I Jornadas Nacionales de Humanidades Digitales. Asociación Argentina de Humanidades Digitales, Buenos Aires, 2014.

# Repositorio colaborativo de comunicaciones digitales: aproximación a un corpus para el español.

Lucía Cantamutto y Cristina Vela Delfa.

Cita:

Lucía Cantamutto y Cristina Vela Delfa (Noviembre, 2014). *Repositorio colaborativo de comunicaciones digitales: aproximación a un corpus para el español. I Jornadas Nacionales de Humanidades Digitales. Asociación Argentina de Humanidades Digitales, Buenos Aires.*

Dirección estable: https://www.aacademica.org/lucia.cantamutto/30

ARK: https://n2t.net/ark:/13683/ptCk/xHw

FILO:UBA
Facultad de Filosofía y Letras
Universidad de Buenos Aires

# Las Humanidades Digitales desde Argentina

## Tecnologías, Culturas, Saberes

Buenos Aires, 2014

Actas de las I Jornadas de Humanidades Digitales

FILO:UBA
Facultad de Filosofía y Letras

FILODIGITAL
Repositorio Institucional de la Facultad de Filosofía y Letras, UBA

Cantamutto, Lucía
Actas de las I Jornadas de Humanidades Digitales / Lucía Cantamutto; Gimena del Rio Riande; Gabriela Striker (eds.). - 1a ed. . - Ciudad Autónoma de Buenos Aires : Editorial de la Facultad de Filosofía y Letras Universidad de Buenos Aires, 2015.
Libro digital, PDF

Archivo Digital: descarga
ISBN 978-987-3617-89-8

1. Ciencias Sociales y Humanidades. 2. Aplicaciones Informáticas.
I. Río Riande, Gimena del II. Título
CDD 301

**LAS HUMANIDADES DIGITALES DESDE ARGENTINA: CULTURAS, TECNOLOGÍAS, SABERES**

**Gimena del Rio Riande, Lucía Cantamutto, Gabriela Sriker (eds.)**

**PRELIMINARES**



**I. *REPRESENTACIONES SOCIALES Y HUMANIDADES DIGITALES***



**II. *REPOSITORIOS, DOCUMENTACIÓN, DIGITALIZACIÓN Y EDICIÓN DIGITAL ACADÉMICA***

*Cóncavo y convexo: Documentación y Humanidades Digitales, punto de inflexión*. BOSCH, Mela (CAICYT, CONICET)

*Repositorio colaborativo de comunicaciones digitales: aproximación a un corpus para el español.* CANTAMUTTO, Lucía (Universidad Nacional del Sur-CONICET)/VELA DELFA, Cristina (Universidad de Valladolid)

*Proyecto Archivo Digital Dr. Alberto Rex González: digitalización y catalogación de un fondo documental en dirección al acceso abierto.* DOMÍNGUEZ, Marcelo Adrián (DILA-CAICYT, CONICET)

*Plataforma Interactiva de Investigación en Ciencias Sociales.* LEFF, Laura (PLIICS, CONICET)/PLUSS, Ricardo (PLIICS, CONICET)

*Propuestas y desafíos para una base de datos de mujeres artistas en Argentina.* GLUZMAN, Georgina (Universidad de San Martín-CONICET)

Un proyecto de edición digital académica en Argentina. Diálogo Medieval. RIO RIANDE, Gimena del (IIBICRIT, CONICET)/ZUBILLAGA, Carina (IIBICRIT, CONICET/Universidad de Buenos Aires)

**III. LA PUBLICACIÓN DIGITAL**

*Herramientas de publicación académica en la web 2.0: ¿tercera vía para el acceso abierto?.* DE GRANDE, Pablo (Proyecto Acta Académica)/QUARTULLI, Diego (Proyecto Acta Académica)/RUSSO, Alejandra (Proyecto Acta Académica)

*Publicaciones digitales: hacia una edición profesional.* DIEZ, María Clara (Universidad de Buenos Aires)/KESSLER KENIG, Carola (Universidad de Buenos Aires)

*Editing de publicaciones digitales.* ESPÓSITO, Cecilia (Universidad de Buenos Aires)

*Políticas editoriales en el entorno digital. El caso de los materiales educativos.* TOSI, Carolina (CONICET – Universidad de Buenos Aires)

*Hypothèses: un aliado para las Humanidades Digitales.* TEJADA-CARRASCO, Beatriz (Universidad Nacional de Educación a Distancia, España)

**IV. REFLEXIONES SOBRE LO DIGITAL**

*Las humanidades en la era del canon digitalizado.* GABRIELONI, Ana Lía (Universidad Nacional de Río Negro-CONICET)

*Ejes para un debate sobre el uso ético de datos interaccionales escritos y orales obtenidos en línea.* DE-MATTEIS, Lorena M. A. (CONICET- Universidad Nacional del Sur)

*Imaginario y Tecnologías Digitales: el sueño del receptor activo.* LESTA, María Laura (Universidad Siglo 21)/ORTEGA VILLAFAÑE, Manuel (Universidad Siglo 21)/RODRIGUEZ, Ana Paula (Universidad Siglo 21)/TORRES, Celeste Rocío (Universidad Siglo 21)

*El conocimiento digital desde una visión foucaulteana.* PIRIZ, Franco (Universidad Nacional de Mar del Plata)/ CAMARA, Ezequiel (Universidad Nacional de Mar del Plata)

## *V. EDUCACIÓN Y DESAFÍOS DIGITALES*

*La implementación de las tecnologías móviles en las escuelas: las ciudadanías digitales.* HANDAL, Boris (Universidad de Notre Dame, Australia)/WATSON, Kevin (Universidad de Notre Dame, Australia)/DENG, Hui Hong (Universidad de Notre Dame, Australia)

*Conectar Igualdad, la política de inclusión tecnológica del Estado argentino. Reflexiones sobre la escolarización en el siglo XXI.* NECUZZI, Constanza (Programa Conectar Igualdad, Universidad de Buenos Aires)

*Construcción de espacios interculturales en la educación superior: un abordaje desde las clases invertidas.* POZZO, María Isabel (Universidad Nacional de Rosario)/TALLEI, Jorgelina (Universidad de Integración Latinoamericana)

*Producción y gestión de contenidos educativos digitales y una nueva agenda.* SAGOL, Cecilia (Ministerio de Educación, Portal educ.ar)

*Comunidades de práctica virtuales: conocimiento compartido para el crecimiento profesional y personal de los docentes.* SCORIANS, Erica Elena (Universidad Nacional de La Plata)/VERNET, Mercedes (Universidad Nacional de La Plata)

## VI. COMUNICACIÓN. TEXTO E IMAGEN DIGITAL (Imagen y comunicación digital)

*La comunicación por mensajes de texto en el español bonaerense: uso y percepción.* CANTAMUTTO, Lucía (Universidad Nacional del Sur-CONICET)

*Las nuevas tecnologías y los estilos comunicacionales de jóvenes universitarios.* GIAMMATTEO, Mabel (Universidad de Buenos Aires)/ PARINI, Alejandro (Universidad de Belgrano)

*La imagen en Facebook y la comunicación visual móvil. El caso de la fotografía celular.* GUREVICH, Ariel (Universidad de Buenos Aires)/SUED, Gabriela (Universidad de Buenos Aires)

*Contenido digital accesible. Accesibilidad de los materiales y entornos virtuales académicos.* MARTINEZ, María del Milagro (Universidad Nacional de Córdoba)

*Los dos Borges. Imágenes de un escritor en YouTube. De la cultura textual a la cultura visual.* SUED, Gabriela (Universidad de Buenos Aires)

*De lo vertical a lo disperso. Apuntes para una historia de la perspectiva*. MENDOZA, JUAN (Universidad de Buenos Aires-CONICET)

**VI. LAS HUMANIDADES DIGITALES EN PRÁCTICA**

*Introducción a la edición digital académica.* MARTÍNEZ CANTÓN, Clara Isabel (Universidad Nacional de Educación a Distancia, España)/RIO RIANDE, Gimena del (IIBICRIT, CONICET)

*Gestores de referencias bibliográficas. Zotero y Mendeley.* CAMPOS, Guadalupe (Universidad de Buenos Aires)/VILAR, Mariano (Universidad de Buenos Aires)

*Introducción a la edición de textos en LaTeX.* DE-MATTEIS, Lorena (Universidad Nacional del Sur-CONICET)

*Archivos y mapas.* NAVARRO, Gustavo (Universidad Nacional de la Patagonia Austral)

*Bibliotecas y archivos digitales con Greenstone.* PICHININI, Mariana (Universidad nacional de La Plata)

*Scrapping visual.* CINGOLANI TRUCCO, Gino (Universidad de Buenos Aires)/RODRÍGUEZ KEDIKIAN, Martín (Universidad de Buenos Aires)/VACCARI, Gonzalo (Universidad de Buenos Aires)/ALONSO, Julio (Universidad de Buenos Aires)

# *Repositorio colaborativo de comunicaciones digitales: aproximación a un corpus para el español*

*CANTAMUTTO, Lucía/ Universidad Nacional del Sur-CONICET – luciacantamutto@gmail.com*

*VELA DELFA, Cristina/ Universidad de Valladolid – vela@fyl.uva.es*

» *Palabras clave: comunicación digital, corpus lingüístico, Sociolingüística, español, repositorio.*

## › Resumen

Dentro del conjunto de investigaciones sobre la comunicación digital, uno de los problemas con los que se enfrenta el investigador es el establecimiento de corpus de datos: no siempre resulta posible la recogida de muestras de lengua que conserven la representatividad necesaria para legitimar cualquier estudio (Torruella & Llisterri, 1999). Para solventar esta carencia resulta necesario establecer corpus estables de muestras de lengua de la comunicación digital. Las herramientas colaborativas que actualmente ofrece internet resultan muy útiles a este fin. Los corpus lingüísticos han encontrado en la web un espacio dinámico y accesible para investigadores de disciplinas diversas que requieren datos primarios no siempre a su alcance. No obstante, hasta donde llega nuestro conocimiento, no existe aún para el español un repositorio de corpus de comunicación digital que recopile muestras de chat, email, SMS, entre otros, para su estudio sociolingüístico. El objetivo de este trabajo es doble. Por un lado, llevaremos a cabo una reflexión sobre los problemas metodológicos de recogida y fijación de datos en los entornos comunicativos digitales, a partir de la reflexión sobre los antecedentes encontrados en inglés, francés y chino; por otro lado, presentaremos los preliminares de un proyecto de creación de un repositorio abierto y colaborativo de comunicaciones digitales (CoDiCE), que permitirá avanzar en los estudios de variación sociolingüística y pragmática intra/interlingüística.

