

# Sociolinguistic Corpus of WhatsApp Chats in Spanish among College Students - Data Paper

Alejandro Dorantes

Gerardo Sierra

Yamín Donohue Pérez

Gemma Bel-Enguix

Mónica Jasso Rosales

Universidad Nacional Autónoma de México

Grupo de Ingeniería Lingüística

{MDorantesCR,GSierraM,TDonohueP,GBelE,MJassoR}@iingen.unam.mx

## Abstract

The aim of this paper is to introduce the Sociolinguistic Corpus of WhatsApp Chats in Spanish among College Students, a corpus of raw data for general use, which was collected in Mexico City in the second half of 2017. This with the purpose of offering data for the study of the singularities of language and interactions via Instant Messaging (IM) among bachelors. This article consists of an overview of both the corpus's content and demographic meta-data. Furthermore, it presents the current research being conducted with it—namely parenthetical expressions, orality traits, and code-switching. This work also includes a brief outline of similar corpora and recent studies in the field of IM, which shows the pertinence of the corpus and serves as a guideline for possible research.

## 1 Introduction

As digital communication technologies grow and spread, computer mediated communication (CMC) (Baron, 1984)—which includes (IM)—changes and becomes a very distinct sort of interaction. According to Álvarez (2011), a new discourse level emerges through such interaction—one that makes the distinction between writing and speaking less and less clear. This discourse style has been previously called both spoken writing (Blanco Rodríguez, 2002) and oralized text (Yus Ramos, 2010).

In order to study such a particular register, it is necessary to gather a robust corpus. The Sociolinguistic Corpus of WhatsApp Chats in Spanish for College Speech Analysis intends to be a resource

that allows researchers to explore and characterize conversations held by college students and their peers, or other kind of participants, via the IM application known as WhatsApp (hereafter WA). This corpus is limited to bachelors studying at Ciudad Universitaria (commonly known as C.U.), the main campus of the National Autonomous University of Mexico (UNAM). The reason for choosing bachelors is because, in Mexico, 94.1% of the population with an undergraduate degree or a higher educational level uses the Internet for communication purposes, this mainly via IM, and generally they access the net on a smartphone. Furthermore, most of IM users are 12 to 34 years old, which is the age group the majority of college students belong to (INEGI, 2016).

## 2 State of the Art

### 2.1 Similar Corpora

Prior to the collection of WA corpora, other databases were created to allow the study of CMC. Examples of said data are the NPS Internet Chatroom Conversations Corpus (Forsyth et al., 2010), an annotated corpus of interactions in English in diverse chatrooms, and the Dortmunder Chat Corpus (Beisswenger, 2013), a robust, annotated corpus in German divided in 4 subcorpora, based on the topic of the chats (free time, learning contexts, consultations, and media). In addition to these corpora, it is worth mentioning the NUS SMS Corpus (Chen and Kan, 2013) which comprises 71,000 messages, both in English and Chinese. Even though the SMS is not an internet-mediated mean of communication, it can be compared to interactions via WA.

Although the study of WA chats is a relatively novel research field, there are several corpora specialized mostly on them. One of the most impor-

tant projects is the one conducted by researchers of the Universities of Zurich, Bern, Neuchâtel and Leipzig. The *What's up, Switzerland?* corpus (Stark et al., 2014-) has as main aim the characterization of WA chats and the comparison of these to SMS. It has 617 chats written by 1,538 participants. Since just 945 of them consented to have their chats used, the total number of messages available for linguistic research is 763,650 comprising 5,543,692 tokens. Only 426 participants shared further demographic information (Überwasser and Stark, 2017). Given the fact that Switzerland is a multilingual country, 46% of the corpus is in German, 34% in French, 14% in Italian, 3% in Romansh and 3% in English. The sociodemographic information saved as metadata comprises age, gender, educational level, and place of residence divided in 9 regions. So far, the publications derived from this project focus not only on the different levels of language, but also the role of complementary items in conversation, such as images, acronyms, emojis, emoticons, and combination or modification of characters.

Verheijen and Stoop (2016) compiled a corpus which is a part of the SoNaR project (STEVIN *Nederlandstalig Referentiecorpus*) of posts and WA chats in Dutch. The corpus has 332,657 words in 15 chats donated by 34 informants. Their metadata encompasses informants' name, birth place and date, age, gender, educational level, and place in which the chats were sent. This corpus was used as one of the bases for a research where WA and other written forms were compared (Verheijen, 2017).

Hilte et al. (2017) compiled a corpus of chats between Flemish teenagers aged 13-20 taken from Facebook Messenger, WA, and iMessage. This, with the purpose of identifying the impact of social variables—namely age, gender and education—in teenagers' non-standard use of language in CMC.

In addition to these, an ongoing project is that of MoCoDa2 conducted by Beisswenger et al. (2017), which is a continuation of the preceding corpus MoCoDa, and has put together 2,198 interactions with 19,161 user posts.

Nevertheless, all of these authors did not define what they conceive as a chat. In order to avoid any misconception, in the making of this corpus we consider a chat an exchange between two users regardless of length or date. Meaning that it does not

matter when the conversation started, but rather the wholeness of the txt file.

Although there are, indeed, corpora of WA chats in Spanish, they are not for general use, but project-related. Besides, they are not as robust as the aforementioned. Said corpora are presented in the following section.

## 2.2 Research on WhatsApp Chats

Because of their peculiarities, virtual interactions through diverse platforms like WA, WeChat, Facebook Messenger, and so forth have drawn the attention of linguists. Some of the previous studies that have been conducted using similar corpora are varied in the topics they approach. Some of the aspects of language that can be studied with sociolinguistic corpora like ours are discourse units and phenomena such as turns and turntaking, speech acts, and interactions (Bani-Khair et al., 2016; Martín Gascueña, 2016; Alcántara Plá, 2014; García Arriola, 2014); linguistic variation from a diaphasic, diastratic or diatopic point of view (Pérez Sabater, 2015; Sánchez-Moya and Cruz-Moya, 2015); multimodal communication (verbal, iconic or hybrid) (Sánchez-Moya and Cruz-Moya, 2015); use of orthotypographic elements (Vázquez-Cano et al., 2015); the role of the so-called emojis in communication (Sampietro, 2016b,a; Dürscheid and Siever, 2017); and even the didactic use of IM for digital and linguistic competence (Gómez del Castillo, 2017).

Another phenomenon that has proved itself to be interesting is code-switching in IM (Nurhamidah, 2017; Zaehres, 2016; Zagoricnik, 2014). As Al-Emran and Al-Qaysi (2013) have stated “WhatsApp is found to be the most social networking App used for code-switching by both students and educators”; which is why authors like Elsayed (2014) have focused on such population.

## 3 Methodology

### 3.1 Sociolinguistic Variables in the Corpus

Considering this is a sociolinguistic corpus, several sociodemographic variables were defined as metadata and divided into two groups:

- (a) Balance axes, which are the two variables that help to keep the balance and representativeness of the corpus:

- Sex: male or female <sup>1</sup>

<sup>1</sup>We chose sex over gender because it is the sociodemo-

- Faculty students are enrolled in: Architecture, Sciences, Political and Social Sciences, Accounting and Administration, Law, Economy, Philosophy and Literature, Engineering, Medicine, Veterinary Medicine, Odontology, Psychology, Chemistry, and the National School of Social Work.

Our goal was to collect at least 1% of the campus's population maintaining the same proportion of men and women as in each faculty.

- (b) Post-stratification criteria, whose relevance will depend on the type of study conducted with this corpus as main data: age, (open answer), sexual orientation, (heterosexual, bisexual or homosexual), birthplace, (any state in Mexico), current place of residence, (post code), other languages, (any indigenous language spoken in Mexico or any language taught at UNAM), education level, (no formal education, elementary school, middle school, high school, bachelor's degree, master's degree, doctorate), major, (any undergraduate program offered at the Ciudad Universitaria campus), occupation, (student, working student, worker, unemployed or retired), profession (open answer), and kinship or type of relationship between speakers.

In overall, our corpus has 12 sociolinguistic variables that contribute to a large degree to the characterization and study of language in IM among youngsters. Furthermore, this allows our data to become a subcorpus of a much larger one in the future.

### 3.2 Data Collection

After establishing the sociodemographic metadata to be collected along with WA chats, the team proceeded to gather the data. In order to ease the data processing we collected chats with two participants only. All chats were donated as text files sent directly from the donors' devices, while metadata was collected manually. At the initial stage, the chats were collected using the directed sampling method. A team approached random students on campus explaining the project to them and inviting them to collaborate donating one or more WA chats. Those who consented to share graphic variable used by UNAM in its statistics.

their chats -the donors- sent them via email to an institutional address, then were asked to answer a survey so the team could gather their and their interlocutor's sociodemographic information. After that, the information provided was entered into a spreadsheet along with a code that made it possible to link it to the corresponding text file. It is worth mentioning that the same metadata was collected with both methods.

