Towards the WhAP Corpus: A Resource for the Study of Italian on WhatsApp

Ilaria Fiorentini, Marco Forlano, Nicholas Nese
University of Pavia, University of Bergamo/Pavia, University of Milan
ilaria.fiorentini@unipv, marco.forlano@unibg.it, nicholas.nese@unimi.it

Abstract

Over the past two decades, the rise of new technologies and social networks has significantly shaped written language, imbuing it with characteristics akin to the spoken language. This study reports on the ongoing initiative to build the WhAP corpus, a resource featuring WhatsApp conversations in Italian, encompassing both written and spoken messages and totaling at present more than 400.000 tokens, 89 conversations, and 194 participants from diverse age groups and geographical regions of Italy. More specifically, this paper focuses on the practical steps involved in the construction of the resource. Once publicly accessible, the WhAP Corpus will enable in-depth linguistic research on the language used on WhatsApp, which shows unique features such as the blending of written and spoken elements.

Keywords: sociolinguistics, pragmatics, corpora

1. Introduction

Over the last two decades, the rise of new technologies and social networks has significantly shaped the evolution of written language, imbuing it with traits similar to those typically found in speech. This transformation is particularly evident in the case of typed messages within chat rooms, which differ markedly from more traditional written forms due to a more informal language use. Crucially, the recent spread of voice messages on instant messaging systems such as WhatsApp (from 2013) or Telegram (from 2020) has led to a "return" to speech, thereby introducing a new dimension within written chats, that has hitherto received limited academic attention. Against this background, our paper introduces an ongoing initiative aimed at constructing the WhAP Corpus, a resource consisting of WhatsApp conversations in the Italian language. The main motivation behind this endeavor is to address a gap in terms of attention towards the use of the Italian language within this communication channel, which has increasingly replaced traditional text messaging on cell phones and web-based chats in everyday communication. The primary objective of this project is to build a comprehensive resource that includes both textual and spoken messages, which can be systematically searched using a range of parameters, including age group, gender of participants, and interaction types. The resource will make it possible to observe in more detail hybrid interactions, in which spoken and written elements (but also graphic and multimedia ones) interact, giving rise to new forms of conversation. Below, we present illustrative examples of the kind of hybrid conversations observable on WhatsApp, followed by the English translation. In (1), images and emoji are employed as graphic representations of meaning, supporting, if not replacing, verbal language. In (2), the transition in conversational turn corresponds to a shift in medium, from vocal to written.

(1)

15/03/21, 10:03 - IB03: Quando vuoi chiedere di recuperare il sabato pomeriggio?
15/03/21, 10:04 - RS01: *immagine omessa [reazione]*
15/03/21, 10:06 - RS01: Ma 🖤 🖤 🖤 🖤

15/03/21, 10:03 - IB03: When do you want to ask to make up for Saturday afternoon?
15/03/21, 10:04 - RS01: *omitted image [reaction]*
15/03/21, 10:06 - RS01: But ❤️ ❤️ ❤️ ❤️

(2)

24/11/22, 10:31 AI01: [AUDIO] *no oggi io sono a casa finalmente perché la settimana scorsa anche sono stata via quattro giorni su cinque […]*
24/11/22, 10:32 LR02: anche io fuori 4 giorni su 5 settimana scorsa.. per fortuna se ne riparla lunedì

4/11/22, 10:31 AI01: [AUDIO] *no today I am home finally because last week also I was away four days out of five* [...].
4/24/22, 10:32 AM LR02: Me also out 4 days out of 5 last week.. luckily we'll talk about it on Monday

The ultimate goal of this project is to make the corpus publicly accessible, following appropriate anonymization procedures, and to make it freely searchable for linguistically oriented research on different levels.
The structure of this paper unfolds as follows. Section 2 offers a concise review of the existing literature on the language used on the web and describes the currently available resources for Italian. Section 3 delineates the practical steps involved in constructing the resource and provides quantitative data regarding the corpus in its current form. In Section 4, we outline an ongoing process of data cleaning to convert data into .xml format. Finally, in Section 5 we conclude the paper and highlight our forthcoming research plans.

