La web como corpus: un esbozo

The Web as Corpus: An Overview

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

Resumen

La irrupción en la sociedad de la *World Wide Web* ha revolucionado la forma de entender el conocimiento y de acceder a él. 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 de 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 en cuanto a la recopilación y análisis de datos de manera exponencial. En esta investigación documental realizamos una aproximación al concepto de *web como corpus* y a las distintas terminologías existentes para denominarlo. Mediante un marco teórico, explicamos también sus características fundamentales, las diferencias que existen con respecto a los corpus tradicionales y las ventajas de su utilización para la investigación lingüística.

Palabras clave: web como corpus, web para corpus, lingüística de corpus, investigación lingüística.

Abstract

During the last years, the World Wide Web has been the main meeting point of information, communication, culture, and commerce. The web, widely known as an inexhaustible and free access resource, is considered an invaluable tool of information for scientific research and knowledge popularization. As any other discipline, Linguistics takes advantage from the web and its hundreds of millions of words and texts there included. Apart from traditional linguistic approaches exploring the web and doing research based on it, corpus linguistics and related areas such as computational linguistics have spread enormously their research scope thanks to machine readable texts, data size, accessibility, and web constant updating. This study explores the concept *web as a corpus* and its terminological variation. Key features of the web and differences with traditional text corpus are also explored. Finally, web advantages for linguistic research are offered.

Keywords: web as corpus, corpus web, big data, corpus linguistics, linguistic research.

1. CONCEPTOS PREVIOS

La *World Wide Web* se ha convertido en el principal punto de encuentro mundial 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 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.

En este estudio pretendemos llevar a cabo una investigación documental, con una revisión crítica de conceptos y visiones teóricas relacionadas con el concepto de la web como corpus y sus principales características, así como una comparación con los corpus más tradicionales y una aproximación a las ventajas que supone su utilización en la investigación lingüística.

Así pues, la Lingüística, como el resto de las 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 presentados 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, 2004a) 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 en la Universidad de Sheffield (Greenwood *et al*., 2002) y en Microsoft (Dumais *et al*., 2002) 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 (2004b), por otra parte, han utilizado las oportunidades 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*[1] (*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 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:

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

> 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. 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 (Hundt *et al*., 2007) 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*"[2]) 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, de las tres desventajas que menciona Rojo ninguna 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

[2] 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.

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.

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 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, traducción propia). 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 (Kilgarriff, 2001: 1):

> 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. LA WEB COMO CORPUS *VS.* LA WEB PARA LOS CORPUS (*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 ejemplos 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 los autores (Baroni y Bernardini, 2006) a este tipo de corpus 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: 237):

> 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.

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 (2014).

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 (1992: 96), 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; o Bowker y Pearson (2002: 9), 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. 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 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).

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 representativeness 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 (2010) 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 (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 Chomsky 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" (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 (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 disposición. 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".

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". Aunque, a pesar de ello, aconsejan no admitir el gran tamaño como algo siempre deseable: "it is very important, however, not to assume 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. McEnery y Wilson (3003: 30) hacen alusión al 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) explican que una de las desventajas de los corpus tradicionales es que presentan pocas o ninguna ocurrencia de muchas palabras, mientras que otras aparecen muy repetidas, debido al fenómeno conocido como la ley de Zipf (1965).[3] 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) tampoco duda en afirmar que: "con toda claridad, es necesario seguir construyendo corpus y su tamaño debe ser lo más grande que podamos conseguir". 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 de trabajo conocido como *corpus-driven linguistics* y defendido por Tognini-Bonelli (2001), Teubert (2005), Sinclair (2004) o Gatto (2008, 2014), 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

[3] 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.

> 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, 2014).

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 (2014), analizaremos cuatro aspectos importantes dentro del contenido de la web.

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 offer a home to all linguistics styles within a language; it offers a home to all language –one their 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.

