

DialettiBot: a Telegram Bot for Crowdsourcing Recordings of Italian Dialects

Federico Sangati
University L’Orientale
Naples, Italy
fsangati@unior.it

Ekaterina Abramova
Nijmegen University
The Netherlands
e.abramova@ftr.ru.nl

Johanna Monti
University L’Orientale
Naples, Italy
jmonti@unior.it

Abstract

English. In this paper we describe DialettiBot, a Telegram based chatbot for crowdsourcing geo-referenced voice recordings of Italian dialects. The system enables people to listen to previously recorded audio and encourages them to contribute to building a collective linguistic resource by sending voice recordings of their own spoken dialects. The project aims at collecting a large sample of voice recordings in order to promote knowledge of linguistic variation and preserve proverbs or idioms typical for different local dialects. Moreover, the collected data can contribute to several voice-based Natural Language Processing (NLP) applications in helping them understand utterances in non-standard Italian.

Italiano.

In questo articolo descriviamo DialettiBot, un chatbot basato su Telegram per raccogliere registrazioni audio geo-referenziate di dialetti italiani. Il sistema permette alle persone di ascoltare le registrazioni precedentemente inserite, e le incoraggia a contribuire alla costruzione di questa risorsa linguistica collettiva, attraverso l’invio di registrazioni audio nel proprio dialetto. Il progetto mira a raccogliere una grande mole di registrazioni che possono aiutare a promuovere la conoscenza delle variazioni linguistiche e la salvaguardia dei proverbi o modi di dire tipici di ogni dialetto locale. I dati raccolti possono inoltre contribuire a diverse applicazioni del trattamento automatico del linguaggio (TAL) che hanno bisogno di essere adattate per comprendere espressioni dialettali.

1 Introduction

It is commonly known that Italy has an abundance of different dialects, such as Florentine, Venetian, and Neapolitan. These dialects are not only characterized by simple phonetic variation as it is usually meant by this term, but they are proper Romance languages, with a fully developed grammar and lexicon. As Repetti puts it:

The Italian ‘dialects’ [...] are daughter languages of Latin and sister languages of each other, of standard Italian, and of other Romance languages, and they may be as different from each other and from standard Italian as French is from Portuguese. (Repetti, 2000)

This dialectical variety is a resource that deserves to be studied and preserved for both cultural and applied reasons. The former, because it is quickly disappearing with less and less people who regularly use dialect at home and in public places. According to UNESCO ‘Atlas of the World’s Languages in Danger’,¹ there are about 2,500 endangered languages worldwide. In Italy, thirty dialects are at risk of extinction, such as friulano, ladino and veneciano.² The applied motivation is that in recent years we have witnessed a significant growth in the number of voice-based NLP applications (such as virtual assistants), which are currently not trained on local dialects and therefore perform poorly with a number of Italian speakers.

In this paper we present a freely available tool that enables geo-referenced recording of Italian dialects: *DialettiBot*, a Telegram based chatbot, whose aim is to collect a large sample of voice recordings, promoting preservation of linguistic

¹<http://www.unesco.org/languages-atlas>

²http://www.culturaitalia.it/opencms/en/contenuti/focus/UNESCO_warns_that_thirty_Italian_dialects_are_at_risk_of_extinction.html?language=en

variation and its use in NLP applications. The rest of the paper is organized as follows: in section 2 we describe related work, in section 3 the implemented system and in section 4 the collected data.

2 Related work

There has been an extensive linguistic research of Italian dialects (Lepschy and Lepschy, 1992; Belletti, 1993; D’Alessandro et al., 2010). Here we summarize a number of projects that relate to the idea of gathering linguistic recordings for producing a map of dialects. We also point out their limitations that inspire our project.

VIVALDI project the “Vivaio Acustico delle Lingue e dei Dialetti d’Italia” is a collection of recordings and transcriptions of fixed phrases in the dialects of different cities from all regions in Italy (Kattenbusch et al., 1998). Unfortunately, the project is no longer active and has mainly focused on a finite set of chosen sentences, as opposed to spontaneous utterances.