## › Presentación

A partir del creciente número de investigaciones que abordan la comunicación digital se ha avanzado en nutrir de reflexiones teóricas y de datos empíricos este campo de estudios. Sin embargo, esta enorme producción de literatura científica no siempre acarrea una perspectiva y un

modelo metodológico apropiados para dar cuenta de la complejidad del objeto de estudio; en particular, para las investigaciones de corte sociolingüístico y sociopragmático. El objetivo de este trabajo es presentar una revisión de la literatura en torno al establecimiento de corpus de comunicaciones mediadas digitalmente y, además, plantear la necesidad de conformar un repositorio digital abierto y colaborativo de comunicaciones digitales para el español. De este modo, se intenta responder a dos preguntas fundamentales: ¿en qué situación metodológica se encuentran las investigaciones sobre la comunicación digital? y ¿es esperable establecer un corpus abierto y colaborativo de comunicaciones digitales en nuestra lengua?

Los corpus lingüísticos encontraron en la web un espacio dinámico y accesible para investigadores de disciplinas diversas que requieren datos primarios, no siempre a su alcance. Sin embargo, hasta donde llega nuestro conocimiento, no existe aún para el español un repositorio de corpus de comunicación digital que recopile muestras de chat, email, SMS, entre otros, apropiadas para su estudio sociopragmático. Por ello, abordamos la necesidad y viabilidad en la constitución de un corpus o repositorio abierto de comunicaciones digitales en español.

Para tal fin, dividimos nuestro trabajo en tres partes. En primer lugar, presentamos y definimos el campo de estudio de la comunicación digital, identificando sus características generales y sus atributos específicos en relación con los diferentes dispositivos mediante los cuales se desarrolla el intercambio (*La comunicación digital: un campo de estudio emergente*), señalando las cuestiones metodológicas pertinentes. En segundo lugar, haremos una revisión de la bibliografía para dar cuenta de las estrategias metodológicas utilizadas en la recolección de datos para diferentes corpus; se presentará una síntesis de algunas investigaciones en torno a cada uno de los tipos discursivos (chat, email, SMS, redes sociales, *WhatsApp*, entre otros) para el español y para otras lenguas (*Sobre los corpus de comunicaciones digitales*). Por último, se defenderá la pertinencia de la creación de un repositorio de comunicaciones digitales del español, al que denominaremos CoDiCE (*Comunicaciones Digitales Corpus del Español*).

## › *La comunicación digital: un campo de estudio emergente*

En las últimas décadas se ha consolidado el interés por el estudio de las comunicaciones digitales. Desde distintas disciplinas -Sociología, Filosofía de la Ciencia, Psicología, Ciencias de la Información y de la Comunicación y Lingüística- han proliferado las investigaciones destinadas a entender este fenómeno comunicativo. Entre ellas destacan las reflexiones de carácter lingüístico, puesto que internet es, antes que cualquier otra cosa, un fenómeno textual: "sea lo que sea la cultura de Internet, sigue siendo un fenómeno textual" (Wilbur, 1996: 6) o, en palabras de Crystal (2001: 53), "si internet es una revolución, desde luego, parece ser una revolución lingüística" [la traducción es nuestra].

Resulta necesario señalar la escasez de obras de conjunto sobre el español en internet, que sirvan como complemento a las existentes en el ámbito anglosajón, algunas consideradas ya clásicas, entre ellas Crystal (2001), Kress (2003) y Herring (1996 y 2004). Más allá de estos

ISBN 978-987-3617-89-8

trabajos, encontramos aportaciones sobre el uso del *ciberlenguaje*, que provienen de diferentes paradigmas epistemológicos. Abundan las aproximaciones que, desde una perspectiva comparatista, ponen en relaciones la comunicación en contextos presenciales con la comunicación digital (Walther, 1996; Gobato, 2014). Sobre todo en sus orígenes, muchas se centraron en la identificación de las limitaciones de esta última frente a la primera, entre ellas, Kiesler, Siegel & McGuire (1984).

Asimismo, se ha ido generalizando una corriente de estudio que se orienta hacia el análisis de los aspectos meramente normativos. Focaliza su interés en los fenómenos de desviación de la norma culta estándar manifestados por los usos lingüísticos de los medios digitales (Ivars, 2007). No deberían confundirse estos con otros estudios que, desde un enfoque sociolingüístico, interpretan el habla en la red desde una orientación variacionista, por ejemplo Baron (1984), Covarrubias (2008), o desde una perspectiva sociocultural (Muñiz Calderón, 2011). Lamentablemente, estos últimos son todavía sorprendentemente escasos, a tenor de la generalización del fenómeno de la comunicación por internet.

Especialmente relevantes se consideran aquellos estudios que asumen una orientación pragmática como los de Yus (2001 y 2010). Algunos sesgan su atención hacia fenómenos relativos a cuestiones de cortesía y en particular a la llamada *netiquette* (Booher, 2001; Dorner, 2002). En el exhaustivo trabajo compilado por Bernete García (2007) sobre la comunicación digital y los jóvenes, se explican estas normas para los participantes de foros. Es interesante que, entre las cuestiones que se señalan (además de evitar mayúsculas porque se *interpretan* como gritos) se sugieren muchas reglas que responden a las máximas de Grice: "no contribuir a llenar los foros de mensajes inservibles", "no desviar los temas", no abrir "el mismo tema varias veces, o pegar el mismo mensaje en todos los demás mensajes" (Bernete García, 2007: 77). Por su parte, Noblia (2001) aborda la negociación de la cortesía y la *netiquette* en los chats.

Otro conjunto de trabajos ponderados provienen del marco del análisis de la conversación. Estos se interesan por cuestiones fundamentalmente estructurales, y tratan de entender en qué medida los intercambios digitales manifiestan propiedades auténticamente interaccionales. Entre estos, destacamos los trabajos sobre la organización de la interacción, turnos de palabra, etc. (Herring, 1999a) y estudios más específicos para el chat (Noblia, 2000), para la mensajería instantánea (Vela Delfa & Jiménez Gómez, 2011), para el correo electrónico (Vela Delfa, 2006), para la comunicación por SMS (Cantamutto, 2013) y, recientemente, para la comunicación por *WhatsApp* (Alcántara Plá, 2014). Asimismo, desde la perspectiva del Análisis Crítico del Discurso, se han abordado las manifestaciones discursivas de cuestiones sociales, es decir, sobre las jerarquías de poder en la red, los estudios sobre las relaciones entre el género y su participación en internet (Herring, 1996 y 2004; Eisenchlas, 2012), o el acoso sexual en la red (Herring, 1999b).

Estas aproximaciones tienen una importante limitación: la falta de organización que posibilitaría la configuración de un área de estudio más unificada y coherente. La multiplicidad de acercamientos ha permitido el enriquecimiento de las reflexiones, pero no han ayudado a su sistematización, principalmente, desde el punto de vista metodológico. Así, los trabajos sobre la interacción comunicativa digital se establecen, necesariamente, a partir de una

ISBN 978-987-3617-89-8

transdisciplinariedad, en la medida en que estos objetos de estudio están atravesados por distintas disciplinas y que han de orientarse al surgimiento de nuevos campos de estudio colindantes a las Ciencias Sociales (Gobato, 2014), manifestando, por tanto, una dimensión interdisciplinar.

Quizá la orientación que más se ajusta a esta perspectiva es la del análisis del discurso mediatizado por ordenador (ADMO), que aplica la metodología del análisis del discurso al estudio de intercambios lingüísticos producidos a través de una computadora u otro dispositivo tecnológico. Una de sus aportaciones clave, fundamental para la delimitación epistemológica de la disciplina, estriba en el establecimiento de un objeto de estudio definido: la *comunicación mediatizada por ordenador* (CMO). Por CMO, siglas adaptadas de su equivalente en inglés *Computer-Mediated Communication* (CMC), se entiende aquella comunicación producida cuando dos o más personas interactúan transmitiendo mensajes a través de un ordenador o de otro dispositivo tecnológico (Herring, 2001: 612). La noción de CMO supone un constructo que, más allá de conformar una clase de elementos, adquiere un trasfondo metodológico importante.

