L'evento

Corpora e lessici di italiano parlato e scritto (*CLIPS*)

I testi che seguono sono la versione modificata delle relazioni tenute a Napoli il 29 maggio 2007, al Convegno organizzato dall'Università "Federico II", in occasione dell'apertura ufficiale del sito (www.clips.unina.it) che ospita il *corpus* detto *CLIPS* (*Corpora* e lessici di italiano parlato e scritto). Di questo progetto l'Università di Napoli "Federico II", nella persona di Federico Albano Leoni, ha avuto la responsabilità complessiva, ma la parte relativa allo scritto fu curata dall'Istituto di Linguistica computazionale (ILC) del CNR di Pisa e da suoi consorziati. In queste pagine si tratterà solo della sezione parlata.

Gli stessi testi che qui si presentano sono stati inseriti, in formato elettronico, anche nel sito www.parlaritaliano.it dell'Università di Salerno. Al Convegno parteciparono con relazioni anche Pier Marco Bertinetto, Francesco Cutugno e Renata Savy, che hanno rinunciato a presentare i loro testi scritti.

Si ringrazia calorosamente il prof. Elio Marciano, direttore del CIRASS, che ha reso possibile quella occasione di incontro e di discussione.

Un frammento di storia recente della ricerca (linguistica) italiana. Il *corpus* CLIPS

di *Federico Albano Leoni*

1
Premessa

Visitando il sito www.clips.unina.it si entra in un *corpus* di italiano parlato: circa 100 ore di registrazioni, tutte liberamente accessibili, ascoltabili, analizzabili e, come si dice, scaricabili. I documenti che le accompagnano consentono di farsi un quadro molto dettagliato delle caratteristiche del *corpus*, delle modalità di raccolta, archiviazione e trattamento dei dati.

Del *corpus* in sé, della sua architettura, delle sue proprietà, scriveranno altri in questo fascicolo. Io qui vorrei ricordarne sommariamente la storia e ricostruirne le premesse: progetti di questo genere, infatti, sono, nel bene e nel male, il risultato dei tempi e del lavoro delle persone e strutture che vi sono state attive. Apparirà così un frammento di storia della ricerca italiana, piccolo ma forse non privo di interesse generale. Penso infatti che la storia di CLIPS sia una storia esemplare di come proceda la ricerca, non solo umanistica, in Italia.

La storia formale è molto semplice. Il progetto CLIPS, finanziato dalla legge 1992, n. 488[1], è iniziato il 5 febbraio 2000, si è concluso il 4 febbraio 2004 ed è stato messo in rete nel 2007, con un ritardo dovuto sia ai necessari ritocchi e aggiustamenti, sia al progressivo esaurirsi delle risorse finanziarie e umane che gli avevano dato vita.

2
Le condizioni al contorno

La storia formale va però situata dentro una storia più ampia, perché è lì che si determinarono le condizioni per la progettazione, il finanziamento e la realizzazione del progetto.

1. CLIPS fu uno degli otto progetti del cluster C18 "Linguistica computazionale: ricerche monolingui e multilingui", finanziato dal Ministero dell'Istruzione, dell'università e della ricerca (MIUR) sulla legge 488. La fonte del finanziamento, cioè una legge a sostegno delle industrie e della innovazione tecnologica, mostra l'importanza accordata alle finalità applicative del *corpus* CLIPS, come si vedrà più avanti. Il progetto era triennale ma, in conseguenza di un ritardo iniziale da parte del Ministero, di cui dirò più avanti, fu necessario un anno di proroga.

Gli anni che vanno dalla fine degli Ottanta del Novecento fino al 2000 sono stati una stagione breve ma fervida per la fonetica italiana, sia generale, sia applicata, e sembrò che questa disciplina potesse uscire, anche in Italia, dallo stato di marginalità in cui si era sempre trovata[2].

L'elemento saliente di questo periodo fu la costituzione, nel 1988, del Gruppo di fonetica sperimentale (GFS) in seno all'Associazione italiana di acustica (AIA), per iniziativa di Franco Ferrero, un fisico, uno studioso di grande rigore, di grande prestigio, di grande passione, fra i rinnovatori di questi studi in Italia, attivo presso il Centro studi di fonetica del CNR di Padova, che all'epoca era il punto più avanzato della ricerca italiana in questo settore, grazie a una illuminata compresenza di studiosi di fonetica, di audiologia, di fisica acustica, di linguistica.

La costituzione del GFS dette migliore visibilità a una situazione interessante che si era determinata già da qualche anno in seno all'AIA. Questa infatti, nata nel 1972 come associazione scientifica dei fisici acustici, aveva inserito tra gli argomenti di proprio interesse, sul modello della Acoustical Society of America e delle altre consorelle europee, la fonetica e le questioni del parlato. Dunque aderivano all'AIA, oltre alla componente accademica, proveniente dalle facoltà di fisica e di ingegneria e da centri CNR, anche numerosi ricercatori e funzionari di aziende e centri attivi nel settore delle tecnologie della voce, anche essi, con rarissime eccezioni, fisici o ingegneri. Nel 1988 Franco Ferrero invitò i fonetisti italiani di formazione linguistica, provenienti in genere dalle facoltà umanistiche, ad aderire all'AIA per costituirvi un gruppo specifico dedicato alla fonetica sperimentale. La risposta, relativamente ai modesti numeri in gioco, fu massiccia.

Si aprì cosi una stagione che, per chi scrive e, credo, per molti altri, fu di grande importanza scientifica, di grande vivacità e di grande crescita. Ai congressi annuali dell'AIA e poi a quelli del GFS si incontravano e si parlavano, anche se a volte faticosamente, persone di formazione diversa. Nella fase iniziale di questo periodo i fonetisti di formazione linguistica incontravano dunque ricercatori e tecnici della Fondazione Ugo Bordoni (FUB) di Roma, dell'Olivetti, dell'IBM, del Centro studi e laboratori telecomunicazioni (CSELT) di Torino, dell'ALCATEL FACE, dell'Istituto per la ricerca scientifica e tecnologica (IRST) di Trento, dell'Istituto superiore poste e telecomunicazioni (ISPT), e con questi si confrontavano: da loro apprendevano tecniche e protocolli di analisi; a loro mostravano, o almeno cercavamo di mostrare, come le lingue fossero oggetti complessi, non sempre riducibili alla serenità di una rappresentazione binaria. La fonetica italiana ne uscì profondamente trasformata e svecchiata: il GFS fu la

2. La fonetica in Italia non ha mai avuto statuto solido: la sua presenza è sporadica sul piano didattico (e in gran parte dipendente dalle inclinazioni personali dei docenti di linguistica), per lo più limitata a rudimenti di fonetica articolatoria e di trascrizione, e non è parte stabile di *curricula*; sono rari i casi in cui profili scientifici di fonetisti siano stati apprezzati dalle commissioni giudicatrici di concorso. Non migliore è la situazione sul versante scientifico-tecnologico: fisici e ingegneri che si dedicano alle tecnologie della voce fanno poca carriera.

palestra nella quale si sono formati, direttamente o indirettamente, i quadri attuali degli studi sulla voce e sul parlato.

L'altro momento saliente nella storia che sto rievocando fu la costituzione, pure negli anni Novanta, del primo gruppo di lavoro sul Trattamento automatico delle lingue (TAL), voluto e animato da Antonio Zampolli, fondatore e direttore dell'Istituto di Linguistica computazionale (ILC), organizzatore di progetti italiani ed europei, uno dei pochi linguisti italiani che univa straordinarie capacità organizzative e manageriali a una lucida visione dell'importanza del trasferimento dei risultati della ricerca nelle applicazioni, come mostrava l'esperienza dei paesi occidentali scientificamente e tecnologicamente più avanzati, dove questo trasferimento costituisce un elemento portante e strategico.

Così, in parallelo a quanto accadeva nell'AIA e nel GFS, intorno al tavolo voluto da Zampolli e patrocinato da qualche illuminato dirigente del Ministero delle Poste e telecomunicazioni, sedettero rappresentanti della ricerca e delle imprese, discutendo, o forse sognando, di progetti per la predisposizione anche per l'italiano delle cosiddette risorse linguistiche, cioè di quegli insiemi di dati (*corpora*, lessici, software robusti per la lemmatizzazione o per la conversione grafema-fonema automatiche ecc.) che in genere forniscono la base per le elaborazioni e lo sviluppo di tecnologie applicative, delle quali la più complessa è quella che consente l'interazione vocale uomo-macchina. Per questo rispetto l'Italia era infatti in ritardo e intorno al tavolo TAL si discuteva di come rimediare[3].

Infine, *si parva licet componere magnis*, quelli furono anche gli anni della costituzione e dello sviluppo del Centro interdipartimentale di ricerca per l'analisi e la sintesi dei segnali (CIRASS) dell'Università di Napoli "Federico II", che nacque formalmente il 1 gennaio 1990, grazie alla volontà di collaborazione di un piccolo gruppo di linguisti, fisici, audiologi e poi anche di ingegneri. Il CIRASS, che sarebbe stato dieci anni dopo l'attuatore del progetto *CLIPS*, nasceva su una ipotesi di fonetica multidisciplinare, si ispirava a centri italiani ed europei, ed era dunque lo specchio delle tendenze e delle illusioni di quel periodo.

In questo quadro, finché durò, uno dei punti di convergenza delle energie che portarono poi a concepire e realizzare il progetto *CLIPS* fu appunto la questione delle cosiddette risorse linguistiche e in particolare dei *corpora* di parlato. Questi infatti sono riconosciuti non solo come strumenti di base per lo sviluppo di ogni tecnologia di riconoscimento o di produzione di voce, ma anche come strumenti che consentono allo studio linguistico del parlato la verificabilità dei risultati e delle procedure, la ripetibilità degli esperimenti, la riutilizzazione delle risorse. Essi sono dunque un oggetto sul quale convergono gli interessi sia della ricerca linguistica generale, sia di quella applicata.

