PIETRO G. BELTRAMI*

Informatica e studi umanistici: qualche appunto linguistico e filologico

(Incontro di discussione *Informatica: cultura e società*, Roma, 24 gennaio 2006)

Informatica e studi umanistici vanno insieme fin dalle origini. Se si parla della linguistica, e in particolare degli studi sul lessico, basti pensare che fu nel 1949 che padre Roberto Busa diede inizio da vero pioniere, con l'IBM, ai lavori dell'*Index Thomisticus*, che oggi è in rete[1]. Suo allievo, a lui inviato dal linguista Carlo Tagliavini, è stato Antonio Zampolli, poi direttore a Pisa della Divisione Linguistica del CNUCE e fondatore dell'Istituto di Linguistica Computazionale del CNR, centro di riferimento internazionale per il trattamento automatico della lingua, molto attivo nei progetti europei. Di Zampolli, tragicamente scomparso nel 2003, ricordo in particolare l'antica collaborazione con l'Accademia della Crusca, quando questa decise, nel 1965, di usare l'informatica (si intende quella di allora) per i lavori del nuovo vocabolario storico italiano, e in particolare di codificare tutti i testi antichi: di qui un patrimonio di testi informatizzati dal quale, fallito il primo progetto, è ripartito il *Tesoro della Lingua Italiana delle Origini*, che l'Opera del Vocabolario Italiano porta ora avanti con strumenti informatici propri.

Oggi è sotto gli occhi di tutti che l'informatica ha trasformato la vita degli studiosi di discipline umanistiche, come del resto quella degli altri. Esiste anzi una percezione diffusa che negli studi umanistici si sia prodotto un salto culturale, o in altre parole che, com'è vero che gli strumenti non sono neutri, così questi studi non siano più gli stessi di prima. Spia ne è l'uso frequente del termine 'informatica umanistica', che per alcuni è un modo di pensare e di lavorare, per altri una vera e propria disciplina o scienza[2], ed è comunque nelle facoltà umanistiche, con variazioni, il nome di insegnamenti e di corsi di laurea, di cui uno triennale, a Pisa, dove se ne sta avviando anche uno specialistico. I laureati di quest'ultimo, sintetizzo dall'ordinamento[3], dovrebbero possedere una formazione di base negli studi afferenti alla Facoltà di Lettere e Filosofia, e insieme saper af-

---

* Pietro G. Beltrami (n. 1951) ha studiato alla Scuola Normale Superiore e alla Facoltà di Lettere di Pisa, dove è ora professore ordinario di filologia romanza. Dal 1992 è direttore dell'Istituto Opera del Vocabolario Italiano del CNR, a Firenze; dal 2003 accademico della Crusca (socio corrispondente dal 1997). Ha scritto di metrica italiana e romanza, su trovatori provenzali e galego-portoghesi, su Chrétien de Troyes e su Brunetto Latini; la sua opera più nota è *La metrica italiana* (Bologna, Il Mulino, 1991, 4ª ed. 2002), la più bella la traduzione di Chrétien de Troyes - Godefroy de Leigni, *Il cavaliere della carretta (Lancillotto)*, Alessandria, Edizioni dell'Orso, 2004. Dal 1992 dedica la maggior parte del proprio lavoro al *Tesoro della Lingua Italiana delle Origini*, elaborato e pubblicato in rete dall'OVI (www.vocabolario.org, http://tlio.ovi.cnr.it/TLIO).

L'Opera del Vocabolario Italiano (OVI) è l'Istituto del CNR che ha il compito di elaborare il vocabolario storico italiano. Ha sede presso l'Accademia della Crusca, dalla quale ha avuto origine e con la quale ha importanti rapporti di collaborazione. Redige attualmente il *Tesoro della Lingua Italiana delle Origini*, per il quale ha elaborato una grande banca dati dell'italiano antico, aperta agli studiosi online; produce in funzione di questi lavori, ma rende disponibili anche agli studiosi, i software GATTO, per la creazione, gestione, lemmatizzazione e interrogazione di banche dati testuali, e GATTOWEB, per l'interrogazione online di corpora creati in GATTO: per tutto ciò cfr. www.vocabolario.org.

[1] All'indirizzo http://www.corpusthomisticum.org/it/index.age.

