

# Linguistica Computazionale

---



## La codifica di alto livello del testo

Salvatore Sorce

Dipartimento di Ingegneria  
Chimica, Gestionale, Informatica e Meccanica

Lucidi Adattati da Alessandro Lenci  
Dipartimento di Linguistica "T. Bolelli"



Informatica**U**manistica

# Codifica di alto livello

## *Interpretazione e codifica*

---

- **Interpretazione del testo**
  - informazioni che caratterizzano la struttura, contenuto, presentazione, natura linguistica, ecc. di un testo e del suo contenuto informativo
  - esistono vari livelli e gradi di interpretazione
    - tipografica, extratestuale, linguistica, ecc.
  - **metadati**
    - “informazione sull’informazione”
- **Codifica di alto livello**
  - processo attraverso cui viene **resa esplicita un’interpretazione del testo**
  - la codifica permette di **rendere machine readable informazioni sul testo e tratti del testo** che altrimenti non sarebbero elaborabili dal computer

# Perché codificare?

## *la gerarchia dell'informazione*

I dati non hanno un significato intrinseco a meno di non inserirli in uno schema o **struttura** che li organizza e li trasforma in **informazione**



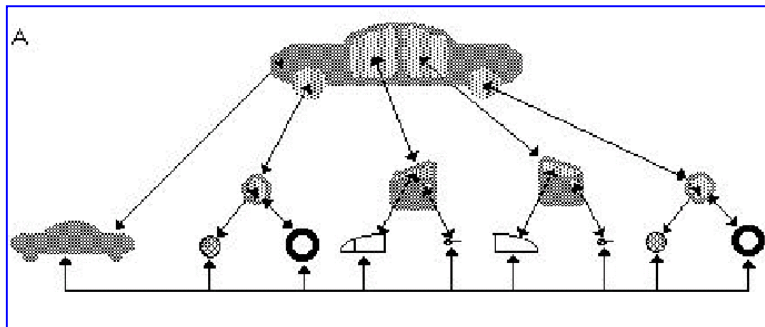
# La gerarchia dell'informazione

benzina, 4 cil. in linea  
1.997 cc  
130 kW (180 CV)  
02 Nm (20,6 kgm)  
manuale a 5 rapporti  
anteriore  
205/50 R 17  
4,35/1,76/1,42 m

dati di un'auto

**Motori:** benzina, 4 cil. in linea  
**Cilindrata:** 1.997 cc  
**Potenza max:** 130 kW (180 CV)  
**Coppia max:** 02 Nm (20,6 kgm)  
**Cambio:** manuale a 5 rapporti  
**Trazione:** anteriore  
**Pneumatici:** 205/50 R 17  
**Dimensioni:** 4,35/1,76/1,42 m

informazione su un'auto



conoscenza sulle auto  
(struttura, funzionamento,  
tipologie, ecc.)

