Sistemi di codifica dei testi

Matteo Pavesi
ELN2006

Testo

- Lat. *textus*, part.pass. di *temere,* intrecciare.
  - Quintiliano denotò così l'intreccio di connotazioni semantiche fra le parole.
- "enunciato complesso, orale o scritto, considerato un'entità unitaria in base a proprietà particolari quali la compattezza morfosintattica e l'unità di significato"
  - De Mauro

Testo

- Testo-testo
  - Contenuto astratto, enunciato
- Testo-documento
  - Testo + Supporto del testo
  - Aspetti visivi
  - disposizione del testo
  - suddivisione in pagine
  - tipo di supporto

Paratesto

- Paratesto: pratiche discorsive, iconiche e materiali che non sono testo, ma che lo accompagnano, sia spazialmente che cronologicamente.
  - Peritesto
    - Titolo, prefazione,capitoli, note
  - Disposizione tipografica
    - *Coup de dés* di Mallarmé, il *Pantagruel* di Rabelais, il *Tristram Shandy* di Sterne, *Alice nel paese delle meraviglie* di Lewis Carrol.

- Mallarmè
  - Paratesto
  - Disposizione grafica (valore semantico)

Modellizzazione

- COSA?
  - individuare i dati pertinenti (elementi significanti del testo)
- COME?
  - sistema di codifica progettato in funzione della natura fisica del canale e del destinatario
  - Evitare/limitare la perdita di informazione
  - Portabilità

MRF

- Machine Readable Form
  - Testo archiviato su medium informatico
  - Testo interpretabile da computer

Strumenti: codifica dei caratteri

- Codice ASCII: ISO 646
  - 7 bit (2^7= 128)
- Codice Latin-1: ISO 8859-1
  - 8 bit (2^8= 256)
- Codice Unicode: ISO 10646
  - UTF-16 a 16 bit (2^{16}= 65.536)
  - UTF-8 a 8 bit

Strumenti: XML

- Markup Language
  - SGML
  - XML
- Document Type Definition che definisca
  - *Element*s : i marcatori per gli elementi strutturali e funzionali
  - *Content model*: il contenuto di ciascun elemento ovvero quali altri elementi possono apparire all'interno di un certo elemento, con quale ordine e con quale frequenza;
  - *Attribute*s: i marcatori per gli attributi degli elementi
  - *Entity*: i simboli per le entità che possono occorrere come contenuto del documento

TEI

- Text Encoding Initiative
- Standard internazionale e interdisciplinare che permette a librerie, musei, editori e utenti di rappresentare una varietà di testi linguistici e letterari per la ricerca online, l'insegnamento e la conservazione storica.

Tei: base

- Ogni documento Tei sono delimitati da tag <teiHeader>
  - Contiene informazioni sulla digitalizzazione del testo, come revisioni, metodi di codifica e descrizione
- Il testo è contenuto nel tag <text>

Tei: base 2

- All'interno del testo si collocano i tag
  - <front> elementi che precedono il testo (titolo, frontespizio, dediche).
  - <group> raggruppa insiemi di testi distinti ma da codificare unitamente
  - <body>
  - <back> ogni elemento che segue il testo, comprese le appendici

Tei: suddivisione del testo

- Suddivisione del testo
  - Paragrafi (in prosa)
    - <p>
  - Capitoli, sottocapitoli
    - (innestati)
    - <div1>…<div8>
- Attributi:
  - type (libro, capitolo,poema…)
  - Id (identificare il paragrafo o capitolo, per riferimenti)
    - Attributo globale (e importante)
- <Milestone>
  - Attributo generico di suddivisione

Tei:interruzioni

- <pb/> fine pagina
- <lb/> inizio di nuova linea di testo
  - *ed*: attributo (di entrambi) che specifica in quale edizione ci sia l'interruzione
  - possono esserci più attributi pb differenziati da *ed*

Tei: titoli

- <title> contiene il titolo
  - Attributo level:
    - M per titoli monografici (titolo volume)
    - S per titolo collana
    - J per titolo giornale
    - U per titolo materiale inedito (tesi)
    - A per titolo analitico (un poema di un'opera)