3. Il gruppo TAL si è poi formalizzato nel 2002, per iniziativa del Ministero delle Comunicazioni, in Forum-TAL. Ne fanno parte istituti di ricerca pubblici e privati nonché rappresentanti di vari ministeri (l'università è presente in modo marginale). Non risulta però che abbia dato avvio a progetti nazionali significativi.

3
Segnali di deterioramento

Questo era il quadro degli anni Novanta, o almeno sembrava. Ma, se qualcuno pensa che qui si stia raccontando una favola bella, si rassicuri: la stagione durò poco, perché negli stessi anni comparvero segnali che andavano in senso contrario.

Forse in conseguenza di una divisione del lavoro internazionale, per la quale all'Italia non toccava sviluppare tecnologie nel settore della voce, o forse, più semplicemente, come conseguenza dello scarso interesse italiano per l'innovazione, sta di fatto che l'incipiente collaborazione scientifica tra università, centri di ricerca e imprese fu progressivamente ridimensionata: chiusero in rapida successione le divisioni e i centri studi della Olivetti (che uscì definitivamente di scena), dell'IBM, e dell'ALCATEL FACE (i cui centri di studio furono trasferiti all'estero) e scomparvero dunque dall'AIA e dal GFS; quelli che rimasero, la FUB, lo CSELT (da una cui costola vocale nacque poi "Loquendo", che fornisce tecnologie vocali alla Telecom), l'IRST, in mancanza di finanziamenti pubblici certi e di commesse private, si ripiegarono su se stessi, preoccupati per la propria sopravvivenza più che proiettati in avanti.

Il TAL di conseguenza si rivelò un'impresa disperata in Italia perché l'idea moderna che lo aveva animato dovette confrontarsi con l'ambiente di casa nostra, dove la scienza è spesso separata (ma forse in questo caso non lo era), l'imprenditoria è timida (e preferisce al rischio e all'investimento il bagnomaria delle provvidenze statali), una politica razionale di sostegno alla ricerca è assente. Tuttavia, come vedremo, il TAL fece fortunosamente in tempo a essere il destinatario di qualche risorsa finanziaria.

Infine è forse da vedere come un segno di queste difficoltà anche la crisi del GFS, iniziata alla fine degli anni Novanta, che portò poco dopo al suo scioglimento: dalle ceneri sorsero due distinte associazioni, una prevalentemente linguistica, una prevalentemente tecnologica.

4
Il tirocinio

Ma, come dicevo all'inizio, i progetti marciano con le persone. Dovendo cercare un punto di inizio simbolico, ma anche concreto, del percorso che in quegli anni portò alla formazione del gruppo che avrebbe costituito la struttura portante di *CLIPS* lo collocherei nel 1995: in quell'anno fu approvato e finanziato un piccolo progetto interuniversitario di ricerca di fonetica sperimentale. Il progetto era coordinato dal CIRASS e vi partecipavano la Scuola Normale Superiore di Pisa e l'Università di Padova con il Centro di fonetica del CNR (i luoghi dove allora era più vivo l'interesse per la fonetica). Fu una prima collaborazione, un rodaggio, che, tra l'altro, consentì il consolidarsi di rapporti che sarebbero durati nel tempo.

L'anno successivo il Ministero, con un provvedimento una volta tanto assennato, cambiò i modi di finanziamento della ricerca universitaria passando

dal sistema detto "a pioggia" (praticamente tutte le ricerche venivano finanziate nella misura media del 10% della cifra richiesta), a un sistema di finanziamento pressoché integrale di progetti interuniversitari, detti progetti di rilevante interesse nazionale (PRIN), presentati secondo regole complesse, con una partecipazione finanziaria delle università che si consorziavano, valutati da giudici anonimi. Era certamente una svolta positiva che consentì al gruppo che era stato finanziato nel 1995, e poi allargato a nuovi compagni di strada, di rafforzarsi e avviare concretamente la raccolta di *corpora* di italiano parlato.

Così, nel 1997 partì un progetto biennale, chiamato "Archivio delle varietà italiane di parlato" (*AVIP*), coordinato dalla Normale di Pisa e a cui parteciparono il CIRASS di Napoli e il Politecnico di Bari. Ne uscì, mi sia consentito dirlo, la prima raccolta veramente pubblica di materiale parlato italiano semispontaneo (dialoghi elicitati con il metodo detto del *map task* a Pisa, Napoli e Bari, nonché parlato infantile di bambini sordi a confronto con quello di normoudenti, riflesso della presenza nel CIRASS di un gruppo di audiologi). I dialoghi, registrati e in parte annotati foneticamente secondo criteri espliciti, furono riversati in CD e distribuiti gratuitamente agli interessati.

I buoni risultati ottenuti ci indussero a proseguire nel 1999 con il progetto detto "Archivio del parlato italiano" (*API*), coordinato dal CIRASS e a cui parteciparono unità di ricerca di Napoli (CIRASS e Policlinico), della Normale di Pisa, dell'Università "Orientale", dell'Università del Piemonte Orientale "Amedeo Avogadro", del Politecnico di Bari, dell'Università "Ca' Foscari". Facevano quindi parte della cordata non più solo fonetisti e informatici, ma anche studiosi di linguistica computazionale. Il *corpus AVIP* venne arricchito di nuovi materiali e di nuovi strumenti e ripubblicato in DVD con il nome di *API*, anche questa volta distribuito gratuitamente.

Il *corpus API*, anche se di dimensioni relativamente limitate, conteneva comunque molto materiale, molto più di quanto noi fossimo riusciti ad analizzare con le risorse di cui disponevamo. Ci sembrò dunque utile proporre un progetto destinato principalmente non alla raccolta di nuovo materiale ma all'analisi e all'approfondimento di quello già disponibile. Così nel 2001 fu finanziato un progetto detto "Italiano parlato" (IPAR), coordinato dal CIRASS e a cui parteciparono 12 università (Napoli "Federico II", Seconda Università di Napoli, Napoli "Orientale", Salerno, Normale di Pisa, Perugia, Piemonte Orientale, Torino, Siena, Roma "La Sapienza", Perugia).

Da questi progetti felicemente portati a termine uscirono, oltre ai *corpora* in sé (*AVIP* e *API* raccolti in DVD), pubblicazioni scientifiche, libri, Atti di congressi, tesi di laurea e di dottorato. Questi progetti furono inoltre il luogo della formazione di giovani studiosi: per molti rappresentarono il percorso formativo fondamentale, per alcuni, purtroppo pochi[4], il percorso si concluse con l'inserimento nell'università.

4. Di giovani studiosi che si possano considerare in qualche modo miei allievi di ambito fonetico solo una è attualmente inserita in una università italiana e altri cinque sono all'estero.

Ma se mi sono soffermato su questo itinerario è stato per sottolineare, in questa sede, come la formazione di quadri che fossero all'altezza della esecuzione del progetto *CLIPS* abbia richiesto a tutti, giovani e meno giovani, un tirocinio di circa cinque anni.

5
Il progetto *CLIPS*

Dunque, ricapitolando e riannodando fili apparentemente sconnessi, nel corso degli anni Novanta era accaduto che: *a*) grazie alla frequentazione degli ambienti TAL e GFS, avevamo imparato a ragionare anche in termini di applicazioni della ricerca fonetica e linguistica; *b*) grazie alla esperienza maturata durante i progetti PRIN avevamo imparato a organizzare risorse linguistiche, in particolare *corpora* di italiano parlato, e a mettere a punto gli strumenti di supporto; *c*) infine, avevamo imparato a lavorare in gruppi di competenza mista e a parlare, noi linguisti, con fisici, con informatici, con ingegneri e, in qualche misura, a capirli e a farci capire: senza questa osmosi nulla sarebbe stato possibile di quanto avevamo fatto nei PRIN e di quanto avremmo fatto in *CLIPS*.

Date queste condizioni, era stato agevole e quasi naturale concepire un'idea e un progetto in sé banali ma nuovi per l'italiano: costruire un *corpus* che tenesse conto, in misura ragionevole, delle complesse articolazioni in cui si manifesta ogni lingua, sotto forma di variazioni geografiche e variazioni stilistiche: un progetto ambizioso, complesso, costoso, che richiedeva competenze diverse, che intendeva andare incontro a esigenze tanto dei tecnologi, quanto dei linguisti, che avrebbe colmato una grave lacuna nelle nostre cosiddette infrastrutture linguistiche[5].

A questa idea fu possibile dare corpo grazie alla felice compresenza, all'interno del gruppo TAL, di Zampolli e del CIRASS (che aveva ormai una esperienza rispettabile nello studio del parlato). Sul finire del 1998 Zampolli mi segnalò il bando per le richieste di finanziamento ai sensi della legge 488/92 e mi invitò a partecipare. Noi ci inserimmo in un sussulto terminale: il bando arrivava quando stavano sparendo i potenziali utilizzatori industriali italiani dei risultati del progetto.

Fra la notizia del bando e l'avvio dei lavori ci fu una faticosa incubazione di circa due anni, dedicati alla predisposizione dei dettagli esecutivi e al reperimento dei partner. A conclusione di questa fase presentammo un progetto, il cui bilancio complessivo, tra erogazioni ministeriali e cofinanziamento degli enti che partecipavano, si aggirava intorno ai quattro miliardi di lire, dei quali più della metà erano per la sezione parlata. Il progetto fu approvato.

Al CIRASS, il cosiddetto "soggetto attuatore" nella terminologia ministeriale, toccava il comando, cioè il coordinamento generale e amministrativo, e anche

5. Infatti, per l'italiano parlato, si disponeva solo di *corpora* di dimensioni variabili, quasi mai collegati fra loro, acquisiti e annotati con criteri diversi, costruiti prevalentemente con finalità applicative o descrittive circoscritte, mai di dominio veramente pubblico.

numerose azioni esecutive; all'Università di Lecce toccava il compito delicato di tracciare la mappa delle località di raccolta (15 città) in modo motivato dal punto di vista linguistico, demografico e socioeconomico, nonché la predisposizione di alcuni testi che sarebbero stati letti; alla Normale di Pisa toccava il coordinamento della raccolta dei materiali nel Centro-Nord; alla FUB toccava, oltre al compito della raccolta del parlato telefonico, quello della predisposizione di alcuni strumenti software, anche in collaborazione con il Politecnico di Bari; all'ISPT (oggi ISCOM), in collaborazione con la FUB, toccò il compito di organizzare la lettura e la registrazione di alcuni testi opportunamente predisposti da Lecce.

