Misurare la Produttività: Esperimenti e Riflessioni

Marco Baroni
SSLMIT (Università di Bologna)
baroni@sslmit.unibo.it

Lunedì 14 febbraio 2005
MorBoConf, Bologna

1 Introduzione

- "Morphological productivity is the property of a morphological process to give rise to new formations on a systematic basis" (Plag 2004).
- Nozione centrale per ragioni teoriche e pratiche:
  - Teoria: identificare processi "vivi" in una lingua; formulazione di regole in acquisizione del linguaggio;
  - Pratica: forme nuove problematiche per tecniche NLP basate su corpora.
- I processi morfologici possono avere *gradi* di produttività diversi.
- L'approccio *quantitativo* (Baayen e altri) usa dati estratti da corpus per misurare gradi di produttività.

1.1 Alcune questioni per l'approccio quantitativo

1.1.1 Elaborazione dei dati

- Automatica: molti errori, spesso in parole a bassa frequenza, forme derivate, i.e., dove possono avere effetti sistematici su computo di produttività (Evert/Lüdeling 2001; Gaeta/Ricca 2003).
- Manuale:
  - Rischio di circolarità (e.g., contare come istanze di prefisso *ri-* solo forme semanticamente trasparenti, formate produttivamente).
  - Troppo lavoro, impossibile andare oltre studi pilota.
- "Robustezza" di misura deve essere criterio di valutazione fondamentale!

1.1.2 Dipendenza di misure da frequenza assoluta

- Misure di produttività dipendono da frequenza assoluta di classi misurate (Gaeta/Ricca 2003).
- Classi tendono ad avere frequenza molto diversa (e.g., *ri-* vs. *meta-*).

1.1.3 Funzioni delle misure di produttività

- Conferma "oggettiva" di intuizione di morfologo (di nuovo: circolarità?)
- Strumento d'esplorazione dei dati?
- Possono *spiegare* la produttività? Forniscono indizi su perché un processo è più produttivo di un altro?
- Misuriamo sintomi o cause della produttività?

1.2 I miei piani per questo talk

- Studio pilota di produttività di 7 prefissi dell'inglese alla luce di questioni appena elencate.
- In particolare:
  - Variazione di misure al variare di livello di elaborazione automatica di dati.
  - Modello parametrico per eliminare effetti della frequenza assoluta.
  - Nuove misure della produttività che cercano di "spiegare" fenomeno in termini di trasparenza semantica.

2 Dati

- I "prefissi di Baayen", con rango medio di produttività assegnato da quattro morfologi di madrelingua inglese:

un	1.500
re	1.625
mis	3.250
de	3.625
be	5.875
en	5.875
in	6.250

- Notare:
  - Tre classi: a produttività libera, a produttività ristretta, non produttivi.

  - Un solo valore per prefisso è forzatura: e.g., *re-* iterativo vs. *re-* rafforzativo; *in-* negativo vs. *in-* locativo (particella?)

- Frequenze e altri dati estratti dal British National Corpus (Aston/Burnard 1998), ∼100M di tokens, "bilanciato", relativamente pulito, categorie morfosintattiche e lemmi assegnati semi-automaticamente.

2.1 Identificazione di parole prefissate

2.1.1 Forme flesse

- Tratta come forma prefissata ciascuna forma nel corpus che inizi con stringa identica a prefisso e la cui base potenziale (la stringa che segue il prefisso) sia di almeno 3 lettere e attestata nel corpus
- Per es., *beads*, ma non *bead* (stemma troppo corto) né *benny* (stemma non attestato come parola autonoma).
- Brutale, ma spesso l'unico approccio possibile.
- Meno distorsioni dovute ad elaborazione automatica.

2.1.2 Lemmi

- Tratta come forma prefissata ciascun lemma nel corpus che inizi con stringa identica a prefisso e la cui base potenziale sia un lemma di almeno 3 lettere attestato nel corpus.
- Per es., *really*, ma non *beads* (stemma sarebbe *ad*, che è troppo corto).
- Statistiche più robuste, meno rumore, ma anche perdita di dati (e.g., le forme di *reacquaint* vengono lemmatizzate come *UNKNOWN*).

2.1.3 Lemmi verbali

- Tratta come forma prefissata ciascun lemma verbale nel corpus che inizi con stringa identica a prefisso e la cui base potenziale sia un lemma verbale di almeno 3 lettere attestato nel corpus.
- Per es., *belong*, ma non *beads* (stemma sarebbe *ad*, che è troppo corto).
- Stessi vantaggi e svantaggi che con i lemmi, ma più accentuati (*really* sparisce, ma sparisce anche *decontamination*).
- *in-* rimane quasi solo come locativo.

