

Indice

Indice generale

Indice.....	1
Lezione 15 Marzo.....	1
1) Definizione di Informatica Umanistica.....	1
2) Storia dell'Informatica Umanistica.....	2
3) Modellizzazione.....	2
4) Formalizzazione.....	2
5) Applicazioni in ambito archeologico e storico-artistico:.....	3
7) La TEI:.....	4
8) Rapporti XML, SGML ed XML-based:.....	4
9) Un esempio di codifica di un testo in XML/TEI:.....	4
Lezione 22 Marzo.....	5
1) Le applicazioni in ambito linguistico e letterario:.....	5
2) Il Trattamento Automatico del Linguaggio (TAL):.....	9
3) Problematiche ed obiettivi	9
Lezione 5 Aprile.....	10
1) IL PARSING:.....	10
2) Word Net:.....	11
3) L'Information Retrieval:.....	11
4) L'Information Extraction:	12
Lezione 12 Aprile.....	13
1) Applicazioni in ambito filologico:.....	13

Lezione 15 Marzo

1) Definizione di Informatica Umanistica

L'Informatica Umanistica, chiamata a livello internazionale *Digital Humanities* (una volta *Humanities Computing*), possiamo considerarla come l'interazione tra gli studi umanistici (storia dell'arte, studio delle letterature classiche, storia, ecc...) e gli strumenti informatici. Le sue finalità cambiano in base alle diverse impostazioni o scuole di pensiero. Tra queste quella di Edoardo Ferrarini, il quale sottolinea che l'applicazione degli strumenti informatici agli studi umanistici non deve essere volta soltanto alla velocizzazione degli studi stessi ma ad acquisire nuovi punti di vista, nuovi approcci e nuove prospettive nei confronti degli studi umanistici stessi. Così, ad esempio, il fatto che la digitalizzazione di un testo permetta di cercare in poco tempo una determinata parola in un autore classico latino (al contrario del grande sforzo a cui si sarebbe sottoposti se si cercasse quella parola sfogliando manualmente le singole pagine di un singolo libro) non è il solo obiettivo della disciplina. L'Informatica Umanistica permette, come dicevo sopra, di approcciarsi con nuovi orizzonti ad esempio anche ad un testo stesso: se prendiamo in considerazione un testo come l'Iliade, di esso ci sono pervenuti numerosi manoscritti. Ora, mentre lo studio filologico su carta stampata tende (inevitabilmente) ad aspirare alla ricostruzione del testo "originale" attraverso l'eliminazione degli errori, invece un approccio digitale ci permette di guardare contemporaneamente (magari in una tabella) tutte le varianti di in un determinato punto di quell'opera facendoci, così, valorizzare i singoli manoscritti con le loro singole varianti perché ognuno di essi ha un proprio valore. Insomma la stampa, proprio perché fissa i testi e li rende immutabili, ha contribuito a canonizzare i testi, mentre un approccio digitale ci permette di guardare "contemporaneamente" tutte le testimonianze di quel testo stesso.

2) Storia dell'Informatica Umanistica

L'Informatica Umanistica nasce negli anni successivi alla Seconda Guerra Mondiale, quando era nata l'Informatica stessa nell'ambito della crittoanalisi, che si impegnava nella decodificazione di messaggi criptati dei Nazisti. Pioniere dell'Informatica Umanistica è stato il padre gesuita Roberto Busa che decise di creare il lessico

dell'opera omnia di Tommaso d'Aquino su supporto digitale: questo permetteva di scrivere in un elenco tutte le parole dei testi dell'Aquinate indicando quante volte esse venivano utilizzate. Questo processo si chiama indicizzazione.

3) Modellizzazione

La digitalizzazione di un testo è una modellizzazione, in quanto che essa consiste nel creare un modello di una realtà (in questo caso, un testo) mantenendone la struttura di partenza con i suoi elementi interni nella gerarchia originale (in questo caso, organizzando quel testo in base ai capitoli e paragrafi, per esempio, che aveva sul supporto cartaceo).

4) Formalizzazione

La formalizzazione è, possiamo dire, uno dei modi in cui può avvenire una modellizzazione. La digitalizzazione di un testo, infatti, prevede che si trasformi il testo di partenza, che è su supporto cartaceo e scritto in una lingua umana, in un linguaggio comprensibile per il computer, come, ad esempio, XML. Insomma, si cambia la forma del testo stesso: si traduce quel testo dal linguaggio umano scritto al linguaggio del computer, che è un linguaggio formale.