El empleo de dispositivos tecnológicos en las interacciones comunicativas provoca innovaciones en el uso del lenguaje, así como también en la forma de conceptualizar los marcos de interacción (Goffman, 1959; Gobato, 2014). Desde esa perspectiva, el papel del analista del discurso consistirá en determinar las características situacionales. La hipótesis de partida asume que se trata de una situación comunicativa en la que la influencia del medio es muy grande y los usuarios, por tanto, deberán adaptarse a él.

Aunque el término ADMO no fue introducido hasta 1995, Susan Herring (2001) estima una antigüedad de unos veinte años para esta rama del análisis del discurso. Según su revisión histórica se pueden identificar tres etapas en el desarrollo de estos estudios:

a) A partir de la segunda mitad de los años ochenta del siglo pasado, momento en que se llevan a cabo los primeros análisis sobre la CMO, que aunque tienen el mérito de inaugurar el área de estudio, son todavía algo limitados y están abordados desde una perspectiva comparativa con la comunicación oral, focalizando principalmente en las limitaciones del medio.
b) La primera mitad de la década de los noventa, que supone el afianzamiento del área de estudios. Se abandona la perspectiva comparativa para caracterizar lo que se denominará el *interactive written discourse* (Ferrara, Brunner & Whittemore, 1991). Esta propuesta venía a zanjar una cuestión que ocupó muchos de los estudios iniciales y que sigue estando presente, de alguna manera, incluso en los más recientes: la relación del discurso electrónico con la dicotomía oral/escrito (Baron, 1998). El discurso electrónico no puede clasificarse fácilmente como una variedad de la modalidad escrita ya que muchas de sus propiedades lo alejan de ella para acercarlo al prototipo del oral, con el que tampoco puede identificarse. Yates (1996) demuestra que el DMO (Discurso Mediatizado por Ordenador) presenta propiedades que hacen de él una variedad diferente del discurso oral y del escrito.
c) A partir de la segunda mitad de la década de los noventa, aumentan exponencialmente los trabajos dedicados a este fenómeno comunicativo. La mejora no es únicamente cuantitativa,

ISBN 978-987-3617-89-8

> sino cualitativa, porque estos acercamientos abordan la variedad de géneros discursivos albergados en la red, subsanado un error fundamental de los trabajos anteriores que tendían a clasificar la CMO en su conjunto como un género unitario. Así, en esta etapa se abordan cuestiones como la textualidad en la red, en lo relativo a los fenómenos de coherencia (Herring, 1999a), conectividad, intertextualidad (Payà, 2000), tipología genérica (Gruber, 2000).

La periodización propuesta por Herring (2001) no contempla, como es lógico, el devenir que los estudios sobre el discurso digital han experimentado en el siglo XXI. En esta última década se han revisado muchas de las caracterizaciones desarrolladas en años previos. Estas actualizaciones, como la revisión propuesta por Yus (2010), el trabajo de Thurlow, Lengel & Tomic (2004), e incluso la obra de conjunto, que se propone como un manual sobre la CMO, editada por Schiffrin, Tannen & Hamilton (2001), han permitido la consolidación de la disciplina, gracias a delimitar su objeto de estudio y a la normalización teórica. Las propias circunstancias contextuales colaboraron con este cambio. Las interacciones mediatizadas se han instalado en todas las esferas de la comunicación interpersonal, y la asiduidad con que los usuarios acceden, a través de dispositivos cada vez más versátiles, ha aumentado considerablemente; hasta el punto que los estudios sobre el discurso mediatizado alcanzan procesos comunicativos de toda índole.

Al tiempo que se consolida la disciplina, identificamos un aumento de los acercamientos de mayor solidez teórica: las teorías y metodologías pragmáticas y sociolingüísticas (Yus, 2001; Yus, 2010; Martín Corvillo, 2014), por ejemplo, han servido de hilo conductor en aproximaciones recientes. La generalización de la comunicación mediatizada ha ayudado a asumir el entorno como un espacio variado en el que proliferan géneros discursivos muy diversos con usos y convenciones específicas (Gobato, 2014). A pesar de esto, llama la atención que el área siga manifestando limitaciones metodológicas. Muchos de los trabajos desarrollados en estos años presentan carencias en aspectos relacionados con la recogida de datos o con la representatividad de los mismos. Aunque estas cuestiones se vienen corrigiendo, y cada vez son menos los trabajos que parten de muestras inestables o poco sistemáticas, no es raro encontrar evidentes limitaciones en lo que respecta a la representatividad necesaria para legitimar cualquier estudio (Torruella & Llisterri, 1999). Justificaciones existen para esta situación. En primer lugar, identificamos aquella implícita en el establecimiento de corpus de datos: tarea ardua, complicada, que requiere mucha dedicación y que no siempre puede ser asumida de forma eficiente por una persona o por un grupo reducido. En segundo lugar, podemos aludir a otra específica de los géneros relativos a la comunicación interpersonal, que atañen a la esfera de la comunicación privada, en los que el acceso a los datos resulta complejo (Vela Delfa, 2006; Alcántara Plá, 2014).

Por las razones arriba señaladas, muchos investigadores se han visto obligados a trabajar con corpus “escasos” (Ling, 2005) o “fortuitos” (Campano Escudero, 2007), por ejemplo, en la comunicación por SMS. Estas circunstancias limitan el avance y la consolidación de la disciplina y justifican la necesidad de establecer corpus de textos de CMO que faciliten el asentamiento de las investigaciones. Como veremos en la siguiente sección, existen intentos en esta dirección en

ISBN 978-987-3617-89-8

diversas lenguas, pero todavía son escasos los avances en este sentido en relación a la lengua española, que constituye, sin embargo, la tercera más usada en las comunicaciones digitales. En este trabajo avanzamos en esta reflexión, al tiempo que proponemos soluciones en el marco del proyecto CoDiCE.

## › *Colecciones de datos lingüísticos: el caso de la comunicación digital*

Diferentes corrientes de estudio lingüístico encontraron en el uso de corpus un complemento para los objetivos de su investigación. El primer antecedente claro remite a trabajos de la década del cincuenta, orientados hacia la adquisición del lenguaje, lingüística comparativa e histórica, la dialectología o la enseñanza de la lengua.

> Dentro de esta tendencia empirista pre-chomskiana destacan los trabajos realizados por lingüistas de la talla de Z. Harris, A. Hill o C. Fries, para los que el uso de un corpus (es decir, una colección lo suficientemente amplia de texto producido de forma espontánea) era condición suficiente y necesaria para el estudio lingüístico (Pérez Hernández, 2002).

Sin embargo, a partir del avance de los dispositivos tecnológicos ha crecido exponencialmente la investigación lingüística basada en corpus por la aparente facilidad para almacenar y analizar los datos digitalizados.

De esta manera, en los últimos años, las colecciones de datos se han tornado pertinentes para la mayoría de las líneas teóricas (Ädel & Reppen, 2008) y fundantes para el trabajo empírico, aún a pesar de algunas dificultades para la sistematización de los datos, su respectivo análisis y respuestas inexistentes para el almacenaje. La definición de los corpus lingüísticos y su relevancia metodológica requiere identificar, en primer lugar, los límites y alcances desde su especificidad y las aproximaciones a través y a partir de ellos, en el marco de los estudios lingüísticos. De hecho, la confusión terminológica en torno a la lingüística de corpus como una rama dentro de los estudios sobre la lengua ha significado reacciones opuestas. En palabras de Leech (1992):

> But is corpus linguistics really comparable with these other hyphenated branches of linguistics? [socio-linguistics, psycholinguistics, text linguistics] No, because "corpus linguistics" refers not to a domain of study, but rather to a methodological basis for pursuing linguistic research.

Siguiendo este planteamiento, consideramos el establecimiento de un corpus de comunicaciones digitales como un complemento metodológico para investigaciones orientadas desde las diversas disciplinas lingüísticas.

Un corpus es, inicialmente, una larga colección de datos almacenados electrónicamente, factibles de ser analizados por algún software específico; cuando este conjunto es demasiado extenso e imposible de analizar en una unidad, se habla de *corpora* (McEnery, 2013). Si bien en los estudios pioneros no se identificaba una necesidad estricta de informatizar los corpus, el desplazamiento actual los coliga a su codificación y procesamiento a través de programas de computadora (Torruella & Llisterri, 1999). Por tanto, estos conjuntos de datos suelen estar

ISBN 978-987-3617-89-8

anotados, de manera manual o automática a través de algún programa específico. Es lícito diferenciar los corpus de otros conjuntos de datos como los archivos/colecciones informatizados y bibliotecas de textos electrónicos, ambos recopilaciones sin criterios lingüísticos (Torruella & Llisterri, 1999: 48).

El tipo de corpus que se utilice en la investigación dependerá, en primera instancia, de la hipótesis, pregunta de investigación y del objetivo que se persiga. De la misma manera, el volumen de datos se relaciona de manera estricta con los intereses del estudio. La etapa de elaboración implica, como se mencionó anteriormente, recolectar una gran cantidad de datos (sea de realizaciones orales o escritas). En este momento, abogar por la representatividad de los datos recolectados favorecerá varios aspectos de la investigación. De esta manera, el uso de corpus como fundamento empírico de una investigación de corte lingüística permite generar los instrumentos de estudio y recolección de datos (Hernández Sampieri, Fernández-Collado & Baptista Lucio, 2006) necesarios para favorecer la objetividad y la confianza en la validez de los resultados de estudios extensos (Parodi, 2008).

