010, *apub* Gatto, 2014a: 119) proponen pasa por la creación de una clasificación más flexible que sea capaz de integrar los nuevos géneros y los tradicionales en un mismo sistema del que se beneficien tanto la búsqueda y extracción de información como la propia Lingüística de Corpus.

2.3.5 Copyright

Kilgarriff y Grefenstette (2003) argumentan que, aunque los abogados se empeñen en asimilar las aplicaciones legales de copyright de los corpus de Internet a los corpus tradicionales, hay dos diferencias fundamentales La primera de ellas es que los investigadores tienen la posibilidad de recopilar un corpus de Internet simplemente accediendo a los documentos y almacenando páginas web sin copiarlas; la segunda, por el contrario, tiene que ver con el aspecto de la insignificancia: si un lingüista de corpus utiliza material para su trabajo infringiendo la ley de copyright, está haciendo lo mismo que un buscador comercial, con la diferencia de que este lo hace a una increíblemente mayor escala.

Fletcher (2004), por su parte, en este ambiente de indeterminación legal, se plantea la misma cuestión y, aunque con cierta intención de no sobrepasar los límites legales, elude responsabilidades desde una perspectiva optimista:

> Optimistically I assume that a Web-accessible corpus for research and education derived from online documents retrieved by a search agent in ad-hoc searches will fall within legal boundaries. Meanwhile, I intend to assert and help establish our profession's rights while scrupulously respecting any restrictions a webpage author communicates via industry-standard conventions. (Fletcher, 2004: 281)

Rock (2001) y McEnery y Hardie (2012) también se plantean la cuestión, enfocándola esta vez desde distintas perspectivas, pero aseguran que los asuntos legales varían entre países y también con el tiempo. En cualquier caso, e independientemente del sistema legal, hay distintas formas de acercarse al problema. Gatto (2014a) distingue cuatro casos: a) se utiliza la información disponible en la web después de haber contactado con los dueños del copyright, b) se utiliza la información disponible solo en dominios públicos, c) se utiliza cualquier tipo de información, pero no se distribuye y d) se redistribuyen exclusivamente las direcciones web y no el contenido de los textos que contienen.

2.3.6 Nuevas características

Hasta ahora hemos visto las bases teóricas sobre las que se asienta la Lingüística de Corpus y que, con las diferencias lógicas que implica el uso de la *World Wide Web*, se pueden aplicar a la web como corpus. Maristella Gatto (2008) enumera tres características más que distinguen a los corpus compilados con material de la web en cualquiera de las formas que presentaban Baroni y Bernardini (2006) de los corpus tradicionales por su relación con Internet. Advierte la autora, no obstante, de que no se trata de características específicas de los web corpus, sino al impacto que las nuevas tecnologías tienen en las fuentes lingüísticas en general. Estas nuevas características son: dinamismo, reproducibilidad, y relevancia y fiabilidad.

El **dinamismo** tiene que ver con la naturaleza cambiante de la web, provocada no solo porque el número de páginas y de datos aumenta diariamente, sino porque estos datos son extremadamente variables, se actualizan constantemente, se modifican o se eliminan. Como ya hemos explicado, la ventaja que esto conlleva es fundamentalmente que, gracias a este dinamismo, la web se convierte en la mayor fuente de información y de conocimiento que se haya conocido nunca.

Por otro lado, esta cuestión está directamente relacionada con el problema de la reproducibilidad de los datos, del que también hemos hablado unas páginas más atrás, al enumerar las posibles desventajas de la web como corpus. No obstante, debemos dejar de ver este último punto como una desventaja insuperable puesto que si la web es un reflejo de la interacción humana y esta es relativamente constante durante un período de tiempo determinado, los resultados van a ser parecidos, independientemente de la fuente de donde obtengamos los datos que vayamos a analizar. Además, aunque es cierto que resulta imposible almacenar todo el contenido de la web, ya hemos visto que es posible almacenar, de una forma u otra, los datos que utilicemos para la investigación, de manera que se pueda recurrir a ellos en cualquier otro momento y las nuevas tecnologías están avanzando mucho en este sentido.

En último lugar, el dinamismo también genera otras características que han de ser tenidas en cuenta a la hora de llevar a cabo cualquier estudio. El "ruido" que se genera en la web puede afectar a los análisis tanto cuantitativos como cualitativos. Es, por tanto, fundamental tener este aspecto en cuenta para que la investigación tenga un alto grado de **relevancia** y de **fiabilidad**. Los problemas de duplicación de datos o de falta de precisión lingüística se acentúan si el estudio se lleva a cabo a través de

buscadores comerciales porque, insistimos, estos no fueron creados en sus inicios con fines lingüísticos. Esta es la razón por la que nos parece tan importante el desarrollo de herramientas específicas capaces de limpiar los datos y que nos permitan obtener la información que necesitamos.

2.4 HERRAMIENTAS PARA LA INVESTIGACIÓN LINGÜÍSTICA EN LA WEB

Como ya hemos explicado, hay muchas maneras de acometer la tarea de la investigación lingüística a través de la web. No solo por el tipo de acercamiento a la misma para construir un corpus o para utilizar la web como corpus en sí misma, sino por la diversidad de herramientas que se pueden utilizar para acceder a la información. En un primer acercamiento y con fines *ad hoc* o poco científicos, los buscadores comerciales son el recurso más inmediato de búsqueda de información. Sin embargo, existen algunas herramientas útiles y más específicas, diseñadas ya con fines lingüísticos, que intentan solventar algunos de los problemas derivados de la utilización de la web como corpus y que acabamos de comentar. Las herramientas que aquí vamos a explicar son las dos más utilizadas en Lingüística de Corpus: *WebCorp Live* y *BootCat*.

2.4.1 *WebCorp Live*

WebCorp Live[26] es la última versión de la herramienta *WebCorp*, creada en 1998 en la Universidad de Birmingham con Antoinette Renouf y Andrew Kehoe a la cabeza (Kehoe y Renouf, 2002). Su objetivo es desarrollar una herramienta de búsqueda que presente ejemplos de uso reales de la web con una interfaz y unos resultados más adecuados para los lingüistas. Con respecto a los buscadores comerciales habituales, permiten realizar una búsqueda más enfocada a la investigación lingüística y con la delimitación más precisa de algunos de sus aspectos. Sus funciones se limitan a búsquedas léxicas con el formato KWIC (*Key Word in Context*), colocaciones y frecuencias de uso de palabras.

En principio, la búsqueda se lleva a cabo como se haría en cualquier buscador, introduciendo una serie de parámetros (el idioma, la API, la aparición de variantes o el

[26] http://www.webcorp.org.uk/live/

número de caracteres que aparecerán a la izquierda y a la derecha de la palabra) en una interfaz muy similar a la de estos:

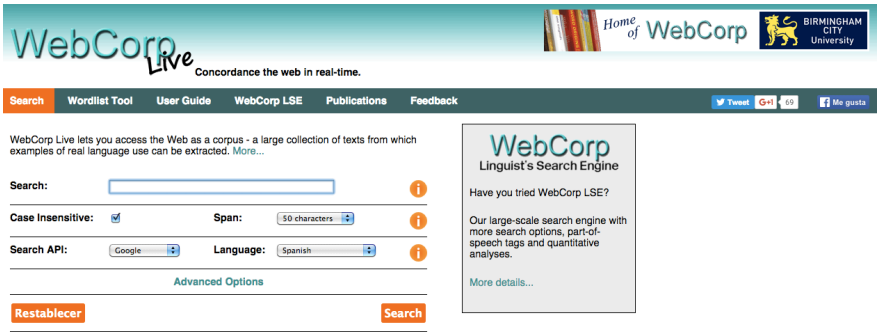


Figura 2.5. Página de inicio de *WebCorp Live*

Fuente: WebCorp.org, consultado el 15 de noviembre de 2015

Los resultados de la búsqueda, como decimos, los devuelve en formato KWIC. Por ejemplo, para la palabra "elecciones", el primero de los resultados es el siguiente:

Results for query "elecciones"

case insensitive,
using the Google API

1) https://es.wikipedia.org/wiki/Elecciones
Text, Wordlist, text/html, UTF8 (Content-type), 2016-01-05 (Server header)

1:		Elecciones	De Wikipedia, la enciclopedia libre Saltar a:
2:	de discusión pegando: {{subst:Aviso referencias	Elecciones}}	~~~~ En política, las elecciones son un
3:	referencias|Elecciones}} ~~~~ En política, las	elecciones	son un proceso de toma de decisiones en el que
4:	políticos en una democracia representativa. Hay	elecciones	generales (las que se convocan para elegir a los
5:	caso al jefe del Estado o del poder ejecutivo) y	elecciones	locales, de ámbito municipal o regional. En
6:	popular otros cargos. Se considera que hay	elecciones	libres cuando el voto se emite en circunstancias
7:	de igualdad. A fin de evitar el fraude en las	elecciones,	se hace uso de la observación electoral. La
8:	observación electoral. La organización de unas	elecciones	libres y competitivas es en la calidad en
9:	total de la función estatal de organizar las	elecciones	por parte del Poder Ejecutivo Federal, y la
10:	electorales.[1] Las características de las	elecciones	en cada país se regulan en la legislación
11:	electoral, como por ejemplo su naturaleza de	elecciones	directas (la totalidad de la ciudadanía elige
12:	ciencia política que analiza científicamente las	elecciones	se denomina psefología (de psephos ψῆφος,
13:	encomiendas. Índice 1 Características de las	elecciones	1.1 Quién puede votar 1.2 Nominación 1.3 Quién
14:	Notas 5 Enlaces externos Características de las	elecciones	[editar] Quién puede votar[editar] La pregunta
15:	quién debe sufragar es un asunto central en las	elecciones.	En el electorado generalmente no se encuentra
16:	y propietarios. Gran parte de la historia de las	elecciones	se trata sobre la lucha y promoción del voto
17:	del gobierno para las cuales se celebran las	elecciones	varían dependiendo de la localidad. En una
18:	algunas posiciones no son llenadas mediante	elecciones,	por ejemplo, los jueces son usualmente
19:	o por su propio partido). Generalmente las	elecciones	directas y aquellas con segundos grados
20:	Véase también[editar] Derecho electoral	Elecciones	por país Electorado Sistema D'Hondt Protocolo nº
21:	Wikiquote alberga frases célebres sobre	Elecciones.	Wikiquote Wikinoticias tiene noticias relacionad
22:	Wikinoticias tiene noticias relacionadas con	Elecciones.	Wikinoticias The Surprising Effect of Facial
23:	de «https://es.wikipedia.org/w/index.php?title=	Elecciones	&oldid=87879466» Categorías: Ciencia política
24:	Ciencia política Derecho constitucional	Elecciones	Teoría de la decisión Categoría oculta:

Figura 2.6. Resultados de búsqueda en KWIC con *WebCorp Live*

Fuente: WebCorp.org, consultado el 5 de febrero de 2016

Otro de los servicios que ofrece esta herramienta es la creación de listas de frecuencias de palabras de un documento concreto o de una determinada página web, introduciendo su URL, así como el análisis de colocaciones:

Collocates

Word	L4	L3	L2	L1	R1	R2	R3	R4	Total
de	24	16	72	22	21	30	32	34	251
las	5	6	20	167	1	6	4	8	217
en	14	11	55	5	26	15	17	10	153
Elecciones	27	21	17	7	7	17	23	24	143
2015	2	2	6	20	38	21	13	10	112
Generales	2	1	7	2	48	6	4	5	75
a	6	7	12	4	25	3	8	8	73
y	16	4	5	3	10	18	8	5	69
la	18	12	3	0	0	9	9	16	67
elecciones	10	7	6	3	3	6	5	11	51
Municipales	2	7	4	2	21	3	4	3	46
el	16	2	2	0	1	7	5	12	45
Política	3	5	7	16	0	5	3	3	42
del	8	2	0	0	6	14	5	6	41
se	5	6	3	0	7	7	4	4	36
que	4	5	7	0	3	0	4	6	29
los	9	2	0	0	0	2	7	6	26
al	4	2	0	0	14	1	1	2	24
El	3	2	1	0	4	8	3	2	23
2016	0	3	0	3	5	6	5	1	23

Figura 2.7. Resultados de colocaciones de la búsqueda de "elecciones" en *WebCorp Live*
Fuente: WebCorp.org, consultado el 5 de febrero de 2016

El principal problema que ofrece esta herramienta es que los resultados de la búsqueda se almacenan en una caché durante siete días a partir del día en que se realiza la consulta, lo que nos lleva de nuevo al problema de la irreplicabilidad.

En una ampliación de la herramienta, en parte para solucionar este problema, se creó *WebCorp Linguist's Search Engine*, pensada para realizar búsquedas de palabras o de frases sobre una serie de corpus ya elaborados, como son: *Synchronic English Web Corpus*, *Diachronic English Web Corpus*, *Birmingham Blog Corpus*, *Anglonormal Correspondence Corpus* y *Novels of Charles Dickens*.

2.4.2 *BootCat*

A diferencia de *WebCorp*, que se trata de una herramienta *online* accesible desde cualquier ordenador, *BootCat*[27] es un programa descargable que se puede ejecutar en cualquier ordenador que tenga un entorno Java instalado. Está diseñado para compilar corpus de la web de manera automática, es decir, cumple las funciones que cualquier lingüista de corpus tendría que llevar a cabo si quisiera elaborar un corpus con textos extraídos de la web, pero con la idea de que se pueda hacer de manera rápida e instantánea y, por supuesto, sin esfuerzo alguno. Realiza la búsqueda, descarga los resultados relevantes y lleva a cabo los procedimientos necesarios para el cambio de formato y el almacenamiento de la información que conformará el corpus; todo ello, en unos minutos.

El primer paso para la construcción del corpus es definir el nombre que llevará por título y el idioma que se utilizará.

Figura 2.8. Pantalla de *BootCat* para definir el nombre del corpus y el idioma

Para darle al programa las directrices necesarias para comenzar la búsqueda del corpus deseado, basta con introducir palabras o expresiones clave relacionadas con el tema en cuestión; así, el propio programa realizará las búsquedas pertinentes en el buscador.

[27] http://bootcat.sslmit.unibo.it/

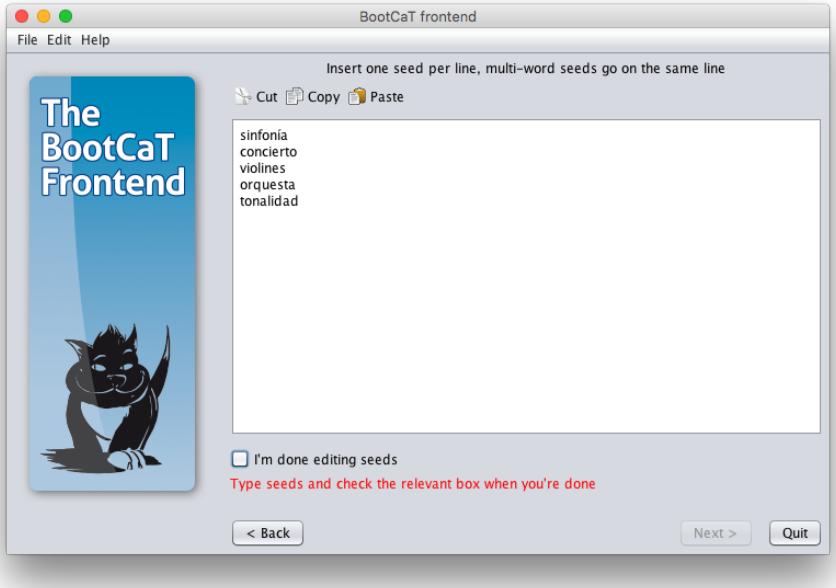


Figura 2.9. Pantalla de *BootCat* para introducir las palabras claves

A continuación, el programa ofrece por defecto las posibles combinaciones de palabras que puede buscar, relacionadas con las palabras claves introducidas anteriormente. Estas combinaciones son editables.

Figura 2.10. Pantalla de *BootCat* para editar las combinaciones de palabras

Tras introducir estas palabras claves, el programa realiza la búsqueda sobre *Bing* para acceder a la información que servirá de base para el corpus. El usuario puede intervenir en el proceso especificando parámetros tales como el número de búsquedas, el número de palabras claves por cada búsqueda o el número de páginas que quiere obtener. Además, también se pueden excluir páginas o dominios no deseados, lo que permite evitar el "ruido" o mejorar la relevancia y la fiabilidad del corpus.

Una vez que el programa ha terminado de seleccionar las páginas web que se ajustan a los parámetros establecidos por el usuario, descarga el contenido de estas y lo convierte en texto, que será el corpus definitivo.

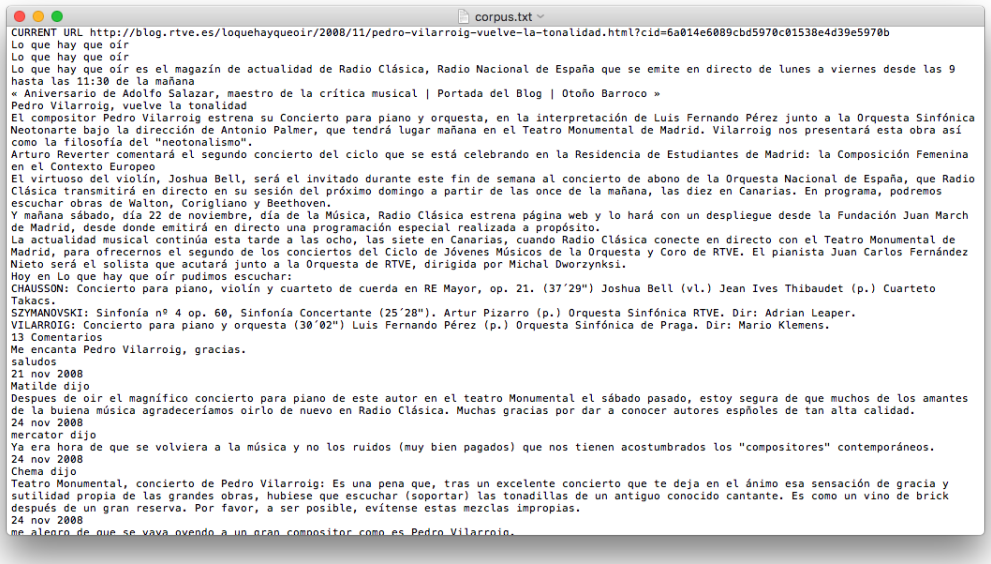


Figura 2.11. Resultado final de corpus obtenido con *BootCat*

La herramienta *BootCat*, por lo tanto, sirve para elaborar corpus con información obtenida de la web en solo unos minutos y sin ningún esfuerzo. Sin embargo, esa es su única función, ya que no ofrece soluciones para analizar la información contenida en los corpus elaborados. Para ello, es necesario utilizar posteriormente otros programas, como *Shared ngram collector*, *Onion*, *justText* o *Web Content Extractor* –para eliminar el posible "ruido" que se haya generado– o herramientas como *WordSmith Tools* –pensadas para establecer listas de frecuencia o de concordancias.

3
Big data

3.1 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 innovación avanza rápidamente y la única constante que podemos encontrar en ella es el cambio, un cambio incesante y acelerado que trae consigo nuevas formas de ver, analizar y comprender el mundo.

Uno de los ámbitos en los que este cambio ha causado un mayor impacto es, sin lugar a dudas, el de la información. La cantidad y la variedad de datos disponibles, así como la velocidad a la que se producen, 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.

La información es poder. Siempre lo ha sido. Pero es cierto que, con la magnitud que está adquiriendo en los últimos tiempos, se ha convertido en uno de los recursos más valiosos existentes. En palabras de los editores del *Global Information Technology Report 2014* del Foro Económico Mundial, "realmente, hoy en día la información es equivalente al oro o al petróleo".

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 solo el número creciente de personas y empresas que utilizan Internet, sino también los pequeños aparatos conectados (Gantz y Reinsel, 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.

Gantz y Reinsel (2012) estimaron en su informe 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 de el doble hasta alcanzar más de 5.200 gigabytes por cada habitante de la Tierra.

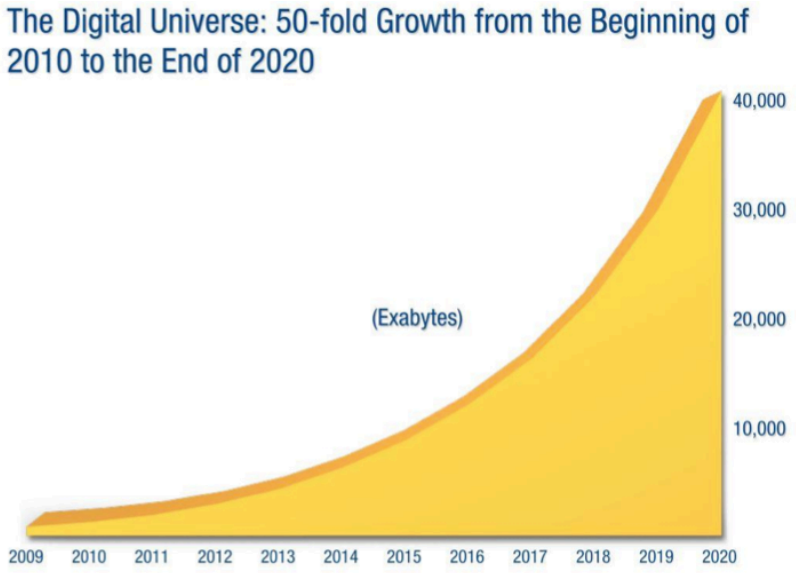


Figura 3.1. Evolución de la cantidad de información disponible en el universo digital
Fuente: Gantz y Reinsel, 2012

En este escenario global de explosión de la información, 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 que sean capaces de analizarla en tiempo real. *Big data* presenta nuevas oportunidades, desafíos y problemas a los que las empresas, las organizaciones y los investigadores debemos hacer frente para conseguir comprender la información y obtener el máximo beneficio de ella. 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 grandes cantidades de dinero para acelerar su investigación y su aplicación. También los medios de comunicación, como *The Ecomist* o *New York Times*, a nivel internacional, o *El País* o *Expansión*, dentro de nuestras fronteras, así como prestigiosas revistas científicas, entre las que encuentran *Nature* o *Science*, se están haciendo eco de las posibilidades que se nos abren gracias a *big data*. Sin lugar a dudas, la era de *big data* ha llegado para instalarse.

En los últimos años, los costes de inversión en un gigabyte han caído drásticamente, con una inversión estimada que va desde los 2,00 dólares a 0,20 dólares por gigabyte desde 2012 a 2020 (Gantz y Reinsel, 2012). Sin embargo la inversión aumenta por parte de las empresas; esto es lo que Gantz y Reinsel (2012) denominan la paradoja del universo digital (Figura 3.2). Actualmente, los costes de almacenaje de un gigabyte para minoristas rondan los 3 céntimos de dólar. Esto nos permite almacenar estas enormes cantidades de información a un precio muy bajo, de manera que la infraestructura del universo digital crecerá más de un 40% durante este período. Para hacernos una idea, en 2011, el espacio necesario para almacenar la música de todo el mundo se podía comprar con 600 dólares (Manyika, Chui, Brown, Bughins, Dobbs, Roxburgh y Byers, 2011). Teniendo en cuenta la reducción de los costes, este precio actualmente se vería muy reducido. En la gráfica siguiente, podemos ver cómo las empresas empiezan a valorar el potencial de este conjunto de bits de información digital, que está en continuo crecimiento, y a invertir –a pesar del descenso en los costes- en tecnologías de *big data*, algoritmos de etiquetado automático, análisis a tiempo real, minería de datos de los medios sociales, etc.

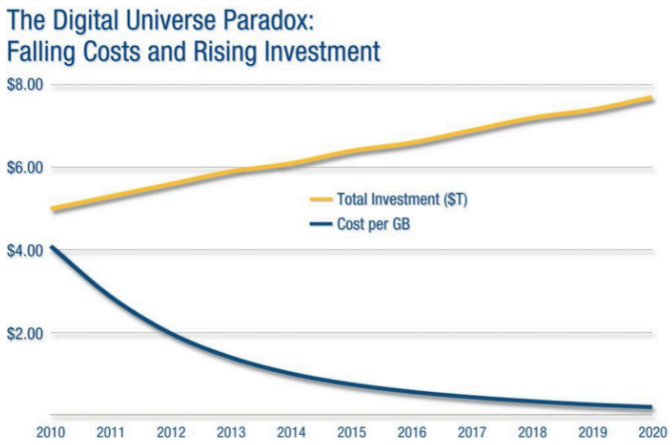


Figura 3.2. Evolución y previsión de costes e inversión e inversión en almacenamiento

Fuente: Gantz y Reinsel, 2012

En el mismo informe de Gantz y Reinsel para IDC, se explica que gran parte del universo digital es efímero, como las llamadas telefónicas que no son grabadas, imágenes de la televisión digital que no se guardan, imágenes de vigilancia que son sustituidas por otras nuevas, información almacenada en los *routers* de manera temporal, etc. A pesar de que estos bits que se quedan sin utilizar se multiplicarán por ocho hasta el año 2020, seguirán suponiendo menos de un cuarto del total del universo digital en esta misma fecha.

Resulta sorprendente el dato de que la información que las personas crean por sí mismas (elaboración documentos escritos, toma de fotografías, descargas de música, etc.) es muy inferior a la cantidad de información que, como consecuencia de ello, se crea en el universo digital acerca de ellas. Esto es lo que los analistas de IDC denominan la "sombra digital". Nuestra sombra digital crece continuamente y cada vez más y, la mayoría de las veces, sin que seamos conscientes de ello. Se construye de información que podemos considerar pública, pero también de aquella perteneciente al ámbito privado. Y es precisamente aquí donde las empresas encuentran las mayores oportunidades de negocio. 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.

La geografía del universo digital también es susceptible de cambios con el paso del tiempo. Mientras que en 2012, más de la mitad de la información se generaba entre EE.UU. (32%) y Europa Occidental (19%), se estima que en 2020 los mercados emergentes (que en 2012 generaban un 36%) sean responsables de un 62% y China pase del 13% al 21% de la producción total (Gantz y Reinsel, 2012).

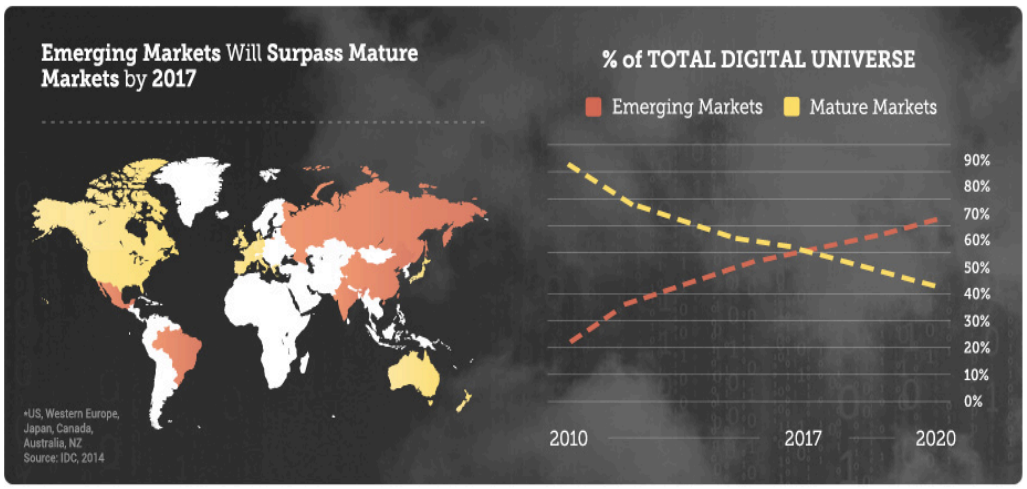


Figura 3.3. Geografía del universo digital

Fuente: IDC, 2014

Por otro lado, Gens (2014), en otro informe para IDC, predice que, en este año, el mercado global de *big data* y los análisis de este 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 volverán mucho más complejas gracias a un aumento en las fuentes de información, en los usuarios y en las aplicaciones, lo que incrementará la ratio de servicios profesionales de la tecnología en un 25% en los próximos cinco años.

En cualquier caso, lo que es indiscutible es que la forma en la que analizamos el mundo para una mejor toma de decisiones está cambiando desde hace tiempo. Por poner solo algunos ejemplos, Google utiliza *big data* para predecir los nuevos brotes de gripe a través de su programa GFT (*Google Flu Trends*); IBM usa la información para optimizar el tráfico en Estocolmo y para obtener una mejor en la calidad de aire (Schaefer, Harrison, Lamba y Srikanth, 2011); el Doctor Jeffrey Brenner, como parte de un programa para reducir los costes el sistema sanitario en una ciudad de Nueva Jersey llamada Camden, utiliza la información de las facturas médicas para diseñar mapas calientes en los que aparecen los casos más complejos y costosos de hospitales; y el *Centro Nacional para la Transformación Académica de Estados Unidos* utiliza la minería de datos para conocer qué alumnos tienen más probabilidades de éxito en la universidad y en qué asignaturas (Parry, 2012). Dentro de nuestras fronteras, el oncólogo cordobés Juan de la Haba, en colaboración con la Sociedad Andaluza de Oncología Médica y la Junta de Andalucía, ha lanzado en enero de 2016 una iniciativa

para prevenir el cáncer basada en técnicas de *big data*. A través de las redes sociales, se ha difundido un cuestionario *online* que permitirá elaborar el primer mapa dinámico de los factores de riesgo de la población. Esta iniciativa, llamada *#TanSolo5Minutos* consiguió registrar más de 23.000 en tan solo veinticuatro horas.

3.2 APROXIMACIÓN AL CONCEPTO

Big data, sin embargo, no es fácil de definir. A pesar de que su importancia parece innegable y es globalmente reconocida, es un concepto abstracto que ha sido definido de diferentes formas y desde distintos puntos de vista.

En general, el término *big data* se refiere a 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.

Autores como Chen, Mao y Liu (2014) ofrecen una definición de *big data* muy similar a esta. Sin embargo, dependiendo de la orientación que se le dé al concepto –económica, científica, teórica, etc.– proliferan distintos enfoques.

En 2011, Manyika et al. (2011: 1) lo definieron para *McKinsey & Company* como "conjuntos de información cuyo tamaño supera la capacidad de las herramientas de *software* de bases de datos tradicionales para capturar, guardar, gestionar y analizar", presentándolo también como "la próxima frontera para la innovación, la competencia y la productividad". Para esta empresa, por tanto, el volumen no es la única característica de *big data*, sino también la creciente velocidad de producción de información que requiere la superación de las tecnologías existentes para su gestión.

En la misma línea aparecen las definiciones de Provost y Fawcett (2013: 54): "conjuntos de datos demasiado grandes para los sistemas de procesamiento de información tradicionales y que, por lo tanto, requieren nuevas tecnologías", Zikopoulos, Eaton, deRoos, Deutsch y Lapis, (2012: 3) para IBM: "información que no puede ser procesada o analizada por medio de herramientas o procesos tradicionales" o Dumbill (2012: 1): "información que excede la capacidad de procesamiento de los sistemas de datos convencionales. La información es demasiado grande, se mueve demasiado rápidamente o no encaja en las estructuras de la arquitectura de tu base de datos. Para obtener valor de esta información, hay que procesarla de una forma alternativa".

En todas estas definiciones, parecidas entre sí, podemos entrever las tres características fundamentales de *big data* que explicaremos más adelante: volumen, velocidad y variedad. Otras empresas, entre las que se incluye IDC –una de las más influyentes del sector- añaden el valor como cuarta característica constitutiva y así lo hacen notar en su definición del término. Así, IDC explicó en 2011 que "la tecnología de *big data* describe una nueva generación de tecnologías y arquitecturas, diseñadas para extraer valor económico de grandes volúmenes de información muy diversa mediante capturas, descubrimiento y/o análisis a gran velocidad" (Carter, 2011: 24). Por su parte, NIST (Cooper y Mell, 2012) se centra más en el aspecto tecnológico del término, afirmando que "podemos hablar de big data cuando el volumen, la velocidad de adquisición o la representación de la información limitan la capacidad para llevar a cabo un análisis efectivo mediante modelos relacionales tradicionales o requieren el uso de escalados horizontales significativos para un procesamiento eficaz".

Para los analistas de IBM (Zikopoulos et al., 2012), no obstante, el término *big data* puede no ser apropiado porque implica o bien que el volumen de información disponible hasta el momento era pequeño –y no lo era–, o bien que el principal reto al que se enfrenta *big data* es el tamaño; y este, aunque es uno de ellos, no es el único. Cada vez más, las empresas se encuentran con enormes cantidades de información desestructurada que se generan como nunca antes se había hecho en la historia y que no les resulta fácil manejar y aprovechar. Con este volumen de información disponible hoy en día, las empresas intentan explotarla y manejarla para adquirir ventaja sobre sus competidores. Estas ingentes cantidades, sin embargo, precisan de ordenadores potentes que sean capaces de gestionarlas y de almacenarlas porque el volumen, la variedad y la velocidad con la que se generan no se pueden manejar de forma manual. Se necesitan también mayores bases de datos que las tradicionales para almacenar la información.

Dejando a un lado las similitudes o las discrepancias en la definición del término, parece claro que podemos encontrar un punto de convergencia entre las distintas opiniones: para un correcto análisis y una adecuada gestión de la información disponible, son necesarios sistemas informáticos potentes que nos ayuden a utilizar el valor de los datos para conseguir beneficios en cualquier ámbito. En palabras de Dumbill (2013: 1), se comparte comúnmente la noción de que debemos "computerizarnos" para una mejora en la toma de decisiones. Según el analista, hasta ahora los ordenadores habían ejercido la función de apoyo al trabajo, una ayuda inestimable como eficaces sustitutos electrónicos de procesos de trabajo con papel o

tareas mecánicas; se habían mantenido como lo que él denomina el "exoesqueleto digital" (Dumbill, 2013: 2). Sin embargo, la irrupción de la web supuso un punto de inflexión a partir del cual los sistemas de información dejaron a un lado su función accesoria para adquirir un papel central dentro de cualquier proceso. Como resultado, la estructura completa de cualquier empresa, organización o centro de trabajo puede y suele consistir en un sistema digital que permite ser analizado en profundidad para comprender el funcionamiento de la empresa y adaptarse a sus necesidades.

Todo esto, sumado a la proliferación de los *smartphones* y de la mecánica física está provocando la desaparición de ese exoesqueleto digital para dar paso a lo que Dumbill (2013: 2) se refiere como "sistema nervioso digital". Por un lado, el incremento masivo del uso de los teléfonos móviles con conexión a Internet y sin cables saca los ordenadores a la calle y fomenta la intercomunicación con cualquier persona o empresa, a cualquier momento y en cualquier lugar, tanto en el mundo real como en el virtual. Las comunicaciones a través de Internet y de aplicaciones móviles involucran a las empresas con los usuarios y les permiten no solo vender sus productos, sino obtener información valiosa acerca de los gustos, las necesidades y la forma de actuar de los consumidores. Por otro lado, el desarrollo de los sensores y de la robótica, unido a los tres vectores en los que avanza la informática (más potencia, menos costes y menos tamaño) posibilitan, gracias al análisis de los datos, la ubicuidad y la disponibilidad de los ordenadores para acciones tan sencillas como la detección de cambios de temperatura o el estado del tráfico.

Parece claro, por tanto, que nos encontramos en una nueva era que está cambiando la manera en la que vivimos y hacemos negocios. Pero conseguir el éxito con *big data* requiere algo más que simplemente información. La obtención de valor basada en la información necesita la identificación de patrones para predecir y tomar decisiones. Es importante decidir qué información utilizar y cómo hacerlo.

No obstante, el mundo de *big data* ha provocado también el surgimiento de algunas preocupaciones, fundamentalmente la de asuntos relacionados con la privacidad de la información. Bilbao-Osorio, Dutta y Lanvin (2014) afirman que, ahora que vivimos en un mundo en el que todo puede ser medido, la información podría convertirse en una nueva ideología, y que nos encontramos en el inicio de un largo viaje en el que seremos capaces de captar, medir y analizar cada vez más datos para tomar mejores decisiones, de forma individual y colectiva, siempre con las técnicas y las pautas adecuadas.

3.3 CARACTERÍSTICAS

Doug Laney, analista de *Gartner* (antiguamente META), presentó en 2001 un modelo de retos y oportunidades que, aunque no se usó originariamente para definir *big data*, aún se sigue utilizando por empresas como Gartner, Microsof o IBM (Chen *et al*. 2014) para caracterizarla. En su informe, definió el modelo de las tres V (o V^3): volumen, velocidad y variedad.

Hoy en día, como hemos apuntado más arriba, estas tres características fundamentales han creado la necesidad de desarrollar nuevas herramientas capaces de optimizar el trabajo para mejorar el rendimiento y obtener el máximo provecho de los dominios de conocimiento existentes, así como la capacidad de actuar sobre ellos.

3.3.1 Volumen

Estamos inundados por la información. La cantidad de información digital generada y almacenada en todo el mundo está reproduciéndose vertiginosamente. Según Zikopoulos et al. (2012), en el año 2000, se llegaron a almacenar 800 mil petabytes (PB) –un petabyte equivale a 10^{15} o mil billones de bytes- y se estima que para el año 2020 este número alcance 35 zettabytes (ZB) –un zettabyte es igual a 10^{21} o mil trillones bytes. Para hacernos una idea, un byte equivale a un carácter escrito y *El Quijote* ocupa unos dos millones de bytes. En 2012, *Twitter* generaba más de 7 terabytes (TB) –un billón de bytes- de información diarios, mientras que, por su parte, *Facebook* alcanzaba los 10 TB; incluso hay empresas que generan terabytes de información cada hora de cada uno de los días del año (Zikopoulos *et al*., 2012). Por nombrar otro ejemplo, solo el 25% de los asistentes a la *SuperBowl XLIX* de 2015 (aquellos que se conectaron a la red Wi-Fi del estadio) retransmitieron 6,23 TB durante el evento deportivo (Diosdado, 2015).

Lo almacenamos todo: información económica, social, medioambiental, sanitaria, política… El simple gesto de usar una tarjeta de crédito, de utilizar una aplicación móvil o de comprar una canción en *iTunes* genera información que se queda almacenada. Disponemos de más información que nunca antes y las organizaciones y empresas que se enfrentan a estas cantidades masivas se ven sobrepasadas. Este es el reto y la oportunidad que presenta *big data*. Con la tecnología adecuada, es posible

analizar toda o casi toda la información para obtener un mayor conocimiento del funcionamiento de cualquier aspecto de la sociedad, de la ciencia o de la empresa.

Los beneficios que pueden producir el procesamiento y el análisis del enorme volumen de información es una de las principales atracciones del *big data*. Una mayor cantidad de datos implica una mejora en los modelos de trabajo. Muchas empresas ya almacenan esta información, pero no tienen la capacidad de procesarla. Por lo tanto, el volumen se presenta como el principal desafío al que se enfrentan las estructuras de las tecnologías de la información. Se requiere no solo almacenamiento escalable (es decir, que maximiza la eficiencia en entornos de almacenamiento masivo), sino también sistemas de acceso distribuido o, lo que es lo mismo, búsquedas en paralelo para aumentar el rendimiento.

La generación masiva de información, sin embargo, no se produce de forma unilateral por los seres humanos. La comunicación denominada *machine to machine* (M2M) o máquina a máquina (Barranco, 2012) también contribuye a las cifras astronómicas de datos que conforman *big data*. Como explica Barranco, esta tecnología M2M dispone de sensores o medidores que capturan algún evento, como velocidad, temperatura, presión, etc. y lo transmiten a otras aplicaciones para que lo conviertan en información. Se estima que estos sensores interconectados alcanzan una tasa anual del 30%. En el siguiente diagrama aparecen representados los tipos de información más comunes que nos podemos encontrar en *big data*:

Figura 3. 4. Tipos de información más comunes en *big data*

Fuente: Barranco, IBM, 2012

1. Web y medios sociales abarca a todo el contenido que podemos encontrar en la web y en las redes sociales.

2. Machine to Machine, como hemos visto antes, se refiere a los sensores o dispositivos que pueden conectarse a otros ordenadores para convertir las señales en información.

3. La información de transacciones está relacionada con los datos de las telecomunicaciones, que se pueden encontrar de forma estructurada o semiestructurada.

4. Biométrica se utiliza en los ámbitos de seguridad e inteligencia e incluye datos biométricos como reconocimiento facial, de retina, huellas dactilares, etc.

5. Información generada por los seres humanos incluye, como hemos señalado anteriormente, documentos electrónicos, llamadas telefónicas, mensajes, compras por Internet, utilización de aplicaciones, etc.

3.3.2.Variedad

El volumen del que hemos hablado trae asociados otros retos distintos para los centros de información que trabajen con él, entre ellos, la variedad, que se refiere a los diferentes tipos de información compleja que se generan. Disponemos de la información tradicional estructurada, pero esta aparece junto con información semiestructurada o no estructurada, proveniente de páginas webs, blogs, correos electrónicos, redes sociales o cualquier otra fuente digital; es decir, no obtenemos información ordenada lista para ser trabajada, sino información sin procesar y sin estructuras relacionales claras, que crece gracias a la proliferación de pequeños aparatos electrónicos o de sensores y que no puede ser gestionada con los métodos tradicionales. Esta información no estructurada supone aproximadamente un 90% del universo digital, según Gantz y Reinsel (2012). Además, las nuevas herramientas diseñadas para estructurar esta información, crean de manera automática información sobre la información o, lo que es lo mismo, metadatos, que, por otra parte, crece el doble que el universo digital. Los metadatos son, por tanto, datos que describen otros datos y que permiten acceder a la información y manipularla. Por ejemplo, en *Twitter*, toda la información adicional que aparece en un tuit, como autor, localización, hora, imágenes incrustadas, etc. se considera metadata; o, en *Spotify*, la clasificación de la música por géneros (jazz, rock, clásica, etc.).

También en la web, donde la relación directa entre ordenadores debería ofrecer garantías, abunda la información no estructurada. Esto se debe a multitud de factores, como la utilización de diferentes buscadores o de diferentes versiones de *software* y, por supuesto, la intervención humana que se requiere en ciertos procesos (Dumbill, 2012). El analista de O'Reilly Radar advierte, sin embargo, de los riesgos de eliminar información durante el proceso de ordenación para su procesamiento porque existe el peligro de romper con uno de los principios de *big data*: siempre que sea posible, hay que guardarlo todo.

Puesto que *big data* se presenta en forma de mensajes, actualizaciones, interacciones en las redes sociales, GPS, señales de teléfonos móviles, etc. (McAfee y Brynjolfsson, 2012). Por tanto, muchas de las fuentes más importantes de *big data* son relativamente nuevas, como las gigantescas cantidades de información procedentes de las redes sociales que, obviamente, tienen la misma antigüedad que ellas. Para hacernos una idea, dos de las redes sociales más importantes en la actualidad, como son *Facebook* y *Twitter*, fueron creadas en 2004 y en 2006. Como consecuencia, conforme la actividad digital, las nuevas fuentes de información y los cada vez más baratos aparatos de almacenamiento crecen, una nueva era se abre ante nosotros en la que Internet, los servicios de localización GPS o las redes sociales –por nombrar algunos– generan torrentes de información en operaciones ordinarias. Sin embargo, todos los datos generados por dispositivos móviles, vídeos, audios, sensores medidores, cámaras, etc. requieren una rápida velocidad de respuesta para obtener la información correcta en el momento adecuado (Barranco, 2012). Cada uno de nosotros, como usuarios, somos generadores de información. Los datos disponibles, generalmente no estructurados, es el gran reto al que se enfrenta *big data* en términos de variedad. En palabras de Peter Norvig, director de investigación de Google: "No disponemos de mejores algoritmos, sino simplemente de más información" (McAfee & Brynjolfsson, 2012: 63).

3.3.3 Velocidad

El tipo de información que manejamos actualmente no solo ha cambiado en términos de volumen y variedad, como acabamos de ver. Hay otro aspecto fundamental que la distingue de la información tradicional: la velocidad. Es decir, la rapidez con la que la información se genera. Zikopoulos *et al*. (2012) señalan, sin embargo, que esta definición de velocidad responde al concepto tradicional que contempla la rapidez con

la que llega la información y a la que se almacena, así como sus correspondientes tasas de recuperación de datos. Desde su punto de vista, aunque reconocen las virtudes de la rapidez en la gestión de la información y admiten que los grandes volúmenes de información que manejamos son posibles gracias a esa velocidad, el concepto debería ir más allá de las definiciones convencionales. Sugieren enfocar el concepto a la idea de información en movimiento, es decir, la velocidad a la que fluye esa información. El creciente uso de pequeños aparatos electrónicos, de aplicaciones, de sensores y de sistemas de transmisión de información, como *Twitter*, hace posible un flujo constante de terabytes de información a una velocidad que resulta imposible manejar con los sistemas tradicionales.

A nivel empresarial, identificar un problema, un suceso o una tendencia segundos antes que la competencia puede ser decisivo. Además, gran parte de la información generada hoy en día tiene una vida muy corta y se elimina fácilmente, con lo que las empresas tienen que analizarla prácticamente a tiempo real para poder obtener beneficio de ella. Por ejemplo, investigadores del MIT, con Alex Pentland a la cabeza, utilizaron la información de localización de los teléfonos móviles para saber cuánta gente había en los aparcamientos del famoso centro comercial estadounidense *Macy's* el día de *Black Friday* (el inicio de la temporada de compras navideñas). Gracias a esto, fue posible estimar las ventas antes de que el propio establecimiento las registrara (McAfee y Brynjolfsson, 2012). Análisis rápidos de este tipo pueden proporcionar ventajas a analistas de bolsa o a empresas del sector.

Por lo tanto, trabajar correctamente con *big data* implica realizar análisis eficaces del volumen y de la variedad de la información, pero siempre con esta en movimiento; esperar a que la información descanse puede suponer el fracaso en términos de obtención de resultados fiables o útiles.

En la figura 3.5 observamos la caracterización de *big data* conforme a estas 3V que acabamos de explicar y que suponen la base de todas las teorías existentes al respecto:

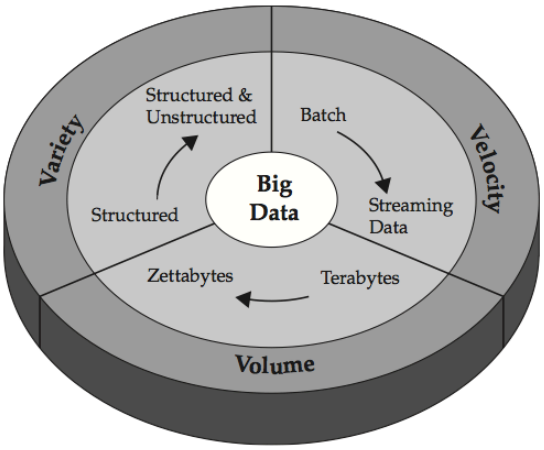


Figura 3.5. Caracterización de V^3 por IBM
Fuente: Zikopoulos *et. al* (2012)

A pesar de que nadie duda de la universalidad de la V^3 (como comúnmente se le conoce) como base para la caracterización de *big data*, no son pocos los autores que añaden otros atributos que también pueden compartir la base entera de esta potencia y simplemente añadirse al exponente. La más importante de todas ellas ya la reflejaba IDC en su definición de *big data* al resaltar la importancia de las nuevas tecnologías para obtener el máximo beneficio económico de la información: el valor. Por ejemplo, y haciendo referencia al concepto anteriormente explicado de sombra digital, lo que el dueño de una librería física puede conocer acerca de sus ventas no va mucho más allá de los libros más vendidos, de aquellos que no interesan al público para su adquisición y, como mucho, de cuáles son las inclinaciones lectoras de cada cliente. Sin embargo, en una librería *online* o con tienda en la web, es posible dibujar un perfil mucho más específico de cada cliente, así como obtener mucha más información. Además de conocer, lógicamente, las compras realizadas por cada uno, también se puede saber cuáles son los libros que han mirado antes o después de la compra, si han tenido en cuenta o no las promociones, si se han interesado por las reseñas de los libros o por los comentarios, el número de veces que visitan la página e incluso qué días de la semana o a qué hora suelen realizar sus compras y realizar así paralelismos entre clientes. Sin duda, la estrategia de negocio se puede ver muy beneficiada si todos estos datos se analizan correctamente.

Esta nueva V se ha reconocido globalmente porque resalta una de las principales necesidades –y oportunidades– de *big data*: la de "explorar los inmensos valores

escondidos" (Chen *et al*., 2014: 171). En una sociedad en la que la información es dinero, la clave está en obtener beneficio a partir de la adecuada extracción de conocimiento de toda la información que se alberga en el universo digital. Se centra, por tanto, en uno de los principales problemas que se presentan, el de cómo descubrir los valores internos que ofrecen los conjuntos de datos a gran escala, de varios tipos y de rápida generación. En palabras de Jay Parikh, subdirector ingeniero de *Facebook*: "Una buena cantidad de información recopilada solo puede no ser big data si no la analizas" (Mayer-Schömberger y Cukier, 2014, *apud* Chen *et al*., 2014: 173).

IBM, por su parte, también añadió posteriormente otra V distinta de valor: la de veracidad. Se refiere al desorden o a la confianza que pueda ofrecer la información, lo que implica aún más dificultades a la hora de controlar la cantidad y la precisión de *big data*. Estudios como el de Higdon *et al*. (2013) ya reconocen la suma de las cinco características como elementos constitutivos de *big data*.

3.4 RETOS DE *BIG DATA*

Por todas las características explicadas hasta ahora, *big data* no solo ofrece grandes oportunidades para el conocimiento, sino también una serie de retos complejos que es necesario superar para abordar una investigación basada en cantidades masivas de información de manera adecuada. Estos retos se relacionan, como ya hemos apuntado, con cuestiones de adquisición, almacenamiento, procesamiento y análisis de los datos.

Torres (2012), los clasifica en cuatro:

-Almacenamiento: la magnitud de los datos disponibles requiere nuevas tecnologías para almacenarlos.

-Bases de datos: las tecnologías existentes no soportan el volumen de información y las bases de datos relacionales tradicionales se quedan atrás. La solución a este problema se encuentra en las llamadas bases de datos NoSQL. En el apartado de almacenamiento de la información explicaremos esto más detenidamente.

-Procesado: también se requieren nuevos modelos de programación para el análisis de la información. Google desarrolló *Map Reduce*, que permite procesar grandes conjuntos de datos y Yahoo! lanzó *Hadoop*, de características similares, pero con un conjunto de herramientas de acceso libre.

-Obtención de valor: los beneficios potenciales de *big data* pueden ser muy numerosos, pero la información masiva por sí sola no es útil; es necesario procesarla para poder realizar un correcto análisis de los datos. Para ello se han desarrollado las técnicas de minería de datos.

Bejerano (2013) califica este último reto de Torres como uno de los retos fundamentales a los que nos enfrentamos: el análisis del inmenso volumen de información del que disponemos, ya que es la única manera de que podamos sacarle provecho a la información. Según Gantz y Reinsel (2013), el panorama del análisis del universo digital en 2012 era bastante desalentador, aunque las previsiones para 2020 son más esperanzadoras (Figura 6):

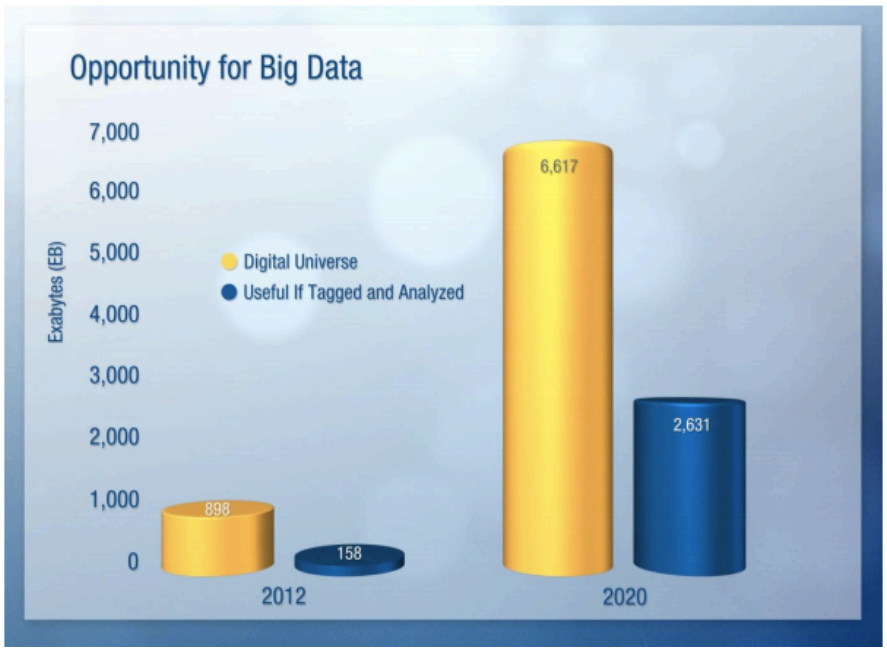


Figura 3.6. Análisis del universo digital

Fuente: Gantz y Reinsel, 2013

En este mismo informe se señala la privacidad como otro de los retos a los que es necesario prestar más atención. Numerosos autores, como Gantz y Reinsel (2011, 2013), Zikopoulos *et al*. (2012), Tole (2013), Bejerano (2013), Hendler (2013) o Chen *et al*. (2014), por nombrar solo algunos ejemplos de ellos hacen especial hincapié en este aspecto. La información es numerosa y de una naturaleza muy variada e incluye no solo la información digital que los usuarios crean, sino también la sombra digital que se dibuja en torno a ellos, de la que ya hemos hablando anteriormente. Los aspectos de seguridad y de privacidad que envuelven a todos estos datos son una de las mayores preocupaciones de las empresas y de los gobiernos. La cantidad creciente de

información que requiere seguridad proviene de dos fuentes principales: empresas (incluyendo a los empleados) y consumidores (Gantz y Reinsel, 2011). En la figura siguiente se muestra los porcentajes de información protegida y de información que necesitaba estarlo en 2012:

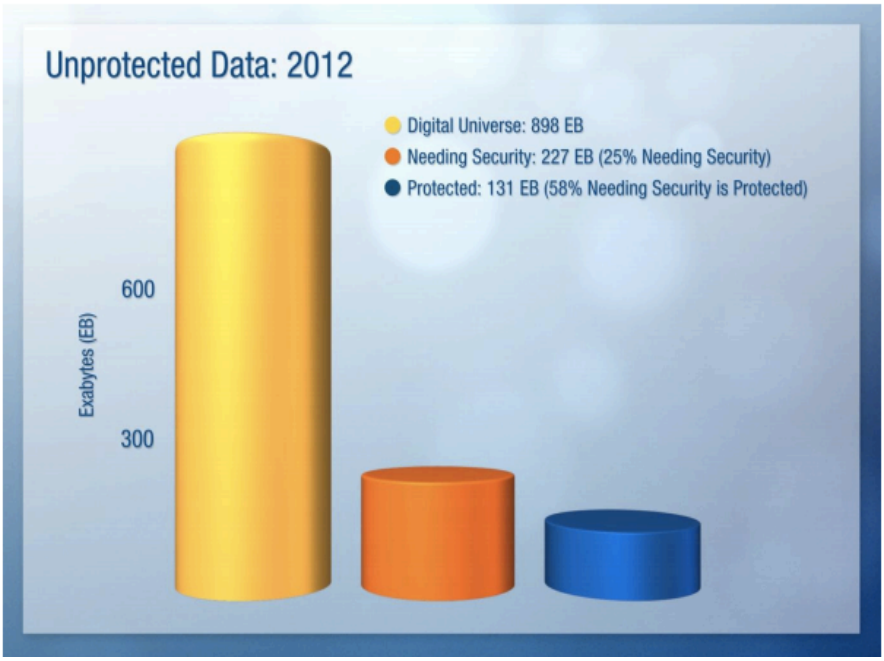


Figura 3.7. Seguridad de la información digital

Fuente: Gantz y Reinsel, 2013

Ammu e Irfanuddin (2013) señalan también la heterogeneidad de la información (también conocida como una de las tres V: variedad), y aseguran que es importante generar de forma automática la metainformación adecuada para describir qué información hay que guardar y cómo se guarda y se cuantifica. También hacen referencia al volumen –lo denomina *escala*– como obstáculo que hay que superar para sacar provecho de la información y a la necesidad a veces inmediata de obtener los resultados de los análisis de la información, algo que puede verse frenado por el enorme volumen del que disponemos.

En un sentido más empresarial, Gantz y Reinsel (2011) explican que la arquitectura de *big data* necesita una renovación completa de los sistemas tecnológicos de las empresas y requiere unas habilidades concretas en los desarrolladores, lo que ha provocado escasez de profesionales cualificados para trabajar en este ámbito. Sin embargo, señalan los analistas de IDC que el mayor reto al que se enfrenta *big data* es el que ellos denominan "reto cultural". Desde su punto de vista, la investigación en este

sentido se encuentra todavía en sus inicios, aunque confían en que, gracias a los avances y al enorme ritmo con el que avanza este campo, los proyectos de *big data* crezcan rápidamente con todas sus consecuencias empresariales, industriales, organizativas y legales. Están convencidos de que las empresas, los gobiernos y las organizaciones se encuentran ante una oportunidad única para el crecimiento y el desarrollo. A pesar de los desafíos, las oportunidades son numerosas y el aumento del universo digital es un potente estímulo para explorar nuevos usos de la información.

3.5 FASES DE *BIG DATA*

Debido a la explosión de *big data*, al valor que se le otorga y a los retos que presenta, existe un creciente interés por esta ciencia que hace que empresas y organismos inviertan grandes cantidades de dinero en su investigación. Esto, junto con el espectacular avance que está sufriendo, provoca que todas las técnicas, herramientas, procesos y sistemas que aquí se presentan de manera general, estén en continuo cambio y varíen de forma considerable con el paso de los meses y dependiendo del tipo de trabajo que se quiera llevar a cabo. Además, es importante señalar que, debido a la naturaleza de nuestra investigación, no hemos considerado oportuno entrar en detalles y especificaciones técnicas, más útiles para un especialista en informática que para los propósitos de nuestro trabajo. Por este motivo, se presenta a continuación una explicación general que pueda ayudar al lector a hacerse una idea completa de las fases que se siguen en cualquier proceso que implique *big data* sin perderse en cuestiones específicas.

3.5.1 Producción

La producción de la información es, como en cualquier ámbito, la primera fase de *big data*. Sin embargo, ya hemos visto que los tipos de información que se generan son muy variados y provienen de diversas fuentes (Barranco, 2012). Los datos generados por la interacción humana son, sin lugar a dudas, una parte esencial del proceso. Gracias a Internet, toda la información obtenida con los motores de búsqueda, las redes sociales, las conversaciones, los foros, etc. se reproduce rápidamente y se queda almacenada y relacionada con la vida diaria de las personas. Pero no podemos olvidar la información que nos proporcionan los sensores, los vídeos, los clicks, etc. Las

empresas, la logística o la investigación científica, entre muchas otras, también son las fuentes que constituyen *big data*. Sin embargo, ninguna caracterización de de este está completa sin mencionar el *Internet of Things* (IoT) –el Internet de las cosas-, que se presenta como otra de las fuentes esenciales de las que se nutre *big data*. El concepto de IoT fue introducido, como él mismo reconoce en un artículo en 2009, por Kevin Ashton, del MIT. Lo acuñó por primera vez en una conferencia que impartió en 1999 en *Procter & Gamble*. La idea de Ashton era muy sencilla: si los ordenadores tuvieran información acerca de los objetos, podríamos conocer todo sobre ellos, si funcionan, si necesitan repararse o cambiarse, etc., los humanos conseguiríamos ahorrar tiempo y esfuerzo porque serían los ordenadores los que harían el trabajo por nosotros.

Weber y Weber (2010: 1), lo definen como:

> Una emergente arquitectura global de información basada en Internet que facilita el intercambio de bienes y servicios. Su propósito es el de proporcionar una infraestructura de tecnología de la información que propicie el intercambio de "cosas" de una forma segura y fiable; es decir, su función es superar la distancia entre los objetos del mundo físico y su representación en los sistemas de información. El IoT servirá para aumentar la transparencia y mejorar la eficiencia de las cadenas de suministro globales.

En otras palabras:

> El IoT se refiere a la interconexión en red de los objetos del día a día que normalmente están equipados con inteligencia ubicua. IoT aumentará la ubicuidad de Internet integrando cada objeto con la interacción vía sistemas incrustados que conducen a una red distribuida de aparatos que se comunican con seres humanos y con otros aparatos. (Xia, Yang, Wang, Vinel, 2012: 1101)

Para *big data*, el IoT consiste en una enorme cantidad de sensores de red incrustados en aparatos y máquinas que los conectan con el mundo real. Estos sensores se encuentran en cualquier ámbito y aportan información de muchos tipos: medioambiental, geográfica, astronómica, logística, etc.

Pepper y Garrity (2014) definen para Cisco las tres características fundamentales del IoT, a saber: comunicación, control y automatización y ahorro de gastos.

Con respecto a la primera, el Internet de las cosas aporta información a las personas y a los sistemas, tales como el estado de los aparatos (si están apagados o

encendidos, cargados o sin batería, etc.) e información de los sensores que pueden monitorizar las constantes vitales de una persona. Antes, en la mayoría de los casos, no teníamos acceso a esta información o teníamos que obtenerla manualmente. Los GPS permiten conocer la localización actual y el movimiento, importante para el seguimiento de medios de transporte, localización de objetos o de personas dentro de una organización.

El control y la automatización son importantes en las empresas y para los consumidores para conocer dónde se encuentran las cosas, o para llevar a cabo acciones de manera remota, encender o apagar aparatos, controlar la temperatura y los equipos de climatización, cerrar el coche, poner la lavadora, etc. o incluso para mandar alertas para notificar anomalías o cambios de funcionamiento.

Por último, la instalación de todo tipo de sensores que conecten los aparatos de las empresas y de los hogares con Internet, permite el ahorro de costes porque prevé situaciones futuras gracias a la información que aporta del estado actual de estos aparatos, lo que haría, por ejemplo, que las empresas no perdieran dinero a la hora de arreglarlos cuando se rompieran, ya que llevarían actualizado su mantenimiento.

Volviendo a la generación de información, el IoT es, por tanto, una parte importante en su producción. La información procedente del mundo empresarial, de las ciudades inteligentes basadas en IoT, de la industria, de la agricultura, del tráfico, de los transportes, de la medicina, etc. construye todo un conjunto de datos útiles para su procesamiento.

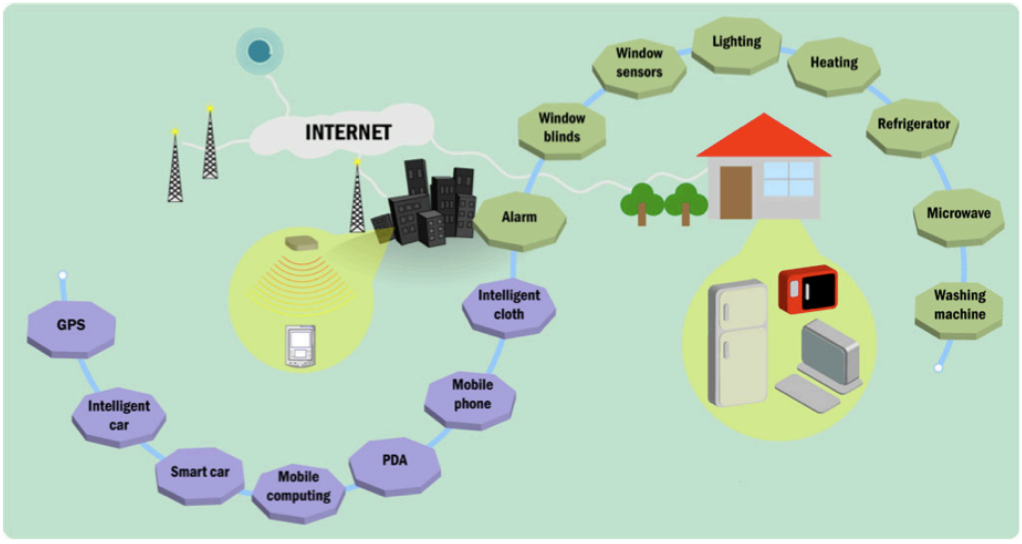


Figura 3.8. Equipamiento de adquisición de información en IoT

Fuente: Chen *et al*., 2014

Por otra parte, Chen *et al.* (2014) también proponen la información bio-médica como una fuente importante de *big data*, que puede relacionarse en algunos puntos con la biométrica que proponía Barranco (2012). Gracias a los avances en biomedicina y el desarrollo de aplicaciones basadas en modelos analíticos y teóricos, se pueden determinar el futuro de esta ciencia y la adopción de los roles necesarios para la elaboración de estrategias industriales relacionadas con la economía nacional, la vida diaria de la gente, la seguridad nacional, la industria farmacéutica, etc. Además, la compleción del *Proyecto Genoma Humano* y las investigaciones derivadas de él también han tenido mucha información que aportar al universo digital. Simplemente una secuencia de un gen humano puede llegar a generar de 100 a 600 GB de información sin procesar. Otras estimaciones de Chen *et al*. (2014) para 2015 indican que el volumen medio de información generado por hospital aumentará de 167 TB a 665 TB. Otros campos como la biología computacional o la física unidos a la proliferación de sensores y de ordenadores en prácticamente todas las áreas de la sociedad están también contribuyendo, como indicaba Barranco, a la producción de estas cifras astronómicas de información.

3.5.2 Adquisición, transporte y preprocesamiento

Durante esta segunda fase, una vez recopilada la información en bruto, es decir, sin procesar, debemos utilizar un mecanismo de transmisión eficiente para enviarla a sistemas de gestión de almacenamiento apropiados para que se lleven a cabo las operaciones de análisis pertinentes. Por otra parte, el preprocesamiento es un paso fundamental para asegurar un almacenamiento y una explotación adecuados de la información. Esto se debe a que la información recopilada incluye muchas veces información redundante o poco útil, que aumenta de forma innecesaria el espacio de almacenamiento requerido y, además, afecta al posterior análisis.

La recopilación de la información puede hacerse de dos formas fundamentales: con sensores conectados directamente a la fuente que produce la información o recibiendo la información ya capturada con anterioridad y generada de otras fuentes de datos, a través de redes privadas o de Internet. En el primero de los casos, como hemos apuntado ya varias veces, los sensores miden todo tipo de aspectos (de tráfico, de voz, de temperatura, de presión, del tiempo, químicos, etc.) que están presentes en nuestra

vida diaria y permiten convertir toda la información física obtenida en información legible. Los datos recopilados se unen en distintos nodos y luego se reenvían a la base para su procesamiento. Los sensores devuelven datos numéricos que se almacenan en lo que se conoce como archivos de registro o *log files*, que contienen grabaciones secuenciales en bases de datos, generadas automáticamente por las fuentes de información. De esta forma, se graban de forma periódica muestras de actividad que luego se usan para caracterizar una actividad concreta dentro del sistema y evaluar posibles nuevos mecanismos que se vayan a usar (Wahab, Mohd, Hanafi & Mohsin, 2008), es decir, se utilizan para su posterior análisis. Un ejemplo de grabaciones de *log files* son los que realizan las páginas web para conocer información acerca de número de visitas, número de clicks, etc.

A través de la red, existen varios métodos para capturar la información, entre los que se encuentran los llamados *web crawlers*, sistemas de segmentación de palabras, sistemas de tareas, de índices, etc. Los *web crawlers* (Cho y García-Molina, 2008) consisten en programas que descargan y almacenan páginas web, generalmente para un buscador. A grandes rasgos, un web *crawler* empieza con la URL de una página inicial para acceder a otras páginas relacionadas. De esta forma, unas vez obtenidas todas las URL, va priorizando y colocando las páginas en cola. El proceso de adquirir y ordenar las URL para posteriormente descargar la página web y almacenarla se produce de manera repetitiva hasta que el *web crawler* se para. Las páginas almacenadas tienen después otras aplicaciones, tales como un motor de búsqueda web o un caché web. Básicamente, son programas que navegan a través de la red y obtienen copia de la información que aparece en ella, generalmente páginas web.

Tras la recopilación de la información sin procesar, la información se transfiere a una infraestructura para su procesamiento y análisis. La mayor parte de la información se almacena en los llamados centros de procesamiento de datos o *data centers*, y es aquí precisamente donde se produce la transmisión de la información, que tiene dos fases: transmisiones inter-DCN (*Dynamic Circuit Network*) y transmisiones intra-DCN. Las primeras de ellas se realizan entre la fuente de información y el centro de procesamiento de datos; las segundas consisten en los flujos de comunicación de información que se producen dentro del centro de procesamiento de datos. Estas últimas dependen del mecanismo de comunicación que posea el centro (chips, memorias internas, placas de conexión, protocolos, etc.)

En lo que al preprocesamiento se refiere, antes de procesar y analizar la información, es necesario pasar por esta fase. El volumen y la variedad de los que ya hemos hablado provocan que los datos recopilados varíen considerablemente entre sí; además, es muy común que presenten ruido, redundancia, inconsistencia, etc. Para evitar que estos factores supongan pérdidas en el almacenamiento en cuestiones de espacio y de dinero y para asegurar un análisis efectivo de la información, es necesario preprocesar la información. Las siguientes técnicas son algunas de las más importantes (Chen *et al*. 2014):

-Integración: consiste en la combinación de la información procedente de las distintas fuentes para que el usuario tenga una visión unificada de esta información (Lenzerini, 2002).

-Limpieza: se trata del proceso de identificación de información inapropiada, incompleta o poco aceptable, para después modificarla o eliminarla, en aras de mejorar la calidad. Según Maletic y Marcus (2000), realizar esta tarea manualmente es prácticamente imposible debido a que conllevaría un gran número de horas, sería extremadamente laborioso y bastante susceptible de error. Se necesitan herramientas potentes para hacerlo de manera automática y para que se obtenga una calidad razonable de la información. Estos dos autores proponen el proceso de limpieza en tres fases: definición y determinación de los tipos de errores, búsqueda e identificación de ejemplos de error y corrección de los errores que hayan quedado sin corregir.

-Eliminación de la redundancia: como su propio nombre indica, permite eliminar información repetida o sobrante que puede desembocar en aumentos innecesarios de costes de transmisión de la información, en defectos en los sistemas de almacenaje o en análisis inadecuados.

3.5.3 Almacenamiento

Hasta el momento, el procesamiento de la información se había limitado a bases de datos tradicionales, conocidas como RDBMS (*Relational database management system*) o bases de datos relacionales. Sin embargo, estas bases de datos no son aptas para el manejo de *big data*. En primer lugar, porque trabajan solamente con información estructurada, pero no con semiestructurada o no estructurada. Por otra parte, no están preparadas para trabajar con tales cantidades de datos, con lo que el proceso se volvería extremadamente lento y, además, tienen costes altos. Para solucionar este problema, la

comunidad científica propone soluciones como la computación en la nube, que ofrece rentabilidad, rapidez, espacio, elasticidad, etc. Para almacenaje más permanente, también son útiles los sistemas de archivo distribuidos, es decir, sistemas que distribuyen la información en varios dispositivos para su acceso en paralelo o concurrente; o los llamados NoSQL (*Not only structured query language*). Estos últimos son bases de datos más evolucionadas que las RDBMS y permiten almacenar grandes cantidades de información no estructurada (generalmente, información documental) para poder realizar operaciones de búsqueda y recuperación de forma eficiente. Algunos de los sistemas de programación más importantes, como *Hadoop*, *MapReduce*, *Cassandra*, *BigTable*, *MongoDB*, *CouchDB* o *SimpleDB* utilizan estas últimas bases de datos.

Con respeto a la computación en la nube (*cloud computing*) es importante señalar que, aunque los costes sean menores, la mayoría de las empresas no disponen de la infraestructura suficiente. A pesar de que las grandes compañías están invirtiendo en sus propios centros de proceso de datos (CPD), las empresas nuevas y las instituciones no cuentan con tales recursos. Para solucionar esto, empresas de gran calado en el mundo de la informática están ofreciendo servicios en la nube mediante los cuales proveen de esa infraestructura de manera remota a las empresas, evitando así que estas realicen esa inversión. Dos de las plataformas en la nube más solicitadas actualmente son *Microsoft Azure* y AWS (*Amazon Web Services*) –la más utilizada hasta el momento-; estas empresas alquilan sus infraestructuras a otras empresas de cualquier tamaño, como la revista *Time*, los laboratorios *Novartis* o *Pfeezer*, *Nokia*, *Adobe* o incluso la NASA. Google, por su parte, también tiene su propia computación en la nube, llamada *Google Cloud Platform*.

3.5.4 Análisis

Se estima que solo la mitad de la información existente es analizada para su transformación en conocimiento (Gantz y Reinsel, 2012). Además, como ya sabemos, la inmensa mayoría de la información existente es no estructurada y generada por máquinas.

El análisis de la información es la etapa final de la cadena de *big data*. Este análisis de resultados juega un papel primordial a la hora de la elaboración de políticas y de planes por parte de los gobiernos, y también para entender la evolución del comercio,

las tendencias de mercado, etc. Los análisis estadísticos aplicados a una gran cantidad de información pueden conducir a aumentos de productividad, crecimiento económico y desarrollo de la sociedad. Por tanto, su propósito es, en último término, obtener valor de esa información y utilizarla de manera apropiada. Los datos por sí mismos no son útiles ni interesantes; solo cuando se usa la información para procesarla es cuando los datos pueden tener un impacto positivo en el funcionamiento de las empresas y en la vida de las personas. El análisis de *big data* implica también métodos analíticos que eran útiles para la información tradicional, pero que aún se ajustan a su arquitectura. El proceso de análisis de la información es un área extensa y compleja que cambia con rapidez. No obstante, mencionamos a continuación algunas de las prácticas más comunes.

En primer lugar, es necesario mencionar los análisis tradicionales de información que utilizan métodos estadísticos para analizar, concentrar, extraer y pulir la información. Estos métodos son también, como acabamos de señalar, los que se utilizan actualmente en *big data*. Algunos de los más comunes son:

-Análisis de agrupación: agrupan objetos y los clasifican según sus características.

-Análisis de correlación estadística: consiste en un método analítico para determinar relaciones, como correlación, dependencia, restricción mutua, etc.

-Análisis de regresión: herramienta matemática para extraer correlaciones entre una variable y varias variables.

-Algoritmos de minería de datos: la minería de datos consiste en extraer conocimiento a partir de información desestructurada, masiva, incompleta y con ruido.

Por otra parte, el análisis puede efectuarse con una doble vertiente: análisis en tiempo real y análisis *offline*. El primero de ellos se utiliza en comercio electrónico y finanzas debido a que la información cambia constantemente y se necesitan análisis rápidos y resultados inmediatos. El análisis *offline* se usa para aplicaciones que no tienen esa necesidad urgente de obtener rápidamente resultados, como el aprendizaje automático, el análisis estadístico, etc.

Existe una amplia variedad de herramientas disponibles para el análisis de *big data*, entre las que podemos encontrar algunas de corte profesional y otras más amateur, herramientas para empresas de altos precios e incluso de *software* libre. Según una encuesta realizada por *KDnuggets* 2014 a 3285 votantes que trabajan con minería de datos, en 2014 aumentó el número de herramientas utilizadas, con una media de un 3,0

en 2013 a un 3,7 el año siguiente. La distancia entre el uso de *software* libres y comerciales se está reduciendo. La encuesta indica que las diez herramientas más utilizadas en 2014 fueron, en orden de utilización: *RapidMiner*, *R*, *Excel*, *SQL*, *Python*, *Weka*, *KNIME*, *Hadoop*, *SAS Base* y *Microsoft SQL Server*.

Las estrategias de análisis de minería de datos y los análisis estadísticos no son nada nuevo en la historia de la investigación informática, sin embargo, continúan siendo unos ámbitos muy activos que se encuentran en una búsqueda constante de nuevos métodos adecuados a la realidad más reciente. Estas técnicas no solo sirven para el análisis de conocimiento estructurado generado en el universo digital, sino también para toda la información semiestructurada o no estructurada. Es aquí y en el inmenso volumen de información disponible donde se están encontrando los mayores retos.

El análisis de información textual es otro aspecto que no podemos olvidar, puesto que se trata del formato de almacenamiento de la información más común (Carbonell, 1992; Chen *et al.*, 2014). La minería de textos es una rama específica de la minería de datos que está relacionada con sistemas de captura de información, aprendizaje automático, estadística, lingüística computacional, etc., e intenta obtener conocimiento de textos no estructurados procedentes de *e-mails*, documentos electrónicos, páginas web y medios sociales. La mayoría de las técnicas de minería de textos están basados en lo que se conoce como PLN (Procesamiento del lenguaje natural). El PLN incluye herramientas de adquisición léxica, desambiguación de palabras, etiquetados del discurso, gramática probabilística, etc. (Manning & Schütze, 1999). Sus principales objetivos, según Carbonell (1999) son: interfaces en lenguaje natural, procesamiento de textos y traducción automática.

Otro campo fundamental de análisis de *big data* es el estudio de la información multimedia, fundamentalmente de imágenes, audios y vídeos, que han ido creciendo a una velocidad vertiginosa y que son una fuente muy rica de información. Se trata también de una información no estructurada, heterogénea y multidisciplinar que está trayendo consigo gran cantidad de investigaciones y de avances para encontrar sistemas de análisis adecuados.

Sin embargo, el área que más está creciendo como fuente de *big data* y que más nos interesa para nuestra investigación es el análisis de la información existente en las redes sociales. En los últimos años, *Twitter*, *Facebook*, *Linkedin*, *Instagram*, *Snapchat*, etc. están ganando popularidad en todos los ámbitos de la sociedad y están generando cantidades masivas de información y de contenido. La naturaleza de cada red social es

diferente y no todas generan el mismo tipo de información ni en la misma cantidad; mientras que algunas, como *Facebook* o *Instagram*, basan gran parte de su contenido (*Facebook*) o prácticamente la totalidad del contenido (*Instagram*) en fotografías, otras, como *Twitter* o *Linkedin* ofrecen mucha más información textual. En general, los contenidos de la información incluyen texto, imágenes y otra información multimedia, como vídeos, hipervínculos, etc. La información asociada a estas plataformas se presenta en forma de estructuras gráficas que describen la comunicación entre dos entidades. El valiosísimo contenido que podemos extraer de ellas para generar conocimiento y las nuevas características emergentes que presentan los datos están planteando, una vez más, desafíos sin precedentes y grandes oportunidades para el análisis de la información. Entre los principales retos que surgen se encuentra, en primer lugar, el crecimiento masivo y continuado de información, que debe ser analizado de forma automática en un tiempo razonable (Chen *et al*., 2014); en segundo lugar, dado que las redes sociales son sistemas dinámicos, hay que tener en cuenta que cambian y se actualizan constantemente. A pesar de que la investigación en el análisis de las redes sociales se encuentra todavía en su fase inicial, no cabe duda de que se trata de una de las fuentes de información más importantes en *big data*. Los analistas también consideran a las redes sociales como una herramienta para promocionar los negocios, gracias a la información que son capaces de transmitir:

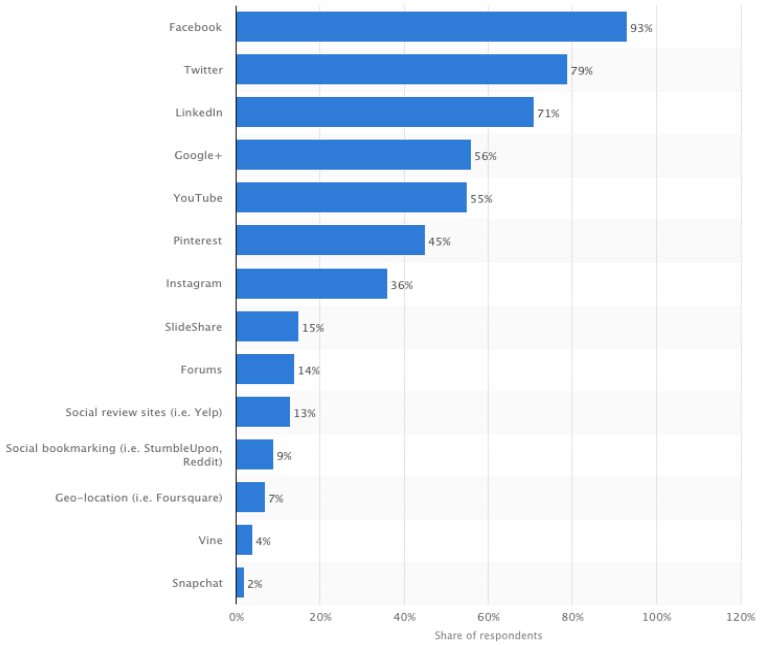


Figura 3.9. Plataformas sociales para promocionar un negocio

Fuente: Statista, 2016

Esta información que aportan las redes sociales está directamente relacionada con la que aportan las aplicaciones móviles, entre otros motivos, porque es en forma de aplicaciones como se instalan en los dispositivos, lo que permite que los usuarios tengan la posibilidad de utilizar la red social en cualquier momento y en cualquier lugar. Por tanto, el análisis de la información presente en las aplicaciones móviles, en cualquier tipo de red social o, en definitiva, de cualquier comunidad de Internet (geográfica, cultural, de ocio, etc.) también supone una fuente de información para su análisis. El incremento de la utilización de teléfonos inteligentes y de descargas de aplicaciones favorecen la proliferación de estas comunidades, que fomentan la interacción entre los usuarios y el intercambio de información en inmensas cantidades y en tiempo real, actualizándose de forma permanente.

