Trotter, David/Bozzi, Andrea/Fairon, Cédrick (éd.) (2016). *Actes du XXVII^e Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 16 : Projets en cours ; ressources et outils nouveaux*. Nancy, ATILF :
http://www.atilf.fr/cilpr2013/actes/section-16.html.

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

1. Introduzione

I *corpora* sono una grande risorsa per le indagini linguistiche poiché permettono di costruire e/o validare le proprie spiegazioni su materiale autentico, strutturato, pubblicamente fruibile. Un *corpus* non è infatti una mera collezione di dati, ma un insieme strutturato sulla base di criteri espliciti e controllabili. I lavori basati su *corpora* consentono quindi di condividere il processo di costruzione delle ipotesi teoriche formulate su dati condivisi e non creati *ad hoc*. È questa una condizione necessaria per la realizzazione di protocolli di ricerca comuni, riproducibili e riutilizzabili da più ricercatori.

I *corpora* sono, d'altro canto, essi stessi il frutto di ipotesi linguistiche poiché sono disegnati sulla base di criteri di rappresentazione in scala di una lingua o di una porzione di essa: ogni *corpus* sottintende un modello di lingua e di aggregazione dei dati. Ogni *corpus* implica, infatti, la scelta delle variabili che si ritengono più significative nel condizionare gli usi linguistici e il peso che si assegna a ciascuna di esse; queste scelte producono *corpora* non solo di dimensioni diverse, ma di contenuti molto diversi. Questa scelta può variare in base alle convinzioni teoriche dei compilatori del *corpus* e/o in base ai livelli linguistici che si vogliono indagare o, ancora, in base agli scopi per i quali il *corpus* è concepito. La grande espansione della linguistica dei *corpora* negli ultimi vent'anni, dovuta soprattutto allo sviluppo straordinario degli strumenti elettronici di raccolta, annotazione e elaborazione del materiale linguistico, ha permesso non solo la crescita delle dimensioni dei *corpora*, ma la moltiplicazione dei tipi di *corpora*. Accanto ai *corpora* che potremmo chiamare di riferimento o generali, si possono consultare *corpora* concepiti per indagare specifici ambiti o per usi determinati, quali per esempio quelli didattici (O'Keeffe, McCarthy 2010).

Oltre alle scelte relative alle dimensioni di variazione e al peso ad esse assegnato e quindi ai tipi di testi da inserire, i *corpora* differiscono molto per ciò che riguarda le scelte di trattamento e annotazione, il tipo e il grado di fruizione pubblica del materiale (Baroni 2010). E' opportuno operare una distinzione tra *corpora* di testi scritti e di testi parlati. Benché anche i testi scritti siano sottoposti a trattamenti di varia natura per renderli elaborabili elettronicamente e poi fruibili in modo uniforme dagli utenti, i testi parlati richiedono un passaggio ulteriore: la trascrizione ortografica più

o meno larga. Si tratta di un passaggio non necessario né concettualmente né teoricamente, ma che metodologicamente aumenta la facilità di elaborazione del parlato, rendendolo permanente e non volatile. E' questo un lavoro molto oneroso in termini di tempo e tutt'altro che neutro, che ancora una volta implica scelte teoriche e metodologiche non banali. E' utile ribadire che la trascrizione di un testo parlato deve trasferire attraverso il canale grafico-visivo, normalmente utilizzato per comunicazioni unidirezionali non sincrone, testi perlopiù dialogici sincroni e quindi trovare soluzioni univoche per registrare fenomeni di dinamica comunicativa (presa dei turni, sovrapposizioni ecc.), non segmentali (pause, velocità di eloquio, andamento prosodico, variazioni vocali significanti, come il *self-talk*), disfluenze (autocorrezioni, cambiamenti di progetto ecc.), elementi verbali non lessicali (segnalazioni di assenso, interiezioni varie, forme contratte ecc.), elementi non verbali (elementi vocali, per esempio risate, colpi di tosse, inspirazioni e simili, ma anche espressioni facciali, movimenti del corpo ecc.) e contestuali ritenuti rilevanti (Edwards 1993, Leech *et al*. 1995). I sistemi di trascrizione oggi a disposizione sono molteplici e in alcuni casi molto ricchi: possono includere per esempio molti livelli di codifica, fino a comprendere anche quello cinesico. Purtroppo più una trascrizione è dettagliata e 'completa' meno è leggibile: nuovamente si è di fronte ad una scelta che dipende dal tipo di analisi a cui si è interessati e per la quale il *corpus* è stato concepito.

