KIParla corpus: a new resource for spoken Italian

Caterina Mauri
Università di Bologna
caterina.mauri@unibo.it

Silvia Ballarè
Università di Torino
silvia.ballare@unito.it

Eugenio Goria
Università di Torino
eugenio.goria@unito.it

Massimo Cerruti
Università di Torino
massimosimone.cerruti@unito.it

Francesco Suriano
Università di Bologna
francesco.suriano2@studio.unibo.it

Abstract

In this contribution we introduce the main features of the KIParla corpus, a newly built resource for the study of spoken Italian. Among other things, KIParla provides access to a wide range of metadata that characterize both the participants and the settings in which the interactions take place. Furthermore, it is built to be shared as a free resource through the NoSketch Engine tool and to be expanded as a monitor corpus (Sinclair 1991).

1 KIParla corpus: an introduction

The aim of this paper is to describe the design and implementation of a newly built resource for the study of spoken Italian. KIParla corpus is the result of a joint collaboration of the Universities of Bologna and Turin and is open to further partnerships in the future.

It is characterized by a number of innovative features: it provides access to a wide range of metadata concerning the speakers and the setting in which the interactions take place; it provides transcriptions time-aligned with audio files; it is designed to be expanded and upgraded through the addition of independent modules, constructed with a similar attention to the metadata; it is completely open-access and makes use of open-access technologies, such as the NoSketch Engine tool.

Section 2 provides a detailed description of the corpus design, aimed at representing the geographic, social and situational variation characterizing spoken Italian. In Section 3 we discuss the corpus implementation, by describing how data have been collected, with a special view to ethical requirements, how they have been treated and transcribed, and how they have been made accessible and searchable through NoSketchEngine. Section 4 focuses on the incremental modularity of the corpus, which makes it an open monitor corpus of spoken Italian. The two modules that constitute the actual core of KIParla are briefly illustrated, namely KIP module and ParlaTO module, and some prospects for future developments are sketched.

2 Corpus design

This section discusses the parameters taken into account for the creation of the KIParla corpus. In particular, we stress the relevance of extralinguistic factors (regarding both the socio-geographical characterization of the speakers and the interactional contexts) in order to build a corpus suitable for investigating the (socio)linguistic variation of contemporary Italian.

2.1 Aims

The KIParla corpus was designed in order to overcome some of the shortcomings that characterize previous resources used in the study of spoken Italian. It is intended to bring major improvements concerning three key aspects of corpus-based research: (i) access to the speakers' metadata, and particularly to those concerning age and social group; (ii) possibility to browse the corpus online as well as to download specific recordings; (iii) text-to-speech alignment.

As for (i), the possibility to recover information about the speakers or about the situation in which a conversational exchange has occurred is central in several fields of linguistics, such as Sociolinguistics and Conversation Analysis, and is potentially relevant in many others such as Second Language Acquisition and Language Teaching. Currently, no other corpus of spoken Italian offers

detailed information about single speakers, while some provide general information about the setting of the interaction. As for (ii), KIParla will be accessible online through the NoStechEngine interface, while on the project website it will be possible to download all the recordings (in .wav or .mp3 format) and transcription, as previously done, among others, for CLIPS (Albano Leoni 2007) and VoLIP (Voghera *et al.* 2014). Moreover, (iii) the research platform will allow to listen and download the results of single queries in .mp3, offering text-to-speech alignment.

The philosophy behind KIParla is to open the way towards a collection of spoken corpora, each built according to shared methodology in order to allow comparability. For this reason, it was designed as an open resource that is able to receive further implementations from external contributors who want to share their data; therefore, it can also be thought of as a monitor corpus (Sinclair 1991) whose size grows over time including an increasingly wide range of materials.

2.2 Geographic dimension: collecting data in different cities with speakers from all over Italy

The diatopic dimension has always been considered to be the most significant one to describe the Italian sociolinguistic scenario (see Berruto 2012 *inter al.*); speech utterances without any regional features are seldom if ever found even among educated speakers and in formal situations. Currently the only corpora that take into account the geographic variation are the LIP corpus and the CLIPS corpus. In the KIParla corpus, so far, we have collected data in Turin and Bologna; the sociolinguistic situation of both urban settings is characterized by the coexistence of Italian and the local dialect, as well as the resulting development of intermediate varieties. Furthermore, even with significant differences, both cities have been and are destinations of internal mobility and thus we are likely to find several varieties of Italian from other parts of Italy, as well as Italo-Romance dialects. For this reason, in addition to the city in which data were collected, information regarding the place of origin of each speaker is accessible.

