CorAIt – A Non-native Speech Database for Italian

Claudia Roberta Combei
FiLeLi - University of Pisa
(on leave at FAU Erlangen-Nürnberg)
roberta.combei@fileli.unipi.it

Abstract

English. CorAIt is a non-native speech database for Italian, which is freely accessible online for academic research purposes. It was especially designed to meet the requirements of a larger research project focused on foreign accented Italian speech. The corpus is aimed at providing a uniform collection of speech samples uttered by non-native speakers of Italian. To date, 105 non-native speakers – whose mother tongues are either French, Romanian, Spanish, English, German, or Russian – have been recorded. The corpus includes also a control group made up of 16 Italian speakers. There are almost 8 hours of audio material, both read speech (first and second reading), and spontaneous speech. This paper emphasizes the necessity for this type of database, it describes the steps involved in its construction, and it presents the features of CorAIt.

Italiano. *CorAIt è un corpus audio di l'italiano L2 liberamente consultabile online per scopi di ricerca scientifica. Il corpus è parte integrante di un progetto di ricerca che affronta l'accento straniero nella lingua italiana da una prospettiva più ampia. E' stato ideato e costruito con lo scopo di fornire una raccolta uniforme di materiale audio prodotto da parlanti di italiano L2. Ad oggi sono stati registrati 105 parlanti stranieri di madrelingua: francese, romena, spagnola, inglese, tedesca, e russa. In aggiunta, il corpus è dotato di un gruppo di controllo composto da 16 parlanti italiani. Sono disponibili circa 8 ore di registrazioni, sia di parlato letto (prima e seconda lettura) che di parlato spontaneo. L'articolo evidenzia la necessità di costruire questo tipo di database, e descrive la progettazione e le caratteristiche di CorAIt.*

1 Introduction

It has become clear that accurately designed speech corpora are of essential importance for the development of efficient speech technologies. Investigating how native and foreign-accented speech differ is a necessary step in non-native speech recognition (Tomokiyo, 2001).

Currently, the number of non-native speech databases seems almost insignificant if compared to corpora of native speech.

Moreover, until recently, the majority of the research has focused on English. Therefore, some of the largest non-native speech databases are available for this language: *TED* (Lamel et al., 1994), *Duke-Arslan* (Arslan & Hansen, 1997), *ISLE* (Menzel et al., 2000), *IBM-Fisher* (Fisher et al., 2003), *ATR-Gruhn* (Gruhn et al., 2004), *CSLU* (Lander, 2007), *NATO M-ATC* (Pigeon et al., 2007), and *Speech Accent Archive* (Weinberger, 2015). Large speech corpora of foreign-accented English are owned by Speechocean, and they are specifically built for commercial purposes, especially for training and testing speech recognizers, but some of them are also available for academic research on the KilingLine Data Center platform (Speechocean, 2017).

Only in the last few years has there been an interest in other languages. Without claiming to be exhaustive, some of the largest non-native speech databases for languages other than English will be mentioned: *BAS Strange I+II* (University of Munich, 1998) for German, *WP Russian* (La Rocca & Tomei, 2003) for Russian, *Tokyo-Kikuko* (Nishina, 2004) for Japanese, *TC-STAR* (van den Heuvel et al., 2006) and *WP Spanish* (Morgan, 2006) for Spanish, *SINOD* (Žgank et al., 2006) for Slovenian, and iCALL (Chen et al., 2015) for Chinese.

However, as a result of that fact that many non-native speech databases are built for commercial

purposes within private research centres, it is actually quite difficult to map all the resources of this type ever built (Cf. Gruhn et al., 2011, for an overview of the non-native speech databases available at the date their study was published).

This paper presents CorAIt, a non-native speech database for Italian. The database is part of a Ph.D. project which intends to study foreign accented Italian speech both from a computational perspective (automatic identification and classification of non-native accent) and a perceptual perspective (interpretation of quantitative and qualitative judgments delivered by expert and naïve native Italian speakers with respect to non-native pronunciations). The design and the development of this corpus were determined by several factors, which are outlined below.

