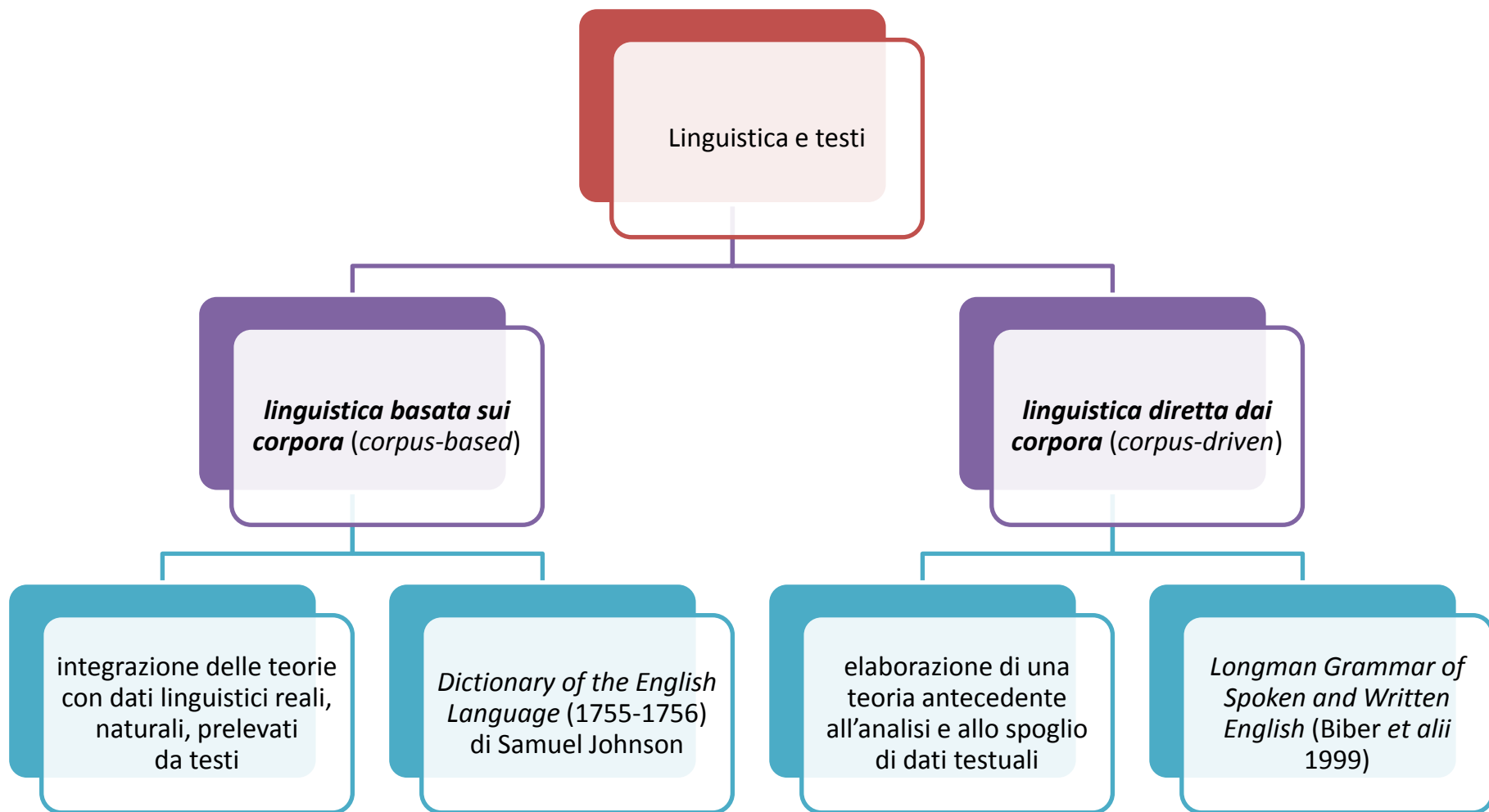


# Linguistica dei corpora

## Analisi del testo letterario 1

# *Elena Tognini Bonelli (2001)*



# Corpus (plur. corpora)

«raccolta completa e ordinata di scritti, di uno o più autori, riguardanti una certa materia» (De Mauro, GRADIT)

«campione di una lingua preso in esame nella descrizione di una lingua» (De Mauro, GRADIT)

## TESTI

- opere di Alessandro Manzoni
- lettere d'amore,
- atti giudiziari
- perizie psichiatriche
- testi di telefonate

