CUADERNOS.INFO Nº 54
Versión electrónica: ISSN 0719-367x
http://www.cuadernos.info
https://doi.org/10.7764/cdi.54.53165

Recibido 26-08-2022 / Aceptado : 22-12-2022

Corpus de interacciones digitales: sistematización de técnicas para recoger datos en WhatsApp

Corpus of digital interactions: systematization of techniques to collect data on WhatsApp

Corpus de interações digitais: sistematização de técnicas para coleta de dados no WhatsApp

Lucía Cantamutto, CIEDIS-Universidad Nacional de Río Negro/ CONICET, Viedma, (Argentina) (lcantamutto@unrn.edu.ar)

Cristina Vela Delfa, Universidad de Valladolid, Segovia (España), (cristina.vela@uva.es)

RESUMEN | La recolección de conjuntos de datos de interacciones reales es un paso ineludible en muchas investigaciones que buscan comprender los usos lingüísticos. En el campo del análisis del discurso digital, esto resulta complejo tanto por las características cambiantes de las aplicaciones como por las decisiones éticas que suponen. Este artículo tiene un doble objetivo. En primer lugar, ofrecer un estado de la cuestión sobre los conjuntos de datos de intercambios digitales por WhatsApp y, en segundo lugar, sistematizar diferentes técnicas de recolección de estas muestras, utilizadas en investigaciones previas. La metodología empleada es el análisis de contenido de cien tesis y artículos de investigación recuperados de portales científicos. Se realizó un análisis descriptivo que consideró, entre otras variables, la cantidad de datos recogidos, la técnica de recolección de datos utilizada, la forma de contacto con los participantes y el acceso en línea a los corpus lingüísticos. Los resultados muestran la existencia de algunos corpus anotados y disponibles en lenguas diferentes a la española. Asimismo, se observa, en la mayoría de los antecedentes, la combinación de diferentes técnicas para recoger un conjunto amplio de datos lingüísticos y multimodales. En tal sentido, se sistematizan las principales alternativas metodológicas con las que es posible recolectar datos de interacciones digitales por WhatsApp.

PALABRAS CLAVE: discurso digital; corpus lingüístico; mensajería instantánea; interacción digital.

FORMA DE CITAR

Cantamutto, L. & Vela Delfa, C. (2023). Corpus de interacciones digitales: sistematización de técnicas para recoger datos en WhatsApp. *Cuadernos.info*, (54), 117-139. https://doi.org/10.7764/cdi.54.53165

ABSTRACT | *The collection of datasets from real interactions is an unavoidable step in many research works aiming to understand language use. In the field of digital discourse analysis, data collection is complex due to the fast-paced changes in the applications and the ethical decisions involved. This work has two goals. First, we seek to show an overview of the literature on datasets of digital exchanges by WhatsApp. Then, we aim to systematize different sampling techniques used in previous research. We thus proceeded by applying content analysis to 100 research articles and theses retrieved from open access portals. We conducted a descriptive analysis that included the amount of data collected, the technique employed in the collection of the data, the method used to contact participants, and the online access to the linguistic corpora, among other variables. The results show the existence of some corpora annotated and available in languages other than Spanish. In addition, most of the literature shows a combination of different techniques to collect a wide set of linguistic and multimodal data. Then, we systematize the main methodological alternatives for data collection from digital interactions by WhatsApp, with the participant observation method standing out.*

KEY WORDS: *digital discourse; corpus linguistics; instant messaging; digital interaction.*

RESUMO | A coleta de conjuntos de dados de interações reais é um passo inevitável em muitas pesquisas que buscam compreender os usos linguísticos. No campo da análise do discurso digital, a coleta de dados é complexa tanto pelas características mutáveis das aplicações quanto pelas decisões éticas envolvidas. O artigo tem um duplo objetivo. Em primeiro lugar, oferecer um estado da questão sobre os conjuntos de dados de trocas digitais por WhatsApp e, em segundo lugar, sistematizar diferentes técnicas de coleta destas amostras usadas em pesquisas anteriores. A metodologia utilizada é a análise de conteúdo de 100 teses e artigos de pesquisa recuperados de portais científicos. Foi realizada uma análise descritiva que levou em consideração, entre outras variáveis, a quantidade de dados coletados, a técnica de coleta de dados utilizada, forma de contato com os participantes e o acesso online ao corpus linguístico. Os resultados mostram a existência de alguns corpus anotados e disponíveis em outros idiomas além do espanhol. Além disso, observa-se, na maioria dos antecedentes, a combinação de diferentes técnicas para coletar um amplo conjunto de dados linguísticos e multimodais. Nesse sentido, são sistematizadas as principais alternativas metodológicas com as quais é possível coletar dados de interações digitais pelo WhatsApp.

PALAVRAS-CHAVE: discurso digital; corpus linguístico; mensagens instantâneas; interação digital.

INTRODUCCIÓN

La función social del lenguaje es objeto de atención de un conjunto de disciplinas en las que estudiar interacciones reales es un paso ineludible (Ädel & Reppen, 2008). Para investigar textos naturales, resulta necesario acceder a corpus de referencia o conformar un conjunto de datos de intercambios reales. Aunque existen numerosos corpus generales o de referencia del español escrito y hablado en diferentes países hispanohablantes, no ocurre lo mismo con datos recogidos en intercambios digitales. Como señala De Benito Moreno (2022), solo están disponibles el *Corpus del Español: Web/Dialects* y el corpus *EsTenTen*, además de la base de datos CoDiCE, que ha permitido intercambiar muestras de lengua entre investigadores. Por lo tanto, los avances en el campo de estudios del discurso digital se realizan a partir de corpus especializados o conjuntos de datos creados *ad hoc* (Collins, 2021), muchas veces reducidos y centrados en alguna plataforma particular (de Benito Moreno, 2022).

Los corpus lingüísticos son una amplia colección de datos (textos) que responden a algún criterio lingüístico, organizados por algún parámetro específico (fecha, género discursivo, procedencia geográfica, situación comunicativa, etc.), almacenados, preferentemente de manera digital, y –en la actualidad– susceptibles de ser analizados por algún software específico (Molina Mejía, 2021). Estas diferencias cualitativas hacen que sea lícito diferenciar los corpus de otros conjuntos de datos, como los archivos/colecciones informatizados y bibliotecas de textos electrónicos, recopilaciones sin criterios lingüísticos (Toruella & Llisterri, 1999).

Rojo Sánchez (2015) manifiesta que la lingüística de corpus equivale a una "lingüística basada en el análisis de corpus" (p. 681), lo que la hace admisible como una herramienta metodológica complementaria a otros estudios lingüísticos y, en tal caso, al análisis del discurso digital, cuyas metodologías para recoger datos lingüísticos han adquirido un creciente interés editorial (Collins, 2021; Vásquez, 2022).

La novedad de la variable digital en el estudio de datos lingüísticos junto con los permanentes cambios de dispositivos y aplicaciones produce que, en los entornos comunicativos digitales, las dificultades, siempre presentes en la recolección de datos lingüísticos, se maximicen. Por esta razón, se suele utilizar un conjunto de técnicas metodológicas complementarias o trabajar con muestras acotadas para mantener rasgos de multimodalidad propios del discurso digital (Thurlow, 2018), en el que se combinan recursos de diferente naturaleza. Llamativamente, las técnicas utilizadas para recoger datos digitales se mantienen relativamente estables desde los estudios iniciales de salas de chat (Pihlaja, 2022) hasta ahora y,

en todo caso, lo que cambian son las opciones brindadas por cada aplicación para acceder, almacenar o exportar esos datos.

