Carlota de Benito Moreno
Ana Estrada Arráez

➲ Introducción. Si Alvar levantara la cabeza: cómo tratar con informantes virtuales

La idea de este monográfico surgió al observar que cada vez resulta más frecuente encontrar trabajos que recurren a datos obtenidos en la red para estudiar o ilustrar los fenómenos lingüísticos de los que se ocupan (Méndez García de Paredes 2011, Di Tullio 2011, y especialmente en inglés, Grieve/Nini/Guo 2016, Jones 2015, Gillen/Merchant 2012, Kilgariff/Grefenstette 2003). Esto puede resultar sorprendente *a priori,* teniendo en cuenta que la cantidad de corpus —escritos, orales; sincrónicos, diacrónicos, etc.—, diseñados a propósito para la investigación lingüística, no deja de aumentar. Sin embargo, a pesar de esta creciente disponibilidad de datos en formatos ya diseñados para la investigación lingüística, la documentación de ciertos fenómenos sigue siendo una tarea complicada con los corpus al uso, pues muchas veces estos resultan insuficientes. Esto puede suceder por varias razones, como el hecho de que, cuanto más subestándar es un fenómeno, más difícil es de documentar, incluso en corpus orales que incluyen diversos perfiles sociolingüísticos. Lo mismo ocurre con las categorías gramaticales que solo afloran en contextos determinados: por ejemplo, aquellos fenómenos que afectan a la 2ª persona del plural son especialmente difíciles de documentar, pues se requiere la existencia de un interlocutor plural (véase Lara Bermejo 2015). Asimismo, y como es sabido, recabar datos para la investigación sintáctica o léxica exige cantidades de texto mucho mayores que para la investigación fonético-fonológica, por el simple hecho de que la frecuencia en el discurso de las unidades de estos distintos niveles gramaticales difiere sustancialmente entre sí.

Desde este punto de vista, por tanto, resulta perfectamente comprensible que los investigadores recurran a fuentes alternativas de datos, aunque no hayan sido diseñadas específicamente para la investigación. La red es una candidata natural para esta tarea, por el enorme caudal de textos (de muy diversa naturaleza) que acumula, a los cuales es relativamente sencillo y rápido acceder (Kilgarriff/Grefenstette 2003).

Sin embargo, las diferencias que existen entre internet y los corpus lingüísticos exigen una reflexión previa sobre dichos datos (cf. Morala 2002, 2003, Villayandre Llamazares 2003). Por poner un ejemplo, no es extraño ver mencionado a Google como fuente de los ejemplos de un trabajo, en un uso similar al que se hace del CREA o el CORDE. Google es, sin embargo, profundamente distinto a estos corpus, siendo la mayor y más importante de estas diferencias que es un motor de búsqueda cuyo alcance no es un conjunto estable y cerrado de textos, sino un universo en continuo movimiento. Por lo tanto, citar "Google" como fuente de un ejemplo no ofrece garantías de que este pueda ser recuperado por el lector o el revisor, siendo la garantía tanto más débil cuanto más tiempo pase entre las búsquedas. Así, aunque Kilgarriff/Grefenstette (2003) proponen tomar la web como corpus, en tanto en cuanto es "a collection of texts" (2003: 334), esta propuesta es cuanto menos polémica, ya que suele entenderse que un corpus debe tener un diseño específico (que, entre otras cosas, justifique su representatividad) (Rojo 2016).

Otra consecuencia de la mutabilidad de los textos de la red (un espacio cuyos contenidos cambian cada día y donde se crean, editan y desaparecen dominios continuamente) es el hecho de que es imposible conocer su composición exacta, lo que dificulta sobremanera la cuantificación de los datos obtenidos (pero véase el trabajo de Octavio de Toledo y Huerta en esta sección temática para una propuesta). De hecho, no solo mutan los textos, sino que varía también su posición en los resultados de las búsquedas, siguiendo los designios del algoritmo de Google, que, a pesar de su utilidad indudable en la vida diaria, supone un impedimento para el lingüista, en cuanto que pone una nueva traba para conocer el conjunto de textos que analiza.

Asimismo, esta mutabilidad y volatilidad de los datos recogidos de la red implican la necesidad de continuar la práctica —ya establecida para datos de corpus en línea, como el CREA o el CORDE— de ofrecer la dirección URL y la fecha en las que se han encontrado (como hace, por ejemplo, Méndez García de Paredes 2011). Yendo un paso más allá, autores como Villayandre Llamazares (2003) proponen incluso hacer una captura de pantalla de cada texto relevante para tratar de sortear estos problemas, solución tremendamente costosa en términos de espacio, por el elevado peso que tienen las imágenes digitales. Un término medio, que no parece descabellado, podría ser descargar las páginas en que se localizan los datos recabados, creando así una base de datos específica para el fenómeno estudiado —y que puede también ponerse a disposición de otros investigadores—.

