



Nuove frontiere informatiche dell'analisi linguistica

Simona Colombo

Con riferimento ad un articolo di Dino Bozzetti[1] "la semplice sequenza di caratteri non è sufficiente a rappresentare tutta l'informazione contenuta nel 'materiale letterario qual'è originariamente scritto da un autore", l'utilizzo di strumenti informatici nello studio ed elaborazione di un testo permettono di inserire dei dati ausiliari all'interno del testo, ma da esso enucleabili,mirati all'ottenimento di una rappresentazione digitale di un testo che sia elaborabile esaustivamente per indagini automatiche di linguistica computazionale.

Padre Roberto Busa, uno dei pionieri dell'informatica linguistica, che ha analizzato l'intera produzione di Tommaso d'Aquino, realizzando un "*Index thomisticus*" (56 volumi) in cui sono elencati tutti i termini usati da Tommaso d'Aquino e la posizione di ciascuno all'interno delle opere del filosofo, in un'intervista rilasciata al *magazine* Mediamente (RAI), afferma in merito all'evoluzione della linguistica nell'ottica della sua integrazione con le discipline informatiche:

> "Ho alcuni desideri.
> Il primo: che non si studino i testi di un sola lingua, ma si facciano assaggi coordinati sui principali sistemi linguistici ed alfabetici di questo mondo.
> Secondo: andrebbe continuata un'analisi più metodica, laboriosissima, per rifare una filologia, ossia una morfologia, una sintassi, un lessico per uso di *computer*.
> La filologia non va gettata via, ma va approfondita, precisata e accompagnata da informazioni quantitative probabilistiche.
> Per le applicazioni pratiche del linguaggio, quello che manca non è da parte del *computer*, non è che manchino rapidità di accesso o ampiezza di memoria o abilità di *software*, mancano informazioni di base, di natura filologica, su come microanalizzare, e quindi mitizzare, le operazioni umane che noi vorremmo delegare alla macchina."[2]

L'uso dell'informatica pertanto diventa interessante per la possibilità di applicare metodi statistici sofisticati che permettano al filologo di individuare un insieme di comportamenti lessicali, sintattici e stilistici significativi senza dover analizzare manualmente tutti i diversi testi.

Lo sviluppo dell'informatica ha portato a notevoli evoluzioni negli ultimi anni nello studio e nell'analisi di un testo assistito dal calcolatore concentrando la sua attenzione sullo studio sistematico e formalizzato di raccolte di testi sui quali operare elaborazioni statistiche, tale insieme di testi che rappresentano in maniera quantitativamente equilibrata le varie tipologie di testi reali, scritti e trascritti dal parlato, di una lingua prende il nome di corpus.

La nozione di testo intesa come sequenza o stringa di caratteri è molto differente da quella del testo inteso in termini letterari. Il testo inteso come sequenza di caratteri non coglie che una piccola parte dell'informazione testuale, occorre perciò aggiungere esplicitamente, attraverso l'inserimento di annotazioni o segni convenzionali, tutta l'informazione testuale

che non è possibile rappresentare con i semplici caratteri, strutturando cioè i dati testuali come informazione.

E' necessario quindi scindere il concetto di contenuto di un testo da quello della sua rappresentazione. Il contenuto di testo non coincide con la sua rappresentazione ma contemporaneamente la rappresentazione di un testo è strumento per esplicitarne il contenuto. Secondo Buzzetti, questa dicotomia tra contenuto e rappresentazione è la stessa, trasposta sul piano della rappresentazione digitale, che si incontra tra dato, inteso come elemento del testo digitale, e modello di dato come forma del contenuto del dato, o ancora la distinzione tra formato, inteso come sintassi secondo cui viene rappresentata una informazione e formalismo che è l'interpretazione della rappresentazione usata[3].

I Linguaggi di formattazione dei testi

Un linguaggio di marcatura (*markup*) è un sistema strutturato che si pone tra il documento "fonte" e la sua visualizzazione. Attraverso l'inserimento nel corpo del testo di specifiche etichette (*tag*), il programma interprete può definire correttamente l'informazione associata, sia questa di natura editoriale (marcando ad esempio una riga come titolo o una frase con il carattere allineato a destra) che di natura informativa (esprimendo ad esempio l'origine del documento analizzato o la sua edizione).

L'origine dei linguaggi di marcatura è lo *Standard Generalised Markup Language* (SGML) in cui non vengono definiti dei vincoli per la generazione dei *tag*: generato un *file* SGML il suo visualizzatore (*browser*) traduce la struttura per fornirne la giusta interpretazione. Questo approccio implica grande versatilità in fase di generazione, ma anche difficoltà di comunicazione tra *file* diversi, spesso composti da strutture eterogenee.

