Il VoLIP una risorsa per lo studio della variazione nel parlato della lingua italiana

sezione 16
Miriam Voghera, Francesco Cutugno, Claudio Iacobini, Renata Savy

Il VoLIP (Voce del LIP) è una risorsa linguistica, liberamente disponibile attraverso il portale *Parlare italiano* (www.parlaritaliano.it), realizzata da Miriam Voghera, Francesco Cutugno, Claudio Iacobini, Renata Savy presso l'università di Salerno all'interno di un progetto finanziato con fondi del Ministero dell'Istruzione, dell'Università e della Ricerca scientifica e tecnologica.

Il VoLIP permette di ascoltare e scaricare i file dei segnali audio da cui sono tratte le trascrizioni ortografiche del Corpus LIP, e permette la ricerca nel corpus sia secondo criteri sociolinguistici sia secondo criteri lessicali e morfo-sintattici.

Il Corpus LIP (composto da 500.000 occorrenze di parole ricavate da 60 ore di registrazione) è stato raccolto nei primi anni '90 del ventesimo secolo (T. De Mauro, F. Mancini, M., Vedovelli, M. Voghera, *Lessico di frequenza dell'italiano parlato*, Milano, Etaslibri, 1993) al fine di produrre un lessico di frequenza dell'italiano parlato che permette di individuare i primi 3000 lemmi in ordine d'uso. Il corpus è stato raccolto tenendo presente criteri di variazione diafasici, diatopici e diamesici.

Per quanto riguarda la variazione diafasica e diamesica, i testi del LIP sono suddivisi in 5 gruppi: a) conversazioni faccia a faccia; b) conversazioni telefoniche; c) scambi comunicativi bidirezionali con alternanza di turno predefinita, come interviste, dibattiti, interazioni in aule scolastiche, esami orali, ecc.; d) monologhi, come letture, sermoni, discorsi, ecc.; e) programmi radiofonici e televisivi. I testi contenuti nei gruppi a) e b) appartengono a registri sia formali sia informali, mentre i testi dei gruppi c), d) ed e) sono registrati prevalentemente in contesti pubblici in cui si adottano registri formali. Per quanto riguarda la variazione diatopica, i testi sono stati raccolti a Milano, Roma, Napoli e Firenze. Le prime tre città sono state scelte per la loro posizione geografica e per il numero di abitanti, perché Milano, Roma e Napoli sono le città più popolate d'Italia. Firenze è stata scelta per la sua grande importanza nella storia della lingua italiana.

Il LIP è a tutt'oggi il corpus di parlato spontaneo più ampio e con la maggiore variabilità diatopica dell'italiano. Poiché è stato originariamente raccolto per ricerche in ambito lessicale, le condizioni di registrazione e la qualità acustica delle sessioni sono disomogenee. La scala di qualità oscilla fra livelli alti e livelli bassi di chiarezza del segnale.

Il VoLIP rende disponibile l'audio del corpus LIP in file wav (Windows PCM, 22050Hz. 16 bit) correlati con: a) metadati testuali e sociolinguistici in formato IMDI; b) la trascrizione ortografica (già pubblicata in De Mauro et al. 1993) in file TXT integrata con la lemmatizzazione.

Il VoLIP permette due strategie di interrogazione: i) per variabili di testo e di registro, in base ad una annotazione in metadati più articolata rispetto al LIP; ii) per criteri lessicali e morfosintattici, in base al lessico (distinto in tipi e forme) e all'analisi in parti del discorso. Le due strategie di interrogazione possono essere incrociate. Le interrogazioni hanno come risultato la trascrizione ortografica associata ai file audio.

La metadatazione VoLIP, basata su IMDI, arricchisce la classificazione adottata nel LIP integrandola con informazioni riguardanti: sesso dei locutori, genere testuale (es. scambi comunicativi in presenza, parlato trasmesso, ecc.), sottogenere (es. conversazione, descrizione di un evento, narrazione, ecc.), grado di interazione, livello di pianificazione, contesto comunicativo, struttura dell'evento comunicativo, canale di trasmissione.

La ricerca lessicale e morfosintattica dà come risultato tutti i testi che presentano la voce (forma della parola o lemma) o la parte del discorso ricercata. I risultati della ricerca sono evidenziati all'interno del loro contesto di occorrenza. Per ogni lessema, forma o parte del discorso cercata è fornita la frequenza di occorrenza distinta per città e per registro. Il sistema di ricerca dei lemmi è basato sulla classificazione in parti del discorso utilizzata per la lemmatizzazione del LIP (aggettivo; articolo; avverbio; cognome; congiunzione; interiezione; nome; nome commerciale;

nome geografico; nome proprio; onomatopea; preposizione; pronome; verbo). E' possibile ottenere sia il numero totale di occorrenze del lemma sia le forme del lemma (eventualmente ripartite nelle cinque sezioni del LIP). E' inoltre possibile la ricerca di una sequenza di parole (forme di parola), ad es. "ho dormito"; "i cani", "è una città che". La ricerca di una sequenza permette di accedere al testo e all'audio contenente la sequenza, ma non permette di visualizzare l'indicazione del numero di occorrenze.

Esempi di interrogazione. Cercando il lemma *correre*, è possibile ottenere anche il numero di occorrenze delle sue diverse forme, e da queste accedere all'elenco dei file (audio e trascrizioni) in cui l'elemento cercato ricorre.

Pos	Lemma	Form	A	B	C	D	E	Total
Verb	CORRERE	corre	2	0	1	9	1	13
Verb	CORRERE	corri	1	2	1	0	0	4
Verb	CORRERE	correre	0	0	1	1	1	3
Verb	CORRERE	correndo	0	1	0	1	0	2
Verb	CORRERE	corrono	1	0	0	1	0	2
Verb	CORRERE	correra'	0	1	0	1	0	2
Verb	CORRERE	@core@	0	1	0	0	0	1
Verb	CORRERE	@curruto@	0	0	1	0	0	1
Verb	CORRERE	correrebbero	0	0	0	1	0	1

Selezionando uno dei file individuati, sarà possibile rintracciare l'occorrenza o le occorrenze della forma nel file trascritto, visualizzare la forma, sottolineata ed evidenziata in giallo, nel contesto di occorrenza ed ascoltarla in maniera mirata, grazie all'allineamento audio (es. *inizia a sparare mentre corre anch'egli verso la salumeria e indirizzando i colpi*).

E' possibile anche incrociare criteri di ricerca, in modo da ottenere informazioni su una data forma, distinguendo le occorrenze in parti del discorso. Si possono così distinguere forme omografe, quali la prima singolare del presente indicativo del verbo "amare" ed il singolare del nome "amo", e poi passare alla visualizzazione e all'ascolto della forma desiderata

Pos	Lemma	Form	A	B	C	D	E	Total
Verb	AMARE	amo	0	3	0	0	0	3
Noun	AMO	amo	0	0	0	0	1	1

.

La ricerca di sequenze permette l'individuazione di contesti utili per analisi fonetiche e fonologiche di tipo sia segmentale sia soprasegmentale, ma soprattutto permette di studiare i fenomeni di interrelazione tra realizzazione fonica e costruzioni semantico-sintattiche: punto cruciale per qualsiasi analisi dei testi parlati.