[2] Cfr. in particolare Gino Roncaglia, *Informatica umanistica: le ragioni di una disciplina*, «Intersezioni», XXIII, 2002, 3, pp. 353-76 (anche in web all'indirizzo http://www.merzweb.com/testi/saggi/informatica_umanistica.htm). Per un panorama di strumenti e risorse cfr. anche *Informatica e scienze umane. Mezzo secolo di studi e ricerche*, a cura di Marco Veneziani, Firenze, Olshki, 2003; *Informatica umanistica: dalla ricerca all'insegnamento*, a cura di Domenico Fiormonte, Roma, Bulzoni, 2003; Maurizio Lana, *Il testo nel computer. Dal web all'analisi dei testi*, Torino, Bolati Boringhieri, 2004; Riccardo Castellana, *Risorse digitali dantesche: testi, commenti, metrica, filologia*, «Allegoria», XVI, 2004, n. 48, pp. 96-124.

[3] Scaricabile dal sito http://infouma.di.unipi.it.

frontare problemi di rappresentazione della conoscenza con strumenti informatici, essere esperti di teoria e tecnica del trattamento informatico di testi, lingue, immagini, suono e video, sapersi occupare di banche dati e di archivi digitali. Diversa è la figura scientifica degli informatici propriamente detti che contribuiscono agli studi umanistici, e diversa ancora la figura degli studiosi umanistici dotati di quel tanto di competenza che consente loro di lavorare con gli informatici. Entro questo quadro molto aperto, che riguarda un campo molto vasto e profondamente diversificato, accennerò solo a un paio di punti più accessibili alle mie competenze.

La prima applicazione dell'informatica alla linguistica è consistita, come ho già accennato, nell'indicizzazione di testi digitalizzati, con metodi che nel tempo si sono molto evoluti. Un uso noto a tutti è la ricerca in un testo o in un corpus dei passi che contengono una determinata parola, come si faceva una volta, solo per pochi testi fondamentali, con le concordanze redatte a mano; e la generazione di liste di tutti i passi che contengono una determinata parola o stringa. Questo è per esempio un uso di base di un software per la lessicografia, con l'ausilio del quale si costituisce un corpus di tutti i testi da spogliare per un vocabolario, e si estraggono i materiali che servono per la redazione di ogni voce. È così che si lavora al *Tesoro della Lingua Italiana delle Origini*, con strumenti in realtà molto più sofisticati, utilizzando un corpus di testi anteriori alla fine del Trecento e scritti in tutte le lingue dell'Italia medievale, implementato in una banca dati lemmatizzata con il software GATTO, sviluppato all'Opera del Vocabolario da Domenico Iorio-Fili. Senza un'efficiente attrezzatura informatica non si potrebbe infatti nemmeno pensare di lavorare di prima mano ad un vocabolario storico, spogliando e interpretando i testi invece delle fonti secondarie.

Un'innovazione importante portata alla linguistica fin dalle prime applicazioni informatiche è la possibilità concreta, prima scarsamente attingibile, di compiere elaborazioni statistiche. Il già citato Zampolli è stato coautore nel 1972 del primo *Lessico di frequenza della lingua italiana contemporanea*[4], basato su un corpus di lingua scritta; successivamente il metodo è stato applicato alla lingua parlata (cioè a trascrizioni della stessa) con il *Lessico di frequenza dell'italiano parlato* di Tullio De Mauro[5]. Gli effetti di questo tipo di ricerche (molto altro si dovrebbe citare) si vedono nei dizionari attuali. Il *Grande Dizionario Italiano dell'Uso* dello stesso De Mauro[6] distingue su basi statistiche e contrassegna esplicitamente il lessico 'fondamentale' (circa 2000 parole che da sole formano il 90% di tutto quanto si dice o si scrive), e il lessico di 'alto uso' (circa 2500 parole che ne formano un ulteriore 6%); nel restante 4% distingue (necessariamente non più su base statistica) fra 'alta disponibilità' (parole poco frequenti, ma note a tutti), lessico 'comune' (parole generalmente note a un certo livello di istruzione) e lessico tecnico-specialistico. Questa articolazione del lessico si sta calando nella conoscenza del pubblico più vasto, tramite i vari dizionari dell'uso che marcano almeno globalmente il lessico più frequente e più noto; ciò si può usare per esempio per l'insegnamento dell'italiano come lingua straniera, oppure, per chi si rivolge al grande pubblico, per stabilire quali parole dare per note e quali invece spiegare quando si usano.

Un elenco anche solo dei più importanti corpora elaborati in Italia e altrove sarebbe molto più lungo che utile[7], anche trascurando quel tipo particolare, ma molto diffuso,

---

[4] Umberta Bortolini, Carlo Tagliavini, Antonio Zampolli, *Lessico di frequenza della lingua italiana contemporanea*, Milano, Garzanti, 1972.

[5] Tullio De Mauro, Federico Mancini, Massimo Vedovelli, Miriam Voghera, *Lessico di frequenza del''italiano parlato*, Milano, ETAS Libri, 1993.

[6] Tullio De Mauro, *Grande Dizionario Italiano dell'Uso*, Torino, UTET, 2000.

[7] Ma si citeranno almeno, in più di quanto detto a testo, il CORIS/CORDIS (Corpus di riferimento di italiano scritto) del CILTA - Centro Interfacoltà di Linguistica Teorica e Applicata (http://corpus.cilta.u-

che sono le numerose biblioteche o raccolte di testi in rete. È tra parentesi un effetto non dell'informatica, ma forse dell'interesse per il mezzo che prevale su quello per i contenuti, il fatto che molto spesso i testi in rete mancano di indicazioni sulla provenienza editoriale, un po' come nel Medioevo si leggeva un testo sul primo o unico manoscritto che si trovava, prima che secoli di filologia ci insegnassero quanto è importante l'edizione in cui si legge. È diverso il caso, importante da rilevare per la tipologia, dei corpora creati per costituire risorse linguistiche (s'intende di lingua contemporanea). Per questo uso servono corpora di dimensioni sempre maggiori, oggi oltre i cento milioni di parole, entro i quali da un lato la ricerca di determinati fenomeni dia risultati statisticamente rilevanti, dall'altro possano emergere anche fenomeni percentualmente rari; conta perciò l'accumulazione di testi, e si possono usare tecniche automatiche di ricerca, scaricamento e indicizzazione nel corpus dalla rete; mentre contano molto meno, o si trascurano, sia il controllo delle fonti, sia la precisa distinzione e indicazione delle stesse in ogni risultato di ricerca. Si ottengono così risorse utilizzabili a più fini; per esempio per costituire dizionari di macchina in servizio della traduzione automatica, che è un settore trainante, punto forte, anche per quanto riguarda i finanziamenti[8], di un insieme di ricerche intorno all'analisi automatica della morfologia, della sintassi e del significato, che possono, per esempio, essere rivolte al fine di reperire o interpretare dati dalla rete cercandoli per significati, o ad altri fini di sicuro interesse economico, come far interagire uomini e macchine in lingua naturale, e di utilità sociale, come consentire ai non vedenti di utilizzare il computer ed esplorare la rete con l'uso della parola.

La lessicografia storica (dove 'storica' allude al metodo, che può valere per la lingua contemporanea come per la lingua antica) è invece più vicina alla filologia, e pur avendo interesse per corpora se possibile di grandi dimensioni, ha prima di tutto bisogno di poter valutare la provenienza e l'affidabilità dei dati in ogni fase; non interessa solo stabilire che una parola o un costrutto sono attestati in un corpus, ma a che data, in che autore, in che testo, e in quale tipo di edizione e di quale affidabilità. La qualità del corpus diventa perciò importante almeno quanto la dimensione, e l'elaborazione è più lunga e complessa. Le risorse che in questo modo si ottengono sono utili in un campo meno esposto ai finanziamenti, per l'elaborazione di dizionari, soprattutto storici, e per la ricerca storico-linguistica e filologica. Vale la pena di citare, in CD e non in rete, la *Letteratura Italiana Zanichelli* di Pasquale Stoppelli e Eugenio Picchi, la cui prima versione è uscita nel 1993, che ha molto inciso sul lavoro degli storici della lingua; come oggi è molto usata dagli storici della lingua antica e dai filologi la banca dati online dell'italiano antico dell'OVI[9]. Per quanto riguarda i dizionari, solo l'informatica consente sia di lavorare di prima mano, sia di puntare allo spoglio esaustivo della documentazione (non

---

nibo.it:8080/coris_ita.html), e sulla lingua parlata i lavori del LABLITA - Laboratorio Linguistico del Dipartimento di Italianistica dell'Università di Firenze (http://lablita.dit.unifi.it/).

[8] Vale la pena di citare la recente *Comunicazione sul multilinguismo* della Commissione Europea del 22 novembre 2005, pubblicata nel sito http://europa.eu.int/languages, dove si parla molto di traduzione e di traduzione automatica, e si afferma (§ II.3) che la Commissione intende «rafforzare nell'ambito del 7° programma quadro di ricerca le attività di ricerca e sviluppo tecnologico sulle tecnologie della società dell'informazione in campo linguistico mettendo in particolare rilievo le nuove tecnologie per la traduzione automatica nonché studiare i modi in cui l'Unione europea potrebbe stimolare un'ulteriore cooperazione nell'ambito delle nuove tecnologie di traduzione e interpretazione. Nello stesso contesto le questioni linguistiche costituiranno parte integrante delle attività sostenute nei settori delle scienze umane e sociali».

[9] L'ultima citazione nella bibliografia, molto elogiativa, è in Claudio Marazzini, *Memoria linguistica e rivitalizzazione digitale dell'antico*, in *Il futuro della memoria: la trasmissione del patrimonio culturale nell'eta digitale*, a cura di Agata Spaziante, Torino, CSI-Piemonte, 2005, pp. 119-31, p. 125 («esempio e modello di una meravigliosa sinergia tra antico e moderno»).

senza qualche problema prodotto dal demone dell'esaustività scatenato dallo strumento); e incide anche più profondamente sul metodo per la necessità di tradurre in forma esplicita e discreta, anche dove l'utente finale non se ne renderà conto, ogni dato di per sé correttamente approssimativo e discorsivo (l'esempio più banale è che i giudizi sulla datazione di testi databili genericamente a un certo periodo devono essere tradotti convenzionalmente in numeri).

A proposito della filologia, meritano un cenno almeno velocissimo le ricerche intorno a strumenti informatici per l'edizione critica, cioè per quel tipo di edizione che mira a ricostruire almeno idealmente l'originale di testi tramandati da copie non d'autore, manoscritti o stampe. In parte si tratta di strumenti per consentire di lavorare in modo integrato, al computer, su immagini dei manoscritti, trascrizioni, testo e apparati, come la 'stazione filologica' elaborata a Pisa da Andrea Bozzi[10]. Ma poiché per stabilire i rapporti fra i manoscritti, e valutare di conseguenza il valore delle lezioni che portano, ci si basa sui passi in cui le lezioni divergono e sulla concordanza negli errori, in varie sedi, soprattutto all'estero, è stata sviluppata la ricerca di algoritmi per confrontare il testo di più manoscritti ed elaborare i risultati del confronto[11]. Il problema non è che le differenze fra i manoscritti possono essere numerosissime, ma che solo alcune, in genere poche, sono rilevanti, e stabilire quali lo siano è materia di interpretazione filologica; e qui si vede un esempio della tensione che c'è sempre, nelle discipline umanistiche, fra la dimensione quantitativa, che non è tutta dalla parte del computer, e quella qualitativa, che non è tutta dalla parte dello studioso. Nei fatti, questi metodi automatici comportano anche un'idea dell'edizione, o addirittura della natura dell'originale, molto diversa da quella rappresentata dall'edizione ricostruttiva basata sul confronto tra i manoscritti, e si avvicinano piuttosto alla scuola di chi ritiene che al di là dello studio della tradizione l'editore debba riprodurre fedelmente il manoscritto giudicato migliore. Oppure si deve consentire all'editore di introdurre come dato nel calcolo anche la sua valutazione su ogni differenza fra i manoscritti; e anche in questo senso si sta lavorando, in particolare a Pisa (lo stesso Bozzi e Maria Sofia Corradini)[12]. Mi limito però a osservare conclusivamente, ed è per questo che ho voluto citare l'esempio della filologia informatica anche se così imperfettamente, che anche in questo caso, e vistosamente, si osserva quanto l'interazione tra informatica e studi umanistici non sia solo questione di usare strumenti comodi o utili o necessari, ma, almeno altrettanto, di ripensare i metodi, forse anche le finalità, del nostro lavoro.

[10] Cfr. Andrea Bozzi, *Aspetti e metodi di critica testuale assistita da calcolatore*, «Studi e Saggi Linguistici», XL-XLI, 2002-2003 [ma 2005] = *Atti del Convegno di Studi in memoria di Tristano Bolelli*, Pisa, 28-29 novembre 2003, a cura di Giovanna Marotta, Pisa, ETS, pp. 69-82.

[11] Cfr. Peter Robinson, *Making electronic editions and the fascination of what is difficult*, in *Digital Technology and Philological Disciplines*, a cura di Andrea Bozzi, Laura Cignoni, Jean-Louis Lebrave, Pisa-Roma, IEPI, 2004, pp. 415-38. Sulla 'critica del testo filologica' e in particolare sui lavori di Robinson cfr. la sezione dedicata alla filologia del saggio di R. Castellana cit.

[12] Cfr. Andrea Bozzi, Maria Sofia Corradini, *Aspects and methods of computer-aided textual criticism*, in *Digital Technology and Philological Disciplines* cit., pp. 49-66.