3.6 *DATA SCIENCE*

El fenómeno de *big data* está relacionado con la aparición de la denominada *data science* (Dumbill, 2012), una disciplina íntimamente ligada a la informática que también combina las matemáticas con la programación. Patil (2011), no obstante, aclara que, a pesar de que el término *data science* tiene una larga historia, es en el de *data scientist* –científico de datos- donde su compañero Hammerbacher y él creen tener la novedad. Usaremos aquí el término *data science* en inglés por la ausencia de un equivalente exacto en español y para evitar la confusión que pueda producirse con los ámbitos relacionados con el periodismo o las ciencias de la comunicación al ejecutar una traducción literal. Tomando prestada la definición de Dhar (2012: 1), por tanto, nos referimos con *data science* al "estudio de la extracción de conocimiento a partir de la información"; *big data*, afirma, es el combustible que alimenta a *data science*. Un par de años más tarde, en 2014, Dhar vuelve a insistir en la fuerte relación entre ambos conceptos y enfatiza que el hecho de que nosotros podamos "crear conocimiento" a partir de la información y de que ese "nosotros" se refiera a ordenadores, a ordenadores asistidos por humanos o a humanos asistidos por ordenadores es, desde su punto de vista, uno de los motivos por los que *big data* es tan importante.

Volviendo al concepto de *data scientist*, Patil (2011) especifica cuáles son las cualidades que lo caracterizan, a saber: pericia técnica profunda en la disciplina científica, curiosidad para no quedarse en la superficie y descomponer el problema en

una serie de hipótesis que puedan ser probadas, capacidad narrativa para ser capaz de usar la información y comunicar de manera efectiva e ingenio para poder mirar el problema desde perspectivas distintas y creativas.

El hecho de tratarse de una profesión extraordinariamente reciente, sin embargo no le ha impedido a los *data scientists* situarse en el puesto número uno entre las mejores profesiones del mundo. Así lo afirma el equipo de la empresa *Glassdoor* y lo publica el *World Economic Forum* en un estudio sobre los veinticinco trabajos mejor considerados atendiendo a aspectos como las posibilidades que ofrecen para conciliar la vida laboral y familiar, el sueldo y la demanda que tiene. Es curioso también que la mayoría de los oficios que suceden al *data scientist* en la clasificación están relacionados con la adquisición, el análisis, la coordinación, la gestión o el desarrollo de *big data*, como los *SEO Managers* (que se encargan de coordinar la aparición de una web determinada en los buscadores), los gestores de medios sociales, los desarrolladores web y de *software*, los ingenieros de *software*, los analistas de programas y de información, etc. Gartner calcula que en el sector de *big data* hay unos 4,4 millones de trabajadores en todo el mundo (Torres, 2015). Podemos ver a continuación la tabla con la posición que ocupan las distintas profesiones en la clasificación y el salario anual de cada una de ellas en el estudio publicado el 20 de octubre de 2015:

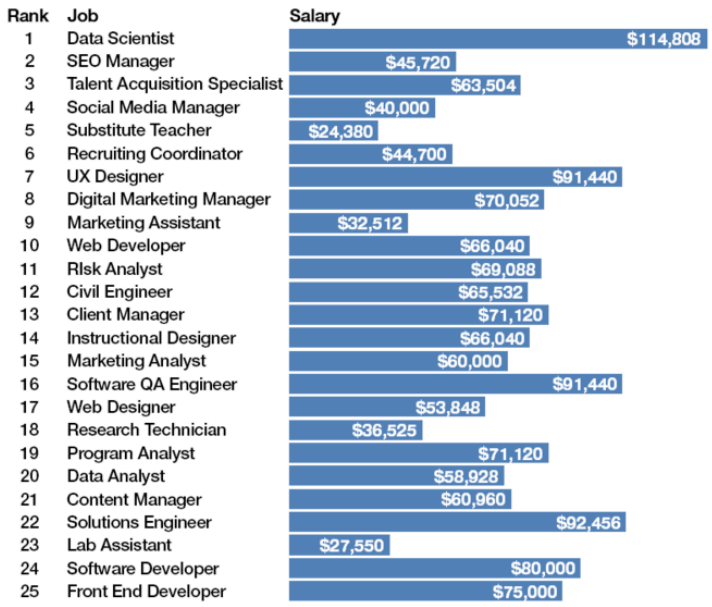


Figura 3.10. Mejores trabajos del mundo según el salario, la conciliación laboral y familiar y su demanda en el mercado. Fuente: Glassdoor, 2015

Al analizar *big data*, para extraer la información, es necesario descomponerla previamente y los analistas deben aprender a comunicar e interpretar los resultados de los estudios. Por otra parte, a la hora de ver si los beneficios de las labores analíticas son analizados por una organización, las capacidad de comunicación y la creatividad son claves. El arte y la práctica de la visualización de la información son cada vez más importantes para salvar el hueco entre humanos y ordenadores y llevar a cabo análisis apropiados.

Provost y Fawcett (2013) también aportan una definición bastante detallada de *data science* en la que la relacionan con el concepto de minería de datos, afirmando que se trata de dos conceptos muy afines. Para ellos, mientras que *data science* es "un conjunto de principios fundamentales que apoyan y guían la correcta extracción de información y conocimiento a partir de los datos", la minería de datos consiste en "la extracción real de conocimiento a partir de los datos mediante las tecnologías que incorporan los principios de *data science*".

No obstante, a pesar de la fuerte relación existente entre los dos conceptos, estos autores explican que *data science* va más allá que los simples algoritmos de minería de datos porque implica que los *data scientists* deben visualizar los problemas de negocios desde la perspectiva de la información. Hay, por tanto, varios campos de estudio tradicionales que se encuentran detrás de esta disciplina, principalmente la estadística. También son importantes otras áreas –en la línea de lo que señalaba Patil–, como son la intuición, la creatividad, el sentido común, etc. En definitiva, *data science* aporta estructuras y principios útiles para resolver el problema de la extracción adecuada de conocimiento a partir de la información.

Los expertos coinciden en que el fin último de *data science* es mejorar la toma de decisiones en cualquier ámbito. Las decisiones que se adoptan basándose en el análisis de *big data* es lo que en inglés se conoce comúnmente como *data-driven decision making* (DDD^3) y aquí lo llamaremos toma decisiones basada en datos.

En un estudio del MIT y de Wharton School (Universidad de Pannsylvania) acerca de cómo afecta el enfoque DDD^3 al rendimiento de las empresas, Brynjolfsson, Hitt y Kim (2011) demuestran que las empresas que llevan a cabo la toma de decisiones basada en datos obtienen mejores resultados financieros. Para ello entrevistaron a ciento setenta y nueve empresas públicas norteamericanas y contrastaron los datos obtenidos con los informes anuales internos y externos. El resultado fue que las empresas que usaban DDD^3 eran, de media, entre un 5% y un 6% más productivas y rentables que sus

competidoras. En el mismo artículo, demuestra cómo una compañía aérea estadounidense de primera línea contrató a una empresa que utilizaba *big data*, llamada *PASSUR Aerospace*, para reducir gastos. La aerolínea se dio cuenta de que la hora estimada de llegada al aeropuerto proporcionada por los pilotos no coincidía con la hora real; de hecho, aproximadamente un 10% de los vuelos llegaban con un margen de más de diez minutos entre la hora estimada y la real, y alrededor de un 30% lo hacían con un margen de cinco minutos. Esto producía pérdidas en dos sentidos. Si el avión llegaba antes de lo previsto y el personal de tierra no estaba preparado, los pasajeros se quedaban virtualmente encerrados hasta que los trabajadores estuvieran preparados para posibilitar la salida del avión. Por el contrario, si el avión aterrizaba más tarde de la hora estimada, el personal de tierra del aeropuerto desperdiciaba valiosos minutos esperando a la aeronave, con las pérdidas económicas que ello conllevaba. Son los pilotos quienes normalmente indican la hora estimada de llegada (ETA) unos minutos antes de la toma de tierra, mientras tienen otros asuntos también importantes que atender relacionados, fundamentalmente, con el aterrizaje. La compañía decidió contratar a *PASSUR Aerospace*, que en 2001 desarrolló un servicio llamado *RightETA*, con el que ofrecía al aeropuerto sus propias estimaciones de la hora estimada de llegada de los aviones, basadas en una combinación de información pública como el tiempo meteorológico y horarios de vuelos. Combinaban esta información con datos propios de la empresa obtenidos fundamentalmente a través de una red de radares instalados en las proximidades del aeropuerto. En 2012, la empresa contaba con 155 instalaciones que recopilaban información de cada uno de los aviones cada cuatro segundos y medio. La información obtenida durante todos estos años se almacena, de manera que la empresa realiza estimaciones de la hora de llegada de los vuelos a partir de un histórico de datos que se basa, esencialmente, en hacer predicciones basadas en lo ocurrido anteriormente. Para hacer la estimación, asemejan, a través de análisis y reconocimiento de patrones, las condiciones de un determinado vuelo a otros casos parecidos. La fórmula es muy sencilla: "usar big data para hacer mejores predicciones, que desembocarán en una mejor toma de decisiones" (McAfee & Brynjolfsson, 2012: 62). Después de contratar a *PASSUR*, la compañía aérea eliminó la diferencia entre la hora estimada de llegada y la hora real y consiguió ahorrar varios millones de dólares al año.

Según Brynjolfsson *et al*. (2011), *big data* como base para la toma de decisiones en la empresa está empezando a sustituir a lo que los investigadores de Microsoft (Kohavi, Longbotham, Sommerfield, & Henne, 2008) denominan HiPPO –*the highest-*

paid person's opinion- o, lo que es lo mismo, la opinión de la persona mejor pagada de la empresa. Este es, como hemos visto, uno de los aspectos en los que más ha influido *big data* dentro de la empresa; en opinión de McAfee & Brynjolfsson (2012), las empresas deben dejar de confiar tanto en la intuición y en la opinión de los altos cargos para dar paso a las garantías que ofrece un análisis de los datos a la hora de tomar decisiones, lo que no quiere decir, apuntan, que *big data* suprima la intervención humana, fundamental, sobre todo, para una correcta interpretación de los resultados obtenidos.

3.7 PANORAMA Y APLICACIONES DE *BIG DATA*

Según un estudio del *Foro Económico M*undial e *INSEAD* (2014) –una de las más importantes y conocidas escuelas de negocios, además de centro de investigación, con sede en Francia– los diez puntos mundiales que más preparados se encuentran para la conexión digital y las tecnologías de la información están dominados por las economías del norte de Europa, los gigantes asiáticos y las economías occidentales más avanzadas. Finlandia, Suecia y Noruega aparecen entre los cinco primeros puestos, liderando así la clasificación de países mejor preparados y con las infraestructuras digitales más potentes, además de fuertes sistemas de innovación. Las grandes potencias asiáticas formadas por Singapur, Hong Kong, la República de Corea y Taiwán figuran entre los mejores preparados y todos, excepto Taiwán, entre los diez primeros. También se sitúan en esta franja Holanda, Suiza, Estados Unidos y Reino Unido, que han reconocido el potencial de las tecnologías de la comunicación y se han sumado a la revolución económica y social invirtiendo en el desarrollo del potencial digital. Centrándonos en Europa, el viejo continente ha estado en la primera línea en cuanto al desarrollo y al fomento del universo digital como estrategia clave para la innovación y la competitividad. Como prueba de ello, un buen número de países europeos se encuentran entre los mejor preparados para las tecnologías de la información y las comunicaciones en red y seis de ellos –Finlandia, Suecia, Holanda, Noruega, Suiza y Reino Unido- se encuentran entre los diez primeros. De hecho, la Comisión Europea ha situado su Agenda Digital como una de las siete iniciativas emblemáticas de la estrategia Europa 2020. No podemos olvidar, no obstante, que, a pesar de todo esto, todavía quedan diferencias considerables entre los distintos países de la Unión Europea, ya que el estudio revela que los países del sur, del este y algunos del centro de Europa

se encuentran más atrás en el grado de desarrollo de innovación y de preparación para las comunicaciones en red. A continuación, exponemos un mapa en el que podemos ver el grado de preparación para las comunicaciones en red que presentan los países a nivel mundial:

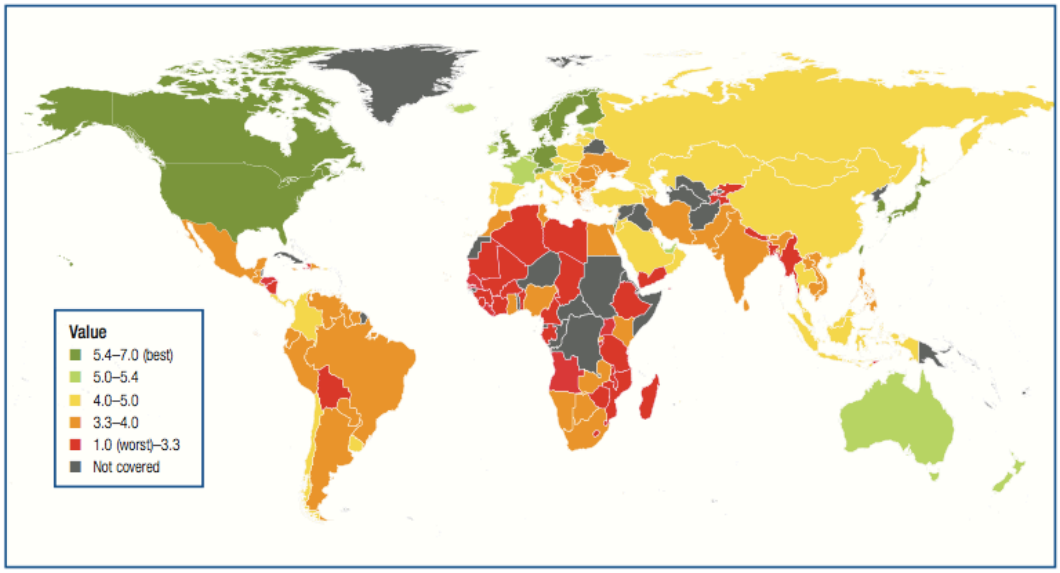


Figura 3.11. Índice de preparación para las comunicaciones en red. Fuente: Bilbao-Osorio, Crotti, Dutta y Lanvin, 2014

En el caso concreto de España, nos encontramos en la posición número treinta y cuatro de la clasificación mundial, lo que demuestra una gran diferencia con respecto a nuestros países vecinos. No obstante, según el mismo informe, tanto el gobierno español, como el portugués (Portugal se encuentra solo un puesto por encima de España), han realizado intentos significativos para aumentar el número de servicios *online*. En España, casi tres cuartos de la población es usuaria de Internet. A pesar de los esfuerzos, el país sigue presentando notables deficiencias en el ámbito de la innovación. Desde el *Foro Económico Mundial*, se propone realizar mayores y mejores inversiones en innovación y en las tecnologías de la comunicación y de la información, necesarias para la transformación económica.

El apoyo de los gobiernos y la elaboración de políticas que favorezcan la innovación y la investigación de *big data* son, por lo tanto, factores fundamentales que deben complementar a la intención de las empresas para trabajar en este campo. Se debe facilitar la elaboración de políticas que favorezcan un ambiente propicio para la viabilidad de negocios basados en *big data* y el desarrollo de medidas educativas para subsanar el déficit de especialistas en este campo. Puesto que *big data* está presente

tanto en organizaciones públicas como privadas, su utilización y su correcta implantación es fundamental para la competitividad económica, cultural y social.

Las empresas, por su parte, deben conocer cuál es su grado de "madurez en *big data*", como indican El-Darwiche, Koch, Meer, Shehadi y Tohme (2014), de Booz y Company. Explican que, aunque es cierto que *big data* tiene el potencial para mejorar y crear nuevas industrias, así como de transformar sectores económicos completos, es importante conocer qué estado de madurez presentan las empresas en términos de *big data* para poder evaluar el progreso e identificar las acciones e iniciativas necesarias para crecer. Para valorar esto, es necesario, primero, una autoevaluación que revise las condiciones en las que se encuentra la empresa para trabajar en *big data*; además, es importante determinar si los gobiernos han aportado los marcos legales necesarios y la infraestructura adecuada. Por otra parte, aunque es indudable la importancia del aspecto técnico, esto no es suficiente para explotar al máximo las posibilidades de *big data*; es fundamental rediseñar la cultura de la empresa a la hora de la toma de decisiones, de manera que estas se basen en los análisis de la información masiva, más que en simples corazonadas. El último nivel de madurez, indican, consiste en la transformación del modelo de negocio para su conversión en un negocio basado en *big data*, pero esto requiere una gran inversión y esfuerzo durante varios años.

En cualquier caso, aseguran Manyika *et al*. (2013: 23) y Gupta (2014: 87), mucho más comedido que los primeros en lo que se refiere a las numerosísimas virtudes atribuidas a *big data*, que la afirmación fundamental es indiscutible: "la información –y las decisiones basadas en ella– constituyen en la actualidad la próxima frontera para la innovación y la productividad".

Es precisamente Gupta (2014) quien, en el *Foro Económico Mundial* admite lo asombroso de las cifras estimadas que calculan los beneficios potenciales que se obtendrían gracias al trabajo con *big data*: en el sector de la sanidad estadounidense alcanzarían los 300 mil millones de dólares al año, mientras que el sector público en Europa llegaría a los 250 mil millones. El campo de la tecnología de almacenamiento no se queda atrás: solo 600 dólares americanos son suficientes para comprar el espacio necesario para almacenar la totalidad de la música del mundo.

También otras empresas, continúa Gutpa, han sacado provecho de las nuevas tecnologías para obtener el valor del que hablábamos antes de *big data*. Por ejemplo, *Visa* anunció recientemente que, gracias al incremento del número de atributos analizados en las transacciones con tarjetas de crédito de 40 a 200, ha conseguido

ahorrar 6 céntimos cada 100 dólares (Gutpa, 2014). El mismo autor, sin embargo, señala que, a pesar de estos datos, hay muchas empresas que todavía quedan lejos de alcanzar un nivel adecuado de análisis de *big data*, fundamentalmente, debido a la falta de inversión. Esta inversión está polarizada entre los sectores que más apuestan por ella, como las telecomunicaciones, los viajes, las ventas, etc. y las que en este sentido se quedan atrás, como la producción o los propios gobiernos.

Este desequilibrio en la financiación de *big data* no evita, sin embargo, que sean numerosos los ejemplos de las aplicaciones que este nuevo recurso encuentra en ámbitos muy variados entre los que se encuentran las ventas, el desarrollo humanitario, la sanidad, las finanzas, las telecomunicaciones y las nuevas tecnologías.

Por nombrar solo algunos de los casos más llamativos, *Amazon* y *Netflix* utilizan una técnica estadística, denominada *filtro colaborativo*, para hacer recomendaciones a los usuarios basadas en los gustos de otros usuarios. Casi dos tercios de las películas seleccionadas por los clientes de *Netflix* provienen de las sugerencias que realiza la página y, gracias a esta técnica, se han obtenido ventas adicionales de millones de dólares (*The Economist*, 2010). Algo parecido ocurre con *EBay*, aunque la empresa que más beneficio obtiene de esta forma es Google, que, gracias al almacenamiento de la inimaginable cantidad de información generada desde sus inicios, había obtenido 170 mil millones de dólares de beneficio cuando solo contaba con doce años de historia.

En el mismo informe, aparecen datos acerca del tiempo que se ahorra en la investigación con las nuevas tendencias en tecnología, como la computación en la nube o los *software* abiertos. Gracias a este tipo de sistemas, el *New York Times* consiguió hace unos años escanear más de cuatrocientas mil imágenes de sus archivos pertenecientes al período de años comprendido entre 1851 y 1922 en solo treinta y seis horas. La compañía *Visa*, por otro lado, tardó trece minutos en procesar 73 mil millones de transacciones registradas durante dos años y que ascendían a 36 terabytes de información. Con métodos tradicionales, la empresa habría tenido que emplear un mes de trabajo.

La oficina ejecutiva del Secretario General de las Naciones Unidas, Ban Ki-moon, ha apostado también por la inversión en la investigación en *big data* con la iniciativa *Global Pulse*, con la que investigan para un futuro en el que *big data* se fomente y se aproveche de manera segura y responsable como un bien público (Kirkpatrick, 2013). Su misión es acelerar el descubrimiento, el desarrollo y la innovación progresiva en este sentido para el desarrollo sostenible y la acción

humanitaria. Esta iniciativa se estableció a partir del reconocimiento de que la información digital nos brinda la oportunidad de entender más profundamente los cambios en el bienestar humano y de conocer en tiempo real si las acciones políticas al respecto están funcionando. Uno de los tres. objetivos fundamentales de esta agencia es reforzar el ecosistema de innovación en *big data* (Global Pulse, 2015). Para ello, *Global Pulse* está apostando por la concienciación de las oportunidades que *big data* ofrece en términos de ayuda y desarrollo, para facilitar alianzas para el intercambio de información pública y privada y para generar herramientas analíticas de alto impacto a través de su red *Pulse Labs* e impulsar así la innovación en las Naciones Unidas.

Podemos concluir que las aplicaciones de *big data* son numerosas y tocan muy diversos ámbitos. Por ejemplo, en el mundo empresarial, la aplicación del análisis de información masiva puede mejorar la eficiencia y la productividad en muchos sentidos, especialmente, en el marketing, ya que las empresas pueden predecir el comportamiento de los clientes y establecer nuevos modelos de negocio. También en la planificación de ventas, ajustando los precios de los productos según la información que tienen; en la cadena de suministros, en las finanzas, en el comercio electrónico y, por supuesto, en la informática y todas las disciplinas relacionadas con ella.

Por otra parte, también es importante mencionar que el Internet de las cosas, del que ya hemos hablado anteriormente, no solo es una fuente para *big data*, sino también uno de los principales mercados de aplicaciones de *big data* (Chen *et al*., 2014). Gracias a la gran variedad de objetos, las empresas de logística han experimentado un enorme crecimiento debido a la necesidad de sensores, adaptadores, GPS, etc. Los mismos autores resaltan las numerosas implicaciones que tiene *big data* en las redes sociales y explican que el análisis de la información contenida en ellas conlleva métodos analíticos para entender las relaciones entre los individuos mediante ciencias como las matemáticas, la informática, la sociología o la psicología, entre otras. Además, la aplicación incluye también analistas de opinión pública, marketing, recopilación y análisis de inteligencia en la red, apoyo para las decisiones gubernamentales, educación en línea, etc. Por ejemplo, el Departamento de Policía de la Ciudad de Santa Cruz, en California, analiza a través de *big data* los crímenes sucedidos hasta el momento e incluso predice los índices de criminalidad en algunas regiones (Mayer-Schönberger y Cukier, 2013).

No podemos olvidar tampoco los beneficios de *big data* en sanidad y en medicina, donde las enormes cantidades de información son una ventaja indiscutible a

la hora de obtener mejores resultados y para favorecer la medicina preventiva, gracias a la obtención de patrones por medio de la estadística.

Podemos concluir que *big data* conlleva un enorme cambio que moldeará el siglo veintiuno (Global Pulse, 2012). Autores como Gray (2009) se atreven a denominarlo "el cuarto paradigma de la ciencia", un nuevo paradigma que trasciende los límites entre la teoría y la práctica.

Es difícil determinar con exactitud la magnitud del cambio que va a conllevar el trabajo con *big data*. En primer lugar, porque no conocemos la naturaleza de la información que se generará en el futuro. En segundo lugar, porque los ordenadores necesitan seguir el ritmo de las exigencias de la velocidad, la variedad y el volumen de la información; hay autores que defienden la idea de que la ley de Moore, mediante la cual el número de transistores de un circuito integrado se multiplica por dos cada dos años, se quedará pronto obsoleta y es necesario dar respuesta al nuevo panorama. Por último, aunque no menos importante, porque es imprescindible aunar fuerzas para la toma de decisiones estratégicas en el futuro que contemplen no solo la gestión con *big data*, sino los posibles usos incorrectos que se puedan hacer de ello.

Si nos preguntamos, sin embargo, cómo podemos aprovechar el inmenso potencial de *big data* en la investigación lingüística, la respuesta es clara. Necesitamos la capacidad y el propósito para trabajar en ello desde la base del reconocimiento de las grades oportunidades, y también de los retos que se nos abren ante nosotros. Por un lado, necesitamos el apoyo institucional y de la comunidad científica para la investigación. Por el otro, no podemos avanzar en este ámbito sin la colaboración transversal con otras áreas científicas imprescindibles en esta tarea, fundamentalmente, la informática.

Big data constituye una oportunidad única para la ciencia de obtener y entender información como nunca antes se había hecho. En palabras de Gantz y Reinsel (2013: 6): "Ningún país, ninguna región ni ninguna empresa puede detener la expansión del universo digital. Solo nos queda prepararnos lo mejor posible".

4
Twitter

4.1 ¿QUÉ ES *TWITTER*?

Las redes sociales, como *Twitter*, *Facebook* o *YouTube*, se han convertido en poderosos medios de comunicación en los que las personas comparten e intercambian información sobre una amplia gama de eventos reales. Estos eventos pueden ser de cualquier tipo, desde eventos públicos que afecten a un gran número de personas (como unas elecciones, un partido de fútbol o un desastre natural) hasta otros a menor escala o personales (una reunión social local, una protesta, un accidente, etc.). Los mensajes cortos que se publican en las redes sociales como *Twitter* reflejan estos eventos conforme ocurren, por lo que el contenido de estos medios es especialmente útil para la identificación en tiempo real de lo que sucede en el mundo y su respuesta social asociada (Becker, Naaman, Gravano, 2011).

Desde que Tim O'Reilly diera su famosa charla sobre una nueva forma de comportamiento humano relacionada con la web en 2004 (O'Reilly, 2005), la web 2.0 se ha convertido en una increíble historia de éxito. Elaborada con contenido generado por los usuarios, el uso de *weblogs*[28], de *wikis*[29] y de *podcasts*[30] ha crecido exponencialmente. Los medios sociales, las redes sociales y las comunidades sociales

[28] Walker (2003) define *weblog* como un sitio web actualizado con frecuencia que consiste en entradas de información en orden contrario al cronológico y cuyo contenido es creado por fundamentalmente por una persona. Antes de conocer el lenguaje HTML, los usuarios no podían escribir ni modificar la WWW, sin embargo, ahora es posible compartir sentimientos, trabajo o conocimiento con todo el mundo y tanto los *weblogs* privados como los públicos están perfectamente establecidos y ampliamente aceptados.

[29] En palabras de Ward Cunningham, el creador de la primera *wiki* en 1994, una *wiki* se define como "the simplest online database that could possible work" (Cunningham, 2002, *apub* Rimmer, 2009: 172). Se trata de un sitio web en el que los usuarios pueden crear, editar, añadir hipervínculos a otras páginas o crear comunidades sociales colaborativas.

[30] Un podcast es un archivo multimedia distribuido en la web mediante un sistema de sindicación RSS, que se utiliza para compartir información a los usuarios que se han suscrito a la fuente de contenidos que genera el podcast. Aunque el término fue introducido por primera vez por un periodista inglés llamado Ben Hammersly (Hammersly, 2004), el pionero en utilizar el servicio de podcasting fue el presentador de MTV Adam Curry para compartir con el público algunos archivos de audio (Pastor, 2009).

representan una nueva forma de colaboración y comunicación. En un corto período de tiempo, la *World Wide Web* ha dejado de ser un medio de información estática para convertirse en una plataforma de comunicación mundial. Esta evolución ha hecho posible que los usuarios sean los que creen, modifiquen, eliminen, tomen parte activa y contribuyan en el contenido de la web, de manera que esta funciona como una plataforma de comunicación y creación de conocimiento.

El concepto de web como se conocía hasta el momento estaba llegando a su ocaso y el surgimiento de la web 2.0 se caracterizó por la creación de más aplicaciones y, sobre todo, de nuevos principios, entre los que destacaban el control de la información por parte de los usuarios, la sustitución de *Netscape Navigator* por Google, el refuerzo de la inteligencia colectiva, el servicio de *blogging* y la sabiduría de las masas, gestión de las bases de datos, utilización de dispositivos electrónicos individuales en lugar de ordenadores, entre otros. La figura que aparece a continuación es el meme que utilizaron los investigadores de O'Reilly Meida para representar lo que para ellos significaba la web 2.0:

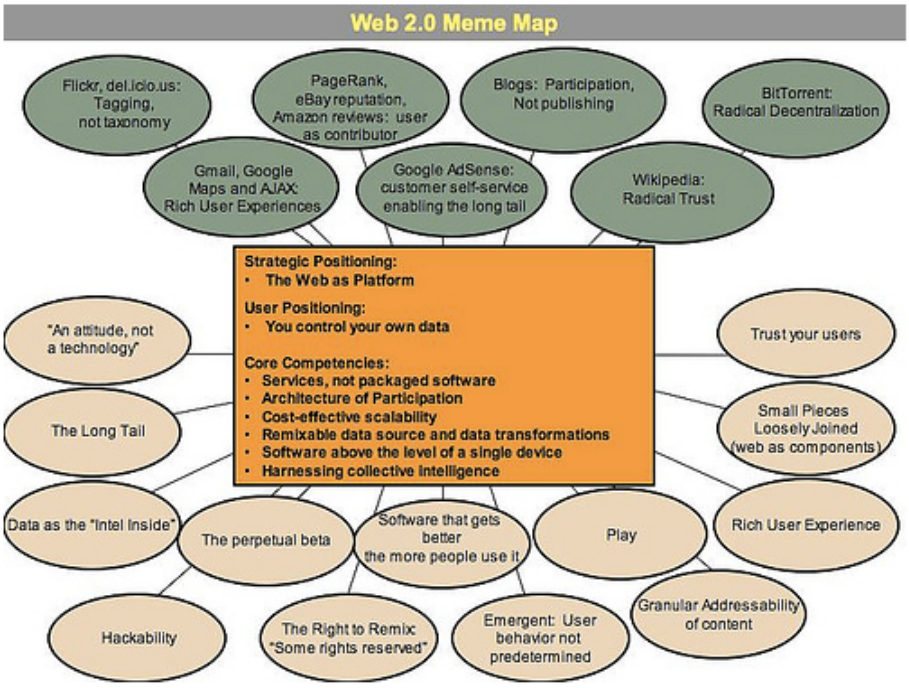


Figura 4.1. Meme desarrollado en 2004 por los investigadores de O'Reilly Media en la Conferencia sobre la web 2.0. Fuente: O'Reilly, 2005

Según algunos autores, como Krishnamurthy, Gill y Arlitt (2008) la web 2.0 ha traído consigo muchas aplicaciones que han provocado la aparición de subconjuntos

arbitrarios de usuarios que se comunican entre sí con un fundamento social. Por ello, las redes sociales *online* se han erigido como la aplicación más popular desde el inicio de la web a principios de los años 90.

La web 2.0, no obstante, ha continuado su evolución hasta llegar en los últimos tiempos a su última versión: la web 3.0. En ella, la información que el propio usuario gestionaba en la web 2.0 se convierte en información masiva, *big data*, que se almacena, fundamentalmente, gracias a la computación en la nube. Así, la información se constituye como una base de datos colosal que permite crear contenidos que puedan ser útiles para distintas aplicaciones y para la Inteligencia Artificial.

Cuando *Twitter* surgió, una nueva forma de *blogs* saltó a escena, conocida como *microblogging*. Y así lo plasmaron Yoon, Elhadad & Bakken (2013: 122) en su definición: "*Twitter* es una red social de *microblogging* con una marca de "¿qué está pasando?" que permite a los usuarios registrarse gratis y publicar posts cortos de 140 caracteres llamados tuits" (*tweets*, en inglés). Todo el mundo se hizo eco de una simple y rompedora idea: la comunicación a través de la web y con mensajes cortos de, como mucho, 140 caracteres.

Para entender el concepto de *blog*, autores como Java, Song, Finin, Tseng (2007: 56) definen *microblogging* de una forma sencilla y expresan que se trata de una nueva forma de comunicación con la que los usuarios tienen la posibilidad de describir su situación actual en *posts* cortos que se distribuyen con mensajes instantáneos, teléfonos móviles o la web. McFedries (2007: 84), por su parte, concreta un poco más y circunscribe el concepto de *microblog* al ámbito de *Twitter*, afirmando que un *microblog* es un *weblog* que se restringe a 140 caracteres por *post* y que está apoyado por las facilidades de las redes sociales. Desde entonces, se ha establecido una polémica sobre si escribir 140 caracteres debería entenderse como una forma de *weblog* o si representa, de hecho, una nueva forma de comunicación.

Los 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 diaria. Los tuits pueden ser publicados y también leídos por otros usuarios, ya sea mediante la aplicación móvil, por 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.

Los tuits son una fuente de información, como decimos, en tiempo real de lo que pasa en el mundo. Según Yoon, Elhadad & Bakken (2013: 122), 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 usuarios o bien tienen seguidores. A diferencia de la mayoría de las redes sociales, como *Facebook* o *MySpace*, estas relaciones de seguidores no requieren reciprocidad. Puede suceder que un usuario siga a otro y que este no lo siga a él. Ser seguidor en *Twitter* significa, por tanto, que el usuario recibe todos los mensajes (tuits) de aquellas personas a las que sigue.

Además, las facilidades que aportan las redes sociales están caracterizadas por la posibilidad de seguir a otra gente, de ser seguidos, de contestar o de enviar mensajes directos. *Twitter* comenzó con la famosa pregunta: "¿qué estás haciendo?" Y otras aplicaciones lo siguieron, como *Jaiku*, *Pownce* o *Plurk*. En cualquier caso, dejando a un lado la herramienta escogida, la publicación de ideas, opiniones o modificaciones rápidas han llevado a una nueva forma de utilizar la web. Quizá uno de los aspectos más poderosos de la plataformas de *microblogging* es su movilidad. Los *microblogs* se pueden escribir o leer mediante interfaces de la web, teléfonos móviles con aplicaciones o a través del navegador, SMS o herramientas de mensajería instantánea. La participación desde cualquier parte del mundo hizo famosa la expresión A^3 (*anytime*, *anywhere*, *anybody*), que cada vez cobra más fuerza. Uno de los primeros estudios científicos (Java, Song, Finin y Tseng, 2007) pretende contestar a la pregunta "¿Cómo usa la gente las plataformas?", y señala que las facilidades del *microblogging* pueden usarse de tres formas: para compartir información (los usuarios de este perfil publican información y suelen tener un gran número de seguidores), para las relaciones de amistad (esta categoría es muy amplia y puede incluir a casi todos los usuarios, incluyendo familiares, compañeros de trabajo y desconocidos) y para buscar información (estos usuarios suelen ser poco activos a la hora de escribir *posts*, pero siguen regularmente a otros). En el mismo estudio, Java *et al*. también identifican varias categorías en cuanto a la intención de los usuarios de *Twitter*. Estas incluyen la de la conversación diaria (los usuarios hablan sobre cualquier evento privado o público y opinan sobre él), el intercambio de información (que también incluye la publicación de URL), la publicación de noticias y la conversación. Ebner y Schiefner (2008: 159)

afirman que el uso del *microblogging* para el rápido intercambio entre personas de intereses similares tiene un gran valor.

La literatura científica ha reconocido desde siempre la importancia de las redes sociales en la difusión de la información (Granovetter, 1973) y de la innovación (Rogers, 2003). Las nuevas tecnologías de las comunicaciones, antes el *e-mail* y, más recientemente, también los medios sociales, han contribuido a favorecer el intercambio de la información y su utilidad en campos tan diversos como la difusión de información (Wu, Huberman, Adamic y Tyler, 2004 o Adamic y Adar, 2005).

El *Darpa Network Challenge* de 2009 demostró la capacidad de las redes sociales *online* para movilizar equipos de personas improvisados a fin de resolver problemas del mundo real, lo que podría mejorar, por ejemplo, la respuesta y la coordinación en desastres naturales o en cualquier otro suceso. Además de haber hecho posible la ubicuidad de las redes sociales, las webs de medios sociales han permitido a los investigadores acceder a grandes cantidades de información para análisis empíricos (Lerman y Ghosh, 2010). Otros autores, como Leskovec y Horviz (2007) o Hogg y Lerman (2009), también coinciden en que esta información es una valiosa fuente de evidencias para el estudio de las redes sociales.

La transmisión de información y de noticias es una de las facetas más importantes de los medios sociales, como hemos visto. Los usuarios publican noticias o enlaces a noticias, las discuten y comparten opiniones en tiempo real. De hecho, a menudo son estas plataformas las primeras en dar una noticia de última hora. Por ejemplo, después de un atentado frustrado que pretendía hacer explotar un avión de *Delta Airlines* en la navidad de 2009, la primera fuente que publicó las nuevas medidas de seguridad para los vuelos internacionales fue *Twitter* (Carr, 2010).

4.2 ORÍGENES

La plataforma de *Twitter* fue creada el 21 de marzo de 2006, cuando el primer tuit de la historia fue publicado por uno de los fundadores de Twitter, Jack Dorsey. El tuit rezaba: "just setting up my twttr" ("estrenando mi twittr").

El 15 de julio de ese mismo año, los creadores de *Twitter* –Dorsey, Evan Williams, Bizz Stone y Noah Glass– lanzaron al público la plataforma. Desde entonces, ha crecido rápidamente y ha ganado una inmensa popularidad, lo que le ha reportado cifras de usuarios y económicas astronómicas. En 2011 ya doblaba en tamaño a la

colección impresa de la Librería del Congreso de Estados Unidos (Hachman, 2011). En 2012, con solo seis años de historia, tenía 140 millones de usuarios activos que publicaban 340 millones de tuits diarios (Twitter, 2012). Dos años más tarde, la página de *Twitter* se encontraba entre las diez páginas web más visitadas (Alexa, sin fecha) y, en mayo de 2015, *Twitter* ha alcanzado más de 500 millones de usuarios y 302 millones de usuarios activos en un mes (Smith, 2015). Esto supone 500 millones de tuits diarios (Twitter, 2015).

El origen de *Twitter* se remonta a una sesión de lluvia de ideas que llevaron a cabo los miembros de *Odeo*, una empresa con sede en San Francisco que había sido fundada por Williams y Glass dos años antes. En aquel momento, la compañía estaba desarrollando un servicio de radio *online* que se vio truncado por el lanzamiento de un servicio similar por parte de *iTunes*, lo que les hizo verse en la necesidad de reinventarse para no caer en el fracaso. Durante aquella reunión, Jack Dorsey propuso la idea de utilizar el servicio de SMS para comunicarse de manera interna entre los miembros de la empresa y poder conocer qué estaba haciendo quién en cada momento.

Al principio, el sistema era utilizado por los miembros de la empresa y los familiares más inmediatos, como explica Dom Sagolla (2011), uno de los trabajadores de la compañía en aquellos momentos. No se le permitía el acceso a ninguna otra empresa de ningún tipo. Durante aquellos meses, *Twitter* se estuvo utilizando en secreto con la versión de prueba (*Alpha*) para mantener al margen a sus competidores; contaba con unos cincuenta usuarios.

Unos cinco o seis años antes, cuando *Twitter* era simplemente un boceto en una libreta del todavía universitario Dorsey, este lo denominó *status* o *stat.us*, aunque reconoce que era simplemente un nombre para referirse a la plataforma. Cuando la idea empezó a tomar forma en *Odeo*, se consideraron varios nombres hasta que se adoptó el definitivo. El propio Dorsey explica, en una entrevista para *Los Angeles Times* (Sarno, 2009), que la intención era capturar en el nombre el carácter de SMS[31] y el hecho de que se pudiera actualizar y recibir desde cualquier sitio; querían transmitir la sensación física de provocar un zumbido en el bolsillo de los amigos, así que la primera opción fue *twitch* ("sacudida"), por el movimiento que hace el teléfono cuando vibra. Sin

[31] Precisamente, el hecho de que se esté abandonando el uso de los SMS para escribir mensajes en *Twitter* está provocando, en los últimos meses, que parte de la Junta Directiva de la empresa se replantee mantener la limitación de los 140 caracteres. Sin embargo, parece ser que la noticia no está siendo bien recibida entre los usuarios de la plataforma, que se encuentran respaldados por un sector importante de la compañía.

embargo, el término no les convenció, y siguieron buscando en el diccionario hasta que finalmente encontraron *Twitter*. Un nombre prácticamente perfecto para los fundadores porque sus dos significados reflejaban exactamente lo que la plataforma significaba para ellos: "una corta ráfaga de información intrascendente", por un lado, y "el gorjeo de un pájaro", por otro.

El paralelismo que Dorsey establecía con los pájaros era el siguiente: los gorjeos de los pájaros no tienen significado para nosotros, sino que son los otros pájaros los que se lo dan. Lo mismo ocurre con *Twitter*: un buen grupo de mensajes puede parecer que no tiene ningún tipo de utilidad ni de significado, pero esto depende totalmente del receptor. Sin embargo, como la pasarela de envío del código corto de SMS era de cinco dígitos, le quitaron las vocales y lo dejaron en "twttr", pero el código pertenecía a otra empresa (10958), así que decidieron utilizar un código fácil de recordar (40404) y devolverle las vocales al nombre. Por este motivo, explica Dorsey en la misma entrevista, "twttr" fue el nombre inicial –de ahí su primer tuit.

Otras fuentes atribuyen el nombre original de *twttr* más directamente a Noah Glass (Carlsonn, 2011), quien, según su compañero, se inspiró en la red social de fotografías *Flickr*, aprovechando también que coincidía con los cinco caracteres de código corto de SMS. Además, el dominio *Twitter.com* ya había sido comprado con anterioridad y no pudieron adquirirlo hasta seis meses más tarde, que fue cuando cambiaron al nombre de *Twitter*.

Según Dorsey, la esencia de *Twitter* ha estado siempre presente en su vida. Desde joven sentía fascinación por saber qué estaba pasando en la ciudad y cómo funcionaba, lo que lo llevó a desarrollar programas que le ofrecieran este tipo de información. El primer obstáculo que se encontró en este viaje fue que, gracias a los GPS, las radios o lo teléfonos móviles, podía conocer cómo se movía la ciudad, pero no sabía nada acerca de la gente; y este fue el verdadero germen de *Twitter*. En un primer momento, sobre el año 2000, desarrolló un servicio llamado *LiveJournal* basado en mensajería instantánea, que permitía tener una lista de amigos y saber lo que hacían o lo que decían que hacían en tiempo real y con la posibilidad constante de actualización, pero se veía limitado por el tipo de dispositivos que se necesitaban para ello, de los que casi nadie disponía No fue, por tanto, hasta la llegada de los SMS, cuando esto fue posible. Este fue el origen de la limitación de los 140 caracteres, ya que la mayoría de los teléfonos móviles están limitados a 160 caracteres por mensaje corto, pero se reservaron 20 caracteres para el nombre de usuario.

Twitter, por tanto, fue concebido originariamente como un servicio móvil de actualización sencillo (Stone, 2009). El objetivo era que la gente se mantuviera en contacto mediante el envío y la recepción de respuestas cortas a la pregunta que aparecía en la interfaz de la aplicación: "¿Qué estás haciendo?". Sin embargo, con el paso del tiempo, los usuarios, las organizaciones y las empresas comenzaron a aprovechar la naturaleza abierta de la red para compartir cualquier otro tipo de contenido, sin tener en cuenta la pregunta original. No quiere decir esto que no hubiera respuestas que contestaran a esa primera pregunta, pero eran mucho más frecuentes las actualizaciones acerca de eventos, noticias de última hora, links, información personal, etc. En opinión de Stone (2009) el modelo abierto de *Twitter* ha creado una nueva red de información superando el concepto de actualización de estatus personal. La red ayuda a los usuarios a descubrir lo que está ocurriendo en tiempo real con las cosas, las personas y los eventos que les importen. Esto fue lo que el 19 de noviembre de 2009 llevó a los creadores a cambiar la pregunta original de "¿Qué estás haciendo?" a "¿Qué está pasando?".

Esta misma idea fue la que desde un primer momento hizo descartar la palabra *watching* ("viendo") y sustituirla por *following* ("siguiendo"). Dorsey (Sarno, 2009) insiste en la idea de que *Twitter* no se concibió como red social, y por lo tanto no había necesidad de llamar a los miembros "amigos" porque no se trataba con ellos personalmente, sino con el contenido que publicaban. De ahí que fuera más apropiado el término *seguir* que *ver*; los usuarios no ven a los demás usuarios, siguen lo que producen.

4.3 DATOS ESTADÍSTICOS

De la misma manera que el contenido que conforma la web ha sufrido un crecimiento exponencial –y también espectacular– provocado también en parte por la cantidad de usuarios que lo utilizan y lo editan diariamente, el número de cuentas registradas en *Twitter* ha seguido la misma tendencia ascendente. En el siguiente gráfico podemos ver cómo aumentó el número de usuarios desde el primer cuatrimestre de 2010 hasta septiembre de 2015:

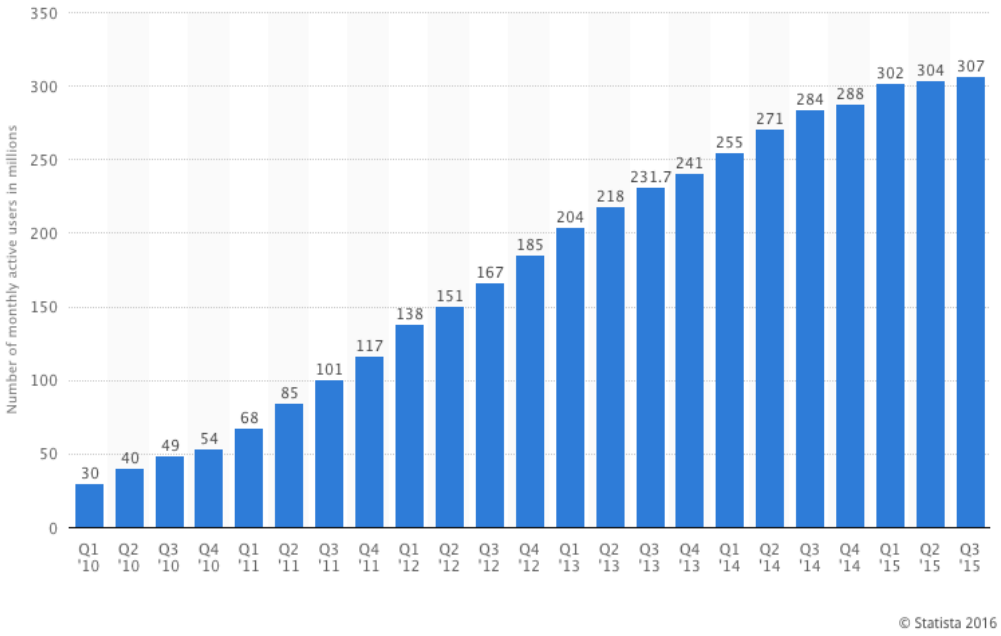


Figura 4.2. Usuarios mensuales activos de *Twitter* en millones desde 2010 hasta 2015

Fuente: Statista, 2016

A pesar de que el crecimiento es más lento en los últimos meses, la previsión se mantiene en la misma línea; se espera, por tanto, que el número de usuarios activos siga aumentando y que sume unos 80 millones más en los próximos dos años:

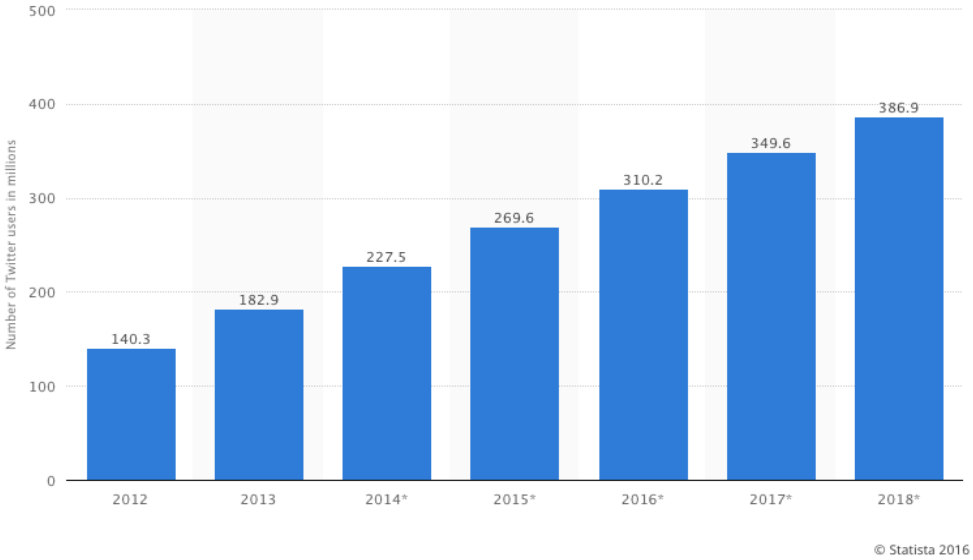


Figura 4.3. Previsión del número de usuarios en millones de *Twitter* hasta 2018

Fuente: Statista, 2016

Obviamente, la distribución de los usuarios entre países e idiomas no se organiza de una manera homogénea. Cuestiones como el número de habitantes de cada país o la conectividad a Internet son factores determinantes a la hora de analizar las estadísticas.

De los casi 387 millones de usuarios previstos para 2018, solo EE.UU. se prevé que alcance unos 66 millones en ese año y que se encuentre rozando los 70 millones de usuarios activos mensuales en 2019. Lo podemos ver a continuación:

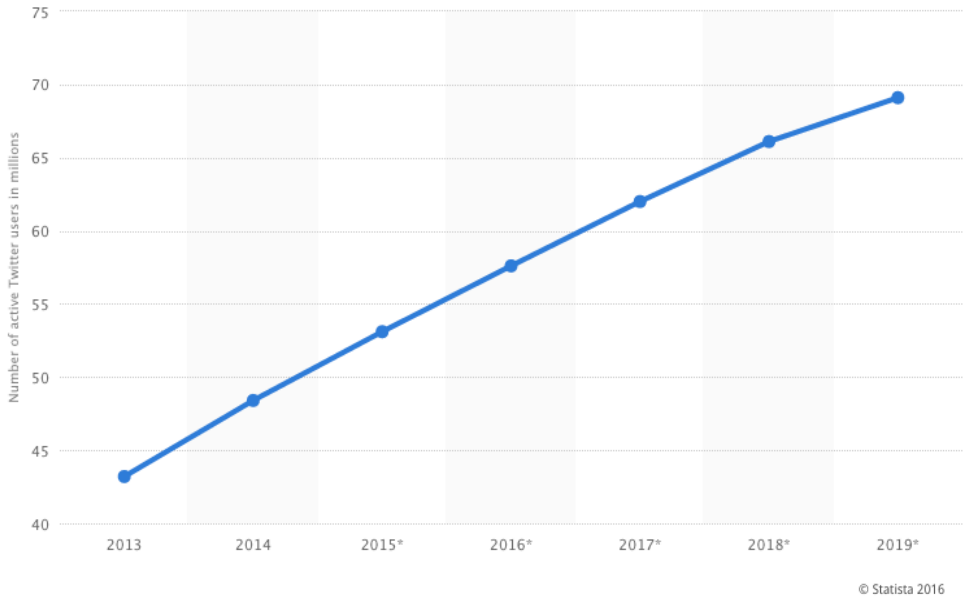


Figura 4.4. Previsión de usuarios activos de *Twitter* en millones en EE.UU. desde 2013 hasta 2019. Fuente: Statista, 2016

Sin embargo, en estas previsiones Estados Unidos no es la región mundial que previsiblemente se encontrará a la cabeza de los países con el mayor número de usuarios. Mientras que en el año 2012 este país se encontraba muy igualado con los países asiáticos, estos están superándolo cada año y se prevé que dentro de dos tengan más del doble que el país americano. El resto de países presentes en la gráfica, por el contrario, se mantienen con una tendencia relativamente estable:

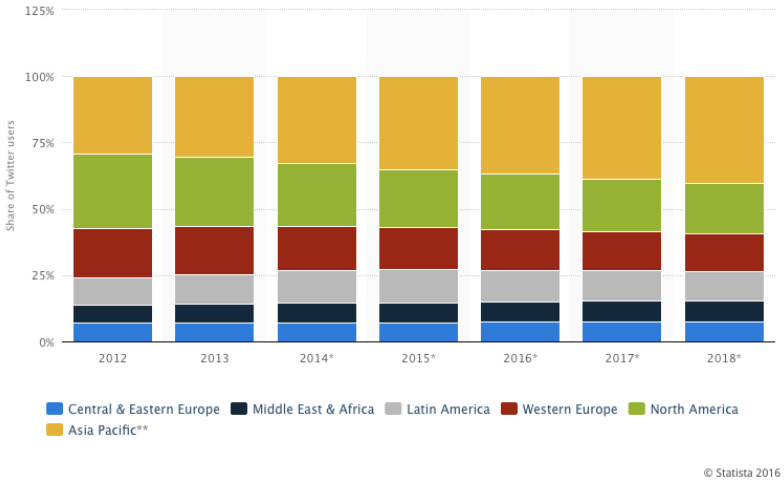


Figura 4.5. Predicción de usuarios de *Twitter* desde 2012 hasta 2018. Fuente: Statista, 2016

En este escenario emergente global, ante el aumento desmesurado de la información, las empresas están cada vez más interesadas en el alto potencial de *big data* en general y de las redes sociales en particular, motivo por el cual, como ya hemos visto, están invirtiendo enormes cantidades económicas para acelerar la investigación en esta línea.

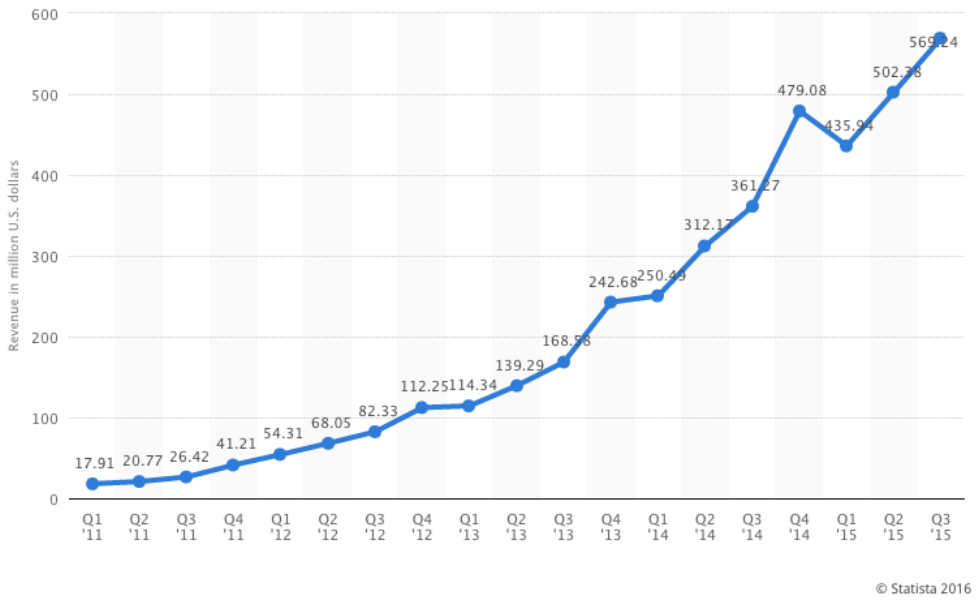


Figura 4.6. Ingresos de *Twitter* en millones de dólares desde el primer cuatrimestre de 2011 hasta el tercer cuatrimestre de 2015. Fuente: Statista, 2016

4.4 APLICACIONES CIENTÍFICAS

La explosión del uso de las redes sociales ha venido acompañada de un incremento de técnicas de minería social para detectar opiniones, tendencias y patrones de consumo. *Twitter*, por su naturaleza abierta, su alto alcance y su constante actualización en tiempo real, ha sido una de las más estudiadas en este sentido. Numerosos estudios demuestran que se trata de una fuente de información que ofrece oportunidades sin precedentes para la investigación en un amplio abanico de ámbitos, que van desde la economía hasta la salud, pasando por la sociología o el desarrollo, entre muchos otros. Por ejemplo, Salathé, Vu, Khandelwal y Hunter (2013) llevaron a cabo un análisis en *Twitter* para entender cómo se puede extender el sentimiento negativo ante las vacunas a través de las comunidades *online*. También UNICEF

publicó un artículo en el mismo año para analizar el sentimiento antivacuna en las redes sociales en Europa. Esta vez, las medios analizados fueron *Facebook*, *Twitter* y algunos foros y blogs. Otros estudios también han demostrado que las redes sociales contienen una sabiduría colectiva que puede ser un indicador de eventos futuros muy preciso y extremadamente poderoso. Por ejemplo, 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 (prestamos, 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:

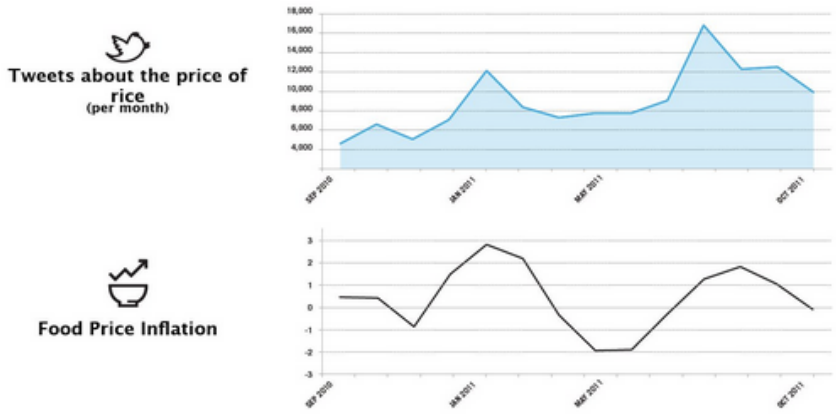


Figura 4.7. Relación entre los tuits referidos al precio del arroz y los precios oficiales de la comida. Fuente: United Nations Global Pulse

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

Otro ejemplo de la utilidad de la información disponible en *Twitter* es el de la previsión de la gripe (Paul, Dredze y Broniatowsky, 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. Se demuestra también 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 Paul *et al*. (2014), la mejora se produce porque GFT realiza una sobreestimación de la prevalencia de la gripe debido a su sensibilidad con los medios de comunicación. Otro de los motivos por los que GFT ha sido criticado ha sido por la falta de transparencia en la información y la poca frecuencia con la que se actualiza. Según Lazer, Kennedy, King y Vespignani (2014), sin embargo, el principal problema es que el análisis de *big data* todavía 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 Paul *et al*. concluye que el número de tuits indicativos de la enfermedad real mantiene una notable coincidencia con los índices de los organismos oficiales de Estados Unidos para el control de las enfermedades (*Centers for Disease Control and Prevention* –CDC en la gráfica), así como el *Departamento de Salud e Higiene* de Nueva York, mientras que el de Lazer *et al*. (2014) demuestra que los análisis de GFT tienen un margen de error considerable con respecto a los CDC.

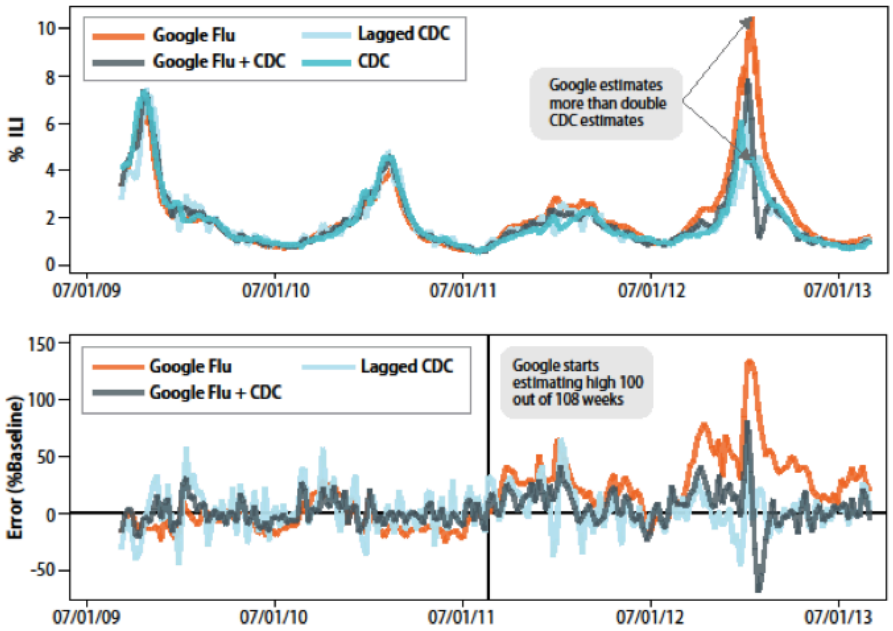


Figura 4.8. Comparación de la estimación de GFT con los índices de los CDC

Fuente: Lazer *et al*., 2014

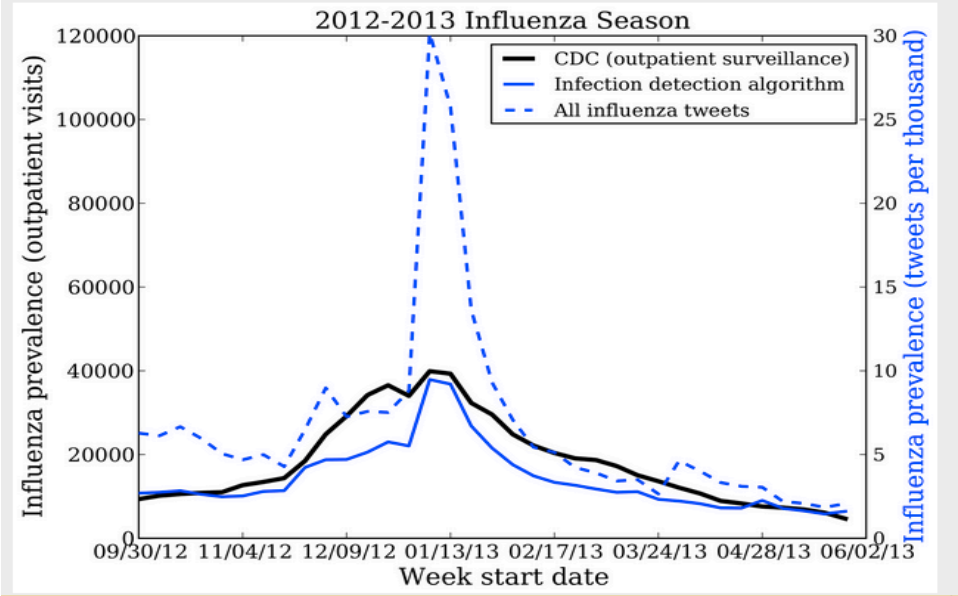


Figura 4.9. Comparación de la estimación de la evolución de la gripe a través de *Twitter* con los CDC. Fuente: Paul *et al*., 2014

El mismo Mark Drezde explica que han desarrollado un método en la Universidad Johns Hopkins de Estados Unidos que proporciona datos reales sobre la gripe y, además, desarrolla un filtro para parar el llamado "parloteo" que se pueda hacer sobre la enfermedad. De esta forma, se pueden evitar *posts* que no contribuyan a medir la cantidad de gente que realmente haya contraído la enfermedad porque simplemente hablan sobre sus temores a contagiarse o mencionan a una figura pública que haya caído enferma. Así, el sistema puede distinguir entre tuits del estilo de "Tengo gripe" y otros como "Estoy preocupado por coger la gripe". Las investigaciones de Drezde demostraron en 2011 que la información de *Twitter* se podía usar para monitorizar enfermedades y para luchar contra ellas de forma más eficaz y, desde entonces, las autoridades sanitarias estadounidenses han patrocinado iniciativas para detectar brotes de enfermedades (Shneiderman, Plaisant y Hesse (2013). Los tuits publicados por los enfermos pueden ayudar a delimitar el alcance y la gravedad de una posible epidemia.

Por otro lado, el Hospital Infantil de Boston ha desarrollado un método para estudiar el perfil de las personas con trastornos del sueño a través de la información obtenida de *Twitter* (Mclver, Hawkins, Chunara, Chatterjee, Bhandari, Fitzgerald, Jain, Brownstein, 2015). En la misma línea, científicos de la Universidad de Michigan han hecho lo propio para investigar la migraña (Nascimento, DosSantos, Danciu, DeBoer, van Holsbeeck, Lucas, Aiello, Khatib, Bender , UMSoD (Under) Graduate Class of

2014, Zubieta, DaSilva, 2014). Por otra parte, Ram, Zhang, Williams y Pegetnze (2015) han demostrado que gracias a *Twitter* es posible hacer una estimación de las personas enfermas de asma que acudirán al hospital, lo que revierte en una mejora del servicio en cuanto a recursos humanos y eficacia del tratamiento. Otros investigadores afirman la viabilidad de utilizar la información contenida en los tuits como método para evaluar y detectar el virus del sida mediante comportamientos de riesgo y su situación en el mapa a través de la localización de los *posts* (Young, Rivers, y Lewis, 2014).

Como es de esperar y ya hemos apuntado unas líneas más arriba, los beneficios que se pueden extraer para la ciencia y su aplicación práctica no solo tienen que ver con cuestiones sanitarias. Sazaki, Okazaki y Matsue (2010), demostraron la utilidad de *Twitter* en la prevención de desastres naturales, como terremotos o tifones. También hay estudios que enumeran las ventajas de su utilización en la educación como herramienta para el fomento del aprendizaje activo de idiomas (Borau, Ullrich, Feng & Shen, 2009) y como plataforma para el aprendizaje en educación superior (Ebner, Lienhardt, Rohs & Meyer, 2009).

Estas nuevas características de la actividad social y de los patrones de comunicación en *Twitter* es lo que Naaman, Boase y Lai (2010: 189) denominan "social awareness streams" (corrientes de concienciación social), que vienen caracterizados por tres factores fundamentales: a) carácter público de la comunicación y de la conversación, b) brevedad del contenido y c) espacio social altamente conectado.

En cualquier caso, no son estos, ni mucho menos, los únicos autores que hablan de la utilidad de *Twitter* para la investigación y el desarrollo. Por nombrar algunos más, otros estudios utilizan nuevas metodologías para utilizar el contenido de *Twitter* en análisis políticos (Bruns y Burgess, 2011), para su aprovechamiento en el ámbito sanitario (Yoon, Elhadad y Bakken, 2013), para obtener mayores beneficios en las grandes empresas (Culnan, McHugh y Zubillaga, 2010).

Sin embargo, a pesar de que algunos trabajos, como el de Kwak, Lee, Park y Moon (2010) resaltan que *Twitter*, gracias a su API[32] 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

[32] Las siglas API responden a la expresión inglesa *Application Programming Interface* (interfaz de programación de aplicaciones) y se refieren a un conjunto de especificaciones informáticas que permiten a los desarrolladores crear programas específicos para otros sistemas operativos. Es decir, estas especificaciones se utilizan como una biblioteca para que las aplicaciones puedan comunicarse entre sí.

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. Como vimos en el segundo capítulo del presente trabajo, las ventajas de utilizar *big data* en la investigación lingüística, es decir, de concebir la web como corpus susceptible de análisis lingüístico, son numerosas y los problemas que esta presenta no son lo suficientemente poderosos como para frenar una tendencia que se va asentando y reforzando paulatinamente, para obtener el máximo potencial que la web nos ofrece y para superar las limitaciones técnicas que encierra. *Big data* ha llegado para quedarse y, probablemente, para seguir evolucionando. Como decía Tognini-Bonelli (2001: 1), la lingüística de corpus cambió la "unidad de moneda" de la investigación lingüística para dar un "salto cualitativo en nuestra forma de entender el lenguaje" (Halliday, 1993: 24, *apud*. Gatto, 2008: IX). Ahora big data lo ha vuelto a hacer; estamos ante la implantación de una nueva unidad monetaria que va a cambiar por completo la manera en la que nos movemos en el mundo y, por supuesto, en la que concebimos la investigación científica.

4.5 FUNCIONAMIENTO EXTERNO 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 dé 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 puede tener acceso a ellos en cualquier momento, incluso sin ser miembro registrado.

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" (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*[33] (el servicio propio de *Twitter*), *tinyurl.com*[34] o *bit.ly*[35]. El funcionamiento de estos servicios consiste en introducir la URL extendida y la herramienta la acorta de forma instantánea y automática, de manera que ocupa un menor número de caracteres.

Los tuits pueden ser contestados, retuiteados (RT + @nombredeusuario + contenido del tuit) o también marcados como favoritos con un icono de un pequeño corazón (hasta noviembre de 2015 se hacía con una 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.

A continuación, mostramos un gráfico con la anatomía de un tuit, en el que podemos ver las partes que acabamos de explicar y que analizaremos pormenorizadamente en las siguientes líneas.

[33] Se puede consultar en https://support.Twitter.com/articles/344713 para más información.
[34] Disponible en: http://tinyurl.com/
[35] Disponible en: https://bitly.com/

Figura 4.10. Anatomía de un tuit. Fuente: Twitter, 2015

4.6 LA UNIDAD DE INFORMACIÓN: EL TUIT

Un tuit es la unidad básica de información con la que *Twitter* trabaja y que va a conformar la base de nuestra investigación. Los miles de millones de tuits manejados por *Twitter* son creados, modificados, almacenados y transmitidos mediante un lenguaje descriptivo denominado JSON. Sin embargo, para poder explicar este lenguaje, realizaremos previamente un recorrido por los dos sistemas de codificación más importantes utilizados hasta la actualidad (ASCII y *Unicode*) y cuyo objetivo ha sido solventar el problema de la internacionalización en Internet.

4.6.1 Sistemas de codificación

Según el *World Wide Web Consortium* (W3C) "la codificación de caracteres refleja la manera en la que el set de caracteres codificados se convierte a bytes para su procesamiento en la computadora" (Ishida, 2010). Mediante la codificación, por tanto, los caracteres de los lenguajes naturales se convierten en símbolos, números u otros sistemas de representación para que puedan ser procesados por el ordenador. Para ello, es necesario establecer unas tablas que indiquen la correspondencia entre el lenguaje natural y el del ordenador, llamadas conjuntos o mapas de caracteres (*charset* o *character map*).

Los tipos más comunes de codificación de caracteres son ASCII y Unicode, aunque Christensson (2010) recuerda que no son los únicos y que existen otros estándares que también se pueden utilizar para la codificación de textos.

4.6.1.1 *American Standard Code for Information Interchange*: ASCII

El conjunto de caracteres ASCII (*American Standard Code for Information Interchange*) fue publicado por primera vez en 1963 por el ANSI (*American National Standard Code for Information*) y se diseñó con solo 7 bits, dejando el octavo para el control de errores por paridad. Con sus posteriores modificaciones e inclusiones de las que hablaremos a continuación, ASCII ha sido el sistema de codificación más utilizado en la *World Wide Web* hasta finales del año 2007, cuando se vio superado por Unicode (Davis, 2008).

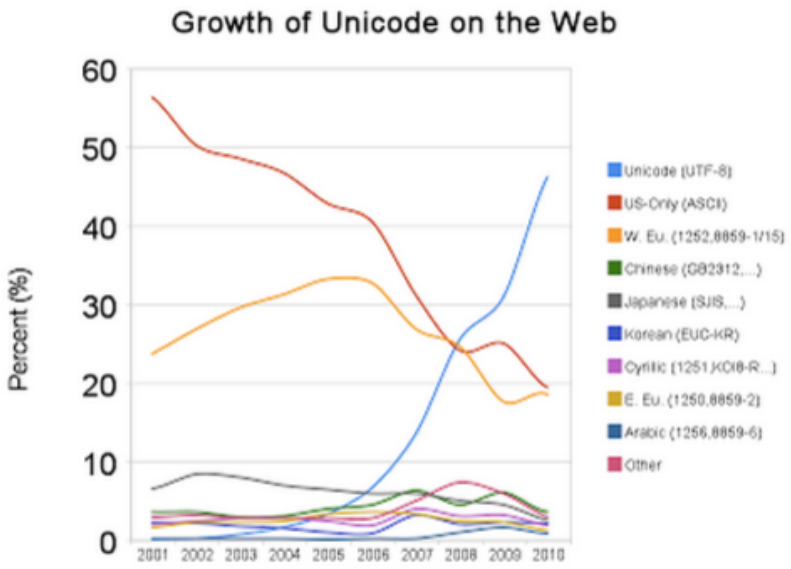


Figura 4.11. Aumento del uso de Unicode (UTF-8) en la Web
Fuente: Davis, Google, 2010

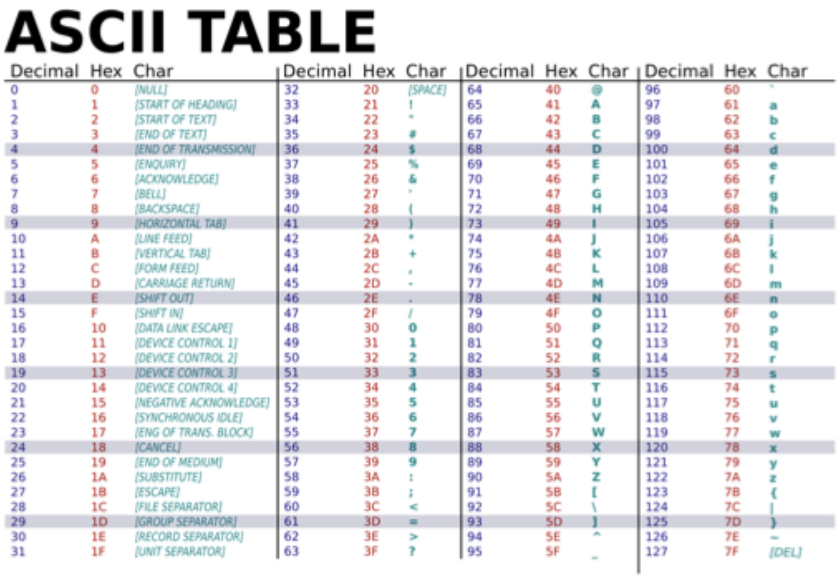
ASCII nació con el desarrollo de códigos telegráficos usados por la compañía Bell y, aunque su primera publicación data del año 63 (Brandel, 99), tres años antes ya se había empezado a trabajar con él. A esta edición le siguió una revisión en 1967 y la última versión en el año 1986 (Jennings, 2004). El 11 de marzo de 1968, el Presidente de los Estados Unidos, Lyndon B. Johnson, publicó un memorándum mediante el que aprobaba y regulaba la utilización de ASCII:

> The adoption of this code as a Federal standard is a major step toward minimizing costly incompatibility among our vast Federal computer and telecommunications data systems... All computers and related equipment configurations brought into the Federal Government inventory on and after July,

> 1969, must have the capability to use the Standard Code for Information Interchange and the formats prescribed by the magnetic tape and paper tape standards when these media are used. The standard code will be used as the basic code in those networks of the Natinal Communications System whose primary function is either the transmission of record communications or the transmission of data related to information processing. (Johnson, 1968)

Como cualquier otro sistema de codificación, ASCII especifica la correspondencia entre patrones de bits y caracteres simbólicos para posibilitar la comunicación de los aparatos digitales entre sí, así como para almacenar y procesar la información contenida en esos caracteres (por ejemplo, la lengua escrita). Con ASCII apareció la primera estandarización de los sistemas de codificación, lo que permitió, entre otras cosas, el nacimiento de la *World Wide Web* (Brandel, 99), algo que no habría sido posible sin una estandarización mundial.

Este sistema se basó originariamente en el alfabeto inglés y codificaba 128 caracteres con una combinación binaria de 7 bits, a la que le añadió un bit más para la corrección de errores. Estos 128 caracteres incluían los números del 0 al 9, las letras de la *a* a la *z*, tanto mayúsculas como minúsculas, algunos signos de puntuación, códigos de control (para el procesamiento de textos: espacios, saltos de página, retroceso, etc.) y el espacio. Con este código no se define ningún mecanismo de formato o estructura de texto, especificados por otros lenguajes, sino simplemente se establece una relación entre caracteres (imprimibles y no imprimibles) y secuencias de bits.



ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

Figura 4.12. Tabla de codificación de caracteres en ASCII. Fuente: Wikimedia, 2015

Sin embargo, 128 caracteres resultaban escasos para un sistema con pretensiones de estandarización mundiales. Como se puede observar en la tabla, por ejemplo, ningún carácter acentuado está incluido, ni tampoco multitud de caracteres y signos ortográficos de cualquier idioma distinto del inglés. Para superar esta limitación, surgió la norma ISO 8859, también conocida como ASCII extendido, que tuvo su primera revisión en 1985. Este nuevo sistema utilizó 8 bits, que admitían 256 caracteres, lo que ampliaba enormemente el espectro y posibilitaba la inclusión de todos los caracteres de un idioma concreto. Se mantuvieron los 128 caracteres primeros del ASCII original y se dejaron los 128 siguientes para los símbolos añadidos (Morales, 2014).

La ampliación, no obstante, a 256 caracteres seguía resultando insuficiente para abarcar todos los alfabetos existentes, lo que llevó a la creación de distintas especificaciones de la norma ISO 8859. Desde 2006, existen 15 especificaciones, después de que se eliminara la número 12. Para los caracteres de Europa occidental, dentro de los cuales, naturalmente, encontramos los del español, se utiliza la norma ISO 8859-15 (también conocida como Latin-9), que es una ampliación de la ISO 8859-1 (Latin-1).

4.6.1.2 *Unicode*

Así las cosas, a pesar de los intentos de estandarización, todavía no era posible utilizar un solo código universal; cada idioma se acogía a su propia especificación. Con la aparición de *Unicode* en 1991 se comenzó a reemplazar los sistemas de codificación de caracteres existentes hasta el momento para subsanar las restricciones en cuanto al tamaño y el obstáculo que suponían en entornos multilingües. *Unicode*, como explica el W3 *Consortium*, surgió como un estándar compuesto por un set de caracteres universal en el que se contemplan los caracteres necesarios para la escritura de la práctica totalidad de los idiomas actuales (Ishida, 2010), así como de otras lenguas muertas, como el griego antiguo, el fenicio o el sumerio, o incluso la escritura cuneiforme. Su diseño se ideó con tres objetivos concretos: la universalidad, para abarcar todos los caracteres posibles; la eficiencia, para que su manejo sea fácil y rápido; y la ausencia de ambigüedad en la correspondencia entre caracteres y códigos (Unicode, Inc., 2015).

El grupo de trabajo que se encarga del mantenimiento constante y de las revisiones del código es el Unicode Technical Committee (UTC). El UTC pertenece al

Unicode Consortium, del que forman parte empresas como Microsoft, Apple, IBM o Google y algunas instituciones y universidades, como el Gobierno de India o la Universidad de Berkeley (Unicode, Inc., 2015). Desde su creación en 1991, Unicode trabaja conjuntamente con ISO/IEC para mantener la sincronización entre estándares (Unicode, Inc., 2007).

La codificación con Unicode, sin embargo, es más compleja que la de otros sistemas de codificación que utilizan correspondencias fijas e invariables, como explica Ushida (2010). Por ejemplo, en la norma ISO 8859-1 (Latin-1), la *A* ocupa el lugar número 65 y el byte que la codifica corresponde al mismo número. Sin embargo, en Unicode, entre otros aspectos, se especifica un nombre y número entero para cada carácter, llamado punto de código, y estos números enteros se pueden representar con diferentes formatos, dependiendo del ordenador; es decir, existen diferentes formatos de codificación para un mismo carácter, que son UTF-8, UTF-16 y UTF-32 (W3 Consortium, 2015).

El primero de ellos y el más utilizado, UTF-8, es de longitud variable –puede ocupar un mínimo de 1 byte (8 bits) y un máximo de 4 bytes. Puede representar cualquier carácter del estándar Unicode y es compatible con el sistema ASCII, por lo que es el formato más utilizado para los correos electrónicos y las páginas web. Los caracteres de ASCII son representados mediante 1 byte (8 bits); los caracteres de otros alfabetos, mediante 2 bytes; y los pertenecientes al plano básico multilingüe, con 3 bytes. Para los caracteres complementarios utiliza 4 bytes.

UTF-16, por su parte, también puede variar en su longitud. Codifica cualquier carácter el plano básico multilingüe con la longitud mínima 2 bytes (16 bits), y los complementarios, con 4 bytes (32 bits). Se usa en los principales sistemas operativos, como Microsoft Windows, Java y .NET.

Por último, UTF-32, al contrario que los dos anteriores, es de longitud fija y utiliza 4 bytes (32 bits) para todos los caracteres. Es el menos utilizado de los tres.

En la siguiente tabla se muestra la codificación de un mismo carácter con los tres sistemas:

	A	א	好	不
Punto de código	U+0041	U+05D0	U+597D	U+233B4
UTF-8	41	D7 90	E5 A5 BD	F0 A3 8E B4
UTF-16	00 41	05 D0	59 7D	D8 4C DF B4
UTF-32	00 00 00 41	00 00 05 D0	00 00 59 7D	00 02 33 B4

Figura 4.13. Codificación de caracteres en Unicode con distintos formato. Fuente: Ushida, 2010

4.6.2 Lenguaje descriptivo JSON

Según *The JSON Data Interchange Format* (Standard ECMA-404) de octubre de 2013, JSON es un formato ligero de intercambio de datos, independiente del lenguaje y basado en el texto que permite la transmisión de información estructurada entre todos los lenguajes de programación. Sus siglas responden a *JavaScript Object Notation* (Notación de Objetos de JavaScript) y presenta una sintaxis basada en llaves, paréntesis, dos puntos y comas fácilmente procesable en la mayoría de contextos. Leerlo y escribirlo no solo es simple para humanos, sino también para que las máquinas lo interpreten y lo generen. Está basado en un subconjunto del lenguaje de programación JavaScript y, a pesar de constituir un formato de texto completamente independiente del lenguaje, utiliza convenciones que son ampliamente conocidas por los programadores de la familia de lenguajes C, incluyendo C, C++, C#, Java, JavaScript, Perl, Python, Ruby (lenguaje de programación de *Twitter*), PHP y muchos otros. Estas propiedades son las que hacen que JSON sea un lenguaje ideal para el intercambio de datos.

