

Recensione

Silvia Ballarè, Eugenio Goria, Caterina Mauri 2022. *Italiano parlato e variazione linguistica. Teoria e prassi nella costruzione del corpus KIParla*. Pàtron

Book review

Silvia Ballarè, Eugenio Goria, Caterina Mauri 2022. *Italiano parlato e variazione linguistica. Teoria e prassi nella costruzione del corpus KIParla*. Pàtron

Daniela Mereu

Università degli Studi di Torino (Torino, Italia)

daniela.mereu@unito.it

Ricevuto il 5/6/2023, accettato il 12/6/2023, pubblicato il 17/10/2023 in base ai termini della licenza *Creative Commons Attribution 4.0 International* (CC BY 4.0)

Come citare questa recensione

Mereu, Daniela 2023. Recensione. Silvia Ballarè, Eugenio Goria, Caterina Mauri 2022. *Italiano parlato e variazione linguistica. Teoria e prassi nella costruzione del corpus KIParla*. Pàtron. *Studia linguistica romanica* 2023.10, 39-44. <https://doi.org/10.25364/19.2023.10.3>.

[1] Il recente lavoro di Silvia Ballarè, Eugenio Goria e Caterina Mauri (128 pp.) si inserisce nella collana *Linguistica e Linguistiche*, diretta da Nicola Grandi e Rema Rossini Favretti. Questa collana ospita volumi di carattere scientifico, ma contraddistinti anche da un taglio divulgativo, in modo da poter essere fruiti non solo da un pubblico strettamente specialistico.

[2] La monografia si articola in quattro capitoli, preceduti da un'introduzione. Ogni capitolo si conclude con uno specchietto in cui vengono ripresi i punti salienti. Il libro si sviluppa a partire dall'esperienza di costruzione del corpus *KIParla*, un corpus di parlato italiano accessibile liberamente dalla piattaforma *No-Sketch Engine* (Rychlý 2007). A partire da questa esperienza, gli autori offrono sia una trattazione teorica sulla rilevanza del parlato per gli studi linguistici e socio-linguistici, sia un'illustrazione delle metodologie impiegate per la pianificazione e la costruzione della risorsa. Il volume presenta infine una sezione maggiormente incentrata sulla ricerca, nella quale trovano spazio i lavori già condotti sui dati del *KIParla*. Da questo resoconto finale emergono le numerose prospettive di ricerca rese possibili dal corpus.

[3] Il primo capitolo (pp. 11-24) affronta il tema della centralità del parlato per gli studi linguistici negli ambiti della variazione inter- e intralinguistica e del cambiamento diacronico. Nel quadro presentato, particolare rilievo assumono la dimensione temporale e la dialogicità del parlato:

[Q]uando leggiamo una frase scritta, non abbiamo la possibilità di vedere sul nascere le strutture linguistiche che la formano, né di seguirne lo sviluppo temporale; vediamo piuttosto un 'progetto' già ultimato e non più modificabile. Al contrario, la lingua parlata, in virtù del suo radicamento nell'interazione faccia a faccia e nella dimensione temporale, offre svariate possibilità di intervenire in vario modo sulle unità che vengono prodotte, per ampliare, correggere, modificare le unità prodotte durante lo scambio comunicativo. (pp. 12-13)

In questa prospettiva, il parlato fornisce in primo luogo un punto di vista privilegiato per la ricerca sugli universali del linguaggio e sul mutamento linguistico, perché consente di descrivere l'emergere della grammatica a partire dall'interazione discorsiva e di rilevare i mutamenti linguistici nelle loro prime fasi. La visione più completa offerta dalla dimensione parlata, rispetto a quella scritta, emerge sia in prospettiva sincronica, perché il parlato consente di osservare aspetti della lingua assenti nello scritto, sia nell'ambito del mutamento linguistico, in quanto attraverso le varietà parlate è possibile cogliere le fasi iniziali di mutamenti che si propagano solo successivamente a livello scritto. In seconda battuta, grazie al parlato è anche «possibile cogliere tratti e costruzioni 'in movimento' che emergono dagli usi concreti e/o si diffondono in più varietà sociolinguistiche. Per questo motivo, i dati di parlato costituiscono la fonte più preziosa per il sociolinguista che voglia ricercare processi di variazione in atto nell'italiano contemporaneo» (p. 22).

