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

Received: 08-26-2022 / Accepted: 12-22-2022

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

Corpus de interacciones digitales: sistematización de técnicas para recoger datos en 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, Spain (cristina.vela@uva.es)

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.

HOW TO CITE

Cantamutto, L., & Vela Delfa, C. (2023). Corpus of digital interactions: systematization of techniques to collect data on WhatsApp. *Cuadernos.info*, (54), 117-139. https://doi.org/10.7764/cdi.54.53165

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.*

RESUMO | A coleta de conjuntos de dados de interações reais é um passo inevitável em muitas investigações 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 arte sobre os conjuntos de dados de trocas digitais por WhatsApp e, em segundo lugar, sistematizar diferentes técnicas de coleta de amostras utilizadas em pesquisas anteriores. A metodologia utilizada é a análise de conteúdo de 100 artigos de pesquisa e teses 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 acesso online ao material 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 pesquisas, 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, dentre as quais se destaca a observação participante.

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

INTRODUCTION

The social function of language is the subject of a number of disciplines in which the study of real interactions is an unavoidable step (Ädel & Reppen, 2008). In order to study natural texts, it is necessary to access reference corpora or to create a dataset of real exchanges. Although there are numerous general or reference corpora of written and spoken Spanish in different Spanish-speaking countries, this is not the case with data collected in digital exchanges. As De Benito Moreno (2022) points out, only the *Corpus del Español: Web/Dialects* and the *EsTenTen* corpus are available, in addition to the CoDiCE database, which has enabled the exchange of language samples between researchers. Therefore, advances in the field of digital discourse studies are made on the basis of specialized corpora or datasets that are created ad hoc (Collins, 2021), often reduced and focused on a specific platform (de Benito Moreno, 2022).

Linguistic corpora are a large collection of data (texts) that meet certain linguistic criteria, are organized according to certain parameters (date, discursive genre, geographical origin, communicative situation, etc.), are preferably stored digitally and –currently– can be analyzed with a specific software (Molina Mejía, 2021). These qualitative differences allow corpora to be distinguished from other data sets, such as computerized archives/collections and electronic text libraries, compilations without linguistic criteria (Toruella & Llisterri, 1999). Rojo Sánchez (2015) states that corpus linguistics amounts to "linguistics based on corpus analysis" (p. 681), which allows it to be a methodological tool complementary to other linguistic studies and, in this case, to digital discourse analysis, whose methods of collecting linguistic data have found increasing editorial interest (Collins, 2021; Vásquez, 2022).

The novelty of the digital variable in the study of linguistic data, together with the constant changes in devices and applications, means that in digital communicative environments the difficulties that are always present in the collection of linguistic data are maximized. For this reason, a range of complementary methodological techniques are usually used or work with limited samples to obtain the multimodality features typical of digital discourse (Thurlow, 2018), combining resources of different types. It is striking that the techniques used to collect digital data have remained relatively stable from the first studies on chatrooms (Pihlaja, 2022) to the present day, and that, if anything, only the options that each application offers for accessing, storing or exporting this data change.

Within the repertoire of digital genres, one of the phenomena that poses the most methodological difficulties is that of private digital interactions that take place through the exchange of short texts (Cantamutto & Vela Delfa, 2020), what

Yus (2021) refers to as smartphone messaging. When linguists are interested in other genres of public digital discourse, such as social networks, there is a large amount of data available to them that is considered public. However, when the focus is on analyzing private digital genres, interesting methodological challenges become apparent that can be overcome with innovative proposals. In addition to privacy, the increasing multimodality of these environments can also pose difficulties.

Taking these limitations into account, the aim of this article is twofold: to provide a first state of the art on datasets of digital communicative exchanges via WhatsApp and to systematize different techniques for collecting WhatsApp interaction samples that have been used in previous research.

METHODOLOGY

Following the suggestions of Beißwenger and Storrer (2008) and Pano Alamán and Moya Muñoz (2016), we carried out a documentary content analysis of a corpus consisting of research articles, master's and doctoral theses and websites of research projects. For their selection, a systematic search was carried out in different academic portals based on the keywords: corpus and WhatsApp. Since we are interested in Spanish-language datasets, the search was carried out in the Dialnet portal (29 results) and then in other databases: DOAJ (24 results), Scielo (3 results) and Redalyc[1] (62 results). Finally, only the 50 most relevant results in other languages were collected in *Google Scholar.*