No somos, desde luego, las primeras en interesarnos por estas cuestiones. Diversos autores, como los ya mencionados Morala (2002, 2003) o Villayandre (2003), han llamado la atención sobre algunas de las dificultades metodológicas que plantea la investigación lingüística a partir de datos de la red. No podemos tampoco ignorar el creciente caudal de trabajos que, desde la década de los 90 —comenzando con los pioneros de Susan C. Herring—, han tratado la comunicación virtual como un objeto de investigación en sí mismo: la llamada comunicación mediada por ordenador o por computadora (CMO o CMC, para el español son esenciales los trabajos de Mancera Rueda/Pano Alamán 2013, Pano Alamán/Moya Muñoz 2016). Es, sobre todo, en el ámbito de este último campo de estudio donde se han llevado a cabo mayores avances en la compilación y sistematización de datos procedentes de la red con el objetivo de crear un corpus (para el español, un destacado ejemplo es el corpus de correos electrónicos compilado por Vela Delfa 2006). En una línea similar, encontramos el repositiorio colaborativo CoDiCE (Cantamutto/Vela Delfa/Boisselier 2016) o el proyecto CorpusRedEs, en proceso de elaboración (Pano Alamán/Moya Muñoz 2015), que buscan recoger muestras del español de diversas plataformas virtuales y tienen un detallado protocolo para la anotación de dichas muestras. El corpus esTenTen (Sketch Engine 2011) se sitúa en el extremo opuesto, recogiendo abundantísimas muestras de webs de todas las regiones hispanohablantes (existen versiones también para otras lenguas), pero con un diseño mínimo: si bien contiene alrededor de 11 000 millones de palabras, el filtrado por metadatos está muy restringido: no se conocen datos como la fecha o el autor, aunque se da el enlace de la web original.

Esta sección temática no busca, sin embargo, indagar en las peculiaridades de la CMO, sino ofrecer una serie de trabajos que persiguen describir la variación lingüística de la lengua *general* (entendida aquí como no exclusivamente virtual) utilizando datos tomados de la red. En las siguientes páginas ofrecemos algunas reflexiones generales sobre la utilidad de este tipo de datos para la investigación de la variación y el cambio lingüísticos, así como sobre las precauciones que debe tomar el lingüista al enfrentarse a ellos.

Los datos obtenidos de la red se prestan, por sus características, a tres funciones esencialmente distintas: a) comprobar la existencia de una estructura lingüística o ejemplificarla, b) crear una base de datos de un fenómeno lingüístico concreto —algunas veces con el objeto de completar la información proporcionada por otros corpus lingüísticos— y c) crear un corpus para la investigación lingüística.

En el primer caso, la inconmensurable cantidad de material lingüístico disponible en internet facilita enormemente la obtención de ejemplos *ad hoc*, ya sea para ejemplificar usos que no se encuentran fácilmente en otros corpus y evitar el empleo de ejemplos fabricados al efecto por el investigador, ya sea para demostrar que una estructura sí existe en la lengua. Lo primero lo hace Di Tullio (2011) en su estudio de deísmo, cuyos ejemplos están tomados de la red, mientras que lo segundo lo hace De Benito (2015), localizando casos del llamado *se* aspectual que se habían considerado agramaticales. Estos ejemplos pueden rastrearse fácilmente en cualquier portal con un buscador que permita realizar búsquedas por material lingüístico, como Google o Twitter. Mientras que el primero accede a una muestra mucho más amplia de la población (tanto textual como demográfica), el segundo presenta la ventaja de permitir más fácilmente localizar las características del emisor del ejemplo.

El siguiente paso lógico en la investigación nos llevaría a las otras dos funciones de los datos de la red, ambas consistentes en compilar un corpus que permita realizar un estudio más en profundidad. Este corpus puede ser de dos tipos y, aunque normalmente se emplee el mismo término para denominarlos a ambos, puede resultar útil diferenciarlos.

En una de estas acepciones, el corpus está compuesto únicamente de ejemplos del fenómeno —ya sea una variable o una variante— estudiado: es al que nos referimos con frases como "Compilamos un corpus de 1673 ejemplos de alternancia entre indicativo y subjuntivo". Cuando esta compilación se realiza a partir de datos de la red, es fundamental tener en cuenta la falta de control que tiene el investigador sobre el conjunto total de textos en que se realiza la búsqueda, pues esta dificulta la evaluación de la representatividad de los datos obtenidos. Tenemos un ejemplo en el trabajo de Octavio de Toledo y Huerta en esta sección temática, en el que se obtienen ejemplos de *algotro* a partir de búsquedas en Google y en Google Libros. Para maximizar el control sobre el corpus total de rastreo, el autor realiza las búsquedas durante periodos fijos, diferenciando tipo de fuente y clasificando por lugar de origen; establece un número fijo de resultados a examinar y decide hacer un filtrado geográfico de estos, excluyendo los de países sobrerrepresentados en cada búsqueda. En el caso de Twitter, el control sobre la población de textos en que se realiza la búsqueda es algo mayor, debido a que existen estadísticas sobre el tipo de usuarios y el volumen de tuiteo dentro de esta red social.[1] Una forma habitual de restringir las búsquedas es hacerlas sobre las publicaciones de un lapso temporal determinado, como hacen en este volumen Estrada/De Benito para el estudio del *puto* intensificador; Tinoco/Perea, al investigar formas léxicas dialectales del catalán, y Gonçalves/Sánchez al ocuparse de las diferencias léxicas del español a gran escala. Esta falta de control sobre la masa de datos en la que se realizan las búsquedas tiene consecuencias en la cuantificación, que deben solucionarse de otra forma. Estrada/De Benito normalizan los datos de acuerdo con la población del territorio estudiado, para evitar la distorsión que pueda producir un mayor

