

# LINGUISTICA DEI CORPORA

Storia, metodi, problemi

2

## Che cos'è un corpus?

definizioni

## Corpus (plur. corpora)

3

### definizioni

«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)

Un **corpus elettronico** è “A corpus which is **encoded in a standardized and homogeneous way for open-ended retrieval tasks**” (Eagles, 1996a: 3).

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Corpora di...

4

### 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

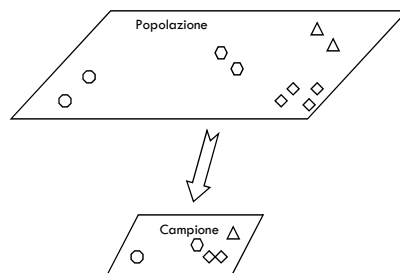
Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Popolazione e campione

5

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



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

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Rappresentatività ed estensione

6

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 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

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

7	<b>Corpus non rappresentativo (insuff.)</b>	< 15.000 parole grafiche
	<b>Corpus piccolo</b>	Da circa 15.000 a 100.000 parole
	<b>Corpus medio</b>	Da circa 100.000 a 1 milione di parole
	<b>Corpus medio-grande</b>	Da circa 1 milione a 50 milioni di parole
	<b>Corpus standard</b>	Da circa 50 milioni a 100 milioni di parole
	<b>Corpus grande</b>	Oltre i 100 milioni di parole
<i>Estensione corpora per analisi lessicali</i>		
Fakulteta za humanistične študije Koper, Univerza na Primorskem, Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari		

## Estensione di alcuni corpora di riferimento

8

### Brown Corpus (1961)

- 1 milione di occorrenze

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

- 500.000 occorrenze

### British National Corpus - 1995

- 100 milioni di occorrenze

### CORIS, Corpus di italiano scritto contemporaneo (CORIS) - 1998

- 100 milioni di occorrenze

### Bank of English

- Circa 500 milioni di occorrenze

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

9

## A cosa servono i corpora?

Usi e applicazioni

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Per il linguista

10

osservare la lingua in uso, in testi autentici e integrali

Linguistica descrittiva

permettere la comparazione – anche quantitativa – delle caratteristiche dei diversi testi

Statistica linguistica

osservare caratteristiche della lingua che sono inosservabili in modo qualitativo e occasionale

Linguistica testuale

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Per la linguistica computazionale

11

### Applicazioni lessicografiche

- dizionari cartacei ed elettronici

### Grammatiche

- *corpus-driven grammar*

### Trattamento automatico del linguaggio

- costruzione di *parsers*, *taggers* e lemmatizzatori che includano moduli di tipo statistico

### Traduzione automatica

- *corpus-based*, *example-based* e statistica

### Speech technologies

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Per l'insegnamento delle lingue

12

### I learners' dictionaries

- Cobuild, Cambridge, Oxford, Longman e MacMillan
- centrati sul testo
  - nell'elaborazione della voce del dizionario stessa, negli esempi e nella possibilità per l'utente di accedere direttamente a concordanze da corpora di riferimento

### Grammatiche

### Ordinamento e progressione dei materiali

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Per l'apprendimento delle lingue

13

### Usi delle concordanze

- estrarre esempi reali
- mettere in luce usi differenti delle parole
- osservare le parole in contesto
- diversificare le accezioni delle parole esaminandone i contesti

### Materiale autentico

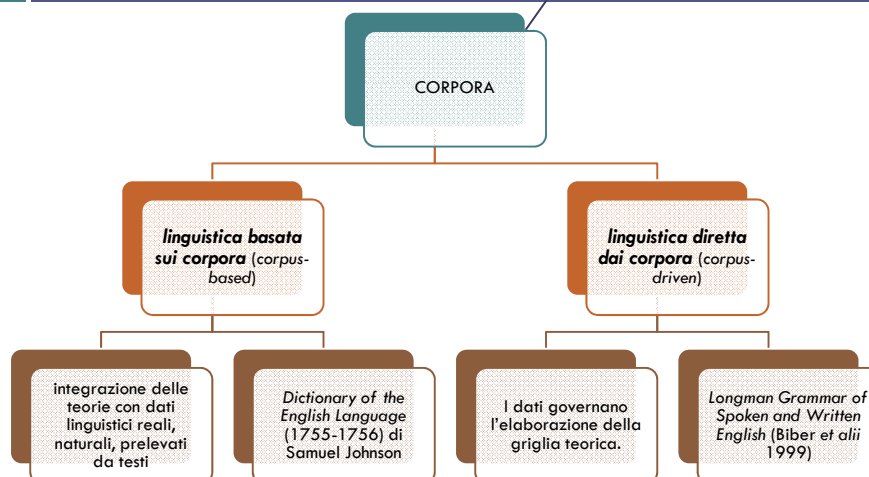
- il discente può trarre inferenze generali sugli usi e sulle variazioni della lingua