En cualquier lenguaje de programación se usa una gran variedad de número: enteros, reales, naturales, imaginarios, etc., que hay programar de manera independiente. Por el contrario, JSON solo ofrece una representación de los mismos números que utilizamos los humanos, es decir, una secuencia de dígitos. Esto permite el intercambio de información porque, a pesar de que el resto de los programas no comparten las mismas representaciones internas para los números, todos entienden la secuencia de JSON. Además, este lenguaje utiliza la conversión de Unicode y eso hace también que sea posible representar información de sets de caracteres de diferentes idiomas.

JSON está constituido por dos estructuras:

- Una colección de pares de nombre/valor. En varios lenguajes esto es conocido como un objeto, registro, estructura, diccionario, tabla *hash*, lista de claves o un vector asociativo.
- Una lista ordenada de valores. En la mayoría de los lenguajes, esto se implementa como vectores, listas o secuencias.

Estas son estructuras universales; virtualmente todos los lenguajes de programación las soportan de una u otra forma. Es razonable que un formato de intercambio de datos que es independiente del lenguaje de programación se base en estas estructuras, que son combinables y recursivas entre sí; es decir, pueden aparecer unas dentro de otras.

En JSON, la información se representa como tipos de datos simples (cadena de caracteres y número) o estructuras de datos (objeto, vector y valor). Para explicar la sintaxis, los datos de valor, cadena de caracteres y número se han dejado para el final porque se ha seguido una metodología que va desde lo más complejo hasta lo más sencillo.

Exceptuando pequeños detalles más técnicos de codificación, estos datos describen completamente el lenguaje JSON.

4.6.2.1 Objeto

Un objeto (*object*) es un conjunto desordenado de pares nombre/valor[36]. Un objeto comienza con { (llave de apertura) y termina con } (llave de cierre). Cada nombre es seguido por : (dos puntos) y los pares nombre/valor están separados por , (coma).

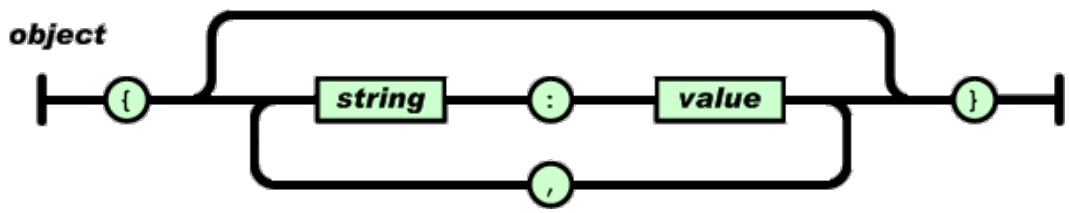


Figura 4.14. Sintaxis de un objeto
Fuente: ECMA International, 2013

[36] Se explica el concepto de valor en la página 151

Para entender esta y el resto de figuras que representan a cada uno de los tipos y estructuras de datos, vamos a explicar, a modo de ejemplo, cuáles serían las tres posibilidades que podrían darse con un objeto:

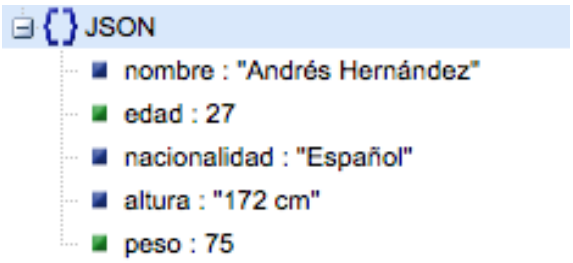
1. Objeto nulo: se da cuando no se escribe nada entre el corchete de apertura y el corchete de cierre. Puede ocurrir, por ejemplo, cuando se le solicita información al sistema y este no devuelve ninguna ocurrencia.
2. Siguiendo la línea central, tendríamos un par nombre/valor. Un ejemplo sencillo de objeto nombre/valor es el siguiente, en el que "texto" es el nombre del atributo como tipo de dato cadena de caracteres o *string*[37] y "El increíble Don Quijote de La Mancha" es el valor del atributo.
3. En lugar de terminar de un solo par, puede haber n-pares porque a través de la tercera línea se pueden crear ciclos indefinidos añadiendo pares nombre/valor al diccionario.

A continuación vamos a ver, primero, un ejemplo de lenguaje JSON para un objeto tipo persona en el que todo lo que está entre comillas y de color azul es una cadena de caracteres, que puede ser el nombre o el valor; la parte negra son tipos de valor número, es decir, son valores numéricos, por lo que van sin entrecomillar.

```
{
"nombre":"Andrés Hernández",
"edad":27,
"nacionalidad":"Español",
"altura":"172 cm",
"peso":75
}
```

Este ejemplo tiene una representación estructurada con diagrama de árbol en la siguiente imagen:

[37] El concepto de cadena de caracteres o *string* se explica en la página: añadir página al final.



4.6.2.2 Vector

Un vector (*array*) es una colección de valores. Un vector comienza con [(corchete izquierdo) y termina con] (corchete derecho). Los valores se separan por , (coma).

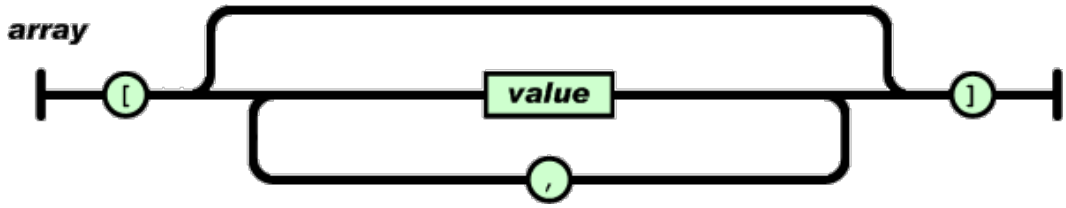


Figura 4.15. Sintaxis de un vector
Fuente: ECMA International, 2013

Las posibles casuísticas que pueden darse con el vector, como podemos ver, son las mismas que las tres que hemos explicado con el objeto. Un ejemplo de vector de números podría ser este:

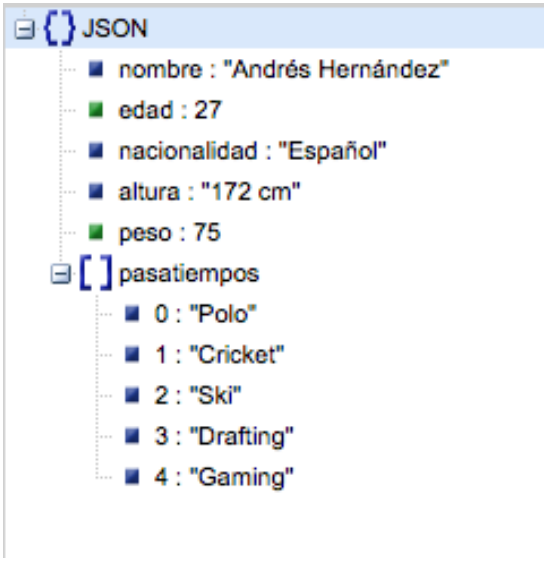
```
[       10,
        20,
        25,
        42      ]
```

Si al ejemplo en JSON del objeto tipo persona del apartado anterior le añadiéramos un vector o una colección de valores referida a los pasatiempos de Andrés Hernández, se vería del siguiente modo:

```
{
"nombre":"Andrés Hernández",
"edad":27,
"nacionalidad":"Español",
"altura":"172 cm",
"peso":75,
"pasatiempos":["Polo","Cricket","Ski","Drafting","Gaming"]
```

}

La parte que se ha añadido de pasatiempos es un vector o un conjunto de valores, en este caso, los valores son cadenas. En un diagrama, quedaría representado así:



4.6.2.3 Valor

Un valor (*value*) puede ser una cadena de caracteres[38] con comillas dobles, un número, un objeto, un vector o valores cuyo resultado solo puede ser verdadero, falso o nulo (*true*, *false*, *null*). Como se puede observar, estas estructuras pueden anidarse.

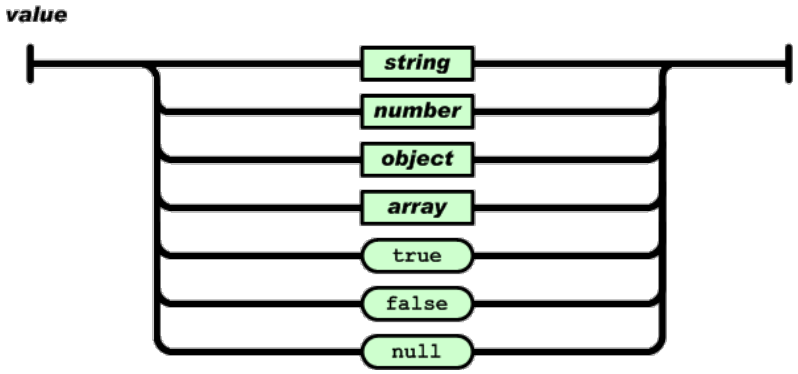


Figura 4.16. Sintaxis de un valor
Fuente: ECMA International, 2013

[38] El concepto de cadena de caracteres se explica en la página 153.

Un ejemplo de lenguaje JSON para el objeto tipo persona anterior, después de añadirle el vector de pasatiempos y más elementos de objeto nombre/valor del diccionario, pero con un tipo de valor binario (*true*, *false* o *null*), además un objeto anidado.

```
{
"nombre":"Andrés Hernández",
"edad":27,
"nacionalidad":"Español",
"altura":"172 cm",
"peso":75            ,
"pasatiempos":["Polo","Cricket","Ski","Drafting","Gaming"],
"soltero":true,
"dirección":{
    "calle":"Ave. Siempre Viva",
    "número":"123",
    "país":"México"
    }
}
```

Aquí se puede ver cómo valores como "soltero", por ejemplo, son valores binarios porque solo admiten verdadero o falso. La calle, el número y el país son objetos anidados.

Su representación en diagrama de árbol quedaría así:

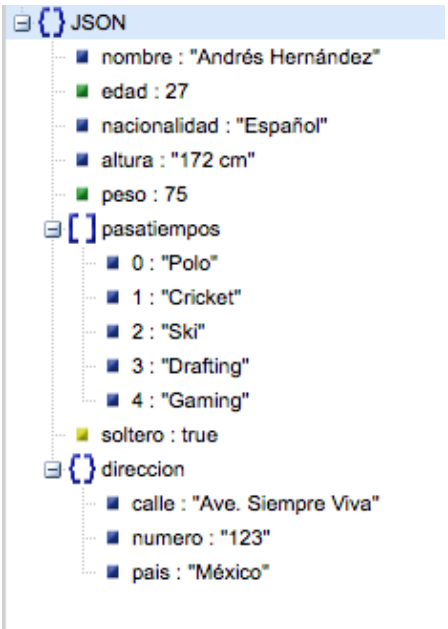


Figura 4.17. Diagrama de árbol del ejemplo en JSON.

4.6.2.4 Cadena de caracteres

Una cadena de caracteres (*string*) es una colección de cero o más caracteres Unicode, encerrados entre comillas dobles, usando barras divisorias invertidas llamadas código escape para introducir caracteres especiales, como intro, tabulador, caracteres específicos de cada idioma, emoticonos, etc.

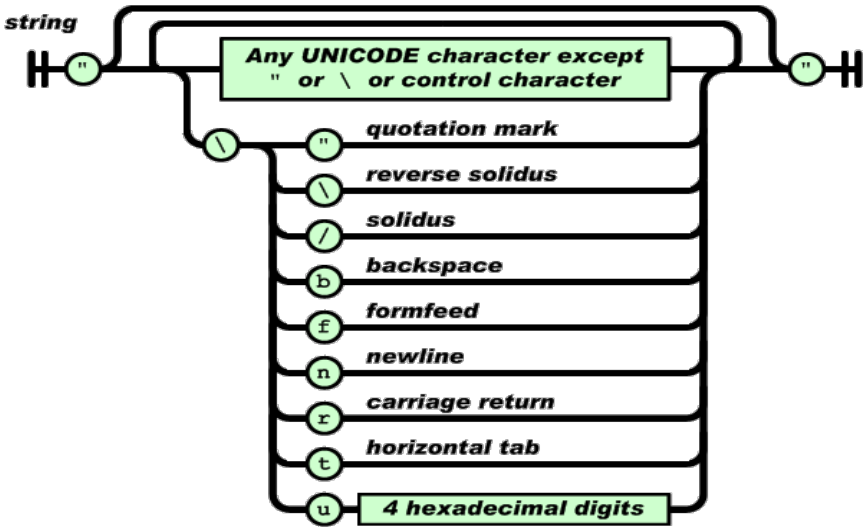


Figura 4.18. Sintaxis de una cadena de caracteres
Fuente: ECMA International, 2013

4.6.2.5 Número

Un número es la combinación de varios símbolos opcionales, el primero es positivo o negativo. Se asume que la ausencia de negativo declara al número como positivo. A continuación, podemos tener uno o más dígitos, que representan la parte entera, para ser continuados opcionalmente por el punto decimal. Después del punto, tenemos la parte digital, que consiste en uno o más dígitos opcionales. Después, aparece la notación en punto flotante, que viene representada por la letra “e” o “E” y simboliza el exponente elevado al valor que viene a continuación y que puede ser positivo o negativo.

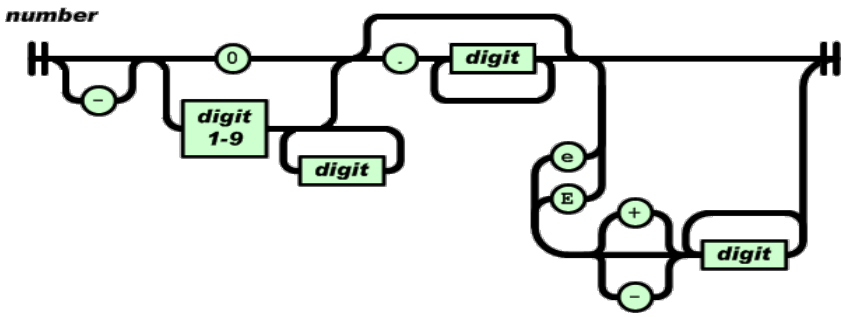


Figura 4.19. Sintaxis de un número
Fuente: ECMA International, 2013

4.6.3 Lenguaje JSON en los tuits

Una vez explicado el lenguaje JSON, vamos a ver cómo se utiliza aplicado a un tuit con un ejemplo real, representado gráficamente con una estructura en forma de árbol, del mismo tipo de la que hemos visto en el apartado anterior, para poder apreciar bien su sintaxis:

Figura 4.20. Diagrama de árbol avanzado ejemplo en JSON.

Esta estructura quedaría representada en el tuit real de la siguiente forma:

Figura 4.21 Representación real de un tuit.

4.6.4 Anatomía de un tuit

Hasta el momento, hemos tratado de explicar el lenguaje JSON a través de ejemplos para comprender su sintaxis. Pero ningún lenguaje de programación está completo sin la semática, y eso es, precisamente, lo que vamos a exponer a continuación.

Para ello, volveremos a poner un ejemplo de un tuit en lenguaje JSON. Como hemos ya podido comprobar, un tuit tiene mucha más información de la que aparece en pantalla a través de la página. Aunque hemos mantenido toda la información almacenada en nuestro sistema, hemos marcado en amarillo solo aquellas partes que nos han resultado útiles para esta investigación y que han permitido obtener conclusiones de interés desde el punto de vista lingüístico.

En primer lugar, mostramos el tuit y a continuación la explicación de los campos subrayados con color amarillo.

```
{
    "created_at": "Thu Aug 13 09:29:00 +0000 2015",
    "id": 631759347585040384,
    "id_str": "631759347585040384",
    "text": "Reckon I'm the only person who could get into uni with only one a level",
    "source": "<a href=\"http:\/\/Twitter.com\/download\/iphone\"
rel=\"nofollow\">Twitter for iPhone<\/a>",
    "truncated": false,
    "in_reply_to_status_id": null,
    "in_reply_to_status_id_str": null,
    "in_reply_to_user_id": null,
    "in_reply_to_user_id_str": null,
    "in_reply_to_screen_name": null,
    "user": {
        "id": 34995672,
        "id_str": "34995672",
        "name": "kiz",
        "screen_name": "kierawebb_",
        "location": "oxford",
        "url": null,
```

```
        "description": "\u0641\u0639\u0644 \u0627\u0644\u062e\u064a\u0631 \u060c
\u0648\u0633\u0648\u0641 \u064a\u0623\u062a\u064a \u0627\u0644\u062e\u064a\u0631
\u0644\u0643",
        "protected": false,
        "verified": false,
        "followers_count": 909,
        "friends_count": 877,
        "listed_count": 23,
        "favourites_count": 2088,
        "statuses_count": 48694,
        "created_at": "Fri Apr 24 18:18:09 +0000 2009",
        "utc_offset": 3600,
        "time_zone": "London",
        "geo_enabled": true,
        "lang": "en",
        "contributors_enabled": false,
        "is_translator": false,
        "profile_background_color": "131516",
        "profile_background_image_url":
"http:\/\/pbs.twimg.com\/profile_background_images\/378800000103212764\/eceaf2fecc7d6c39
4907e054e02908c8.png",
        "profile_background_image_url_https":
"https:\/\/pbs.twimg.com\/profile_background_images\/378800000103212764\/eceaf2fecc7d6c3
94907e054e02908c8.png",
        "profile_background_tile": true,
        "profile_link_color": "009999",
        "profile_sidebar_border_color": "FFFFFF",
        "profile_sidebar_fill_color": "FFFFFF",
        "profile_text_color": "FF0000",
        "profile_use_background_image": true,
        "profile_image_url":
"http:\/\/pbs.twimg.com\/profile_images\/619076317955387392\/3dWFm9wy_normal.jpg",
        "profile_image_url_https":
"https:\/\/pbs.twimg.com\/profile_images\/619076317955387392\/3dWFm9wy_normal.jpg",
        "profile_banner_url":
"https:\/\/pbs.twimg.com\/profile_banners\/34995672\/1436434271",
        "default_profile": false,
        "default_profile_image": false,
        "following": null,
        "follow_request_sent": null,
        "notifications": null
    },
    "geo": null,
    "coordinates": null,
    "place": {
        "id": "754a8cac7b60c7d9",
        "url": "https:\/\/api.Twitter.com\/1.1\/geo\/id\/754a8cac7b60c7d9.json",
        "place_type": "city",
        "name": "Kidlington",
        "full_name": "Kidlington, England",
        "country_code": "GB",
        "country": "United Kingdom",
        "bounding_box": {
            "type": "Polygon",
            "coordinates": [
                [
                    [
                        -1.3070607,
                        51.807681
                    ],
                    [
                        -1.3070607,
                        51.831936
                    ],
                    [
                        -1.269216,
                        51.831936
                    ],
                    [
                        -1.269216,
                        51.807681
                    ]
                ]
            ]
        },
        "attributes": []
    },
```

```
    "contributors": null,
    "retweet_count": 0,
    "favorite_count": 0,
    "entities": {
        "hashtags": [],
        "trends": [],
        "urls": [],
        "user_mentions": [],
        "symbols": []
    },
    "favorited": false,
    "retweeted": false,
    "possibly_sensitive": false,
    "filter_level": "low",
    "lang": "en",
    "timestamp_ms": "1439458140642"
}
```


- "created_at" se refiere a la hora UTC (Tiempo Universal Coordinado) a la que se escribió el tuit.
- "id" es el número entero que identifica a cada tuit; es único e irrepetible, incluso aunque el tuit haya sido eliminado. Este número es mayor de 53 bits y esto provoca dificultades a la hora de interpretarlo en ciertos lenguajes de programación. Para evitar estos problemas, se utiliza "id_str", que está adaptado al resto de los lenguajes de programación.
- "text" es el contenido propio del tuit, codificado en lenguaje UTF-8.
- "user" no solo se refiere al usuario que ha generado el tuit, sino también a los atributos de este usuario, como pueden ser: foto de perfil, número de seguidores, número de usuarios a los que sigue, etc. La información específica del usuario se representa mediante otro diccionario JSON anidado, es decir, se añaden valores a la clave "user". De estos valores, los que nos resultan útiles para esta investigación son "id" (el número entero identificativo de cada usuario); "screen_name" (el nombre de usuario en el formato interno de *Twitter* y distintivo de cada cuenta–es el nombre que sigue a la @, distinto del alias del usuario, que puede cambiarse en cualquier momento); el valor "location" es la ubicación que cada usuario ha especificado en su perfil, generalmente se refiere a su lugar da nacimiento o de residencia, aunque también con frecuencia se realizan indicaciones con tintes humorísticos que no hacen referencia a la realidad); "geo-enabled" indica si el tuit tiene o no información de localización o, lo que es lo mismo, si podemos saber dónde se ha escrito porque el usuario haya activado la función de geolocalización; el idioma en el que el tuit está escrito aparece en la clave "lang" y es detectado por

el sistema de forma automática (en caso de no poder identificar el idioma, *Twitter* lo señala como "indeterminado").

- "geo", como clave independiente, indica si el tuit en concreto (y no el usuario, como en el caso anterior) está geolocalizado.
- "coordinates" se refiere a las coordenadas geográficas del lugar donde se haya publicado el tuit, si está geolocalizado.
- "place", al igual que "user", también tiene una serie de valores asociados, que son "id" (número identificativo del lugar desde el que se haya escrito el tuit); "url" (una dirección web que provee *Twitter* con información sobre el tipo de lugar que es: una ciudad, un edificio, un barrio…); "place_type" se refiere precisamente al tipo de lugar; "name" es el nombre de ese lugar (nombre de la ciudad, del restaurante, etc.); "full_name" da una información más completa del lugar (por ejemplo, la ciudad y el país).
- "bounding_box" es un cuadro envolvente de coordenadas que delimita el espacio geográfico. Esta clave tiene dos claves anidadas con información más específica, que son "type" (el tipo de información codificada en el trazado coordinadas de coordinadas –en *bounding boxes* es "polígono") y las coordenadas propiamente dichas. Estas coordenadas se expresan como una serie de puntos de longitud y latitud que se usan para definir el cuadro delimitado al que se refieren.
- "time_stamp" es el número de milisegundos desde el 1 de enero de 1970 (tiempo estándar de los sistemas informáticos) e indica el momento en el que se ha escrito el tuit.

PARTE II
MARCO METODOLÓGICO Y DE APLICACIÓN

5
Metodología

5.1 CONSIDERACIONES PREVIAS

En este trabajo nos proponemos demostrar la utilidad del análisis de la información en cantidades masivas –*big data*–, en particular, de aquella generada por *Twitter*, para estudiar el lenguaje. Para ello, hemos desarrollado una herramienta web que nos permite trabajar con *big data*, puesto que manualmente sería imposible, y que ofrece a los investigadores y científicos pertenecientes al ámbito de la Lingüística la posibilidad de llevar a cabo estudios basados en este nuevo paradigma.

Hasta ahora, las investigaciones empíricas sobre el lenguaje, enmarcadas dentro de la Lingüística de Corpus, han encontrado las limitaciones propias de la elaboración de los propios corpus, entre las que se encuentran, principalmente, el tiempo y los recursos económicos necesarios para ello.

Desde la aparición de los primeros corpus de fichas construidos de forma manual en la primera mitad del siglo pasado, hasta los últimos corpus informatizados de millones de palabras, las técnicas de recopilación y de análisis de la información han sufrido una enorme evolución que, en cambio, no han sabido superar estos problemas.

Por otro lado, la recopilación de material para el corpus es un proceso costoso y que viene determinado por numerosos condicionantes, como las posibilidades de obtener material, la accesibilidad a esos materiales o la cantidad que hay de ellos disponible; y, además, estos factores influyen en el diseño del corpus y en el producto final (Hunston, 2002 y Gatto, 2008). Sin embargo, los dos factores más determinantes a la hora de sopesar la viabilidad y la rentabilidad de la elaboración de corpus –incluso informatizados– son el tamaño y el tiempo. Ambos aspectos son reconocidos por Rojo (2008) cuando valora, por un lado, la conveniencia de diseñar corpus cuya construcción no sea capaz de avanzar al mismo ritmo con el que evoluciona la lengua y esto

desemboque en que la información que contenga esté desactualizada para el momento de su uso; además, se plantea el autor si merece la pena mantener las inversiones económicas necesarias para la construcción y el mantenimiento de los corpus cuando se nos abre ante nosotros un nuevo horizonte de información accesible a través de la web.

No hay duda de que la última etapa y más innovadora de la Lingüística de Corpus se basa en la concepción de la *web como corpus*, en todas las modalidades que hemos visto en el segundo capítulo del presente trabajo. Tampoco podemos negar, no obstante, algunas de sus limitaciones más importantes, como el tamaño, que, paradójicamente y a pesar de ser su mayor virtud, presenta algunos problemas para los investigadores que se ven sobrepasados por las cifras astronómicas de datos.

Los programas informáticos desarrollados a la luz de las necesidades sobrevenidas en este sentido tratan de resolver este tipo de cuestiones. Su diseño está enfocado al manejo de la información proveniente de la *World Wide Web*, bien directamente, o bien a través de subcorpus construidos con este material, como ocurre, por ejemplo, con *BootCat*, que genera corpus *ad hoc* a través de la web.

Indudablemente, estos programas y otros del mismo estilo ofrecen al científico la posibilidad de un acercamiento más cómodo, fácil y, sobre todo, más enfocado a la investigación lingüística, puesto que no solo devuelven información, sino que lo hacen en un formato adecuado para sus necesidades y que respeta el entorno y la forma de trabajar propios del ámbito de las ciencias del lenguaje. En otras palabras, si pretendemos realizar una búsqueda sobre la web a través de un buscador comercial cualquiera, podemos llegar a obtener los mismos resultados a los que llegamos con los programas específicos. Sin embargo, la forma en la que se nos presentan los datos no está diseñada para su aprovechamiento lingüístico, sino para una mera búsqueda de información. Esto quiere decir que la inversión de tiempo y de esfuerzo, así como la posibilidad de cometer errores aumentan considerablemente en comparación con la utilización de un programa especializado.

Estas nuevas herramientas, por el contrario, están diseñadas para trabajar con la web y para analizar la información desde un punto de vista lingüístico, lo que ha permitido avances en la investigación. Sin embargo, a pesar de tratarse de instrumentos innovadores, las técnicas de análisis que nos ofrecen no difieren de aquellas que se realizan con los corpus que no provienen de la web.

Es aquí donde creemos que es necesario avanzar para aprovechar al máximo las posibilidades que nos ofrecen tanto la web como los avances tecnológicos. Gracias al

desarrollo de la ciencia, ahora es posible dar un paso más allá en las fuentes de donde proviene la información y en los tipos de investigación que se pueden llevar a cabo, para poder así exprimir los datos, realizar análisis y extraer conclusiones como nunca antes se había hecho.

Como hemos explicado detenidamente en la primera parte del trabajo, los estudios realizados con *big data* están a la vanguardia de las investigaciones científicas tanto en las grandes empresas, como en los gobiernos, universidades u organizaciones de todos los ámbitos. Los beneficios no solo se están viendo en términos de avances científicos, sino también desde el punto de vista económico, puesto que un correcto análisis de la información permite el ahorro de grandes sumas de dinero y la obtención de mayores ingresos. Sin olvidar, por supuesto, la drástica disminución del tiempo invertido a la hora de llevar a cabo la extracción y la manipulación de la información.

No obstante, y a pesar de sus múltiples y evidentes ventajas, el campo de la Lingüística no ha sabido o no ha podido explotar, hasta el momento, las ventajas de *big data* y los avances científicos, metodológicos y prácticos que podrían derivarse de ella. En este sentido, estamos convencidos de que una aproximación a la investigación desde este punto de vista y con los medios adecuados supondría dar un gran paso en los estudios lingüísticos y una revolución para los científicos del lenguaje y su modo de trabajo. Este es el punto desde el que partimos en esta investigación y en torno al cual basaremos nuestro estudio.

Como ya hemos explicado, no es objetivo de este trabajo llevar a cabo una investigación de un aspecto concreto de la lengua, sino desarrollar y presentar una aplicación que permita a los lingüistas investigar el lenguaje basándose en *big data,* con técnicas novedosas que aporten nuevos resultados, basados en miles de millones de datos, desde múltiples perspectivas, con mayor agilidad y con el menor esfuerzo e inversión posibles.

La herramienta que aquí presentamos obtiene, almacena y analiza la información contenida en *Twitter* para realizar estudios sobre la lengua escrita que emplean los usuarios de esta plataforma desde varios puntos de vista, que explicaremos más adelante.

El principal motivo por el que hemos escogido *Twitter* se debe a que esta se trata de una red social extendida por todo el mundo, de fácil acceso y utilizada por personas de cualquier país, raza, edad, sexo o profesión. Además, resulta especialmente relevante si tenemos en cuenta que, a diferencia de otras redes sociales, prácticamente la totalidad

de los tuits contiene texto, escrito directamente por los usuarios. Esta ventaja es fundamental para análisis de corte lingüístico ya que, aunque se puede incluir material multimedia, suele venir acompañado por lenguaje escrito. Por otra parte, la variedad de usuarios y la opción de escribir sobre cualquier tema abre enormemente el abanico y amplía las posibilidades de la investigación.

Ya explicamos, al desarrollar la anatomía de los tuits en el capítulo dedicado a *Twitter*, cómo se construyen los mensajes y qué tipo de información nos aportan. De todos los datos contenidos en cada tuit, escrito en lenguaje JSON, solo algunos son específicamente relevantes para una investigación lingüística, como el idioma en el que está escrito, la hora, la ubicación geográfica, el usuario y, por supuesto, el texto.

La tipología de estudios que se pueden llevar a cabo con estos datos es muy variada porque, además de permitir los análisis tradicionales susceptibles de ser llevados a cabo en cualquier otro tipo de corpus, amplían los límites de la investigación basándose en dos características diferenciadoras y esenciales: el tiempo real y la extraordinaria cantidad de información manejada. Dos aspectos que, como indican Yoon *et al*. (2013), son una ventaja adicional y aportan una información más real. A estas características, hay que añadirles una más, no menos importante: la geolocalización de los tuits, que permite, por primera vez en la historia, conocer con exactitud dónde se producen los textos en un momento determinado.

Uno de los objetivos de esta herramienta es poder crear uno o varios corpus[39] 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 A^3 *–anytime, anywhere, anybody*.

Para la construcción de los corpus, obtenemos la información directamente de la API[40] de *Twitter*, que sirve de interfaz entre este y la herramienta, y facilita la interacción humano-software. Según Merino (2014), estas API, que se han sumado en los últimos años a las redes sociales, como *Twitter*, *Facebook*, *Youtube*, *Flickr* o *LinkedIn* y a otras plataformas *online*, entre las que se encuentran *Google*, *Maps*,

[39] Es importante insistir en que, como reza en el título del trabajo, el concepto de *corpus* que aquí utilizamos no se emplea en el sentido habitual del término, sino que se concibe como una evolución de este, puesto que no mantiene los estándares de construcción de los corpus tradicionales. De esta manera, nos referimos con *corpus* a todo el material textual que extraemos de *Twitter* y que utilizamos para su análisis, así como para el establecimiento de conclusiones lingüísticas.

[40] Para más información, véase la nota al pie número 31.

WordPress, etc. han hecho que el marketing de las redes sociales sea mucho más sencillo, rastreable y, por consiguiente, rentable.

Así pues, la existencia de una API abierta facilita en gran medida la tarea de la recopilación de tuits, que es el primer paso para la construcción de nuestro corpus (refiriéndonos de forma general al término, puesto que la mayoría de las investigaciones se basarán en corpus más pequeños, obtenidos también de *Twitter*).

El principal problema al que nos enfrentamos, como ocurre en cualquier trabajo realizado con *big data*, es la gestión de la información. Al tratarse de cantidades de tuits tan monumentales y de tanta información dentro de cada uno, el primer escollo que nos encontramos es la gestión de la propia información, sobre todo cuando se trata esta se produce en tiempo real, puesto que en este caso se requiere un procesamiento lo más rápido posible.

Este es uno de los motivos por los que consideramos importante complementar el análisis en tiempo real con técnicas que permitan el almacenamiento de los datos, puesto que, si no recogemos esos tuits, terminarán perdiéndose y este carácter efímero limitaría en gran medida las posibilidades de la investigación. La otra razón está relacionada con los distintos tipos de análisis que se pueden llevar a cabo y que explicaremos a continuación.

Por otra parte, dado que estamos trabajando con *big data*, también es fundamental obtener una capacidad de proceso estadístico que permita extraer conclusiones fiables sin que se vean obstaculizadas por el volumen de información. Ya hemos hablado de que las técnicas de gestión de la información deben evolucionar y superarse cuando se trata de *big data* y que son necesarias nuevas tecnologías que creen formas de procesamiento alternativas. De hecho, esta idea de grandes conjuntos de información que no pueden ser gestionados por sistemas tradicionales está presente en todas las definiciones de *big data* que hemos analizado hasta el momento (Manyika *et al*., 2011; Zikopoulos *et al*., 2012; Dumbill, 2012; Provost y Fawcett, 2013; Chen *et al*., 2014, etc.).

No debemos caer, pues, en una comparación de esta metodología con sistemas anteriores de análisis, ya que el trabajo con *big data* aporta mejoras, sobre todo en la inmediatez y en el tamaño de la muestra, pero ello conlleva una serie de necesidades que no existían hasta el momento.

Por tanto, y dadas las características de la información aportada por *Twitter* y de las posibilidades técnicas, cabe plantearse en qué puede beneficiarse un investigador del

ámbito de la Lingüística y qué uso puede darle a la información obtenida a través de esta herramienta.

La idea fundamental en torno a la cual gira el proyecto trata de combinar las técnicas de análisis de corpus lingüísticos utilizadas hasta el momento con otras más novedosas que, por razones obvias, no se han realizado nunca antes. De esta forma, no queremos dejar atrás análisis basados en frecuencias de uso, colocaciones o KWIC, porque consideramos que siguen siendo útiles, y más ahora que el volumen de información es mucho mayor.

El propósito, por tanto, no es elaborar un estudio concreto de un aspecto determinado de la lengua, sino evaluar la capacidad de utilizar *big data* para el análisis lingüístico en sus diferentes modalidades. Para ello, se ha llevado a cabo la creación de una herramienta que permita a investigadores, fundamentalmente del campo de la Lingüística, procesar millones de datos de forma rápida, precisa y con total eliminación de costes, ya que no existe, por el momento ninguna otra que realice estos tipos de análisis y que cumpla con los objetivos que nos hemos marcado.

Así pues, aunque seguiremos trabajando con algunas de las técnicas de corpus tradicionales, las combinaremos con otras que puedan complementar en análisis de la información. La visualización instantánea de los tuits que se están escribiendo en tiempo real, su localización, el idioma, el número de publicaciones o el análisis histórico, entre otros que explicaremos con más detenimiento a continuación, son los puntos fuertes de esta herramienta y los que la distinguen de cualquier otra que se haya desarrollado hasta ahora.

Es importante señalar que el análisis de los resultados debe ser llevado a cabo por parte del especialista, que será quien interprete y determine el significado de los datos obtenidos a través de la herramienta; y, aunque reconocemos que no es la herramienta única y definitiva para el análisis lingüístico, sí hay que reconocer que aporta resultados basados en millones de datos que pueden complementar cualquier otro estudio sobre el lenguaje y las lenguas que pueda realizarse.

Además, consideramos que puede resultar útil para otro tipo de investigaciones relacionadas con otros ámbitos, basadas también en el lenguaje, en las que sea posible relacionar variables demográficas, culturales o sociológicas que permitan realizar estudios más amplios. Todo ello, gracias a la manipulación de *big data* en tiempo real o en diferido que nos proporciona *Twitter*.

5.2 ANÁLISIS DE REQUISITOS

En los trabajos lingüísticos actuales, en general, y en los que se enmarcan dentro de la Lingüística de Corpus, en particular, la relación entre el lingüista y el desarrollador informático es indispensable y se encuentra en una continua comunicación.

Para que la relación funcione y el resultado final sea el esperado, es primordial que el primero de ellos –el lingüista– sea capaz de expresarle al desarrollador cuáles son sus objetivos, expectativas y necesidades en lo que a la herramienta se refiere. Solo si este primer paso se lleva a cabo con claridad y organización, el desarrollador informático podrá hacerse una idea de cómo enfocar su trabajo para poder dar respuesta a las exigencias del lingüista. A partir de aquí, e incluso habiendo superado esta fase con éxito, la comunicación entre ambos debe ser constante para añadir, eliminar, modificar o perfilar todos los aspectos que tengan que ver con la herramienta que va a permitirnos trabajar con *big data* para obtener beneficios en el ámbito de la Lingüística.

Esta primera fase es lo que se conoce, en una relación entre un cliente y un informático, como "análisis de requisitos". Para elaborar la herramienta que se presenta en este trabajo, las especificaciones que se hicieron llegar al desarrollador en una entrevista en la que se establecieron las necesidades fundamentales fueron las que se detallan a continuación.

a) Para comenzar, lo primero que había que tener en cuenta es que se pretendía servirse de *big data* como fuente para la investigación lingüística y, para ser más concretos, solo una parte de esa monumental cantidad de información: la que se encuentra en *Twitter*.

Como investigadores en ciencias del lenguaje, consideramos que esta plataforma social ofrece numerosas ventajas para estudiar cómo la utilizan los hablantes de las distintas lenguas para hablar sobre sí mismos o sobre acontecimientos externos, para reaccionar frente a algo o para comunicarse entre sí. Por lo tanto, necesitamos que nuestra herramienta sea capaz de analizar la información textual proveniente de *Twitter*.

Puesto que pretendemos crear corpus basados en los tuits que se publican, es necesario obtener la mayor cantidad de información posible acerca de ellos. Si, además, podemos ser nosotros quienes seleccionemos los parámetros del corpus o de los tuits que vamos a analizar para poder establecer las características que más nos interesen de cara a una investigación específica, conseguiríamos optimizar mucho más el tiempo y afinar los resultados. De esta manera, hemos solicitado que la herramienta ofrezca al

usuario la opción al usuario de seleccionar la información según los siguientes criterios: fecha de publicación del tuit, ubicación geográfica (si la hubiere), idioma (si estuviere disponible) y contenido.

b) Por otra parte, como ya hemos visto al analizar los retos de *big data*, uno de los principales problemas a los que nos enfrentamos trabajando en este ámbito es el hecho de que el procesamiento y el análisis de los datos, que se presentan en en cifras tan astronómicas, supone una ralentización en el proceso. Aunque es cierto que el mero hecho de poder procesar millones de fragmentos textuales en solo unas horas o, incluso, en unos días, ya supone de por sí un avance si lo comparamos con el tiempo que necesitaríamos para ello con métodos tradicionales, hemos considerado un elemento importante que la herramienta presentara los resultados en el menor tiempo posible y que, en el peor de los casos, tan solo nos llevara unos pocos minutos. El mayor ahorro posible de tiempo fue, por lo tanto, otro de los requisitos exigidos en esta primera aproximación a la herramienta.

c) También tuvimos en consideración que los datos obtenidos pudieran ser intercambiados y utilizados con otros programas y herramientas conocidos de corte estadístico, textual, de presentaciones, etc. Esto es lo que se conoce como "interoperabilidad" y creemos que es uno de los requisitos para facilitar el intercambio de comunicación y el trasvase de resultados entre la comunidad científica.

d) En relación con este último se encuentra el de desarrollar una herramienta que pueda ser compatible con diferentes plataformas, sin que dependa de una infraestructura específica. En otras palabras, buscamos la mayor universalidad y accesibilidad posible; y para ello es necesario que se trate de una aplicación web y no de una aplicación móvil o un programa informático descargable en el ordenador. De esta manera, cualquier usuario con acceso a Internet podrá utilizarla.

e) Siguiendo con este objetivo de lograr la mayor comodidad y facilidad posible para el usuario, era nuestra intención que cualquier persona que se acercara a la herramienta fuera capaz de utilizarla sin gran esfuerzo y sin la necesidad de poseer amplios conocimientos informáticos. Para ello, pensamos que había que prestar especial atención a la interfaz a la que el lingüista se iba a enfrentar, para que fuera intuitiva, cómoda y flexible en cuanto al tipo de información y de análisis que se le demandaran.

f) El último requisito que quisimos tener en cuenta, fue que la herramienta tuviera la capacidad de distinguir entre los idiomas de los textos recopilados.

Una vez especificados –desde el punto de vista de nuestras necesidades y nunca desde un punto de vista técnico– los anteriores requisitos de funcionamiento interno y de apariencia que considerábamos importantes, nos centramos en el tipo de estudios que pretendíamos realizar con la información obtenida.

Como ya hemos apuntado unas líneas más arriba, dado que las bases teóricas de este trabajo se encuentran ancladas en la Lingüística de Corpus, creímos necesario llevar a cabo algunos de los análisis más importantes que se siguen realizando hoy en día en los corpus informatizados y que nos pueden seguir resultado útiles incluso con los corpus obtenidos con información procedente de la web, como es nuestro caso. Estas funciones más tradicionales, presentes en los estudios de corpus informatizados y en los programas diseñados para la utilización de la web como corpus, son: a) el estudio de las palabras en contexto –con la técnica conocida como KWIC–, b) el análisis de las colocaciones lingüísticas, c) la representación de las ocurrencias de palabras, términos o expresiones y d) el análisis de la relación entre tipos y casos de palabras para determinar la densidad lingüística de un texto.

Este tipo de utilidades, aunque enormemente rentables para cualquier estudio que nos planteemos, son, sin embargo, insuficientes si nos enfrentamos a un corpus con las características que van a presentar los que esperamos recopilar con esta herramienta. Dado que la naturaleza, la cantidad y el tipo de información con que contamos son radicalmente distintos a los que estamos acostumbrados, se presta atención a otros tipos de análisis más amplios que se pueden combinar con los anteriores.

f) Por esta razón, otro de nuestros requisitos era poder consultar la información en tiempo real, sin olvidar aquellos tuits que han sido publicados en el pasado para poder así realizar estudios tanto diacrónicos como sincrónicos de la lengua. Además, era importante poder sacar partido de la geolocalización de los tuits para poder identificar el lugar desde el cual se comunican las personas así como el número de personas que lo hacen, bien en un momento concreto, o bien a lo largo del tiempo.

Por último, solicitamos también que la herramienta tuviera capacidad para representar los resultados de los análisis de forma visual, mediante gráficos, tablas y mapas, y que estas representaciones pudieran exportarse a distintos formatos compatibles que permitieran la interoperabilidad con otras herramientas y sistemas existentes de la que hemos hablado más arriba.

5.3 *WORDICS SUITE*

En este epígrafe procederemos a explicar en detalle la herramienta que hemos desarrollado para probar la hipótesis de que es posible utilizar *big data* para investigar determinadas cuestiones relacionadas con el lenguaje y las lenguas. Puesto que la herramienta maneja millones de datos de forma simultánea, la metodología de desarrollo nos obliga a dividirla en bloques independientes para facilitar el procesamiento de la información. La herramienta lleva por nombre genérico *Wordics Suite* y los cuatro módulos que la componen son *Wordics Live*, *Wordics One*, *Wordics Archive* y *Wordics Data*. Cada uno de ellos está especializado en una función determinada, de manera que el usuario puede priorizar sus necesidades dependiendo de si está más interesado en una alta velocidad de respuesta o, por el contrario, en un análisis más exhaustivo de los datos.

Para ofrecer una visión completa de las características y utilidades de la herramienta que hemos denominado *Wordics Suite*, dividiremos nuestras explicaciones y comentarios en cuatro apartados, comenzando por la presentación de sus características generales y de su arquitectura, para finalizar con la explicación individualizada de cada uno de los cuatro módulos.

El acceso a nuestra herramienta se puede realizar a través de la página web www.wordics-suite.com/desde cualquier ordenador con acceso a Internet.

5.3.1 Características generales

Las funciones o utilidades que nos ofrece la herramienta en su última versión –aunque no la definitiva, puesto que queda la puerta abierta a futuras mejoras y desarrollos– son el resultado del trabajo del desarrollador informático basándose en el análisis de requisitos que llevamos a cabo en el apartado anterior. Por lo tanto, una vez diseñada la herramienta, presentamos las funciones generales que es capaz de llevar a cabo y que hacen de ella una aplicación útil para nuestros propósitos:

1. Analiza la información de *Twitter* a través de la API de *Twitter*, lo que nos permite obtener un *streaming* de información de 52 tuits por segundo (esto es, 3.120 tuits al minuto, 187.200 a la hora y 4.492.800

al día). Estos datos se van almacenando en un servidor web conforme se van recibiendo.

2. Ofrece la posibilidad de seleccionar la información y de visualizarla en función del segmento temporal, la ubicación geográfica (en términos de latitud y longitud) e idioma previamente determinados.
3. Lleva a cabo un preprocesado de la información en el momento de su almacenamiento, lo que permite realizar los análisis posteriores en una cantidad razonable de tiempo, algo que no sería posible si la herramienta guardara la información para analizarla en el momento en el que se solicitara para un estudio concreto.
4. Realiza análisis y elabora la visualización de los resultados en tiempo real, a partir de la información obtenida directamente desde *Twitter*.
5. Exporta la información obtenida en distintos formatos, dependiendo de si se trata de información en gráficos o de información textual. La primera podrá ser exportada en formato JPG o PNG, mientras que la segunda, en JSON, XML, CSV, TXT, SQL o MS-Excel.
6. Utiliza una interfaz especialmente gráfica y dinámica, basada en los estándares de HTML 5 y JavaScript, con una arquitectura cliente-servidor web que permite el acceso a la herramienta desde cualquier plataforma con navegador web (Google Chrome, Opera, Safari, Internet Explorer, Mozilla Firefox, etc.).
7. Lleva a cabo análisis del tipo KWIC, colocaciones, frecuencias de uso y densidad léxica.
8. Posibilita nuevos tipos de análisis de la información, basados en el filtrado de tuits en función del idioma, de las coordenadas geográficas y de la fecha exacta que se deseen.

Filtrado

En los diferentes apartados de *Wordics Suite* tendremos que filtrar los resultados o realizar análisis en función de una o varias palabras. Es por tanto de vital importancia definir qué entiende Wordics por palabra, y qué posibles criterios aplica en cuanto a la selección de una determinada. Dado que trabajamos con información que utiliza caracteres específicos (hash # o arroba @), es importante tanto para el investigador

como para la herramienta delimitar claramente el funcionamiento de cada uno de los módulos en cuanto al aspecto de selección.

Tipos generales de selección de palabras (combinables)

1. Selección de palabra completa o subcadena

En una selección de palabra completa, esta queda seleccionada completamente cuando la cadena de búsqueda concuerda en número y caso de letras. La longitud de ambas cadenas ha de ser exacta, y las letras en cada posición han de ser las mismas.

Ejemplo:

España -> España : Verdadero

España -> Español: Falso

Sin embargo, en una selección de subcadena, una palabra generará un estatus de encontrada si la subcadena provista coincide en algún lugar dentro de la palabra buscada.

Ejemplo:

cata -> recatado: Verdadero

cata -> alicatas: Verdadero

2. Selección sensible a mayúsculas y minúsculas

Llamamos selección estricta o sensible a la comprobación exacta de la palabra con otra dada, incluyendo la distinción entre mayúsculas y minúsculas.

España -> españa: Falso

España -> España: Verdadero

Por otro lado, una selección neutra o no sensible da como válida la comprobación de dos palabras aunque no coincidan en la distribución de mayúsculas y minúsculas.

España -> españa: Verdadero

España -> España: Verdadero

Tipos avanzados de selección de palabras

1. Selección por expresiones regulares

Una **expresión regular**, también llamada *regex*, es una secuencia de caracteres que forma un patrón de búsqueda, principalmente utilizada para la búsqueda de patrones de cadenas de caracteres u operaciones de sustituciones. La mayoría de las formalizaciones proporcionan los siguientes constructores: una expresión regular es una forma de representar a los lenguajes regulares y se construye utilizando caracteres del alfabeto sobre el cual se define el lenguaje.

Este tipo de expresiones regulares, aunque potente, es demasiado complejo (cita con web o texto sobre *regex*) para el sistema que nos compete, puesto que podemos optar a otro mucho más sencillo e igualmente eficaz, como es la selección por comodines.

2. Selección por comodines

Este tipo de selección se basa en una serie de criterios de tipo y número definidos por un conjunto de caracteres denominados 'comodines'. Un **carácter comodín** es un carácter que representa cualquier otro carácter o cadena de caracteres. Los comodines se utiliza habitualmente en los buscadores de Internet. Veamos a continuación la tabla de comodines implementada en algunas de las secciones de análisis de *Wordics Suite*:

NOMBRE	EJEMPLO	REPRESENTACIÓN	SELECCIONES
Asterisco	Espa*	Cero o más caracteres	España, Esparce, etc.
Interrogación	Cas??	Un carácter por cada interrogación	Casos, Casas, Casón, etc.
Clase	[CPR]aso	Cualquier carácter dentro de la lista	Caso, Paso, Raso, etc.
Conjunto	Españ{a,ol,ola}	Cualquier conjunto dentro de la lista	España, Español, Española

Este tipo de selección es mucho más avanzado y más adecuado para los propósitos de la herramienta, puesto que permite buscar palabras de cuya existencia u ortografía no se tiene certeza. Es fácil presumir el paralelismo entre la selección por comodines y la separación entre morfemas y lexemas de las diferentes palabras.

Ejemplos de selección por comodines

a) Definición: buscar todas las palabras con diferentes prefijos derivadas de historia.

Patrón de selección: *historia.

Selección: historia, microhistoria, plastihistoria, prehistoria, protohistoria.

b) Definición: buscar todas las palabras que contengan el morfema 'stud'.

Patrón de selección: *stud*.

Selección: estudiado, estudiamos, estudiante, estudiantina, studium, etc.

c) Definición: buscar todas las palabras que comiencen por 'anti'.

Patrón de selección: anti*

Selección: antiadhererente, antialérgico, antiamericanistas, antibalas, etc.

d) Definición: buscar todas las palabras que comiencen por 'ca' y contengan dos caracteres adicionales.

Patrón de selección: ca??

Selección: cabe, cabo, caco, cada, caes, caja, cama, etc.

El tipo de búsqueda en cada módulo de la herramienta es el siguiente:

-*Wordics Live*: búsqueda de cadena completa no sensible a mayúsculas y minúsculas .

-*Wordics One*: búsqueda de subcadena no sensible a mayúsculas y minúsculas en el corpus de tuits principal. La tabla de frecuencias presenta el mismo tipo de búsqueda que el corpus principal. Las colocaciones y las KWIC, por el contrario, presentan una búsqueda de cadena completa y sensible a mayúsculas y minúsculas.

-*Wordics Archive*: el buscador de palabras funciona con comodines y con una comparación no sensible a mayúsculas y minúsculas. La búsqueda sobre el término objeto de estudio y en KWIC es de cadena completa y no sensible a mayúsculas y

minúsculas. Las colocaciones, sin embargo, también son de cadena completa, pero sensibles a mayúsculas y a minúsculas.

5.3.2 Arquitectura

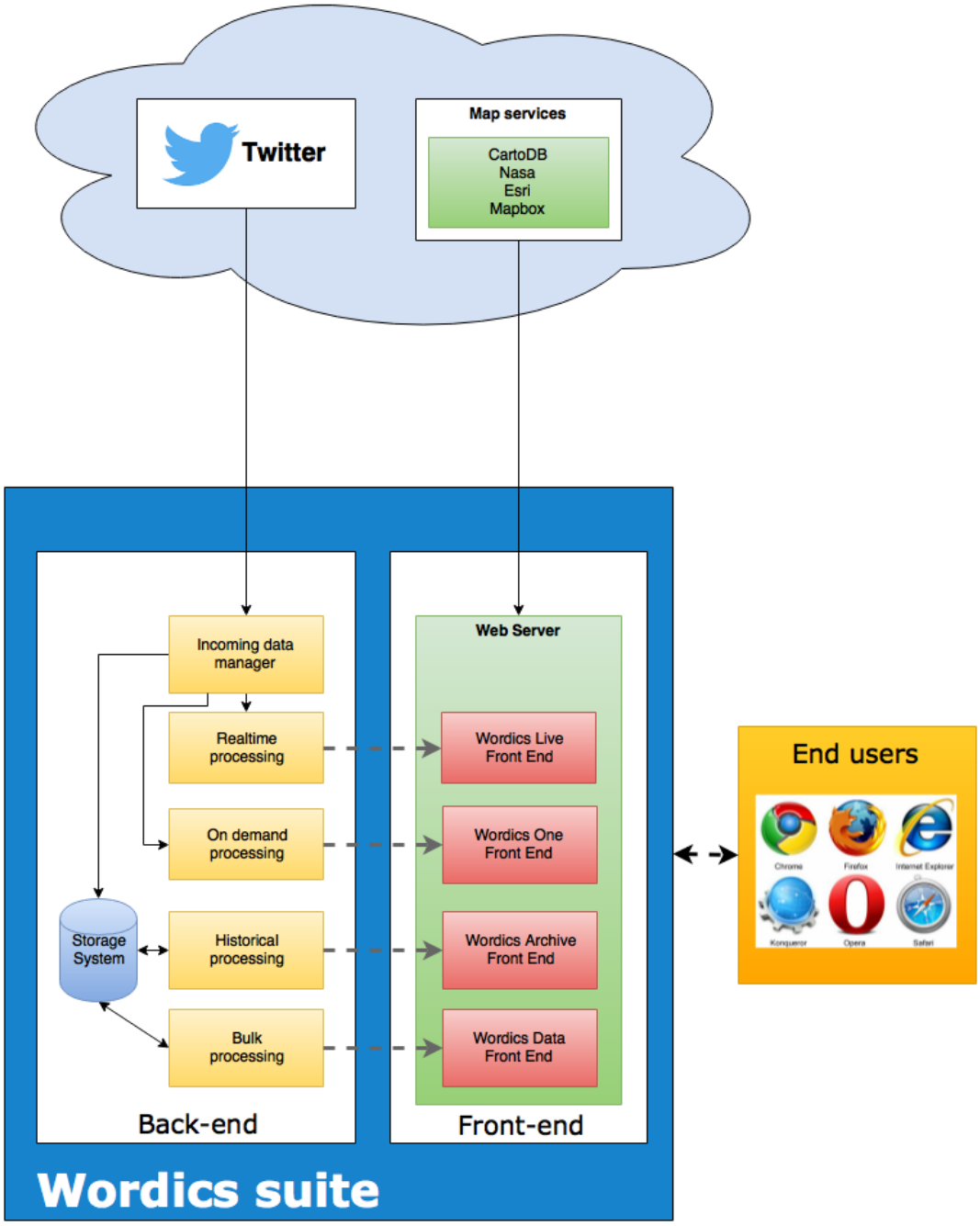


Figura 5.1. Arquitectura de *Wordics Suite*

El funcionamiento de la herramienta se basa en dos partes diferenciadas: por un lado, la nube, dentro de la cual se encuentra Internet y a partir de donde se obtiene toda

la información necesaria; por el otro lado, los sistemas y programas que adquieren, almacenan, procesan y analizan los datos con los que el lingüista va a trabajar.

Comenzaremos explicando la nube, representada en el diagrama en la parte superior de este. Dentro de la nube está Internet, la fuente de *big data* que recoge toda la información que necesitamos para poder poner en funcionamiento la herramienta.

Como ya hemos avanzado, para nuestro propósito no es necesario rastrear toda la información posible en Internet, puesto que nos vamos a limitar a la información contenida en *Twitter*. Esta es la primera fuente de la que bebe la herramienta, la que nos va a aportar los tuits y toda la información contenida en ellos.

Por otro lado, también dentro de la nube, tenemos los servicios de cartografía y los servicios de conocimiento. Estos dos tipos de servicios no aportan datos para ser procesados, sino que se trata de los proveedores de mapas –en el primer caso– y de programas de traducción automática o de sugerencias de palabras –en el segundo– que vamos a utilizar para representar la información u ofrecer otro tipo de servicios en la interfaz de la herramienta. Así, cada uno de los tipos de mapas que ofrecemos para visualizar los tuits (nocturno satélite, diurno satélite, mapa base, base invertida, etc.) pertenecen a una empresa u organización distinta (*CartoDB*, *Nasa*, *Esri*, *Mapbox*…).

Centrándonos en *Wordics Suite*, se trata de una herramienta con dos módulos, denominados *back-end* y *front-end*, que se encuentran interconectados con la nube (es de aquí de donde toma toda la información) y con los distintos buscadores desde los que los usuarios van a poder acceder a ella.

Mientras que el *back-end* se puede considerar la parte de funcionamiento interno de la herramienta, esto es, el motor que gestiona e impulsa las distintas funciones que la herramienta ofrece, el *front-end* se trata de la interfaz que utiliza el usuario y que le permite interactuar con la información del *back-end*.

El primer módulo, es decir, el *back-end*, es, por tanto, el corazón de la herramienta. Por explicarlo de una forma sencilla, recoge la información de *Twitter*, la procesa y la almacena en una base de datos. Como se puede apreciar en el diagrama, está subdividido en otra serie de módulos más pequeños, interrelacionados entre sí, que se encargan de la gestión de toda la información. El submódulo denominado *incoming data manager* es el encargado de recoger los tuits y almacenarlos en una base de datos (*storage system*), después de haber realizado un filtrado para limpiar la información. Además, tiene dos funciones adicionales. La primera de ellas consiste en seleccionar los últimos tuits y enviarlos a otro submódulo, denominado *realtime processing*, que será el

encargado de analizar la información en tiempo real. La segunda conecta con el proceso denominado *on-demand processing*, encargado de recopilar los tuits de la cuenta concreta de *Twitter* que se vaya a analizar. El tercer submódulo del *back-end –historical processing–* es el que utilizaremos para el análisis histórico y será el que permitirá procesar, a petición del usuario, la información almacenada en el *storage system*. Es decir, gracias a esta función, el usuario podrá visualizar y realizar un estudio de los tuits que se hayan publicado a lo largo del tiempo sin la urgencia que pueda provocar la visualización en tiempo real. El último de los submódulos pertenecientes al *back-end*, denominado *bulk processing*, obtiene la información sin procesar de la base de datos en un formato compatible con otras herramientas para facilitar la interoperabilidad con ellas.

El segundo módulo de la herramienta, el llamado *front-end* es, como decimos, el que permite la visualización por parte del usuario de todo lo que ocurre dentro del programa. Puesto que los usuarios potenciales de esta herramienta serán en su mayoría lingüistas a los que no se les presupone, en principio, amplios conocimientos informáticos, consideramos fundamental que la interfaz con la que ellos interactúen con el servicio sea sencilla, intuitiva y diseñada específicamente para la investigación lingüística. Esta es, precisamente, la función desempeñada por el *front-end*; este módulo está dividido en tres submódulos que se corresponden, cada uno de ellos, con los tres apartados en los que hemos dividido la herramienta y que explicaremos en profundidad más adelante. Estos tres apartados consisten, *grosso modo*, en análisis de la información en tiempo real, análisis histórico de la información e información general contenida en los tuits. Como hemos explicado unas líneas más arriba, el módulo de análisis en tiempo real *–Wordics Live–* interactúa de forma directa con el procesador en tiempo real del *back-end*, mientras que el de información relativa a cuentas individuales *–Wordics One–* conecta con *on-demand processing*; el de análisis histórico *–Wordics Archive–* está relacionado con el submódulo de *back-end* denominado *historical processing*, que, recordemos, procesaba la información a petición del usuario, sin necesidad de ofrecer todo lo que estuviera sucediendo en un instante determinado. Por último, *Wordics Data* permite la exportación de información en formatos compatibles.

Dado que uno de los requisitos que se establecieron como fundamentales es el acceso fácil y directo a la herramienta, esta ha sido desarrollada en una plataforma de difusión de información estándar, como es la web. Esto quiere decir que no es necesario descargarse el programa para poder ser utilizado, sino que, al ser una aplicación web, se

puede acceder a ella desde cualquier ordenador y con cualquier sistema operativo, siempre que esté conectado a Internet, ya que este tipo de plataformas tienen un acceso universal. De esta forma, conseguimos que al investigador se le impongan las mínimas restricciones de acceso. Aun así, debido a que la herramienta ha sido desarrollada utilizando los estándares más actuales de programación y acceso a datos, es altamente recomendable acceder a la misma utilizado las últimas versiones de sistemas operativos y navegadores.

5.3.3 Módulos de *Wordics Suite*

En este apartado, desarrollaremos pormenorizadamente las características de los cuatro módulos constitutivos de *Wordics Suite* citados anteriormente: *Wordics Live*, *Wordics One*, *Wordics Archive* y *Wordics Data*. Como ya hemos explicado, puesto que estamos trabajando con información masiva, aunque el origen de esta información siempre sea *Twitter*, este diseño ha permitido optimizar el procesamiento de la información en función del tipo de análisis y visualización requeridos.

5.3.3.1 *Wordics Live*

Wordics Live constituye el primero de los módulos que componen nuestra herramienta. Su característica esencial es la obtención de tuits en tiempo real para un análisis preliminar rápido e instantáneo.

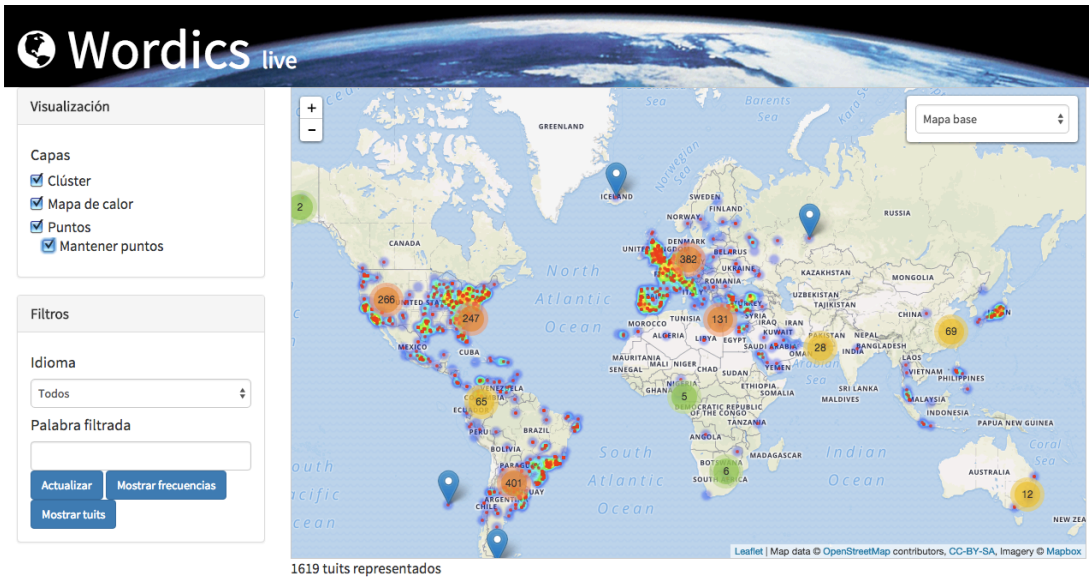


Figura 5.2. Interfaz principal de *Wordics Live*

Las funciones que desempeña este módulo, por tanto, están diseñadas para poder realizar un estudio de lo que está pasando en un momento determinado en *Twitter*, es decir, de lo que están escribiendo los usuarios de *Twitter* (los denominados *tuiteros*). Dada la inmediatez característica de este apartado de la aplicación, no hemos considerado oportuno incluir aquí análisis muy pormenorizados de la información, por varios motivos:

a) El primero de ellos tiene que ver con la cantidad de información y con su permanencia en el tiempo. Aunque es cierto que en unos segundos tenemos la posibilidad de obtener miles de tuits, una cantidad más que considerable si tenemos en cuenta el tiempo empleado en recopilarlos y el que habría hecho falta con cualquier otro sistema, este es un número insignificante si lo comparamos con lo que podemos llegar a obtener solo con unas cuantas horas de búsqueda –por no hablar de días, meses o años–. Ello nos lleva a considerar este módulo sumamente útil para obtener un fotograma de lo que está ocurriendo en tiempo real y poder analizarlo sin necesidad de acudir a un histórico.

b) En segundo lugar, los análisis lingüísticos más pormenorizados implican más procesamiento y ralentizarían la capacidad de análisis. Por otro lado, este tipo de investigaciones requieren una mayor estabilidad de los datos y se beneficiarían de mayores cantidades de información, algo que se puede conseguir realizando un análisis histórico de la información a lo largo del tiempo. Por este motivo, los análisis estadísticos y con un corte más lingüístico se llevarán a cabo en el tercer módulo de la herramienta: *Wordics Archive*.

5.3.3.1.1 Filtrado

La función principal de *Wordics Live* es la visualización de los tuits y de las palabras que estos contienen en tiempo real. Esta representación de los tuits, no obstante, se puede llevar a cabo desde distintos puntos de vista. Para ello, el usuario tiene la posibilidad de ajustar una serie de parámetros que determinarán y especificarán la búsqueda de manera que pueda adaptarse a sus necesidades. Así, existe la posibilidad de filtrar diversos campos, en el caso de que el investigador no desee visualizar todos los tuits que se publican en todo el mundo y en cualquier lengua en un momento dado.

El primero de los parámetros que el usuario puede ajustar es el idioma. Aunque el sistema soporta todos los idiomas del estándar ISO 639-1[41], la realidad es que muchos de ellos son dialectos que se hablan en zonas en las que ni siquiera existe conectividad a Internet. Esto implica que muchos de estos idiomas, a pesar de estar contemplados en el estándar, no se utilizan en *Twitter* prácticamente nunca.

En segundo lugar, existe la posibilidad de filtrar una o varias palabras para que la herramienta realice la búsqueda sobre ellas, eliminando así todos aquellos tuits que no contengan esas palabras. En efecto, se puede filtrar una sola palabra, pero también es posible introducir en el buscador más de una palabra que conforme el objeto de la búsqueda. Esta opción ofrece, por tanto, cuatro alternativas:

1. Introducción de una sola palabra en el buscador.
2. Introducción de más de una palabra separadas por espacios, de manera que el sistema realice una búsqueda de esas palabras en bloque. Los tuits que aparezcan después del filtrado mostrarán los resultados con las palabras tal y como se hayan escrito y en el mismo orden.
3. Introducción de más de una palabra –hasta ocho– separadas entre sí por comas y sin espacios. De esta forma, se llevará a cabo una búsqueda de todas las palabras filtradas que aparecerán en los resultados con un color distinto para cada una de ellas, de manera que se distinga dónde se ha publicado cada una.
4. Combinación de las dos anteriores. Esta última posibilidad permite filtrar varios bloques de palabras distintos. Para ello solo es necesario que estén separados entre sí, igual que en el caso anterior, por comas y sin espacios.

En la figura 5.3 podemos ver un ejemplo de filtrado de varias palabras en el que cada una de ellas se muestra con un color diferente. Así, la palabra *amor* aparece en rojo; *semana*, en azul; *noche*, en verde y *menos*, en amarillo.

[41] La norma ISO 639-1 pertenece a la primera parte del estándar 639 y se utiliza para asignar un código de dos letras a cada uno de los principales idiomas del mundo de forma que puedan ser reconocidos internacionalmente.

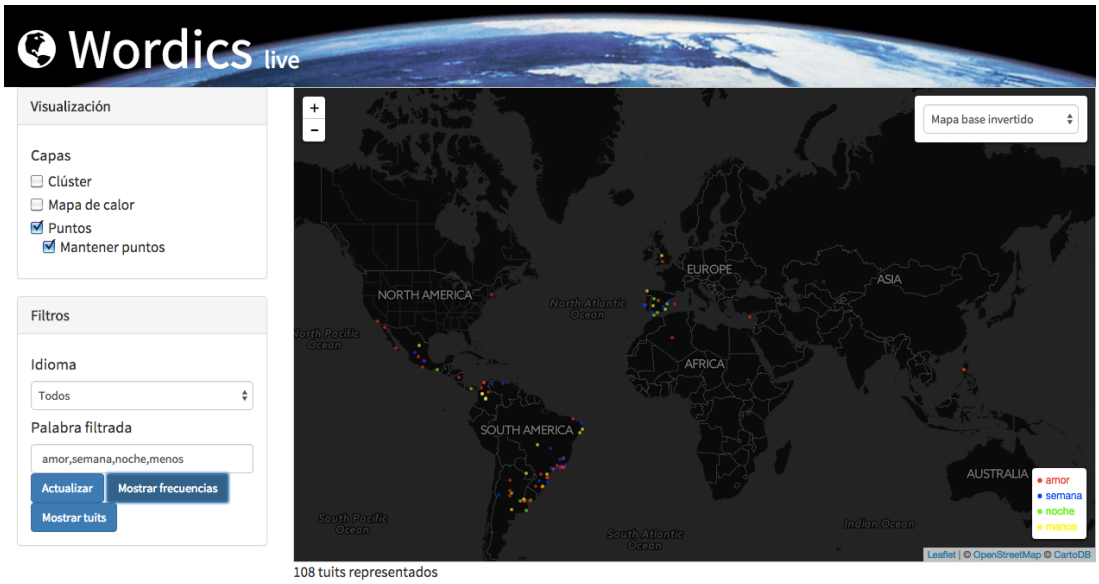


Figura 5.3. Ejemplo de filtrado con varias palabras separadas por comas

Otra de las funcionalidades de *Wordics Live* consiste en mostrar las frecuencias de uso, bien de la totalidad de las palabras que se estén escribiendo en tiempo real, bien de las filtradas con los sistemas anteriormente descritos. En este módulo, por motivos de optimización de la plataforma, solamente se muestran las cien palabras más frecuentes de todos los tuits. *Wordics Live* permite exportar esta lista en formato *Microsoft Excel* para que se pueda ser utilizada en otros programas o en cualquier otro tipo de análisis ajeno a la herramienta.

Figura 5.4. Tabla de frecuencias de palabras

Por otro lado, *Wordics Live* puede mostrar también la lista de tuits que el sistema está procesando en tiempo real (todos ellos, o bien solo los resultantes de haber aplicado algún filtro de los que hemos explicado), con indicación de sus coordenadas geográficas, cronológicas y el idioma utilizado, así como el país de procedencia. De nuevo esta lista puede ser exportada a *Excel*, al igual que en el caso anterior.

Figura 5.5. Tabla de detalles de los tuits filtrados

5.3.3.1.2 Visualización

Por un lado, obtenemos distintos tipos de mapas en los que se van a situar los tuits que nos ha devuelto la aplicación la aplicación tras el filtrado. La elección de mapa no altera en absoluto los resultados, ya que es una cuestión estética dirigida a optimizar la visualización en función del tipo de búsqueda. Para cambiar el tipo de mapa, basta con desplegar las flechas del cuadro de selección que aparece en la parte superior derecha del mapa. Los tipos de mapas son:

a) Mapa base:

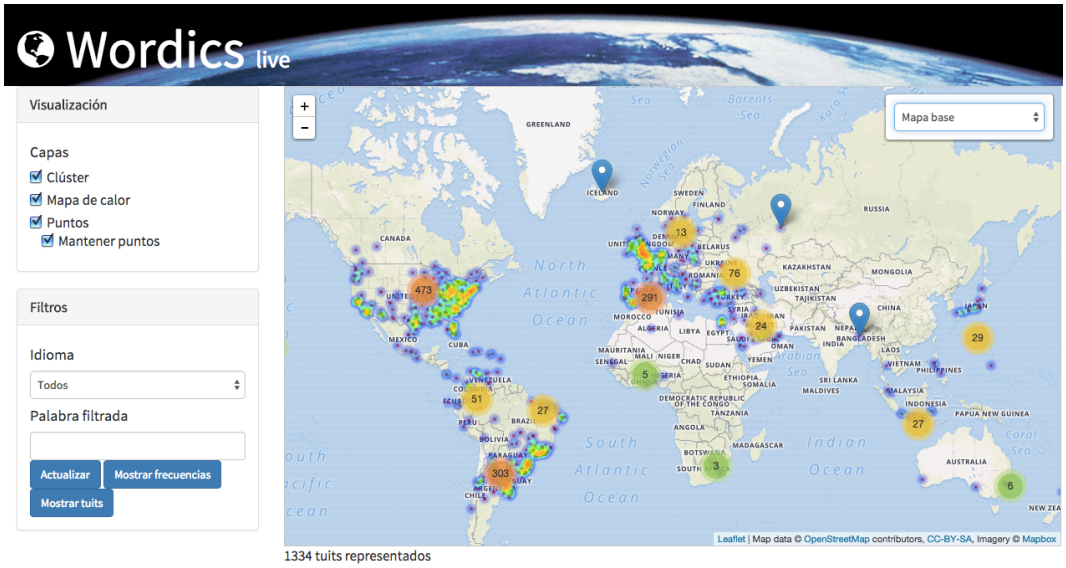


Figura 5.6. Interfaz de *Wordics Live* con mapa base

b) Mapa base invertido:

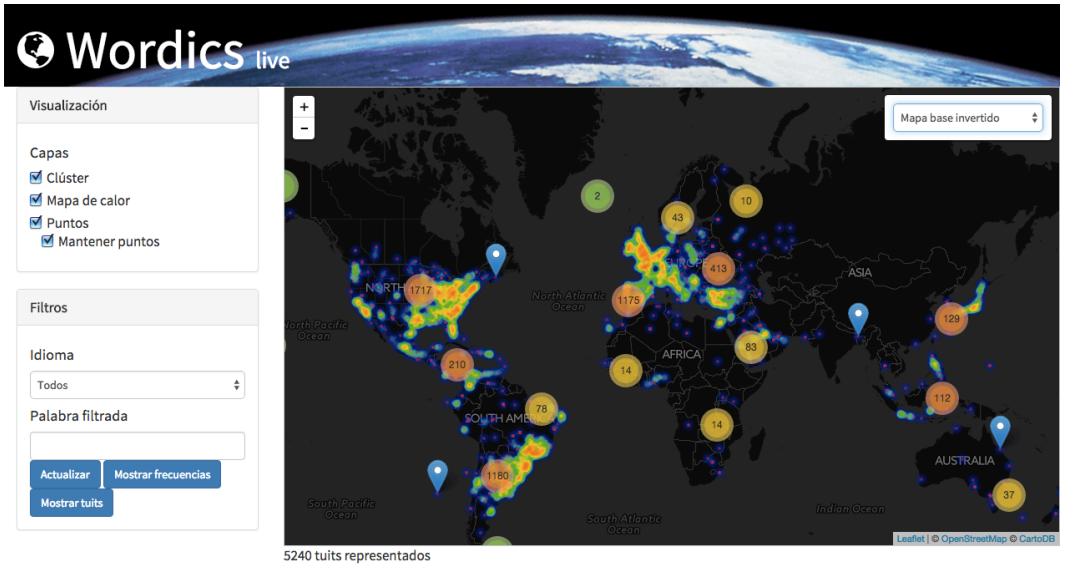


Figura 5.7. Interfaz de *Wordics Live* con mapa base invertido

c) Satélite:

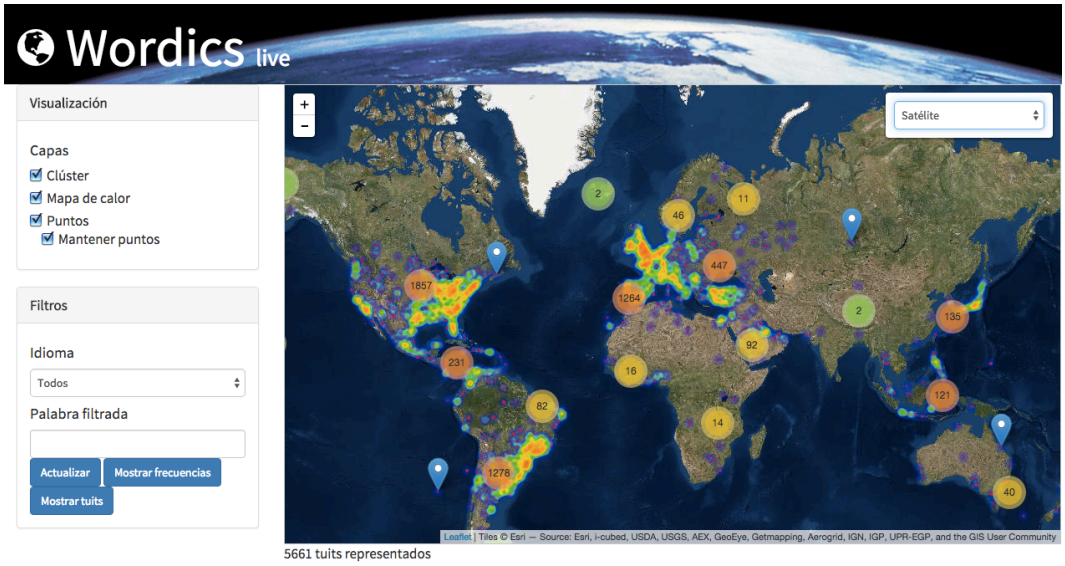


Figura 5.8. Interfaz de *Wordics Live* con mapa satélite

d) Satélite nocturno:

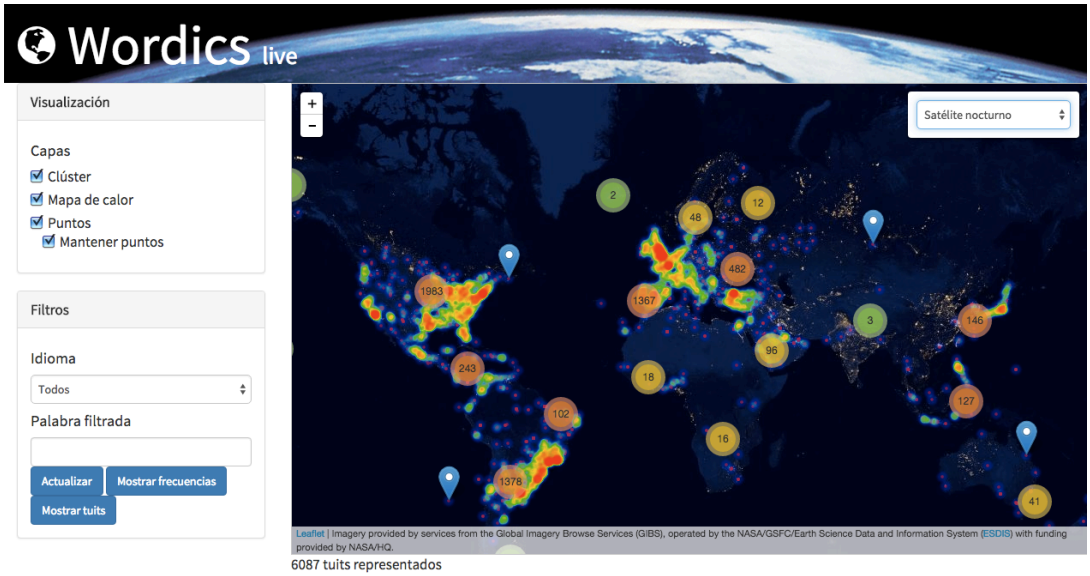


Figura 5.9. Interfaz de *Wordics Live* con mapa satélite nocturno

En lo que respecta a la visualización de los datos, también existen distintas posibilidades. En la esquina superior izquierda podemos seleccionar tres capas distintas de representación de datos: clúster, mapa de calor y puntos.

Un clúster representa un área geográfica determinada en la que se agrupa un conjunto de tuits, de forma que se optimiza la visualización por zonas. Cada clúster está representado por un círculo que va cambiando de color según la densidad de tuits que se van publicando en ese clúster concreto. Al situar el ratón sobre la circunferencia, se dibuja una silueta que delimita la zona abarcada por el clúster. Conforme nos movemos por diferentes niveles de zoom del mapa, los clústeres se van subdividiendo en otros de menor tamaño hasta llegar a la unidad, representada por una marca azul en forma de lágrima invertida denominada "marca de posición". Si clicamos sobre la marca de posición, podremos ver el tuit original completo.

En la figura 5.10 hemos seleccionado un clúster de 219 palabras, generado automáticamente, que engloba a la mitad sur de Reino Unido, la mitad norte de Francia y los Países Bajos. El resto de los círculos de distintos colores también representan clústeres que se corresponden con otras zonas geográficas.

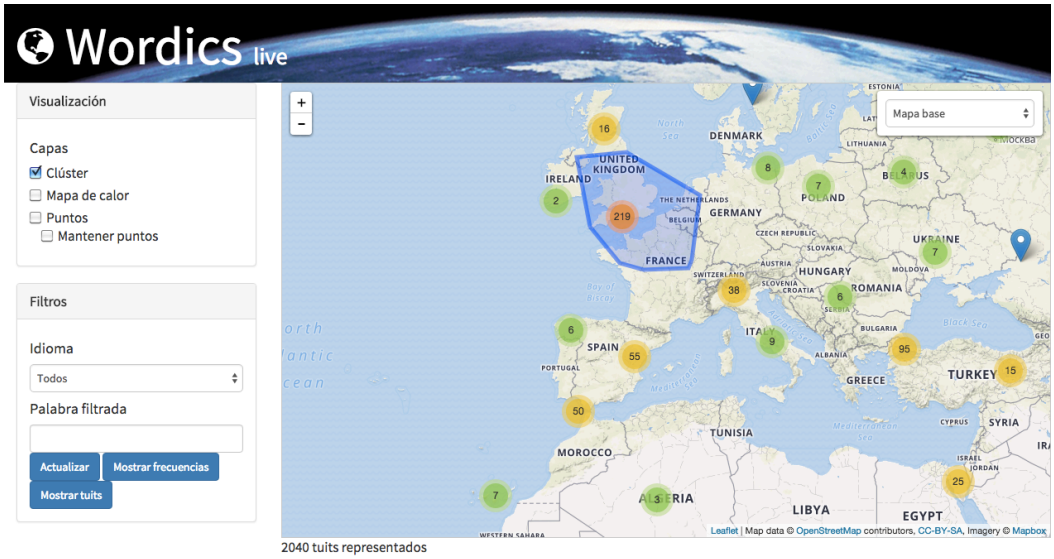


Figura 5.10. Visualización de los tuits con clústeres

En segundo lugar, se puede seleccionar la visualización mediante mapas de calor (figura 5.11). De esta forma podemos ver rápidamente en el mapa la densidad de publicaciones en todo el mundo. Conforme los colores se vayan acercando al rojo, esto significa que hay mayor densidad de tuits, mientras que los tonos más cercanos al violeta indican menor densidad. En el caso de que no se publique ningún tuit en una zona determinada, no aparecerá en ella ningún color.