2.3 Diastratic dimension: a perspective on Italian society

Speakers involved in the recordings are primarly differentiated by their age and level of education; the latter are traditionally deemed to be the most relevant social factors for the analysis of sociolinguistic variation in Italian (see Berretta 1988). Part of the KIParla corpus (see KIP module in §4.1) is focused on educated speakers, i. e. undergraduate and graduate students and academic professors. During the second data collection (see ParlaTO module in §4.2), far more social factors have been taken into account and both the age range and the level of education of the informants have been widened.
Ideally, the incremental nature of the corpus will allow to explore the various dimensions of variation in depth.

2.4 Types of interaction: settings and activities

Building on a central assumption in the conversation analytic framework, linguistic practices are often related to specific social activities. Hence, we gave particular attention to including different types of situations, expecting to find considerable differences between the structures involved in each of these.

In order to narrow down the field of analysis, for the first bulk of the KIParla corpus we chose to consider various types of interaction occurring in a single sociolinguistic domain (Fishman 1972), namely the academic context. The different activities were thus classified according to external factors, such as: (i) the symmetrical vs asymmetrical relationship between the participants; (ii) the presence *vs* absence of previously established topics; (iii) the presence *vs* absence of constraints on turn-taking. We believe indeed that using three very general features is particularly helpful in the task of integrating into the corpus new data that were recorded in other situations, without losing comparability with the other parts of the corpus. For example, interviews collected with different types of speakers in the ParlaTO section (§ 4.2) will result as comparable to those collected in the academic setting, regardless of any other difference between the two sets.

3 Building the corpus: data collection, transcription, publication and accessibility

3.1 Data collection: praxis and ethics

All data have been collected by professional researchers; students and interns of the Universities of Bologna and Turin have been involved in the process, after a period of specific training. Increasing the number of people who collected data is crucial to avoid snowball effects and thus including informants that belong to the same social

network. Furthermore, they acted as second-order contacts (see *friend of a friend* in Tagliamonte 2006: 21-22) and thus played an intermediary role in recording spontaneous speech and interviews.

Before every data collection, speakers were informed of the main aims of the project and the reasons why we needed to record the interaction. They agreed to the recording and signed a consent form that complies with the General Data Protection Regulation (G.D.P.R.) of the European Union. The consent form allowed us to collect linguistic material for scientific purposes, to store hardware located in Europe and/or via cloud services provided by universities, and to make it available online.

All the collected data are transcribed (see § 3.2) and, before being publicly available, are anonymized. The voice of the speakers is the only sensitive data that remains directly accessible.

3.2 Transcription: challenges and solutions

All the recordings have been transcribed by professional researchers, and trained students or interns using ELAN software (Sloetjes and Wittenburg 2008). This tool is designed specifically to handle multi-level annotations related to different speakers in a conversation. Among other things, it makes it possible to link each annotation with the media timeline. Thanks to this feature of the software, it was possible to implement text-to-speech alignment within the NoSketchEngine interface (§3.3).

Every tier in the transcription is referred to an alphanumeric code that links the spoken production of a single speaker with his/her metadata (e.g. age and level of education); similarly, each transcription file is associated to a code that allows to trace back its metadata (e.g. type of activity, number of participants, time and place of collection).

The most challenging aspect of transcribing spoken data is to strike a balance between a faithful representation of oral productions and the "serchability" of the written texts. For this reason, we decided to adopt a simplified version of the Jefferson's (2004) conventions used in Conversation Analysis; see Figure 1.

,	Rising intonation
.	Falling intonation
:	Prolonged sound (each : corresponds to ca. 20ms)
(.)	Short pause
> hello <	Bracketed speech is delivered more rapidly
<hello>	Bracketed speech is delivered more slowly
[hello]	Overlap between participants
(hello)	Hardly intelligible speech (transcriber's best guess)
xxx	Unintelligible speech
((laughs))	Non-verbal behavior
=	Prosodically attached units

Figure 1: Signs used in the transcription based on Jefferson (2004)

The choice of a conversational transcription was mainly due to the fact that it allows us to obtain a sufficient level of precision, without forcing the researcher to make interpretive choices. This is crucial in the handling of both performance-related phenomena occurring in spoken language (reformulations, truncated words, …) and non-standard variants. However, as will be explained in the next section, we decided to make the data searchable based on the simple orthographic transcription, while the conversational transcript is accessible as an additional option.