Dentro del repertorio de géneros digitales, unos de los fenómenos que más dificultades metodológicas plantea es el de las interacciones digitales privadas llevadas a cabo mediante el intercambio de textos breves (Cantamutto & Vela Delfa, 2020), lo que Yus (2021) denomina *Smartphone messaging*. Cuando los lingüistas se interesan por otros géneros del discurso digital público, por ejemplo, las redes sociales, tienen a su disposición una gran cantidad de datos considerados públicos. Sin embargo, cuando el foco es analizar los géneros digitales privados se evidencian interesantes retos metodológicos que pueden sortearse desde propuestas innovadoras. Además de la privacidad, la creciente multimodalidad de estos entornos también puede acarrear dificultades.

Teniendo en cuenta estas condicionantes, el objetivo de este artículo es doble: por un lado, ofrecer un primer estado de la cuestión sobre los conjuntos de datos de intercambios comunicativos digitales por WhatsApp y, por otro, sistematizar diferentes técnicas para recolectar muestras de interacción por WhatsApp utilizadas en investigaciones previas.

METODOLOGÍA

Siguiendo las propuestas de Beißwenger y Storrer (2008) y Pano Alamán y Moya Muñoz (2016), realizamos el análisis documental de contenido de un corpus formado por artículos de investigación, tesis de maestría y doctorado y webs de proyectos de investigación. Para su selección, se llevó a cabo una búsqueda sistemática en diferentes portales científicos a partir de las palabras claves: corpus y WhatsApp. Dado nuestro interés en los conjuntos de datos en lengua española, se realizó la pesquisa en el portal *Dialnet* (29 resultados) y, posteriormente, en otras bases de datos: *DOAJ* (24 resultados), *Scielo* (3 resultados) y *Redalyc*[1] (62 resultados). Por último, en *Google Scholar* se recogieron únicamente los 50 resultados más relevantes en otras lenguas.

A partir de los resultados, se realizó un primer cribado manual para seleccionar una muestra no probabilística de cien documentos, número que permitió saturar las categorías analizadas. Luego, se volcó en una base de datos la información del documento reseñado, en relación tanto con los corpus que en ellos se analizan

1. En esta base de datos, se seleccionaron únicamente los resultados de la búsqueda WhatsApp con un filtro por disciplina (Lengua y Literatura). Al emplear ambos términos (incluso utilizando operadores booleanos), recuperaba textos que tuvieran uno de los dos e incrementaba notablemente la cantidad (5844).

como con los referenciados dentro del texto. Los datos de estos últimos fueron completados tras buscar las referencias originales. Así, se elaboró una planilla con información sobre las siguientes variables: tipo de contacto, cantidad de datos (conversaciones, mensajes, palabras o tokens), cantidad de participantes, técnica empleada, resguardos éticos (consentimiento, anonimización), año de recogida de los datos, ámbito de uso y tipo de conversación (completa o fragmento). No obstante, tal y como señala Kreis (2022), muchas investigaciones no describen en profundidad el conjunto de datos utilizados ni las técnicas con las que los recogieron. De la base de datos completa, fueron seleccionados algunos ejemplos prototípicos para su comentario en estas páginas.

RESULTADOS

A partir de la reseña de algunos de los corpus lingüísticos de intercambios comunicativos en entornos digitales privados, se sistematizan las principales técnicas utilizadas para recoger datos en WhatsApp.

Corpus de interacciones digitales escritas

Si el acceso y el almacenamiento de muestras de lenguas del discurso digital en general es difícil, más complicado es crear corpus de interacciones digitales del ámbito privado e íntimo, como lo son los intercambios a través de plataformas de mensajería instantánea.

El acceso a interacciones reales ocurridas mediante estas aplicaciones implica desafíos para la investigación lingüística. Estas dificultades provocan un importante sesgo en la elección del objeto de estudio, pues gran parte de las investigaciones sobre comunicaciones digitales se enfoca en muestras de lengua recogidas en redes sociales públicas (como Twitter o plataformas de *e-commerce*) o utilizan la web como corpus (Collins, 2021; de Benito Moreno, 2022; González Fernández, 2017). También hay una cierta debilidad metodológica en trabajos que trabajan con mensajería privada, ya que muchas veces deben conformarse con muestras de lenguas construidas por conveniencia y limitadas por importantes carencias (generalmente, por el tamaño de la muestra).

Las consecuencias negativas de esta situación son dobles. Por una parte, las redes sociales públicas no siempre permiten conocer con precisión las características socioculturales de los hablantes (de Benito Moreno & Estrada Arráez, 2018), lo que implica limitaciones en lo que concierne, por ejemplo, a la variación sociolingüística o sociopragmática. Por otra parte, las redes sociales no dan cuenta de algunos fenómenos interaccionales que no pueden rastrearse de igual manera en intercambios cuya fragmentación conversacional es mucho más acusada. En este contexto, merece la pena enfrentarse a las dificultades metodológicas del análisis

de la mensajería instantánea, dado que nos ofrece la oportunidad de estudiar la lengua en uso desde una perspectiva contextualmente densa.

Los conjuntos de datos utilizados en investigaciones sobre mensajería instantánea son, en general, pequeñas muestras de interacciones de redes sociales de familiares y amigos o a través de contactos con adolescentes y jóvenes en instituciones educativas, tal como sucedía con otros géneros del discurso digital (Cantamutto & Vela Delfa, 2016; Pano Alamán & Moya Muñoz, 2015, 2016). La principal carencia se registra en la lengua española, tercera más utilizada en los intercambios digitales, después del inglés y del chino. En otras lenguas, encontramos diferentes propuestas que recogen datos de interacciones públicas o privadas a través de plataformas de mensajería (incluidas salas de chat).

A continuación, comentamos algunas de estas propuestas, seleccionadas debido a que responden a los principales criterios que diferencian un corpus de un conjunto de datos que, también denominado corpus, sirve como sustento empírico de una investigación.

En una recopilación de corpora de CMC, sobre los intercambios prototípicos de Internet, Beißwenger y Storrer (2008) clasifican los corpus en cuatro tipos. Los corpus compilados para uso general se dividen entre los de datos crudos (como el *Apache SpamAssasin Project* con 6000 correos electrónicos spam, además de otros tres que ya no están accesibles) y los de datos anotados, de los que se mencionan solo dos: *The Dusseldorf CMC Corpus*, con muestras de diferentes géneros sincrónicos y asincrónicos, y *The Dortmund Chat Corpus*, compuesto de sesiones de chat anotadas. El *Dortmunder Chat Corpus* (Beißwenger et al., 2013a, 2013b) está disponible en formato electrónico desde 2005. Este conjunto de datos cuenta con 140.000 chats y aproximadamente un millón de tokens en alemán. A través de diferentes proyectos, se han anonimizado sus datos y, actualmente, cuenta con etiquetado TEI. La mayoría de los corpus reseñados es relativa a proyectos (es decir, no de uso general) y de datos crudos (sin anotar) (Beißwenger & Storrer, 2008). Esta tendencia se ha revertido levemente.

El procesamiento natural del lenguaje requiere de conjuntos de datos para entrenar sus herramientas. Ese fue el objetivo del *NPS Internet Chatroom Conversations Corpus* (Forsythand, Lin & Martell, 2007). El corpus recogió, entre octubre y noviembre de 2006, 10.567 entradas/comentarios, en inglés, en salas de chat, que representan 45.068 tokens. Estos datos han sido utilizados por investigadores de diferentes disciplinas (por ejemplo, Kim et al., 2021), evidenciando la necesidad de contar con muestras de interacciones conversacionales digitales y la vigencia de estos datos a pesar de ser antiguos.