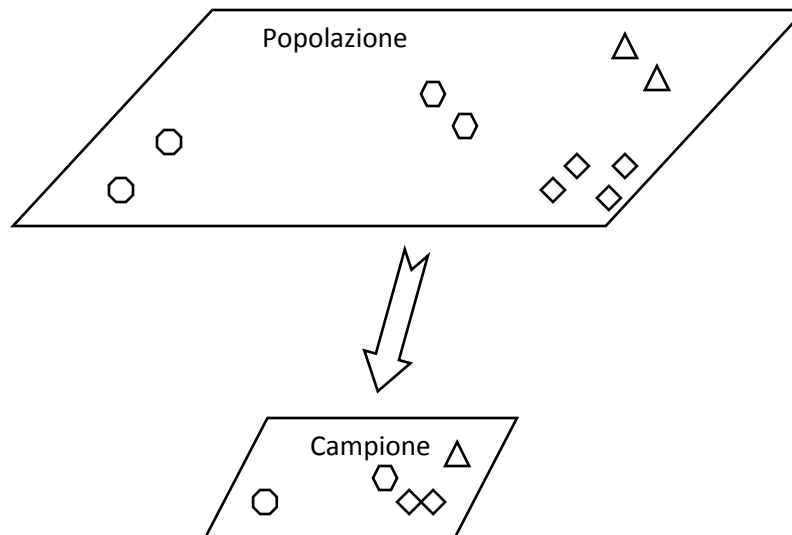
## SCOPI

- usare le osservazioni condotte su un corpus campionario per estenderle all'intera popolazione
- comparare le osservazioni condotte su diversi corpora per confrontarle infine con un corpus di riferimento, individuandone le deviazioni

# «Popolazione» e «Campione»

Una *popolazione* è un insieme di tutte le possibili osservazioni di un tipo su un dato campo

Un *campione*, invece, è una sezione, una parte della popolazione, che include solo alcune delle possibili osservazioni



# *La rappresentatività*

Il campione deve, per l'aspetto che si intende studiare, essere atto a esibire lo stesso tipo di informazioni (**qualitative**) con la stessa probabilità di occorrenza (**quantitativa**) della popolazione

La rappresentatività è una caratteristica **relativa**

- varia secondo l'aspetto linguistico che si intende studiare
- un corpus rappresentativo per caratteristiche lessicali potrebbe non esserlo per caratteristiche di tipo sintattico oppure stilistico

Un campione non è mai comunque «di per sé» rappresentativo

# *L'estensione*

L'estensione è una variabile che influenza il grado di rappresentatività di un campione testuale

Esistono diverse estensioni standard a seconda del livello di analisi linguistica obiettivo del design del corpus stesso

- per le analisi di tipo lessicale, di gran lunga le più frequenti condotte su corpora, si sono individuate soglie indicative minime per determinare un'estensione ragionevole per i corpora

Un indicatore globale più agevole può essere considerato anche il numero di occorrenze (*token*) di parole grafiche presenti nel corpus

# *Estensione corpora per analisi lessicali*

Corpus non rappresentativo (insuff.)	<i>&lt; 15.000 parole grafiche</i>
Corpus piccolo	<i>Da circa 15.000 a 100.000 parole</i>
Corpus medio	<i>Da circa 100.000 a 1 milione di parole</i>
Corpus medio-grande	<i>Da circa 1 milione a 50 milioni di parole</i>
Corpus standard	<i>Da circa 50 milioni a 100 milioni di parole</i>
Corpus grande	<i>Oltre i 100 milioni di parole</i>

# *Estensioni di alcuni corpora di riferimento*

## *Brown Corpus (1961)*

- 1 milione di occorrenze

## *LIF, Lessico di frequenza della lingua italiana contemporanea (1971)*

- 500.000 occorrenze

## *British National Corpus*

- 100 milioni di occorrenze

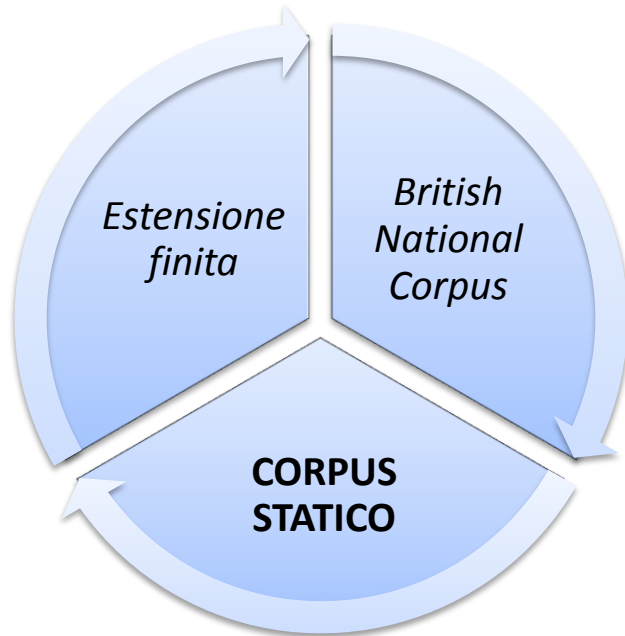
## *CORIS, Corpus di italiano scritto contemporaneo (CORIS)*