5) Applicazioni in ambito archeologico e storico-artistico:

L'applicazione dell'informatica agli studi del settore archeologico e storico-artistico permette ad esempio:

- a)** l'uso di tecniche innovative nell'analisi dei reperti (raggi X, raggi laser, onde sonore, telerilevamento aereo e satellitare);
- b)** l'uso del clustering, ossia di un raggruppamento di dati omogenei sulla base di algoritmi prestabiliti;
- c)** la catalogazione dei beni storico-artistici (ad esempio, quella del SIGEC: Sistema Informatizzato Generale del Catalogo, realizzato dall'ICCD: Istituto Centrale per il Catalogo e la Documentazione);
- d)** l'interazione tra banche dati e GIS (Geographical Information System);
- e)** i virtual tours, ossia dei viaggi virtuali nei musei o nelle ricostruzioni in 3D di siti archeologici, ad esempio, come quelli realizzati dall'HOC: Hypermedia Open Center).

6) Il GIS:

Il GIS (Geographical Information System) è nato negli anni '70 e permette di creare delle mappe geografiche digitalizzate su cui poter effettuare l'associazione di informazioni di vario genere sui vari punti di esse.

Un GIS, quindi, si compone di una parte grafica (con immagini raster, formate da input di colore, ed immagini vettoriali come, ad esempio, le curve di livello che permettono, per esempio, la resa grafica di un palazzo visto dall'alto e di 3/4) e di un database, ovverossia di una banca dati contenente le varie informazioni che vengono, appunto, associate ai vari punti della mappa digitale. Ora, le fonti per la parte grafica e per il database possono essere, ad esempio, la scannerizzazione di fotografie, il telerilevamento aereo, gli input da tastiera e la fotogrammetria digitale.

In italiano il termine GIS viene tradotto con il corrispondente termine SIT (Sistema informativo Territoriale). Tuttavia, secondo alcuni studiosi di questo campo, ci sono delle differenze tra GIS e SIT: infatti, mentre il GIS ha a che fare con dati fisici e geografici, invece, il SIT ha a che fare con dati legati a chi abita nei territori analizzati.

Esempi di applicazioni di un GIS sono Google Maps, Pleiades e Pelagios.

7) La TEI:

La TEI (Text Encoding Initiative) è un progetto internazionale volto alla realizzazione di un modello di codifica standardizzato per i testi elettronici. Il termine "encode" in inglese significa proprio "codificare". Quando scriviamo un testo nel linguaggio di codifica TEI, esso si compone di due parti: il <teiHeader>, che contiene tutte le informazioni sul testo, ed il testo vero e proprio. Il modello di codifica TEI si adatta alla sintassi XML (eXtensible Markup Language). Il TEI è basato su una DTD, ossia su un vocabolario di tag XML.

8) Rapporti XML, SGML ed XML-based:

Da SGML (Standard Generalized Markup Language), markup language dichiarativo e metalinguaggio, si sono originati:

- XML, che è un markup language dichiarativo perché, basato su un markup generico, che "dichiara" la collocazione di ogni elemento nella struttura logica del testo. Esso è un metalinguaggio nel senso che può creare tanti linguaggi quanti vocabolari (DTD) si inventano.

Esso è per definizione “extensible” nel senso che permette di creare dei tag personalizzati.

- HTML, che è un markup language procedurale perché, basato su un markup specifico, non rappresenta e non elabora la struttura logica del testo ma si limita a definire le informazioni tipografiche e compositive della pagina.

Invece, non sono XML-based (ovvero markup language non fondati su XML) per esempio: *LaTeX*, *RTF* (Rich Text Format) e *Word*, anche se, in realtà, quest’ultimo, nelle sue recenti versioni, è diventato un linguaggio XML-based (infatti, è chiamato *WORD DOCX*, dove la X finale sta, per l’appunto, per “extensible”).