Based on the results, a manual screening was first performed to select a non-probabilistic sample of one hundred documents, a number that allowed a saturation of the analyzed categories. Subsequently, the information of the documents studied was transferred to a database, both in terms of the corpus analyzed and the documents referenced in the text. The data of the latter were completed after searching for the original references. Thus, a spreadsheet was created with information on the following variables: type of contact, amount of data (conversations, messages, words or tokens), number of participants, technology used, ethical guarantees (consent, anonymization), year of data collection, extent of use and type of conversation (complete or fragmentary). However, as Kreis (2022) notes, many studies do not describe in detail the dataset used or the techniques

[1] In this database, only the results of the WhatsApp search with a filter by discipline (language and literature) were selected. If both terms were used (also using Boolean operators), texts were found that contained one of the two terms and significantly increased the number (5844).

used to collect it. Some prototypical examples were selected from the complete database and are commented on in these pages.

RESULTS

Based on the review of some linguistic corpora of communicative exchanges in private digital environments, the main techniques for collecting data in WhatsApp are systematized.

Corpus of written digital interactions

If accessing and storing patterns of digital discourse languages is difficult in general, it is even more complicated to create a corpus of digital interactions in the private and intimate sphere, such as exchanges via instant messaging platforms.

Access to real interactions that take place via these applications is a challenge for linguistic research. These difficulties lead to a significant bias in the choice of the object of study, as much of the research on digital communication focuses on language samples collected in public social networks (such as Twitter or e-commerce platforms) or uses the web as a corpus (Collins, 2021; de Benito Moreno, 2022; González Fernández, 2017). Works that work with private messages also present a certain methodological weakness, as they often have to make do with language samples constructed for convenience and limited by significant shortcomings (generally due to sample size).

The negative consequences of this situation are twofold. On the one hand, public social networks do not always allow us to know precisely the sociocultural characteristics of speakers (de Benito Moreno & Estrada Arráez, 2018), which imposes limitations in terms of sociolinguistic or sociopragmatic variation, for example. In addition, social networks cannot capture some interactional phenomena that cannot be reproduced in the same way in conversations, which are much more fragmented. In this context, it is worth addressing the methodological difficulties of analyzing instant messaging, as it offers us the opportunity to study language use from a contextually dense perspective.

The datasets used in research on instant messaging are generally small samples of interactions in social networks of family and friends or through contacts with adolescents and young people in educational settings, as has been the case with other genres of digital discourse (Cantamutto & Vela Delfa, 2016; Pano Alamán & Moya Muñoz, 2015, 2016). The greatest lack is in Spanish, which is the third most used language in digital exchanges after English and Chinese. In other languages, we find different proposals that collect data on public or private interactions through messaging platforms (including chat rooms).

Below we discuss some of these proposals, which we have selected because they fulfill the main criteria that distinguish a corpus from a dataset, also called a corpus, that serves as the empirical basis for a research project.

In a CMC corpora compilation of prototypical Internet exchanges, Beißwenger and Storrer (2008) classify corpora into four types. The corpora compiled for general use are divided into raw data corpora (such as the Apache SpamAssasin Project with 6000 spam emails, plus three others that are no longer accessible) and annotated data corpora, of which only two are mentioned: The Dusseldorf CMC Corpus with examples of various synchronous and asynchronous genres and the Dortmund Chat Corpus, which consists of annotated chat sessions. The Dortmund Chat Corpus (Beißwenger et al., 2013a, 2013b) has been available in electronic form since 2005. This data set comprises 140,000 chats and around one million tokens in German. The data has been anonymized as part of various projects and is currently tagged with TEI tags. Most of the corpora examined are project-related (i.e., not generally used) and raw (unannotated) data (Beißwenger & Storrer, 2008). This trend has reversed slightly.

Natural language processing needs data sets to train its tools. This was the goal of the NPS Internet Chatroom Conversations Corpus (Forsythand, Lin & Martell, 2007). The corpus collected 10,567 entries/comments in English language chatrooms between October and November 2006, representing 45,068 tokens. This data has been used by researchers from various disciplines (e.g., Kim et al., 2021), demonstrating the need for samples of digital conversational interactions and the validity of this data despite its age.

As the use of WhatsApp became popular, various projects collected data on this application, but as we will see below, the amount of data differs greatly from the samples from the chat corpus. The architecture of the application did not favor the accessibility of the data for research purposes. While there were ways to export the conversations, contact with the participants was an unavoidable step to obtain these speech samples.