LOCALINGUAL A web application for crowdsourcing recordings from around the world. This project is the one that most closely relates to ours. The main difference is that it is not restricted to a specific country, does not use geo-locations and works via a web application, which makes it difficult to be used on mobile devices or in case of poor data connection.³

ALF Atlas Linguistique de la France: an influential dialect atlas of Romance varieties in France published in 13 volumes between 1902 and 1910 (Gilliéron and Edmont, 1902). An example of more recent work of this type is Hall, Damien (2012).⁴

ALD Linguistic Atlas of Dolomitic Ladinian and neighbouring Dialects (Skubic, 2000). The project studies the linguistic variation between dialects of the region which covers the Grisons and Friuli region.⁵

IDEA The International Dialects of English Archive was created in 1998 as the internet’s first archive of primary-source recordings of English-language dialects and accents

as heard around the world. With roughly 1,400 samples from 120 countries and territories, and more than 170 hours of recordings, IDEA is now the largest archive of its kind.⁶

MICROCONTACT aims at developing a theory of syntactic change by observing the evolution of the dialects spoken by Italians who have migrated to North and South America during the 20th century.⁷

SPEAKUNIQUE and VOCALID are two similar projects that aim at collecting English voice sample from different regions for creating personalized digital voices for communication text to speech devices.⁸

Our project aims to be an updated and continuously evolving initiative that can capture spontaneous (living) dialectal variation over the whole Italian territory by being freely accessible and easy to use for a variety of non-specialists. As such, the project follows methodological practices similar to other citizen-science projects (Gurevych and Zesch, 2013; Simpson et al., 2014; Hosseini et al., 2014), it incorporates a GWAP⁹ feature (Lafourcade et al., 2015), and fits within the line of ‘explicit crowdsourcing’ as defined by the EnetCollect¹⁰ COST¹¹ action.

3 System description

In order to crowdsource recordings from Italian dialects, we have built a Telegram chatbot: DialettiBot.¹² As shown in the screenshot in figure 1, the user can interact with the bot via a standard dialogue chat interface in a Telegram application which is freely available for all mobile or desktop operating systems.¹³ Apart from textual input, the interface provides a small keyboard of buttons that changes during the dialogue flow to simplify the interaction. In addition, the bot is able to accept vocal recordings and GPS locations.

The bot gives the possibility to the user to listen to approved recordings or to add new ones.

In the **listening mode**, it is possible to search for recording based on location or view the list

³<https://localingual.com>

⁴<http://cartodialect.imag.fr/cartodialect/accueil>

⁵<https://www.micura.it/en/activities/ald-linguistic-atlas>

⁶<https://www.dialectsarchive.com>

⁷<https://microcontact.sites.uu.nl/project>

⁸<https://www.speakunique.org>, <https://www.vocalid.co>

⁹Game with a purpose.

¹⁰<http://enetcollect.eurac.edu>

¹¹European Cooperation in Science and Technology.

¹²<https://t.me/dialettibot>

¹³<https://telegram.org/apps>



Figure 1: Screenshot of the DialectiBot system.

of the most recent recordings. As an element of gamification (Lafourcade et al., 2015), there is the possibility to ask for a random recording and try to guess its location. The user would then receive a feedback about the distance between the guessed location and the correct one. With this simple game we gather valuable data that would enable us to plot a type of confusability matrix between dialects, i.e., how much a dialect of place A resembles a dialect of place B.

In the **recording mode**, the user is asked to submit a freely chosen vocal recording of a sentence, that can be a simple phrase or a proverb, typical for their dialect. In addition, the user is asked to indicate the place where the dialect comes from (either by sending a GPS location or inputting the name of the place – in case the user is not currently located in the place associated with the dialect), and optionally the translation of the recording in Italian.

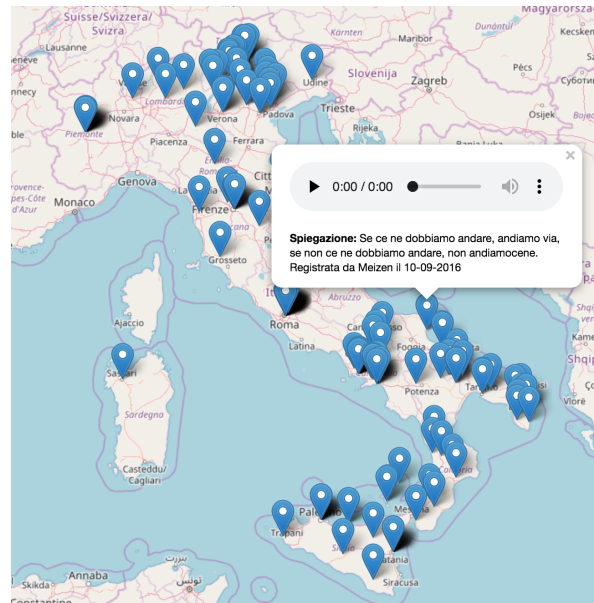


Figure 2: Screenshot of the web application displaying the audio map of the approved recordings.