## *Tipología*

Los tipos de corpus existentes pueden ser elaborados desde, al menos, dos perspectivas distintas: por un lado, a partir de su estructura formal y su organización (puntos 1 a 4) (Beißwenger & Storrer, 2008a) y, por otro, desde su contenido (puntos 5-9)[1].

1. Corpus derivados de proyectos de investigación: compilados en la elaboración de datos para investigaciones particulares a partir de preguntas de investigación.
2. Corpus para uso general: no se integran con alguna pregunta de investigaciones particulares y sirven para distintas hipótesis.
3. Corpus de datos sin procesar o simples: se accede a los datos tal como fueron recolectados inicialmente.
4. Corpus anotados: los datos están anotados ya sea de manera manual o a través de algún software específico.
5. Corpus generales: es una muestra de la lengua en diversos ámbitos, con el fin de mostrar datos generales sobre las comunidades de habla.
6. Corpus especializados: a diferencia del corpus general, recogen datos en torno a un tipo particular de lengua. Se diferencian de los subcorpus, que recogen variedades (Torruella & Llisterri, 1999).
7. Corpus orales: hay que distinguir entre corpus de datos orales y transcripciones de datos de la oralidad. Los corpus orales se conforman de registros orales que pueden estar acompañados o no de sus respectivas transcripciones (Torruella & Llisterri, 1999).
8. Corpus textuales: muestras de lengua escrita, que pueden pertenecer a diferentes registros,

[1] La siguiente lista no contempla todas las clasificaciones posibles. Para ampliar, véase Torruella & Llisterri (1999).

ISBN 978-987-3617-89-8

géneros y formatos (Torruella & Llisterri, 1999).

9. Corpus de comunicaciones digitales: a diferencia de los corpus textuales, diferenciamos aquellos repositorios que disponen de datos extraídos de entornos virtuales, que mantienen alguna de las características de este soporte (puede ser a través de la notación). Tal como señalamos anteriormente, el soporte condiciona y determina la interacción en estos medios, por lo tanto, por más que sea de orden textual u oral.

Recientemente las posibilidades tecnológicas, sumadas a la trayectoria de estudios en la materia, han dado muestras de avances dentro de la lingüística de corpus. De esta manera, su empleo no se circunscribe únicamente a campos como el de la Lexicografía y afines, sino que, como metodología, la Lingüística de corpus ofrece beneficios para complementar la mayoría de las áreas de estudio lingüístico. Una recolección de datos apropiada dará información no solo sobre cuestiones léxicas, sino también permitirá caracterizar "diferentes niveles del lenguaje (vulgar, culto, literario, etc.), datos, estos últimos, muy interesantes no solo para los estudios lexicográficos sino también para los estudios sociolingüísticos y estilísticos" (Torruella & Llisterri, 1999: 4).

## › *Sobre los corpus de comunicaciones digitales*

### Breve panorámica de los corpus del discurso digital en el ámbito internacional

En este apartado presentamos una panorámica general de los corpus de discurso digital recogidos en el ámbito internacional. En la siguiente tabla se esquematizan los más representativos:

| **Enlace o Bibliografía** | **Descripción** | **Tipo de corpus** | **Idioma/s** | **Nombre del corpus** |
|---|---|---|---|---|
| Yates, 1996 | 50 presentaciones de 152 conferencias de informática | Corpus simple | Inglés | *CoSy:50 Corpus* |
| http://www.linguistik-online.de/15_03/Pow.pdf | Chats | Corpus simple | Sueco y alemán | *German-Swedish IRC-Corpus* |
| http://spamassassin.apache.org/publiccorpus/ | > 6000 mensajes de correo electrónico spam | Corpus simple para uso general | Inglés | SpamAssassin Public Corpus |
| http://www.coli.unisaarlan | 160 emails | Corpus de | Alemán | *E-Mail corpus from the* |

ISBN 978-987-3617-89-8

| | | | | |
|---|---|---|---|---|
| d.de/publikationen/softcopies/Declerck:1997:EKE.pdf | | proyectos de investigación | | *COSMA project* |
| http://www.chatkorpus.tu-dortmund.de/ | >5000 chats | Corpus anotados para uso general | Alemán | *Dortmund Chat Corpus* |
| http://www.sud4science.org/ | >88.000 SMS | Corpus de proyectos de investigación anotado | Francés y otras lenguas de Francia | *Sud4science* |
| http://www.lsi.upc.edu/~nlp/wikicorpus/ | Español: 120 millones de palabras<br>Catalán: 50 millones de palabras<br>Inglés: 600 millones de palabras | Corpus anotado para uso general | Español, Catalán e Inglés | *Wikicorpus* |

*Tabla 1. Elaboración propia sobre la lista provista por CMC-Corpora disponible en http://www.cmc-corpora.de/, Beißwenger & Storrer (2008a; 2008b) y Panckhurst & Moïse (2012).*

Algunas investigaciones toman sus datos a partir de espacios públicos, es decir, mediante muestras que no han sido elicitadas (De-Matteis, *en este volumen*). Así, por ejemplo, Herring y Zelenkauskaite (2009) trabajan con un repositorio abierto de interacciones por SMS, tomadas en programas de televisión; otros estudios trabajan con datos tomados de páginas web, tal es el caso del proyecto *Wikicorpus* (http://www.cs.upc.edu/~nlp/wikicorpus/) que, sin embargo, se restringe a los géneros monologales y no recopila datos interaccionales. Si bien estos sitios ofrecen muestras de lengua para el investigador, descuidan cuestiones de ética al no informar a los usuarios sobre el uso de sus producciones y resultan poco representativas por ser contextos demasiado específicos.

Un antecedente interesante por la metodología empleada para la recolección de datos y para su presentación es el corpus de comunicación por SMS del proyecto *Sud4Science*. A partir de una plataforma online, se recogieron más de 90.000 SMS en Francia (Panckhurst & Moïse, 2012). Aún en fase de sistematización de los datos, este colosal proyecto reunió, entre 2011 y 2012, sus muestras gracias a colaboradores que transcribían sus SMS en la plataforma provista por el grupo de investigación. La mayoría de ellos completaron una encuesta sociolingüística con datos sobre

ISBN 978-987-3617-89-8

edad, sexo, variedades lingüísticas (materna y bilingüismo), nivel de estudio actual, profesión. Asimismo, se recogió información sobre tipo de teléfono, paquete de mensajes contratado y representaciones sobre prácticas relacionadas al envío de SMS. Esta propuesta resulta muy interesante para el análisis de múltiples aspectos lingüísticos, tal como señalan las autoras. Sin embargo, las manifestaciones recolectadas atienden a una población no representativa de la sociedad general, ya que todos los colaboradores son personas que utilizan internet con frecuencia, tienen acceso a él y, por tanto, pertenecen al grupo de los info-ricos (Quevedo, 2012). En *Sud4Science*, se dejan de lado cuestiones culturales que atañen a la competencia comunicativa de los usuarios en este medio particular. Esta técnica podría ser replicada, aunque con esta metodología no se atiende a la variable socioeducativa ni estrato sociocultural (Cantamutto, 2015).

## Los corpus de discurso digital en lengua española

Llegados a este punto, cabe preguntarse qué presencia tienen los datos del español en los corpus de comunicaciones digitales. Para ello analizaremos su manifestación en dos tipos de corpus. Por un lado, en corpus generales y, por otro, en los corpus originados en el seno de proyectos de investigación específicamente orientados al discurso digital.

Dentro del conjunto de corpus generales, en el español se destaca el CORPES (Corpus del Español del Siglo XXI). Está integrado por datos recolectados entre 2001 y 2012, consta de 180 millones de formas y responde a los parámetros establecidos por la Real Academia Española. Entre estos, se destacan los siguientes:

a) La composición diatópica es de 70% de formas latinoamericanas y 30% de formas de España.
b) El 10% corresponde a formas de la oralidad y el 90% a material escrito.
c) Está separado en dos grandes bloques temáticos: ficción y no ficción.

Un aspecto relevante es la clasificación de los datos, consiste en combinar el criterio de género discursivo con criterios derivados del medio y del soporte en el que se produjeron las muestras de lengua. El corpus es semi-abierto y está aún en construcción; sin embargo, llama la atención que un corpus del siglo XXI no incorporara soportes más novedosos. En los datos estadísticos, se evidencia la escasa representatividad del material proveniente de internet (no representa siquiera un 1%), respecto a libros y a la prensa (véase *Tabla 2*).