Il 5 febbraio del 2000, data di inizio ufficiale stabilita dal decreto di finanziamento, eravamo tutti pronti a partire ma, secondo un costume proprio di tutti gli enti pubblici italiani, il Ministero tardava a trasferire i fondi. Dopo alcuni mesi di attesa, quando cominciava a diventare tangibile il rischio che il complesso gruppo messo insieme da noi si dissolvesse, un provvidenziale anticipo di mezzo miliardo di lire, disposto dalla "Federico II" a giugno 2000, sbloccò la situazione[6].

Il lavoro partì e fu ovviamente complesso. Si trattava infatti di addestrare le persone ai rispettivi compiti; di coordinare gli operatori nelle 15 località prescelte che organizzassero le sedute di registrazione di dialoghi con le persone giuste e nel modo giusto e che organizzassero le registrazioni radio-televisive; di rispettare la concatenazione temporale delle diverse azioni; di archiviare il materiale, controllandone la qualità e l'omogeneità, effettuando le eventuali revisioni; di adattare i criteri generali di trascrizione, annotazione, segmentazione ed etichettatura ai problemi pratici continuamente insorgenti; di controllare i flussi finanziari in entrata e in uscita. Tutto ciò fu realizzato da una squadra alla quale parteciparono, in modo e misura diversi, 68 persone, delle quali 17 appartenevano agli organici delle strutture coinvolte e 51 erano giovani studiosi a contratto.

Ora il *corpus* è pronto, disponibile, accessibile. Delle sue qualità e dei suoi difetti diranno, dai rispettivi punti di vista, gli altri contributi di questa sezione e, più ancora, diranno coloro che lo stanno utilizzando e lo utilizzeranno. Alcuni riscontri sono comunque confortanti. *CLIPS* in effetti aveva cominciato a essere utilizzato per singole ricerche e per tesi di laurea e di dottorato prima ancora che fosse completato ed è quindi presente da tempo nelle bibliografie. Da quando è stato aperto il sito è stato visitato da diverse centinaia di visitatori, dei quali sono italiani meno della metà. *CLIPS* ha inoltre richiamato l'attenzione di enti pubblici e privati stranieri (come la Microsoft e il laboratorio LINCOLN del MIT)[7].

6. È superfluo ricordare che lo stesso Ministero che a giugno del 2000, cinque mesi dopo l'avvio ufficiale del progetto, non aveva ancora trasferito una lira, al 5 luglio esigeva inflessibilmente la rendicontazione delle spese effettuate nel primo semestre, senza la quale non avrebbe erogato le rate successive. Ma di questo tipo di difficoltà, che furono numerose e tutte surreali, non dirò, per non essere ripetitivo, se non di una che mi sembra esemplare: il progetto era articolato in azioni di durata, costi e complessità diversi, ma la rendicontazione era inesorabilmente semestrale (non ci fu verso di far capire che i tempi e i ritmi della rendicontazione erano convenzioni umane, modificabili secondo le convenienze operative, e non cicli lunari).

7. Non risulta invece che *CLIPS* abbia richiamato l'attenzione di coloro a cui era destinata

6
Conclusioni

Tutto bene dunque? Sembrerebbe di sì, ma per esprimere un giudizio più compiuto è necessario considerare ancora due aspetti riguardanti la politica della ricerca.

Il progetto ha superato tutti i vagli contabili. Il giudizio finale del giudice ministeriale, nominato da molto tempo (nella persona di un autorevole studioso di intelligenza artificiale), più volte sollecitato, è arrivato con molto ritardo[8]. Bisogna dunque riconoscere che quello che interessa veramente al Ministero (anzi, ai ministeri che si sono succeduti in questi anni), a giudicare dai contenuti dei controlli, è solo la contabilità. Infatti, mentre il soggetto attuatore è stato letteralmente tormentato, nella fase preliminare, in quella esecutiva e in quella conclusiva, dai funzionari ministeriali e bancari su questioni contabili che qui non voglio elencare, nessuno ha mai chiesto nulla circa i contenuti scientifici del progetto, sui metodi proposti, le finalità, le dimensioni, le applicazioni, i protocolli. Come nessuno entrò mai nel merito del progetto quando lo presentammo, nessuno, almeno fino al momento in cui questo scritto viene licenziato (ottobre 2007), ha mai chiesto di vederne il risultato: in questi sette anni, può essere utile saperlo, il Ministero, o chi per esso, non ha mai chiesto di ascoltare un brano registrato, di controllare una trascrizione, di leggere un documento di lavoro, insomma di vedere cosa c'era in una scatola che, secondo i miei ingenui parametri, era costata all'erario una cifra enorme. Per quanto ne sa il Ministero, la scatola potrebbe anche essere vuota, e forse il Ministero si meriterebbe davvero che così fosse, in una perversa armonia tra governanti e governati. Ma, come ho detto prima, *CLIPS* non è una scatola vuota.

Il secondo aspetto, ma che forse è il risvolto del primo, è che il patrimonio di risorse umane, di competenze, di energie, che si era condensato intorno al CIRASS a partire dal 1995 nei progetti PRIN che ho ricordato e poi nel progetto *CLIPS*, si è dissipato. Quel patrimonio era costituito in piccola parte da persone in organico e per il resto da giovani laureati, dottorandi, dottori di ricerca, appassionati, generosi, bravi e precari. L'esaurirsi delle risorse finanziarie e la mancanza di prospettive di integrazione in strutture di ricerca hanno generato la diaspora: alcune persone sono uscite di scena, altre (poche) hanno trovato sistemazioni precarie altrove, qualcuno all'estero, una è entrata nell'organico universitario.

Io credo che qui si osservi il riflesso del modo casuale in cui queste vicende si sono snodate e in cui se ne snodano altre consimili.

Bisogna infatti riconoscere che in fondo quel patrimonio si era costituito gra-

la legge che lo ha finanziato, cioè gli imprenditori italiani. È uno dei paradossi che costellano questa storia.

8. Di conseguenza il Ministero ha liquidato il saldo del finanziamento a novembre 2007, e quindi a sua volta il CIRASS ha potuto saldare il debito di mezzo miliardo con la "Federico II".

zie a una molteplicità di fattori positivi ma casuali ed era il risultato di un progetto razionale voluto non da una facoltà, da una università, da un Ministero (strutture che permangono e che dovrebbero garantire continuità nel realizzare una politica), ma voluto da un piccolo gruppo di persone, che passano. I fattori positivi, ma casuali, erano stati: *a*) la pervicacia con cui alcuni avevano tenuto in vita il CIRASS e avevano partecipato, a titolo personale, alle vicende degli anni Novanta; *b*) una breve stagione di finanziamenti ragionevoli; *c*) il fortunoso accesso ai fondi della legge 488/92. Questi fattori positivi, altrettanto casualmente, si sono rovesciati in negativo: qualche trasferimento di singole persone ad altre università (ovviamente senza sostituzione), il calo della pervicacia di alcuni, una chiusura dei cordoni della borsa ministeriale, qualche malumore.

7
Breve congedo

Dunque, il *corpus* CLIPS esiste, ha una sia pur elettronica corposità, ed esistono, sia pure in parte disperse, le persone che lo hanno ideato e costruito. Per tutto questo non posso non manifestare la mia soddisfazione personale e soprattutto la mia gratitudine e ammirazione per la squadra che ha realizzato il progetto e per la "Federico II" che ci ha sostenuto. Considerando le condizioni in cui lavoriamo in questo paese, e considerando in particolare le condizioni in cui lavora la ricerca umanistica, l'esistenza di CLIPS mi sembra un miracolo.

Però un sito che ospita un *corpus* non è come un libro che, una volta pubblicato, va per la sua strada e non richiede altro, ma è un oggetto dinamico, bisognoso di cure: continua manutenzione, aggiornamenti, arricchimenti, senza di che rischia di diventare presto obsoleto e inutile. Credo che questa sia la sfida che dobbiamo raccogliere tutti, e in particolare quelli che in qualche modo hanno preso il testimone di quelle ricerche e quelle attività e quelli che oggi rappresentano l'università che lo ospita.

Il *CLIP* negli studi sul parlato

di *Alberto A. Sobrero*

Per l'italiano parlato disponiamo a tutt'oggi solo di corpora di dimensioni variabili, non sempre o quasi mai collegati fra loro, acquisiti e annotati con criteri diversi, costruiti prevalentemente per finalità applicative o descrittive circoscritte. 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 del collaudo dei sistemi automatici di riconoscimento in condizioni variabili, si capisce l'importanza di un corpus generale e stratificato di italiano parlato.

Questo scriveva Albano Leoni nel 2000[1]. Per questa occasione mi è sembrato utile verificare queste affermazioni, aggiornandole all'anno di grazia 2007.

Partendo dalla "Banca dati dell'italiano parlato" di Graz ho provato ad allestire una lista dei *corpora* di italiano parlato che sono stati resi disponibili – in varie forme, sia cartacee (libro, articolo) che elettroniche (CD-ROM, banche dati) – negli ultimi quarant'anni (a parte il *CLIPS*, e salvo errori od omissioni)[2].

