

I corpora linguistici

- Cos'è un “corpus”?
 - È un termine che può indicare qualunque testo in forma scritta o parlata.
 - Nella linguistica moderna il termine indica un insieme molto grande di testi, in formato leggibile da un calcolatore, che rappresentino una particolare varietà od utilizzo di una lingua.
- Esistono tipi diversi di corpora:
 - possono concernere la lingua scritta o parlata (trascritta), antica o moderna;
 - possono consistere di interi libri, giornali, riviste, discorsi o di estratti di lunghezza variabile.

Classificazioni dei corpora

- Esiste la seguente suddivisione dei corpora:
 - **corpora generali**: non riguardano un singolo tipo di testo o argomento (es.: **British National Corpus**, <http://www.natcorp.ox.ac.uk/>);
 - **corpora di “sottolinguaggi”**: sono composti da testi che riguardano una varietà di una lingua (es.: varietà dialettale o inerente ad un particolare argomento);
 - **corpora paralleli**: sono composti dagli stessi testi tradotti in più lingue;
 - **corpora confrontabili**: sono composti da testi “simili” (il criterio di similarità è variabile) scritti in più lingue o varietà di una lingua: lo scopo è quello di confrontare l’uso di lingue diverse in circostanze comunicative simili.

Corpora elettronici

- I corpora elettronici possono includere il testo in due forme:
 - in forma semplice, senza informazioni aggiuntive,
 - con informazioni linguistiche aggiuntive (annotazioni o marcatori).
- Le informazioni in un corpus annotato possono essere di varia natura:
 - grammaticali,
 - sintattiche,
 - semantiche,
 - storiche,
 - ...

Origini storiche

- L'idea di utilizzare collezioni di testi nel campo della linguistica risale al Medioevo, quando si iniziò a:
 - produrre delle liste di tutte le parole di certe categorie di testi unitamente al loro contesto di utilizzo (processo che oggi viene chiamato “concordancer”),
 - calcolare le frequenze d'uso delle parole nei testi e produrre una lista di quelle maggiormente utilizzate.
- Il vantaggio degli studiosi dei nostri tempi è la disponibilità del calcolatore per gestire ed analizzare i corpora.

La prima generazione

- La prima generazione di corpora elettronici iniziò con il **Brown Corpus of Standard American English**.
- Tale corpus consiste di **un milione di parole** di testi americani stampati nel 1961. I testi furono scelti in proporzioni diverse da varie categorie:
 - stampa (reportage, editoriali ecc.),
 - letteratura scientifica,
 - ...
- Al giorno d'oggi tale corpus viene ormai ritenuto “piccolo” e leggermente datato, anche se per lungo tempo è stato uno dei pochi corpora elettronici disponibili.

La prima generazione

- Il **London-Lund Corpus of Spoken British English** (LLC) è un altro esempio di corpus “piccolo” (**100** testi di circa **5.000** parole ognuno).
- LLC è stato il primo corpus elettronico di una lingua parlata.
- I testi sono stati classificati in categorie come:
 - conversazioni spontanee,
 - orazioni,
 - ...
- Il tutto è stato trascritto in modo ortograficamente corretto e annotato con informazioni prosodiche.

Corpora di grandi dimensioni

- Nonostante i primi corpora si fossero resi utili per molti studi linguistici, cominciava a nascere l'esigenza di avere a disposizione grandi collezioni di parole (ad es. per la compilazione di dizionari).
- Nel 1987 nasce il Collins Cobuild English National Dictionary con circa 20 milioni di parole.
- Nel 1991 nasce la Bank of English (BoE) che nel 1996 contava 320 milioni di parole e rappresenta un monitor corpus (ovvero, un corpus che cresce con i cambiamenti della lingua corrente).
- Nel 1995 nasce il British National Corpus (BNC) con 100 milioni di parole (il BNC tuttavia è finito, ovvero, non vengono aggiunte nuove parole).

Corpora specializzati

- **Corpora “storici”:**

- The Helsinki Corpus of English Texts: 1,5 milioni di parole estratte da testi inglesi antichi, medievali e dei primi secoli dell’era moderna.
- Lampeter Corpus of Early Modern English Tracts: opuscoli e versi pubblicati dal 1640 al 1740.