Según Lewis, Paul, Simons y Fennig (2015), los tres idiomas más hablados en el mundo en 2015 eran, en este orden, el chino, el español y el inglés, mientras que en Internet, el idioma más utilizado es el inglés, seguido del chino y del español. Si comparamos estos datos con cifras de cinco años atrás, vemos que, a pesar de que estos tres idiomas siguen a la cabeza, el número de usuarios de Internet en otras distintas lenguas ha crecido desde 2010. Además, 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% (Internet World Stats, 2016). En este estudio, podemos observar que 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".

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 nos recuerda que: "organizing knowledge into ontologies is an ancient art, descended from philosophy and epistemology" (2003: 7). 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, 2014). 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.

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, 2014).

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 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" (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, 2014: 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.

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, 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" (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 (2014) 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.

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.

5. PRESENTE Y FUTURO DE LA WEB COMO CORPUS

Con este panorama, parece lógico pensar que el desarrollo de la web como corpus, con el apoyo de nuevas metodologías y tecnologías, como las basadas en *big data*, se encuentran en permanente evolución y expansión. Naturalmente, como acabamos de apuntar, todo desarrollo científico requiere una serie de herramientas metodológicas de apoyo, al igual que ocurre en este caso. Más aún, teniendo en cuenta la inevitable relación con la tecnología de la materia en cuestión. Hace ya bastantes años que existen herramientas para el trabajo con grandes corpus textuales, como, por ejemplo, *WebCorp Live* o *BootCat*, ambas diseñadas bien para construir corpus textuales o para utilizar la web como un corpus en sí mismo. Como mejora de estas herramientas y línea de trabajo futura, estamos trabajando en una herramienta informática con soporte web para el trabajo lingüístico a través de *big data* que permite la extracción, gestión, almacenamiento y análisis de la información textual contenida en *Twitter* (González Fernández, 2016). Consideramos que solo con el trabajo interdisciplinar entre la informática y la lingüística obtendremos los mejores resultados en la gestión de grandes corpus y de la web como corpus.

Referencias bibliográficas

Agirre, Eneko, Olatz Ansa, Eduard Hovy y David Martínez. 2000. Enriching very large ontologies using the WWW. En *Proceeding of the Ontology Learning Workshop of the European Conference of AI (ECAI)*, Berlin, Germany.

Araujo, María Helena y Silvia Melo. 2003. Del caos a la creatividad: los chats entre lingüistas y didactas. En Covadonga López Alonso y Arlette Séré (eds.). *Nuevos géneros discursivos: los textos electrónicos*, 45-61. Madrid: Biblioteca nueva.

Baroni, Marco y Motoko Ueyama. 2006. Building general -and special- purpose corpora by Web crawling. *Proceedings of the NIJL International Symposyum, Language Corpora: Their Compilation and Application*, 31-40, http://home.sslmit.unibo.it/~baroni/publications/bu_wac_kokken_formatted.pdf (3 de febrero de 2016)

Baroni, Marco y Silvia Bernardini (eds.). 2006. *Wacky! Working Papers on the Web as Corpus*. Bologna: Gedit, http://wackybook.sslmit.unibo.it/ (9 de septiembre de 2015)

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi y Eros Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43, 3: 209-226.

Calero Vaquera, María Luisa. 2014. El discurso del Whatsapp: entre el Messenger y el SMS. *Oralia: Análisis del discurso oral* 17: 87-116.

Chklovsky, Timothy y Patrick Pantel. 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.

Crystal, David. 2006. *Language and the Internet*. Cambridge: University Press.

De Groc, Clement. 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 Schryver, Gilees Maurice. 2002. Web for / as corpus: a perspective for the African languages. *Nordic Journal of African Studies* 11: 266–282.

Dumais, Susan, Michele Banko, E. Brill, Jimmy Lin y Andrew Ng. 2002. Web question answering: is more always better? En *Proceedings of the 25th ACM SIGIR*, 291-298, Tampere, Finland: ACM. Edinburgh University Press.