- 100 milioni di occorrenze

## *Bank of English*

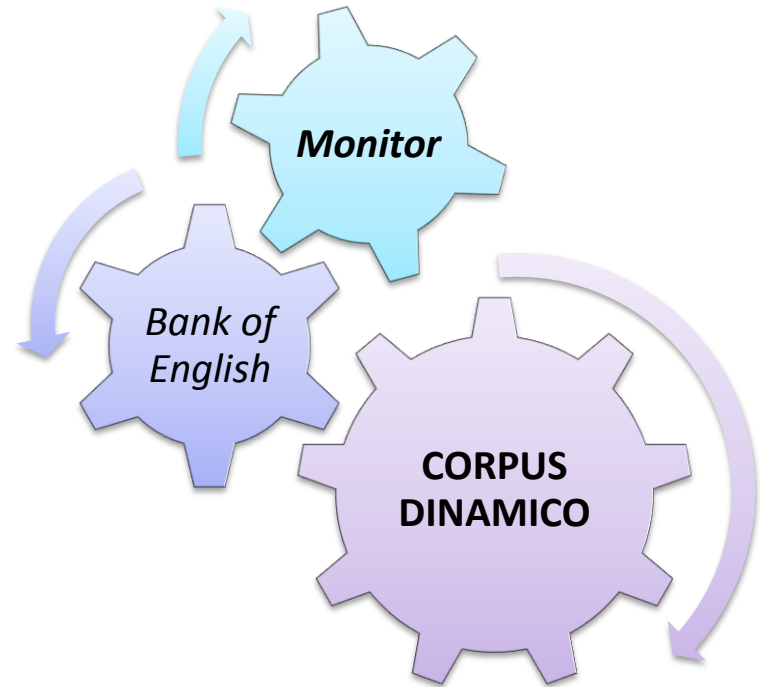
- Circa 500 milioni di occorrenze

# Corpora statici e corpora dinamici



## Vantaggi

- analisi finite e ripetibili
- comparabilità



## Vantaggi

- aggiornamento
- analisi diacroniche

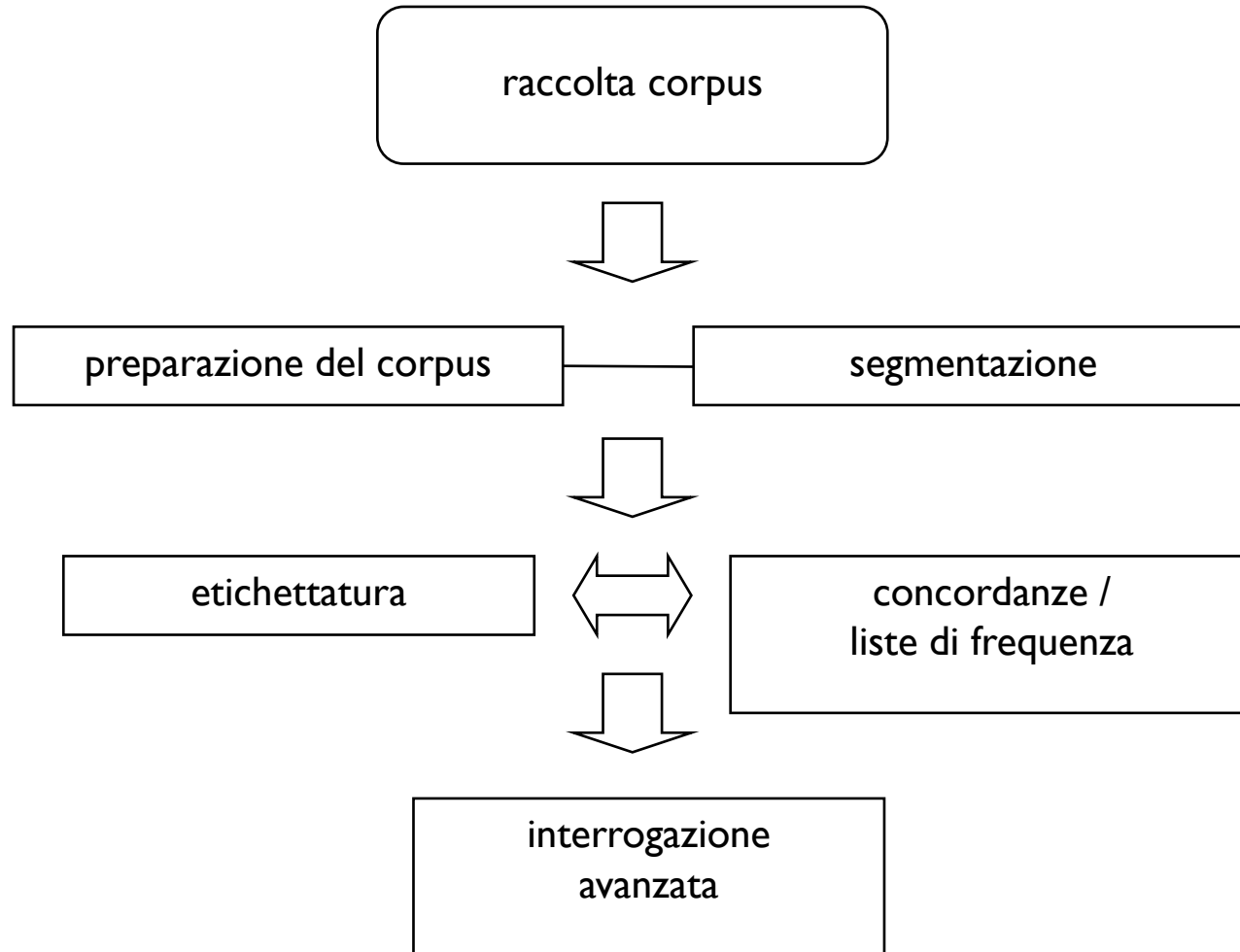
## *Formato elettronico (machine-readable form)*

