Modulo A

Valutare i sistemi automatici di interazione linguistica

16. Valutazione e rappresentatività

30 novembre 2016

Linguistica italiana II
Mirko Tavosanis
A. a. 2016-2017

Calendario

- Da oggi a venerdì 2 dicembre alle 13: scegliere tra Letto e Dialogo
- All'inizio della prossima settimana vi chiederò di scegliere anche i file (che rivedrò caso per caso)

Corpus

- Un corpus è una collezione di testi selezionati e organizzati per facilitare le analisi linguistiche (o di altro genere)
- Arbitrariamente, è diventato di moda il plurale *corpora*
- Quando si parla di “corpus” oggi di solito si intende un “corpus elettronico”
- Un corpus può essere fatto di testi:
  - scritti (per esempio, CORIS)
  - parlati (CLIPS); in questo caso, possono essere presenti, o no, audio e trascrizione

Usi dei corpora

- I corpora servono allo studio della lingua (la “linguistica dei corpora” è un tipo particolare di studio, basato su corpora, che esamina la lingua proprio da questa angolazione)
- I corpora servono inoltre all’addestramento dei sistemi automatici

- In generale: i corpora delle lingue moderne hanno di regola l’ambizione di essere rappresentativi di qualcosa (= essere una specie di equivalente in scala ridotta di qualcosa di più grande)
- Per esempio: un corpus di conversazioni telefoniche può avere l’obiettivo di documentare come sono fatte un po’ tutte le conversazioni telefoniche

Corpora di una lingua

- Alcuni corpora hanno l'ambizione di rappresentare una lingua nel suo assieme
- Naturalmente, non ci riescono, perché:
  - la variazione linguistica si fonda su un numero altissimo di parametri (molti dei quali sfuggono all'analisi)
  - anche nel migliore dei corpora, solo alcuni di questi parametri sono tenuti in considerazione in modo adeguato (anche se alcuni corpora sono progettati con immensa intelligenza)
- E se sono rappresentati, lo sono nel modo corretto?

Parallelamente, occorre ricordare che…

- … l'uso documentato dal corpus non riflette completamente la conoscenza della lingua da parte dei parlanti
- Un esempio classico: il lessico di "alta familiarità" o "alta disponibilità" degli studi di De Mauro, basati su corpora ma non solo
- Questo lessico è compost da parole rarissime nell'uso normale ma "legate ad atti e oggetti della vita quotidiana (da *aceto, avvitare* o *forchetta* a *vomito* o *zuppa*), che abbiamo continuamente in mente": 1.800 parole circa da aggiungere al vocabolario di base
- Nella preparazione di un dizionario non ci si può quindi basare solo sulla semplice frequenza, considerando un assieme di testi "rappresentativo" delle conoscenze

Rappresentatività rispetto ai diversi tipi di lingua

- Nozione intuitiva: i contenuti di un corpus dovrebbero "rappresentare" un determinato tipo di lingua (un po' come una carta topografica rappresenta un territorio)
- Un punto di partenza moderno: D. Biber, "Representativeness in corpus design", *Literary and Linguistic Computing*, 1993, 8, pp. 243-257
- Metodo: procedere a cicli, controllare la variazione e poi modificare la composizione del materiale
- Il risultato dovrebbe essere utilissimo: fornisce una salda base empirica per formulare generalizzazioni scientifiche

Scelta dei testi o delle sezioni

- Si può decidere se prendere testi interi o sezioni
- Prendere solo l'inizio di un testo? O la fine?
- Pensiamo a un corpus di lettere che comprenda *solo* il vocativo d'inizio o che lasci fuori i saluti…
- … o a un corpus di conversazioni telefoniche che non comprenda le fasi di apertura e conclusione
- Ma sui testi scritti lunghi (monografie, romanzi…), per esempio, ha senso prendere il testo intero? Non basta una sezione?

Brown Corpus

- Il primo corpus moderno (inglese): Realizzato nel 1967 da Henry Kucera e W. Nelson Franklin
- Comprende un milione di “parole” prese da 500 testi diversi pubblicati a stampa nel 1961
- I testi sono stati scelti in proporzione a quanto pubblicato nel 1961
- Ogni campione parte da una frase presa a caso nel testo e termina con la frase che si conclude dopo la duemillesima parola successiva

Alcuni materiali inclusi nel Brown Corpus

* A. PRESS: Reportage (44 texts)
    - o Political
    - o Sports
    - o Society
    - o Spot News
    - o Financial
    - o Cultural
* B. PRESS: Editorial (27 texts)
    - o Institutional Daily
    - o Personal
    - o Letters to the Editor
* C. PRESS: Reviews (17 texts)
    - o theatre
    - o books
    - o music
    - o dance
* D. RELIGION (17 texts)
    - o Books
    - o Periodicals
    - o Tracts
* E. SKILL AND HOBBIES (36 texts)
    - o Books
    - o Periodicals
* F. POPULAR LORE (48 texts)
    - o Books
    - o Periodicals
* G. BELLES-LETTRES - Biography, Memoirs, etc. (75 texts)
    - o Books
    - o Periodicals
* H. MISCELLANEOUS: US Government & House Organs (30 texts)
    - o Government Documents
    - o Foundation Reports
    - o Industry Reports...