For example, the What's up, Switzerland? Corpus (Uerberwasser & Stark, 2017) collects 617 chats between 1538 German-, French-, Italian-, Romanian- and English-speaking participants from 2014, comprising about five and a half million tokens and 350,000 emojis. For two months (June and July 2014), Swiss citizens were asked to send their WhatsApp conversations as attachments via an email address provided by the project. Upon receipt, participants received an informed consent questionnaire in which additional socio-demographic data was collected. This corpus is anonymized and has search tools. Its direct predecessor is the corpus of the Sud4Science project[2], which focuses on SMS. The corpus What's up, Deutschland?[3] (Wyss, 2015 in Ayan, 2020) was also created.

An interesting aspect of the corpus of What's up, Switzerland? is that, due to the methodology used, it is possible to observe the discrepancy between the conversations donated and the consents received. The number of chats received was 967, but in the cases where the consent form was not completed, the messages were replaced with expressions such as redactedQ51tokens248characters (Ueberwasser & Stark, 2017: 108). Similarly, for the German language, for example, 93 chats with consent forms were received. However, as Ueberwasser & Stark (2017) illustrate using a table, the consent of all participants is only available for 44 conversations and the complete demographic information of the participants is only available for 25 chats. Following these suggestions, data from 218 WhatsApp conversations in the Verheijen WhatsApp corpus were collected in the Netherlands between 2012and 20144 . The corpus is anonymized and all participants gave their consent for their exchanges to be used for academic purposes. the age group of adolescents and young adults between the ages of 12 and 23 was prioritized. Prizes were raffled off among those who sent their conversations via a project email (as was done in the Sud4Science project).

Another precursor is data collected in public chats. The PoliWAM corpus retrieves data from 281 public WhatsApp groups with a total of 223,404 messages and 31,078 participants before, during and after the 2019 Indian general elections (Srivastava & Singh, 2020) and compiles data from 26 political parties. Part of it (3848 messages) is publicly available. In this sense, messages from 127 public groups in Brazil were collected based on the WebWhatsAppAP tool (Resende et al., 2018).

For the Spanish language, some corpus-based research should be highlighted. For example, Vázquez Cano and colleagues (2015) collected 417 WhatsApp conversations (101,401 words) through the voluntary participation of students from five secondary education centers in four provinces of Spain. Although the collection method is not specified, it is explicitly stated that real conversations on a topic were requested. The interest of this study is the lexicometric analysis of digital writing. Similarly, Pérez-Sabater (2015) analyzes linguistic variations from chats of adolescents and adults in three varieties (Spanish, Catalan and English) collected in collaboration with students of a master's course who created subcorpora of 500 words. The total number of words analyzed is 41,000.

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

[3] According to Verheijen and Stoop (2016), this corpus was hosted at http://www.whatsupdeutschland.de/ . However, at the last consultation (11 August 2022) it was not accessible.

[4] https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:112987 (Accessed August 11, 2022).

Bach and Costa Carrera (2020) also asked for the collaboration of university students in the data collection and conducted a survey among 200 students, of which they selected 23 whose mother tongue was Catalan or Spanish and Catalan. After contacting them by email and conducting a series of interviews, "the students [were] asked to send us fragments of their WhatsApp, freely chosen by them, in text format (.txt) by email" (Bach & Costa Carrera, 2020, p. 574).

In this sense, the sociolinguistic corpus of WhatsApp chats in Spanish for linguistic analysis in higher education (Dorantes et al., 2018) consists of conversations of students from the National Autonomous College of Mexico. The linguistic data, donated by randomly selected students on campus, were sent via email. The corpus includes 835 chats from 1325 participants. After the process of anonymization and removal of the messages from the platform, the corpus includes 66,465 messages and 756,066 tokens.

Finally, the CoDiCE database (Cantamutto, Vela Delfa and Boisselier, 2015) is a collaborative project to provide language samples from written digital interactions (email, SMS and WhatsApp) to academic researchers. This database is based on various corpora collected by researchers and contains around 14,000 SMS and emails and 3000 WhatsApp messages. The data has been anonymized and can be analyzed using database tools. Currently, this project is being expanded to include new functionalities and new data from WhatsApp interactions in Spanish from Buenos Aires and Spanish from Patagonia.

In this brief summary, we note a mixed situation in terms of the amount of data available for the different languages, as well as a large concentration of conversations collected in the adolescent and young adult age group (just as with the SMS data). In other cases, these are small datasets extracted from groups created for academic purposes but subsequently used as research material (e.g. Ayan, 2020; García Gómez, 2020).

the following section presents the methodological alternatives used in the corpora described above.