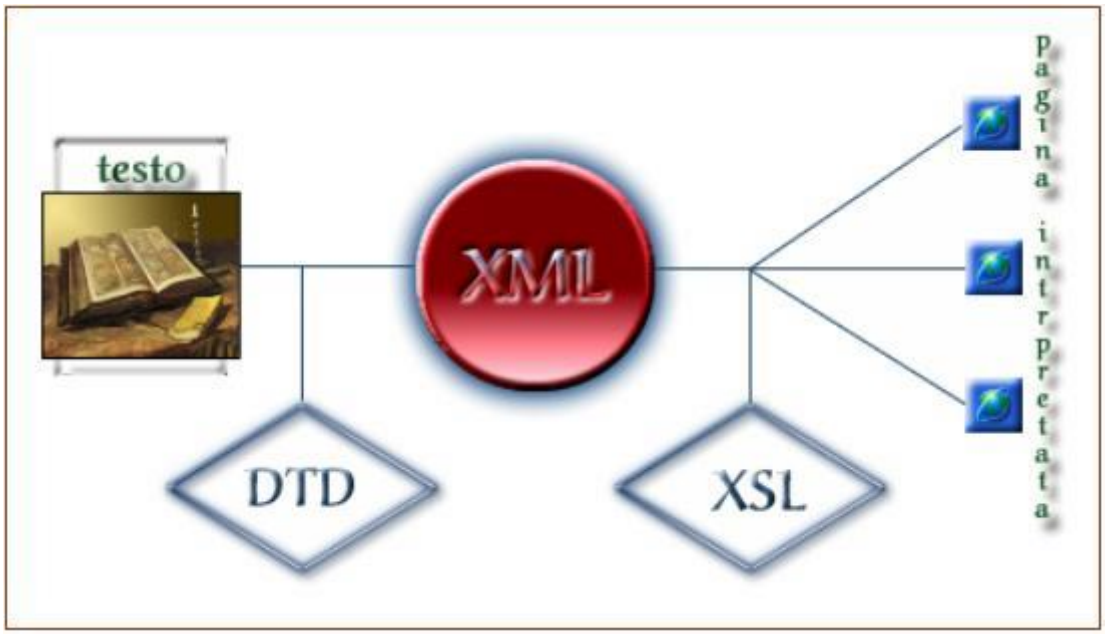
Nell'ottica di semplificarne l'utilizzo, sulla base delle prime esperienze di pubblicazione di informazioni via *internet*, nasce l'HTML (*Hypertext Markup Language*) come sottoinsieme dell'SGML, ma il suo obiettivo è la descrizione della struttura fisica di un documento e della sua rappresentazione visiva, impostata sviluppando fogli di stile (*Cascade Style Sheet* –CSS) che permettono di filtrare un documento secondo la veste tipografica desiderata. I *tag* utilizzati dichiarano come gli elementi testuali sono disposti all'interno di una pagina, quando il *browser* processa la pagina, la semantica viene ignorata, la rappresentazione dei dati non tiene conto della loro natura.

Un primo livello di analisi permette di interpretare ad esempio un testo letterario filtrandolo con differenti fogli di stile per ottimizzare la sua resa in relazione all'uso che l'utente desidera fare del documento in oggetto: è possibile creare un foglio di stile per l'edizione elettronica del testo consultabile a video ovvero un foglio di stile per l'ottimizzazione della stampa su stampante, o ancora per una pubblicazione multimediale.

L'XML (*Extensible Markup Language*), i cui standard sono sviluppati e approvati dal *"World Wide Web Consortium"* (W3C) del M.I.T. di Boston, nasce per permettere la definizione e la descrizione delle informazioni testuali prescindendo dalle indicazioni relative alla loro visualizzazione e mantiene, come l'SGML, la caratteristica di non possedere un'insieme predefinito di *tag* (infatti è definito linguaggio a *tagset* aperti), ma la libertà della loro definizione è vincolata da uno schema del documento DTD (*Document Type Definition*) sviluppato contestualmente al *file* XML stesso. Il DTD, inteso nella sua accezione generale (testo letterario, compendio di regole chimiche, dispense matematiche, raccolta di indirizzi telefonici) definisce gli elementi ed i loro attributi che verranno esplicitati nel file xml: a ciascuno di questi assegna una relazione logica con l'elemento che lo precede e con uno o più elementi che lo seguono, nella definizione di una struttura gerarchica (ad albero),

componendo dei "vocabolari". Il *tag* XML rappresenta quindi il testo contenuto al suo interno e non la posizione fisica nella pagina, codificando pertanto la semantica e la struttura dei dati e mantenendo il loro eventuale ordinamento gerarchico.

Analogamente a quanto permette il CSS per il *file* HTML, le regole editoriali per il *file* XML sono definite con l'*Extensible Style Sheet Language* (XSL), che però nasce con maggiori potenzialità rispetto al CSS che gli permettono ad esempio di operare delle estrazioni sul documento XML che si sta interrogando permettendone quindi una sua prima rimanipolazione logica.



E' possibile generare un unico file XML che rappresenti un testo e contenga in esso le citazioni critiche di autori differenti (sulle modalità di generazione un struttura analoga si tornerà a breve).

Attraverso la stesura di un opportuno XSL sarà possibile, a seguito di una specifica richiesta dell'utente, la visualizzazione dell'intero testo, o della sua critica riconducibile un autore o secondo un intervallo temporale o secondo la natura stessa della critica, il tutto con un solo XSL ed un solo documento XML senza ulteriori integrazioni di algoritmi di programmazione

Per accedere dinamicamente alla manipolazione della struttura e la formattazione di una stringa in XML è disponibile la piattaforma di strumenti DOM (*Document Object Model Specification*) che permette di navigare all'interno di un *file* XML,seguendo l'albero nella sua struttura, individuandone i nodi, gli attributi, le relazioni gerarchiche e permettendo di aggiungere, modificare o cancellare gli stessi o di utilizzare uno specifico linguaggio di interrogazione che permetta di estrarre informazioni all'interno del documento stesso senza doverlo navigare nella sua struttura ad albero ma interpretando i singoli nodi come entità sulle quali operare le indagini.

