Linguistica Computazionale
Corpora

Salvatore Sorce
Dipartimento di Ingegneria
Chimica, Gestionale, Informatica e Meccanica
Ludici Adattati da Alessandro Lenci
Dipartimento di Linguistica “T. Bolelli”

InformaticaUmanistica

Dati linguistici

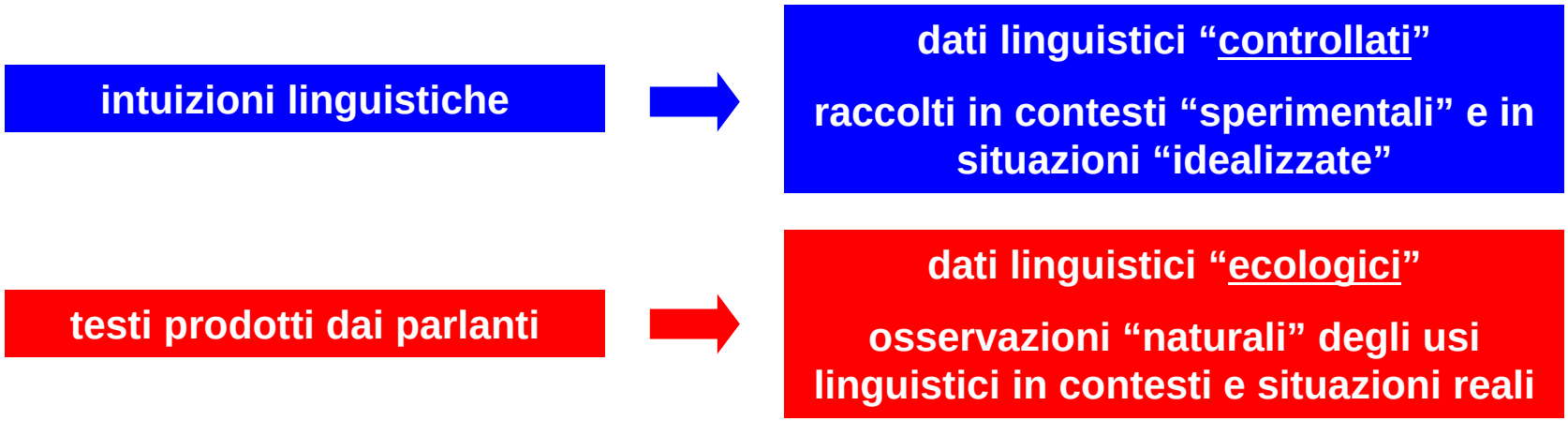
- Dati linguistici
  - i prodotti del linguaggio che sono oggetto di un processo di analisi (computazionale) e che formano l'evidenza empirica per lo sviluppo di modelli e teorie linguistiche
    - come funziona il linguaggio, qual è la sua organizzazione, come viene usato, come viene appreso

Il linguista computazionale:

- raccoglie dati linguistici
- usa metodi formali (logici, matematici, statistici, ecc.) e strumenti informatici per analizzare i dati raccolti e ricostruire l'organizzazione e struttura del linguaggio

Dati linguistici

- Le fonti dei dati linguistici



Dati linguistici controllati

- Fonte di dati primaria per la linguistica formale "razionalista" di derivazione chomskiana
  - obiettivo dell'indagine linguistica è ricostruire le conoscenze che i parlanti hanno della lingua (competenza) indipendentemente dal modo in cui la usano (esecuzione o performance)
    - i fenomeni tipici dell'uso linguistico sono considerati "rumore" da cui è necessario fare astrazione
- Fonte di dati primaria per la linguistica computazionale e Intelligenza Artificiale di I^{a} generazione
  - sistemi generalmente in grado di operare in ambienti circoscritti (toy models)
- Limiti e problemi dei dati controllati
  - le intuizioni dei parlanti non sono sempre "chiare e distinte"
    - "la ragazza che ci sono uscito ieri" (???)
    - "c'è la maggior parte di noi che non leggono abbastanza" (???)
  - esperimenti "in vitro"
  - eccessivo grado di idealizzazione e astrazione rispetto all'uso effettivo del linguaggio
  - i sistemi computazionali sono scarsamente adattabili e "robusti"

Dati linguistici “ecologici”