Tools and techniques for collecting instant messaging linguistic data

In line with the growing number of WhatsApp users, research into various pragmatic-discursive phenomena has increased significantly in recent years. Prominent researchers in the field of digital discourse have dealt with specific phenomena of this type of exchange. This applies, for example, to the classic text by Manuel Alcántara-Plá (2014) on WhatsApp conversation units, the proposal for discursive characterization by Calero Vaquera (2014), which places this type of text between Messenger and SMS, or the analysis of the values and functions of emoticons and emojis in Sampietro (2016) and in Cantamutto and Vela Delfa (2019), to name but a few.

All of these studies used different methods to create a small corpus or random sample. In this section, we present the main techniques, which are not mutually exclusive but, on the contrary, are often used in a complementary way. We briefly discuss the form of contact, the preservation of participants' identities and the storage of voice samples, and explain the methodological alternatives that were used during this short but intensive period in which WhatsApp was studied.

Export chat

The main tool for collecting data donated by speakers has changed since its introduction. In the early versions of the WhatsApp application, users could save conversations in a folder in the phone's memory or on an external card and extract them that way. Later, the option to export them via email as text in the body of the email or as a text file in attachment format (with the .txt extension) was enabled. In these cases, the multimodal information was not attached (figure 1). With this technique, for example, neither the photos nor the emojis were captured, but a document was accessed in which the conversation was presented in the order of the interventions as they were arranged in the account from which the conversation was shared (sometimes in ascending order and sometimes in reverse order). The transcription of the conversations performed automatically by the application results in a "visualization similar to that of a theatrical script or the traditional transcription of oral conversations" (Sampietro, 2016, p. 150), losing the "visual format of the application, as well as the multimedia attachments sent and most of the emoticons included in WhatsApp" (Bach & Costa Carreras, 2020, p. 574).

Gradually, and depending on the phone's operating system, the multimedia elements were restored during export: first emojis and later also images, audio and other non-verbal content. In the previous example, the photo sent is only labeled with the word "picture" (intervention 10 by "Martín") and the emoji is labeled with a □ (intervention 9 by "Usted5"). This technique was used by asking gatekeepers (usually students from middle or high school institutions or social networks of family and friends who serve as intermediaries) to export the conversations. This was done, for example, for the corpus by Dorantes and colleagues (2018) or by Maíz-Arévalo, 2018, who explicitly thanks his collaborators (Maíz-Arévalo, 2002).

[5] In 2012, when the conversations were exported, You and the name of the other participant appeared. In this case, it was decided to change the name of the participant Martín and keep the form You as the conversation was exported at that time.

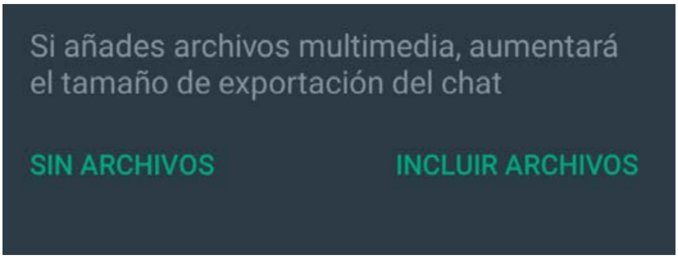
15:51, 2012/6/20 - Martin: I'm going to be picked up by a friend or girlfriend.
15:52, 2012/6/20 - *You: :D*
15:53, 2012/6/20 - You: The photo never arrived 🙁
15:56, 2012/6/2- - Martín: I resent it....
15:56, 2012/6/2- - Martín: It's a random photo
15:56, 2012/6/2- - Martín: Mine from today ...
15:58, 2012/6/2- - You: :D
15:59, 2012/6/2- - You: I wish it arrived.
15:59, 2012/6/2- - *You:* □
16:01, 2012/6/2- - *Martín: image*
16:01, 2012/6/2- - You: I got it.

Note: the conversation is translated.

Figure 1. Export of WhatsApp conversations (year 2012)

Source: Own elaboration.

In the case of the WhatsApp Corpus Verheijen, they note, among other things, that the project website contained many instructions for exporting conversations, depending on the device and operating system (Verheijen & Stoop, 2016, p. 251). After several updates to the application, the export tool works similarly on each phone, which favors this type of data collection. The user must access the drop-down menu at the top right of the three dots and first select More and then Export chat. Once this action has been performed, the following banner will appear (figure 2):