Analisi e archiviazione dei testi

Come già anticipato, premessa necessaria per affrontare l'analisi di un testo è la sua disponibilità formato elettronico, tale rappresentazione può essere o meno accurata ed esaustiva, a seconda del metodo utilizzato per generarla, considerando l'edizione elettronica

di un testo come problema fulcro per lo studio linguistico assistito da un calcolatore. In un'intervista Dino Buzzetti afferma:

> "Come fare ricerche lessicali sull'uso delle espressioni dialettali, sul rapporto tra diversi usi linguistici, lavorando su di un materiale costruito artificialmente e non rifacendoci alle fonti originali? Quindi il *database* può essere una forma di rappresentazione adeguata di forme di testualità diverse da quella canonizzata dal libro a stampa. Altre forme di questo tipo sono, per esempio, i manoscritti d'autore lasciati incompiuti. Qual è il testo? Le varianti sostitutive dell'autore? Non è un testo che deve ancora arrivare a compimento. E' un testo compiuto, costituito di posizioni funzionali, ha più valori, perché l'autore non ha ancora scelto tra le diverse varianti. Quindi lo studio di documenti di questo genere pone le varianti tutte sullo stesso piano. Il concetto di edizione critica scarta le lezioni discordanti dal testo, per definizione, come le lezioni di secondo piano, come lezioni che possono avere qualche interesse. Ma la rappresentazione come *database* è un'edizione? Secondo noi può essere un'edizione solo con determinati requisiti. Un archivio di tutta la tradizione testuale, non è un'edizione. Allora quali requisiti deve avere questa forma di rappresentazione, per potere essere considerata una edizione? Secondo noi, introdurre procedure computazionali, che svolgano su questo materiale la stessa funzione che in una edizione critica svolge l'apparato, un filtro per potere filtrare l'informazione e presentare un'opzione possibile, ma un'opzione sempre rivedibile, perché consente il confronto diretto con la fonte. Le fonti sepolte negli apparati che nessuno riesce a decifrare o a leggere. Già ricostruire la situazione testuale a partire da un apparato di un'edizione scritta è un'operazione che, anche a persone molto addestrate, crea notevoli difficoltà, in secondo luogo non c'è possibilità di rivedere la scelta che ha fatto l'editore in quel particolare punto. Queste erano le linee che ispiravano il nostro lavoro e quindi il problema a questo punto diventa un problema di informatica. Qual è il modello computazionale adeguato a risolvere questo tipo di problema? Allora ci si rende conto che il tipo di dati, il modello di dati, attraverso cui si rappresenta in forma elettronica il testo, diventano importanti per potere prevedere procedure di elaborazione richieste. Il problema informatico diventa quello di trovare un modello computazionale adeguato"[4]

A partire dal 1987 le tre maggiori associazioni mondiali di studiosi di scienze umane attraverso metodologie informatiche, la Association for Computers and the Humanities (ACH) la Association for Computational Linguistics (ACL) e la Association for Literary and Linguistic Computing (ALLC) hanno deciso di utilizzare l'SGML, ovvero un markup dichiarativo piuttosto che procedurale, per la codifica dei testi umanistici dando il via ad un progetto (TEI) Text Encoding Initiative che ha portato alla definizione di un DTD appositamente sviluppato per i testi letterari, storici e filosofici. le cui caratteristiche sono state per la prima volta presentate nella pubblicazione uscita nel 1994 col titolo *Guidelines for Electronic Text Encoding and Interchange* (TEI P3).

La TEI ha definito delle linee guida per descrivere la struttura di un testo e propone dei nomi per individuare i suoi componenti (titoli, capitoli, note, riferimenti bibliografici...) favorendo uno degli aspetti fondamentali del lavoro di ricerca sui testi letterari che è l'interscambiabilità del materiale di studio.

La codifica proposta dal TEI permette il superamento dei tipici problemi di edizione di un testo; esempi tipici sono la numerazione di pagina impostata in modo difficilmente

riconoscibile dai programmi di trattamento testi, o ancora la mancata distinzione tra apice ed apostrofo o tra i simboli che introducono un discorso diretto e la linea di sillabazione.

Lo scopo di questa codifica è fornire uno standard di rappresentazione dei testi umanistici e normalizzare i diversi formati di memorizzazione di un testo, infatti affronta specificatamente i diversi generi letterari per fornire per ciascuno di questi uno standard rappresentativo (opera drammatica, prosa, poesia, atto teatrale...).

Attraverso un opportuno linguaggio di marcatura, sia questo conforme o meno alle specifiche TEI, è pertanto possibile aggiungere al testo un'insieme di informazioni che ne permettono un'indagine accurata.

Laddove le informazioni inseribili all'interno del linguaggio di marcatura non fossero esaustive o laddove il tipo di informazione da inserire non sia gestibile a livello di informazione testuale (ad esempio aggiunta di supporti multimediali) o laddove le informazioni da aggiungere fossero ripetitive (ad esempio la valorizzazione …) allora le informazioni testuali possono essere affiancate da ulteriori informazioni catalogate in una base dati esterna.