Fletcher, William H. 2001. Concordancing the Web with KWiCFinder. *American Association for Applied Corpus Linguistics Third North American Symposium on corpus Linguistics and Language Teaching*, 23-25. Boston, MA.

Fletcher, William H. 2004a. Facilitating the Compilation and Dissemination of Ad-Hoc Web Corpora. En G. Aston, S. Bernardini y D. Stewart (eds.). *Corpora and Language Learners*, 271–300. Amsterdam: John Benjamins.

Fletcher, William H. 2004b. Making the web more useful as a source for linguistic corpora. En U. Connor y T. Upton (eds.). *Applied corpus linguistics: a multidimensional perspective*, 191–205. Amsterdam: Rodopi.

Fletcher, William. H. 2007. Concordancing the web: promise and problems, tools and techniques. En Marianne Hundt, Nadja Nesselhauf y Caroline Biewer (eds.). *Corpus linguistics and the web*, 25-46. Amsterdam: Rodopi.

Fletcher, William. H. 2012. Corpus Analysis of the World Wide Web. En C. Chapelle (ed.). *Encyclopedia of Applied Linguistics*. London: Wiley-Blackwell.

Fujii, Atsushi y Tetsuya Ishikawa. 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*, 448-495. Hong Kong: ACM.

Galán, Carmen. 2007. Cnct kn nstrs: los SMS universitarios. *Revista de Estudios de Juventud, Instituto de la Juventud (InJuve)* 78: 63-73.

Gatto, Maristella. 2008. *From body to web: an introduction to the web as corpus*. Bari: Laterza.

Gatto, Maristella. 2014. *The web as corpus: theory and practice*. London: Bloomsbury Academic.

Gómez Torrego, Leonardo. 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, http://congresosdelalengua.es/valladolid/ponencias/nuevas_fronteras_del_espanol/4_lengua_y_escritura/gomez_l.htm (30 de noviembre de 2015)

González Fernández, Adela. 2016. *Más allá del corpus: Big data en la investigación lingüística. Evolución, análisis y predicción del uso de la lengua a través de Twitter.* Tesis doctoral, Universidad de Córdoba, Córdoba, España.

Greenwood, M., I. Roberts y R. Gaizauskas. 2002. University of sheffield trec 2002 q & a system. En E. M. Voorhees y L. P. Buckland (eds.). *The Eleventh Text REtrieval Conference (TREC-11)*, Washington. U.S. Government Printing Office. NIST Special Publication.

Grefenstette, Gregory. 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.

Guevara, Emiliano. 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.

Hundt, Marianne, Nadja Nesselhauf y Caroline Biewer (eds.). 2007. *Corpus linguistics and the web*. Amsterdam: Rodopi.

Internet Live Stats. 2015. *Top Ten Languages Used in the Web*, http://www.internetlivestats.com/ (8 de agosto de 2015)

Jones, Rayid y Rosie Ghani. 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.

Kehoe, Andrew y Antoinette Renouf. 2002. WebCorp: Applying the web to Linguistics and Linguistics to the Web. *WWW2002 Conference*. Honolulu, Hawaii.

Keller, F., M. Lapata y O. Ourioupina. 2002. Using the Web to Overcome Data Sparseness. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 230-237. Philadelphia: ACM.

Kilgarriff, Adam y Gregori Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational linguistics* 29, 3: 333-347. http://acl.ldc.upenn.edu/J/J03/J03-3001.pdf (3 de septiembre de 2015)

Kilgarriff, Adam. 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, http://ucrel.lancs.ac.uk/publications/CL2003/CL2001%20conference/papers/kilg arri.pdf (3 de febrero de 2016)

Lapata, Mireia y Frank Keller. 2005. Web-based Models for Natural Language Processing. *ACM Transactions on Speech and Language Processing* 2, 1: 1-31.

