Università Ca' Foscari di Venezia

Linguistica Informatica Mod. 1

Anno Accademico 2010 - 2011

TEI

Rocco Tripodi
rocco@unive.it

Esempi di corpora annotati

Il primo corpus annotato automaticamente a livello morfo - sintattico fu il Brown Corpus

Treebanks

- Corpora annotati a livello sintattico
- Rappresentazione dell'albero sintattico delle frase
- Penn Treebank Link
- Venice Italian Treebank Link (il visore funziona con IE)

Annotazione semantica

- Framenet nasce come annotazione dei ruoli semantici del *British Nationl Corpus*

Informazione relazionale

- Codifica delle relazioni tra le unità linguistiche. Anafora – pronomi clitici
- Relazioni di dipendenza: soggetto → oggetto
- Coreferenza: usare *the president* per riferirsi ad una entità specifica
- Utile per la definizione dei ruoli semantici

Codifica elettronica dei testi

Obiettivi

Sviluppare teorie e modelli formali del testo (o di alcuni suoi livelli) individuare formalismi atti a esprimerli in modo computazionalmente accettabile

Soluzioni

Adozione dei *markup language* descrittivi basati su XML

Testo visto come oggetto linguistico astratto organizzato secondo una struttura gerarchica ordinata (oggetti di contenuto)

Gli oggetti per essere descritti vengono serializzati ed inseriti all'interno di un albero gerarchico

Text Encoding Initiative 1

Schema di codifica indirizzato a chi vuole produrre e diffondere testi in formato elettronico, a fini scientifici, in particolare nel dominio umanistico

Mancanza di uno standard internazionale condiviso per quei contenuti che all'epoca venivano chiamati *machine readable text*

Il progetto nasce nel novembre 1987 in una conferenza programmatica tenutasi al Vassar College (New York)

Originariamente pensata in SGML poi convertita e sviluppata in XML

Text Encoding Initiative 2

Struttura astratta a cui diversi tipi di testo possono essere ricondotte (prosa, poesia, manoscritto, testo drammaturgico, ecc)

Consente di rappresentare le caratteristiche testuali rilevanti di diverse aree di ricerca (filologia, linguistica, narratologia, ecc.)

Schema di codifica flessibile che rintraccia le caratteristiche comuni di diversi testi e che allo stesso tempo consente l'aggiunta di elementi specialistici
Restrizioni sugli elementi comuni e opzionalità degli elementi speciali

Possono essere descritti e annotati anche elementi non testuali, come immagini e suoni

TEI: versioni speciali

TEI Lite

Rappresenta una versione ridotta della TEI
Facilitare lo scambio e l'integrazione di dati testuali di natura scientifica
Sottoinsieme di marcatori che rappresenta lo schema di partenza dal quale ogni utente dovrebbe partire

Proprietà

contiene elementi rilevanti virtualmente per tutti i testi e per tutti i tipi di elaborazione testuali (generalità)
consente di creare documenti *compliant* con l'intera DTD TEI

XCES

XML Corpus Encoding Standard
Standard per la codifica di corpus

Strutturazione dei documenti

Categorie di informazione

Documentazione
informazioni generali sul testo (codifica, contenuto, descrizione bibliografica, ecc)

Dato linguistico primario
Macrostruttura del testo (capitolo, verso, … , capoverso)
Microstruttura del testo (periodi, citazioni, nomi propri, ecc.)
Annotazione linguistica
Annotazione delle strutture del linguaggio (morfologica, sintattica, ecc)

TEI Lite: struttura

Tutti i documenti conformi contengono due elementi fondamentali intestazione <teiHeader> e testo <text>

Intestazione

Elemento composto che contiene informazioni analoghe a quelle contenute nel frontespizio di un testo a stampa. Si articola in quattro elementi:

1. descrizione bibliografica del testo digitale
2. descrizione delle caratteristiche inerenti il metodo di codifica
3. descrizione generale del testo
4. elenco delle revisioni

Testo

Elemento che accoglie la digitalizzazione del testo vero e proprio. Si articola in tre elementi generali:

1. <front> peritesto iniziale
2. <body> testo unitario
3. <back> peritesto finale

TEI Lite: esempio XML

```
<TEI>
   <teiHeader>
        <fileDesc/>
        <encodingDesc/>
        <profileDesc/>
        <revisionDesc/>
     </teiHeader>
     <text>
        <front> [Indice per esempio]</front>
        <body/>
        <back>[apparati finali, appendici, ecc] </back>
     </text>
</TEI>
```

TEI Lite: esempio DTD

<!ELEMENT TEI.2 (teiHeader, text) >

<!ELEMENT teiHeader
(fileDesc, encodingDesc*, profileDesc*, revisionDesc?) >

<!ELEMENT text
((anchor | gap | figure | index | interp | interpGrp | lb
| milestone | pb)*, (front, (anchor | gap | figure | index
| interp | interpGrp | lb | milestone | pb)*)?, (body |
group), (anchor | gap | figure | index | interp | interpGrp
| lb | milestone | pb)*, (back, (anchor | gap | figure |
index | interp | interpGrp | lb | milestone | pb)*)?) >