Questo amplia in maniera ragguardevole lo spettro di indagine sull'analisi del testo, perché la trasversalità dell'interrogazione può permettere l'inserimento di qualsivoglia collegamento anche interdisciplinare.

Il tipo di informazioni testuali che vengono inserite sono fondamentali per la strutturazione della successiva indagine.

Nello sviluppo di un corpus, il tipo di informazioni che si decide di inserire all'interno dei tag del markup permettono e guidano le successive indagini testuali sul corpus in oggetto.

A tal proposito è importante suddividere il tipo di marca che si associa ad un testo per crearne differenti livelli di interpretazione in quanto è fondamentale stabilire rigorosamente una classificazione del tipo di metainformazione aggiunta al testo per poter agevolare lo studio di corpora differenti sviluppati da gruppi di ricerca distinti.

Inserire degli identificatori all'interno di un testo permette di isolare in esso tre diversi piani di analisi:

- Documentazione (bibliografia, lingua, autore..)
- Dati primari (paragrafi, titoli, note)
- Annotazioni linguistiche (*POS Tagging*)

Secondo quanto detto il *markup* si presenta come uno degli strumenti utilizzabili per rappresentare la struttura di un testo; si possono distinguere due diversi tipi di markup: vincolato e sciolto.

Il *markup* vincolato è inserito nel testo e la posizione in cui è inserito è una delle informazioni che il *markup* stesso rappresenta. Il fatto che dipenda pertanto dalla forma della rappresentazione del testo comporta che abbia dei limiti per esprimere la struttura di questo. Quindi il *markup* vincolato appare più indicato per esprimere le strutture dei dati che non per esprimere un modello di dati, serve per descrivere la forma dell'espressione dei dati e non gli elementi strutturali dell'informazione rappresentata.

"Un sistema di markup vincolato serve fondamentalmente per esprimere proprietà strutturali riguardanti le caratteristiche 'notazionali' della rappresentazione dell'informazione"[5].

Un esempio di questo tipo di *markup* è SGML che nasce per descrivere la struttura dell'espressione del testo, basandosi inoltre sul presupposto che tutte le strutture logiche di un testo siano di tipo gerarchico e pertanto rappresentabili con una rappresentazione ad

albero i cui nodi sono ordinati linearmente tanto da poter essere contrassegnati con dei marcatori chiusi e annidati.

Il *markup* non vincolato invece, non si presenta come portatore dell'informazione della sua posizione all'interno del testo analizzato. Essendo un marcatore che non si colloca all'interno della forma scelta come rappresentazione del testo appare più idoneo a rappresentarne un modello del suo contenuto, esso può addirittura essere collocato esternamente al testo stesso.

Theodor Nelson[6], inventore dell'ipertesto, sostiene l'utilizzo di un *markup* non vincolato come unica forma di rappresentazione esaustiva del tipo di pensiero espresso con le parole e la scrittura, i cui marcatori possono essere simili a quelli dell'SGML ma non sono inseriti in maniera vincolata nel testo ma "trattati separatamente".

Il modello di *markup* sviluppato facendo uso di marcatori non vincolati prevede la definizione di un insieme di contrassegni (*token*) cui vengono associati un insieme di attributi.

In questo modo è possibile, secondo un processo di assegnazione degli attributi lineare, rappresentare diversi livelli di strutture sul testo, siano queste gerarchiche o meno.

In questo modo, sviluppando opportunamente i filtri di interrogazione è possibile leggere il testo interrogando solo il livello di indagine desiderato.

Ovvero per ciascun *token* è possibile individuare attributi di categoria distinta, ad esempio si può analizzare morfologicamente il *token* individuandone il suo tipo grammaticale e ancora etichettarlo a seguito dell'identificazione nel testo delle frasi nominali ed analogamente i gruppi verbali.

Allo scopo di riconoscere le frasi nominali, si può definire un'espressione regolare che ne descriva il profilo. Estendendo le tecniche di riconoscimento dei sintagmi nominali, si possono definire delle espressioni regolari in grado di riconoscere dei gruppi verbali (porzioni di frase che raggruppano uno o più verbi consecutivi) e di identificare l'inizio sia dei sintagmi nominali che dei gruppi verbali, in questo modo è possibile tracciare delle relazioni sintattiche a basso livello tra le diverse strutture sintattiche.

L'interrogazione dei *tag* di *markup* porterà ad indagare circa la struttura, la forma ed il registro del testo analizzato, permetterà altresì di indagare sulle informazioni puramente "anagrafiche" del testo interrogato quali autore, anno, luogo di pubblicazione.

La reale informazione linguistica, sulla quale effettuare l'indagine statistica sarà effettuata sui tag del testo che ne trasmettono l'informazione.

Quindi ad esempio nel caso di un *corpus* "postaggato" (in cui sono state etichettate opportunamente con metodi anche statistici le differenti parti del discorso) sarà possibile operare un'indagine che coinvolga la distribuzione ad esempio di verbi o di coppie nomi – aggettivi.

Lo scopo per cui, affiancata a questo tipo di indagine, si può disegnare, costruire e popolare una base dati è la razionalizzazione delle informazioni necessarie e disponibili, l'arricchimento dell'archivio stesso per permettere analisi sempre più raffinate, l'aggiornamento i dati senza dispersione o ridondanza e soprattutto un'interrogazione logica.