[1] Incluso existen trabajos que muestran las diferencias en los perfiles de los usuarios según estos hayan o no permitido la geolocalización de sus tuits y por qué medios (Pavalanathan/Eisenstein 2015).

número de habitantes (y. por tanto. de tuiteros) en los datos. Algo diferente es el caso del trabajo de Pato/De Benito (en prensa) sobre la metátesis que se produce en las combinaciones de pronombres *nos lo/os lo*: por tratarse de un fenómeno de bajísima frecuencia, obtienen exhaustivamente todos los ejemplos que ofrece el buscador de Twitter. Si bien esto parece aumentar el control sobre el volumen total de textos de los que se extraen los ejemplos, las reglas que sigue el buscador de Twitter son difusas y no se encuentran especificadas en ningún lugar. El problema surge en la comparación con las formas estándar, de frecuencia muchísimo mayor: Pato/De Benito (en prensa) minimizan la cantidad de resultados que se obtienen restringiendo la búsqueda a un periodo de un año, algo que complica la comparabilidad de ambas muestras y que los investigadores evitan comparando únicamente frecuencias relativas dentro de cada subcorpus.

Cuando el número de ejemplos que componen el corpus es especialmente alto, como ocurre con los millones de ejemplos utilizados en los trabajos de Gonçalves/Sánchez (2014, y en esta sección temática), puede restársele importancia a la cuestión de la representatividad, porque la profundidad del análisis es esencialmente distinta. Así, estos autores no se centran en la distribución de las distintas variantes de una misma variable, sino en la de un conjunto de variables. Los trabajos de Gonçalves/Sánchez iluminan nuestra comprensión de la distribución de la variación léxica atendiendo a variables como el eje urbano-rural, íntimamente relacionado con la densidad de población. Estos autores demuestran que en las zonas urbanas —y, por ende, más pobladas— la variación es mayor, pues en ellas se documenta el mayor número de variantes, incluso de una misma variable. En las zonas rurales, sin embargo, se observa un grado de variación menor; es decir, en ellas se documentan menos variantes de cada variable. De esto se deduce que son todavía las zonas rurales las que mejor se prestan al estudio dialectal, pues preservan mejor las variantes propias. En cambio, en entornos urbanos las diferencias dialectales pueden estar más diluidas, seguramente debido a la confluencia de hablantes de diversas zonas.

La segunda acepción de *corpus* a la que nos referimos, que es en realidad la primaria, es aquella que ya hemos mencionado: una compilación de datos idealmente representativa de uno o varios medios virtuales. Ya hemos hablado anteriormente de los corpus CoDiCe, CorpusRedEs y esTenTen, que son corpus con una vocación general, pero también existe la posibilidad de crear un corpus cerrado *ad hoc* para el estudio de un fenómeno concreto y que permita la cuantificación. Esto es lo que hace Vela Delfa (2006) en su tesis sobre las características de la comunicación por correo electrónico, o Chariatte (2014) cuando estudia las características fonético-gráficas del español malagueño en Facebook. Igual técnica emplea Morala (2003), que compila un corpus de crónicas periodísticas en línea sobre el Mundial de fútbol de 2002, procedentes de 13 países diferentes, con el objetivo de comparar la variación léxica en el campo léxico del fútbol entre las distintas regiones que hablan español. En este volumen, Estrada/De Benito también compilan una serie de pequeños corpus extrayendo datos de Twitter, restringidos únicamente por su geolocalización, para investigar la frecuencia del uso de *se* con valor de 2ª persona del plural tanto en español oriental como en catalán. Iglesias (2016) compila dos corpus con la producción de sendas blogueras en su investigación sobre la subida de clíticos. La ventaja más obvia de este modo de proceder es que permite cuantificar las diferentes variantes de una variable con más facilidad, pero la creación de un corpus representativo y suficiente es una tarea notablemente más traba-

josa. La red ofrece, además, numerosas posibilidades en cuanto a los niveles diafásicos a documentar, ya sean elevados (como los textos periodísticos y algunos blogs, esencialmente dependiendo de su temática) o más coloquiales y conversacionales (como el habla documentada en Facebook o Twitter). Aunque todos estos ejemplos pertenecen al medio escrito, en internet también se encuentran abundantes muestras de lengua oral que pueden utilizarse para compilar un corpus de estas características, ya sea gracias a los vídeos de Youtube o a los conocidos como "vlogs" —videoblogs—, tarea que resulta, desde luego, mucho más exigente para el investigador, debido a la necesidad de transcribir los materiales para poder realizar búsquedas lingüísticas.

