KIParla Corpus: A New Resource for Spoken Italian[1]

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 paper we introduce the main features of the KIParla corpus, a new resource for the study of spoken Italian. In addition to its other capabilities, 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 designed to be shared as a free resource tool through the NoSketch Engine interface 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 new resource tool for the study of spoken Italian. The KIParla corpus is the result of a joint collaboration between the Universities of Bologna and Turin and is open to further partnerships in the future.

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

Section 2 provides a detailed description of the corpus design, aimed at featuring the geographic, social and situational variation that characterizes spoken Italian. In Section 3 we discuss corpus implementation, describing how data have been collected in adherence with ethical requirements, how they have been treated and transcribed, and how they have been made accessible and searchable through NoSketch Engine. 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 current core of KIParla, namely KIP and ParlaTO, are then briefly illustrated, and some prospects for future developments are outlined.

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-geographic profile/status of the speakers and the interactional contexts) in order to build a corpus suitable for investigating (socio)linguistic variation in contemporary Italian.

2.1 Aims

The KIParla corpus is designed to overcome some of the shortcomings that characterize previous resources used in the study of spoken Italian. It is intended to bring about major improvements concerning three key aspects of corpus-based research: (i) access to the speakers' metadata, particularly to those concerning age and social group; (ii) the 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. While some corpora provide general information about the setting of the interaction, at present there is no other corpus of spoken Italian that offers detailed information about single speakers. As for (ii), KIParla will be accessible online through the NoSketch Engine interface, and on the project website it will be possible to download all the recordings (in .wav or .mp3 format) and transcriptions, as previously done for CLIPS (Albano Leoni 2007),VoLIP (Voghera *et al.* 2014), and other corpora. Moreover, with regard to (iii) the research platform will enable users to listen to the results of single queries and download them in .mp3 format, offering text-to-speech alignment.

The philosophy behind KIParla is to pave the way for a collection of spoken corpora, each compiled according to a shared methodology in order to facilitate 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) which grows in size over time thanks to an increasingly wide range of materials.

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

The diatopic dimension has always been considered to be of greatest significance when describing the Italian sociolinguistic scenario (see Berruto 2012 *inter al.*); in fact, speech utterances without any regional features are seldom if ever found even among educated speakers and in formal situations. Currently, the only spoken corpora that take into account geographic variation are the LIP corpus and the CLIPS corpus. In the KIParla corpus, thus far we have collected data in Turin and Bologna; the sociolinguistic situation in 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. One good example of such a scenario is provided in (1); the conversation, recorded in Turin, has two speakers using the progressive periphrasis *stare* + *a* + infinitive combined with the apocopated form of the lexical verb, which are two typical features of regional varieties of Italian spoken in central Italy.

(1) GF_TO091: ho capito ma tu sei entrata troppo nella parte **stai a fa'** l'attrice
"I see but you are getting too much into this, you're putting on an act"

BC_TO089: sì
"yes"

SF_TO090: no non **sto a fa'** l'attrice io parlo così normalmente come potete notare ragazze

"no, I'm not putting on an act. This is the way I usually speak, as you can see girls"

(KIP corpus, TOA3012)

In order to have a deeper understanding of the situation, information regarding both the city in which the data were collected and the place of origin of each speaker can be retrieved.

2.3 The diastratic dimension: a perspective on Italian society

The speakers involved in the recordings are distinguished primarily 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. undergraduates, graduate students, and university professors. In the second data collection sample (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 broadened. Ideally, the incremental nature of the corpus will make it possible 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, i.e. that linguistic practices are often related to specific social activities, we dedicated particular attention to including dif-

ferent types of situations, expecting to find considerable differences between the structures involved in each.

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 the following external factors: (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 these three very general features is particularly helpful in the task of integrating new data 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 be 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 also been involved in the process, but only after a period of specific training. Increasing the number of data collectors is crucial to avoid unwanted bias caused by the inclusion of 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.

Whenever data were being collected, speakers were first 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 European Union's General Data Protection Regulation (G.D.P.R.). The consent form allowed us to collect linguistic material for scientific purposes, to store it in 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 anonymized before being made available to the public. 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 relating to different speakers in a conversation. It also makes it possible to link each annotation to the media timeline. Thanks to this feature of the software, it was possible to implement text-to-speech alignment within the NoSketch Engine interface (§3.3).

Every tier in the transcription refers to an alphanumeric code that links the spoken production of a single speaker to his/her metadata (e.g. age and level of education); similarly, each transcription file is associated with a code that allows its metadata to be traced (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 production and the "searchability" of the written texts. For this reason, we decided to adopt a simplified version of the Jefferson (2004) conventions used in conversation analysis (see Figure 1). An example of this transcription convention is provided in Figure 2.