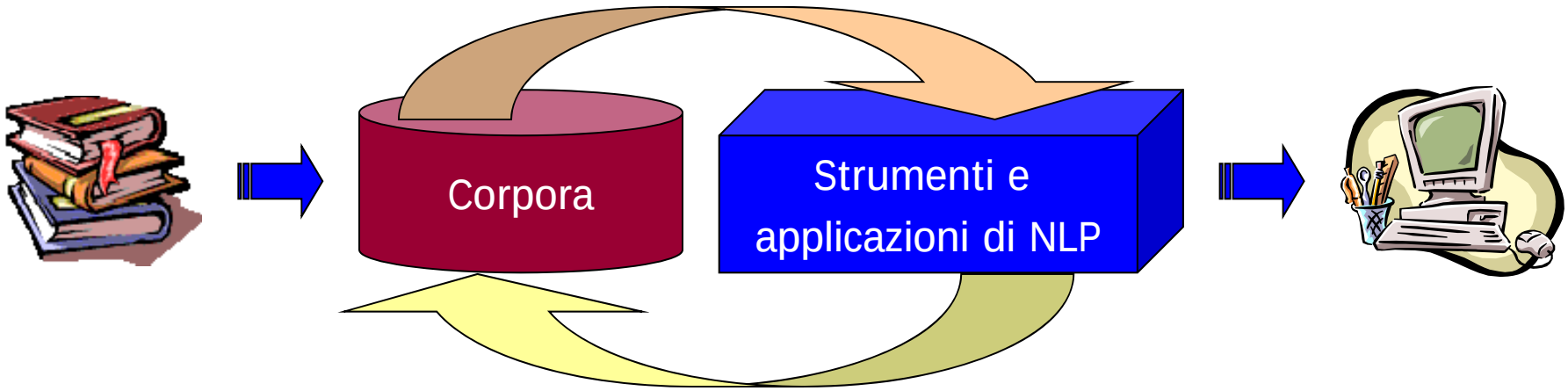
(Dal lat. *corpus*, “corpo”, pl. *corpora*)

Un *corpus* è una collezione di testi selezionati e organizzati in maniera tale da soddisfare specifici criteri che li rendono funzionali per le analisi linguistiche

- I corpora rappresentano fonti di dati linguistici “ecologici”, ovvero raccolti nei loro “habitat naturali”
  - lingua scritta
    - libri (saggistica, narrativa, poesia, ecc.), giornali, riviste, pagine Web, produzioni “effimere” (e-mail, pubblicità, chat, volantini, ecc.)
  - lingua parlata (tipicamente trascritta)
    - notiziari radio-televisivi, conversazioni telefoniche, conversazioni faccia-a-faccia, interviste, ecc.

Corpora e linguistica computazionale

I corpora testuali rappresentano la principale (anche se non esclusiva) fonte di dati in linguistica computazionale



Corpora elettronici

- L'avvento dell'era informatica ha rivoluzionato la natura, il ruolo e l'uso stesso dei corpora
- Il computer permette di:
  - immagazzinare quantità di dati testuali prima inimmaginabili
  - interrogare in maniera avanzata il contenuto del corpus
  - compiere nuove forme di elaborazione e computazione sui dati lingustici

corpus = corpus elettronico

i testi sono in formato digitale (machine-readabale)

Tipologia ed uso

Ogni corpus è per sua definizione il risultato di un'opera di selezione

i criteri che guidano questa scelta determinano la natura stessa del corpus e condizionano lo spettro dei suoi usi possibili

- Parametri rilevanti per classificare i corpora
  - generalità
  - modalità
  - cronologia
  - lingua
  - integrità dei testi
  - codifica digitale dei testi

Tipi di corpora

generalità

- corpus specialistico (o verticale)
  - orientato alla descrizione di una particolare varietà del linguaggio (sublanguage) o a un ristretto dominio applicativo
    - linguaggio giornalistico
    - linguaggio infantile
    - linguaggio giuridico, medico, ecc.
    - linguaggio dei controllori di volo, ecc.
- corpus generale o di riferimento (reference corpus)
  - trasversale rispetto alle diverse varietà di un linguaggio *L*
  - plurifunzionale
  - orientato a rappresentare tutti gli aspetti caratteristici di *L*, proponendosi come risorsa di riferimento per la descrizione di *L*
  - può essere organizzato in vari sottocorpora specializzati per varietà di *L*

Tipi di corpora

modalità

- corpus di scritto
  - solo testi di linguaggio scritto
- corpus di parlato
  - solo trascrizioni di linguaggio parlato
- corpus misto (e.g. British National Corpus (BNC))
  - testi scritti e trascrizioni di parlato (in proporzioni variabili)
- speech database (corpus audio, e.g. TIMIT)
  - campioni di linguaggio parlato in forma di segnale acustico (più eventualmente la trascrizione ortografica)
- corpus multimediale (audio-video)
  - testi scritti, video, parlato in forma di segnato acustico, ecc.