### 3.3 Data Processing

The processing of data was done in two different stages. First, by means of a Python script, the collected data was saved into a spreadsheet. In the same stage, it was organized in JSON format and sent to the database as a document file. Second, a program allowed the users anonymity by changing their names in every chat to USER1 and USER2, and by deleting sensitive information —such as names, addresses, emails, phone numbers, bank accounts, and so forth.

Currently, queries can be done with both with Python scripts and MongoDB. Said tools permit the filtering of results depending on the metadata, allowing also the possibility of selecting relevant sociological variables and determining their ranges. In the future, we will develop an interface that makes the access and consultations to the database possible.

## 4 The Corpus

Although the corpus is still being processed, it has reached a mature stage which allows us to offer a general panorama of its content and demographics. The following figures represent the corpus state by March 2018. Should some changes be made, the final figures will be presented in future publications.

### 4.1 Content

Nowadays, we have 835 chats with 1,325 informants. After deleting dates, user names and all messages generated automatically by the app, we got 66,465 messages, 756,066 tokens and 45,497 types available for linguistic research.

Despite the fact that the vast majority of our informants are Mexican native Spanish speakers, texts in some other languages were found as well. Most of the messages in a language other than Spanish were written in English, however there are also texts in French, Japanese, Italian, German,

Korean, Greek and Chinese.

Other than that, we were also able to pinpoint which are the most frequently used lexical words among the informants. Students seem to be keen on using the ones displayed in Table 1.

Lexical Words		
bien	“good/well”	608,401
bueno	“good/well”	45,700
amor	“love”	44,900
bebé	“baby”	40,302
solo	“just/alone”	39,563

Table 1: Most frequent lexical words.

As it was previously mentioned, communication via IM shares several features with oral communication. However, since it lacks physical co-presence, it is necessary to develop some compensation strategies. Which is why emojis and emoticons are so widespread. The most frequent of these icons found in the corpus are shown in Table 2.






Emojis		Emoticons	
	2,221	xd	1,516
	1,015	:V	489
	445	:(	453
	435	:3	235
	249	:)	117

Table 2: Most frequent emojis and emoticons.

## 4.2 Demographics

As stated above, our corpus was built with the collaboration of 1,325 informants (51% women and 49% men), between ages 14 and 60, born in 23 of the 32 states in Mexico. Such a wide range of informants’ age is due to the fact that some donors shared chats, held not with peers, but with people in their families, coworkers, or friends. Of all informants, 84.9% are undergraduates studying at C.U. Out of these students, 51.2% are women and 48.8% are men.

Henceforth, all figures refer only to bachelor informants. 80.7% of bachelors in the corpus were born in Mexico City, while 11.7% were born in Estado de México, the biggest state surrounding the capital. The rest were born in 20 other states —particularly Hidalgo, Guerrero and Michoacán.

Our corpus have also informants born in Chile (2), Colombia (1), The United States (1), and 3 that did not report their birthplace. 77.4% of our informants live in the city, while 19.4% live in Estado de México. The remaining 3.2% did not state their post code.

As to sexual orientation, 88.9% of students in the corpus declared themselves as heterosexual, 5.5 % bisexual and 5.4 % homosexual. Just .2% chose not to share such information.

Although the purpose of this corpus is to collect data from Mexican native Spanish speakers, some informants donated chats with people from other countries: Chile, Colombia, Costa Rica, Italy, Lebanon, and the United States, to name a few. All of these conversations were conducted mostly in Spanish. As second language, informants claimed to speak Arabian, Bulgarian, Chinese, English, French, German, modern Greek, Italian, Japanese, Korean, Nahuatl, Portuguese, Russian, or Swedish.

The students who donated their chats and their interlocutors belong to different faculties. The following table presents both the faculty roster at C.U. and the number of informants by sex.

Faculty	Male	Female	Total
Engineering	144	33	177
Accounting and Administration	71	52	123
Sciences	40	60	100
Political and Social Sciences	37	56	93
Chemistry	53	40	93
Philosophy and Literature	33	54	87
Medicine	27	50	77
Law	26	50	76
Architecture	32	42	74
Economy	39	22	61
Psychology	10	41	51
Veterinary	16	31	47
Medicine			
Odontology	15	25	40
National School of Social Work	7	20	27
<b>Total</b>	<b>550</b>	<b>576</b>	<b>1,126</b>

Table 3: Informants by faculty.