Sigla	Autori	Anno	Parole	Località/Area
STA-1	Harro Stammerjohann	1965/1970	13.000	Firenze S. Spirito
ARNU	Anna Maria Arnuzzo	1976		Basso Monferrato (italiano popolare)
ROV	Giovanni Rovere	1977	60.000	emigrati (italiano popolare)
BIAN	Sandro Bianconi	1980		Canton Ticino
FP	Sandro Fontana / Maurizio Pieretti (a cura di)	1980		Bergamo-Brescia-Cremona (italiano popolare-italiano regionale)
LOY	Nanni Loy	1981		
FMR	Fabio Foresti /	1982	40.000	S. Giovanni in Persiceto (BO)

(*segue*)

1. F. Albano Leoni, *Tre progetti per l'italiano parlato*, in *Italia linguistica anno Mille. Italia linguistica anno duemila*, Atti del XXXIV Congresso SLI (Firenze, 19-21 ottobre 2000), a cura di N. Maraschio, T. Poggi Salani, Bulzoni, Roma 2003, pp. 675-83: 676.

2. Per ogni *corpus* riporto la sigla, il nome dell'autore o curatore, l'anno di pubblicazione, il numero approssimativo delle entrate lessicali (quando disponibile), la/le località o l'area in cui i dati sono stati raccolti.

(*seguito*)

Sigla	Autori	Anno	Parole	Località/Area
	Paola Morisi / Maria Resca (a cura di)			(italiano regionale-italiano popolare)
BRAN	Luciana Brandi	1987	5.800	Toscana (a scuola)
CRES	Emanuela Cresti	1987	2.700	Firenze (a colazione)
RV	Elena Rizzi / Giuseppe Vincenzi	1987		Bologna (italiano regionale)
PIXI	Laura Gavioli / Gillian Mansfield (a cura di)	1990	22.000	Bologna / Nord / Centro
LIP	Tullio De Mauro / Federico Mancini / Massimo Vedovelli / Miriam Voghera	1993	490.000	Napoli, Roma, Firenze, Milano
FRA	Rita Franceschini	1998	6.650	Svizzera (analisi conversazionale)
CIP (LABLITA)	Emanuela Cresti (a cura di)	2000	58.300	Toscana
AVIP/API	Federico Albano Leoni *et al.*	2000-2003	35.000	Pisa, Napoli, Bari
CAFF	Claudia Caffi	2001	13.500	Nord (contesti terapeutici)
VIN	Stefan Rabanus	2001		Lecco-Reggio Calabria
BERT	Cristina Bertoli Sand	2002	19.500	Centro-Sud (pragmatica)
C-ORAL-ROM	Emanuela Cresti / Massimo Moneglia (a cura di)	2005	1.200.000	Toscana
STA-2 (LABLITA)	Harro Stammerjohann / Emanuela Cresti / Massimo Moneglia (a cura di)	2006	100.000	Firenze (v. *STA-1*)
CPT	Antonella Giannini / Massimo Pettorino / Ilaria Vitagliano (a cura di)	2007		Parlato telegiornalistico
LIPS	Massimo Vedovelli (a cura di)	2007	500.000	Parlato di stranieri
ROM	Antonella Stefinlongo / Riccardo Cimaglia / Silvia Di Baia	2007		Roma

Risultano subito evidenti due tipi di eterogeneità:
– per dimensione (cfr. colonna 4)
– per localizzazione (cfr. colonna 5).
a) Per dimensione: si va dalle 2.000-3.000 parole di piccoli *corpora* finalizzati a ricerche specifiche, di area ovviamente circoscritta, a più di un milione, per *corpora* di dimensione nazionale, con campioni ampi e variegati.
b) Per localizzazione. I criteri sono i più vari. Dei 22 *corpora*[3] elencati:
– 6 sono stati raccolti in una o più città di un'area di dimensione regionale o subregionale (3 in Toscana, 1 in Canton Ticino, 1 nel Basso Monferrato, 1 a Bergamo-Brescia-Cremona);

3. *STA-2* comprende e assorbe i testi di *STA-1*: i due *corpora* sono perciò conteggiati una sola volta.

– 4 sono stati raccolti in singole città o paesi (Roma, Bologna, Firenze, S. Giovanni in Persiceto);
– 3 in due o più città che non appartengono alla stessa area linguistica (Napoli/Roma/Firenze/ Milano, Pisa/Napoli/Bari, Lecco/Reggio Calabria);
– 2 con criteri misti, o addirittura vaghi, di localizzazione (Bologna/Nord/Centro; Centro/Sud);
– 7 non considerano la variabile diatopica in quanto i brani di parlato sono stati selezionati con criteri diversi dalla localizzazione: la variabile indipendente era di volta in volta costituita dalla condizione di emigrati, dalla condizione di immigrati, da una situazione comunicativa specifica (conversazione in esercizi commerciali, a colazione, in viaggio, in contesto terapeutico ecc.), da una varietà diamesica (parlato telegiornalistico).

Alcuni, inoltre, sono mirati a varietà specifiche e ad analisi di vario tipo: italiano popolare, italiano regionale, analisi conversazionale, analisi pragmatica.

Guardando al nostro elenco, dunque, possiamo confermare, anche con certificazioni aggiornate al 2006, che, come diceva l'Albano Leoni citato prima, i *corpora* disponibili sono «di dimensioni variabili, non sempre o quasi mai collegati fra loro, acquisiti e annotati con criteri diversi, costruiti prevalentemente per finalità applicative o descrittive circoscritte»[4].

Ma l'elenco che abbiamo visto prima è anche la prova provata, l'ennesima dimostrazione dell'inconciliabilità pratica dei due mega-obiettivi di ogni raccolta di dati linguistici: la *rappresentatività* su larga scala e la *comparabilità*. I *corpora* oggi a disposizione privilegiano di volta in volta l'uno o l'altro obiettivo, in modi vari, condizionati da specifiche linee di ricerca. E non potrebbe essere diversamente: la disputa sulle tecniche di raccolta dati ha salde radici nella metodologia, e addirittura nell'epistemologia dialettologica, dove ha chiamato in causa le esigenze da una parte dell'investigazione grammaticale (sistemica) o sociolinguistica (variazionista) e dall'altra le esigenze dell'indagine diatopica (geolinguistica), portando a privilegiare di volta in volta la monografia relativa al punto (il *case study*) o l'atlante linguistico. Che, come ora sappiamo, non sono alternativi ma complementari. Esattamente come i *corpora* rappresentativi e quelli comparabili.

Torniamo ai nostri *corpora*. Quanto alla *rappresentatività*, è distribuita su due parametri: *sociale* (rappresentatività dell'universo della popolazione italiana, o di parti significative di esso) e *varietistico* (rappresentatività delle varietà del repertorio linguistico italiano). Diamo pure per scontata la dolorosa rinuncia alla fascia della produzione dialettale, e accantoniamo – per ora – la spinosa questione del parlato mistilingue e del cambio/alternanza di codice: impostando le scelte sul criterio della rappresentatività bisogna quantomeno dar conto delle variazioni diafasica, diatopica, diastratica, diamesica.

Esplorando l'elenco appena presentato la nostra attenzione si ferma su alcuni *corpora*, che si sono posti in modo più acuto questo problema e hanno ri-

4. Albano Leoni, *Tre progetti*, cit., p. 676.

sposto – in modi, ahimè, sempre diversi – a una o più di queste complesse esigenze.

Il *corpus* Stammerjohann, il più antico, è diatopicamente invariabile (presenta materiale fiorentino, anzi specificamente del rione di S. Spirito) ma presenta una ricca articolazione su parametri diamesici (parlato naturale, parlato telefonico, parlato radiofonico) e diafasici (tipo di contesto sociale, tipo di interazione, struttura dell'evento), con intuizioni sociolinguistiche ampiamente precorritrici (non dimentichiamo che il materiale è del 1965, anche se è stato ripreso e riedito in forma più completa e moderna nel 2006 in ambiente LABLITA).

L'*AVIP*, com'è noto, contiene testi (semi)spontanei ottenuti con le mappe, con la lettura dei toponimi e con una lista di parole organizzata intorno ai principali fatti di interesse fonosintattico. Offre un primo, interessante ma insufficiente avvicinamento alla rappresentatività diatopica, presentando campioni di quelli che – con larga approssimazione – si potrebbero definire rispettivamente l'italiano regionale toscano (in realtà è toscano occidentale), campano (propriamente: napoletano) e pugliese (in realtà: barese). Assente l'italiano regionale settentrionale. Punti di forza: la durata complessiva dei materiali registrati (circa 20 ore), la quantità degli informatori (88 locutori complessivi), la ricchezza lessicale (35.000 parole). Punti di debolezza: le altre dimensioni della variazione (classi di situazioni, variabili sociolinguistiche ecc.) e il grado di "naturalezza" dei testi.

Un altro *corpus* di tutto rispetto è il *CIP* di Emanuela Cresti, che non affronta il problema della variabilità diatopica ma affronta molto bene quello della rappresentatività diafasica, diastratica e diamesica, con impostazione pragma-socio-linguistica. Punti di forza: i parametri principali utilizzati (dominio d'uso e struttura del dialogo); la naturalezza del parlato spontaneo, riconducibile per il 50% a parlato informale e per il resto a parlato formale (ma in contesto naturale) e a parlato dei *media*.

Il *LIP* (1993) gode di una progettazione attenta alla pluralità di contesti e di tipi di discorso: le 500.000 entrate lessicali provengono da interazioni raggruppate in 5 macroclassi (tre tipi di scambio comunicativo bidirezionale, uno di scambio unidirezionale e uno di parlato trasmesso: circa un quinto delle parole provengono da emissioni radio e TV) e numerose sottoclassi, che danno luogo a una vasta tipologia di testi. È anche il *corpus* che affronta con più consapevolezza degli altri il problema della rappresentatività della variazione diatopica: è infatti concepito – dice De Mauro – come «rappresentativo sia dal punto di vista dei generi di parlato [...] sia dal punto di vista geolinguistico»[5]. Milano, Firenze, Roma, Napoli sono selezionati come punti di rilevamento con specifiche motivazioni di rappresentatività areale, al termine delle quali l'autore conclude:

5. T. De Mauro, *Dai vincoli statistici alle scelte sociolinguistiche e geolinguistiche*, in T. De Mauro, F. Mancini, M. Vedovelli, M. Voghera, *Lessico di frequenza dell'italiano parlato*, Etaslibri, Milano 1993, pp. 23-32: 30.

