Las redes sociales como corpus de estudio para el Análisis del discurso mediado por ordenador

Ana Mancera Rueda
Universidad de Sevilla
anamancera@us.es

Ana Pano Alamán
Università di Bologna
ana.pano@unibo.it

1. INTRODUCCIÓN

El *Análisis del discurso mediado por ordenador* (ADMO) es el acercamiento a la comunicación en la Red desde la metodología del Análisis del discurso. Su principal objetivo es analizar las propiedades lingüísticas y pragmáticas de este tipo de interacciones. Uno de los aspectos que más atención ha recibido por parte de los analistas del discurso mediado por ordenador es la naturaleza híbrida, entre oral y escrita, de los textos que se publican en el entorno digital. De ahí que en el ámbito de la lengua española Francisco Yus (2001) y Leonardo Gómez Torrego (2001), entre otros, hablen, por ejemplo, de *textos escritos oralizados* para hacer alusión a algunos de los que se difunden en la Red. Desde esta perspectiva, Ana Pano (2008) explora cómo los condicionantes situacionales pueden explicar por qué los hablantes utilizan una modalidad coloquial en contextos donde lo que prima es la inmediatez comunicativa y, más recientemente, Ana Mancera (2011) afronta el estudio de este aspecto en el periodismo digital.

Sin embargo, en la actualidad, donde de manera más recurrente se hace uso de la modalidad coloquial es en las redes sociales virtuales, ámbito en el que se centran nuestros trabajos más recientes, cuyo objeto es verificar cuáles son los rasgos coloquializadores más frecuentes en los mensajes que se publican en estos canales. Para ello, en primer lugar, hemos constituido un corpus de textos publicados en Facebook, Tuenti y Twitter considerando algunas cuestiones metodológicas relacionadas con la recopilación de datos extraídos de Internet para un análisis de tipo lingüístico-discursivo (Susan C. Herring, 2001; Jannis Androutsoupoulos y Michael Beißwenger, 2008). Algunas de estas cuestiones son la de la representatividad –que nos ha obligado a recopilar un número similar de mensajes en las tres redes sociales–, la cantidad de información contextual necesaria, y la privacidad –cuestión ética que tiene que ver con el anonimato de los usuarios de estas redes–. Vamos a examinarlas con más detenimiento.

2. LA LINGÜÍSTICA DE CORPUS Y EL ANÁLISIS DEL DISCURSO MEDIADO POR ORDENADOR

En los últimos años, con el desarrollo de la Lingüística de corpus (LC), se ha impuesto una tendencia a recopilar textos naturales y completos, tratando de alcanzar una relativa extensión y diversidad. La hoy denominada LC es un área de la lingüística que ha adquirido un espacio autónomo, y que se define como "una metodología para la investigación de las lenguas y el lenguaje, que permite llevar a cabo investigaciones empíricas en contextos auténticos, y que se constituye en torno a ciertos principios reguladores" (Giovanni Parodi, 2010: 15)[1]. Por su parte, John Sinclair (1991: 171) sostiene que un corpus es "una colección de textos de ocurrencias de lenguaje natural, escogidos para caracterizar un estado o una variedad de lengua". La LC es una técnica, cuyo fundamento es el corpus mismo. En este sentido, y de acuerdo con distintos estudios (Geoffrey Leech, 1992; Francisco Marcos Marín, 1994; Josse de Kock, 2001; Julia Lavid, 2005, entre otros), este no sería un campo ni un área de estudio, sino un ámbito de investigación focalizado en los corpus con base en metodologías diferentes, por ejemplo, la del Análisis del discurso (Paul Baker, 2006), una cuestión que también concierne a las Humanidades digitales entendidas bien como disciplina, bien como método de trabajo en el ámbito de diversas disciplinas. Entre las ventajas de utilizar corpus se han señalado las siguientes: permiten una adecuada representación del discurso en muestras amplias y representativas de textos originales; posibilitan la exploración de textos etiquetados y no etiquetados; por medio del procesamiento (semi)automático de los textos, es posible realizar análisis más amplios y detallados; ofrece mayor fiabilidad en los análisis cuantitativos y cualitativos; los resultados son acumulativos y confrontables con posteriores investigaciones. Un aspecto relevante que persiguen hoy en día los trabajos desde la LC radica en el interés por la variabilidad lingüística. Por ello, existe una tendencia a realizar investigaciones multirregistros y/o multigéneros, en las cuales es posible establecer comparaciones entre variedades distintas de una misma lengua.