Tipi di corpora

cronologia e lingua

- corpus sincronico (e.g., Brown Corpus)
  - descrive un particolare stadio del linguaggio (i testi appartengono tutti ad una stessa finestra temporale)
- corpus diacronico (e.g. Tesoro della Lingua Italiana delle Origini)
  - descrive il mutamento linguistico (i testi appartengono a diverse finestre temporali)
- corpus monolingue
  - contiene testi di una sola lingua
- corpus bi/plurilingue (traduttori automatici)
  - corpus parallelo – lo stesso testo è rappresentato (in traduzione) in più di una lingua
    - corpus allineato – ciascuna frase (parola) della lingua L1 è esplicitamente collegata col suo traducente nella lingua L2
  - corpus comparabile – testi in più lingue (non in traduzione) appartenenti alle stesse tipologie (ciascuna lingua è rappresentata da testi diversi)

Tipi di corpora

integrità e codifica dei testi

- Un corpus può contenere testi interi o porzioni di testi di lunghezza prefissata (bilanciamento del corpus vs. naturalezza dei dati)

- Corpora codificati
  - i testi sono arricchiti con etichette (codici) che ne rendono esplicite vari tipi di informazione (es. struttura testuale, composizione, ecc.)
- Corpora annotati
  - le informazioni codificate sul testo riguardano la struttura linguistica del testo a livelli diversi di rappresentazione (es. morfologica, sintattica, semantica, ecc.)

Dimensione del corpus

- Numero di parole contenute nel corpus
  - numero di ore di registrazione, per corpora di parlato
- Regola generale: "*The larger, the better!*"

Evoluzione della dimensione dei corpora

corpora di prima generazione

anni 60-70	milioni di parole

corpora di seconda generazione

anni 80-90	decine di milioni di parole
2000-oggi	centinaia di milioni di parole

corpora di ultima generazione

oggi - …	miliardi di parole

Dimensione del corpus

- corpus chiuso
  - corpus standard tradizionale
  - la quantità di testi e di parole è prefissata all'inizio del progetto
  - corpus statico : "fotografa" un particolare stadio linguistico
- corpus aperto (monitor corpus, Sinclair 1991, e.g. Bank of English)
  - nuovi testi sono continuamente aggiunti alla collezione, secondo le proporzioni decise in fase progettuale
  - corpus dinamico, ideale per studiare l'evoluzione del linguaggio (dinamiche del lessico)

CORPORA

ALCUNI ESEMPI

Corpora di prima generazione

Brown Corpus

- Il primo corpus computazionale in formato elettronico, iniziato nel 1961
  - Francis e Kucera (Brown University)
  - corpus standard di American English contemporaneo
- Dimensione
  - 1 milione di parole tratte da materiale pubblicato nel 1961 appartenente a vari generi
- Tratti caratteristici:
  - generale
  - sincronico
  - monolingue
- Registrato su 100.000 schede perforate e trasferito su nastri magnetici nel 1964. Disponibile su CD-ROM
- Modello di riferimento per tutti i corpora di prima generazione

Corpora di prima generazione

Brown Corpus

A01 0010 1 The Fulton County Grand Jury said Friday an investigation
A01 0020 1 of Atlanta's recent primary election produced "no evidence"
A01 0020 9 that any irregularities took place.
A01 0030 5 The jury further said in term-end presentments that
A01 0040 3 the City Executive Committee, which had over-all charge
A01 0050 2 of the election, "deserves the praise and thanks of
A01 0050 11 the City of Atlanta" for the manner in which the election
A01 0060 11 was conducted.
A01 0070 1 The September-October term jury had been charged
A01 0070 9 by Fulton Superior Court Judge Durwood Pye to investigate
A01 0080 8 reports of possible "irregularities" in the hard-fought
A01 0090 6 primary which was won by Mayor-nominate Ivan Allen
A01 0100 5 Jr.