A medida que el uso de WhatsApp se popularizó, diferentes proyectos han recogido datos en esta aplicación, pero, como se verá a continuación, la cantidad de datos difiere mucho de los ejemplos de corpus de chat. La arquitectura de la aplicación no favoreció la accesibilidad de los datos con fines de investigación. Si bien existían formas de exportar las conversaciones, el contacto con los participantes era un paso ineludible para obtener esas muestras de lengua.

Por ejemplo, el *What's up, Switzerland? Corpus* (Uerberwasser & Stark, 2017) recoge 617 chats entre 1538 participantes hablantes de alemán, francés, italiano, rumano e inglés, recogidos en el año 2014. Cuenta con aproximadamente cinco millones y medio de tokens y 350.000 emojis. Durante dos meses (junio y julio de 2014), se invitó a los ciudadanos suizos a enviar como archivos adjuntos sus conversaciones de WhatsApp a través de una dirección de correo electrónico provista por el proyecto. Una vez recibida, se enviaba a los participantes un cuestionario de consentimiento informado que recogía, además, los datos sociodemográficos. Este corpus está anonimizado y cuenta con herramientas de búsqueda. Su antecedente directo es el corpus del proyecto *Sud4Science*[2], centrado en los SMS. También se conformó el corpus *What's up, Deutschland?*[3] (Wyss, 2015 en Ayan, 2020).

Un aspecto interesante del corpus *What's up, Switzerland?* es que, por la metodología empleada, es posible observar el desgranamiento existente entre las conversaciones donadas y los consentimientos conseguidos. El número de chats recibidos fue de 967. Sin embargo, para aquellos casos en los que no se completó el formulario de consentimiento informado los mensajes fueron reemplazados por expresiones tales como *redactedQ51tokens248characters (*Ueberwasser & Stark, 2017: 108). Asimismo, para el alemán, por ejemplo, se recibieron 93 chats con consentimiento. No obstante, como Ueberwasser & Stark (2017) lo ilustran a través de una tabla, solo se cuenta con el consentimiento de todos los participantes en 44 conversaciones e información demográfica completa de los participantes de 25 chats.

Siguiendo estas propuestas, en Holanda se recogieron datos de 218 conversaciones de WhatsApp en el corpus *WhatsApp Corpus Verheijen*, recolectados entre 2012-2014[4]. El corpus está anonimizado y todos los participantes dieron su consentimiento para que sus intercambios fueran utilizados para fines académicos.

2. http://sud4science.org/ (consulta: 11 de agosto de 2022).

3. Según señalan Verheijen y Stoop (2016), este corpus estaba alojado en http://www.whatsup-deutschland.de/ . Sin embargo, en la última consulta realizada (11 de agosto de 2022) no se pudo acceder.

4. https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:112987 (Consulta: 11 de agosto de 2022).

Se priorizó el grupo etario de adolescentes y jóvenes entre 12 y 23 años. Entre quienes enviaron sus conversaciones a través de un correo electrónico del proyecto se sortearon premios (tal como sucediera en el proyecto *Sud4Science*).

Otro antecedente son los datos recogidos en chats públicos. El corpus *PoliWAM* recupera datos de 281 grupos públicos de WhatsApp con un total de 223.404 mensajes y 31.078 participantes antes, durante y después de las elecciones generales en India en 2019 (Srivastava & Singh, 2020) y compiló datos de 26 partidos políticos. Una parte de estos (3848 mensajes) está disponible de manera pública. En esta línea, a partir de la herramienta WebWhatsAppAP, se recogieron mensajes de 127 grupos públicos de Brasil (Resende et al., 2018).

En lengua española, destacamos algunos trabajos de investigación basados en corpus. Por ejemplo, Vázquez Cano y sus colegas (2015) recogieron 417 conversaciones de WhatsApp (101.401 palabras) a través de la colaboración voluntaria de estudiantes de cinco centros educativos de nivel secundario de cuatro provincias de España. Si bien no se indica la técnica de recogida, se explicita que se solicitaron conversaciones reales sobre un tema. El interés de esta investigación es el análisis lexicométrico de la escritura digital. Asimismo, Pérez-Sabater (2015) analiza la variación lingüística a partir de chats de adolescentes y adultos en tres variedades (español, catalán e inglés), recogidos con la colaboración de estudiantes de una asignatura de máster que preparaban subcorpus de 500 palabras. La cantidad total de palabras analizadas es de 41.000.

También solicitando colaboración a estudiantes universitarios en la recolección de datos, Bach y Costa Carrera (2020) implementaron una encuesta a 200 estudiantes de los cuales seleccionaron 23 cuyas lenguas maternas eran el catalán o el español y el catalán. Tras contactarlos por correo electrónico y una serie de entrevistas, se solicitó a "los estudiantes [que] nos hicieron llegar fragmentos de sus wasaps en formato de texto (.txt) por correo electrónico, seleccionados libremente por ellos" (Bach & Costa Carrera, 2020, p. 574).

En esta línea, el *Sociolinguistic Corpus of WhatsApp Chats in Spanish for College Speech Analysis* (Dorantes et al., 2018) está compuesto por conversaciones de estudiantes universitarios de la Universidad Nacional Autónoma de México. Los datos lingüísticos, donados por estudiantes elegidos al azar en el campus, fueron enviados por correo electrónico. El corpus tiene 835 chats de 1325 participantes. Tras el proceso de anonimización y eliminación de mensajes de la plataforma, el corpus cuenta con 66.465 mensajes y 756.066 tokens.

Por último, la base de datos *CoDiCE* (Cantamutto, Vela Delfa y Boisselier, 2015) es un proyecto colaborativo para poner a disposición de investigadores

académicos muestras de lengua de interacción digital escrita (correo electrónico, SMS y WhatsApp). Esta base de datos se nutre de diferentes corpus recogidos por investigadores y contiene aproximadamente 14.000 SMS y correos electrónicos y 3000 mensajes de WhatsApp. Los datos han sido anonimizados y, a partir de herramientas de la base de datos, pueden ser analizados. Actualmente, este proyecto está siendo ampliado e incorpora nuevas funcionalidades, además de datos nuevos procedentes de interacciones por WhatsApp del español bonaerense y del español de la Patagonia.

En esta breve síntesis, se observa una situación disímil en relación con la cantidad de datos disponibles para las diferentes lenguas, así como una gran concentración de conversaciones recogidas en el grupo etario de jóvenes y adolescentes (del mismo modo que sucediera con los datos sobre SMS). En otros casos, se trata de pequeños conjuntos de datos extraídos de grupos creados con fines académicos pero que se utilizan, posteriormente, como material de investigación (entre otros, Ayan, 2020; García Gómez, 2020).

En la siguiente sección, se exponen las alternativas metodológicas utilizadas en los corpus reseñados previamente.

Instrumentos y técnicas para recoger datos lingüísticos de mensajería instantánea

En concordancia con el creciente número de usuarios de WhatsApp, las investigaciones sobre diferentes fenómenos pragmático-discursivos han proliferado en los últimos años. Destacados investigadores e investigadoras del campo del discurso digital han atendido a fenómenos específicos de este tipo de intercambios. Tal es el caso, por ejemplo, del texto clásico de Manuel Alcántara-Plá (2014) sobre las unidades conversacionales de WhatsApp, la propuesta de caracterización discursiva de Calero Vaquera (2014), que sitúa a este tipo textual entre el Messenger y los SMS, o el análisis de los valores y funciones de los emoticones y emojis en Sampietro (2016) y en Cantamutto y Vela Delfa (2019), entre otras.