ISBN 978-987-3617-89-8

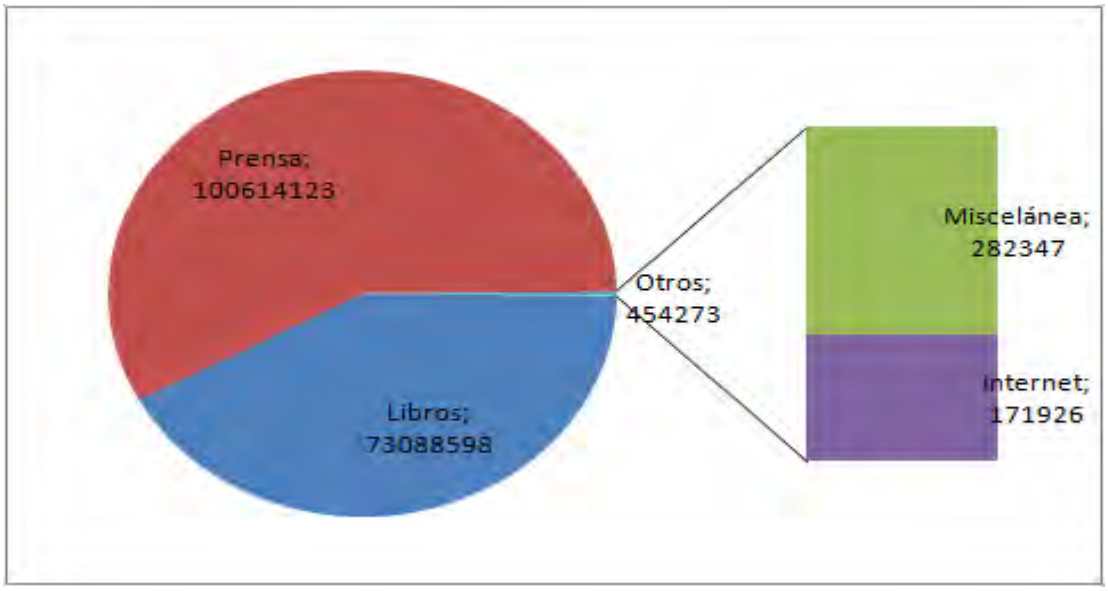


*Gráfico 1. Distribución de formas por soporte. Datos extraídos de http://web.frl.es/CORPES/org/publico/pages/ayuda/informacion.view.*

Por su parte, el *Corpus del Español* (http://www.corpusdelespanol.org/x.asp) ofrece 100.000.000 de palabras y herramientas para su análisis, también de acceso online. En la sección relativa a los siglos XX y XXI, es posible consultar textos extraídos de diversos portales de internet[2] sin ser clasificados como característicos de determinado soporte. Al observar ambos corpus, llama la atención no solo la infrarrepresentación de las muestras de lengua procedentes de internet, sino también el hecho de que estos textos no se correspondan en ningún caso con muestras de la CMO. Ninguno de estos corpus incluye datos interaccionales, limitándose a muestras de géneros monologales en soporte digital (tal como son los portales de la prensa online).

Existen, sin embargo, investigaciones que abordan la comunicación digital a través de trabajo empírico de recolección de datos y conformación de corpus en español. Tal es caso de la investigación sobre correo electrónico (Vela Delfa, 2006), chat (Sanmartín Sáez, 2007) y plataforma Galanet para el aprendizaje de L2 (Álvarez Martínez, 2008) en España, la comunicación por SMS en Argentina (Cantamutto, 2013). A continuación, se sistematizan algunos de los antecedentes encontrados de investigaciones particulares que utilizan datos primarios.

| **Referencia** | **Descripción del corpus** | **País** | **Dominio** | **Carácter** | **Tipo de interacción** |
|---|---|---|---|---|---|

[2] Por ejemplo, *Encarta* y *El nuevo Herald*.

ISBN 978-987-3617-89-8

| | | | | | |
|---|---|---|---|---|---|
| Vela Delfa, 2006 | >330 emails recogidos entre 2001 y 2004 | España | Privado | Monolingüe | Correo electrónico |
| Mariottini, 2006 | 100.000 palabras de chat recogidas en 2004 | España | Privado | Bilingüe | Chat |
| Álvarez Martínez, 2008 | 55 chats de entre 15 minutos y 2 horas entre 2004 y 2007 | España | Semi-público | Monolingüe | Chat |
| Noblia, 2009 | 20 chats grupales y de persona a persona de *Messenger* y ICQ, entre 2001 y 2002 | Argentina | Privado y público | Monolingüe | Chat |
| Kaul de Marlangeon y Cordisco, 2014 | >70.000 palabras de 1897 comentarios recogidos en 2013 | Argentina | Público | Monolingüe | Redes sociales |
| Cantamutto, 2013; Cantamutto, 2014 | >6000 SMS de diferentes grupos etarios recogidos entre 2011 y 2014 | Argentina | Privado | Monolingüe | SMS |

*Tabla 2. Relación de corpus de CMO en español de investigaciones privadas.*

En particular, no interesa comentar el trabajo de Kaul de Marlangeon & Cordisco (2014) para hacerse una representación del estado metodológico del área. Estos autores manejan un corpus de comentarios de *Facebook*, *Twitter* y otros sitios electrónicos, recolectado en 2013. El material está disponible online, a través de un enlace que, sin embargo, direcciona a la red social de donde fue extraída la muestra (Kaul de Marlangeon & Cordisco, 2014: 148). De esta manera, tenemos una selección de datos en su contexto sin tratamiento ni sistematización: una página de *Twitter* recortada sobre una temática particular.

Tal como se desprende de la *Tabla 2,* existen datos primarios sobre la CMO en lengua española, pero, no son sistemáticos, no están coordinados y no responden a ningún tipo de estándar. Esta situación no permite la retroalimentación de las investigaciones ni favorece la optimización de los recursos. La recogida de un corpus es siempre una labor ardua y compleja, pero

ISBN 978-987-3617-89-8

cuando estos forman parte de la comunicación interpersonal -géneros de la CMO, en general- e incluso de la esfera privada -SMS, correo electrónico, chat, redes sociales, entre otras- el trabajo se torna todavía más difícil.

En este marco, CoDiCE surge como un intento de compensar estas carencias, con el objetivo de generar un espacio de recogida y sistematización de datos abierto a la comunidad científica general. En los siguientes apartados describiremos las principales características de este proyecto en marcha.

## › *CoDiCE: un corpus posible*

Hasta aquí hemos presentado la necesidad de disponer de datos estables y, en la medida de lo posible, masivos y relativos a los diferentes géneros discursivos interpersonales desarrollados en los entornos de comunicación mediatizada en español. Por ello, esta sección está destinada a diseñar una estrategia para subsanar esta necesidad que derive en el proyecto CoDiCE.

En un contexto como el que acabamos de describir, parece indicado que, a falta de un proyecto más ambicioso y sistemático, se propusiera la creación de un repositorio colaborativo de acceso abierto. Entendemos por repositorio un depósito de archivos digitales de diferentes tipologías, creados con el fin de difundirlos y preservarlos para hacerlos accesibles a la comunidad científica. Un modelo colaborativo permite optimizar recursos, puesto que podrían compartirse muestras recogidas para investigaciones concretas, al tiempo que se generalizan estándares, al proponerse unos principios metodológicos comunes para la recogida de datos. En este marco, el proyecto CoDiCE busca la creación de un repositorio digital abierto de muestras de datos de interacciones mediatizadas en español con orientación pragmática y sociolingüística.

### *Antecedentes de repositorios digitales abiertos y colaborativos de datos lingüísticos: el caso de THE TALKBANK y CHILDES*

La iniciativa de desarrollar repositorios para compartir muestras de datos tiene una larga tradición en áreas de conocimiento marcadamente empíricas, como por ejemplo, la psicología[3]. Esta tendencia se suma a otra más general, conocida como *data sharing*, definido por Torres-Salinas, Robinson-García & Cabezas-Clavijo (2012): "consiste en compartir los datos de las investigaciones de los científicos, con el objetivo de unir esfuerzos y optimizar el uso de los recursos". Además, algunos consejos de investigación públicos insisten en que los datos recolectados a través de fondos y subsidios estatales deben estar accesibles para otros investigadores. Algo similar ocurre en Argentina tras la implementación de la Ley 26.899/2013 de Repositorios Digitales Científicos.

Dentro de esta tendencia, situamos el banco de datos conocido como *Child Language Data*

[3] Se puede consultar la página web de la Asociación Americana de Psicología, para comprobar que estos repositorios abarcan áreas tan dispares como los hábitos de consumo o salud adolescente. Véase la lista ofrecida en http://www.apa.org/research/responsible/data-links.aspx.

ISBN 978-987-3617-89-8

*Exchange System* (CHILDES), que se integra dentro de un repositorio de mayor alcance y de temática más generalizada, *The TalkBank*[4]. Como se describe en el propio proyecto:

> […] el objetivo de TalkBank es fomentar la investigación fundamental en el estudio de la comunicación humana y animal. Para ello propone construir muestras de datos de todos los subcampos que constituyen el estudio de la comunicación. Se pretende utilizar esta base de datos para avanzar en el desarrollo de estándares y herramientas para crear, compartir, buscar y hacer comentarios en torno a datos primarios, a través de ordenadores conectados en red (*la traducción es nuestra*)[5].

A pesar de la intención manifestada por el proyecto de abarcar todos los ámbitos de la comunicación, la realidad muestra que este repositorio dista todavía de alcanzar ese objetivo. No contiene, por ejemplo, datos relativos a comunicación mediatizada.

Una de las áreas más completa en *The TalkBank* es la de adquisición del lenguaje, a través del proyecto CHILDES. Los objetivos de este programa son los siguientes:

a) Proporcionar más datos de más niños hablantes de más lenguas.
b) Obtener mejores datos mediante un sistema de transcripción consistente y documentado.
c) Automatizar el proceso de análisis de los datos (Diez-Itza *et al.*, 1999: 519).