Tanto si decidimos compilar un corpus específico de un fenómeno como uno más general, una división que conviene tener en cuenta es la de contextos conversacionales frente a no conversacionales. El tipo de elementos lingüísticos que aparecerán en unos y otros no será, por supuesto, comparable, y en casos como el que ya hemos mencionado sobre el estudio de la 2ª persona del plural esto puede ser trascendental.[2] Dado que nos interesa especialmente la variación lingüística y que esta tiende a ser más acusada en los contextos de inmediatez comunicativa (Koch/Oesterreicher 2007), centraremos nuestro interés en los datos conversacionales virtuales, de cuyas peculiaridades hacemos algunas breves observaciones a continuación.

Una de las grandes revoluciones que ha traído internet es la posibilidad de establecer nuevos contactos sin necesidad de movernos del sofá. Ciertas plataformas, como Youtube, Twitter o Curious Cat, todas con evidente vocación interaccional, propician la comunicación con personas con la que los usuarios no habían tenido contacto previo (constituyendo redes centrífugas, de acuerdo con Cantamutto/Vela Delfa 2016). Dentro de este grupo, al que podríamos denominar "de comunicación abierta", cabría hacer una división entre aquellas plataformas que tienen una motivación sincrónica y las que tienen una motivación asincrónica. En las primeras están englobadas las que buscan el contacto con otros usuarios en un periodo relativamente reducido de tiempo, como pueden ser Twitter, Facebook o Snapchat (siendo esta última el arquetipo de sincronicidad, puesto que, en general, su contenido solo está disponible un corto periodo de tiempo tras haberse visto —entre unos segundos y 24 horas— y se borra automáticamente una vez que ese tiempo ha transcurrido). Es evidente también esta intención en aquellos sitios, como Facebook, que animan a felicitar el cumpleaños a otros usuarios, y también se puede observar fácilmente en Twitter, cuyo propósito primero era la difusión inmediata de información y donde las búsquedas simples de tuits devuelven únicamente los más recientes. Las que tienen una motivación asincrónica, por su parte, no buscan necesariamente un contacto rápido entre usuarios. Es el caso de los vídeos de Youtube o de los blogs en general, donde se publican contenidos que pueden tardar años en llegar a usuarios interesados, que pueden elegir responder con un comentario a pesar del lapso temporal transcurrido.

Este tipo de comunicación abierta, en busca de nuevos contactos, contrasta con una comunicación que podríamos denominar "cerrada" y que se corresponde con herramien-

---

[2] Debemos aclarar aquí que nuestro concepto de contextos conversacionales se corresponde a grandes rasgos con lo que Cantamutto/Vela Delfa (2016) denominan "transitividad", es decir, la existencia de dinamismo dialógico y la posibilidad de que el interlocutor pueda responder al emisor través de la misma plataforma. Estas autoras proponen una clasificación del discurso digital según el modo de realización, las relaciones interpersonales y la enunciación, clasificación que puede resultar extremadamente útil a la hora de enfrentarse a los datos digitales.

tas como los mensajes de Whatsapp, el *e-mail* o los SMS, pensadas sobre todo para mantener comunicación con personas que, de una u otra manera, ya están dentro de alguno de nuestros círculos. De nuevo aquí se podría hacer la distinción entre motivación sincrónica (Whatsapp o SMS)[3] y asincrónica (*e-mail*).

Esta diferenciación entre contextos abiertos y cerrados dentro de las situaciones conversacionales virtuales puede desgranarse más siguiendo la clasificación que Cantamutto/Vela Delfa (2016) proponen para los datos digitales, especialmente atendiendo al apartado en el que definen los diferentes tipos de relaciones interpersonales. Queremos insistir, en cualquier caso, en que estas distinciones entre comunicación abierta y cerrada o sincrónica y asincrónica no son, por supuesto, rígidas, sino que se trata de un continuo que depende esencialmente de características individuales de los usuarios. En lo que hemos denominado "comunicación cerrada", por ejemplo, nuestras expectativas en cuanto a la fluidez conversacional son mayores, ya que, en general, suele haber una relación previa con la persona a la que nos dirigimos y, por tanto, es de esperar que los límites de la asincronía en la comunicación sean menores que para la comunicación abierta.

