Problemas de recogida y fijación de muestras del discurso digital

Cristina Vela Delfa°, Lucía Cantamutto*
°Universidad de Valladolid, *Universidad Nacional del Sur - CONICET

One of the challenges faced by the analyst of digital communication consists on the establishment of a corpus that preserve the representativeness and which responds to the nature of these samples of language, in particular as regards two parameters: multimodality and the multi-simultaneity. The article has three objectives. On the one hand, we review the situation of samples of digital speech at international level and specifically for the Spanish. Then, we show a methodological reflection on the problems of collecting and setting of data in the digital speech, from four cornerstones (communicative situation, nature of the data, representativeness of the sample and ethical issues). Finally, we propose guidelines to collect and transcript samples of digital speech for the repository feed; textual or enriched with multimodal data to the repository code.

Keywords: Digital Discourse Analysis, linguistic corpus, methodology, samples of language, Spanish.

1. Introducción

La interacción digital deja disponibles huellas, en forma de datos de naturaleza diversa, accesibles para los investigadores: tanto es así que la gran revolución actual es, precisamente, el salto cuantitativo -y cualitativo- que suponen los Grades Datos (Big Data) y, en parte, las Humanidades Digitales (Berry 2012). Sin embargo, esta afirmación resulta adecuada para determinados campos de estudio, pero niega la realidad de otros, entre ellos, las disciplinas de corte interaccional -como la sociolingüística, la pragmática y el análisis del discurso y de la conversación- que requieren estudios de orden cualitativo a través de caudales de información contextual amplios. En este proceso centrífugo contemporáneo, en el cual los usuarios generan información constantemente, las prácticas comunicativas deben ser observadas con detenimiento a fin de

comprender su complejidad y distinguir los elementos dados por las interfaces de las estrategias comunicativas de los usuarios o diferenciar los procesos de conformación de unidades de aquellos otros aspectos que permiten la co-construcción de la interacción. En otras palabras, en línea con el "computacional turn" (Berry 2012) de los estudios humanísticos, algunas inquietudes no encuentran respuestas acordes en las metodologías actuales. Si el objeto de estudio es el discurso digital, cabe preguntarse: ¿cómo se aborda la dimensión multimodal inherente a la comunicación digital en las diferentes etapas de recolección y fijación de muestras estables de lengua? ¿Es posible fijar aspectos intrínsecamente digitales, por tanto, virtuales, mediados por plataformas e interfaces, dispositivos particulares, que funcionan como vectores de la comunicación digital?

Ante este panorama, el trabajo se plantea tres objetivos. En primer lugar, repasamos de forma breve la situación de las muestras de lengua disponibles para el estudio del discurso digital, tanto en el ámbito internacional como en lo que concierne específicamente a la lengua española (Apartado 2). En segundo lugar, presentamos un acercamiento a los problemas metodológicos de recogida y fijación (es decir, una suerte de transcripción que refleje las características multimodales) de muestras de lengua de interacciones en entornos comunicativos digitales (Apartado 3 y Apartado 4). Todo ello nos lleva, en tercer lugar, a la propuesta de ciertas pautas de recogida y transcripción de muestras de lenguas del discurso digital (Apartado 5) -diseñadas a partir de los antecedentes encontrados (principalmente en inglés, pero también en francés, alemán y chino, véase Beißwenger & Storrer 2008)- con el fin de alimentar un base de datos colaborativa (CoDiCE[1]).

[1] La base de datos CoDiCE (http://codice.aplicacionesonline.com.ar/), desarrollada por Leandro Boisselier, está implementada como una Aplicación Web en lenguaje PHP con el soporte del framework YII. La persistencia de los datos se hace mediante el Gestor de bases relacionales MYSQL. Se encuentra alojado en un servidor con sistema operativo GNU/Linux corriendo el servicio Apache. Para poder llegar a una audiencia con dispositivos heterogéneos se utiliza la codificación HTML5 con soporte de BootStrap y JQuery, permitiendo obtener una interfaz flexible que se adapta a cualquier tamaño de pantalla (tanto de móvil como de computadora). El código está organizado mediante el patrón de diseño Modelo Vista Controlador y la persistencia aprovecha el patrón ActiveRecord. La organización del sistema incluye un sistema de Roles y Permisos para poder asignar granularmente capacidades de uso a los distintos usuarios. Actualmente, se encuentran definidos roles de Usuario y Administrador, pudiendo los usuarios tener acceso solo a su información y los administradores a la totalidad de los registros. En etapas próximas se habilitará la posibilidad de compartir datos entre investigadores particulares, a fin de favorecer incipientes trabajos de variación intralingüística. En esta primera etapa, los mecanismos de gestión de los registros

2. Estado de la cuestión: los corpus en el análisis del discurso digital

En las últimas décadas, desde distintas disciplinas, se ha consolidado el interés por el estudio de las comunicaciones digitales: la sociología, la filosofía de la ciencia, la psicología, las ciencias de la información y de la comunicación y la lingüística, entre otras. Esto ha llevado a diversos avances de índole teórica que dan respuesta a la conceptualización de la comunicación digital (entre ellas, las clásicas de Crystal 2001; Jewitt & Kress 2003; Herring 1996; Herring 2004). Sin embargo, las experiencias demuestran que el proceso de recogida y elaboración de datos para la conformación de corpus lingüísticos no avanza de forma tan acelerada. Cuando se intenta atender a la magnitud integral de la comunicación digital, resulta imprescindible disponer de corpus consistentes, que reparen en ciertas cuestiones metodológicas. Este aspecto, tan allanado en las investigaciones relativas a la interacción cara a cara, no ha logrado la madurez requerida en la comunicación digital (véase *Infra*): dificultades metodológicas repercuten en la falta de representatividad de las muestras manejadas en muchos estudios, así como en cuestiones de ética en el acceso y utilización de estos datos, aspectos que han sido problematizados, entre otros, por Herring (1996), Hine (2000), Buchanan (2004), Crystal (2005) y De-Matteis (In press).

La revisión de bibliografía y de proyectos en marcha da cuenta, sin embargo, de cierto avance en la constitución de corpus de comunicación digital en lenguas como el inglés, el francés y el chino. Muestras de ello son, por ejemplo, los trabajos realizados en torno a la recolección masiva de muestras de comunicación por teléfono móvil (el proyecto Sud4science[2] para recolectar SMS en francés, NUS SMS Corpus[3] sobre mensajes en chino e inglés) o respecto al

están diseñados para facilitar la tarea del investigador proveyendo utilidades de validación, auto completado y definición dinámica de etiquetas. En etapas posteriores se unificarán las etiquetas en función de los fenómenos mayormente estudiados. Se propone una homologación de fenómenos a partir de hipónimos o términos amplios (por ejemplo, "fórmulas_de_tratamiento" (vs. "formas_nominales_de_tratamiento"; "formas_de_saludo" (vs. "saludo_de_bienvenida", "saludo_de_despedida"), entre otros. Las herramientas de análisis actualmente permiten realizar una exploración de los registros que puede ser regulada mediante filtros que afinan la búsqueda hasta recuperar los elementos sobre los que se quiere operar.

Una vez recuperados se pueden listar y exportar a distintos formatos para un análisis con herramientas externas o resguardo personal de entradas. Las herramientas implementadas en la actualidad son de Listado, Cuenta de Palabras y Acumulador de Etiquetas.

[2] Véase http://sud4science.org/ (consulta: febrero de 2014).

[3] Véase https://www.comp.nus.edu.sg/entrepreneurship/Corpus_kanmy.html (consulta: noviembre de 2014).

estudio del correo electrónico (ApacheSpamAssassin[4], que reúne correos electrónicos no deseados en inglés).