Corpora paralleli

- Canadian Hansard Corpus (2001)
  - 1,3 milioni di frasi francesi-inglesi allineate a livello di frase, tratte dagli atti del Parlamento Canadese

Monday, September 22, 1997	Le lundi 22 septembre 1997
FIRST SESSION-36TH PARLIAMENT	PREMIÈRE SESSION-36E LÉGISLATURE
The 35th Parliament having been dissolved by proclamation on Sunday, April 27, 1997, and writs having been issued and returned, a new Parliament was summoned to meet for the dispatch of business on Monday, September 22, 1997, and did accordingly meet on that day.	La trente-cinquième législature ayant été prorogée et les Chambres dissoutes par proclamation le dimanche 27 avril 1997, puis les brefs ayant été émis et rapportés, les nouvelles Chambres ont été convoquées pour l'expédition des affaires le lundi 22 septembre 1997 et, en conséquence, se sont réunies le jour dit.
Monday, September 22, 1997	Le lundi 22 septembre 1997.
This being the day on which Parliament was convoked by proclamation of His Excellency the Governor General of Canada for the dispatch of business, and the members of the House being assembled:	Le Parlement ayant été convoqué pour aujourd'hui, par proclamation de Son Excellence le Gouverneur général du Canada pour l'expédition des affaires, et les députés étant réunis:
Robert Marleau, Esquire, Clerk of the House of Commons, read to the House a letter from the Administrative Secretary to the Governor General informing him that the Right Honourable Antonio Lamer, in his capacity as Deputy Governor General, would proceed to the Senate chamber to open the first session of the 36th Parliament of Canada on Monday, September 22 at Ottawa.	M. Robert Marleau, greffier de la Chambre, donne lecture d'une lettre du directeur administratif du Gouverneur général annonçant que le très honorable Antonio Lamer, à titre de suppléant du Gouverneur général, se rendra à la salle du Sénat le lundi 22 septembre 1997, à Ottawa, pour ouvrir la première session de la trente-sixième législature.

Corpora specialistici

- Switchboard Corpus (1992)
  - 2.400 conversazioni telefoniche registrate in varie regioni degli USA e trascritte (ca. 3 milioni di parole)
  - applicazioni: Automatic Speech Recognition (ASR), Speaker Identification, ecc.
- Child Language Data Exchange (CHILDES) (B. MacWhinney)
  - database di interazioni conversazionali di bambini in fase di apprendimento linguistico o di soggetti con patologie del linguaggio
  - finalità: *studio dell'apprendimento linguistico*
  - "meta-corpus":
    - sistema per la raccolta, trascrizione e trattamento di di dati linguistici
    - collezione di dati aperta
  - http://childes.psy.cmu.edu/

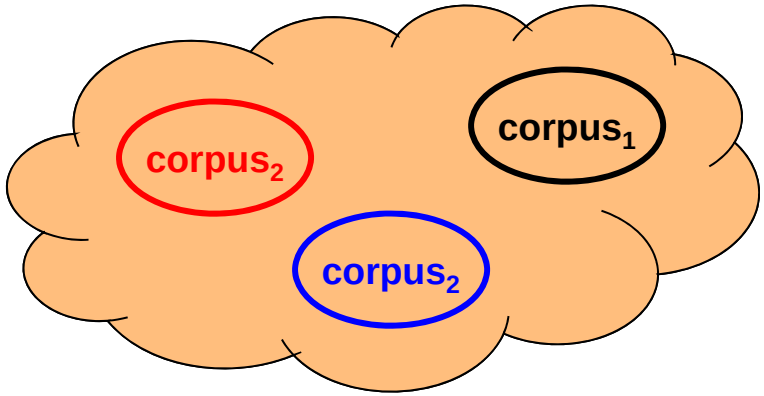
Risorse di corpora

Corpora di grandi dimensioni e di varie tipologie esistono per un numero crescente di lingue

- Agenzie per la distribuzione di corpora
  - Language Data Consortium (LDC)
    - http://www.ldc.upenn.edu/
  - European Language Resources Association (ELRA)
    - http://www.elra.info/
- Consultazione on-line di corpora (a pagamento)
  - Sketchengine (http://www.sketchengine.co.uk)
- Liste di corpora:
  - http://devoted.to/corpora

Il corpus come campione

- Il linguaggio è un sistema potenzialmente illimitato
  - è possibile comprendere e generare un numero potenzialmente infinito di frasi
  - in termini statistici:
    - le frasi di un linguaggio formano una popolazione infinita
- Un corpus è una porzione finita di un linguaggio dalla quale cerchiamo di ricostruire le proprietà dell'intero sistema
  - in termini statistici il corpus è un campione di un linguaggio