Para los investigadores de la lengua es especialmente relevante, además, la facilidad de acceso a los datos. Muchas de las aplicaciones que hemos mencionado tienen el inconveniente de ser privadas, como Facebook o Whatsapp, por lo que es necesario pedir permiso a los usuarios para acceder a ellas. El acceso a estos contenidos, aunque trabajoso, no es, sin embargo, imposible, como demuestra la existencia de proyectos dedicados a su estudio.[4] Esta dificultad se ve, además, compensada por el hecho de que, al ser espacios privados, el grado de inmediatez comunicativa es mayor, lo que nos hace esperar un lenguaje más cercano a la oralidad. Asimismo, estas fuentes de datos ofrecen la posibilidad de seleccionar el tipo de hablante y, en casos como Facebook, de obtener cierta información sobre ellos difícilmente accesible con otras fuentes de datos (edad, nivel educativo, lugar de origen y residencia, etc.).

Frente a las aplicaciones de uso privado, existen otras diseñadas para ser públicas, donde cabe destacar la red social Twitter. El inmenso éxito que está teniendo esta plataforma en los estudios sobre el lenguaje se debe sobre todo a la facilidad de acceso a los datos, así como en la creación de diversas herramientas de búsqueda y extracción de los mismos (también dependientes de su carácter público). Los artículos de Perea/Ruiz Tinoco, Sánchez/Gonçalves y Estrada/De Benito en esta sección monográfica son una clara muestra de dicho éxito. Pueden ponerse, sin embargo, dos grandes *peros* a los datos de esta red: en primer lugar, pueden no ser tan conversacionales como esperaríamos, ya que existen numerosas cuentas institucionales o empresariales que no buscan la interacción, sino meramente ofrecer información, y, en segundo lugar, hay cierta complejidad técnica a la hora de extraerlos, ya que exige tener tanto conocimientos de programación (ya sea en R, en Python, en Java, etc.) como de manejo de bases de datos complejas (como MySQL), tanto más esenciales cuanto mayor sea la cantidad de datos manejada. Los artículos de Perea/Ruiz Tinoco y Estrada/De Benito recogidos en esta sección temática explican en detalle las distintas técnicas empleadas para extraer tuits, usando los lenguajes de programación

---

[3] Hoy en día existen diversos proyectos para documentar e investigar la lengua de estas plataformas, para varios idiomas, como *sms4science* (<http://www.sms4science.org>) y *What's Up Switzerland?* (<http://www.whatsup-switzerland.ch>).

[4] Además de los proyectos dedicados a Whatsapp y los SMS ya mencionados, hay estudios dedicados a datos de otras fuentes privadas, como el de Nadine Chariatte sobre Facebook (2014).

Java y R, respectivamente, lo que esperamos que sea de utilidad a aquellos investigadores que decidan iniciarse en la recolección de este tipo de datos.[5]

Otra de las desventajas que debemos mencionar sobre el uso de datos obtenidos directamente de la red se refiere al hecho de que estos, frente a lo que ocurre con una gran parte de los corpus lingüísticos, no están lematizados o etiquetados, lo que redunda en una mayor complejidad de las búsquedas. Por un lado, no hay que olvidar que esta dificultad nos ha acompañado a los lingüistas durante años (y sigue haciéndolo): en el caso del español, por ejemplo, no ha sido hasta hace pocos años que la RAE —en 2013— ha puesto a disposición del público general corpus lematizados y etiquetados, como el CORPES XXI y las versiones del CREA y el CORDE, consultables ahora en el *Corpus del Diccionario Histórico*, CDH, y sigue siendo habitual que los corpus a disposición del público no estén lematizados, al menos en sus primeras fases. Esta es, por tanto, una dificultad que podemos relativizar. Sí debemos subrayar el hecho de que, puesto que es el propio investigador el que recoge los datos en bruto, no existe siquiera una interfaz que facilite las búsquedas, por lo que el manejo de expresiones regulares puede resultar útil (véase aquí también el empleo que hacen de estas Estrada/De Benito en este trabajo). En este sentido, debemos recordar también la elevada frecuencia de ortografías subestándar que se observan en la lengua de la red (Mancera Rueda/Pano Alamán 2013), algo que debe ser tenido en cuenta para tratar de incluir todas las posibilidades relevantes en las búsquedas (a este respecto, son especialmente interesantes las consideraciones de Jones (2015: 414) sobre la obtención de formas ortográficamente subestándar en *AAVE —African American Vernacular English—*).

Una posibilidad, especialmente útil cuando se trabaja con grandes cantidades de datos, es la de lematizar y etiquetar el corpus (así hacen, por ejemplo, Kulkarni/Perozzi/Skiena 2016 para estudiar la variación léxica y semántica de diferentes variedades del inglés), tarea también compleja —y en la que deben tenerse en cuenta las mismas precauciones relacionadas con las realizaciones ortográficas subestándares—, que creemos que solo merece la pena para la compilación de corpus en su sentido primario: es decir, de corpus general para el estudio diversos fenómenos.