Il Coris

- CORpus di Italiano Scritto
- Realizzato a partire dal 1998 presso l'Università di Bologna come corpus "generale" per l'italiano scritto ("scritto-scritto")
- In linea presso l'Università di Bologna http://corpora.dslo.unibo.it/coris_ita.html
- Contiene 120 milioni di parole provenienti da testi scritti (in buona parte degli anni Ottanta e Novanta), ed è stato aggiornato ogni tre anni con un corpus di controllo

Corpus CORIS, annotated version (2011, 130Mw)
- Corpus query form -

User Authentication

CORIS access is now free for research purposes
(Please, read the footnote carefully).

Query

(Query Language Help).

Subcorpus All **Section** All

Concordance Options

Show (•) 30 ( ) 100 ( ) 300 ( ) 1000 **lines.**

Sort position: Unsorted

Collocations

Get Collocates? (•) NO! ( ) Yes.

Sort using (•) Log-Likelihood Ratio. ( ) Mutual Information. ( ) T-score. ( ) Raw frequency.

Esegui Cancella

Interrogando il corpus CORIS/CODIS tramite questa procedura, l'utente dichiara e accetta che l'interrogazione è volta unicamente a scopi di ricerca scientifica e che non ne sarà tratto alcun beneficio eco
è concesso esclusivamente per scopi di ricerca scientifica.

Designed by Fabio Tamburini.

Coris: fasi di allestimento

1. **Corpus design**
   1. Corpus typology
   2. Corpus size
   3. **Representativeness**
2. Design of source text framework
   1. Text typology
   2. Text unit size
   3. Definition of selection criteria
3. Corpus structure
   1. Subcorpora definition
   2. Subcorpora-to-subcorpora ratio
   3. Definition of sampling criteria
4. Source data collection and corpus building
5. Part-of-speech tagging and lemmatisation

Scelte di base del Coris

- Testi interi o sezioni di testo?
  La scelta di sezioni “may lead to a stronger influence of the researcher's subjective judgment and implies that the selected sequence is taken out of context. This **could mean** that the larger size invalidates the very representativeness of the corpus. It was therefore decided that, **where possible**, the entire text would be entered, rather than standardising sample size.”
  Si tratta di una scelta dimostrabile? Anche qui, si procede molto sulla base di “impressioni” e vincoli pratici
- Delimitazione di sezioni e sottosezioni
- Quali sono i testi più interessanti?
- Come si opera la selezione? (Di sicuro non sulla base dei “migliori”)
- Eccetera…

Altri criteri

- A later step was the definition of **linguistic varieties** used to create the corpus. These are considered as a collection of documents identifiable on the basis of both external and internal features, in which the peculiarity of the single variety fades away in comparison to the mass of data. This constituted one of the most important points. Although the corpus included specialist areas, such as legal, scientific and bureaucratic-administrative language, an attempt was made to bring together not so much a collection of specialist texts as a **variety of types which, according to our investigations, can be placed within a continuum**, overlapping and integrating one and another.
- When defining the selection and creation criteria reference was made to both external and internal criteria in order to reduce the researcher's interference to a minimum. Furthermore, considering the scientific context of CORIS as well as the wide availability of existing and planned corpora, a further criterion was introduced, that of "comparability", in order to offer scholars the possibility of interlinguistic comparison of corpora.

Per esempio:

- Subcorpus: PRESS
  Sections: newspapers, periodic, supplement
  Subsections:
  - national, local
  - specialist, non specialist
  - connotated, non connotated
- Subcorpus: FICTION
  Sections: novels, short stories
  Subsections:
  - Italian, foreign
  - for adults, for children
  - crime, adventure, science-fiction, women literature

Libri o stampa periodica?