> integrato il criterio del peso demografico con quello del prestigio linguistico e bloccando a quattro i centri di prelievo si ottiene una complessiva documentazione che accoglie equilibratamente le testimonianze centrali sia toscane sia non toscane, romane, le settentrionali e le meridionali[6].

Ma si tratta pur sempre di quattro soli punti, perché, come osserva lo stesso De Mauro,

> una raccolta di documenti fatta tenendo conto, come pure sarebbe desiderabile, di almeno tutti i grandi centri urbani e "capitali" della policentrica Italia, ci avrebbe portato a un frazionamento dei materiali tale da renderli sì qualitativamente interessanti, ma con dati statisticamente poco significativi dal punto di vista della confrontabilità interna delle liste di frequenza[7].

Inoltre, si può dire sottorappresentata la variabilità legata alla tipologia urbana (non sono rappresentati centri medi e piccoli).

Appare molto ricco il *LIR* (LABLITA, a cura di Nicoletta Maraschio e Stefania Stefanelli), un insieme di lessici di frequenza dell'italiano radiofonico, ottenuti con la registrazione e trascrizione di 50 ore di parlato radiofonico, trascritto ortograficamente e allineato all'audio mediante software, lemmatizzato e pubblicato in CD-ROM. Il vocabolario contemplerà circa 500.000 occorrenze lessicali. Per definizione, è un "*corpus* di sottolinguaggi", limitato al parlato radiofonico, e in quanto tale considera unicamente variabili legate al mezzo e agli scopi: i generi radiofonici, le tipologie comunicative, il progetto complessivo dell'emittente.

Analogamente, il *CIT* (*Corpus* di italiano televisivo), organizzato presso l'Università per Stranieri di Perugia da Stefania Spina, prevede una base dati di 250.000 parole, da incrementare fino a 500.000, ma è anch'esso un "*corpus* di sottolinguaggi", per definizione limitato al parlato televisivo (non di fiction). Considera le variabili specificamente legate al mezzo e ai suoi generi.

Il *C-ORAL-ROM*, ricchissimo *corpus* multimediale del parlato romanzo, realizza un progetto europeo: la sezione italiana di questo "*corpus* plurilingue" è costituita da oltre 1.000.000 di parole, raccolte in tre *subcorpora*: *a*) italiano parlato spontaneo adulto: situazioni comunicative diafasicamente diverse; *b*) italiano registrato nella fase del primo apprendimento (18-36 mesi); *c*) italiano dei *media* (radio-TV-cinema). Le cifre sono imponenti: 122 ore di registrazione, 1.427 parlanti, 600.000 parole solo nel corpus di italiano parlato spontaneo, e così via. I parametri di scelta dei testi – selezionati in quanto ritenuti i maggiori responsabili della variabilità del parlato – sono tre:

– struttura del testo (monologhi, dialoghi, conversazioni);
– dominio (famiglia, vita privata, vita pubblica);
– caratteri individuali (sesso, età, istruzione, occupazione).

6. Ivi, p. 31.
7. Ivi, p. 30.

Non compare la variabile diatopica né la variabile "tipologia del punto". Punto di forza, anzi di eccellenza: la comparabilità di ogni *corpus* – o meglio *sottocorpus* – nel suo complesso con altri *corpora* di lingue europee, sul piano lessicale, sintattico, intonativo, comunicativo.

Questo per quanto riguarda la rappresentatività sociolinguistica.

Quanto alla *comparabilità* dei dati, è assicurata in molti – anzi, in quasi tutti – i *corpora* una comparabilità più o meno "larga": i testi sono selezionati tenendo conto di volta in volta del dominio, dell'intenzione comunicativa, del tipo di testo ecc., e sono comparabili a un livello che va dal (sub)regionale all'internazionale. La comparabilità in senso stretto è invece assicurata in parte dal LIR e dal CIT e in modo pieno dall'AVIP, il solo, se non erro, ad assicurare omogeneità del campione (tutti 20-30enni, studenti universitari, nati e cresciuti nel centro investigato, da genitori preferibilmente indigeni) e omogeneità tipologica dei dati.

In sintesi, osservando a volo d'uccello il quadro dei *corpora* dell'italiano oggi (o fra breve) disponibili si rileva una priorità quasi generalizzata per la *rappresentatività* rispetto alla *comparabilità* in senso stretto. All'interno delle dimensioni della variabilità mi sembra che ci sia un'attenzione generalizzata – direi quasi unanime – per la resa della variazione diafasica e subito dopo, in un'ideale graduatoria, per le dimensioni diamesica e diastratica. È meno rappresentata nei campioni utilizzati la varietà diatopica dell'italiano. I motivi sono tanti, e tutti ottimi (a partire dalle dimensioni di un campionamento significativo e dalle difficoltà organizzative e logistiche). Sta di fatto che proprio la compresenza di varietà regionali, che caratterizza l'italiano in modo più marcato rispetto alle altre lingue nazionali romanze, risulta sottorappresentata.

Tenendo conto di tutto questo e guardando, ora, al CLIP, si può dunque dire che esso si colloca nella fascia alta di un'ideale classifica di innovatività e di sperimentalità dei *corpora* di parlato d'Italia, in quanto è caratterizzato: *a*) da un'attenzione specifica per la comparabilità dei dati, superiore alla grande maggioranza – per non dire alla totalità – degli altri *corpora*; *b*) dalla ricerca della rappresentatività non solo per la stratificazione diafasica e diamesica ma anche per quella diatopica, che negli altri *corpora* trova nel complesso un numero basso di testimonianze. In altre parole, il CLIP orienta le scelte proprio sui settori che sono in genere sottorappresentati. Da una parte conferma l'utilizzazione – già presente nell'AVIP – di tecniche di elicitazione orientate verso la comparabilità puntuale dei dati raccolti, a qualunque livello (fonetico, prosodico, morfosintattico, lessicale, testuale, pragmatico ecc.), dall'altra offre dati relativi alla variabilità diatopica, in riferimento tanto alle principali varietà di italiano regionale quanto alla variazione di tipo demografico (tipologia dei centri di raccolta).

Vorrei qui richiamare l'attenzione sul metodo e sulle tecniche di campionamento.

Per assicurare un alto grado di *comparabilità* si è scelto un campione sostanzialmente omogeneo, per classe di età, *status* socioeconomico, livello di istru-

zione, residenza in centri grandi e medi. Per quanto riguarda la *rappresentatività*, la selezione dei punti è stata fatta utilizzando parametri diversi:
– geolinguistici, per assicurare che fossero rappresentate nel campione le principali aree linguistiche d'Italia, e dunque le principali varietà diatopiche d'italiano;
– sociolinguistici, per tener conto dei principali fenomeni sociolinguistici in atto nella popolazione;
– socioeconomici, per selezionare località che fossero rappresentative delle realtà socioeconomiche più significative del paese.

Quest'ultimo criterio introduce un correttivo, o meglio integra il criterio dimensionale che di solito si applica utilizzando come unico indicatore la popolazione. Si sono utilizzati allo scopo più indicatori:
– *indicatori di sviluppo*, sia statici (peso percentuale di agricoltura, industria e servizi, quota di reddito prodotto) che dinamici (tasso di incremento del valore aggiunto);
– *indicatori di dinamicità* del cluster socioeconomico: consistenza e dinamica demografica, infrastrutture (economiche e sociali), tipologia urbana, ruolo della città nel processo di sviluppo economico territoriale e nazionale.

Ai diversi indicatori sono stati assegnati "pesi" leggermente diversi: ad esempio, un peso minore per quote di reddito, un peso maggiore per dotazione e domanda di infrastrutture.

Come si vede, il taglio è innovativo, rispetto ai *corpora* esistenti, caratterizzandosi con un deciso ancoraggio a criteri di scelta delle località di tipo geo-socio-linguistico, con una particolare attenzione per la struttura socioeconomica del punto di prelievo, esaminata in modo più articolato rispetto ai consueti indicatori di tipo dimensionale.

L'attenzione specifica per il campionamento e per la rappresentatività, soprattutto al livello diatopico, unita a una progettazione specificamente funzionale alla comparabilità dei dati (per metodi, strumenti e procedure di elicitazione) inducono a vedere nel CLIPS uno strumento ricco, duttile, ampiamente fruibile per analisi sul parlato diverse e complementari – in prospettiva, integrate – condotte, sui diversi livelli della lingua, con le finalità e con gli approcci più disparati (geo-socio-pragmalinguistici ecc.).

La prima conferma della correttezza di questa valutazione è costituita da un recente, originale, ricco, stimolante studio a più voci, curato dallo stesso Federico Albano Leoni e da Rosa Giordano[8], nel quale uno dei dialoghi del *corpus* è stato analizzato, a titolo sperimentale, da prospettive diverse: fonetico-fonologica segmentale, pragmatico-discorsiva, morfosintattica, intonativo-informativa. Con risultati molto, molto interessanti.

Tuttavia, la formazione e l'esperienza ormai quarantennale di dialettologo – che vuol dire quarant'anni di convivenza fianco a fianco coi problemi, teorici e pratici, di raccolta e analisi di parlato-parlato – mi inducono ad aggiunge-

8. F. Albano Leoni, R. Giordano, *Italiano parlato. Analisi di un dialogo*, Liguori, Napoli 2005.

re almeno una riflessione, ad uso degli utilizzatori del *CLIPS*. È un invito a trattare sempre con molta cautela dati che sono comunque il frutto di un'elicitazione, la quale induce a produrre testi la cui distanza dalla produzione reale non è di molto inferiore rispetto ai questionari di traduzione che usiamo – o abbiamo usato – noi dialettologi.