Tei: note

- <note> contiene una nota
  - Resp: attributo per il responsabile della notazione
  - Place: la posizione della nota nel testo originale
  - Target: punto di aggancio

Tei: riferimenti

- <ref> è un riferimento ad un'altra parte del testo
  - Contiene testo (anchor)
  - identificata mediante l'attributo *target*
- <ptr> è un riferimento vuoto

Tei: riferimenti estesi

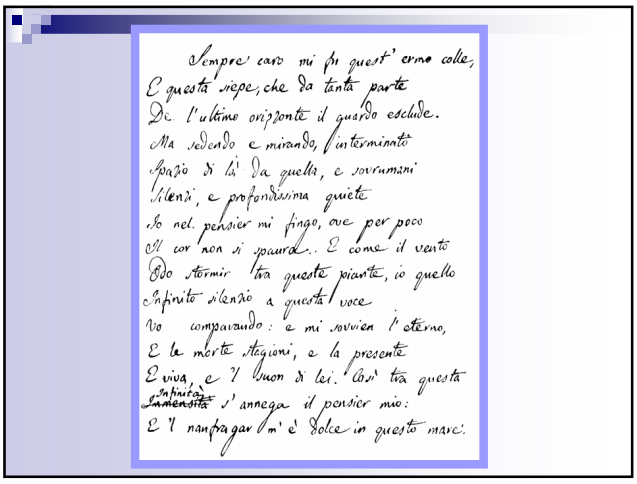
- <xptr>
  - puntatore a un'altra posizione nel documento corrente **o in un documento esterno**.
- <xref>
  - definisce un puntatore ad un'altra posizione nel documento corrente **o in un documento esterno**, eventualmente corredato da testo supplementare o da un commento.
- Attributi
  - Doc: indica l'id del documento

Tei: riferimenti estesi 2

- Per riferire un elemento in un altro testo viene usata la *Sintassi per puntatori estesi TEI*
- Attributi from e to per identificare tag mediante id
- Sintassi basata su valori come:
  - Previous, next, anchestor, child, proceding, following
  - Per esplorare l'albero dei tag Xml
  - <xptr doc='P4' from='id (SA) child (3 p)'>
    - Selezionerà il terzo figlio del tag con id "SA" del documento con id "P4"

Tei: correzioni critiche

- <corr> contiene la forma di un passaggio errato
  - sic: forma originale
- <orig> contiene la forma originale di una lezione
  - reg: forma regolarizzata
- <reg> contiene una lezione che è stata regolarizzata
  - orig: forma originale



Sempre caro mi fu quest' ermo colle,
E questa siepe, che da tanta parte
De l'ultimo orizzonte il guardo esclude.
Ma sedendo e mirando, interminato
Spazio di là da quella, e sovrumani
Silenzi, e profondissima quiete
Io nel pensier mi fingo, ove per poco
Il cor non si spaura. E come il vento
Odo stormir tra queste piante, io quello
Infinito silenzio a questa voce
Vo comparando: e mi sovvien l'eterno,
E le morte stagioni, e la presente
E viva, e il suon di lei. Così tra questa
~~Infinità~~ Immensità s'annega il pensier mio:
E 'l naufragar m'è dolce in questo mare.

Tei: correzioni critiche

E viva, e 'l suon di lei. Così tra questa
<corr sic='immensità' resp='Leopardi'>Infinità</corr> s'annega il pensier mio:
E 'l naufragar m'è dolce in questo mare.