Esto es más evidente, si cabe, cuando se utilizan las páginas de la web actual para constituir corpus de estudio lingüístico y discursivo (Mike Rundell, 2000; Philip Resnik y Noah Smith, 2003; William H. Fletcher, 2004). Como señalan Adam Kilgarriff y Gregory Grefenstette (2003: 33),

> "language scientists and technologists are increasingly turning to the Web as a source of language data, because it is so big, because it is the only available source for the type of language in which they are interested, or simply because it is free and instantly available".

De hecho, los principales estudios dedicados a la metodología basada en corpus aplicada al Análisis del discurso mediado por ordenador (Michael Beißwenger

[1] En el ámbito de la lengua española, destacan los corpus: Proyecto de Estudio coordinado de la norma lingüística culta de las principales ciudades de España e Ibero América; PRESEEA (Proyecto para el Estudio Sociolingüístico del Español de España y de América); CREA (Corpus de Referencia del Español Actual) y CORDE (Corpus Diacrónico del Español), de la Real Academia Española de la lengua, que prepara ahora el CORPES XXI (Corpus del español del Siglo XXI); Val.Es.Co, centrado en la lengua oral, el registro coloquial y la variedad conversacional de la lengua; COVJA, de la Universidad de Alicante y COLA, de la Universidad de Bergen, sobre las variedades juveniles del español.

y Angelika Storrer, 2008; Brian King, 2009) coinciden en que la constitución de un corpus para investigar el lenguaje en la comunicación mediada por ordenador (CMO) es una tarea aparentemente sencilla, ya que los datos están digitalizados y son accesibles fácilmente en la Red. Sin embargo, en seguida surgen incompatibilidades entre algunos parámetros de la LC y los datos recogidos en los distintos canales. Y es que los datos extraídos a partir de la web presentan un elevado grado de heterogeneidad, lo cual dificulta la obtención de un conjunto de documentos de aspecto homogéneo, estructurado y libre de "ruido"[2] (Fernando Martínez Santiago *et al.*, 2001). Por tanto, además de tener en cuenta aspectos fundamentales a la hora de constituir un corpus, como la extensión del mismo, la representatividad, o la ética en la recolección de datos, Claudia Claridge (2007) y Michael Beißwenger y Angelika Storrer (2007) advierten de que para establecer corpus de estudio de la CMO es necesario distinguir entre los datos que genera el sistema –por ejemplo, las líneas de entrada y salida de un canal de chat, o la fecha de envío de un *tuit*– y los que genera el usuario, que pueden incluir texto, enlaces, emoticones e imágenes.

A la hora de recolectar datos y constituir un corpus de textos extraídos de las redes sociales virtuales Facebook, Twitter y Tuenti hemos tenido en cuenta todos estos aspectos. Pero quizás una de las cuestiones más importantes y más controvertidas[3] que hemos considerado es la de la representatividad. Un corpus puede ofrecer información detallada acerca de una lengua particular, pero es imposible constituir uno que abarque toda una lengua. Por tanto, cabe buscar soluciones para establecer una proporcionalidad adecuada del corpus que conduzca a ciertas proyecciones, aun renunciando a realizar generalizaciones. Además, un corpus no es una única instancia comunicativa y tampoco cuenta con un cierre de ningún tipo, sino que presenta una organización predeterminada en torno a categorías identificables para la descripción de una variedad de lengua. Partiendo de estas consideraciones, nuestro objetivo ha sido el de constituir lo que Parodi llama un "pre-corpus" de mensajes extraídos de redes sociales virtuales, "con el fin de proponer hipótesis de trabajo y de explorar ciertas características o categorías para una posterior recolección más amplia y robusta" (Giovanni Parodi, 2010: 27).

Por otra parte, y siguiendo a Elena Tognini-Tonelli (2001), cabe distinguir entre el enfoque c*orpus-based* y *corpus-driven* en función de si las categorías de análisis lingüístico están previamente determinadas y enmarcadas en una opción teórica, o si emergen del análisis y dan sustento a la construcción de una teoría. En la primera opción el corpus actúa como un método de indagación y corroboración de ideas preexistentes, así, en nuestro caso, hemos adoptado ese primer enfoque para indagar sobre las marcas de la oralidad en las tres redes mencionadas y sus implicaciones en la escritura que se emplea en ellas. Con este objetivo hemos tratado de recoger una muestra lo más representativa posible de los enunciados publicados en estos espacios. El primer paso ha sido constituir dicho corpus por medio de la recopilación de 600 mensajes publicados

[2] "Conjunto de términos no pertinentes o de palabras no significativas que se obtiene en una búsqueda automática a una base de datos documental" (*Glosario de Terminología Informática*, <http://www.tugurium.com/gti>).