Occorrerà *non cedere alla tentazione di considerare questi come campioni di parlato spontaneo*. Non lo sono, per quattro buoni motivi.

a) Il parlato spontaneo – fatti salvi i casi di parlato trasmesso, che hanno regole loro specifiche – avviene in situazione di faccia-a-faccia, con conseguenze importanti sulla distribuzione dell'informazione su più canali (gestualità, prossemica, sguardi ecc.). Nel test delle differenze e nella *map task* l'interazione è solo fittiziamente *in praesentia*, perché la mancanza di visibilità reciproca dei due parlanti toglie loro la possibilità di una comunicazione multicanale, che a sua volta influenza l'articolazione verbale, le scelte linguistiche e il sincrono gestuale-verbale.

b) Il comportamento linguistico in ambienti controllati e in risposta a sollecitazioni standardizzate e fortemente indirizzate è altamente prevedibile (e questa è una qualità importante, per la paragonabilità dei dati), ma il parlato spontaneo è per definizione non prevedibile. Controprova: se gli informatori "sbrigano" il loro compito in pochi minuti, il loro comportamento è considerato insoddisfacente perché non forniscono materiale abbastanza ricco per le nostre analisi. Questa rigidità è molto utile ai fini della consistenza del *corpus* ma è anche, direi, "ecologicamente scorretta".

c) Nella registrazione le procedure, altamente standardizzate, sono controllate da un mediatore esterno, che assiste con la consegna di evitare di intervenire, ovvero di limitare all'essenziale i suoi interventi. È l'incarnazione del "paradosso dell'osservatore", problema cruciale nella sensibilità del dialettologo, forse meno sentito nella linguistica "di laboratorio": la presenza di un estraneo alla conversazione, in qualche modo sovraordinato ai due attori, è di per sé fatto di turbamento della naturalezza della situazione. A questo bisogna aggiungere l'azione del fattore imponderabile – perché sconosciuto al fruitore del dialogo – costituito dal quadro relazionale che lega i soggetti intervistati all'assistente al dialogo, e dall'interpretazione che ogni assistente dà del ruolo direttivo che gli è assegnato.

d) La qualità acustico-fonetica delle registrazioni è ottima, ma la sua rappresentatività del parlato spontaneo è in parte condizionata – e spesso ridotta – proprio dal fatto che, nella realtà, il parlato spontaneo non è mai "pulito": la sua qualità acustico-fonetica è, come sappiamo, quanto meno variabile. E quanto più cerchiamo di modificare il comportamento del parlante per avere un messaggio "pulito", tanto più ci allontaniamo dal parlato spontaneo.

In generale, si può dire che variabilità e non-prevedibilità sono caratteri costitutivi specifici del parlato, scambi comunicativi multicanale *in praesentia* e "correttezza ecologica" (senza perturbazioni) sono costitutivi della conversazione: gli uni e gli altri vengono necessariamente sacrificati sull'altare della comparabilità. I dati raccolti in ambienti così controllati hanno dunque un al-

to grado di comparabilità, ma un grado non elevato di rappresentatività del parlato prodotto *in quanto parlato spontaneo*, ai livelli più avanzati dell'analisi linguistica. O, quanto meno, ai livelli sui quali più si è sacrificato in fatto di naturalezza.

Ad esempio: il dialettologo ha molta esitazione a ritenere che si possa condurre uno studio raffinato sulle sovrapposizioni, utilizzando dialoghi elicitati con un apparato di registrazione sofisticato, che – anche in assenza di istruzioni precise, che pure invece vengono date – induce chissà quanti parlanti a seguire quella che ha imparato come regola aurea di ogni trasmissione radiofonica (o televisiva): *non sovrapporre le voci* (il telespettatore appassionato di calcio ricorderà l'urlaccio "non sovrapponetevi !!!!" di Aldo Biscardi). Quello che si studia, in questo caso, è un parlato semi-controllato (o, per qualcuno, "controllato"). Che è utilissimo studiare, ma non come campione di parlato spontaneo.

Il *corpus* CLIPS e il Trattamento automatico delle lingue

di *Andrea Paoloni*

1
Premessa

Si ritiene opportuno premettere che con il termine TAL (Trattamento automatico delle lingue) o TL (Tecnologie linguistiche, in inglese HLT – Human Language Technologies) vengono designate quelle discipline che si occupano di modelli, metodi, tecnologie, sistemi e applicazioni relativi all'elaborazione automatica della lingua, sia parlata sia scritta (Di Carlo, Paoloni, 2004). Il TAL comprende dunque sia lo "Speech Processing" (SP) o elaborazione del parlato, sia il "Natural Language Processing" (NLP) o elaborazione del testo e si pone obiettivi strettamente connessi, quali l'interazione vocale uomo-computer o la comprensione del linguaggio umano per servizi come traduzione automatica e il reperimento di informazioni.

La prima area (SP) è volta a riprodurre la capacità umana di comunicare attraverso la parola e comprende la codifica del segnale vocale, il cui obiettivo è la riduzione della quantità di informazione da trasmettere o memorizzare, la sintesi da testo, ovvero la realizzazione della macchina in grado di leggere un testo qualsiasi, il riconoscimento del parlato, ovvero la macchina in grado di scrivere e, infine, il riconoscimento del parlante.

La seconda area (NLP) tende a riprodurre la capacità umana di comprendere il linguaggio e prevede, dal punto di vista dei componenti e metodi utilizzati, analizzatori sintattici e semantici, basati su moduli algoritmici o statistici, modelli di rappresentazione della conoscenza, basati su dizionari o enciclopedie, metodologie di apprendimento automatico e tecniche di annotazione e classificazione quale punto di partenza per il reperimento dell'informazione, mentre, dal punto di vista delle applicazioni, oltre alla traduzione automatica, che riveste una importanza particolare nell'Europa dalle molte lingue, comprende i temi della gestione del dialogo, della produzione di sommari, dei motori di ricerca, della gestione della conoscenza.

Queste tecnologie, quella della elaborazione del parlato e quella della elaborazione dello scritto, sono vissute per lungo tempo separate, anche a cagione della diversa origine degli studi di base, prevalentemente orientati alla fisica nel primo caso, all'applicazione dell'informatica, alla linguistica ovvero agli studi di linguistica computazionale, nel secondo.

Nella convinzione che una migliore integrazione tra i due settori possa portare ad importanti progressi scientifici e tecnologici, il Ministero delle Telecomunicazioni e la Fondazione Bordoni hanno promosso la costituzione di un forum, denominato Forum-TAL (www.forumtal.it), con i seguenti obiettivi: monitorare l'attività degli enti coinvolti nel TAL per ottenere maggiori sinergie; promuovere la ricerca e lo sviluppo di strumenti linguistici innovativi; studiare le iniziative dirette all'ampliamento del mercato e allo sviluppo della competitività dell'industria del settore; promuovere gli investimenti pubblici e privati, anche per la salvaguardia della lingua italiana e la sua diffusione nel mondo; studiare il fenomeno dell'evoluzione del TAL con particolare attenzione allo sviluppo di iniziative in ambito europeo; promuovere l'uso della lingua italiana all'estero con particolare riferimento alla sua utilizzazione nelle sedi europee.

2
Tipologia dei *corpora*

Chiarito cosa si intende con il termine TAL si ritiene ora opportuno illustrare il concetto di *corpus*. Un *corpus* non è altro che una raccolta di materiale e nel caso particolare dei *corpora* linguistici, si tratta di raccolte di testi o di segnali vocali. Per quanto attiene alle raccolte di testi, esse non sono certo cosa nuova; cos'era la biblioteca di Alessandria se non un insieme di *corpus*, e cosa sono i vocabolari se non uno strumento realizzato a partire da *corpora*? Per quanto attiene alle raccolte di parlato la possibilità di memorizzare la voce è divenuta reale solo nell'Ottocento con l'invenzione del fonografo dovuta a T. A. Edison. Possiamo certamente affermare che è la possibilità di fissare su supporti durevoli la lingua a consentire la realizzazione dei *corpora*.

L'insieme dei *corpora*, opportunamente integrati in una struttura informatica comprendente procedure di definizione, di ritrovamento e di manutenzione dei dati, costituisce una "base di dati" che può essere messa a disposizione di una comunità tecnico-scientifica.

Esistono *corpora* di diverse tipologie in funzione di diverse applicazioni. Vi sono *corpora* riconducibili ad una particolare applicazione: ad esempio, per la costruzione di un call center per la prenotazione ferroviaria è necessario disporre dei dati audio relativi alle interazioni tra centralinista e utente; vi sono poi i *corpora* adibiti alla misura delle prestazioni dei sistemi, ad esempio i *corpora* di cifre connesse utilizzati per valutare le prestazioni dei sistemi che appunto riconoscono le cifre.

I *corpora* possono essere caratterizzati in vari modi e sotto vari punti di vista. Una prima caratterizzazione può essere quella sociolinguistica: diacronia, diatopia, diastratia, diafasia, diamesia. Per illustrare meglio la rilevanza che i sopraelencati effetti hanno sul segnale vocale si propongono qui alcuni esempi del loro "impiego" nella ricerca, nello sviluppo e nelle applicazioni.

La diacronia, ovvero l'evoluzione temporale del segnale, è presente nel *corpus* FOCUS per il riconoscimento del parlante (Falcone, Barone, 2003), contiene

campioni di voce dello stesso informatore prelevati a distanza di tempo (un anno); la diatopia, presente anche nel *corpus* CLIPS, consiste nella registrazione di parlato ottenuto da informatori di origine geografica diversa; la diastratia rappresenta la variabile sociale, il grado di istruzione del parlante, non ha particolare impiego nelle basi di dati più diffuse; la diafasia invece ha notevole rilievo e rappresenta la variazione delle diverse tipologie del parlato che vanno dal parlato letto al parlato formale, così sino al parlato spontaneo informale; infine la diamesia rappresenta l'influenza diretta e indiretta che il mezzo utilizzato per comunicare ha sul segnale: si pensi al peggioramento di qualità prodotto dall'uso del telefono. L'altro asse che differenzia tra loro i *corpora* è quello legato all'obiettivo della "raccolta": i *corpora* possono essere collazionati a scopo di studio, come è probabilmente avvenuto per le prime raccolte, possono essere collazionati a scopo didattico, possono essere collazionati per preservare una cultura, come sta avvenendo nei confronti di lingue che si stanno spegnendo e come è già avvenuto per lingue diffusissime come il latino (a chi non piacerebbe avere la registrazione di un'orazione di Cicerone); possono essere collazionati per costruire una macchina, e di questo impiego discuteremo nel seguito, infine possono essere collazionati per poter valutare le prestazioni di un sistema, e questo è stato uno dei primi obiettivi dei *corpora* di dati vocali.