El propósito transversal fue crear una base de datos informatizada nutrida por las transcripciones que los investigadores pudieran aportar. De esta manera, en poco tiempo se obtuvieron gran cantidad de corpus del inglés aunque no así del español (Diez-Itza et al., 1999: 520). La plataforma provee, además, herramientas para el análisis de las bases de datos. CLAN es un paquete de programas específicamente diseñados para analizar esos archivos que contienen transcripciones de muestras de habla, incluye recuentos de frecuencias, búsqueda de palabras, análisis de la interacción, etc. (Diez-Itza *et al*., 1999: 520). Además, propone un sistema de estándares de codificación CHAT para la elaboración de las transcripciones.

La posibilidad de acceder a los recursos enumerados favoreció el empleo de esta plataforma y la extensión de su uso reforzó la validez de los estándares que constituyen, actualmente, una norma en los estudios de adquisición del lenguaje. ¿Qué ventajas supone un proyecto de este tipo para un área de conocimiento? Desde nuestro punto de vista, estas iniciativas ayudan a mejorar la orientación metodológica de un área, al tiempo que permite la rápida difusión de los datos de investigación (en el apartado *Ground Rules*, este proyecto invita a compartir todos los resultados de las investigaciones obtenidos a través de datos primarios provenientes del repositorio).

A pesar de que los objetivos de CoDiCE difieren de los aquí comentados, hemos considerado apropiado introducirlos porque, en buena medida, representan un modelo para el desarrollo de nuestro proyecto.

[4] Disponible en http://talkbank.org/.
[5] En el original: "The goal of TalkBank is to foster fundamental research in the study of human and animal communication. It will construct sample databases within each of the subfields studying communication. It will use these databases to advance the development of standards and tools for creating, sharing, searching, and commenting upon primary materials via networked computers".

ISBN 978-987-3617-89-8

## *Características de CoDiCE*

Como se observa a lo largo del trabajo, CoDiCE propone la creación de un repositorio de comunicaciones digitales en español, a partir de las aportaciones de los trabajos parciales de investigadores de este campo disciplinar. Es decir, se intenta optimizar los recursos invertidos en la recopilación de muestras de lenguas. De esta manera, se pondrán a disposición tanto datos de fuentes primarias como trabajos que aborden aspectos teóricos y metodológicos sobre la comunicación digital. Asimismo, se plantea como objetivo complementario la creación de unos estándares comunes en la recogida de los datos, en lo que concierne principalmente a los factores contextuales y situacionales, a fin de facilitar los análisis sociopragmáticos. Por ello, CoDiCE debe arrancar con una reflexión metodológica que permita el diseño de un repositorio eficaz de interacciones con las particularidades que presentan los entornos de CMO y que presentamos con detalle en el apartado siguiente.

### Reflexiones metodológicas

Para la constitución de un corpus de comunicaciones digitales destinado a su estudio lingüístico, es necesario atender a diversos factores derivados de la especificidad del medio en que tienen lugar estas interacciones. Así, deben observarse los condicionantes derivados del soporte, las interfaces que mediatizan el intercambio imponen condiciones al mensaje, las particularidades de la situación de comunicación. Cada entorno comunicativo presenta condiciones de enunciación particulares que deben considerarse en la recogida de datos. La relación y familiaridad de los usuarios con los soportes también puede influir en la naturaleza de los datos recogidos. A continuación nos detenemos a explicar la influencia de cada uno de estos factores:

1. *Condicionantes del soporte (interfaz)*: Cada una de las plataformas y dispositivos donde se desarrolla la comunicación tiene características propias que repercuten en las muestras de lengua con las que se encuentra el investigador. De esta manera, una emisión enviada desde una computadora a través de una red social puede ser recibida en un teléfono móvil como mensaje de texto. En cada interacción, las características de la plataforma y del dispositivo intervienen en ambas direcciones de la díada imponiendo ciertas condiciones al mensaje, tanto desde el punto de vista paratextual, como a las posibilidades específicas de la interacción, a saber tipo y cantidad de datos que pueden enviarse. Todos estos elementos pueden determinar características que influyan en los distintos niveles de lengua.
2. *La trama de interfaces que median la comunicación* (Yus, 2010; Vela Delfa & Jiménez Gómez, 2011; Cantamutto, 2013; Gobato, 2014): En la comunicación cara a cara se obtienen la participación de ambos interactuantes de manera relativamente sencilla en el mismo momento. En cambio, en la comunicación digital pueden ocurrir diferentes situaciones que entorpecen la tarea del investigador: a) la respuesta es inmediata, b) la respuesta se dilata, c) no hay respuesta, d) la respuesta ocurre en otro medio/soporte. En determinadas interacciones, es necesario

ISBN 978-987-3617-89-8

explicitar la necesidad de respuesta mediante interrogaciones u otras formas lingüísticas y paralingüísticas apropiadas a cada soporte. Además, la interacción no comparte un espacio físico común de desarrollo, sino que se representa de modo paralelo, a veces simultáneo, aunque no necesariamente, en la aplicación que cada usuario esté empleando. Este condicionante provoca que, en no pocas ocasiones, la representación de los signos lingüísticos no sean equivalentes en las distintas interfaces -interfaz de producción *vs* interfaz de producción- rompiéndose criterios de linealidad que afectan, por ejemplo, a los procedimientos de cohesión discursiva.

3. *La representación de las situaciones de comunicación:* Los datos lingüísticos, producidos en situaciones mediatizadas tecnológicamente, se corresponden con situaciones de comunicación donde los interlocutores no comparten el contexto. El proceso de producción y de recepción se desarrolla en entornos diferentes, tanto espacial como temporalmente. Estos elementos (nula o escasamente reflejados en los datos recogidos), pueden tener influencia en cuestiones como la gestión temporal del intercambio, en un intervalo cada vez más difuso entre sincronía y asincrónica, o la retroalimentación mutua entre interlocutores. Algunas aplicaciones ofrecen reflejos de la situación de comunicación, como las marcas de conectado/desconectado o las indicaciones de "x está escribiendo"; marcas que se pierden en la mayoría de los corpus que no son recogidos en el proceso y, en la mayoría de los casos, se corresponden con una representación en pantalla del intercambio. Así, Anderson, Beard & Walther (2010) se plantean la necesidad de grabar el proceso de desarrollo conversacional para contrastar estos datos con los representados en las pantallas de los interlocutores, ya que presumen que existe un alto grado de variación entre la representación local del intercambio percibida por cada interlocutor. La propuesta es interesante, pero poco factible. No obstante, reflexiona sobre la necesidad en el proceso de recolección de incluir, en la medida de las posibilidades de cada contexto, referencias a la situación de comunicación.
4. *La reconstrucción de los contextos*: Por tratarse de entornos interaccionales que configuran discursos co-construidos, resulta muy común que, en las diferentes experiencias de recolección de este tipo de datos, parte del contenido de la interacción esté elidido y se vincule con el entorno cognoscitivo compartido por los hablantes o con una interacción cuyo desenlace se realiza en otra interface artefactual (véase punto 2).
5. *Variables sociolingüísticas:* La mayoría de la bibliografía revisada se centra en adolescentes y jóvenes, atendiendo a este grupo etario donde se centra la mayor riqueza para el análisis (Domínguez Cuesta, 2005; Avendaño, 2007; Andrade Hidalgo, 2008). El habla adolescente presenta particular interés para cualquier sociolingüista, ya que en esta etapa de grandes cambios también se producen modificaciones en el plano de la dinámica lingüística con recurrencias al carácter lúdico y críptico (Sobrero, 1993: 95). Sin embargo, si es natural en los jóvenes realizar elecciones lingüísticas que innovan en distintos niveles de lengua y construyen discursos identitarios (Zimmermann, 2003), ¿por qué la discusión sobre las consecuencias

ISBN 978-987-3617-89-8

*lingüísticas* de las nuevas tecnologías recae sobre las prácticas de este grupo? No debe desestimarse que otros grupos etarios imiten esta posible jerga juvenil (Palazzo, 2005; Campano Escudero, 2007; Andrade Hidalgo, 2008) al igual de lo que ocurre en rasgos de la interacción cara a cara (Rígano, 1998). Por tanto, definir adecuadamente la variable edad, atendiendo al uso que los adultos hacen de la lengua, e incorporar entrevistas o tests de hábitos sociales permitirá recabar información complementaria para comprender el fenómeno en su totalidad. Asimismo, es necesario delimitar cuáles rasgos detectados responden a restricciones del dispositivo, cuáles a la competencia comunicativa y cuáles al carácter lúdico y críptico propio del habla adolescente e imitado por otros grupos. Por último, para un análisis sociolingüístico de la comunicación digital es necesario atender a todas las variables sociodemográficas (Cantamutto, 2015), omitir otros perfiles de interactuantes puede conducir a observaciones erróneas: en la variación intragrupal hay una amplia riqueza de elementos para considerar. Realizar un muestreo intencionado reconociendo los años de escolarización de los informantes elegidos permite obtener datos más fiables en relación a la variación sociolingüística. Asimismo, la incorporación de la variable sexo permitirá registrar los usos diferenciados que puedan estar también vinculados a cuestiones identitarias de género (Ling, 2002; Herring & Zelenkauskaite, 2009).