<group> raggruppa un insieme di testi unitari

TEI Lite: il frontespizio elettronico

Contiene le informazioni bibliografiche dell'edizione cartacea che si codifica e del file XML stesso

L'elemento principale è <fileDesc>

<!ELEMENT fileDesc
(titleStmt, editionStmt?, extent?, publicationStmt, seriesStmt?, notesStmt?, sourceDesc+) >

<titleStmt>
contiene gli elementi <title>, <author>, <name> (autore della codifica)

< publicationStmt>
informazioni bibliografiche del file

TEI Lite: peritesto

Il peritesto iniziale contiene elementi come il frontespizio e la prefazione

Frontespizio

rappresentato dall'elemento <titlePage>

<!ELEMENT titlePage
((anchor | gap | figure | index | interp | interpGrp | lb
| milestone | pb)*, (byline | docAuthor | docDate | docEdition
| docImprint | docTitle | epigraph | titlePart), (byline
| docAuthor | docDate | docEdition | docImprint | docTitle
| epigraph | titlePart | anchor | gap | figure | index |
interp | interpGrp | lb | milestone | pb)*) >

byline: contiene la dichiarazione di responsabilità di un'opera (Es: a cura di…)

titlePart: contiene una suddivisione del titolo. Ha attributi che ne specificano il tipo (principale, sottotitolo, ecc)

TEI Lite: body – sezioni e paragrafi

Partizioni testuali (dato linguistico primario)

Paragrafi <p>
Sezioni
Sotto sezioni <div1> - <div2> - ecc

Attributi di sezioni e paragrafi

<!ATTLIST div

next IDREF #IMPLIED	Riferimento alla sezione successiva
prev IDREF #IMPLIED	Riferimento alla sezione precedente
id ID #IMPLIED	Identificatore univoco
n CDATA #IMPLIED	Identificatore mnemonico
lang IDREF #IMPLIED	Lingua in cui è scritta la sezione
type CDATA #IMPLIED	Nome convenzionale che specifica
TEIform CDATA "div" >	Wordform del tag

Il valore di *id* deve essere unico in tutto il documento.

TEI Lite: prosa e poesia

Testi in prosa

non c'è differenza nell'uso dei tag <p> e
Gli elementi <*div*> possono contenere
<head> qualsiasi tipo di titolazione
<trailer> formula di chiusura o elemento a piè di pagina

Poesia

<l> singola riga, se incompleta si usa dall'attributo booleano *part*
<lg> codifica unità formali come la strofa
<sp> battuta di un testo drammatico (attributo *who*)
<speaker> fornisce il nome del parlante

```
<lg type="terzina">
   <l>Allor fu la paura un poco queta,</l>
   <l>che nel lago del cor m'era durata</l>
   <l n="21">la notte ch'i' passai con tanta pieta.</l>
</lg>
```

TEI Lite: struttura editoriale 1

Le interruzioni di riga e di pagina sono indicate con i tag vuoti <pb> e <lb> e specificate con l'attributo *n*

Stili e caratteri

cambio della fonte tipografica, nello stile di scrittura, nel colore dell'inchiostro
<hi> tag segnala una differenza di aspetto senza specificala
<hi rend="italic"> specifica i dettagli dell'aspetto grafico
<emph> codifica espressioni evidenziate per effetto retorico
<foreign> identifica una parola evidenziata perché in lingua diversa

Titoli

<title> con attributi che specificano a quale tipo di testo appartiene

```
<!ATTLIST title
  level (a | m | j | s | u) #IMPLIED
```