9) Un esempio di codifica di un testo in XML/TEI:

La versione P5 della TEI, uscita nel 2007, permette di interagire meglio con altri linguaggi che usano la sintassi XML. Vediamo il DTD di tag:

```
<div type="poem" n="42">
<head>Tanto gentile e tanto onesta pare</head>
<lg>
  <l n="1">Tanto gentile e tanto onesta pare</l>
  <l n="2">la donna mia quand'ella altrui saluta,</l>
  <l>ch'ogne lingua deven tremando muta, </l>
  <l>e gli occhi no l'ardiscon di guardare.</l>
</lg>
<lg>
  <l>Ella si va, sentendosi laudare, </l>
  <l>benignamente d'umiltà vestuta; </l>
  <l>e par che sia una cosa venuta </l>
  <l>da cielo in terra a miracol mostrare. </l></lg>
</div>
```

N.B:

La struttura generale di un documento XML/TEI è la seguente:

- TEI.2
 - `teiHeader`
 - `text`
 - `front`
 - `body`
 - `back`

Lezione 22 Marzo

1) Le applicazioni in ambito linguistico e letterario:

L'Informatica Umanistica si suddivide in diversi ambiti:

- ambito storico-artistico;
- ambito storico-archeologico;
- ambito linguistico (NLP);
- ambito letterario (HTP);
- ed altri ancora;

Ora, per quanto concerne l'ambito linguistico e quello letterario, possiamo stabilire (come sopra) una differenza tra i due. Infatti, al primo corrisponde l'*NLP, Natural Language Processing*, che si occupa del modo in cui il computer può elaborare automaticamente la struttura della lingua umana. Invece, al secondo corrisponde l'*HTP, Humanistic Text Processing*, che è un processo volto alla rielaborazione automatica di un testo ed una delle sue applicazioni è la ricerca di occorrenze. Entrambi i rami fanno parte della Linguistica Computazionale, che naturalmente è parte della stessa Informatica Umanistica, ed entrambi si servono (tra gli altri) di metodi statistico-matematici.

Ora, l'*Humanistic Text Processing* si compone di più fasi:

- il filologo codifica il testo in linguaggio XML (od in altro);
- il computer analizza il testo, ad esempio tramite lo string-matching o tramite altri metodi, seguendo le istruzioni del filologo.

L'*Humanistic Text Processing* consiste in alcuni procedimenti formali. Alla base vi sono:

- lo string-matching;
- la tokenizzazione.

N.B: per fare uno string-matching si presume che ci sia di già stata una tokenizzazione!!!

Seguono quindi altri procedimenti formali, ad esempio:

- a)** la creazione di un indice;
- b)** la creazione di concordanze (KWOC e KWIC);

- c)** la creazione di una lista di frequenza;
- d)** l'individuazione di collocations ("sparare un colpo");
- e)** l'individuazione di clusters ("conferenza-stampa");
- f)** la determinazione di un rango;
- g)** la determinazione di una lemmatizzazione ("pensando", "pensiamo", "pensai" riconducibili a "pensare");
- h)** lo stemming ("pensando", "pensiamo", "pensai" riconducibili a "pens-").

Ora vediamo singolarmente in che cosa consiste lo string-matching e le varie applicazioni dell'HTP, per l'appunto.

- **STRING-MATCHING:** attraverso lo string-matching ("ricerca di stringa") possiamo chiedere al computer di individuare in un testo digitalizzato una sequenza di caratteri conforme alla stringa digitata. A differenza nostra, che possiamo cogliere le parole nella loro complessità, il computer agirà per via algoritmica, cioè, attraverso piccoli passaggi, confronterà il carattere iniziale di ogni parola nel testo col primo carattere della stringa digitata e ripeterà tale procedimento fino a quando verranno riconosciuti tutti i caratteri della stringa digitata. Lo string-matching presenta però dei limiti, in quanto non è in grado di fare l'analisi morfologica e quindi di lemmatizzare (se digito *hominis* cercherà solo il genitivo) e non è in grado di fare l'analisi semantica, cioè non consente di compiere la disambiguazione tra gli omografi ("tasso" può indicare un albero, un animale ecc.).

- **TOKENIZZAZIONE:** la tokenizzazione (o tokenization) è un procedimento che consiste nella separazione o distinzione, all'interno di un testo digitalizzato, delle stringhe corrispondenti alle singole parole. Ciascuna parola corrisponde ad una stringa che inizia e che finisce con dei whitespaces, che possono corrispondere ad elementi di interpunzione o spazi, per esempio. Tuttavia, non dobbiamo dimenticarci che esistono anche dei casi complessi come "dammi" o "senatusque" che non facilmente possono essere tokenizzati.