3.3 Data publication: From ELAN to NoSketchEngine

Since the transcriptions obtained through ELAN are in the XML format and are automatically time-aligned to the speech audio files, they are ready to be treated and parsed by XML-compatible technologies. Since one of our aims was to make the corpus fully accessible, we decided to make data available through the NoSketchEngine online platform (Rychlý 2007).

The NoSketch Engine is an open-source tool for corpus management providing a powerful and user-friendly interface to perform corpus KWIC queries, generate word/keyword lists, retrieve collocations based on several statistical measures and much more. In order to adapt the XML output of ELAN to the format required by NoSketchEngine, we wrote a python script that allows to: (i) make the metadata available both as query filters and text information; (ii) search the orthographic and the Jefferson transcriptions; (iii) directly link every occurrence with the time-aligned portion of the connected media file; (iv) search each module of the corpus separately.

The corpus will be released in September 2019 and will be available at the website www.kiparla.it. For now, the corpus has not been lemmatized nor POS-tagged, but these steps are planned for the near future.

4 Incremental modularity: an accessibile and open monitor corpus of spoken Italian

A key feature that makes KIParla corpus particularly innovative is its incremental modularity, namely its internal organization in independent modules and the ability to add new modules over time.

Modules are different corpora of Spoken Italian sharing the same design and a common set of metadata (see §2), transcribed by ELAN and made available through NoSketch Engine, by running the same script (see §3). The modules may focus on different dimensions of linguistic variation and may collect data from different geographical areas. However, the shared procedure of data collection and treatment guarantees a high level of mutual comparability.

The full accessibility of metadata makes the corpus easily *expandable*, through the addition of further modules focusing on different geographical, socio-cultural or communicative aspects, and *upgradable*, through the addition of new data for existing modules. Such a dynamic nature of KIParla corpus makes it a potential monitor corpus, open to integrations and upgrades across time. In the following sections, we provide a brief description of the two modules which at present constitute the core of the KIParla corpus.

4.1 KIP module

The KIP subcorpus is the first section that was designed within KIParla, and was originally thought of as a self-sufficient unit. It consists of around 70 hours of recorded speech collected in Turin and Bologna (35 hours per city approximately) and transcribed between 2016 and 2019.

The subcorpus is domain-specific in that it includes various types of interactions occurring within the academic setting; moreover, from a sociolinguistic perspective, it includes only speakers with major educational achievements, namely university students and professors.

The structure of this subcorpus is intended to maximize the diaphasic variability, according to the parameters described in 2.4. (symmetrical *vs* asymmetrical relation; presence *vs* absence of a moderator; presence *vs* absence of a fixed topic). This resulted in the selection of the contexts given in Figure 2, which represent ideal combinations between such parameters.

Activity	Bologna	Turin
spontaneous conversation	10:00:37	06:22:24
exams	03:09:34	03:10:48
lessons	12:19:39	13:25:33
interviews	06:18:37	07:47:38
office hours	02:59:11	03:49:08
TOTAL	34:47:38	34:35:30

Figure 2: hours recorded per each activity in Torino and Bologna

4.2 ParlaTO module

ParlaTO is a corpus of spontaneous speech collected in Turin between 2018 and 2019. The corpus has been compiling with the aim of portraying a contemporary multilingual urban setting. In fact, Turin has been, and still is, the scene of contact between different languages, partly because of the endogenous coexistence of Italian and Piedmontese and partly as the result of both internal and external migrations. Basically, the corpus contains actual speech data coming from three categories of individuals: (i) speakers of Piedmontese origin, (ii) speakers from other parts of Italy and (iii) speakers of foreign origin, i.e. first and second-generation immigrants. The corpus amounts to approximately 60 hours of speech; one third of which is contributed by speakers of foreign origin. The collection of data accounts for different languages and language varieties; namely, Italian – either as L1 or L2 – and, to a lesser extent, immigrant minority languages and Piedmontese, as well as other Italo-Romance dialects. Data has been collected through semi-structured interviews about city life and personal experiences (urban initiatives, policies for neighborhoods, leisure time activities, etc.). The corpus is supplemented with a rich set of metadata, with the purpose of fostering the investigation of linguistic variation across socio-economic classes and social groups. The set includes indeed, among others, age, level of education, gender, employment status, place of birth (of both the individual and their parents), mother tongue, knowledge of other languages, and, only as regards first and second-generation immigrants, duration of stay and duration of study in Italy. The occurrence of Italo-Romance dialects and/or foreign languages in speech utterances has been tagged as well. ParlaTO is thus meant to fill some crucial gaps in the *panorama* of Italian

speech corpora. In particular, the spontaneous speech of social groups such as young speakers with low educational qualifications and first and second-generation immigrants can, for the first time, be the subject of targeted corpus-based searches online.