2.1.4 Tabella riassuntiva (*types*)

prefissi	forme flesse	lemmi	lemmi verbali
un	3612	845	119
re	2846	601	322
mis	409	121	66
de	1202	307	141
be	366	95	50
en	372	118	62
in	1119	555	62

3 Misure classiche della produttività (Baayen)

3.1 V

- Numero di parole distinte (*types*) che contengono un certo affisso.
- Misura della produttività "storica" dell'affisso.

3.2 $V(1)$

- Numero di *hapax legomena* (parole con frequenza 1) che contengono un certo affisso.
- Ovvero, parole "nuove" (nel corpus, non necessariamente nella lingua) che contengono l'affisso.

3.3 $\mathscr{P}$

$$\mathscr{P} = \frac{V(1)}{N}$$

- Proporzione di hapax legomena su tutte le parole (*tokens!*) che contengono un certo affisso.
- La misura più popolare.
- Giustificazione teorica: derivata di curva di crescita di V.
- "Ritmo" a cui cresce V, l'insieme di tipi contenenti l'affisso di interesse.

3.4 Risultati

- Vedi figure 1, 2 e 3.
- In tutte le figure, valori normalizzati.
- Cambi anche radicali a seconda di data-set (ma non è chiaro chi sia il vincitore).

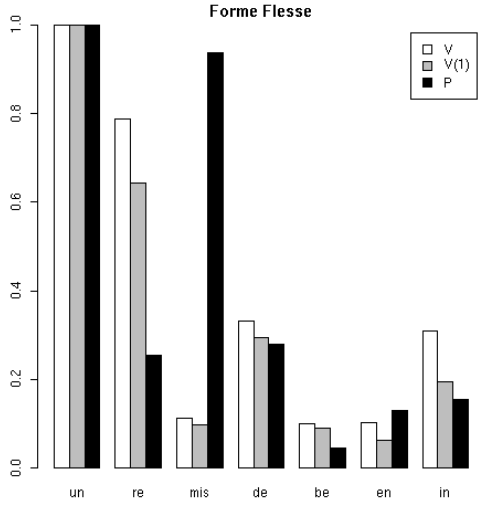

Figura 1: Misure classiche di produttività, forme flesse.

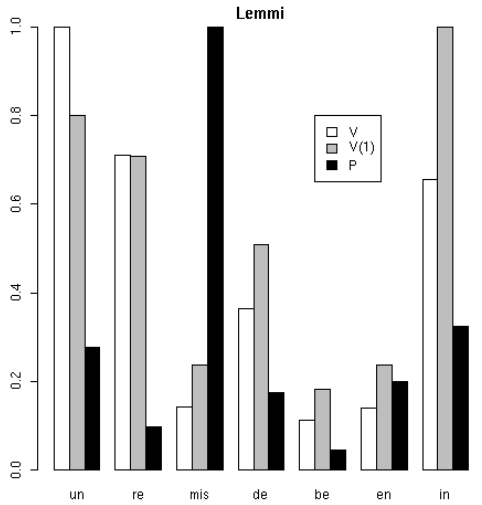

Figura 2: Misure classiche di produttività, lemmi.

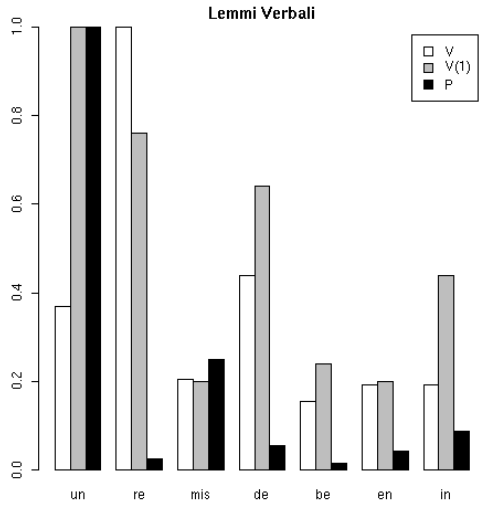

Figura 3: Misure classiche di produttività, lemmi verbali.

- $\mathscr{P}$ sottostima *re-*/sovrastima *mis-*.
- *in-* negativo+locativo vs. *in-* locativo: differenza netta tra data-set lemmi e data-set lemmi verbali.