- **INDICE:** si ottiene mettendo un dizionario-macchina (ossia un elenco delle parole del testo, in cui ogni type compare una sola volta) in fila ponendo accanto ad ogni parola il numero di volte in cui essa compare (ricorre) nel testo. In questo senso, l'opera pionieristica del padre gesuita Roberto Busa, relativa all'opera omnia di Tommaso d'Aquino (ossia l'*Index Thomisticus*) è un indice. N.B: qui naturalmente le definizioni possono essere soggettive.

- **CONCORDANZE:** una concordanza si ottiene mettendo in fila un dizionario-elenco e ponendo accanto ad ogni type tutti i suoi token. Se accanto ad ogni type pongo tutti i suoi token senza il contesto (ossia senza inserire le parole che immediatamente precedono il type e quelle che lo seguono immediatamente), magari riportando l'indicazione del libro, del capitolo, del paragrafo e del verso, allora ho una concordanza KWOC (Key Word Out of Context). Se, invece, accanto ad ogni type del dizionario-elenco pongo anche il contesto, allora ho una concordanza KWIC (Key Word In Context).

- **LISTE DI FREQUENZA:** si ottiene aggiungendo per ogni type di una concordanza la frequenza del type stesso. Tale frequenza è calcolabile dividendo il numero di volte in cui compare il type nel testo per il numero complessivo delle parole dell'intero testo. Quindi, ad esempio, se il testo è costituito da 100 parole, ed il type preso in considerazione compare (ricorre) 10 volte, ecco che allora divido 10 per 100, ottenendo 0,1 che è appunto il 10%. Quindi la frequenza di quel determinato type nel testo digitalizzato preso in considerazione è del 10%, per l'appunto, o, se vogliamo esprimerci in termini di decimali, dello 0,1 per un totale di 1.

- **COLLOCATIONS:** le collocations (o collocazioni) sono delle combinazioni standardizzate di parole, ad esempio "sparare un colpo". In questo caso, notiamo che il verbo "sparare" ammette come complemento oggetto quasi sempre la parola "colpo". Quindi, se cerchiamo "sparare", un sistema basato sull'HTP, sulla base di dati statistici, ci suggerirà l'intera stringa corrispondente all'espressione "sparare un colpo".

- **CLUSTERS:** sono dei gruppi di parole che si ripetono in maniera identica all'interno di un determinato testo e possono essere costituiti dalle due alle otto parole. Un esempio è "conferenza-stampa". Quindi, se io cerco, attraverso lo string-matching, una stringa corrispondente alla parola "conferenza", ecco che un sistema basato sull'HTP mi suggerisce automaticamente la stringa corrispondente all'intera espressione "conferenza-stampa", perché quelle due parole ricorrono più volte, all'interno del testo digitalizzato considerato, come un'unica espressione, per l'appunto.

- **RANGO:** è la posizione della parola all'interno della "top ten" (ma con ben più di dieci parole) delle parole più frequenti all'interno di

un determinato testo digitalizzato. Ad esempio, in Dante una parola di rango 1 è probabilmente la congiunzione “e” che ricorre molto spesso, per l’appunto.

- LEMMATIZZAZIONE: la lemmatizzazione è il procedimento attraverso cui ogni token di un determinato testo digitalizzato viene ricondotto ad un lemma. Con questo procedimento, quindi, se cerco una stringa corrispondente alla parola “penso”, oppure “pensiamo” od ancora “pensando”, tutte queste stringhe vengono ricondotte al lemma “pensare”. Ciò accade perché, una volta digitalizzato il testo, si è anche creato un elenco con tutte le parole del testo in cui si indica, per l’appunto, il lemma corrispondente.

- STEMMING: lo stemming è il procedimento tramite cui da una stringa (token) si ottiene la radice della parola (stem). Ad esempio, “pensai”, “penso” e “pensiero” derivano tutti dalla stessa radice “pens-”.

N.B:

Lemmatizzazione in TEI

```
<w lemma="rex">Rex</w>
```

```
<w lemma="rex">regum</w>
```

2) Il Trattamento Automatico del Linguaggio (TAL):

Consiste, ad esempio, nel controllo ortografico e grammaticale del linguaggio.

