

# Il web come corpus

Il world wide web può essere considerato un corpus?  
Quali sono i suoi limiti? E le sue potenzialità?

# WWW

## Enorme deposito di materiale testuale

- accessibile, gratuito, variato negli stili, nelle tipologie e nei contenuti linguistici rappresentati

Ma il web può essere considerato un enorme corpus linguistico di riferimento?

Quali sono i limiti e le possibilità dell'uso di materiale fornito da internet come base per studi di linguistica dei corpora?

Il web come corpus è rappresentativo delle lingue ivi presenti?

# *Il web è un corpus?*

## *Sì*

- sì, poiché raccoglie testi in varie lingue e tali testi sono una selezione di occorrenze comunicative appartenenti a diverse tipologie testuali (manuali, notizie, blog, dialoghi in forum, ecc.)

## Estensione

- il web è forse uno dei corpora più grandi mai raccolti
- l'esatta composizione in termini quantitativi è ancora indeterminata

## Lingue

- l'inglese si stima copra il 70% circa delle pagine, seguito da giapponese, tedesco, francese e cinese

# *Il web è un corpus rappresentativo?*

## *Corpus di riferimento*

- il web non presenta le diverse varietà della lingua
- vi sono privilegiati (quantitativamente) alcuni domini
  - tecnologie, news
- ne sono quasi del tutto assenti altri
  - parlato spontaneo, varietà regionali

## Interrogazione attraverso i motori di ricerca

- come per esempio Google o Altavista
- essi utilizzano algoritmi di selezione che privilegiano determinati tratti (come il numero di accessi degli utenti e il numero di citazioni e riferimenti esterni al sito) per ordinare e selezionare i documenti
- presenza/assenza del materiale sul web
  - le pagine spariscono, ma soprattutto aumentano vertiginosamente
- frequenti duplicazioni del materiale stesso
- capacità di recupero dei documenti e delle informazioni
- costituiscono un limite alla rappresentatività del web come corpus.

No, non è un corpus rappresentativo delle varietà della lingua

# L'errore

Il tasso di errori è significativamente più alto rispetto ad altri testi che possono essere raccolti e controllati

- non è solo l'errore di battitura o di scannerizzazione che si ritrova in qualunque testo elettronico
- la presenza massiccia di pagine amatoriali, spesso scritte da utenti che non governano bene la lingua perché costituisce una lingua seconda (specialmente per l'inglese), finisce per costituire una rappresentazione molto sbilanciata delle caratteristiche linguistiche dei testi.

## Esempio

- e digito *information retrieval* (invece di *retrieval*) ottengo da Google
- 15.900 risultati
- e ben 181.000 per il solo *retrieval*
- questo non solo significa che, con l'interrogazione scorretta, ottengo risultati fuorvianti, ma anche che con l'interrogazione corretta mi perdo una quantità di risultati utilizzabili

# *Dinamicità incontrollata*

## Dimensione indeterminata

- ...e forse indeterminabile
- molte pagine spariscono
- ...e molte appaiono ogni istante
- l'estrazione, comparazione, ripetibilità dell'analisi dei dati linguistici estratti risulta dunque particolarmente aleatoria

## Limiti all'uso dei motori di ricerca comuni

- l'insufficienza dei risultati visualizzati rispetto a quelli individuati dal motore;
- l'insufficienza del contesto presentato
- la selezione viene fatta sulla base di algoritmi di rilevanza che non usano criteri di tipo linguistico;
- non è possibile condurre ricerche con requisiti linguistici (come la categoria grammaticale, per esempio)
  - la mancanza di criteri linguistici non permette di estrarre lessemi, ma solo forme grafiche, non permette la disambiguazione degli omografi (testuali e assoluti), non permette l'ordinamento per rilevanza esclusivamente linguistica.
- le statistiche di frequenza presentate non sono valide dal punto di vista linguistico per via dei criteri di rilevanza degli algoritmi usati. Killgarriff e Greffentette (2003, pp.12-13)

# Progetti

WebCorp (<http://www.webcorp.org.uk/>)

- di Andrei Kehoe e Antoinette Renouf dell'Università di Liverpool

KWiCFinder (<http://miniappolis.com/>)

Gsearch (<http://www.hcrc.ed.ac.uk/gsearch/>)

The Linguist's Search Engine (<http://lse.umiacs.umd.edu:8080/>)

- di Philip Resnik e Aaron Elkiss

Wacky project (<http://wacky.sslmit.unibo.it/>)

GoogleLing

- di Joseph Smarr e Tim Grow

# *Questioni sul Web as a Corpus*

Sistemi di selezione dei risultati in modo che appaiano bilanciati

Accuratezza e validità dei risultati ottenuti

Possibilità di vedere realizzati motori che possano trattare lingue diverse dall'inglese

Possibilità di operare ricerche avanzate su categorie linguistiche di vario genere