- trattamento informatizzato dei dati testuali
- archiviazione dei testi in forma digitale
- interrogabile

## *Riferimento standard*

- punti di riferimento per lo studio della varietà che rappresentano
- mediante l'esplicitazione delle metodologie di analisi
- facilitando la comparazione tra corpora diversi

# *La costruzione di un corpus elettronico*



# *Le tappe (1)*

Design del corpus



*Acquisizione del materiale*

biblioteche  
digitali

cd-rom

digitazione

dettatura

scannerizzazione



Correzione degli errori

# Le tappe (2)

Preparazione del corpus (pre-trattamento)

Individuazione di ALFABETO E SEPARATORI (.,:;/?!) 

Etichettatura e annotazione 

Interrogazione

Esplorazione con liste di  
frequenza

Concordanze

Ricerca avanzata

# *L'acquisizione del materiale*

## Testi scritti disponibili in formato digitale

- testi scritti letterari
- testi giornalistici
- biblioteche digitali online o su cd-rom

## Testi scritti non disponibili in formato digitale

- digitazione
- dettatura
- scannerizzazione

## Testi parlati

- acquisizione del segnale audio (analogico o digitalizzato)
- standardizzazione delle procedure di trascrizione del parlato (CES, Eagles, Chat)

# *Gestione degli errori*

**Tutti i corpora contengono errori**

Lapsus, sviste, errori di battitura, ambiguità e forme disomogenee di trascrizione

Correttore ortografico

Correzione manuale o semi-automatica

# I corpora di riferimento

delle principali lingue europee  
moderne

# Un corpus di riferimento

## *Reference corpus*

- testi appartenenti a diverse varietà sociolinguistiche, diafasiche e diatopiche

Mira a rappresentare «la lingua», non una sua varietà

## Standard di estensione

- da 500.000 occorrenze
- a 500 milioni

# «*Brown Corpus of Standard American English*»

W. N. Francis e H. Kučera, della Brown University

1961

## Corpus di lingua scritta

- primo corpus linguistico elettronico dell'inglese americano
- corpus più usato nella ricerca
- lessico di frequenza abbinato

## Composizione

- 500 testi
- ciascun testo è composto da 2000 parole (*sample corpus*)
- 15 categorie testuali diverse
- un totale di un milione di parole

## Sistema di trascrizione e annotazione proprio

- programma automatico CLAWS dell'Università di Lancaster
- L'etichettatura rispetta le *Guidelines* della [Text Encoding Initiative](#) (TEI)

# «*British National Corpus*»

## *Oxford University Press*

- interrogabile dal sito di Mark Davies: <http://view.byu.edu/>
- raccolta 1991 – uscita 1995

## Lingua parlata e lingua scritta

- inglese contemporaneo

100.106.008 parole

- 4.124 testi

## Software di interrogazione SARA

## Composizione

- 4.124 testi
- 90% deriva da testi scritti
  - romanzi e saggi, e testi tecnico-scientifici
- 10% da trascrizioni di parlato
  - 863 testi
  - programmi radiofonici, conversazioni telefoniche, parlato spontaneo

# «British National Corpus»

File Modifica Visualizza Cronologia Segnalibri Strumenti 2

http://view.byu.edu variations in english

Alphabit.net Literary Encyclopedia: ... Glottophilia Live Mail TuttoCittà DMP Forum DelCamp Forum Porcheddu Catemario Forum GSCP FSU Danilo Cartia