3) Problematiche ed obiettivi

Possiamo notare che l’HTP/NLP si servono di calcoli matematico-statistici per aiutare l’utente nelle ricerche di parole all’interno di un testo digitalizzato e che certamente le loro varie applicazioni costituiscono un supporto significativo. Tuttavia, così come lo string-matching presenta il limite di non poter effettuare analisi morfologiche né, tantomeno, analisi semantiche, d’altro canto anche le altre applicazioni dell’HTP/NLP hanno, a loro volta, dei limiti. Pensiamo, in tal senso, alla differenza tra la sinonimia perfetta e la sinonimia parziale (l’HTP/NLP hanno la possibilità di riconoscere facilmente una sinonimia tra parole come “università” ed “ateneo”, la sinonimia perfetta, ma hanno difficoltà a riconoscere una sinonimia del genere “laurea” e “titolo di studio”, la

sinonimia parziale, che in questo caso è anche una iponimia: la laurea è infatti uno dei tanti titoli di studio).

In questo campo semantico ecco che allora interviene il WordNet, per l'appunto, il Web semantico.

Inoltre, non tutti i corpora digitalizzati sono lemmatizzati. A tal proposito, effettueremo un confronto tra Intratext, Perseus e TLG.

- INTRATEXT: non è un corpus lemmatizzato, perché se clicco sulla stringa corrispondente alla parola “fosse” non la mette in relazione al verbo “essere”, ma si limita ad indicarmi tutti i passi in cui “fosse” ricorre nel testo da me selezionato, creando le concordanze KWIC, ossia dandomi anche il contesto in cui quelle parole si trovano, ovvero le parole che le precedono e quelle che le seguono.

- PERSEUS: è un corpus lemmatizzato, perché se voglio cercare “armis”, posso selezionare sia la ricerca della parola esatta “armis”, per l'appunto, sia la ricerca delle parole che derivano dallo stesso lemma, ovverossia tutte le forme inflesse.

- TLG (CD-Rom): non è un corpus lemmatizzato ma ha un indice di parole che permette di raffinare le ricerche a partire da una stringa. Se cerco, ad esempio, “efialt” mi darà 3 opzioni (efialten, efialtes, efialtou) e spetta a me, a questo punto, selezionare quale delle tre mi interessa. Così, una volta selezionata la parola dal *word index*, ecco che il TLG mi indicherà 5 occorrenze, ad esempio, per la forma efialtou.

- TLG (online): è un corpus con una lemmatizzazione secondaria, ovverossia fatta *a posteriori* da dei programmi (software).

Lezione 5 Aprile

1) IL PARSING:

In inglese il verbo “to parse” significa “analizzare”. Il parsing può essere anche un’analisi nell’ambito linguistico-grammaticale. Il termine parsing, in informatica, (e quindi il parsing digitale), indica proprio l’analisi di testi digitalizzati. Ora, a livello basico, il parsing si limita ad individuare le parti del discorso. A livelli avanzati, invece, il parsing analizza la morfologia.

Inoltre, il parsing può funzionare in modo automatico, laddove il computer lavora senza essere coadiuvato dall’uomo. Viceversa, esso può funzionare in modo semi-automatico nel caso in cui sia coadiuvato dall’uomo.

Questi diversi modi si possono applicare sia al livello basico che a livelli avanzati. Uno dei modi è il CROWDSOURCING, consistente nella distribuzione della parte del lavoro umano tra moltissimi utenti.

Facciamo subito degli utili esempi:

- Perseus, che effettua il parsing e che è crowdsourcing;
- Ancient Lives Oxford, che trascrive papiri, ma che non fa il parsing e che è crowdsourcing;
- Transcribe Betham, che, come il precedente, trascrive papiri ma non fa il parsing ed è crowdsourcing;
- Wikipedia, che realizza una enciclopedia, non fa parsing ed è crowdsourcing.

In particolare, notiamo che Persues effettua il parsing su testi greci e latini e presenta anche il crowdsourcing.

Uno dei metodi che il computer ha di fare il proprio lavoro (automatico o semi-automatico che sia) è l’uso di calcoli statistici, metodo facilmente applicabile ad una lingua dalla sintassi rigida come l’inglese, che gli permette di indicare in quale posizione tende a comparire una determinata parola. Infatti, in inglese, ad esempio, è facile stabilire che “he” sia un pronome in base alla posizione (che è sempre la stessa) che ha rispetto al verbo (ovverossia lo precede). La tendenza di “he” è quindi quella di ricorrere come pronome.