[4] La variazione sociolinguistica e le modalità in cui questa si manifesta costituiscono i temi principali del secondo capitolo (pp. 25-35). Le dimensioni di variazione diatopica, diafasica e diastratica vengono illustrate nei loro punti essenziali e messe in relazione alla situazione dell'italiano contemporaneo. Un elemento chiave del corpus, che lo rende adatto a essere usato per indagini sociolinguistiche, è rappresentato dai metadati, che consentono di fare delle ricerche definendo le variabili macro-sociolinguistiche di riferimento (p. es. origine, sesso, età, occupazione lavorativa) o situazionali (p. es. tipo di evento comunicativo, come intervista, esame, lezione universitaria, etc.). La loro presenza rende possibile, perciò, sia la mera consultazione dei dati, sia l'associazione del dato linguistico alle caratteristiche sociali del parlante che l'ha prodotto e, di conseguenza, consente di individuare eventuali correlazioni tra fenomeni linguistici e categorie sociali o situazionali.

[5] Il terzo capitolo (pp. 37-75) presenta un quadro dettagliato delle fasi di progettazione e costruzione del corpus *KIParla*. In queste pagine vengono descritte le scelte metodologiche operate durante la realizzazione del corpus, dalla raccolta dati fino alla loro pubblicazione. Attualmente il *KIParla* è costituito da due moduli, il *KIP* e il *ParlaTO*, e comprende materiali raccolti nelle città di Bologna e Torino, per le quali fino a questo momento non esistevano delle basi di dati di parlato spontaneo.

[6] Nello specifico, i dati del modulo *KIP*, circa 70 ore di parlato (661.175 tokens), sono stati raccolti nelle città di Bologna e Torino e consistono in registra-

zioni di situazioni comunicative caratterizzate da diversi gradi di formalità e di pianificazione (p. es. lezioni universitarie, esami, ricevimento studenti), ma accomunate dal contesto universitario. In ragione della sua struttura, questo modulo si rivela particolarmente adatto allo studio della variazione diafasica. I dati del *ParlaTO*, circa 49 ore di parlato (552.461 tokens) registrate a Torino e provincia, sono stati raccolti mediante interviste semi-strutturate. Visto che i parlanti intervistati provengono da gruppi socioculturali diversi, il *ParlaTO* si presta a essere usato per dare conto della differenziazione diastratica del parlato torinese.

[7] I dati del *KIParla* sono stati trascritti ortograficamente con l'uso del software *ELAN* (Sloetjes & Wittenburg 2008) adottando una versione semplificata del sistema Jefferson (2004). I file sonori e le rispettive trascrizioni, una volta anonimizzati, sono stati pubblicati e resi disponibili per la consultazione attraverso l'interfaccia *NoSketch Engine*, che permette di usare i metadati come filtri di ricerca e come informazioni associate alle singole occorrenze. Ogni occorrenza è poi collegata al segmento di file sonoro che la contiene.

[8] Una caratteristica che rende il *KIParla* innovativo è la sua struttura modulare e incrementale, ovvero la divisione interna in moduli indipendenti e la possibilità di poter aggiungere nuovi componenti nel corso del tempo. Tale struttura, oltre a rendere possibile un ampliamento costante del corpus, favorisce il confronto dei dati contenuti in moduli diversi e la consultazione tra le diverse sezioni in modo separato o congiunto, a seconda dei fini specifici di ricerca, attraverso la piattaforma *NoSketch Engine*:

[1]a filosofia che sta alla base della modularità è quindi mettere insieme corpora di piccole/medie dimensioni, costruiti per scopi diversi e spesso complementari, raccogliendo dati in aree geografiche diverse o presso tipi di comunità diverse, mantenendo però una sostanziale comparabilità di fondo nella struttura e nell'accessibilità. (p. 38)

Ai due moduli attualmente disponibili se ne affiancheranno a breve altri due: il *KIPPasti* e il *ParlaBO*. Il primo, il *KIPPasti*, contiene registrazioni effettuate durante i pasti tra persone legate da un rapporto confidenziale. Considerato che questi dati provengono da diverse città e i protagonisti delle interazioni sono eterogenei per quanto riguarda l'età, il titolo di studio e l'occupazione lavorativa, questo modulo potrà essere impiegato per lo studio della variabilità diatopica e diastratica. Il secondo modulo, il *ParlaBO*, mostra una struttura e una modalità di discussione dei dati simili a quelle del *ParlaTO*; in questo caso però l'unico punto di inchiesta è la città metropolitana di Bologna. Un impianto modulare di questo tipo potrebbe consentire nel medio termine di catturare le principali dimensioni di variazione sociolinguistica dell'italiano contemporaneo.