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## L'uso dei corpora

14

Elena Tognini Bonelli  
*Corpus Linguistics at Work*  
John Benjamins, Amsterdam-  
Philadelphia 2001



Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

15

## Tipologie di corpora

Corpora di riferimento, specialistici, multilingui e paralleli

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## 1. I corpora di riferimento

16

### *Reference corpus*

- testi appartenenti a diverse varietà sociolinguistiche, diafasiche e diatopiche
- Lingua scritta e lingua parlata

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

### Standard di estensione

- da 500.000 occorrenze
- a 500 milioni

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## L'esempio dell'inglese

17

### «Brown Corpus of Standard American English» 1961

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

#### 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

### «British National Corpus» 1995

#### Oxford University Press

- interrogabile dal sito di Mark Davies: <http://view.byu.edu/>

#### Lingua parlata e lingua scritta

- inglese contemporaneo
- 100.106.008 parole

#### Composizione

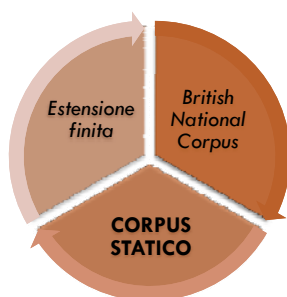
- 4.124 testi
- 90% deriva da testi scritti
  - romanzi e saggi, e testi tecnico-scientifici
- 10% da trascrizioni di parlato
  - 863 testi
  - programmi radiofonici, conversazioni telefoniche, parlato spontaneo

#### Software di interrogazione SARA

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

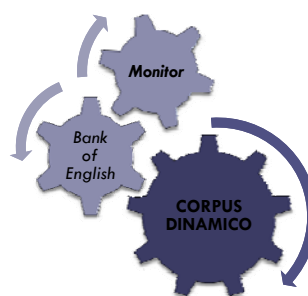
## Corpora statici e corpora dinamici

18



### Vantaggi

- analisi finite e ripetibili
- comparabilità



### Vantaggi

- aggiornamento
- analisi diacroniche

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## La bank of English

19

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

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## 2. Corpora specialistici o settoriali

20

Rappresentano una particolare varietà di lingua

- Soprattutto nei linguaggi settoriali: economia, scrittura accademica, ecc.

*Corpus di sogni*  
di Michel Santacroce

- raccolta di narrazioni di sogni (scritte e parlate) in lingua francese

Fromkin Speech Error  
Database

- Raccoglie lapsus in numerose lingue incluso l'italiano

Corpus di italiano  
televisivo (CIT)

- attualità, intrattenimento, pubblicità, sport e telegiornali

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

### 3. I corpora multilingui e paralleli

21

#### 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

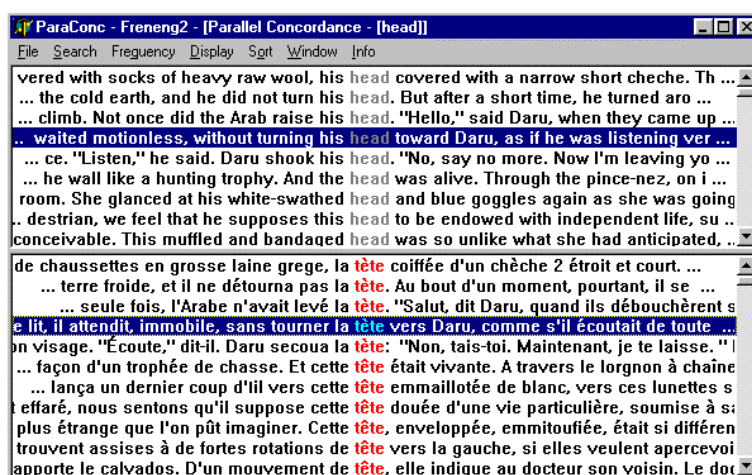
#### Esempi

- C-ORAL-ROM
- progetto MULTEX (Multilingual Text Tools and Corpora)
- progetto CHILDES (Child Language Data Exchange System)

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## ParaConc

22



Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

23

## La costruzione di un corpus

### Le tappe

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

### Le tappe (1)

24

#### Design del corpus

Criteria di rappresentatività: tipologie  
testuali

Estensione (assoluta e relativa)