2 Word Net:

Il Word Net (uno per ciascuna lingua) è letteralmente una “rete di parole”. Si tratta di un progetto iniziato negli anni Ottanta da un gruppo di linguisti e di psicologi che volevano creare un archivio lessicale che contenesse, per le parole della lingua inglese, relazioni di tipo lessicale. Il progetto si limitò inizialmente ad organizzare le parole in base alla sinonimia/antinomia (“bello”/“carino” e “bello”/“brutto”) raggruppandole in uno stesso Synonym Set, detto anche Synset. Il progetto, in seguito, ha organizzato le parole sulla base anche di altre relazioni, come l’iperonimia/iponimia (“animale” è un iperonimo di “cane” e “gatto”, viceversa “cane” e “gatto” sono iponimi di “animale”).

In definitiva, il Word Net permette di organizzare una lingua a livello semantico. Così, per esempio, una volta creato il testo digitalizzato ed il relativo database con i set di sinonimi, ecco che sarà possibile per l’utente trovare automaticamente, durante la ricerca di una parola, i sinonimi di essa.

Ora, il progetto del Word Net si è spinto oltre, nel corso degli anni, ed è stato affiancato da altre Semantic Technologies, cioè da altre tecnologie che cercano di codificare/formalizzare la semantica. In questo senso, alcuni linguaggi semantici che fanno parte, per l’appunto, delle Semantic Technologies, sono RDF, OWL e Linked Data, che sono utilizzati per descrivere relazioni tra concetti, cioè a dire, ontologie.

Diverse applicazioni di Word Net, invece, sono, ad esempio, Word Net Search 3.1 ed ItalWordNet.

3) L’Information Retrieval:

L’Information Retrieval è un recupero di informazioni di già formalizzate secondo operazioni formali. Un esempio concreto è la ricerca degli orari dei treni su un sito come *Trenitalia*, laddove le informazioni sono state formalizzate su una tabella di orari, per esempio, che l’utente visiona, recuperando le informazioni prima formalizzate, per l’appunto.

4) L’Information Extraction:

L’Information Extraction è un procedimento che consiste proprio nella estrapolazione di informazioni. Ciò coincide, di fatto, anche con lo stesso Data Mining, ovvero con quel procedimento che ultimamente effettua il famoso motore di ricerca Google quando,

nel creare una scheda o profilo di un autore (come Alessandro Manzoni, per esempio), ecco che “estrae” varie informazioni prese da diversi siti (probabilmente in maniera automatica).

N.B:

Per intenderci meglio, e comprendere a fondo la differenza tra *l'Information Retrieval* e *l'Information Extraction*, possiamo riportare la “parabola” del maestro e dei due allievi, presente nella micro-dispensa. Immaginiamo che il maestro sia l'utente che interroga due allievi: uno effettua *l'Information Retrieval*, perché ripete “a pappagallo” le informazioni date dal maestro durante la lezione, l'altro effettua *l'Information Extraction*, perché, essendo più intelligente e più capace di interagire con la realtà, rielabora le informazioni ricevute, esprimendosi in maniera autonoma ma riportando sostanzialmente le stesse informazioni dell'allievo-“pappagallo”. Così, il maestro, nel parlare di un celebre passo dei Promessi Sposi dell' A. Manzoni, dice ai due allievi: “L'Innominato, dopo il colloquio con Lucia, domanda a se stesso: *invecchiare, morire e poi?*”. Il maestro poi domanda ai due allievi che cosa fa l'Innominato dopo il colloquio con Lucia. L'allievo-“pappagallo” risponde semplicemente: “L'Innominato domanda a se stesso: *invecchiare, morire e poi?*”. L'allievo più intelligente, invece, risponde: “L'Innominato si interroga sul senso della vita e sul valore di essa”.

Mi permetto di sottolineare quanto sia calzante a questo proposito l'aforisma, tratto da *Detti e Contraddetti* di Karl Kraus, posto all'inizio del Syllabus:

“Se, dopo tanti anni, uno sa ancora da quale dramma classico e da quale atto è presa una certa citazione, la scuola ha fallito. Ma se uno ha idea di dove potrebbe stare quella citazione, allora è una persona veramente colta e la scuola ha raggiunto appieno il suo scopo.”

Tutto torna, insomma.