2. Theoretical background

2.1 The languages of the web

The proliferation of social networks and the advent of new instant messaging technologies have sparked a growing scholarly interest in the linguistic aspects of online communication (cf. e.g. Antonelli, 2007; Cerruti

LREC-COLING 2024, pages 16659–16663

et al., 2011; Cerruti and Onesti, 2013, Fiorentino, 2013; Pistolesi, 2004, 2014; Tavosanis, 2011, 2013). From the earliest studies on this subject, scholars have suggested that the languages used on the Internet show unique features and thus represent linguistic varieties of their own, which, consequently, have been assigned various labels, both global (cf. e.g. Netspeak; Crystal, 2004) and language-specific (cf. e.g. E-taliano for Italian; Antonelli, 2007). However, it is now recognized that the notion of a singular "language of the web" is unrealistic. Instead, scholars tend to acknowledge the existence of diverse varieties actively shaped and negotiated by language users participating in various communities of practice (cf. eg. Fiorentino, 2018).

In the Italian context, the debate over the languages of the web has been mostly concerned with their classification within the so-called "architecture of contemporary Italian" as formulated by Berruto (1987). Presently, there is a consensus that these varieties are most prominently marked along the diamesic axis of variation, which pertains to the variation through the medium used for communication (ie., written *vs* spoken; Mioni, 1983). However, as pointed out by Antonelli (2016), the varieties under investigation may also exhibit variations along the diaphasic and diastratic axes, respectively linked to the communicative situation and the demographics of the speakers, depending on the individuals' repertoire. For educated speakers, the varieties employed on the web may represent specific registers, while for less educated people, they me represent their only option, and may consequently be susceptible to social stigma. Regarding the diamesic aspect, scholars have observed that the varieties used on the web cannot be neatly categorized either as prototypically spoken or prototypically written, falling at different points along a continuum between these two extremes. Indeed, aside from the written medium, they appear to be closer to orality than to writing in the strict sense, due to their greater immediacy and limited planning at the expense of accuracy, much like in face-to-face conversation (cf. eg. Pistolesi, 2004; Gheno, 2011; Antonelli, 2016). In this context, which nowadays features a well-established research tradition, we believe that a promising area for expanding the debate could be the analysis of the linguistic features of voice messages, which have recently entered instant messaging chats.

2.2 Available resources for Italian

The online language usage in the Italian context is documented by a range of accessible online resources.

First, it is important to mention ItTenTen20[1], which encompasses an impressive 12 billion tokens collecting texts from websites, forums, and blogs. The corpus Paisà (Lyding et al., 2013) consists of texts gathered between 2009 and 2012 from various online sources, including Wikipedia, Indymedia, and blogs, totalling 250 million tokens. Finally, the resource NUNC (Barbera, 2013), dating to the year 2004, compiles texts extracted from newsgroups. In addition, there exist corpora related to social networks, particularly Twitter (Cignarella et al., 2018), which are often designed for specific purposes, such as the annotation of irony. In the realm of WhatsApp, the "What's up, Switzerland?" corpus (2016-2018) is noteworthy (Ueberwasser-Stark, 2017); however, while it also contains messages in Italian, it exclusively focuses on text-based chats.

3. Building the corpus

In this section, we delineate the resource's construction process. Section 3.1 illustrates the conception of the project and the organizational framework. Section 3.2 describes the data management, while Section 3.3 focuses on the corpus current composition, both in terms of the quantity of tokens, messages, and conversations, and of the demographics of the participants.

3.1 Project conception and initial steps

The conception of this project dates to the academic year 2020/2021, namely within the classes of Sociolinguistics and Pragmatics and Texts Linguistics at the University of Pavia, Department of Humanities. The research group currently includes as active members more than thirty students enrolled in the Bachelor of Arts and Languages and in the Master's program in Theoretical, Applied and Modern Language Linguistics. It also includes doctoral candidates, postdoctoral fellows, researchers, and professors affiliated with the Linguistics Section of the Department.