E' possibile classificare un dato definendo per esso molteplici attributi, che ne permetteranno la sua interrogazione secondo differenti prospettive.

Per un termine può ad esempio essere fornita una forma regolarizzata secondo un dialetto, i nomi propri potrebbero essere riferiti ad un lessico specializzato, potrebbe essere associato ad ogni termine un appropriato codice per definirne il contesto linguistico, quindi essere associato un collegamento con la critica o con le traduzioni del testo, potrebbero essere associate immagini o suoni.

Questo lascia intendere la molteplicità di punti di vista secondo i quali sarà poi possibile ricercare nel contesto di indagine il termine oggetto di analisi, che diventerà pertanto cartina tornasole di differenti fenomeni linguistici e stilistici del testo.

Indagine linguistica e reti di conoscenza

Dal momento in cui le informazioni sono catalogate nella base dati ed nel *markup* del testo è possibile pensare alle differenti metodologie e finalità di interrogazione e di confronto dei dati registrati, nucleo del lavoro di indagine linguistica su di un testo.

Un ambito in cui la possibilità di effettuare indagini statisticamente rigorose grazie all'utilizzo dell'informatica e della potenza degli algoritmi di estrazione è quello dell'ambiguità di interpretazione di un testo.

In merito J.L. Austin evidenzia la difficoltà di attribuzione dei significati sostenendo che un'affermazione non si può univocamente definire vera o falsa ma la decisione è influenzata dalla situazione e dal livello di accuratezza ricercato:

> "La Francia è esagonale.
> Lord Raglan vinse la battaglia di Alma.
> Oxford è a 100 km da Londra.
> È ben vero che per ciascuna di queste asserzioni ci si può porre la questione " vera o falsa". Ma non è che nei casi abbastanza favorevoli che dobbiamo aspettarci una risposta Sì o No una volta per tutte. Ponendo la domanda si comprende che l'enunciato deve essere confrontato con i fatti in un modo o in un altro. Indubbiamente. Che dire? È vero oppure no? Domanda, lo si vede, semplicistica. Bene, se volete, fino ad un certo punto si può vedere quello che volete dire; vero forse per alcuni scopi e in alcuni contesti: per l'uomo della strada potrebbe andare, ma non per i geografi. E cosí via. È un'asserzione abbozzo [*rough statement*] , che volete? , ma non si può dire semplicemente che sia falsa. E la battaglia di Alma, battaglia del semplice soldato , se mai ve ne fu una? È vero che Lord Raglan aveva il comando dell'armata alleata e che questa armata riportò in una certa misura una specie confusa di vittoria; sì questo sarebbe giustificato, anche meritato, per degli scolari almeno, sebbene veramente un po' esagerato. E Oxford ?, sì, è vero che questa città è a 100 km da Londra, se non volete che un certo grado di precisione. Sotto il titolo di " vero " ciò che abbiamo di fatto non è affatto una qualità semplice, né una relazione, né una cosa qualunque, ma piuttosto tutta una dimensione di critica."[7]

Consideriamo a titolo esemplificativo che a ciascuna parola può essere associata non una singola entrata lessicale, ma un insieme di tali entrate, sia perché in ogni lingua le parole hanno spesso più di un significato, sia perché può accadere che una stessa forma flessa possa avere differenti interpretazioni ("pesca" non solo presenta l'ambiguità tra l'interpretazione come sostantivo e quella come voce del verbo "pescare", ma anche quella tra diverse voci del verbo quali l'imperativo e l'indicativo).

L'analisi automatizzata dei corpora permette di ottenere risultati significativi per l'implementazione di algoritmi statistici che, partendo e imparando dai risultati ottenuti indagando la parola ambigua nell'insieme dei contesti in cui il corpus la produce, permettono di elaborare strategie di disambiguazione sofisticate e con risultati soddisfacenti.

Interrogando il *file* "taggato" ed indagando sulle frequenze riscontrate con i diversi significati si può arrivare a costituire un algoritmo che associa al vocabolo ambiguo un

significato selezionato in base al calcolo statistico.

Reinserire nella *corpus* questo nuovo testo disambiguato significa aver aumentato la numerosità dei campioni di riferimento per la analisi successive.

Un problema affascinante in tale ambito è lo scambio di testi "arricchito" di informazioni tra diversi centri di studio.

Quanto esposto in precedenza può essere inserito in un unico *file* XML e permette di avere l'intera base dati raccolta in un semplice *file* di testo che mantiene e rispetta la classificazione tematica e gerarchica della base dati che lo ha generato.

Questo significa che due poli di studio possono scambiarsi le basi dati integre di relazioni e attributi, sviluppate in seguito ai loro studi, tra due diversi punti di una rete, sia questa locale o geografica, semplicemente scambiandosi due *file* di testo, ossia in maniera del tutto indipendente dalla specifica struttura utilizzata per generare le informazioni.

Acquisito il documento di testo di un diverso polo di studio è possibile, applicando la logica inversa a quella descritta, inserire nella base dati le informazioni nuove presenti nel documento XML ricevuto, integrare le proprietà delle informazioni già censite e quindi arricchire il bacino di informazioni sul quale operare gli studi successivi. Essendo una parte degli studi linguistici operati su di un testo di natura statistica, avere a disposizione un congruo numero di "casi linguistici" sui quali rilevare le occorrenze aumenta l'affidabilità dei risultati ottenuti.