- **Corpora relativi a campi specifici:**

- Air Traffic Control Corpus,
- TRAINS Spoken Dialogue Corpus.

- **Corpora internazionali/multilingue:**

- English-Norwegian Parallel Corpus
- English Turkish Aligned Parallel Corpora.

Modi di utilizzo

- Cercare il significato di una parola.
- Confrontare due sinonimi/parole simili.
- Studiare gli usi di una parola in contesti differenti.
- Capire che preposizione usare.
- Controllare l'ortografia di una parola.

W3-Corpora

- Un insieme di corpora elettronici online (con motore di ricerca) è disponibile all'URL

<http://clwww.essex.ac.uk/cgi-bin/w3c/w3c:>

HELP

Corpus Access

W3CORPORA

PROTOTYPE F

1. Corpus Selection

Press the **CORPUS** button, to enter a corpus.

CORPUS

Selezione del corpus.

2. Search String

Press the **STRING** button, to enter a search string.

STRING

3. Submit Query

Once a corpus has been selected and a search string entered you may proceed further.

Comments and Suggestions

If you have any comments or suggestions please let us know, thank-you.

COMMENTS

Esempio di utilizzo

HELP

Corpus Selection

ACCESS

1. Corpus

The Gutenberg Project, a collection of electronic texts.

2. Confirm Choice

Press the **CONFIRM** button, once you have made your selection.

CONFIRM

Esempio di utilizzo

HELP

Subcorpus Selection

ACCESS

The **Gutenberg Project**, a collection of electronic texts.

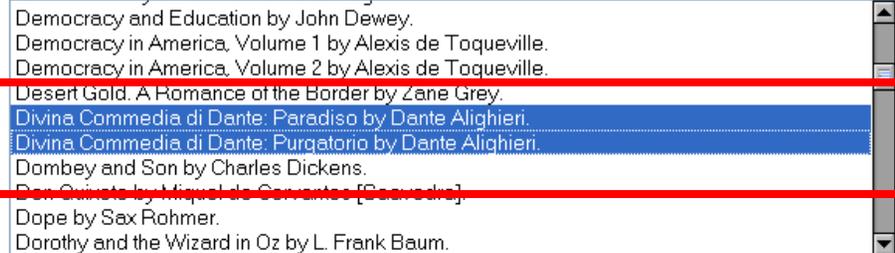
CORPUS

1. Subcorpus

CLEAR

- Please select one or more subcorpora -

1



Democracy and Education by John Dewey.
Democracy in America, Volume 1 by Alexis de Toqueville.
Democracy in America, Volume 2 by Alexis de Toqueville.
Desert Gold. A Romance of the Border by Zane Grey.
Divina Commedia di Dante: Paradiso by Dante Alighieri.
Divina Commedia di Dante: Purgatorio by Dante Alighieri.
Dombey and Son by Charles Dickens.
Don Quixote by Miguel de Cervantes [Cervantes].
Dope by Sax Rohmer.
Dorothy and the Wizard in Oz by L. Frank Baum.

2. Confirm Choice

Press the **CONFIRM** button, once you have made your selection.

2

CONFIRM

Esempio di utilizzo

HELP

Corpus Access

W3CORPORA

PROTOTYPE F

GUTENBERG

1. Corpus Selection

CORPUS

The **Gutenberg Project**, a collection of electronic texts.

2. Search String

STRING

Press the **STRING** button, to enter a search string.

3. Submit Query

Once a search string has been entered you may proceed further.

Comments and Suggestions

COMMENTS

If you have any comments or suggestions please let us know, thank-you.

Esempio di utilizzo

HELP

Search String

ACCESS

1. String Type

CLEAR

- A regular expression
- within a word
- Exact match

- The beginning of a word
- The end of a word

1

2. String

Please enter search string below:
Virgilio

2

3. Confirm Choice

Press the **CONFIRM** button, once you have selected a string type and string.

3

CONFIRM

Esempio di utilizzo

HELP

Corpus Access

W3CORPORA

PROTOTYPE F

GUTENBERG

1. Corpus Selection

The **Gutenberg Project**, a collection of electronic texts.

CORPUS

2. Search String

Virgilio

STRING

3. Submit Query

Press the **SEARCH** button to start the search ...