# Perché codificare?

## *i motivi della codifica di alto livello*

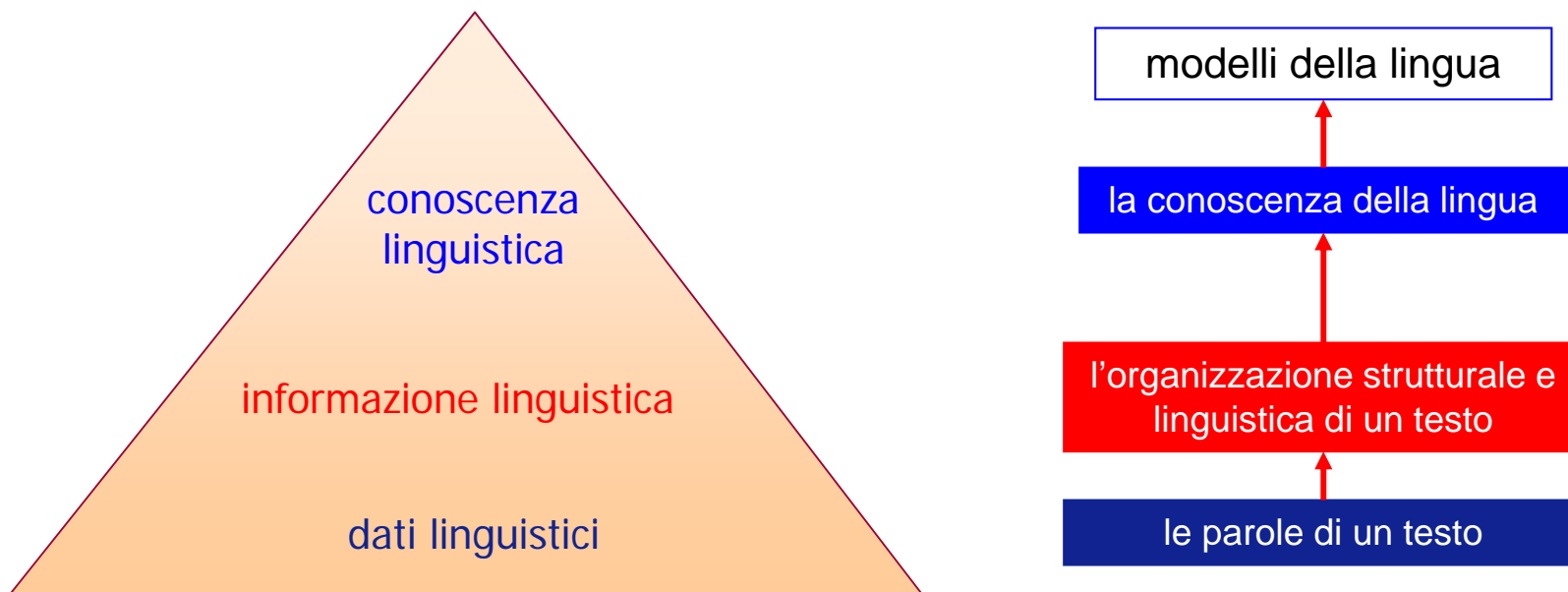
---

- Un testo come flusso di caratteri e parole è una fonte di **dati linguistici**
- Il testo è un'entità altamente strutturata, nella quale i dati linguistici sono correlati secondo piani di organizzazione multipli
  - **struttura del testo**
    - l' articolazione in sezioni, capitoli, titoli, ecc.
  - **struttura del contesto**
    - l'autore, la data di produzione, la finalità del testo, ecc.
  - **struttura linguistica** (*implicita nel testo!!*)
    - informazioni morfologiche, sintattiche, semantiche, ecc.

# Perché codificare?

*La gerarchia dell'informazione linguistica*

---



La codifica di alto livello permette di rendere **espliciti** e **accessibili al computer** i livelli di organizzazione strutturale di un testo e lo trasforma in una **fonte di informazione linguistica**

# Cosa codificare?

## *i contenuti della codifica di alto livello*

---

- Individuare il livello di informazione da codificare
  - strutturale, linguistica, ecc.
  - la codifica esplicita di informazione linguistica viene detta **annotazione del testo**
    - morfologica, semantica, sintattica, ecc.
- Definire il **repertorio dei tratti giudicati rilevanti per la codifica**
  - un esempio: *la codifica morfo-sintattica*
    - oggetto: codificare esplicitamente la categoria grammaticale e le proprietà morfologiche delle parole di un testo
    - da definire:
      - **quali attributi codificare** (cat. grammaticale, persona, genere, numero, caso, ecc.)
      - **quali valori possono avere i diversi attributi** (numero = SING, PLUR; caso = NOM, GEN, ecc.)

# Cosa codificare?

## *i contenuti della codifica di alto livello*

---

- **Schema di codifica**
  - un **repertorio di categorie per la codifica**, corrispondenti alla tipologia dei tratti da rappresentare nel testo
    - generalmente espresso nella forma di attributi e dei loro possibili valori
  - la **definizione delle regole** di compatibilità tra categorie
    - es. l'aggettivo non possiede un attributo di persona, o un nome quello di tempo
  - la **specificazione accurata dei criteri di applicazione al testo** delle categorie selezionate
- **Schema di annotazione linguistica**
  - schema di codifica di informazione linguistica

Uno schema di codifica (annotazione) definisce il **contenuto** linguistico dell'annotazione, non il **modo** in cui la codifica (annotazione) è proiettata sul testo



# Come codificare?

## *i formati digitali del testo*

---

- **Formato solo testo** (plain text o txt)
  - un file solo testo è costituito da una sequenza di bytes dove ciascun byte rappresenta **un carattere** secondo un particolare codice
  - gli **editori di testo** sono programmi in grado di creare e leggere files di tipo solo testo
    - Emacs, Blocco Note, Word Pad, ecc.
      - » quando un editore di testo legge un file cerca di associare ogni **sequenza di bits a un carattere secondo un particolare codice**
- **Vantaggi**
  - formato “aperto”, indipendente dal sistema operativo e dal programma che lo ha creato
  - massima portabilità e interscambiabilità dei testi
- **Svantaggi**
  - non può rappresentare aspetti relativi alla codifica di alto livello
    - minima espressività