- In base ai dati disponibili, il rapporto di testi pubblicati in Italia risultava:
  Stampa periodica: circa 4 miliardi di parole
  Libri: circa 300 milioni di parole
- Conclusioni: “The ratio of 1:12 established, more or less, between texts from the mass media and texts from the book market **could not be accepted** as being reproducible in the samples. On the other hand, it appeared to be **too relevant to ignore**, even bearing in mind the comparability of the corpus under construction. Within the ratio allowed by the sales volumes, which, on the basis of the data, is represented as an interval, it was decided to set the ratio between the different areas of circulation as the smallest allowed value in order not to penalise certain textual varieties, such as letters.”
- Insomma: arbitrio al massimo (ma non è un caso isolato)

Scelte finali del Coris

- PRESS - 38 million words
- FICTION - 25 million words
- ACADEMIC PROSE - 12 million words
- LEGAL AND ADMINISTRATIVE PROSE - 10 million words
- MISCELLANEA -10 million words
- EPHEMERA - 5 million words

Il Codis

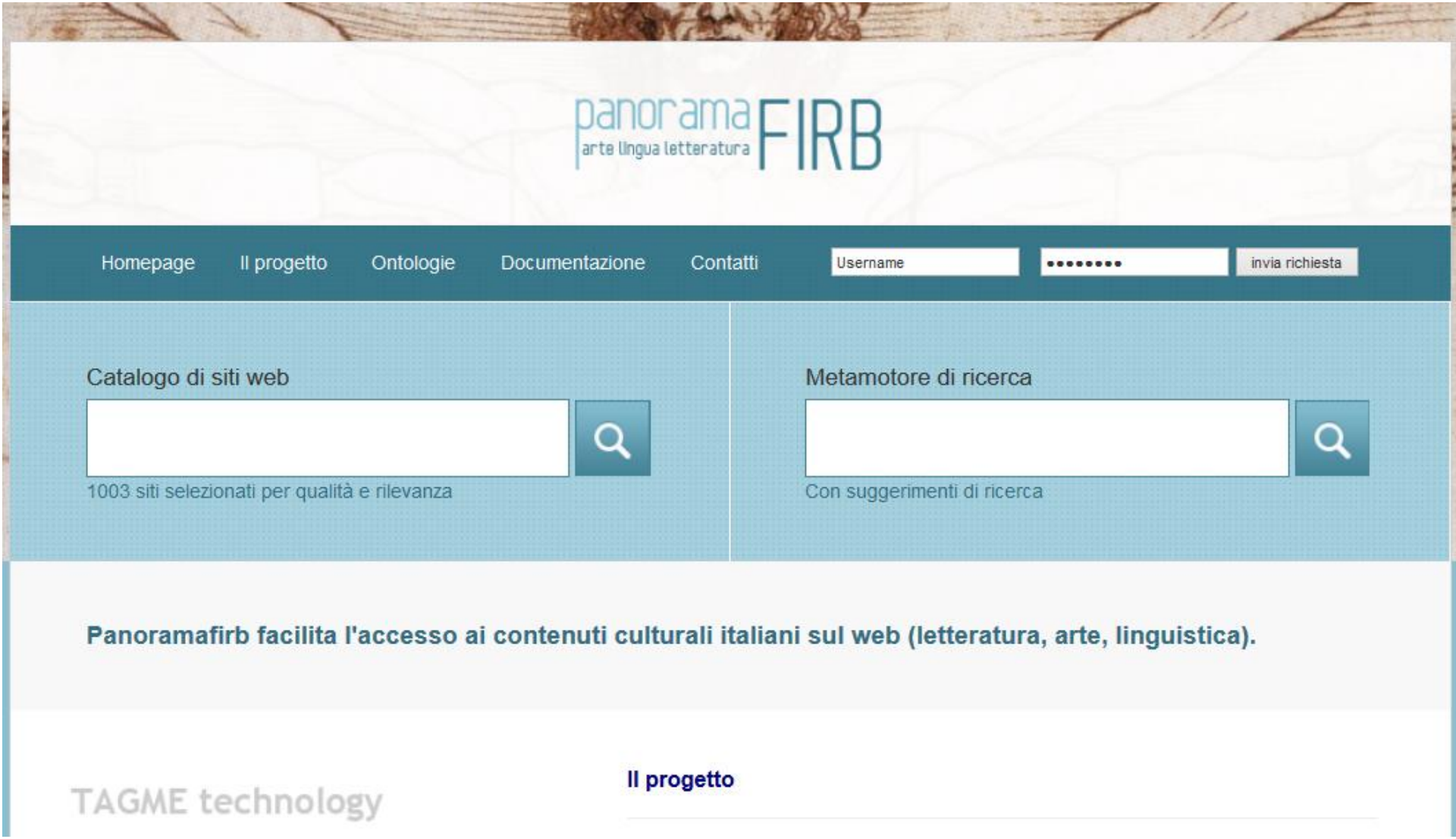
- COrpus Dinamico dell'Italiano Scritto
- Permette di selezionare sottoinsiemi (precompilati) all'interno dei materiali presenti nel Coris
- Per esempio: una sezione di 20 milioni di parole tratte da quotidiani e periodici (utile per esaminare la lingua dei giornali)