Todas estas investigaciones utilizaron diferentes métodos para construir un pequeño corpus o muestras por conveniencia. En esta sección, reseñamos las principales técnicas que no se excluyen mutuamente, sino que, por el contrario, se suelen utilizar de manera complementaria. Se comenta, brevemente, la forma de contacto, de resguardo de identidades de los participantes, de almacenamiento de las muestras de lengua, y se profundiza en alternativas metodológicas que se han utilizado en este breve, pero intenso período en el que WhatsApp ha sido estudiado.

Exportar chat

La principal herramienta para recoger datos donados por hablantes se ha ido modificando desde su creación. En las primeras versiones de la aplicación de WhatsApp, los usuarios podían guardar las conversaciones en una carpeta en la memoria del teléfono o en una tarjeta externa y, de ese modo, extraerlas. Posteriormente, se habilitó la opción de exportarla a través del correo electrónico como texto en el cuerpo del correo o como archivo de texto en formato adjunto (con extensión .txt). En estos casos, la información multimodal no se adjuntaba (figura 1). Con esta técnica no se recogían las fotografías, por ejemplo, ni los emojis, pero se accedía a un documento en el que se presentaba la conversación siguiendo el orden de las intervenciones, con la disposición que tenían en la cuenta desde la cual se comparte la conversación (en ocasiones, de manera ascendente y en otras de manera inversa). La transcripción de las conversaciones realizada automáticamente por la aplicación resulta en una "visualización parecida a la de un guion de teatro o a las tradicionales transcripciones de conversaciones orales" (Sampietro, 2016, p. 150), en las que se perdía el "formato visual de la aplicación, así como los adjuntos multimedia enviados y la mayoría de los emoticonos incluidos en los wasaps" (Bach & Costa Carreras, 2020, p. 574).

De manera paulatina y dependiendo del sistema operativo del teléfono, en la exportación se comenzaron a recuperar los elementos multimedia: primero, emojis y, posteriormente, imágenes, audios y otros contenidos no verbales. En el ejemplo precedente, la fotografía enviada solo se indica mediante la palabra "imagen" (intervención 10 de "Martín") y el emoji se consigna con un □ (intervención 9 de "Usted[5]").

Esta técnica fue utilizada a partir de solicitar a "porteros" (en general, estudiantes de instituciones de nivel medio o superior o redes sociales de familiares y amigos que sirven como intermediarios) que exportaran las conversaciones. Así lo hicieron, por ejemplo, para el corpus de Dorantes y sus colegas (2018) o en Maíz-Arévalo, 2018, quien agradece de manera explícita a sus colaboradores (Maíz-Arévalo, 2002).

En el caso de *WhatsApp Corpus Verheijen*, un aspecto que remarcan es que la página web del proyecto tenía muchas indicaciones, según dispositivo y sistema operativo, para exportar las conversaciones (Verheijen & Stoop, 2016, p. 251). Tras varias actualizaciones de la aplicación, la herramienta de exportación funciona de modo similar en cualquier teléfono, favoreciendo esta técnica de recolección de datos. El usuario debe ingresar al menú desplegable que se presenta arriba a la derecha en los tres puntos y seleccionar, primero, "Más" y luego "Exportar chat". Una vez realizada esta acción, aparece el siguiente cartel (figura 2):

5. En el año 2012, al exportarse las conversaciones aparecía "Usted" y el nombre del otro participante. En este caso, se optó por cambiar el nombre del participante "Martín" y dejar la forma "Usted" tal como se exportó la conversación en ese momento.

15:51, 2012/6/20 – Martín: Me va a buscar un amigo o la novia
15:52, 2012/6/20 – *Usted: :D*
15:53, 2012/6/20 – Usted: Nunca llegó la foto ☹
15:56, 2012/6/2– - Martín: La reenvie...
15:56, 2012/6/2– - Martín: Es una foto cualquirea
15:56, 2012/6/2– - Martín: Mia de hoy ...
15:58, 2012/6/2– - Usted: :D
15:59, 2012/6/2– - Usted: Ojalá llegué
15:59, 2012/6/2– - *Usted:* 🔲
16:01, 2012/6/2– - *Martín: imagen*
16:01, 2012/6/2– - Usted: Ahí llegó la foto

Figura 1. Exportación de conversaciones de WhatsApp (año 2012)

Fuente: Elaboración propia.

Figura 2. Mensaje automático de la aplicación WhatsApp para exportar las conversaciones con o sin archivos multimedia

Fuente: WhatsApp.

Si se selecciona la segunda opción, se genera un archivo de texto (con extensión .txt) que se acompaña de los archivos multimedia presentes en la conversación. El usuario puede elegir entre múltiples aplicaciones para enviar el historial y, en todos los casos, el archivo generado conservará relación entre los archivos multimedia que no aparecen embebidos en el documento .txt pero que se pueden detectar por su nombre (como sucede en la figura 4). En la primera intervención y en la segunda se hace referencia a dos archivos de audio enviados que, si bien no están disponibles en el historial de la conversación exportada, lo están entre los archivos adjuntos del correo electrónico (figura 3).

Es decir, con la herramienta para exportar es posible conservar, por un lado, un archivo de texto con la conversación estructurada según fecha y hora de envío, en las que los emojis se visualizan en el mensaje correspondiente, denominada "Chat de WhatsApp con + nombre del contacto/grupo". En cambio, las notas de audio, fotografías, archivos de texto y stickers aparecen adjuntos, pero indicados en el cuerpo de la conversación como se observa en la siguiente imagen (intervención 1 y 2 de la figura 4).

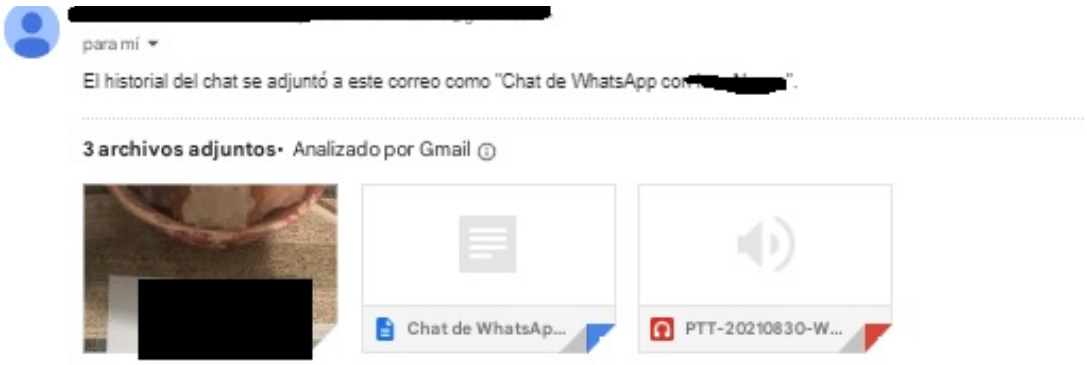


Figura 3. Captura de correo electrónico de exportación de conversación con la herramienta Exportar de WhatsApp (año 2020)

Fuente: Elaboración propia.

19/6/20 14:18 - Docente: PTT-20200619-WA0080.opus (archivo adjunto)
20/6/20 18:08 - Estudiante: PTT-20200620-WA0044.opus (archivo adjunto)
20/6/20 18:23 - Estudiante: jaja bueno gracias por la sinceridad
20/6/20 18:23 - Docente: no hay problema
20/6/20 18:23 - Estudiante: te recomiendo que hagas el 3 y el 4 que son de la unidad 1.
20/6/20 18:24 - Estudiante: Aah el 2 es una película no?
20/6/20 18:24 - Docente: el 2 es un trabajo de la escuela que nos pidieron con esa película. Si en algún momento podes, lo haces.
20/6/20 18:24 - Docente: sisi
20/6/20 18:24 - Estudiante: Bueno voy a arrancar con el 3 entonces
20/6/20 18:24 - Estudiante: dale! cuando lo hagas me decís si hay algo que no entendes
2/9/21 15:50 - Docente: https://forms.gle/xxxxxxxxxx
2/9/21 15:50 - Docente: AUD-20210902-WA0068.opus (archivo adjunto)
2/9/21 15:50 - Estudiante: Hacen click en el enlace y los lleva directo 🙂

Figura 4. Fragmento de conversación exportado a través de la herramienta de WhatsApp (año 2020)

Fuente: Elaboración propia.