I *corpora* di scritto e di parlato presentano di fatto delle differenze anche per ciò che riguarda l'accesso al materiale. Mentre i primi di solito rendono disponibile il materiale raccolto e, se interrogabili, si può di norma risalire dall'interrogazione al cotesto e quindi all'intero testo originario, molto raramente questo è possibile con i *corpora* di parlato. Benché esistano *corpora* etichettati a vari livelli, da quello fonetico a quello pragmatico (Lüdeling, Kytö 2008), non tutti sono interrogabili e sono pochi i *corpora* di parlato che danno accesso al materiale audio e ancora meno quelli che permettono qualche forma di interrogazione diretta o indiretta, attraverso la trascrizione ortografica, del materiale audio.

Un *corpus* è quindi prima di tutto il risultato di un insieme di scelte complesse e integrate, fortemente condizionate anche da fattori metodologici e di organizzazione della ricerca. Per questo motivo, così come normalmente abbiamo bisogno di diverse condizioni sperimentali sui cui testare le nostre ipotesi, è utile poter disporre di più *corpora*. Ciò è tanto più vero quando si tratta di parlato, almeno per due ragioni. Innanzi tutto i *corpora* di parlato sono ancora molto pochi se confrontati con quelli di testi scritti. Ciò vale per tutte le lingue, ma ancor di più per le lingue romanze. Anche nella linguistica del *corpus*, la riflessione teorica sulla specificità del parlato è relativamente recente e ancora più recente è il riconoscimento di una rilevanza della variabile diamesica in sé (Baker 2010). Ciò rende necessario uno studio della comunicazione parlata in diverse condizioni enunciative e, probabilmente, anche in diversi contesti settoriali per aver ben chiara l'integrazione tra tutte le componenti del processo di significazione.

In secondo luogo, più *corpora* permettono di prendere in considerazione un maggior numero di contesti diafasici e diatopici. La questione è di grande rilevanza soprattutto per aree linguistiche come quella italiana in cui il parlato non solo è associato al registro informale e colloquiale, ma anche a varietà geografiche locali. È ben nota la storia linguistica italiana caratterizzata fino al secondo dopoguerra da una forte diglossia, con uso del dialetto nelle situazioni informali parlate e uso dell'italiano nello scritto formale. Sebbene dagli inizi degli anni Ottanta ad oggi sia cresciuta la consapevolezza che il canale ha un ruolo suo proprio nel determinare l'uso delle lingue storico-naturali e quindi le caratteristiche dei testi, l'associazione prevalente/preponderante di un canale con certe situazioni e/o varietà diatopiche rende difficile individuare le proprietà attribuibili alle diverse dimensioni di variazione. Ciò ingenera un'indebita equivalenza tra proprietà del canale e registri e varietà regionali o dialetti, tra variazione diamesica, diafasica, diatopica (Voghera 2010a). I tre tipi di variabili vanno evidentemente tenuti separati e un *corpus* di parlato deve prevedere una variazione lungo l'asse sia diafasico sia diatopico.

Partendo da queste considerazioni, da quasi un decennio abbiamo iniziato un percorso di costruzione di una base di dati varia e condivisa del parlato italiano, creando nuove risorse e valorizzando e sviluppando quelle già esistenti, grazie alla creazione di un osservatorio permanente del parlato italiano, il portale Parlaritaliano (‹www.parlaritaliano.it›, Voghera 2010b). Si tratta di un osservatorio permanente del parlato che ha due obiettivi fortemente interrelati. Il primo consiste nell'allargare la base conoscitiva dei principali meccanismi enunciativi e grammaticali della comunicazione parlata attraverso studi basati su *corpora*. Il secondo obiettivo consiste nel valutare come e quanto l'allargamento e l'approfondimento della base documentaria possano contribuire a una migliore comprensione del sistema linguistico nel suo complesso. È evidente infatti che il parlato, ad un primo livello, si caratterizza come un sottoinsieme di strutture linguistiche tipiche non, o solo parzialmente, osservabili in altri contesti; ad un secondo livello, il parlato ci permette di scoprire relazioni tra porzioni diverse della grammatica, altrimenti nascoste, ma comunque centrali nell'architettura generale del sistema.

Agli obiettivi più strettamente linguistici si aggiungono obiettivi informatici rivolti allo studio della struttura dei metadati e dei *database* linguistici, e all'analisi e all'implementazione, anche in ambienti di calcolo ad alte prestazioni, di tecniche di segmentazione automatica per segnali audio-video, con specifiche finalità di supporto alla gestione di *corpora* linguistici multimediali e multimodali.

In questa cornice è nato il VoLIP (Voce del LIP), finanziato dal Ministero italiano dell'Istruzione, dell'Università e della Ricerca, una nuova risorsa linguistica che permette l'ascolto e l'interrogazione dei file audio del *corpus* LIP (De Mauro *et al.* 1993), secondo criteri sociolinguistici, lessicali e morfosintattici. Il VoLIP è stato creato in una prospettiva ecologica delle risorse linguistiche che mira al pieno utilizzo e potenziamento, laddove è possibile, di risorse già disponibili, ma sottoutilizzate per vari motivi. Il VoLIP, mettendo a disposizione i file audio del LIP e rendendoli

interrogabili, ha potenziato l'utilizzazione della risorsa offrendo alla comunità scientifica nuove funzionalità di fruizione non esistenti precedentemente e la possibilità di accesso libero e diretto all'intero *corpus* sul portale Parlaritaliano.it.