2 Motivations

Currently, the automatic speech recognition systems for Italian which are integrated into generally available virtual assistant software (e.g. *Google Now*, *Google Assistant*, *Siri*, *Cortana*, etc.) perform quite well on native speech. However, despite recent advances in this field, non-native accents still represent a challenge. This may be due to fact that there is significantly less training data available for automatic speech recognition systems on non-native pronunciations. Considering that Italy is a multicultural country, with over 5 million foreign citizens, representing 8.3% of the entire population residing on its territory[1], it would be desirable to provide services to users who speak Italian with non-native accents.

Apart from acting like training sets for automatic speech recognition systems or for text-to-speech systems, non-native speech databases might be beneficial in the fields of computer assisted language learning (CALL) and mobile-assisted language learning (MALL), as well as for linguistic profiling tasks. Glottologists and scholars working on Italian as a foreign language might also benefit from the presence of these resources.

At the date this research began, there was only one audio corpus for foreign-accented Italian speech, freely available for online consultation, namely *DILS - Dialoghi in Italiano Lingua Straniera* (Savy et al., 2012), consisting of semi-spontaneous audio material obtained by means of the task-oriented dialogue elicitation technique. *DILS* contains 9 large audio samples (for a total duration of 100 minutes) uttered by 18 speakers: 12 Dutch females, 3 Spanish females and 3 Spanish males.

It is worthwhile to mention that there are several other learner corpora for Italian: *VALICO - Varietà Apprendimento Lingua Italiana Corpus Online* (Barbera & Marello, 2004), which is a collection of non-native written Italian; *LIPS - Lessico dell'italiano parlato da stranieri* (Vedovelli et al., 2006); and *Corpus Parlato di Italiano L2* (Spina et al., 2006). The last two corpora consist of transcriptions of audio samples produced by non-native speakers.

In addition to the above-mentioned corpora, there exists a database of written and spoken non-native Italian, entitled *ADIL2 - Archivio Digitale di Italiano L2* (Palermo, 2009*),* which is purchasable in the form of a DVD. However, despite the sophistications of its search tool, the accurate transcription, as well as the admirable amount of data collected, *ADIL2* presents a series of issues that cannot be ignored, such as: imbalance with respect to the speakers' mother tongues (i.e. some languages are underrepresented while others are overrepresented) and elicitation technique used for some samples (i.e. interviews repeated various times over variable time-frames to the same subjects). These aspects render *ADIL2* unsuitable for the type of research to be taken on. Therefore, it became necessary to collect a database of non-native Italian speech.

3 Data Collection

The corpus was designed, collected and developed from January 2016 through July 2017, and it was aimed at providing a uniform collection of audio material produced by adult non-native speakers of Italian residing in Bologna[2].

Initially, the intent was to collect data for 11 different mother tongues (L1s): Maghrebi Arabic, Urdu, Mandarin Chinese, Albanian, Russian, English, German, French, Romanian, Spanish, and Italian (as a control group). The first 10 groups correspond to the L1s spoken by some of the major foreign populations residing in Italy. However, recruiting speakers for all these groups

[1] The data were provided by the National demographic balance (year 2016) produced by the Italian National Institute for Statistics (ISTAT). The full report is available at: http://www.istat.it/it/files/2017/06/bilanciodemografico-2016_13giugno2017.pdf

[2] To simplify the data collection process, the author chose to recruit people that studied or worked in Bologna (her city of residence).

proved to be a challenging task. This may be result of the fact that participation was entirely voluntary and no material reward was provided to informants. Since it was not possible to recruit enough speakers of Maghrebi Arabic, Urdu, Mandarin Chinese and Albanian, these four groups were abandoned.

3.1 Speakers' recruitment

Specific criteria of quality, quantity and diversity were observed, as much as possible, for each L1 when the participants were selected (Cf. section 4.1).