El grupo etario predilecto es el de adolescentes y jóvenes que, por estar dentro de instituciones educativas, resultan accesibles como participantes voluntarios; además suelen encontrarse entre la red de familiares y amigos del investigador/a que se comentará en el siguiente apartado. Esta tendencia a enfocar la atención en este grupo etario, sin embargo, no carece de problemas asociados con la falta de participación. Por ejemplo, Sampietro (2016) señala que optó por dos vías de contacto. Primero a su red de contactos el envío de chats o fragmentos (tanto exportando como a través de capturas) y, luego, a un grupo de estudiantes. Sin embargo, solamente una estudiante envió una conversación.

Este aspecto nos enfrenta a uno de los principales problemas de exportar conversaciones directamente a casillas de correo institucionales (como en el caso de Sampietro, 2016 y Verheijen & Stoop, 2016): el poco control de los y las participantes sobre sus datos. En este sentido, señala Sampietro (2016) que en su recogida de datos dos participantes optaron por enviar el historial de la conversación a su correo personal, eliminar un fragmento que no querían compartir y luego enviar parte del chat a la investigadora.

Exportar la conversación a un correo que no sea el propio genera un archivo con todo el historial de conversación entre dos o más participantes que se puede remontar al primer intercambio entre esas personas en la aplicación de WhatsApp en el dispositivo. Si se trata de intercambios entre familiares y amigos, por ejemplo, que interactúan a diario, una sola conversación puede durar varios meses. Sin embargo, en las sucesivas actualizaciones, WhatsApp, además de incorporar la opción de eliminar mensajes y de enviar mensajes que se autoeliminan en un lapso, recupera únicamente una parte de la conversación con la opción de exportar. En este momento, por restricciones en el tamaño del archivo, la técnica más utilizada recorta la conversación según el peso de los archivos adjuntos. De este modo, si se envía el historial por correo electrónico sin incluir archivos multimedia se exporta toda la conversación, mientras que si se lo hace incluyendo archivos se obtiene un fragmento dependiente del tamaño de los archivos.

Capturas

A medida que los teléfonos móviles han ido evolucionando, se han incorporado diferentes funciones en los dispositivos, como las capturas o pantallazos (*screenshots*): fotografías de la pantalla del celular, similar a la función imprimir pantalla de las computadoras. Las ventajas de esta técnica son varias. Por un lado, conserva gran cantidad de datos multimodales (incluso, por ejemplo, la imagen del fondo de pantalla que puedan usar los interlocutores o la estructura conversacional a partir del empleo de las funciones responder o reaccionar con emojis). Por otro lado, resulta un método sencillo y rápido para los colaboradores, dado que tienen mayor control sobre los datos que comparten, al seleccionar los enunciados que envían. Si bien los datos enviados no están modificados, los participantes salvaguardan su imagen al seleccionar fragmentos de la conversación o al enviar comentarios posteriores que reflexionan sobre sus usos (figura 5).

En una investigación previa (Cantamutto & Vela Delfa, 2019), utilizando la técnica de bola de nieve, conseguimos un gran número de capturas de pantalla de los emojis usados recientemente. En la primera recogida de datos (2015-2016), los colaboradores preguntaban cómo realizar la captura. Algunos, desconociendo cómo hacerlo, copiaban los emojis que aparecían en el menú.

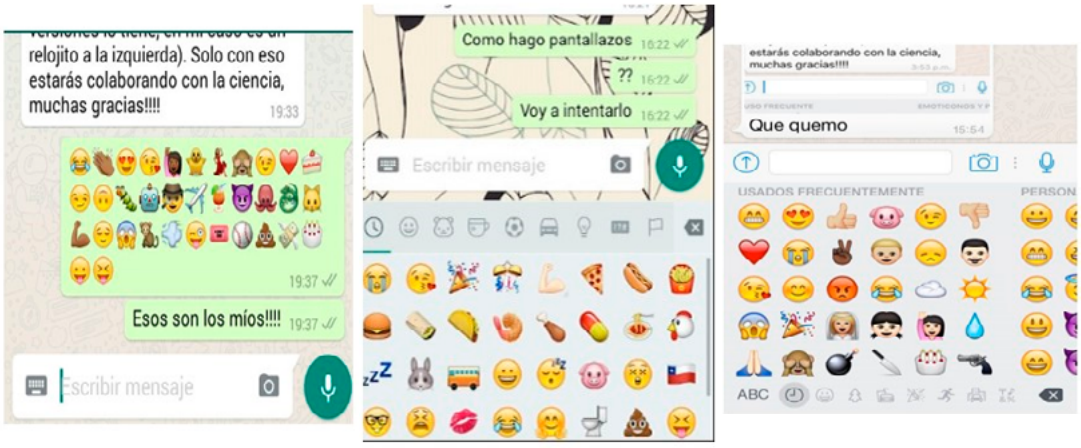


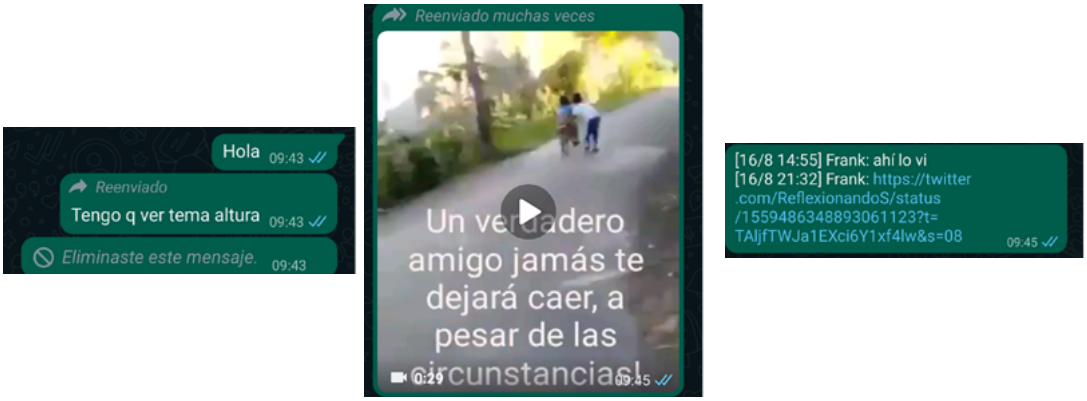
Figura 5. Ejemplos de capturas de pantalla recogidas entre 2015-2016 (datos utilizados en Cantamutto y Vela Delfa, 2019)

Fuente: Elaboración propia.

En otros casos, tras averiguar qué modelo y marca de teléfono tenía el colaborador, era posible darle instrucciones sobre cómo hacer la fotografía de la pantalla. En la siguiente imagen, se observan los tres casos reseñados: el usuario que transcribe los emojis frecuentes; quien pregunta cómo realizar la captura; y el que reflexiona metalingüísticamente sobre sus emojis.

En resumen, cada decisión metodológica tiene consecuencias en los datos obtenidos. Optar por las capturas favorece una menor manipulación de las muestras de lengua por parte del colaborador (y ese era el interés perseguido al pedir los emojis frecuentes). Aunque tiene mayor control –al poder seleccionar y observar qué envía–, no modifica los intercambios. Sin embargo, el uso de este tipo de imágenes tiene consecuencias en su tratamiento posterior. En primer lugar, es necesario transcribir las conversaciones manualmente. En este sentido, se manipulan los datos y se pueden producir modificaciones involuntarias. En segundo lugar, se accede a una imagen estática de los videos y tampoco es posible reproducir los audios. De este modo, resulta necesario combinar el envío de capturas con la exportación del historial de conversación o con los reenvíos.

Sin embargo, las capturas ayudan a comprender mejor el desarrollo de la conversación, dado que, con la exportación de la conversación como con los reenvíos, se pierden las marcas paratextuales de la estructura conversacional de WhatsApp: responder, reenviado, reenviado muchas veces y las reacciones a los mensajes (figura 6).



1/8/22 09:43 - A: Hola 1/8/22 09:43 - A: Tengo q ver tema altura 1/8/22 09:43 - A: Eliminaste este mensaje.	1/8/22 09:45 - A: VID-20220720-WA0038.mp4 (archivo adjunto)	1/8/22 09:45 - A: [16/8 14:55] Frank: ahí lo vi [16/8 21:32] Frank: https://twitter.com/xxxxxx

Nota: Los mensajes son ilustrativos y no corresponden a intercambios reales.

Figura 6. Diferencias de visualización de los mensajes extraídos con capturas o mediante exportar chat

Fuente: Elaboración propia.

Reenvíos y copiar/pegar

Una tercera vía para solicitar a colaboradores que donen conversaciones es pedirles que reenvíen las distintas intervenciones a un número de teléfono. De hecho, esta alternativa, presente desde los inicios de la aplicación, permitía enviar fragmentos de conversaciones de WhatsApp a través de SMS. En la actualidad, solo se puede reenviar al mismo WhatsApp o utilizar la función de copiar/pegar para exportarlo a otra aplicación.

Esta práctica es habitual entre los usuarios, más allá de las posibilidades que le brinda al investigador. Por ello, en las sucesivas actualizaciones la herramienta introdujo nuevas funcionalidades para reenviar los mensajes con información paratextual que permitiera saber si el enunciado ha sido escrito en otra conversación. Un mensaje reenviado suele aparecer con una leyenda superior que señala la intertextualidad (reenviado) o con marcas similares a didascalias que indican fecha, hora y emisor del mensaje, cuando varios mensajes son copiados y pegados en otro chat (figura 6).

En los proyectos con una cuenta de WhatsApp, usar esta técnica permite solicitar a los colaboradores que compartan fragmentos de una conversación, por ejemplo, en torno a un tema, o mensajes que funcionan como ejemplo (el empleo de algún

emoji específico o alguna expresión). Esta técnica fue empleada para el estudio del voseo en Bolivia. Contactaron, a través de correo electrónico, a 80 personas (estratificadas por grupo etario) y les solicitaron que reenviaran un audio: "A los interesados, se les pidió que enviaran sus propias grabaciones de mensajes dirigidas a uno de sus padres" (Castedo et al., 2022, p. 398).

Esta técnica resulta productiva en determinados dominios de uso (como el comercial), en los que es posible reponer los papeles enunciativos a partir de las intervenciones enviadas. Como se observa, resulta complementaria de otras alternativas para recoger datos, ya que se pierde gran cantidad de información sobre la estructura conversacional, principalmente.

Participante observador

Esta opción resulta diferente de las anteriores, porque ya no es el colaborador quien elicita las muestras de lengua, sino que el investigador participa en la conversación. Esta participación puede ser desde un punto ciego, a partir de la observación-participante en grupos de WhatsApp, o como participante-observador (Vela Delfa & Cantamutto, 2016). Esta técnica ha sido utilizada con éxito en recientes investigaciones que, entre otras técnicas etnográficas, incluyen la observación en grupos de WhatsApp.

Tal es el caso, por ejemplo, del trabajo de Lucía Godoy (2021) sobre prácticas letradas con tecnologías digitales en las clases de lengua y literatura de nivel secundario. Además de la observación-participante en las clases presenciales, la investigadora se sumó a los grupos escolares de WhatsApp. De hecho, Godoy (2021) señala que esta técnica "permite recoger datos reales que surgen en contextos digitales auténticos (Hine, 2000, 2015), sin la necesidad de elicitarlos, limitando el grado de invasión en el espacio de participación y considerando de manera profunda y amplia los contextos en los que surgen esos datos sin depender del envío de información complementaria por parte de los hablantes (...)" (pp. 123-124). Este es un caso en el que se logra la observación desde punto ciego (Vela Delfa & Cantamutto, 2016).

Otro antecedente es el estudio de los marcadores conversacionales *ahre* y *tipo* (de Luca, 2021). De manera complementaria con otras técnicas etnográficas, parte de los datos de esta investigación fue recolectada mediante la creación de grupos de WhatsApp en los que estudiantes enviaban memes con estos marcadores. La investigadora formaba parte de un grupo creado *ad hoc* que funcionaba para que los participantes compartan y reflexionen sobre sus propios usos. Así, se combina la observación-participante con los reenvíos.

	Exportar chat	Capturas	Reenviar	Participante observador
Texto plano disponible	Sí	No	Sí	Sí
Archivos multimedia: fotos	Sí	Depende	Sí	Sí
Archivos multimedia: audios	Sí	No	Sí	Sí
Archivos multimedia: videos	Sí	No	Sí	Sí
Graficones: emojis	Sí	Sí	Sí	Sí
Graficones: stickers	Sí	Depende	Sí	Sí
Marcas paratextuales: mensaje eliminado	Sí	Sí	No	Sí
Marcas paratextuales: reenviado/reenviado muchas veces	No	Sí	Depende	Sí
Marcas paratextuales: responder a mensaje	No	Sí	No	Sí
Marcas paratextuales: reacciones	No	Sí	No	Sí

Tabla 1. Resumen de técnicas para extraer datos lingüísticos de WhatsApp

Fuente: Elaboración propia.

Por otro lado, se utilizan interacciones en las que el investigador cumple un rol doble: interlocutor y observador. Esto se alinea con una técnica muy común: el envío de muestras por parte de la red de familiares y amigos. Por ejemplo, Sampietro (2016, p. 149) justifica su empleo para recoger los datos de su investigación reseñando la metodología usada por Tannen (1984), quien registró una cena de amigos para, posteriormente, analizarla. Es decir, al solicitar a familiares y amigos el consentimiento para donar conversaciones ocurre, muchas veces, que las propias intervenciones del investigador/a sean analizadas.

Como se observa, recoger datos con esta técnica implica dos posibilidades según el rol que asume el investigador: participante o participante observador. La primera alternativa es posible en los grupos de WhatsApp, mientras que la segunda implica la utilización de intercambios en los que el investigador sea uno de los interlocutores. En WhatsApp, la técnica de participante-observador resulta sumamente productiva en intercambios diádicos y en conversaciones grupales en las que es posible evitar la paradoja del observador.

En la tabla 1, exponemos sintéticamente las opciones que, cada una de las técnicas, brinda en la recogida de datos. La combinación de ellas redundará en la generación de corpus ricos en datos lingüísticos y multimodales, mayor información sobre los perfiles sociolingüísticos y la obtención del consentimiento informado de las personas colaboradoras.

DISCUSIÓN

En primer lugar, este artículo ha procurado ser un aporte a las consideraciones metodológicas requeridas para desarrollar una investigación en profundidad sobre muestras de lengua de interacciones digitales. El relevamiento de los conjuntos de datos utilizados por investigaciones previas pone en evidencia un área de vacancia en torno a la constitución de corpus amplios, sistemáticos y de consulta general sobre interacciones digitales en lengua española. Asimismo, estos datos tampoco tienen presencia en los corpus de referencia del español. Sin embargo, el creciente interés por conversaciones digitales del ámbito privado ha dado lugar a diferentes investigaciones sobre WhatsApp que, para sortear las dificultades metodológicas existentes, han creado muestras por conveniencia a partir de la combinación de técnicas.

En segundo lugar, hemos presentado las principales técnicas para recoger datos interaccionales de WhatsApp. En tal sentido, destaca la técnica de participante-observador como aquella que permite recuperar mayor cantidad de datos lingüísticos y multimodales, así como información sobre los perfiles sociolingüísticos y sobre las situaciones comunicativas en las que se insertan estos intercambios. Este tipo de materiales puede complementarse con otras técnicas para recoger fragmentos de conversaciones, como la captura de pantalla y el uso de la opción reenviar/copiar y pegar. Asimismo, la alternativa más utilizada en las investigaciones previas –la herramienta de exportación– presenta múltiples ventajas para el tratamiento de los datos, ya que están en formato de texto. Estos datos podrían ser analizados mediante herramientas de análisis lingüístico (como AntConc) de manera muy simple.

Por todas estas dificultades, diferentes proyectos de corpus de WhatsApp han propuesto la combinación de técnicas, así como la inclusión de otros instrumentos para recuperar la percepción de los hablantes sobre sus usos lingüísticos. Entre estas técnicas complementarias para recoger datos de la percepción destacamos, nuevamente, la importancia de evitar la mediación de otras aplicaciones. La utilización de WhatsApp en la realización de entrevistas, encuestas o test de hábitos sociales resulta muy productiva para indagar, precisamente, en los usos de esta aplicación.

El dinamismo de WhatsApp, por un lado, y los cambios en los hábitos de los usuarios, por otro, ponen de relieve la necesidad de recoger muestras de lengua para atender a los cambios (micro)diacrónicos en las prácticas discursivas. La descripción proporcionada ha intentado dar cuenta del estado actual de las alternativas para recoger datos lingüísticos que ofrece WhatsApp. En trabajos futuros se atenderá a las poblaciones que han sido relegadas en la agenda de investigación (personas adultas, por ejemplo), así como a diferentes maneras de resguardar éticamente a los participantes.

FINANCIACIÓN

"Diseño e implementación de un corpus sobre comunicación digital del español bonaerense y de la Patagonia". PICT-2019-02093-PRÉSTAMO BID, Agenia Nacional de promoción de la investigación, el desarrollo tecnológico y la innovación (Argentina). IP: Lucía Cantamutto

REFERENCIAS

Ädel, A. & Reppen, R. (Eds.). (2008). *Corpora and Discourse. The challenges of different settings*. John Benjamins Publishing.

Alcántara-Plá, M. (2014). Las unidades discursivas en los mensajes instantáneos de wasap (The discursive units in WhatsApp instant messages). *Estudios de Lingüística del Español, 35*, 2014. https://infoling.org/elies/35/elies35.1-9.pdf

Ayan, E. (2020). Descriptive Analysis of Emoticons/Emoji and Persuasive Digital Language Use in WhatsApp Messages. *Open Journal of Modern Linguistics*, *10*(4), 375-389. https://doi.org/10.4236/ojml.2020.104022

Bach, C. & Costa Carreras, J. (2020). Las conversaciones de wasap: ¿un nuevo género entre lo oral y lo escrito? (Whatsapp conversations: A new genre between orality and writing?) *Revista Signos. Estudios De Lingüística*, *53*(104), 568-591. http://revistasignos.cl/index.php/signos/article/view/329

Beißwenger, M. & Storrer, A. (2008). Corpora Of Computer-Mediated Communication. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook* (pp. 292-308). Mouton de Gruyter.

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L., & Storrer, A. (2013a). Dortmunder Chat-Korpus. [Data Set] https://www.uni-due.de/germanistik/chatkorpus/ (consulta: 11 de agosto de 2022).

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L., & Storrer, A. (2013b). *DeRiK*: A German reference corpus of computer-mediated communication. *Literary and Linguistic Computing*, *28*(4), 531-537. https://doi.org/10.1093/llc/fqt038

Cantamutto, L., Vela Delfa, C. & Boisselier, L. (2015). *Comunicaciones Digitales: Corpus del español (CODICE).* [Data Set]. Disponible en: aplicacionesonline.codice.com.ar.

Cantamutto, L. & Vela Delfa, C. (2016). El discurso digital como objeto de estudio: de la descripción de interfaces a la definición de propiedades (Digital Discourse As A Subject Of Study: From The Interfaces Description To The Properties Definition). *Aposta. Revista de Ciencias Sociales, 69*, 296-323. http://apostadigital.com/revistav3/hemeroteca/cvela2.pdf

Cantamutto, L. & Vela Delfa, C. (2019). Emojis frecuentes en las interacciones por WhatsApp: estudio comparativo entre dos variedades de español (Argentina y España) (Frequent emojis in WhatsApp interactions: a comparative study between two Spanish varieties (Argentina and Spain). *Círculo de Lingüística Aplicada a la Comunicación, 77*, 171-186. https://doi.org/10.5209/CLAC.63282

Cantamutto, L. & Vela Delfa, C. (2020). Mensajes, publicaciones, comentarios y otros textos breves de la comunicación digital (Messages, Publications, Comments and other brief Texts of the Digital Communication). *Tonos Digital: Revista Electrónica de Estudios Filológicos*, (38), 1-27. http://www.tonosdigital.es/ojs/index.php/tonos/article/view/2394/

Calero Vaquera, M. L. (2014). El discurso del WhatsApp: entre el Messenger y el SMS. Oralia, 17, 85-114.

Castedo, T. M., de Marques Lucena, R., & Gomes da Silva, C. (2022). Vos: Young, Poor and Vulgar in Eastern Bolivia? A Corpus Study on Voseo in WhatsApp Exchanges. *Íkala, Revista De Lenguaje Y Cultura*, *27*(2), 393–410. https://doi.org/10.17533/udea.ikala.v27n2a06

Collins, L. C. (2019). *Corpus Linguistics For Online Communication: A Guide For Research*. Routledge.

de Benito Moreno, C. (2022). Uso de los medios digitales de comunicación como corpus de español (Use of digital communication media as a corpus of Spanish). In G. Parodi, P. Cantos-Gómez, & C. Howe (Coords), *Lingüística de corpus en español* (The Routledge Handbook of Spanish Corpus Linguistics) (pp. 481-493). Routledge.

de Benito Moreno, C. & Estrada Arraéz, A. (2018). Aproximación metodológica al estudio de la variación lingüística en las interacciones digitales (A methodological approximation to the study of linguistic variation in digital interactions). *Revista de Estudios Del Discurso Digital*, (1), 74–122. https://doi.org/10.24197/redd.1.2018.74-122

De Luca, N. (2021). El marcador conversacional ahre en memes: hacia la definición del marcador-meme en interacciones digitales de dos comunidades de práctica juveniles (The conversational marker *ahre* in memes: towards the definition of the marker-meme in digital interactions of two youth communities of practice). *Pragmática Sociocultural/ Sociocultural Pragmatics*, 9(1), 76–95. https://doi.org/10.1515/soprag-2021-0008

Dorantes, A., Sierra, G., Donohue Pérez, T. Y., Bel-Enguix, G., & Jasso Rosales, M. (2018). Sociolinguistic Corpus of WhatsApp chats in Spanish among College Students. In L.W. Ku & C. T. Li (Eds.), *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media* (pp. 1-6). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-3501

Forsythand, E. N., Lin, J. y Martell, C. (2007). *NPS Internet Chatroom Conversations Corpus*. [Data Set] Release 1.0 LDC2010T05. https://doi.org/10.35111/eqdj-ta72

Forsythand, E. N. & Martell, C. H. (2007). Lexical and Discourse Analysis of Online Chat Dialog. *International Conference on Semantic Computing (ICSC 2007)*, 19-26. IEEE. https://doi.org/10.1109/ICSC.2007.55

García-Gómez, A. (2020). Intercultural and interpersonal communication failures: analyzing hostile interactions among British and Spanish university students on WhatsApp. *Intercultural Pragmatics, 17*(1), 27-51. https://doi.org/10.1515/ip-2020-0002

Godoy, L. F. (2021). Interacción colaborativa escolar en WhatsApp: entre la tarea y las bromas (Collaborative school interaction: between homework and jokes). *Revista Estudios del Discurso Digital*, (4), 115-145. https://doi.org/10.24197/redd.4.2021.115-145

González Fernández, A. (2017). The Web as Corpus: An Overview. *Lengua y Habla*, (21), 126-150.

Kim, J. Y., Calvo, R. A., Enfield, N. J., & Yacef, K. (2021). A Systematic Review on Dyadic Conversation Visualizations. In Z. Hammal & C. Busso (Eds.), *ICM'21 Companion: Companion Publication of the 2021 International Conference on Multimodal Interaction* (pp. 137–147). ACM. https://doi.org/10.1145/3461615.3485396

Kreis, R. (2022). Data Collection, Preparation, and Management. In C. Vásquez (Ed.), *Research Methods for Digital Discourse Analysis* (pp. 73-90). Bloomsbury

Maíz-Arévalo, C. (2018). Emotional Self-Presentation on Whatsapp: Analysis of the Profile Status. *Russian Journal of Linguistics, 22*(1), 144-160. https://doi.org/10.22363/2312-9182-2018-22-1-144-160

Molina Mejía, J. M. (2021). *Lingüística computacional y de corpus: Teorías, métodos y aplicaciones* (Computational and corpus linguistics: Theories, methods and applications). Universidad de Antioquia.

Pano Alamán, A. & Moya Muñoz, P. (2015). CorpusRedEs. Proyecto de creación y anotación de un corpus de comunicación mediada por ordenador en español (CorpusRedEs. Project for the creation and annotation of a corpus of communication mediated by computer in Spanish). *CHIMERA. Romance Corpora and Linguistic Studies*, *2*, 117–129. https://revistas.uam.es/chimera/article/view/1042

Pano Alamán, A. & Moya Muñoz, P. (2016). Una aproximación a los estudios sobre el discurso mediado por ordenador en lengua española (An approach to studies on computer-mediated discourse in Spanish). *Tonos Digital: Revista Electrónica de Estudios Filológicos*, (30), 1-30.

Pérez-Sabater, C. (2015). Discovering language variation in WhatsApp text interactions. *Onomázein*, (31),113-126. https://doi.org/10.7764/onomazein.31.8

Pihlaja, S. (2022). Data Sampling and Digital Discourse. In C. Vásquez (Ed.), *Research Methods for Digital Discourse Analysis* (pp. 55-72). Bloomsbury

Resende, G., Messias, J., Silva, M., Almedia, J., Vasconcelos, M., & Benevenuto, F. (2018). A System for Monitoring Public Political Groups in WhatsApp. In M. Carvalho Marques Neto, R. Lima Novais, C. Ferraz, & W. Viana (Chairs), *WebMedia'18: Proceedings of the 24th Brazilian Symposium on Multimedia and the Web* (pp. 387–390). ACM. https://doi.org/10.1145/3243082.3264662

Sampietro, A. (2016). *Emoticonos y emojis: análisis de su historia, difusión y uso en la comunicación digital actual*. Tesis doctoral: Univerdidad de Alicante

Srivastava, V. & Singh, M. (2020). PoliWAM: An Exploration of a Large Scale Corpus of Political Discussions on WhatsApp Messenger. *arXiv preprint arXiv:2010.13263*. https://doi.org/10.48550/arXiv.2010.13263

Tannen, D. (1984). *Conversational style: Analyzing talk among friends*. Ablex

Thurlow, C. (2018). Digital discourse: Locating language in new/social media. In J. Burgess, A. Marwick, & T. Poell (Eds.), *The SAGE Handbook of Social Media* (pp. 135-145). SAGE. https://doi.org/10.4135/9781473984066

Toruella, J. & Llisterri, J. (1999). Diseño de corpus textuales y orales (Design of textual and oral corpora). In J. M. Blecqua, G. Clavería, C. Sánchez, & J. Toruella (Eds.), *Filología e informática. Nuevas tecnologías en los estudios filológicos* (Philology and computer science. New technologies in philological studies). Editorial Milenio.

Ueberwasser, S. & Stark, E. (2017). What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik online*, *84*(5). https://doi.org/10.13092/lo.84.3849

Vásquez, C. (2022). *Research Methods for Digital Discourse Analysis*. Bloomsbury

Vázquez-Cano, E., Mengual-Andrés, S., & Roig-Vila, R. (2015). Análisis lexicométrico de la especificidad de la escritura digital del adolescente en WhatsApp (Lexicometric Analysis of the Specificity of Teenagers' Digital Writing In WhatsApp). *Revista de Lingüística Teórica y Aplicada*, *53*(1), 83-105. https://doi.org/10.4067/S0718-48832015000100005

Vela Delfa, C. & Cantamutto, L. (2016). De participante a observador: el método etnográfico en el analisis de las interacciones digitales de WhatsApp (From Participant to Observer: The Ethnographic Method In The Analysis Of Whatsapp Digital Interactions). *Tonos Digital: Revista Electrónica de Estudios Filológicos*, (31), 1-22. http://www.tonosdigital.com/ojs/index.php/tonos/article/view/1531

Ueberwasser, S. & Stark, E. (2017). What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik online* 84(5), 105-126. https://doi.org/10.13092/lo.84.3849

Verheijen, L. & Stoop, W. (2016). Collecting Facebook Posts and WhatsApp Chats. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, Speech, and Dialogue* (pp. 249-258). Springer. https://doi.org/10.1007/978-3-319-45510-5_29

Yus, F. (2021). *Smartphone Communication: Interactions in the App Ecosystem*. Routledge.

SOBRE LAS AUTORAS

LUCÍA CANTAMUTTO, es Doctora en Letras por la Universidad Nacional del Sur (Argentina). Actualmente, es investigadora del CONICET en el Centro Interdisciplinario de Estudios sobre Derechos, Inclusión y Sociedad de la Universidad Nacional de Río Negro. Sus investigaciones están centradas en la comunicación digital. En 2015, junto con Cristina Vela Delfa y Leandro Boisselier, creó la base de datos CoDiCE (comunicación digital: corpus del español). Es vicepresidenta de la Red de Estudios sobre Comunicación Digital.

https://orcid.org/0000-0001-5868-7608

CRISTINA VELA DELFA, es Doctora en Ciencias del Lenguaje por la Universidad Complutense de Madrid (España). En la actualidad, es profesora en el Departamento de Lengua Española de la Universidad de Valladolid. Lleva veinte años investigando aspectos pragmáticos y discursivos de la interacción digital escrita, especialmente el correo electrónico. Es presidenta de la Red de Estudios sobre Comunicación Digital y dirige con Lucía Cantamutto la Revista REDD (Revista de Estudios del Discurso Digital).

https://orcid.org/0000-0002-4915-5260