The design of this resource actively engaged the group across all phases of its development: from data collection, to organizing summary tables with information about the chats and participants, to transcribing audio files. More specifically, each participant undertook the following tasks, which are detailed in the next paragraphs: (a) provision of personal chats and voice messages while ensuring appropriate anonymization; (b) solicitation of informed consent from all participants engaged in the conversations; (c) annotation of participants and conversation metadata; (d) transcription of audio files using the software ELAN (Sloetjes and Wittenburg, 2008).

3.2 Data management

The data constituting the corpus, which is hybrid in nature, encompass both written and oral texts, derived from WhatsApp conversations. The collection, uploading and organization of written messages are still ongoing, as is the transcription of the voice messages.

3.2.1 Selection and privacy of written data

As mentioned before, project members personally select the data from their personal chats and subsequently export them in .txt format, post the collection of informed consent from all involved participants. Whenever consent is not obtained from a participant, all chats containing that participant are excluded.

[1] https://www.sketchengine.eu/ittenten-italian-corpus/

The .txt format allows the potential presence of emoji to be preserved. For other content types such as images and GIFs, specific conventions are used to indicate their omission in the exported text and to recover their function. More specifically, it was decided to annotate whether the omitted content performs a reaction (i.e., it is used in response to previous messages) or a sharing function (i.e., the speaker wants to show something to the interlocutor). For privacy reasons, proper names found in the message texts are replaced with other names, as well as the names of non-public places. The substitutive names are maintained for one and the same referent throughout conversations to ensure consistency. It might be argued that substitutive names could introduce biases due to potential associations with stereotypes. However, it is worth noting that the selected substitutive names are among the most used in Italian and are not intended to carry any connotations.

3.2.2 Transcription and privacy of spoken data

Voice messages are transcribed through the software ELAN, which allows exporting the transcription in .txt format. The transcription is carried out based on a slightly simplified version of the Jefferson Transcription System (Jefferson, 2004). Although the transcription of the recordings is verbatim, certain arrangements were agreed upon to ensure homogeneity of the data while accounting for potential variations in individual pronunciation. The audios are stored in .opus format (see https://opus-codec.org/). Before the corpus is publicly available, audios will be anonymized by introducing white noise in correspondence with person names and places of origin.

3.2.3 Metadata annotation

To facilitate the future search of the data in the corpus, metadata pertaining to participants and conversations are registered by assigning to each an alphanumeric code.
As for conversations, the following metadata are annotated: start and end dates, number of participants, formal/informal conversation, number of voicemails within the conversation, languages used, and presence of media files (images, videos, etc.). Since all collected conversations as of now are informal, the parameter formal *vs* informal might ultimately be abandoned.
As for participants, the annotated metadata pertains to: age group, gender, place of birth, place of residence, educational qualification, occupation, native language, and other known languages.

3.3 Corpus composition

The resource currently contains 89 conversations and totals more than 414,000 tokens. More detailed information on the composition of the corpus is given in Tab. 1.

Total number of conversations	89
Total number of participants	194
Total number of messages	79.488
Total number of tokens (chat)	414.177
Total number of voiced messages	280
Total duration of voiced messages	71'52''

Table 1: current composition of the corpus

The 194 participants in the chats are distributed by age group and gender as shown in the following figures (in percentage).

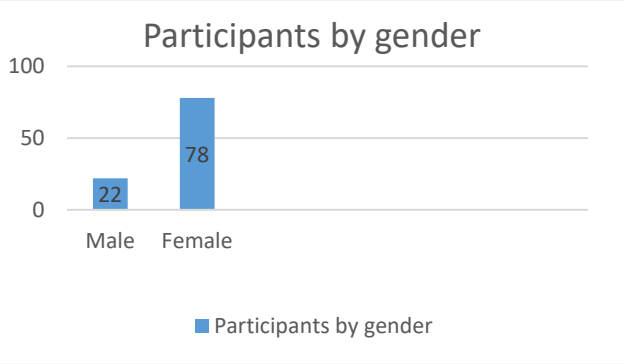

Figure 1: participants' distribution by gender (%)

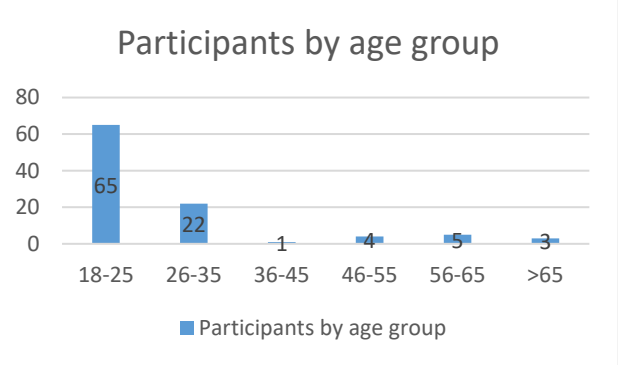

Figure 2: participants' distribution by age group (%)

As can be observed, participants between the ages of 18 and 25 currently stand out as the largest participant demographic, with women outnumbering men. Regarding the participants' geographic area of origin, more than 70% come from Northern Italy, while less than 30% come from Central and Southern Italy. As for the educational status, 1.7% of participants have completed their middle school education, 36,6% feature a high school diploma, and 60% have earned either a bachelor's or a master's degree. To obtain a more balanced corpus, we are planning targeted data collection to fill this gap, directly engaging speakers from the least represented age groups and

geographical area. It is also noteworthy that, in terms of register, all conversations are informal. Finally, while Italian is obviously the predominant language in the corpus, other languages are found as well, including regional Italian varieties like Sicilian or Neapolitan (often used for expressive purposes within conversations in Italian) and international languages like English, French, and Spanish.

4. Chat formatting

To ensure the accessibility of the database, the collected data must undergo conversion into an .xml format. This necessitates an ongoing process of data cleaning aimed at generating files that can be transformed into .xml format using specialized scripts. This section provides an overview of the procedures involved.
For each chat, a dedicated folder is created, structured to include:
a) A .txt file housing the chat;
b) A subfolder containing the original audio file;
c) Another subfolder with the transcription of the audio in .txt format based on established conventions;
d) A subfolder housing the textual content of the audio files in .txt format.
The .txt file hosting the chat is structured such that each line corresponds to an individual message. The line format is as follows: [when] who: rawtext.
The "when" field features a timestamp in the format [day/month/year, hh:mm:ss]. The "who" field records the unique identifier of the message sender. Finally, the "rawtext" field can be one of four types:
a) Plain text message;
b) Visual media (image, sticker, video, gif);
c) Audio;
d) Reply to a chat message.
In the case of visual media, a description of the omitted media is inserted between pipe characters (|). This description specifies whether the omitted content is an image, sticker, video, or gif. Relevant keywords are also provided within square brackets to describe the omitted media. If the media includes a caption, it is reported after the pipe, separated by a space. In case of audio, The alphanumeric code under which the audio file was stored in its relevant subfolder is indicated to incorporate audios into the chat. Lastly, when replies to a chat message incorporate the answered message via the dedicated option, the text and visual media from the initial message are indicated as well.
Examples of the resulting files are provided below. In (3), the user is replying to a text message, while in (4), she is replying to an image.

(3)

[11/05/23, 14:38:25] AB01: Ciao Barbara, come stai?
[11/05/23, 15:22:47] BR01: |CITAZIONE AB01: Ciao Barbara, come stai?| ciaooooo, tutto bene tu?

(4)

[08/08/22, 10:43:17] BR01: |immagine omessa [condivisione; selfie]| oggi mareeee
[08/08/22, 10:44:18] AB01: |CITAZIONE BR01: immagine omessa [condivisione;
selfie]| Ma è il costume nuovo?

5. Conclusions and future work

In this paper, we set out to describe a current initiative to build the WhAP corpus, the first corpus of the Italian language as used on the instant messaging platform WhatsApp. Specifically, we focused on two main aspects: on the one hand, the methods employed for data collection, transcription, and processing, and, on the other, the internal composition of the corpus and the socio-demographic characteristics of the participants in the chats. As the corpus is still under construction, it is presently accessible by request only. An ad hoc interface will be developed to allow the consultation of the data. In this regard, a script in python has been developed to make the data searchable by the final user. A partial release of the corpus is expected by June 2024. We believe that a resource like this can prove valuable not only for shedding light on the peculiar language variety used on WhatsApp, but also for researching emerging linguistic phenomena in contemporary Italian.

6. Acknowledgments

We would like to thank the three anonymous reviewers whose valuable suggestions contributed to the improvement of this paper. In addition to all the members of the WhAP project, we would like to thank Luca Brigada Villa for technical assistance with chat formatting. This paper is the result of close collaboration between the authors. For academic purposes, Ilaria Fiorentini is responsible for Sections 1 and 2, Marco Forlano for Sections 3 and 5, and Nicholas Nese for Section 4.

7. Bibliographical references

Antonelli, G. (2007). *L'italiano nella società della comunicazione*. Bologna, Il Mulino.

Antonelli, G. (2016). L'italiano nella società della comunicazione 2.0. Bologna, Il Mulino.

Berruto, G. (1987). Sociolinguistica dell'italiano contemporaneo. Roma, Carocci.

Cerruti, M. & Onesti, C. (2013)., Netspeak. A language variety? Some remarks from an Italian sociolinguistic perspective. In Miola, E. (ed.) (2013) Languages Go Web. Standard and non-standard languages on the Internet. Alessandria, Edizioni dell'Orso: 23-40.

Cerruti, M., Corino, E. and Onesti, C. (eds.) (2011). *Formale e informale. La variazione di registro nella comunicazione elettronica*. Roma, Carocci.

Cignarella, A.T., Frenda, S., Basile, V., Bosco, C., Patti, V. and Rosso, P. (2018). Overview of the Evalita 2018 Task on Irony Detection in Italian Tweets (IronITA). In Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final

Workshop (EVALITA 2018). CEUR.org.
Crystal, D. (2004). A glossary of netspeak and textspeak. Edinburgh, Edinburgh University Press.
Fiorentino, G. (2013). 'Wild language' goes Web: new writers and old problems in the elaboration of the written code. In Miola, E. (ed.) (2013) Languages Go Web. Standard and non-standard languages on the Internet. Alessandria, Edizioni dell'Orso: 67-90.
Fiorentino, G. (2018). Variabilità linguistica. Roma, Carocci.
Gheno, V. (2011). Socializzare in rete. In Stefanelli, S. and Saura, A.V. (eds.), I linguaggi giovanili, Firenze, Accademia della Crusca: 41-112.
Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. H. Lerner (Ed.), Conversation analysis: studies from the first generation. (pp. 13-23). Philadelphia, John Benjamins.
Mioni, A. (1983). Italiano tendenziale: osservazioni su alcuni aspetti della standardizzazione. In AA.VV., Scritti linguistici in onore di G. B. Pellegrini. Pisa, Pacini, pp. 495-517.
Pistolesi, E. (2004*).* Il parlar spedito. L'italiano di chat, e-mail e SMS. Padova, Esedra.
Pistolesi, E. (2014). *Scritture digitali*. In Antonelli, G., Motolese, M. & Tomasin, L. (eds.i) (2014). Storia dell'italiano scritto. Vol. III: Italiano dell'uso. Roma, Carocci. 2014: 349-375.
Sloetjes, H., & Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. In Proceedings of the 6th International Conference on Language Resources and Evaluation.
Tavosanis, Mirko (2011) L'italiano del Web. Roma, Carocci.
Tavosanis, M. (2013). Non-standard rules: innovation you cannot find on the Italian Web. In Miola, E. (ed.) (2013) Languages Go Web. Standard and non-standard languages on the Internet. Alessandria, Edizioni dell'Orso: 141-151.
Ueberwasser, S. & Stark, E. (2017). What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik online* 84/5, 105-126.