Encontramos, pues, diversas dificultades técnicas en el manejo de grandes cantidades de datos. El *big data* ha llegado a la lingüística para quedarse, lo que sobre todo supone una dificultad para aquellos cuya formación originaria es filológica o lingüística, pues solemos carecer de los conocimientos técnicos necesarios para enfrentarnos a estas grandes cantidades de datos. Esta dificultad no es tal, sin embargo, para especialistas de otros ámbitos científicos, especialmente en el ámbito de las ciencias infor-

---

[5] En este sentido nos gustaría incidir sobre los problemas que pueden surgir al exportar los datos obtenidos de la API de Twitter (en formato .json) a los formatos de hojas de cálculo con los que solemos estar familiarizados (.xls, .csv). La existencia de saltos de línea (cambios de renglón) dentro de los tuits o de los perfiles del usuario (los campos *text* y *description*) causan con frecuencia problemas en la conversión de estos campos, que pasan a dividirse en varias celdas, descuadrando así la tabla en que se organizan los datos. Cuando el número de datos no es excesivo y estos pueden analizarse manualmente, el problema puede corregirse también de forma manual (como hacen Estrada/De Benito), pero, para cantidades más abundantes de datos, es recomendable realizar la exportación desde los archivos en formato .json a una base de datos MySQL, donde estas complicaciones no surgen (véase el procedimiento seguido por Perea/Ruiz Tinoco). Si bien la jerga técnica puede resultar descorazonadora para el novicio, esperamos que los trabajos aquí contenidos muestren que las alegrías producidas por llegar a dominar estas técnicas superan a las dificultades iniciales.

máticas, que están haciendo trabajos interesantísimos sobre la lengua de la red (para el caso del español, son fundamentales los trabajos de Gonçalves/Sánchez 2014 y en esta sección temática; para el inglés hallamos una auténtica proliferación de este tipo de trabajos, algunos de los cuales son Doyle 2014, Eisenstein *et al.* 2014, Kulkarni/Perozzi/Skiena 2016).

A las complejidades técnicas se unen, además, algunas de carácter lingüístico. En primer lugar, nos enfrentamos a la cuestión de la autoría: parece paradójico que una de las cuestiones más antiguas de la filología resurja con tanta importancia en los datos del siglo XXI, pero la lingüística moderna, gracias a los aportes de la sociolingüística y dialectología variacionistas a la cuestión del cambio y la variación lingüísticas, ha demostrado la necesidad de conocer bien tanto la procedencia geográfica y social de los autores como el contexto de empleo para poder comprender correctamente un fenómeno lingüístico.[6] El trabajo de Octavio de Toledo y Huerta en esta sección temática es un ejemplo modélico de cómo la labor filológica cuidadosa es esencial para desgranar la información obtenida en la red e interpretarla correctamente.

En segundo lugar, también debemos ocuparnos de la cuestión de las motivaciones tras la utilización de formas no normativas en la lengua documentada en internet: ¿son usos lúdicos o pueden considerarse *genuinamente dialectales*? La reutilización de formas subestándar con motivos lúdicos es un fenómeno común en el habla y también se documenta con abundantísima frecuencia en la comunicación virtual, cuestión de la que se ocupan con cierto detalle Estrada/De Benito en esta sección temática.

En tercer lugar, otra preocupación común a los investigadores que emplean datos de la red es la de qué hacer con los datos repetidos, ya procedan de Google —donde reflejan normalmente redundancias del buscador— o de Twitter —donde suele tratarse de retuits, "plagios" o, incluso, *spam*—. Las soluciones a este problema pueden ser variadas: Octavio de Toledo y Huerta no considera las redundancias producidas por Google, igual que Estrada/De Benito deciden excluir las repeticiones obtenidas de la API de Twitter, pero no faltan los autores que consideran que estas son muestras de la difusión del fenómeno buscado y, por ello, deben incluirse (Grieve/Nini/Guo 2016). Esta cuestión debe dilucidarse teniendo en cuenta los objetivos de la investigación.

Todas estas cuestiones, ya sean de carácter técnico, ya sean de carácter lingüístico, nos parecen irrenunciables para el estudio de la variación con datos obtenidos de internet. Así, queremos finalizar esta introducción no con una conclusión, sino con una reflexión. La doble complejidad de los datos, técnica por un lado y lingüística y filológica por otro, parece estar produciendo una bifurcación de la investigación: un camino explota las posibilidades del *big data*, con la idea de que un número suficientemente elevado de datos permite observar generalizaciones interesantes y considerar como "ruido" los ejemplos que no se ajustan a estas, mientras que el otro prefiere aplicar las precauciones propias de la filología y combina un cuidadoso análisis cualitativo con el análisis cuantitativo (que, necesariamente, se ve más limitado por el hecho de que el número de datos ha de ser menor). Creemos que ambas vías