All speakers were recruited locally in Bologna. Most informants were enrolled as regular or exchange students in B.A., M.A. and Ph.D. programmes at the University of Bologna and they were contacted on their personal e-mail address. The e-mail message contained a description of the research project and informed potential participants about the tasks they would have performed. Nearly one fourth of them replied positively to the call.

3.2 Experimental protocol

All informants were aware that they were recorded. They gave informed consent in writing to the use of their speech samples and their sociolinguistic data for research purposes.

In order to guarantee uniformity, the same experimental protocol was employed for all subjects. Before each recording session, speakers were asked to fill in a detailed form regarding their sociocultural and sociolinguistic background. The digital recordings were performed with a Samson METEOR MIC cardioid pickup microphone (condenser diaphragms: 25 mm) on the Praat software (Boersma & Weenink, 2017). The sampling parameters were the following: mono channel, 16-bit, 44,100 Hz, linearly encoded WAV.

Each recording session lasted around 60 minutes. The sessions were individual-based, and they were guided and monitored by the author.

3.3 Speech modalities

The speakers were asked to perform two tasks: reading an article excerpt published on the Italian newspaper *Corriere della Sera*[3]; and describing spontaneously how they spent their last holidays. That specific reading fragment was chosen because it presented various levels of complexity and it contained all Italian phonemes. The reading task was necessary for triggering difficulties that could emerge as a result of conflicting orthographic conventions between the speakers' mother tongues and Italian (Wottawa & Adda-Decker, 2016). Moreover, it could allow speakers comparisons and analyses on the same type of material.

All participants had two reading attempts and they were asked to read and speak as naturally as they could.

4 Description of CorAIt

CorAIt is a non-native speech database for Italian, which has become fully and freely available online for academic research consultation[4]. It contains 2,244 audio samples produced by 105 non-native speakers of Italian. It also includes 300 audio samples obtained from 16 native Italian speakers. In total there are almost 8 hours of speech, consisting in roughly 72,000 words.

4.1 Speakers' statistics

Originally, it was planned to recruit at least 15 speakers for each L1. This threshold was reached for French, and it was exceeded for the other six groups.

Regarding the age distribution of the informants, the range is 19-40 years, but most speakers are older than 20 and younger than 30 years (Cf. Table 1).

Mother tongue	Number of speakers	Age means (±S.D.)
Russian	17	27.06 (5.94)
English	16	23.75 (4.69)
German	17	23.53 (4.06)
French	15	23.33 (2.72)
Romanian	20	25.40 (2.22)
Spanish	20	23.65 (2.95)
Italian	16	29.00 (4.62)

Table 1: Speakers' basic statistics.

Despite the efforts to call up the same number of male and female speakers, the corpus is not perfectly gender-balanced. Apparently, it was

[3] The newspaper article is available online at: http://cinquantamila.corriere.it/storyTellerArticolo.php?storyId=0000002228555. The excerpt was included in this frame: "Don Geretti è un grande affabulatore […] Pietro terminò il suo cammino terreno e quello, tormentatissimo, verso la fede."

[4] CorAIt database is accessible for consultation (prior to registration) at: www.corait.it

easier to engage female participants. In fact, 80 females (66%) and 41 males (34%) were recorded for this project. Nonetheless, according to literature, gender has not been reported as a major source of pronunciation issues, and for this reason it was assumed that gender imbalance had minor effects on this type of studies (Gruhn et al., 2011).

For the corpus design, the age of Italian language onset was taken into consideration, and it was distributed as follows: childhood (23%), adolescence (26%), and adulthood (51%). The data are predictable, considering that most informants learnt Italian naturalistically (64%) soon after they moved to Italy. In fact, only 36% of them claimed that they used mainly scholastic methods for learning Italian, and that they had already spoken the language at their arrival in Italy.

Since most informants were exchange students, 59% of them had spent 6-12 months in Italy at the time they were recruited for this research. The remaining part had lived in Italy for 12-24 months (15%), or for more than 24 months (26%). Not surprisingly, the great majority of speakers claimed that they had been exposed only to the Bolognese variety of Italian.