Il corpus come campione

Un corpus è una raccolta di testi selezionati e organizzati secondo espliciti criteri, allo scopo di essere usata come campione rappresentativo del linguaggio o di una sua particolare varietà

(McEnery & Wilson 2001, *Corpus Linguistics*)

- Una biblioteca digitale o un archivio di testi elettronici NON è un corpus
- Concetto chiave:
  - rappresentatività del corpus

Il corpus come campione

- Popolazione
  - l'insieme di tutte le entità, con particolari caratteristiche, che sono oggetto di indagine
    - es. *gli studenti dell'Università di Palermo*
- Campione
  - sottoinsieme della popolazione
    - es. A = {studenti maschi di Lettere con meno di 20 anni}
    - *Problema*: il campione A non è rappresentativo della popolazione!!
- Popolazione linguistica
  - insieme di tutti i testi che appartengono ad un linguaggio *L* o a una sua varietà, oggetto di studio
    - es. il linguaggio sportivo, l'italiano, ecc.
- Corpus
  - un campione (rappresentativo) del linguaggio

Corpus e rappresentatività

- Un corpus è un campione rappresentativo di una data popolazione linguistica se e solo se:

  – fornisce un modello delle proprietà linguistiche della popolazione, ovvero

  1. è in grado di restituire un quadro il più accurato possibile delle varietà e tendenze linguistiche della popolazione, rispettandone le proporzioni
  2. permette di generalizzare (induttivamente) le proprietà linguistiche del corpus (es. distribuzione dei termini lessicali, tipologia di strutture sintattiche, ecc.) all'intera popolazione

Corpus e rappresentatività

A corpus seeks to represent a language or some part of a language. The appropriate design for a corpus therefore depends upon what is meant to represent. Representativeness of the corpus, in turn, determines the kind of research questions that can be addressed and the generalizability of the results of the research.

Biber (1998): 246

- Per essere rappresentativo di una lingua o varietà un corpus deve tenere traccia dell'intero ambito di variabilità dei suoi tratti e proprietà
- Corpus linguistics
  - tenta di definire criteri scientifici per la selezione di campioni di testi rappresentativi.

Rappresentatività e tipi di corpora

La complessità dell'operazione di selezione dipende dalla generalità della lingua che il corpus deve rappresentare

- Corpora specialistici
  - varietà ristrette di lingua
    - i corpora per lo studio della lingua di un autore
    - i corpora di domini linguistici settoriali (ad es. il gergo dei controllori del traffico aereo, ecc.)
    - i corpora di testi che appartengono a generi particolari (ad es. sms, bollettini meteorologici, notiziari stampa, ecc.)
  - la variabilità interna limitata e l'elevata omogeneità linguistica garantiscono la possibilità di ottenere un alto grado di rappresentatività

Rappresentatività e tipi di corpora

- Corpora generali (Biber 1993)
  - devono essere diversificati rispetto a un ampio spettro di tipi testuali
    - corpora bilanciati
  - 200 milioni di parole di uno stesso tipo testuale non costituiscono un corpus di riferimento
- Bilanciamento di un corpus
  - presuppone la creazione di una "mappa" che fornisca una descrizione accurata della popolazione linguistica di riferimento:
    - confini spaziali e temporali (quali testi sono inclusi o esclusi dalla popolazione)
    - tipologia dei testi (l'articolazione in strati della popolazione)
  - "random sampling" di testi appartenenti alle varie categorie individuate nella popolazione
    - ogni categoria deve essere rappresentata

Network of European Reference Corpora (NERC 1995)

- Criteri per il design di un corpus di riferimento plurifunzionale:
  - ampie dimensioni
  - sia testi scritti che testi di parlato trascritto
  - “argomento” come criterio privilegiato di selezione dei testi
  - testi completi
  - documentato in maniera estensiva
  - la sua organizzazione dovrebbe facilitare la creazione di sotto-corpora
  - distribuzione delle proporzioni di testi determinata da considerazioni pragmatiche di disponibilità del materiale testuale in formato digitale

“Knowing that your corpus is unbalanced is what counts”

(Atkins et al. 1992)