[3] En palabras de Adam Kilgarriff y Gregory Greffenstette: "'Representativeness' begs the question 'representative of what?' Outside very narrow, specialized domains, we do not know with any precision what existing corpora might be representative of" (2003: 333).

entre noviembre de 2011 y abril de 2013 en perfiles muy diferentes, creados tanto por jóvenes como por políticos, escritores, periodistas, profesores o empresarios. Estos datos preliminares constituyen un primer panorama singular, pues se cuenta con pocos antecedentes similares –corpus de mensajes publicados en distintos canales– producto de investigaciones de naturaleza comparativa en lengua española.

En la LC, el segundo paso prevé una descripción lo más detallada posible del contexto de recopilación de los mensajes, que contribuye a justificar la selección de estas redes frente a otros canales de comunicación en la Red, y a identificar las principales semejanzas y diferencias entre los tres espacios, parámetros esenciales de cara al análisis comparativo. Para ello hemos consultado el informe *5ª Oleada del Observatorio de redes sociales* (abril, 2013). En este documento se señala que Facebook, Twitter y Tuenti son las redes más utilizadas y con el mayor número de usuarios en España. La primera es líder, con un 83% de los entrevistados que tiene una cuenta personal y la utiliza, frente a Twitter (42%) y Tuenti (27%). Los motivos principales que llevan a utilizar estas redes son, en el caso de Facebook y Tuenti, reforzar el contacto con el círculo social cercano, actualizando el perfil y compartiendo fotografías, y en Twitter, informarse y promover o apoyar causas sociales o solidarias, además de seguir a personajes famosos. La media de edad de los usuarios es de 31 años en Facebook y de 28 en Twitter. En cambio, en Tuenti, el 59% tiene menos de 25 años. Así, las redes sociales se distinguen básicamente por su función y por el tipo de usuario, sus características y objetivos.

Para Inmaculada Berlanga y Estrella Martínez (2010), la función del lenguaje que predomina en Facebook es la conativa o apelativa, el elemento preponderante es el receptor y la intención comunicativa influir en él por medio de instrucciones, consejos y preguntas. De hecho, abundan aquí los enunciados exhortativos e interrogativos, el modo imperativo e indicativo, la segunda persona verbal y los vocativos. Decíamos que Tuenti es la red social más utilizada por adolescentes y jóvenes universitarios españoles quienes, según Alba Torrego (2010 y 2011), están ahí para tener una constante presencia virtual y para interactuar con sus amistades. Esta investigadora señala que en dicho espacio los jóvenes juegan con el lenguaje prescindiendo en cierta medida de la ortografía y de la gramática normativas. Las razones que impulsan a los jóvenes a hacer un uso no normativo de la lengua en Tuenti pueden explicarse por la comodidad o la rapidez del proceso de escritura, esto es, por la escasa planificación en la elaboración del discurso, mientras que otros fenómenos pueden obedecer a factores de naturaleza pragmática, como el deseo de llamar la atención del interlocutor o de configurar la propia identidad diferenciándose de los demás. Por último, Twitter es una plataforma de *microblogging* o *nanoblogging*, es decir, un servicio en línea que permite enviar y publicar mensajes de no más de 140 caracteres. En este canal los mensajes tienen distintas funciones comunicativas (Tíscar Lara, 2012), como las de reconocimiento, cuando se *retuitean* los textos de otros y se reconoce su autoridad sobre la información que se comparte; dialógica, que permite conversar con alguien insertando "@usuario" en el mensaje o simplemente haciendo *clic* sobre el botón *Respuesta*; discursiva, mediante la incorporación de etiquetas, facilitando el seguimiento de distintos *tuits* sobre un mismo tema; e identitaria, pues en el perfil personal del usuario aparecen fotografías y una breve descripción que le identifican.