SEARCH

Comments and Suggestions

If you have any comments or suggestions please let us know, thank-you.

COMMENTS

Esempio di utilizzo

HELP

DISPLAY

FREQUENCY

SEARCH

OPTIONS

Frequency Screen

MATCH FREQUENCY

SUBCORPUS FREQUENCIES

LEXICAL FREQUENCIES

Match Frequency

Search String : Virgilio
String Type : Exact match.
Corpus : The Gutenberg Project, a collection of electronic texts.
Frequency : 27

Frequenze degli elementi trovati.

Esempio di utilizzo

HELP

DISPLAY

FREQUENCY

SEARCH

OPTIONS

Frequency Screen

MATCH FREQUENCY

SUBCORPUS FREQUENCIES

LEXICAL FREQUENCIES

Subcorpus Frequencies

| Subcorpus | Frequency |
|--|-----------|
| Divina Commedia di Dante: Purgatorio by Dante Alighieri. | 25 |
| Divina Commedia di Dante: Paradiso by Dante Alighieri. | 2 |
| Total | 27 |

Frequenze degli elementi trovati nei vari subcorpora.

Esempio di utilizzo

HELP

DISPLAY

FREQUENCY

SEARCH

OPTIONS

Frequency Screen

MATCH FREQUENCY

SUBCORPUS FREQUENCIES

LEXICAL FREQUENCIES

Lexical Frequencies

| Lexical Items | Frequency |
|-----------------|-----------|
| Virgilio | 27 |
| Total | 27 |

Frequenze lessicali degli elementi trovati.

Esempio di utilizzo

HELP

DISPLAY

FREQUENCY

SEARCH

OPTIONS

<

KWIC FRAME

>

1..10/ 27

son presenti ; mentre ch'io era a Virgilio congiunto su per lo monte che
segno . Quindi onde mosse tua donna Virgilio , quattromila trecento e due volumi di
per far colei confusa . ½ O Virgilio , Virgilio , chi ^ questa ? © , fieramente dicea ; ed el
Noi ci volgemmo s · biti , e Virgilio rend %o li Æ l cenno ch Æ a
esser vivuto di l € quando visse Virgilio , assentirei un sole pi · che non
al mio uscir di bando © . Volser Virgilio a me queste parole con viso
in alto li occhi miei , ^ quel Virgilio dal qual tu togliesti forte a
in s · li spiriti veloci ; quando Virgilio incominci ø : ½ Amore , acceso di virt · , sempre
quivi convien che senza lui rimagna . Virgilio ^ questi che cos Æ mi dice © , e
non so chi diceva ; per che Virgilio e Stazio e io , ristretti , oltre

Contesti in cui la parola è usata.

Espressioni regolari

Search String

HELP

ACCESS

1. String Type

CLEAR

- A regular expression
- Within a word
- Exact match

- The beginning of a word
- The end of a word

2. String

Please enter search string below.

3. Confirm Choice

CONFIRM

Press the **CONFIRM** button, once you have selected a string type and string.

Espressioni regolari

HELP

DISPLAY

FREQUENCY

SEARCH

OPTIONS

Frequency Screen

MATCH FREQUENCY

SUBCORPUS FREQUENCIES

LEXICAL FREQUENCIES

Match Frequency

Search String : `((Cc]ittà)((Cc]ittade)`
String Type : A regular expression.
Corpus : The **Gutenberg Project**, a collection of electronic texts.
Frequency : 2

Espressioni regolari

HELP

DISPLAY

FREQUENCY

SEARCH

OPTIONS

<

KWIC FRAME

>

1..2/2

principio fu del mal de la cittade , come del vostro il cibo che
aver , che discernesse de la vera cittade almen la torre . Le leggi son

Alcuni costruttori di espressioni regolari ricorrenti

- **[stringa]**: un qualunque carattere che compaia in “stringa”.
- *****: zero o più ripetizioni dell’elemento precedente.
- **?**: zero o un’occorrenza dell’elemento che precede.
- **(exp)**: le parentesi tonde servono a raggruppare un’espressione.
- **.**: un carattere qualsiasi
- **Exp1 | Exp2**: l’espressione Exp1 oppure (in alternativa) l’espressione Exp2.