Because it was almost impossible to predict the speakers' proficiency level[5] in Italian before meeting them, the balancedness is not guaranteed for all accent groups (e.g. no Romanian speaker had A2 waystage/elementary level in Italian). For the sake of brevity, at the general level, this variable is represented as follows in the database: waystage/elementary level - A2 (12%), threshold/intermediate level - B1 (28%), vantage/upper-intermediate level - B2 (28%), and advanced/proficiency levels - C1 and C2 (32%).

4.2 Speech samples

An average of 4 minutes and 25 seconds of raw audio material consisting of read and spontaneous speech were recorded for each speaker. Some speakers had to terminate the registration session earlier than planned, so in those cases it was possible to record only their first reading attempt. Regardless of that, the spontaneous speech collected (23%) is, however, inferior to the reading speech material (77%). All raw samples were segmented manually into utterances corresponding to grammatical sentences for the reading material, and to phonological sentences for the spontaneous speech. In most cases, the material was not qualitatively altered, so hesitation phenomena and disfluencies were generally left as they were.

4.3 Webapp architecture[6]

One contribution of this project is that of making the corpus available to the research community. Following the model of similar tools, a website that would host the database was created. Then, the embedded webapp could extrapolate and classify the audio files from the dataset, according to specific criteria.

For the creation of the webapp, the web framework *Django* as well as several *Python* libraries (*MySQL-python*, *django-treebeard*, *django-filer*, *html5lib*, *sorl*, *wsgi*, *polymorphic*, *classy-tags*, *audiofield*, *appconf*, etc.) were employed. That allowed the use of a powerful *ORM* system, equipped with a web interface for storing multiple data types into our *MySQL* database.

Moreover, the *Django* web framework *ORM* favoured the realization of data collection models: a model is the only final data source containing the fields and the essential behaviours of the dataset and of the reference objects.

Generally, each model is mapped to a single database table and each attribute represents a database field. The queries are performed by means of ad-hoc *API*s for each model.

The project is hosted on a server with a *CentOS 7* operating system. CortAIt is already configured for various types of *SQL* and *NoSQL* databases (*PostgreSQL*, *MongoDB*, *Cassandra*, etc.). It also supports the execution of some cloud computing platforms, such as *Amazon Web Services* (*AWS*), which could improve its performance in case of an exponential growth of the computational complexity.

4.4 Front-end presentation

The web database is queryable from the dedicated section of CorAIt website, prior to registration and approval. Due to storage issues, and observing the design of *Speech Accent Archive* (Weinberger, 2015), the format of the audio files available on the online version of CorAIt is .mp3. Samples coded in other formats (e.g. .wav, .flac, etc.) are freely available under request.

[5] All participants self-assessed their Italian level based on the Common European Framework of Reference for Languages, available at: http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

[6] The section 4.3 was written with the contribution of Antonio Maria Tenace, who provided support on the graphical implementation of the webapp and was in charge with the technical aspects of its architecture.

To enable advanced queries, various layers of metadata were added to each audio file: the speaker's mother tongue, gender, age of Italian language onset, age at the time the sample was recorded, level of Italian proficiency, Italian learning method, length of residence in Italy, proficiency in other foreign languages. Moreover, information on the type of sample and its quality was included (Cf. Figure 1).

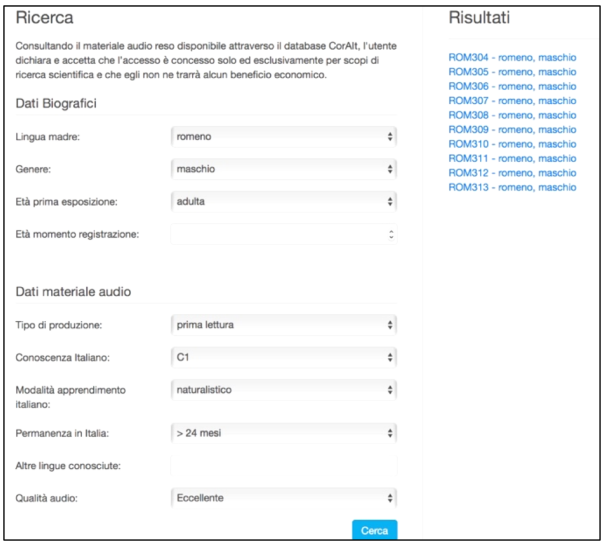

Figure 1: Search tool.

The corpus has not been transcribed nor annotated yet. However, following the example of the *Speech Accent Archive* (Weinberger, 2015), the sole grammatical sentences from the reading excerpts were inserted under the samples corresponding to the reading task.

Besides the embedded audio player – which allows to listen and download the audio sample – the window where the single result is displayed provides biographical and quantitative information with respect to the speaker who uttered that speech sample, as well as qualitative information regarding the audio file (Cf. Figure 2).

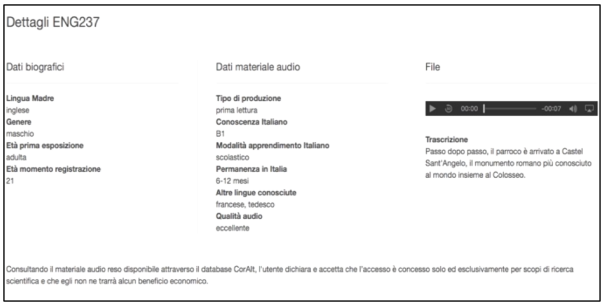

Figure 2: Results window.

5 Conclusion and future work

Considering that currently this non-native speech database presents some imbalance issues as regards the speakers' age of Italian language onset, their proficiency level, as well as the length of their residence in Italy, the data collection will be further extended.

In the future, the web database might also be enhanced with orthographic and phonetic transcriptions. Disfluencies (i.e. false starts, filled and silent pauses, phoneme lengthening, mispronounced words), mouth clicks, and external noise could be annotated.

6 Acknowledgements

The author gratefully acknowledges the informants who participated in this research, and A. M. Tenace, whose expertise and assistance were fundamental for the architecture and the implementation of the webapp.

Reference

L. M. Arslan & J. H. Hansen. 1997. Frequency characteristics of foreign accented speech. In *Proc. of ICASSP*, pp. 1123-1126, Munich, Germany.

E. Atagi & T. Bent. 2013. Auditory free classification of non-native speech. In *J. Phon.*, 41(6).

M. Barbera & C. Marello. 2004, VALICO (Varietà di Apprendimento della Lingua Italiana Corpus Online): una presentazione. In *ITALS 4*, Guerra Edizioni, Perugia, Italy.

P. Boersma & D. Weenink. 2017. Praat: doing phonetics by computer [Computer program]. Version 6.0.33. [http://www.praat.org]

N. F. Chen et al. 2015. iCALL Corpus: Mandarin Chinese Spoken by Non-Native Speakers of European Descent. In *Proc. of Interspeech*, pp. 801-805, Dresden, Germany.

C. Cieri et al. 2004. The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text. In *Proc. of LREC*, pp. 69-71, Lisbon, Portugal.

T. Cincarek et al. 2004. Speech Recognition for Multiple Non-Accent Groups with Speaker-Group-Dependent Acoustic Models. In *Proc. of Interspeech*, pp. 1509-1512, Jeju Island, Korea.

C. Cucchiarini & H. Van Hamme. 2013. The JASMIN Speech Corpus: Recordings of Children, Non-natives and Elderly People. In P. Spyns & J. Odijk (editors), *Essential Speech and Language Technology for Dutch*, pp. 43–59. Springer, Heidelberg, Germany.

J. Durand et al. (editors). 2014. *The Oxford Handbook of Corpus Phonology*. Oxford University Press, Oxford, UK.

V. Fischer et al. 2003. Recent progress in the decoding of non-native speech with multilingual acoustic models. In *Proc. of Eurospeech*, pp. 3105-3108, Geneva, Switzerland.