En lo que se refiere a la privacidad, Lori Kendall (2002: 60) ha puesto de manifiesto cómo en el Análisis del discurso mediado por ordenador cabe entender dicho concepto de otro modo, quizás porque en estos espacios "talk tends to blur the distinction between public and private". Por otro lado, la cuestión de la identidad digital es muy compleja, y tiene que ver también con el carácter virtual y público del medio, y con la tendencia a reelaborar la propia identidad en las descripciones personales, por ejemplo, en clave humorística (Ana Pano y Ana Mancera, en prensa). En todo caso, sí es posible adoptar algunas soluciones relativas a la privacidad de los usuarios en la recolección de datos. Por ejemplo, hay que tener en cuenta que en el caso de Tuenti, al tratarse de una red social dirigida principalmente a adolescentes, son estos los autores y los destinatarios de la mayor parte de los mensajes. Esto ha hecho necesario proteger su privacidad de cara al análisis discursivo. Por eso hemos suprimido cualquier tipo de información que pudiera arrojar pistas sobre la identidad de los autores o los destinatarios de tales textos. Pero el caso de Twitter es diferente, ya que aquí los perfiles no suelen ser de acceso restringido y cualquier internauta puede consultar lo publicado por otros. Sin embargo, por coherencia con nuestra manera de proceder con los mensajes extraídos de Facebook y Tuenti, hemos optado por eliminar también en el caso de Twitter el nombre de la mayor parte de los *tuiteros*. Sólo mantenemos su identidad cuando se trata de personas que desempeñan una actividad pública, y cuyo conocimiento resulta relevante para interpretar los resultados del análisis.

Por otra parte, y a pesar de que somos conscientes de que la lectura de este tipo de mensajes no siempre resulta fácil, hemos optado por reproducirlos todos tal y como fueron publicados. Es decir, sin omitir los errores ortográficos ni aquellas otras muestras representativas del subcódigo escrito que está difundiéndose en Internet.

3. ANÁLISIS DEL CORPUS: ASPECTOS ORTOGRÁFICOS Y SELECCIÓN LÉXICA

La "ciber-ortografía" presenta desafíos enormes para la LC, sobre todo a la hora de etiquetar este tipo de textos, aunque empiezan ya a aparecer algunos programas especialmente diseñados para el Análisis del discurso mediado por ordenador (Brian King, 2009: 313). En su afán por recrear la lengua oral, muchos internautas redactan enunciados seseantes como el siguiente:

> "negroooo felicidadeeees!! De regalo otra vueltesita en el coche jajajaja" (Tuenti, 21-04-2013).

Y son también frecuentes los casos de aféresis,

> "Norabuena por duplicado, Antonio…" (Facebook, 20-04-2012).

o la supresión de algunos sonidos finales e incluso de sílabas enteras:

> "ya tio pero tu ere mas grande ynmas fuerte…jaja es k ere especia" (Tuenti, 18-03-2013).

De esta forma, los internautas parecen querer recrear los rasgos dialectales del español meridional, a lo que parece responder también el uso de la hache en palabras que no la llevan, tal vez para simbolizar una aspiración:

> "sevillanita miaa,ya esty en mi tierra,q *hartera* de autobus tia,como llevas esa fiestecilla??un besazoo muak" (Tuenti, 29-09-2012).

El hecho de que la representación del sonido africado palatal sordo se lleve a cabo en lugar de con el dígrafo ch, con una equis,

> "yo exo asta la encuenta y to y por eso me lo quite y se lo dixo tiaa XD" (Twitter, 22-04-2013).

o el trueque de qu por la letra k para representar al fonema oclusivo velar sordo,

> "[…] yo creo que me voy a pasar, y ya que hoy estamos todos medio muertos lo propongo como *kedada* para mañana […]" (Facebook, 09-03-2013).

e incluso la omisión de la hache

> "Xurraaaa papa ma dxo k si este año kiero aser snow! Jaja le dxo k sii! Lo ases cnmgo?" (Tuenti, 25-01-2013).

son rasgos que suelen atribuirse a la economía lingüística. Sin embargo, de las redes sociales analizadas, solamente Twitter cuenta con condicionamientos técnicos que limitan la cantidad de grafemas, lo que justificaría el uso recurrente de abreviaturas y el aspecto "jibarizado" de los mensajes (José Ramón Morala, 2001). No obstante, creemos que este tipo de escritura puede interpretarse como una marca grupal, como una muestra de que el internauta está al tanto del peculiar código comunicativo vigente en las redes sociales. Por tanto, su empleo constituye también una prueba de su integración en la comunidad virtual.