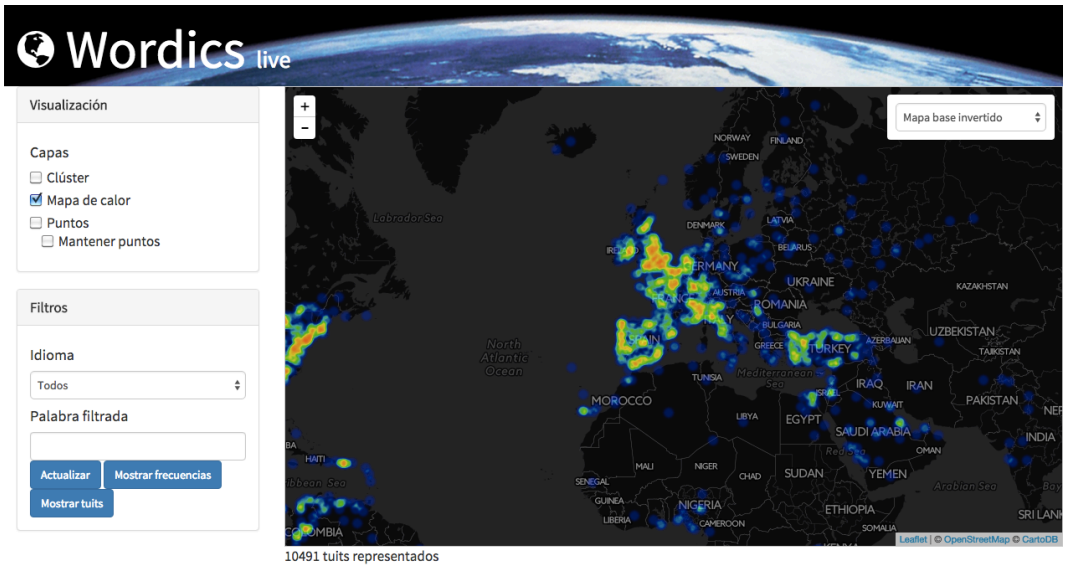


Figura 5.11. Visualización de los tuits con mapa de calor

Por último, podemos utilizar puntos para representar tuits individuales que aparecerán por las distintas zonas del mapa, según se vayan publicando. Aquí tenemos la posibilidad o bien de marcar una pestaña para que los tuits se mantengan en el mapa conforme aparecen –desde el momento en el que comienza la búsqueda– o bien de desactivar esta opción, de forma que los puntos se encienden cuando se detecte la publicación y se van apagando progresivamente. Mientras que con la opción de mantener los puntos podemos apreciar lo que está ocurriendo desde el momento en el que se empieza a filtrar, la segunda opción es más dinámica y nos permite ver de un solo golpe de vista los lugares desde los que se envían los mensajes y el número mensajes producidos.

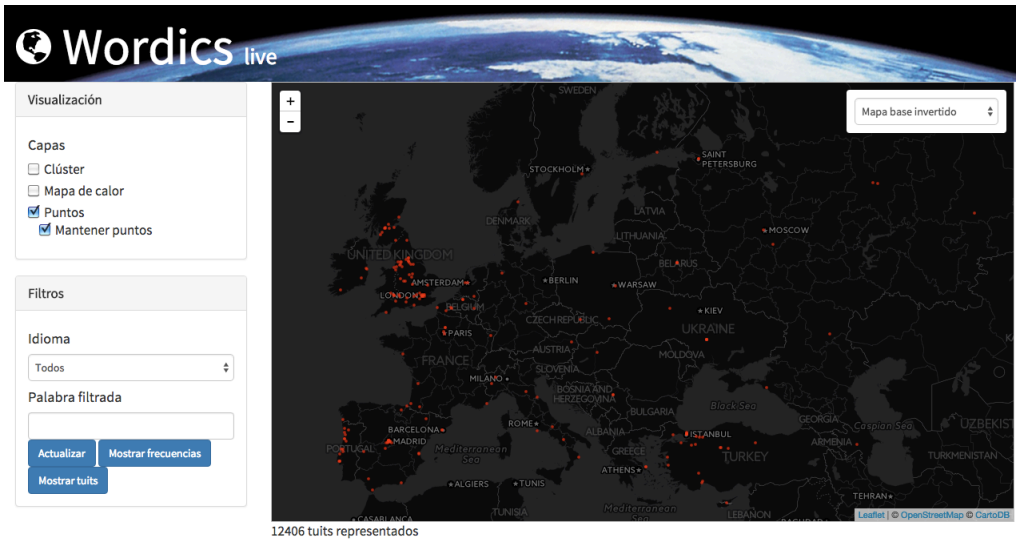


Figura 5.12. Visualización de los tuits con puntos

5.3.3.2 *Wordics One*

El segundo módulo de *Wordics* toma se denomina *Wordics One* (ver pantalla de inicio en figura 5.13) y nace con el objetivo de llevar a cabo estudios lingüísticos –en la línea de los otros dos módulos– centrados, ya no en la utilización general de *Twitter*, sino en cuentas de usuarios concretas, específicamente escogidas para el objeto de estudio en cuestión.

De esta forma, mientras que en *Wordics Live* y en *Wordics Archive* los distintos tipos de análisis se realizan sobre todos los usuarios que publican posts, en tiempo real o no, y filtrados o no por idioma, región, palabra, fecha, etc., con *Wordics One* tenemos la posibilidad de seleccionar una o varias cuentas determinadas sobre las que pretendemos, de manera específica, realizar los distintos tipos de análisis. Así, podemos seleccionar la cuenta de cualquier usuario particular (conocido o no), asociación, empresa, organización, partido político, etc., para poder llevar a cabo estudios específicos para ellos e, incluso, poder compararlos entre sí.

A diferencia de los otros dos módulos del programa, *Wordics One* no presenta mapas para la visualización de los tuits. La razón por la que se ha diseñado de esta manera se debe a que, ya que el objeto de estudio en este caso se centra en un usuario concreto, no tiene sentido la localización en el mapa.

Por el contrario, la primera pantalla que aparece consiste simplemente en un buscador en el que debemos introducir el nombre de usuario, bien precedido del símbolo "@"[42] o bien sin él, para que el sistema proceda a la búsqueda de la cuenta en cuestión. Antes de ello, es posible filtrar el tamaño de la muestra que deseemos, en función del tipo de análisis requerido. De esta forma, podemos seleccionar tamaños de muestra pequeño, mediano o grande. Estos valores calculan hasta doscientas peticiones de búsqueda, hasta mil ochocientas y hasta tres mil seiscientas, respectivamente. Esta limitación viene impuesta por la API de *Twitter*; la herramienta elimina los retuits para que cuando analicemos un usuario concreto nos aseguremos de que es él quien escribe. El motivo por el cual se lleva a cabo esta división es que, para estudios sencillos o preliminares, no sea necesario hacer grandes cargas de datos, lo que aumentará la velocidad de procesamiento.

[42] Véase página 156.

Figura 5.13. Pantalla de inicio de *Wordics One*

5.3.3.2.1 Información sobre la cuenta y los tuits

Una vez filtrado el usuario para ser analizado, encontramos, en primer lugar, información acerca de la cuenta concreta objeto de estudio (ver figura 5.14). Esta información responde al nombre real o ficticio que el administrador de la cuenta haya especificado y todos los datos que quiera aportar sobre sí mismo. Estos pueden abarcar desde información personal, hasta gustos, trabajo, aficiones u otros datos de contacto. En el caso de que se trate de una institución o empresa, se suelen aportar los datos básicos para su identificación. También la herramienta nos permite ver la foto de perfil y el número total de tuits publicados en esa cuenta.

Figura 5.14. Información acerca de la cuenta de usuario en *Wordics One*

Al mismo tiempo, podemos ver una tabla de texto en la que aparecen los tuits publicados en esa cuenta. En esta tabla, se muestra la información perteneciente a la fecha de publicación –a la izquierda–, el cuerpo de texto del tuit –en el centro– y el idioma en el que está escrito –a la derecha. Nótese que, dentro de la misma tabla, es posible realizar un filtrado de texto para que los tuits que aparezcan sean

exclusivamente los que contengan una determinada palabra, así como un filtrado de idioma. En la parte inferior de la tabla se puede personalizar el tipo de visualización, especificando el número de resultados que se desea que se muestren por página. Las posibilidades son: diez, veinticinco, cincuenta o cien tuits.

En la tabla siguiente (figura 5.15) podemos ver la información que nos devuelve la aplicación cuando se le solicitan los tuits publicados por la cuenta de la Real Academia Española. Los tuits aparecen en orden cronológico inverso, de manera que el primero que leemos es el último que se ha escrito. En este ejemplo mostramos diez ejemplos por página para que resulte más cómoda la investigación.

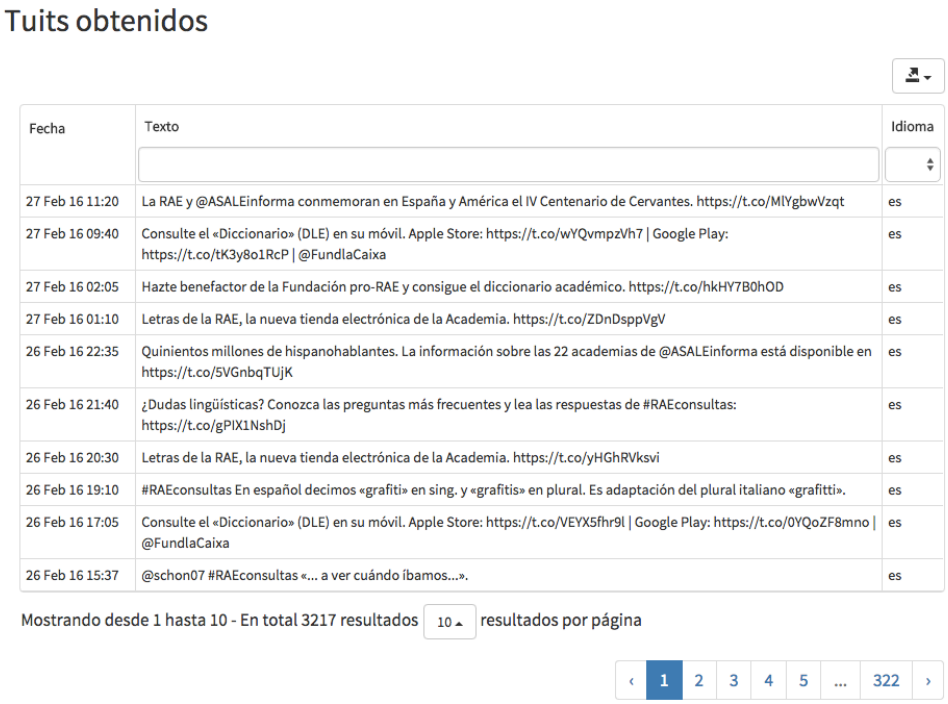


Figura 5.15. Tabla de tuits obtenidos con *Wordics One* de la cuenta de la Real Academia Española

Como hemos explicado, en esta tabla es posible filtrar una palabra, o bien el idioma en que se escribe. En este caso (figura 5.16), hemos filtrado la palabra *asín* para ver los comentarios de la RAE respecto de su uso, o si la utiliza en algún contexto concreto. No olvidemos que la herramienta realiza los filtrados de palabras por expresiones regulares, con lo que es posible que la herramienta esté devolviendo resultados de la propia palabra o de otras que contengan esas mismas letras en orden.

Figura 5.16. Tabla de tuits de la cuenta de la Real Academia Española filtrando la palabra *asín*

En la esquina superior derecha de la tabla anterior observamos un icono que volverá a aparecer en el resto de tablas y gráficos que obtendremos a lo largo de todo el proceso. El icono sirve para exportar los resultados a distintos formatos que podremos seleccionar según nuestras necesidades. Como el usuario de *Wordics* podrá comprobar, existen dos iconos distintos para la exportación: uno de ellos es el que acompaña a las tablas que contienen información textual, y el otro aparece para las gráficas que pueden ser exportadas como imagen. Así, recordemos que los cuadros de texto podrán ser exportados en los siguientes formatos: JSON, XML, CSV, TXT, SQL y MS-EXCEL. Por otro lado, los dos formatos entre los que tendremos la posibilidad de elegir para exportar imágenes son: JPEG y PNG.

A continuación (figura 5.17) mostramos una gráfica que representa la evolución de la frecuencia de publicación de tuits de ese usuario desde la fecha en la que comienza el recuento de posts hasta la actualidad, lo que nos facilita una rápida visualización de la actividad de la cuenta. Además, si situamos el ratón del ordenador sobre las columnas de la gráfica, aparecerá un cuadro de información acerca de la fecha exacta de la publicación (en el color correspondiente) y el número de tuits escritos a lo largo de ese día concreto (en color negro). El botón de exportar, en este caso, también está ubicado arriba a la derecha pero, como ya hemos explicado, difiero del botón que utilizamos para exportar texto porque se trata de una imagen y los formatos a los que se exporta son distintos. En el caso que indicamos a continuación (figura 5.17), la imagen se ha

guardado como PNG. Mostramos la frecuencia de los tuits que hemos filtrado anteriormente; es decir, los tuits publicados por la RAE que contienen la palabra *asín*.

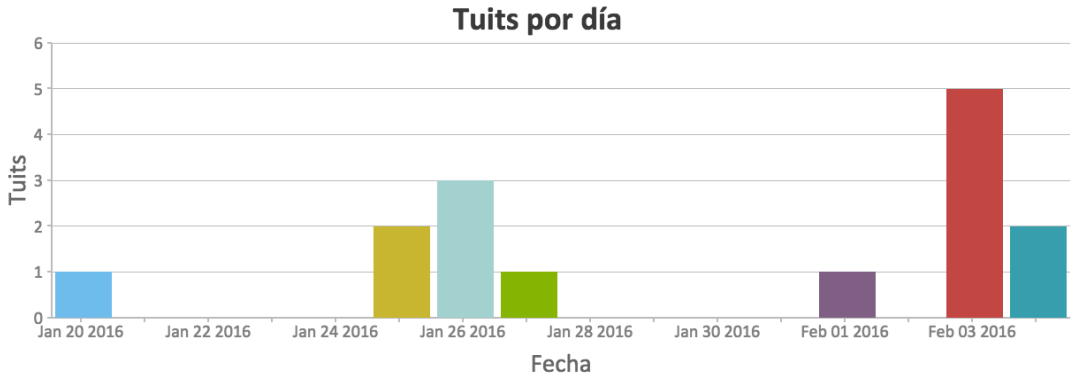


Figura 5.17. Gráfica de frecuencias de aparición de tuits diarios de la Real Academia Española que contienen la palabra *asín*, exportada en formato PNG.

Para ampliar la gráfica, basta con seleccionar el fragmento deseado y su tamaño aumentará de forma automática. Utilizar esta función no tendría mucho sentido en una gráfica como la de la RAE para la palabra *asín* puesto que el número de tuits es bastante reducido; sin embargo, para una cuenta muy activa o una palabra que se utilice frecuentemente, resulta de gran utilidad, puesto que la gráfica general tiene la información muy condensada. Por ejemplo, si observamos la actividad en la cuenta de Jack Dorsey (@jack), del que ya hemos hablado en el capítulo cuarto, podemos ver las ventajas de realizar esta ampliación. Esta gráfica (figura 5.18) muestra todos los tuits publicados por el cofundador de *Twitter* desde enero de 2015 hasta febrero de 2016, con más de 1200 resultados:

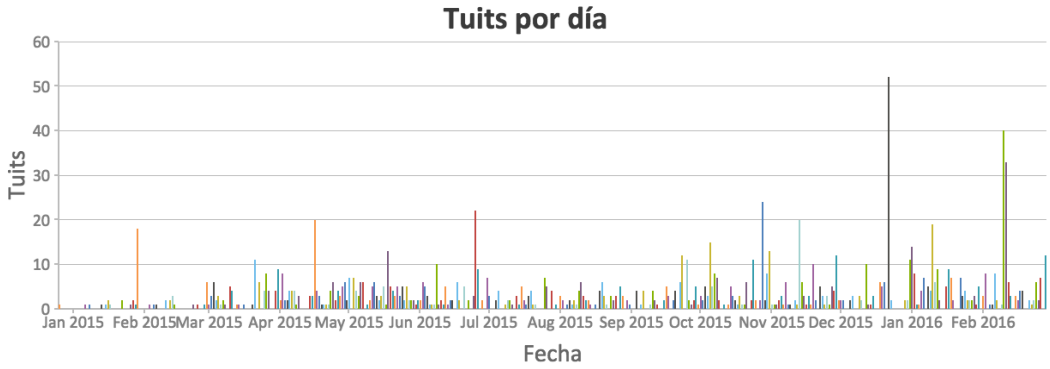


Figura 5.18. Tuits publicados por Jack Dorsey desde enero de 2015 hasta el 28 de febrero de 2016

En la figura 5.19, en cambio, hemos ampliado la franja correspondiente a febrero de 2016 –simplemente seleccionándola con el ratón–, para que resulte más fácil y cómodo interpretar los resultados:

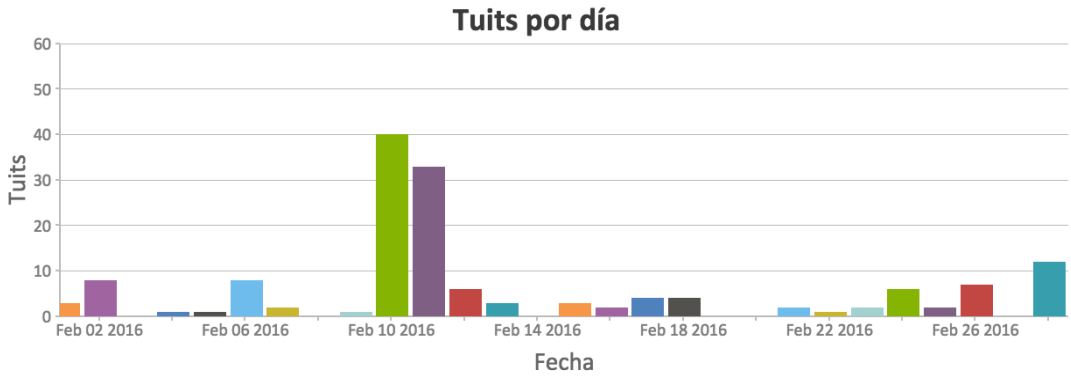


Figura 5.19. Ampliación de los tuits publicados por Jack Dorsey, correspondientes al mes de febrero de 2016

5.3.3.2.2 Análisis lingüísticos

Hasta aquí, la información que aporta *Wordics One* se ha limitado a presentar el corpus con el contenido, la fecha o el idioma de los tuits obtenidos, ya sea en la tabla de texto o en la gráfica que muestra la frecuencia de utilización de *Twitter* por ese usuario. Además del propio corpus, la información aportada por las gráficas anteriores (figura 5.17, figura 5.18 o figura 5.19) es fundamental para la contextualización del estudio, pero que no es específica para un análisis lingüístico, a pesar de que ya se puede comenzar a analizar el texto mostrado en la primera tabla (figura 5.16).

Las funciones de este módulo diseñadas específicamente para la investigación lingüística no difieren de las que podremos realizar en el tercer módulo de la herramienta –*Wordics Archive*–, aunque en este último se ampliarán las posibilidades, puesto que se podrá analizar mucha más información y realizar filtrados por zonas y por idiomas. La diferencia fundamental, como decimos, estriba en el objeto de estudio: mientras que en el último módulo (*Wordics Archive*) el estudio se centra en la totalidad de usuarios que escriben en un idioma, lugar o período de tiempo determinados, aquí (*Wordics One*) se analizan los tuits que hayan sido publicados por una o varias cuentas de *Twitter* concretas.

Por lo tanto, una vez seleccionados los tuits que vamos a estudiar, podemos empezar a realizar un análisis más estrictamente lingüístico. Este estudio se llevará a

cabo a partir de los parámetros que tengamos establecidos desde el principio. Por lo tanto, se ajustarán al tamaño de muestra seleccionado, así como a los tuits que nos aparezcan en la tabla inicial y en la gráfica que le sucede. Es decir, si no hemos predeterminado ningún filtro, ni por idioma ni por palabra de la totalidad de los tuits de la cuenta, los análisis que realizaremos a partir de ahora se llevarán a cabo sobre toda la muestra. Si, por el contrario, filtráramos alguna palabra o algún idioma, la herramienta lo tendrá en cuenta para el estudio.

Así, las funciones que nos ofrece *Wordics One*, desde este punto de vista, son:

a) Análisis de las frecuencias de palabras.
b) Análisis de los idiomas utilizados.
c) Análisis de la densidad léxica (relación tipo/caso).
d) Análisis de las colocaciones.
e) Análisis de las palabras clave en contexto (KWIC).

5.3.3.2.2.1 Análisis de las frecuencias de palabras

El programa realiza una búsqueda de la frecuencia de aparición de las palabras contenidas en los tuits que hayan sido filtradas con anterioridad. La forma de visualización se plantea con un doble formato que permite, por un lado, reconocer fácil y rápidamente de forma gráfica cuáles son las palabras más utilizadas y, por el otro, tener un análisis completo de todas las palabras que encontramos en esos tuits.

La primera de las funciones las realiza un gráfico de colores en forma de rueda que solo muestra las treinta palabras más utilizadas, asignándole a cada una de ellas un color (que viene detallado en una leyenda en la parte inferior). El hecho de que, en este caso, únicamente se muestren las treinta primeras palabras se justifica con el propósito de este primer gráfico, que no es otro que realizar una primera aproximación de manera visual. Al situar el ratón del ordenador sobre las distintas franjas coloreadas, podremos ver el número exacto de veces en las que se da la palabra correspondiente a ese color. En la figura siguiente (figura 5.20) se muestra un análisis de la frecuencia de palabras solamente en los tuits de la RAE que contienen la palabra *asín*, gráfico que hemos exportado a formato PNG:

Figura 5.20. Gráfico en forma de rueda que representa las palabras más utilizadas, exportado en formato PNG

Para un análisis más exhaustivo, al lado del gráfico de rueda aparece un cuadro de texto con todas las palabras contenidas en los tuits analizados y el número de ocurrencias de cada una de ellas. Con esta función se completa el estudio, porque el recuento se realiza sobre la totalidad de las palabras. Como en el resto de los cuadros de texto, tenemos la opción de buscar la palabra que nos interese o de leerlas en su conjunto, una por una (seleccionando el número de resultados que deseamos ver por página), teniendo siempre en cuenta que aparece, en primer lugar, la palabra que mayor número de ocurrencias posee. En el siguiente ejemplo (figura 5.21) vemos la tabla de frecuencias de los tuits en los que aparece la palabra *asín* y observamos que, entre todos los que hay publicados, suman 86 palabras distintas (tipos), cada una, como es lógico, con un índice de aparición determinado. Como podemos ver, tenemos también la posibilidad de exportar los datos para almacenarlos más cómodamente y de acuerdo con nuestras necesidades, de cara a la investigación.

Figura 5.21. Tabla del total de ocurrencias de palabras en los tuits filtrados con *asín*

5.3.3.2.2.2 Análisis de los idiomas utilizados

En la misma pantalla, a la derecha del cuadro de frecuencias, podemos ver una figura en forma de rueda, muy similar a la primera en la que veíamos las treinta palabras más utilizadas pero, esta vez, con la distribución de idiomas en los que se escriben los tuits filtrados. Recordemos que, cuando aparece *idioma indeterminado*, se debe a que el sistema no es capaz de reconocer el texto de ningún tuit en concreto debido a que posiblemente haya más de un idioma junto o a la presencia de algún emoticono o símbolo no reconocible. En la gráfica que se produce con *asín* (figura 5.22) lógicamente, solo aparece un idioma, puesto que todos los tuits de la RAE están escritos en español:

Figura 5.22. Gráfico en forma de rueda que representa los idiomas utilizados en el filtrado realizado, exportado en formato PNG

5.3.3.2.2.3 Análisis de la densidad léxica (relación tipo/caso)

Al igual que ocurrirá en el siguiente módulo, *Wordics One* ofrece la posibilidad de establecer la relación entre el número total de palabras (caso-*token*) y el número de palabras distintas que aparecen (tipo-*type*). Esta relación se establece en forma de porcentaje en la misma tabla en la que aparece la información sobre los tipos y los casos de forma independiente. Como ya hemos explicado, cuanto mayor sea el porcentaje final, mayor será la diversidad léxica del corpus. Aunque es importante saber interpretar bien los datos para no caer en las limitaciones propias de esta técnica. En el caso de la cuenta de la RAE y de los tuits que contienen *asín* (figura 5.23), vemos que:

a) el número de casos es 288
b) el número de tipos es 87, como ya vimos en la tabla de frecuencias
c) la densidad de la lengua, en esta muestra, es del 30,21%

Esta información, en la página web, se muestra de la siguiente forma:

Descripción	Valor
Tipos (Type)	87
Casos (Token)	288
Densidad de la lengua	30.21%

Figura 5.23. Relación tipo/caso de los tuits de la Real Academia Española que contienen la palabra *asín* en *Wordics One*

No olvidemos que esta información, como en casos anteriores, puede ser exportada a distintos formatos para facilitar la investigación y el almacenamiento de la información.

Para obtener datos más precisos acerca de la densidad y poder establecer comparaciones entre distintos tipos de lenguajes, usuarios, etc., hemos elaborado una gráfica que permite segmentar los casos. Mediante esta gráfica (figura 5.25, figura 5.26 y figura 5.27), es posible situar el ratón del ordenador en un punto concreto de la línea que representa a los casos y se nos mostrará la cantidad de tipos que se producen dentro de esos casos. La línea diagonal roja representa el total, desde el momento en el que se

empieza a recopilar palabras del corpus hasta que se termina; siempre, por lo tanto, representa el cien por cien. Cuanto más cerca de esta línea diagonal de color rojo se encuentre la línea verde, más densidad léxica existirá en ese momento.

Por ejemplo, en el caso de la RAE, el total de tuits recopilados alcanza la cifra de 3204, que suman un total de 49.211 casos y 8.124 tipos. La densidad, por tanto, quedaría así:

Descripción	Valor
Tipos (Type)	8124
Casos (Token)	49211
Densidad de la lengua	16.51%

Figura 5.24. Relación tipo/caso de todos los tuits de la Real Academia Española analizados con *Wordics One*

Como explicaremos más adelante (sección 7.1), en el estudio que realizaremos de la densidad léxica en varios escritores españoles, las cuestiones teóricas relacionadas con este aspecto son varias y requieren tener en cuenta una serie de consideraciones. Una de las más importantes está relacionada con la longitud del texto, que, por la naturaleza del método que obtiene los datos de densidad, provoca que esta se vea influida por la cantidad de palabras.

Por ello, y para poder obtener una información más precisa, sobre todo a la hora de tener la posibilidad de comparar con otras cuentas, es necesario acudir a la gráfica de densidad segmentada. Aquí tenemos la posibilidad, como acabamos de mencionar, de marcar un punto determinado para conocer la variedad léxica en el momento seleccionado.

En el mismo ejemplo de los 49.211 casos de la RAE, hemos examinado la gráfica en tres puntos distintos: al principio, coincidiendo con los 10.000 casos; algo más hacia delante de la mitad, con 30.000 casos; y casi al final, con 49.000 casos:

Densidad de la lengua con 10.000 casos:

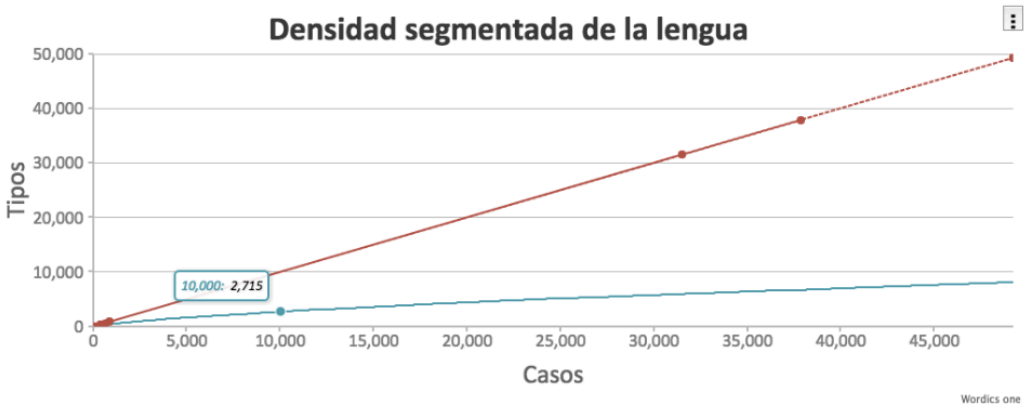


Figura 5.25. Densidad segmentada de los tuits de la RAE con 10.000 casos

Densidad de la lengua con 30.000 casos:

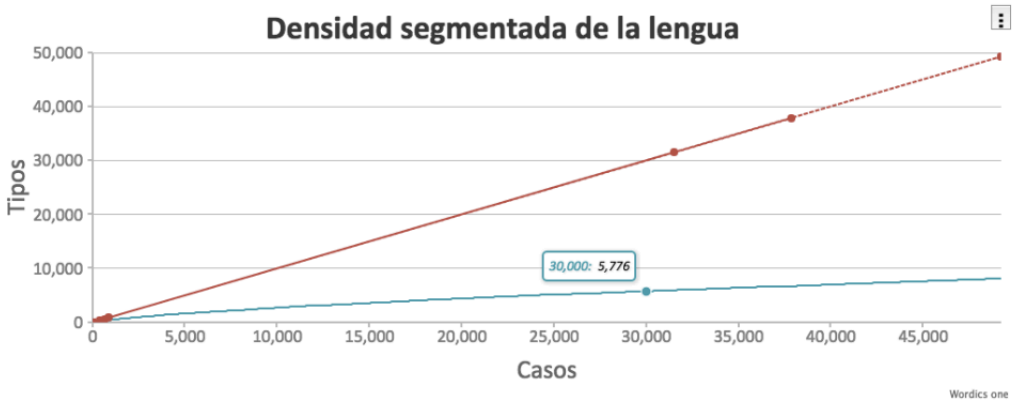


Figura 5.26. Densidad segmentada de los tuits de la RAE con 30.000 casos

Densidad de la lengua con 49.000 casos:

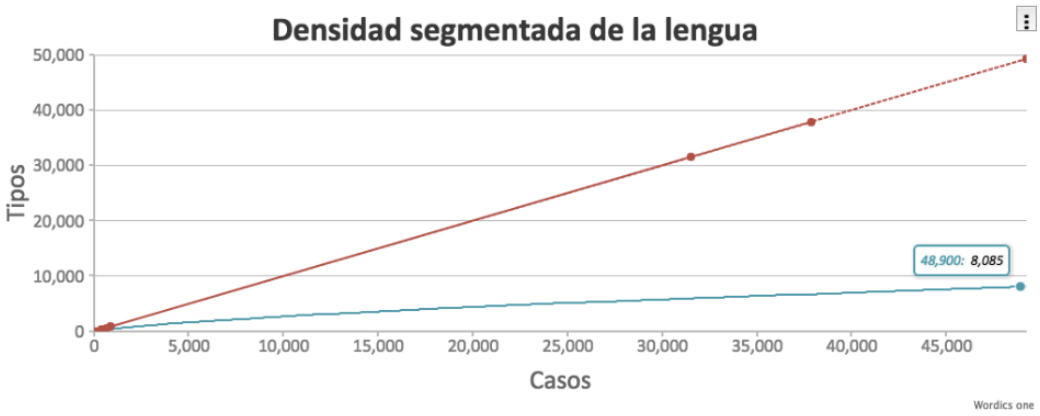


Figura 5.27. Densidad segmentada de los tuits de la RAE con 49.000 casos

Como vimos en la primera tabla, para el total de palabras analizadas (49.211) la densidad léxica era del 16,51%, puesto que el número de tipos era 8.124. Sin embargo, con las gráficas segmentadas obtenemos los siguientes datos:

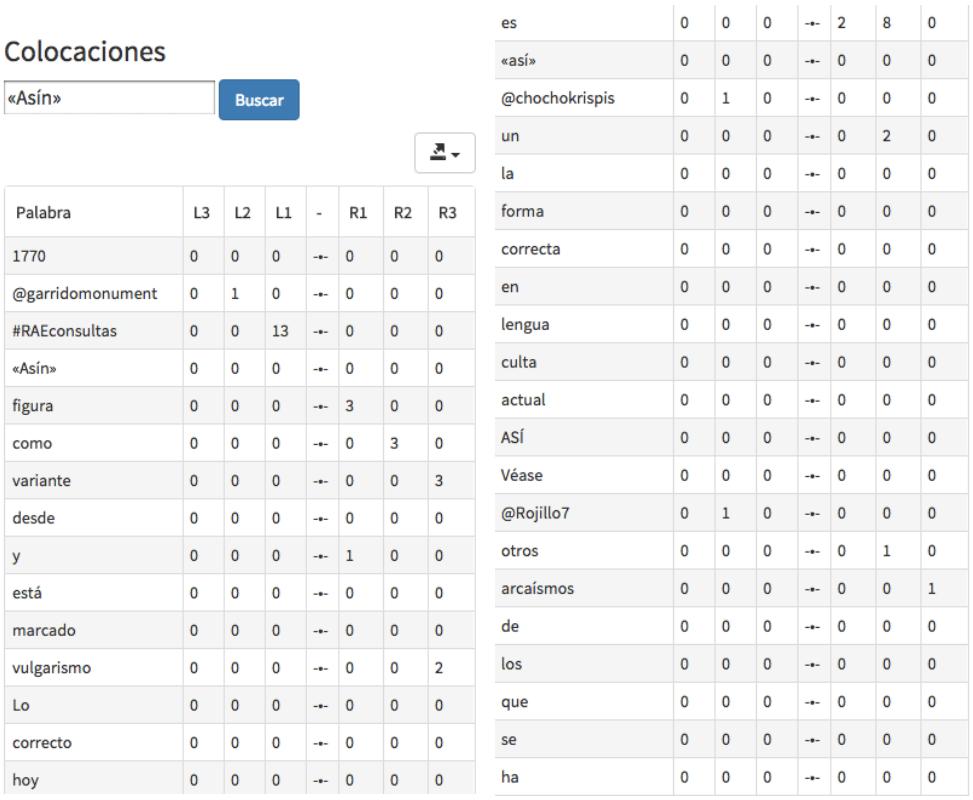
Casos	Tipos	Densidad
10.000	2.715	27,1%
30.000	5.776	19,3%
49.000	8.085	16,5%

Figura 5.28. Comparación de las densidades segmentadas de los tuits de la RAE

Podemos observar, tras estos resultados, cómo la relación entre la longitud del texto y la densidad es inversamente proporcional, lo que hace que consideremos extremadamente útil esta utilidad de la herramienta porque nos permite hacer comparaciones entre distintos textos en cuanto a su densidad.

5.3.3.2.2.4 Análisis de las colocaciones

Una vez más, tenemos la posibilidad de estudiar la manera en que las palabras se distribuyen dentro de la oración en relación con las palabras que la rodean. Mientras que las funciones anteriores muestran los resultados de manera automática, en la función de análisis de las colocaciones y de las palabras clave en contexto no ocurre lo mismo, de modo que, si queremos utilizar estas funciones, debemos solicitárselo al programa. En cualquier caso, la forma de hacerlo es sumamente sencilla: para ello, solamente debemos introducir en el buscador la palabra que vayamos a estudiar y, a continuación, pulsar la tecla *intro* o el botón de *buscar*. La información obtenida puede, claro está, ser exportada a otro formato. A continuación (figura 5.29), mostramos el ejemplo de la palabra asín con las 36 primeras colocaciones:



Colocaciones

«Asín» Buscar

Palabra	L3	L2	L1	-	R1	R2	R3
1770	0	0	0	-•-	0	0	0
@garridomonument	0	1	0	-•-	0	0	0
#RAEconsultas	0	0	13	-•-	0	0	0
«Asín»	0	0	0	-•-	0	0	0
figura	0	0	0	-•-	3	0	0
como	0	0	0	-•-	0	3	0
variante	0	0	0	-•-	0	0	3
desde	0	0	0	-•-	0	0	0
y	0	0	0	-•-	1	0	0
está	0	0	0	-•-	0	0	0
marcado	0	0	0	-•-	0	0	0
vulgarismo	0	0	0	-•-	0	0	2
Lo	0	0	0	-•-	0	0	0
correcto	0	0	0	-•-	0	0	0
hoy	0	0	0	-•-	0	0	0
es	0	0	0	-•-	2	8	0
«así»	0	0	0	-•-	0	0	0
@chochokrispis	0	1	0	-•-	0	0	0
un	0	0	0	-•-	0	2	0
la	0	0	0	-•-	0	0	0
forma	0	0	0	-•-	0	0	0
correcta	0	0	0	-•-	0	0	0
en	0	0	0	-•-	0	0	0
lengua	0	0	0	-•-	0	0	0
culta	0	0	0	-•-	0	0	0
actual	0	0	0	-•-	0	0	0
ASÍ	0	0	0	-•-	0	0	0
Véase	0	0	0	-•-	0	0	0
@Rojillo7	0	1	0	-•-	0	0	0
otros	0	0	0	-•-	0	1	0
arcaísmos	0	0	0	-•-	0	0	1
de	0	0	0	-•-	0	0	0
los	0	0	0	-•-	0	0	0
que	0	0	0	-•-	0	0	0
se	0	0	0	-•-	0	0	0
ha	0	0	0	-•-	0	0	0

Figura 5.29. 36 primeras colocaciones de la palabra *asín* en *Wordics One*

5.3.3.2.2.5 Análisis de palabras clave en contexto (KWIC)

De igual forma que ocurre con las colocaciones, para poder ver una palabra en contexto, introduciremos la palabra deseada en el buscador y obtendremos los resultados correspondientes, que podrán ser exportados en cualquiera de los formatos que ya se han explicado. Veamos, por último, el ejemplo del análisis de *asín* realizado con la técnica de KWIC:

Figura 5.30. KWIC de la palabra *asín* con *Wordics One*

5.3.3.3 *Wordics Archive*

Wordics Archive es el nombre del tercer y último módulo de la herramienta que aquí presentamos. Como ya hemos anunciado en varias ocasiones, la idea en torno a la cual gira su construcción no es sino la de poder realizar análisis históricos de los datos contenidos en *Twitter*. En otras palabras, mientras que las dos primeras partes de *Wordics* –*Wordics Live* y *Wordics One*– nos ofrecen la posibilidad de llevar a cabo estudios del lenguaje en tiempo real y de usuarios concretos de *Twitter*, respectivamente, ahora se nos abre una puerta al análisis diacrónico del lenguaje producido en esta plataforma social, gracias al almacenamiento constante de la información en una base de datos.

Mediante esta base de datos propia, las investigaciones con *Wordics Archive* son menos restrictivas que en los módulos anteriores, puesto que ya no es necesario ajustarse a las limitaciones impuestas por la API de *Twitter* para acceder a la información de las cuentas individuales. Por otra parte, a pesar de las numerosas ventajas que ofrece el análisis en tiempo real (*Wordics Live*), su propia naturaleza basada en millones de datos efímeros nos hace cuestionarnos la conveniencia de llevar a cabo estudios lingüísticos exhaustivos como los que ofrecen los dos módulos

posteriores (*Wordics One* y *Wordics Archive*). No obstante, como ya hemos explicado, las tablas de Excel generadas en *Wordics Live*, junto con las capas de visualización que se mantienen, permiten rebajar, en cierta medida, este carácter fugaz de los datos, puesto que van acumulando toda la información generada desde el momento en que comienza a realizarse el filtrado, con la posibilidad de recuperar posteriormente esa información guardada.

En cualquier caso, consideramos que el módulo que nos ocupa en este apartado, *Wordics Archive*, debido a su acceso a nuestra propia base de datos construida a partir de la API de *Twitter,* amplía enormemente el abanico de posibilidades en cuanto a la investigación lingüística y puede dar lugar, por tanto, a una mayor diversidad de estudios, como demostraremos en el capítulo siguiente. En este módulo de análisis histórico tenemos la posibilidad de realizar filtrados por idiomas, zonas geográficas y una o varias palabras de manera simultánea. Una vez seleccionados los parámetros previos a la búsqueda y obtenidos los datos, podremos proceder a análisis más tradicionales y habituales de los corpus informatizados (ya explicados anteriormente: frecuencias de palabras, KWIC, etc.).

El módulo *Wordics Archive* se divide en dos partes diferenciadas, dependiendo del tipo de análisis que se desee llevar a cabo. Si el objetivo se centra en estudiar de forma general la información textual generada en *Twitter* o la relativa a una palabra o término específico, debemos seleccionar el tipo de *análisis simple* que aparece en la página de inicio (figura 5.31). Si, por el contrario, nos interesa poder contrastar, comparar o estudiar más de una palabra de forma paralela, escogeremos la opción de *análisis comparativo*:

Figura 5.31. Pantalla de inicio de *Wordics Archive*

Con cualquiera de las dos posibilidades, la pantalla de inicio es la misma. Encontraremos en ella un buscador para introducir, en primer lugar, el término o los términos de la búsqueda. Si dejamos en blanco este campo y pulsamos el botón *Analizar* sin determinar la búsqueda, *Archive* devolverá todos los tuits publicados pertenecientes a la zona y a la fecha seleccionadas (en caso de que se seleccionen).

Como podemos observar en la imagen siguiente (5.32), es posible delimitar el área de estudio. Esta tarea se puede desempeñar por dos vías distintas. Por un lado, podemos seleccionar la zona de interés dibujándola dentro del mapa que se despliega al pulsar en la opción *Mostrar mapa*. Explicar aquí el mapa:

Figura 5.32. Mapa de selección de región geográfica objeto de estudio en *Wordics Archive*

Por otro lado, si conocemos con exactitud las coordenadas geográficas de la zona en la que vamos a delimitar nuestro estudio, podemos introducir directamente los datos relativos a la latitud y la longitud, y la herramienta nos dirigirá a ella.

Además, la opción de *elegir palabra* nos permite realizar búsquedas mediante comodines, como hemos explicado unas páginas más arriba.

El último parámetro que podemos establecer antes de que el sistema proceda a la búsqueda de información es el relativo a las fechas. Una vez más, tenemos la posibilidad de establecer un período concreto, fijando manualmente la fecha de inicio y la de fin, o de comenzar la búsqueda sin delimitar fechas. En este último caso, *Archive*

recuperará la información almacenada en nuestro servidor que, como hemos explicado unas líneas más arriba, comienza en agosto de 2015.

Las utilidades de este módulo son numerosas y los estudios de índole lingüística que se pueden derivar de él, enormemente variados. Gracias a *Archive*, no solo es posible determinar la evolución del uso de una o varias palabras, sino que se pueden establecer relaciones entre el comportamiento de las formas léxicas y la zona geográfica donde son utilizadas, y lo mismo en relación con el factor tiempo. Estos datos pueden ser de gran utilidad, por ejemplo, para conocer dónde y cuándo ha empezado a utilizarse un determinado neologismo y cuál ha sido su evolución geográfica y temporal. Por otra parte, si no introducimos ninguna palabra en el buscador inicial, podemos también conocer qué idiomas se hablan en distintas zonas o cuál es la tendencia de uso de determinadas lenguas.

5.3.3.3.1 Análisis simple

Como hemos explicado, el diseño de esta opción está enfocado al análisis, o bien de todo el contenido de *Twitter*, o bien de una palabra o de un término específico. Por ejemplo, si filtramos la palabra *selfi*[43], sin especificar zona geográfica ni rango de fechas, se nos dibuja la siguiente gráfica:

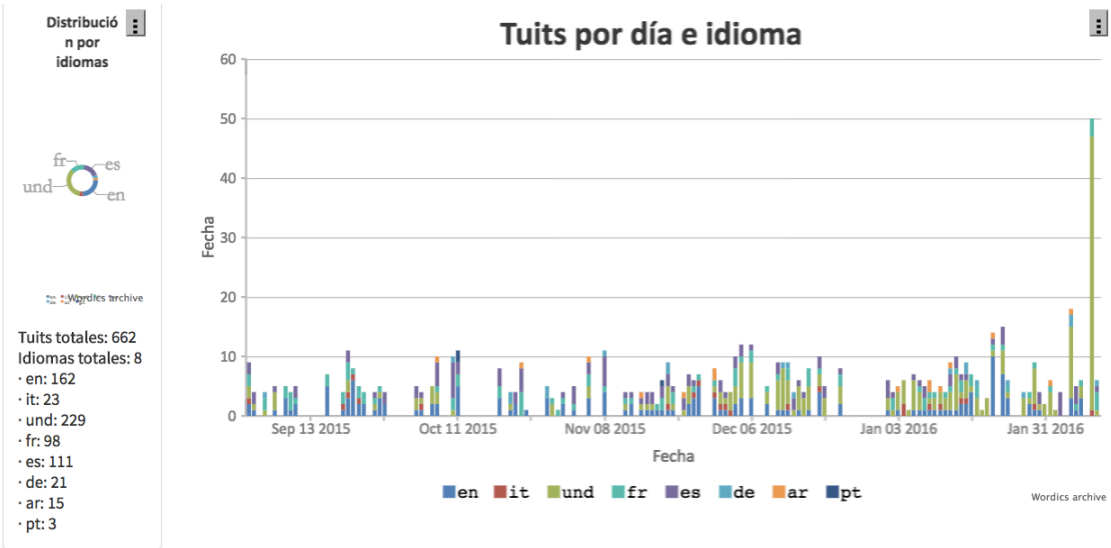


Figura 5.33. Gráfica de la frecuencia de uso de la palabra *selfi*

[43] Marcamos la forma "selfi" en cursiva para indicar que estamos haciendo una reflexión metalingüística de la palabra y es en ese término en el que nos estamos centrando. Cuando utilicemos *selfi* integrado en el discurso, lo haremos en redonda, por indicación de la Fundéu BBVA.

En la gráfica anterior (figura 5.33) aparecen, como se puede observar distintas barras, cada una de ellas con varios colores, pertenecientes a los distintos idiomas en los que aparece esta palabra concreta con esta ortografía determinada. Es lógico que el español, representado por la barra morada, sea el idioma que alcanza los picos más altos, puesto que *selfi* es una adaptación a la ortografía española de la voz inglesa *selfie*. La barra verde supera a la línea morada que representa el español, pero recordemos que se trata de tuits que *Twitter* (y no *Wordics*) es incapaz de reconocer, debido a que utiliza más de un idioma de forma simultánea, o a que contiene enlaces a páginas web, emoticonos, etc. También podemos ver a la izquierda de la figura 5.33 el mismo gráfico en forma de rueda que aparecía en *Wordics One*, gráfico que muestra la distribución por idiomas de los tuits que contienen esta palabra.

La gráfica central es interactiva y nos va señalando el número exacto de tuits conforme vamos pasando el ratón del ordenador por encima de los puntos de las líneas. Además, puesto que la información de la gráfica, por defecto, representa a todos los idiomas en los que, repetimos, se han escrito tuits con la palabra de la búsqueda, la gráfica también nos permite decidir si deseamos ver toda esa información o, por el contrario, preferimos visualizarla por idiomas. Para ello, basta con activar o desactivar los botones de idiomas que aparecen debajo y que, además, sirven de leyenda de la gráfica. Conforme estos idiomas se desactivan, va ampliándose el tamaño de los idiomas que permanecen, para que se pueda ver con mayor claridad. A continuación (figura 5.34), mostramos el resultado de desactivar todos los idiomas, excepto el español:

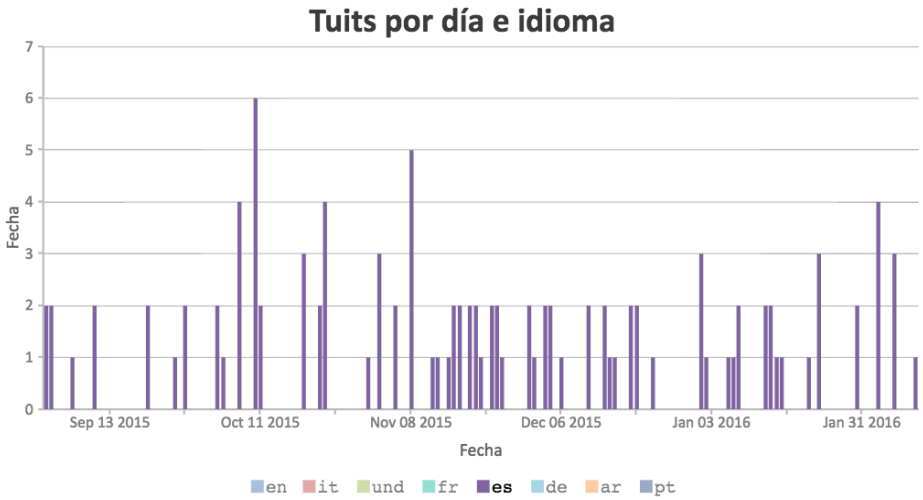


Figura 5.34. Gráfica de la frecuencia de uso de la palabra *selfi* en español

Más abajo (figura 5.35), podemos ver otro mapa que recrea, día por día, los movimientos de la palabra en cuestión. Para ello debemos deslizar la bolita azul que aparece debajo del mapa a lo largo de la línea. De esta manera, se nos muestra la información de la tabla superior en movimiento. En este mapa, también es posible seleccionar el idioma. Debido a la imposibilidad de mostrar este movimiento sobre el papel, mostramos a continuación una captura del mapa, para que el lector sepa a qué nos referimos, aunque lo explicaremos más detenidamente en el DVD que se adjunta al trabajo.

Figura 5.35. Mapa de movimientos de *selfi* en todos los idiomas

A partir de este momento, el resto de utilidades del módulo de análisis simple de *Wordics Archive* son las mismas que llevaban a cabo análisis lingüísticos mediante *Wordics One*, es decir:

a) Lista de frecuencias de aparición de la palabra en cuestión.

b) Cuadro de texto con el corpus obtenido, en el que se puede seleccionar idioma. Como es habitual, esta información se puede exportar a cualquiera de los formatos de texto que hemos visto anteriormente.

c) Palabras clave en contexto (KWIC). En este módulo, se añade la funcionalidad de poder introducir otra palabra en el buscador de KWIC. De esta forma, podemos ver con qué frecuencia aparecen ambas palabras en el mismo tuit y en qué posición.

5.3.3.3.2 Análisis comparado

Esta segunda opción de *Wordics Archive* consiste en una ampliación del módulo de análisis simple que acabamos de explicar. La única diferencia que presenta con respecto a este se resume en la posibilidad de estudiar dos o más términos de forma paralela. Por ejemplo, si queremos conocer qué forma de saludo se utiliza más en el idioma inglés entre *hello* y *hi*, introducimos estos dos términos en el buscador, sin filtros geográficos ni temporales y estos son los resultados:

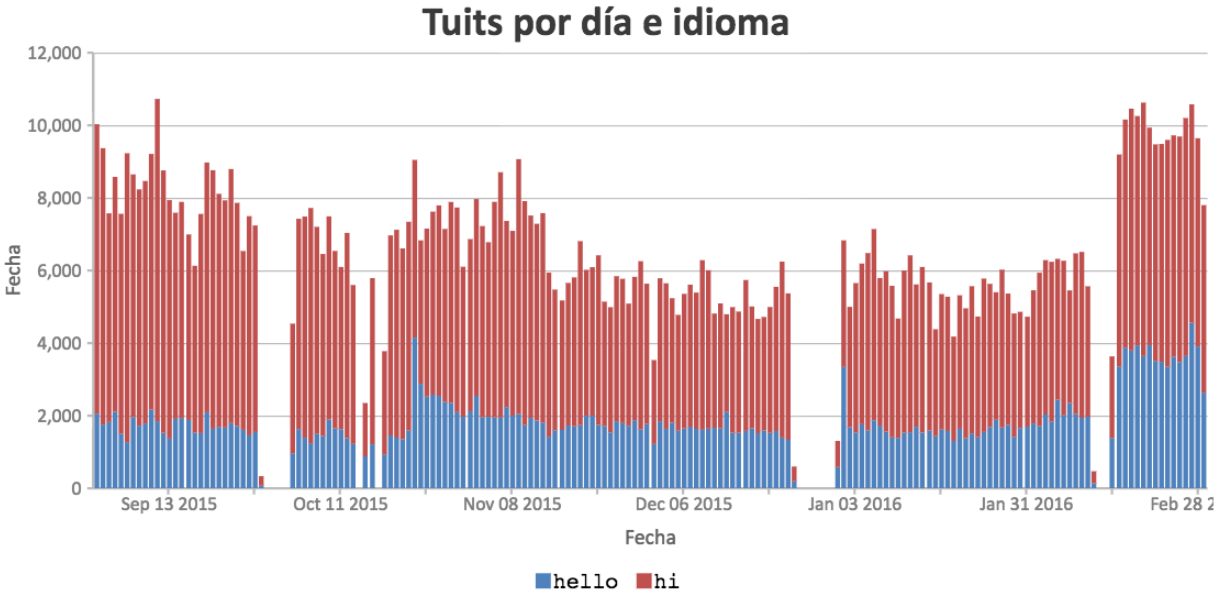


Figura 5.36. Gráfico con la frecuencia de aparición de las formas inglesas *hello* y *hi* en todos los idiomas

Como podemos comprobar en la figura 5.36, el uso del saludo más informal *hi* es mucho más habitual que el de *hello*. Puesto que no hemos aplicado ningún filtro, la gráfica nos muestra todas las ocurrencias de estos dos términos sea cual sea el idioma del tuit en el que aparecen. Para conocer con detalle en qué idiomas se han utilizado, podemos consultar los dos gráficos que aparecen a la izquierda de la pantalla correspondientes a cada una de las búsquedas. Los resultados, por idioma, de estos dos términos, son los siguientes:

IDIOMA	*hello*	*hi*
en	279100	671449
es	8722	16296
ar	679	1621
pt	3248	3378
fr	13897	7038
und	4960	81229
it	2970	3054
De	2536	5760
Nº TOTAL DE TUITS	316112	789825

También podemos ver la distribución geográfica y temporal de los tuits en el mapa interactivo que encontramos debajo del gráfico. Mediante el deslizador de color azul, podemos movernos a través de los días y ver cómo evoluciona el uso de los términos objeto de estudio. Además, las capas asociadas al mapa se pueden activar y desactivar en función de las necesidades. En la figura siguiente (figura 5.37) mostramos una captura concreta de este mapa con los términos *hello* y *hi*:

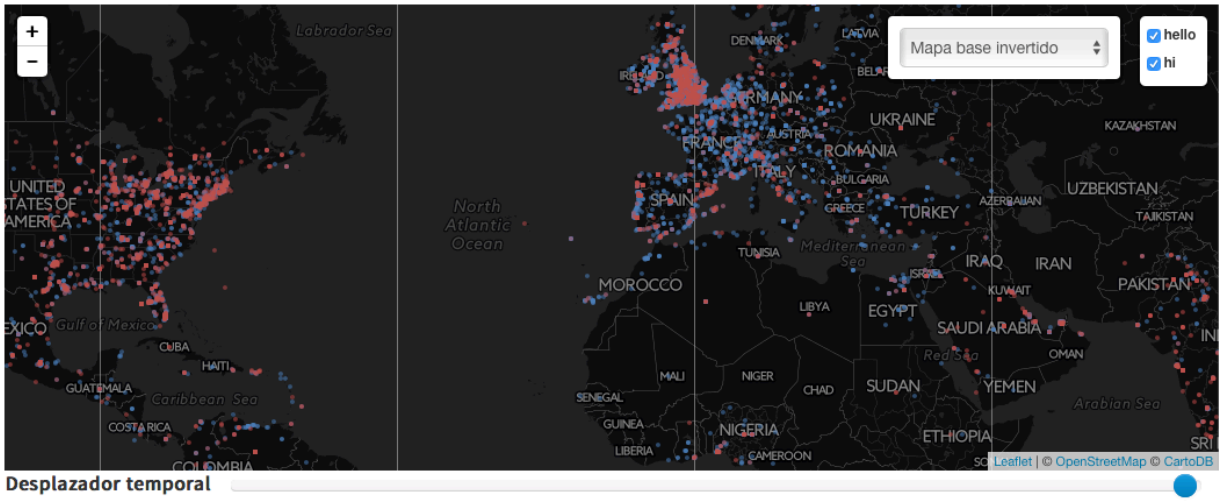


Figura 5.37. Mapa con la distribución geográfica y temporal de los tuits

Debido a que los términos estudiados en el análisis comparativo de *Wordics Archive* son siempre más de uno, este apartado no ofrece las herramientas de análisis por colocaciones o KWIC, que podrán siempre efectuarse sobre la palabra deseada en el análisis simple.

5.3.3.4 *Wordics Data*

Wordics Data constituye el último módulo de la herramienta. Su función, a diferencia de los módulos anteriores, no es realizar búsquedas o análisis de información, sino servir de intermediario entre la información almacenada en la base de datos y otras aplicaciones externas. De este modo, el analista podrá seleccionar bloques de información delimitados por cualquiera de los filtros anteriormente descritos y obtener el resultado en formato de corpus digital (XML, JSON, CSV, Excel, SQL o TXT).

5.3.4 *Wordics Suite* en cifras

Los datos numéricos que aquí presentamos corresponden al período de tiempo comprendido desde septiembre de 2015 hasta febrero de 2016.

Número de tuits

El sistema ha almacenado un total de 305.304.324 tuits en diferentes idiomas. Por cuestiones de optimización y teniendo en cuenta el propósito de nuestra investigación, el sistema no ha mostrado los tuits que no iban a resultarnos útiles, por ejemplo, aquellos publicados en determinados idiomas, como el chino, el japonés, etc. En caso de que sea necesario para un estudio de distintas características, estos idiomas pueden habilitarse.

Número de casos

El sistema almacena un total de 1.959.812.910 palabras, pertenecientes solo a los ocho idiomas filtrados. Esto quiere decir que, en el caso de que se habilite la función de analizar todos los idiomas, el número total de tuits disponibles para ser analizados aumentará considerablemente.

Número de tipos

El sistema ha indizado un total de 1.935.551 palabras no sensibles a mayúsculas y minúsculas. El contenido total, por tanto, es mucho mayor si atendemos a esa

diferenciación. La herramienta, aunque solo haya contabilizado las palabras no sensibles a mayúsculas y a minúsculas, accede y manipula todos los casos.

Número de idiomas

Los tuits almacenados pertenecen a un total de unos 40 idiomas, dependiendo del uso que los hablantes hagan de la plataforma. Por el motivo citado anteriormente, se han reducido a ocho los idiomas tenidos en cuenta para la realización de los análisis, aunque el sistema mantiene almacenada la información de todos los tuits. Esta situación solo se da en los módulos de *Wordics One* y *Wordics Archive*. *Wordics Live*, al ser un reflejo de la realidad en tiempo real, refleja todos los idiomas en los que se escriba en *Twitter* en un momento determinado.

Espacio ocupado

El volumen total de los tuits almacenados por el sistema en sus estado original (previo al preprocesado) se corresponde con 873 GB o, lo que es lo mismo, 893.952 MB. Valga como comparación considerar que la obra completa de Shakespeare ocupa 5 MB y un libro estándar de 300 páginas ocupa 0,5 MB. Esto quiere decir que hemos obtenido el equivalente a 178.790 veces la obra completa de Shakespeare y a 1.787.904 libros de unas 300 páginas.

6

Muestras de uso de *Wordics Live*

El primer módulo de la herramienta *Wordics Suite* tiene como objetivo analizar los tuits que se publican en todo el mundo y en cualquier idioma en tiempo real. *Wordics Live*, que es el nombre que recibe esta primera parte, permite, por tanto, obtener una visión global del comportamiento de la lengua en directo, un fotograma que no solo aporta información acerca de qué se está escribiendo en *Twitter*, sino de cómo, cuándo y dónde está ocurriendo.

La casuística de estudios lingüísticos que se puede llevar a cabo con *Wordics Live* es muy variada; estos pueden dar respuesta a un amplio número de objetivos que converjan, en último término, en el análisis en tiempo real de la lengua. Estos análisis, naturalmente, pueden comprobarse, refutarse o complementarse con las funciones disponibles en el resto de módulos de la herramienta.

Para hacer una demostración de cómo puede usarse *big data* para la investigación lingüística y cuáles son las posibilidades que se nos abren gracias a su consideración en el tiempo real, llevaremos a cabo una serie de estudios prácticos en los que expondremos una mínima muestra de las posibilidades de aplicación de nuestra herramienta a distintos aspectos de la lengua. Dividiremos en dos los grupos de investigaciones, agrupadas según el objetivo y la tipología, para ejemplificar algunas de las utilidades de *Wordics Live*. Así, el orden de las investigaciones será el siguiente:

a) Estudios acerca del uso de algunos términos futbolísticos.
b) Estudios acerca de las distintas variantes ortográficas de la conjunción causal del español *porque*.

6.1 Estudios acerca del uso de algunos términos futbolísticos

En este estudio nos hemos propuesto comprobar cuál es el uso que los hablantes del español hacen de algunos términos relacionados con el ámbito del fútbol y que,

aunque ya están aceptados por la Real Academia Española (DRAE, edición 23ª, 2014), fueron inicialmente introducidos en nuestro idioma como prestamos del inglés. Presentaremos tres investigaciones independientes entre sí, pero coincidentes en cuanto a la metodología. De esta forma, pretendemos determinar si los hispanohablantes, en sus mensajes de *Twitter*, siguen las normas establecidas por la RAE o, por el contrario, utilizan los términos con su escritura original. Los términos estudiados son:

a) *Derbi* y *derby*.
b) *Penalti* y *penalty*.
c) *Córner*, *corner* y *saque de esquina*.

6.1.1 Metodología

La metodología seguida para estudiar estos tres casos ha sido la misma en todos ellos. Puesto que el objetivo era tratar de conocer el uso de estos términos en el idioma español y en tiempo real, se establecieron los siguientes parámetros:

a) Se seleccionó el idioma español, ya que el estudio se centra en este idioma y en las variantes que se utilizan dentro de él.
b) Se filtraron de manera simultánea los binomios o trinomios que pretendíamos comparar, separados por comas y sin espacios entre sí. De tal manera que, para introducir los términos en el buscador, en el primero de los casos escribimos: "derbi,derby"; en el segundo, "penalti,penalty"; y, en el último, "corner,córner,saque de esquina".
c) Se analizaron los datos provenientes de los tuits obtenidos durante la tarde del 27 de febrero de 2016, desde las 15:30 horas, hasta las 23:00 horas, coincidiendo con la jornada número 26 de la liga española de fútbol. Uno de los partidos que se disputaron ese día fue el Real Madrid-Atlético de Madrid, uno de los derbis más seguidos de la competición y que fue determinante para nuestro estudio.
d) En cuanto a las opciones de visualización, se escogió el mapa base invertido, por la comodidad de su diseño para visualizar mejor los datos, y los puntos mantenidos, para que resultara más sencillo poder situar el lugar exacto de publicación del tuit. Posteriormente, se añadieron los clústeres que nos indicaron el número de tuits publicados en cada zona.

Una vez seleccionados estos ajustes y transcurrido el tiempo de recogida de datos, se obtuvieron los resultados que a continuación mostramos.

6.1.2 Casos

6.1.2.1 *Derbi* y *derby*

La palabra *derbi*, procedente del término inglés *derby*, está aceptada en nuestro idioma con *i latina* y, aunque su primera acepción en el Diccionario de la Real Academia Española es "competición hípica, especialmente aquella que se celebra anualmente y en la que corren ejemplares de pura sangre de tres años de edad" (DRAE, 23ª ed.), el uso más común se enmarca en el ámbito deportivo, cuyo significado es: "encuentro, por lo común futbolístico, entre dos equipos cuyos seguidores mantienen constante rivalidad, casi siempre por motivos regionales o localistas" (DRAE, 23ª ed.).

Para estudiar los usos de los términos *derbi* y *derby*, se activó la herramienta la tarde del 27 de febrero de 2016, día en el que se jugaba el partido de fútbol entre el Real Madrid y el Atlético de Madrid, como se ha dicho. La búsqueda comenzó, como también hemos explicado en la metodología, a las 15:30 horas y finalizó a las 23:00 horas de esa misma tarde. En total, se recopilaron 260 tuits en todo el mundo, de los cuales, aproximadamente 140 fueron escritos en España.

Los resultados que devolvió la herramienta a nivel mundial quedaron representados en el mapa de la siguiente forma (teniendo en cuenta que los puntos de color rojo corresponden a *derbi* y los de color azul a *derby*):

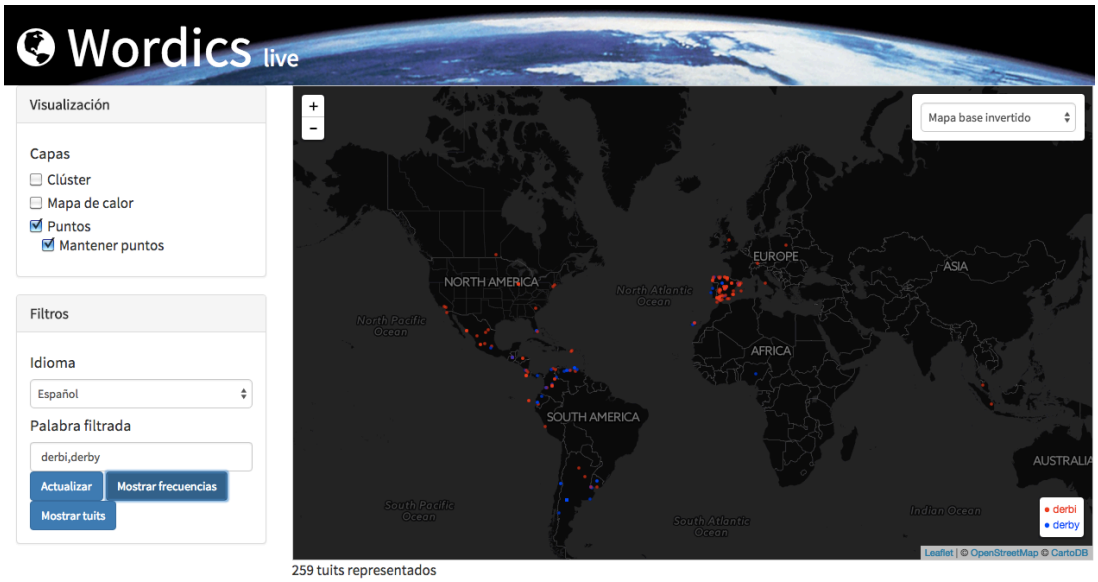


Figura 6.1. Mapamundi con los resultados de la búsqueda *derbi*/*derby* con puntos, el 27 de febrero de 2016, de 15:30 h. a 23:30 h. en *Wordics Live*

Como podemos apreciar, teniendo siempre en cuenta que hemos filtrado el idioma español, mientras que en España y en algunos puntos de Europa –Reino Unido, Italia, Suiza y Polonia–, así como en la zona de Malasia y en Portugal, utilizan *derby*, el continente americano se encuentra más dividido en este sentido, aunque también predomina el uso de *derbi*, incluso en zonas de Estados Unidos o Canadá, donde la influencia del inglés se supone que es mayor.

Esto no quiere decir, sin embargo, que en los tuits generados en nuestro país no se utilizara la forma inglesa durante la tarde del 27 de febrero; antes bien, la forma *derby* se registró en diversos puntos de la geografía peninsular, aunque a mucha menor escala que la forma adaptada al idioma español (*derbi*). Lo podemos apreciar en la figura siguiente (6.2), donde, de nuevo, *derbi* está representado por puntos de color rojo y *derby*, de color azul:

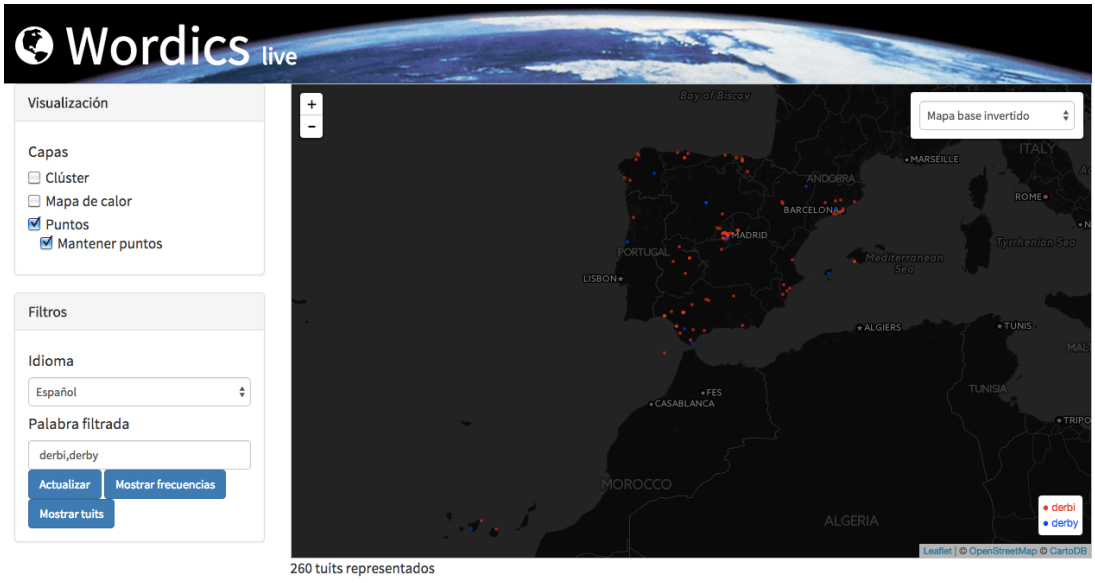


Figura 6.2. Mapa de España con los resultados de la búsqueda *derbi*/*derby* con puntos, el 27 de febrero de 2016, de 15:30 h. a 23:30 h. en *Wordics Live*

Los datos obtenidos como resultado de esta búsqueda son un fotograma de las horas durante las cuales *Wordics Live* estuvo en funcionamiento, que se correspondieron con la jornada 26 de la Liga BBVA. Además, conviene recordar que los puntos solo representan aquellos tuits escritos en español en todo el mundo en los que apareciera alguno de los términos objeto de estudio.

Como acabamos de mostrar, *Wordics Live* nos ofrece no solo la posibilidad de situar en el mapa los puntos en los que se publican tuits de forma general o con alguna palabra o expresión concretas: también podemos obtener la lista de ocurrencias de todas las palabras que aparecen en los tuits filtrados, así como la lista de esos tuits. Ambas listas se pueden exportar a formato *Excel*, para una mejor gestión de los datos obtenidos.

A continuación (figura 6.3), se muestran las 19 palabras más utilizadas de los tuits obtenidos.

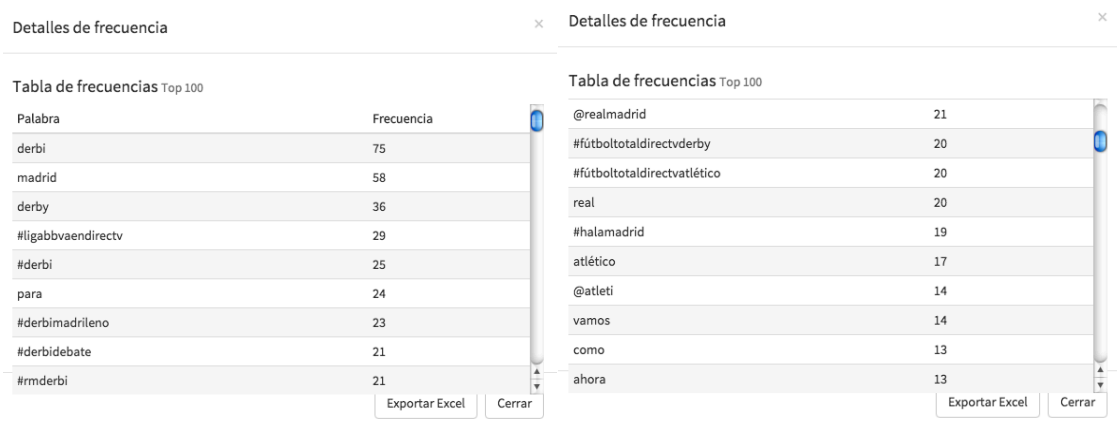


Figura 6.3. Tabla de frecuencias de las 20 primeras palabras de los tuits con los términos *derbi*/*derby*

Por otro lado, podemos ver en la siguiente figura (6.4) los tuits que se corresponden con nuestra búsqueda.

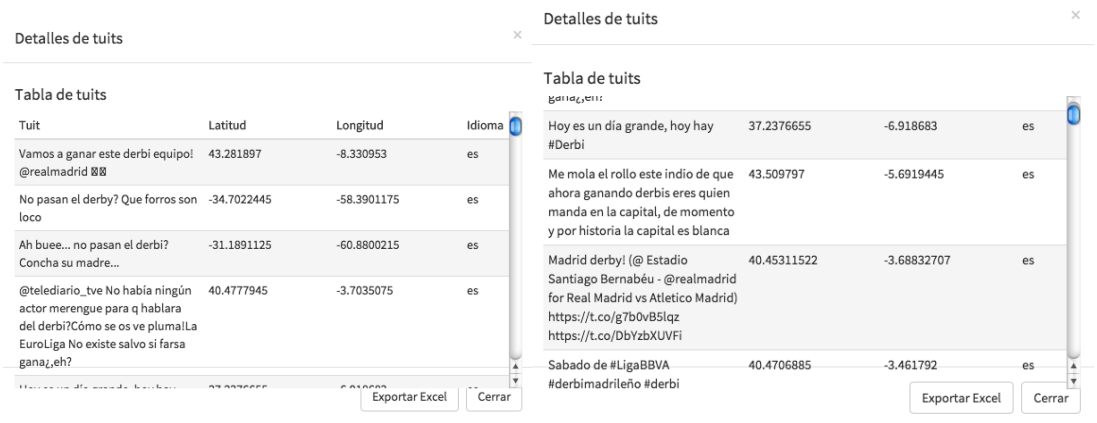


Figura 6.4. Tabla de los 8 primeros tuits filtrados con los términos *derbi*/*derby*

6.1.2.2 Penalti y penalti

La misma tarde (27 de febrero de 2016) en la que se llevaron a cabo los estudios para determinar qué palabra es la más usada en español del binomio *derbi* y *derby*, la herramienta *Wordics Live* estuvo también realizando una recopilación de información similar, pero en este segundo caso con relación a las palabras *penalti* y *penalty*. Al igual que ocurre con los dos términos anteriores (*derby*/*derbi*), en español se utiliza *penalti* como adaptación gráfica del anglicismo *penalty*, según la Fundéu BBVA. De hecho, en el Diccionario de la Real Academia Española esta adaptación está recogida con la definición de "en el fútbol y otros deportes, máxima sanción que se aplica a ciertas faltas del juego cometidas por un equipo dentro de su área" (DRAE, 23ª ed.). No es de extrañar, por otro lado, que términos pertenecientes al campo semántico de *fútbol* y procedentes del inglés abunden en nuestro idioma y que, gracias a su continuo uso, hayan quedado adaptados a la ortografía española y, en general, a la estructura morfosintáctica de la lengua española.

Aun así, como hemos podido comprobar, y a pesar de que los hispanohablantes siguen, por regla general, la norma establecida, todavía persiste en algunos casos la utilización de la grafía inglesa frente a la española, y así pudimos constatarlo.

Para poder conocer la utilización en tiempo real, por tanto, de los términos *penalti* y *penalty*, llevamos a cabo el mismo procedimiento que en el caso anterior. De manera que filtramos el idioma español y, a continuación, ambas palabras, separadas por una coma, pero sin espacios. Hicimos coincidir la fecha y la franja horaria con la búsqueda de tuits realizada para *penalti* y *penalty*, para aprovechar la jornada de liga, que es cuando presuponemos que se va a hacer un mayor uso de este tipo de léxico futbolístico. Puesto que las condiciones y los parámetros de la investigación son los mismos que en los ejemplos anteriores, hemos vuelto a seleccionar los puntos para la visualización de los datos y el mapa satélite invertido.

Comenzamos, por tanto, con el mapa de puntos a nivel mundial, que se puede ver en la figura siguiente (6.5). En este caso, la palabra con ortografía española estará representada con el color rojo y la inglesa, en azul:

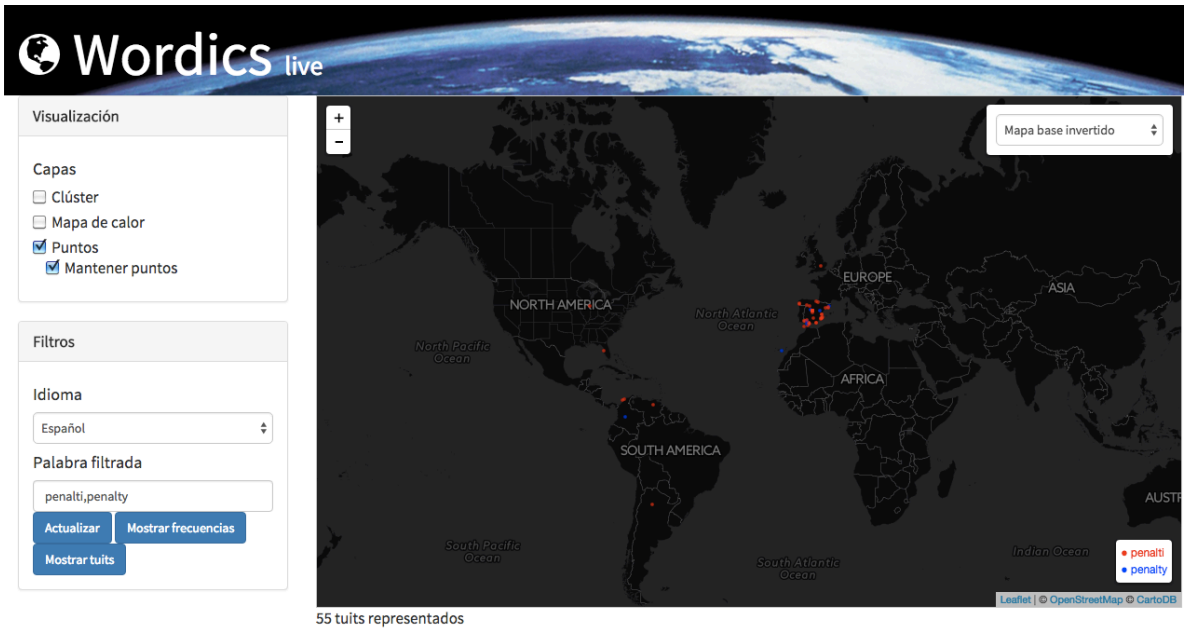


Figura 6.5. Mapamundi con los resultados de la búsqueda *penalti*/*penalty* con puntos, el 27 de febrero de 2016, de 15:30 h. a 23:30 h. en *Wordics Live*

Aunque los resultados no se apartan de lo esperable, debido a que el número de penaltis en un partido no suele ser elevado, resulta curioso comprobar que, mientras que fueron obtenidos 260 tuits para *derbi* y *derby*, para *penalti* y *penalty* solo se recogieron 55, durante la misma franja horaria. Por otra parte, mientras que en todo el continente americano se utilizó *derbi* durante esa tarde (mañana para ellos), los usuarios de *Twitter* no se hicieron mucho eco de los penaltis, puesto que solo aparecen algunos tuits repartidos por Venezuela, Argentina, Florida e Illinois.

En lo que se refiere a España (figura 6.6), también podemos apreciar un notable descenso en el número de resultados obtenidos aunque, lógicamente, con un índice mayor de utilización. La mayor actividad de la plataforma con estos términos, como indica el mapa, se concentra en Madrid, ciudad donde se disputaba el derbi y durante el cual se produjo una jugada polémica con posibilidad de terminar en penalti. Conviene advertir que, como el lector puede apreciar, el número total de tuits indicado en la parte inferior del mapa varía entre el mapa mundial y el mapa de España. Este desajuste se debe a que, puesto que el análisis se realiza en tiempo real, en los minutos transcurridos entre la captura de pantalla de un mapa y la del otro se publicaron nuevos tuits, lo que fue registrado por el propio sistema.