R. Gruhn et al. 2004. A multi-accent non-native English database. In *Proc. of Acoustical Society of Japan*, Kyoto, Japan.

R. Gruhn et al. 2011. *Statistical Pronunciation Modelling for Non-native Speech Processin*g. Springer, Heidelberg, Germany.

H. Heuvel et al. 2006. TC-STAR: New language resources for ASR and SLT purposes. In *LREC*, pp. 2570-2573, Genoa, Italy.

L. M. Tomokiyo. 2001. *Recognizing non-native speech. Characterizing and adapting to non-native usage in LVCSR*. PhD thesis. Carnegie Mellon University, Pittsburgh, USA.

L. F. Lamel et al. 1994. The translanguage English database TED. In *ICSLP*, Yokohama, Japan.

T. Lander. 2007. *CSLU: Foreign Accented English*. Release 1.2. LDC2007S08. Linguistic Data Consortium, Philadelphia, USA.

A. La Rocca and C. Tomei. 2003. *West point Russian speech corpus*. Tech. Rep., LDC, Philadelphia, Pennsylvania, USA.

Ludwig Maximilian University of Munich. 1998. Bavarian Archive for Speech Signals. [http://www.phonetik.uni-muenchen.de/Bas]

W. Menzel et al. 2000. The ISLE corpus of non-native spoken English. In *LREC*, pp. 957-963, Athens, Greece.

W. Minker et al. (editors). 2014. *Spoken dialogue systems. Technology and design*. Springer, Heidelberg, Germany.

J. Morgan. 2006. *West point heroico Spanish speech*. Tech. Rep., LDC, Philadelphia, Pennsylvania, USA.

K. Nishina. 2004. Development of Japanese speech database read by non-native speakers for constructing CALL system. In *ICA*, pp. 561-564, Kyoto, Japan.

M. Palermo. 2009. *Percorsi e strategie di apprendimento dell'italiano lingua seconda: sondaggi su ADIL2*. Guerra Edizioni, Perugia, Italy.

S. Pigeon et al. 2007. Design and characterization of the non-native military air traffic communications database. In *ICSLP*, Antwerp, Belgium.

M. Raab et al. 2007. Non-native speech databases. In *Proc. of IEEE-ASRU*, pp. 413 – 418, Kyoto, Japan.

T. Robinson et al. 1995. WSJCAM0: A British-English speech corpus for large vocabulary continuous speech recognition. In *Proc. of IEEE-ICASSP*, pp. 81-84, Detroit, USA.

R. Savy et al. 2012. *DILS - Dialoghi in Italiano Lingua Straniera*. [http://www.parlaritaliano.it/index.php/it/corpora-di-parlato/794-corpus-dils-dialoghi-in-italiano-lingua-straniera]

Speechocean. 2017. *King-ASR-L-190: Chinese English Speech Recognition Database*. [http://kingline.speechocean.com/exchange.php?id=13873&act=view]

S. Spina et al. 2006. *Corpus Parlato di Italiano L2*. [http://elearning.unistrapg.it/osservatorio/corpus/frames-cqp.html]

M. Vedovelli. 2006, LIPS - Lessico di frequenza dell'Italiano Parlato dagli Stranieri. In C. Bardel, J. Nystedt (editors), *Progetto Dizionario Italiano-Svedese*, pp-55-58. Acta Universitatis Stockholmiensis 22, Romanica Stockholmiensia, Stockholm, Sweden.

S. Weinberger. 2015. *Speech Accent Archive*. George Mason University. [http://accent.gmu.edu]

J. Wottawa & M. Adda-Decker. 2016. French Learners Audio Corpus of German Speech (FLACGS). In *LREC*, pp. 3215-3219, Portorož, Slovenia.

A. Žgank et al. 2006. SINOD – Slovenian non-native speech database. In *LREC*, pp. 1620-1623, Genoa, Italy.