DOI: 10.17469/O2106SLI000006

CATERINA MAURI, SILVIA BALLARÉ, EUGENIO GORIA, MASSIMO CERRUTI

Il corpus KIParla

In questo articolo presentiamo le principali caratteristiche del CORPUS KIParla, una nuova risorsa per lo studio dell'italiano parlato, liberamente accessibile al sito www.kiparla.it. Il corpus è stato progettato per essere gratuitamente consultato attraverso l'interfaccia NoSketch Engine e per essere espanso nel tempo tramite l'aggiunta di nuovi moduli. Il corpus KIParla fornisce l'accesso a una vasta gamma di metadati che caratterizzano sia i partecipanti che le interazioni, utilizzabili come filtri di ricerca. Al momento il KIParla consiste di due moduli (KIP e ParlaTO), che permettono di effettuare ricerche sulla variazione diafasica, diatopica e diastratica dell'italiano contemporaneo.

Parole chiave: corpus, italiano parlato, sociolinguistica.

1. *Introduzione: il corpus KIParla*

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 video della DEMO è accessibile online: https://underline.io/lecture/33036-d12---il-corpus-kiparla

Il KIParla si distingue da altre risorse attualmente disponibili per lo studio dell'italiano parlato per alcune proprietà; fra le altre, la possibilità di avvalersi di 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. Le due città 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, oltre che di italiano di non nativi e lingue di recente immigrazione; entrambe le città, infatti, sono e sono state meta di mobilità interna e migrazione.

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 dei dati, trascrizione e accessibilità*

La raccolta dati è stata effettuata da ricercatori, ricercatrici, studenti e studentesse delle Università di Bologna e di Torino. Tutte le interazioni sono state registrate a microfono palese e i soggetti coinvolti hanno firmato un consenso informato (conforme alle norme europee di protezione dati – v. G.D.P.R.), che autorizza il gruppo di lavoro a utilizzare i dati raccolti per finalità di ricerca, ad archiviarli e condividerli in forma parzialmente anonimizzata. Per questo motivo, 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.

Le trascrizioni sono state effettuate utilizzando il software ELAN (Sloetjes & 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.

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.

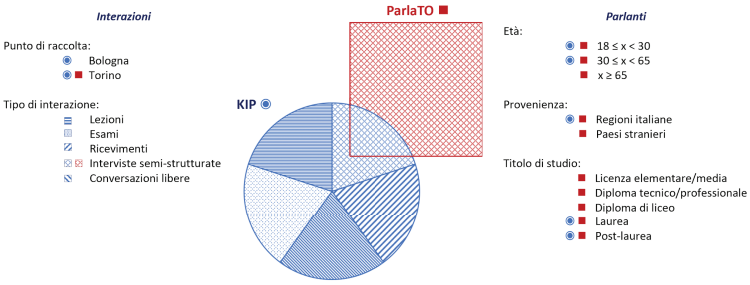
4. *Struttura modulare e incrementale del corpus KIParla*

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), il KIP e il ParlaTO.

Il KIP offre innanzitutto la possibilità di indagare fenomeni di variazione diafasica, specialmente di registro, dell'italiano nel parlato di soggetti colti; mentre il ParlaTO offre in primo luogo l'opportunità di esplorare aspetti di diversificazione diastratica dell'italiano parlato. Entrambi i corpora, poi, includono produzioni di parlanti con provenienza geografica diversa; consentono perciò di osservare almeno alcune manifestazioni della variazione diatopica dell'italiano. Con il KIParla, nel complesso, si ha quindi la possibilità di indagare aspetti di diversificazione geografica (KIP e ParlaTO), sociale (ParlaTO) e, limitatamente a parlanti colti, situazionale (KIP) dell'italiano parlato.

Figura 1 - *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 sua costruzione è iniziata nel 2016 e si è conclusa nel 2019.

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 consente in primo luogo di condurre ricerche su aspetti e fenomeni di variazione diafasica nel parlato di soggetti colti. Nella Tab. 1 si riportano la struttura del KIP, le ore registrate per ciascun contesto, il numero di informatori, le dimensioni e i metadati raccolti.

