MANUELA SASSI - EUGENIO PICCHI

ELABORAZIONI TESTUALI: UN'ESPERIENZA ITALIANA

L'**Istituto di Linguistica Computazionale** ha una tradizione storica nel settore dell'analisi testuale e della lessicografia. Sono stati prodotti: standard, procedure, materiali testuali e lessicali.

Fra questi: **PiSystem** è un sistema integrato per il trattamento di materiali testuali e lessicali.

Il **DBT** -*Data Base Testuale*- costituisce il motore e lo standard di riferimento per l'intero progetto *PiSystem*. La elaborazione testuale resa disponibile dal DBT e dal suo *Query System* può essere applicata a testi (ASCII) provenienti da qualunque tipo di *software* e prevede le seguenti funzioni:

- testo non annotato;
- testo strutturato;
- testo lemmatizzato;
- testo annotato morfo-sintatticamente;
- integrazione con dizionari macchina (morfologia);
- integrazione con tassonomia;
- testi con apparato critico;
- testi con iconografia o sorgenti manoscritte;
- testi con fonti sonore di riferimento;
- testi contrastivi.

A partire da un archivio in DBT si può accedere al **DBT-Corpus,** che presenta le seguenti caratteristiche:

- possibilità di costruire velocemente *corpora* testuali molto grandi a partire da archivi già strutturati ed indicizzati;
- selezione *corpus* di ricerca selezionando informazioni bibliografiche o di classificazione associate a ciascun testo;
- funzione di interrogazione DBT disponibili, in tempo reale, su tutto il *corpus* selezionato.

Su ciascun archivio possono essere applicate le varie funzioni che integrano il *PiSystem,* fra cui citiamo brevemente:

- indici Vari (alfabetici, decrescenti, *index locorum*, ecc.);
- concordanze (Kwick, con vari ordinamenti);
- ricerca Co-occorrenze statistiche;
- ricerca sequenziale;
- lemmatizzazione/*tagging;*
- stazione di lavoro lessicografica;
- integrazione fra DBT & SGML/TEI;
- client per interrogazioni via Internet;
- inserimento a visualizzazione di immagini collegate.

In dettaglio e molto schematicamente, descriviamo le più importanti.

Ricerca Co-occorrenze statistiche

- Utilizzo della formula di *Mutual Information*;
- applicazione della formula ad un intero testo;
- selezione delle parole attive nel calcolo delle co-occorrenze;
- parametri: *window size*, utilizzo di una propria lista di parole vuote, soglia minima di frequenza, tipo di ordinamento;
- ricerca delle preposizioni in una finestra di testo a partire da parole attive selezionate;
- possibilità di vedere la finestra dei contesti di una coppia di parole ritrovata;
- allargamento dei contesti desiderati;

Ricerca Sequenziale

Strumento per la ricerca nel testo di *Pattern*, integrando il *Query System.*

- Testi non lemmatizzati e/o annotati;
- ricerca su tipologia di informazione (parola, punteggiatura, informazioni topografiche, ecc.);

- ricerca sulla grafia della parola;
- ricerca di vocali e/o consonanti e/o lettere specifiche;
- analisi sulla lunghezza della parola;
- analisi sulla sequenza degli elementi;
- generazione o meno dei contesti;
- testi lemmatizzati e/o annotati;
- ricerca sugli elementi del testo come per i testi non lemmatizzati;
- utilizzazione del lemma associato come elemento di ricerca;
- utilizzazione del codice grammaticale (TAG) associato come elemento di ricerca;
- utilizzazione delle informazioni morfo-sintattiche associate come elemento di ricerca.