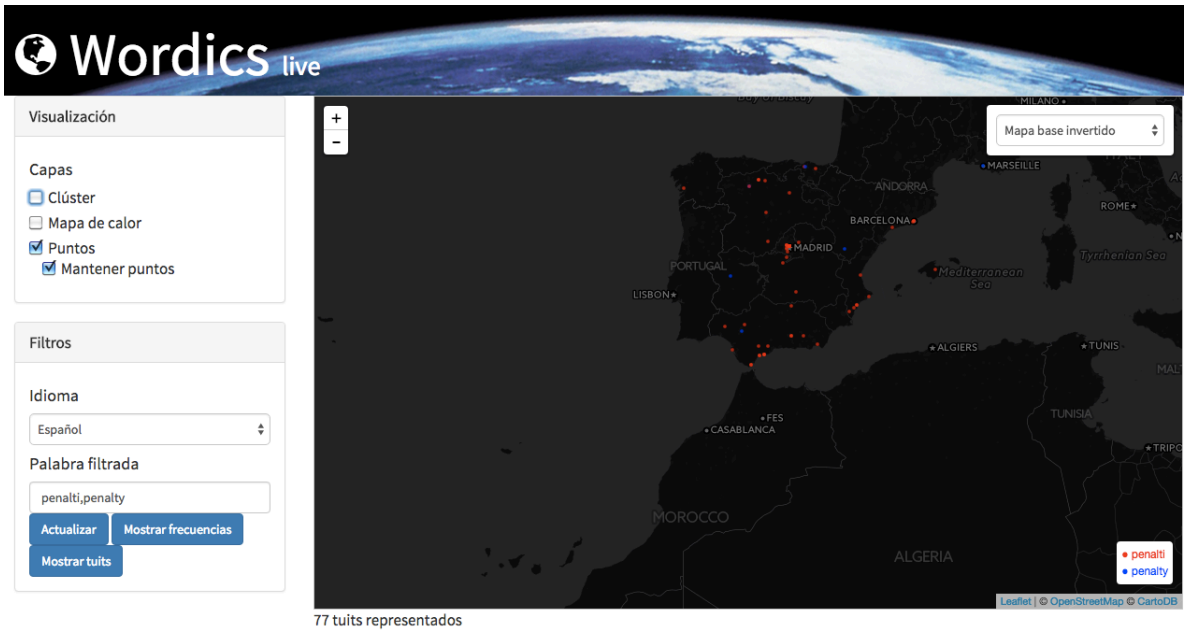


Figura 6.6. Mapa de España con los resultados de la búsqueda *penalti*/*penalty* con puntos, el 27 de febrero de 2016, de 15:30 h. a 23:30 h. en *Wordics Live*

Al igual que ocurría en el caso anterior, a pesar de la muy mayoritaria utilización de la alternativa española del término (*penalti*), persisten algunos casos de escritura inglesa (*penalty*) en ciertas zonas de la Península. Podemos ver de forma preliminar las frecuencias de palabras que aportamos a continuación en las tablas (6.7), donde aparecen las 18 más usadas en los tuits que conforman nuestra búsqueda y el número exacto de utilización de ambas posibilidades (*penali* y *penalty*).

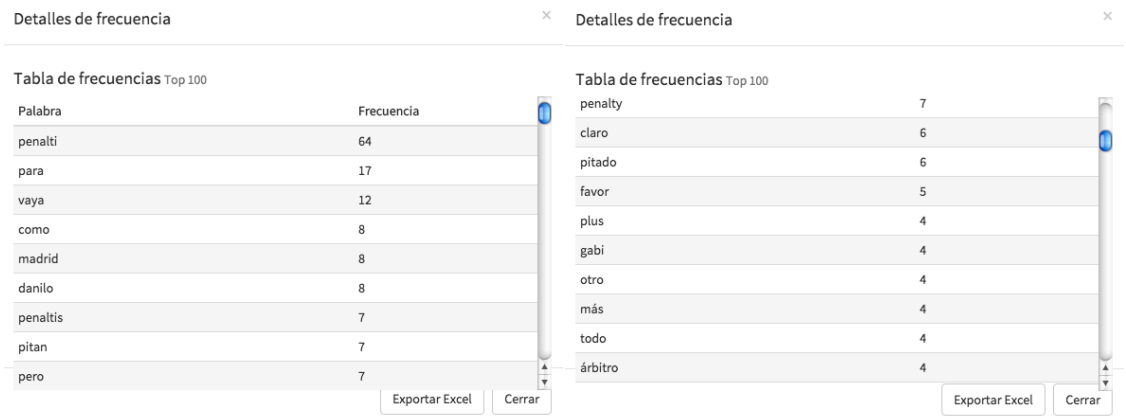


Figura 6.7. Tabla de frecuencias de las 18 primeras palabras de los tuits con los términos *penalti*/*penalti*

En esta primera aproximación a las ocurrencias de *derbi* y de *derby*, así como de las palabras que las acompañan en el texto del tuit, observamos cómo *penalti* es la forma más utilizada, con 64 apariciones, mientras que *penalty* aparece la número 10,

con 7 apariciones. *Penaltis*, en plural, tiene también 7 apariciones y, entre las palabras más ocurrentes, encontramos *madrid*, *danilo*, *pitan*/*pitado*, *claro*, *favor*, *gabi* o *árbitro*, por nombrar las que mayor carga semántica tienen.

Los tuits que contienen estas palabras los podemos ver a continuación (figura 6.8) en una pequeña muestra del total:

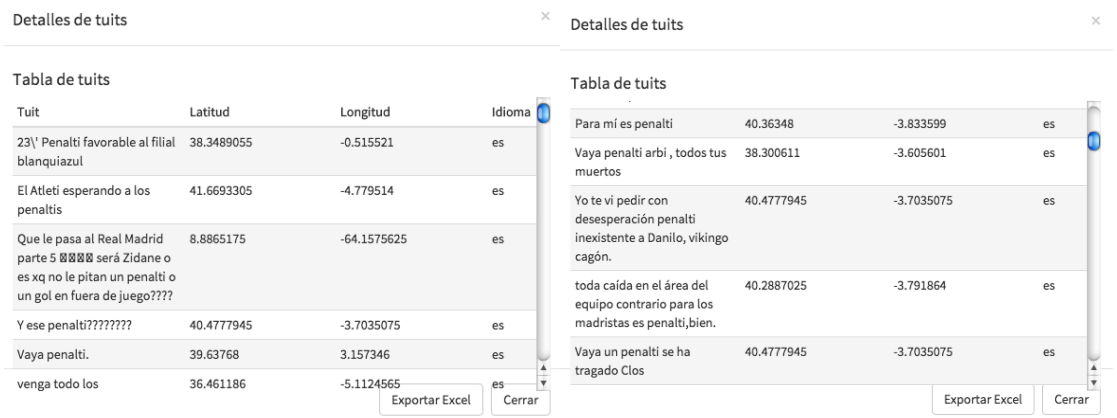


Figura 6.8. Tabla de los 8 primeros tuits filtrados con los términos *penalti*/*penalty*

6.1.2.3 Córner, corner y saque de esquina

El último conjunto de términos del ámbito del fútbol que analizamos se trata de *corner*, en inglés, con su versión española adaptada ortográficamente y recogida en el DRAE (23ª ed.), *córner*, así como la expresión española equivalente *saque de esquina*.

Para la fase de recogida de información volvimos a establecer los mismo parámetros que en los dos casos anteriores en cuanto a idioma, horas y criterios de visualización.

En la figura siguiente (6.9) vemos el mapamundi con la representación de los tuits que contienen los términos *corner*, *córner* o *saque de esquina*, representados en color rojo, azul y verde, respectivamente:

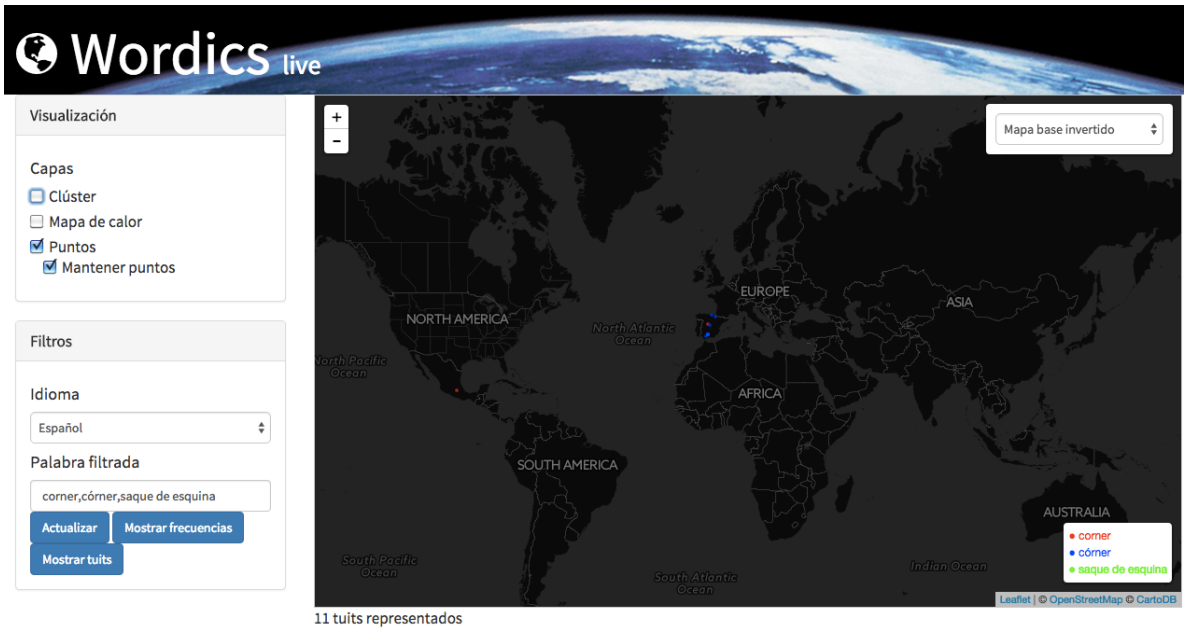


Figura 6.9. Mapamundi con los resultados de la búsqueda *corner*/*córner*/*saque de esquina* con puntos, el 27 de febrero de 2016, de 15:30 h. a 23:30 h. en *Wordics Live*

En esta muestra, en comparación con el ya visto de *penalti*, son menos los casos de tuits publicados que contienen alguno de los términos *corner*/*córner*/*saque* de esquina. Este hecho puede deberse a tres motivos, fundamentalmente: a) que no se produjera un gran número de saques de esquina a lo largo de la tarde, b) que los saques de esquina no sean, en realidad, tan determinantes para el resultado de un partido como lo puedan ser los penaltis y c) la combinación de ambos factores. Desde nuestro punto de vista, creemos que se puede descartar la primera de las posibilidades, puesto que el número medio de saques de esquina producidos en un partido es mucho mayor que el de la pena máxima y, sin embargo, en el caso estudiado aparecen muchos más tuits referidos a penaltis que a saques de esquina. Creemos, por tanto, que el motivo por el que el número de resultados en este caso es menor se basa en la menor trascendencia de cara a los resultados que tienen los saques de esquina frente a los penaltis.

Mientras que la mayoría de los tuits que contienen estos tres términos se sitúan en España, el único post publicado en el continente americano se localiza en México con la palabra escrita con ortografía inglesa (*corner*). Para poder apreciar mejor el caso de nuestro país, podemos observar el siguiente mapa ampliado (figura 6.10). Aquí también, como ocurría en el caso anterior (*penalti*/*penalty*), existe una diferencia entre el total de tuits representados en todo el mundo y en España, que pasa de 11 a 16, por el

mismo motivo de antes: los segundos transcurridos entre la realización de ambas capturas de pantalla.

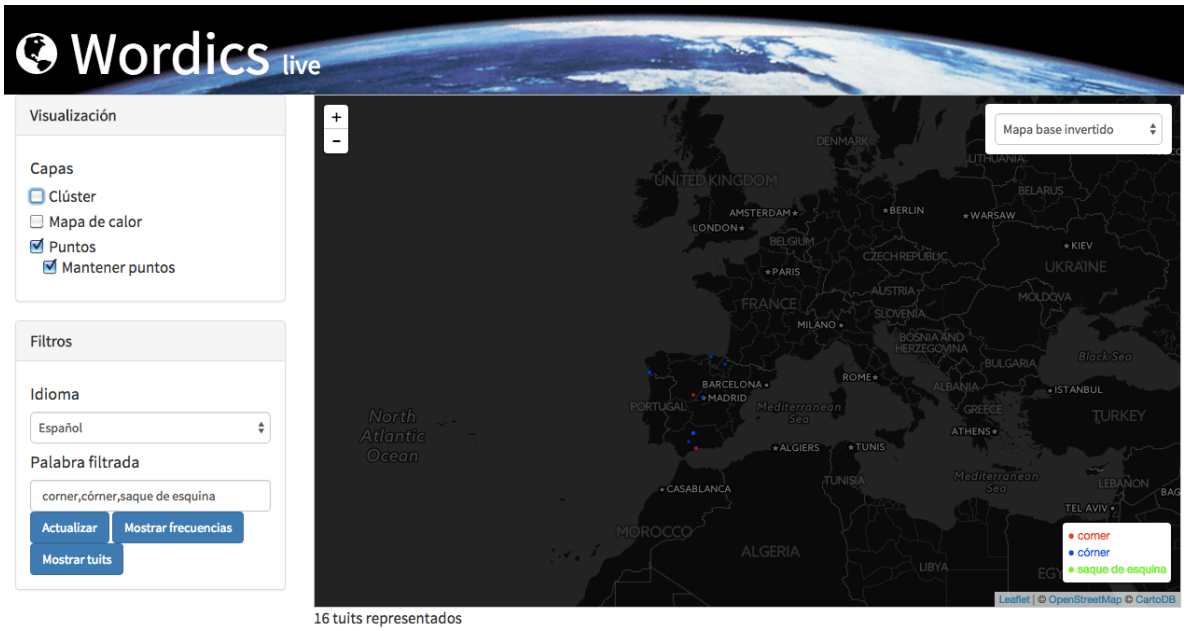


Figura 6.10. Mapa de España con los resultados de la búsqueda *corner/córner/saque de esquina* con puntos, el 27 de febrero de 2016, de 15:30 h. a 23:30 h. en *Wordics Live*

En España, durante la franja horaria indicada, se registran 14 tuits representados en total (los dos restantes se sitúan en Estados Unidos), de los cuales 12 contienen la palabra *córner* y 2 *corner*. Resulta significativo el hecho de que, durante las horas en que la herramienta estuvo recopilando datos, en ningún momento se utilizó la unidad lexemática *saque de esquina*.

A continuación, al igual que en los dos casos anteriores, introducimos las tablas (figura 6.11) de las primeras ocurrencias de palabras contenidas en estos tuits, así como de los tuits en cuestión que se han escrito en el tiempo establecido (figura 6.12).

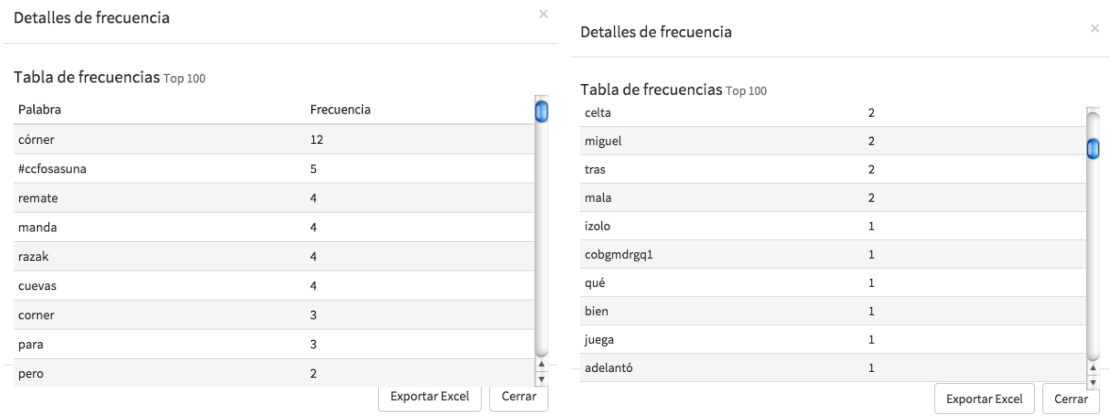


Figura 6.11. Tabla de frecuencias de las 19 primeras palabras de los tuits con los términos *corner/córner/saque de esquina*

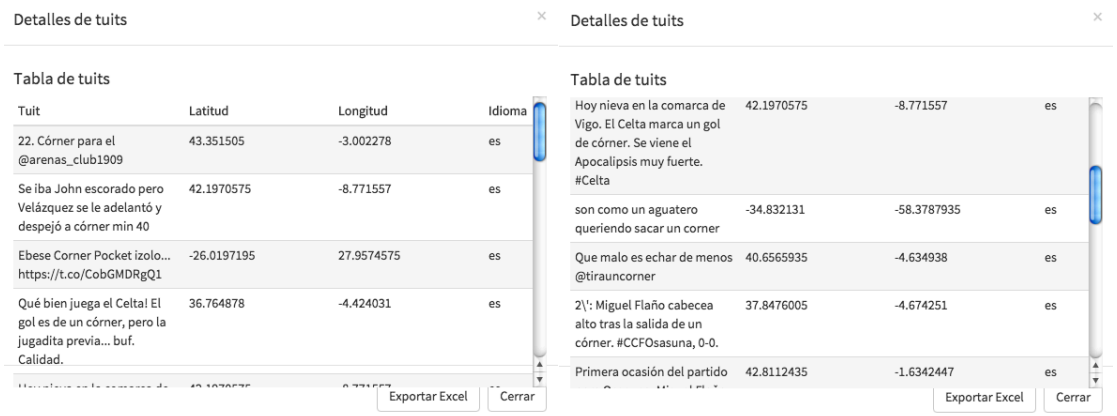


Figura 6.12. Tabla de los 8 primeros tuits filtrados con los términos *corner/córner/saque de esquina*

6.1.3 Resultados

Tras analizar la información lingüística –o aquella que puede tener utilidad para el estudio, a pesar de tratarse de información extralingüística– contenida en los tuits predeterminados, podemos concluir, en términos generales, que, aunque las voces inglesas o con ortografía anglosajona no han dejado de usarse por completo en el idioma español, predomina ampliamente el término adaptado y aceptado por la Real Academia Española en la última versión de su Diccionario (23ª ed.).

Por otra parte, las dos zonas mundiales en las que más se utiliza en español (al menos en el español de los tuits) el léxico futbolístico son Europa –y, mucho más concretamente, España– y el continente americano. En este último vemos un uso más extendido que el del viejo continente, aunque predomina, sobre todo, en la zona de Centroamérica.

Además, también pudimos comprobar que, incluso perteneciendo todos los términos al campo futbolístico, *derbi* y *derby* se utilizan con mucha más frecuencia que el resto de las palabras examinadas:

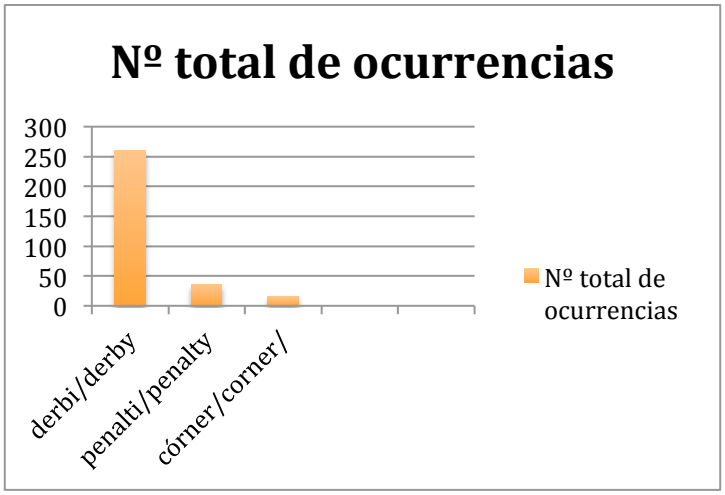


Figura 6.13. Número total de ocurrencias de los términos de los distintos casos estudiados el 27 de febrero de 2016, de 15:30 h. a 23:30 h.

Aunque, aparentemente, estos resultados entran dentro de lo esperable, teniendo en cuenta que el análisis se realizó en el transcurso de un derbi, no es menos cierto también que es también durante la celebración de los partidos cuando se producen los hechos que desencadenan la utilización de los otros términos (*penalti* y *córner*).

En los tres casos vistos, el porcentaje de uso del término en español frente a la grafía inglesa es, por orden de estudio, del 68%, 90% y 80%, respectivamente (figura 6.14). Merece especial atención el caso de la expresión *saque de esquina*. Mientras que en todos los casos estudiados vemos cómo se impone la ortografía española a la extranjera, es curioso observar cómo la expresión *saque de esquina*, la más española de todos ellos puesto que en sus inicios no tenía origen extranjero, no se escribió ni una sola vez en los tuits a lo largo de toda la tarde, tal vez a causa de su mayor longitud formal:

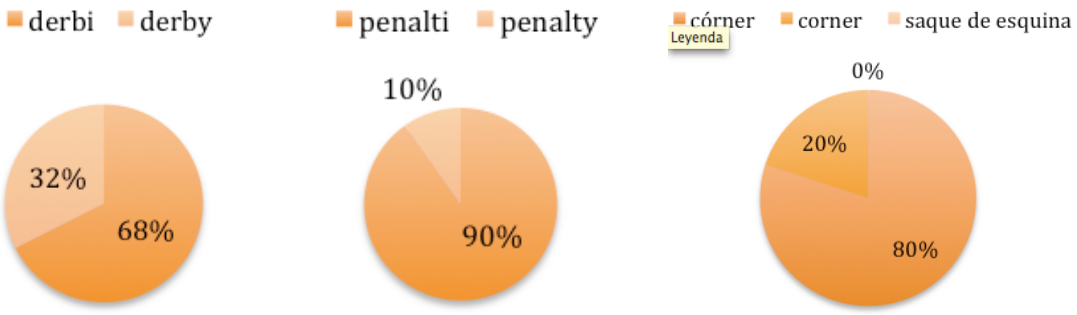


Figura 6.14. Gráficos con el porcentaje de ocurrencias de los distintos términos

Si analizamos en profundidad las tablas de ocurrencias y de tuits generadas, podremos comprobar que los términos cuyo uso estamos mostrando se han utilizado en más ocasiones. Sin embargo, puesto que la herramienta no descarta los *hashtags* (dado que, al fin y al cabo, estos también son palabras) y, además, busca con la técnica de expresiones regulares[44], podemos encontrar, efectivamente, los términos incluidos en los *hashtags* o en su forma del plural. En la figura siguiente (6.15) hemos mostrado el número total de veces que aparecen los términos en estas formas en los tuits analizados que se analizaron durante la tarde del 26 de febrero de 2016. Tengamos en cuenta que en esta tabla ya no están contemplados los datos que aparecen en la figura 6.14.

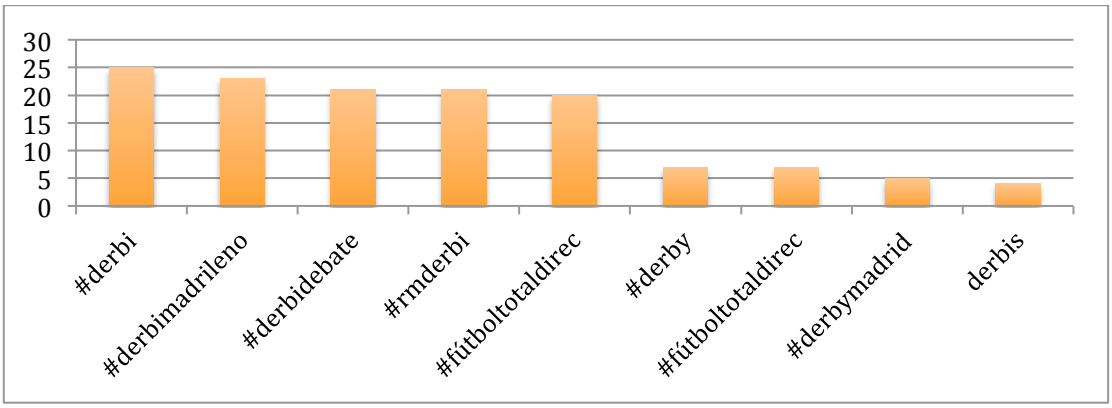


Figura 6.15. Variantes de *derbi* y *derby* y número total de apariciones

La variante encontrada para el segundo caso –*penalti* y *penalty*– es solo la forma plural (*penaltis*), escrita con *i* latina y con 7 apariciones; mientras que para el último caso, solo aparece en una ocasión un nombre de usuario *(@tirauncorner*), sin acento.

Como podemos observar, incluso dentro de los *hashtags* y, lógicamente, los plurales, sigue predominando la variante española normalizada por el DRAE.

Para concluir, mostramos a continuación una tabla (figura 6.16) en la que sintetizamos la mayor parte de la información obtenida en los tres estudios realizados, en los que se ha trabajado con los grupos de términos *derbi*/*derby*, *penalti*/*penalty* y *córner*/*corner*/*saque de esquina*, localizados en los tuits generados en la tarde del día 27 de febrero de 2016. En dicha tabla podemos contrastar de manera rápida los datos aportados hasta el momento:

[44] Este concepto de expresiones regulares ya lo hemos explicado en el capítulo 5, dedicado a la metodología.

Casos		1º: DERBI/DERBY		2º: PENALTI/ PENALTY		3º: CÓRNER/CORNER SAQUE DE ESQUINA		
Nº TOTAL DE OCURRENCIAS		260		36		16		
ZONA	EUROPA	147		69		14		
	AMÉRICA	110		6		1		
	OTROS	4		2		1		
Nº INDIVIDUALIZADO DE OCURRENCIAS		derbi	derby	penalti	penalti	córner	corner	saque de esquina
		75	36	64	7	12	3	0
VARIANTES Y Nº DE OCURRENCIAS		#derbi (25) #derbimadrileno (23) #derbidebate (21) #rmderbi (21) #fútboltotaldirect-derby (20) #derby (7) #fútboltotaldirectv-derby (7) #derbymadrid (5) derbis (4)		penaltis (7)		@tirauncorner (1)		

Figura 6.16. Resumen de datos obtenidos en el análisis de términos futbolísticos con *Wordics Live,* la tarde del 27 de febrero de 2016

6.2 ESTUDIOS ACERCA DE LAS DISTINTAS VARIANTES ORTOGRÁFICAS DE LA CONJUNCIÓN CAUSAL DEL ESPAÑOL *PORQUE*

Otro de los estudios que hemos realizado para ejemplificar algunas de las utilidades que nos ofrece *Wordics Live* tiene como objetivo investigar el uso escrito que los hispanohablantes hacen de la conjunción causal del español *porque*.

Aunque a lo largo de la historia la costumbre de abreviar y acortar palabras ha estado presente en los sistemas de escritura, parece ser que esta práctica se ha visto fomentadas por el uso de las nuevas tecnologías, donde la economía lingüística suele tener un papel dominante frente a la ortodoxia gráfica.

En las próximas líneas explicaremos la metodología que hemos seguido para llevar a cabo esta investigación y los resultados obtenidos a partir de ella.

6.2.1 Metodología

En este nuevo estudio de caso nos proponemos conocer qué posibilidad, de las varias que se utilizan actualmente, es la preferida por los hablantes del español hoy en día, gracias a la oportunidad que nos brinda la herramienta de obtención de datos en tiempo real y geolocalizados. Es decir, pretendemos realizar un estudio sincrónico en el que se pueda mostrar qué escriben los usuarios de *Twitter*, pero también, dónde y a qué hora.

El estudio se ha realizado a lo largo de dos semanas, en distintos momentos del día, para asegurarnos de que la información obtenida en un momento concreto no es fruto de algún factor externo que provoque esos resultados, sino que es la tendencia habitual.

La metodología empleada ha seguido los mismos pasos que en los estudios anteriores: se ha seleccionado el idioma español y se han introducido en el buscador los distintos términos, separados por comas y sin espacios. Estos términos han sido, en orden: *porque*, *porq*, *pq*, *xq*, *xk*, *xque*. Para una mayor claridad y coherencia, todos los días en los que se ha realizado la búsqueda se ha introducido la misma secuencia de términos, con la intención de que los colores de los puntos dibujados en el mapa representen siempre al mismo elemento.

Las dos semanas en las que se ha llevado el análisis están comprendidas entre el 15 de febrero de 2016 y el día 28 del mismo mes. La recogida de información se ha llevado a cabo durante una hora los lunes, miércoles, viernes y domingos de cada una de las dos semanas, a distintas horas del día, pero de manera que coincidieran los horarios en los mismos días. Queda, por tanto, el calendario, con la siguiente distribución:

-Lunes, 15 de febrero: de 12:00 h. a 13:00 h.
-Miércoles, 17 de febrero: de 15:00 h. a 16:00 h.
-Viernes, 19 de febrero: de 17:00 h. a 18:00 h.
-Domingo, 21 de febrero: de 22:00 h. a 23:00 h.
-Lunes, 22 de febrero: de 12:00 h. a 13:00 h.
-Miércoles, 24 de febrero: de 15:00 h. a 16:00 h.
-Viernes, 26 de febrero: de 17:00 h. a 18:00 h.
-Domingo, 28 de febrero: de 22:00 h. a 23:00 h.

6.2.2 Resultados

Para mostrar los resultados, mostramos a continuación un mapa para cada uno de los días estudiados con clústeres y puntos. El resto de la información, relativa a las frecuencias de aparición de las distintas variantes de la conjunción *porque* en español, así como el corpus de tuits obtenido de cada día, los podemos encontrar en el Anexo 4. Como podemos observar en las imágenes, la secuencia de términos introducida en el buscador ha seguido siempre el mismo orden con el objetivo de que los colores identificativos de cada variante fueran siempre los mismos. De esta forma, *porque* está representado por puntos de color rojo; *porq*, en azul; *pq*, en verde; *xq*, en amarillo; *xk*, en rosa y *xque*, en turquesa (se puede consultar en la leyenda de todos los mapas que aportamos a continuación: figuras 6.17 a 6.24).

El color predominante en todos los mapas obtenidos es, sin lugar a dudas, el rojo; es decir, la variante más utilizada por los usuarios de *Twitter* en español (porque este ha sido el idioma filtrado) de todas las posibilidades que hemos introducido en el buscador es *porque*. Mostramos a continuación (figuras 6.17 a 6.24) los ocho mapas correspondientes a cada uno de los estudios, con mapa base y clústeres que nos indican el número de tuits correspondiente a cada región.

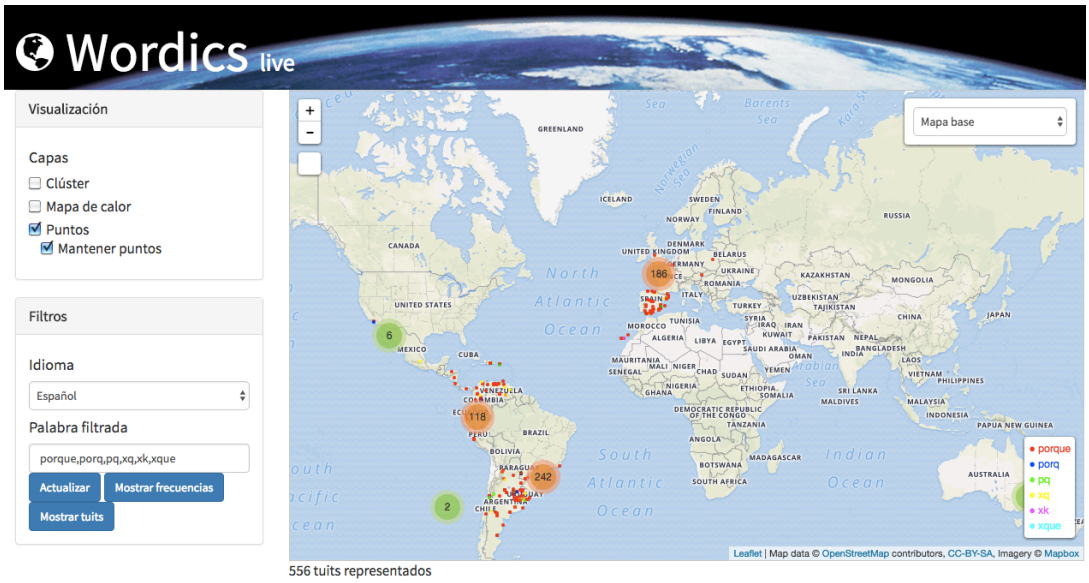


Figura 6.17. Mapa con clústeres de las variantes ortográficas de la conjunción *porque* en español, 15 de febrero, de 12:00 h. a 13:00 h.

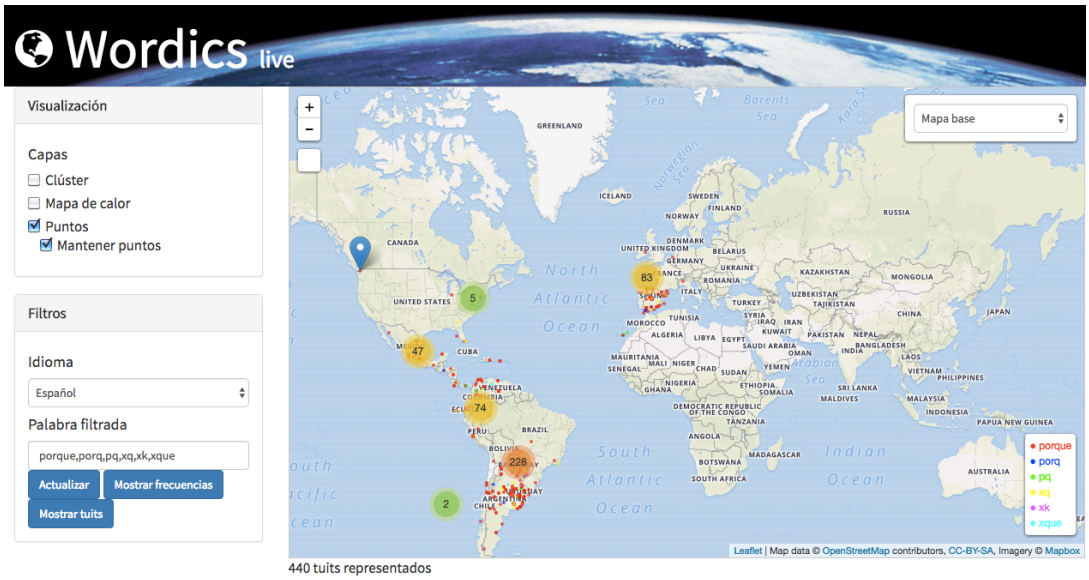


Figura 6.18. Mapa con clústeres de las variantes ortográficas de la conjunción *porque* en español, 17 de febrero, de 15:00 h. a 16:00 h.

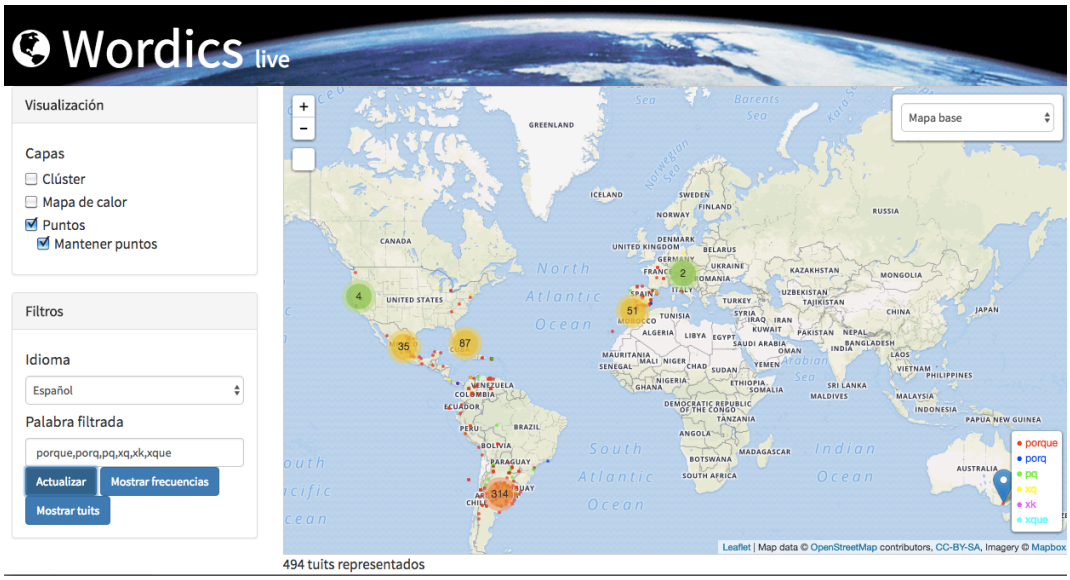


Figura 6.19. Mapa con clústeres de las variantes ortográficas de la conjunción *porque* en español, 19 de febrero, de 17:00 h. a 18:00 h.

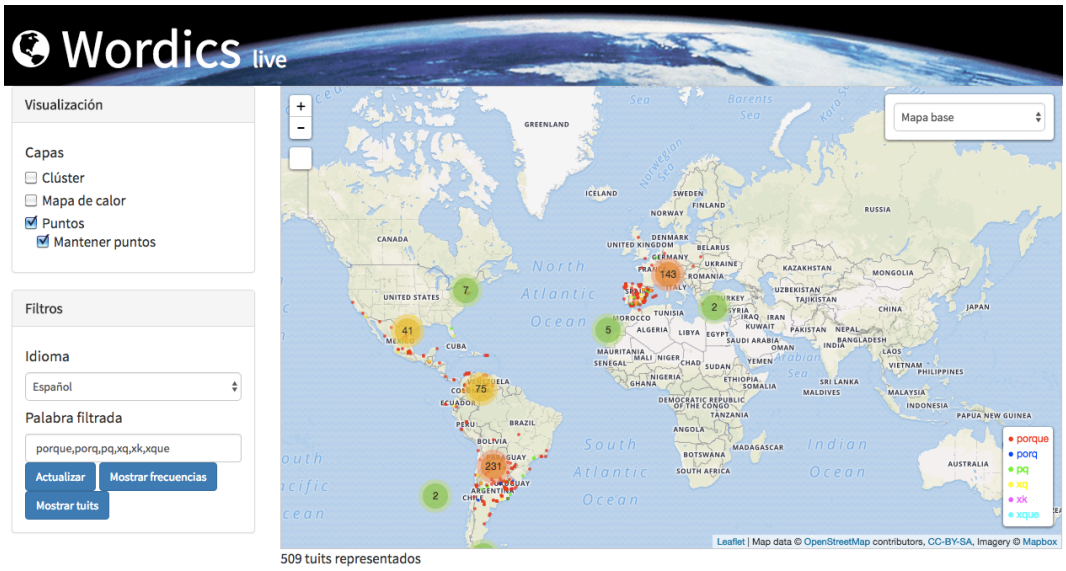


Figura 6.20. Mapa con clústeres de las variantes ortográficas de la conjunción *porque* en español, 21 de febrero, de 22:00 h. a 23:00 h.

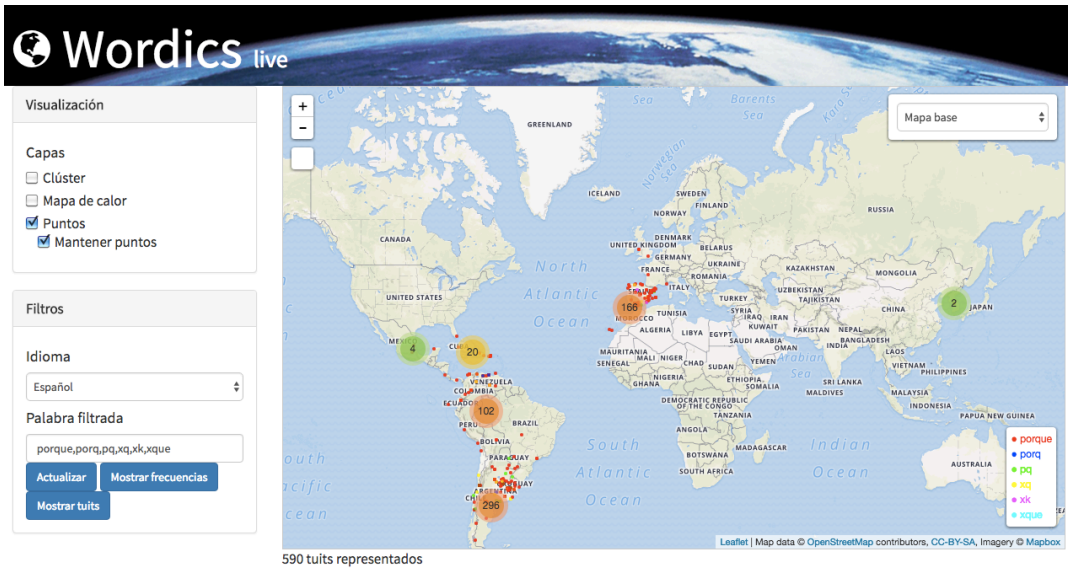


Figura 6.21. Mapa con clústeres de las variantes ortográficas de la conjunción *porque* en español, 22 de febrero, de 12:00 h. a 13:00 h.

Figura 6.22. Mapa con clústeres de las variantes ortográficas de la conjunción *porque* en español, 24 de febrero, de 15:00 h. a 16:00 h.

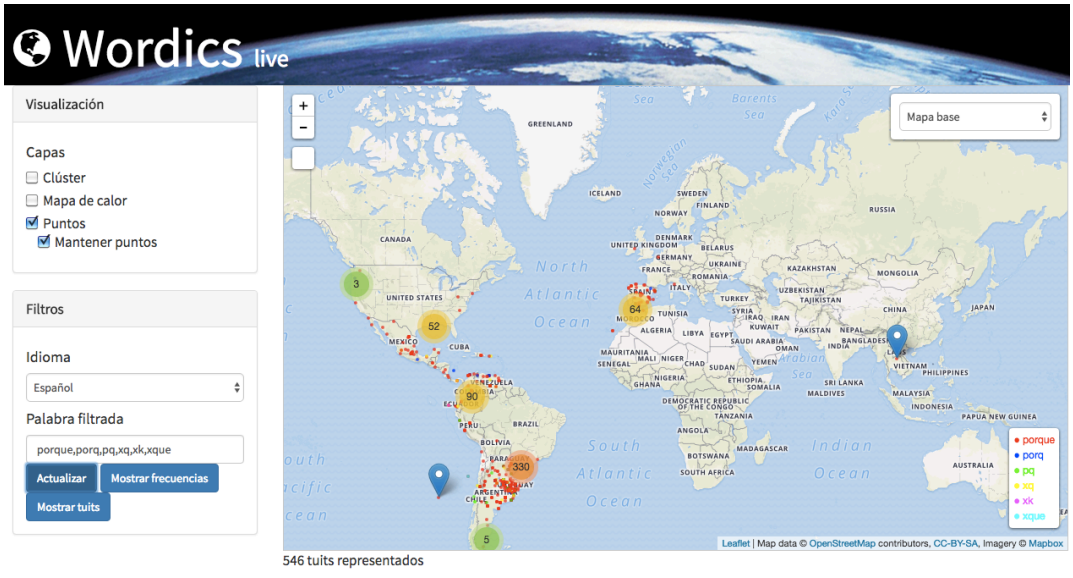


Figura 6.23. Mapa con clústeres de las variantes ortográficas de la conjunción *porque* en español, 26 de febrero, de 17:00 h. a 18:00 h.

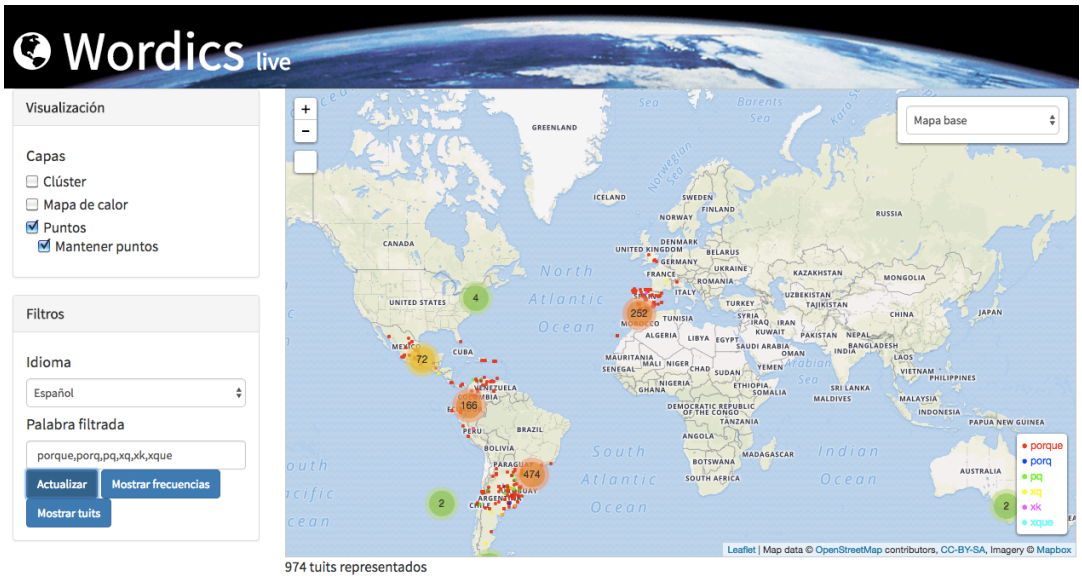


Figura 6.24. Mapa con clústeres de las variantes ortográficas de la conjunción *porque* en español, 28 de febrero, de 22:00 h. a 23:00 h.

Los resultados, como hemos apuntado más arriba y podemos comprobar, son contundentes: en todos los casos estudiados predomina el uso de la forma íntegra *porque* sobre el resto de variantes ortográficas estudiadas. El resto de las variantes, en orden de aparición son: *xq*, *pq*, *porq*, *xk* y *xque*. A continuación (figura 6.25) ofrecemos una tabla que recoge toda la información aportada en los mapas anteriores y donde podemos ver los días en los que se ha realizado la recogida de información, el número total de tuits analizados y la frecuencia de aparición de cada variante gráfica de la conjunción causal *porque*. Conviene señalar que la suma de todas las variantes de la conjunción no se corresponden exactamente con el número total de tuits obtenidos cada día porque se da el caso de que en algunos tuits se escribe alguna de estas palabras en más de una ocasión.

	lunes		miércoles		viernes		domingo	
Fecha	15 feb.	22 feb.	17 feb.	24 feb.	19 feb.	26 feb.	21 feb.	28 feb.
Total de tuits	556	590	440	442	494	546	509	487
porque	466	494	370	357	414	465	442	427
porq	18	10	12	18	15	15	11	11
pq	32	42	31	34	41	33	40	26
xq	60	66	48	46	50	39	83	37
xk	4	4	1	1	1	2	2	4
xque	-	-	-	1	-	1	2	-

Figura 6.25. Tabla resumen del uso de la conjunción causal *porque* en español

6.2.3 Conclusiones

La ventaja de poder llevar a cabo un estudio con *Wordics Live* se resume, fundamentalmente, en la posibilidad que brinda al investigador de obtener información lingüística en tiempo real y con indicación de su ubicación geográfica, lo que abre nuevas puertas a este tipo de investigación como nunca antes se había visto. Para poder obtener un mínimo de datos cuantificables, hemos establecido un período de una hora en cada una de las búsquedas de información que hemos llevado a cabo en este trabajo. Los resultados, sumando todos los datos obtenidos, alcanzan la cifra de 4.064 tuits. En la figura 6.26 podemos comprobar cómo se distribuyen cada una de las posibilidades estudiadas y sus porcentajes respecto a este total:

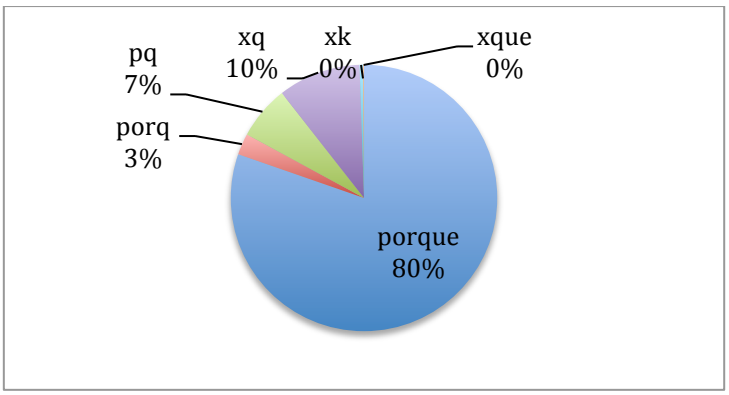


Figura 6.26. Distribución en porcentajes del uso de las variantes de la conjunción *porque* en español

Los resultados no dejan lugar a dudas. El patrón de uso se repite en cada uno de los días estudiados, al igual que ocurre con la frecuencia aproximada de apariciones. El uso de la conjunción causal *porque*, escrita sin acortamientos o abreviaturas, predomina sobre el resto de las posibilidades abreviadas, a pesar de las restricciones de espacio que presenta *Twitter*, cuyos mensajes debe limitarse al uso de 140 caracteres.

7
Muestras de uso de *Wordics One*

7.1 ESTUDIO DEL LENGUAJE DE VARIOS ESCRITORES ACTUALES

La principal ventaja que nos ofrece el módulo de *Wordics One* deriva de la posibilidad de acceder a la información de cuentas individuales de usuarios de *Twitter*. Así, podemos la forma de escribir de uno o varios personajes concretos y también de organizaciones y colectivos. Esto nos permite extraer conclusiones no solo acerca de usos individuales de la lengua, sino también establecer generalizaciones por grupos sociales, laborales, políticos, de comunicación, etc.

Con la intención de demostrar alguna de estas utilidades, hemos llevado a cabo un estudio del lenguaje de distintos escritores españoles activos en sus cuentas de *Twitter*, para ver sus formas de utilizar la lengua y compararlas entre sí.

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

Algunos autores, como Daller, van Hout y Treffers-Daller (2003) intercambian indistintamente el uso del término *diversidad léxica* con el de *riqueza léxica*, mientras que otros, como Malvern *et al*. (2004), en Johansonn (2008), indican que la *diversidad léxica* es solo una parte del concepto multidimensional de *riqueza léxica*. En esta línea, Laufer y Nation (1995: 309) enumeran varios elementos para medir la riqueza léxica, que son: originalidad léxica, densidad léxica, sofisticación léxica y variación léxica. En este trabajo, siguiendo a Daller, van Hout y Treffers-Daller (2003), así como a Gregori Signes y Clavel Arroitia (2015) y al *Proyecto Aracne* de la Fundéu BBVA (2016),

utilizaremos los términos *densidad* léxica, *riqueza léxica* y *diversidad léxica* como sinónimos.

Tradicionalmente, la riqueza léxica de los textos se ha medido a través de la denominada TTR (*Type-Token Ratio*), donde *token* se refiere al número total de palabras que un texto contiene (*casos*, en español) y *type* al repertorio de palabras distintas (*tipos*, en español) (Bergman y Paavola, 2003). Cuanto más variado sea el vocabulario de un texto, mayor diversidad léxica tendrá, lo que significa que estará compuesto por un alto número de palabras distintas y poco repetidas. La relación tipo-caso se obtiene de dividir los primeros entre los segundos:

TTR=tipo/caso

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

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

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

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

Tiene una ratio de 0,83, puesto que, de 12 palabras que la componen, hay una ("se") que se repite dos veces. Por lo tanto: 10/12=0,83.

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

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

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

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

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

7.2.1 Metodología

Para poder llevar a cabo el estudio, elaboramos un corpus con la información de las cuentas de cuatro escritores españoles de reconocido éxito, con el objetivo de estudiar su densidad léxica. Se trata de escritores contemporáneos y aficionados a compartir públicamente en *Twitter* ideas, sentimientos, relatos, análisis o noticias, prácticamente a diario; estos autores son: Mónica Carrillo (@MonicaCarrillo), Arturo Pérez Reverte (@perezreverte), Daniel Sánchez Arévalo (@sanchezarevalo) y Lucía Etxebarria (@LaEtxebarria).

Para ello, introdujimos los nombres de usuario de cada uno de ellos en el buscador y seleccionamos la opción de tamaño de muestra grande. Recordemos que la cantidad que *Twitter* nos provee en el tamaño de muestra grande es de, aproximadamente, los últimos 3.500 tuits, de donde se excluyen los retuits, con lo que, lo más habitual, es que esa cifra resulte en una menor. Una vez realizado este paso, *Wordics* nos devuelve de manera automática los datos correspondientes a cada usuario. Podemos consultar el corpus de cada una de estas cuentas en los anexos 1.1, 1.2, 1.3 y 1.4.

7.1.2 Resultados

Los resultados obtenidos con la herramienta los mostramos a continuación, de forma individual para cada escritor:

Mónica Carrillo:

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

Figura 7.1. Relación tipo/caso en la cuenta de Mónica Carillo

Arturo Pérez-Reverte:

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

Figura 7.2. Relación tipo/caso en la cuenta de Arturo Pérez-Reverte

Daniel Sánchez Arévalo:

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

Figura 7.3. Relación tipo/caso en la cuenta de Daniel Sánchez Arévalo

Lucía Extebarria:

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

Figura 7.4. Relación tipo/caso en la cuenta de Lucía Etxebarria

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

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

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

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

Figura 7.5. Comparación densidad léxica entre los autores analizados

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

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

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

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

En el momento de utilizar esta función de la herramienta, hemos seleccionado el punto de la gráfica coincidente con 20.000 palabras en los cuatro casos. El motivo de esta cifra es que se buscaba un valor alto, pero que todos los escritores compartieran, para que pudiera hacerse la comparación. Como Pérez-Reverte es el que menos casos tiene, se ha escogido una cifra próxima a su máximo de palabras. A continuación mostramos las gráficas individuales, donde la línea roja diagonal representa el 100% de los casos y que la línea indica el número de tipos en relación a los casos concretos, según vayan aumentando a lo largo del tiempo. Como se puede observar, aparece, en verde, el número de casos seleccionado –20.000– y, en color negro, los tipos contabilizados dentro de ese número de casos. El porcentaje de la densidad es el resultado de la división del número de tipos entre el número de casos y su multiplicación por 100.

Mónica Carrillo:

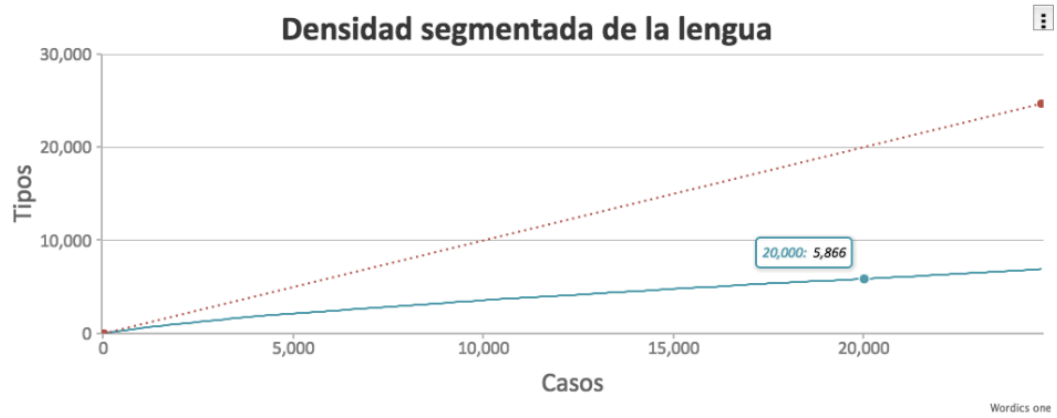


Figura 7.6. Densidad léxica segmentada de Mónica Carrillo

Arturo Pérez-Reverte:

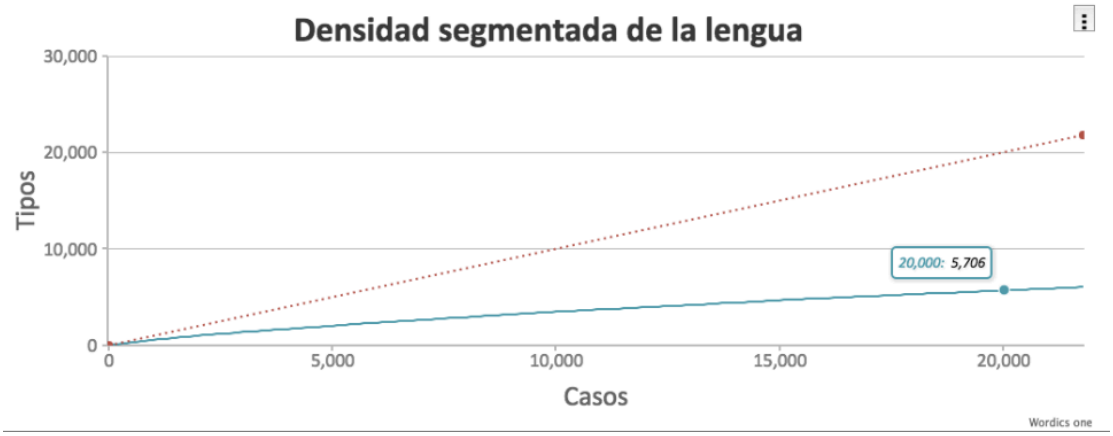


Figura 7.7. Densidad léxica segmentada de Arturo Pérez-Reverte

Daniel Sánchez Arévalo:

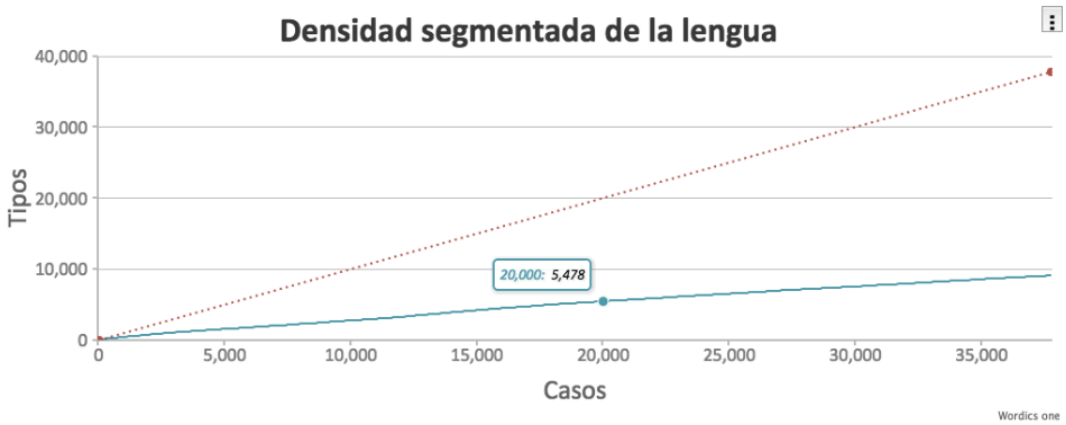


Figura 7.8. Densidad léxica segmentada de Daniel Sánchez Arévalo

Lucía Etxebarria:

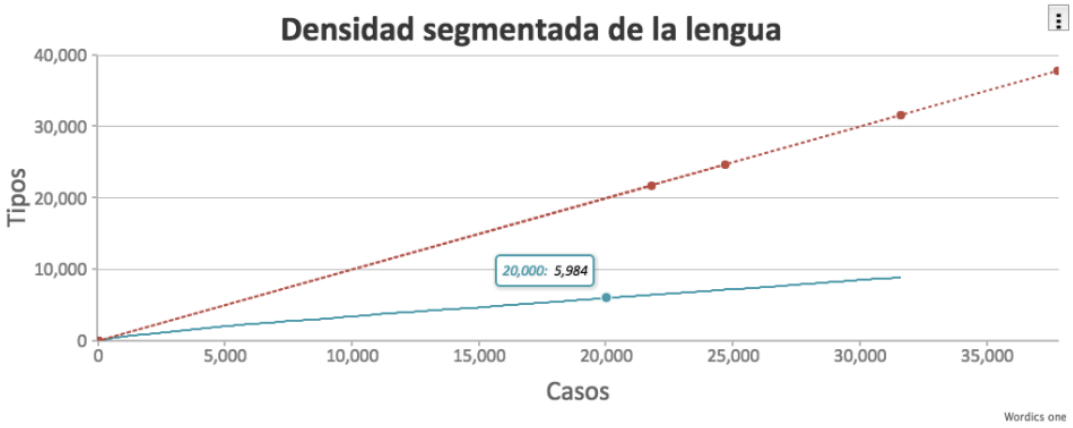


Figura 7.9. Densidad léxica segmentada de Lucía Etxebarria

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

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

Figura 7.10. Comparación densidad léxica segmentada entre los autores analizados

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

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

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

7 nov 2015, 13:39 h.	@faustianovich Te he tocado la patata? ;) Me alegro. (Me acabo de dar cuenta que este tweet nunca se envió. Qué cosas)
25 may 2014, 22:18 h.	Las radios del Spotify siempre empiezan bien, con buen criterio en base a tu selección, pero luego se les empieza a ir la olla cosa fina.
01 mar 2016, 21:09 h.	@Lucia_Alvarez_ Jajaja, qué maja! Llévatela a la isla a jugar con Noesunponi!
15 feb 2016, 21:24 h.	@YUSAN_5 Jajajaja. A mí también me falta lamentablemente mi talento
20 mar 2013, 20:57 h.	@SallyBurton @ariadnunii @quimyo A nosotros aunque vayamos de guays y tal en plan estrellitas del celuloide, también nos animais los días ;)
30 jun 2014, 23:37 h.	@iguardans Pues sí, deberían estar en plan SGAE, que te pillan siempre en cualquier lugar remoto del planeta donde pongas un chunda chunda

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

Es posible que aquí resida la explicación a la menor variedad léxica del escritor, ya que los porcentajes de densidad varían según se trate de material oral o escrito. Ure

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

7.2 ESTUDIO DEL LENGUAJE PERIODÍSTICO

En esta sección, con el objetivo, una vez más, de ejemplificar algunos de los posibles usos de *Wordics One*, nos acercamos al lenguaje periodístico para llevar a cabo un estudio sobre el lenguaje sexista en los medios de comunicación, analizando la información obtenida de las cuentas de cinco periódicos españoles, tres nacionales y dos locales de la ciudad de Córdoba. Los periódicos estudiados son: *El País*, *El Mundo*, *ABC* –en el ámbito nacional– y *Diario Córdoba* y *Cordópolis* (que solo ofrece versión digital) –en el ámbito local.

Para el estudio, nos hemos guiado por las características sexistas presentes en el lenguaje de los medios de comunicación enumeradas por la profesora Guerrero Salazar (2000, 2006, 2006, 2007a)[45], que son, en líneas generales:

1. Abuso del masculino genérico.
2. Utilización del término *hombre*.
3. Oficios y profesiones en femenino.
4. Androcentrismo y salto semántico.
5. Duales aparentes y vocablos ocupados.
6. Disimetría en el tratamiento de los sexos.

Antes de comenzar, no obstante, aclaramos el concepto de sexismo lingüístico, según la definición de Álvaro García Meseguer, quien asegura que:

[45] Seguimos aquí a Guerrero Salazar (2007a) para analizar las características sexistas del lenguaje en los medios de comunicación, pero, debido a las características de nuestro trabajo, no nos ocupamos de cuestiones teóricas sobre tan complejo problema, con implicaciones ideológicas, sociales y culturales. Para ello, se pueden consultar las obras de López y Morant (1991); Lozano (1995); Calero Fernández (1999); Calero Vaquera (2003a y 2003b); Lledó (coord.), Calero Fernández y Forgas (2004); Márquez, (2013) o Bengoechea (2015).

un hablante incurre en sexismo lingüístico cuando emite un mensaje que, debido a su forma (es decir, debido a las palabras escogidas o al modo de enhebrarlas) y no a su fondo, resulta discriminatorio por razón de sexo. Por el contrario, cuando la discriminación se debe al fondo del mensaje y no a su forma, se incurre en sexismo social. (García Meseguer, 2001: 20)

7.2.1 Metodología

Se han estudiado, como decimos, las cuentas de *Twitter* de cinco periódicos españoles distintos y se ha seleccionado el tamaño de muestra grande para obtener un mayor número de tuits. Recordemos que, para acceder a una cuenta en particular, es necesario introducir en el buscador el nombre de usuario exacto, con o sin @. Para poder detectar la cuenta más fácilmente, el buscador que da acceso a las cuentas no discrimina entre mayúsculas y minúsculas, con lo que no es necesario ser exhaustivo en lo que al tipo de letra se refiere. Los nombres de usuario y el total de tuits obtenidos de cada perfil son los siguientes:

PERIÓDICO	NOMBRE DE USUARIO	NÚMERO TOTAL DE TUITS
El País	@el_pais	1691
El Mundo	@elmundoes	1956
ABC	@abc_es	3018
Diario Córdoba	@CORDOBA_diario	2666
Cordópolis	@cordopolis_es	3180

Figura 7.11. Datos básicos de las cuentas de periódicos analizadas

En las figuras siguientes, mostramos una gráfica con el número de tuits publicados por cada cuenta en el período de tiempo filtrado. Es importante tener en consideración que, aunque se haya seleccionado el tamaño de muestra grande para todos los casos, el número de resultados varía entre unas cuentas y otras, fundamentalmente porque no se han incluido los retuits. Por ello, en algunos casos, los resultados corresponden a un período de tiempo más largo, mientras que en otros están más concentrados en menos meses.

Periódico *El País*:

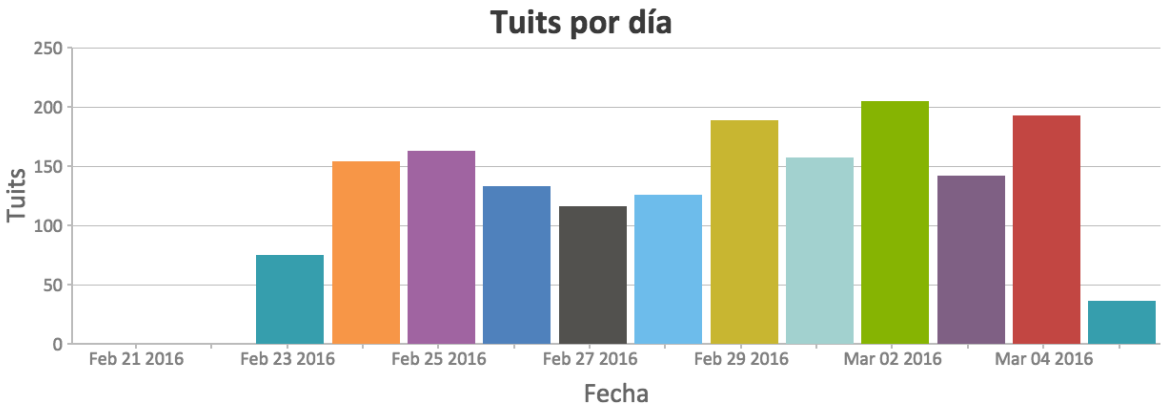


Figura 7.12. Distribución de tuits del periódico *El País*

Diario *El Mundo*:

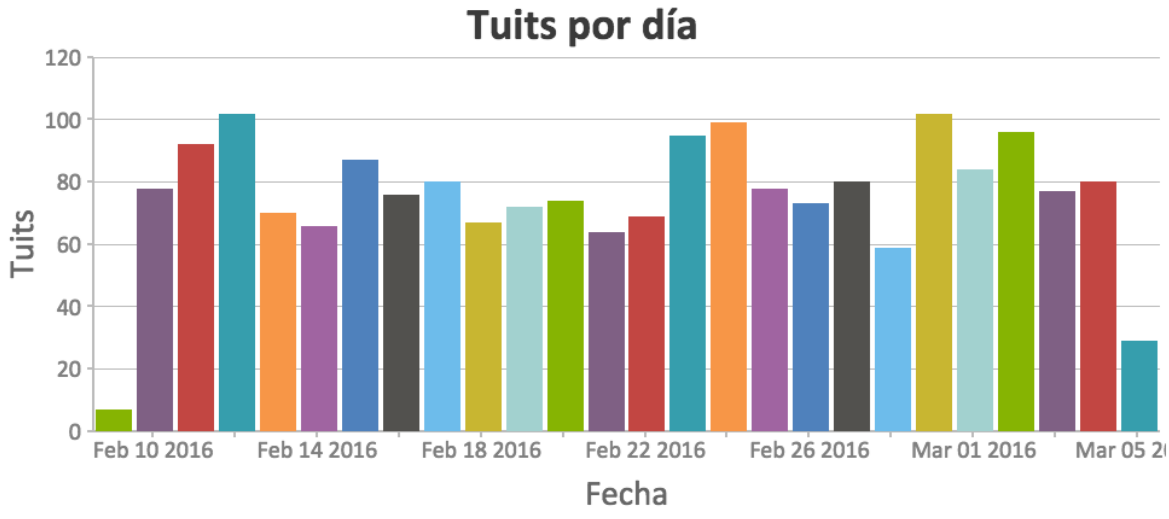


Figura 7.13. Distribución de tuits del diario *El Mundo*

Diario *ABC*:

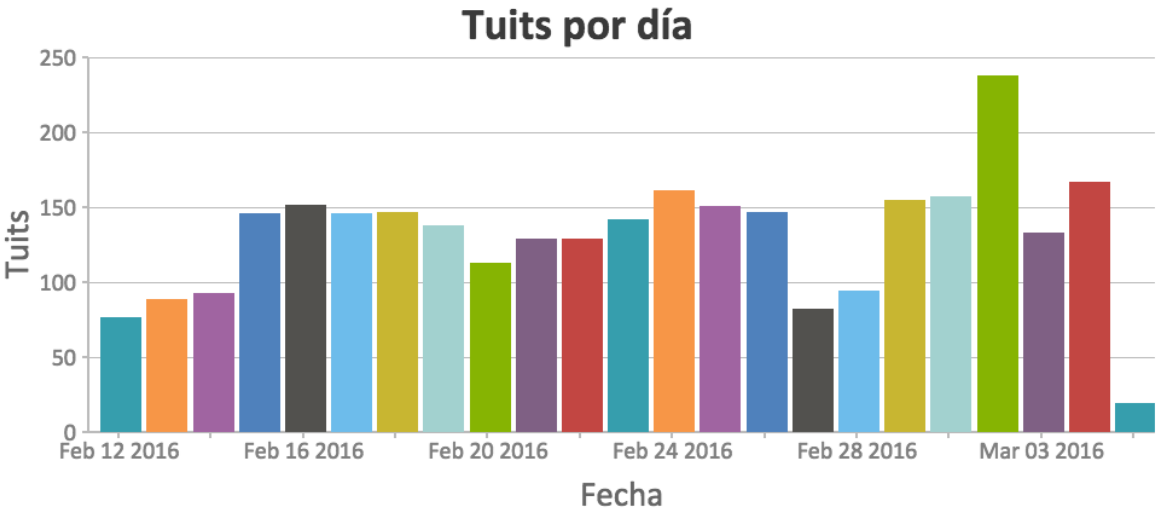


Figura 7.14. Distribución de tuits del diario *ABC*

Diario Córdoba:

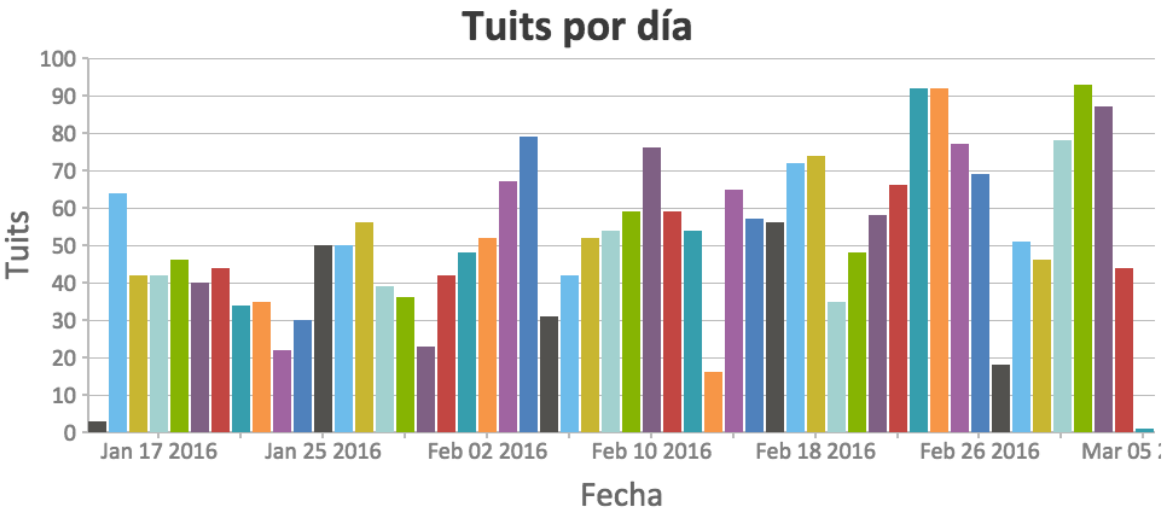


Figura 7.15. Distribución de tuits del *Diario Córdoba*

Periódico *Cordópolis:*

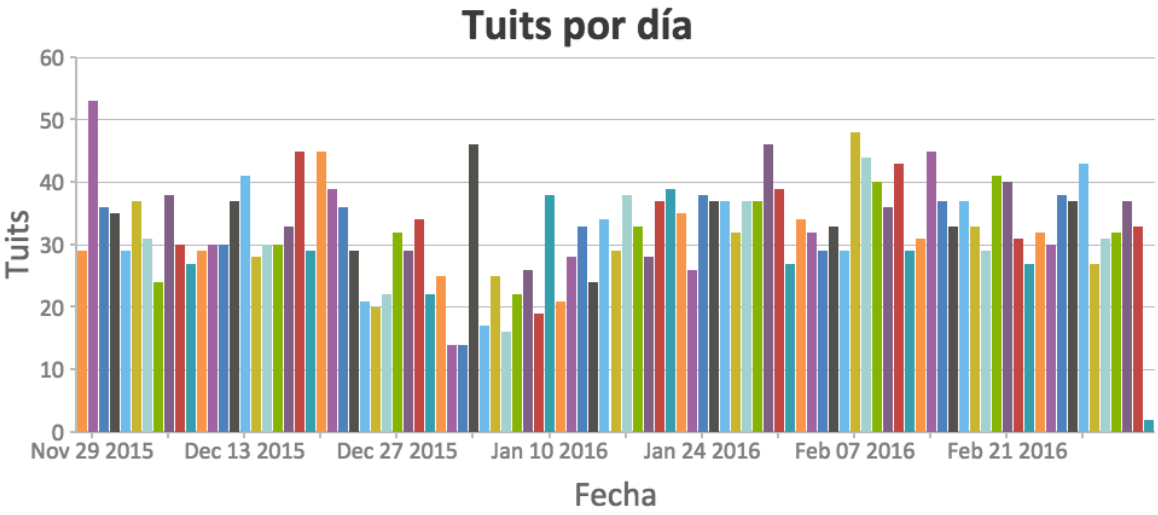


Figura 7.16. Distribución de tuits del periódico *Cordópolis*

La lista total de los tuits generados por las distintas cuentas se pueden consultar en los anexos 2.1, 2.2, 2.3, 2.4 y 2.5.

Para el estudio, hemos utilizado la opción de filtrado de palabras que aparece en la parte superior del cuadro que muestra los tuits, de manera que hemos ido buscando los distintos elementos uno por uno, comparando en todos los periódicos. Conviene recordar, una vez más, que el programa trabaja con expresiones regulares, lo que nos facilita mucho la tarea a la hora realizar búsquedas por lemas o fragmentos de palabras. Así, si introducimos en el buscador la combinación de letras "s-e-ñ-o-r", la herramienta nos devolverá todos los resultados que contengan esa combinación: *señor*, *señora*, *señores*, *señoras*, *señoría*, *señorías*, *señorita*, *señorito*, *señoritas*, *señoritos*, *señorío*, etc.

Dado que el objetivo de este estudio es analizar el lenguaje periodístico general, y no comparar el lenguaje que utilizan los distintos periódicos entre sí, mostraremos aquí la suma de los resultados. Para profundizar en los diarios de forma individual, se pueden consultar los anexos señalados unas líneas más arriba.

7.2.2 Resultados

7.2.2.1 Abuso del masculino genérico

La utilización en español del género masculino puede hacerse desde dos perspectivas distintas. La primera de ellas se refiere al sexo masculino y la segunda engloba a los dos sexos de forma genérica. Esto se debe al doble valor que el género masculino adopta en nuestro idioma. Guerrero (2007a) advierte de que el lenguaje periodístico abusa del uso del masculino genérico debido a este carácter integrador y que esto puede provocar, no solo una discriminación hacia el sexo femenino, sino también cierta confusión y ambigüedad en algunos casos. Estos problemas morfosintácticos, como los denominan Ayala Castro, Guerrero Salazar y Medina Guerra (2002), se manifiestan en el uso de términos como *empleados*/*desempleados*, *trabajadores* o *ciudadanos* con un sentido genérico para referirse a ambos sexos.

En el estudio, confirmamos que así es en los cinco periódicos consultados, después de analizar la utilización de estos tres términos –*empleados*, *trabajadores*, *ciudadanos*– y sus derivados, y contrastarlos con sus formas femeninas correspondientes.

Del total de tuits analizados, 12.511, los resultados obtenidos para el término *(des) empleados* ha sido de 25 ocurrencias; para *(ex)trabajadores*, 30; y, para *ciudadanos*, 196. Frente a estas cifras, las formas femeninas se han utilizado en muy pocos casos y, en ningún momento, para referirse a ambos sexos o para complementar al género gramatical masculino y facilitar así la inclusión del sexo femenino. Estos son los resultados:

PERIÓDICO	*(DES)EMPLEADOS/ (DES)EMPLEADAS*			*(EX)TRABAJADORES/ (EX)TRABAJADORAS*			*CIUDADANOS/ CIUDADANAS*		
	Masc.	Fem.	Ambos	Masc.	Fem.	Ambos	Masc.	Fem.	Ambos
El País	3	0	0	5	0	0	36	0	0
El Mundo	4	0	0	5	1	0	49	0	0
ABC	5	2	0	3	0	0	52	0	0
Diario Córdoba	10	0	0	10	0	0	20	0	0
Cordópolis	3	0	0	7	0	0	40	0	0

Figura 7.17. Resultados obtenidos para el masculino genérico

Los casos en los que se utilizan las formas femeninas –solo 2 para *empleadas*, en el *ABC* y 1 para *trabajadoras*, en *El Mundo*– podemos ver cómo se refieren, exclusivamente, a trabajadoras, sin ningún ánimo de inclusión o de complementación:

2 mar 2016, 23:00 h.	Una empresa inglesa dará descanso a sus empleadas en sus días de regla https://t.co/fOZ1zvc0jc https://t.co/UXQIKdEcay
26 feb 2016, 12:15 h.	El Gobierno reconoce el nuevo permiso de gestación también a las empleadas de la Administración https://t.co/MkGyMaWqj9
3 mar 2016, 7:11 h.	¿Una baja especial para trabajadoras con la menstruación? https://t.co/CbGqpn9POj https://t.co/LsMUsbDdo7

Por otra parte, el número considerablemente más alto de ocurrencias de *ciudadanos* con respecto a los otros dos casos estudiados se explica porque la inmensa mayoría de los tuits que contienen esta palabra se refieren al partido político liderado por Albert Rivera. He aquí dos ejemplos en los que la palabra *ciudadanos* no aparece como el nombre propio del partido y se utiliza como masculino genérico:

4 mar 2016, 14:55 h.	"Hay 15 millones de ciudadanos que han pedido cambio". En esta cifra, Cándido Méndez incluye a Ciudadanos https://t.co/NK9jlorK8J
5 mar 2016, 7:16 h.	Una 'app' implica a los ciudadanos para prevenir la llegada del Zika a España https://t.co/Kc7qhaGWL1 https://t.co/NzawxdoSmm

A continuación, mostramos también algunos ejemplos para que se pueda ver el uso de los otros dos términos en contexto.

(Des)empleados:

27 feb 2016, 10:37 h.	Urdangarin declara desconocer cómo funcionaba su empresa https://t.co/5avUR3xHKq Y dice que se acaba de enterar de los empleados falsos
26 feb 2016, 7:28 h.	¿Qué es una diputación? ¿Cuántas hay? ¿Cuántos empleados tienen? Radiografía de un organismo en peligro de extinción https://t.co/aSKAdkajAb
4 mar 2016, 16:11 h.	El Gobierno de Barcelona expulsa a los empleados públicos de la mutua municipal, ahogada por el déficit https://t.co/yFiqCRwfdO
25 feb 2016, 19:07 h.	Revés judicial a la obligación de que los empleados de la Generalitat se hablen en catalán https://t.co/uE4lkpnEDH https://t.co/ui9vVwSThV
26 feb 2016, 18:20 h.	Denuncian el «sádico» maltrato de los empleados de un matadero «ecológico» en Francia https://t.co/pxY7Zvibaw

	https://t.co/2aRqYP7S6g
26 feb 2016, 14:55 h.	El presidente de Abengoa comunica a los empleados que no hay liquidez para pagar las nóminas de febrero https://t.co/2NXrKGO8n9
18 feb 2016, 11:00 h.	@panasonic iguala en beneficios laborales a sus empleados homosexuales https://t.co/78MsPsp3t8
31 ene 2016, 10:36 h.	Unos 250.000 empleados públicos cobrarán a finales de febrero la paga extra de 2012 #CórdobaEsp https://t.co/q7nPHyc0UN
29 ene 2016, 19:06 h.	La Junta posibilita la contratación de 165 desempleados en 42 entidades sin ánimo de lucro https://t.co/k65MOG7lD1

(Ex)trabajadores:

3 mar 17:05 h.	Infraestructuras negociará con los trabajadores si no puede con los sindicatos https://t.co/qs6cGGyZoN https://t.co/1uvm1XyiGZ
1 mar 16:35 h.	Uno de cada seis trabajadores sufre acoso y no lo denuncia, según CSIF https://t.co/VAcFBnU6qm https://t.co/MvMmjbTITe
8 dic 9:05 h.	La justicia obliga a Aneri a readmitir a sus trabajadores https://t.co/NVKN8AAmL7 https://t.co/0T0VM7ztdA
30 nov 15:05 h.	Los trabajadores de los contratos contra la exclusión limpiarán la ciudad https://t.co/uAOK6sYa7I https://t.co/zb5bZBRKyQ
4 mar 20:25 h.	#FCC planea despedir a 750 trabajadores en España #Desempleo https://t.co/3rM7qIsxcQ
1 mar 12:22 h.	Uno de cada seis trabajadores sufre acoso laboral, según CSIF #CórdobaEsp https://t.co/A5QgzyFKAx
1 mar 16:20 h.	.@Renfe contratará a 465 nuevos trabajadores en 2016 https://t.co/VqvmcZ9vvB
21 feb 2:29 h.	La reunión entre trabajadores y TMB acaba sin acuerdo y se mantiene la huelga en Barcelona https://t.co/P1ZlehQwGb https://t.co/PKxIpVaLBD
3 mar 14:30 h.	Mercadona tiene 75.000 trabajadores y el 90% recibe más de 1.400 euros, según la empresa https://t.co/BGrdLNFRBI

Como podemos ver, en los tuits se usa el masculino genérico para referirse a ambos sexos y no se utiliza ningún recurso alternativo para evitar su asociación exclusiva con el sexo masculino; entre esos recursos podrían encontrarse: el uso de metonimias; el uso de sustantivos abstractos, colectivos, epicenos o comunes; pronombres, determinantes y adjetivos no marcados; perífrasis o desdoblamientos (Guerrero, 2007a: 317-318). No obstante, para una visión completa de la utilización de estas palabras en los periódicos consultados, se puede realizar la búsqueda con la opción de KWIC y la de colocaciones.