## 5 Current Research

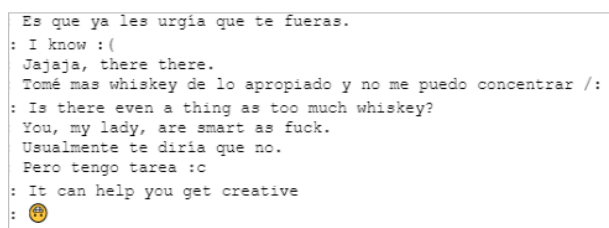
At the time of the writing, there are three lines of research in the study of our corpus. One of them is parenthetical expressions that can work as repairs, instructions for interpretation, onomatopoeic expressions, surrogate prosodic cues to indicate how an utterance should be read, or surrogate proxemic cues such as emotes —sentences that indicate imaginary actions taking place at the moment of texting (Christopherson, 2010).

1. \* se pone a llorar \*  
“Starts crying.”
2. (léase como si fuera eco)  
“Read as if it were an echo.”

Another research line is the study of oral (phonic) traits in WA chats, for instance: the emulation of children’s speech, repetition of vowels to indicate elongation of sounds, omission of letters to indicate consonant and vowel reduction, haplology, use of upper case for emphasis (volume), etc. (Yus Ramos, 2001)

3. kesestoooo  
Standard Spanish: ¿Qué es esto?  
“What is this?”

There is also the quantitative approach to code-switching from a sociolinguistic perspective, followed by a qualitative study of the forms and functions of it (Elsayed, 2014).



```
Es que ya les urgía que te fueras.
: I know :(
Uajaja, there there.
Tomé mas whiskey de lo apropiado y no me puedo concentrar /:
: Is there even a thing as too much whiskey?
You, my lady, are smart as fuck.
Usualmente te diría que no.
Pero tengo tarea :c
: It can help you get creative
: 🤔
```

Figure 1: Code-switching among bachelors.

## 6 Conclusion and Future Work

We presented a corpus that will make the study of language usage by college students via an Instant Messaging application possible. Its metadata will allow research, not only on mere linguistic phenomena, but also the establishment of correlation between these and sociodemographic variables. Some of the phenomena that can be studied in interactions, such as the ones via IM, are phonic

traits, parenthetical expressions, code-switching, turn-taking, speech acts, linguistic variation, and usage of emojis and emoticons.

Since the processing of data is still a work in progress. As next step, we plan to perform an evaluation of the anonymization process.

The objective of this corpus is to be used by both scholars and students in our group for the research of the aforementioned phenomena and others, and it is our intention to make it available upon request for others, with academic purposes only.

## Acknowledgments

This work was supported by CONACYT project 002225 (2017) and CONACYT Redes 281795, as well as two DGAPA projects: IA400117 (2018) and IN403016 (2018). We would also like to show our gratitude to students Ana Laura del Prado Mota, Mayra Paulina Díaz Rojas, and Paola Sánchez González for their participation in the collection and processing of data.

## References

- Mostafa Al-Emran and Noor Al-Qaysi. 2013. *Code-switching usage in social media: A case study from oman*. *International Journal of Information Technology and Language Studies(IJITLS)*, 1(1):25–38.
- Manuel Alcántara Plá. 2014. *Las unidades discursivas en los mensajes instantáneos de wasap*. *Estudios de Lingüística del Español*, 35:223–242.
- Baker Bani-Khair, Nisreen Al-Khawaldeh, Bassil Mashaqba, and Anas Huneety. 2016. *A corpus-based discourse analysis study of whatsapp messenger’s semantic notifications*. *International Journal of Applied Linguistics & English Literature*, 5(6):158–165.
- Naomi Baron. 1984. Computer-mediated communication as a force in language change. *Visible Language*, 18(2):118–141.
- Michael Beisswenger. 2013. Das dortmunder chat-korpus. *Zeitschrift für germanistische Linguistik*.
- Michael Beisswenger, Marcel Fladrich, Wolfgang Imo, and Evelyn Ziegler. 2017. *Mocoda 2: Creating a database and web frontend for the repeated collection of mobile communication (whatsapp, sms & co.)*. In *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora17)*, pages 11 – 15. Eurac Research.
- Maria José Blanco Rodríguez. 2002. *El chat: la conversación escrita*. *ELUA. Estudios de Lingüística*, (16):43–87.