6. *Cuestiones de ética. Consentimientos informados. Anonimización*: El estudio de la comunicación por interfaces artefactuales presenta una dificultad inmediata para el investigador: cómo recoger los datos de intercambios realizados en un medio que se lo suele preferir por considerarse íntimo y privado. Desde una perspectiva ética, acceder a estas interacciones implica estar adentrándonos en su vida privada e involucrando información que puede perjudicar a otras personas. En tal sentido, se deben implementar diversas estrategias para proteger a los participantes de nuestra investigación, respetando su autonomía y cuidando de no afectar su privacidad. En particular, a partir de la firma de consentimientos informados por parte de todos los interactuantes y, si son menores, también por parte de sus padres. La propuesta del CoDiCE no ofrece riesgos potenciales para los participantes ya que los resultados publicados han sido a partir de datos totalmente anonimizados (Christians, 2000: 145).

Esta lista no agota todos los condicionantes que deberían considerarse. Queda fuera, por ejemplo, la recogida de datos multimodales tan característicos en la comunicación a través de sistemas de mensajería instantánea, que afectan directamente al sistema de implementación del corpus y que, por tanto, resultan cruciales en una segunda fase del proyecto.

## › *Palabras finales*

En este artículo hemos justificado la necesidad de una reflexión precisa y sistemática sobre los métodos de recogida de datos en un área de estudio emergente: el análisis de los discursos

ISBN 978-987-3617-89-8

digitales. Hemos comprobado que, aunque existen propuestas más o menos amplias en distintas lenguas, todavía dista de ser un ámbito discursivo suficientemente representado. En el caso específico del español, la situación todavía es extremadamente deficiente (Danet & Herring, 2007). A pesar de tratarse de la tercera lengua más empleada en los intercambios mediatizados, todavía no se ha llevado a cabo una recogida sistemática de datos relativos a este tipo de comunicación.

En este trabajo hemos revisado la presencia de estas muestras de lenguas tanto en corpus generales del español, en los que no están representadas, como en corpus recogidos para investigaciones concretas, orientadas al estudio de la CMO. Estas últimas, aunque numerosas, carecen de sistematicidad. Esta situación es la que nos lleva a defender la idoneidad del diseño de un repositorio abierto y colaborativo para la compilación de un corpus de interacciones digitales. El modelo a seguir sería el impulsado en otras áreas, como por ejemplo el desarrollado por *The TalkBank*. Con este marco de referencia surge CoDiCE, que busca generar el entorno propicio para el desarrollo de ese repositorio de datos. Por ello, dos son los objetivos concretos del proyecto: 1) el establecimiento de una serie de estándares que guíen la recogida de datos y 2) la alimentación del repositorio a partir de las contribuciones de investigadores concretos que aporten datos primarios.

El primero de estos objetivos constituye un requisito previo para el segundo, por ello se aborda de forma inicial. Como paso previo para el desarrollo de CoDiCE, es necesaria una reflexión metodológica sobre las condicionantes de este tipo de corpus. En este sentido, presentamos seis elementos que deben tenerse en cuenta a la hora de abordar la especificidad de estos datos:

1. Condicionantes del soporte
2. La trama de interfaces que median la comunicación
3. La representación de las situaciones de comunicación
4. La reconstrucción de los contextos
5. Variables sociolingüísticas
6. Cuestiones de ética, consentimientos informados, anonimización

Una vez establecida la reflexión en torno a los principios sobre los que será diseñado el corpus comienza una segunda etapa donde se proceda a la compilación propiamente dicha. Esta fase de desarrollo dirigida a la puesta en funcionamiento de un repositorio abierto de comunicaciones digitales permitirá el avance de investigaciones sobre variación pragmática y sociolingüística intra e interlingüística, estudios diacrónicos y sincrónicos de la lengua española. En la medida en que el proyecto crezca se comenzará a subsanar las aproximaciones metodológicas deficientes que se han observado en estudios sobre la comunicación digital en el español.

## › *Bibliografía*

Ädel, A. & Reppen, R. (Eds.) (2008). *Corpora and Discourse: The Challenges of Different Settings*, 31. Amsterdam: John Benjamins Publishing.

ISBN 978-987-3617-89-8

Alcántara Plá, M. (2014). Las unidades discursivas en los mensajes instantáneos de wasap. *Estudios de Lingüística del Español*, 35, 223-242.

Alonso, E. & Perea, M. (2008). SMS: Impacto social y cognitivo. *Escritos de Psicología*, 2, 22-29.

Álvarez Martínez, S. (2008). *Interactions synchrones écrites en ligne et apprentissage de l'espagnol: caractérisation, potentialités et limites.* (Tesis doctoral internacional en programa de cotutela). Université Stendhal Grenoble 3, Grenoble-Universidad de Lleida, Lleida.

American Psychological Association (s/f). *Links to Data Sets and Repositories.* Recuperado de http://www.apa.org/research/responsible/data-links.aspx el 12/03/2015

Anderson, J. F., Beard, F. K. & Walther, J. B. (2010). Turn-Taking and the Local Management of Conversation in a Highly Simultaneous Computer-Mediated Communication System. *Language@Internet*, 7(7). Recuperado de http://www.languageatinternet.org/articles/2010/2804 el 12/03/2015

Andrade Hidalgo, L. (2008). *Los SMS: nuevas formas de interacción juvenil.* (Tesis de maestría). Flacso sede Ecuador, Quito. Recuperado de www.flacsoandes.org/comun el 12/03/2015

Avendaño, V. (2007). El lenguaje del chat los SMS: ¿un nuevo género discursivo? *Educ.ar: El portal educativo del Estado argentino*. Recuperado de http://portal.educ.ar/debates/eid/lengua/debate/el-lenguaje-del-chat-los-sms-un-nuevo-genero-discursivo.php el 12/03/2015

Baron, N. S. (1984). Computer-Mediated Communication as a Force in Language Change. *Visible Language*, 18(2), 118-141.

---- (1998). Letters by Phone or Speech by Other Means: The Linguistics of Email. *Language and Communication*, 18, 133-170.

Beißwenger, M. & Storrer, A. (Comps.) (2008a). *Corpora of Computer-Mediated Communication.* Recuperado de http://www.cmc-corpora.de/ el 12/03/2015

---- (2008b). Corpora of Computer-Mediated Communication. En Lüdeling, A. & Kytö, M. (Eds.), *Corpus Linguistics: An International Handbook.* Berlin: Mouton de Gruyter.

Bernete García, F. (Coord.) (2007). *Comunicación y lenguajes juveniles a través de las TIC.* Madrid: Instituto de la Juventud, Ministerio de Trabajo y Asuntos Sociales.

Booher, D. (2001). *E-Writing: 21st-Century Tools for Effective Communication.* Nueva York: Pocket Books.

Campano Escudero, B. (2007). Análisis lingüístico-pragmático de un corpus de mensajes SMS. *Ferrán*, 8, 185-210. Recuperado de www.educa.madrid.org/web/ies.jaimeferran.colladovillalba/revista2 el 12/03/2015

Cantamutto, L. (2013). La recursividad de las interacciones contemporáneas. Límites teórico-metodológicos del estudio de los SMS como conversación. *Revista de Ciencias Sociales, segunda época. Dossier: Al abordaje de la comunicación contemporánea. Cultura, lenguaje y sociedad en los mundos de la mediación digital*, 5(23), 83-104.

---- (2014). El discurso de los mensajes de texto en el habla adolescente del español bonaerense. En Parini, A. & Giammatteo, M. (Eds.), *Lenguaje, discurso e interacción en los espacios*

*virtuales* (65-81). Mendoza: UNCuyo-SAL.

---- (2015). Aspectos pragmáticos de la literacidad digital: la gestión interrelacional en la comunicación por teléfono móvil. *Revista Internacional de Tecnología, Conocimiento y Sociedad*, 4, 95-111.

Christians, C. G. (2000). Ethics and Politics in Qualitative Research. En Denzin, N. K. & Lincoln, Y. S. (Eds.), *Handbook of Qualitative Research*, 2, 133-155. Thousand Oaks, California: SAGE.

Covarrubias, J. I. (2008). La ciberhabla juvenil en los Estados Unidos. En López Morales, H. (Coord.), *Enciclopedia del español en los Estados Unidos* (512-538). Madrid: Instituto Cervantes/Santillana.

Crystal, D. (2001). *Language and the Internet*. Cambridge: Cambridge Press.

Danet, B. & Herring, S. C. (Eds.). (2007). *The Multilingual Internet: Language, Culture, and Communication Online.* Oxford: Oxford University Press.

Davies, M. (s/f). *Corpus del Español*. Recuperado de http://www.corpusdelespanol.org/x.asp el 12/03/2015

De-Matteis, L. M. A. (en este volumen). Ejes para un debate sobre el uso ético de datos interaccionales escritos y orales obtenidos en línea.

Diez-Itza, E., Snow, C. E. & MacWhinney, B. (1999). La metodología *RETAMHE* y el proyecto *CHILDES*: breviario para la codificación y análisis del lenguaje infantil. *Psicothema*, 11(3), 517-530.

Domínguez Cuesta, C. (2005). El lenguaje de los SMS y del chat en las aulas. *Cuadernos de pedagogía*, 343, 65-69. Recuperado de www.cuadernosdepedagogia.com/verpdffree.asp?idArt=8870 el 12/03/2015

Dorner, J. (2002). *Writing for the Internet.* Oxford: Oxford University Press.

Eisenchlas, S. A. (2012). Gendered Discursive Practices On-Line. *Journal of Pragmatics,* 44(4), 335-345.

Ferrara, K., Brunner, H. & Whittemore, G. (1991). Interactive Written Discourse as an Emergent Register. *Written Communication*, 8, 8-34.