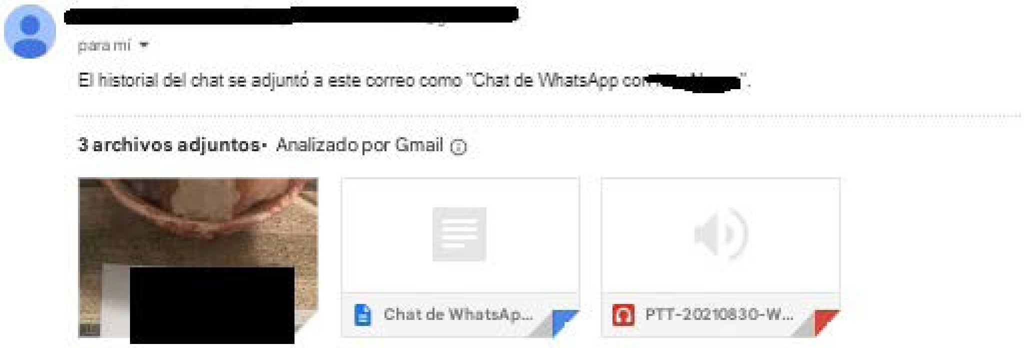
Si añades archivos multimedia, aumentará el tamaño de exportación del chat

Figure 2. Automatic message from WhatsApp application to export conversations with or without multimedia files

Source: WhatsApp.

If the second option is selected, a text file (with the extension .txt) is created, which is accompanied by the multimedia files in the conversation. The user can choose between several applications to send the history. In all cases, the generated file retains the relationship between the multimedia files, which are not embedded in the .txt document but can be recognized by their name (as in figure 4). In the first and second intervention, reference is made to two sent audio files that are not present in the exported conversation history but are present in the e-mail attachments (Figure 3).

This means that with the export tool it is possible, on the one hand, to keep a text file with the conversation structured by date and time of sending, in which the emojis are displayed in the corresponding message labeled “WhatsApp chat with + name of contact/group”. On the other hand, audio notes, photos, text files and stickers are attached but displayed in the body of the conversation, as shown in the following image (engagement 1 and 2 of figure 4).



[redacted]

para mí ▾

El historial del chat se adjuntó a este correo como "Chat de WhatsApp con [redacted]".

3 archivos adjuntos · Analizado por Gmail ⓘ

[redacted]

Chat de WhatsAp...

PTT-20210830-W...

Figure 3. Capture of conversation export email with the WhatsApp Export tool (year 2020)

Source: Own elaboration.

19/6/20 14:18 - Teacher: PTT-20200619-WA0080.opus (attached file)
20/6/20 18:08 - Student: PTT-20200620-WA0044.opus (attachment)
20/6/20 18:23 - Student: haha well thanks for the honesty.
20/6/20 18:23 - Teacher: no problem
6/20/20 18:23 - Student: I recommend you do 3 and 4 which are from unit 1.
20/6/20 18:24 - Student: Aah the 2 is a movie, isn’t it?
20/6/20 18:24 - Teacher: 2 is a school assignment that we were asked to do with that movie. If you can at some point, you do it.
20/6/20 18:24 - Teacher: yes, yes.
6/20/20 18:24 - Student: Well, I’ll start with 3 then.
20/6/20 18:24 - Student: Great! When you do it tell me if there is something you don’t understand.
2/9/21 15:50 - Teacher: https://forms.gle/xxxxxxxxxx
2/9/21 15:50 - Teacher: AUD-20210902-WA0068.opus (attachment)
2/9/21 15:50 - Student: They click on the link and it takes them straight 🙂.

Note: the conversation is translated.

Figure 4. Fragment of conversation exported through the WhatsApp tool (year 2020)

Source: Own elaboration.

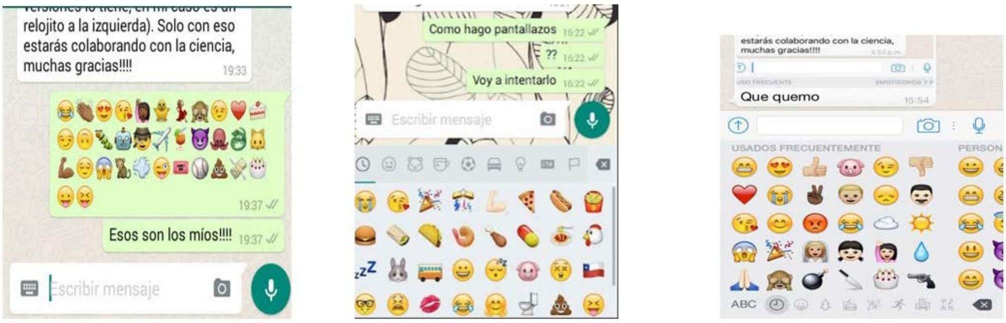
The preferred age group is adolescents and young people who, being in educational institutions, are accessible as voluntary participants; they are also usually found in the researcher's circle of family and friends, which will be discussed in the next section. However, this tendency to focus attention on this age group is not without its problems when it comes to lack of participation. For example, Sampietro (2016) notes that he chose to approach them in two ways. First, to her network of contacts by sending chats or snippets (both when exporting and through screenshots), and then to a group of students. However, only one student sent a conversation.