2. Il passaggio dal LIP al VOLIP

Come abbiamo già detto, la costruzione del VoLIP ha comportato un processo di valorizzazione di una risorsa già esistente, ma solo parzialmente utilizzata. Ciò ha implicato varie fasi nel trattamento del materiale originario.

2.1. Trattamento dei testi

Il *corpus* LIP è costituito da circa 60 ore di registrazione per un totale di circa 500.000 occorrenze, ricavate da testi raccolti in diverse città italiane e disposti su una scala di progressiva dialogicità e spontaneità. Nella Tabella 1 si dà la distribuzione delle occorrenze in rapporto alle variabili diatopica e diafasica.

Città	Conversazioni faccia a faccia	Conversazioni telefoniche	Interviste Dibattiti	Monologhi	Radio/ TV	Totale
Milano	~25.000	~25.000	~25.000	~25.000	~25.000	~125.000
Firenze	~25.000	~25.000	~25.000	~25.000	~25.000	~125.000
Roma	~25.000	~25.000	~25.000	~25.000	~25.000	~125.000
Napoli	~25.000	~25.000	~25.000	~25.000	~25.000	~125.000
Totale	~100.000	~100.000	~100.000	~100.000	~100.000	~500.000

Tabella 1: composizione del *corpus* LIP

Le trascrizioni ortografiche originarie evitavano qualsiasi intervento di normalizzazione. Nel passaggio dal LIP al VoLIP i materiali audio sono stati digitalizzati in file wav (Windows PCM, 22050Hz, 16 bit) e il riascolto ha inevitabilmente portato a revisioni nelle trascrizioni. Si è quindi deciso di rendere disponibili sia la vecchia sia la nuova trascrizione: ogni porzione revisionata richiama una finestra di *pop-up* con la trascrizione emendata.

2.2. Metadatazione

La scelta di procedure standardizzate, adeguate e sufficientemente accurate di metadatazione, incide notevolmente sulla gestione, fruibilità e diffusione dei *corpora*.

Al fine di arricchire la catalogazione già prevista nel *corpus* LIP, per generi discorsivi e provenienza geografica dei parlanti, ma soprattutto con lo scopo di adottare un formato standardizzato, nel passaggio al VoLIP si è deciso di adottare la metadatazione in formato IMDI (*ISLE Metadata Initiative*[1]).

Tale formato, proposto per descrivere risorse linguistiche multimediali e multimodali, è stato scelto anzitutto per favorire il confronto con altri *corpora* (data la sua diffusione in vari progetti di ricerca su *corpora* di diverse lingue), ma anche per la garanzia della correttezza formale dei metadati grazie ai *tool* di supporto (Broeder *et al*. 2001).

Attualmente è possibile consultare ed interrogare in rete il VoLIP sia in base alla catalogazione adottata nel *corpus* LIP, sia in base ai metadati IMDI. La figura 1 mostra la maschera di interrogazione dei metadati in VoLIP.

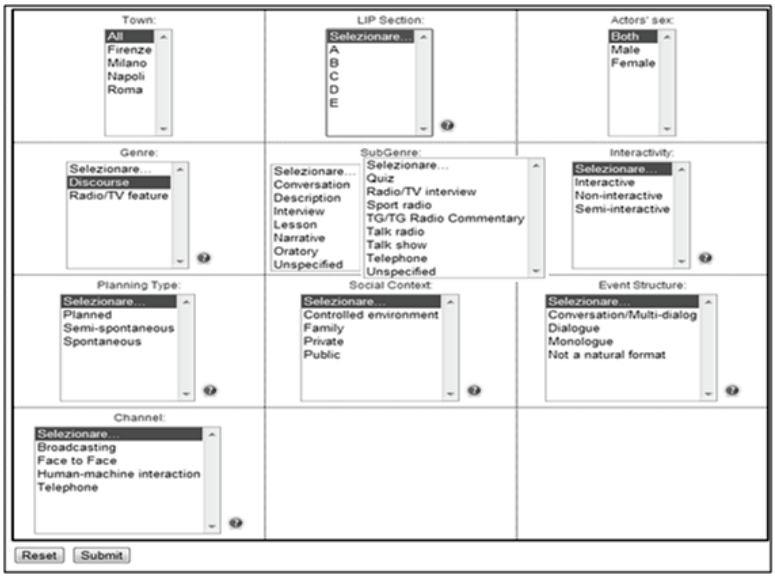

Figura 1: maschera di interrogazione per metadati.

