Il corpus CLIPS,
presentazione del progetto a cura di Federico Albano Leoni

Il *corpus* di italiano parlato che qui si presenta e si rende integralmente pubblico, è parte di un progetto finanziato da MURST, poi MIUR, partito il 5 febbraio 1999 e concluso nel 2004. Il progetto, come mostra il suo acronimo (*Corpora e Lessici di Italiano Parlato e Scritto – CLIPS*), era finalizzato alla messa a punto di strumenti per lo studio generale e per il trattamento automatico dell'italiano, tanto nella sua forma scritta quanto nella sua forma parlata.

Per quanto riguarda la sezione relativa al parlato, il progetto ha consentito di colmare una lacuna negli strumenti per lo studio dell'italiano dal punto di vista linguistico e da quello applicativo, in un momento in cui, da ambedue i punti di vista, l'interesse per la comunicazione parlata è in forte crescita.

La necessità di disporre di strumenti di base per lo studio delle lingue, e in particolare per la loro dimensione parlata, è ormai largamente nota (McEnery & Wilson 1996). Tra questi strumenti, i *corpora* di parlato, acquisiti in varie condizioni, sono di fondamentale importanza da due punti di vista: a) per la descrizione e la conoscenza del funzionamento della lingua parlata in tutte le condizioni di impiego; b) per la predisposizione di strumenti applicativi che servano come base per la realizzazione di sistemi di riconoscimento del parlato e di produzione di voce sintetica di buona qualità, con particolare riferimento all'intonazione.

Questi due obiettivi sono connessi tra di loro molto più strettamente di quanto si potrebbe pensare, e il secondo dipende dal primo. Infatti, l'esperienza di molti paesi, e in primo luogo degli USA, mostra che tanto da un punto di vista operativo, quanto da quello economico, è sbagliato procedere solo alla predisposizione di strumenti di ambito circoscritto, immediatamente ed esclusivamente finalizzati ad una determinata applicazione. Infatti le caratteristiche di funzionamento di una lingua sono generali ed è antieconomico, oltre che metodologicamente scorretto, tentare di descriverle solo su piccoli sottoinsiemi specifici. Inoltre, la somma di questi strumenti settoriali non sarà mai pari alla struttura nel suo complesso.

Ne consegue che una strategia efficace di predisposizione di strumenti, in grado di soddisfare tanto le esigenze di conoscenza generale della lingua, quanto

quelle della produzione di applicazioni di buona qualità, è una strategia che intreccia strettamente gli aspetti generali con quelli particolari.

Il primo strumento per l'attuazione di questa strategia integrata è la costituzione di *corpora* stratificati e calibrati. Le lingue naturali sono infatti caratterizzate da una fortissima variabilità in tutte le loro manifestazioni (Sobrero 1993; Berruto 1995), ed è noto da tempo che questa caratteristica si manifesta in modo drammatico proprio nel parlato (Brown 1990). Questa constatazione, di per sé banale, non affiora sempre al nostro senso comune perché la percezione del fenomeno è offuscata da due fattori: il filtro della rappresentazione scritta, che dà una impressione di stabilità e facile segmentabilità alla lingua; il tipo di atteggiamento normativo nei confronti della lingua a cui siamo addestrati dalla scuola. Ma le cose non stanno così, perché in realtà ogni nostro atto comunicativo si colloca in una determinata posizione rispetto alle seguenti variabili.

A) variabile regionale: il modo in cui parliamo dipende anche dalla nostra regione di provenienza.

B) variabile sociale: il modo in cui parliamo dipende anche dal nostro grado di istruzione, dal nostro mestiere, dall'ambiente sociale a cui apparteniamo.

C) variabile stilistica: il modo in cui parliamo dipende anche dalla situazione in cui ci troviamo di volta in volta;

D) variabile individuale: il modo in cui parliamo dipende anche da caratteristiche anatomiche e idiosincratiche di ciascuno di noi.