Tabella 1 - *Il modulo KIP*

Attività	*Bologna*	*Torino*	*TOT*
Conversazioni libere	10:00:37	06:22:24	16:23:01
Esami	03:09:34	03:10:48	6:20:22
Lezioni	12:19:39	13:25:33	25:45:12
Interviste semistrutturate	06:18:37	07:47:38	14:06:15
Ricevimento studenti	02:59:11	03:49:08	6:48:19
TOT	34:47:38	34:35:30	**69:23:08**
Informatori	150	123	273

Metadati	*Parlanti:* classe d'età, sesso, regione in cui si sono frequentate le scuole superiori, occupazione (studente/professore). *Interazioni:* tipo di interazione, relazione tra i partecipanti (simmetrica/asimmetrica), presenza di un moderatore (sì/no), argomento (libero o no), numero di partecipanti.
Dimensioni	661.175 tokens

Con conversazioni libere si intende la registrazione di parlato conversazionale spontaneo tra studenti, raccolto senza alcuna indicazione da parte del gruppo di ricerca. Per ottenere un tipo di evento comunicativo che si avvicinasse il più possibile alla situazione desiderata, la registrazione delle conversazioni è stata svolta nella maggior parte dei casi coinvolgendo direttamente studenti e studentesse delle Università coinvolte: in particolare è stato chiesto loro di registrare autonomamente momenti ricreativi di vario tipo (ad esempio pause durante lo studio, cene, ...) in cui fossero coinvolti altri membri della loro rete sociale, anch'essi studenti universitari.

Esami e ricevimento studenti fanno riferimento ai due contesti più tipici in cui è possibile osservare uno scambio comunicativo a due tra studente e docente. Si tratta in entrambi i casi di attività con un grado maggiore di strutturazione, che prevedono il raggiungimento di scopi concreti da entrambe le parti. Per il ricevimento studenti cfr. ad esempio il lavoro in prospettiva conversazionale di Limberg (2010). Da un punto di vista sociolinguistico si considerano questi eventi come caratterizzati da un maggiore grado di formalità rispetto al parlato spontaneo colloquiale, e da una maggiore presenza dei sottocodici legati alle discipline trattate di volta in volta.

Le lezioni costituiscono il tipo di interazione in cui sono maggiormente coinvolti i docenti, e sono presenti pochissimi interventi da parte degli studenti presenti a lezione, che, in questo caso, sono stati anche esclusi dalla raccolta dei metadati. Il parlato è perlopiù monologico, e dunque caratterizzato da una maggiore pianificazione del turno di parola, e da un maggiore apporto della varietà scritta, ad esempio nel caso di brani letti e commentati o della presenza delle slides.

Infine, l'intervista semi-strutturata è stata inserita in modo da poter includere nel corpus anche interazioni studente-studente caratterizzate da un maggiore grado di strutturazione e da ruoli esplicitamente formalizzati, appunto l'intervistato e l'intervistatore. Infatti, è stato chiesto a studenti e studentesse del tirocinio di intervistare membri della propria rete sociale sulla base di una traccia prestabi-

lita di temi e questioni. In una prima parte, l'intervistatore chiede di descrivere la propria casa facendo un paragone con le altre case in cui gli intervistati hanno vissuto. Successivamente, viene chiesto agli intervistati di raccontare un evento particolarmente significativo legato a una delle case in cui hanno vissuto, in modo da ottenere anche una parte di parlato monologico. Nella Tab. 2 vengono mostrati i metadati che possono essere usati come filtri di ricerca:

Tabella 2 - *KIP: i metadati usabili come filtri di ricerca*

Parlanti	Classe d'età:	under25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-55, 56-60, over60
	Sesso:	M, F
	Regione delle scuole superiori:	Abruzzo, Basilicata, Calabria, Campania, Emilia-Romagna, Estero, Friuli-Venezia Giulia, Lazio, Liguria, Lombardia, Marche, Molise, Piemonte, Puglia, Sardegna, Sicilia, Toscana, Trentino-Alto Adige, Umbria, Valle d'Aosta, Veneto
	Occupazione:	p (professore), s (studente)
Interazioni	Tipo:	conversazione libera, esami, interviste semistrutturate, lezioni, ricevimento studenti
	Luogo:	Bologna, Torino
	Relazione:	Simmetrica, asimmetrica
	Partecipanti:	1, 2, 3, 4, 5, 6
	Moderatore:	Sì, no
	Topic:	Fisso, libero