This aspect confronts us with one of the main problems of exporting conversations directly to institutional mailboxes (as in the case of Sampietro, 2016 and Verheijen & Stoop, 2016): the participants' lack of control over their data. In this regard, Sampietro (2016) notes that two participants in her data collection opted to send the conversation history to their personal email, delete a fragment they did not want to share, and then send part of the chat to the researcher.

Exporting the conversation to an email other than their own creates a file with the entire conversation history between two or more participants, which can be traced back to the first exchange between these people in the WhatsApp application on the device. For example, if it is an exchange between family and friends who communicate daily, a single conversation can last several months. In subsequent updates, however, WhatsApp has not only integrated the option to delete messages and send messages that delete themselves after a certain period of time, but also restores only part of the conversation with the export option. Due to file size restrictions, the most commonly used technology currently truncates the conversation according to the weight of the attachments. So if you send the history by e-mail without multimedia files, the entire conversation is exported, while including files results in a fragment that depends on the size of the files.

Captures

With the further development of cell phones, various functions have been integrated into the devices, such as screenshots: Photos of the cell phone screen, similar to the screen-printing function on computers. This technology has several advantages. On the one hand, it preserves a large amount of multimodal data (e.g., the image of the wallpaper used by the interlocutors or the structure of the conversation resulting from the use of the Reply or React with emojis functions). On the other hand, it is a simple and quick method for the participants, as they have greater control over the data, they exchange by selecting the statements they send. Although the data sent is not altered, participants protect their image by selecting fragments of the conversation or sending comments afterwards reflecting on their use (figure 5).


relojito a la izquierda). Solo con eso estarás colaborando con la ciencia, muchas gracias!!!! 19:33

19:37

Esos son los míos!!!! 19:37

Como hago pantallazos 16:22

?? 16:22

Voy a intentarlo 16:22

estarás colaborando con la ciencia, muchas gracias!!!!

Que quemo 15:54

Figure 5. Examples of screenshots collected between 2015-2016 (data used in Cantamutto & Vela Delfa, 2019).

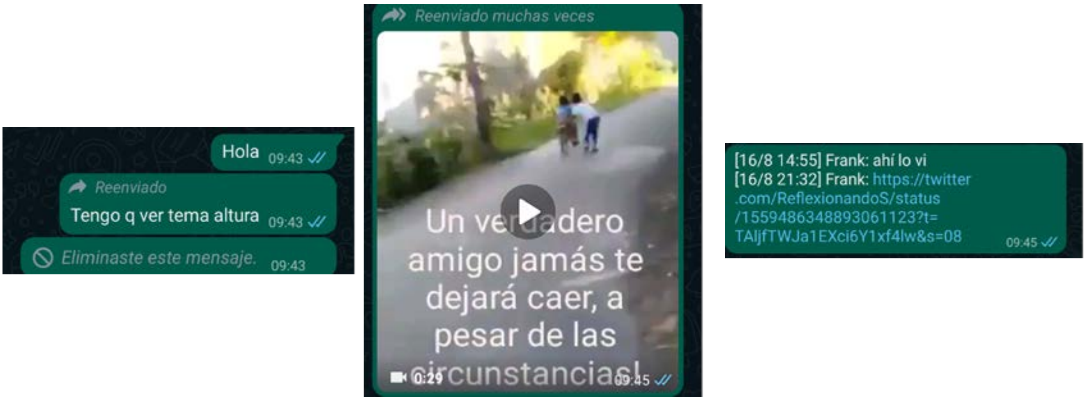
Source: Own elaboration.

In a previous study (Cantamutto & Vela Delfa, 2019), we obtained a large number of screenshots of recently used emojis using the snowballing technique. In the first data collection (2015-2016), participants asked how they should go about capturing them. Some who did not know how to do it copied the emojis that appeared in the menu.

In other cases, once we found out which phone model and brand the employee had, it was possible to give him/her instructions on how to take the screenshot. The following figure shows the three cases described above: the user transcribing the frequent emojis; the user asking how to take the screenshot; and the user thinking metalinguistically about his/her emojis.