Tei: poesia

- <l> rappresenta un verso
  - Metricamente completo o incompleto
- <lg> rappresenta una unità formale della poesia (stanza, paragrafo in versi, quartina, terzina)

Tei:poesia

<lg>
<l> Spesso il male di vivere ho incontrato</l>
<l> era il rivo strozzato che gorgoglia</l>
<l> era l'incartocciarsi della foglia</l>
<l> riarsa, era il cavallo stramazzato. </l>
</lg>

Tei: testi drammatici

- <sp> singola battuta in un testo drammatico
  - Attributo who: identifica il parlante con un id
- <speaker> contiene l'etichetta del parlante
- <stage> per tutte le didascalie e direttive di scena

Tei: interpretazione

- Interpretazione e analisi (intrinsecamente soggettiva)
  - Frasi ortografiche
  - Elementi di analisi

Tei: interpretazione generica

- <seg> attributo di segmentazione generica del testo
  - Attributo type: può assumere qualsiasi valore
- <seg> viene usato per attribuire ad una porzione di testo una interpretazione arbitraria

Tei: interpretazione generica

- Specializzazione di <seg>
- <s>
  - Identificazione di s-unità: unità delimitate da segni ortografici (interpunzione)
  - Attributo type: DICHIARATIVA, INTERROGATIVA
- <w>
  - Parola grammaticale
- Ecc…

Tei: interpretazione generica 2

- Limite di Xml: nessun overlap
  - <seg> Ma sedendo e mirando, <seg> interminato
    Spazio di là da quella, e sovrumani
    Silenzi, e profondissima quiete</seg>
    Io nel pensier mi fingo</seg> , ove per poco
    Il cor non si spaura.
- Altro limite di Xml: valore singolo
  - Gli attributi type che descrivono l'interpretazione devono essere un valore singolo

Tei: interp

- Lo strumento più generico e potente è il tag <interp>.
  - <interp> non marca parti di testo, descrive una interpretazione da connettere a elementi del testo
  - Si richiama da elementi del testo mediante l'attributo *ana*

Interp: esempio

<interp id='cor-ogg' type='dolore esistenziale' value='correlativo oggettivo' inst='co1' />

…

Spesso il male di vivere ho incontrato

<l>era il **<seg id='co1' ana='cor-ogg'>**rivo strozzato</seg> che gorgoglia</l>

era l'incartocciarsi della foglia

riarsa, era il cavallo stramazzato.

Interp: esempio

<interp id='Np01' type='pos' value='sintagma-nominale, singolare' />

<interp id='vv01' type='pos' value='coniugazione verbo, terza pers. singolare tempo presente>

…

<w ana='Np1'>Alvaro</w>

<w ana='vv01'>pesca</w>

NB: disambigua il verbo dal frutto

NB: il tag <w> è una specializzazione di <seg> per rappresentare una parola grammaticale

OSIS

- **Open Scripture Information Standard**
  - Standard specifico per la codifica dei testi sacri
  - Specializzazione del TEI
- Necessario affrontare il problema dell'overlap
  - Nella bibbia i versi spesso superano i paragrafi

OSIS:differenze

- A titolo esemplificativo.
- Osi introduce il tag <divineName> che delimita tutte le occorrenze di un qualsiasi nome di divinità.
- I nomi di angeli, demoni o simili vengono raggruppati nel tag standard tei <name> usando l'attributo type='nonhuman'

Trojan Milestone

- <q> contiene una citazione o un brano testuale simile ad una citazione.
- E ripeteva <q who='antonio'> bla bla bla </q>
- E ripeteva <q who='antonio' sID='tm1'/> bla bla bla<q eID='tm1'/>

Trojan Milestone

- Tutti gli elementi milestoneable (che permettono contenuto vuoto) supportano questa tecnica
- Le applicazioni XML standard non capiscono che questi segmenti fanno parte di un insieme.
- Il modello è chiamato CLIX

Limite del TEI

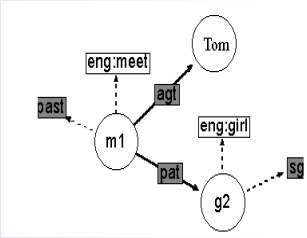
- Ipertrofia del sistema (400 tag)
- Testo-Documento centrica
  - Adatta ad archiviazione
  - Inadatta ad una elaborazione al computer
    - Difficile una analisi semantica
- A livello informatico, preferibile soffermarsi sul Testo-testo