---

[6] La lingüística histórica hispánica moderna parece estar especializándose en demostrar la importancia de estas características, como muestran las abundantes revisiones de explicaciones tradicionales gracias al uso de variables dialectales y sociolingüísticas más detalladas, por ejemplo, en el caso del origen del leísmo, laísmo y loísmo (Fernández-Ordóñez 2001, Matute 2004); del supuesto triunfo del femenino en los posesivos (Espinosa Elorza 2002, Del Barrio 2014) o de la gramaticalización de los tiempos compuestos (Rodríguez Molina 2010).

ofrecen resultados sumamente interesantes y que estos lo serán todavía más si ambos esfuerzos se combinan. Así lo sugieren los trabajos de Jones (2015), en que los casi 18 000 tuits empleados para estudiar las diferencias dialectales del inglés vernáculo afroamericano fueron examinados a mano, lo que resultó esencial para identificar algunas variantes ortográficas relevantes, o de Kulkarni/Perozzi/Skiena (2016), que proponen un método para analizar la variación dialectal en el ámbito de la semántica a partir de los términos más frecuentes con que se relacionan las palabras investigadas, lo que no solo tiene un interés en sí mismo —el método permite *identificar* las palabras que muestran variación semántica—, sino que también puede resultar útil para refinar los trabajos sobre variación léxica a partir de *big data*.

Bibliografía

Cantamutto, Lucía/Vela Delfa, Cristina (2016): "El discurso digital como objeto de estudio: de la descripción de interfaces a la definición de propiedades", en: *Aposta: Revista de ciencias sociales* 69, 296-323.

Cantamutto, Lucía/Vela Delfa, Cristina/Boisselier, Leandro (2016): *COmunicación DIgital Corpus del Español (CoDiCE)*, < http://www.codice.aplicacionesonline.com.ar/> (25 de octubre de 2016).

Chariatte, Nadine (2014): "'Facebook Style': The use of non-standard features in virtual speech conditioned by the medium Facebook", en: Brumme, Jenny/Falbe, Sandra: *The Spoken Language in a Multimodal Context*. Berlin: Frank & Timme, 93-117.

CDH = Instituto Rafael Lapesa de la Real Academia Española (2013) : *Corpus del Nuevo diccionario histórico,* <http://www.web.frl.es/CNDHE>.

CORDE = Real Academia Española: *Corpus Diacrónico del Español*, <http://www.corpus.rae.es/cordenet.html>.

CORPES XXI = Real Academia Española: *Corpus del Español del Siglo xxi,* <http://www.rae.es>.

CREA = Real Academia Española: *Corpus de Referencia del Español Actual*, <http://www.corpus.rae.es/creanet.html>.

De Benito Moreno, Carlota (2015): *Las construcciones con* se *desde una perspectiva variacionista y dialectal*, tesis doctoral. Madrid: Universidad Autónoma de Madrid, <http://hdl.handle.net/10486/670892>.

Del Barrio de la Rosa, Florencio (2014): "Factores externos y cambio lingüístico: la pérdida de la distinción genérica en los posesivos del español antiguo", en: *Revista de Historia de la Lengua Española* 9, 3-26.

Di Tullio, Ángela (2011) : "Infinitivos introducidos por *de*", en: *Cuadernos de la ALFAL* 3, 176-187.

Doyle, Gabriel (2014): "Mapping dialectal variation by querying social media", en: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, <http://www.web.stanford.edu/~gdoyle/papers/doyle-2014-eacl.pdf>.

Eisenstein, Jacob/O'Connor, Brendam/Smith, Noah A./Xing, Eric P. (2014): "Diffusion of Lexical Change in Social Media", en: *Plos One* 9, 11, e113114.

Espinosa Elorza, Rosa María (2002): "¿Alguna vez triunfó el femenino? Revisión de los posesivos en castellano medieval", en: Veiga, Alexandre/Suárez Fernández, Mercedes (eds.): *Historiografía lingüística y gramática histórica. Gramática y léxico*. Madrid/Frankfurt a. M.: Iberoamericana/Vervuert, 9-18.

Fernández-Ordóñez, Inés (2001): "Hacia una dialectología histórica. Reflexiones sobre la historia del leísmo, el laísmo y el loísmo", en: *Boletín de la Real Academia Española* 81, 389-464.

Gonçalves, Bruno/Sánchez, David (2014): "Crowdsourcing Dialect Characterization Through Twitter", en: *PLoS ONE* 9, 11, e112074.

Grieve, Jack/Nini, Andrea/Guo, Diansheng (2016): *Analyzing lexical emergence in Modern American English*, en: *English Language and Linguistics*, 1–29. DOI: < https://doi.org/10.1017/S1360674316000113>.

Gillen, Julia/Merchant, Guy (2012): "Contact calls. Twitter as a dialogic social and linguistic practice", en: *Language Sciences* 35, 47-58, <http://www.dx.doi.org/10.1016/j.langsci.2012.04.015>.