Ciascun menù a tendina corrisponde a un campo interrogabile che consente di filtrare e selezionare tutti i testi che presentano le caratteristiche richieste. È quindi possibile effettuare diversi tipi di interrogazioni, il cui risultato consiste nelle trascrizioni ortografiche associate ai file audio.

La casella *Town* permette di selezionare la provenienza geografica dei parlanti (delle quattro città Firenze, Milano, Napoli, Roma), mentre il campo *Lip section* fa

[1] Per le specifiche adottate, rimandiamo alla documentazione disponibile all'indirizzo: ‹www.mpi.nl/imdi/›.

riferimento ai generi discorsivi della catalogazione adottata nel LIP, riportati nella tabella 1.

Le altre voci in figura corrispondono, invece, a campi dei metadati IMDI. La casella *Actors'sex* consente di filtrare voci maschili o femminili. A *Genre* corrisponde la prima macrocatalogazione tra parlato radiotelevisivo (*Radio/TV feature*) e tutti gli altri tipi (*Discourse*). Per ciascuno dei due, è poi possibile mediante il campo *SubGenre* selezionare un sottotipo specifico: in *Radio/TV feature*, si possono richiedere, ad esempio, interviste radiofoniche o notiziari televisivi; in *Discourse*, è possibile effettuare una ricerca che individui, ad esempio, unicamente lezioni scolastiche o universitarie. Gli altri campi sono relativi al grado di interattività (*Interactivity*) e di pianificazione dei testi (*Planning type*), al contesto in cui avviene lo scambio comunicativo (*Social Context*), alla struttura dell'evento rispetto al numero di partecipanti (*Event Structure*) e al tipo di canale di trasmissione (*Channel*).

La possibilità di incrociare diverse selezioni consente di effettuare ricerche particolareggiate relative a più aspetti contemporaneamente.

3. L'allineamento della voce con le trascrizioni

L'obiettivo principale del VoLIP è quello di dare voce al LIP, cioè rendere disponibile e interrogabile il *corpus* di parlato nella sua interezza[2]. Lo scopo del VoLIP è quindi non solo quello di fornire la versione audio dei testi del LIP, ma di poter fare ricerche che permettano di trovare e ascoltare singole parole in contesto. Abbiamo quindi creato un sistema *online* che permette di ascoltare una data forma all'interno del suo contesto e contemporaneamente leggere la relativa trascrizione ortografica senza che sia necessario scaricarla sul proprio computer.

Ciò si è ottenuto con l'uso di uno strumento di allineamento forzato che ha elaborato le trascrizioni ortografiche e i file audio del *corpus*. La procedura di *Forced Alignment* (FA) fra testo e audio si basa su un approccio semplificato di quanto accade normalmente nel caso del processo di *Automatic Speech Recognition* (ASR). Il caso FA è semplificato dal fatto che il testo è noto, dunque la procedura deve solo ottimizzare la suddivisione del segnale in piccole parti e la loro assegnazione ai singoli eventi fonici che il sistema è in grado di riconoscere contemporaneamente nella loro forma simbolica (testo) e nella loro forma acustica (audio). Il testo da allineare viene quindi suddiviso (tokenizzato) in parole, e per ogni parola un trascrittore automatico grafema > fonema descrive la composizione in eventi simbolici unitari. Il sistema di ASR analizza il segnale audio ogni 20 millisecondi e determina quante porzioni consecutive di segnale possono essere associate ad ogni evento simbolico nella catena. Successive fasi di ottimizzazione completano il processo.

---

[2] Ricordiamo che le trascrizioni originarie dei testi erano state già pubblicate insieme al lessico di frequenza (De Mauro *et al.* 1993) e che sono interrogabili sul sito dell'Università di Graz ‹badip.uni-graz.at/›.

Poiché il materiale originale era di qualità molto varia e raccolto in condizioni audio non ottimali, spesso molto disturbate (mercati, assemblee, classi scolastiche, ecc.), è stato necessario integrare il lavoro svolto tramite il programma di allineamento con un lavoro di divisione manuale a partire dalla trascrizione ortografica. Tale soluzione è consistita nell'individuare porzioni della durata media di 30 secondi. Il lavoro si è svolto in due fasi: nella prima, le trascrizione ortografiche sono state suddivise in porzioni di senso compiuto cercando di rispettare, per quanto possibile, l'intervallo temporale fissato; nella seconda fase si sono ascoltati i file audio e tramite un software appositamente creato, basato sulla libreria SoundManager 2 (‹www.schillmania.com/projects/soundmanager2/›), si sono salvati il tempo di inizio e fine di ogni intervallo precedentemente segnalato sulla trascrizione. Il risultato di queste due fasi è un file di tempi ed un file contenente la trascrizione ortografica divisa in parti. A questo punto, per verificare la correttezza del lavoro manuale e allineare ogni coppia porzione-di-trascrizione/tempi in un unico file che fosse nello stesso formato finale dei file ottenuti in precedenza grazie al software di FA, è stato sviluppato un altro script che fosse anche in grado di rilevare e segnalare differenze tra le porzioni della trascrizione e il numero di intervalli salvati.