Gobato, F. (2014). *La escritura secundaria: Oralidad, grafía y digitalización en la interacción contemporánea*. Bernal: Universidad Nacional de Quilmes.

Goffman, E. (1959). *The Presentation of Self in Everyday Life*. New York: Anchor.

Gruber, H. (2000). Scholarly Email Discussion List Postings: A Single New Genre of Academic Communication? En Pemberton, L. & Shurville, S. (Eds.), *Words on the Web: Computer-Mediated Communication* (36-43). Bristol: Intellect Books.

Hernández Sampieri, R., Fernández-Collado, C. & Baptista Lucio, P. (2006). *Metodología de la investigación*. México: McGraw-Hill.

Herring, S. C. (Ed.). (1996). *Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives* (Pragmatics & Beyond New Series, 39). Amsterdam: John Benjamins Publishing.

---- (1999a). Interactional Coherence in CMC. *Journal of Computer-Mediated Communication*, 4(4). Recuperado de http://www.ascusc.org/jcmc/vol4/issue4/herring.html el 12/03/2015

---- (1999b). The Rhetorical Dynamics of Gender Harassment On-Line. *The Information*

*Society,* 15(3), 151-167.

---- (2001). Computer-Mediated Discourse. En Schiffrin, D., Tannen, D. & Hamilton, H. (Eds.), *The Handbook of Discourse Analysis* (612-634). Oxford: Blackwell Publishers.

---- (2004). Computer-Mediated Discourse Analysis: An Approach to Researching Online Behavior. En Barab, S. A., Kling, R. & Gray, J. H. (Eds.), *Designing for Virtual Communities in the Service of Learning* (338-376). Cambridge: Cambridge University Press.

Herring, S. C. & Zelenkauskaite, A. (2009). Symbolic Capital in a Virtual Heterosexual Market: Abbreviation and Insertion in Italian iTV SMS. *Written Communication*, 26, 5-31.

Ivars, O. G. (2007). El chat y la reconfiguración enunciativa. En *Actas de las I Jornadas Nacionales de Investigación Educativa. II Jornadas Regionales- VI Jornadas Institucionales*. Mendoza: Universidad de Cuyo. Recuperado de http://www.feeye.uncu.edu.ar/web/posjornadasinve/area3/Lengua%20-%20Didactica%20de%20la%20lengua%20-%20TICs/275%20-%20Ivars%20-%20FEEyE.pdf el 12/03/2015

Kaul de Marlangeon, S. & Cordisco, A. (2014). La descortesía verbal en el contexto político-ideológico de la redes sociales. *Revista de Filología*, *32*, 145-162.

Kiesler, S., Siegel, J. & McGuire, T. W. (1984). Social Psychological Aspects of Computer-Mediated Communication. *American Psychologist*, 39(10), 1123-1134.

Kress, G. (2003). *Literacy in the New Media Age.* New York: Routledge.

Leech, G. (1992). Corpora and Theories of Linguistic Performance. Svartvik, J. (Ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991* (105-122). Berlin: Mouton de Gruyter.

Ling, R. (2002). Chicas adolescentes y jóvenes adultos varones: dos subculturas del teléfono móvil. *Revista de Estudios de Juventud, 57,* 33-46. Recuperado de http://www.injuve.es/ el 12/03/2015

---- (2005). The Sociolinguistics of SMS: An Analysis of SMS Use by a Random Sample of Norwegians. *Mobile Communications: Re-negotiation of the Social Sphere* (335-349). London: Springer. DOI: 10. 1007/1-84628-248-9_22

MacWhinney, B. (Coord.) (1999). *TalkBank*. Recuperado de http://talkbank.org/

Martin, M. V. (2006). Jóvenes, identidad y telefonía móvil: algunos ejes de reflexión. *Revista TEXTOS de la CiberSociedad,* 1. Recuperado de http://www.cibersociedad.net/congres2006/gts/comunicacio. php?id=693 el 12/03/2015

Martín Corvillo, J. M. (2014). Propuesta metodológica para el estudio del lenguaje de la protesta y su transmisión a las redes sociales. *LinRed*, 12. Recuperado de http://www.linred.es/articulos_pdf/LR_articulo_17052014.pdf el 12/03/2015

McEnery, T. (2013). *Corpus: Some Key Terms*. UK: Lancaster University.

Muñiz Calderón, R. (2011). *Análisis contrastivo de las variaciones lingüísticas y culturales en la comunicación digital entre hablantes no nativos.* (Tesis inédita de doctorado). Universidad Politécnica de Valencia, Valencia. Recuperado de https://riunet.upv.es/bitstream/handle/10251/14573/tesisUPV3707.pdf?sequence=1 el

12/03/2015

Noblia, M. V. (2000). Conversación y comunidad: Las chats en la comunidad virtual. *Revista Iberoamericana de Discurso y Sociedad*, 2, 77-99.

---- (2001). Más allá de la *netiquette*: La negociación de la cortesía y del español en las *chats*. *Oralia: Análisis del discurso oral*, 4, 149-178.

---- (2009). Modalidad, evaluación e identidad en el chat. *Discurso & Sociedad*, *3*(4), 738-768.

Palazzo, G. (2005). ¿Son corteses los jóvenes en el chat? Estudio de estrategias de interacción en la conversación virtual. *Revista TEXTOS de la CiberSociedad*, 5. Recuperado de http://www.cibersociedad.net el 12/03/2015

Panckhurst, R. y Moïse, C. (2012). French Text Messages: From SMS Data Collection to Preliminary Analysis. *Lingvisticae Investigationes,* 35(2), 289-317.

Parodi, G. (2008). Lingüística de corpus: una introducción al ámbito. *Revista de Lingüística Teórica y Aplicada*, 46, 93-119.

Payà, M. (2000). Com responem els missatges de correu electrònic? Noves formes de diàleg. En *Actes de la I Jornada sobre Comunicació Mediatitzada per Ordinador en Català*. Barcelona: Universitat de Barcelona.

Pérez Hernández, M. C. (2002). Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento. *Estudios de Lingüística del Español*, 18. Recuperado de http://elies.rediris.es/elies18/index.html el 12/03/2015

Quevedo, L. A. (2012). Los medios de la comunicación en la era de las TICs. *Posgrado Gestión Cultural y Comunicación*. Argentina: FLACSO-Virtual.

Rígano, M. (1998). El léxico de los adolescentes. Rojas Mayer, E. (Ed.), *La Oralidad. Actas del VI Congreso de la Sociedad Argentina de Lingüística*, 2. Tucumán: INSIL.

Sanmartín Sáez, J. (2007). *El Chat: la conversación tecnológica*. Madrid: Arco Libros (Cuadernos de Lengua Española).

Schiffrin, D., Tannen, D. & Hamilton, H. (Eds.). (2001). *The Handbook of Discourse Analysis*. Oxford: Blackwell Publishers.

Sobrero, A. (1993). Costanza e innovazione nelle varietà linguistiche giovanili. En Radtke, E. (Ed.), *La lingua dei giovani* (95-108). Tübingen: Gunter Narr Verlag.

Thurlow, C., Lengel, L. B. & Tomic, A. (2004). *Computer Mediated Communication: Social Interaction and the Internet*. London: SAGE.

Torres-Salinas, D., Robinson-García, N. & Cabezas-Clavijo, Á. (2012). Compartir los datos de investigación en ciencia: introducción al data sharing. *El profesional de la información,* 21(2), 173-184.

Torruella, J. & Llisterri, J. (1999). Diseño de corpus textuales y orales. En Blecua, J. M., Clavería, G., Sánchez, C. & Torruela, J. (Eds.), *Filología e informática: Nuevas tecnologías en los estudios filológicos* (45-77). Barcelona: Departamento de Filología Española, Universidad Autónoma de Barcelona-Editorial Milenio.

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

ISBN 978-987-3617-89-8

Vela Delfa, C. & Jiménez Gómez, J. J. (2011). El sistema de alternancia de turnos en los intercambios sincrónicos mediatizados por ordenador. *Pragmalingüística*, 19, 121-138.

Walther, J. B. (1996). Computer-Mediated Communication: Impersonal, Interpersonal and Hyperpersonal Interaction. *Communication Research,* 23, 3-43. London: SAGE. DOI: 10.1177/009365096023001001

Wilbur, S. P. (1996). An Archaeology of Cyberspaces: Virtuality, Community, Identity. En Porter, D. (Ed.), *Internet Culture* (5-22). New York: Routledge.

Yates, S. J. (1996). Oral and Written Linguistic Aspects of Computer Conferencing. En Herring, S. C. (Ed.), *Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives*. Amsterdam: John Benjamins Publishing (Pragmatics & Beyond New Series, 39, 29-46).

Yus, F. (2001). *Ciberpragmática: El uso del lenguaje en Internet*. Barcelona: Ariel.

---- (2010). *Ciberpragmática 2.0: Nuevos usos del lenguaje en Internet*. Barcelona: Ariel.

Zimmermann, K. (2003). Constitución de la identidad y anticortesía verbal entre jóvenes masculinos hablantes de español. En Bravo, D. (Ed.), *La perspectiva no etnocentrista de la cortesía: Identidad sociocultural de las comunidades hispanohablantes* (47-59). *Actas del Primer Coloquio del Programa EDICE.* Estocolmo: Universidad de Estocolmo. Recuperado de www.edice.org el 12/03/2015

ISBN 978-987-3617-89-8