Questo approccio permette di costruire una fonte di informazioni tematica che si arricchisce a seguito di nuovi studi e che permette di migliorare la qualità della ricerca grazie alla disponibilità crescente di materiale, soddisfacendo l'attuale esigenza di centralizzare la conoscenza di un gruppo di lavoro specializzato per mezzo della condivisone delle singole conoscenze acquisite. Questo obiettivo è perseguito in molte realtà didattiche e lavorative per ottimizzare i benefici ottenuti dalle singole esperienze ed individuare gli aspetti migliorabili o assenti al fine di capitalizzare l'uso delle risorse intellettuali.

Un approccio alternativo è utilizzare i risultati ottenuti da un indagine su di un testo come nuovi dati da inserire nella base dati per gli studi sui testi successivi, quindi è possibile ipotizzare una metodo di indagine che attraverso la rimanipolazione di documenti XML a seguito di indagini linguistiche, inserisca i risultati ottenuti (nuovi documenti XML ottenuti dal documento di partenza opportunamente interrogato per indagare un preciso fenomeno linguistico) all'interno della base dati, non più secondo gli attributi di partenza ma secondo i risultati ottenuti dall'indagine.

Un ulteriore applicazione derivante dallo studio dei testi ed in particolare della loro collezione è la stesura di glossari specialistici.

La stesura di un glossario con i *corpus* presenta indubbi vantaggi di velocità e di varietà di risultati rispetto all'approccio tradizionale (che prevedeva la stesura delle liste di parole per mano di studiosi del linguaggio, che affiancati da specialisti nella dottrina di riferimento, cercavano di estrapolare da un numero elevato di testi le parole ed i contesti più significativi).

Innanzi tutto è necessario individuare l'insieme di testi che possono rappresentare adeguatamente la disciplina sulla quale si sta operando l'indagine.

Stilando la lista di frequenza dei termini presenti in questi testi è possibile individuare un set di termini che si può considerare candidato a popolare il glossario.

Si può ad esempio supporre di non includere nel glossario i termini che compaiono nei testi con un numero di frequenze inferiore ad un certo limite prefissato in modo da non disperdere troppo i risultati. Chiaramente questo tipo di semplificazioni va discriminato, ovvero è importante un'analisi manuale dei termini scartati per evitare di eliminare dall'elenco dei termini significativi proprio quelli che, talmente specializzati, appaiono un numero limitato di volte ma sono fortemente rappresentativi della dottrina analizzata.

I termini così selezionati possono poi venire ampliati andando ad analizzare i termini che compiono nelle righe di concordanza di quelli scaltri, ed analizzandone la loro frequenza. Analogamente si possono considerare le occorrenze dei termini nella lista in relazione alle parole con cui più frequentemente appaiono accoppiati e valutare queste nuove parole come possibili candidati per i termini del glossario stesso.

Una volta reperita una lista, ritenuta soddisfacente per rappresentare il dominio analizzato, l'utilizzo dei *corpus* permette di associare a ciascuna parola non solo la sua definizione, che può anch'essa essere reperita dal contesto in cui i singoli termini sono stati estrapolati, ma anche informazioni circa la posizione della parola nel contesto della frase, in pratica è possibile catalogare informazioni circa il suo uso.

Modelli di analisi distribuita – le reti geografiche

Per utilizzare una risorsa disponibile sotto forma di *file* XML non è indispensabile il caricamento di tutto il suo contenuto nella base dati singola di ciascun ente di ricerca. Per un certo tipo di indagine può essere utile confrontare il *file* XML prodotto da un altro centro con quello disponibile: con il termine confronto si intende anche la possibilità di includere il *file* in quello principale, facendolo magari diventare un nodo di uno specifico passaggio e facilitando il confronto tra i dati in esso contenuti. Al termine dello studio si può immagazzinare l'informazione necessaria al compimento dell'indagine ed evitare di inserire nella base dati delle informazioni non necessarie o ridondanti.

L'approccio presenta due vantaggi, uno di natura tecnica, poiché operando in questo modo si evita di sovrappopolare una base dati a danno delle prestazioni e dell'affidabilità, l'altro, più rilevante, è la possibilità di creare dei *file* XML specializzati e la cui rilevanza di contenuti nell'ambito di un'indagine linguistica può essere definita con maggiore accuratezza.

L'XML dispone di metodologie di comunicazione aperte fra le applicazioni : i *Web Services* "canali" che consentono alle applicazioni di comunicare attraverso la rete geografica, indipendentemente dal sistema operativo o dal linguaggio di programmazione utilizzato e attraverso cui possono condividere i dati, utilizzare funzionalità o moduli messi a disposizione da altre applicazioni indipendentemente da come queste sono costruite.

I *Web Services* possono essere immaginati come siti *internet*: raggiungibili con normali indirizzi web ed invocabili attraverso l'uso di XML e di semplici codifiche utili alla definizione del formato del messaggio, della sintassi con la quale il *web service* pubblica la funzione che espone e un indice in cui sia possibile sapere dove reperire su *Internet Web Services* utili per particolari necessità di studio o di analisi, ad esempio per lo studio critico di un testo.