Con frecuencia, las faltas de ortografía en las redes sociales suelen atribuirse también a la falta de concentración, o a la dejadez de quien escribe desde la proximidad comunicativa, consciente de que el destinatario de su mensaje –generalmente un amigo– se encuentra al tanto de las convenciones que rigen la comunicación en las redes sociales. Pero, en realidad, estos internautas hacen uso de una ortografía diferente de la de los textos convencionales, de una "antiortografía" (Gabriela Palazzo, 2005), que no impide que el enunciador y su enunciatario se comprendan, dado que comparten las mismas competencias lingüísticas. En este sentido, José Martínez de Sousa (2004) distingue entre "faltas de ortografía" y "heterografías". Las primeras son fruto de la ignorancia de las normas que rigen la grafía del español, mientras que las segundas constituyen lo que podríamos llamar desviaciones intencionadas. A este último grupo parece responder la duplicación de vocales y consonantes que encontramos en muchos mensajes:

> "#ParejasQueMeGustan peter lanzani y lali esposito definitivamente, *vuelvann mierdaaa*!" (Twitter, 20-04-2013).

Tales transgresiones pueden suponer un problema para la competencia lingüística de algunos jóvenes. Por ejemplo, al aumentar su inseguridad sobre la escritura del verbo auxiliar *haber*:

> "Si mo fuera x @yomismaeirene no *abria* entendido la película ocean's eleven" (Facebook, 24-03-2013).

Del análisis de estos mensajes se desprende que la mayor parte de las "acometidas" contra la ortografía del español realizadas tiene lugar en Tuenti y, en

menor medida, en Facebook. Sin embargo, no podemos decir lo mismo de los mensajes redactados en Twitter, donde parece existir un mayor respeto a las normas ortográficas. Tal vez porque sus usuarios son más conscientes de que en la mayor parte de los casos se dirigen a internautas para los que resultan desconocidos, y para los que, más que las fotografías, su carta de presentación son las palabras. Por eso, quien no sabe escribirlas con corrección se arriesga a duros ataques, como los que recibió María Antonia Trujillo –ex ministra de Vivienda del primer gobierno de José Luis Rodríguez Zapatero–, quien publicó en su perfil de Twitter el siguiente texto, con el que trataba de mostrar su indignación después de haberse hecho pública su declaración de bienes:

> "Mi anti #ff para todos los q han informado tanto de nuestro patrimonio y tampoco de nuestro trabajo en #Congreso" (Twitter, 09-09-2011).

La falta de ortografía presente en este mensaje desencadenó toda una avalancha de críticas en Twitter, y además fue objeto de burla en numerosos medios de comunicación, de modo que la ex ministra no tuvo más remedio que rectificar y aclarar en otro mensaje que no había querido decir "tampoco" sino "tan poco". Por tanto, es importante respetar la ortografía incluso en las redes sociales, especialmente si se es un personaje público.

Por otro lado, la selección léxica resulta también de considerable relevancia a la hora de redactar estos mensajes. De hecho, Sigfrid Soria, ex diputado autonómico del PP por la isla de Lanzarote, fue apartado de sus responsabilidades en este partido el pasado mes de abril a raíz de la publicación en su perfil de Twitter de una serie de textos como los siguientes, en los que amenazaba a los miembros de la Plataforma de Afectados por la Hipoteca que se manifestaban frente a los domicilios de algunos políticos:

> "Son tan cobardes los tuiteros perroflautas como lo eran los del ku klux klan cuando linchaban a seres humanos con sus identidades ocultas" (Twitter, 10-04-2013).

> "Eso sí, como un perroflauta me acose por la calle, me intimide o agreda, la ostia que se lleva ni se la va a creer" (Twitter, 10-04-2013).

La selección léxica aquí resulta poco adecuada para el rol público que debe desempeñar un político. Pero además es muy llamativa la falta de ortografía cometida en la redacción del sustantivo "hostia". Y es que sorprende mucho el empleo por parte de un político de este término, en su acepción vulgar y malsonante. Con estas palabras trataba de justificar en un programa radiofónico su escritura "antiortográfica":

> "No la escribo con 'h' pues no quiero semejanza alguna con la de la principal acepción de la RAE".

Desconocemos si en este caso nos encontramos ante una verdadera muestra de lo que hemos denominado *heterografía*, es decir, de un uso innovador del lenguaje. O si esta es en realidad una mera falta de ortografía que luego el político ha tratado de disimular con sus declaraciones.

