Lingua, statistica e computazione

Isabella Chiari
Facoltà di Scienze Umanistiche
Università “La Sapienza” di Roma
Roma, 25 novembre 2005

UNIVERSITÀ DEGLI STUDI DI ROMA
“LA SAPIENZA”
DIPARTIMENTO DI MECCANICA E AERONAUTICA

Via Eudossiana, 18 – 00184 Roma

Obiettivi

- Osservare alcune interazioni tra linguistica generale, linguistica dei corpora e linguistica computazionale
- Vedere come la linguistica e la statistica entrano nella costruzione di strumenti computazionali efficaci
- Riflettere sulle ricadute che l’incontro tra statistica e computazione ha sulla nostra rappresentazione delle lingue natura[li]

Lingua e calcoli: gli ostacoli

- *Creatività linguistica*:
  - Apertura del sistema
    - Inventario
    - Regole
  - Apertura degli usi
  - Ridondanza
- *Asimmetrie*:
  - Omografie/omofonie
  - Sinonimie
  - Ambiguità sintattiche
  - Riferimenti anaforici
  - Polirematiche e collocazioni

Che ruolo ha la statistica nel linguaggio?

- Dal punto di vista descrittivo
  - Che cosa e quanto può essere colto delle lingue attraverso l'osservazione dei fenomeni dal punto di vista quantitativo e statistico?
- Dal punto di vista dei locutori
  - Negli usi produttivi
  - Negli usi ricettivi
- Si modifica la nostra nozione di grammatica?
- Si modifica la nostra nozione di competenza o conoscenza linguistica?

La struttura statistica dei suoni

- sistemi stenografici e crittografici
  - della Porta, Wilkins, marchese de Viaris, Sacco, Piccoli
- tecniche di compressione dei testi
- fonologia statistica
  - scuola di Praga (Mathesius, Trnka e Vachek) Zipf
- grafemica statistica
  - Teoria dell'informazione (Shannon e Weaver, Miller, Chapanis, ecc.)
- Frequenze assolute e relative di foni e fonemi
- Struttura statistica della fonotassi (restrizioni nelle sequenze fonologiche)

Esempi

1) Parole crociate crittografate: niente definizioni, numero al posto delle lettere

*2) _L L_NG_ _GG_ _ / _ / G_V_RN_T_ / D_LL_ / PR_B_B_L_T_*

- La capacità di inserimento dipende dalle regole della grafemica, della fonotassi e dalla prevedibilità a livello morfologico e testuale

3)Una strufia dutra ha scriciato predumente un ciutro e parpa un ciutrino

- Le parole sono tutte possibili per la fonotassi italiana. Non è violata alcuna restrizione
- Serbatoio potenziale

La struttura statistica del lessico

- Zipf, Guiraud, Muller, Herdan
- Lessici di frequenza
  - Kaeding (1897) Thorndike (1921, 1931-32) Vander Beke (1930) Kučera e Francis (1967)
  - Italiano: LIF - (1971) LIP (1993)
- I dizionari fondamentali
- Il Vocabolario di base
  - Italiano: De Mauro (1980)
- Glottocronologia (Swadesh)

Lista di frequenza del primo capitolo dei *Promessi Sposi*

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

Isabella Chiari - Lingua, statistica e computazione (2005)

Statistica della morfologia e della sintassi

- G. U. Yule e di Herdan
- quantificazione della presenza delle diverse parti del discorso in varie tipologie testuali
- lunghezza della frase
- tipi di sintagmi e dell'ordine delle parole
- produttività morfologica (Baayen)

La struttura statistica dei testi

- ricchezza lessicale (Guiraud, Cossette)
- valutazione della leggibilità (Flesh, Gulpease)
- relazione tra lunghezza delle parole o delle frasi e testo (Zipf, Mandelbrot)
- concentrazione delle frequenze lessicali – *textual profiling*
- stilometria
- analisi delle specificità

Linguistica statistica e linguistica dei corpora

- L'estrazione di dati e strutture statistiche ha come dominio privilegiato il TESTO
  - *Linguistica corpus-based*
  - *Linguistica corpus-driven*
- Corpora di riferimento e special-purpose
- Web come corpus

Linguistica computazionale

- **Lessicografia computazionale**
  - Dizionari macchina
  - Dizionari elettronici
  - I dizionari basati su corpora
- ***Natural Language Processing* e *Natural Language Generation***
  - Tagging grammaticale
  - Parsing sintattico
- **Traduzione automatica dei testi**
- **Tecnologie della lingua parlata**
  - Sintesi del parlato (*Speech Synthesis*)
  - Riconoscimento del parlato (*Speech Recognition*)

La linguistica dei corpora è linguistica computazionale?