Iglesias, Olivier (2016): "'Se le quedó mirando': la atracción de clíticos en un corpus de idiolectos (s. xix-xxi)", en: Kabatek, Johannes (ed.): *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín/Boston: De Gruyter, 424-448.

Jones, Taylor (2015): "Toward a Description of African American Vernacular English Dialect Regions Using 'Black Twitter'", en: *American Speech* 90, 4, 403-440. DOI: 10.1215/00031283-3442117.

Kilgariff, Adam/Grefenstette, Gregory (2003): "Introduction to the Special Issue on the Web as Corpus", en: *Computational Linguistics* 29, 3, 333-347.

Koch, Peter/Oesterreicher, Wulf (2007): *Lengua hablada en la Romania: español, francés, italiano*. Madrid: Gredos.

Kulkarni, Vivek/Perozzi, Bryan/Skiena, Steven (2016): "Freshman or Fresher? Quantifying the Geographic Variation of Language in Online Social Media", en: *Procceedings of the Tenth International AAAI Conference on Web and Social Media*, <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13121/128>10 (25 de octubre de 2016).

Lara Bermejo, Víctor (2015): *Los tratamientos de 2pl en Andalucía occidental y Portugal. Estudio geo- y sociolingüístico de un proceso de gramaticalización*, tesis doctoral. Madrid: Universidad Autónoma de Madrid, <http://hdl.handle.net/10486/667216>.

Mancera Rueda, Ana/Pano Alamán, Ana (2013): *El español coloquial en las redes sociales*. Madrid: Arco Libros.

Matute Martínez, Cristina (2004): *El sistema referencial de los pronombres personales en la documentación castellana medieval. Intento de reconstrucción histórica*, tesis doctoral. Madrid: Universidad Autónoma de Madrid, <http://www.corpusrural.es/publicaciones/2004/2004_sistemas.pdf>.

Méndez García de Paredes, Elena (2011): "*Si yo fuera estado allí, no fuera pasado eso*. Pervivencia de un aparente arcaísmo en la lengua de Internet", en: Bustos Tovar, José Jesús de (ed.): *Homenaje a Antonio Narbona*, vol. II. Sevilla: Universidad de Sevilla, 1009-1032.

Morala, José Ramón (2002): "Nuevas tecnologías y recursos lexicográficos: *fuereño*", en: Clavería, Gloria (ed.): *Filología en Internet*. Barcelona: Universitat Autònoma de Barcelona, 45-53.

Morala, José Ramón (2003): "Variantes del español en red", en: Perdiguero Villarreal, Hermógenes/Álvarez, Antonio A. (coords.): *Actas del XIV Congreso Internacional de ASELE*. Burgos: Universidad de Burgos, 73-87.

Pano Alamán, Ana/Moya Muñoz, Patricio (2015): "CorpusRedEs. Proyecto de creación y anotación de un corpus de comunicación mediada por ordenador en español", en: *CHIMERA. Romance Corpora and Linguistic Studies* 2, 117-129.

Pano Alamán, Ana/Moya Muñoz, Patricio (2016): "Una aproximación a los estudios sobre el discurso mediado por ordenador en lengua española", en: *Tonos Digital* 30, 1-30.

Pato, Enrique/De Benito Moreno, Carlota (en prensa): "*Tráenolos para comérnolos* o la transposición del clítico en español actual", en: *Philologica Jassyensia* 1/2017.

Pavalanathan, Umashanthi/Eisenstein, Jacob (2015): "Confounds and consequences in geotagged twitter data", en: Màrquez, Lluís/Callison-Burch, Chris/Su, Jian (eds.): *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Lisbon, Portugal: arXiv, 2138-2148, <http://www.arxiv.org/abs/1506.02275> (28 de octubre de 2015).

Rodríguez Molina, Javier (2010): *La gramaticalización de los tiempos compuestos en español antiguo: cinco cambios diacrónicos*, tesis doctoral. Madrid: Universidad Autónoma de Madrid, < https://repositorio.uam.es/handle/10486/6279?show=full>

Rojo, Guillermo (2016): "Citius, maius, melius: del CREA al CORPES XXI", en: Kabatek, Johannes (ed.): *Lingüística de corpus y lingüística histórica iberorrománica,* Berlín/Boston: De Gruyter, 197-212.

Sketch Engine (2011): *esTenTen corpus*, <https://www.the.sketchengine.co.uk/>.

Vela Delfa, Cristina (2006): *El correo electrónico: el nacimiento de un nuevo género,* tesis doctoral. Madrid: Universidad Complutense de Madrid.

Villayandre Llamazares, Milka (2003): "Internet como corpus: el caso de 'Bidibí'", en: *Contextos* 41-44, 205-231.

Zimmer, Michael (2015): "The Twitter Archive at the Library of Congress: Challenges for information practice and information policy", en: *First Monday*, <http://www.firstmonday.org/ojs/index.php/fm/article/view/5619/4653> (28 de julio de 2016).