Lemmatizzazione/*Tagging*

- Lemmatizzazione assistita dal calcolatore (stazione di lavoro specializzata);
- senza dizionari di riferimento;
- utilizzando dizionari di riferimento;
- dizionari e morfologia preesistenti (DMI);
- auto-dizionari che si creano e si arricchiscono automaticamente lemmatizzando più testi;
- procedura di *Tagging* e lemmatizzazione automatica;
- morfologia della lingua italiana;
- approccio statistico;
- post-editor interattivo;

Caratteristiche della funzione di Lemmatizzazione/*Tagging*

- Convergenza delle due procedure in un'unica struttura;
- inserimento nel DBT della classificazione introdotta sul testo;
- DBT può:
- registrare/contenere la classificazione ottenuta;
- interrogare il testo anche su queste nuove informazioni introdotte;
- generare indici e concordanze x lemma e/o forma;
- effettuare ricerche nel testo *tagged* di *pattern* sintattici;
- servire come supporto per nuove procedure di analisi;
- essere integrato nel sistema "Stazione di lavoro lessicografico".

Stazione di lavoro Lessicografica

E' un sistema integrato per l'aiuto alla redazione lessicografica che comprende le seguenti funzioni:

- gestione di entrate lessicali;
- editor specializzato delle entrate lessicali con tutta una serie di funzioni ad hoc per facilitare e guidare il lavoro redazionale;
- "*context cutter*" funzione per il "taglio" ottimizzato dei contesti;
- integrazione del DBT per analisi ed utilizzazione di testi e *corpora;*
- integrazione del "Polmone" deposito di contesti selezionati, classificati ed ordinati utilizzabili come un *framework*;
- inserimento automatico del lessico costruito nella DBT strutturato a costituire un data base lessicale;
- diversi approcci operativi;
- EDITOR per la creazione e l'aggiornamento della entrata lessicale strutturata;
- import/export per estrazione e l'inserimento di materiale di archivio lessicale con ampia possibilità di selezione;
- WSindici per la generazione di indici a stampa (anche complessi);
- WSmacro per la creazione e l'applicazione di macro (serie di interventi sulle entrate lessicali) da applicarsi analiticamente su l'intero archivio lessicale;
- WS-DBT per inserire l'archivio lessicale nella procedura DBT come materiale attivo;

DBT & SGML/TEI

DBT possiede un proprio sistema di codifica e strutturazione interna dei dati, derivata da decenni di esperienze nel settore dell'analisi testuale. Per rendere esportabile tale esperienza si è implementata una procedura di import/export tra DBT e SGML/TEI.

Nel passaggio da SGML/TEI a DBT tutte le informazioni contenute nel testo possono essere tradotte in codifiche DBT. Si sta cercando di generalizzare al massimo l'interazione tra le due codifiche senza lasciarle in qualche modo Pisa-dependent.

E' in corso di realizzazione una versione di DBT la quale può utilizzare materiale SGML/TEI senza ristrutturalo ma creando delle indicizzazioni DBT - la fase di segmentazione può essere pre-eseguita con altre procedure o compiuta dal DBT al momento della indicizzazione.

DBT & Internet

La messa in linea su Internet tramite il sistema DBT viene realizzata secondo due diverse metodologie:

- *Mosaic/Netscape*;
- creazione di un server Html che cerca di sfruttare al massimo le funzionalità DBT adattata alla modalità di tali sistemi;
- estrema facilità di uso ed estrema capacità di penetrazione nel mercato e di servizio agli utenti;
- *Client/Server*;
- *Splitting* del programma di interrogazione in due componenti *interneticamente* comunicanti avendo due obiettivi:
- esegue localmente la maggior quantità di operazioni possibili per non appesantire i tempi di connessione;
- dà all'utente la stessa interfaccia e tutte le stesse funzionalità del sistema standalone.

DBT e la iconografia

Il sistema DBT integra al suo interno la possibilità di gestire delle immagini, siano esse rappresentazioni grafiche originali dello stesso testo che iconografie inserite nel testo stesso.

Le immagini sono inserite nel testo, la loro presenza è segnalata ed è possibile richiamarle quando lo si desidera.

L'insieme di tutte le immagini di un testo possono essere automaticamente raccolte in una libreria di immagini consultabile in DBT e dalle immagini è possibile risalire automaticamente al testo che le contiene.

Le immagini visualizzate possono essere elaborate graficamente utilizzando una sofisticata libreria di funzioni grafiche (*LEADtools for Windows*).