Corpora disponibili in rete

- Una lista commentata si trova nel libro di Barbera
- Una lista più ampia (25) si trova sul sito di un FIRB coordinato a Pisa

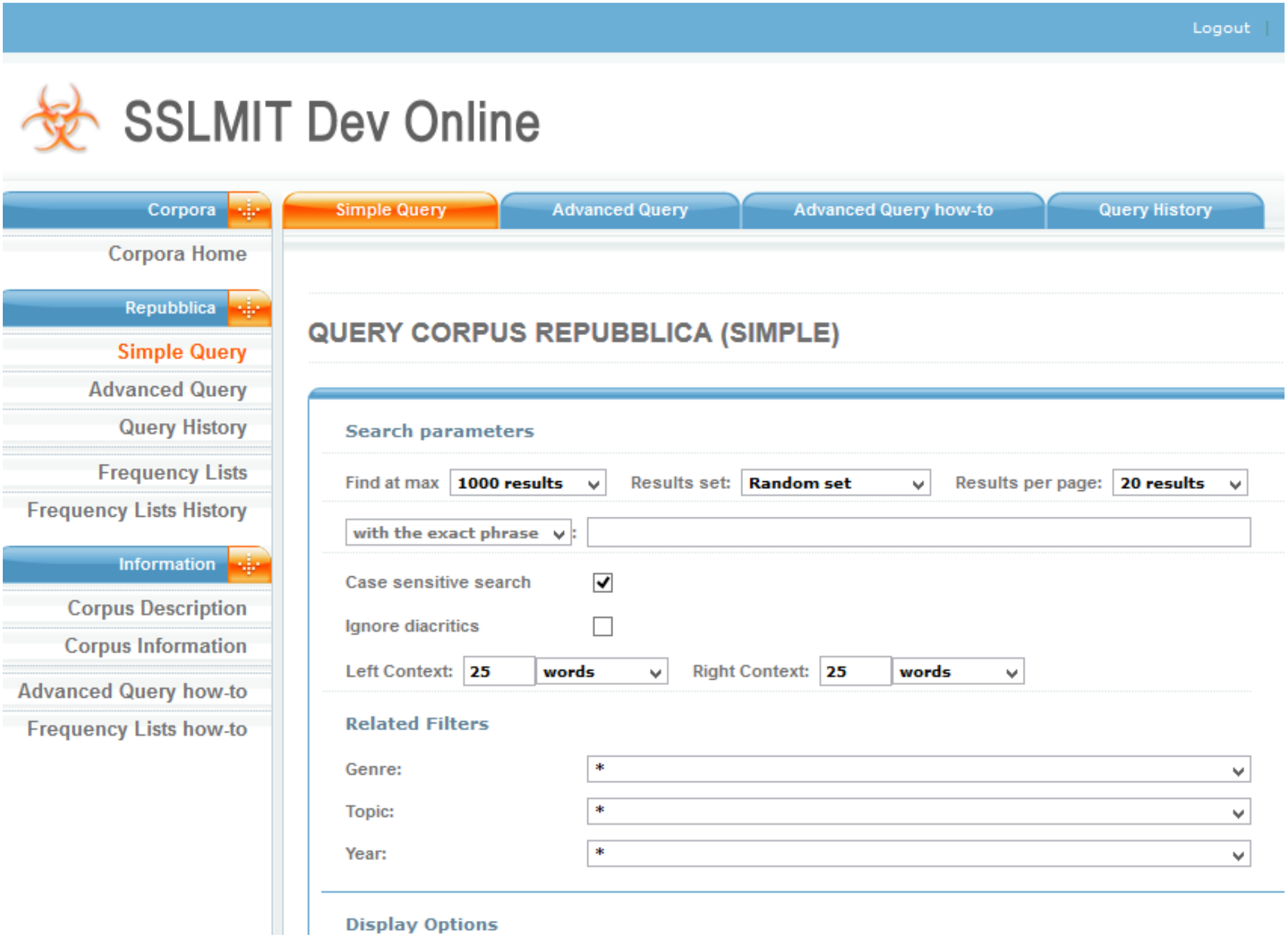
http://www.panoramafirb.it/

- In effetti, quelli contemporanei non sono molti…


panorama FIRB
arte lingua letteratura
Homepage
Il progetto
Ontologie
Documentazione
Contatti
Username
invia richiesta
Catalogo di siti web
1003 siti selezionati per qualità e rilevanza
Metamotore di ricerca
Con suggerimenti di ricerca
Panoramafirb facilita l'accesso ai contenuti culturali italiani sul web (letteratura, arte, linguistica).
TAGME technology
Il progetto