GDA

- Global Document Annotation tag set
  - Descrizione del piano proposizionale
  - Annotazione del testo e della struttura semantica

GDA: strutture semantiche

- Tom met a girl
- Rettangoli: **concept identifier**
- Nodi e frecce ne sono istanze
- I rettangoli in grigio sono **operator identifier**



GDA: codifica

- Tom met a girl

```
<su>
<persnamep opr="agt">Tom </persnamep>
<v sem="past.eng:meet">met <v>
<np opr="obj">
<adp sem="sg">a </adp>
<n sem="eng:girl">girl</n>
</np>.
</su>
```

<su> sentential unit, descrive una frase
<persnamep> contiene un nome di persona
<v> descrive un verbo
<np> descrive un sintagma nominale
<adp> descrive molti elementi linguistici, come avverbi, preposizioni, frasi avverbiali e proposizionali, articoli determinativi e indeterminativi.
<n> descrive un nome generico

GDA:dipendenze

- Opr (e sem)
  - Definiscono le dipendenze standard fra elementi
  - Numerosissimo valori per definire le tipologie di relazione
- Eq
  - Attributo di eguaglianza
  - Valore è un ID di tag
  - <np id="j0">John </np>beats <adp eq="j0">his </adp> wife.

GDA: dipendenza generica

- Attributo dep
  - Valore è l'id di un tag
  - Disambigua l'associazione di un elemento del testo che specifica o descrive un altro elemento all'elemento descritto.
- <su>
- <np>I </np>
- saw
- <np id="m0">a man </np> <np>yesterday </np>
- <adp **dep="m0"**>with a binocular </adp>
- </su>

MathML

- Linguaggio derivato da XML per marcare espressioni matematiche.
- Due livelli di codifica
  - marcatura di presentazione: ricalca la struttura bidimensionale e visibile della notazione matematica.
  - marcatura di contenuto: codifica della struttura matematica sottostante di un'espressione.

MathML: esempio

- $(a+b)^2$
- Codifica di presentazione

```
<msup>
<mfenced>
<mrow>
<mi>a</mi>
<mo>+</mo>
<mi>b</mi>
</mrow>
</mfenced>
<mn>2</mn>
</msup>
```

MathML: codifica di presentazione

- <mrow> rappresenta una espressione matematica piana
- <msup> rappresenta una espressione ad apice, composta da un espressione base ed una espressione a potenza
- <mfenced> rappresenta una espressione parentesizzata (default: tonde)
- <ci> sono variabili
- <co> è l'operatore

MathML: esempio

- Codifica di contenuto

```
<apply>
<power/>
<apply>
<plus/>
<ci>a</ci>
<ci>b</ci>
</apply>
<cn>2</cn>
</apply>
```

MathML: esempio

- <apply> è il tag che indica l'esecuzione di una funzione
- <power/> e </plus> rappresentano la funzione
- Albero di sintassi astratta

Bibliografia

- http://crilet.scu.uniroma1.it/ciotti/pubblicazioni/tes_el.htm
  - Testi elettronici e banche dati testuali: problemi teorici e tecnologie,
  - Fabio Ciotti
- http://www.tecnoteca.it/howto/marcaturaxml/testi
  - I testi in formato elettronico e i linguaggi di codifica
  - Pietro Bortoluzzi
- http://www.tei-c.org/Lite/teiu5_it.xml.ID=U5-it-pref
  - TEI Lite: introduzione alla codifica dei testi
  - Fabio Ciotti
- http://www.bibletechnologies.net/
  - OSIS website
- http://www.mulberrytech.com/Extreme/Proceedings/html/2004/DeRose01/EML2004DeRose01.html
  - Markup Overlap: A Review and a Horse
  - Steven DeRose
- http://www.w3c.it/traduzioni/MathML2/overview.html
  - MathML
- http://infouma.di.unipi.it/corsi/Pierazzo/home.html
  - Corso di Codifica di Testi, università di Pisa.
  - Elena Pierazzo
- http://i-content.org/GDA/
  - GDA