[9] Il quarto capitolo (pp. 77-114) illustra alcuni studi condotti sui dati del *KIParla*. Questa rassegna, che evidenzia le potenzialità di ricerca del corpus, si snoda intorno a quattro linee direttrici per le quali l'analisi del parlato assume un

ruolo determinante: l'individuazione di percorsi diacronici emergenti, l'identificazione di costruzioni linguistiche poco descritte, la descrizione di varietà di lingua o di fenomeni di contatto linguistico, e l'analisi dei meccanismi soggiacenti allo sviluppo temporale della conversazione. Riportiamo dunque cursoriamente gli studi presentati in quest'ultima sezione, raggruppandoli sulla base del loro posizionamento rispetto alle prospettive accennate sopra.

[10] La centralità del parlato nel processo di individuazione di percorsi diacronici emergenti è dimostrata in questo capitolo da quattro diverse ricerche: a) l'analisi di costruzioni emergenti nelle quali *come* mostra funzioni legate alla gestione dell'informazione nell'interazione, e che quindi si discostano da quelle individuate in letteratura; b) lo studio di caso che prende in esame le funzioni di *dove*, sulla base del diverso grado di formalità delle interazioni analizzate; c) la ricerca sulla particella modale *solo* e sulle nuove funzioni soggettive e intersoggettive che emergono a partire da espressioni dotate inizialmente di un significato oggettivo; d) l'analisi delle costruzioni condizionali controfattuali introdotte da *se*, in relazione alle dimensioni di variazione diastratica e diafasica. Che il parlato rivesta un ruolo fondamentale nell'osservazione di costruzioni linguistiche poco descritte e nella descrizione delle loro proprietà è esemplificato da due studi: a) l'analisi del connettivo *senza* con funzione anticircostanziale; b) lo studio sulle strategie di negazione non canonica nelle produzioni di parlanti colti (*non è che e mica*). L'opportunità offerta dai dati di parlato di descrivere fenomeni di contatto linguistico è portata all'attenzione del lettore grazie alla discussione di esempi di *code-switching* (italiano e dialetto, non solo piemontese) tratti dal *ParlaTO*, mentre la possibilità di descrivere varietà di lingua emerge concretamente nella sezione dedicata all'italiano popolare. Infine, l'ultimo paragrafo presenta un approfondimento su alcuni fenomeni connessi alla sintassi del parlato. Una proprietà, ripresa più volte dagli autori nel corso del volume, è l'incrementalità, ovvero la modalità secondo la quale il messaggio linguistico viene prodotto e processato in modo additivo dai partecipanti all'interazione. Le co-costruzioni rappresentano emblematicamente questo aspetto della sintassi del parlato.

[11] Come sarà emerso dal quadro sin qui delineato, il volume si caratterizza sia come il prodotto di un'esperienza pluriennale di ricerca nell'ambito degli studi sui corpora di parlato, sia come un testo che ben si presta a essere usato con finalità didattiche. A una parte teorica, incentrata sul ruolo del parlato nelle scienze linguistiche, si affianca un'anima più pratica, che guida il lettore all'uso del corpus e all'acquisizione dei metodi da seguire nel processo di costruzione di una risorsa di questo tipo.

[12] Data (anche) la scarsità di basi di dati di parlato disponibili per l'italiano, il *KIParla* rappresenta senza dubbio una risorsa preziosa per tutti i linguisti (non solo sociolinguisti) che siano interessati a lavorare con dati di parlato italiano. Se è vero infatti che i dati rappresentano la base sostanziale della ricerca sociolinguistica, occorre assodare però che, ripercorrendo la storia di questa discipli-

na, l'interesse dei sociolinguisti nel corso dei decenni si è indirizzato maggiormente verso gli aspetti metodologici più spiccatamente legati alle strategie di escussione dei dati messe in atto dal ricercatore (p. es. Milroy & Gordon 2003; Schilling-Estes 2013) e alle tecniche di analisi quantitativa di singole variabili sociolinguistiche (p. es. Tagliamonte 2006), mentre poca attenzione hanno ricevuto le procedure di trasformazione delle registrazioni sonore in dati consultabili, accessibili e interrogabili (Kendall 2008). Per l'affermazione e il consolidamento di queste pratiche, sono stati determinanti i contributi offerti alla sociolinguistica da parte di altre discipline, quali la *language documentation* e la *corpus linguistics*¹.