4.2 Il modulo ParlaTO

Il corpus ParlaTO è stato allestito nell'ambito di un progetto omonimo (*ParlaTO – Corpus plurilingue del parlato di Torino*, Fondazione CRT, E.O. 2018, ID63411) ed è confluito nel più ampio KIParla a settembre 2020. Il ParlaTO è composto essenzialmente da una serie di conversazioni registrate a Torino per mezzo di interviste semi-strutturate. Le interazioni hanno coinvolto parlanti d'età diversa, alcuni di origine italiana (piemontese e non), altri di origine straniera, e con livelli d'istruzione e tipi di occupazione differenti. Le interviste hanno affrontato esperienze personali di vita in città (studio, lavoro, attività nel tempo libero o in pensione, ricordi del passato, ecc.) e hanno visto all'opera più intervistatori, ciascuno dei quali quasi sempre apparte-

nente alla rete sociale dell'intervistato; a una stessa intervista, inoltre, hanno spesso partecipato più intervistati. Il che ha favorito modalità di conversazione tipiche del comportamento *in-group*, caratterizzate dall'impiego di varietà spontanee (e, in certi casi, dall'uso alternato di lingue diverse).

Il corpus comprende due sezioni, contenenti l'una le interazioni con parlanti di origine italiana e l'altra le interazioni con parlanti di origine straniera. Al momento è interrogabile online soltanto la prima sezione, che è costituita di circa 50 ore di parlato e include produzioni in italiano, piemontese e altri dialetti (specialmente di area meridionale). L'uso dell'italiano è largamente prevalente in tutte le conversazioni, ed esclusivo nella maggior parte di esse; il piemontese e gli altri dialetti compaiono invece soltanto occasionalmente, secondo la fenomenologia del discorso bilingue. La seconda sezione, che è in via di allestimento, ammonterà anch'essa a circa 50 ore di parlato e verterà sull'italiano di nativi e non nativi e sulle lingue di recente immigrazione. Le Tabelle 3 e 4 forniscono alcune informazioni essenziali riferite alla prima sezione del corpus.

Tabella 3 - *La prima sezione del corpus ParlaTO*

Informatori	88 parlanti	
Raccolta dati	Torino, 2018-2020	
Metodo	Interviste semi-strutturate (individuali e di gruppo)	
Lingue	Italiano, dialetto piemontese, altri dialetti	
Metadati	Parlanti: classe d'età, sesso, regione di nascita, titolo di studio, occupazione, lingua materna, competenza (attiva o passiva) di dialetto/i italo-romanzo/i, competenza (attiva o passiva) di altre lingue, città e quartiere di residenza, città di nascita del padre e della madre. Interazioni: numero di partecipanti, lingue impiegate	
Ore di parlato	Giovani ($18 \leq x \leq 30$):	17:33:20
	Adulti ($30 < x \leq 60$):	14:49:53
	Anziani ($60 < x \leq 89$):	16:15:31
Dimensioni	552.461 tokens	

Il corpus è dotato di metadati relativi alle caratteristiche sociodemografiche dei parlanti e ad aspetti dell'interazione (v. Tabella 3). Alcuni metadati, quali ad es. la classe d'età, il sesso, la regione di nascita, il titolo di studio e l'occupazione del parlante, sono usabili a tutti gli effetti come filtri di ricerca (v. Tabella 4); altri, quali il quartiere di

residenza dell'informatore o la città di nascita del padre e della madre, sono accessibili soltanto come informazioni supplementari (contenute in tabelle Excel scaricabili dal sito web del corpus). La seconda sezione del corpus avrà inoltre alcuni metadati pertinenti ai parlanti di origine straniera, quali il tempo di permanenza e gli anni di studio in Italia.