Il risultato è che, in un certo senso, ogni nostro atto comunicativo è irripetibile. Infatti bisogna ricordare che la variabilità di cui stiamo parlando investe tutti i livelli della comunicazione parlata, dai suoni alla grammatica e al lessico, ed è proprio questo aspetto che rende così complessi i processi naturali della decodifica e che costituisce il principale ostacolo alla produzione di sistemi di riconoscimento del parlato di validità ed efficacia generali, come ben sanno ricercatori e produttori di questo settore. Da ciò consegue la necessità che nello studio del parlato si introducano strumenti di tipo probabilistico, costruiti, appunto, a partire da basi di dati significative.

Un *corpus* stratificato è dunque un *corpus* nel quale siano presenti, in proporzioni opportune che riflettano anche le varietà regionali, le diverse varietà della lingua parlata, da quella di laboratorio (parlato controllato di *speakers* professionisti), a quelle via via meno formali, fino alle varietà più spontanee,

includendo voci maschili, voci femminili, nonché campioni di parlato telefonico e radiotelevisivo.

Un *corpus* di questo genere tende a coprire una gamma significativa dei possibili tipi di situazione comunicativa, dal punto di vista della fonologia, della prosodia, della morfologia, della sintassi e del lessico di base, e costituisce il punto di partenza per la descrizione dei modi concreti in cui avviene la comunicazione. Ma esso è anche la base a partire dalla quale diventa più agevole e economico predisporre gli strumenti per le applicazioni specifiche, perché disponendo di modelli statistici della distribuzione dei fenomeni nei vari tipi di comunicazione parlata, riguardanti la fonetica, la prosodia, la morfologia e la sintassi, i problemi residui sarebbero quasi esclusivamente connessi al lessico raro o di bassa frequenza, e, in quanto tali, legati prevalentemente all'ambito su cui si sta lavorando.

Gli utenti di un *corpus* di questo genere sono dunque la comunità scientifica dei linguisti e quella degli *speech engineers*. Per questi ultimi, se si considera che nell'ambito del trattamento automatico delle lingue un obiettivo molto importante è quello della conversione automatica dello scritto in parlato e del parlato in scritto, da realizzare con il minor numero possibile di vincoli (sia testuali, sia legati al parlatore), nonché quello della valutazione e del collaudo dei sistemi automatici di riconoscimento in condizioni variabili, appare che la disponibilità di un *corpus* generale e stratificato è uno strumento irrinunciabile.

In questo quadro, si osservava per l'italiano una anomalia rispetto alla situazione di altre lingue di cultura. E' utile ricordare qui in modo sommario la storia della vicenda che ha portato alla costituzione di *CLIPS*.

Il parlato italiano aveva cominciato ad essere oggetto di studi specifici a partire dagli anni Ottanta del Novecento, grazie ad importanti iniziative dell'Accademia della Crusca ed alla pubblicazione di lavori nei quali esso veniva visto non come la mera versione fonica della lingua, né come la sua varietà .'popolare' o 'informale' o 'viva', o 'ridotta', ma come una sua modalità d'uso peculiare, complessa e stratificata, al punto che si discusse la questione se parlato e scritto avessero due grammatiche diverse o avessero invece, come poi è sembrato ai più, la stessa grammatica profonda, con la conseguente messa a fuoco dei problemi, anche teorici, dell'esecuzione.

A partire da quegli anni le indagini cominciano a essere condotte su parlato tendente al naturale, raccolto e registrato da ciascuno studioso in modo diverso, più o meno sistematico. Per fare alcuni esempi, Sornicola (1981), Berretta (1985), Voghera (1993), Bazzanella (1994) allestiscono ciascuna proprie raccolte, finalizzate alla ricerca da svolgere.

Per commentare questa nuova situazione mi servo delle parole di Alberto Sobrero (1985), nel suo contributo a un volume di Holtus e Radtke (1985):

> Sulle caratteristiche dell'italiano parlato disponiamo oggi di molti studi parziali, relativi ad aree geografiche limitate e/o a singoli fenomeni linguistici […]. Manca tuttavia uno studio d'insieme, e la ragione è evidente: i *corpora* rilevati sono parziali, differenti, occasionali, e non possono dar luogo ad analisi complessive.