Concluse le fasi appena descritte, si è giunti alla fase finale del lavoro. In particolare, è stato sviluppato un innovativo software che guida l'utente nella scelta dei testi. Una volta cercata una forma all'interno del *corpus*, il software permette di visualizzare tutti i testi in cui occorre, di evidenziarne le occorrenze nella trascrizione dei singoli file e di ascoltarle in contesto online. Questa procedura rende molto veloce l'accesso al *corpus* perché non è necessario scaricare alcuno strumento o file. Il programma consente anche di scaricare i contesti audio oppure l'intero file audio corredato di trascrizione per una futura elaborazione[3].

4. Le interrogazioni nel VoLIP

Il VoLIP prevede diversi tipi di interrogazione a partire dalla ricerca per lemmi o per forme, permettendo di ottenere sia le liste di frequenza delle occorrenze sia il contesto dell'elemento lessicale ricercato in trascrizione ortografica e sonora.

4.1. Frequenza di lemmi e forme

Nella ricerca per lemmi si può selezionare una specifica parte del discorso, tra quelle su cui è basata la lemmatizzazione del LIP (*Adjective, Adverb, Article*, *Company name*, *Conjunction*, *Geographic name*, *Interjection*, *Name*, *Noun*, *Onomato-*

[3] Il menù di selezione che appare su ogni occorrenza individuata e che consente la scelta di scaricare il file o una sua porzione, e di ascolto della sezione audio relativa alla forma cercata è basata su jQuery (‹jquery.com/›), libreria creata da John Resig, ed il plugin Audero Context Menu (‹github.com/AurelioDeRosa/Audero-Context-Menu›) creato da Aurelio De Rosa. Infine, l'estrazione vera e propria di una certa porzione di audio di un file WAV è resa possibile grazie alla libreria Audero Wav Extractor (‹github.com/AurelioDeRosa/Audero-Wav-Extractor›) creata da Aurelio De Rosa.

poeia, *Preposition*, *Pronoun, Surname, Verb*), o selezionarle tutte. Sia l'interrogazione per lemmi sia quella per forme offre la possibilità di raffinare la ricerca selezionando uno dei cinque genere discorsivi distinti nel LIP o il loro insieme (si veda Tabella 1).

La maschera di interrogazione è riprodotta in figura 2.

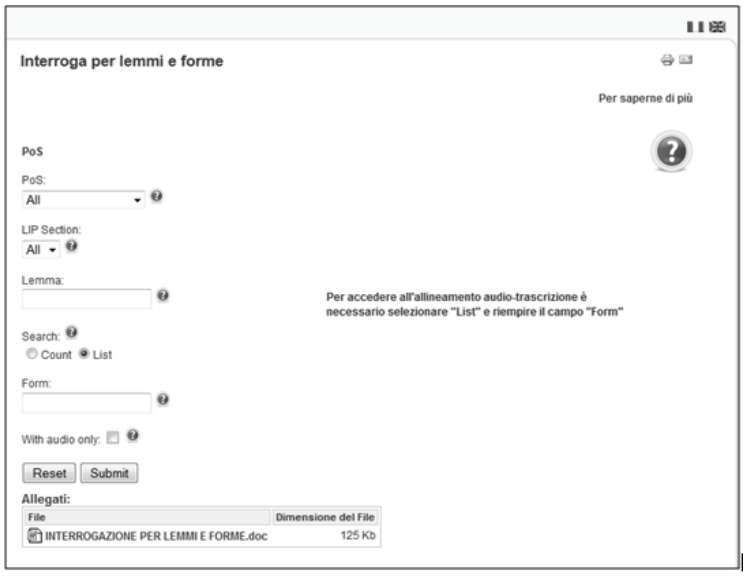

Figura 2: maschera di interrogazione per lemmi e forme.

La casella "PoS" permette di restringere le parti del discorso da ricercare. Una volta digitata nella casella "Lemma" la parola che si intende cercare, la scelta fra le due opzioni "Count" e "List" permette di orientare la ricerca in due direzioni.

L'opzione "Count" consente di ottenere sia il numero totale di occorrenze del lemma sia la loro distribuzione nei cinque generi discorsivi: il conteggio è quello originario del LIP.

Come si può notare dall'esempio (1), se si cerca il lemma correre, viene, viene anche fornita l'indicazione della parte (o delle parti del discorso) a cui può essere ricondotto il lemma cercato.

(1)

Pos	A	B	C	D	E	Total
Verb	45	71	25	22	73	236
Noun	0	0	0	0	2	2

La scelta dell'opzione "List" permette invece di ottenere l'elenco delle forme e la loro distribuzione suddivisa nei cinque generi di discorso (come mostrato nell'esempio numero 2). Sono riportate tra @ le forme locali e/o dialettali.