Además, los resultados varían considerablemente en función del grupo objeto de estudios. Mientras que los medios de comunicación centran su atención en la eficacia

comunicativa y en la economía del lenguaje, combinadas con una lectura sencilla y adecuada, el lenguaje político, por el contrario, pretende llegar a la mayor parte de público posible, lo que se refleja en un uso mucho más habitual del desdoblamiento de las palabras para simultanear los géneros masculino y femenino. Simplemente a modo de muestra –puesto que el estudio se centra en los medios de comunicación–, aportamos a continuación (figura 7.19) algunas concordancias de palabras utilizadas por algunos líderes políticos españoles, con las que nos podemos hacer una idea de la frecuencia de uso. En el buscador de concordancias hemos introducido las palabras *todos*, en masculino, para comprobar su concordancia y el número de ocurrencias con la forma femenina. Puesto que no es el objeto de estudio, solo mostramos el uso de la expresión *todos y todas*, pero la realidad es que el desdoblamiento se usa con mucha más frecuencia que en el lenguaje periodístico. La búsqueda se ha realizado sobre las cuentas de Susana Díaz (@susanadiaz), Pedro Sánchez (@sanchezcastejon), Albert Rivera (@Albert_Rivera), Pablo Iglesias (Pablo_Iglesias_) y Mariano Rajoy (@marianorajoy). Para hacer más sencilla la lectura, situamos las concordancias una debajo de otra, este es el motivo por el que aparecen las palabras repetidas y los cuadros cortados. El orden de políticos es el establecido unas líneas más arriba.

Político	L3	L2	L1	hombre	R1	R2	R3
Susana Díaz	0	0	0	#	0	1	0
Pedro Sánchez	0	1	0	#	0	3	1
Albert Rivera	0	0	0	#	0	2	1
Pablo Iglesias	0	0	0	#	0	0	0

Figura 7.18. Colocaciones de *todos* y *todas* en los políticos estudiados

La primera colocación, perteneciente a la presidenta de la Junta de Andalucía, Susana Díaz, indica que en una ocasión utiliza la expresión *todos y todas*, puesto que *todas* aparece dos puestos a la derecha de todos. Pablo Iglesias –líder de *Podemos*– que se corresponde con la segunda línea, es el que más veces utiliza las dos palabras cerca una de otra; como podemos ver, en una ocasión sitúa *todas* dos palabras a la izquierda de *todos*, en tres ocasiones, dos a la derecha y en una ocasión tres espacios a la derecha. El presidente del *Partido Socialista*, Pedro Sánchez, el siguiente, escribe en dos ocasiones *todas* dos puestos a la derecha de *todos* y, en una ocasión, tres puestos a la

derecha. Por último, Albert Rivera, número uno de *Ciudadanos*, utiliza una vez las dos palabras en la misma oración, pero alejadas la una de la otra, puesto que no aparece ni en los tres puestos inmediatamente anteriores ni en los tres inmediatamente posteriores. Esto quiere decir que, aunque estén relativamente cerca, no las utiliza con la intención de obtener un lenguaje más inclusivo, y así lo confirmamos después de buscar el tuit en cuestión:

Nótese que hemos seleccionado tuits de cinco líderes y solo aparecen cuatro líneas de concordancias; esto se debe a que el presidente del *Partido Popular*, Mariano Rajoy, no introduce en la misma oración en ningún caso los términos *todos* y *todas* juntos. Aparentemente, podríamos decir que los partidos de izquierdas están más a favor de la utilización de un lenguaje inclusivo y no sexista, mientras que los que se encuentran más escorados hacia la derecha no le prestan tanta atención a este aspecto. Pero, una vez más, insistimos en que no es el lenguaje político el objeto de nuestro estudio, por lo que no consideramos conveniente detenernos más en este aspecto.

7.2.2.2 El caso del término *hombre*

La profesora Guerrero hace mención especial al uso, por parte de los medios de comunicación, del término *hombre*, que, igual que el masculino genérico del que acabamos de hablar, también posee un doble valor. Por un lado, el que se refiere exclusivamente al sexo masculino y, por el otro, el que encierra el significado de "ser humano". Afirma Guerrero (2007a) que es habitual un uso sistemático del término *hombre* como genérico de ambos sexos en el lenguaje de la prensa y que eso conlleva un comportamiento sexista que es necesario evitar.

Tras el filtrado del término *hombre* (así la búsqueda también incluirá la forma en plural, gracias a las expresiones regulares) en el buscador, podemos constatar que la afirmación de Guerrero aún sigue teniendo vigencia hoy en día, aunque con menos frecuencia que la utilización de un término alternativo más inclusivo. De hecho, de las 131 ocurrencias registradas en los cinco periódicos con los términos *hombre* u *hombres*, en solo 5 se da esta circunstancia. Además, una de ellas es una cita literal del Papa Francisco –el último ejemplo de la tabla–, con lo que no puede atribuírsele al periódico.

A continuación mostramos una tabla con los cinco casos, elaborada después de la revisión en los cinco periódicos:

1 mar 2016, 11:41 h.	Desde Gagarin al Sputnik pasando por Laika. Los 10 grandes hitos del hombre en el espacio exterior, en FOTOS https://t.co/uZFnnPjHAK
5 mar 2016, 5:18 h.	80 años del desastre del #Hindenburg, el dirigible más espectacular construido por el hombre https://t.co/QI62zcUHDT https://t.co/mi0QXXtksX
25 feb 2016, 10:18 h.	'Peca el cura, delinque el hombre'. https://t.co/sA5SMS9zBU Por @PERYRIERA https://t.co/UbZVrdC8vN
3 mar 2016, 2:01 h.	Hombre y coche compiten para ser los más rápidos sobre nieve en este vídeo https://t.co/fKxM324K0d ¿Quién ganará? https://t.co/4pD55i83uu
14 feb 2016, 9:45 h.	El Papa Francisco, a la Iglesia católica mexicana: "si tienen que pelearse peléense como hombres, a la cara". https://t.co/oVzu7CqCB5

Por el contrario, está bastante generalizado el uso de *persona(s)* o *humano(s)*, que resultan en un lenguaje menos sexista y más inclusivo. De *persona(s)*, en primer lugar, hay un total de 116 ocurrencias y todas ellas están referidas al conjunto de hombres y mujeres que integran un grupo. En la tabla siguiente podemos observar algunas de los usos en contexto del término.

Previo	Palabra	Posterior
... Dos traficantes de	personas	sirios, condenados a cuatro años de cár ...
... México se han reportado 2.818 casos de	personas	con influenza. 98 han fallecido https:/ ...
... SD88N El paro registrado crece en 2.231	personas	en febrero ...
... Photoshop para cambiar el físico de las	personas	". La carta de @tentaciones a Lena Dunha ...
... Unas 200	personas	se concentran ante la puerta de la cárc ...
... amigos en el aeropuerto hasta ayudar a	personas	ciegas en su aventura ...
... 82	personas	evacuadas del puerto de Cotos por la ni ...
... Cuatro	personas	fueron apuñaladas durante una protesta ...
... s que causaron la muerte de casi 10.000	personas	, Nepal sigue existiendo https://t.co/RS ...

En segundo lugar, contabilizamos 28 apariciones entre *humano* y *humanos*, utilizadas de la misma forma que *persona* y *personas*. Aquí es necesario tener en cuenta que, en algunos de estos casos, la palabra funciona como adjetivo, a diferencia de la anterior, que solo puede funcionar como sustantivo. Veamos algunas de las palabras clave en contexto:

Previo	Palabra	Posterior
... El informe de derechos	humanos	de la CIDH molestó a México https://t.c ...
... 50 años que una máquina construida por	humanos	se posó sobre un planeta extraterrestre ...
... ¿Qué pasaría si los	humanos	hibernáramos? Unos, más felices que otr ...

Por otro lado, también es frecuente encontrar el recurso de la utilización simultánea de los términos masculino y femenino, para evitar la discriminación. Sin embargo, observamos que esto ocurre, principalmente, cuando ambos términos se oponen semánticamente y se hace necesaria la aclaración, como ocurre en los siguientes ejemplos:

29 feb 2016, 2:30 h.	Muchos están decidiendo no tener hijos https://t.co/jxzdFU7HRr ¿Qué lleva a hombres y mujeres a no querer procrearse?
2 mar 2016, 12:30 h.	UGT reivindica un reparto de poder igualitario entre hombres y mujeres #CórdobaEsp https://t.co/uTkejBlFt9
23 feb 2016, 13:05 h.	. @UGTCordoba sitúa la brecha salarial entre hombres y mujeres en un 24% https://t.co/vB8oFgTPgy
30 ene 2016, 13:45 h.	Fiebre runner: 65 carreras en 2016 y los mismos premios para hombres y mujeres. @IMDECO https://t.co/msUcBVmN4m https://t.co/I2B5dtpEZG

Además, ni siquiera en estos casos es muy frecuente encontrar esta combinación (*hombres y mujeres*); lo podemos comprobar con las tablas de colocaciones generadas tras la búsqueda en los distintos periódicos. En estos ejemplos, buscamos el término *hombres* para comprobar la frecuencia y el lugar en los que lo acompaña el término *mujeres* (el término *hombres* aparece representado por el símbolo almohadilla –#).

Periódico *Cordópolis*:

Palabra	L3	L2	L1	Palabra	R1	R2	R3
20	0	0	0	#	0	0	0
54	0	0	0	#	0	0	0
65	0	0	0	#	0	0	0
92	0	0	0	#	0	0	0
2016	0	0	0	#	0	0	0
Las	0	0	0	#	0	0	0
mujeres	0	0	0	#	0	1	0

Diario Córdoba:

mujeres	0	0	0	#	0	2	0

Diario *ABC*:

mujeres	0	0	0	#	0	0	0

Periódico *El País*:

mujeres	0	1	0	#	0	1	0

No se incluye información referente al diario *El Mundo* porque este no muestra ninguna concordancia entre estos dos términos (*hombres*, *mujeres*). En el resto de periódicos, como podemos ver, aunque hay algunas ocurrencias, son muy escasas. Así, mientras que *Cordópolis* solo escribe una vez la palabra *mujeres* dos lugares a la derecha de la palabra *hombres*, el *Diario Córdoba* lo hace dos. *El País*, por su parte, muestra una ocasión en la que se manifiesta de igual modo que en los casos anteriores y otra en la que la palabra *mujeres* se sitúa dos lugares a la izquierda de *hombres*. El caso del *ABC*, donde vemos que *mujeres* está en el cuadro de colocaciones, pero los tres valores de la izquierda y de la derecha dan cero, ocurre que la el término *mujeres* aparece en la misma oración que *hombres*, pero no en los tres lugares inmediatamente anteriores ni posteriores.

En conclusión, comprobamos que, a diferencia del uso muy extendido del género gramatical masculino genérico para hacer referencia tanto a hombres como a mujeres, el término *hombre* con significado de "persona" es mucho menos frecuente.

7.2.2.3 Oficios y profesiones en femenino

El hecho de que tradicionalmente los trabajos públicos hayan sido desempeñados por hombres ha provocado que los sustantivos referidos a ellos hayan estado siempre dominados por el género masculino. La incursión laboral progresiva de la mujer, sin embargo, junto con los avances en sus derechos, han removido los cimientos de la sociedad, alcanzando también las reglas lingüísticas. Así, conforme las mujeres han ido conquistando los oficios históricamente masculinos, la lengua ha ido amoldándose a los cambios con el objetivo de reflejar a la realidad.

No obstante, estos cambios no están siendo sistemáticos ni integrales. A pesar de que el Diccionario de la Real Academia ha aceptado la inmensa mayoría de ellos (*médica*, *bedela*, *edila*, *música*, *arquitecta*, *jueza*, *fiscala*, etc.) (ver última edición, 23ª, 2014) aún quedan algunos, como *conserja* o *albañila*, que no se recogen en la última edición. Además, la aceptación académica no garantiza, casi en ningún caso, su adopción efectiva por parte de los hablantes del español, como veremos enseguida. Su uso o la ausencia de él puede estar determinada por dos factores fundamentales, como son el peso de la tradición y el hecho de que algunas de esas profesiones (como la de *albañil*) siguen siendo desempeñadas mayoritariamente por hombres. Sin embargo, el hecho de que, siguiendo la lógica anterior, términos como *conserja* no estén admitidos, a pesar del gran número de mujeres dedicadas a este oficio, nos lleva a afirmar que la aceptación académica de los nuevos femeninos no está siendo sistemática[46].

En cualquier caso, cuando se opta por la utilización de la forma masculina, se suele adoptar la forma común del sustantivo en cuanto al género, es decir, se cambia el artículo o el determinante para diferenciar el sexo al que designa la palabra en cuestión: *el arquitecto*/*la arquitecto*, *el músico*/*la músico*.

En los periódicos analizados se refleja muy bien esta situación, ya que no encontramos un patrón concreto de uso; incluso algunos periódicos son ambiguos y adoptan indistintamente una forma u otra. Por ejemplo, si tomamos el caso de *juez* y *jueza*, *El País*, *El Mundo* y el *Diario Córdoba*, usan sistemáticamente el género común:

Periódico *El País*:

Previo	Palabra	Posterior
... El policía que dijo ante el	juez	que los manifestantes mantuvieron "una ...
... ÚLTIMA HORA | La	juez	cita a la infanta Cristina en el 'caso ...
... nuestra obra". Los titiriteros piden al	juez	que archive el caso https://t.co/1wukcC ...
... DIRECTO https://t.co/gy1SZT232Y La	juez	reprende a la abogada de Manos Limpias ...
... DIRECTO https://t.co/gy1SZTjEry La	juez	ha impedido contestar a Urdangarin porq ...
... Un	juez	respalda la negativa de Apple al desblo ...
... laya de Tarifa llegaron a Gibraltar. Un	juez	investiga su posible robo https://t.co/ ...
... La	juez	archiva por segunda vez el caso de Dieg ...
... Un	juez	investiga si Ignacio González recibió d ...
... El	juez	lleva a los primeros cargos del PP vale ...
... repartieron https://t.co/xfQYP2O9Zx El	juez	les ha condenado ...
... llevado al artista Abel Azcona ante el	juez	https://t.co/8OGuqnjKBz ...
... Pujol, al	juez	: "No podía afrontar el riesgo político ...

[46] Para un análisis detallado sobre la inclusión de lo femenino en el DRAE, se puede consultar la obra de Lledó (coord.), Calero y Forgas (2004).

Diario *El Mundo*:

Previo	Palabra	Posterior
... El	juez	archiva el caso de Diego, el niño que s ...
... El	juez	archiva el caso de Diego, el niño que s ...
... El	juez	archiva el caso de Diego, el niño que s ...
... El	juez	amplía el 'caso Rato' con pagos de Laza ...
... #ÚltimaHora La	juez	envía a prisión al dueño y a tres direc ...
... #ÚltimaHora El	juez	rechaza imputar a Carmena y Celia Mayer ...
... ece Antonin Scalia, el ultraconservador	juez	del Tribunal Supremo de EEUU. https://t ...
... ece Antonin Scalia, el ultraconservador	juez	del Tribunal Supremo de EEUU. https://t ...
... #ÚltimaHora El	juez	Pedraz propone juzgar a Zapata por sus ...
... #Ampliamos El	juez	lanza una operación contra la financiac ...

Diario Córdoba:

Previo	Palabra	Posterior
... capitó a una niña en Moscú dice ante el	juez	que se lo "ordenó Alá" https://t.co/0J8 ...
... El	juez	cita a declarar a un forense por uno de ...
... El	juez	cita a un forense por uno de los casos ...
... El	juez	lleva a juicio la #financiación ilegal ...
... La	juez	Alaya resolverá el recurso de Emerita c ...
... El	juez	inicia los trámites para imputar a Barb ...
... Un	juez	declara nulo el despido de una trabajad ...
... La	juez	cita a declarar por prevaricación a dos ...
... Rato, de nuevo ante el	juez	para aclarar el origen de su fortuna #C ...
... El	juez	ordena 5 registros por posibles pagos d ...
... El 23% de menores internados por el	juez	están por maltrato familiar https://t.c ...
... La	juez	Núñez aplaza al día 16 la declaración d ...
... La	juez	Núñez Bolaños confirma la división en c ...
... Un	juez	de Jaén no considera maltrato animal la ...
... Anticorrupción pide al	juez	que cite a declarar como investigados a ...
... El	juez	deja en libertad provisional a #Rus con ...
... ÚLTIMA HORA / El	juez	manda a prisión provisional a los deten ...
... lanos detenidos esperan la decisión del	juez	tras declarar: Tres diputados de Syriza ...
... La	juez	Núñez confirma la pieza separada Chaves ...

De hecho, la herramienta no devuelve, en ninguno de los tres casos, concordancias para el término *jueza*. Por el contrario, los otros dos diarios –*ABC* y *Cordópolis*– son más laxos a la hora de seguir las normas y se permiten no mantener la coherencia en cuanto al uso de ambas posibilidades. Ambos devuelven, por tanto,

resultados, tanto al introducir el término *juez*, como *jueza*, aunque mucho menos numerosos en el segundo caso.

Diario *ABC*:

Resultados para *juez*

Previo	Palabra	Posterior
... El TSJ valenciano refuerza al	juez	de la operación Taula ante la gravedad ...
... Un	juez	de Málaga se inhibe a favor del Supremo ...
... El	juez	rechaza dejar al PSPV fuera del caso Va ...
... El abogado que recomendó a un	juez	pedirse un traslado https://t.co/RCpN9n ...
... Un	juez	de Madrid pregunta al fiscal si debe in ...
... El	juez	sienta en el banquillo a veinte persona ...
... lamentable y grotesca» la actitud de la	juez	de #Podemos https://t.co/wXqp75rQWk htt ...
... RT@abc_familia: «El	juez	no permitió que mi #madre viera a mi #h ...
... La	juez	decreta prisión provisional para los se ...
... □ #ÚLTIMAHORA El	juez	no ve delito en la pitada al himno de l ...
... Rita #Maestre se confiesa ante el	juez	https://t.co/YxlY2Kv6pl https://t.co/iD ...
... La	juez	cita a declarar al responsable de infor ...
... #AMPLIACIÓN https://t.co/y2VtNHcaBe El	juez	absuelve a «los ocho de Airbus» por fal ...
... #ÚLTIMAHORA El	juez	absuelve a «los 8 de Airbus» por falta ...
... #ÚLTIMAHORA El	juez	rechaza la querella contra Carmena y Ma ...
... Un	juez	obliga a CaixaBank a devolver 3,4 millo ...
... #HaSidoNoticia El	juez	propone juzgar a Zapata por sus tuits o ...

Resultados para *jueza*

Previo	Palabra	Posterior
... Una	jueza	francesa autoriza la demolición parcial ...

Periódico *Cordópolis*:

Resultados para *juez*

Previo	Palabra	Posterior
... a no retira una cerca como le obliga un	juez	https://t.co/fcqJdzJooz https://t.co/52 ...
... Un	juez	cuestiona cómo se multa por la zona azu ...
... Un	juez	declara nulo el despido de una trabajad ...
... El	juez	rechaza que CTA se persone en la invest ...
... P, Miguel Reina y Navas, citados por el	juez	https://t.co/QXr6rHkg1K https://t.co/dU ...
... ncejales del PP, Reina y Navas, ante el	juez	por prevaricación https://t.co/QXr6rHkg ...
... ser readmitido en 2016 por orden de un	juez	https://t.co/UznbPERyDW https://t.co/3P ...
... Una	juez	llama a declarar a dos altos cargos de ...
... La Iglesia lleva al	juez	la titularidad municipal en la Fuensant ...
... El	juez	desestima el recurso del policía trafic ...
... ntalbán pasará mañana a disposición del	juez	https://t.co/a2CeoFFbDr https://t.co/kU ...
... Un	juez	investiga a Mercacórdoba por un supuest ...

Resultados para *jueza*

Previo	Palabra	Posterior
... Una	jueza	condena a Bankia por provocar daños psi ...
... Una	jueza	ordena investigar todo el patrimonio de ...
... La	jueza	divide en nueve piezas el fraude en el ...

Analizando estos mismos medios de comunicación observamos que, mientras que para algunos oficios está más extendido el uso de la forma femenina, como *abogada*, hay otros, como el caso de *música* (para referirse a la mujer que se dedica a la Música de manera profesional) que apenas se utilizan. En otros casos, como el de *médica*, por ejemplo, aparece el término en femenino, pero no para referirse a la mujer que desempeña la profesión de la Medicina, sino como adjetivo.

Nos advierte Gómez Torrego (2008) de que esta forma de proceder de la prensa, reticente a los cambios académicos, no debería sorprendernos cuando es la actitud habitual de los medios de comunicación ante los cambios normativos propuestos, ya que suelen tardar más tiempo del deseable en aceptarlos.

7.2.2.4 Androcentrismo y salto semántico

Para Guerrero (2007a), el androcentrismo en la lengua se produce como consecuencia del salto semántico resultante de emplear el género masculino supuestamente con valor genérico, para descubrir posteriormente el receptor que el término en cuestión estaba refiriéndose exclusivamente al sexo masculino. Esto, según la autora, provoca malentendidos y confusiones que deberían ser evitados. Tras el análisis de la información de nuestro corpus, hemos podido comprobar, no obstante, que en estos periódicos no ocurren estos saltos semánticos. Aun así, hemos encontrado un tuit publicado en *El País* que, aunque no se trata estrictamente de un salto semántico, sí puede dar lugar a confusión; lo mostramos a continuación:

28 feb 2016, 19:13 h.	ÚLTIMA HORA | Halladas muertas dos mujeres senderistas perdidas en Castellón https://t.co/PjH8YlDfMY Una tercera persona ha sobrevivido

Efectivamente, aunque el orden de los elementos ha sido a la inversa de como lo explica Guerrero, la segunda oración es ambigua a la hora de interpretar su significado. Sabemos por el resto de noticias que la tercera persona a la que se alude es un hombre, sin embargo, esto no queda claro debido a la utilización del término *persona* para referirse al sexo masculino y de la palabra *mujer*, para el sexo femenino.

También en este tuit vemos una redundancia en la aposición, ya que no es necesario utilizar el término *mujeres* para indicar el género del sustantivo *senderista* porque existen más elementos en la oración que concuerdan con el femenino y desambiguan el género. Veremos este fenómeno con más profundidad un poco más adelante.

7.2.2.5 Duales aparentes y vocablos ocupados.

Ayala, Guerrero y Medina (2002: 59) definen duales aparentes como: "términos que adquieren significados diferentes según el sexo al que se refieran, como ocurre con *señorito*/*señorita*, *hombre público*/*mujer pública*, *individuo*/*individua*, *verdulero*/*verdulera*, *prójimo*/*prójima*, etc. Casi siempre resultan peyorativos para la mujer." La existencia de duales aparentes provoca que haya términos en la lengua con género femenino que ya están ocupados con esos significados de connotaciones

negativas, lo que dificulta un uso igualitario entre ambos géneros. Sin embargo, estos términos, que deben sus distintas connotaciones semánticas a la tradición y a la sociedad machista de la época, son, quizá por este mismo motivo, generalmente evitados en la lengua de los medios de comunicación. Además, debido precisamente a los cambios sociales, cada vez se hace un uso más simétrico de las formas masculina y femenina, de manera que los duales aparentes están perdiendo fuerza y se les está ganando terreno a los vocablos ocupados.

En el pasado ocurría, por ejemplo, que el término femenino adoptaba el significado de "esposa de", como era el caso de *alcaldesa*; o se refería a cargos de menor importancia que el masculino, como *secretaria*. También, por supuesto, se daba el caso en el que el término femenino tenía connotaciones peyorativas, mientras que el masculino era neutro o incluso positivo; por ejemplo, en *zorra*. En los siguientes ejemplos, se observa claramente cómo se van dejando atrás estas desigualdades:

3 mar 2016, 13:33 h.	ÚLTIMA HORA: La alcaldesa de Madrid reclama el apoyo a la investidura de Pedro Sánchez https://t.co/aPyQctvQDq
29 feb 2016, 8:54 h.	Siete alcaldesas para 2,4 millones de habitantes. Mujeres al poder en la gran Barcelona, por @clarablanchar https://t.co/1ArIzmOHem
25 feb 2016, 15:35 h.	La exalcaldesa de Valencia y senadora, Rita Barberá, ha negado esta mañana todas las acusaciones que le implican https://t.co/Rdic2LXVro
3 mar 2016, 20:23 h.	La delegada de Educación en Córdoba presenta una campaña de 'excolarización' sin percatarse de una errata... https://t.co/Q9b1Dj6iLg
18 feb 2016, 13:37 h.	#ÚLTIMAHORA Ruiz-Gallardón ofreció a Urdangarin dirigir Madrid 2016, según la exconsejera delegada del proyecto https://t.co/5BuzcbLFyq
5 mar 2016, 00:26 h.	Rousseff considera la detención de Lula "innecesaria" https://t.co/mrDHIEzhNo La presidenta se defiende a sí misma y a su predecesor
14 feb 2016, 14:38 h.	Finaliza la comparecencia de Esperanza Aguirre donde ha anunciado su dimisión como presidenta del PP de Madrid. https://t.co/lQMaLFXkXI

7.2.2.6 Disimetría en el tratamiento de los sexos

La disimetría en el tratamiento de los sexos es un fenómeno que puede tener su origen en distintos focos, según Guerrero (2007a), como, por ejemplo, al nombrar al hombre por el apellido y a la mujer por su nombre de pila; al presentar a la mujer como

dependiente del hombre, al emplear aposiciones redundantes o al emplear un tratamiento heterogéneo de los sexos.

Debemos admitir que, en este estudio, no hemos encontrado demasiados casos de disimetría a la hora de referirse a los hombres y a las mujeres. Creemos, por ello, que los medios de comunicación son cada vez más conscientes de este aspecto sexista y este es el motivo por el que los ejemplos sobre algunos de los tratamientos desiguales escasean.

No obstante, sería inadecuado asegurar que la disimetría ha desaparecido por completo. Donde más abunda es, sin duda, en el uso del masculino genérico que, además, parece bastante difícil de erradicar teniendo en cuenta la idiosincrasia de nuestro idioma y la falta de economía que supondría duplicar, por ejemplo, los adjetivos o los sustantivos cuando acogen a los dos sexos:

6 mar 2016, 10:35 h.	Doce emprendedores culminan sus planes de empresas en el Imdeec https://t.co/TrvOd5H76o https://t.co/neviYrfDu1
3 mar 2016, 20:35 h.	Un centenar de alumnos protesta por la falta de docentes en Económicas https://t.co/h37Hwu1d2P https://t.co/NGMOQeRnXV
5 mar 2016, 19:55 h.	Ambrosio y Pernichi visitan las obras reclamadas por los vecinos de la avenida del Corregidor #CórdobaEsp https://t.co/TY17hcxaTe
6 mar 2016, 11:14 h.	¿Reconoces a estos famosos españoles de pequeños? #Hemeroteca https://t.co/nAjjvGK3Wn
5 mar 2016, 16:49 h.	Malas noticias para los seguidores de Bertín Osborne y su programa 'En tu casa o en la mía' https://t.co/UHd3nP9QVP

Afirma Guerrero (2007) que otro de los signos sintomáticos del carácter sexista del lenguaje consiste en la diferencia de tratamiento dirigido a los hombres y a las mujeres. Según la autora, es frecuente nombrar a los hombres por su apellido o apellidos y a las mujeres, por el contrario, por su nombre de pila. Además, considera que también se produce disimetría cuando los medios se refieren al sexo masculino por su condición profesional o social, mientras que a la mujer, simplemente por su condición sexuada. En nuestro estudio, sin embargo, hemos encontrado escasos ejemplos que demuestren esta desigualdad, por ejemplo, en lo que a cargos públicos se refiere. En la mayoría de las ocasiones, los medios de comunicación se refieren a ellos indistintamente con el apellido o con el nombre de pila más el apellido, sean hombre o mujeres. En la tabla siguiente, mostramos tres ejemplos de políticas a las que se les nombra por el apellido y otros tres referidos a las mismas mujeres, en los que aparece

también su nombre de pila. A continuación, los mismos ejemplos para casos de políticos:

28 feb 2016, 16:00 h.	Díaz hace bandera del 28-F mientras asfixia a la Fundación Blas Infante https://t.co/wVEQm0HqyH #DiadeAndalucia https://t.co/SRyZhUZg93
17 feb 2016, 1:28 h.	Los trabajadores de Telemadrid denuncian el "despilfarro" cometido en la cadena en la época de Aguirre https://t.co/ORyb28Jsys
24 feb 2016, 14:59 h.	Cifuentes quiere traer el #MWC a Madrid. https://t.co/UBjY74rYBq
1 dic 2016, 10:45 h.	Susana Díaz apela en Montilla a la movilización el próximo 20D https://t.co/hTRQilXfAE https://t.co/gudonSJvqV
14 feb 2016, 15:48 h.	Esperanza Aguirre dimite como presidenta del PP de Madrid https://t.co/5j3TF5YCOb
15 feb 2016, 20:21 h.	#ÚltimaHora Cristina Cifuentes presidirá la comisión gestora del PP madrileño https://t.co/8NhMvIscZ5 https://t.co/zHT8Y8zkr3
23 feb 2016, 16:49 h.	Sánchez acepta la propuesta de reforma de la Constitución de Ciudadanos https://t.co/xWXHXmbzb0
22 feb 2016, 12:09 h.	Rivera no descarta ahora entrar en un gobierno del PSOE https://t.co/lEsDtiSWsw
4 mar 2016, 6:30 h.	Trump ha logrado sacar de sus casillas al resto de aspirantes https://t.co/ovc4i6aeIa Y ninguno consigue parar al magnate
7 feb 2016, 10:33 h.	Pedro Sánchez aventura que hay "mimbres" para el pacto del cambio https://t.co/1uUAQuid5p
2 mar 2016, 11:35 h.	#ELPAÍSinvestidura Albert Rivera espera su turno en su escaño https://t.co/qvuyWZiRlZ https://t.co/RIhIXKurIN
5 mar 2016, 20:28 h.	Morderse la lengua y esperar a que el fuego se apague. La postura oficial del Gobierno mexicano ante Donald Trump https://t.co/pxoWKANOxU

En cuanto al tratamiento de *señor* o *señora*, todos los ejemplos analizados que mantienen estas fórmulas de tratamiento delante del apellido están incluidos dentro de citas textuales de políticos, que son los que más las utilizan. Además, ya hemos visto que el lenguaje político suele ser más inclusivo, lo que significa no hacen, por lo general, distinciones, entre hombres y mujeres. El número más elevado de ocurrencias de *señor* se debe, simplemente, al mayor número de hombres que ostentan cargos políticos:

12 feb 2016, 14:15 h.	'A usted no hay quien la regañe ni quien la eche señora Aguirre'. #EnDirecto https://t.co/f7wN6kIzKR https://t.co/5aFFz9FqHN
19 feb 2016, 19:10 h.	Mónica Oltra: "El gobierno del señor Sánchez será progresista o no será" https://t.co/JM2O5hOVZT
4 mar 2016, 20:32 h.	«¿Queremos evitar que el señor Rajoy siga al frente del Gobierno, sí o

	no?», y otras diez frases de Pedro Sánchez https://t.co/5M5VgQDlV8
2 mar 2016, 10:09 h.	#DIRECTO Rajoy: «No creo que haya que cambiar de política, señor Sánchez. La política económica ha funcionado» https://t.co/MZ7ArwJklJ

También advierte Guerrero (2007a) del peligro de presentar a la mujer con un papel dependiente del hombre, cuando se habla de *mujer de* o de *esposa de*. Recomienda evitar estas expresiones que, desde su punto de vista, hacen que la mujer pierda su propia identidad. En nuestro corpus hemos encontrado, efectivamente, algunos casos con estas locuciones a los que, sin embargo, no les atribuimos rasgos sexistas porque consideramos que no tendrían relevancia mediática por sí mismas:

25 feb 2016, 3:33 h.	La esposa de Trump es inmigrante y apoya hasta en el mensaje xenófobo de su marido [Salvo cuando dice palabrotas] https://t.co/4SgjLwTj2h
26 feb 2016, 8:40 h,	Marta Ferrusola, esposa del expresidente catalán Jordi Pujol, eludió en Panamá el pago de impuestos andorranos https://t.co/IhcMLpQ6yu
23 feb 2016, 00:38 h.	La esposa de Bill Cosby, obligada a testificar por los supuestos abusos https://t.co/LKqjDhd6e3 https://t.co/VfnLLSGGLy
29 feb 2016, 22:52 h.	La mujer del Chapo Guzmán asegura que en la prisión no dejan descansar a su marido para así matarlo. https://t.co/OHu0dDSf0N
27 feb 2016, 5:15 h.	Detienen a la exnovia de Evo Morales por presunto tráfico de influencias https://t.co/M8Wxq28Dnk
21 feb 2016, 22:25 h.	La novia de Albert Rivera y la mujer de Pedro Sánchez disfrutan de la moda española https://t.co/oyX6UBNT8w https://t.co/3NFK9f9Efb

Por ejemplo, los casos de "la novia de Albert Rivera", "la mujer de Pedro Sánchez", "la mujer de Chapo Guzmán" o "la esposa de Trump" no serían noticia si sus parejas fueran anónimas. Quizá el caso de "Marta Ferrusola" o "la esposa de Bill Cosby" se acercan más a la disimetría de la que habla Guerrero aunque, en todo caso, vemos que es poco frecuente.

En último lugar, otro de los motivos que suelen ser causa de desigualdades en el tratamiento de los sexos en el lenguaje es lo que Guerrero (2007a) denomina aposiciones redundantes. En ellas, el sustantivo *mujer* viene acompañado de términos que aluden a su profesión o a otras características, ocultado así su identidad profesional y destacando su condición sexuada. Estos ejemplos, al contrario de los anteriores, sí aparecen con frecuencia en el corpus de los cinco periódicos analizados.

28 feb 2016, 19:13 h.	ÚLTIMA HORA | Halladas muertas dos mujeres senderistas perdidas en Castellón https://t.co/PjH8YlDfMY Una tercera persona ha sobrevivido
26 feb 2016, 19:14 h.	Las mujeres marroquíes que quedan embarazadas fuera del matrimonio son acusadas de prostitución https://t.co/Q9WGaOThU1 En @Planeta_Futuro
4 mar 2016, 10:04 h.	La mujer española ya no quiere ser perfecta https://t.co/7Riz6lQal9 #maternidad #trabajo https://t.co/jvCPC6LZ2R
23 feb 2016, 10:58 h.	Las mujeres olvidadas de la Generación del 27: Ellas, el género neutro https://t.co/h73Hh5yzmE #LasSinSombrero https://t.co/e7XTtId8g8
24 feb 2016, 12:08 h.	La Junta lanza campaña para "visibilizar" a las mujeres andaluzas en la Historia https://t.co/E44QaQ6w5m
29 nov 2015, 16:05 h.	Un proyecto para empoderar a mujeres costureras gana un premio del Urban Sur https://t.co/CJOs378NfB https://t.co/NJxH3XoUJJ

Ocurre, como se observa en los tuits, que el término *mujer* es redundante en todos los casos puesto que se sobreentiende gracias al género gramatical del resto de palabras que lo acompañan.

7.2.3 Conclusiones

Aunque autores como García Meseguer (2001) defienden que el español –a diferencia del inglés– no es una lengua sexista[47], la literatura científica ha venido defendiendo que el uso sexista del lenguaje es intrínseco a nuestro idioma desde muy antiguo. El motivo es fundamentalmente que la lengua es un elemento cultural que refleja la sociedad y la forma de pensar de sus hablantes. El hecho de que la sociedad esté evolucionando y las mujeres estén encontrando poco a poco la tan deseada igualdad está provocando que el idioma cambie con ella para seguir siendo el espejo de sus hablantes. No obstante, erradicar por completo una forma tan arraigada como el masculino genérico en el uso del lenguaje no es nada fácil.

Aun así, después de analizar el corpus de los medios de comunicación que hemos elaborado para este estudio, podemos afirmar que muchos de los rasgos sexistas del lenguaje que enumera la profesora Guerrero (2007a) son cada vez menos frecuentes en nuestro idioma. Por lo tanto, a pesar de que aspectos como la utilización del

[47] Para García Meseguer (2001: 20) existen tres agentes responsables del sexismo lingüístico: a) el hablante y su contexto mental, b) el oyente y su contexto mental y c) la lengua como sistema, y, en su opinión, en español solo actúan los dos primeros, mientras que en inglés actúan los tres: "el español, como sistema lingüístico, no es una lengua sexista, a diferencia de otras, como el inglés, cuyo sistema lingüístico sí presenta elementos sexistas".

masculino genérico, de aposiciones redundantes o la poca aceptación social del género gramatical femenino de algunos nombres de oficios siguen estando presentes en el español, los resultados demuestran que la concienciación acerca este aspecto por parte de los medios de comunicación es cada vez mayor, algo fundamental teniendo en cuenta el carácter mediático y la influencia que pueden ejercer sobre los lectores.

8

Muestras de uso de *Wordics Archive*

Wordics Archive es el tercer y último módulo de la herramienta que hemos diseñado para estudiar el uso del lenguaje a través de *big data*. Como ya hemos explicado en el capítulo dedicado a la metodología, es el más completo de los tres módulos de la herramienta, no solo porque permite realizar análisis lingüísticos útiles (como los demás módulos), sino porque ofrece la posibilidad de delimitar la búsqueda a través de muy diversos parámetros. Además, su diseño, enfocado hacia un análisis histórico de la información, sortea las limitaciones impuestas por la API de *Twitter* en cuanto a la accesibilidad de la información. Este inconveniente se ve superado gracias a la obtención y almacenamiento de información continuos que la herramienta realiza sin descanso, con la ayuda de un servidor que la custodia

De la misma manera que hemos tratado de ejemplificar algunos de los posibles usos que se pueden hacer de la herramienta, en este capítulo nos proponemos presentar tres muestras que den cuenta de cómo y para qué puede utilizarse *Wordics Archive* en la investigación lingüística y cuáles son las ventajas que lo sitúan a la vanguardia de los trabajos en este campo y que lo distinguen de las herramientas disponibles actualmente.

Los estudios son, como decimos, tres, y sus objetivos se centran en la utilización de *big data* para:

a) Estudiar la evolución del bilingüismo en diez países latinoamericanos.
b) Determinar las necesidades traductológicas en cinco capitales europeas.
c) Comprobar la utilidad de *Twitter* para el estudio de neologismos en español.

8.1 Evolución del bilingüismo en diez países latinoamericanos

En este primer estudio, pretendemos analizar la progresión de la lengua inglesa y de la española en diez países latinoamericanos para determinar el uso de ambas y tratar de predecir su evolución en los próximos años. Los datos recogidos abarcan desde el

año 2010 hasta 2014 y los resultados muestran las distintas realidades que nos encontramos en estos países y cuál es su situación lingüística.

8.1.1 Metodología

La investigación que presentamos es una adaptación de un artículo publicado en el año 2015 por la autora de este trabajo (González Fernández, 2015), realizado con una primera versión de *Wordics*.

Para el estudio se seleccionaron diez países de Latinoamérica con distintas condiciones socioculturales, políticas y geográficas, pero con una característica en común: el español como lengua oficial.

El período de tiempo establecido para el análisis de la información queda comprendido entre el 1 de enero de 2010 y el 31 de diciembre de 2014. Conviene recordar, no obstante, que dadas las facilidades de la herramienta, la recogida de la información perteneciente a estos cinco años se ha realizado automáticamente transcurridos unos meses desde la finalización de este período. Los países seleccionados son: Argentina, Chile, Colombia, República Dominicana, Ecuador, México, Panamá, Paraguay, Perú y Venezuela. En cada uno de ellos, se ha estudiado la presencia del inglés y se ha comparado con el uso del español.

La búsqueda incluye el número total de tuits publicados durante esos cinco años, con una granularidad diaria. De esta forma, ha sido posible observar la evolución de la implementación de *Twitter* en cada país.

Tras esta primera aproximación, necesaria para asegurarnos de que el tamaño de muestra sería adecuado, hemos llevado a cabo dos búsquedas más, similares a la anterior, pero filtradas por idioma. Cada una de estas dos búsquedas se corresponden con el total de tuits escritos en cada idioma y en cada país durante el período de tiempo establecido.

Los parámetros establecidos en *Wordics* para la recogida de información son, por tanto:

a) Intervalo temporal de cinco años: desde el 1 de enero de 2010 hasta el 31 de diciembre de 2014.

b) Granularidad[48] de un día.

[48] Aspecto que define el nivel de detalle de representación de la información.

c) Búsqueda de los tuits totales publicados.

d) Búsqueda de los tuits escritos solo en inglés.

e) Búsqueda de los tuits escritos solo en español.

A la hora de elaborar las gráficas, debemos tener en cuenta que se han establecido tres líneas distintas de representación de datos (tuits) y –recordemos– con granularidad diaria. La primera de ellas, de color azul, representa al idioma español; la segunda, anaranjada, al inglés; y la tercera, de color verde, la suma de todos los tuits escritos. Esta última línea no solo incluye los tuits publicados en inglés y en español, sino también los que hayan sido producidos en idiomas minoritarios o irreconocibles por el sistema debido a la presencia de enlaces, símbolos, emoticonos, vídeos, fotos e incluso tuits sin texto. Puesto que la granularidad es de un día, el movimiento de la línea de tuits resulta muy irregular, así que, para una visualización más sencilla y cómoda, se ha superpuesto una línea de tendencia en cada una de las tres líneas de datos, del mismo color que estas. Para la línea del total de tuits, la línea de tendencia es una media móvil de 100 puntos, mientras que para los otros dos casos, se ha utilizado una línea de tendencia polinomial cuadrática o cúbica, dependiendo de la conveniencia según la información analizada.

8.1.2 Resultados

Como ya hemos mencionado, el período de tiempo analizado ha sido de cinco años y hemos obtenido un valor de frecuencia diario, lo que nos ha dado como resultado un número total de entradas analizadas de 1826 días (365*5 con 1 año bisiesto). Cada una de estas entradas contiene información acerca de los tuits escritos en ambos idiomas.

A continuación, presentamos un mapa coroplético[49], en el que el dato correspondiente al número total de tuits analizados en cada país está representado con una intensidad distinta de la gama de color naranja. Cuanto mayor sea la cantidad de tuits en una zona, más oscura estará esta representada.

[49] Tipo de mapa temático que representa información cuantitativa mediante gamas de colores.

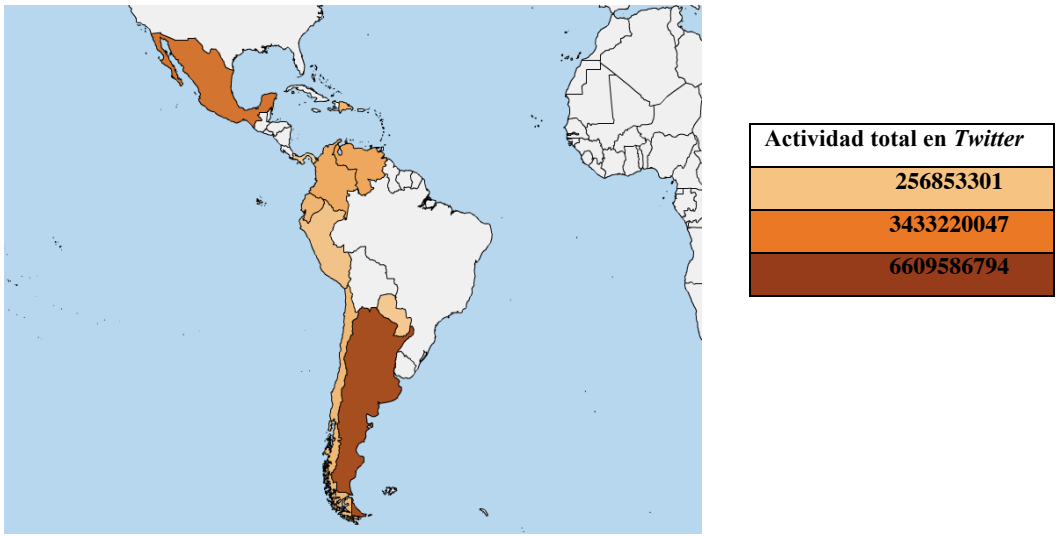


Figura 8.1. Actividad total en *Twitter* en los países analizados

Con los resultados obtenidos, hemos elaborado un informe detallado para cada país, en el que, en primer lugar, aportamos información general acerca de aspectos demográficos y económicos, así como datos relativos al uso de Internet y de *Twitter* en el país en cuestión (esta información ha sido obtenida de fuentes externas a la herramienta). En segundo lugar presentamos un gráfico que representa la evolución de los tuits con las líneas de tendencia de las que ya hemos hablado. Por último, el informe de cada país concluye con un gráfico en forma de tarta con la distribución media de los idiomas utilizados en diciembre de 2014, el ultimo mes de la investigación.

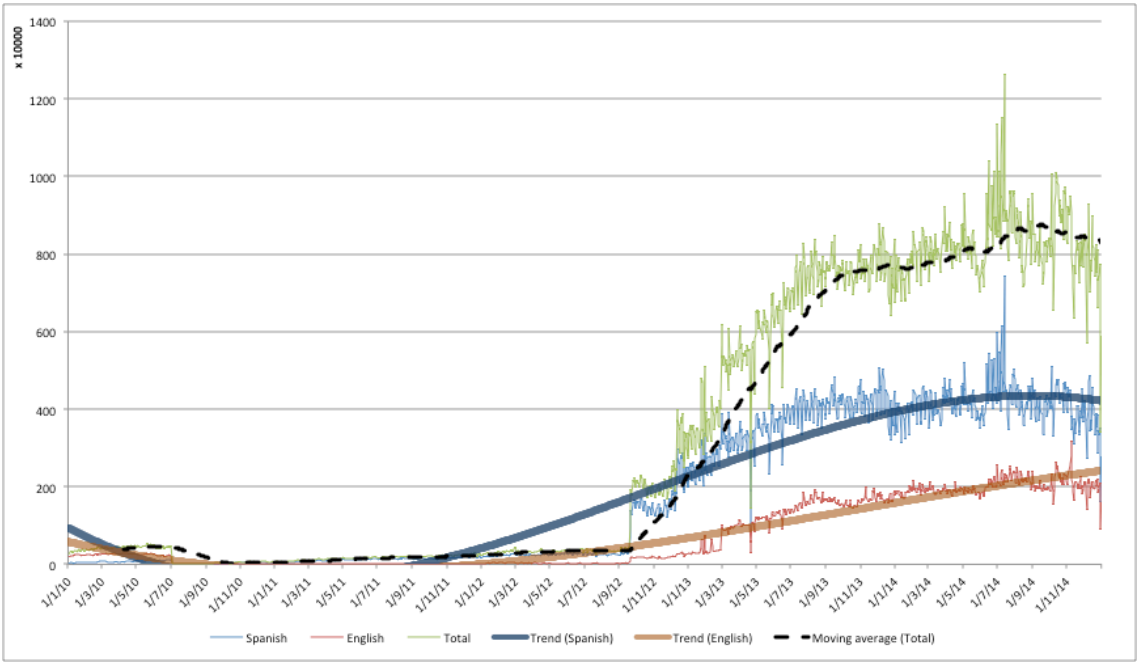
Argentina

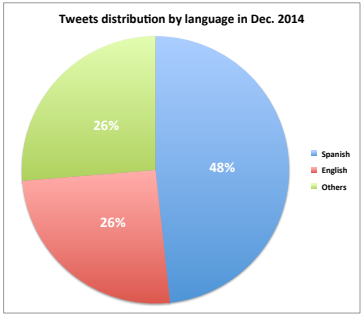
Información general

Período	01/01/2010 a 12/31/2014
Número de tuits	375.042.140
Número de tuits y RT	5.807.966.578
Media de edad *	31,2
Población *	41.803.125
Usuarios de Internet *	24.973.660
% mundial de usuarios de Internet *	0.86%

**información correspondiente al año 2014. Información correspondiente a la población, el número de usuarios de Internet y el porcentaje mundial de cada país de usuarios de Internet obtenida de Internet Live Stats. Información correspondiente a la media de edad obtenida de la CIA*

Frecuencia de tuits por idioma





	Absoluto	Relativo
Español	3.217.848	48%
Inglés	1.680.483	26%
Otros	1.357.272	26%
Total	6.255.603	100%

Figura 8.2. Informe general de los datos obtenidos de Argentina

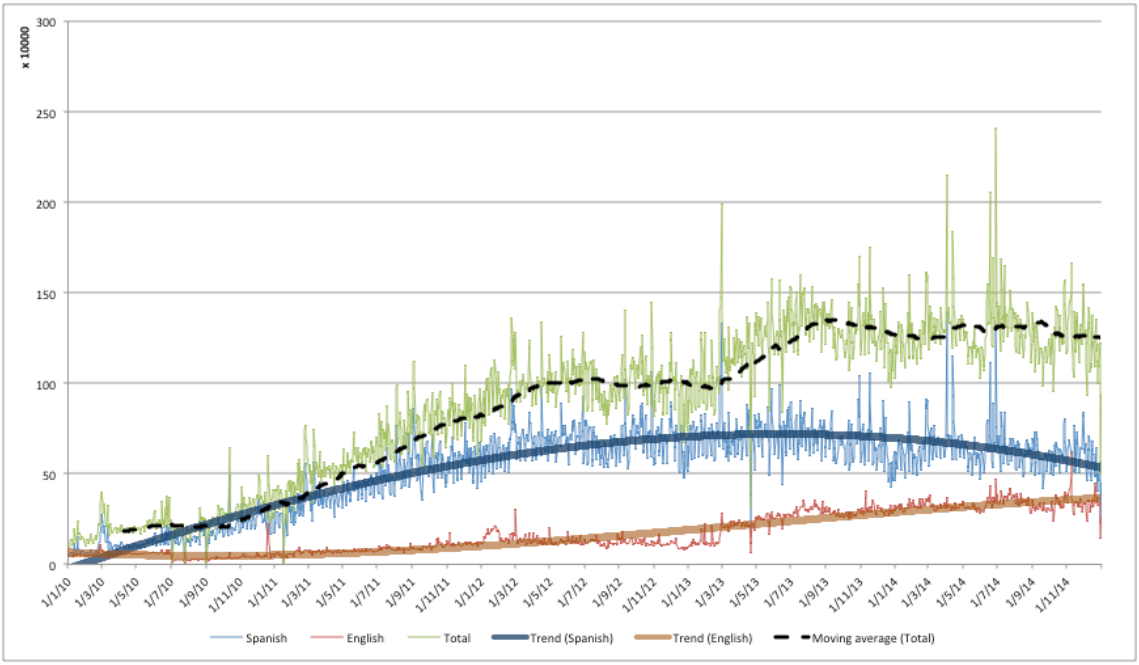
Chile

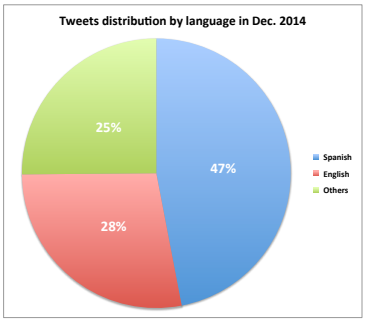
Información general

Período	01/01/2010 a 12/31/2014
Número de tuits	611.711.173
Número de tuits y RT	1.693.996.206
Media de edad *	33,3
Población *	17.772.871
Usuarios de Internet *	11.686.746
% mundial de usuarios de Internet *	0,40%

**información correspondiente al año 2014. Información correspondiente a la población, el número de usuarios de Internet y el porcentaje mundial de cada país de usuarios de Internet obtenida de Internet Live Stats. Información correspondiente a la media de edad obtenida de la CIA.*

Frecuencia de tuits por idioma





	Absoluto	Relativo
Español	17.109.275	47%
Inglés	10.155.046	28%
Otros	9.152.545	25%
Total	36.146.866	100%

Figura 8.3. Informe general de los datos obtenidos de Chile

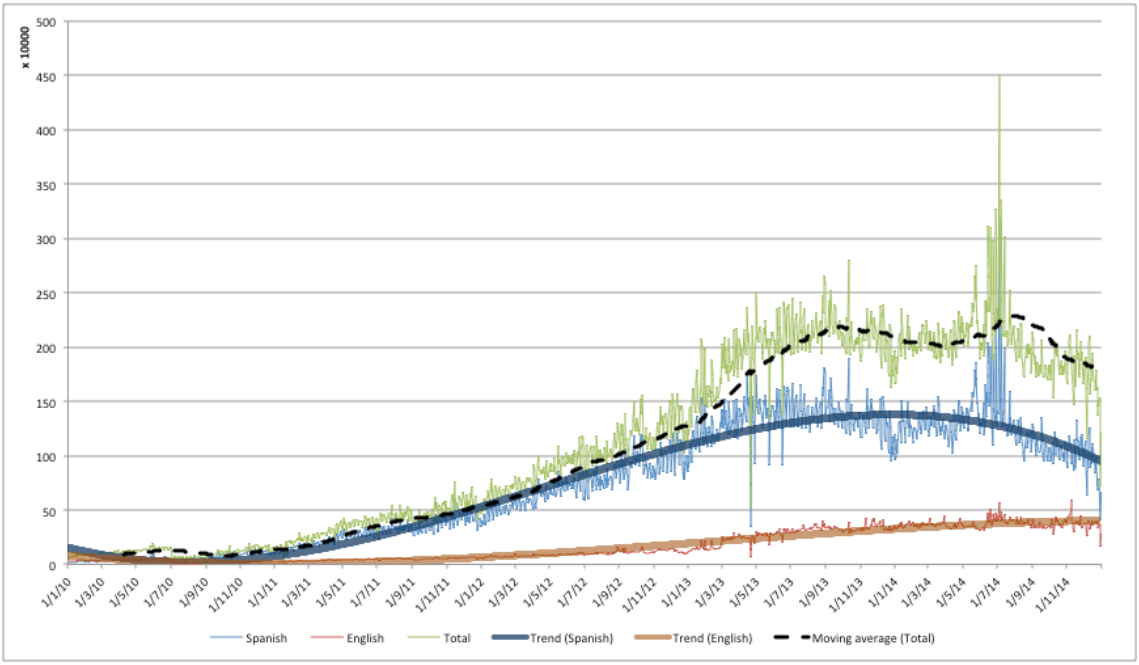
Colombia

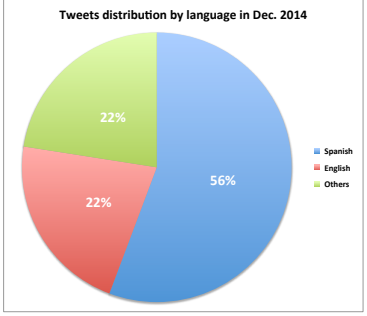
Información general

Período	01/01/2010 a 12/31/2014
Número de tuits	96.456.159
Número de tuits y RT	2.009.918.380
Media de edad *	28,9
Población *	48.660.725
Usuarios de Internet *	25.583.953
% mundial de usuarios de Internet *	0,88%

**información correspondiente al año 2014. Información correspondiente a la población, el número de usuarios de Internet y el porcentaje mundial de cada país de usuarios de Internet obtenida de Internet Live Stats. Información correspondiente a la media de edad obtenida de la CIA.*

Frecuencia de tuits por idioma





	Absoluto	Relativo
Español	28.478.980	56%
Inglés	11.116.060	22%
Otros	11.512.520	22%
Total	51.107.560	100%

Figura 8.4. Informe general de los datos obtenidos de Colombia

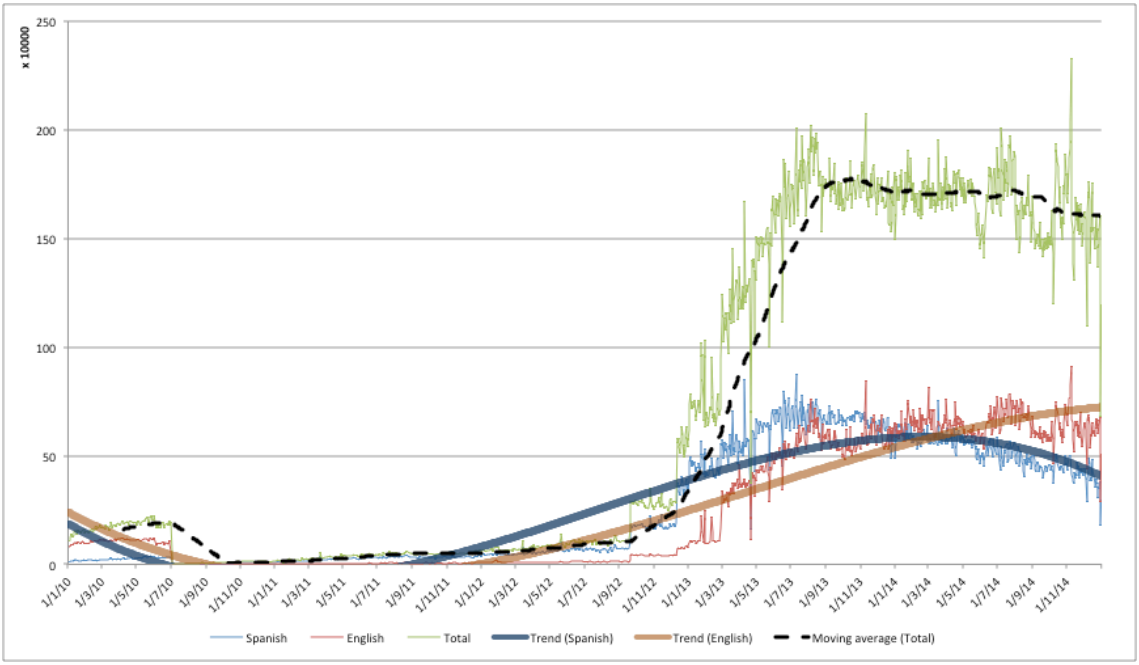
República Dominicana

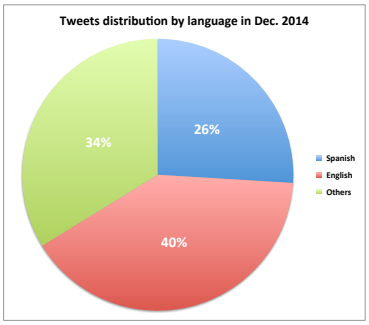
Información general

Período	01/01/2010 a 12/31/2014
Número de tuits	79.004.775
Número de tuits y RT	1.247.816.952
Media de edad *	27,1
Población *	10.528.954
Usuarios de Internet *	5.072.674
% mundial de usuarios de Internet *	0,17%

**información correspondiente al año 2014. Información correspondiente a la población, el número de usuarios de Internet y el porcentaje mundial de cada país de usuarios de Internet obtenida de Internet Live Stats. Información correspondiente a la media de edad obtenida de la CIA.*

Frecuencia de tuits por idioma





	Absoluto	Relativo
Español	12.169.926	26%
Inglés	18.884.736	40%
Otros	15.813.234	34%
Total	46.867.896	100%

Figura 8.5. Informe general de los datos obtenidos de República Dominicana

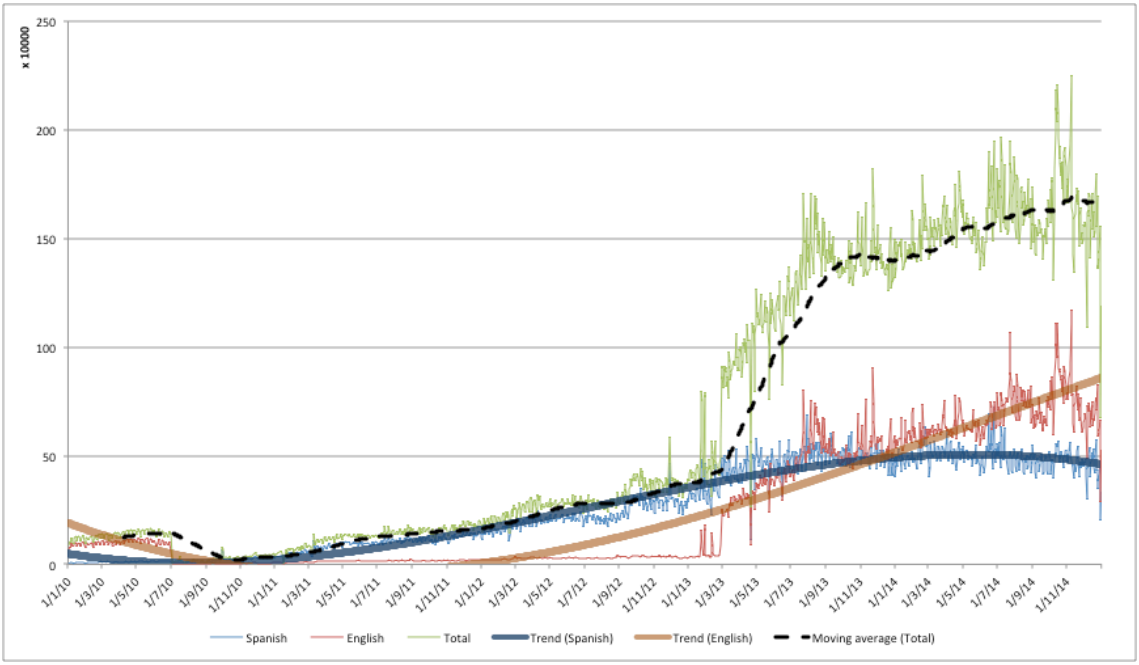
Ecuador

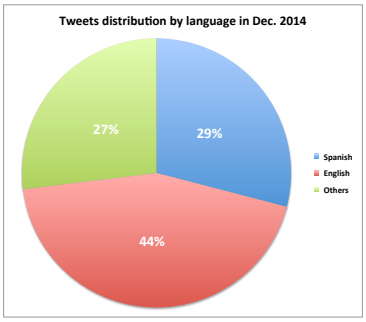
Información general

Período	01/01/2010 a 12/31/2014
Número de tuits	68.881.758
Número de tuits y RT	1.179.867.690
Media de edad *	26,7
Población *	17.772.871
Usuarios de Internet *	6.012.003
% mundial de usuarios de Internet *	0,21%

**información correspondiente al año 2014. Información correspondiente a la población, el número de usuarios de Internet y el porcentaje mundial de cada país de usuarios de Internet obtenida de Internet Live Stats. Información correspondiente a la media de edad obtenida de la CIA.*

Frecuencia de tuits por idioma





	Absoluta	Relativa
Español	13.614.915	29%
Inglés	20.624.925	44%
Otros	12.626.580	27%
Total	46.865.070	100%

Figura 8.6. Informe general de los datos obtenidos de Ecuador

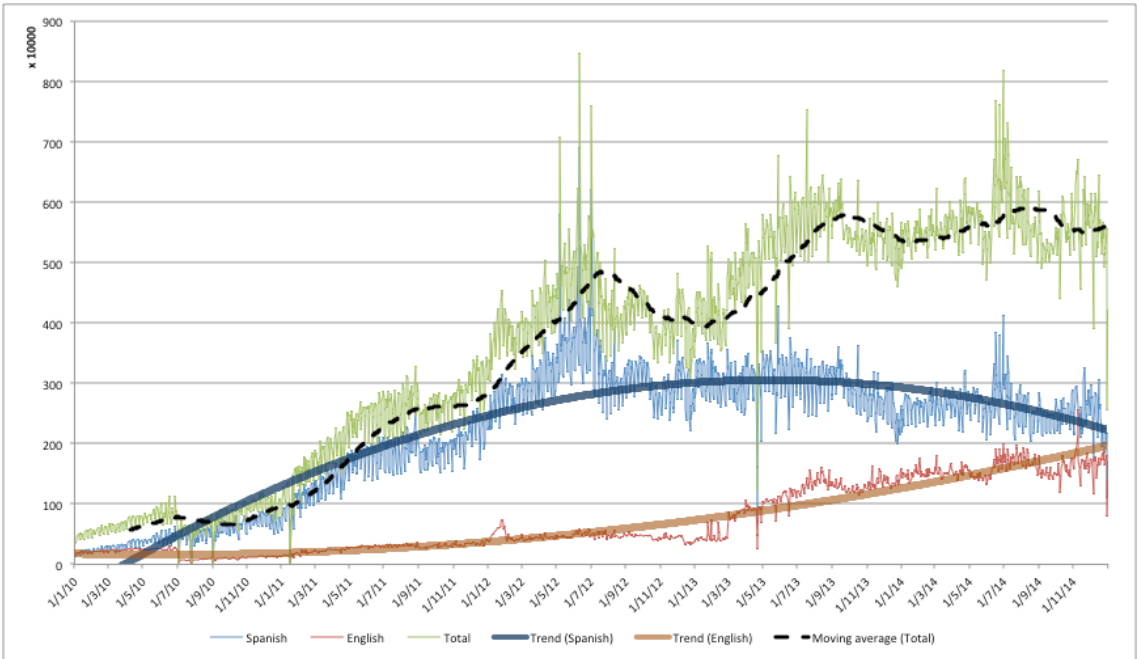
México

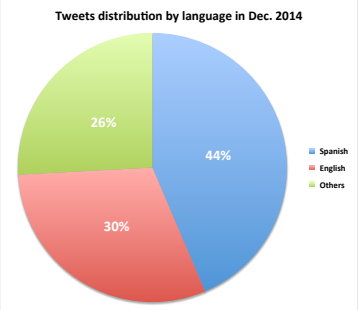
Información general

Período	01/01/2010 a 12/31/2014
Número de tuits	26.354.620
Número de tuits y RT	6.609.586.794
Media de edad *	27,3
Población *	123.799.215
Usuarios de Internet *	50.923.060
% mundial de usuarios de Internet *	1,74%

**información correspondiente al año 2014. Información correspondiente a la población, el número de usuarios de Internet y el porcentaje mundial de cada país de usuarios de Internet obtenida de Internet Live Stats. Información correspondiente a la media de edad obtenida de la CIA.*

Frecuencia de tuits por idioma





	Absoluto	Relativo
Español	73.067.907	44%
Inglés	51.211.667	30%
Otros	43.300.774	26%
Total	167.579.748	100%

Figura 8.7. Informe general de los datos obtenidos de México

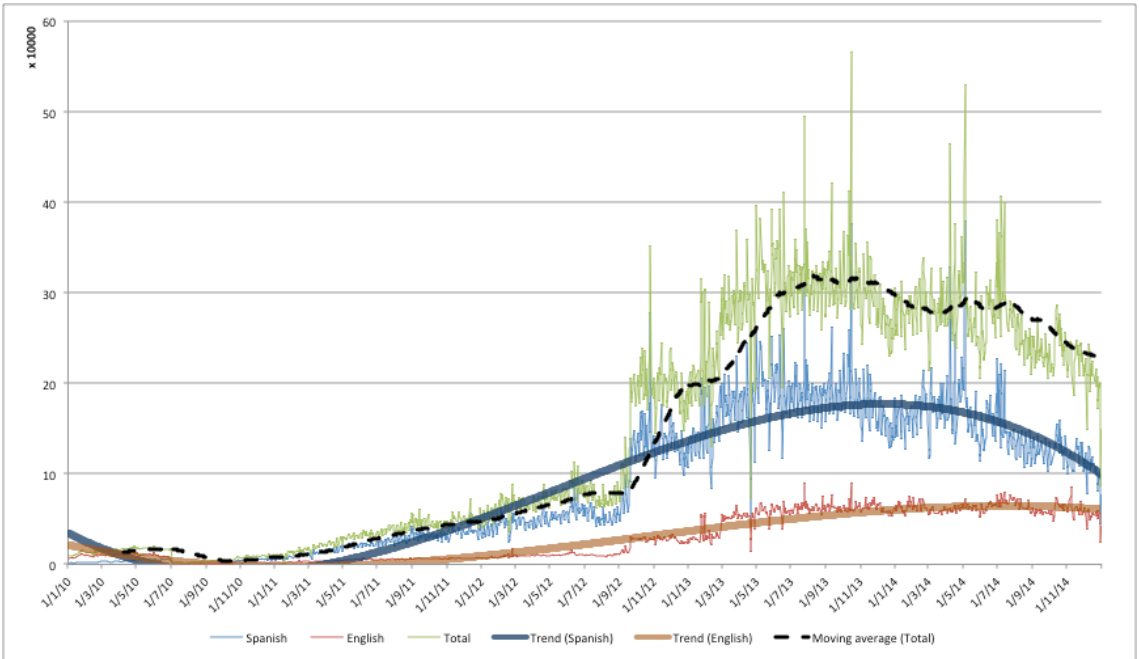
Panamá

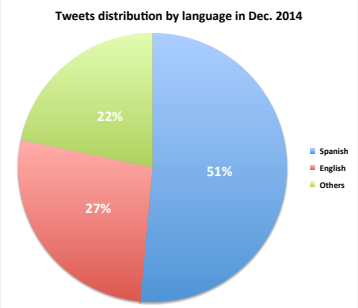
Información general

Período	01/01/2010 a 12/31/2014
Número de tuits	12.928.175
Número de tuits y Rt	256.853.301
Media de edad *	28,3
Población *	3.926.017
Usuarios de Internet *	1.899.892
% mundial de usuarios de Internet *	0,07%

**información correspondiente al año 2014. Información correspondiente a la población, el número de usuarios de Internet y el porcentaje mundial de cada país de usuarios de Internet obtenida de Internet Live Stats. Información correspondiente a la media de edad obtenida de la CIA.*

Frecuencia de tuits por idioma





	Absoluto	Relativo
Español	3.217.848	51%
Inglés	1.680.483	27%
Otros	1.357.272	22%
Total	6.255.603	100%

Figura 8.8. Informe general de los datos obtenidos de Panamá

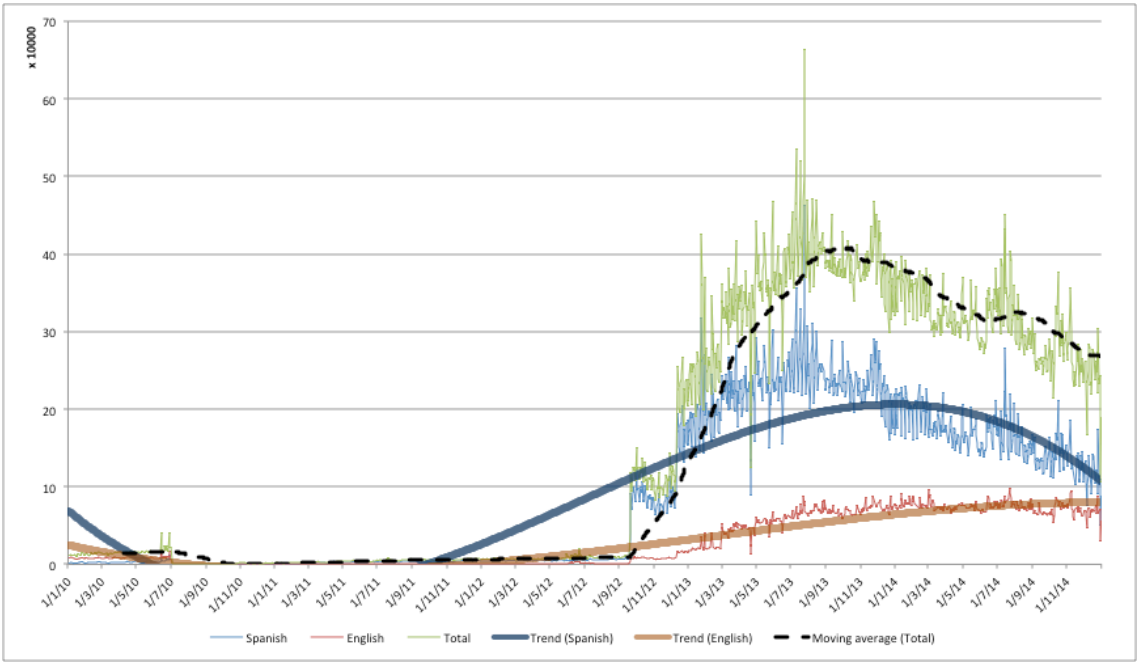
Paraguay

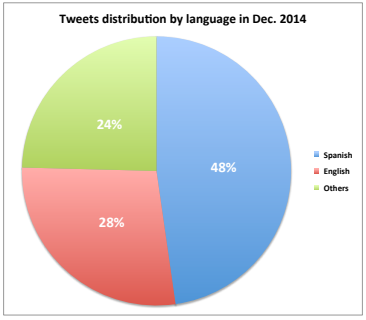
Información general

Período	01/01/2010 a 12/31/2014
Número de tuits	18.405.504
Número de tuits y RT	263.354.928
Media de edad *	26,8
Población *	6.917.579
Usuarios de Internet *	2.005.278
% mundial de usuarios de Internet *	0,07%

**información correspondiente al año 2014. Información correspondiente a la población, el número de usuarios de Internet y el porcentaje mundial de cada país de usuarios de Internet obtenida de Internet Live Stats. Información correspondiente a la media de edad obtenida de la CIA.*

Frecuencia de tuits por idioma





	Absoluto	Relativo
Español	3.632.844	48%
Inglés	2.105.528	28%
Otros	1.871.328	24%
Total	7.609.700	100%

Figura 8.9. Informe general de los datos obtenidos de Paraguay

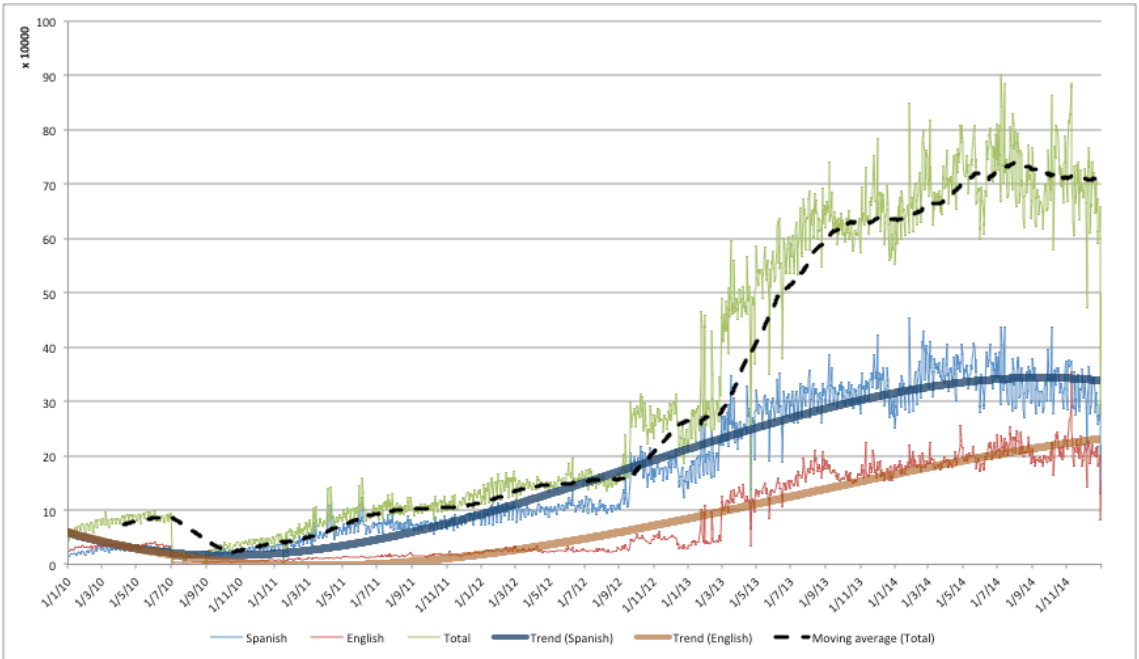
Perú

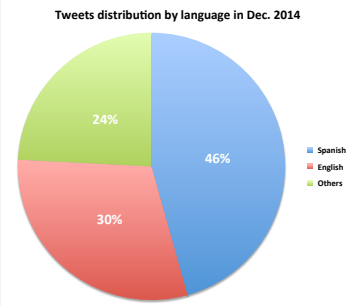
Información general

Período	01/01/2010 a 12/31/2014
Número de tuits	28.689.750
Número de tuits y RT	574.687.874
Media de edad *	27
Población *	30.769.077
Usuarios de Internet *	12.583.953
% mundial de usuarios de Internet *	0,43%

**información correspondiente al año 2014. Información correspondiente a la población, el número de usuarios de Internet y el porcentaje mundial de cada país de usuarios de Internet obtenida de Internet Live Stats. Información correspondiente a la media de edad obtenida de la CIA.*

Frecuencia de tuits por idioma





	Absoluto	Relativo
Español	9.284.409	46%
Inglés	6.132.492	30%
Otros	4.920.714	24%
Total	20.337.615	100%

Figura 8.10. Informe general de los datos obtenidos de Perú

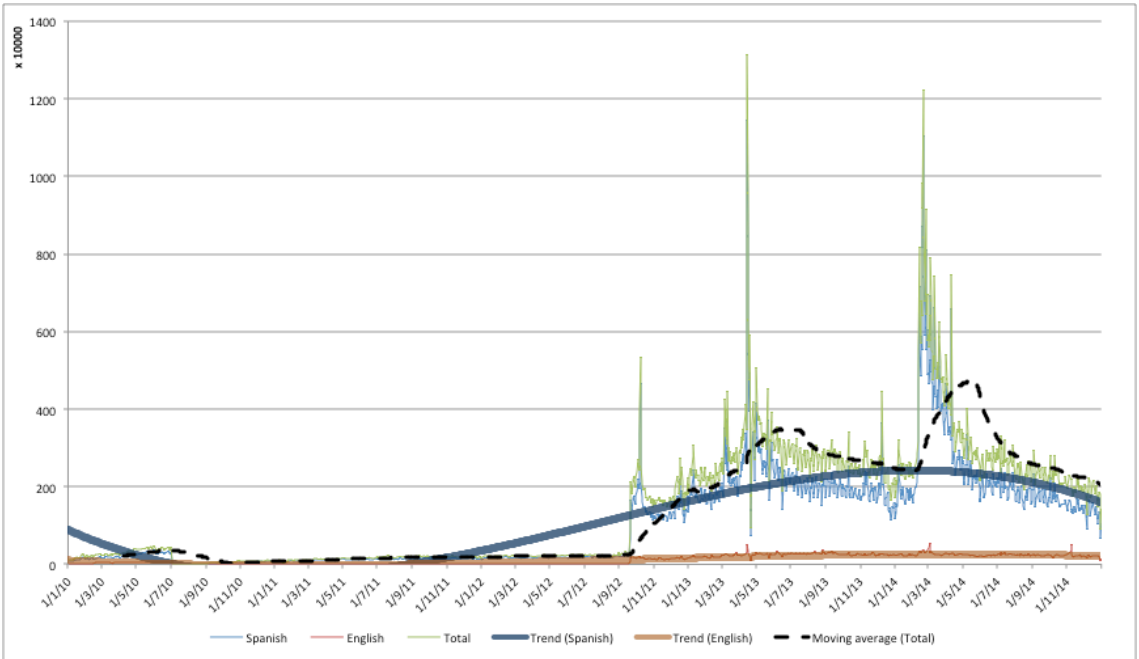
Venezuela

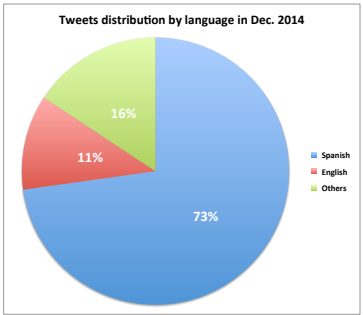
Información general

Período	01/01/2010 a 12/31/2014
Número de tuits	109.360.332
Número de tuits y RT	2.507.541.140
Media de edad *	26,9
Población *	30.851.343
Usuarios de Internet *	12.583.953
% mundial de usuarios de Internet *	0,43%

**información correspondiente al año 2014. Información correspondiente a la población, el número de usuarios de Internet y el porcentaje mundial de cada país de usuarios de Internet obtenida de Internet Live Stats. Información correspondiente a la media de edad obtenida de la CIA.*

Frecuencia de tuits por idioma





	Absoluto	Relativo
Español	41.666.716	73%
Inglés	6.572.592	11%
Otros	9.007.908	16%
Total	57.247.216	100%

Figura 8.11. Informe general de los datos obtenidos de Venezuela

8.1.3 Conclusiones

Como muestran los resultados, durante la recogida de información hemos obtenido millones de tuits que cubren un porcentaje bastante significativo de la población (figura 8.1). En esta figura podemos observar que el porcentaje más bajo corresponde a Paraguay, con casi el 30% del total, mientras que, en el otro extremo, encontramos a Chile (que alcanza el 65,7%) y a Argentina (con un 59,7%). Consideramos, por tanto, que estas cifras son lo suficientemente elevadas como para poder extraer conclusiones fiables, incluso en aquellos países con menor actividad en *Twitter*, como Paraguay.

	Tuits	Población	Usuarios de Internet	Porcentaje mundial de usuarios de Internet
Argentina	375.042.140	41.803.125	24.973.660	0,86%
Chile	611.711.173	17.772.871	11.686.746	0,40%
Colombia	96.456.159	48.660.725	25.583.953	0,88%
República Dominicana	79.004.775	10.528.954	5.072.674	0,17%
Ecuador	68.881.758	17.772.871	6.012.003	0,21%
México	26.354.620	123.799.215	50.923.060	1,74%
Panamá	12.928.175	3.926.017	1.899.892	0,07%
Paraguay	18.405.504	6.917.579	2.005.278	0,07%
Perú	28.689.750	30.769.077	12.583.953	0,43%
Venezuela	109.360.332	30.851.343	12.583.953	0,43%

Figura 8.12. Número de tuits, población, usuarios de Internet y porcentaje mundial de usuarios de Internet

Si consideramos el alto porcentaje de hablantes que utilizan el inglés para escribir en *Twitter* en estos países cuya lengua oficial es el español, podemos concluir que la población está cada vez más en contacto con el idioma anglosajón.

Según Baker (2001), son muchas y muy diversas las posibles vías para alcanzar el bilingüismo en una sociedad. El hecho de que un idioma determinado –en este caso, el inglés– sea utilizado de manera habitual entre la población ejerce una notable influencia en el resto de los hablantes que conviven con esa población. Este fenómeno es lo que Baker denomina "aprendizaje informal de la segunda lengua" (Baker, 2001: 94). En este sentido, afirma el autor que este tipo de aprendizaje "informal", por contagio del resto de habitantes, puede llegar a ser tan influyente como la propia educación.

Conviene no olvidar que la línea de tendencia total incluye otros idiomas minoritarios distintos al inglés y al español, e incluso información no descifrable como idioma por el sistema, lo que explica que, en algunos casos, exista diferencia entre la suma del inglés y el español, y el número total de tuits.

A la hora de analizar los resultados, hemos realizado una comparación entre la línea de tendencia de cada idioma y la total. A continuación, se ha establecido una clasificación de cinco grupos de países, según su situación lingüística.

Primer grupo: aumento de la línea de tendencia del inglés, descenso de la línea de tendencia del español

En esta primera clasificación hemos agrupado aquellos países en los que el uso de la lengua inglesa está aumentando en comparación con la tendencia del número total de tuits, mientras que el uso del español está disminuyendo. Esta situación se está dando en México y en Chile. No obstante, a pesar de estas tendencias, sigue habiendo más tuits escritos en español que en inglés, aunque podemos predecir que esta situación podría invertirse en un futuro cercano.

En México hay un punto de inflexión evidente que se puede situar en septiembre de 2012. A partir de este momento, la tendencia del español decrece en comparación con la línea de tendencia total. Por el contrario, unos meses más tarde, a principios de 2013, podemos observar cómo la tendencia del inglés comienza a subir gradualmente, superando la línea de tendencia principal. Para intentar comprender las posibles causas de este fenómeno, debemos centrarnos en el período comprendido entre el tercer trimestre de 2012 y el primer trimestre de 2013, que es cuando se producen los cambios más significativos.

Si analizamos los máximos en México, observamos que los puntos más altos se corresponden con el día del concierto de Justin Bieber en El Zócalo, Ciudad de México. Encontramos otro máximo relativo el 1 y el 2 de julio de 2012, las fechas de las elecciones federales en las que hubo un cambio de gobierno. Este podría ser un buen punto de partida para analizar si las políticas lingüísticas y educativas posteriores pudieron ejercer influencia en la línea de tendencia de algún idioma. Finalmente, hay otro máximo relativo el 21 de junio de 2014, coincidiendo con el ultimo partido de la selección mexicana en el mundial de fútbol.

Segundo grupo: aumento de la línea de tendencia del inglés y mantenimiento de la línea de tendencia del español

El número de tuits publicados en español en este segundo grupo de países se mantiene, pero los usuarios están incrementando su uso del inglés a la hora de publicar tuits. Países como Colombia, Panamá, Paraguay y la República Dominicana se encuentran en esta situación. En República Dominicana, de hecho, el número total de tuits en inglés ya supera al del español. Con respecto a los otros tres países –Colombia, Panamá y Paraguay–, nos atrevemos a decir que pronto se alcanzará el mismo número de tuits escritos en ambos idiomas, aunque en Colombia ocurrirá algo después que en Panamá y en Paraguay.

La mayor parte de los máximos producidos en Panamá coinciden con eventos deportivos, como el partido de fútbol entre Panamá y Estados Unidos el 16 de octubre de 2013. La producción total de tuits, sin embargo, desciende a partir del 5 de mayo de 2014, después de las elecciones presidenciales en las que resultó elegido Juan Carlos Valera.

Por otro lado, en Colombia encontramos una subida drástica de los tuits el 4 de julio de 2014, pero se trata solo de tuits en español. Esta fecha coincide con la eliminación de la selección colombiana del mundial de fútbol.

En términos absolutos, la línea de tendencia general de Paraguay no es significativa debido a que el número total de tuits es mucho menor que en los demás países, por lo que pequeños cambios generan gráficos muy acentuados. No obstante, debemos señalar cómo se aprecia un mayor uso de *Twitter* a partir de octubre de 2012.

La República Dominicana posee la tasa más alta de bilingüismo de los países estudiados. La irrupción de *Twitter* también se sitúa en la misma fecha que en Paraguay, solo unos meses después de la llegada al gobierno del presidente Danilo Medina. Se puede apreciar un incremento repentino desde esta fecha hasta, aproximadamente, un año más tarde, donde este aumento comienza a estabilizarse. Algunas de las razones que pueden explicar la enorme presencia del inglés en el país son la influencia de la economía americana, la importancia del sector servicios, centrado en el turismo, o los flujos migratorios. En cualquier caso, es evidente que el bilingüismo está muy presente en esta parte de la isla, como ya señalaba Alvar (1985). Resulta significativo el hecho de que, a pesar de ser la lengua oficial, el español es el idioma menos utilizado en *Twitter*,

comparado con otros idiomas o con tuits no identificados y con el inglés, que es el idioma dominante.

Tercer grupo: descenso de la línea de tendencia del español y mantenimiento de la línea de tendencia del inglés

Este es el caso de Ecuador. En la figura 8.6 se muestra que el número de tuits escritos en inglés es más alto que el de aquellos escritos en español. Sin embargo, con respecto a la línea de tendencia total, podemos observar que el uso del inglés no está aumentando; lo que sí está ocurriendo es un descenso en la utilización del español, con respecto al total. Se produce también un incremento considerable en la producción total de tuits, coincidiendo con la tercera reelección de Rafael Correa como presidente del gobierno. A partir de este momento, podemos ver cómo el inglés y otros idiomas son los responsables del crecimiento.

Cuarto grupo: las tres líneas de tendencia se mantienen

Argentina y Perú se encuentran dentro de este grupo de países en los que no hay evidencia de cambios significativos a la hora de los idiomas de escritura de los tuits. En ambos países, la línea de tendencia del total de tuits publicados aumenta, pero lo hace de forma paralela a los dos idiomas estudiados, lo que significa que los habitantes hacen un mayor uso en general de la plataforma, pero el uso del inglés y del español se mantiene estable.

Quinto grupo: la línea de tendencia del español se mantiene estable y existe una baja presencia de otros idiomas

Encontramos el caso particular de Venezuela, donde la existencia del bilingüismo es prácticamente nula. Desde un punto de vista puramente descriptivo, y a partir de los tres máximos presentes en el gráfico, podemos ver que hay ciertos eventos sociales y políticos relevantes que aumentan la actividad en *Twitter*, pero esta actividad se produce, casi en su totalidad, en español. Estos tres puntos importantes los situamos con la victoria de Hugo Chávez en las elecciones presidenciales, el 7 de octubre de 2012; con su muerte y funerales, entre el 5 y el 8 de marzo; y con la elección de Nicolás

Maduro como presidente del gobierno el 14 y 15 de abril de 2013, respectivamente. Se produce también un aumento significativo el 20 de febrero de 2014, coincidiendo con una oleada de protestas violentas contra el gobierno. Sin embargo, como decimos, la práctica totalidad de los tuits están escritos en español.

Si tenemos en cuenta la información relativa al mes de diciembre de 2014, el porcentaje de uso del español es de un 73%, mientras que el del inglés se corresponde con n 11%. Existe también un 16% en el que se incluyen otros idiomas e información no textual. Convendría, quizá, llevar a cabo un estudio sociopolítico para llegar a comprender las posibles causas de esta situación, tan distinta a la del resto de países analizados en los que, en mayor o menor grado, hay una clara evolución favorable del bilingüismo.

Tras recopilar toda la información de tipo geográfico, demográfico, económico y textual, observamos que no hay una relación directa entre estos aspectos y el grado de implementación del bilingüismo. Así lo demuestra, por ejemplo, el caso de México, que a pesar de tener una población de casi 124 millones de habitantes y de su proximidad con Estados Unidos, posee una tasa de implementación del bilingüismo más baja que la de República Dominicana. Chile, por otro lado, también experimenta un mayor crecimiento del uso del inglés que sus vecinos Perú y Argentina, a pesar de ser países limítrofes. Por el contrario, otro países más alejados entre sí, como Panamá y Perú, presentan tendencias similares en la evolución del bilingüismo. A modo de síntesis, podemos afirmar que, partir de los datos obtenidos con *big data*, el bilingüismo en los países estudiados está implantándose con fuerza y creciendo gradualmente en la mayoría de ellos, con la curiosa excepción de Venezuela.

8.2 ANÁLISIS DE LAS NECESIDADES TRADUCTOLÓGICAS EN CINCO CAPITALES EUROPEAS

En este segundo estudio con el que tratamos de ejemplificar algunos de los usos potenciales de *Wordics Archive*, nos marcamos como objetivo conocer las necesidades traductológicas en cinco capitales europeas a partir de la información que nos ofrece *big data*. Esta investigación es también, al igual que la anterior, una adaptación de un artículo de la autora de este trabajo que se encuentra aún por publicar (González Fernández, en prensa). El modo de trabajo empleado ha consistido en utilizar *Wordics Archive*, cuando todavía estaba en la fase inicial de su diseño, para obtener los tuits

publicados en estas ciudades y determinar cuáles son los idiomas más utilizados, situándolos también en el mapa.

8.2.1 Metodología

Las capitales seleccionadas en esta investigación han sido Berlín, Bruselas, París, Madrid y Londres. Para el estudio, hemos establecido un período de tiempo de un mes, comprendido entre el 21 de agosto y el 21 de septiembre de 2015. *Wordics Archive* ha recopilado los tuits publicados durante este período de tiempo y ha almacenado la identificación única de cada uno de ellos, 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 solo 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.

En esta etapa inicial, en la que la herramienta se dedicaba fundamentalmente a la recogida de información de *Twitter*, se diseñó también, al igual que en la investigación anterior, una versión preliminar de visualizador, que permite mostrar una representación gráfica para los fragmentos temporales y espaciales seleccionados.

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

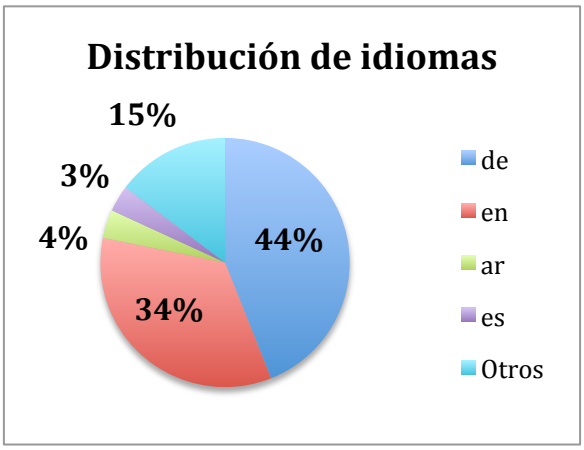
un color distinto. Los colores se indican en los cuatro siguientes mapas, donde vienen desglosados los tuits por idioma.

8.2.2 Resultados

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

Figura 8.13. Distribución de idiomas en Berlín

Región seleccionada y distribución geográfica de idiomas

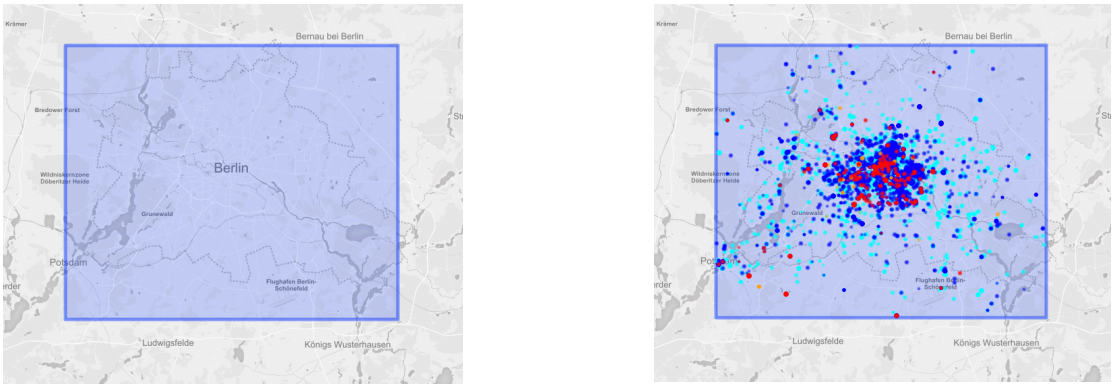


Figura 8.14. Región seleccionada y distribución geográfica de idiomas en Berlín

Distribución detallada por idioma

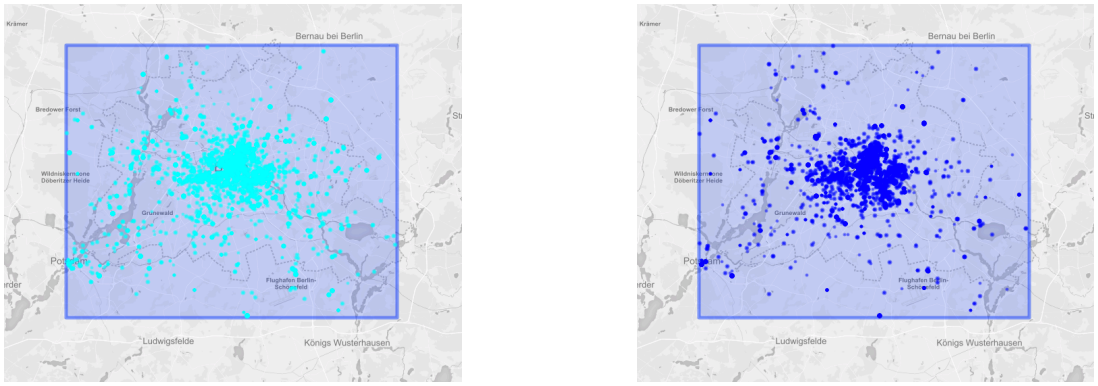


Figura 8.15. Región seleccionada y distribución geográfica de inglés y alemán en Berlín

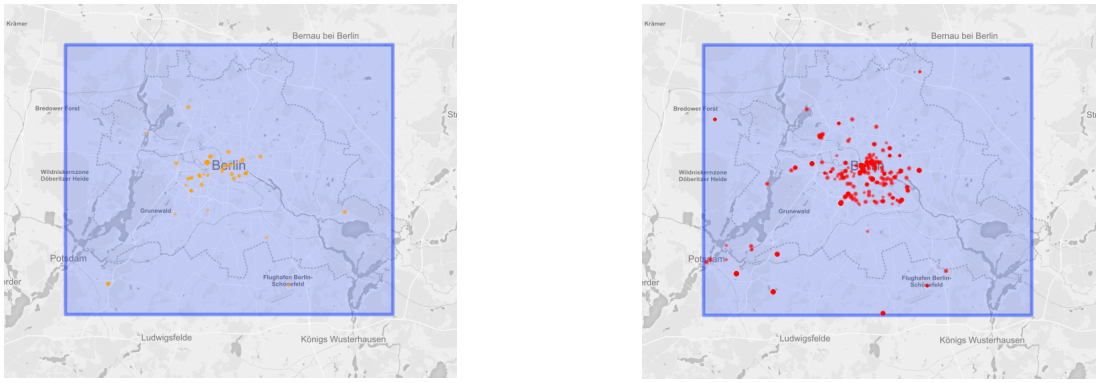


Figura 8.16. Región seleccionada y distribución geográfica de árabe y español en Berlín

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 aproximadamente 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.

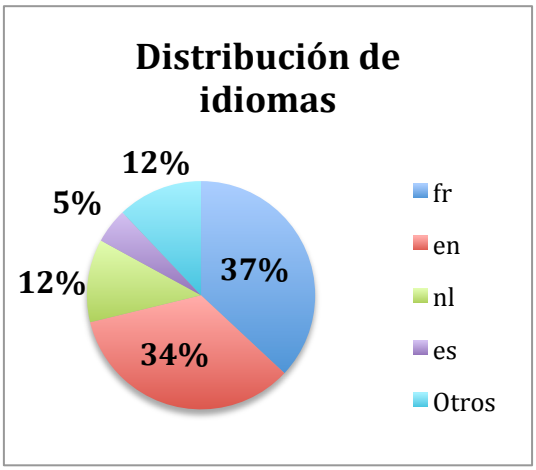
Así, la utilización del árabe se limita fundamentalmente a la parte más céntrica de la ciudad y solo en algunos puntos concretos, aunque también se utiliza en algún otro punto de la región, sobre todo al sur. Por el contrario, el español, aunque menos hablado, tiene una distribución más amplia hacia 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

Figura 8.17. Distribución de idiomas en Bruselas

Región seleccionada y distribución geográfica de idiomas

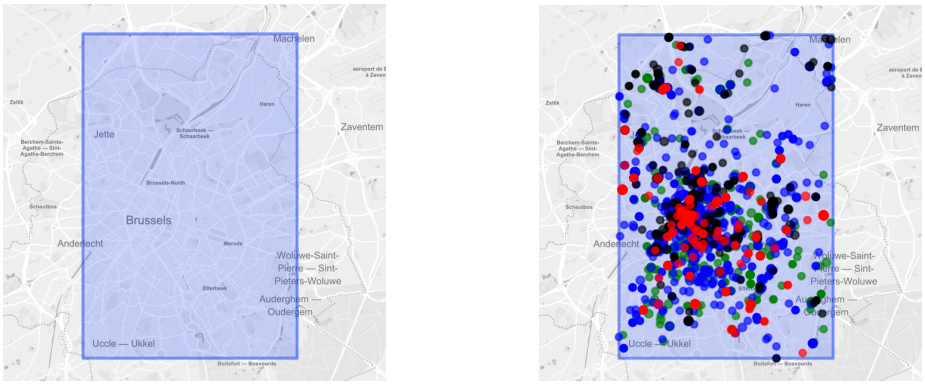


Figura 8.18. Región seleccionada y distribución geográfica de idiomas en Bruselas

Distribución detallada por idioma

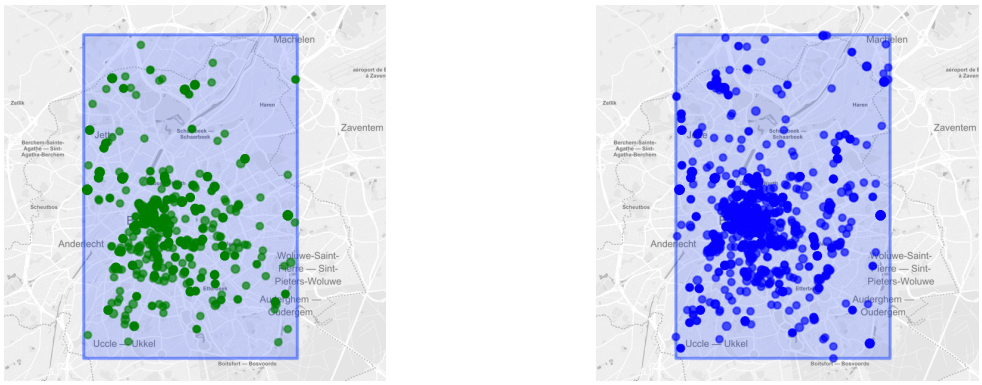


Figura 8.19. Región seleccionada y distribución geográfica de francés e inglés en Bruselas

Figura 8.20. Región seleccionada y distribución geográfica de neerlandés y español en Bruselas

Comentarios

En el caso de 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.

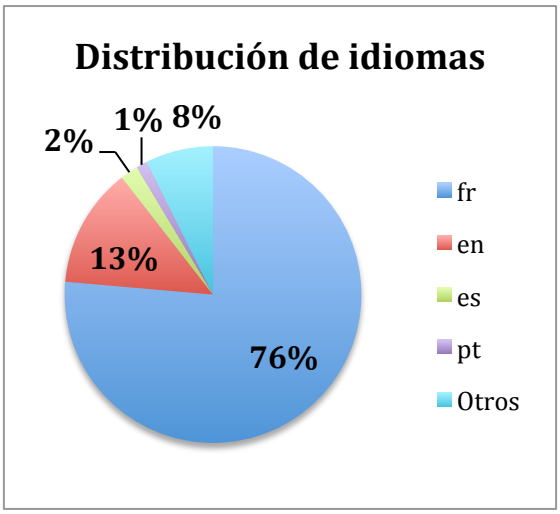
Los mapas reflejan esta situación y muestran que tanto el francés como el inglés se utilizan de forma aproximada por las mismas zonas de la región, sobre todo en el centro, aunque también observamos cómo el inglés se extiende por algunas partes del norte y del noreste. El rastro que dejan los tuits escritos 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 otro

lado, 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, es únicamente en el oeste 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 este último se encuentra más repartido por el sur.

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

Figura 8.21. Distribución de idiomas en Bruselas

Región seleccionada y distribución geográfica de idiomas

Figura 8.22. Región seleccionada y distribución geográfica de idiomas en París

Distribución detallada por idioma

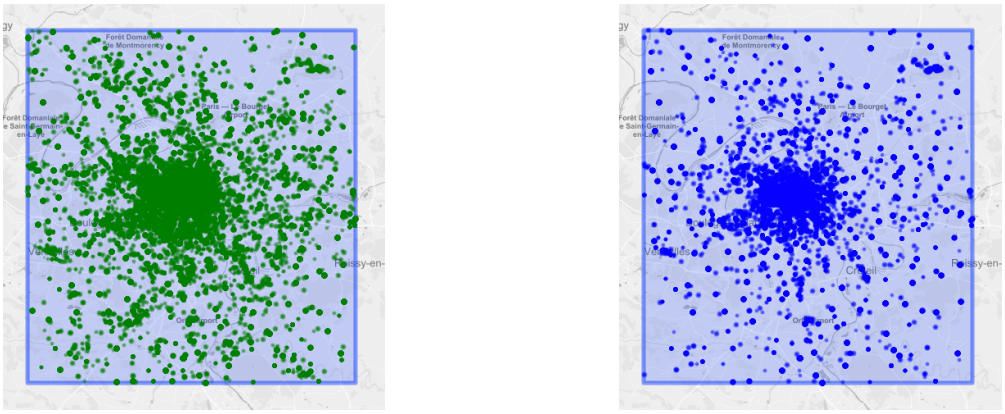


Figura 8.23. Región seleccionada y distribución geográfica de francés e inglés en París

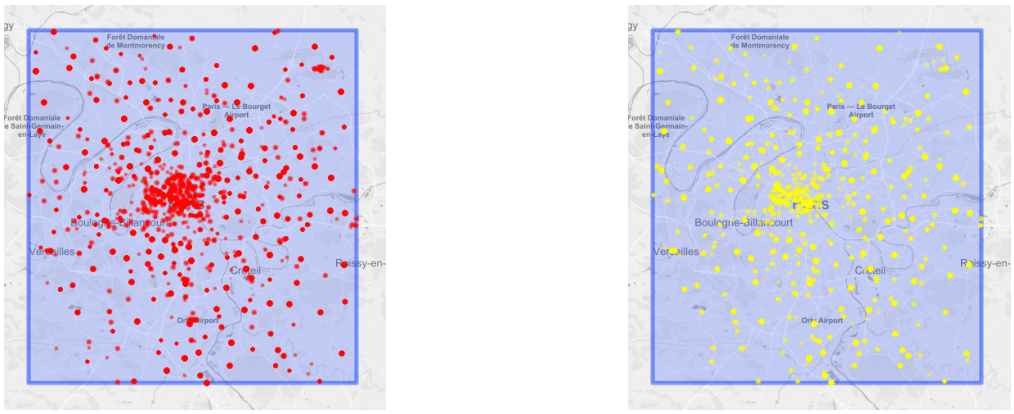


Figura 8.24. Región seleccionada y distribución geográfica de español y portugués en París

Comentarios

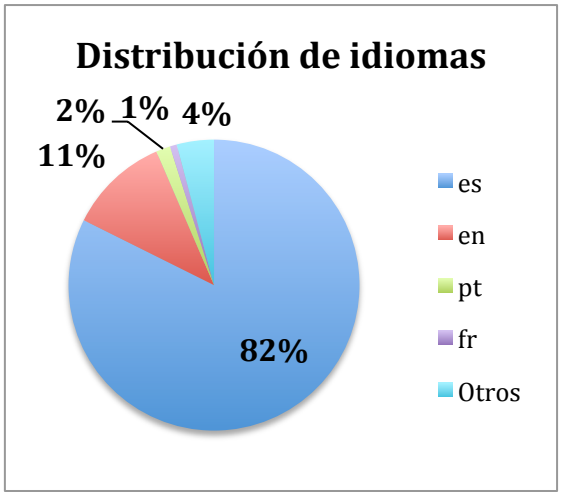
En el caso de la capital francesa, la lengua oficial presenta un claro predominio sobre las restantes, encabezadas por el inglés, que solo representa algo más de un octavo del porcentaje de tuits escritos en francés. En una situación similar se encuentra el español con respecto a su antecedente más inmediato, con solo 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, solo cubre 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 solo 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 homogénea 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

Figura 8.25 Distribución de idiomas en Madrid

Región seleccionada y distribución geográfica de idiomas

Figura 8.26. Región seleccionada y distribución geográfica de idiomas en Madrid

Distribución detallada por idioma

Figura 8.27. Región seleccionada y distribución geográfica de español e inglés en Madrid

Figura 8.28. Región seleccionada y distribución geográfica de portugués y francés en Madrid

Comentarios

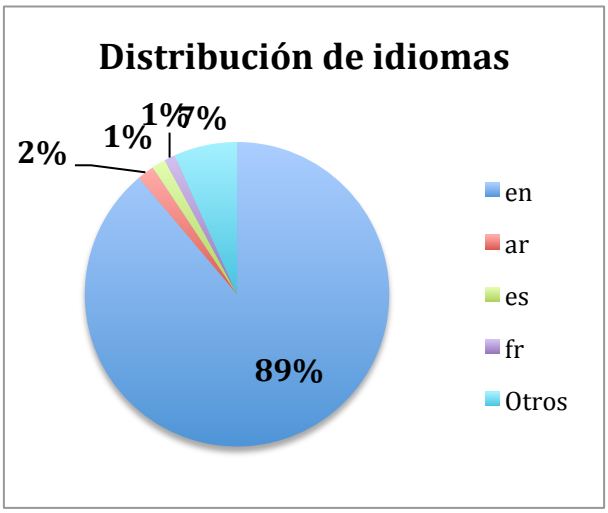
Los cuatro idiomas más utilizados en Madrid coinciden con los de París, aunque con distinto orden de frecuencia. Naturalmente, el español es la lengua más usada entre los usuarios de *Twitter*. Existe aquí una diferencia aún mayor entre esta primera lengua y la segunda más usada –el inglés–, que solo 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 que presentaba 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: 1/8 – 21/9

en	2416695
ar	49125
es	40073
fr	30953
otros	183223

Figura 8.29. Distribución de idiomas en Londres

Región seleccionada y distribución geográfica de idiomas

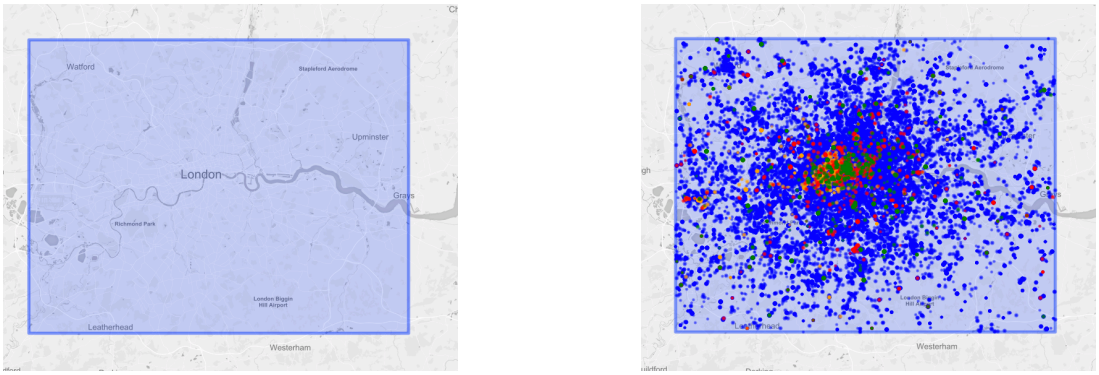


Figura 8.30. Región seleccionada y distribución de idiomas en Londres

Distribución detallada por idioma

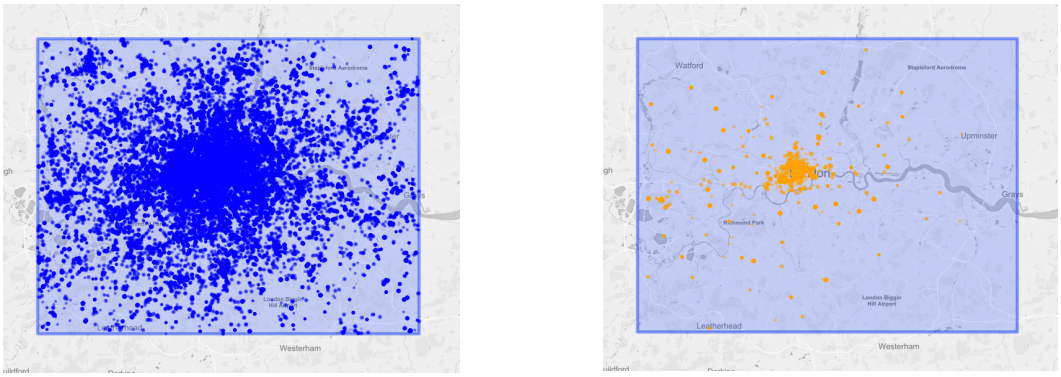


Figura 8.31. Región seleccionada y distribución geográfica de inglés y árabe en Londres

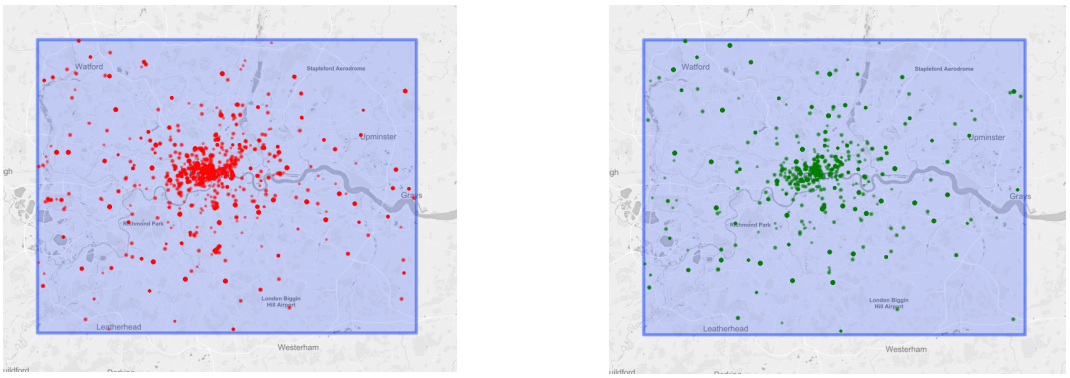


Figura 8.32 Región seleccionada y distribución geográfica de español y francés en Londres

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 de idiomas 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– solo representan un 2% (árabe) y un 1% (español y francés). 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 que parte 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 registran 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.

8.2.3 Conclusiones

En este ejemplo de investigación, hemos podido demostrar la eficacia de *Wordics Archive* a la hora de llevar a cabo estudios de lingüística aplicada útiles no solo para obtener información relacionada con el lenguaje y su uso, sino también para poder enfocarla a otros aspectos prácticos del ámbito profesional, como el sector de la traducción.

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 así el resto en una sola categoría.

Desde un punto de vista profesional traductológico comprobamos, por ejemplo, que el inglés predomina como segunda lengua más usada en todas las capitales en las que no actúa 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 solo 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. Otros 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 realista 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 solo ahorra tiempo y costes con respecto a los métodos tradicionales, 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.

8.3 *TWITTER* COMO HERRAMIENTA PARA EL ESTUDIO DE NEOLOGISMOS

Este tercer estudio de caso y último ejemplo de nuestro trabajo pretende, una vez más, dar cuenta 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 determine. En este caso, aportaremos solo algunos ejemplos de este fenómeno en la lengua española, aunque, gracias a *Wordics Archive*, sería posible llevar a cabo este tipo de investigación en cualquiera de los idiomas integrados en *Twitter*.

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, puesto que sería una tarea completamente inabarcable en un estudio de estas características, 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 los neologismos[50]. Nos limitaremos, por tanto, a hacer referencia al estado de la cuestión en la investigación neológica y 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)[51] y cuyo interés principal se 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). *Wordics Archive* 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 los otros dos módulos de la herramienta ya vistos: *Wordics Live* –para el estudio en tiempo real–, y *Wordics One* –para el estudio de un usuario determinado 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 al*., 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).

8.3.1 Metodología

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

[51] 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 Hormingo, 2015).

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.). Nos proponemos aportar algunos ejemplos de neología léxica, siguiendo la clasificación de Guerrero Ramos (1995).

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 al*. (2004). La metodología seguida ha consistido en seleccionar la opción de *análisis simple* o la de *análisis comparado* –dependiendo de las necesidades en cada caso– que facilita de *Wordics Archive* para introducir palabras en el buscador. En este caso, a diferencia de los dos estudios anteriores, no ha sido necesario utilizar el filtrado por coordenadas, puesto que nuestro objetivo consiste en comprobar la utilización de determinadas palabras a nivel mundial. Sí hemos hecho uso, por el contrario, de la selección de idioma, de manera que, una vez que la herramienta nos ha devuelto la gráfica con la frecuencia de utilización de la palabra en cuestión, hemos desactivado todos los idiomas, excepto el español, para poder ceñirnos al estudio de los neologismos en nuestra lengua.

El rango de fechas que la herramienta ofrece por defecto tampoco ha sido modificado, de manera que el análisis se ha realizado sobre toda la información que tenemos almacenada hasta el momento procedente de *Twitter*; esto es, desde septiembre de 2015.

Los corpus de tuits obtenidos para cada búsqueda realizada podemos encontrarlos en el Anexo 3 y los documentos que este contiene, mientras que, a continuación, mostramos los mapas 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.

Por otro lado, no debemos olvidar 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, así como las KWIC y las colocaciones de cada palabra (todos estos datos se pueden consultar con más detalle en el Anexo 3). Estas dos últimas se pueden filtrar también por idioma e incluso por una palabra contenida en los tuits del corpus en cuestión, pero distinta a la palabra objeto de búsqueda.

8.3.2 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[52], 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*:

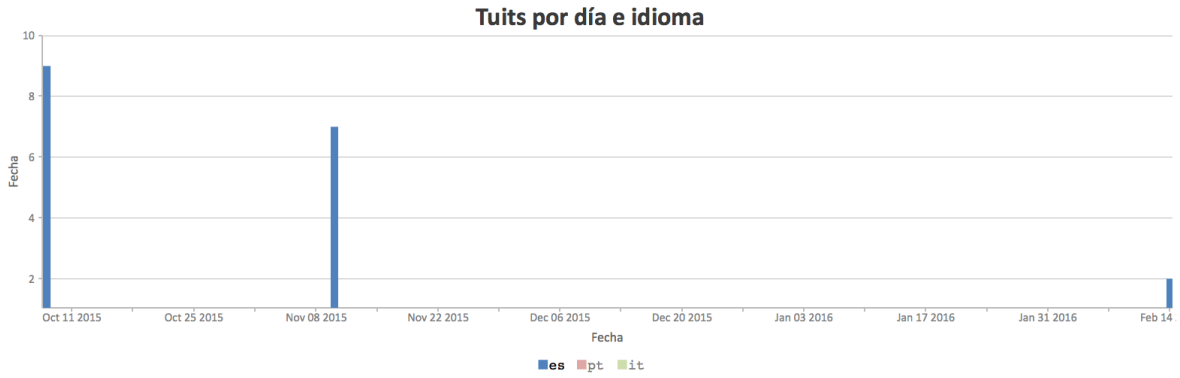


Figura 8.33. Gráfica de la frecuencia de uso de *poliamor*

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 (como nos indica la herramienta en el cuadro que aparece a la izquierda del gráfico). Estas son dos de ellas en contexto, para consultar el corpus completo, con información también acerca de la latitud y la longitud donde se ha publicado, véase el Anexo 3.1:

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.
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.

[52] Se puede consultar Alvar Ezquerra (1973), Lang (1992), Varela Ortega (2015) o Díaz Hormigo (2015), para ahondar más en esta cuestión.

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 número*), se han formado palabras nuevas añadiendo un sufijo 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 roja representa a *beticismo*, la azul, a *madridismo* y la verde, a *sevillismo*, como se puede observar en la leyenda:

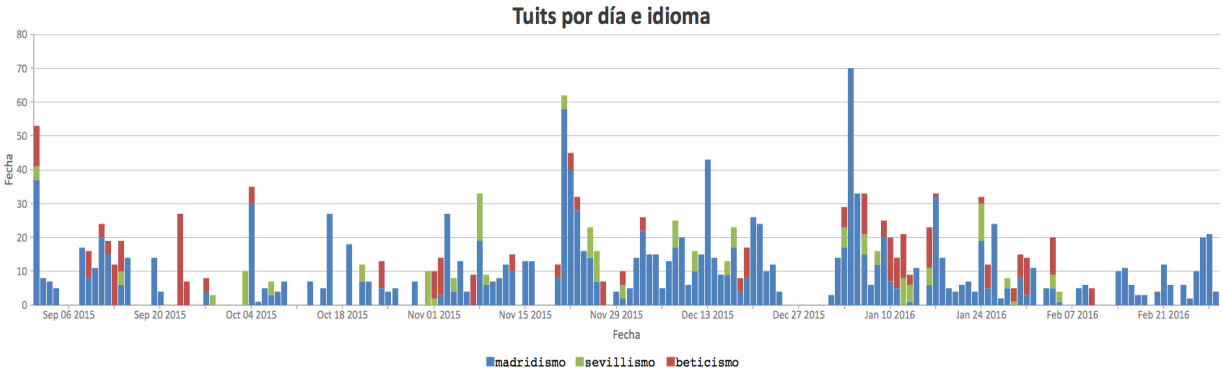


Figura 8.34. Gráfica de la frecuencia de uso de *madridismo*, *sevillismo* y *beticismo*

Los resultados numéricos en español son:

-*Beticismo*: 288 apariciones.

-*Madridismo*: 1.269 apariciones.

-*Sevillismo*: 163 apariciones.

Veamos algunos ejemplos de uso:

1 sept 2015, 00:47:43	Que grande!!! Gran ilusión para todo el beticismo! #LaMejorAficion https://t.co/wy5wZjDVgc
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
25 ene 2016, 00:05:26	Bonito día para el beticismo. Ya era hora! https://t.co/sa5Xr7Cvko
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.
9 feb 2016, 00:37:06	que gozada ver el madridismo patas arriba #ChiringuitoMadrid
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
8 nov 2015, 22:17:54	Gran Sevilla ante un Madrid mediocre. Las críticas del sevillismo a su equipo son desmesuradas, un año más.
4 feb 2016, 22:55:49	Sevillismo puro! La giralda presume orgullosa...? https://t.co/hLx2pSzVAl
5 feb 2016, 00:56:04	Feliz noche a todo el sevillismo ! Sevilla una vez más me llevas a una final ..!

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:

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
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
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 un ejemplo comparado de ambos términos en el que se pueden observar sus frecuencias de uso (*trolear* aparece en color azul y *troleo* lo hace en rojo):

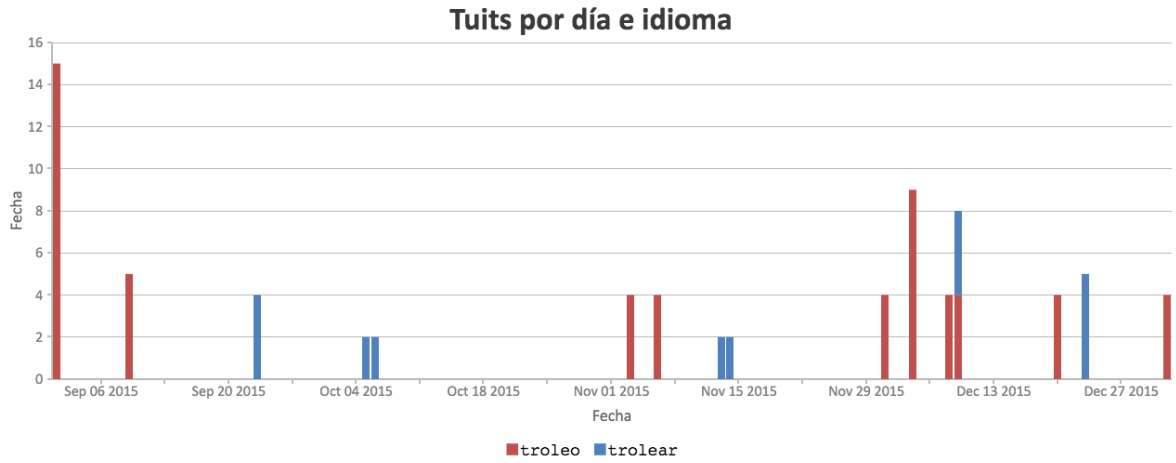


Figura 8.35. Gráfica de la frecuencia de uso de *troleo* y *trolear*

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

23 sept 2015, 23:19:23	He sido formado por mi tío el calvo en el arte de trolear y vacilar
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
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
9 sept 2015, 22:30:29	Se esta quedando con todos vosotros, vaya troleo jajajajajajaja #CantizanoEH @ElHormigueroMx
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
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 mismo patrones de comportamiento que *trolear* y *troleo* siguen

los neologismos *posturear* y *postureo*, 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*:

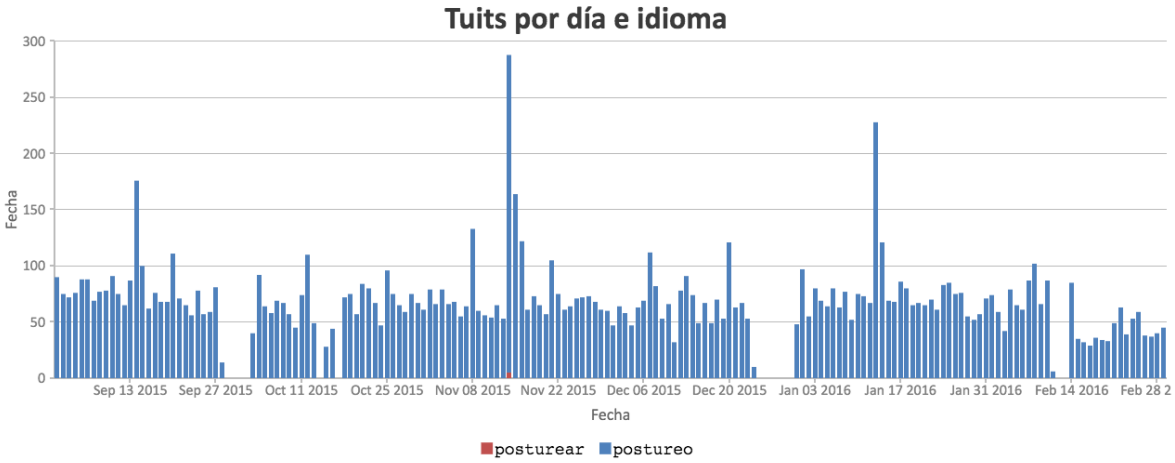


Figura 8.36. Gráfica de la frecuencia de uso de *postureo* y *posturear*

Podemos observar algunos ejemplos de uso a continuación:

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
1 sept 2015, 18:14:59	Ahora mismo esto llega a niveles de postureo que ni la Nasa conoce... http://t.co/9TnnbXoDsp
5 sept 2015, 00:45:42	Me da coraje toda esa gente que se compra cámaras reflex por postureo
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
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?
14 nov 2015, 16:07:11	Yo Postureo del verbo posturear #YoPostureo #MiguelitoStyle #DeMayorQuieresSerComoYo #Malaga #Style #Sabado https://t.co/Tz8PEJo4qB

Los mecanismos de formación de palabras nuevas a partir da 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 (número x) muestra un total de 152 apariciones de *veroño*:

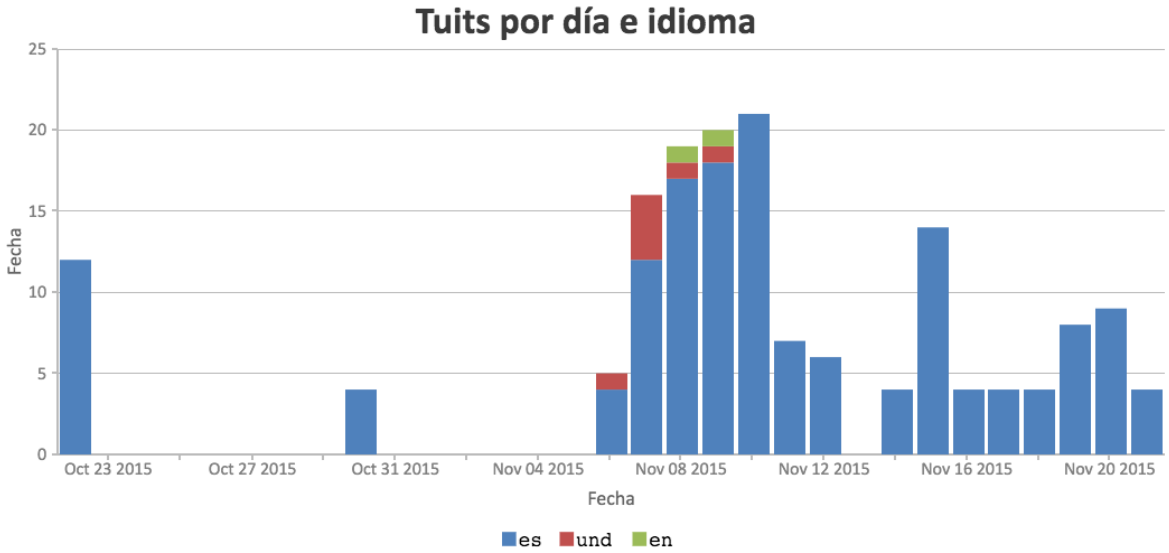


Figura 8.37. Gráfica de la frecuencia de uso de *veroño*

Estos son algunos ejemplos reales de uso de este término:

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 □□□
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
9 nov 2015, 19:58:26	¿Quien dice Noviembre? En Málaga seguimos con el veroño ..y parece que se quiere quedar #DíasCalurosos

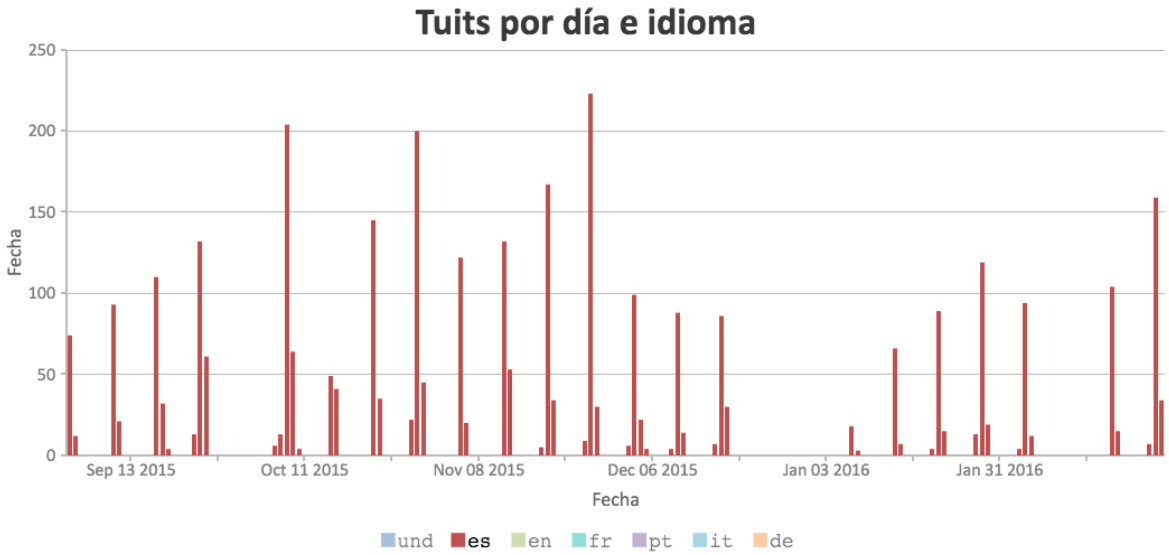


Figura 8.38. Gráfica de la frecuencia de uso de *juernes*

Resulta muy llamativa la gráfica de *juernes* (figura 8.38) si tenemos en cuenta que todos los máximos relativos que aparecen en ella coinciden en 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 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. En el período analizado registramos un total de 3.317 apariciones de *juernes*:

3 sept 2015, 20:17:50	Alguien sale de juernes
4 sept 2015, 13:20:27	Juernes con mis amores 💏❤ @ Bora Bora Polinesian Bar https://t.co/l3ulh1IGHt
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 prestamo 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 posible salida de Gran Bretaña de la Unión Europea y formado a partir de las palabras *Britain+exit*.

Desde el 1 de septiembre hasta el 19 de febrero, *Wordics Archive* registró 15.988 ocurrencias de *brexit* en todo el mundo, de las cuales 252 fueron en español:

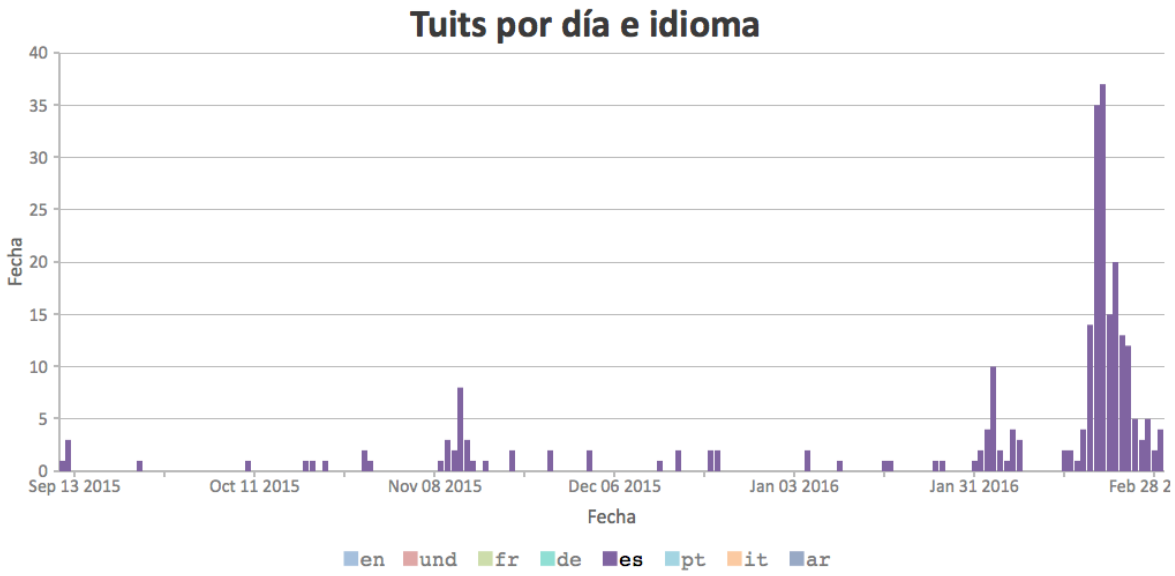


Figura 8.39. Gráfica de la frecuencia de uso de *brexit* en español

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 a la derecha del buscador. Para llevar a cabo la búsqueda, hemos 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 1.634 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 8.40) la evolución de esta palabra:

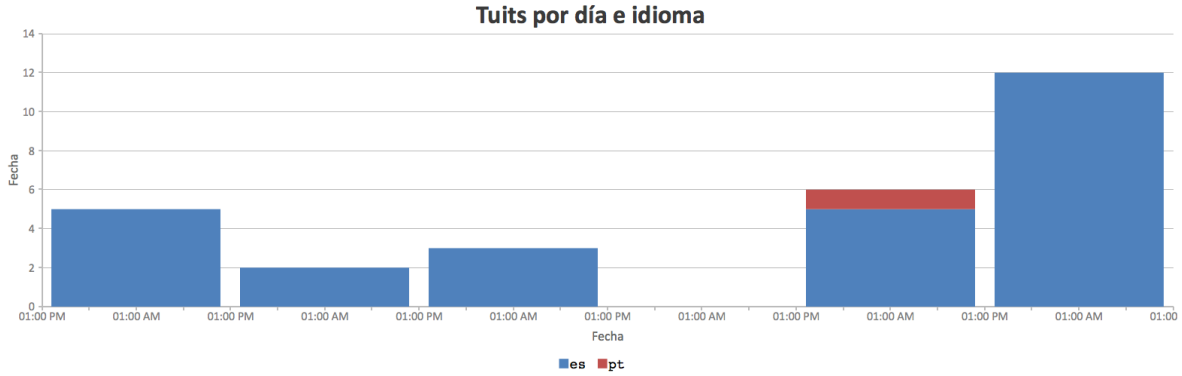


Figura 8.40. Gráfica de la frecuencia de uso de *googlear*

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:

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.
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.
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 4.686 apariciones en español, como podemos comprobar en la figura 8.41:

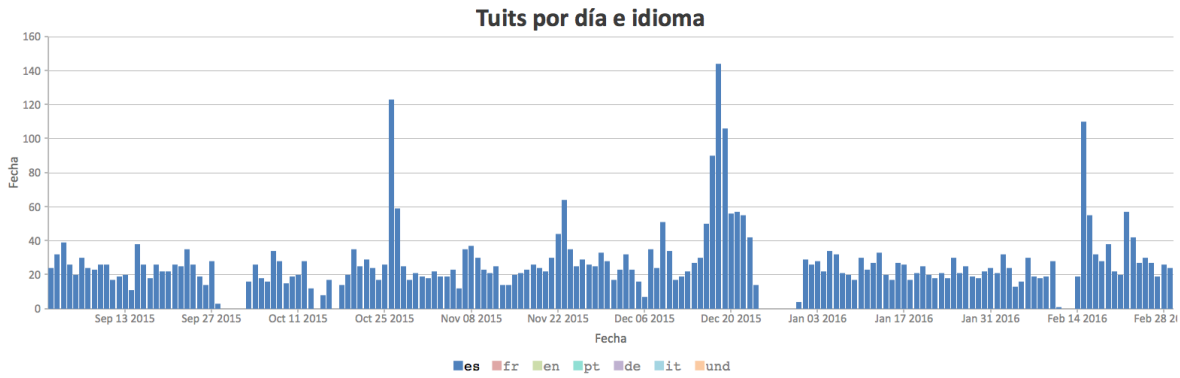


Figura 8.41. Gráfica de la frecuencia de uso de *spoiler* en español

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:

1 sept 2015, 00:16:04	Cuando tu madre te hace el spoiler del último capítulo de la mejor serie del mundo...□□□ #breakingbad
2 sept 2015, 03:43:51	Quiero un spoiler sobre lo que va a pasar con el mundo
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 *Wordics Archive*, cómo todavía sigue predominando en español la forma gráfica inglesa, con 11.554 apariciones, frente a las 111 de la forma adaptada al español, *selfi*:

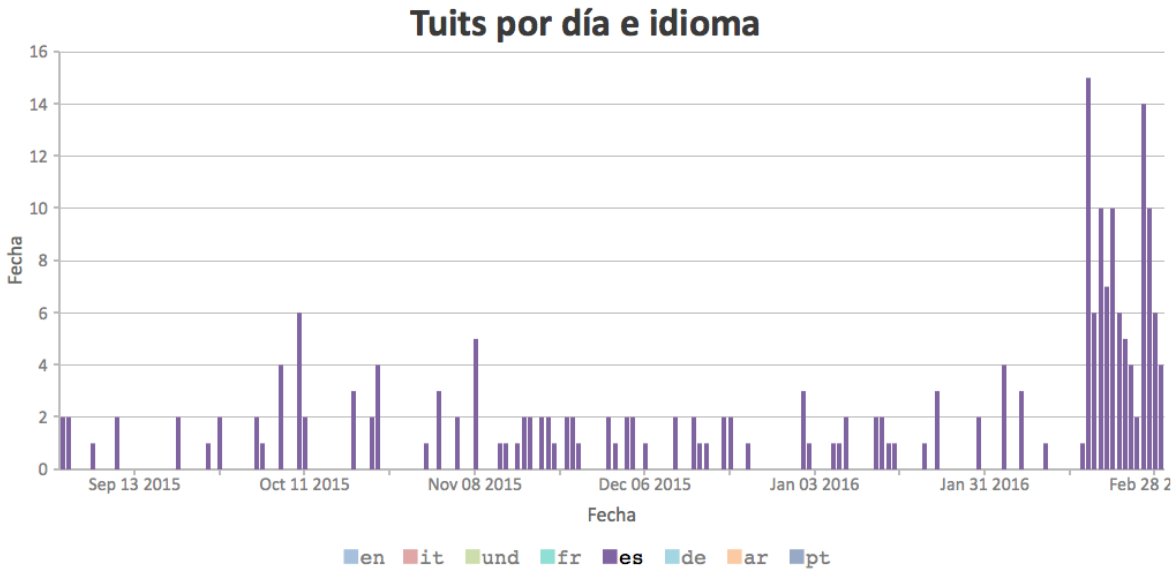


Figura 8.42. Gráfica de la frecuencia de uso de *selfi* en español

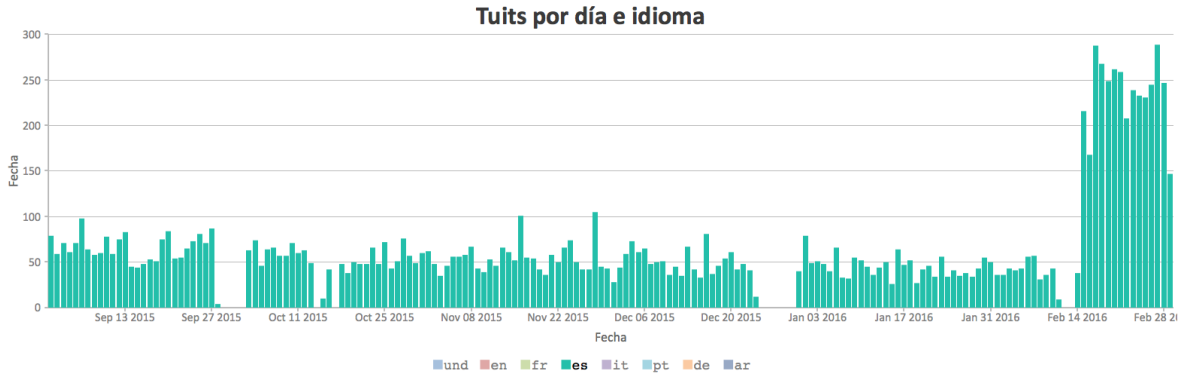


Figura 8.43. Gráfica de la frecuencia de uso de *selfie* en español

A pesar de 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.

8.3.3 Conclusiones

Tras analizar los resultados, creemos que es posible utilizar *Twitter* para poder obtener información acerca de los comportamientos que en cuestión neológica adopta lengua determinada, que en este caso ha sido el español. Además, consideramos que,

gracias a investigaciones de este tipo, es posible contribuir 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).

Conclusiones generales

Los beneficios de la utilización de *big data* para la obtención de conocimiento y para el avance de la ciencia quedan avalados por el creciente número de estudios e investigaciones que se están publicando al respecto desde hace varios años, en su mayoría enfocados a los ámbitos de la economía, del comercio o de la sanidad. En la línea de estas aportaciones, el objetivo de este trabajo ha sido, desde el primero momento, presentar la posibilidad de investigar el lenguaje y la lengua en el marco de este reciente concepto de *big data* y, en concreto, a través de *Twitter*.

Para ello, hemos creído oportuno llevar a cabo una fundamentación teórica que siente las bases de nuestra investigación y, a su vez, proponer una nueva metodología adecuada para acometer la investigación lingüística basada en *big data*.

Esta investigación, como esperamos haber demostrado a lo largo del trabajo, habría sido impensable con una metodología tradicional. Para poder llevar a cabo los estudios que nos permitieran probar la veracidad de nuestra hipótesis, ha sido absolutamente necesario disponer de una herramienta informatizada que posibilitara la extracción, el almacenamiento, la gestión y el análisis de la información. Puesto no que existía hasta el momento ninguna otra herramienta, software o aplicación web que pudiera ajustarse a los requerimientos de nuestra investigación, nos propusimos crear la herramienta que en este trabajo presentamos: *Wordics Suite*. Esta aplicación nace con el objetivo de servir al lingüista en investigaciones de muy diversa índole sin que sean necesarios conocimientos informáticos previos, y con la intención de obtener mejores resultados en menos tiempo. *Wordics Suite* ha sido, por tanto, creada y diseñada específicamente para nuestro trabajo, involucrando a un desarrollador informático con el que hemos trabajado para hacer posible la elaboración de la herramienta.

En consecuencia con esto, las conclusiones obtenidas a partir de nuestra investigación presentan una doble perspectiva. Por un lado, el trabajo nos conduce a conclusiones de índole teórica y disciplinar, puesto que podemos afirmar que la relación entre las dos disciplinas en torno a las cuales se construye esta tesis –la Lingüística y la Informática– no solo es necesaria para la obtención de un mayor y más profundo conocimiento de nuestro sistema lingüístico, sino que también resulta deseable para el avance y el progreso de la ciencia, sin dejar de perder de vista el papel que ocupa cada

científico en esta unión. En segundo lugar, también hemos llegado a conclusiones de carácter metodológico, puesto que hemos corroborado la utilidad, la pertinencia y las ventajas de la utilización de *big data* y de *Twitter*, para ser más concretos, en la investigación lingüística, al mismo tiempo que enfatizamos la necesidad de disponer de las herramientas y soportes informáticos adecuados para que se puedan efectuar los distintos análisis lingüísticos.

Con la intención de situar y de delimitar nuestro campo de estudio, consideramos primordial llevar a cabo una revisión teórica que abarcara las distintas disciplinas desde las que abordamos la investigación, que son, a grandes rasgos la Lingüística y la Informática. La unión entre estas dos ciencias, no obstante, no es nueva, y mucho se ha escrito y trabajado al respecto, como hemos visto en el primer capítulo del presente trabajo. Las investigaciones en este sentido desembocaron, hace más de medio siglo, en la denominada Lingüística Computacional. Los motivos por los que la unión entre la Lingüística de Corpus, en concreto, y la Lingüística computacional ha sido tan estrecha durante estos últimos años están relacionados con la gran cantidad de ejemplos reales de uso del lenguaje que la primera almacena sistemáticamente en bases de datos en forma de corpus. La Lingüística Computacional, a su vez, nos brinda sofisticadas herramientas informatizadas que extraen y analizan la información textual que contienen los corpus para después crear distintos productos relacionados con el lenguaje.

El advenimiento de la *World Wide Web* a finales del siglo veinte supuso la mayor revolución de la historia en todos los ámbitos del conocimiento. La Lingüística de Corpus encontró en la web la llave para la construcción de corpus de mayor tamaño y accesibilidad, lo que desembocaría en la última y más reciente etapa de esta disciplina, que adoptó por nombre la expresión *The web as corpus* (la web como corpus). Por su parte, la Lingüística Computacional vio en la web la oportunidad de resurgir con fuerza de las dificultades por las que había pasado unos años atrás y comenzó así una nueva andadura más allá de las aplicaciones básicas con material textual, en la que la mejora de las técnicas de etiquetado, de reconocimiento de voz, de recuperación de información o de traducción automática, así como el procesamiento del lenguaje natural (PNL) ocupan un lugar central.

Los resultados del trabajo entre la Lingüística Computacional y los corpus ha dado lugar a avances científicos en los que la gestión de la información ha sido la mayor

beneficiada, debido al desarrollo de la capacidad de almacenamiento, de análisis y de velocidad de respuesta.

A pesar de tratarse de una de las principales líneas de investigación en la actualidad y uno de los sectores en los que más recursos se están invirtiendo a nivel mundial, hemos podido constatar que, en los estudios sobre el lenguaje, no se había planteado hasta el momento la posibilidad de trabajar en este ámbito, con la colaboración imprescindible, claro está de la Lingüística Computacional. Esta ha sido nuestra principal motivación a la hora de emprender una investigación de estas características.

Desde el primer momento, nos hemos propuesto presentar un trabajo innovador que tuviera la capacidad de aportar una nueva perspectiva a los estudios lingüísticos y que ofreciera a los científicos la posibilidad de emprender investigaciones con una metodología novedosa y útil.

Para validar la hipótesis, recordamos las dos ideas en las que esta se sustentaba al principio de nuestro trabajo.

3. La utilización de *big data* –y, en concreto, de la información contenida en *Twitter*– asistida por las tecnologías incorporadas a su manipulación, supone una mejora en la investigación lingüística, ya que enriquece las características de la metodología tradicional referidas al volumen de datos, al tiempo de recopilación y al tiempo de procesamiento de la información.

Esta realidad ha quedado demostrada a partir de los distintos ejemplos de estudios llevados a cabo a través de la herramienta en la parte de aplicación práctica del trabajo. No solo *big data*, sino también 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 los usuarios. Este hecho resulta de gran utilidad para la investigación lingüística, ya que tenemos la posibilidad de acceder a las producciones textuales de los usuarios sean cuales sean su idioma, su sexo, su edad o su situación social. Además, el enorme número de personas que se comunican y que nos aportan ejemplos lingüísticos reales para la investigación hace que los resultados extraídos estén basados en millones

de muestras, lo que aporta veracidad a la investigación. Por otro lado, es fundamental tener en cuenta que una parte importante de los tuits que se publican diariamente son expuestos en la televisión o comentados en los medios de comunicación. Esto implica que el tipo de lenguaje que se utiliza en esta plataforma tiene la capacidad de llegar más allá de las cuentas de los propios usuarios.

Los materiales textuales que hemos extraído a partir de *Twitter* están en la línea de las predicciones que Svartvik (1992) hacía hace más de veinte años, cuando vaticinaba que los corpus del futuro serían mucho mayores que sus predecesores y que este aumento de tamaño sería posible gracias a los desarrollos de nuevos *hardware* y a la creación de *software* más cercanos al lingüista, lo que provocaría recopilaciones de textos en forma de monitor corpus, es decir, corpus dinámicos, sin un número límite de palabras y en continuo movimiento que pudieran ser analizados en tiempo real. También apuntaba Svartvik que la línea entre la gramática y el léxico se difuminaría y que el énfasis en el discurso hablado y en las técnicas de Lingüística Computacional afectarían a uso de los corpus. Por otra parte, creemos también haber demostrado la última de las predicciones de este lingüística, mediante la cual afirmaba que el fortalecimiento de la Lingüística de Corpus estaría ligado a un "ambiente teórico más liberal" (Svartvik, 1992: 11), así como a un acceso más fácil, barato y económico a los programas informáticos, debido al desarrollo técnico experimentado por los ordenadores, y a la disponibilidad de crecientes fuentes textuales electrónicas y *online*.

En otro orden de cosas, y a la luz de los resultados obtenidos en los distintos estudios llevados a cabo, consideramos que el análisis del lenguaje a través de *Twitter* permite obtener conclusiones correctas por varios motivos. En primer lugar, como ya hemos comentado, el elevado número de usuarios que se comunican a través de esta plataforma hace que los resultados reflejen en gran medida la realidad de las lenguas. Por otra parte, como hemos podido comprobar, el lenguaje utilizado *Twitter*, muchas veces a medio camino entre el registro oral y el escrito no es sino un reflejo del uso que los hablantes hacen de su propio idioma. Los resultados reflejan que el español que se utiliza en *Twitter*, en contra de lo que se pueda pensar en un primer momento, es el mismo que se utiliza en otros foros. Al igual que ocurre en otros contextos, hay una parte del lenguaje más cuidada y otra que obedece más a un diálogo espontáneo o a una intervención puntual, a la que se le presta menos atención. Es decir, *Twitter* nos permite utilizar distintos registros, dependiendo del objetivo del proceso comunicativo y el contexto de producción lingüística. Los errores que se puedan ver en esta plataforma,

por tanto, son un reflejo de los errores que esos mismos hablantes cometen al margen de las redes sociales. La economía del lenguaje que en algunas ocasiones nos exige el límite de los 140 caracteres obliga a los usuarios a utilizar técnicas y recursos que les permitan expresarse con claridad. Sin embargo, esta forma de utilizar el lenguaje no es algo reciente ni consecuencia de las limitaciones de espacio de los medios técnicos. Recordemos, por ejemplo, el origen de la letra *ñ*, que evolucionó a partir de la unión de la doble *n* a raíz de los esfuerzos de los escribas por ahorrar espacio en el pergamino, o los frescos de las iglesias en los que se escribían textos, muchos siglos más tarde.

En cualquier caso, lo que en este trabajo nos proponemos no es demostrar y analizar el uso normativo y correcto del lenguaje; ni tampoco defender la ortodoxia o no del lenguaje utilizado en *Twitter*. Nuestra intención, por el contrario, no es otra que estudiar de forma descriptiva la manera en la que los hablantes se expresan mediante esta plataforma –que cuenta con miles de millones de usuarios y ejerce una fuerte influencia en la sociedad de hoy en día– y de qué manera pueden influir los distintos comportamientos en la evolución de las lenguas. David Crystal expresa con claridad esta idea cuando afirma que:

> Language is indeed «at the heart of the Internet» and as a «social fact», rather than simply a «technological fact», where «the chief stock-in-trade is language», the web may paradoxically have been brought to the attention of many linguists as the largest text collection in the world almost against their will (Crystal, 2006: 271).

Por otra parte, la herramienta no solo lleva a cabo análisis del lenguaje general, sino que también nos permite basar nuestras investigaciones en corpus más especializados mediante el módulo *Wordics One*, lo que posibilita, como hemos comprobado en los estudios correspondientes a esta parte, centrarnos en lenguajes más específicos, como es el caso del lenguaje periodístico o del político.

En palabras de Armonk (2015), para IBM: "*Twitter* is like no other data source in the world. It is a real-time, public, conversational and global information platform where voices from around the world are speaking about every topic imaginable". Nuestra intención es sacar el máximo provecho de esta fuente de información para obtener beneficios en la investigación lingüística.

Sin embargo, dadas las características de esta información y el volumen de datos disponible, sería impensable emprender la labor investigadora sin el apoyo de un soporte informático que sea capaz de gestionarla. Por ello, consideramos que también queda probada la afirmación contenida en la hipótesis que asegura que esta nueva metodología que presentamos debe estar asistida por las tecnologías oportunas. Estudiar el lenguaje y las lenguas a partir de los datos que *Twitter* genera no solo supone una novedad por el hecho de cómo se produce y se transmite la información, sino que requiere una serie de especificaciones técnicas que sean capaces de ajustarse a la naturaleza de esta información para poder analizarla. Por este motivo, como lingüistas, necesitamos una herramienta que obtenga, almacene, analice y devuelva todo el material que posteriormente vamos a utilizar como investigadores del lenguaje.

Puesto que lo más habitual es que los científicos del campo de la Lingüística sean expertos en este ámbito y no dominen con maestría las técnicas informáticas, la relación entre estas dos disciplinas –Lingüística e Informática– se vuelve absolutamente necesaria, por un lado, y enriquecedora, por otro. Como ya hemos hablado en capítulos anteriores, una ciencia necesita de la otra y viceversa para que el conocimiento avance en la línea en que se mueve la sociedad de hoy en día.

En este trabajo, la necesidad de trabajar con una herramienta específica surge en el momento en el que pretendemos demostrar el inmenso potencial contenido en *big data* para los estudios lingüísticos.

Es precisamente en este punto donde más dificultades hemos encontrado a lo largo del trabajo. Por un lado, como ya hemos explicado, ha sido necesario un intercambio de comunicación constante con el desarrollador informático que se ha encargado de codificar la herramienta. No ha sido fácil cumplir con las peticiones y los requisitos que, desde un punto de vista lingüístico, hemos solicitado para la construcción de la herramienta; fundamentalmente, debido a las limitaciones propias de *big data* de las que hemos hablado en profundidad en el segundo capítulo. Sin embargo, gracias a numerosas pruebas y reestructuraciones a lo largo de un largo proceso, consideramos que hemos obtenido una herramienta de gran utilidad en el mundo de la Lingüística y que presenta una enorme versatilidad a la hora de diseñar los diversos estudios que se pretendan llevar a cabo.

Consideramos que la versión final de *Wordics Suite* respeta los principios sobre los que se asienta su construcción; desde el primer momento pretendimos que se tratara de una herramienta accesible, dinámica y sencilla de utilizar. Además, creemos que los

resultados obtenidos con ella confirman, de nuevo, la idea contenida en la primera parte de nuestra hipótesis referida al enriquecimiento de las posibilidades de la metodología tradicional, atendiendo a tres factores distintos:

1. Volumen de los datos
2. Tiempo de recopilación
3. Procesamiento de la información

La información recogida por la herramienta hasta el momento de redacción de estas líneas ha alcanzado un volumen de 873 GB, como ya hemos explicado. Es obvio, que realizar un acopio de información de este volumen (recordemos que es equivalente a casi dos millones de libros de unas trescientas páginas) con las técnicas tradicionales habría supuesto una inversión de tiempo y dinero, así como de espacio de almacenaje prácticamente imposible de asumir. El tercer factor característico, relacionado con el procesamiento de la información es también de suma importancia, puesto que plantear un estudio tradicional, sin la intervención de los medios técnicos, y basado en tal cantidad de material textual es inimaginable.

4. La metodología de trabajo lingüístico con *big data* incorpora nuevas dimensiones de análisis de textos, como son el etiquetado temporal y la geolocalización, lo que permite la realización de estudios sobre el lenguaje y las lenguas que no había sido posible llevar a cabo hasta el momento.

La segunda parte de nuestra hipótesis da un paso más en lo que a los beneficios de la metodología con *big data* se refiere. Los estudios realizados no solo han servido para demostrar que se pueden mejorar los resultados de la investigación atendiendo a las variables de estudio tradicionales, como son el tamaño de la muestra, el tiempo de recopilación de información o el tipo de procesamiento que se realiza sobre ella, sino que, gracias a la información que aporta *Twitter* y a la herramienta, podemos añadir dos variables más: el etiquetado temporal y la localización geográfica.

La primera implicación que se deriva de estas mejoras es la posibilidad de conocer en qué momento exacto se produce un determinado texto. En segundo lugar, podemos añadirle a este factor temporal la geolocalización, lo que nos permite ubicar

con casi total exactitud el punto geográfico en el que se publican los tuits. Gracias a este hecho, es posible determinar la relación causal de los comportamientos lingüísticos porque no solo tenemos información acerca de cuándo empieza a usarse un determinado término o expresión, cuál es frecuencia de uso y en qué momento deja de utilizarse, sino que podemos saber dónde se produce todo esto, lo que nos ayudará a entender las posibles causas que lo provoquen. Además, la posibilidad de ubicar los tuits en tiempo real supone uno de los aspectos más novedosos de este trabajo, puesto que no solo permite analizar la naturaleza dinámica de las lenguas, sino conocer esa evolución al mismo tiempo que se está produciendo, algo impensable con los corpus clásicos.

Estos dos factores, por tanto, amplían enormemente el campo de trabajo y las perspectivas de la investigación porque el incremento en el volumen y en la variedad de la información nos permite seleccionar una mayor cantidad de variables sobre las que elaborar nuestros estudios. Así lo expresa Kilgarriff en su histórico artículo sobre la web como corpus:

> As corpus linguists, we are in the fortunate position of having a particular perspective and channel of attack for examining the web –perhaps the most extraordinary phenomenon of our time– which also just happens to provides solutions to many of our practical problems and an endless stream of new data (Kilgarriff, 2001: 243).

Por otra parte, consideramos que estas características ofrecen mejoras para la investigación lingüística desde el momento en el que superan las limitaciones de la metodología clásica, que ofrece un evidente carácter estático y, a su vez, se encuentran encerradas en los límites temporales marcados por el momento histórico de la elaboración del corpus.

Por todos estos motivos, consideramos que hemos alcanzado el objetivo general que nos marcamos al inicio de esta investigación, cuya intención principal consistía en demostrar la doble vertiente de nuestra hipótesis. Como consecuencia de esto, creemos haber cumplido también con el resto de objetivos específicos, centrados en: a) la creación de una herramienta que permitiera trabajar con la información textual de *Twitter* y que aportara beneficios al trabajo del lingüista, b) el desarrollo de diversos estudios que demostraran algunas de las aplicaciones que nos ofrece esta nueva

metodología de investigación y c) la demostración de la conveniencia de la unión entre la Lingüística y la Informática.

Creemos, además, haber superado las expectativas en lo que al primer objetivo específico se refiere. Desde el primer momento, pretendimos que la herramienta se caracterizara por la facilidad y la sencillez a la hora de su utilización por parte del lingüista o de cualquier tipo de usuario que se enfrentara a ella, por un lado; por otro lado, la accesibilidad rápida, fácil y cómoda era otro de nuestros requisitos para la elaboración de nuestra herramienta. Desde el punto de vista informático, este aspecto quedó resuelto mediante el recurso de utilizar una aplicación web a la que se pudiera acceder desde cualquier ordenador con acceso a Internet. Este formato le aporta una serie de ventajas con respecto a una aplicación estándar o de tipo *standalone* (es decir, un programa independiente específico que necesita ser instalado previamente a su uso en un ordenador o dispositivo electrónico).

Por otra parte, el amplio abanico de posibilidades que nos ofrecía la información de *Twitter* y los recursos informáticos han hecho posible que la herramienta presentara un mayor rango de aplicaciones de las que en un principio nos planteamos. Este hecho derivó en la creación de los tres primero módulos de análisis de la *Wordics Suite* y en la especialización de cada uno de ellos en una técnica de análisis determinada. La elaboración del último módulo, *Wordics Data*, tiene como única finalidad la obtención fácil y directa de los corpus generados en los módulos anteriores.

La eficiencia conseguida con la herramienta, el amplio número de aplicaciones que se derivan ella y la enorme diversidad de estudios que permite nos ha llevado a superar las expectativas generadas en un primer momento y a confirmar la conveniencia y los beneficios de *big data* en nuestro campo de estudio.

No obstante, debemos admitir que el sistema que aquí presentamos no es perfecto, como no lo es ningún otro conjunto de datos al que podamos acceder. Aunque es innegable el hecho de que no todos los hablantes de un determinado idioma o de una región geográfica específica participan en *Twitter*, también es cierto que esta plataforma le ha dado voz a gran parte de la población que antes, bien por procedencia, por estatus social o, simplemente por falta de medios, no tenía la capacidad de expresarse públicamente y dejar evidencia de ello. Los datos lingüísticos que nos brinda *Twitter* permiten a los investigadores de cualquier ámbito abordar toda una serie de estudios con más posibilidades que la metodología tradicional.

Las limitaciones que hemos encontrado a lo largo de este camino son, fundamentalmente, técnicas, y están relacionadas con las restricciones que impone el propio servicio de *Twitter* a la hora de aportar la información solicitada.

En primer lugar, debemos señalar los problemas a la hora de la detección del idioma en el que se escriben algunos tuits. Como hemos podido observar en los estudios, el sistema devuelve un cierto número de publicaciones que etiqueta como pertenecientes a "idioma indeterminado". Este hecho puede provocar que perdamos información contenida en esos tuits debido a que el propio texto que se envía no es lo suficientemente concluyente como para que *Twitter* sea capaz de identificarlo y clasificarlo en un idioma concreto. En este sentido, tampoco tienen los sistemas de *Twitter* la capacidad de distinguir, por ejemplo, el catalán o el valenciano del español, lo que puede suponer un impedimento a la hora de estudiar estos idiomas individualmente. Sin embargo, *Twitter* está avanzando en este sentido y ampliando el soporte para los diferentes idiomas; dado que está basado en la norma ISO 639-1, la herramienta detectará de forma automática los nuevos idiomas que se vayan implementando.

En segundo lugar, nos enfrentamos también a algunas restricciones relacionadas con la cantidad de información que *Twitter* devuelve para que esta sea analizada por la herramienta. Actualmente, dado que *Twitter*, en su modelo de explotación da acceso a un porcentaje limitado de los tuits, podemos acceder a una ratio máxima de 52 tuits por segundo. No obstante, en el caso de que, en un momento determinado, se requiriera una mayor cantidad de información, sería posible realizar una subscripción mediante un contrato de acceso pagado a la información contenida en *Twitter*, para que la herramienta pudiera devolvernos la totalidad de los tuits publicados. Este problema solo lo encontramos en *Wordics Live* y en *Wordics Archive*, mientras que el inconveniente que presenta *Wordics One* es que se nos restringe el acceso a los –aproximadamente– 3.500 últimos tuits de una cuenta concreta. La superación de esta limitación pasaría por la misma solución que en el caso anterior.