To summarize, each methodological decision has an impact on the data obtained. Opting for captures favors less manipulation of the speech samples by the contributor (and this was the interest pursued by asking about frequent emojis). Although they have more control - they can select and monitor what they send –they do not alter the exchange. However, the use of this type of image has consequences for subsequent processing. Firstly, it is necessary to manually transcribe the conversations. In this way, the data is manipulated and unintended changes can occur. Secondly, a static image of the videos is accessed and it is also not possible to play back the audio. This makes it necessary to combine the sending of screenshots with the export of the call history or forwarding.

However, the screenshots help to better understand the development of the conversation, as the paratextual markers of the WhatsApp conversation structure are lost when the conversation is exported, as with the forwards: Reply, forwarded, forwarded multiple times and the reactions to the messages (figure 6).



Hola 09:43

Reenviado

Tengo q ver tema altura 09:43

Eliminaste este mensaje. 09:43

Reenviado muchas veces

Un verdadero amigo jamás te dejará caer, a pesar de las circunstancias!

0:29 09:45

[16/8 14:55] Frank: ahí lo vi
[16/8 21:32] Frank: https://twitter.com/ReflexionandoS/status/1559486348893061123?t=TAljfTWJa1EXci6Y1xf4lw&s=08
09:45

1/8/22 09:43 - A: Hello

1/8/22 09:43 - A: I have to see the height issue

1/8/22 09:43 - A: You deleted this message.

1/8/22 09:45 - A: VID-20220720-WA0038.mp4 (file attachment)

1/8/22 09:45 - A: [16/8 14:55] Frank: I saw it

[16/8 21:32] Frank: https://twitter.com/xxxxxx

Note: The messages are illustrative and do not correspond to actual exchanges.

Figura 6. Differences in the visualization of messages extracted with captures or by exporting chat

Source: Own elaboration.

Forwarding and copy/paste

A third way of asking participants to donate conversations is to ask them to forward the various contributions to a telephone number. This alternative, which has been around since the beginning of the application, made it possible to send fragments of WhatsApp conversations via SMS. Today, you can only forward them to WhatsApp itself or use the copy/paste function to export them to another application.

This practice is widespread among users, apart from the possibilities it offers the researcher. For this reason, in successive updates, the tool has introduced new functions for forwarding messages with paratextual information, making it possible to determine whether the statement was written in another conversation. A forwarded message usually appears with a legend at the top indicating intertextuality (forwarded), or with didactic markers indicating the date, time and sender of the message when multiple messages are copied and pasted into another chat (Figure 6).

In projects with a WhatsApp account, you can use this technique to ask collaborators to share fragments of a conversation, e.g., on a topic or messages that serve as an example (the use of a specific emoji or expression). This technique was used for the study on Voseo in Bolivia. They contacted 80 people by email (stratified by age group) and asked them to forward an audio: "Interested parties were asked to send their own recordings of messages addressed to one of their parents" (Castedo et al., 2022, p. 398).

This technique is productive in certain applications (e.g., commercial) where it is possible to replace the speaking parts from the broadcast messages. As can be seen, it is a supplement to other alternatives for data collection, as a lot of information about the structure of the conversation is lost.

Participant observer

This option differs from the previous ones, as it is no longer the employee who takes the speech samples, but the researcher who participates in the conversation. This participation can be done from a blind spot, through participant observation in WhatsApp groups or as a participant observer (Vela Delfa & Cantamutto, 2016). This technique has been successfully used in recent research, which includes observation in WhatsApp groups alongside other ethnographic techniques.

This is the case, for example, of Lucía Godoy's (2021) work on reading practices with digital technologies in the teaching of language and literature in secondary education. In addition to participant observation in face-to-face lessons, the researcher took part in school WhatsApp groups. Godoy (2021) points out that this technique "makes it possible to collect real data that emerge in authentic digital contexts (Hine, 2000, 2015) without having to elicit them, limiting the degree of invasion into the space of participation and looking deeply and broadly at the contexts in which these data emerge, without relying on the transmission of complementary information by speakers (...)" (pp. 123-124). This is a case of achieving observation from a blind spot (Vela Delfa & Cantamutto, 2016).

Another precursor is the study of the conversational markers ahre and tipo (de Luca, 2021). Complementing other ethnographic techniques, part of the data for this research was collected through the creation of WhatsApp groups in which students sent memes with these markers. The researcher was part of a group that was created ad hoc and in which participants could share and reflect on their own use. In this way, participant observation is combined with the forwarding of messages.

On the other hand, interactions are used in which the researcher plays a dual role: interlocutor and observer. This corresponds to a very common technique: sending samples through the network of family and friends. For example, Sampietro (2016, p. 149) justifies the use of this technique to collect his research data with the methodology of Tannen (1984), who recorded a dinner party of friends in order to analyze it afterwards. That is, when asking family and friends for consent to donate conversations, what often happens is that the researcher's own interventions are analyzed.