López Quero, Salvador. 2003. *El lenguaje de los "chats". Aspectos gramaticales*. Granada: Port-Royal Ediciones [Colección Lingüística].

Lüdeling, Anke, Stefan Evert y Marco Baroni. 2007. Using Web Data for Linguistic Purposes. En Hundt, Marianne, Nadja Nesselhauf y Caroline Biewer (eds.). *Corpus linguistics and the web*, 7-24. Amsterdam: Rodopi.

McEnery, Tony, y Andrew Hardie. 2012. *Corpus linguistics: Theory, Method, and Practice*. Cambridge: University Press.

McEnery, Tony y Andrew Wilson. 2003. *Corpus linguistics* (2nd Edition*)*. Edinburgh: University Press.

Milhacea, Rada y Dan I. Moldovan. 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, https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.acl99.pdf (3 de marzo de 2016)

Nakov, Preslav y Marti Hearst. 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.

Renouf, Antoinette y Andrew Kehoe (eds.). 2006. *The changing face of corpus linguistics*. Amsterdam: Rodopi.

Renouf, Antoinette, Andrew Kehoe y Jay Banerjee. 2007. WebCorp: an integrated system for web text search. En M. Hundt, N. Nesselhauf y C. Biewer (eds.). *Corpus linguistics and the web*, 47-68. Amsterdam: Rodopi.

Resnik, Philip. 1999. Mining the Web for Bilingual Text. *Proceedings of the AMTA-98 Conference*, http://www.aclweb.org/anthology/P99-1068 (15 de diciembre de 2015)

Rigau, German, Bernardo Magnini, Eneko Agirre y John Carroll. 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.

Rock, Frances. 2001. Policy and practice in the anonymization of linguistic data. *International Journal of Corpus Linguistics* 6, 1: 1-26.

Rojo, Guillermo. 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, http://gramatica.usc.es/~grojo/Publicaciones/Lgca_corpus_lgca_espanol.pdf (28 de enero de 2016)

Schäfer, Roland, Adrien Barbaresi y Felix Bildhauer. 2014. Focused web corpus crawling. *Proceedings of the 9th Web as Corpus Workshop,* 9-15. Stroudsburg, PA: ACM.

Sinclair, John. 2005. Corpus and Text - Basic Principles. En M. Wynne (ed.). *Developing Linguistic Corpora: a Guide to Good Practice,* 1-16. Oxford: Oxbow Books.

Suchomel, Vít y Jan Pomikálek. 2012. Efficient Web Crawling for Large Text Corpora. *Proceedings of the 7th Web as Corpus Workshop,* 39-43. Lyon, France.

Turney, Peter David. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. *European Conference on Machine Learning*, 491-502.

Varantola, Krista. 2002. Disposable corpora as intelligent tools in translation. En S. E. O. Tagnin (ed.). *Cadernos de Traduçao: Corpora e Traduçao* 1, 9: 171-189.

Villaseñor Pineda, Luis, Manuel Montes Gómez, Manuel Pérez Coutiño y Dominique Vaufreydaz. 2003. A Corpus Balancing Method for Language Model Construction. *Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico, http://www-prima.inrialpes.fr/Vaufreydaz/Telechargement/Villasenor03a.pdf (15 de marzo de 2016)

Volk, Martin. 2001 . Exploiting the WWW as a corpus to resolve PP attachment ambiguities. En *Proceedings of Corpus Linguistics 2001*, Lancaster, UK. http://www.zora.uzh.ch/20269/2/Volk_2001V.pdf (3 de marzo de 2017)

Yus, Francisco. 2001. *Ciberpragmática. El uso del lenguaje en Internet*. Barcelona: Ariel.

Zheng, Zhiyong. 2002. AnswerBus Question Answering System. *Proceedings of the 2nd International Conference on Human Language Technology Research.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Zipf, George. 1965. *The Psycho-Biology Of Language*. Cambridge: MIT Press.