Además, desde el punto de vista teórico existen algunas obras que abordan los procesos de recogida de corpus del discurso digital, aunque se trata, en realidad, de publicaciones aisladas: por ejemplo, un gran aporte lo constituyen los capítulos recopilados en *Corpus Linguistics and the Web* (Hundt, Nesselhaulf & Biewer 2007), otros trabajos exhaustivos aportan información sobre cómo construir un corpus de comunicación mediada por computadora (por ejemplo, King 2009) y, más específicamente sobre tipos discursivos particulares, la tesis doctoral de Tagg (2009) y de Cougnon (2015) sobre la comunicación por SMS.

Fuera de estos macroproyectos no siempre es posible la recogida de muestras de lengua que conserven la representatividad necesaria para legitimar un estudio (Toruella & Llisterri 1999). La respuesta a estas dificultades debiera venir de los corpus lingüísticos generales disponibles para las diferentes lenguas; sin embargo, en estos no se brinda atención suficiente al discurso digital y, menos aún, a sus géneros conversacionales (Vela Delfa & Cantamutto 2015; Cantamutto & Vela Delfa In press). Esta carencia resulta especialmente acuciante en una lengua como el español que, según datos del Instituto Cervantes (Moreno Fernández 2012), es la tercera más usada en los intercambios digitales: en los corpus generales de reciente compilación no se incluyen datos procedentes de interacciones mediatizadas. Así sucede tanto en *CORPES XXI*[5] como en otros corpus del español coloquial. Por ejemplo, en el repositorio *Corpus del Español*[6] se ofrecen 100.000.000 palabras, entre las cuales es posible consultar algunos textos extraídos de portales de internet que, sin embargo, no responden a muestras extraídas de entornos interaccionales del discurso digital.

Esta ausencia puede justificarse desde dos perspectivas complementarias. Por un lado, la comunidad científica no le otorga la importancia que merece el fenómeno de la comunicación digital, lo que se refleja en su infra-representación en los corpus generales. Por otro lado, la dificultad en la recogida y fijación de los datos disuade de su inclusión. En este punto actual de la difusión real del fenómeno, ambas cuestiones requieren de atención para ser revertidas en el corto plazo.

[4] Véase https://spamassassin.apache.org/publiccorpus/ (consulta: noviembre de 2014).
[5] Véase http://www.rae.es/recursos/banco-de-datos/corpes-xxi (consulta: noviembre de 2014).
[6] Véase http://www.corpusdelespanol.org/ (consulta: noviembre de 2014).

3. Las muestras de lengua en el discurso digital: características particulares

Entendemos por *discurso digital* aquel producido en entornos de mediación tecnológica, principalmente a través de internet, que suele desarrollarse en plataformas que permiten los intercambios y que delimitan, de algún modo, las posibilidades y potencialidades comunicativas del usuario. En particular, en este trabajo nos interesa el subgrupo de los géneros del discurso digital dialógico o conversacional, caracterizado por la construcción colaborativa, la alternancia de voces (a través de la sucesión de turnos) y la negociación colectiva del significado.

Como sucede con muchos de los fenómenos contemporáneos emergentes relativos a la eclosión digital, la denominación fluctuante se superpone constantemente. Es así que se suele usar como sinónimos *interacción digital*, *comunicación digital* y *discurso digital* (además de las formas que incorporan la idea de virtual y que ponen el foco en la no presencialidad sin atender al entorno específico de estas nuevas prácticas comunicativas). Si bien reconocemos los puntos de encuentro entre estas denominaciones, optamos por *discurso digital,* para referir al resultado, frente a *comunicación digital* que alude más al proceso. Así, acorde a la propuesta pionera de Susan Herring (1996), el discurso digital debe atender tanto a las interacciones desarrolladas en internet y computadoras (epítomes de la computación digital) como a toda la serie de dispositivos que se han desarrollado y se desarrollan (teléfonos, tabletas, relojes) y que transfieren sus datos a través de internet o redes como GSM.

Por otra parte, el discurso digital cubre cada vez más necesidades sociales, por lo que se concreta en un abanico de géneros discursivos (Swales 1990; Eggins & Martin 2000) muy diversos, entre los cuales destaca el género conversacional digital o *interacción digital* que agrupa los intercambios de carácter dialógico que se acercan, conceptualmente, al tipo de intercambios conversacionales (con toda la dificultad analítica que esto supone).

En este contexto, el analista al discurso digital trabaja con un conjunto de datos que manifiestan una naturaleza muy particular. Dos de estas propiedades resultan especialmente interesantes desde la perspectiva de la elaboración de un corpus: la multimodalidad y la multisimultaneidad. Además, una tercera cuestión resulta problemática desde la perspectiva teórica interaccional: la falta de copresencia.

La *multimodalidad* (Jewitt & Kress 2003) refiere a la propiedad de ciertos discursos de integrar datos multisensoriales, que superan lo que tradicionalmente se ha considerado lenguaje verbal. Con la incorporación de este aspecto, la

naturaleza del discurso es abordada de forma integradora, colocando al lenguaje verbal en relación de interdependencia con otros elementos que habían sido considerados marginales y tildados, por ello, de extralingüísticos. Así, la visión tradicional por la que el lenguaje verbal se situaba como código semiótico hegemónico es sustituida por otra en la que este se analiza en relación al resto de los sistemas de signos con los que interactúa. Pardo[7] (2007: 2) señala que "el amalgamiento de códigos y de recursos verbales y no verbales en los discursos es la forma de expresión natural y compleja mediante la cual los seres humanos construimos significado".

Todo discurso es esencialmente multimodal (Kress & Van Leeuwen 2001) y dicha característica no afecta únicamente al flujo de códigos semióticos, sino que incide en los mecanismos comunicativos de producción y comprensión que intervienen en la construcción de los esquemas de significado (Obando 2012: 881). La integración de los diferentes códigos resulta necesaria en el abordaje de casi la totalidad de los discursos. Razón por la cual los estudios sobre la multimodalidad han crecido vertiginosamente, generado nuevos objetos de estudio; por ejemplo, el código no verbal para la enseñanza de segundas lenguas a través de recursos audiovisuales (Quintero García & Vicente-Rasoamala 2007).

Además, la propia evolución de las interfaces justifica ese giro multimodal. Inicialmente, Herring (1999, 2001) concebía y caracterizaba el discurso digital como *text-only* (es decir, como un fenómeno estrictamente textual) mientras que, en trabajos más recientes (Bourlai & Herring 2014; Herring 2015), opta por una perspectiva multimodal. La primitiva concepción monomodal del discurso digital, cuyo foco central consistía en el hecho de que estos intercambios prescindían de los recursos paralingüísticos propios de la oralidad, se ha transformado en una visión que pone su interés en la creciente multimodalidad del medio y que requiere un abordaje holístico[8]. Desde los juegos estrictamente paratextuales (Vela Delfa 2006) hasta la integración de fotografías, videos, audios, etc., los códigos semióticos incorporados en el discurso digital son cada vez más variados. Como señala Scolari (2009), en este tipo de intercambios confluyen diferentes modos que dan lugar a una nueva textualidad: "las pantallas interactivas integran diferentes sistemas de significación (verbal,

[7] N. Pardo A., *Mediatización, multimodalidad y significado*. Ponencia presentada en PROSUL 2007 y en el *X Congresso Internacional de Humanidades - Palavra e cultura na América Latina: Heranças e desafios*, Brasilia, 17-19 de octubre 2007.

[8] C. Jewitt, *Aprendizaje y comunicación en los escenarios digitales multimodales*. Conferencia plenaria, *IV Foro Internacional de Educación Superior en Entornos Virtuales*, Bernal, 23 y 24 de octubre 2014.

icónico, audiovisual e interactivo) que llevan a la conformación de textualidades sincréticas, donde el significado es algo más que la suma de una serie de contenidos autónomos" (Scolari 2009: 54). Con la integración de aplicaciones, proceso que arranca en el diseño de las herramientas de mensajería instantánea (Vela Delfa & Jiménez Gómez 2005) y culmina en las redes sociales incluyendo microbloging, chat, correo electrónico y otras funcionalidades, la multimodalidad se configura como una propiedad fundamental del discurso digital.

Dado que esta propiedad intrínseca no alcanza a todas las plataformas de igual modo, el grado de multimodalidad constituye, en la actualidad, un criterio para clasificar las aplicaciones. Así Herring (2015) distingue entre las plataformas en las que el intercambio sigue siendo principalmente textual, aunque permitan la multimodalidad (por ejemplo, los SMS), de las Plataformas Multimodales Interactivas (PIM), que permiten a los usuarios integrar el contenido multimodal a través de múltiples canales en un solo sitio web, e incluso dentro de un mismo hilo conversacional y que orientan el intercambio hacia esa combinación[9].

Otra característica de este discurso es la *multisimultaneidad* transversal a dispositivos, plataformas y usuarios. La consecuencia inmediata es el modo en que se desdibujan los límites interaccionales tanto por la confluencia de dispositivos en una interacción como a la identificación de unidades para la segmentación. Según señala Alcántara Plá (2014: 228):

> El concepto de simultaneidad cobra así un significado particular: muchas conversaciones pueden considerarse simultáneas sin que haya coincidencia exacta en las intervenciones. Podemos -y veremos que parece lo habitual- iniciar una conversación en WA (Wasup), iniciar seguidamente otra en el chat de Facebook, retomar después la primera, consultar mientras tanto alguna entrada de un blog y dejar allí un comentario que nos será contestado, etc. Lo interesante, porque nos ayuda a entender diferencias con la comunicación oral, es que la lectura de la entrada del blog puede llevarnos a escribir un

[9] En palabras de Herring (2015: 2): "Sites on which messages are mainly textual (excluding multimedia attachments), such as Wikipedia and Twitter, are not IMPs in their current form. One of the first IMPs was YouTube, which allowed users to comment on a shared video asynchronously, via either text or video. Facebook became an IMP when it added video chat to its suite of textual communication options. Another example is the multiplayer online game World of Warcraft, which for several years incorporated synchronous audio chat (Voice-over-IP) in addition to text chat. The messaging service WhatsApp is arguably an example of an IMP on a mobile device: In addition to text messaging, it enables smart phone users to exchange images, video, and audio media messages in a single "conversation".

> comentario en él para que lo lea su autora, pero también puede incitarnos a realizar una intervención en alguna de nuestras conversaciones paralelas en WA.

En idéntico sentido, Baudrillard (1994: 13) tempranamente anticipaba que el usuario se asemeja a una terminal de múltiples redes de las que parten nodos orientados a interacciones en diversas aplicaciones. Yus (2010) sostiene que estos intercambios se desarrollan de manera progresiva y simultánea en diferentes entornos, identificando al usuario como un nodo de interacciones. Así "más que conectarse a las comunidades virtuales, hoy en día, las personas viven múltiples posibilidades de interacción y socialización en forma de redes personales que forman una interacción en el individuo como nodo" (Yus 2010: 51).

El mantenimiento de interacciones en coocurrencia, es decir, donde los interlocutores se involucran de forma simultánea en varios intercambios paralelos, constituye una condición intrínseca del medio digital, que es posible gracias a la confluencia de varios factores: a) la ausencia de copresencia física, b) la persistencia textual, c) el carácter diferido del intercambio.

La interacción digital presupone que la presencialidad no es una condición obligatoria para su desarrollo. Si bien puede suceder, y como tal es una práctica habitual en los jóvenes (Tjora 2011), lo paradigmático de la comunicación mediada es la ausencia de copresencia física de los interactuantes. En particular, Gobato (2014: 119) expone cómo la mediación ha intercedido para que la coprensencia no sea un requerimiento en la comunicación en las interfaces artefactuales. En tal sentido, el autor diferencia tres momentos históricos en los que la mediación tecnológica modifica la experiencia interaccional cara a cara. Estos son: i) la mediación de la escritura, ii) la de los medios electrónicos de comunicación de masas, iii) la de la comunicación digital o virtual. De este modo, tras los cambios sucesivos, los interactuantes buscan la manera de producir un nuevo orden interaccional. Al mismo tiempo, sumadas a las estrategias de los usuarios, las aplicaciones han incorporado herramientas para transmitir cierta presencialidad y que contribuyen al progresivo aumento de la retroalimentación y que alimentan la "ilusión de instantaneidad" (Gobato 2014), como las marcas de "conectado/desconectado" o las indicaciones de "x está escribiendo", "leído/recibido", marcas que se constituyen, algunas veces, en la intervención que completa el par de adyacencia.

Paradójicamente, el desarrollo de varios intercambios al mismo tiempo no podría ocurrir sin la persistencia textual. Los foros de debate, por ejemplo, suelen almacenar los mensajes. En los chats, la interacción está disponible en la pantalla hasta que llegan nuevas intervenciones que desplazan las más antiguas.

En las herramientas de microblogging las intervenciones se almacenan y se ordenan cronológicamente, lo mismo sucede con la mensajería instantánea y los SMS. Así, aunque existen diferentes niveles de permanencia, hasta el menos persistente de los sistemas de comunicación es más persistente que la comunicación oral (Gobato 2014: 220-221).

El carácter diferido de los intercambios es otra de las condiciones necesarias para la multisimultaneidad. La gestión temporal de los intercambios digitales ofrece un abanico de posibilidades muy amplio en un intervalo cada vez más difuso entre la sincronía y la asincrónica. De este modo, las reacciones a una intervención pueden presentar distintos esquemas: a) la respuesta es inmediata, b) la respuesta se dilata, c) no hay respuesta, d) la respuesta ocurre en otro medio/soporte. Esta diversidad permite a los interlocutores realizar de forma simultánea esquemas y secuencias interaccionales, con un coste cognitivo menor del que supone la ejecución de diferentes tareas al mismo tiempo en la comunicación presencial.

4. Condicionantes, características y respuestas: las pautas de recolección y transcripción

Las características inherentes a la comunicación digital (multimodalidad, multisimultaneidad y no copresencia) deben ser especialmente consideradas durante la constitución de un corpus del discurso digital. Por ello, en el marco de estudios de índole cualitativa, los interrogantes sobre este tipo de datos nos han llevado de forma indirecta a una serie de preguntas que atañen a las diferentes fases de recolección de una muestra y constitución de un corpus. Algunos de ellas se ocupan de dificultades relativas al proceso de recolección de muestras de lengua, por ejemplo: ¿los datos recogidos muestran los fenómenos de interacciones en paralelo?, ¿se puede o se debe rastrear la implicación de los usuarios en los diferentes intercambios?, ¿es posible identificar los desajustes en el ordenamiento lineal del intercambio?, ¿se puede contar con múltiples representaciones visuales del intercambio en los diferentes dispositivos de acceso usados por los distintos interlocutores?, ¿se puede recoger la información de retroalimentación que se va incorporando a la interacción? Otras se refieren, más específicamente, a la manera en que los muestras de lengua se sistematizan y transcriben, por ejemplo, ¿cómo se refleja la multimodalidad en los corpus?, ¿es posible ofrecer pautas para la recogida del tradicionalmente llamado componente paralingüístico en su variedad para el discurso digital?, entre otras cuestiones.

Estos condicionantes, más que restricciones, son características inherentes a la comunicación digital. Por tanto, reconocerlos implica para el investigador encontrar la forma de dar respuesta y resolver las dificultades que impone la etapa de recolección y elaboración de estas bases de datos lingüísticos. La experiencia de los investigadores en comunicación digital, como usuarios de múltiples plataformas, favorece la evaluación de los diferentes procesos que ocurren paralelamente a la interacción entre los usuarios: estos procesos (que tienen su principio y su fin en la decodificación de los bits que generan la comunicación) producen modificaciones en todas las terminales intervinientes.

A continuación, consideramos aspectos propios de la etapa de recogida de datos. Si bien algunos atañen al proceso de recolección y otros refieren principalmente al proceso de transcripción, la naturaleza multimodal es el pivot entre ambas etapas que evidencia el desafío que reviste intentar disociar fenómenos de esta naturaleza.

Así, por ejemplo, en la recogida de datos de mensajería instantánea los elementos multimodales suelen perderse en las acciones de almacenaje o envío que ofrecen las herramientas, con frecuencia utilizadas cuando las muestras son elicitadas por informantes. Por ello, se suele trabajar con datos que se corresponden con archivos de texto plano en los cuales se pierde referencia de aspectos multimediales. Esta situación puede solucionarse a través de un costoso pero efectivo proceso de captura de pantallas que daría cuenta, en parte, de la multimodalidad, pero que afectaría al posible futuro tratamiento automático de los datos y cuya anonimización es muy dificultosa. Además, aunque con las capturas se obtuviera mayor cantidad de información gráfica, aún estarían exentos de registro las notas de audio y los videos que estas plataformas ofrecen como opciones para la comunicación así como todo el material de retroalimentación que suelen ir ofreciendo las plataformas y que, a veces, se constituyen como intervenciones en sí mismas. Este es el tipo de problemas concretos que el analista del discurso digital enfrenta[10].

En términos generales, en cualquier proceso de recogida de datos, resulta necesario atender a los condicionantes derivados del soporte, de las plataformas, de las interfaces y de los dispositivos que median el intercambio, puesto que estos imponen condiciones al mensaje y a las particularidades de los marcos y

[10] Recientemente llevamos adelante una investigación sobre la comunicación grupal a través de la plataforma Whatsapp utilizando dos métodos simultáneos: por un lado, el archivo en texto plano y con capturas de pantalla. Como hemos señalado, los datos multimodales fueron almacenados por separado, conservando la referencia en el archivo de texto plano. En la aplicación de carga de datos de CoDiCE se optó por un casillero donde se pudiera describir los fenómenos multimodales relativos a la muestra de lengua.

rituales de interacción (Gobato 2014). Concretamente, cada entorno comunicativo presenta unas condiciones de enunciación que el investigador debe considerar en su diseño metodológico. La relación de los usuarios con el medio y su familiaridad con los soportes también puede influir en la naturaleza de los datos recogidos y requiere ser atendida como variable. Los soportes condicionan la gestión temporal del intercambio y generan espacios difusos entre la sincronía y la asincronía, con sesgos más o menos marcados hacia uno u otro polo, que han de ser considerados al establecer las fronteras de las unidades de segmentación. Así, volviendo al ejemplo de las conversaciones de mensajería instantánea, los investigadores no siempre pueden establecer con facilidad los límites entre las unidades, puesto que el criterio de la continuidad temporal no opera como en los entornos presenciales. En las formulaciones clásicas de Goffman (1959), la interacción empieza y acaba con el encuentro físico y sincrónico de los interlocutores: ¿qué ocurre en una interacción en la que los intercambios se suceden en una transición en el tiempo sin intervalos de continuidad prefijados? Ante este reto, los investigadores suelen reformular viejas unidades para adaptarlas a las nuevas condiciones. Vela Delfa (2006: 364) proponía el término *historia interaccional* para solventar los obstáculos en el establecimiento de la unidad máxima, que repercute en el análisis y en la recolección de las muestras. Alcántara-Plá (2014: 231) reserva el término conversación para las unidades comprendidas entre secuencias de apertura y cierre explícitas, como él reconoce, muy escasas, y propone el término relación para referirse "al vínculo que unas personas establecen en una aplicación informática concreta de manera que puedan comunicarse (con conversaciones en el caso de los mensajes instantáneos) siempre que lo deseen".

Asimismo, y como un factor a considerar dentro de los ítems contextuales, el *soporte* o la interface artefactual (Gobato 2014) interviniente en la comunicación tiene características propias que se proyectan en las intervenciones de los actantes y, por tanto, en las muestras de lengua con las que se encuentre el investigador. En cada interacción, las características de la plataforma y del dispositivo (entre ellas, el teclado y el tamaño de la pantalla) condicionan el mensaje, tanto desde el punto de vista paratextual, como de las posibilidades específicas de la interacción, por ejemplo, tipo y cantidad de datos que pueden enviarse. Este aspecto, además, encuentra su contracara en el conocimiento que los participantes tienen de las características operativas del soporte que les lleva a optar por una u otra plataforma, en aras de satisfacer las metas comunicativas personales y que se constituye como una variable de investigación.

De este modo, mientras que en la comunicación cara a cara se obtiene registro de manera relativamente sencilla de la participación de ambos interactuantes en el mismo momento de producción de la muestra. En la comunicación digital, en cambio, los intercambios no se establecen en un espacio físico ni temporal común. Estos se producen de forma paralela, a veces simultánea, y no necesariamente en la misma aplicación. Por lo tanto, la representación de los signos lingüísticos, con frecuencia, no es equivalente 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. Esta disimetría puede incidir, entre otras cuestiones, en la ordenación lineal de las intervenciones que resulta del solapamiento temporal de los intercambios: los interactuantes participan de una comunicación que se va configurando, a menudo, con diferencias cronológicas. Llamativamente, participantes de la comunicación son capaces de reponer la intervención como réplica a un par anterior (Herring 1999; Vela Delfa & Jiménez Gómez 2011; Cantamutto 2013). Sin embargo, la mirada del investigador, que la mayoría de las veces accede a la comunicación a través de lo producido y recibido por uno de los interactuantes, no siempre detecta los matices que se producen indefectiblemente en una comunicación *casi* sincrónica.

Estas cuestiones refractan, además, en las categorías que se utilizan en el análisis. El desplazamiento de las perspectivas más lingüísticas hacia los lineamientos más semióticos (Jewitt & Kress 2003) en la comprensión de los prácticas comunicativas en entornos digitales ha encontrado fricciones con la utilización de las categorías teóricas propuestas para el estudio de la conversación cara a cara. En tal sentido, las unidades prototípicas del Análisis de la Conversación (Briz 2003) deben ser repensadas en función de los problemas de segmentación que se encuentran en la comunicación digital. Muestra de ello han sido algunos trabajos que presentan críticamente los problemas persistentes en utilizar la interacción cara a cara como paradigmática. Por ejemplo, para la mensajería instantánea o chat, Vela Delfa & Jiménez Gómez (2011) señalan el constante solapamiento de intervenciones, propiedad estructurante de este tipo de interacción, que genera, entre otras cuestiones, "una estructura de encadenamiento de pares de adyacencia que no responde a un modelo lineal externo" (Vela Delfa & Jiménez Gómez 2011: 134). Algo similar se replica para el análisis de las unidades conversacionales en el Whatsapp propuesta por Alcántara-Plá (2014). En definitiva, si la cuestión es cómo delimitar las unidades dialógicas (aquello que constituye un intercambio) se requiere comprender de manera amplia el contexto de enunciación (Cantamutto 2013);

no solo porque el usuario expande sus canales de comunicación, sino porque la interacción ya no ocurre necesariamente de un modo lineal.

A fin de dar respuesta a algunos de estos aspectos, en el proceso de recogida de estas muestras de lengua es oportuno solicitar a los colaboradores que brinden parte de la información pragmática compartida, sobre la cual se sustenta la construcción del significado. Además, es esperable, cuando sea posible, transcribir la interacción desde los diferentes dispositivos intervinientes en la comunicación, para confrontar cómo recibe la intervención cada uno de los hablantes. Esta cuestión afecta particularmente a la tipografía de los signos lingüísticos y al establecimiento de los turnos y pares de adyacencia, tal como señalamos. Por ello, es conveniente que los colaboradores brinden la información contextual necesaria para la comprensión del intercambio. Más allá de la información pragmática compartida, que debe ser provista por el hablante, parte del contexto comunicativo se puede reponer a partir de los datos que la plataforma da (por ejemplo, hora y día, una suerte de aquí y ahora de la interacción).

Otros condicionantes se nuclean en torno a la representatividad de los datos y sus condiciones éticas. Uno de los problemas radica en atender a todas las variables sociolingüísticas, principalmente, en torno a los niveles socioeducativos más bajos (Cantamutto 2014). Con frecuencia, en la discusión metodológica sobre las muestras de lengua de comunicación digital, las diferentes restricciones que el investigador enfrenta lo colocan en un debate interminable. La dificultad en el acceso a los datos lleva a la mayoría de los investigadores a trabajar con redes sociales de familiares y amigos (por ejemplo, Tagg 2009) o con técnicas de introspección u observación participante, a través del estudio de interacciones que el investigador lleva adelante. De este modo, la ausencia de otros perfiles de interactuantes (que no sean de nivel educativo alto o universitario) puede conducir a observaciones sesgadas, ya que en la variación intragrupal radica la mayor riqueza de elementos para han de ser considerados.

Por último, es fundamental atender a la privacidad y anonimización de las muestras. Una cuestión transversal de los datos emergentes de la comunicación digital deriva en la dificultad para establecer límites entre su carácter público o privado. El discurso digital ofrece una multiplicidad de datos que parecieran estar a libre disposición para el investigador debido a que, por una condición impuesta por las plataformas, las producciones ahí emitidas son públicas y tienen gran permanencia (en tal sentido, es interesante la actual discusión en torno al "derecho al olvido"). Esta situación no tiene una contrapartida en los intercambios presenciales orales cuya permanencia es nula salvo que algún dispositivo esté procediendo su fijación (tal el caso, por ejemplo, del grabador

de un investigador). Y, en tales circunstancias, el registro será, generalmente, de una única dimensión de la interacción: es decir, la voz. Estas cuestiones hacen que sea relativamente imposible identificar a los interlocutores a partir de las muestras de lengua extraídas por el investigador de la interacción cara a cara. De este modo, no se suele problematizar demasiado el tratamiento como intercambios públicos a aquellos que ocurren en espacios considerados, justamente, públicos, como sucede con la interacción comercial o los intercambios en ámbitos institucionales.

Por el contrario, la problemática de la oposición entre lo público y lo privado tiene una larga tradición en los estudios sobre el dicurso digital y los posicionamientos de los distintos investigadores pueden ser muy diferentes. Tal y como sintetiza Estalella[11], algunos como Herring (1996b) consideran que algo que se encuentra en un espacio público manifiesta carácter público aunque las expectativas de privacidad que suponen los participantes no siempre coinciden con la visión de los investigadores (Walther 2002). Otros investigadores proponen criterios más concretos –relacionados con la arquitectura tecnológica- que aluden a factores como presencia/ausencia de contraseña para el acceso a la información, políticas de resguardo, o sensibilidad del tópico que se trata. Un tercer aspecto a considerar son las cláusulas legales que las páginas, redes sociales y dispositivos tengan respecto a los contenidos que ahí se producen.

En tal sentido, De-Matteis (In press) propone atender a dos tipos diferentes de datos: los disponibles y los elicitados. Los primeros, derivados de la permanencia textual de los intercambios multimodales, parecen estar a disposición de cualquiera, aspecto que requiere atención. Los segundos, aquellos que el investigador solicita o conforma, deben ser examinados a fin de evitar que los emisores pudieran ser reconocidos. Un aspecto que los grandes datos buscan matizar a través de la homologación o hibridación de las emisiones a partir de la creación, por ejemplo, de una suerte de puzzle con fragmentos más o menos similares para evitar la posible identificación de los emisores.

Esta trama de condicionantes obliga a los investigadores de la comunicación digital a una justificación constante a causa de las insuficiencias metodológicas que los datos puedan tener y, por tanto, las conclusiones. Ante ello, la propuesta de elaboración de un repositorio que reúna muestras de lengua perfectibles permite, gracias a la confluencia de datos fragmentarios, subsanar estas carencias referidas. Al mismo tiempo, la confrontación con los datos que los investigadores puedan recolectar en sus propias interacciones enriquece el

[11] *Dilemas morales y desafíos empíricos*. In A. Estalella (*sine data*), *Etnografías de lo digital. Una monografía metodológica*.

debate desde una perspectiva variacionista y sociopragmática. El cúmulo de datos provenientes de diferentes interactuantes (colaboradores expertos y colaboradores voluntarios) matiza aspectos etnocentristas que puedan derivar de una investigación introspectiva.

5. La confluencia de investigadores en CoDiCE: datos y etiquetas en el proceso de transcripción

Como adelantamos en el apartado anterior, la enorme dificultad reconocida por los analistas del discurso en la recolección individual de datos representativos y fiables conduce a que muchos de los trabajos publicados llegan a conclusiones que se asumen "precarias", "exploratorias", "no representativas". A través de CoDiCE se busca subsanar dicha limitación gracias a la confluencia de datos primarios de investigaciones particulares, recogidos mediante un protocolo uniforme, que fue avanzado en Vela Delfa y Cantamutto (2015).

CoDiCE no se corresponde con un intento ambicioso de resolver todos los problemas que los investigadores enfrentan en la elaboración de un corpus de discurso digital pero busca asumir y paliar los problemas que se desprenden de su particular naturaleza. Este cambio nos coloca en una nueva postura: si los datos recogidos son, *per se*, fragmentarios, poco sistemáticos y no representativos, una gran cantidad de ellos logrará minimizar muchos de estos conflictos metodológicos que empobrecen las posibilidades de análisis de los investigadores. Además, responde a un requisito actual del campo de las Humanidades Digitales: la disposición de datos primarios y la incorporación de nuevas técnicas para investigar y analizar los fenómenos contemporáneos (Manovich 2012).

Por todo ello, en la base de datos CoDiCE se encuentran dos perfiles de muestras:

a. Las resultantes de recogidas masivas de datos a través colaboradores voluntarios. Estas serán probablemente pobres en recursos multimodales y datos sociolingüísticos, pero tendrán la ventaja de la representatividad.
b. Las recogidas en procesos de introspecciones sistematizadas y *fijadas* a partir de las prácticas comunicativas de los propios investigadores a través de métodos de observación participante. Serán más ricas en datos contextuales y más pobres en representatividad.

Para satisfacer ambos perfiles, proponemos unas pautas de sistematización y trascripción de las muestras de lengua con una serie de categorías no obligatorias. Estas pautas son el resultado de la reflexión metodológica en torno a la recogida de los datos, que referíamos en el apartado anterior, pero se orienta especialmente hacia las dificultades que radican en el modo en que este tipo de muestras son copiadas, almacenadas y resguardadas (Cantamutto & Vela Delfa In press). En tal sentido, hemos discriminado cuatro núcleos principales a la hora de desarrollar la estiquetas específicas de la trascripción: a) la situación comunicativa, b) la naturaleza de los datos, c) la representatividad y d) las cuestiones éticas.

5.1 Situación de comunicación

En las pautas de transcripción se consigna un espacio para indicar desde qué dispositivo y en qué plataforma se produce la interacción, especificando -siempre que se pueda- la situación de cada uno de los interactuantes. Cuando sea posible, y siempre que se trate de dispositivos móviles (tablets y teléfonos), se describe el tipo de teclado y sistema de escritura. Además de indicarse la aplicación empleada (por ejemplo, Facebook o la plataforma de Windows 8 para recibir mensajes de las redes sociales), se solicita una descripción de la situación de enunciación lo más completa posible ya que una misma herramienta puede albergar interacciones con propiedades enunciativas diversas (Vela Delfa & Jiménez Gómez 2005).

Una descripción icónica de la aplicación permite contar con datos paratextuales. Esta información puede remitirse en archivos adjuntos, a partir de capturas de pantallas. No obstante, si se quiere hacer pública incluyéndola en el repositorio, es necesario aplicar los mismos criterios de anonimización que en el texto escrito, a través de zonas difuminadas en la imagen, por ejemplo. En función de registrar el proceso de producción y la información de retroalimentación que algunas aplicaciones ofrecen (por ejemplo, la forma "leído" en WhatsApp) se propone alternativas como las capturas de pantalla o videos.

5.2 Naturaleza de los datos

Atendiendo a la nueva textualización detectada en los intercambios digitales (Scolari 2009), y siguiendo los lineamientos propuestos por Herring (2015), se trabaja siempre con dos niveles de datos: tal como señalamos, el texto limpio o

plano (con información sobre la contextualización básica[12]) y el nivel enriquecido con otros archivos complementarios, tantos como se hayan podido recoger, con videos, audios, capturas de pantalla y otras formas de fijar las manifestaciones multimediales.

Una vez más, emergen las ventajas de la versatilidad de una base de datos colaborativa puesto que a partir de la disposición de muestras recogidas por investigadores en observación participante ricas en información contextual, se pueden abordar fenómenos de variación pragmática inter e intra lingüística, que pasan desapercibidos en investigaciones que se ven obligadas a prescindir de la multimodalidad. Al tiempo que, mediante muestras amplias ofrecidas por colaboradores de diversa índole, la representatividad estará asegurada. De este modo, se prioriza una estrategia en la que, a partir de datos fragmentarios, se recompone el rompecabezas, relativamente completo, de la comunicación digital.

5.3 Representatividad de la muestra

Tal como señalamos, a fin de asegurar la representatividad, la información contextual que acompañan a la trascripción (datos sociolingüísticos de los participantes, de los dispositivos y plataformas intervinientes, de la situación comunicativa) permite el cruce de muestras parciales o fragmentarias. A través de etiquetas que recuperan aspectos como el nivel de formación, el grado de familiaridad con los medios digitales, la frecuencia de uso del medio digital, en general, y de la aplicación, en particular, la edad, el sexo y otras variables relevantes -relación entre los interlocutores (distancia social, poder relativo) o las referencias a la situación de comunicación (registro, tono)- podrán cruzarse las diferentes variables para su estudio sociopragmático.

5.4 Cuestiones éticas

Por último, para satisfacer las necesidades éticas y legales de cualquier proceso de recogida de datos, las pautas sostienen una serie de categorías que invitan a la reflexión sobre el origen de los datos. Así, debe indicarse si la muestra está conformada por datos públicos o privados y aquellos pertenecientes a las plataformas en las cuales se desarrolla el interacción (i.e. Twitter). En el caso de

[12] En esta etapa del proyecto estamos utilizando etiquetas propias, procesables por nuestras herramientas de análisis, que en un futuro podrán converger hacia lenguajes de etiquetado estandarizados como TEI, ISLE o similares y para el tratamiento automático con herramientas de análisis de corpus.

muestras elicitadas por colaboradores, estos deben estar acompañados de los correspondientes consentimientos informados firmados y se recuerda a los colaboradores que el repositorio requiere la anonimización de los interlocutores (con diferentes recursos como el cambio de nombres, el uso de iniciales, entre otros). Las muestras deben enmascarar los detalles que facilitan la identificación personal de los informantes.

A continuación ofrecemos un ejemplo ilustrativo sobre dos muestras de lengua tomadas en la plataforma de mensajería instantánea Whatsapp. En el ejemplo (1), se puede ver el texto plano recogido a través del envío del historial de conversación (metadatos en el Cuadro 1), en algunas intervenciones los emoticones multimodales se han transformado en cuadrados vacíos (por ejemplo, en la intervención de Dorne a las 22.18). En el ejemplo (2), se observa el texto plano junto con algunas capturas de pantalla (Figura 1) y la información que indica las imágenes y audios que conforman la interacción (última fila del Cuadro 2; presentado en Vela Delfa & Cantamutto 2015).

Cuadro 1. Metadatos de interacción por WhatsApp elicitado por un colaborador, ejemplo (1)

Variedad del español principal	Español de España[13]
Otras variedades presentes	Desconocido
Plataforma	Whatsapp
Dispositivo de recolección	Desconocido
Tipo de teclado	Desconocido
Participantes	[número] 4 participantes: [simetría/jerarquía] Roles simétricos entre todos los participantes, menos uno que ostenta el papel de administrador del grupo. [grado de involucración] 4 participantes activos
Metodología	Elicitación por colaborador enviando el historial por correo electrónico
Ámbito	Personal
Consentimiento informado	Un participante. El resto no está en conocimiento de que la muestra fue tomada.

[13] Siguiendo las clasificaciones propuestas por el CREA y el CORPES.

Descripción situación comunicativa	Los intercolucutores se conocen personalmente. Se ven con regularidad. Su vinculación se debe a que son compañeros de la universidad. [Ends] Los temas tratados en el grupo es la organización de un viaje a Barcelona. [Key-registro] Coloquial/informal [Temporalidad] participación diaria, marcada con fecha y hora en el archivo de texto.
TEXTO PLANO	véase ejemplo (1)

(1) **Texto plano**

29/02/2012, 19:47 - Dorle: Edurne M ha cambiado el asunto a Barcelona"
29/02/2012, 19:50 - Dorle: Edurne M se ha unido
29/02/2012, 19:50 - Dorle: Jorge Lappo se ha unido
29/02/2012, 19:50 - Dorle: Sandra se ha unido
29/02/2012, 19:50 - Dorle: No hay prisa
29/02/2012, 19:50 - Dorle: irms x tren
29/02/2012, 19:50 - Dorle: m tnn cntstar ls d cmun
29/02/2012, 19:51 - Edurne M: Lo e creado para hablar los cuatro mas trankis
29/02/2012, 19:51 - Edurne M: Kex les as preguntau pues?
29/02/2012, 20:16 - Jorge Lappo: Xke x tren??
29/02/2012, 21:31 - Dorle: Xq avion sale bastant caro
29/02/2012, 21:31 - Dorle: y cn trn ns hacn dscuento
29/02/2012, 21:31 - Dorle: ls he pedido la tarjeta
29/02/2012, 21:39 - Jorge Lappo: Ok bueni..
29/02/2012, 21:46 - Edurne M: :)
29/02/2012, 21:51 - Jorge Lappo: Y saldra parecido??
29/02/2012, 22:13 - Edurne M: Sandra a muertoo
29/02/2012, 22:13 - Edurne M: Si kreo k si jorge
29/02/2012, 22:17 - Dorle: Cn l dscuento sale mas barato, y n tren llevas tdo lo q qieras...
29/02/2012, 22:17 - Jorge Lappo: Y ay para familia numerosa??
29/02/2012, 22:18 - Dorle: Sii
29/02/2012, 22:18 - Dorle: □

Cuadro 2. Muestra recolectada a través de las pautas de CoDiCE, ejemplo (2)

Variedad del español principal	Español peninsular
Plataforma	Whatsapp
Dispositivo de recolección	Teléfono Android
Tipo de teclado	Qwerty + predictivo
Participantes	[número]13 participantes: [simetría/jerarquía] Roles simétricos entre todos los participantes, menos uno que ostenta el papel de administrador del grupo. [grado de involucración] 11 participantes activos y 2 espectadores
Metodología	Observación participante: una de las investigadoras participa en el grupo. Muestra elicitada desde el dispositivo de una de las investigadoras.
Ámbito	Personal
Consentimiento informado	Todos los participantes
Descripción situación comunicativa	Los interlocutores se conocen personalmente. Se ven con regularidad. Su vinculación se debe a que sus hijos asisten a la misma clase del mismo colegio. [Ends] Los temas tratados en el grupo son principalmente cuestiones relativas a la vida cotidiana de los niños y a temas relativos a la crianza. [Key-registro] Coloquial/informal [Temporalidad] participación diaria, marcada con fecha y hora en el archivo de texto.
TEXTO PLANO	véase ejemplo (2)
Archivos adjuntos	17 archivos de voz, 48 imágenes

(2) **Texto plano**

Total de la muestra: 2611 intervenciones de texto
[fragmento]
28 may 13:56 - Informante1: <Archivo omitido>
28 may 14:53 - Informante5: Osotros iremos el sábado por la tarde....
28 may 15:01 - Informante1: Yo no se si sábado o domingo aun. Pero por la tarde a eso de las 6.

28 may 15:02 – Informante1: A animación deportiva
28 may 15:07 – Informante6: Nosotros tb. Sábado x la tarde, desde las 17h o asÃ¬...
28 may 15:28 - Informante2: Pues nosotros no sabemos aún, porque el peque lleva pachuchillo desde ayer y encima hoy le han puesto la vacun y le ha subido fiebre.
28 may 15:29 - Informante2: Bueno, esperemos que se le pase y podamos ir un ratillo
28 may 15:29 - Informante6: Ay pobrete!!! Mucho ánimo!!
28 may 15:30 - Informante2: Gracias! □
28 may 20:17 - Informante5: Que se mejore el campeón!
28 may 20:18 - Informante1: Sí que se mejore!
28 may 20:21 - Informante2: Gracias!! □
28 may 20:28 - Informante5: Qué tal la vacuna a nivel de impacto emotivo?...en la última eran bebés
28 may 20:33 - Informante2: Pues este pobre lo ha llevado un poco mal. Ha visto el frasquito y se ha puesto a temblar como una hoja. Daba una penita... Pero bueno, ha sido rápido, porque no me acuerdo cuál de las otras eran tres pinchazos y esta es solo uno. La siguiente a los 6 añs
28 may 20:34 - Informante5: Grande!
[/comienza texto de captura]
28 may 20:34 – Informante1: Si ahora lo pasan peor yo creo
28 may 20:37 – Informante2: Yo creo que sí. Pobres! Además como dicen que no se muevan, que les sujetemos, etc., pues yo creo que se asustan más
28 may 20:37 - Informante1: Ya
28 may 20:56 – Informante3: Que se mejore Lito
28 may 20:59 - Informante2: Gracias! □
[/fin texto de captura]
31 may 10:15 – Informante4: <Archivo imagen>
31 may 10:51 – Informante5: Planazo! Lo consulto con el peque...
[/fragmento]

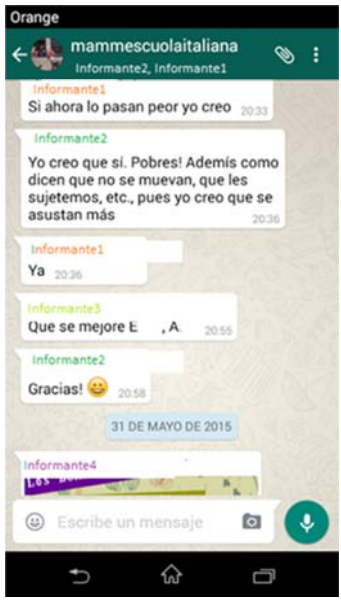

Figura 1. Captura de pantalla (ejemplo 2)

6. Palabras finales

A lo largo del artículo, hemos tratado de establecer ciertos puntos de partida para guiar una reflexión que sabemos difícil y extensa. La ausencia de mayor bibliografía idónea y proyectos activos para la lengua española denota la dificultad propia del estudio de estas formas de comunicación pero, al mismo tiempo, nos instala frente a la necesidad de comprender el fenómeno y abordarlo a través de herramientas de uso compartido entre los investigadores del campo disciplinar. La opción de un repositorio colaborativo, frente al proyecto de recogida de un corpus completo, resulta una solución, precaria quizás, pero que aúna esfuerzos de muchos profesionales. Al tiempo que, como hemos señalados en las páginas precedentes, permite poner en relación muestras con carencias parciales, que al unirse se complementan. Es decir, los dos niveles de muestras: el texto plano y las muestras ricas.

Este artículo ha abordado aspectos del discurso digital que se constituyen como desafíos para el investigador durante la recogida de muestras de lengua, a saber: a) los soportes, plataformas, interfaces y dispositivos, b) la situación de comunicación, c) la multimodalidad, d) la representatividad de los datos y e) las

condiciones éticas de las muestras. A partir de esta reflexión metodológica se han propuesto unas pautas para proceder a la alimentación del repositorio a través de una plantilla lo suficientemente amplia para que satisfaga, al menos, la mayoría de las características de las comunicaciones digitales actuales (véase Cuadro 2). Es probable que, a medida que se desarrollen nuevas plataformas y se actualicen las existentes, sea necesario acondicionar tanto las pautas como la plantilla.

La hipótesis de partida propuso que las dificultades de recogida y fijación de estas muestras de lengua han justificado, con frecuencia, las limitaciones de los corpus utilizados en diversas investigaciones. En este sentido, estamos seguros de que la creación de un repositorio colaborativo en el que se compartan estos datos, permite la suma de muestras fragmentarias con las que lograr un corpus suficientemente representativo. Este artículo delineó los preliminares de un proyecto de creación de un repositorio de interacciones digitales denominado CoDiCE (Comunicación Digital: Corpus del Español). En la medida en que se aumenten la cantidad de datos disponibles en esta base de datos, al menos dos líneas futuras de trabajo serán favorecidas: por un lado, el avance de las investigaciones sobre variación pragmática y sociolingüística inter e intralingüística, de gran repercusión en los estudios lingüísticos de los últimos años, en la comunicación digital. Por otro, en futuras etapas del proyecto, se podrán abordar desde una perspectiva diacrónica estas muestras de lengua, imperativo para la comprensión de fenómenos comunicativos con breve permanencia en el tiempo.