Il concetto di accesso a una determinata funzionalità erogata da un ente didattico costituisce la base del modello *Web Services*: il polo di studio sviluppa ed eroga un servizio condiviso da una molteplicità di utenti, altri poli dislocati in differenti strutture, e propone questo servizio sulla base di un abbonamento, in modo da filtrare gli accessi e monitorarli, da una postazione centralizzata attraverso *Internet* o una rete privata.

Pensare che enti distinti possano contemporaneamente compiere studi analoghi o, cosa ancora più interessante per le enormi possibilità di approfondimento che apre, complementari, ovvero la cui rispettiva integrazione fornisce elementi altrimenti mancanti o incompleti, permette di creare un bacino di informazioni estremamente ricco e in continua evoluzione. Si viene perciò a creare una rete con "sportelli" di servizi virtuali utilizzabili

come risorse di studio e di ricerca, vicendevolmente arricchibili e modificabili a seguito dell'interazione con gli altri soggetti.

Aspetto fondamentale della ricerca diventa pertanto la possibilità di localizzare efficacemente le informazioni disponibili sulla rete geografica al fine di reperire lo strumento idoneo alla ricerca in corso, ovvero codificare un motore di ricerca in grado di "comprendere" il contesto della richiesta in modo da filtrare in base al contesto logico di appartenenza i risultati esposti, i documenti non dovrebbero più risultare come delle "isole di dati", ma piuttosto come dei database aperti nei quali un "applicativo" possa distinguere le informazioni contenute, ricavandone solo quelle richieste: questo approccio alla catalogazione ed estrazione delle informazioni prende il nome di "*Web* Semantico".

Il raggiungimento di tale obiettivo è perseguibile introducendo un ulteriore livello di marcatura dei documenti, un linguaggio gestibile dall'utente che li realizza e che, grazie all'utilizzo di informazioni strutturate e di regole di deduzione, possa condurre alla soluzione dell'interrogazione, generando documenti che possano al tempo stesso essere letti da esseri umani, ma anche interpretabili da agenti automatici alla ricerca di contenuti.

Il *Web* Semantico basa la sua struttura su tre distinti livelli:al primo livello i dati , informazioni sui dati e relazioni che intercorrono tra essi, ossia i metadati, al secondo livello la descrizione dello schema dei metadati, al terzo livello i vocabolari (ontologie) che definiscono il ruolo semantico dei metadati.

Il primo livello è secondo le indicazioni del W3C, strutturato secondo il linguaggio RDF (*Resource Description Framework*), un linguaggio per la descrizione di metadati pensato per affiancare XML stesso. I suoi elementi costitutivi e la sua sintassi sono XML. Tramite l'utilizzo di vocabolari specifici possono essere esplicitate le funzioni dei suoi elementi.

RDF consente di costruire delle asserzioni relativamente ai contenuti di una pagina *web*. Esse si creano in base a "dichiarazioni triple" costituite da soggetto, predicato e complemento oggetto. Tali asserzioni individuano relazioni fra i dati di cui trattano, ma non esplicitano ancora il loro significato.

Per definire il significato delle relazioni serve un ulteriore livello, il secondo, quello delle cosiddette ontologie, dei vocabolari nei quali collezioni di frasi sono associate a concetti, a regole logiche, il linguaggio utilizzato per esprimerle è l'RDF *Schema*, che, ulteriormente astratto con regole semantiche porta alla definizione del terzo livello con strumenti quali DAML (DARPA *Agent Markup Language*) e OIL (*Ontology Interchange Language*), oggi entrambi poco diffusi nelle applicazioni comuni.

Punto di forza del sistema è la sua "non rigidità". Ogni soggetto interrogato può definire, in uno schema, un insieme di relazioni fra i diversi elementi ma nulla vieta che in un futuro si possa collegare quello schema ad un altro che introduca nuove relazioni.

A tal proposito è esemplificativo l'intervento di Massimo Parodi e Alfio Ferrara che in un articolo spiegano:

> "*Semantic Web*: l'ipotesi cioè di arrivare a una sorta di formalizzazione – la più generale possibile – di aspetti semantici, di contenuto, di molteplicità di documenti e soprattutto di tipologie di documento reperibili in rete. Con una mossa teorica, che ricorda

altri passaggi della storia della filosofia occidentale, si sono venuti sviluppando metodi (e linguaggi) che intendono descrivere la semantica dei documenti, introducendo un impianto descrittivo definito su tre livelli : un linguaggio relazionale del tipo "soggetto – predicato – oggetto", per descrivere metadati e modelli dei dati, l'RDF (*Resource Description Framework*); un linguaggio e una sintassi che insieme definiscano e descrivano il vocabolario delle rappresentazioni desiderate, ossia un'estensione dell'RDF pensata per la rappresentazione di strutture più generali, di carattere classificatorio (classi e sottoclassi, per esempio). Queste strutture prendono il nome di schemi e il linguaggio è denominato RDF *Schema*. Infine, un livello nel quale viene definita formalmente la semantica e gli strumenti di supporto per l'interpretazione automatica. A questo livello, che riguarda la descrizione della realtà di riferimento, si collocano proposte di formalismi quali OIL (*Ontology Interchange Language*) e DAML (*Darpa Agent Mark-Up Language*)+OIL . Ma soprattutto, è a questo punto che compare un termine che in ambito informatico assume un significato quasi tecnico e che suscita invece una grande attenzione da parte di chi si accosta a questi temi da una prospettiva filosofica. La definizione infatti dei possibili oggetti di un mondo in base ai quali articolare la descrizione semantica di una realtà di interesse viene definita ontologia. Non è difficile da parte nostra osservare che un'ontologia, anche nel senso informatico, ha la pretesa di determinare con precisione quali predicati definiscano un soggetto, ne precisino cioè l'essenza, quali predicati siano possibili di un soggetto e, infine, quali oggetti necessariamente esistano nel mondo possibile descritto."[8]