SEARCH STRING (HELP)  
WORD/PHRASE  
(INSERT TAG) -select- CUSTOMIZED LISTS

DISPLAY (HELP)  
 TABLE  
 CHART  
 SURROUNDING (HELP)  
-select- 5 5

SORT BY (HELP)  
 FREQUENCY  
 PERCENT

REGISTER 1 (HELP)  
-- IGNORE -- MIN  
SPOKEN FREQ  
FICTION 5  
NEWS  
ACADEMIC LIMIT?  
NON-FICTION MISC  
OTHER MISC

REGISTER 2  
-- IGNORE -- MIN  
SPOKEN FREQ  
FICTION 5  
NEWS  
ACADEMIC LIMIT?  
NON-FICTION MISC  
OTHER MISC

OPTIONS (HELP)  
# HITS 100  
SEE POS TAGS NO

SEARCH RESET

## VIEW: VARIATION IN ENGLISH WORDS AND PHRASES

Mark Davies / Brigham Young University

(Site optimized for 1024x768)  
Fonts too large/small?

### OVERVIEW: INTRODUCTION

More information...

This website allows you to quickly and easily search for a wide range of words and phrases of English in the **100 million word British National Corpus**. As with some other BNC interfaces, you can search for words and phrases by **exact word** or **phrase**, **wildcard** or **part of speech**, or combinations of these. You can also **search for surrounding words** (collocates) within a ten-word window (e.g. all nouns somewhere near *paper*, all adjectives near *woman*, or all nouns near *spin*).

One unique aspect of the corpus is the ability to find the frequency of words and phrases in any combination of **registers** that you define (spoken, academic, poetry, medical, etc). In addition, you can **compare between registers** -- for example, verbs that are more common in legal or medical texts, or nouns near *break* that are more common in fiction than in academic writing.

Completato

Adesso: Generalmente coperto, 12° C Lun: 12° C Mar: 15° C

# «*Bank of English*»

Diretta dal linguista John Sinclair

Corpus dinamico di testi scritti e parlati in inglese britannico

- con monitor corpus

Circa 500 milioni di occorrenze

Obiettivi lessicografici

- il progetto procede insieme al lavoro lessicografico del *Collins Cobuild English Dictionary for Advanced Learners* (2001) e dell'Università di Birmingham

Annotazione

- **ENGTWOL lexical analyser**
  - Statistical tagger
  - Error rate 0,5%

# LIP – «Corpus del Lessico di frequenza dell'italiano parlato»

A cura di Tullio De Mauro, Federico Mancini, Massimo Vedovelli e Miriam Voghera (1993)

57h di registrazione di parlato (1990-1992)

- 475.883 parole grafiche
- 496.335 occorrenze di lemmi

*Rappresentatività geografica*: Milano, Firenze, Roma e Napoli: ogni città 125.000 occorrenze

Tipologie testuali

- 1) scambio *bidirezionale faccia a faccia* con presa di parola **libera**
- 2) scambio *bidirezionale non faccia a faccia* con presa di parola **libera** (conversazioni telefoniche)
- 3) scambio *bidirezionale faccia a faccia* con presa di parola **non libera** (dibattiti, interviste, interrogazioni)
- 4) scambio *unidirezionale in presenza di destinatario/i* (lezioni, conferenze, omelie, comizi, ecc.)
- 5) scambio *unidirezionale o bidirezionale a distanza* (trasmissioni radiofoniche e televisive)

Interrogazione

- sito BADIP (banca dati dell'italiano parlato)
- [http://languageserver.uni-graz.at/badip/badip/20\\_corpusLip.php](http://languageserver.uni-graz.at/badip/badip/20_corpusLip.php)

# Interrogazione BADIP

Microsoft Internet Explorer fornito da FastWeb

Indirizzo [http://languageserver.uni-graz.at/badip/badip/24\\_genSearch.php](http://languageserver.uni-graz.at/badip/badip/24_genSearch.php)

## badip

banca dati dell'italiano parlato

[home](#) [corpus](#) [lip](#) [collaboratori](#) [consulenti](#) [contatto](#) [sponsor](#) [lista di corpora](#)

ricerca  
testi  
tipologia dei testi  
classi di parola  
simboli e notazioni  
durata delle registrazioni  
parlanti

Cerca tutte le sequenze che contengano: (aiuto)

digita la prima parola oppure clicca

digita la seconda parola oppure clicca

digita la terza parola oppure clicca

e che non contengano:

digita la prima parola

digita la seconda parola

digita la terza parola

nelle città:  Firenze  Milano  Napoli  Roma

nei tipi di testo:  A  B  C  D  E

© BADIP  
ultima modifica 20/04/2005 11:47

Operazione completata

Internet

# *Il CORIS/CODIS*

*Corpus di riferimento dell'italiano scritto (CORIS)*

*COorpus dinamico dell'italiano scritto (CODIS)*

- CILTA (Centro interfacoltà di linguistica teorica e applicata “Luigi Heilmann”, Bologna)
- a cura di R. Rossini Favretti
- (1998)

