# Il corpus KIParla.
# Tra linguistica dei corpora e sociolinguistica dell'italiano

*Caterina* MAURI[*], *Silvia* BALLARÈ[**], *Eugenio* GORIA[**], *Massimo* CERRUTI[**]

**ABSTRACT** • ***The KIParla Corpus. Between Corpus Linguistics and Sociolinguistics of the Italian Language***. In this paper we introduce the main features of the KIParla corpus, a new resource for the study of spoken Italian. Among other specific features, 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.

**KEYWORDS** • Corpora; Corpus Linguistics; Spoken Italian; Italian language; Sociolinguistics.

Il corpus KIParla è una risorsa elettronica per lo studio dell'italiano parlato di recente pubblicazione, frutto della collaborazione tra l'Università di Torino e l'Università di Bologna, e aperto a futuri contributi provenienti da altri gruppi di ricerca.

Il KIParla si distingue da altre risorse attualmente disponibili per lo studio dell'italiano parlato per alcune proprietà; fra le altre, la possibilità di avere accesso a una serie di metadati relativi alle caratteristiche socio-demografiche dei parlanti e al tipo di interazione in cui essi sono coinvolti, e l'opportunità di consultare i dati sia in formato audio sia in formato testuale. La risorsa è costruita, inoltre, in maniera tale da rendere possibili futuri ampliamenti, sotto forma di nuovi moduli parzialmente indipendenti ma che condividano uno stesso nucleo di metadati e lo stesso sistema di raccolta e gestione dei dati. Infine, il KIParla è una risorsa di libero accesso che si avvale della piattaforma di interrogazione NoSketch Engine (Rychlý 2007).

## 2. Progettazione del corpus

Il corpus KIParla è costituito da materiali linguistici registrati, fino ad ora, nelle città di Bologna e di Torino. I due punti di inchiesta presentano una situazione sociolinguistica per certi versi analoga, caratterizzata dalla compresenza non soltanto delle varietà locali di italiano e dialetto ma anche di altri italiani regionali e dialetti italiani, dal momento che entrambe le città sono e sono state meta di mobilità interna.

In fase di raccolta dati sono state registrate diverse informazioni relative ai parlanti, come ad es. luogo di origine, età, titolo di studio, occupazione. Il corpus comprende poi vari tipi di interazione verbale, corrispondenti a diverse situazioni comunicative, classificate essenzialmente secondo i seguenti parametri:

- relazione simmetrica/asimmetrica tra i partecipanti;
- presenza/assenza di un argomento predefinito;
- presenza/assenza di norme per la presa dei turni di parola.

### 3. La costruzione del corpus: raccolta dati, trascrizione e accessibilità

La raccolta dati è stata effettuata da ricercatori e studenti (debitamente formati) delle Università di Bologna e di Torino. Tutte le interazioni sono state registrate a microfono palese e gli informanti coinvolti hanno firmato un consenso informato (conforme alle norme europee di protezione dati – v G.D.P.R.).

Le trascrizioni sono state effettuate utilizzando il software ELAN (Sloetjes and Wittenburg 2008), che permette l'allineamento delle trascrizioni alle relative tracce audio; inoltre, per dare conto di alcune caratteristiche intrinseche della comunicazione parlata (ad esempio l'uso dell'intonazione e la sovrapposizione tra turni di diversi parlanti), si è scelto di seguire una versione semplificata del sistema Jefferson (Jefferson 2004), frequentemente impiegato nell'analisi della conversazione.

Prima della pubblicazione, i materiali linguistici (sia i file audio sia le trascrizioni) sono stati anonimizzati: l'unico dato sensibile direttamente accessibile è la voce stessa del parlante.

Una volta ultimata la raccolta e la trascrizione dei dati, è stato elaborato uno script in python che permette di consultare i dati sulla piattaforma NoSketch Engine, consentendo all'utente di:

- utilizzare i metadati (relativi ai parlanti e alle conversazioni) sia come filtri di ricerca sia come informazioni relative alle singole registrazioni;
- collegare l'occorrenza ricercata con l'unità intonativa in cui si trova;
- avere accesso all'intera trascrizione (ortografica e secondo il sistema Jefferson) della conversazione in cui si trova l'occorrenza cercata;
- effettuare ricerche considerando la semplice trascrizione ortografica;
- consultare separatamente ogni modulo.

### 4. Modularità incrementale

Il corpus KIParla è caratterizzato da una modularità incrementale, ovvero è organizzato al suo interno in moduli fra loro indipendenti: è dunque possibile aggiungere progressivamente nuovi moduli a quelli esistenti. I moduli sono da intendere come (sotto)corpora di parlato che condividono (almeno) un *core set* di metadati, presentano una trascrizione effettuata originariamente tramite ELAN, e offrono la consultazione attraverso NoSketch Engine. I vari (sotto)corpora possono concentrarsi su diverse varietà di lingua e/o diversi punti di inchiesta; il disporre di una procedura condivisa per la raccolta e il trattamento dei dati garantisce del resto un alto livello di comparabilità tra i moduli.

Ad oggi, il corpus KIParla è costituito da due (sotto)corpora (v. Fig. 1):

- KIP;
- ParlaTO.