(2)

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	@curruto@	0	0	1	0	0	1
Verb	CORRERE	correrebbero	0	0	0	1	0	1
Noun	CORRERE	correre	0	0	0	1	0	1
Verb	CORRERE	@core@	0	1	0	0	0	1

4.2. L'ascolto di forme o sequenze

Le interrogazioni possono avere come esito l'ascolto di singole forme o sequenze. L'accesso al contesto dei file audio e di trascrizione avviene attraverso la compilazione del campo "Form", nel quale si può richiedere una delle forme presenti nel *corpus*, cfr. Figura 3

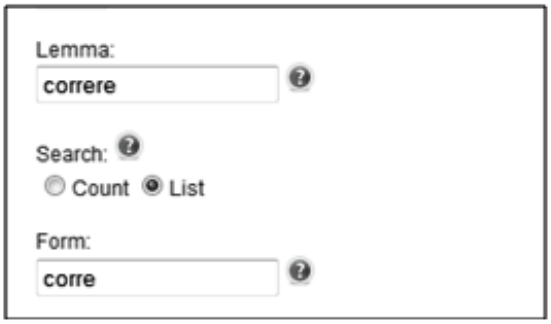

Figura 3: esempio di ricerca di una forma.

Attraverso il campo “Form” si possono ricercare anche sequenze di parole, ad es. “ho dormito”; “i cani”, “e’ una citta’ che”.

La ricerca di una forma permette di visualizzare l’indicazione del numero di occorrenze totali e ripartite nei cinque generi di discorso del LIP, di accedere all’intero testo grafico e audio contenente la forma, e, soprattutto, di accedere direttamente alla porzione di testo e di audio contente la forma ricercata. E’ possibile ricercare anche una sequenza; in questo caso il risultato non fornisce il numero di occorrenze, ma i file audio in cui occorre la sequenza.

Forniamo qui di seguito un esempio di ricerca del segnale audio dei contesti in cui occorre la forme “corre”. La ricerca, illustrata nell’esempio numero (3) dà come risultato il numero delle occorrenze totali della forma e la loro ripartizione nei cinque generi e l’elenco dei file (audio e trascrizioni) in cui l’elemento compare (si veda la figura 4)

(3)

Pos	Lemma	Form	A	B	C	D	E	Total
Verb	Correre	corre	2	0	1	9	1	13

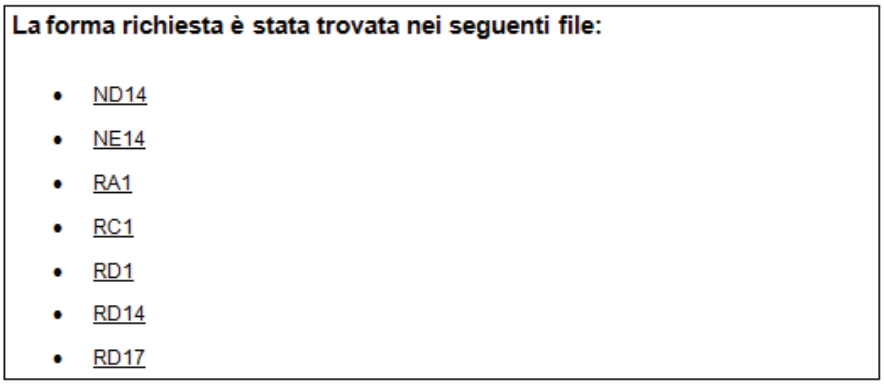

Figura 4: risultato della ricerca della forma “corre” (elenco dei file in cui occorre).

E’ possibile scaricare ciascun file sia in forma di testo che di audio, ma soprattutto individuare velocemente le varie occorrenze della forma (come mostrato in figura 5).

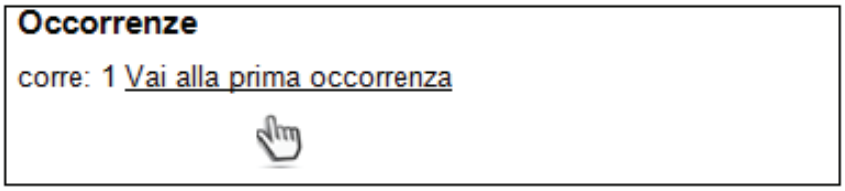

Figura 5: individuazione delle varie occorrenze della forma cercata.

La forma ortografica viene visualizzata in giallo e cliccando sulla forma è possibile ascoltare il contesto in cui appare, grazie all'allineamento tra trascrizione e audio, come mostrato nella figura 6. Oltre ad ascoltare il contesto, è possibile scaricare il frammento di file audio, l'intero file o procedere nella ricerca di altre occorrenze della forma all'interno del file.