100 milioni di parole

- aggiornato tramite un corpus di monitoraggio con cadenza biennale
- testi: prevalentemente di narrativa prodotta negli anni Ottanta e Novanta

Accesso

- [http://corpus.cilta.unibo.it:8080/coris\\_ita.html](http://corpus.cilta.unibo.it:8080/coris_ita.html)

# CODIS

http://corpus.cilta.unibo.it:8080/CODISCorpQuery.html - Microsoft Internet Explorer fornito da FastWeb

Indirizzo http://corpus.cilta.unibo.it:8080/CODISCorpQuery.html

**User Authentication**

Username   
Password

**Query**

[Query Language Help.](#)  
 Case insensitive search

**Subcorpora selection**

Subcorpus	Size (in Mw)			
STAMPA	<input type="checkbox"/> 20	<input type="checkbox"/> 10	<input type="checkbox"/> 5	<input type="checkbox"/> 3
NARRATIVA	<input type="checkbox"/> 13	<input type="checkbox"/> 7	<input type="checkbox"/> 3	<input type="checkbox"/> 2
PROSA ACCADEMICA	<input type="checkbox"/> 5	<input type="checkbox"/> 4	<input type="checkbox"/> 2	<input type="checkbox"/> 1
PROSA GIURIDICO-AMM.	<input type="checkbox"/> 4	<input type="checkbox"/> 3	<input type="checkbox"/> 2	<input type="checkbox"/> 1
MISCELLANEA	<input type="checkbox"/> 4	<input type="checkbox"/> 3	<input type="checkbox"/> 2	<input type="checkbox"/> 1
EPHEMERA	<input type="checkbox"/> 2	<input type="checkbox"/> 1	<input type="checkbox"/> 1	<input type="checkbox"/> 1

**Concordance Options**

Reduce to max  30  100  300 lines.  1 every n-th  Random

Unsorted

Context of  80  120  160 characters.

**Collocations**

Get Collocates?  NO!  Yes, before reduction.  Yes, after reduction.

Sort using  Mutual Information.  T-score.  Raw frequency.

Operazione completata

Internet

# *I corpora multilingui e paralleli*

## *Scopi*

- facilitare la costruzione di risorse didattiche, sistemi di traduzione, basi dati terminologiche, dizionari elettronici, ecc.

## *Corpora paralleli*

- costituiti da testi originali in una lingua (SL, *source language*) e da traduzioni di questi testi in una o più altre lingue (TL, *target language*)
- allineamento

## *Corpora multilingui*

- i testi non sono in traduzioni reciproche, ma vertono su ambiti disciplinari corrispondenti permettendo così la costituzione di banche dati terminologiche
- linguaggi settoriali come linguaggio giuridico, economico, commerciale

## *Esempi*

- BAF (French-English Parallel Corpus)
- progetto MULTEX (Multilingual Text Tools and Corpora)
- progetto CHILDES (Child Language Data Exchange System)

# ParaConc

The screenshot shows the ParaConc software window titled "ParaConc - Freneng2 - [Parallel Concordance - [head]]". The menu bar includes File, Search, Frequency, Display, Sort, Window, and Info. The main text area is divided into two sections. The top section displays an English concordance for the word "head", with the following text: "vered with socks of heavy raw wool, his head covered with a narrow short cheche. Th ... .. the cold earth, and he did not turn his head. But after a short time, he turned aro ... .. climb. Not once did the Arab raise his head. "Hello," said Daru, when they came up ... .. waited motionless, without turning his head toward Daru, as if he was listening ver ... .. ce. "Listen," he said. Daru shook his head. "No, say no more. Now I'm leaving yo ... .. he wall like a hunting trophy. And the head was alive. Through the pince-nez, on i ... .. room. She glanced at his white-swathed head and blue goggles again as she was going .. .. destrian, we feel that he supposes this head to be endowed with independent life, su ... .. conceivable. This muffled and bandaged head was so unlike what she had anticipated, ..". The second section displays a French concordance for the word "tête", with the following text: "de chaussettes en grosse laine grege, la tête coiffée d'un chèche 2 étroit et court. ... .. terre froide, et il ne détourna pas la tête. Au bout d'un moment, pourtant, il se ... .. seule fois, l'Arabe n'avait levé la tête. "Salut, dit Daru, quand ils débouchèrent s .. .. e lit, il attendit, immobile, sans tourner la tête vers Daru, comme s'il écoutait de toute ... .. on visage. "Écoute," dit-il. Daru secoua la tête: "Non, tais-toi. Maintenant, je te laisse. "I ... .. façon d'un trophée de chasse. Et cette tête était vivante. A travers le lorgnon à chaîne ... .. lança un dernier coup d'il vers cette tête emmaillotée de blanc, vers ces lunettes s ... .. t effaré, nous sentons qu'il suppose cette tête douée d'une vie particulière, soumise à s ... .. plus étrange que l'on pût imaginer. Cette tête, enveloppée, emmitoufiée, était si différen ... .. trouvent assises à de fortes rotations de tête vers la gauche, si elles veulent apercevoi ... .. apporte le calvados. D'un mouvement de tête, elle indique au docteur son voisin. Le doc ...".