También es muy frecuente que los internautas publiquen mensajes en los que utilizan expresiones disfemísticas,

> "qeee diseee locoo d la morciyaa!!!! q laa staba scuxaando y la e pstoo d stadooo soo mierdaa asik caya la bocaaa!! ajajjajaj" (Tuenti, 15-12-2012).

incluso adjetivos de significación peyorativa como *feo* o *cabrón* que, en virtud del contexto, adquieren un sentido ponderativo:

> "Felicidades cabrooooon !!!dejate veeeeeeer feooo y pasatelo flamaaa jajajaja" (Tuenti, 06-04-2013).

Tal uso resulta muy habitual en esta red social, lo que podría explicarse por dos razones: por una parte, porque Tuenti cuenta con mayor número de perfiles de adolescentes, y para estos tales prácticas comunicativas son muy recurrentes; por otra parte, porque, además, en esta red social el acceso es restringido; es decir, sólo pueden acceder los "amigos", esto es, personas que aceptan una invitación por parte del usuario a formar parte de su red de amistades en este espacio. Por el contrario, quizás el carácter menos restrictivo de Facebook y Twitter, así como la pluralidad de edades y profesiones de sus usuarios expliquen el hecho de que hayamos encontrado un menor número de insultos y expresiones disfemísticas en estas dos redes sociales.

4. CONCLUSIONES

A través de esta rápida panorámica, esperamos haber mostrado cómo los textos extraídos de las redes sociales virtuales para llevar a cabo análisis del discurso mediado por ordenador en lengua española constituyen un interesante corpus de datos lingüísticos y de discursos heterogéneos que presentan una gran variedad de registros y de estilos. Un aspecto que plantea numerosos desafíos para el Análisis del discurso y sus métodos de investigación, en el contexto de las Humanidades digitales.

Asumiendo los principales postulados de la Lingüística de corpus relacionados con el método de recolección de datos, la representatividad de los mismos y la privacidad, en este trabajo hemos presentado un "pre-corpus" constituido por textos extraídos de tres redes sociales distintas: Facebook, Twitter y Tuenti. Este conjunto de textos nos ha permitido explorar, a través de un análisis cualitativo comparativo, las semejanzas y diferencias relativas a la ortografía y a la selección léxica en el español utilizado en las tres redes sociales, y establecer hipótesis de trabajo para futuras investigaciones centradas en las posibles causas del carácter antiortográfico y heterográfico de la escritura digital, o en los distintos grados de coloquialización de la lengua en estos espacios.

Desde un punto de vista metodológico, en próximos análisis cabrá considerar la cuestión de la anotación lingüística con etiquetadores morfológicos (*tagger*) y sintácticos (*parser*), de cara a afinar el análisis lingüístico y poder tratar los datos de modo automatizado para obtener resultados no sólo cualitativos sino también cuantitativos –frecuencia de uso de determinadas palabras o comparación estadística de los fenómenos estudiados–. De hecho, para la anotación y marcaje estructural de textos extraídos de las redes sociales, contamos ya con un modelo de etiquetado de discursos mediados por ordenador propuesto por Michael Beiβwenger *et al.* (2012) y basado en los presupuestos de la Text Encoding Initiative, que puede resultar de gran utilidad para nuestros propósitos.

Por último, cabe decir que la manera de entender un corpus ha evolucionado y que, sobre todo en el caso de textos digitales publicados en la web, la explotación del mismo obliga a llevar a cabo conceptualizaciones y análisis multi-

dimensionales. Por ejemplo, en nuestro caso es necesario reflexionar en torno a aspectos como el contexto de publicación de los mensajes, pues estos se insertan en entornos semióticos complejos constituidos por fotografías, videos, etiquetas o enlaces que están estrechamente relacionados con los enunciados, y que en muchos casos determinan el sentido de los mismos.

BIBLIOGRAFÍA

Androutsoupoulos, Jannis y Beißwenger, Michael, "Introduction. Data and Methods in Computer-Mediated Discourse Analysis", *Language@Internet*, 5 (9) (2008), <http://www.languageatinternet.org/articles/2008/1609/introduction.pdf> [21/09/2013].

Baker, Paul, *Using Corpora in Discourse Analysis,* London, Continuum, 2006.

Beißwenger, Michael y Storrer, Angelika, "Corpora of computer-mediated communication", en A. Lüdeling y M. Kytö (eds.), *Corpus linguistics. An international handbook*, vol. 1, Berlin/New York, Mouton de Gruyter, 2008, pp. 292-308.