#### Acquisizione del materiale

biblioteche digitali o cd-  
rom

digitazione / dettatura /  
scannerizzazione

audio / trascrizione



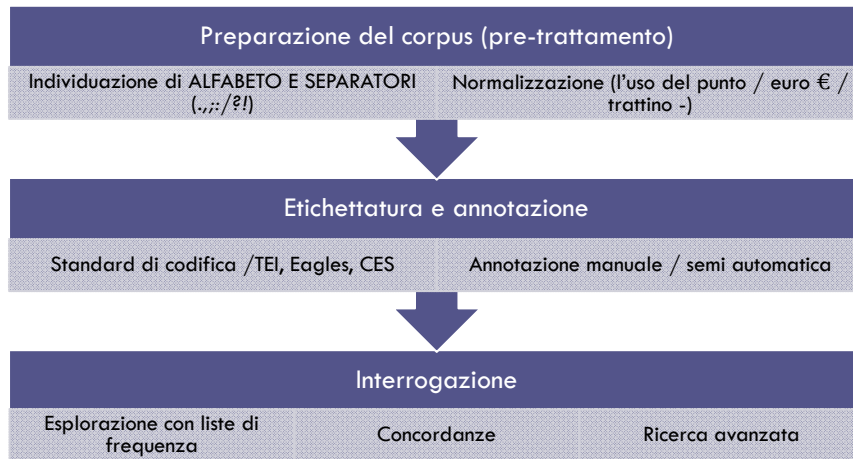
#### Correzione degli errori

correzione manuale / automatica / semi-automatica

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Le tappe (2)

25



Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Esempio di marcatura

26

```

<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>of 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

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

27

## Usare il corpus

L'interrogazione

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Modi di interrogazione

28

Liste di  
frequenza

Concordanze

Interrogazione  
avanzata

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Liste di frequenza

29

### Forma

- elenco di tutte le forme (*types*, tipi di parole)
- indici di frequenza (ossia il numero di occorrenze nel testo)
  - *frequenza relativa* ( $F_w/N$ )

### Presentazione

- per *frequenza decrescente*
  - al primo posto compare la parola testuale più frequente, all'ultimo la meno frequente.

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

### Lista di frequenza del primo capitolo dei «Promessi Sposi»

30

255	4,1255%	e	41	0,6633%	come
195	3,1548%	di	39	0,6310%	una
162	2,6209%	che	38	0,6148%	ma
146	2,3621%	a	38	0,6148%	più
109	1,7635%	il	34	0,5501%	o
100	1,6179%	in	31	0,5015%	gli
100	1,6179%	un	28	0,4530%	don
97	1,5693%	non	28	0,4530%	da
80	1,2943%	la	26	0,4206%	due
78	1,2619%	per	25	0,4045%	se
55	0,8898%	le	24	0,3883%	poi
53	0,8575%	con	24	0,3883%	della
47	0,7604%	si	24	0,3883%	era
44	0,7119%	del	23	0,3721%	al
42	0,6795%	i	22	0,3559%	abbondio

I Frequenze assolute II frequenza relative III tipi di parole

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Le concordanze

31

### Il cotesto

- estrazione di informazioni linguistiche essenziali sugli usi della parola
- individuazione delle sequenze di parole che occorrono più abitualmente
  - *a guisa di, restare con un palmo di naso, giacenza di cassa*

### La concordanza

- è la presentazione delle parole di un testo, con l'indicazione della frequenza con la quale la parola occorre e il contesto linguistico precedente e successivo (cotesto).

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Le concordanze

32

### Funzioni

- osservare i diversi usi di una parola
- esaminare i diversi contesti (semantici, sintattici o testuali) in cui occorre una parola
- analizzare la regolarità con la quale una parola è accompagnata ad altre nel suo cotesto

### KWIC (*keyword in context*)

- la **parola chiave** (*keyword*) è la parola di cui si cerca l'uso, solitamente si trova nella **colonna centrale**.
- il cotesto (precedente e successivo) è stabilito dall'utente:
  - n° fisso di parole (3 – 3, ecc.)
  - frase o verso

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Concordanza di «anima» nella «Divina Commedia»

33

[In.1.122] **anima** fia a ciò più di me degna  
 [In.2.45] **l'anima** tua è da viltade offesa  
 [In.2.58] O **anima** cortese mantoana  
 [In.3.88] E tu che se' costì, **anima** viva  
 [In.3.127] Quinci non passa mai **anima** buona  
 [In.5.7] Dico che quando **l'anima** mal nata  
 [In.6.55] E io **anima** trista non son sola  
 [In.10.15] che **l'anima** col corpo morta fanno  
 [In.12.74] saettando qual **anima** si svelle  
 [In.12.90] non è ladron, né io **anima** fuia  
 [Pu.4.3] **l'anima** bene ad essa si raccoglie  
 [Pu.18.44] e **l'anima** non va con altro piede

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
 Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## L'interrogazione avanzata

34

### Per forma specifica

- *psicologico*: si ottengono tutti i token di questo type

### Per lemma e/o categoria grammaticale

- *psicologico* (su corpus lemmatizzato): si ottengono tutte le occorrenze di tutte le forme flesse

### Con l'uso di caratteri jolly (wildcards) o espressioni regolari

- *Psicologic\** o *conpsicologic[aoih]?*

### Con l'uso delle etichette

- Integrazione e combinazione di tags grammaticali, semantici, e criteri formali

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
 Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari