UNIVERSITÀ DEGLI STUDI DI BOLOGNA

FACOLTÀ DI LETTERE E FILOSOFIA

Corso di Laurea in Lettere Moderne
Curriculum Filologico - Letterario

DAL BIT AL TESTO: PROBLEMI DI MARKUP

Tesi di laurea in
Informatica per le scienze umane

Relatore:

Prof. TOMMASO DEL VECCHIO

Correlatore:

Prof.ssa FRANCESCA TOMASI

Presentata da:

GLORIA GAGLIARDI

Sessione III
Anno Accademico 2005-2006

“Homo sum: humani nihil a me alienum puto”

TERENZIO, *Heautontimorúmenos*

INDICE

INTRODUZIONE

Le applicazioni dell'informatica e della telematica in ambito umanistico sono, ad oggi, molteplici e ragguardevoli: numerosi sono i campi di applicazione dei nuovi strumenti digitali, che esercitano il loro influsso ed estendono i loro approcci innovativi dalla letteratura alla filologia, dalla linguistica alla storia, alla paleografia, alla biblioteconomia, all'archivistica, all'arte; numerose sono le pubblicazioni periodiche e le riviste specializzate[1] nonché i progetti avviati dagli Atenei e dai centri di ricerca;[2] innumerevoli i prodotti digitali già realizzati.[3]

Tuttavia il connubio tra informatica e scienze testuali è ancora guardato con sospetto e diffidenza, come uno stravagante e bizzarro ossimoro.

Oltre che alle problematiche metodologiche insite in questa interdisciplina, per sua stessa natura "liquida",[4] i dubbi e gli scetticismi sono in parte legati all'assenza di un paradigma teorico

1 Come ad esempio «Computers and the Humanities», prima rivista in ordine cronologico di pubblicazione; ma anche «Journal of the Association for History and Computing» e «Literary and Linguistic Computing».

2 Le principali associazioni del settore sono ACH (*Association for Computer in the Humanities*) e ALLC (*Association for Literary and Linguistic Computing*). In ambito nazionale i centri di ricerca Cisadu (<http://rmcisadu.let.uniroma1.it/>) e Crilet (<http://crilet.scu.uniroma1.it>) di Roma; ma anche Cribecu di Pisa (<www.cribecu.sns.it/>).

3 Il portale Internet Culturale promosso dal Ministero per i Beni e le Attività culturali (<http://www.internetculturale.it/>); biblioteche digitali come nel caso del progetto BibIt (<http://www.bibliotecaitaliana.it/>) e raccolte di varianti testuali di autori viventi nel progetto Digital Variants (<http://www.digitalvariants.org>); anche se entrambi un po' datati, il sito <http://www.univ.trieste.it/~nirital/lughi/homepage.html> e l'articolo di Tito Orlandi "Informatica umanistica: realizzazioni e prospettive" (1992, p. 1-22) forniscono una interessante panoramica del settore. È inoltre in uscita un "Annuario di Informatica Umanistica" con i contributi relativi alla situazione dell'I.U. in Italia, compresi i prodotti della ricerca.

4 Fiormonte, <http://www.griseldaonline.it/informatica/fiormonte_risposta.htm>.

unificante, che sia condiviso dall'intera comunità degli studiosi.

Il dibattito sullo statuto epistemologico dell'informatica umanistica, che in ambito italiano ha visto prendere posizione numerosi studiosi tra i quali Tito Orlandi, Gino Roncaglia, Fabio Ciotti, Domenico Fiormonte, Dino Buzzetti, lo stesso Padre Roberto Busa (pioniere in questi studi con il suo *Index Thomisticus*), è indubbiamente complesso e articolato.

Ma oltre alla definizione di uno statuto disciplinare e di un fondamento scientifico condiviso, alla delimitazione del campo di indagine e all'individuazione di specificità tematiche, gli studiosi hanno dovuto far fronte ad attacchi rivolti contro la stessa dignità accademica dell'Humanities Computing:[5] ne è un esempio l'articolo di Giulio Benedetti apparso il 6 giugno 2002 sul «Corriere della Sera»[6] in cui il giornalista, ironizzando sull'abbondante e fantasiosa proliferazione di proposte formative, riportava una dichiarazione del ministro Letizia Moratti nella quale l'informatica umanistica veniva avvicinata alle "scienze del fiore e del verde"; oppure è possibile far riferimento ad un più recente articolo in cui Pietro Citati,[7] polemizzando sul caos delle università italiane, proponeva un sarcastico accostamento tra i computer adottati per l'analisi letteraria e i corsi di gelato artigianale, cappellini per signora, sandali per i tropici

5 La disciplina è chiamata in area anglofona anche con il nome di "Computers in the Humanities": si faccia però attenzione al fatto che, sebbene i due termini vengano usati per lo più intercambiabilmente, e vengano assunti come sinonimi, la coincidenza tra le due denominazioni non è totale.

6 Benedetti, "Troppe lauree brevi, 400 saranno cancellate", «Corriere della Sera», 6 giugno 2002.

7 Citati, "Finanziamenti, crediti, laurea breve: perché i nostri Atenei sono al collasso", «La Repubblica», 23 maggio 2006.

e retto uso dei pannolini.[8]

Sebbene il controverso statuto epistemologico della disciplina si trovi per il momento nell'epicentro di questo dibattito, è però innegabile l'esistenza di un legame tra informatica, in quanto scienza che ragiona sulla manipolazione di simboli, e attività umanistiche.

L'informatica non è infatti una disciplina di interesse unicamente tecnico e ingegneristico: attraverso il paradigma conoscitivo dell'algoritmo propone "uno sguardo sul reale, un approccio alla comprensione della realtà".[9]

Le procedure computazionali sono "basate su formalizzazioni rigorose, elaborate a partire dalla costruzione di modelli simbolici dell'oggetto di studio".[10] E ogni modellizzazione implica necessariamente un'interpretazione del reale.

Oltre che impossibile, è anacronistico restare indifferenti di fronte alle modificazioni indotte dalla multimedialità nella creazione e nella fruizione dei prodotti culturali e alle potenzialità della testualità digitale: dal momento che gli strumenti materiali di produzione della cultura non sono neutrali (con un ben noto aforisma, "il medium è il messaggio"),[11] le tecnologie influenzano profondamente le forme di scrittura e i processi cognitivi, nonostante se ne abbia una coscienza parziale e saltuaria.

L'elettricità in sé, "intesa essa stessa come medium, come modalità fondamentale di comunicazione, su cui si innesta la foltissima

8 Il testo di replica "In risposta a Pietro Citati. Sull'informatica umanistica" è consultabile all'URL <http://www.griseldaonline.it/informatica/5citati_risp.htm>.

9 Ferrarini, <http://www.griseldaonline.it/informatica/5ferrarini.htm>.

10 Roncaglia, <http://www.griseldaonline.it/informatica/roncaglia_secondo.htm>.

11 McLuhan, 1964, p. 15.

vegetazione dei contenuti più espliciti che da essa germogliano"[12] è fattore imprescindibile di condizionamento.

Anche se concepite in una società pre-multimediale, la modalità comunicativa odierna invera le affermazioni di Marshall McLuhan: realmente "dall'avvento dell'elettricità in poi, la rivoluzione dell'informazione è diventata una rivoluzione permanente...Si tratta prima di tutto della trasformazione globale dell'*hardware* in *software*, cioè dell'oggetto in informazione".[13]

"Una rivoluzione che riguarda innanzitutto – ma non solo – il modo di produrre, elaborare, raccogliere, scambiare informazione. Una rivoluzione che porta con sé conseguenze culturali, sociali, politiche, economiche di immenso rilievo".[14]

Ma il computer non può – e non deve – essere solo uno strumento: condividere l'impostazione teorica dell'I.U.[15] comporta innanzitutto riflettere sui modelli, i metodi, i formalismi e i paradigmi concettuali con cui i dati testuali possono essere gestiti e trattati in modo automatico.

Nonostante il computer negli ultimi anni si sia imposto come macchina 'intelligente' per eccellenza,[16] in grado di elaborare quantità enormi di dati strutturati, le sue capacità comunicative sono rimaste infatti molto rudimentali e il suo accesso ai testi è tutt'altro che

12 Gamaleri, 1976, p. 25.
13 Gamaleri, 1976, p. 24.
14 Ciotti - Roncaglia, 2005, p. V.
15 Acronimo di Informatica Umanistica.
16 Da un punto di vista strettamente scientifico il computer è solo apparentemente intelligente: questa fallace impressione è solo legata alla enorme mole di operazioni complesse che riesce ad eseguire in poche frazioni di secondo. Viceversa, in situazioni non previste dai *software* in esecuzione, è incapace di trovare vie d'uscita in modo intuitivo.

immediato.

Il testo è una struttura molto articolata, in cui le informazioni sono organizzate su vari livelli di complessità: per l'elaboratore anche l'identificazione di unità logiche semplici come il capitolo, il titolo, il capoverso è estremamente complessa.

Agli occhi dell'elaboratore un testo si presenta come un insieme di righe, ciascuna formata da una sequenza di caratteri che terminano con un ritorno a capo, e la materia del testo non è nient'altro che un flusso ininterrotto di stringhe di 0 e 1: l'accesso ai molteplici piani della struttura testuale, la gestione delle varianti ortografiche delle parole, la corretta interpretazione della punteggiatura, perfino l'individuazione dei confini di una parola (operazioni che a prima vista possono sembrare tra le più banali) richiedono un enorme bagaglio di conoscenze astratte e competenze specifiche.

Non esiste infatti nessuna corrispondenza diretta e necessaria tra una sequenza di caratteri ed una parola: solo volendo implementare un *software* che segmenti un testo per individuarvi le unità minime linguisticamente plausibili, definite in linguistica computazionale *token*, andremmo incontro ad innumerevoli eccezioni e una smisurata quantità di varianti.[17]

Come mostrano questi brevi esempi quando ci accostiamo all'informatica umanistica siamo costretti a prendere in considerazione aspetti e condizioni che solitamente accogliamo *a priori*, per lo più

17 Intuitivamente la segmentazione potrebbe essere effettuata individuando tutte le stringhe di caratteri che corrispondono agli spazi: si tratterebbe però di una soluzione riduttiva che, solo per fare un esempio, ignorerebbe tutte le parole divise graficamente dall'apostrofo o da altri segni di punteggiatura (la sequenza *un'amica* pur non comprendendo uno spazio, corrisponde a due *token* diversi).

ricomprendo il ruolo di consumatori passivi.

Ma, come sottolineano Ciotti e Roncaglia,[18] consumare senza capire è un lusso che per moltissimi motivi non possiamo permetterci.

Lo scopo di questo lavoro è pertanto tentare di offrire una panoramica delle tecniche di gestione digitale dell'informazione testuale, proponendo un percorso che, partendo dal testo inteso come semplice sequenza di caratteri (o, più propriamente, come stringhe di cifre binarie), giunge ad analizzare gli strumenti di una più esaustiva veicolazione dei significati morfosintattici e semantici: i linguaggi di *markup* dichiarativi.

18 Ciotti - Roncaglia, 2005, p. V.