[13] Come sottolinea Kendall (2011), la maggior parte dei dati registrati per ricerche sociolinguistiche non viene trascritta, né resa consultabile e, di conseguenza, neanche analizzata. Nella conduzione di un'indagine sociolinguistica, visto che l'interesse primario del singolo ricercatore è rivolto verso singole variabili piuttosto che al complesso dei dati inteso come corpus, la pratica più frequente consiste nell'impiegare i dati raccolti solo per analizzare una certa quantità di tokens, quelli riferibili alla variabile oggetto di indagine. Naturale conseguenza di questa prassi è che dati che potrebbero essere usati per indagare più variabili oppure fenomeni riferibili a diversi livelli linguistici, da parte di diversi gruppi di ricerca e con approcci teorici differenti, in realtà, vengono sprecati. D'altra parte, lo sviluppo di corpora di parlato rappresenta un'operazione altamente dispendiosa, in termini di tempo e di risorse umane (ed economiche) e, per questo motivo, non sempre essa risulta praticabile. A maggior ragione, perciò, la costruzione di una base di dati di parlato messa a disposizione della comunità scientifica rappresenta un'iniziativa rilevante, perché consente, a chiunque sia interessato, di poter usufruire della risorsa per condurre ricerche sull'italiano parlato da prospettive diverse.

Abbreviazioni e riferimenti bibliografici

ELAN = Max Planck Institute for Psycholinguistics 2023. *ELAN. Versione 6.5*. <https://archive.mpi.nl/tla/elan>.

Jefferson 2004 = Gail Jefferson 2004. Glossary of transcript symbols with an introduction. Gene H. Lerner (a cura di). *Conversation analysis. Studies from the first generation*. Benjamins, 13-31.

Kendall 2008 = Tyler Kendall 2008. On the history and future of sociolinguistic data. *Language and Linguistics Compass* 2.2, 332-351.

Kendall 2011 = Tyler Kendall 2011. Corpora from a sociolinguistic perspective. *Revista brasileira de linguística aplicada* 11.2, 361-389. <https://www.scielo.br/j/rbla/i/2011.v11n2/>.

KIParla = Caterina Mauri et al. (a cura di) 2023. *Corpus KIParla. L'italiano parlato e chi parla italiano*. <http://kiparla.it>.

Mereu & Vietti 2021 = Daniela Mereu, Alessandro Vietti 2021. Dialogic ItAlian: the creation of a corpus of Italian spontaneous speech. *Speech Communication* 130, 1-14.

¹ Per un approfondimento sul tema dei corpora di parlato per la ricerca sociolinguistica e sul contributo specifico delle altre discipline in questa direzione, si rinvia alla discussione presente in Mereu & Vietti (2021).

- Milroy & Gordon 2003 = Lesley Milroy, Matthew Gordon 2003. *Sociolinguistics. Method and interpretation*. Blackwell.
- NoSketch Engine* = Natural language processing centre, Masarykova univerzita 2023. *NoSketch Engine*. <https://nlp.fi.muni.cz/trac/noske/wiki>.
- Rychlý 2007 = Pavel Rychlý 2007. Manatee/Bonito – A modular corpus manager. Petr Sojka, Aleš Horák (a cura di). *First workshop on recent advances in Slavonic natural language processing, RASLAN 2007*. Masaryk University, 65-70. <https://nlp.fi.muni.cz/raslan/2007/papers/12.pdf>.
- Schilling-Estes 2013 = Natalie Schilling-Estes 2013. *Sociolinguistic fieldwork*. Cambridge University Press.
- Sloetjes & Wittenburg 2008 = Han Sloetjes, Peter Wittenburg 2008. Annotation by category: ELAN and ISO DCR. *Proceedings of the 6th International conference on language resources and evaluation (LREC 2008)*. Marrakech, Morocco, 28-30 May 2008. http://www.lrec-conf.org/proceedings/lrec2008/pdf/208_paper.pdf.
- Tagliamonte 2006 = Sali A. Tagliamonte 2006. *Analysing sociolinguistic variation*. Cambridge University Press.