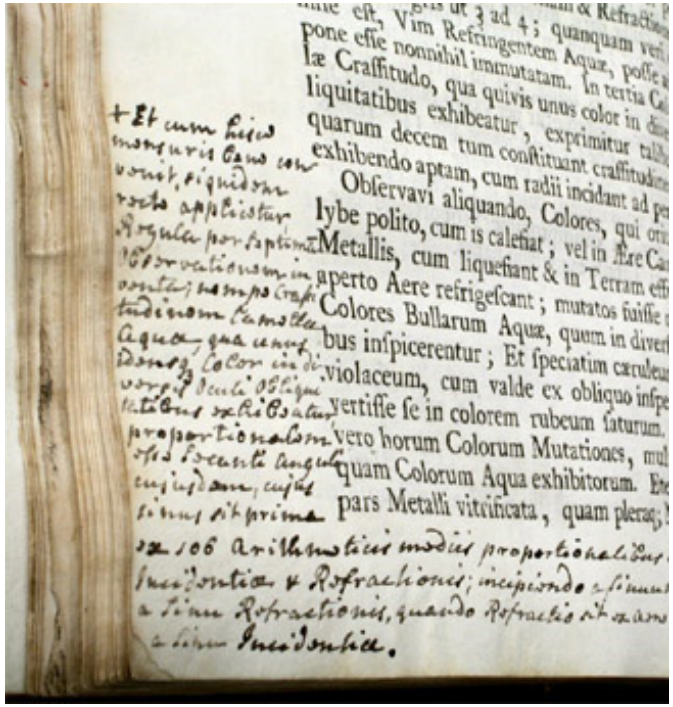
TEI Lite: struttura editoriale 2

Citazioni

<q> citazione (indipendentemente da segni che la introducono).
Attributi *type* e *who* (discorso diretto)
<mentioned> espressioni citate e riportate

Note <note>

attributi
type
resp: responsabile dell'annotazione
place: indica dove appare la nota
I *marginalia* che non possono essere inseriti in un luogo preciso, per convenzione vengono inserite prima del paragrafo cui si riferiscono



TEI Lite: riferimenti incrociati

Riferimenti incrociati semplici

Riferimenti impliciti o collegamenti da un punto ad un altro dello stesso testo
Vengono indicati con i tag
<ref> riferimento ad una posizione (elemento)
<ptr> puntatore (è un elemento vuoto)
Con attributi
target: destinazione del puntatore con uno o più identificatori XML
type: serve a categorizzare il puntatore (note, indice, ecc)
targType: specifica il tipo di elemento a cui si punta
crDate: data
resp: creatore

Poiché l'attributo Id è globale si utilizza per risolvere il riferimento

Es: vai alla <ref target="pag2"> seconda pagina </ref>

<anchor> specifica una posizione affinché possa essere puntata

<xptr> e <xref> consentono di puntare a documenti esterni

TEI Lite: dati strutturati

Espressioni referenziali

<rs> contiene un nome o un'espressione referenziale generica
<name> contiene un nome proprio o un sintagma nominale
attributi
key: fornisce un identificatore alternativo
reg: contiene una forma regolarizzata dell'entità nominata

Date e ore

<date> contiene date in qualunque formato
<time> contiene riferimenti ad orari in qualunque formato
Es: <date value="06-1996"> Maggio 2006 </date>

Numeri

Come per le date anche per i numeri è difficile avere una forma normalizzata.
<num> contiene un numero scritto in qualunque formato
type: indica il sistema di riferimento
value: normalizza il formato

Gerarchie sovrapposte

Dati due oggetti logici presenti in un testo, le coppie di tag bilanciati che li rappresentano non si annidano propriamente ma si sovrappongono

Complessità del testo

Es: alla struttura metrica si sovrappone il discorso diretto

Soluzione

Un elemento logico che si sovrappone ad uno o più elementi, viene segmentato in *n* elementi dello stesso tipo, con attributi che ne identifichino l'ordine. In questo modo gli elementi sono rintracciabili e riassemblabili.

Si ottiene ricorrendo ad un elemento di congiunzione (*joint*) che ha la funzione di esprimere l'unità logica di un fenomeno segmentato

È un artificio sintattico che mantiene la conformità con XML

Gerarchie sovrapposte: esempio

<l n="111">Gridò a gran voce:<q id="part1">Messeri pescatelo di grazia,</q></l>
<!-- inizio sovrapposizione -->
<l n="112"> <q id="part2">Ché la sua barba è quasi tutta inzaccherata:</q> </l>
<l n="113"> <q id="part3">O per lo meno reggetegli una scala"</q> </l>
<!-- fine sovrapposizione -->
<join targets="part1 part2 part3" result="q"/>

Un'altra soluzione è quella di segnare le sovrapposizione e le interruzioni con elementi vuoti come <milestone/> o ricorrere a markup esterno o ad altri formalismo *simil XML*

TEI corpus

```
<teiCorpus>
    <teiHeader type="corpus"/>
      <TEI>
        <teiHeader type="text"/>
        <text>…<text/>
      </TEI>
      <TEI>
    <teiHeader type="text"/>
        <text>…<text/>
     </TEI>
</teiCorpus>
```

Biblioteca digitale

Il sistema di annotazione TEI è utilizzato ampiamente da biblioteche digitali e centri di ricerca

Un esempio è Biblioteca Italiana (BibIt - link)

- accoglie testi della tradizione italiana dal Medioevo al Novecento
- codifica XML a partire dalle edizioni cartacee
- documenti sono disponibili in versione HTML, XML e stampa
- set di metadati per la descrizione bibliografica dei testi digitali
- interfaccia di ricerca per autore e titolo
- ricerche contestuali: sezioni testuali, genere, periodo
- ricerche di prossimità: stringhe separate da un numero fissato di tokens
- ricerche full text
- concordanze (kwic): *keyword in context*