El resto de las limitaciones que podemos encontrar son inherentes al trabajo con big data; es decir, la gestión del volumen de información y de la reducción de la velocidad de proceso y gestión de esta información debido a su gran tamaño. Sin embargo, debido a que la herramienta se ha diseñado específicamente para trabajar con *big data*, su funcionamiento es óptimo, puesto que los posibles problemas técnicos se han evitado gracias a la utilización de las técnicas clásicas de escalado de sistemas de proceso de *big data*. Si el volumen de la información aumentara drásticamente en el

futuro, la herramienta está preparada para distribuirse en varios servidores en paralelo y solventar el problema de la velocidad de respuesta.

En cualquier caso, no debemos olvidar que *Wordics Suite* ha sido creada y diseñada para permitir el análisis del lenguaje y de la lengua a través de *Twitter* y demostrar que, efectivamente, esta metodología es posible. Se trata, pues, de una herramienta dinámica y abierta a futuras mejoras que amplíen las posibilidades de la investigación. Nuestra intención con este trabajo es aportar una contribución que siente las bases de futuras investigaciones en el ámbito de las ciencias del lenguaje, basadas en *Twitter*, como parte de *big data*, y que sea de interés y utilidad para cualquier investigador interesado en este ámbito.

Bibliografía

Aarts, J. (1991). Intuition-based and observation-based grammars. En Aijmer, K. & Altemberg, B. (Eds.), *English Corpus Linguistics* (pp. 44-65). London: Longman.

Abaitúa, J. (2002). Tratamiento de corpora bilingües. En Martí, M. A. & Llisterri, J. (Eds.), *Tratamiento del lenguaje natural: tecnología de la lengua oral y escrita* (pp. 61-90). Soria/Barcelona: Fundación Duques de Soria/Edicions de la Universitat de Barcelona.

Abercrombie, D. (1965). *Studies in Phonetics and Linguistics*. London: Oxford University Press.

Adamic, L. A., & Adar, E. (2005). How to search a social network. *Social Networks,* 27(3),187–203.

Agirre, E., Ansa, O., Hovy, E. & Martínez, C. (2000). Enriching very large ontologies using the WWW. En *Proceeding of the Ontology Learning Workshop of the European Conference of AI (ECAI)*, Berlin, Germany.

Aijmer, K. & Altemberg. B (Eds.). (2004). *Advances in Corpus Linguistics: Papers from ICAME 23, Göteborg 22-26 May 2002 (The Theory and Use of Corpora)*. Amterdam: Rodopi

Aijmer, K. & Altenberg, B. (Eds.). (1991). *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. New York: Longman.

Allen, J. D. *et al*. (Ed.). (2006). Appendix C: Relationship to ISO/IEC 10646. Westford: Unicode, Inc. Recuperado de http://www.unicode.org/versions/Unicode5.0.0/appC.pdf

Altenberg, B. (Ed.) (1986). *Newsletter of the International Computer Archive of Modern English*, 10.

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

Alvar Ezquerra, M. & Corpas Pastor, G. (1994). Criterios de diseño de un corpus europeo. En Alvar Ezquerra, M. y Villena Ponsoda, J. A. (Coords.), *Estudios para un corpus del español* (pp. 31-40). Málaga: Servicio de Publicaciones de la Universidad.

Alvar Ezquerra, M. Blanco Rodríguez, M. J. & Pérez Lagos, F. (1994). Diseño de un corpus español en el marco de un corpus europeo. En Alvar Ezquerra, M. y J. A.

Villena Ponsoda, J. A. (Coords.), *Estudios para un corpus del español* (pp. 9-29). Málaga: Servicio de Publicaciones de la Universidad.

Ammu, N. & Irfanuddin, M. (2013). Big Data Challenges. *International Journal of Advanced Trends in Computer Science and Enfineering, Special Issue of ICACSE*, 2(1), 613-615.

Anthony, L. (2012). *AntConc* (Version 3.3.5) [Computer Software]. Tokyo, Japan: Waseda University. Recuperado de http://www.antlab.sci.waseda.ac.jp/

Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, 30(2), 141-161,

Araujo, M. H. & Melo, S. (2003). Del caos a la creatividad: los chats entre lingüistas y didactas. En López Alonso, C. & Séré, A. (Eds.), *Nuevos géneros discursivos: los textos electrónicos* (pp. 45-61). Madrid: Biblioteca nueva.

Ashton, K. (2009). That "Internet of Things" Thing. *Rfid Journal*, 1. Recuperado de http://www.rfidjournal.com/articles/pdf?4986

Aston, G., Bernardini, S. & Stewart, D. (Eds.). (2004). *Corpora and Language Learners*. Amsterdam: Benjamin.

Asur, S. & Huberman, B. (2010). Predicting the future with Social Media. *Proceedings of the 2010 IEEE/WIC/ACM International Conference of Web Intelligence and Intelligent Agent Technology*, 1 (pp. 492-499). Washington, DC: IEE Computer Society.

Atkins, S., Clear, J. & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*, 7(1), 1–16.

Auger, P. & Rousseau, L. J. (2003). *Metodología de la Investigación Terminológica*. Edición y traducción de Guerrero Ramos, G. & Bermúdez Fernández, J. M. Málaga: Universidad de Málaga.

Austin, J. L. (1975). *How to do things with words.* Urmson, J. O. & Sbisà, M. (Eds.). Cambridge, MA: Havard University.

Ayala Castro, M. C., Guerrero Salazar, S. & Medina Guerra, A. (2006). *Guía para un uso igualitario del lenguaje periodístico*. Málaga: Diputación.

Baker, C. (2001). *Foundations of Bilingual Education and Bilingualism*, 3rd edition. Clevedon/Buffalo/Toronto/Sydney: Multilingual Matters Ltd.

Baker, M., Francis, G., Tognini-Bonelli, E., & Sinclair, J. (Eds.). (1993). *Text and technology*. Philadelphia: J. Benjamins Pub. Co.

Banko, M., & Brill, E. (2001). Scaling to very very large corpora for natural language

disambiguation. *Proceedings of 39th Annual Meeting on ACL-01* (pp. 26-33). Strousburg, PA: ACM.

Baroni, M. & Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. *Proceedings of LREC 2004* (pp. 1313-1316). Lisbon: ELDA.

Baroni, M. & Bernardini, S. (Eds.). (2006). *Wacky! Working Papers on the Web as Corpus*. Bologna: Gedit. Recuperado de http://wackybook.sslmit.unibo.it/

Baroni, M. & Kilgarriff, A. (2006). Large linguistically-processed Web corpora for multiple languages. En Baroni, M. & Kilgarriff, A. (Eds.), *Proceedings of the 2nd International Workshop on Web as Corpus (EACL06)* (pp.87-90). Trento, Italy.

Baroni, M. & Ueyama, M. (2006). Building general -and special- purpose corpora by Web crawling. *Proceedings of the NIJL International Symposyum, Language Corpora: Their Compilation and Application* (pp.31-40). Recuperado de http://home.sslmit.unibo.it/~baroni/publications/bu_wac_kokken_formatted.pdf

Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation,* 43(3), 209-226.

Barrett, M. A., Humblet, O., Hiatt, R. & Adler, N. (2013). Big data and disease prevention. From Quantified Self to Quantified Communities. *Big Data*, 1(3), 168-175. doi: 10.1089/big-2013.0027.

Baru, C., Bhandarkar, M., Nambiar, R., Poess, M. & Rabl, T. (2013). Benchmarking Big Data Systems and the Bigdata top 100 list. *Big Data*, 1(1), 60-64. doi: 10.1089/big.2013.1509.

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

Becker, H., Naama, M. & Gravano, L. (2011). Beyond Trending Topics: Real-World Event Identification on Twitter. *Proceedings of the Fifth International Conference on Weblogs and Social Media* (pp. 438-441). Recuperado de http://www.cs.columbia.edu/~gravano/Papers/2011/icwsm11-a.pdf

Bengoechea Bartolomé, M. (2015). *Lengua y género*. Síntesis: Madrid.

Bergman, M. & Paavola, S. (Eds.). (2003). The commens dictionary of Peirce's terms. Peirce's terminology in his own words. Recuperado de http://www.commens.org/dictionary/term/token

Bernardini, S., Baroni, M. & Evert, S. (2006). A WaCky Introduction. En Baroni, M. y Bernardini, S. (Eds.), *Wacky! Working Papers on the Web as Corpus* (pp.9-40). Bologna: GEDIT.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing* 8(4), 243-257.

Biber, D. (1998). *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, D. & Finegan, E. (1992). The linguistic evolution of five written and speech-based genres from the 17th to the 20th centuries. En Rissanen, M., Ihalainen, O., Nevalainen, T: & Taavitsainen, I. (Eds.), *History of Englishes: New methods and interpretations in historical linguistics* (pp. 688-704). Berlin/New York: Mouton.

Biber, D. & Kurjian, J. (2007). Towards a taxonomy of web registers and text types: a multidimensional analysis. En Hundt, M., Nesselhauf, N., & Biewer, C. (Eds.) *Corpus linguistics and the web* (pp.109-132). Amsterdam: Rodopi.

Biber, D., Reppen, R., Clarck, V. & Walter, J. (2001). Representing spoken language in university settings: The design and construction of the spoken component of the T2K-SWAL Corpus. En Simpson, R., Swales, J. (Eds.), *Corpus Linguistics in North America* (pp. 48-57). Ann Arbor: University Michigan Press.

Bilbao-Osorio, B., Dutta, S. & Lanvin, B. (Eds.). (2014). *The Global Information Technology Report 2014. Rewards and Risks of Big Data*. Geneva: World Economic Forum.

Borau, K., Ullrich, C., Feng, J. & Shen. R. (2009). Microblogging for Language Learning: Using Twitter to Train Communicative and Cultural Competence. En Spaniol, M. *et al*. (Eds.), *ICWL 2009* (pp.78-87). Berlin: Springer-Verlag Berlin Heidelberg.

Bourigault, D., Jacquemin, C. & L'Homme, M. C. (Eds.). (2001). *Recent Advances in Computational Terminology*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Bowker, L. & Pearson, J. (2002). *Working with Specialized Language. A practical guide to using corpora*. London/New York: Routledge.

Broniatowski, D.A., Paul, M.J. & Dredze, M. (2013). National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. *PLoS ONE,* 8(12): e83672. doi:10.1371/journal.pone.0083672.

Bruns, A. & Burgess, J. (2011). The Use of Twitter Hashtags in the formation of Ad Hoc Publics. *Proceedings of the 6th European Consortium for Political Research, General Conference, 2001. University of Iceland, Reykjavik*. Recuperado de http://eprints.qut.edu.au/46515/

Brynjolfsson E., Hitt L.M., & Kim H.H. (2011). Strength in numbers: How does data-driven decision making affect firm performance? *Social Science Research Network*. Recuperado de http://papers.ssrn.com/sol3/papers.cfm?abstract_id=%201819486

Bud, R. (2003). Introduction to elementary probabilistic theory and formal stochastic language theory. En Bod, R., Hay, J. & Jannedy, S. (Eds.), *Probabilistic Linguistics* (pp. 11-37). Londres: MIT Press.

Buyse, K. (2010). ¿Qué corpus en línea utilizar para qué fines en la clase de ELE? *XXI Congreso Internacional de la ASELE. Del texto a la lengua: la aplicación de los textos a la enseñanza-aprendizaje del español L2-LE, Salamanca* (pp. 277-288). Recuperado de http://cvc.cervantes.es/ensenanza/biblioteca_ele/asele/asele_xxi.htm

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

Cabré, M. T. *et al*. (2002). Evaluación de la vitalidad de una lengua a través de la neología: a propósito de la neología espontánea y de la neología planificada. En Cabré, M. T., Freixa, J. & Solé, E. (Eds.), *Lèxic i neología* (pp. 159-201). Barcelona: Universitat Pompeu Fabra.

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

Cabré, M. T., Estopà, R. & Vivaldi, J. (2001). Automatic term detection: a review of current systems. En Bourigault, D., Jacquemin, C. & L'Homme, M. C. (Eds.), *Recent Advances in Computational Terminology* (pp.53-87). Amsterdam/Philadelphia: John Benjamins.

Cafarella, M., Halevy, A. & Khoussainova, N. (2009). Data Integration for the Relational Web. *Proceedings of the VLDB Endowment,* August 2009, 2(1) (pp. 1090-1101). doi: 10.14778/1687627.1687750.

Calero Fernández, M. A. (1999). *Sexismo lingüístico: análisis y propuestas ante la discriminación sexual en el lenguaje*. Madrid: Narcea.

Calero Vaquera, M. L. (2003a). *Guía de estilo 1: Lengua y discurso sexista* (En colaboración con M. Lliteras y Mª Á. Sastre) pp. 155-230. Junta de Castilla y León, Valladolid.

Calero Vaquera, M. L. (2003b). *Guía de estilo 2: Sexismo y redacción periodística* (En colaboración con M. Lliteras y M. Bengoechea) pp. 131-206. Junta de Castilla y León, Valladolid.

Calero Vaquera, M. L. (2014). El discurso del Whatsapp: entre el Messenger y el SMS. *Oralia: Análisis del discurso oral*, 17, 87-116.

Calvo Revilla, A. M. (2002). Cambios lingüísticos ante el proceso de innovación tecnológica de la comunicación digital. Espéculo. *Revista de estudios literarios (UCM)*. Recuperado de http://pendientedemigracion.ucm.es/info/especulo/numero20/digital.html

Carter, P. (2011). Big data analytics: Future Arquitectures, Skills and Roadmaps for the CIO. *White Paper, IDC* (pp.1-15). IDC: Go-To-Market Services. Recuperado de: http://www.sas.com/resources/asset/BigDataAnalytics-FutureArchitectures-Skills-RoadmapsfortheCIO.pdf

Catlett, C. & Ghani, R. (2015). Big Data for Social Good. *Big Data*, 3(1), 1-2. doi: 10.1089/big.2015.1530.

Cerny, J. (2006). *Historia de la lingüística*. Cáceres: Universidad de Extremadura.

Chafe, W. (1992). The importance of corpus linguistics to understanding the nature of language. En Svartvik, J. (Ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991* (pp. 79-97). Berlin/New York: Mouton de Gruyter.

Chafe, W. (1994). *Discourse, consciousness and time*. Chicago: The University of Chicago Press.

Chakrabarti, S. (2003). *Mining the Web. Discovering Knowledge from Hypertext Data*. San Francisco: Morgan Kaufmann.

Chakrarbati, S. (1999). Hypersearching the Web. *Scientific American*, 280(6), 54-60. Recuperado de http://econ.tepper.cmu.edu/ecommerce/hypersearch.pdf

Chen, M., Mao, S., Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications*, 19(2), 171-209. doi: 10.1007/s11036-013-0489-0.

Cheng, W. (2012). *Exploring corpus linguistics*. London: Routledge.

Chklovski, T. and Pantel, P. (2004). VerbOcean: Mining the Web for fine-grained semantic verb relations. *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, 33-40. Barcelona, Spain.

Cho, J. & García-Molina, H. (2002). Parallel Crawlers. *Proceedings of the 11th international conference on World Wide Web*, Honolulu, Hawaii (pp. 124-135). New York: ACM.

Chomsky, N. (1970). *Aspects of the theory of syntax*. Cambridge, MA: MIT.

Chomsky, N. (1971). *Syntactic Structures*. The Hague/Paris: Mouton.

Church, K. W. & Mercer, R. (1993). Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics*, 19(1), 1-24.

Corpas Pastor, G. (2001). Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. *TRANS. Revista de traductología*, 5, 155-184.

Corpas Pastor, G. (2002). Traducir con corpus: de la teoría a la práctica. En García Palacios, J. & Fuentes, M. T. (Eds.), *Texto, terminología y traducción* (pp. 189-226). Salamanca: Almar.

Corpas Pastor, G. (2012). Corpus, tecnología y traducción. En Casas Gómez, M. & García Antuña, M. (Coords.), *XII Jornadas de Lingüística*, 2, 75-98.

Crystal, D. (2000). *Diccionario de lingüística y fonética*, traducción y adaptación de X. Villalba. Barcelona: Octaedro.

Crystal, D. (2006). *Language and the Internet*. Cambridge: Cambridge University Press.

Crystal, D. (2011). *Internet linguistics*. Abingdon, Oxon: Routledge.

Culnan, M. J., McHugh, P. J. & Zubillaga, J. I. (2010). How large US Companies can use Twitter and Other social Media to Gain Business Value. *MIS Quarterly Executive*, 9(4), 243-259.

Cummis, J. & Swain, M. (1998). *Bilingualism in Education: Aspects of Theory, Research and Practice (6th. Impression*). New York: Routledge.

Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, 24/2, 197-222.

De Groc, C. (2011). Babouk: Focus web crawling for corpus compilation and automatic terminology extraction. En *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 1, 497-498. Washington DC: ACM.

De Kock, J. (1990). *Gramática española: Enseñanza e investigación I. Gramática, apuntes metodológicos*. Salamanca: Ediciones Universidad de Salamanca.

De Kock, J. (2001). *Gramática española. Enseñanza e investigación. Lingüística con corpus, catorce aplicaciones sobre el español*. Salamanca: Ediciones Universidad de Salamanca.

De Schryver, G.M. (2002). Web for / as corpus: a perspective for the African languages. *Nordic Journal of African Studies*, 11, 266–282.

Dhar, V. (2012). *Working paper CdCER-12-01*. New York University, Leonard N. Stern School of Business. Recuperado de https://archive.nyu.edu/bitstream/2451/31553/2/Dhar-DataScience.pdf

Dhar, V. (2014). Why Big Data = Big Deal. *Big Data*, 2(2), 55-56. doi: 10.1089/big.2014.1522.

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

Díaz Hormigo, M. T. (2004b). Neología y tecnología: a propósito de los programas de detección automática de neologismos. *Español Actual*, 82, 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, 5-60.

Díaz Hormigo, M. T. (2010). Revisión historiográfica de los conceptos "neología" y "neologismo". En Assunção, C., Fernandes, G. & Loureiro, M. (Eds.), *Ideias Linguísticas na Península Ibérica (séc. XIV a séc. XIX)*. Münster: Nodus Publikationen, I, 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. En *Revista de lingüística y lenguas aplicadas*, 10, 12-20.

Dipper, S. (2008). Theory-driven and corpus-driven computational linguistics, and the use of corpora. En Lüdeling, A. y Kytö, M. (Eds.), *Corpus Linguistics: an International Handbook, v. 1* (pp. 68-97). Berlin: Mouton de Gruyter.

Domínguez Burgos, A. (2002). Lingüística computacional: un esbozo. *Boletín de Lingüística*, 18, 104-109.

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

Dumais, S., Banko, M., Brill, E., Lin, J. & Ng, A. (2002). Web question answering: is more always better? En *Proceedings of the 25th ACM SIGIR*, 291–298, Tampere, Finland: ACM.

Dumbill, E. (2013). Making sense of Big Data. *Big Data*, 1(1), 1-3. doi: 10.1089/big.2012.1503.

Dumbill, E. (2013). The human face of big data: An Interview with Rick Smolan. *Big Data*, 1(1), 5-9. doi: 10.1089/big.2013.1511.

Dumbill, E., Liddy, E. D., Stanton, J., Mueller, K. & Farnham, S. (2013). Educating the next generation of data scientists. *Big Data*, 1(1), 21-28. doi: 10.1089/big.2013.1510.

Ebner, M., & Schiefner, M. (2008). Microblogging-more than fun? En Arnedillo Sánchez, I. & Isaías, P. (Eds.), *Proceeding of IAIDS Mobile Learning Conference 2008* (pp.155-159). Algarve, Portugal.

Ebner, M., Lienhardt, C., Rohs, M. & Meyer, I. (2010). Microblogs in Higher Education – A chance to facilitate informal and process-oriented learning? *Computers & Education*, 55, 92-100.

ECMA. (2013). Standard ECMA-404. The JSON data interchange format. Geneva: ECMA International. Recuperado de http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf

El-Darwiche, B., Koch, V., Meer, D., Shehadi, R. & Tohme, W. (2014). Big Data maturity: an action plan for policymakers and executives. En Bilbao-Osorio, B., Dutta, S. & Lanvin, B. (Eds.), *The Global Information Technology Report 2014. Rewards and Risks of Big Data* (pp. 43-52). Geneva: World Economic Forum.

Evert, S., Kigarrigg, A. & Sharoff, S. (2008). Can we beat Google? *Proceedings of the 4th Web as Corpus Workshop (WAC-4)*. Marrakech, 1 June 2008.

Eysenbach, G. (2011). Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact. *Journal of Medical Internet Research*, *13*(4), e123. http://doi.org/10.2196/jmir.2012.

Facchinetti, R. (Ed.). (2007). *Corpus Linguistics: 25 Years On*. Amsterdam/New York: Rodopi.

Fairon, C., Naets, H., Kilgarriff, A. & De Schryver, G. (Eds.). (2007). *Building and Exploring Web Corpora* (WASC3-2007). Louvain-la-Neuve: presses Universitaires de Louvain.

Fillmore, C. J. (1992). “Corpus linguistics” or “Computer-aided armchair linguistics”. En Svartvik, J. (Ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991* (pp. 35-60). Berlin/New York: Mouton de Gruyter.

Fletcher, W. (2004a). Facilitating the Compilation and Dissemination of Ad-Hoc Web Corpora. En Aston, G., Bernardini, S. & Stewart, D. (Eds.), *Corpora and Language Learners* (pp. 271–300). Amsterdam: Benjamin.

Fletcher, W. H. (2001). Concordancing the Web with KWiCFinder. *American Association for Applied Corpus Linguistics Third North American Symposium on corpus Linguistics and Language Teaching* (pp. 23-25). Boston, MA.

Fletcher, W. H. (2004b). Making the web more useful as a source for linguistic corpora. En Connor, U. & Upton, T. (Eds.), *Applied corpus linguistics: a multidimensional perspective* (pp. 191–205). Amsterdam: Rodopi.

Fletcher, W. H. (2007). Concordancing the web: promise and problems, tools and techniques. En Hundt, M., Nesselhauf, N., & Biewer, C. (Eds.) *Corpus linguistics and the web* (pp.25-46). Amsterdam: Rodopi.

Fletcher, W. H. (2012). Corpus Analysis of the World Wide Web. En Chapelle, C. (Ed.), *Encyclopedia of Applied Linguistics*. London: Wiley-Blackwell.

Francis, N. W. (1982). Problems of assembling and computerizing large corpora. En Johansson, S. (Ed.), *Computer corpora in English language research* (pp. 4-24). Bergen: Norwegian Computing Centre for the Humanities.

Francis, N. W. (1985). Dinner Speech. En Altenberg, B. (Ed.), (1986), *Newsletter of the International Computer Archive of Modern English*, 10, 5-7.

Francis, N. W. (1992). Language corpora B. C. En Svartvik, J. (Ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991* (pp. 17-34). Berlin/New York: Mouton de Gruyter.

Fries, C. (1973). *The Structure of English: an introduction to the construction of English sentences*. London: Longman.

Fuentes, M., Gerding, S., Constanza, S., Pecchi, A., Kotz, G. & Cañete, P. (2009). Neología léxica: reflejo de la vitalidad del español de Chile. RLA, Revista de lingüística teórica y aplicada, 47(1), 103-124. Recuperado de http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-48832009000100006&lng=es&tlng=es. 10.4067/S0718-48832009000100006

Fujii, A. & Ishikawa, T. (2000). Utilizing the world wide web as an encyclopedia: Extracting term descriptions from semi-structured text. En *Proceedings of the 38th Meeting of the Association for Computational Linguistics* (pp. 448-495). Hong Kong: ACM.

Galán Rodríguez, C. (2002). En los arrabales de la comunicación: los mensajes SMS. *Anuario de estudios filológicos*, 25, 103-117.

Galán, C. (2007). Cnct kn nstrs: los SMS universitarios. *Revista de Estudios de Juventud, Instituto de la Juventud (InJuve)*, 78, 63-73.

Gant, J. & Reinsel, D. (2012). The digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. *IDC*. Recuperado de https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

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

Ganz, J. & Reinsel, D. (2013). The digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East-United States. *IDC*. Recuperado de https://www.emc.com/collateral/analyst-reports/idc-digital-universe-united-states.pdf

García Ces, P. (2007). Terminología y terminótica en la formación de traductores e intérpretes en Argentina. *Panace@*, 8(26), 158-161.

García Meseguer, Á. (2001). ¿Es sexista la lengua española? *Panacea*, 2(3), 20-34.

Garside, R. (1987). The CLAWS Word-tagging System. En Garside, R., Leech, G. & Sampson, G. (Eds.), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Gatto, M. (2008). *From body to web: an introduction to the web as corpus*. Bari: Editori Laterza.

Gatto, M. (2014a). *The web as corpus: theory and practice*. London: Bloomsbury Academic.

Gatto, M. (2014b). *Corpus and discourse*. London: Bloomsbury Academic.

Gelbukh, A., Sidorov, G. & Chanona, L. (2002). Corpus virtual, virtual: Un diccionario grande de contextos de palabras españolas compilado a través de Internet. En Gonzalo, J., Peñas, A., Fernández, A. (Eds.), Multilingual Information Access and Natural Language Processing, International Workshop (November 12) at

IBERAMIA-2002, VII Iberoamerican Conference on Artificial Intelligence (pp. 7-14). Sevilla: IBERAMIA, ELSNET y RITOS.

Gens, F. (2014). IDC Predictions 2015: Accelerating Innovation – and Growth –on the 3rd Platform. *IDC*. 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

Giles, J. (2010). Cellphones reveal emerging disease outbreaks. *NewScientist*, 1782, 1.

Gold, M., McClarren, R. & Gaugham, C. (2013). The lessons Oscar taught us: Data Science and Media & Entertainment. *Big Data*, 1(2), 105-109. doi: 10.1089/big.2013.0009.

Gómez Guinovart, J. (1998). Fundamentos de Lingüística Computacional: bases teóricas, líneas de investigación y aplicaciones. En Baró i Queralt, J. & Cid Leal, P. (Eds.), *Anuari SOCADI de Documentació i Informació* (pp. 135-146). Barcelona: Societat Catalana de Documentació i Informació.

Gómez Guinovart, J. (2000). Perspectivas de la lingüística computacional. *Novatica*, especial 25 aniversario, 85-87.

Gómez Guinovart, J. (Ed.). (1999). *Panorama de la investigación en lingüística informática*. Logroño: Asociación Española de Lingüística Aplicada.

Gómez Torrego, L. (2001). La gramática en Internet. En *II Congreso Internacional de la Lengua Española. El español en la Sociedad de la Información*. Valladolid, 16-19 de octubre de 2001. Recuperado de http://congresosdelalengua.es/valladolid/ponencias/nuevas_fronteras_del_espanol/4_lengua_y_escritura/gomez_l.htm

González Fernández, Adela. (2015). Big Data as a Tool for Linguistic Research: Approaches to Trends in Bilingualism in Ten Latin-American Countries. *International Journal of Language and Applied Linguistics (IJLAL)*, special issue "Bilingual Education", 1, 1-12. Khate Sefid Press. ISSN: 2383-5014.

González Fernández, Adela. (2016). Análisis de las necesidades traductológicas en Europa a través de big data. *Skopos*, 7, 45-74. ISSN: 2255-3703 [en prensa].

Granovetter, M. S. (1973). The Strengh of Weak Ties. *American Journal of Sociology*, 78(6), 1360-1380.

Gray, J. (2009). eScience: A transformed scientific method. En Hey, T., Tansley, S. & Tolle, K. (Eds.), *The fourth paradigm. Data-intensive scientific discovery* (pp. XVII-XXXI). Redmond, Washington: Microsoft Research.

Greenbaum, S. (1991). Towards a new corpus of spoken American English. En Aijmer, K. & Altemberg, B. (Eds.), *English Corpus Linguistics* (pp. 83-94). London: Longman.

Greenbaum, S. (1992). A new corpus of English: ICE. En Svartvik, J. (Ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991* (pp. 171-179). Berlin/New York: Mouton de Gruyter.

Greenwood, M., Roberts, I. & Gaizauskas, R. (2002). University of sheffield trec 2002 q & a system. En Voorhees, E. M. & Buckland, L. P. (Eds.), *The Eleventh Text REtrieval Conference (TREC-11)*, Washington. U.S. Government Printing Office. NIST Special Publication.

Grefenstette, G. (1999). The WWW as a resource for example-based MT tasks. *Translating and the Computer: Proceedings of the 21st. International Conference on Translating and the Computer*. Londres: ASLIB.

Grefenstette, G. & Nioche, J. (2000). Estimation of English and non-English language use on the WWW. En Proceedings of the RIAO (Recherche d'Informations Assistée par Ordinateur), 237-246. Paris, 12-14, April 2000. Recuperado de http://arxiv.org/ftp/cs/papers/0006/-0006032.pdf

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

Gries, S. T. (2009). What is corpus linguistics? *Language and Linguistics Compass*, 3, 1–17.

Gries, S. T. (2010). Corpus linguistics and theoretical linguistics. A love-hate relationship? Not necessarily… *International Journal of Corpus Linguistics*, 15(3), 327-342.

Grishman, R. (1991 [1986]). *Introducción a la lingüística computacional*, traducción de A. Moreno Sandoval. Madrid: Visor.

Grosseck, G. & Holotescu, C. (2008). Can we use Twitter for educational purposes? *The 4th International Scientific Conference eLSE (eLearning and Software for*

Education), Bucharest. Recuperado de http://es.scribd.com/doc/2286799/Can-we-use-Twitter-for-educational-activities

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

Guerrero Salazar, S. (2006). El discurso sexista de los medios de comunicación. Cremades García, R. & Núñez Cabezas, E. A. (Coords.), *Lectura, escritura y comunicación* (pp. 81-106). Málaga: VG Ediciones.

Guerrero Salazar, S. (2007a). Alternativas al lenguaje sexista de los medios de comunicación. Novedades legislativas y otras actuaciones. En Loscertales, F. & Núñez, T. (Coords.), *La mirada de las mujeres en la sociedad de la información* (pp. 309-326). Madrid: Siranda editorial.

Guerrero Salazar, S. (2007b). *La creatividad en el lenguaje periodístico*. Madrid: Cátedra.

Guerrrero Salazar, S. (2000). Sexismo lingüístico en los medios de comunicación. *I, II ciclo y curso de conferencias abiertas. Aula de Formación Abierto 2000* (pp. 59-62). Málaga: SPICUM.

Guevara, E. (2010). NoWac: A large web-based corpus for Norwegian. *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, 1-7. Stroudsburg, PA: ACM.

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

Gupta, A. (2014). Making Big Data something more than the "Next Big Thing". En Bilbao-Osorio, B., Dutta, S. & Lanvin, B. (Eds.), *The Global Information Technology Report 2014. Rewards and Risks of Big Data* (pp. 87-94). Geneva: World Economic Forum.

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

Halliday, M. A. K. (1991). Corpus studies and probabilistic gramar. En Aijmer, K. & Altemberg, B. (Eds.), *English Corpus Linguistics* (pp. 30-43). London: Longman.

Halliday, M. A. K. (1992). Language as system and language as instance: The corpus as a theoretical construct. En Svartvik, J. (Ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991* (pp. 61-78). Berlin/New York: Mouton de Gruyter.

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

Halliday, M.A.K. (2007). Afterwords. En Hunston, S. & Thompson, G. (Eds.). (2007). *System and corpus: exploring connections*. London; Oakville, Connecticut: Equinox.

Hare, J. (2014). Bring it on, Big Data: Beyond the Hype. *Big Data*, 2(2), 73-75. doi: 10.1089/big.2014.1520.

Hassani, H., Saporta, Gl, Silva, E. S. (2014). Data Mining and Official Statistics: the Past, the Present and the Future. *Big Data*, 2(1), 34-44. doi: 10.1089/big.2013.0038.

Hendler, J. Broad Data: Exploring the emerging web of data. *Big Data*, 1(1), 18-21. doi: 10.1089/big.2013.1506.

Hey, T., Tansley, S. & Tolle, K. (Eds.). (2009). *The fourth paradigm. Data-intensive scientific discovery*. Redmond, Washington: Microsoft Research.

Higdon, R., Haynes, W., Stanberry, L., Stewart, E., Yandl, G., Howard, C., Broomall, W., Kolker, N. & Kolker, E. (2013). Unraveling the Complexities of Life Sciences Data. *Big Data*, 1(1), 42-51. doi: 10.1089/big.2012.1505.

Hogg, T., & Lerman, K. (2009). Stochastic models of user- contributory web sites. *Proceedings of the 3rd Intl Conf on Weblogs and Social Media (ICWSM2009)* (pp. 50-57). Recuperado de http://arxiv.org/pdf/0904.0016.pdf

Honey, C. & S. C. Herring. (2009). Beyond Microblogging: Conversation and Collaboration via Twitter. 42nd Hawaii International Conference on System Sciences, January 5-8, 2009. (pp. 1-10). Big Island: IEEE.

Honeycutt, C. & Herring, S.C. (2009). Beyond Microblogging: Conversation and Collaboration via Twitter. En *Proceedings of the 42nd Hawaii international Conference on System Sciences* (January 05 – 08, 2009) (pp. 1-10). Washington, CD: IEEE Computer Society.

Hu, Y., Talamadupula, K. & Kambhampati, S. (2013). Dude, srsly?: The Surprinsingly Formal Nature of Twitter's Language. *Proceedings of the 7th Annual International AAAI Conference on Weblogs and Social Media*.

Huberman, B.A., Romero, D.M. & Wu, F. (2009). Social Networks that Matter: Twitter Under the Microscope. *First Monday*, 14, 1.

Hundt, M., Nesselhauf, N., & Biewer, C. (Eds.). (2007). *Corpus linguistics and the web*. Amsterdam: Rodopi.

Hunston, S. (1989). *Evaluation and in experimental research articles*. (Tesis doctoral). Univeristy of Birmingham. Recuperado de http://etheses.bham.ac.uk/912/

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunston, S. & Thompson, G. (Eds.). (2007). *Functional linguistics. System and corpus: exploring connections*. London; Oakville, Connecticut: Equinox.

Hunston, S., y Oakey, D. (2010). *Introducing applied linguistics*. New York, NY: Routledge.

Hutchins, J. (2003). Alpac: the (in)famous reportEn Niremburg, S., Somers, H. y Wilks, Y. (Eds.), *Readings in Machine Translation* (pp. 131-136). Cambridge: MIT.

Jansen, B.J. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11), 1-20.

Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. En *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis* (San Jose, California, August 12 – 12, 2007). (pp. 56-65). New York, NY: ACM.

Johansson, S. (1991). Times change, and so do corpora. En Aijmer, K. & Altemberg, B. (Eds.), *English Corpus Linguistics* (pp. 305-314). London: Longman.

Johansson, S. (Ed.). (1990). *Newsletter of the International Computer Archive of Modern English*, 14.

Johansson, S., Stenström, A. B. (Ed.). (1992). *Newsletter of the International Computer Archive of Modern English*, 16.

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

Johnson, K. & Johnson, H. (Eds.). (1998). *Encyclopedic Dictionary of Applied Linguistics: A Handbook for Language Teaching*. Oxford: Blackwell.

Johnson, S. (2000). How Twitter will change the way we live. *Time, CNN*. Recuperado de http://individual.utoronto.ca/kreemy/proposal/04.pdf

Jones, R. & Ghani, R. (2000). Automatically building a corpus for a minority language from the web. En *Proceedings of the Student Workshop of the 38th Annual Meeting of the Association for Computational Linguistics*, 29–36. Strousburg, PA: ACM.

Kälgren, Gunnel. (1990). Review of Garside *et al*. (1987). *ICAME Journal,* 14, 98-103.

Kehoe, A. & Renouf, A. (2002). WebCorp: Applying the web to Linguistics and Linguistcis to the Web. *WWW2002 Conference*. Honolulu, Hawaii.

Keller, F., Lapata, M. & Ourioupina, O. (2002). Using the Web to Overcome Data Sparseness. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp.230-237), Philadelphia: ACM.

Kilgariff, A. (2013). *SketchEngine* [Computer Software]. Recuperado en http://www.sketchengine.co.uk/

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography* 1(1), 7-36.

Kilgarriff, A. (2001). Web as corpus. *Proceedings of the Corpus Linguistics Conference (CL 2001)*. University Centre for Computer Research on Language Technical Paper Vol. 13, Special Issue, Lancaster University, 342-344. Recuperado de: http://ucrel.lancs.ac.uk/publications/CL2003/CL2001%20conference/papers/kilgarri.pdf.

Kilgarriff, A. (2006). Googlelology is bad science. *Computational Linguistics* 33(1), 147–151.

Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the Special Issue on the Web as as Corpus. Computational linguistics, 29(3), 333-347. Recuperado de http://acl.ldc.upenn.edu/J/J03/J03-3001.pdf

Kilgarriff, A. & Tugwell, D. (2002). Sketching Words. En Marie-Hélène Corréard (Ed.), *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins* (pp.125-137). EURALEX.

Kilgarriff, A., Rychly´, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. En *Proceedings of the Eleventh EURALEX International Congress* (pp. 105-115). Lorient, France.

Kintsch, W. (1998). *Comprehension. A paradigm for cognition*. Cambridge: Cambridge University Press.

Kirkpatrick, R. (2013). Big Data for Development. *Big Data*, 1(1), 3-4. doi: 10.1089/big.2012.1502.

Kohavi, R., Longbotham, R., Sommerfield, D. & Henne, RM. (2009). Controlled Experiments on the Web: Survey and Practical Guide. *Data mining and knowledge discovery*, *18*(1), 140-81.

Krishnamurthy, B., Gill, P. & Arlitt, M. (2008). A few chirps about Twitter. *Proceedings of the First Workshop on Online Social Networks* (pp. 19-24). New York, NY: ACM.

Kučera, H. (1992). The odd couple: The linguist and the software engineer. The struggle for high quality computerized language aids. En Svartvik, J. (Ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991* (pp. 401-420). Berlin/New York: Mouton de Gruyter.

Kwak, H., Lee, Cl, Park., H. & Moon, S. (2010). What is Twitter, a Social Network or a News Media? *Proceedings of the 19th International conference on World Wide Web* (pp.591-600). New York, NY: ACM. doi: 10.1145/1772690.1772751

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. En *Application Deliverty Strategies*. Stamford, CT: META Group. Recuperado de http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Laney, D. (2012). Big Data Strategy Components: IT Essentials. *Gartner*, 1-11. Chicago: Gartner, Inc.

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

Lapata, M. & Keller, F. (2005). Web-based Models for Natural Language Processing. *ACM Transactions on Speech and Language Processing*, 2(1), 1-31.

Laufer, B. & Nation P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics,* 16(3), 307-322.

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014). The Parable of Google Flu: Tramps in Big Data Analysis. Science, 343, 1203-1205. doi: 10.1126/science.1248506

Leech, G. (1991). The state of the art in corpus linguistics. En Aijmer, K. & Altemberg, B. (Eds.), *English Corpus Linguistics* (pp. 8-29). London: Longman.

Leech, G. (1992). Corpora and theories of linguistics performance. En Svartvik, J. (Ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991* (pp. 105-122). Berlin/New York: Mouton de Gruyter.

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

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

Leech, G. & Johansson, S. (2009). The coming of ICAME, *ICAME Journal,* 33,5-20.

Leech, G., Garside, R. & Bryant, M. (1994). CLAWS4: The tagging of the British National Corpus. *Proceedings of the 15th International Conference on Computational Linguistics, Kyoto, Japan* (pp. 622-628).

Lenzerini, M. (2002). Data integration: a theoretical perspective. *Proceedings of the twety-first ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems*, 233-246. New York: ACM.

Lerman, J. (2013). Big Data and Its Exclusions. *Standford Law Review*, 66, 55-63. Recuperado de http://www.stanfordlawreview.org/online/privacy-and-big-data/big-data-and-its-exclusions

Lerman, K. & Ghosh, R. (2010). Information Contagion: An Empirial Study of the Spread of News on DIgg and Twitter Social Networks. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (pp. 90-97). Recuperado de https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1509/1839

Leskovec, J. & Horviz, E. (2007). Planetary-scale views on an Instant-Messaging-Network. *Microsoft Research Technical Report MSR-TR-2006-186*. Recuperado de http://research.microsoft.com/pubs/70389/tr-2006-186.pdf

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. En *KDD '05: Proceedings 11th International Conference on Knowledge discovery in data mining*, 177–187.

Lewis, M, Paul, G., Simons, F. & Fennig, C. D. (Eds.). (2015). *Ethnologue: Languages of the World, Eighteenth edition.* Dallas, Texas: SIL International. Recuperado de **http://www.ethnologue.com**

Lin, J. (2013). MapReduce is good enough? *Big Data*, 1(1), 28-37. doi: 10.1089/big.2012.1501.

Liu, Y., Kliman-Silver, C. & Mislove, A. (2014). The tweets they are a-Changin': Evolution of Twitter users and behavior. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. (306-314). Palo Alto, California: The AAAI Press.

Lledó Cunill, E. (coord.), Calero Fernández, M. A y Forgas Berdet, E. (2004). *De mujeres y diccionarios. Evolución de lo femenino en la 22ª edición del DRAE*. Madrid: Ministerio de Trabajo (Instituto de la Mujer).

Llisterri, J. (2003). Las tecnologías del habla: Entre la ingeniería y la lingüística. *Actas del I Congreso internacional "La ciencia ante el público. Cultura humanística y desarrollo tecnológico"* (pp. 44-67). Salamanca: Instituto Universitario de Estuduios de la ciencia y la tecnología.

Llisterri, J. (2003). Lingüística y tecnologías del lenguaje. Lynx. Panorámica de Estudios Lingüísticos, 2, 9-71. Recuperado de http://liceu.uab.cat/~joaquim/publicacions/Llisterri_03_Linguistica_Tecnologias_Lenguaje.pdf

Llisterri, J. Carbó, C., Machuca, M. J., Mota, M. C., Riera, M. & Ríos, A. (2003). El papel de la lingüística en el desarrollo de las tecnologías del habla. En Casas Gómez, M. (Dir.), Var Varo, C. (Ed.), *VII Jornadas de Lingüística* (pp. 137-191). Cádiz: Servicio de publicaciones de la Universidad de Cádiz.

Llorens Largo, F., Castel de Haro, M. J. (1996-2001). *Prácticas de Lógica, Prolog*. Universidad de Alicante. Recuperado de http://www.infor.uva.es/~teodoro/PrologAlicante.pdf

López García, A. y Morant Marco, R. (1991). *Gramática femenina*. Madrid: Cátedra.

López Quero, S., Calero Vaquera, M. L. & Zamorano Aguilar, A. (2004). Foro de debate vs. otros discursos electrónicos. *Español actual: Revista de español vivo*, 82, 53-76.

López Quero, Salvador. (2003). *El lenguaje de los "chats". Aspectos gramaticales*. Granada: Port-Royal Ediciones [Colección Lingüística].

López Research. (2013). An introduction to the Internet of Things (IoT). *Part 1 of "The IoT Series"*. San Francisco, CA: Lopez Research LLC.

Lozano Domingo, I. (1995). *Lenguaje femenino, lenguaje masculino*. Madrid: Minerva.

Lüdeling, A. y Kytö, M. (Eds.). (2008). *Corpus Linguistics: an International Handbook, v. 1* (pp. 68-97). Berlin: Mouton de Gruyter.

Lüdeling, A., Evert, S. & Baroni, M. (2007). Using Web Data for Linguistic Purposes. En Hundt, M., Nesselhauf, N., & Biewer, C. (Eds.) *Corpus linguistics and the web* (pp.7-24). Amsterdam: Rodopi.

Mackenzie, C. E. (1980). *Coded Character Sets, History and Development*. Reading, MA: Addison-Wesley.

Macskassy, S. A. (2012). On the study of social interaction in Twitter. *Proceedings of the Sixth International AAI Conference on Weblog and Social Media*. (226-233). Palo Alto, California: The AAAI Press.

Mahlberg, M. (2005). *English General Nouns. A corpus theoretical approach*. Amsterdam/Philadephia: Benjamins.

Mair, C. (1992). Comments on Wallace Chafe. En Svartvik, J. (Ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991* (pp. 98-104). Berlin/New York: Mouton de Gruyter.

Mair, C. (2006). Tracking ongoing grammatical change and recent diversification in present-day standard English: the complementary role of small and large corpora. En Renouf, A., y Kehoe, A. (Eds.), *The changing face of corpus linguistics* (pp. 355-376). Amsterdam: Rodopi.

Maletic, J. & Marcus, A. (2000). Data Cleansing: Beyon Integrity Analysis. *Proceedings of the 5th International Conference on Information Quality (IQ'00)* (pp. 200-209). Cambridge: Massachusetts Institute of Technology.

Manning, C., y Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.

Manyika, J., Chui, M., Brown, J. Dobbs, R., Roxburgh, C. & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. Mckinsey Global Institute.

Márquez, M. (2013). *Género gramatical y discurso sexista*. Síntesis: Madrid.

Martí Antonín, M. A. y Castelló Masallels, I. (2000). *Lingüística computacional*. Barcelona: Universitat de Barcelona.

Mayer-Schönberger, V. & Cukier, K. (2013). *Big data: a revolution that will transform how we live, work, and think*. New York: Eamon Dolan/Houghton Mifflin Harcourt.

McAfee, A., Brynjolfsson, E., Davenport, T., Patil, D. J., Barton, D. & Court, D. (2012). Big Data. *Harvard Business Review*, October, 2012, 59-68.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Theory, Method, and Practice*. Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (2003). *Corpus linguistics*. 2nd. Edition. Edinburgh: Edinburgh University Press.

McFedries, P. (2007). Technically speaking: All a-Twitter. *IEEE Spectrum*, 44(10), 84.

Mclver, D. J., Hawkins, J. B., Chunara, R., Chatterjee, A., Bhandari, A., Fitzgerald, T. P., Jain, S. H., Brownstein, J. S. (2015). Characterizing sellpe issues using Twitter. *J Med Internet Res*, 17(6), e140. DOI: 10.2196/jmir.4476.

Meya Llopart, M. y Huber, W. (1896). *Lingüística computacional*. Barcelona: Editorial Teide.

Meyer, C. F. (2002). *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Mihalcea, R. & Moldovan, D. (1999). A method for word sense disambiguation of unrestricted text. En *Proceedings of the 37th Meeting of ACL*, 152–158, Maryland: ACM.

Milhacea, R. & Moldovan, D. (1998). Word Sense Disambiguation Based on Semantic Density. Proceedings of COLING-ACL '98 Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal, Canada, August 1998. Recuperado de https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.acl99.pdf

Minsky, M. (1968). *Semantic information processing*. Cambridge, Mass.: MIT Press.

Minsky, M. L. (1961). Steps Toward Artificial Intelligence. *Proceedings of the Institute of Radio Engineers*, 48, 8-30.

Montes Giraldo, J. J. (1998). El diccionario de construcción y régimen de Cuervo. Boceto histórico. *Thesaurus*, 53(2), 314-324.

Morales, A., Pastor, D., Torres, Y., Frías-Martínez, V., Frías-Martínez, E., Oliver, N. Rutherfors, A., Logar, T., Clausen, R., De Backer, O. & Luego-Oroz, M. A. (2015). Studying Human Behavior through the Lens of Mobile Phones during Floods, *Netmob*. Recuperado de http://enriquefrias-martinez.info/yahoo_site_admin/assets/docs/abstractnetmob15.10433832.pdf

Moreno Fernández, F. (1990). Lingüística informática e informática lingüística. *Lingüística Española Actual*, 12(1), 5-16.

Moreno, A. (1998). *Lingüística computacional: Introducción a los modelos simbólicos, estadísticos y biológicos*. Madrid: Síntesis.

Mortureux, M. F. (2011). La néologie lexicale: de l'impasse à l'ouverture. *Langages* 3, 183, 11-24.

Murray, K. M. E. (1977), *Caught in the web of words: James Murray and the Oxford English Dictionary*. New Haven/London: Yale University Press.

Naaman, M., Boase, J. & 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 (pp.189-192). New York, NY: ACM. doi: 10.1145/1718918.1718953

Nakov, P. & Hearst, M. (2005). Search engine statistics beyond the n-gram: Application to noun compound bracketing. *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, 17–24. Michigan: Ann Arbor.

Nascimento, T. D., DosSantos, M. F., Danciu, T., DeBoer, M., van Holsbeeck, H., Lucas, S. R., Aiello, C., Khatib, L., Bender, M. A., UMSoD (Under)Graduate Class of 2014, Zubieta, J. K., DaSilva, A. F. (2014). Real-Time Sharing and Expression of Migraine Headache Suffering on Twitter: A Cross-Sectional Infodemiology Study. *J Med Internet Res*, 16(4): e96. DOI: 10.2196/jmir.3265.

Nilsson, N. (2009). *The Quest for Artificial Intelligence. A history of ideas and achievements*. Cambridge: Cambridge University Press.

Núñez Cabezas, E. A. & Guerrero Salazar, S. (2002). *El lenguaje político español*. Madrid: Cátedra.

Pak, A. & Paroubek, P. (2010). Twitter as Corpus for Sentiment Analysis and Opinion Mining. Proceedings of LREC (pp. 1320-1326). Recuperado de https://pdfs.semanticscholar.org/ad8a/7f620a57478ff70045f97abc7aec9687ccbd.pdf

Panchadsaram, R. (2014). Untapped potential. *Big Data*, 2(2), 63-64. doi: 10.1089/big.2014.1523.

Parodi, G. (2008). Lingüística de corpus: una introducción al ámbito. *RLA: Revista de Lingüística Teórica y Aplicada*, 46(1), 93-119.

Parodi, G. (2010). *Lingüística de Corpus: de la teoría a la empiria*. Madrid: Iberoamericana.

Paul, M. J., Dredze, M., Broniatowski, D. (2014). Twitter Improves Influenza Forecasting. *PLOS Current Outbreaks*, 1, 1-14. doi: 10.1371/current.outbreaks.90b9ed0f59bae4ccaa683a39865d9117

Pepper, R. & Garriti, J. (2014). The Internet of Everything: how the network unleashes the benefits of Big Data. En Bilbao-Osorio, B., Dutta, S. & Lanvin, B. (Eds.), *The Global Information Technology Report 2014. Rewards and Risks of Big Data* (pp. 35-42). Geneva: World Economic Forum.

Pérez Hernández, C. & Moreno Ortiz, A. (2009). Lingüística computacional y lingüística de corpus. Lingüística computacional y Lingüística de corpus.

Potencialidades para la investigación textual. En Nuria Rodríguez Ortega (Dir.), *Teoría y literatura artística en la sociedad digital: construcción y aplicabilidad de colecciones textuales informatizadas* (pp. 67-96). Gijón: TREA.

Piñol, M. C. (1999). ESPAN-L, un foro de debate en la Internet sobre la lengua española. *Estudios de Lingüística Española*. 1. Recuperado de http://elies.rediris.es/elies1/64.htm

Pitkowsky, E. F., Vásquez Gamarra, J. (2009). El uso de los corpus lingüísticos como herramienta pedagógica para la enseñanza y aprendizaje de ELE. *Tinkuy: La enseñanza del español como lengua extranjera en Quebec, Section d'Études hispaniques*. Université de Montréal y CEDELEQ III, 11, 31-52.

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

Provost, F. & Fawcett, T. (2013). Data Science and Its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, 1(1), 51-59. doi: 10.1089/big.2013.1508.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. New York: Longman.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1987) *A Grammar of Contemporary English*. London: Longman.

Radio Engineers, 49, 8–30. Recuperado de: http://web.media.mit.edu/ minsky/papers/steps.html

Ram, S., Zhang, W., Williams, M. & Pengetnze, Y. (2015). Predicting Asthma-Related Emergency Department Visits Using Big Data. *IEE Journal of biomedical and health informatics*, 19(4), 1216-1223.

Rayson, P., Walkerdine, J., Fletcher, W. H. & Kilgarriff, A. (2006). Annotated web as corpus. *Proceedings of the 2nd International Workshop on Web as Corpus* (pp.27-33). Stroudsburg, PA: ACM.

Renouf, A. (2003). WebCorp: Providing a renewable data source for Corpus Linguistics. En Granger, S. & Petch-Tyson, S. (Eds.), *Extending the scope of corpus-based research: new applications, new challenges* (pp. 39-58). Amsterdam: Rodopi.

Renouf, A. (2007). En Facchinetti, R. (Ed.), *Corpus Linguistics: 25 Years On* (pp. 27-50). Amsterdam/New York: Rodopi.

Renouf, A., & Kehoe, A. (Eds.) (2006). *The changing face of corpus linguistics*. Amsterdam: Rodopi.

Renouf, A., Kehoe, A. & Banerjee, J. (2007). WebCorp: an integrated system for web text search. En Hundt, M., Nesselhauf, N., & Biewer, C. (Eds.) *Corpus linguistics and the web* (pp.47-68). Amsterdam: Rodopi.

Resnik, P. (1999). Mining the Web for Bilingual Text. Proceedings of the AMTA-98 Conference. Recuperado de http://www.aclweb.org/anthology/P99-1068

Rigau, G., Magnini, B., Agirre, E. & Carroll, J. (2002). Meaning: A roadmap to knowledge technologies. En *Proceedings of COLING Workshop on A Roadmap for Computational Linguistics*. *Taipei, Taiwan*, 13, 1-7. Stroudsburg, PA: ACM.

Rimmer, M. (2009). "Wikipedia, collectve authorship and the politics of knowledge." En Arup, C. & van Caenegem, W. (Eds.). Intellectual Property Policy Reform. Fostering Innovation and Develpment. Cheltenham, Northampton: Edward Elgar Publishing Limited.

Robins, R. H. (2000). *Breve historia de la lingüística*. Madrid: Cátedra.

Rock, F. (2001). Poliby and practice in the anonymization of linguistic data. *International Journal of Corpus Linguistics*, 6(1), 1-26.

Rogers, E. (2003). *Diffusion of Innovations, 5th Edition*. New York: Simon and Schuster.

Rojo, G. (2001). La explotación de la basde de datos sintácticos del español actual (BDS). En De Kock (Ed.), *Gramática española. Enseñanza e investigación. Lingüística con corpus, catorce aplicaciones sobre el español* (pp. 255-286). Salamanca: Ediciones Universidad de Salamanca.

Rojo, G. (2002). Sobre la lingüística basada en análisis de corpus. Ponencia plenaria en las Jornadas sobre corpus lingüísticos. San Sebastián, Uzei.

Rojo, G. (2005-2006). Informática y Lingüística: Las lenguas en la sociedad del conocimiento. *Boletín de RedIRIS*, 74-75, 1-8. Recuperado de: http://www.rediris.es/rediris/boletin/74-75/ponencia1.pdf

Rojo, G. (2008). Lingüística de corpus y lingüística del español. Ponencia plenaria en el XV Congreso de la Asociación de Lingüística y Filología de América latina, Montevideo, 2008. Recuperado de http://gramatica.usc.es/~grojo/Publicaciones/Lgca_corpus_lgca_espanol.pdf

Rojo, G. (2009). Sobre la construcción de diccionarios basados en corpus. Revista tradumàtica. Traducció i Tecnologies de la Informació i la Comunicació, 7, 1-7.

Rojo, G. (2010). Aguja de navegar corpus. En Castel, V. M. & Cubo de Severano, L. (Eds.), *La renovación de la palabra en el bicentenario de la Argentina. Los colores de la mirada lingüística* (pp. 1151-1163). Mendoza: Editaorial FFlL, UNCuyo.

Rondeau, G. (1984). *Introduction à la terminologie*. Chicoutimi (Québec): Gaëtan Morin

Russell, S., y Norvig, P. (1995). *Artificial intelligence: A Modern Approach*. Englewood Cliffs, N.J.: Prentice Hall.

Salathé, M., Vu, D. Q., Khandelwal, S., & Hunter, D.R. (2013). The dynamics of health sentiments on a large onine social network. *EPJ DATA Science*, 2(1), 1-12.

Santini, M. (2007). Characterizing Genres of Web Pages: Genre Hy-bridism and Individualization. *40th Annual Hawaii International Conference on System Sciences (HICSS'07)*. Recuperado de http://csdl2.comput-er.org/comp/proceedings/hicss/2007/2755/00/27550071.pdf.

Saussure, F. (2002). *Curso de lingüística general*. Madrid: Akal Ediciones.

Sazaki, T. Okazaki, M. & Matsuo, Y. (2010). Earthquake shake Twitter Users: Real-time Event Detection by Social Sensors. *Proceedings of the 19th International Conference on the World Wide Web* (pp. 851-860). New York, NY: ACM.

Schaefer, S., Harrison, C., Lamba, N. & Srikanth, V. (2011). Smarter cities series: Understanding the ibm approach to traffic management. *Redguides for Business Leaders*. IBM.

Schäfer, R., Barbaresi, A. & Bildhauer, F. (2014). Focused web corpus crawling. *Proceedings of the 9th Web as Corpus Workshop* (pp. 9-15). Stroudsburg, PA: ACM.

Scott, M. (2012). *WordSmith Tools* (Version 5.0) [Computer Software]. Recuperado de http://www.lexically.net/software/index.htm

Scott, M., & Johns, T. (1993). *MicroConcord* [Computer Software]. Recuperado de http://www.lexically.net/software/index.htm

Sekhar, N. (2008). *Corpus linguistics: An introduction*. Nueva Delhi: Pearson Education.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27 (julio y octubre), 379-423 y 623-656.

Shannon, C. E. (1950). A Chess-Playing Machine. *Scientific American*, 182(2), 48-52.

Shannon, C. E. & Weaver, W. (1998). *The mathematical theory of communication*. Urbana: University of Illinois.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. En Baroni, M. & Bernardini, S. (Eds.), *WaCky! Working Papers on the Web as Corpus* (pp. 63-98). Bologna: GEDIT.

Sharoff, S. (2007). Classifying Web corpora into domain and genre using automatic feature identification. En Fairon, C., Naets, H., Kilgarriff, A. & De Schryver, G. (Eds.), *Building and Exploring Web Corpora* (WASC3-2007) (pp. 83-94). Louvain-la-Neuve: presses Universitaires de Louvain.

Shneiderman, B., Plaisant, C. & Hesse, B. (2013). Improving health and healthcare with interactive visualization tools. *IEEE Computer*, 46(5), 58-66.

Signorini A., Segre A.M. & Polgreen, P.M. (2011). The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLoS ONE,* 6(5), e19467. doi:10.1371/journal.pone.0019467.

Sills, J. (2014). Twitter: Big data opportunities. *Science*, 345(6193), 148-149.

Sin-wai, C. (Ed.). (2015). *The Routledge Encyclopedia of Translation Technology*. New York: Routledge.

Sinclair, J. M. (1991): *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. M. (1992). The automatic analysis of corpora. En Svartvik, J. (Ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991* (pp. 378-397). Berlin/New York: Mouton de Gruyter.

Sinclair, J. M. (2004). *Trust the text: language, corpus and discourse*. London/New York: Routledge.

Sinclair, J. M. (2005). Corpus and Text - Basic Principles. En Wynne, M. (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice* (pp. 1-16). Oxford: Oxbow Books: 1-16. Recuperado de: http://ahds.ac.uk/linguistic-corpora/

Stubbs, M. (2001). *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell.

Stubbs, M. (2006). Corpus analysis: the state of the art and three types of unanswered question. En Hunston, S. & Thompson, G. (Eds.), *System and corpus: Exploring connections* (pp. 15-36). London: Equinox.

Stubbs, M. (2007a). Corpus analysis: the state of the art and three types of unanswered questions. En Hunston, S. & Thompson, G. (Eds.). (2007). *System and corpus: exploring connections*. London; Oakville, Connecticut: Equinox.

Stubbs, M. (2007b). On texts, corpora and models of language. En Hoey, M. (Ed.), *Data, description, discourse* (pp. 127-162). London: HarperCollins.

Stubbs. M. (1996). *Text and corpus analysis. Computer-assisted studies of language and culture*. Massachusetts: Blackwell.

Suchomel, V. & Pomikálek, J. (2012). Efficient Web Crawling for Large Text Corpora. *Proceedings of the 7th Web as Corpus Workshop*, Lyon, France. (pp. 39-43).

Svartvik, J. (1992). Corpus linguistics comes of age. En Svartvik, J. (Ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991* (pp. 7-16). Berlin/New York: Mouton de Gruyter.

Svartvik, J. (2007). En Facchinetti, R. (Ed.), *Corpus Linguistics: 25 Years On* (pp. 11-26). Amsterdam/New York: Rodopi.

Svartvik, J. (Ed.). (1992). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin/New York: Mouton de Gruyter.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics*, 10(1), 1-13.

The Economist. (2010). *Data, data everywhere: a special report on managing information*. San Mateo: The Economist Newspaper Ltd. Recuperado de https://www.emc.com/collateral/analyst-reports/ar-the-economist-data-data-everywhere.pdf

Thompson, G. & Hunston, S. (2006). System and corpus: two traditions with a common ground. En Hunston, S., & Thompson, G. (Eds.), Functional linguistics. *System and corpus: exploring connections* (pp. 1-14). London: Equinox Publishing Ltd.

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

Tole, A. A. (2013). Big Data Challenges. *Database Systems Journal*, IV(3), 31-40.

Torruela, J. & Llisterri, J. (1999). Diseño de corpus textuales y orales. En Blecua, J. M., Clavería, G., Sánchez, C. & Torruella, J. (Eds.), *Filología e informática. Nuevas tecnologías en los estudios filológicos* (pp. 45-47). Barcelona: Seminario de Filología e Informática, Departamento de Filología Española, Universidad Autónoma de Barcelona. Editorial Milenio. Recuperado de http://latel.upf.edu/traductica/lc/material/torruella_llisterri_99.pdf

Tucker, G. (2006). Systemic in*corpora*tion: on the relationship between corpus and systemic functional grammar. En Hunston, S., & Thompson, G. (Eds.), *Functional linguistics. System and corpus: exploring connections* (pp. 81-102). London: Equinox Publishing Ltd.

Turing, A. M. (1996). Intellingent Machinery, A Heretical Theory. *Philosohia Mathematica*, 3(4), 256-260.

Turney, P. D. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. *European Conference on Machine Learning*, 491–502.

United Nations Global Pulse. (2012). *Big Data for Development: Challenges & Opportunities*. New York: Global Pulse.

United Nations Global Pulse. (2014). Mining Indonesian Tweets to Understand Food Price Crises. *UN Global Pulse, Methods Papers, February 2014*. Recuperado de http://www.unglobalpulse.org/sites/default/files/Global-Pulse-Mining-Indonesian-Tweets-Food-Price-Crises%20copy.pdf

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

Van Essen, A. (1983). *E. Kruisinga ; a chapter in the history of linguistics in the Netherlands*. Leiden.

Vandelanotte, L., Davidse, K., Gentents, C. & Kimps, D. (Eds.). (2015), Review in ICAME Journal. *De Gruyter Open*, 39, 171-177. doi: 10.1515/icame-2015-0012.

Váradi, T. (2001). The linguistic relevance of Corpus Linguistics. En Rayson, P., Wilson, T., McEnery, A., Hardie, A. & Khoja, S. (Eds.), *Proceedings of the Corpus Linguistics 2001 Conference* (pp. 587-593). Lancaster University: UCREL.

Varantola, K. (2000). Translators and disposable corpora. En *Proceedings of CULT (Corpus Use and Learning to Translate)*. Bertinoro, Italy, November.

Varantola, K. (2002). Disposable corpora as intelligent tools in translation. En Tagnin, S. E. O. (Ed.), *Cadernos de Traduçao: Corpora e Traduçao*, 1(9), 171-189.

Varela Ortega, S. (2005). *Morfología lexica: la formación de palabras*. Madrid: Gredos.

Viana, V., Zyngier, S., & Barnbrook, G. (2011). *Perspectives on corpus linguistics*. Amsterdam: J. Benjamins Pub.

Villasenor-Pineda, L., Montes y Gómez, M., Pérez-Coutino, M. & Vaufreydaz, D. (2003). A corpus balancing method for language model construction. En *Fourth*

International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2003), 393–401. Mexico City.

Villaseñor Pineda, L., Montes Gómez, M., Pérez Coutiño, M. & Vaufreydaz, D. (2003). A Corpus Balancing Method for Language Model Construction. *Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico. Recuperado de http://www-prima.inrialpes.fr/Vaufreydaz/Telechargement/Villasenor03a.pdf

Villayandre Llamazares, M. (2010). Aproximación a la lingüística computacional. Tesis doctoral. Universidad de León. Disponible en el repositorio institucional abierto de la Universidad de León.

Volk, M. (2001). Exploiting the WWW as a corpus to resolve PP attachment ambiguities. En *Proceedings of Corpus Linguistics 2001*, Lancaster, UK. Recuperado de http://www.zora.uzh.ch/20269/2/Volk_2001V.pdf

Wahab, M. H. A., Mohd, M. N. H., Nahafi, H. F. & Mohsin, M. F. M. (2008). Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm. *Proceedings of World Academy of Science, Engineering and Technology*, 36, 970-977.

Wang, T., Rudin, C., Wagner, D. & Sevieri, R. (2015). Finding Patterns with a Rotten Core: Data Mining for Crime Series with Cores. *Big Data*, 3(1), 3-21. doi: 10.1089/big.2014.0021.

Warden, P. (2011). *Big Data Glossary*. Sebastopol, CA: O'Reilly Media, Inc.

Weber, R. H. & Weber, R. (2010). *Internet of things. Legal perspectives*. London/New York: Springer.

White House. (2012). Fact Sheet: Big Data Across the Federal Government, Department of Defense. *Office of Science and Technology Policy, Executive Office of the Pressident*. Recuperado de https://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final.pdf

Wright, S. E. & Budin, G. (2001). *Handbook of terminology management, vol 1, Basic Aspects of Terminology Management*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Wright, S. E. & Budin, G. (2001). *Handbook of terminology management, vol 2, Application-Oriented Terminology Management*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Wu, F., Huberman, B., Adamic, L. A. & Tyler, J. R. (2004). Information flow in social groups. *Physica A: Statistical Mechanics and its Applications*, 337 (1-2), 327-335. doi:10.1016/j.physa.2004.01.030.

Xia, F., Yang, L., Wang, L. & Vinel, A. (2012). Internet of Things. *International Journal of Communication systems*, 25, 1101-1102.

Yoon, S., Elhadad, N. & Bakken, S. (2013). A Practical Approach for Content Mining of Tweets. *American Journal of Preventive Medicine*, 45(1), 122-129.

Young, S. D., Rivers, C. & Lewis, B. (2014). Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Preventive Medicine*, 63: 112-115. doi: 10.1016/j.ypmed.2014.01.024.

Yus, F. (2001). *Ciberpragmática. El uso del lenguaje en Internet*. Barcelona: Ariel.

Zheng, Z. (2002). AnswerBus Question Answering System. *Proceedings of the 2nd International Conference on Human Language Technology Research,* California, United States (pp.399-404). New York, NY: ACM.

Zikopoulos, P. C., Eaton, C., deRoos, D., Deutsch, T. & Lapis, G. (2012). *IBM. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. New York: McGraw Hill.

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

Webgrafía

Alpert, J. & Hajaj, N. (2008). We knew the web was big… *Google Official Blog* [fecha de consulta: 2 de febrero de 2015] Recuperado de https://googleblog.blogspot.com.es/2008/07/we-knew-web-was-big.html

Alvar, M. (1985). La influencia del inglés en la República Dominicana. Valoración de una encuesta oral. *Biblioteca virtual Miguel de Cervantes*. Retrieved from http://www.cervantesvirtual.com/obra-visor/la-influencia-del-ingls-en-la-repblica-dominicana-valoracin-de-una-encuesta-oral-0/html/00ec2fb4-82b2-11df-acc7-002185ce6064_2.html

American Standards Association. (1963). American Standard Code for Infomation Exchange. *WPS* [fecha de consulta: 3 de agosto de 2015] Recuperado de http://worldpowersystems.com/J/codes/X3.4-1963/

Armonk, N. Y. (2015). IBM Delivers First Cloud Data Services with Twitter built-in for Business Professionals and Developers. *IBM* [fecha de consulta: 30 de febrero de 2015] Recuperado de http://www-03.ibm.com/press/uk/en/pressrelease/46344.wss

Barranco Fragoso, R. (2012). ¿Qué es big data? *IBM Developer Works* [fecha de consulta: 20 de diciembre de 2014] Recuperado de https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

Bejerano, P. (2013). Big data gap: los retos del futuro. *Blog Thinking Big* [fecha de consulta: 22 de diciembre de 2014] Recuperado de http://blogthinkbig.com/big-data-gap-retos-futuro/

Bergman, M., Paavola, S. & Queiroz, J. (2016). Token. *The Commens Dictionary: Peirce's Terms in His Own Words. New Edition* [fecha de consulta: 5 de febrero de 2016] Recuperado de http://www.commens.org/dictionary/term/token

Brandel, M. (1999). 1963: The debut of ASCII. *CNN* [fecha de consulta: 3 de agosto de 2015]Recuperado de http://edition.cnn.com/TECH/computing/9907/06/1963.idg/

Carbonell, J. (1992). El procesamiento del lenguaje natural, tecnología en transición. *Centro Virtual Cervantes: La lengua española y las nuevas tecnologías* [fecha de consulta: 30 de julio de 2015] Recuperado de

http://cvc.cervantes.es/obref/congresos/sevilla/tecnologias/ponenc_carbonell.htm

Carlsson, N. (2011). The Real History of Twitter. *Business Insider Tech* [fecha de consulta: 30 de febrero de 2015] Recuperado de http://www.businessinsider.com/how-twitter-was-founded-2011-4

Carr, D. (2010). Why Twitter will endure. *The New York Times* [fecha de consulta: 2 de marzo de 2015] Recuperado de http://www.140characters.com/2009/01/30/how-twitter-was-born/

Christensson, P. (2010). Character Encoding Definition. *Tech Terms* [fecha de consulta: 2 de agosto de 2015] Recuperado de http://techterms.com/definition/characterencoding

Constable, P. (2001). Character set encoding basics. *NRSI: Computer & Writing Systems* [fecha de consulta: 2 de agosto de 2015] Recuperado de http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=IWS-Chapter03#79e846db

Cooper, M. & Mell, P. (s. f.). Tackling Big Data. NIST Information Technology Laboratory, *Computer Security Division* [fecha de consulta: 3 de marzo de 2015] Recuperado de http://csrc.nist.gov/groups/SMA/forum/documents/june2012presentations/fcsm_june2012_cooper_mell.pdf

Davenport, T. H. & Patil, D. J. (2012). Data scientist: the sexiest job of the 21st century. *Harvard Business Review* [fecha de consulta: 30 de septiembre de 2015] Recuperado de https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

Davis, M. (2008). Moving to Unicode 5.1. Google Official Blog. [fecha de consulta: 3 de agosto de 2015] Recuperado de https://googleblog.blogspot.com.es/2008/05/moving-to-unicode-51.html

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

Diosdado Rivera, D. (2015). 6,23 Terabytes de información fueron transmitidos durante el SuperBowl XLIX. *Addictware*. [fecha de consulta: 27 de julio de 2015] Recuperado de http://addictware.com.mx/comunicaciones/infraestructura/7052-6-23-terabytes-de-informacion-fueron-transmitidos-durante-el-super-bowl-xlix

Dumbill, E. (2012). What is big data? An introduction to the big data landscape. *O'Reilly*. [fecha de consulta: 30 de enero de 2015] Recuperado de https://www.oreilly.com/ideas/what-is-big-data

EAGLES. 1996. "Text Corpora Working Group Reading Guide". *Documento Eagles (Expert Advisory Group on Language Engineering)* EAG-TCWG-FR-2. [fecha de consulta: 4 de agosto de 2015] Recuperado de http://www.ilc.cnr.it/EAGLES/corpintr/corpintr.html

Fundación del Español Urgente. (2016). *Fundéu BBVA*. Recuperado de http://www.fundeu.es/

Gawande, A. (2011). Doctor Hotspot. *PBS Frontline. WGBH Educational Foundation*. [fecha de consulta: 5 de agosto de 2015] Recuperado de http://www.pbs.org/wgbh/pages/frontline/ doctor-hotspot/

Hachman, Mark. (2011). Humanity's tweets: Just 20 terabytes. *PC Magazine* [fecha de consulta: 2 de febrero de 2015] Recuperado de http://www.pcmag.com/article2/0,2817,2382347,00.asp

Hammersley, B. (2004). Audible revolution. The Guardian [5 de septiembre de 2015] Recuperado de http://www.theguardian.com/media/2004/feb/12/broadcasting.digitalmedia

Hays, C. L. (2004). What Wal-Mart knows about customers' habits. *New York Times* [10 de enero de 2015] Recuperado de http://www.nytimes.com/2004/11/14/business/yourmoney/what-walmart-knows-about-customers-habits.html?_r=0

Henderson, T. (2014). Ancient Computer Character Code Tables – and Why They're Still Relevant. SmartBear [2 de agosto de 2015] Recuperado de http://blog.smartbear.com/development/ancient-computer-character-code-tables-and-why-theyre-still-relevant/?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+SmartBear+%28SmartBear+Software+Quality+Matters+Blog%29

IBM. (2015). *Insights Using Twitter Data* [30 de enero de 2015] Recuperado de http://www.ibm.com/big-data/us/en/big-data-and-analytics/ibmandtwitter.html.

IDC. (2014). The digital universe of opportunities: rich data and the increasing value of the internet of things. IDC, EMC2. [3 de febrero de 2015] Recuperado de http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

Internet Live Stats. (2015). [25 de julio de 2015] Recuperado de http://www.internetlivestats.com/

Ishida, R. (2010). Codificación de caracteres: conceptos básicos. *W3 Consortium*. [2 de agosto de 2015] Recuperado de http://www.w3.org/International/articles/definitions-characters/

Jennings, T. (2004). An annotated history of some character codes or ASCII: American Standard Code for Information. *WPS*. [3 de febrero de 2015] Recuperado de http://worldpowersystems.com/archives/codes/#ASCII-1967

Johnson, L. B. (1968). 127 - Memorandum Approving the Adoption by the Federal Government of a Standard Code for Information Interchange. *The American Presidency Project*. [2 de agosto de 2015] Recuperado de http://www.presidency.ucsb.edu/ws/index.php?pid=28724

KDNuggets. (2014). What Analytics, Data Mining, Data Science software/tools you used in the past 12 months for a real project poll [25 de julio de 2015] Recuperado de http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-software-used.html

Kowalski, H. (2014). Franz Och, Ph.D., Expert in Machine Learning and Machine Translation, Joins Human Longeviti, Inc. As Chief Data Scientist. Human Longevity, Inc. [24 de julio de 2015] Recuperado de http://www.humanlongevity.com/franz-och-ph-d-expert-in-machine-learning-and-machine-translation-joins-human-longevity-inc-as-chief-data-scientist/

Laney, D. (2012). Big data strategy components: business essentials [24 de julio de 2015] Recuperado de http://www.iab.fi/media/tutkimus-matskut/gartner_big_data_strategy_components.pdf

Loukides, M. (2010). What is data science? The future belongs to the companies and people that turn data into products. *O'Reilly*. [1 de febrero de 2015] Recuperado de https://www.oreilly.com/ideas/what-is-data-science

Merino, M. (2014). ¿Qué es una API y para qué sirve? *Ticbeat*. [8 de octubre de 2015] Recuperado de http://www.ticbeat.com/tecnologias/que-es-una-api-para-que-sirve/

Morales, R. (2014). Qué es la codificación de caracteres. *Ticarte*. [30 de julio de 2015] Recuperado de http://www.ticarte.com/contenido/que-es-la-codificacion-de-caracteres

O'Reilly, T. (2005). What is web 2.0. Design Patterns and Business Models for the Next Generation of software. *O'Reilly*. [15 de octubre de 2015] Recuperado de http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html?page=1

Parry, M. (2012). Big data on campus. The New York Times. [10 de enero de 2015] Recuperado de http://www.nytimes.com/2012/07/22/education/edlife/colleges-awakening-to-the-opportunities-of-data-mining.html

Parry, M. (2012). College Degrees, Designed by the Numbers. *The Chronicle of Higher Education*. Recuprado de https:// chronicle.com/article/College-Degrees-Designed-by/132945/

Pastor, R. (2009). El poscasting II: una breve historia. Anexo M blog [30 de octubre de 2015] Recuperado de http://www.anexom.es/tecnologia/el-podcasting-ii-una-breve-historia/

Patil, D. J. (2011). Building Data Science Teams. *O'Reilly Radar* [15 de enero de 2015] Recuperado de http://radar.oreilly.com/2011/09/building-data-science-teams.html

Proyecto Aracne. (2015). *Fundéu BBVA*. Recuperado de http://www.fundeu.es/aracne/

Sagolla, D. (2009). How Twitter was born. *140 Characters*. [30 de enero de 2015] Recuperado de http://www.140characters.com/2009/01/30/how-twitter-was-born/

Sanz, E. (2016). ¿Es Twitter una herramienta útil para la ciencia? *El País* [1 de febrero de 2015] Recuperado de http://elpais.com/elpais/2016/01/04/ciencia/1451922060_370301.html

Sarno, D. (2009). Twitter creator Jack Dorsey illuminates the site's founding document. Part I. *Los Angeles Times, Technology* [2 de febrero de 2015] Recuperado de http://latimesblogs.latimes.com/technology/2009/02/twitter-creator.html

Segall, L. (2010). Buy a vowel? How Twtter became Twitter. *CNN Money* [10 de febrero de 2015] Recuperado de http://money.cnn.com/galleries/2010/technology/1011/gallery.Startup_Domain_Names/

Setalvad, A. (2015). Google Translate adds 20 new languages to video text translation. *The verge* [2 de noviembre de 2015] Recuperado de http://www.theverge.com/2015/7/29/9061135/google-translate-update-new-languages-word-lens

Shamama, J. (2015). How much of the Internet is Hidden. *TestTube*. [10 de octubre de 2015] Recuperado de http://testtube.com/testtubeplus/how-much-of-the-internet-is-hidden

Smith, C. (2015). By the numbers: 150+ amazing Twitter statistics. *DMR*. [10 de febrero de 2015] Recuperado de http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/

Sneiderman, P. (2013). Using Twitter to track the flu. *Hub* [10 de febrero de 2015] http://hub.jhu.edu/2013/01/24/using-twitter-to-track-flu

Statista. (2015). [24 de julio de 2015] Recuperado de http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Stone, B. (2009). What's happening. *Twitter Blog, Twitter Inc*. Recuperado de https://blog.twitter.com/2009/whats-happening

Stonebraker, M. Big data means at least three different things. Database Group, MIT Computer Science and Artificial Intelligence Lab. Recuperado de http://www.nist.gov/itl/ssd/is/upload/NIST-stonebraker.pdf

The Economist. (2011). Drowning in numbers. [10 de enero de 2015] Recuperado de http://www.economist.com/blogs/dailychart/2011/11/big-data-0

The IEEE and the Open Group. (2001-2004). Regular Expressions. *The Open Group Base Specifications Issue 6* [30 de noviembre de 2015] Recuperado de http://pubs.opengroup.org/onlinepubs/009696899/basedefs/xbd_chap09.html

The top 500 sites on the web (n). Alexa. [22 de julio de 2015] Recuperado de http://www.alexa.com/topsites

The World Factbook. (2015). *Central Intelligence Agency*. [22 de julio de 2015] Recuperado de https://www.cia.gov/library/publications/the-world-factbook/fields/2177.html.

Torres, A. (2015). Lo que tu móvil sabe de ti. *El País*. [11 de febrero de 2015] Recuperado de http://economia.elpais.com/economia/2015/05/19/actualidad/1432029128_644860.html

Torres, J. (2012). Big data 2.0: retos y tendencias tecnológicas del big data. IIR España: an informa business. [22 de diciembre de 2015] Recuperado de

http://www.jorditorres.org/wp-content/uploads/2012/06/iiR2012.BigData.slides.post_.pdf

Torres, J. (2012). Retos del big data. *Barcelona Supercomputing Center* [23 de diciembre de 2015] Recuperado de http://es.slideshare.net/jorditorres/retos-del-big-data

Twitter (2012, March 21). Twitter turns six. Twitter. [12 de febrero de 2015] Recuperado de https://blog.twitter.com/2012/twitter-turns-six

UCL. (2011). A brief history of the Survey of English Usage. *UCL* [30 de octubre de 2014] Recuperado de http://www.ucl.ac.uk/english-usage/about/history.htm

Unicode (1991-2016). About the Unicode Standard. *Unicode, Inc.* [5 de agosto de 2015] Recuperado de http://www.unicode.org/standard/standard.html

Unicode (1991-2016). History of Unicode: Summary Narrative. *Unicode, Inc*. [5 de agosto de 2015] Recuperado de http://www.unicode.org/history/summary.html

Unicode (1991-2016). The Unicode Consortium Members.. *Unicode, Inc*. [5 de agosto de 2015] Recuperado http://www.unicode.org/consortium/members.html

Unicode Technical Report, 17. (1999-2008). *Unicode Character Encoding Model. Technical Report* [5 de agosto de 2015] Recuperado de http://www.unicode.org/reports/tr17/

Uso de Twitter/Datos de la empresa. (2015) [30 de julio de 2015] Recuperado de https://about.twitter.com/es/company

W3Consortium. (2005). Character Model for the World Wide Web 1.0: Fundamentals. *W3 Consortium*. [17 de abril de 2015] Recuperado de https://www.w3.org/TR/charmod/

Walker, J. (2003). Weblog. Routledge Encylcopedia of Narrative Theory [5 de septiembre de 2015] Recuperado de http://jilltxt.net/archives/blog_theorising/final_version_of_weblog_definition.html

Wikimedia. (2015). ASCII Table. *Wikipedia Commons* [6 de agosto de 2015] Recuperado de https://upload.wikimedia.org/wikipedia/commons/1/1b/ASCII-Table-wide.svg

Índice de anexos (incluidos en CD)