# Come codificare?

## *i formati digitali del testo*

---

- **Formati proprietari** (doc, pdf, ecc.)
  - possono essere creati, letti e interpretati solo da uno specifico programma (es. Word, Adobe)
  - oltre a sequenze di bits codificano caratteri, il file contiene sequenze binarie che corrispondono a **istruzioni di formattazione**, codificate secondo le convenzioni di un certo programma
- **Vantaggi**
  - massima capacità espressiva e fruibilità per l'utente umano
    - rappresenta aspetti relativi alla codifica di alto livello
      - » ma solo per quanto riguarda la struttura testuale!!
- **Svantaggi**
  - formato “chiuso”, con minima portabilità e interscambiabilità
  - codifica non per categorie testuali “astratte”, ma per modalità di visualizzazione
  - le informazioni linguistiche rimangono comunque implicite nel testo

# Come codificare?

## *i linguaggi di marcatura*

---

- **Codifica di alto livello con linguaggi di mark-up** (linguaggi di marcatura) come XML
  - dal punto di vista del formato digitale un testo codificato in XML è in **formato solo testo**
  - l'informazione strutturale è rappresentata attraverso l'aggiunta al testo di **etichette** (o **tag**) di marcatura
    - sequenze di caratteri visibili secondo una convenzione standard, intercalati nel testo seguendo precise regole di combinazione
    - “marcano” blocchi di testo a cui viene assegnata una determinata interpretazione
  - **codici in formato testo** vengono usati per specificare informazioni sul testo
    - *il testo e i suoi metadati sono entrambi in formato “plain text”*
  - **Vantaggi**
    - portabilità e interscambiabilità dei testi codificati
    - massimo grado di espressività
      - è possibile esprimere tutti gli aspetti della codifica di alto livello, compresa l'informazione linguistica

# Come codificare?

## *i linguaggi di marcatura*

mark-up per la codifica di informazione strutturale

```
<libro>
<titolo>Le avventure di Pinocchio
<sottotitolo>Storia di un burattino</sottotitolo>
</titolo>
<autore>Carlo Collodi</autore>
<parte p_id="1">
<titolo>Parte prima</titolo>
<capitolo c_id="I">
<titolo> Come andò che maestro Ciliegia, falegname, trovò un pezzo
di legno, che piangeva e rideva come un bambino.</titolo>
<capoverso num="plc1c1">C'era una volta...</capoverso>
<capoverso num="plc1c2">- Un re! - diranno subito i miei piccoli
lettori.</capoverso>
<capoverso num="plc1c3">No, ragazzi, avete sbagliato. C'era una
volta un pezzo di legno.</capoverso>
<capoverso num="plc1c4">Non era un legno di lusso, ma un semplice
pezzo da catasta, di quelli che d'inverno si mettono nelle stufe e
nei caminetti per accendere il fuoco e per riscaldare le
stanze.</capoverso> </capitolo></parte>
</libro>
```

# Come codificare?

## *I linguaggi di marcatura*

---

mark-up per la codifica di informazione strutturale

+

mark-up per la codifica di informazione linguistica

```
<libro>
...
<parte>
<capitolo c_id="I">
<titolo>Come <parola cat="V" tempo="passRem">andò</parola> che
<parola cat="N" genere="m" num="s">maestro</parola> Ciliegia,
falegname, trovò <sintagma tipo="nominale"><parola cat="artInd"
genere="m" num="s">un</parola> pezzo di legno</sintagma>, che
piangeva e rideva come un bambino.</titolo>
</capitolo></parte>
</libro>
```

# Possibili Domande di Esame

---

- Il candidato risponda alle seguenti affermazioni con **vero** o **falso**:
  - I Computer riescono ad immagazzinare i dati in qualsiasi forma **FALSO**
  - La codifica di alto livello permette di rendere esplicita qualsiasi interpretazione associata al testo **VERO**
  - La codifica serve a rendere espliciti i dati linguistici presenti in un testo trasformandoli in informazioni **VERO**
  - Uno schema di codifica (annotazione) definisce il contenuto linguistico dell'annotazione **VERO**
  - Una codifica effettuata mediante linguaggi di mark-up non permette portabilità e interscambiabilità dei testi codificati **FALSO**