Tabella 4 - *La prima sezione del corpus ParlaTO: i metadati usabili come filtri di ricerca*

Parlanti	Classe d'età:	16-20, 21-25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-55, 56-60, 61-65, 66-70, 71-75, 76-80, 81-85, 85-90
	Sesso:	M, F
	Regione di nascita:	Abruzzo, Basilicata, Friuli-Venezia Giulia, Lazio, Lombardia, Piemonte, Puglia, Sardegna, Sicilia, Trentino-Alto Adige, Veneto
	Titolo di studio:	*elem* (diploma di scuola elementare), *medie* (licenza media), *it* (diploma di istituto tecnico o professionale), *lic* (diploma di liceo), *laurea* (laurea triennale, magistrale e a ciclo unico), *phd* (dottorato di ricerca)
	Occupazione:	*artig* (artigiani, operai specializzati e agricoltori), *comm* (professioni qualificate nelle attività commerciali e nei servizi), *disocc* (disoccupati), *impr* (legislatori, imprenditori e alta dirigenza), *intell* (professioni intellettuali, scientifiche e di elevata specializzazione), *nonq* (professioni non qualificate), *oper* (conduttori di impianti, operai di macchinari fissi e mobili e conducenti di veicoli), *pens* (pensionati), *stud* (studenti)
Interazioni	Partecipanti:	2, 3, 4, 5, 6
	Lingue:	italiano, italiano e dialetto

4.3 Nuovi moduli

Attualmente, grazie alla collaborazione degli studenti e delle studentesse dell'Università di Bologna che hanno preso (e stanno prendendo) parte al tirocinio curricolare KIParla, sono in fase di allestimento due nuovi moduli.

Il primo, denominato KIPPasti, è costituito da registrazioni di parlato spontaneo effettuate durante il corso di pranzi e cene. L'ideazione del modulo è avvenuta nella primavera 2020 quando, a causa della situazione pandemica, non era possibile raccogliere dati linguistici di parlato che prevedessero il contatto con membri esterni al proprio nucleo ristretto di conviventi. Il quadro complessivo, dunque, ci ha

portati a pensare ad un tipo di raccolta dati che potesse essere svolto in sicurezza dagli studenti coinvolti nel progetto e, parallelamente, che potesse essere di interesse per ricerche future.

Inoltre, nella larga maggioranza dei casi, i tirocinanti coinvolti nella raccolta dati si trovavano nei loro luoghi di origine. Per questa ragione, si è deciso di cercare di bilanciare il corpus in base all'area geografica di registrazione (nord, centro, sud e isole). In questo momento, sono in via di raccolta le ultime registrazioni necessarie per giungere al bilanciamento e sono in fase avanzata le trascrizioni del materiale raccolto. Ad oggi, il modulo è composto da oltre 50 registrazioni per un totale di oltre 30 ore.

Il secondo modulo, per cui attualmente si sta procedendo alla raccolta dati, è denominato ParlaBO e mira ad avere una struttura analoga a quella del ParlaTO ma ha come unico punto di inchiesta la città metropolitana di Bologna. Fino ad oggi sono state raccolte oltre 20 interviste per un totale di circa 19 ore di registrazione.

4.4 Prospettive future

Al momento, sono in corso collaborazioni con colleghe e colleghi di altri atenei, e dunque si prevede la prossima pubblicazione di ulteriori moduli con dati raccolti in diverse città italiane. Le dimensioni e la rappresentatività del corpus KIParla sono destinate a crescere nel tempo.

In futuro, inoltre, è prevista la lemmatizzazione e il pos-tagging dei dati del corpus (v. già Bosco *et al.* 2020).

Riferimenti bibliografici

Bosco, Cristina & Ballarè, Silvia & Cerruti, Massimo & Goria Eugenio & Mauri Caterina. 2020. "KIPoS @ EVALITA2020: Overview of the Task on KIParla Part of Speech Tagging". In Basile, Valerio & Croce, Danilo & Maro, Maria & Passaro Lucia C. (eds.), *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, 489-495. CEUR.org, http://ceur-ws.org.

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

Limberg, Holger. 2010. *The interactional organization of academic talk: office hour consultations*. Amsterdam , John Benjamins.

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

Sloetjes, Han & Wittenburg, Peter. 2008. Annotation by category – ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 816-820.