Sì, non potrebbe esistere senza strumenti di LC

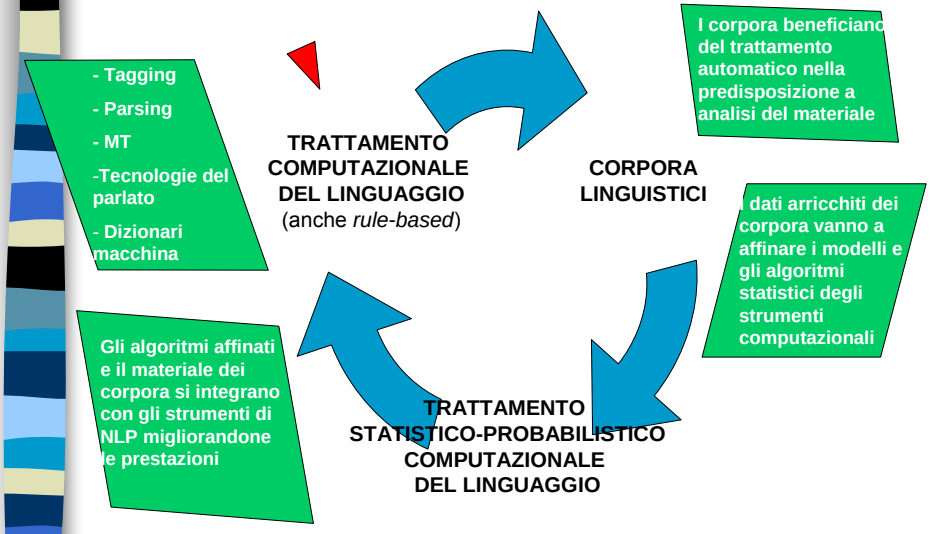
- Strumenti computazionali per la **costruzione** e **preparazione** dei corpora
  - Tagging
  - Parsing
  - lemmatizzazione
- Strumenti computazionali per **l'analisi dei dati**
  - interrogazioni avanzate
  - liste di frequenza e concordanze
  - analisi statistica dei dati testuali
  - information retrieval e extraction

La linguistica dei corpora è linguistica computazionale?

Sì, la LC non potrebbe esistere senza l'accesso ai corpora

- **Lessicografia elettronica** *corpus-based*
- **Training corpora** per il **NLP**
  - Taggers e parsers con training corpora
- **Traduzione automatica**
  - Corpus-based
  - Example-based Machine Translation
- **Tecnologie del parlato**
  - Addestramento allo Speech Recognition
  - Sintesi corpus-based
- **Machine Learning**
  - Individuazione automatica di pattern estratti dai dati

Il circolo virtuoso della LC



Questioni finali

- Se, di fatto, la linguistica computazionale basata su regole è meno potente di quella di tipo statistico e corpus-based, è lecito domandarsi anche se:
  - La produzione e la ricezione siano meccanismi eminentemente probabilistici?
  - Quanta parte delle nostre attività linguistiche è diretta da operazioni di tipo basato su regole e quanta da operazioni probabilistiche?
- Esiste una competenza linguistica di tipo probabilistico?
  - Se sì, è una competenza grammaticale?
  - Quanta di questa competenza riposa su riflessioni di tipo metalinguistico?

Questioni finali

- La struttura statistica delle lingue e la competenza statistica delle lingue non può essere un semplice prodotto dei modi di acquisizione delle lingue e dell'abitudine? (innatismo/culturalità)
- Come si configurerebbe un modello della produzione e ricezione linguistica di tipo probabilistico?
  - Non assomiglierebbe ai modelli di descrizione usati per le scienze della vita (biologia, neuroscienze)?
  - La ridondanza, come risultato della non-aleatorietà dei sistemi, non è forse presente egualmente in qualunque forma vivente? (ridondanza genetica, ridondanza biologica, ecc…)?