,	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: Symbols used in the transcription based on Jefferson (2004)

AG_BO097: e mi ha guardato, io l'ho guardato pero' cioe'
GG_BO095: ti ha riconosciuta [si e' visto,]
AG_BO097: [si':: pero',]
AG_BO097: cioe' non non c'e' stato uno sguardo come dire:::
AG_BO097: oh mio dio sei tu della lezione

Figure 2: Conversational transcription as shown in the corpus page

The decision to implement conversational transcription was mainly due to the fact that it enables 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 (e.g. reformulations and 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 NoSketch Engine

The transcriptions obtained through ELAN are in XML format and are automatically time-aligned to the speech audio files; thus, 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 NoSketch Engine interface (Rychlý 2007).

NoSketch Engine is an open-source tool for corpus management which provides a powerful and user-friendly interface to perform corpus searches, 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 NoSketch Engine, we wrote a python script that allows the user to: (i) make the metadata available both as query filters and text information; (ii) search the orthographic and Jefferson transcriptions; (iii) directly link every occurrence with the time-aligned portion of the media file associated with it; (iv) search each module of the corpus separately.

Users can perform a query either by browsing the whole corpus or by selecting one or more metadata concerning the participants or the conversation in which they appear. Figure 3 shows how the metadata can be selected in the corpus. As reported in Figures 4 and 5 respectively, with regard to the KIP module (§ 4.1) conversation metadata include the type of conversation, the city in which it was recorded and the year, the number of participants, and the relationship between them; the participants' metadata include occupation, gender, age, and the region of origin. During data collection, the participants indicated both the city of birth and the city in which they attended high school; however, we decided to retain only the latter information as an indicator of the speakers' region of origin.

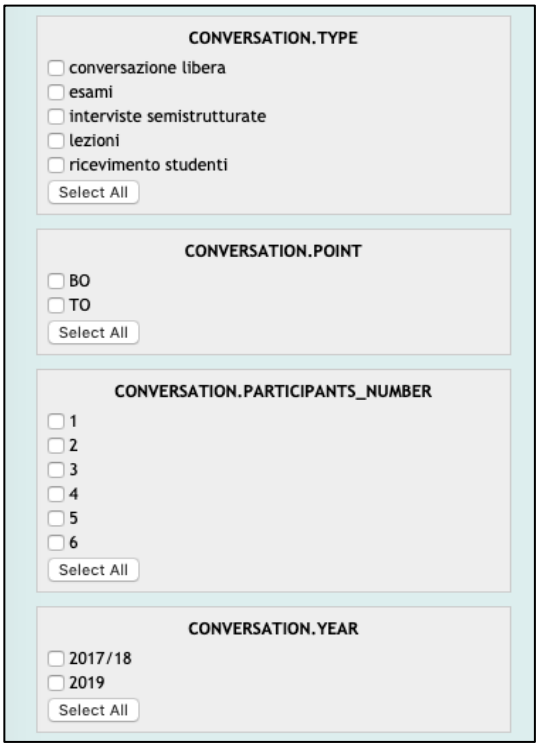

Figure 3: Metadata selection

Type of conversation	Spontaneous conversation
	Exams
	Interviews
	Lessons
	Office hours
City	Bologna
	Turin
Number of participants:	1
	2
	3
	4
	5
	6
Year	2017/18
	2019
Relation between the participants	Asymmetrical
	Symmetrical

Figure 4: Conversation metadata

Figures 6 and 7 provide an example of a query in the NoSketch Engine interface; the results appear in KWIC (Keyword-In-Context) format, in which each token is presented within a string of characters containing the words that precede and follow it. By clicking on the conversation name reported in blue in the left portion of the screen, users can access the conversation's metadata, a full transcription of the file, both in Jefferson and text-only format, and a link to the corresponding

audio file (see Figure 6). By clicking on the token, in red, users can open a text box which provides further context (see Figure 7).

Occupation	Professor
	Student
Gender	Male
	Female
Region	Abruzzo
	Basilicata
	Calabria
	...
Age bracket	Under 25
	26-30
	31-35
	36-40
	41-45
	46-50
	51-55
	56-60
	Over 60