3
Applicazioni delle basi di dati vocali

Lo studio del linguaggio articolato ebbe un grande impulso con la nascita dei primi sistemi di analisi e di memorizzazione del segnale acustico. Per analizzare il parlato ciascuno sperimentatore provvedeva ad emettere, e in seguito anche registrare, i suoni e le parole oggetto del suo studio. Le metodiche utilizzate erano così lente e le analisi in fase così preliminare che non si sentiva l'esigenza di disporre di un *corpus* di dati realmente rappresentativi dei fenomeni in studio. Quando, intorno agli anni Cinquanta, si cominciarono a proporre macchine per il riconoscimento del parlato e del parlante, si cominciò a voler valutare le loro prestazioni sulla base di un materiale audio di dimensioni adeguate. Nacquero così le prime basi di dati vocali. È proprio nella fase della valutazione, fase alla quale contribuirono le prime basi di dati "pubbliche" che ci si accorse che il segnale reale differiva in modo sostanziale da quanto si ritenesse. Il parlato è un fenomeno molto complesso che non si lascia descrivere con un numero limitato di regole: i sistemi cosiddetti "knowledge based" non riuscivano a raggiungere tassi di errore sufficientemente piccoli, tali da poter sperare di costruire sistemi commerciali con la tecnologia del riconoscimento vocale. La soluzione proposta negli anni Ottanta e ancora oggi adottata si basa su un modello statistico del linguaggio, i modelli markoviani. Si tratta di modelli probabilistici che usano un numero finito di stati per modellare le variazioni del segnale vocale. I diversi parametri che definiscono il modello probabilistico vengono "stimati" durante l'addestramento e utilizzati nella fase di riconoscimento (Lee, 1989).

Una base di dati adatta a questo scopo è costituita da un insieme di file che contengono i parametri cepstrali ed energetici relativi al segnale di ogni parola, ai quali è associato un *file* di etichette che contiene la definizione dei limiti temporali delle parole all'interno della finestra di registrazione. L'etichettatura è effettuata manualmente da un operatore, che si avvale di ausili grafici e dell'ascolto delle registrazioni, o automaticamente da un algoritmo che individua i limiti temporali delle parole con opportune tecniche. Ovviamente la completezza della base di dati, cioè la sua adeguatezza a render conto del maggior numero di varianti di pronuncia e intonazione di una parola, avrà risultati rilevanti sulla bontà dell'esito del riconoscimento. Ne consegue che un sistema di riconoscimento acustico ha bisogno di dati accuratamente etichettati e bilanciati, e ovviamente rappresentativi di una competenza linguistica "media". Nel processo di riconoscimento vocale la conoscenza linguistica viene utilizzata anche a livello di modelli statistici del linguaggio (*language models*).

I modelli del linguaggio non impongono di specificare insiemi di frasi lecite che il parlante potrà pronunciare, ma apprendono la probabilità che una certa sequenza ben formata sia riconoscibile in certi contesti di dialogo. I modelli del linguaggio vengono addestrati a partire da insiemi di testi e frasi, sia generali della lingua in cui avviene il riconoscimento, sia specifici del contesto applicativo.

4
AIDA e i primi *corpora* vocali

Nel 1989, anche a seguito di esperienze condotte in varie parti del mondo sulla standardizzazione dei metodi di valutazione e sulla costruzione di basi di dati vocali, esperienze che avevano condotto alla realizzazione di basi di dati in lingua inglese quali TIMIT (frasi lette da 630 diversi parlanti) e ATIS (frasi lette da 630 diversi parlanti) (ESCA, 1989), il Ministero delle Comunicazioni istituì la Commissione nazionale per le basi di dati vocali italiani. Gli obiettivi della Commissione erano i seguenti: creare una base dati vocale fruibile da tutti gli enti; uniformare la raccolta dei dati; uniformare la distribuzione dei *corpora* e coordinarne la produzione; sviluppare un'attività specifica per la lingua italiana.

La Commissione progettò e produsse, nel 1990, una base di dati vocali denominata *AIDA* (Acoustic Italian Database) orientata alla valutazione dei sistemi per il riconoscimento del parlato. *AIDA* costituisce un insieme di dati vocali distribuibile a tutti coloro che abbiano necessità e convenienza ad utilizzare dei dati comuni, esigenza sempre più sentita a livello internazionale, sia a livello di ricerca, sia in compiti più tecnologici, quali ad esempio le omologazioni.

Il *corpus AIDA* è diviso in due *database*, ciascuno contenuto in tre CD-ROM: uno dipendente dai parlanti e uno indipendente. Per le registrazioni sono stati selezionati 40 parlanti (20 maschi e 20 femmine) scelti in parte a Torino e in parte a Roma, sulla base di un criterio di eguale distribuzione sul territorio nazionale. Tutti i parlanti hanno pronunciato una volta il materiale vocale per la parte "Speaker independent", mentre 8 soggetti, selezionati tra i 40, hanno poi

ripetuto altre 5 volte la registrazione dello stesso materiale per la parte "Speaker dependent". Per quanto riguarda il materiale registrato, sono presenti: un brano di calibrazione, bisillabi CV'/ta/ e /t/V'CV (essendo C e V tutte le possibili consonanti e vocali dell'italiano), i principali gruppi bi- e triconsonantici iniziali e intervocalici e le cifre da 0 a 9. In questo lavoro sono state considerate esclusivamente le vocali estratte dai bisillabi con variazione di consonante iniziale e intervocalica. I segnali contenuti in *AIDA* sono stati campionati a 20 kHz, 16 bit (Loj, Di Carlo, Paoloni, 1992).

Il gruppo "Comunicazioni vocali" della Fondazione Ugo Bordoni, proseguendo il lavoro iniziato con la realizzazione del *database AIDA*, ha progettato e realizzato nel 1995 il *database SIVA* (Speaker Identification and Verification Archives) (Contino, Falcone, 1995; Falcone, Gallo, 1996). I *corpora* vocali per l'identificazione del parlatore erano pochi (Godfrey, Holliman, McDaniel, 1995), specialmente se si confronta il loro numero e la loro varietà con quello dei *corpora* disponibili per il riconoscimento del parlato, e *SIVA* è stato il primo *corpus* italiano per il riconoscimento del parlante (Ariyaeeina, Paoloni, 1998). Sullo stesso tema furono in seguito prodotti altri due *corpora*, uno chiamato *CALÌ* (CALler Identification on mobile and fixed network), progettato allo scopo di valutare dei sistemi di identificazione del parlatore su linea telefonica fissa e mobile, l'altro chiamato *FOCUS* (Forensic *corpus*) progettato per le applicazioni forensi (Di Carlo, Falcone, Paoloni, 1994; Paoloni, 2002).

5
Corpora on-line

Superata la fase pionieristica oggi sono disponibili numerosi *corpora* testuali e vocali molti dei quali scaricabili on-line. Naturalmente esistono anche *corpora ad hoc* collazionati da laboratori per usi specifici e che non sono a disposizione del pubblico in forma libera, ad esempio i vocabolari, più o meno indicizzati, le raccolte di lemmi per i sistemi di dettatura ecc. Inoltre non sono disponibili al pubblico i *corpora* dedicati alla valutazione dei sistemi, almeno sino ad valutazione avvenuta.

In questo paragrafo vogliamo brevemente descrivere alcuni dei più importanti *corpora* per l'italiano parlato e scritto. Per il parlato: *API*, *CALÌ*, *CIT*, *CLIPS*, *EUROM.0*, *EUROM.1*, *FOCUS*, *LABLITA*, *LIP*, *LIR*, *POLYCOST*, *SIVA*, *VIPS*; per lo scritto *CORIS*, *COLFIS*, *LIF*, *LIZ*, *TLIO*, *VELI*.

API è un progetto di raccolta di materiale fonico spontaneo composto da circa 14 ore di parlato (di cui circa 3,5 trascritte ortograficamente e un'ora e un quarto trascritta foneticamente) (http://www.cirass.unina.it/).

CALÌ (CALler Identification on Mobile and Fixed Network) raccoglie il parlato di 100 maschi e di 100 femmine. Ciascun parlante ha effettuato 20 chiamate, di cui 10 da GSM e 10 da telefono fisso, nell'arco di due o più mesi. In ciascuna chiamata sono stati pronunciati 18 "item" che comprendono tra l'altro sequenze di cifre, brevi frasi e "spelling" di nomi. La caratteristica principale del

database è la disponibilità di un uguale numero di "item" per le chiamate effettuate dalla rete telefonica mobile GSM e dalla rete fissa (PSTN).

COLFIS (*Corpus* e lessico di frequenza dell'italiano scritto) è costituito da 3.150.075 occorrenze lessicali tratte da quotidiani, periodici e libri di varia natura bilanciate secondo le letture degli italiani (http://www.istc.cnr.it/material/database/colfis/).

CIT (*Corpus* di italiano televisivo) è composto da 250.000 parole, ma è programmato un ampliamento del *corpus* a 500.000 parole per una maggiore omogeneità con altri *corpora* italiani di lingua scritta (*LIF*) e parlata (*LIP*). I testi prescelti sono tratti da trasmissioni originali non di fiction, tratte da diverse categorie. Il *CIT* è annotato secondo gli standard della Text Encoding Initiative (TEI) (http://www.sspina.it/cit/cit.htm).