Sobrero poneva dunque il problema molto serio della mancanza di un *corpus* complessivo che fornisse la base per lo studio del parlato in sé, in grado di coglierne tanto la naturalezza in atto, quanto la variabilità e l'indeterminatezza che lo caratterizzano.

Un passo deciso in questa direzione si ebbe nel 1993, con la pubblicazione del *Lessico di frequenza dell'italiano parlato* (De Mauro e altri 1993), computato a partire dal primo vero e proprio *corpus* di italiano parlato: un *corpus* pubblico, costruito secondo criteri espliciti e per finalità esplicite, stratificato sia dal punto di vista regionale (era stato raccolto materiale parlato a Milano, Firenze, Roma e Napoli), sia dal punto di vista delle varietà stilistiche (era articolato in conversazioni faccia a faccia di vario tipo, telefonate, parlato ufficiale o comunque pubblico, parlato radiotelevisivo). Il *corpus* era inoltre accessibile (il volume era accompagnato dai dischetti che contenevano le registrazioni sbobinate del materiale raccolto), verificabile, in parte riutilizzabile, anche se, date le finalità lessicografiche della raccolta, la qualità delle registrazioni era per lo più troppo scadente per essere oggetto di analisi fonetiche. Per la prima volta un *corpus* poteva essere utilizzato anche da soggetti diversi da quelli che lo avevano predisposto, come mostra il gran numero di lavori analitici che ne sono nati.

Il LIP aveva mostrato una strada e gli anni successivi, fino ad oggi, hanno visto la nascita di numerosi *corpora* di parlato per iniziativa di istituzioni e centri, come l'Accademia della Crusca, il Lablita dell'Università di Firenze, il CIRASS dell'Università di Napoli Federico II, la Fondazione Ugo Bordoni di Roma, il

CNR di Padova (prescindendo dalle raccolte di parlato di laboratorio, curate in genere da é*quipes* tecnologiche a fini applicativi).

Il *corpus* che qui si rende pubblico nasce, in questo quadro, con le caratteristiche di un *corpus* tendenzialmente generale e certamente molto stratificato. Qui ne riassumiamo i caratteri principali, rinviando, per tutti i dettagli, ai *Documenti* raccolti in una apposita sezione di questo sito.

CLIPS consiste di circa 100 ore di parlato, equamente ripartito tra voci maschili e voci femminili, in parte trascritto, segmentato ed annotato dal punto di vista fonetico segmentale, ed è caratterizzato da una duplice stratificazione, diatopica e diafasica.

Per quanto riguarda la variazione diatopica, una indagine sociolinguistica preliminare, condotta su tutto il territorio nazionale dall'Università di Lecce, ha stabilito i punti di raccolta dei materiali, scelti in modo da essere rappresentativi tanto dal punto di vista della varietà di italiano, quanto da quello della significatività demografica e socioeconomica delle località.

Le località prescelte sono state le città di Bari, Bergamo, Bologna, Cagliari, Catanzaro, Firenze, Genova, Lecce, Milano, Napoli, Palermo, Parma, Perugia, Roma, Venezia

Per quanto riguarda la variazione diafasica, il materiale è così articolato:

a) parlato radiotelevisivo (notiziari, interviste, *talk shows*);

b) parlato raccolto sul campo (dialoghi raccolti secondo le modalità del *map task* e del 'gioco delle differenze');

c) parlato letto;

d) parlato telefonico.

I materiali indicati ai punti a), b) e d) sono stati raccolti in pari misura nelle località indicate; i materiali indicati al punto c) sono stati in parte registrati dagli stessi partecipanti ai dialoghi (e quindi ripartiti regionalmente), in parte da parlatori professionisti di una società di doppiaggio di Roma.

Il parlato radiotelevisivo è stato raccolto da 4 tipologie di programmi (intrattenimento, informazione e servizio, cultura e divulgazione, pubblicità), suddivisi equamente tra emittenti radiofoniche e televisive e proporzionalmente tra emittenti locali e nazionali, per ciascuna delle località prescelte.