Bibliografía

Austin J. (1971)	"Performative-Constative.", in The Philosophy of Language. Ed. John R. Searle. Oxford: Oxford UP
Boschetti F. (- - -)	*Informatica e analisi dello stile* - UNITN n. 32
Bowker L. - Pearson J. (2002)	*Working with specialized language. A practical guide to use corpora*, London and New York Rouledge
Busa, Padre Roberto (1995)	"Informatica e scienze umane", *Biblioteca Digitale Univ. La Sapienza*, Roma, 24/11/95 http://www.mediamente.rai.it/home/bibliote/intervis/b/busa.htm
Buzzetti D. (- - -)	"Rappresentazione digitale e modello del testo", in corso di stampa
(1995)	"Stereotipi della comunicazione", *Biblioteca Digitale Univ. La Sapienza*, Roma, 24/11/95, http://www.mediamente.rai.it/home/bibliote/intervis/b/buzzetti.htm
(1996)	"Il testo 'fluido'. Sull'uso dell'informatica nella critica e nell'analisi testuale", *Atti del primo incontro italiano sulle applicazioni informatiche e multimediali nelle discipline filosofiche*, in *Filosofia e informatica*, a cura di L. Floridi, Torino, Paravia.
Ciotti, F. (1997)	*Testo, rappresentazione e computer. Contributi per una teoria della codifica informatica dei testi, in Internet e le Muse*, a cura di P. Nerozzi Bellman, Milano, Mimesis.
Goldfarb C.F. (1990)	*The SGML Handbook*, Oxford, Oxford University Press.
Landow G. P. (1998)	*L'ipertesto. Tecnologie digitali e critica letteraria*, a cura di P. Ferri, Milano, Mondadori.

Leonardi P. (2003)	*Filosofia del linguaggio, Appunti del corso di Filosofia del Linguaggio presso la facoltà di Scienze della comunicazione di Bologna*
Parodi M. - Ferrara A. (2002)	"XML, Semantic Web Rappresentazione della Conoscenza" , *Mondodigitale*, n . 3 , settembre 2002.
Negri M. (1999/2000)	Tesi di Laurea in Filosofia del Linguaggio : LA VALUTAZIONE DI MODULI NELL'ELABORAZIONE DEL LINGUAGGIO NATURALE: PROBLEMI E METODI - Anno Accademico 1999/2000
Nelson T. H. (1990)	*Literary Machines*, Swarthmore.
Orlandi T. (1990)	*Informatica umanistica*, Roma, Nuova Italia Scientifica.
Roncaglia A. (1996)	«Procedimenti formali e "divinatio" nell'ecdotica», in *Lingua letteratura computer*, a cura di M. Ricciardi, Torino, Bollati Boringhieri.
Sue Atkins B. T. (2002)	*Then and Now: Competence and Performance in 35 years of Lexicography*– Copenhagen 2002 , *Atti del 10° Congresso Internazionale EURALEX*
TEI P3 (- - -)	*Guidelines for Electronic Text Encoding and Interchange*
Tomasi F. (2003)	*Manuale di informatica umanistica per l'applicazione delle pratiche computazionali ai testi letterari*, Edizione elettronica –2003, Portale web di letteratura "Griseldaonline" del Dipartimento di italianistica dell'Università di Bologna
Vital F. - Gentilucci R. (2002)	*Metadati: RDF e RDFS, Appunti del corso di Tecnologie Web* presso la Facoltà di Informatica dell'Università di Bologna.

Note

[1] "Rappresentazione digitale e modello del testo", in corso di stampa

[2] Padre Roberto Busa, "Informatica e scienze umane", *Biblioteca Digitale Univ. La Sapienza*, Roma, 24/11/95, http://www.mediamente.rai.it/home/bibliote/intervis/b/busa.htm

[3] D. Buzzetti, "Rappresentazione digitale e modello del testo", art. cit.

[4] D. Buzzetti, "Stereotipi della comunicazione", *Biblioteca Digitale Univ. La Sapienza* Roma, 24/11/95, http://www.mediamente.rai.it/home/bibliote/intervis/b/buzzetti.htm

[5] D. Buzzetti, "Rappresentazione digitale e modello del testo", cit.

[6] T. H. Nelson "Literary Machines", Swarthmore, 1990.

[7] J. Austin, "Performative-Constative", in *The Philosophy of Language*, John R. Searle (ed.), Oxford, Oxford UP, 1971.

[8] M. Parodi, A. Ferrara, "XML, Semantic Web Rappresentazione della Conoscenza" , *Mondodigitale*, n . 3, settembre 2002.

– per citare questo articolo:

Artifara, n. 4, (gennaio - giugno 2004), sezione Addenda, http://www.artifara.com/rivista4/testi/frontiere.asp

ISSN: 1594-378X