- María-Teresa Gómez del Castillo. 2017. *Whatsapp use for communication among graduates*. *REICE. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, 15(4):51–65.
- Tao Chen and Min-Yen Kan. 2013. *Creating a live, public short message service corpus: the nus sms corpus*. *Language Resources and Evaluation*, 47(2):299–335.
- Laura Christopherson. 2010. *What are people really saying in world of warcraft chat?* In *Proceedings of the ASIST Annual Meeting*, volume 47. Learned Information.
- Christa Dürscheid and Christina Siever. 2017. *Beyond the alphabet – communication with emojis*. pages 1–14.
- Ahmed Samir Elsayed. 2014. *Code switching in whatsapp messages among kuwaiti high school students*.
- Eric Forsyth, Jane Lin, and Craig Martell. 2010. *Nps internet chatroom conversations, release 1.0 ldc2010t05*.
- Manuel García Arriola. 2014. *Análisis de un corpus de conversaciones en whatsapp. aplicación del sistema de unidades conversacionales propuesto por el grupo val.es.co*.
- Lisa Hilde, Reinhold Vandekerckhove, and Walter Daelemans. 2017. *Modeling non-standard language use in adolescents' cmc: The impact and interaction of age, gender and education*. In *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora17)*, page 71. Eurac Research.
- INEGI. 2016. *Encuesta nacional sobre disponibilidad y uso de tecnologías de la información en los hogares, 2016*.
- Rosa Martín Gascuña. 2016. *La conversación guasap*. *Sociocultural Pragmatics*, 4(1):108–134.
- Idha Nurhamidah. 2017. *Code-switching in whatsapp-exchanges: cultural or language barrier?* In *Proceedings Education and Language International Conference*, pages 409–416. Center for International Language Development of Unissula.
- Carmen Pérez Sabater. 2015. *Discovering language variation in whatsapp text interactions*. *Onomázein*, 31:113–126.
- Agnese Sampietro. 2016a. *Emoticonos y emojis: Análisis de su historia, difusión y uso en la comunicación digital actual*. Ph.D. thesis, Universitat de València, La Coruña, Spain.
- Agnese Sampietro. 2016b. *Exploring the punctuating effect of emoji in spanish whatsapp chats*. *Lenguas Modernas*, 47:91–113.
- Elisabeth Stark, Simone Ueberwasser, and Anne Göhring. 2014-. *Corpus "what's up, switzerland?"*.
- Alfonso Sánchez-Moya and Olga Cruz-Moya. 2015. *Whatsapp, textese, and moral panics: discourse features and habits across two generations*. *Procedia - Social and Behavioral Sciences*, 173:300–206.
- Lieke Verheijen. 2017. *WhatsApp with social media slang?: Youth language use in Dutch written computer-mediated communication*. Ljubljana University Press, Ljubljana, Slovenia.
- Lieke Verheijen and Wessel Stoop. 2016. *Collecting facebook posts and whatsapp chats: Corpus compilation of private social media messages*. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9924, pages 249–258.
- Esteban Vázquez-Cano, Andrés Santiago-Mengual, and Rosabel Roig-Vila. 2015. *Análisis lexicométrico de la especificidad de la escritura digital del adolescente en whatsapp*. *RLA. Revista de Lingüística Teórica y Aplicada*, 53(1):83–105.
- Francisco Yus Ramos. 2001. *Ciberpragmática. El uso del lenguaje en Internet*. Ariel Lingüística. Ariel, Barcelona, Spain.
- Francisco Yus Ramos. 2010. *Ciberpragmática 2.0: nuevos usos del lenguaje en Internet*. Ariel letras. Ariel, Barcelona, Spain.
- Frederic Zaehres. 2016. *A case study of code-switching in multilingual namibian keyboard-to-screen communication*. *10plus1: Living Linguistics*.
- Jelena Zagoricnik. 2014. *Serbisch-schweizerdeutsches code-switsching in der whatsapp-kommunikation*.
- Isabel Álvarez. 2011. *El ciberespañol: características del español usado en internet*. In *Selected Proceedings of the 13th Hispanic Linguistics Symposium*, pages 1–11. Cascadilla Proceedings Project.
- Simone Überwasser and Elisabeth Stark. 2017. *What's up, switzerland? a corpus-based research project in a multilingual country*. *Linguistik Online*, 84(5).