CORIS (*Corpus* di riferimento dell'italiano scritto) è stato raccolto con lo scopo di costruire un *corpus* generale dell'italiano scritto. Il *corpus* contiene 100 milioni di parole e verrà aggiornato ogni due anni con nuovo materiale di controllo. I testi ivi contenuti sono prevalentemente narrativa prodotta negli anni Ottanta e Novanta (http://corpus.cilta.unibo.it:8080/coris_ita.html).

EUROM.0 contiene 15 frasi pronunciate da 8 diversi parlanti. Il segnale è stato registrato in camera anecoica alla frequenza di campionamento di 16 KHz, 16 bit.

EUROM.1, registrato con le stesse modalità di *EUROM.0*, contiene 10 frasi pronunciate da 30 diversi parlanti ("training") e 10 frasi pronunciate da latri 29 parlanti ("test").

FOCUS (Forensic *corpus*) contiene due diversi tipi di parlato: un testo letto uguale per tutti e un breve dialogo elicitato attraverso un'intervista. Quattro parlanti effettuano sei ripetizioni e otto parlanti due ripetizioni. Complessivamente, il segnale utile per ogni sessione registrazione è di circa 3 min. I parlanti sono scelti con la stessa estrazione dialettale e la stessa età nella prospettiva di avere voci relativamente "simili".

LABLITA dal 1973 si occupa della raccolta e gestione di *corpora* con lo standard di trascrizione chat. Si tratta di un insieme di *corpora* composto da: 1) un *corpus* di italiano parlato spontaneo adulto che raccoglie circa 120 testi che riguardano situazioni comunicative diafasiche diverse per un totale di 60 ore; 2) un *corpus* della lingua dei *media* (cinema, radio e televisione); 3) un *corpus* di 100 ore di italiano registrato nella fase del primo apprendimento (in bambini di 18-36 mesi). In questo *corpus* i testi sono trascritti, e l'audio è disponibile in formato digitalizzato (WAV) (http://lablita.dit.unifi.it/).

LIF (Lessico di frequenza della lingua italiana contemporanea) costituisce il primo grande progetto di costruzione di un lessico di frequenza per la lingua italiana. Esso contiene circa 5.000 lemmi ordinati per frequenza e secondo l'ordine alfabetico, tratti dallo spoglio di testi per un complesso di 500.000 parole.

LIP (Lessico di frequenza dell'italiano parlato): il *corpus* da cui è tratto è costituito da circa 500.000 parole grafiche, trascrizioni di registrazioni effettuate a Milano, Firenze, Roma e Napoli, pari a quasi 57 ore di parlato. I lemmi sono consultabili secondo frequenza e secondo ordine alfabetico, vi è anche una li-

sta di frequenza dei fonosimboli e delle polirematiche (http://languageserver.uni-graz.at/badip/badip/20_corpusLip.php).

LIR (Lessico di frequenza dell'italiano radiofonico) è un progetto di analisi del lessico e del *corpus* del parlato radiofonico nato nel 1998 e gestito da LABLITA. Il *corpus* di circa 60 ore è trascritto ortograficamente, allineato all'audio mediate software apposito, lemmatizzato e pubblicato su CD-ROM (http://lablita.dit.unifi.it/).

SIVA, che come detto è stato il primo *corpus* italiano per il riconoscimento del parlante, comprende 18 sessioni di registrazione attraverso le rete telefonica di 20 voci maschili. Ciascun parlante ha pronunciato 28 parole isolate, le risposte ad alcune domande e infine la lettura di un testo breve (1') foneticamente bilanciato.

TLIO (Tesoro della lingua italiana delle origini) è un *database* testuale composto da circa 1.780 testi per circa 20 milioni di parole, tratte da scritti in lingua italiana prima del 1375, in prosa e in poesia. Il *database* è interrogabile online (http://tlio.ovi.cnr.it/TLIO/).

VELI (Vocabolario elettronico della lingua italiana) è costituito da circa 10.000 lessemi per frequenza nella lingua italiana (tratto da un *corpus* di più di 20 milioni di parole).

6
I *corpora* di *CLIPS*

I *corpora* linguistici per l'italiano parlato e scritto (*CLIPS*), risultato di un progetto dell'Università "Federico II" di Napoli diretto da Federico Albano Leoni, saranno la più vasta raccolta di segnale vocale per la sezione sul parlato (raccolto tra il 2000 e il 2003). *CLIPS* consiste in circa 100 ore di parlato di diverse varietà, equamente ripartito tra voci maschili e voci femminili, in parte trascritto ortograficamente ed etichettato foneticamente. Le registrazioni sono state effettuate in 15 località italiane scelte in base a criteri di rappresentatività linguistica e socioeconomica: Bari, Bergamo, Bologna, Cagliari, Catanzaro, Firenze, Genova, Lecce, Milano, Napoli, Palermo, Parma, Perugia, Roma, Venezia.

Per ogni località è stato raccolto: *a*) parlato radiotelevisivo (notiziari, interviste, *talk shows*); *b*) parlato dialogico (240 dialoghi raccolti secondo le modalità del *map task* e del "gioco delle differenze", dei quali 30 etichettati foneticamente, 90 trascritti ortograficamente); *c*) parlato letto da parlanti non professionisti (20 frasi atte a garantire la copertura delle frequenze medio-alte del lessico italiano); *d*) parlato telefonico (conversazioni tra circa 300 parlanti e un portiere d'albergo simulato); *e*) parlato letto da 20 parlanti professionisti (160 frasi atte a garantire la copertura delle sequenze fonotattiche dell'italiano e delle frequenze medio-alte del lessico italiano registrato in camera anecoica).

La documentazione, i protocolli di raccolta e di annotazione del materiale tengono conto delle direttive del progetto "EAGLES" (cfr. rif.). Informazioni si possono consultare al sito: http://www.clips.unina.it/. Bisognerebbe ora illustrare quale contributo *CLIPS* può fornire al trattamento automatico della lin-

gua. Per ovvie ragioni, la ricerca sul TAL è storicamente evoluta per ciascuna nazione nella propria lingua. Queste tecnologie non possono essere semplicemente acquistate, come un computer o un'automobile, ma richiedono una attenta opera di progettazione per "funzionare" in una determinata lingua, e la progettazione e lo sviluppo richiedono basi di dati di adeguata consistenza. Ovviamente in molti casi si deve ricorrere a basi di dati *ad hoc* perché sia il lessico sia la grammatica dipendono dalla lingua; tuttavia non ci sentiamo di smentire Albano Leoni (Albano Leoni, 2006) quando sostiene: «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 a una determinata applicazione».

È certamente utile poter disporre di materiale vocale eterogeneo sul quale poter confrontare alcuni strumenti e semmai integrare con i vocaboli specialistici il materiale audio generico. *CLIPS* pertanto è uno strumento potente utile anche per lo sviluppo delle applicazioni tecnologiche dell'ambito del trattamento automatico della lingua.

7
Prospettive future

In tempi recenti, anche per la maggiore facilità con cui si hanno a disposizione i supporti di memorizzazione, vi è un proliferare di *corpora*, vocali o testuali, collazionati per diversi fini in diverse modalità. Diventa pertanto importante la ricerca del materiale desiderato all'interno di questi *corpora*, sempre più numerosi.

La disponibilità dei dati, essenziale nello sviluppo delle tecnologie dell'informatica e della comunicazione, è in qualche misura limitata dalla difficoltà di sapere dove reperirli. Ad esempio se un ricercatore volesse studiare l'eventuale dipendenza di alcuni valori formantici dall'età del parlante, potrebbe trovare uno o più *corpora* contenti le informazioni di suo interesse o almeno i segnali acustici con cui ricavarli? Un primo passo per trovare l'informazione su diversi "repository" è interpretarne i diversi dialetti e le diverse strutture e disporre dei necessari diritti di accesso. Una volta acquisite le fonti è necessario identificare i contenuti seguendo opportuni modelli o tassonomie. Si propone a tal fine la costituzione di un portale in grado di rendere più accessibili i dati allo studioso come al tecnologo e facilitare l'accesso ad essi attraverso opportune strategie di "Information Retrieval".

Riferimenti bibliografici

Albano Leoni F. (2006), *Il* corpus *CLIPS presentazione del progetto*, www.clips.unina.it.

Ariyaeeina A. M., Paoloni A. (1998), *Speaker Recognition in Telephony: Activities within the Framework of COST 250* (invited Paper), Conference on Identification Technology (ID TECH), June, Virginia (USA).

Contino U., Falcone M. (1995), *SIVA the MUSER: un database vocale per il riconoscimento del parlatore*, XXIII Convegno nazionale AIA (Bologna, 12-14 settembre).
Di Carlo A., Falcone M., Paoloni A. (1994), *Corpus Design for Speaker Recognition Assessment*, ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, April, Martigny.
Di Carlo A., Paoloni A. (2004), *Libro bianco sul trattamento automatico della lingua*, Eurografica, Roma.
Falcone M., Gallo A. (1996), *The SIVA Speech Database for Speaker Verification: Description and Evaluation*, ICSLP, ottobre, Philadelphia, pp. 1902-5.
Godfrey J. J., Holliman E. C., McDaniel J. (1995), *SWITCHBOARD: telephone speech corpus for research and development*, in Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, San Francisco, pp. 517-20.
Laj E., Di Carlo A., Paoloni A. (1992), *Realizzazione di una banca dati vocali italiani*, in "La comunicazione", XLI, pp. 115-20.
Lee Kai-Fu (1989), *Automatic Speech Recognition*, Kluwer, Boston *et al.*
Paoloni A. (2002), *La voce come elemento di identificazione della persona*, in *La voce come bene culturale*, a cura di A. De Dominicis, Carocci, Roma, pp. 125-39.
Proceeding of ESCA. Workshop on Speech Input /Output Assessment and Speech Databases – Noordwijkerhout, The Netherlands, sept. 1989.