Repubblica corpus

- 380 milioni di parole (= quasi 4 volte il Coris), all'indirizzo http://sslmit.unibo.it/repubblica
- Le parole sono etichettate in base alla parte del discorso (POS-tagging) e al lemma
- Per i testi sono fornite classificazioni strutturali
- Il linguaggio di ricerca è sofisticato
- Invece di essere "rappresentativo", questo corpus è completo, poiché contiene tutti gli articoli disponibili
- Però, riguarda solo i giornali (anzi, solo *un* giornale)


Logout |
SSLMIT Dev Online
Corpora
Corpora Home
Repubblica
Simple Query
Advanced Query
Query History
Frequency Lists
Frequency Lists History
Information
Corpus Description
Corpus Information
Advanced Query how-to
Frequency Lists how-to
Simple Query
Advanced Query
Advanced Query how-to
Query History
QUERY CORPUS REPUBBLICA (SIMPLE)
Search parameters
Find at max 1000 results
Results set: Random set
Results per page: 20 results
with the exact phrase :
Case sensitive search
Ignore diacritics
Left Context: 25 words
Right Context: 25 words
Related Filters
Genre: *
Topic: *
Year: *
Display Options

Scarsa oggettività anche nel Brown Corpus?

- Le scelte del Brown corpus erano basate solo in parte su criteri oggettivi (anche per la difficoltà di avere dati editoriali sufficientemente precisi sulla produzione del 1961)
- Per esempio: la percentuale di *humorous writing* era davvero l'1,8%?
- Le decisioni sul numero esatto di testi (80 di "prosa accademica", 6 di fantascienza) sono state prese in modo informale, durante riunioni, basandosi sulla sensibilità individuale
- Ma se fosse possibile raggiungere la rappresentatività sul pubblicato, il corpus sarebbe a posto?

Critica della rappresentatività

- Oltre alle scelte discutibili e ai problemi nella determinazione oggettiva…
- … non si può *dimostrare* che un corpus sia veramente rappresentativo di una determinata sezione del linguaggio
- Sappiamo *davvero* solo che un corpus rappresenta sé stesso
- Le estrapolazioni dal corpus sull'assieme di una sezione del linguaggio sono tutt'al più ragionevoli assunzioni
- In fin dei conti, *language is never, ever ever random*…

Un punto chiave: destinatari

- Il puro dato bibliografico sui testi pubblicati dà indicazioni sulla **produzione** ma non sul **consumo**
- In altri termini: un libro letto da mille persone può "pesare" quando un best-seller letto da un milione di persone?
- Geoffrey Leech propone il concetto di Atomic Communicative Event – **ACE**
- Ogni volta che un testo viene letto, si ha un ACE
- Un libro letto da mille persone avrà 1000 ACE, uno letto da un milione 1.000.000 ACE, ecc.
- Problemi simili sono molto sentiti, anche a livello commerciale, per l'Auditel, le "metriche" dei siti web, e così via

Metriche migliori: G. Leech

- Geoffrey Leech, “New resources, or just better old ones?”, in *Corpus linguistics and the web,* Rodopi, Amsterdam – New York, 2007, pp. 133-139
- Gli ACE sono un primo passo
- La misurazione del “prestigio” linguistico di un testo (nei limiti del possibile) è un altro passo ancora
- E in ogni caso, “even if the absolute goal of representativeness is not attainable in practical circumstances, we can take steps to approach closer to this goal on a scale of representativity” (p. 140)

Diverso: l'esame di ciò che è elettronico

- In alcuni casi e per alcuni generi di testo i motori di ricerca permettono già oggi di indicizzare tutto o quasi: per esempio, gli articoli scientifici pubblicati in rete
- Diventa quindi possibile esaminare in tempo reale **tutti** i testi appartenenti a un genere di ampia diffusione
- Al confronto: tutte le iscrizioni etrusche si possono inserire in un unico libro, tutti i testi dell'antichità greca stanno in una stanza – e *non* si aggiornano
- Il problema della rappresentatività sparisce, o meglio, viene sostituito da problemi tecnici e di diritto d'autore… o di privacy, nel caso di interazioni tipo Google
- Occorre però essere nella posizione di Google o Amazon