	Export chat	Captures	Resend	Participant observer
Plain text available	Yes	No	Yes	Yes
Multimedia files: photos	Yes	Depends	Yes	Yes
Multimedia files: audios	Yes	No	Yes	Yes
Multimedia files: videos	Yes	No	Yes	Yes
Graphics: emojis	Yes	Yes	Yes	Yes
Graphics: stickers	Yes	Depends	Yes	Yes
Paratextual marks: message deleted	Yes	Yes	No	Yes
Paratextual marks: forwarded/forwarded many times	No	Yes	Depends	Yes
Paratextual marks: reply to message	No	Yes	No	Yes
Paratextual marks: reactions	No	Yes	No	Yes

Table 1. Summary of techniques to extract linguistic data from WhatsApp Source:

Own elaboration.

As can be seen, there are two options when collecting data using this technique, depending on the role the researcher takes: participant or participant observer. The first alternative is possible in WhatsApp groups, while the second involves the exchange of data in which the researcher is one of the interlocutors. In WhatsApp, the participant-observer technique is extremely productive in dyadic exchanges and in group conversations where the observer paradox can be avoided.

Table 1 summarizes the possibilities offered by each of the data collection techniques. Their combination leads to the creation of corpora rich in linguistic and multimodal data, more information on sociolinguistic profiles, and obtaining informed consent from participants.

DISCUSSION

The primary purpose of this article is to contribute to the methodological considerations required to develop an in-depth study of language samples of digital interactions. The review of the datasets used by previous research reveals a gap in the compilation of broad, systematic and general reference corpora on digital interactions in Spanish. Likewise, these data are missing in Spanish reference corpora. However, the growing interest in digital conversations in the private sphere has led to various studies on WhatsApp that have created samples based on a combination of techniques to overcome the existing methodological difficulties.

Secondly, we have presented the main techniques for collecting interaction data via WhatsApp. In this regard, the technique of participant observers stands out as the one that allows us to obtain the largest amount of linguistic and multimodal data, as well as information on sociolinguistic profiles and on the communicative situations in which these exchanges are inserted. This type of material can be complemented by other techniques for collecting fragments of conversations, such as screen capturing and the use of the forward/copy and paste option. the alternative most commonly used in previous research –the export tool– also offers several advantages for data processing as they are in text format. This data can be analyzed with linguistic analysis tools (such as AntConc) in a very simple way.

Due to all these difficulties, several WhatsApp corpus projects have proposed the combination of techniques as well as the inclusion of other tools to capture speakers' perceptions of their language use. Among these complementary techniques to collect perceptual data, we reemphasize the importance of avoiding mediation by other applications. The use of WhatsApp in interviews, surveys or tests on social habits is very productive to study precisely the use of this application.

The dynamics of WhatsApp on the one hand and the changes in user habits on the other emphasize the need to collect language samples to account for (micro-) diachronic changes in discursive practices. In the present description, an attempt has been made to consider the current state of the art of the alternatives offered by WhatsApp for the collection of linguistic data. Future work will address populations that were not considered in the research agenda (e.g., adults), as well as different ways of ethically protecting participants.

FUNDING

Diseño e implementación de un corpus sobre comunicación digital del español bonaerense y de la Patagonia (Design and implementation of a corpus on digital communication of Buenos Aires and Patagonian Spanish). PICT-2019-02093PRÉSTAMO BID, National Agency for the Promotion of Research, Technology Development, and Innovation (Argentina). Lead researcher: Lucía Cantamutto

REFERENCES

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.

ABOUT THE AUTHORS

LUCÍA CANTAMUTTO, Ph.D. in Literature from the Universidad Nacional del Sur (Argentina). She is currently a CONICET researcher at the Interdisciplinary Center for Studies on Rights, Inclusion and Society at the Universidad Nacional de Río Negro. Her research focuses on digital communication. In 2015, together with Cristina Vela Delfa and Leandro Boisselier, she created the database CoDiCE (comunicación digital: corpus del español). She is vice-president of the Digital Communication Studies Network.

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

CRISTINA VELA DELFA, Ph.D. in Linguistics from the Universidad Complutense de Madrid (Spain). She is currently a professor at the Department of Spanish Language at the College of Valladolid. For 20 years she has been researching pragmatic and discursive aspects of written digital interaction, especially email. She is president of the Red de Estudios sobre Comunicación Digital and co-edits the REDD Journal (Revista de Estudios del Discurso Digital) with Lucía Cantamutto.

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