Lezione 12 Aprile

1) Applicazioni in ambito filologico:

I 2 scopi della filologia:

- a) ricostruire la storia del modo in cui un testo è giunto sino a noi, come è stato diffuso e come è stato fruito;
- b) ricostruire la forma più vicina possibile a quella decisa dall'autore del testo stesso.

Il punto b) prevede l'utilizzo, tra gli altri, di due metodi differenti:

- il metodo di Lachmann;
- il metodo di Dom Henri Quentin.

Il metodo *lachmanniano* o *stemmatico* si basa sulla ricostruzione dell'albero genealogico, lo *stemma codicum*, che rappresenta i rapporti genetici tra i manoscritti, rapporti che vengono ricostruiti in base agli errori che ciascun testimone porta con sé. Si seguono quindi, nella ricerca della forma più corretta possibile, le seguenti fasi:

- *recensio*, ossia la ricerca dei testimoni;
- *collatio*, ossia il confronto tra i testimoni;
- *emendatio* (eventualmente), ossia la correzione degli errori;

Il metodo *statistico*, invece, prevede che i rapporti tra i testimoni non vengano stabiliti, come prima, sulla base della valutazione degli errori, ma sulla base della distribuzione statistica di varianti specifiche. Per ogni variante (ad es. bonus/optimus al verso 1), si vede quali manoscritti sono vicini tra loro, cioè hanno la stessa lezione. Lo stesso si fa per tutte le varianti (ad es. est/sit al verso 2; alter/ater al verso 3 etc.), e alla fine si valuta statisticamente quali manoscritti nel complesso sono risultati più spesso vicini tra loro, e si individuano così dei gruppi all'interno dell'insieme dei manoscritti. Questo sembra essere l'unico modo di procedere in tradizioni ampiamente interpolate (vd. il Nuovo Testamento).

Ora, questo metodo venne proposto negli anni '20 da Dom Henri Quentin e venne ripreso negli anni '60 da J. Froger, al tempo in cui esistevano già gli elaboratori elettronici. In realtà, l'applicazione del metodo statistico alla filologia ha incontrato l'opposizione da

parte di molti studiosi legati al metodo di Lachmann. Inoltre, parecchie opposizioni sono state mosse dai filologi anche per l'uso degli strumenti informatici nell'ambito di:

- *recensio*;
- *collatio*;
- ricostruzione automatica dello *stemma codicum*;

Di fatto, l'unico strumento informatico applicato alla filologia che è approvato all'unanimità dagli studiosi di essa è la stampa del testo critico con apparato.

Quanto alla *recensio* digitale, si crea un database con le differenti varianti. Successivamente si procede alla *collatio* digitale che, nel caso di una *recensio* chiusa, può seguire il modello *cladistico*, che è stato ripreso dagli studi della genetica e che si riavvicina all'impostazione lachmanniana: in pratica, essa si basa sulla creazione automatica di un modello ad albero (*κλαδος* significa "ramo" in greco, per l'appunto) che indica i rapporti di parentela tra i testimoni, ma sta al filologo stabilire da dove parte l'errore e quindi da dove "appendere" o fare iniziare l'albero, insomma. Nel caso di una *recensio* aperta, invece, come per i testi di tradizione orale che solamente in seguito sono stati canonizzati venendo messi per iscritto (ad esempio, i *Canterbury Tales*), ecco che recentemente sono stati creati dei siti, come lo stesso *Princeton Charrette Prototype*, che permettono una lettura "dinamica" di testi di questo genere, attraverso il confronto continuo con le versioni presenti nei differenti testimoni.

N.B: naturalmente il filologo deve effettuare delle formalizzazioni degli aspetti automatizzabili del lavoro filologico.

Allora, alcune attività dell'ambito filologico che il computer può svolgere in maniera semi-automatica sono:

- la *eliminatio codicum descriptorum*, ovvero l'eliminazione dei codici che costituiscono delle "copie" derivate dai codici più antichi, (che fa parte sempre della *recensio*);
- usare dei programmi in grado di ricostruire uno *stemma codicum* attraverso il metodo statistico. Sta poi al filologo stesso assegnare ad ogni *lectio* il suo peso specifico (operazione che fa parte della *collatio*);

c) la realizzazione di un programma che sia capace di individuare i *loca variantia*, ovverossia tutti i punti in cui i vari testi differiscono tra loro (che è sempre della *collatio*).