- Tecniche statistiche di campionamento possono aumentare il grado di rappresentatività di un corpus, ma…
- Gran parte delle scelte è condizionata da fattori pragmatici
  - budget, limiti temporali o tecnologici
  - disponibilità di materiale, ecc.
- La rappresentatività e la nozione di corpus bilanciato restano concetti limite, valori di riferimento ideali
  - definire i limiti di una lingua è spesso estremamente arduo
  - nessun corpus è bilanciato in senso assoluto
- Regola fondamentale: Conosci il corpus!!!
  - ruolo cruciale della documentazione che accompagna il corpus per conoscere la composizione interna del corpus e i criteri di campionamento dei testi

“More data is better data”

il Web come corpus

Kilgarriff e Grefenstette (2003)

Language	Web size	Language	Web size
Albanian	10,332,000	Catalan	203,592,000
Breton	12,705,000	Slovakian	216,595,000
Welsh	14,993,000	Polish	322,283,000
Lithuanian	35,426,000	Finnish	326,379,000
Latvian	39,679,000	Danish	346,945,000
Icelandic	53,941,000	Hungarian	457,522,000
Basque	55,340,000	Czech	520,181,000
Latin	55,943,000	Norwegian	609,934,000
Esperanto	57,154,000	Swedish	1,003,075,000
Roumanian	86,392,000	Dutch	1,063,012,000
Irish	88,283,000	Portuguese	1,333,664,000
Estonian	98,066,000	Italian	1,845,026,000
Slovenian	119,153,000	Spanish	2,658,631,000
Croatian	136,073,000	French	3,836,874,000
Malay	157,241,000	German	7,035,850,000
Turkish	187,356,000	English	76,598,718,000

Table 3
Estimates of web size in words as indexed by Altavista for various languages

- Il Web è veramente un corpus?
  - non è un campione rappresentativo, ma è indubbiamente una risorsa inesauribile di dati linguistici

Il Web come corpus

- Il Web come risorsa di materiali testuali per costruire corpora
  - particolarmente utile per costruire corpora rapidamente e per particolari domini specialistici
    - cf. BootCat (Baroni e Bernardini 2003)

- Il Web come fonte di dati linguistici
  - i motori di ricerca commerciali possono essere usati per ottenere dati quantitativi su vari tipi di espressioni linguistiche
    - permette di ottenere informazioni su neologismi o mutamenti lessicali
      - 176.000 hits di pagine web che contengono il verbo *messaggiare* su Yahoo!!
  - sono attestati anche errori ortografici …
    - 29.200.000 hits su Yahoo!! per *coscienza*, "appena" 364.000 per *coscenza*

“Googleology is bad science”

(Kilgarriff 2007)

- L’uso dei motori di ricerca commerciali (Google, Yahoo!!, ecc.) per raccogliere dati linguistici deve essere fatto con cautela e non è sempre affidabile
  - i risultati dipendono dalle specifiche caratteristiche dei motori di ricerca
    - algoritmi di indicizzazione e di recupero delle pagine
  - sono spesso non replicabili
    - motori di ricerca differenti danno risultati molto diversi
    - lo stesso motore di ricerca produce numeri differenti a breve distanza di tempo
  - le statistiche non sono affidabili per la presenza di duplicati, “headers”, ecc.
  - i risultati sono in termini di “hits” (pagine web) e non dell’effettiva frequenza dell’espressione linguistica

Corpora di ultima generazione
web corpora

- Web 1T 5-gram (Google Inc.)
  - dimensione: ca. 1 tera (1.000 miliardi) di parole
  - testi inglesi derivati dal Web

- It-Wac (Baroni & Ueyama 2006)
  - dimensione: ca. 2 giga (miliardi) di parole
  - testi italiani scaricati dal Web
  - annotati automaticamente
    - lemmatizzati e annotati a livello morfosintattico

Possibili Domande di Esame

- Il candidato risponda alle seguenti affermazioni con **vero** o **falso**:
  - I corpora specialistici hanno il massimo grado di generalità
  - Un corpus misto contiene le registrazioni audio-video di scambi comunicativi
  - Un corpus sincrono include testi che appartengono a una stessa finestra temporale
  - Un corpus dinamico non è sicuramente adatto per studiare l'evoluzione del linguaggio
  - In un corpus parallelo i testi sono arricchiti con etichette relative alla struttura sintattica del testo
- Il candidato risponda alle seguenti affermazioni con **vero** o **falso**:
  - Il Web non è propriamente un corpus
  - Un corpus è bilanciato solo e soltanto il numero di informatori di sesso maschile è uguale a quello degli informatori di sesso femminile
  - Google è un corpus per la lingua inglese
  - Un corpus è un dato linguistico controllato
  - I dati controllati sono raccolti in condizioni ideali