Beißwenger, Michael *et al.*, "A TEI Schema for the Representation of Computer-mediated Communication", *Journal of the Text Encoding Initiative*, 3 (2012), <http://jtei.revues.org/476> [21/09/2013].

Berlanga Fernández, Inmaculada y Martínez, Estrella, "Ciberlenguaje y principios de retórica clásica. Redes sociales: el caso Facebook", *Enl@ce: Revista Venezolana de Información*, *Tecnología y Conocimiento*, 7 (2) (2010), pp. 47-61.

Claridge, Claudia, "Constructing a Corpus from the Web: Message Boards", en M. Hundt, N. Nesselhauf y C. Biewer (eds.), *Corpus Linguistics and the Web*, Amsterdam, Rodopi, 2007, pp. 87-108.

Fletcher, William H., "Facilitating the compilation and dissemination of *ad hoc* web corpora", en G. Aston, S. Bernardini y D. Stewart (eds.), *Corpora and Language Learners*, Amsterdam, John Benjamins, 2004, pp. 273-300.

Gómez Torrego, Leonardo, "La gramática en Internet", *Lengua y escritura en Internet: Tres décadas de "red-acción",* 2001, <http://congresosdelalengua.es/valladolid/ponencias/nuevas_fronteras_del_espanol/4_lengua_y_escritura/gomez_l.htm> [21/09/2013].

Herring, Susan C., "Computer-mediated discourse", en D. Schiffrin, D. Tannen y H.E. Hamilton (eds.), *Handbook of Discourse Analysis,* Oxford, Blackwell, 2001, pp. 612-634.

Kendall, Lori, *Hanging out in the virtual pub: Masculinities and relationships online,* Berkeley, University of California Press, 2002.

Kilgarriff, Adam y Grefenstette, Gregory, "Introduction to the special issue on the web as corpus", *Computational Linguistics*, 29 (3) (2003), pp. 333-348.

King, Brian, "Building and analyzing corpora of Computer-mediated communication", en P. Baker (ed.), *Contemporary Corpus Linguistics*, London, Continuum, 2009, pp. 301-320.

Kock, Josse de, *Lingüística con corpus. Catorce aplicaciones sobre el español*, Salamanca, Universidad de Salamanca, 2001.

Lara, Tíscar, "Twitter y sus funciones comunicativas", *Blog Tíscar.com,* 2012, <http://tiscar.com/2012/03/11/twitter-y-sus-funciones-comunicativas> [21/09/2013].

Lavid, Julia, *Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI*, Madrid, Cátedra, 2005.

Leech, Geoffrey, "Corpora theories of linguistic performance", en J. Svartvik (ed.), *Directions in Corpus Linguistics*, Berlin/New York, Mouton de Gruyter, 1992, pp. 105-122.

Mancera Rueda, Ana, *¿Cómo se "habla" en los cibermedios? El español coloquial en el periodismo digital*, Bern, Peter Lang, 2011.

Marcos Marín, Francisco, *Informática y Humanidades*, Madrid, Gredos, 1994.

Martínez de Sousa, José, *Ortografía y ortotipografía del español actual*, Gijón, Trea, 2004.

Martínez Santiago, Fernando *et al.*, "WWW como Fuente de Recursos Lingüísticos para su Uso en PLN", *Procesamiento del lenguaje natural*, 27 (2001), pp. 1-7.

Morala Rodríguez, José Ramón, "Entre arrobas, eñes y emoticonos", *II Congreso Internacional de la Lengua Española. El español en la Sociedad de la información*, Valladolid (2001), <http://congresosdelalengua.es/valladolid/ponencias/nuevas_fronteras_del_espanol/4_lengua_y_escritura/morala_j.htm> [21/09/2013].

Palazzo, Gabriela, "¿Son corteses los jóvenes en el chat? Estudio de estrategias de interacción en la conversación virtual", *TEXTOS de la CiberSociedad,* 5 (2005), <http://www.cibersociedad.net> [21/09/2013].

Pano Alamán, Ana, *Dialogar en la Red. La lengua española en chats, e-mail, foros y blogs*, Bern, Peter Lang, 2008.

Pano Alamán, Ana y Mancera Rueda, Ana, "Identidades y cuentas parodia en Twitter: análisis de la ironía y del humor verbal", comunicación presentada en el Congreso Internacional de Lingüística Hispánica sobre "Lenguaje e identidad en el mundo hispanohablante", Asociación Internacional para el Estudio del Español en la Sociedad (SIS/EES), Londres, 3-5 de julio, 2013.