Referencias

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

Baudrillard, J. 1994. *El otro por sí mismo*. Buenos Aires: Anagrama.

Beißwenger, M. & Storrer, A. 2008. Corpora of Computer-Mediated Communication. In A. Lüdeling & M. Kytö (eds), *Corpus Linguistics. An International Handbook.*. Berlin: Mouton de Gruyter.

Berry, D. 2012. *Understanding Digital Humanities*. New York: Palgrave/Macmillan.

Bourlai, E. & Herring, S.C. 2014. Multimodal communication on Tumblr: "I have so many feels!". In *WebSci '14: Proceedings of the 2014 ACM Conference on Web Science*. New York: ACM. http://ella.slis.indiana.edu/~herring/tumblr.pdf (accessed December 10, 2015)

Briz, A. 2003. Un sistema de unidades para el estudio del lenguaje coloquial. *Oralia* 6: 7-61.

Buchanan, E. (ed.) 2004. *Readings in virtual research ethics. Issues and controversies*. London: Information Science Publishing.

Cantamutto, L. 2013. La recursividad de las interacciones contemporáneas. Límites teórico-metodológicos del estudio de los SMS como conversación. *Revista de Ciencias Sociales de la Universidad Nacional de Quilmes "Al abordaje de la comunicación contemporánea. Sociedad y cultura en los mundos de la mediación digital"* 23: 83-104.

Cantamutto, L. 2014. El discurso de los mensajes de texto en el habla adolescente del español bonaerense. In A. Parini & M. Giammatteo (eds), *Lenguaje, discurso e interacción en los espacios virtuales*. Mendoza: FFyL-UNCuyo-SA, 65-82.

Cantamutto, L. & Vela Delfa, C. (In press). Repositorio abierto de comunicaciones digitales: hacia la construcción de un corpus para el español. In *Las Humanidades Digitales desde Argentina: Tecnologías, Culturas, Saberes*. Buenos Aires: AAHD FyL-UBA.

Cougnon, L. 2015. *Langage et SMS*. Louvain: CIACO.

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

Crystal, D. 2005. The scope of Internet linguistics. Paper given online to the *American Association for the Advancement of Science* for the Annual Meeting, February 18, 2005. http://www.davidcrystal.com/?fileid=-4113 (accessed December 10, 2015).

De-Matteis, L. In press. Ejes para una discusión del uso ético de datos interaccionales escritos y orales obtenidos en línea, *Las Humanidades Digitales desde Argentina: Tecnologías, Culturas, Saberes*. Buenos Aires: AAHD-FyL-UBA.

Eggins, S. & Martin, J.R. 2000. Géneros y registros del discurso. In T. Van Dijk (ed.), *El discurso como estructura y proceso (Estudios sobre el discurso I)*. Barcelona: Gedisa.

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

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

Herring, S.C. (ed.) 1996. *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives*. Amsterdam: John Benjamins Publishing.

Herring, S.C. 1996b. Linguistic and critical research on computer-mediated communication: Some ethical and scholarly considerations. *The Information Society* 12(2): 153-168.

Herring, S.C. 1999. Interactional Coherence in CMC. *Journal of Computer-Mediated Communication* 4(4).

Herring, S.C. 2001. Computer-mediated Discourse. In D. Schiffrin, D. Tannen & H. Hamilton (eds), *The Handbook of Discourse Analysis.* Oxford: Blackwell, 612-634.

Herring, S.C. 2004. Computer-mediated discourse analysis: An approach to researching online behavior. In S.A. Barab, R. Kling & J.H. Gray (eds), *Designing for virtual communities in the service of learning.* New York: Cambridge University Press, 338-376.

Herring, S.C. 2015. New frontiers in interactive multimodal communication. In A. Georgapoulou & T. Spilloti (eds), *The Routledge handbook of language and digital communication*. London: Routledge, 398-402.

Hine, C. 2000. *Virtual ethnography*. London/Thousand Oaks/New Delhi: Sage Publications.

Hundt, M., Nesselhauf, N. & Biewer, C. (eds) 2007. *Corpus Linguistics and the Web*. Amsterdam: Rodopi.

Jewitt, C. & Kress, G. (eds) 2003. *Multimodal literacy*. New York: Peter Lang.

King, B. 2009. Building and Analysing Corpora of Computer-Mediated Communication. In P. Baker (ed.), *Contemporary Corpus Linguistics*. London: Continuum International Publishing Group, 303-322.

Kress, G. & Van Leeuwen, T. 2001. *Multimodal discourse: the modes and media of contemporary communication.* London: Edward Arnold.

Manovich, L. 2012. Trending the promises and the challenges of Big Social Data. In M. Gold (ed.), *Debates in the Digital Humanities*. Minnesota: University of Minnesota Press, 460-475.

Moreno Fernández, F. (ed.) 2012. *El español, una lengua viva. Informe 2012*. España: Instituto Cervantes.

Obando, L. 2012. Semiótica cognitiva y multimodalidad en la interacción pedagógica. In P.C. Cantero, G.E. Veloso, A. Passeri & J.M. Paz Gago (eds), *Proceedings of the 10th World Congress of the International Association for Semiotic Studies (IASS/AIS)*. A Coruña: Universidade da Coruña, 879-886.

Quintero García, D. & Vicente-Rasoamalala, L. 2007. Teaching Spanish nonverbal communication through soap opera. In K. Bradford-Watts (ed.), *JALT 2006 Conference Proceedings*. Tokyo: JALT.

Scolari, C. 2009. Alrededor de la(s) convergencia(s). Conversaciones teóricas, divergencias conceptuales y transformaciones en el ecosistema de medios. *Signo y Pensamiento* 28(54): 44-55.

Swales, J. 1990. *Genre analysis: English in academic and research settings*. New York: Cambridge University Press.

Tagg, C. 2009. *A Corpus Linguistics Study of SMS Text Messaging*. Birmingham: University of Birmingham.

Torruella, J. & Llisterri, J. 1999. Diseño de corpus textuales y orales. In J.M. Blecua, G. Claveria, C. Sanchez & J. Torruella (eds), *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Barcelona: Editorial Milenio, 45-77.

Tjora, A. 2011. Invisible whispers: account of SMS communication in shared physical space. *Convergence: the International Journal of Research into New Media Technologies* 17(4): 193-211.

Vela Delfa, C. 2006. *El correo electrónico el nacimiento de un nuevo género*. Madrid: Universidad Complutense de Madrid.

Vela Delfa, C. & Cantamutto, L. 2015. Methodological Approach to the desing of Digital Discourse Corpora in Spanish. *Procedia - Social and Behavioral Sciences* 198: 494-499.

Vela Delfa, C. & Jiménez Gómez, J. 2005. La transformación de la experiencia virtual a través de la mensajería instantánea, *Actas del II Congreso del Observatorio para la cibersociedad*. Madrid, Observatorio para la Cibersociedad. http://www.cibersociedad.net/congres2004/grups/fitxacom_publica2.php?id=407 (accessed December 10, 2015).

Vela Delfa, C. & Jiménez Gómez, J. 2011. El sistema de alternancia de turnos en los intercambios sincrónicos mediatizados por ordenador. *Pragmalingüística* 0(19): 121-138.

Walther, J.B. 2002. Research Ethics in Internet-Enabled Research: Human Subjects Issues and Methodological Myopia. *Ethics and Information Technology* 4(3): 205-216.

Yus, F. 2010. *Ciberpragmática 2.0. Nuevos usos del lenguaje en Internet*. Barcelona: Ariel.