5 Conclusions and future prospects

The ParlaTO corpus has been added to the KIP corpus, turning them into two modules within the larger KIParla corpus. We aim at making this resource grow over time by subsequent additions and upgrades. The leading idea is that the more different types of interactions, speakers, and geographical areas are recorded in the KIParla data, the more the corpus becomes representative of the language(s) and language varieties spoken in Italy. Moreover, the more the corpus gets upgraded over time, the more it may tell us something about the sociolinguistic situation of the peninsula.

We imagine the future development of the corpus along two main directions. On the one hand, we aim at collaborating with existing projects, in order to verify whether data already collected for different purposes may be adapted to constitute new modules of the KIParla corpus. The only requirement in such cases is the traceability and accessibility of a core set of metadata for the speakers (gender, age, geographical information, level of education and occupation) and for the interaction (interview, free conversation, etc.). Further metadata would of course be welcome and searchable. Moreover, new data collections have already started, or are planned to start, in different regions (e.g. in Lombardy). A data collection parallel to ParlaTO is also planned for Bologna.

The second direction along which KIParla will grow has to do with data annotation. For the moment, KIParla data are available as prosodic and orthographic transcriptions, time-aligned with the speech audio file and associated to the metadata of speakers and interactions. Further functionalities are offered by NoSketch Engine, such as word sketches, thesaurus, and keyword computation.

We plan two further stages of annotation, namely lemmatization and POS-tagging, which will significantly increase the searchability of data. For reasons of space limits, here we will not discuss the problems that lemmatization and POS-tagging raise when applied to spoken data (cf. Panunzi, Picchi, Moneglia 2004), and leave such a crucial discussion to future work.

References

Albano Leoni, Federico (2007), "Un frammento di storia recente della ricerca (linguistica) italiana. Il corpus CLIPS". In: *Bollettino d'Italianistica*, IV, (2), 122-130.

Berretta, Monica (1988), "Italienisch: Varietätenlinguistik des Italienischen/Linguistica delle varietà". In: *Lexicon der Romanistischen Linguistik*, vol. IV 762-774.

Berruto, Gaetano (2012), *Sociolinguistica dell'italiano contemporaneo*. *Seconda edizione*, Roma, Carocci.

De Mauro, Tullio, Federico Mancini, Massimo Vedovelli and Miriam Voghera (1993), *Lessico di frequenza dell'italiano parlato*, Milano, Etaslibri.

Jefferson, Gail (2004), "Glossary of transcript symbols with an introduction". In: Lerner, Gene H. (ed.), *Conversation Analysis: studies from the first generation*, Amsterdam, John Benjamins, 13-31.

Tagliamonte, Sali A. (2006), *Analysing sociolinguistic variation*, Cambridge, Cambridge University Press.

Fishman, Joshua (1972), "Domains and the relationship between micro- and macrosociolinguistics. In: Gumperz, John and Dell Hymes (eds.), *Directions in sociolinguistics. The ethnography of communication*, New York, Holt, Rinehart and Winston, 435-453.

Panunzi, Alessandro, Eugenio Picchi and Massimo Moneglia (2004), "Using PiTagger for Lemmatization and PoS Tagging of a Spontaneous Speech Corpus: C-Oral-Rom Italian". In: *Proceeding of Fourth Language Resources and Evaluation Conference* (LREC 2004).

Rychlý, Pavel (2007), "Manatee/Bonito – A Modular Corpus Manager". In: *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, Brno, Masaryk University, 65-70.

Sinclair, John (1991), *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

Voghera, Miriam, Claudio Iacobini, Renata Savy, Francrsco Cutugno, Aurelio De Rosa and Iolanda Alfano (2014), "VoLIP: A searchable Italian spoken corpus". In: Vaselovská, Ludmila and Markéta Marjanebová (eds.), *Complex visibles out there. Proceedings of the Olomouc Linguistics Colloquium: Language use and linguistic structure*, Olomouc, Palacký University, 628-640.