Parodi, Giovanni, *Lingüística de corpus: de la teoría a la empiria*, Frankfurt, Iberoamericana Vervuert, 2010.

Resnik, Philip y Smith, Noah A., "The Web as a parallel corpus", *Computational Linguistics*, 29 (3) (2003), pp. 349-380.

Rundell, Mike, "The biggest corpus of all", *Humanising Language Teaching,* 2 (3) (2000), <http://www.hltmag.co.uk/may00/idea.htm> [21/09/2013].

Sinclair, John, *Corpus, concordance, collocation*, Oxford, Oxford University Press, 1991.

Tognini-Tonelli, Elena, *Corpus linguistics at work*, Amsterdam, John Benjamins, 2001.

Torrego González, Alba, "'Eskriibo en el Tuenti komo pronunciioh'. Apuntes sobre la ortografía en una red social", *Tarbiya: Revista de investigación e innovación educativa*, 41 (2010), pp. 33-51.

Torrego González, Alba, "Algunas observaciones acerca del léxico en la red social tuenti", *Tonos digital: Revista electrónica de estudios filológicos*, 21 (2011), <http://www.tonosdigital.es/ojs/index.php/tonos/article/view/659/470> [21/09/2013].

Yus Ramos, Francisco, *Ciberpragmática 2.0. Nuevos usos del lenguaje en Internet*, Barcelona, Ariel, 2001.

RESUMEN

El propósito de este artículo es demostrar cómo las interacciones que tienen lugar en las redes sociales pueden aportar multitud de datos lingüísticos y de discursos heterogéneos que presentan una gran variedad de registros y de estilos. Todo ello convierte a este tipo de textos en un corpus ideal para su estudio desde la metodología del Análisis del discurso mediado por ordenador, una disciplina cuyo principal objetivo es la investigación de las propiedades lingüísticas y pragmáticas de los productos discursivos de impronta digital. Además, en este trabajo nos hemos propuesto realizar una revisión de los postulados centrales de la Lingüística de corpus, haciendo especial hincapié en las ventajas de establecer un corpus para estudiar documentos extraídos de la Red, pero sin dejar de lado los problemas que ello puede conllevar, como la selección del método más adecuado para la recolección de datos, la necesidad de que estos sean lo suficientemente representativos, y la importancia del respeto a la privacidad de los usuarios. Teniendo en cuenta estas cuestiones hemos constituido un corpus de mensajes extraídos de las redes sociales Facebook, Twitter y Tuenti, que nos permitirán analizar cómo se manifiesta en ellos la variación lingüística y, en concreto, la modalidad de uso coloquial, por medio de una serie de recursos ortográficos y de una peculiar selección léxica. En definitiva, en esta investigación pretendemos ofrecer una panorámica que recoja los principales desafíos a

los que debe hacer frente el Análisis del discurso mediado por ordenador y sus métodos de investigación, en el contexto de las Humanidades digitales.

Palabras clave: Análisis del discurso mediado por ordenador, Lingüística de corpus, redes sociales, variación lingüística, español coloquial.

ABSTRACT

The aim of this paper is to show how the interactions that take place in social networks can provide plenty of linguistic data and heterogeneous discourses presenting a variety of registers and styles. This type of texts can constitute a useful corpus for the research, using the methodology of Computer-Mediated Discourse Analysis, a discipline whose main objective is the study of linguistic and pragmatic properties of digital discursive products. Moreover, in this study we aimed to make a review of the main theories of Corpus Linguistics, with particular emphasis on the advantages of establishing a corpus to study documents extracted from the Internet, but without neglecting the problems this can lead, such as selecting the most appropriate method for data collection, the need for these data to be sufficiently representative, and the importance of respecting the privacy of users. According to these issues we have created a corpus of messages extracted from three social networks: Facebook, Twitter and Tuenti. It will allow us to explore in them the linguistic variation and, in particular, the samples of colloquial use, by a number of resources of a peculiar spelling and lexical selection. In conclusion, in this research we want to give an overview to the most important challenges that must face the Computer-Mediated Discourse Analysis and its research methods in the context of Digital Humanities.

Keywords: Computer-Mediated Discourse Analysis, Corpus Linguistics, Social Networks, Linguistic Variation, colloquial Spanish.