# La codifica e l'etichettatura dei corpora

## Standard di codifica e modalità di annotazione

# Etichettatura

L'annotazione o etichettatura linguistica di un corpus è l'aggiunta di informazioni di tipo linguistico (o meglio metalinguistico) alle diverse porzioni di un testo

- l'annotazione è una forma di codifica (esistono diversi tipi di annotazione non linguistica)

L'annotazione consiste nell'attribuzione di una **etichetta** (*tag* o *mark-up*) a una porzione specifica e limitata di testo

- qualunque aspetto dell'analisi linguistica può essere etichettato (fonologia e fonetica, morfologia, sintassi, semantica, pragmatica, testo, ecc.)

linguaggio di marcatura (*markup language*)

- SGML (Standard Generalized Markup Language)
- XML (Extensible Markup Language)
- HTML (HyperText Markup Language)

# Pratiche di annotazione

Annotazione morfo-sintattica, detta anche *grammatical tagging* o spesso POS (*part-of-speech*) tagging

- a ogni *word token* viene associata la relativa categoria grammaticale (nome, verbo, aggettivo, ecc.)
  - *Palla* > N
  - *Tornò* > V

## Disambiguazione

- per tutti quei casi in espressioni linguistiche che sono passibili di diverse letture (morfologiche, morfo-sintattiche, sintattiche e lessicali, ecc.)
- *folle*
  - aggettivo maschile singolare (*l'uomo folle*) oppure sostantivo femminile plurale (*le grandi folle*)

## Omografi

- De Mauro (1994) riferisce per l'italiano un tasso di omografi che varia dal 38% al 46%, rispettivamente in testi economico-finanziari e nella lingua parlata, ma stime diverse giungono per alcuni tipi testuali sino a circa 57%

# Metodi di annotazione

## Annotazione manuale

- di tipo tradizionale è svolta da persone che appongono le specifiche etichette alle porzioni di testo sulla base di valutazioni metalinguistiche più o meno condivise e standardizzate

## Annotazione automatica

- procede senza l'intervento umano, attraverso applicazioni del *Natural Language Processing* basate su regole (*rule-based parsing*) oppure su sistemi probabilistici (*statistical parsing*)

## Annotazione semi-automatica

- I fase automatica
- II fase manuale

# *Gli standard di codifica e annotazione linguistica*

## Perché?

- necessità e utilità della comparazione di corpora diversi
- usi dei corpora in ambito scientifico e commerciale

## Requisiti di standardizzazione

- separazione e autonomia del materiale grezzo del corpus dalle codifiche e annotazioni linguistiche
- esplicitazione di tutte le fasi di predisposizione, standardizzazione e annotazione del corpus in un file di documentazione
- forma standard per la codifica e l'annotazione
- indicazione esplicita dei criteri di annotazione (regole di attribuzione di un'etichetta a un *token*)
- neutralità e condivisione generale dei criteri linguistici alla base dell'annotazione
- possibilità di eseguire specifiche ricerche sulle annotazioni
- indipendenza della fruibilità del corpus e dell'annotazione dagli specifici sistemi operativi e da costrizioni sulle caratteristiche dell'hardware

# *Text Encoding Initiative (TEI)*

consorzio nato ufficialmente nel 2000

- anche se già operante a tutti gli effetti dal 1987
  - ACL – Association for Computational Linguistics
  - ALLC – Association for Literary and Linguistic Computing
  - ACH – Association for Computers and the Humanities

Linee guida - TEI *Guidelines*

- per i principali problemi connessi alla codifica e all'annotazione di corpora
- formato di marcatura dei documenti SGML
- codifiche di caratteri ISO 646 (a 7-bit) e Unicode

Documenti

- «testi di ogni lingua, ogni tempo, ogni genere letterario o tipologia testuale, senza restrizioni su forma o contenuto» (Sperberg-McQueen e Burnard 2002)

# Le etichette

È necessario segnalare l'inizio e la fine della porzione testuale da marcare

- **marca di inizio** (*start tag*): è un nome tra parentesi uncinate `<name1>`
- **marca di fine** (*end tag*): è come la marca di inizio, preceduta dal segno / `</name1>`
  - *Paragrafo*
    - `<p> xyz yzxx </p>`
  - *le marche possono anche essere annidate*

## Tipologie di etichetta

- **obbligatorie**
- **raccomandate**
- **opzionali**
  - e infine:
- **personalizzate** dall'utente
- il testo è **suddiviso** in unità testuali (dipendenti dal tipo di etichettatura che si intende effettuare)