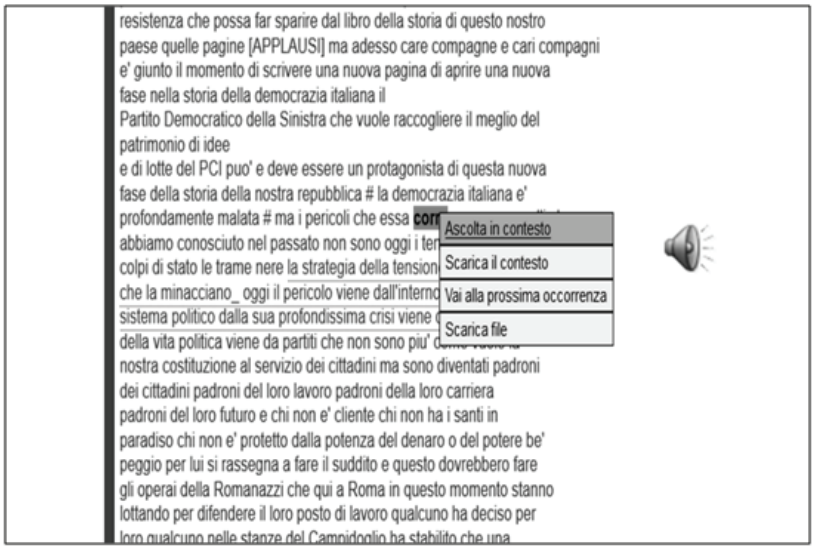

Figura 6: contesto di occorrenza della forma cercata con allineamento tra trascrizione ed audio.

E' ovviamente possibile raffinare la ricerca attraverso l'incrocio dei parametri dei criteri qui brevemente illustrati, in modo, ad esempio, da ottenere informazioni su una data forma, distinguendo le occorrenze in relazione alle parti del discorso a cui possono essere ricondotte. Un tipico esempio è quello della risoluzione dell'omografia tra la prima persona singolare del presente indicativo del verbo "amare" ed il singolare del nome "amo" (si veda l'esempio numero 4).

(4)

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

Una volta individuata la forma appartenente alla parte del discorso voluta, è possibile osservare e ascoltare i contesti di occorrenza (come illustrato negli esempi 5 e 6).

(5) Nome amo

io lo vado a fare cosi' con un amo in maniera tale praticamente da farvi capire che non e' difficile corrente non ce n'e'

(6) Verbo amo

A: ah senti una cosa ma adesso cosa fai fino all'una stai li' a
B: e' quello che non amo affatto 'st' idea di star qua in
cucina da sola o anche in salotto ad aspettare

5. Ipotesi applicative

Nonostante la grande espansione della linguistica del *corpus* negli ultimi decenni, il numero di *corpora* di parlato è ancora molto limitato. Basti pensare che sia nel *British National Corpus* (‹www.natcop.ox.ac.uk›) sia nel *Corpus de Referencia del Español Actual* (‹www.rae.es›), due tra le più importanti imprese per la costruzione di *corpora* nazionali di riferimento, il rapporto tra testi parlati e scritti è di 1: 10. Ciò dà la misura di quanto il parlato sia poco rappresentato anche quando si parla di lingue che hanno una dimensione di diffusione intercontinentale.

La situazione non è diversa, naturalmente, per l'italiano. Mentre negli ultimi anni si sono sviluppate numerose iniziative di raccolta di *corpora* scritti, i *corpora* di parlato disponibili sono ancora molto poco numerosi (Cresti, Moneglia a cura di 2005; Baroni 2010). Ancora di meno sono quelli che consentono il libero accesso ai materiali audio *online* (Savy, Cutugno 2009). Anche per questo motivo il VoLIP rappresenta una risorsa importante che integra varie funzionalità di ricerca sia nei metadati sia all'interno del *corpus*.

La possibilità di incrociare i criteri della metadatazione IMDI con l'originaria suddivisione in tipi di discorso parlato fatta nel LIP consente di selezionare parlati diversi dal punto di vista diafasico e diatopico e oltre ad avere un'ovvia rilevanza per le ricerche di tipo sociolinguistico, può consentire anche un'utilizzazione didattica del VoLIP.

La ricerca sociolinguistica per forme, lemmi e contesti comparabili si arricchisce del contemporaneo accesso al segnale audio: sebbene il VoLIP non sia stato con-

cepito sul piano tecnico-metodologico per ricerche di tipo fonetico, la sua struttura stratificata ne consente l'utilizzo come *corpus* di controllo o di 'verifica' per analisi socio-fonetiche condotte su altri materiali (parlato di laboratorio, *corpora* speciali, parlato elicitato semi-spontaneo o letto) rispetto ai quali il VoLIP presenta l'indubbio vantaggio di un parlato decisamente spontaneo sotto tutti i punti di vista.