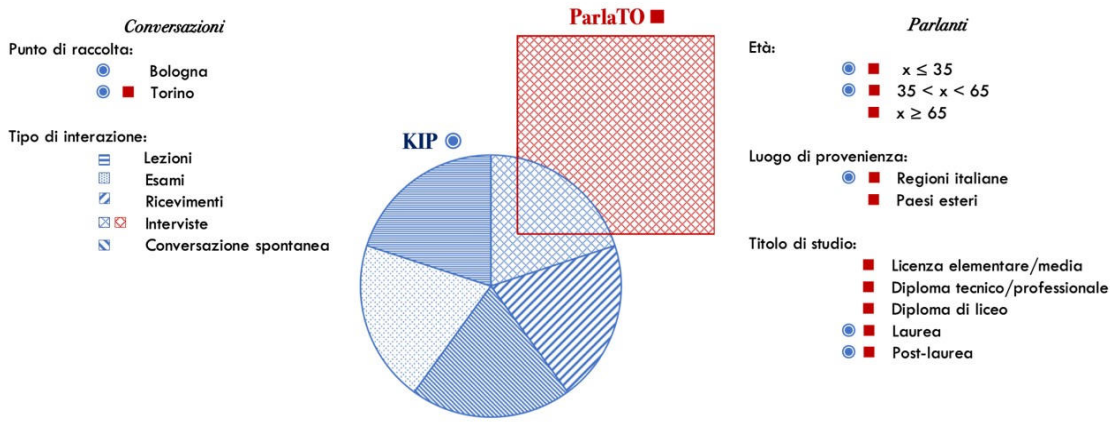


Figura 6. I moduli attuali del corpus KIParla

### *4.1. Il modulo KIP*

Il modulo KIP, concepito inizialmente come unità autosufficiente, è stato allestito nell'ambito del progetto *LEAdhoC – Linguistic expression of ad hoc categories* (2015-2019, SIR n. RBSI14IIG0) e rappresenta il nucleo originario del corpus KIParla. La risorsa è costituita da circa 70 ore di parlato raccolte a Bologna e a Torino in contesto universitario; le interazioni sono state registrate in diverse situazioni comunicative (v. Fig. 1) e hanno coinvolto studenti e professori universitari. In virtù della gamma dei contesti interazionali considerati, il corpus KIP offre in primo luogo l'opportunità di condurre ricerche su aspetti e fenomeni di variazione diafasica (specialmente di registro) nel parlato di soggetti colti. In Tab. 1 si riportano la struttura del KIP e le ore registrate per ciascun contesto.

| Attività | Bologna | Torino |
|---|---|---|
| conversazioni | 10:00:37 | 06:22:24 |
| esami | 03:09:34 | 03:10:48 |
| lezioni | 12:19:39 | 13:25:33 |
| interviste | 06:18:37 | 07:47:38 |
| ricevimenti | 02:59:11 | 03:49:08 |
| TOT | 34:47:38 | 34:35:30 |

Tabella 7. Il modulo KIP

La costruzione del corpus KIP è iniziata nel 2016 e si è conclusa nel 2019. Attualmente, il KIP è il solo modulo accessibile e consultabile online.

### *4.2. Il modulo ParlaTO*

Il corpus ParlaTO è in via di allestimento dal 2018 nell'ambito di un progetto omonimo (*ParlaTO – Corpus plurilingue del parlato di Torino*, Fondazione CRT, E.O. 2018, ID63411). Le produzioni linguistiche che confluiranno nel corpus sono state raccolte a Torino per mezzo di interviste semi-strutturate, individuali o di gruppo, a più di un centinaio di informatori con diversa provenienza geografica (essenzialmente: parlanti di origine piemontese, parlanti originari di altre regioni d'Italia, e parlanti di origine straniera) e diversa collocazione sociale (v.

Fig. 1). Il corpus consisterà in oltre 70 ore di parlato e sarà provvisto di un ampio set di metadati relativi alle caratteristiche socio-demografiche dei parlanti, come l'età, il titolo di studio, il genere, l'occupazione, il luogo di nascita (dell'informatore e dei genitori), la lingua materna e, per i parlanti di origine straniera, il tempo di permanenza e gli anni di studio in Italia. Il corpus ParlaTO offrirà quindi in primo luogo la possibilità di indagare aspetti di differenziazione sociale dell'italiano parlato, oltre all'opportunità di condurre ricerche 'mirate' su categorie sociali specifiche.

La consultazione on line del corpus ParlaTO è prevista per la primavera del 2020.

### *4.3. Prospettive future*

Le dimensioni e la rappresentatività del corpus KIParla potrebbero crescere nel corso del tempo grazie alla collaborazione di altri ricercatori; nuovi (sotto)corpora potranno via via essere aggiunti a quelli esistenti, in virtù della condivisione di una serie di caratteristiche operative e metodologiche.

In futuro, inoltre, è prevista la lemmatizzazione e il pos-tagging dei dati del corpus.

### RIFERIMENTI BIBLIOGRAFICI

Jefferson, G. (2004), Glossary of transcript symbols with an introduction, in G.H. Lerner (ed.), Conversation Analysis: studies from the first generation, Amsterdam, John Benjamins, pp. 13-31.

Rychlý, P. (2007), *Manatee/Bonito – A Modular Corpus Manager*, in *1st Workshop on Recent Ad-vances in Slavonic Natural Language Processing*, Brno, Masaryk University, pp. 65-70.

Sloetjes, H. e P. Wittenburg (2008), *Annotation by category – ELAN and ISO DCR*, in *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC 2008).