# Esempio di marcatura

```
<anthology>
  <poem><title>The SICK ROSE</title>
    <stanza>
      <line>0 Rose thou art sick.</line>
      <line>The invisible worm,</line>
      <line>That flies in the night</line>
      <line>In the howling storm:</line>
    </stanza>
    <stanza>
      <line>Has found out thy bed</line>
      <line>0f crimson joy:</line>
      <line>And his dark secret love</line>
      <line>Does thy life destroy.</line>
    </stanza>
  </poem>
```

Da Sperberg-McQueen e Burnard 2002, § 2.3.2

# DTD (SGML «Document Type Definitions»)

## Tipo di documento

- informazioni che descrivono la struttura di un documento appartenente a una data tipologia
- individuazione di un insieme di etichette ammesse e di regole di attribuzione delle etichette ai fenomeni testuali
- definizioni per il corpo del testo (*core*) e definizioni per l'intestazione (TEI *header*)
- insieme di *etichette di base* (*base tag-set*)

## TEI header

- informazioni sul documento
  - descrizione bibliografica del documento elettronico
  - descrizione della codifica
  - informazioni sulle eventuali correzioni, normalizzazioni, segmentazioni e interpretazioni
- note non bibliografiche
  - situazione, ambiente, partecipanti, ecc.

# *Esempio di TEI «header»*

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>Thomas Paine: Common sense, a
        machine-readable transcript</title>
      <respStmt>
        <resp>compiled by</resp>
        <name>Jon K Adams</name>
      </respStmt>
    </titleStmt>
    <publicationStmt>
      <istributor>Oxford Text Archive</istributor>
    </publicationStmt>
    <sourceDesc>
      <bibl>The complete writings of Thomas Paine, collected
and edited
        by Phillip S. Foner (New York, Citadel Press,
1945)</bibl>
    </sourceDesc>
  </fileDesc>
</teiHeader>
```

Da Sperberg-McQueen e Burnard 2002, § 5.6

# Etichettatura del parlato

<text>

- quando è caratterizzata da una certa coerenza e coesione, ed è ininterrotta

<u> (*utterance*)

- enunciato
- sequenze di testo delimitate da silenzi o da passaggi di turno

<pause>

- pause di varia durata

<vocal>

- elementi vocali non linguistici (come le pause piene, tipo *ehm*)

<kinesic>

- elementi gestuali

<event>

- eventi esterni

# EAGLES

## Expert Advisory Group on Language Engineering Standards

- Unione europea
- annotazioni linguistiche per la descrizione di tutte le lingue appartenenti alla Ue

## Obiettivi

- codifica dei testi e **annotazione** linguistica
- elaborazione degli strumenti di analisi (**software**) con particolare attenzione alla progettazione di corpora di parlato
- ottenere *riusabilità, interscambiabilità ed estensibilità* per corpora differenti

# Tipologie di etichette di EAGLES

## Obbligatorie

- l'etichettatura morfo-sintattica **obbligatoria** è quella per le **categorie sintattiche** (nome, verbo, avverbio, aggettivo, congiunzione), applicabile in modo uniforme al di là della lingua specifica del corpus
- l'insieme delle etichette è chiuso e definito, e costituito da 13 categorie

## Raccomandate

- livello di etichettatura **raccomandata** che riguarda caratteristiche grammaticali *language-dependent* (come genere, persona, ecc.)
- anche in questo caso l'insieme delle etichette è chiuso, ampio e distinto a seconda delle etichette obbligatorie di riferimento

## Estensioni speciali

- livello delle **estensioni speciali** che indica specifiche grammaticali tipiche di un numero ridotto di lingue europee, oppure annotazioni particolari introdotte a fini specifici
- la classe delle etichette è aperta (ed estensibile a seconda dei bisogni di annotazione) (ad es. aspetto verbale, riflessività, ecc.)

# *Corpus Encoding Standard (CES)*

## Progetto

- conforme a TEI e EAGLES
- pensato per le elaborazioni *del Natural Language Processing*, della lessicografia e della traduzione automatica
- Elabora standard per:
  - **dati primari**, ossia i corpora elettronici non annotati
  - **annotazione linguistica**

## Etichette

- 1) raccomandazioni di livello metalinguistico
- 2) etichette e raccomandazioni per la documentazione dei corpora
- 3) etichette e raccomandazioni per l'annotazione dei dati primari
- 4) etichette e raccomandazioni per l'annotazione linguistica

# *CES – fasi di standardizzazione*

## Il metalinguaggio di annotazione

- (*markup metalanguage*) definisce la sintassi delle etichette

## Livello sintattico

- determina le etichette (*tag names*) e le regole sintattiche per il loro l'uso

## Livello semantico

- predisporre le procedure di applicazione di una determinata etichetta a diversi fenomeni linguistici