L'allineamento di trascrizione ortografica e audio permette per esempio un lavoro di analisi e decostruzione di ciò che rappresenta uno dei punti più significativi e complicati nell'apprendimento di una lingua da parte di non-nativi: la relazione inversa tra complessità dei contenuti e delle costruzioni morfosintattiche e quella della forma fonetica, in conseguenza del grado di pianificazione e di controllo dell'enunciazione. Infatti, i testi usati nei primi livelli di insegnamento, che coincidono di solito con testi poco pianificati e informali, sono in realtà i più complessi sul piano fonetico perché ricchi di riduzioni foniche e fenomeni di coarticolazione. Al contrario, i testi pianificati, tendenzialmente più articolati sintatticamente e nell'estensione lessicale e semantica, tendono ad essere iperarticolati sul piano fonetico e quindi di più facile comprensione perché si avvicinano maggiormente alla forma fonetica prototipica di fonemi e giunture. Ciò, come si può facilmente sperimentare, fa sì che, dal punto di vista della forma fonica, sia più comprensibile una lezione universitaria piuttosto che una conversazione tra amici. Benché la didattica delle lingue, soprattutto la didattica delle lingue seconde, abbia cercato di sviluppare le abilità di ascolto e parlato, questo genere di questioni non è affrontato in modo sistematico, ma ci si affida piuttosto alla naturalità del parlato e dell'ascolto che è, implicitamente, ritenuta la base sufficiente per garantire il raggiungimento degli obiettivi.

Al contrario, l'insegnamento e l'apprendimento si gioverebbero molto di esercizi che uniscano l'ascolto con attività di osservazione e analisi. Il VoLIP poiché consente interrogazioni mirate che isolano singole forme e piccoli contesti offre spunti utili per la creazione di materiale didattico *on* e *off line*, contribuendo alla costruzione di sillabi più strutturati per la didattica dell'ascolto e del parlato.

Università degli Studi di Salerno — Miriam VOGHERA
Università degli Studi di Salerno — Claudio IACOBINI
Università "Federico II" di Napoli — Francesco CUTUGNO
Università degli Studi di Salerno — Renata SAVY
Università degli Studi di Salerno — Iolanda ALFANO
Università degli Studi di Salerno — Aurelio DE ROSA

Riferimenti bibliografici

Baker, Paul, 2010. *Sociolinguistics and Corpus Linguistics*, Edimburgh, Edimburgh University Press.

Baroni, Marco, 2010. *Corpora di italiano, voce della Enciclopedia dell'Italiano diretta da R. Simone*, Istituto dell'Enciclopedia Italiana G. Treccani.

Broeder, Daan / Offenga, Freddy / Willems, Don / Wittenburg, Peter, 2001. «The IMDI Metadata set, its Tools and accessible Linguistic Databases», *Proceedings of the IRCS Workshop on Linguistic Databases*.

De Mauro, Tullio / Mancini, Federico / Vedovelli, Massimo / Voghera, Miriam, 1993. *Lessico di frequenza dell'italiano parlato*, Milano, Etaslibri.

Cresti, Emanuela/Moneglia, Massimo (ed.), 2005. *C-Oral-Rom. Integrated reference corpora for spoken romance languages*, Amsterdam /Philadelphia, Benjamins.

Edwards, Jane A., 1993. «Principles and Contrasting Systems of Discourse Transcription», in: Edwards, Jane A. / Lampert, M. D. (ed.), *Talking Data: Transcription and Coding in Discourse Research*, Hillsdale, N. J., Lawrence Erlbaum Associates, 3-32.

Leech, Geoffrey / Meyers, Greg / Thomas, Jenny (ed.), 1995. S*poken English on Computer. Transcription, Mark-up and Applications*, New York, Longman Publishing, 82-98.

Lüdeling, Anke / Kytö, Merja (ed.), 2008. *Corpus Linguistics. An International Handbook*, Berlin/New York, De Gruyter.

O'Keeffe, Anne / McCarthy, Michael (ed.), 2010. *The Routledge Handbook of Corpus Linguistics*, London / NY, Taylor & Francis Ltd Routledge.

Voghera, Miriam, 2010a. *Lingua parlata, voce della Enciclopedia dell'Italiano diretta da R. Simone*, Istituto dell'Enciclopedia Italiana G. Treccani.

Voghera, Miriam, 2010b. «Parlare italiano: towards a multidimensional description and a multidisciplinary explanation», in: Pettorino, Massimo / Giannini, Antonella / Dovetto, Francesca M. (ed.), *La comunicazione parlata 3*, Napoli, Università degli Studi di Napoli L'Orientale, 603-617.

Savy, Renata / Cutugno, Francesco, 2009. CLIPS. «Diatopic, diamesic and diaphasic variations in spoken Italian», in: Mahlberg, Michaela / González-Díaz, Victorina / Smith, Catherine, *Proceedings of Vth Corpus Linguistic Conference (CL2009)*, 1-24 Liverpool, University of Liverpool.