As soon as the recording is submitted, the administrator of the system receives a notification (via the bot) with the new recording and is asked to approve or reject the contribution. Typical causes of rejection are too much background noise and explicitly offensive utterances. In case of approval, the recording is inserted in the database and becomes readily available to other users in the listening mode.¹⁴

In addition to the bot application, we developed a web application¹⁵ (see figure 2) for visualizing the approved recordings in a map and giving the possibility to click on each of them to listen to the audio and read the translation.

3.1 Technical Specification

The bot is implemented in Python using the telegram bot API.¹⁶ We chose to deploy the system via a chatbot (as opposed to a mobile app or web application) because it is much faster to build and to maintain since all the major functionalities (voice recordings, GPS location) are already embedded in the chat application and immediately accessible via simple API calls. Moreover, the system works on all mobile and desktop platforms without the need to build system-specific versions. Fi-

¹⁴In the future, there is a possibility to implement an additional validation step where other users or experts might flag some contribution as not being representative of a dialect.

¹⁵<http://dialectbot.appspot.com/audiomap/mappa.html>

¹⁶<https://core.telegram.org/bots/api>

nally, the simplified interface of a chatbot is particularly suitable to elderly people which are one of the most valuable target groups of the project, and can be easily used for recording other people while traveling also in case of no data connection (recordings are saved locally and uploaded to the server when data connection is again available).

The server behind DialettiBot is hosted by the Google Application Engine (GAE) framework and the data is stored in the integration datastore. The GAE technology guarantees full scalability up to an unrestricted number of users which could enable producing a significantly large volume of recordings. The same system also serves the web application with the map of the recordings illustrated in figure 2, which has been implemented in javascript using the *Leaflet*¹⁷ library.

4 Collected data

The first version of DialettiBot has been deployed in January 2016. Since then, 1,886 users have interacted with the system and have submitted 255 voice recordings out of which 220 have been approved.¹⁸ About 14% of users who interacted with the system contributed a recording.

Figure 3 shows the bar chart with the distribution of the approved recordings over time. The plot shows that the number of contributions in 2017 (31) has been significantly lower than in 2016 (117), whereas in 2018 the number is increasing again (72 in the first 3 quarters of the year).

Figure 4 shows the distribution of the approved recordings on the map of Italy, with the counts clustered by proximity (heat map). Campania is the region with most recordings (38), followed by Lazio (35), Trentino-South Tyrol and Sicily (27), Puglia (22), Veneto (15), Piedmont and Tuscany (12), Calabria and Lombardy (9), Basilicata (5), Emilia-Romagna, Friuli-Venezia Giulia and Marches (2), Abruzzo, Molise and Sardinia (1). Currently we have no recordings from Liguria, Umbria and Valle d’Aosta.

5 Conclusions and future work

We have presented DialettiBot, a chatbot system based on Telegram for crowdsourcing geo-referenced recordings of Italian dialects.

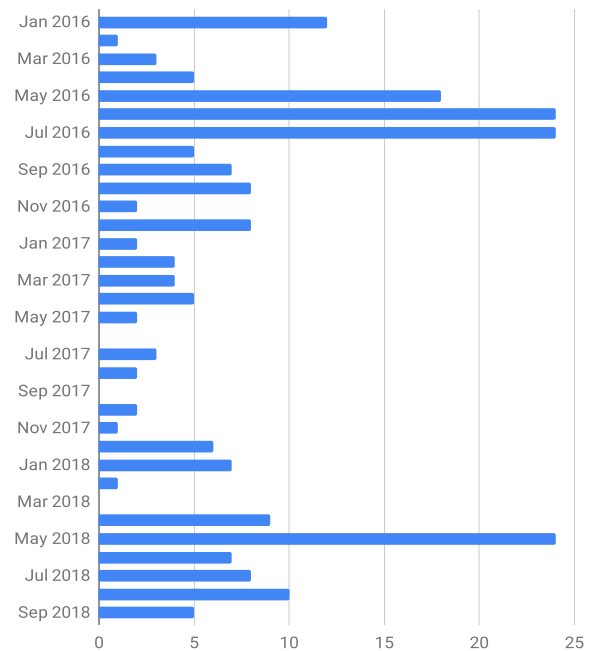


Figure 3: Frequency of approved recordings collected over time.



Figure 4: Heat map of the approved recordings.¹⁹

¹⁷<https://leafletjs.com>

¹⁸As of 31st of September 2018.

¹⁹Created via <https://mapmakerapp.com>.

Preliminary tests show that the system can be easily used by anyone who wishes to collect data in the field as well as the dialect speakers themselves. The recording quality is good and the data is easily exportable to be used for further processing in the service of linguistic research or NLP applications. At the same time, the current state of the project suffers from a number of limitations that need to be addressed in future work and that we discuss next.

First, the preliminary tests have not been informed by a detailed linguistic study of dialectal variation nor have we implemented a methodology for data collection. This is because the tests have been carried out as a proof-of-concept for the technology used to collect linguistic resources rather than a full-fledged linguistic project. Future tests will require a more careful consideration for dialect characteristics in the Italian language, the type of data that would be most valuable (spontaneous speech vs a set of set sentences etc.) and a construction of precise, reproducible instructions for the contributors.

Second, as described in section 3, we make use of a centralized validation procedure to approve a subset of recordings. However, since we have no complete knowledge of all Italian dialects we may end up accepting recordings which are not mapped to the correct location. In the future, we would like to decentralize the procedure, by delegating the approval to a higher number of volunteers spread out in all the regions, so that each new recording will get validated by the closest volunteer.

Finally, the number of users and recordings collected so far is relatively modest. This is due to the fact that no effort has been undertaken so far to promote its use by researchers or the general public. Accordingly, the current goal of the project is to get support from cultural institutions (both at a local and at a national level) to help us engage the citizens in this crowdsourcing effort, as well as academic partners to further refine the methodology and extend the chatbot capabilities.

We believe this project could contribute to help safeguard the Italian dialectic richness and collect useful resources for NLP applications, as we intend to make all recordings openly available for other researchers to use.²⁰

²⁰We are planning to upload the data to the Common Language Resources Infrastructure (CLARIN).

Acknowledgments

We kindly acknowledge all users who have so far contributed to the project by providing audio recordings of their dialects, and the three anonymous reviewers for their useful feedback.

References

- A. Belletti. 1993. *Syntactic Theory and the Dialects of Italy*. Volume 9 of *Linguistica* (Turin, Italy). Rosenberg & Sellier.
- Roberta D’Alessandro, Adam Ledgeway, Ian Roberts, and Frank Nuessel. 2010. *Syntactic Variation: The Dialects of Italy*. Cambridge University Press.
- Jules Gilliéron and Ed. Edmont. 1902. *Atlas linguistique de la France*. H. Champion,, Paris,.
- Iryna Gurevych and Torsten Zesch. 2013. Collective intelligence and language resources: Introduction to the special issue on collaboratively constructed language resources. *Lang. Resour. Eval.*, 47(1):1–7.
- Hall, Damien. 2012. Vers un nouvel atlas linguistique de la france. *SHS Web of Conferences*, 1:2171–2189.
- Mahmood Hosseini, Keith Phalp, Jacqui Taylor, and Raian Ali. 2014. The four pillars of crowdsourcing: a reference model. *IEEE Eighth International Conference on Research Challenges in Information Science*.
- Dieter Kattenbusch, Carola Köhler, Marcel Lucas Müller, and Fabio Tosques. 1998. VIVALDI project: Vivaio acustico delle lingue e di dialetti d’italia. <https://www2.hu-berlin.de/vivaldi>.
- M. Lafourcade, A. Joubert, and N.L. Brun. 2015. *Games with a Purpose (GWAPS)*. Focus Series in Cognitive Science and Knowledge Management. Wiley.
- A.L. Lepschy and G.C. Lepschy. 1992. *The Italian Language Today*. Hutchinson university library. Routledge.
- Lori Repetti. 2000. *Phonological Theory and the Dialects of Italy*. John Benjamins Publishing Company.
- Robert Simpson, Kevin R. Page, and David De Roure. 2014. Zooniverse: Observing the world’s largest citizen science platform. In *Proceedings of the Companion Publication of the 23rd International Conference on World*

Wide Web Companion, WWW Companion '14, pages 1049–1054. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland.

Mitja Skubic. 2000. Ladinia linguistica in una monumentale opera: Atlante linguistico del ladino dolomitico e dei dialetti limitrofi - ald-1, dr. ludwig reichert verlag, wiesbaden 1998. *Linguistica*, 40(1):188–195.