Figure 5: Participants' metadata

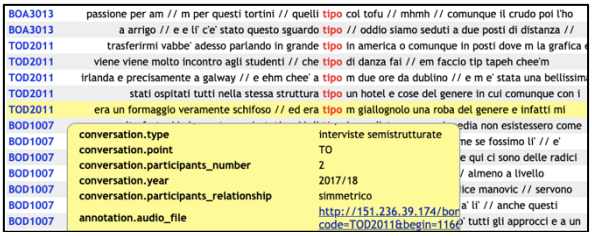

Figure 6: Conversation metadata

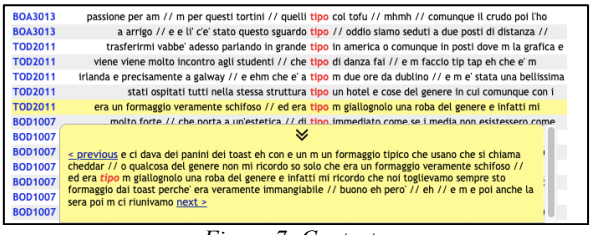

Figure 7: Context

As of September 2019, the corpus can be accessed online at the website www.kiparla.it. At present, it only consists of the KIP module (see 4.1), but further modules are already being processed and will be uploaded to the same website (see below). The corpus has not yet been lemmatized or POS-tagged, but such steps are planned for the near future.

4 Incremental modularity: an accessible open monitor corpus of spoken Italian

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

Modules contain different corpora of Spoken Italian sharing the same design and a common set of metadata (see §2) which have been 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.

Easy access to all of the metadata makes the corpus *expandable*, through the addition of further modules focusing on different geographical, socio-cultural, or communicative aspects, and *upgradable*, through the addition of new data to existing modules. Such a dynamic nature of the KIParla corpus makes it a potential monitor corpus, open to additions and upgrades over 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 conceived as a self-sufficient unit. It consists of approximately 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 only includes speakers whose achievements pertain to higher education, namely university students and professors. The social characteristics of the speakers are clearly reflected in speech data, e.g. in the highly educated use of the relative clause in example (2).

(2) LB_BO100: abbiamo una struttura di dati, abbiamo un algoritmo **attraverso il quale** ci muoviamo tra queste strutture di dati

"we have a data structure, we have an algorithm **through which** we move among these data structures."

(KIP corpus, BOD1007)

The structure of this subcorpus is intended to maximize diaphasic variability, according to the parameters described in 2.4 (symmetrical *vs* asymmetrical relations; presence *vs* absence of a

moderator; presence *vs* absence of a fixed topic). This resulted in the selection of the contexts listed in Figure 8, 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 8: Hours recorded for each interaction type in Turin and Bologna

The complete KIP module is currently available on the www.kiparla.it website.

4.2 ParlaTO module

ParlaTO is a corpus of spontaneous speech collected in Turin between 2018 and 2019. The corpus is being compiled in an effort to portray 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 migration patterns.

Basically, the corpus contains 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. Accordingly, 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. Therefore, the corpus makes it possible to investigate a wide range of phenomena.Below are just a couple of examples of Italian as L1: a case of substratum interference in (3), i.e. the absence of a preverbal negative marker (which characterizes most Northern Italo-Romance dialects), and a typical feature of uneducated speech in (4), i.e. the use of *ci* as 3pl indirect object clitic pronoun.

(3) PST035: in quei tempi q- c'era proprio niente da mangiare

"in those days there was really nothing to eat"

(ParlaTO corpus, PTB009)

(4) PMM017: c'erano gli altri ragazzi **ci** ho fatto dei nomi

"the other boys were there, I gave **them** some names"

(ParlaTO corpus, PTB002)

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 provides a rich set of metadata, geared to fostering the investigation of linguistic variation across socio-economic classes and social groups. It includes such categories as age, level of education, gender, employment status, place of birth (of both the individual and their parents), mother tongue, and knowledge of other languages, as well as duration of stay and duration of study in Italy for first and second-generation immigrants. The occurrence of Italo-Romance dialects and/or foreign languages in speech utterances is being 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 such social groups as young speakers with limited educational qualifications and first and second-generation immigrants can, for the first time, be the subject of targeted corpus-based searches online.

The corpus currently amounts to approximately 60 hours of speech, one third of which is from speakers of foreign origin. However, ParlaTO is still under construction and will not be available online until early 2020.

5 Conclusions and future prospects

The ParlaTO corpus has been added to the KIP corpus, thereby creating two modules within the larger KIParla corpus. We aim to make this resource grow over time through subsequent additions and upgrades. The leading idea is that the greater the variety of interactions, speakers, and geographical areas recorded in the KIParla data, the more the corpus will become representative of the language(s) and language varieties spoken in

Italy. Moreover, as the corpus is upgraded over time, it will tell us more and more about the sociolinguistic situation in the Italian peninsula.

We envision the future development of the corpus to proceed in two main directions. On the one hand, we intend to collaborate with existing projects, in order to verify whether data already collected for different purposes may be adapted into new modules of the KIParla corpus. The only requirement in such cases is the ability to trace and access 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. Moreover, new data collection efforts have already started or are scheduled to start in different regions (e.g. in Lombardy). A data collection project 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 linked to the metadata of speakers and interactions. Further functions 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 enhance data retrieval. Due to space constraints, we are unable to 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.

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.

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.

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.