Il parlato dialogico è stato acquisito organizzando dialoghi di tipo *map task* e 'gioco delle differenze'. Vi hanno partecipato 12 coppie di parlatori per ciascuna

delle 15 località, per un totale di 360 parlatori, equamente suddivisi tra uomini e donne, prevalentemente studenti e studentesse universitari, di età compresa tra i venti e i trenta anni).

Il parlato letto è stato prodotto a) dagli stessi parlatori che hanno partecipato ai dialoghi (per consentire il confronto tra parlato spontaneo e parlato letto da parte della stessa persona) e b) da un gruppo di parlatori professionisti di una compagnia di doppiaggio cinematografico (per fornire una esecuzione parlata che si approssimi al parlato iperarticolato standard). I testi letti sono atti a garantire la copertura lessicale e fonotattica di base e sono costruiti con parole scelte dalle frequenze alte e medie dei lessici di frequenza correnti.

Il parlato telefonico, acquisito a fini prevalentemente applicativi, consiste in conversazioni tra circa 300 parlatori e un operatore turistico simulato.

Tutti i protocolli di raccolta e di analisi del materiale tengono conto, naturalmente, oltre che della esperienza acquisita nella raccolta e organizzazione di altri *corpora* (come *API* 2003), delle direttive del progetto *EAGLES* per l'acquisizione di *corpora* di parlato. Uno dei requisiti fondamentali dei *corpora* di questo genere e della loro analisi è infatti che essi siano confrontabili con i *corpora* prodotti in altri paesi.

Chi ha scritto questa presentazione lo ha fatto perché ha avuto l'onore e l'onere di essere il responsabile del progetto. Tuttavia, nel momento in cui il *corpus* è reso pubblico, il progetto terminato e il suo ruolo esaurito, gli corre il gradito obbligo di ricordare che ciò che viene qui presentato è il risultato di un vero lavoro di squadra, nel quale tutti i partecipanti, giovani e meno giovani ricercatori e personale amministrativo, hanno svolto con intelligenza e dedizione compiti di cruciale importanza, ma nel quale tuttavia Francesco Cutugno e Renata Savy hanno occupato un posto preminente e insostituibile in tutte le fasi di questo lungo e spesso faticoso lavoro, così che essi vanno considerati, e sono *optimo iure,* coautori e corresponsabili di *CLIPS*.

Federico Albano Leoni

Bibliografia

API, 2003, *Archivio del Parlato Italiano,* progetto coordinato da Federico Albano Leoni, *DVD* a c. di Claudia Crocco e altri, CIRASS-Università di Napoli Federico II.

Bazzanella, Carla, 1994, *Le facce del parlare. Un approccio pragmatico all'italiano parlato*, Scandicci, La Nuova Italia.

Berretta, Monica, 1985, *I pronomi clitici nell'italiano parlato*, in Holtus-Radtke 1985, pp. 185-224.

Berruto, Gaetano, 1995, *Fondamenti di sociolinguistica*, Laterza, Bari-Roma

Brown, Gillian, 1990, *Listening to Spoken English*, Longman, London (seconda ediz.).

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

Holtus, Günter, Edgar Radtke (a c. di), 1985, *Gesprochenes Italienisch in Geschichte und Gegenwart*, Narr, Tübingen.

McEnery, Tony & Andrew Wilson, 1996, *Corpus Linguistics*, Edinburgh University Press, Edinburgh.

Sobrero, Alberto (a c. di), 1993, *Introduzione all'italiano contemporaneo*. II. *La variazione e gli usi*, Laterza, Bari-Roma

Sobrero, Alberto, 1985, *Per una prima raccolta sistematica di dati sull'italiano parlato in Salento*, in Holtus-Radtke, 1985, pp. 77-85.

Sornicola, Rosanna, 1981, *Sul parlato*, Bologna, il Mulino.

Voghera, Miriam, 1992, *Sintassi e intonazione nell'italiano parlato*, il Mulino, Bologna.