4 Dipendenza da N delle misure di Baayen

- Gaeta/Ricca (2003).
- Come mostrato in figura 4, V cresce con il numero di tokens di una categoria (dunque, sottostima di morfema raro *mis-*).
- $V(1)$ può avere patterns diversi per classi diverse, ma per suffissi in figura 4 tende a diminuire con N.
- Man mano che N aumenta, il ritmo di crescita di V diminuisce.
- Dunque $\mathscr{P}$, derivata di V, diminuisce (illustrato per tre punti nei grafici in figura 4), onde la sottostima di morfemi frequenti/sovrastima di morfemi rari.

4.1 La soluzione di Gaeta e Ricca

- Paragona $\mathscr{P}$ solo per valori comparabili di N (lo stesso si può fare con V).
- Problemi:

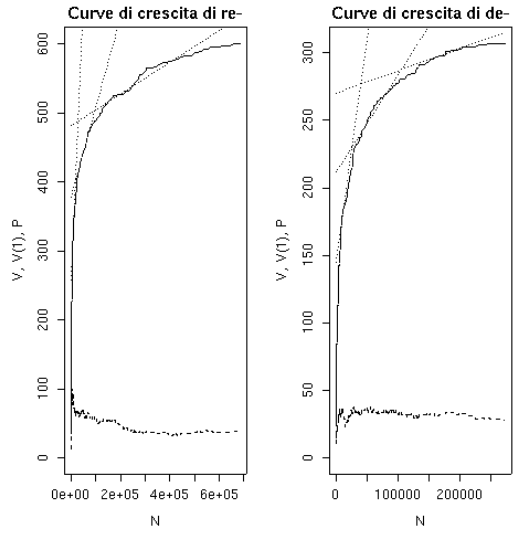

Figura 4: Curve di crescita di V e $V(1)$ e $\mathscr{P}$ come coefficiente angolare della retta tangente a V in N per *re-* e *de-*.

- Si può interpolare, ma non estrapolare.
- Se ordine di parole in corpus non è randomizzato, crescita di V e $V(1)$ è irregolare per effetti di "coagulazione" delle parole (vedi andamento delle curve in figura 4).
- Problemi tecnici nella creazione di sub-corpora?

5 Usare un modello statistico per stimare misure di produttività per N arbitrari

5.1 L'idea di base

- Assumi che data-set sia campione random da popolazione di parole con distribuzione di probabilità la cui forma precisa vada determinata fissando un certo numero (basso) di parametri liberi.
- Stima i parametri della distribuzione sulla base di campione.
- Calcola valori attesi (expectations, medie) di V e $V(f)$ predetti da modello per N arbitrari.

5.2 Il modello Zipf-Mandelbrot finito (fZM)

- Evert (2004a,b).

- Appartiene alla famiglia di distribuzioni LNRE (Large Number of Rare Events), appropriate per frequenze di parole (Baayen 2001).
- Assume che probabilità di parole sia distribuita secondo legge di Zipf-Mandelbrot (Zipf 1949, 1965; Mandelbrot 1953).
- La probabilità π_i della parola in posizione i nella lista di tutte le parole ordinate per probabilità decrescente è data da:

$$\pi_i = \frac{C}{(i+b)^a}$$

- fZM assume inoltre che popolazione sia finita, con parametro extra per probabilità minima che parola può avere.
- Espressioni fZM per $E(V)$ e $E[V(f)]$ con N arbitrario:

$$E[V] = C \cdot N^\alpha \cdot \frac{\Gamma(1-\alpha, NA)}{\alpha} + \frac{C}{\alpha \cdot A^\alpha}(1 - e^{-NA})$$

$$E[V(f)] = \frac{C}{f!} \cdot N^\alpha \cdot \Gamma(f-\alpha, NA)$$

- Parametri di modello possono venire stimati applicando regressione non-lineare direttamente a queste formule con N uguale alle dimensioni del corpus (Evert 2004a).

5.2.1 Il modello fZM è adatto a classi di parole prefissate?

- Non proprio...
- Curve rango/frequenza empiriche dovrebbero assomigliare a quella nel primo riquadro di figura 5 (Brown corpus, Kučera/Francis 1967), che rappresenta la tipica curva di distribuzione Zipf-Mandelbrotiana in spazio logaritmico.
- Inoltre, modello finito è appropriato per affissi produttivi?
- La soluzione ideale: distribuzioni diverse per classi diverse.
- La soluzione pratica: facciamo finta di niente.

5.3 Risultati usando $E(V)$ e $E[V(1)]$

- NB: nel momento in cui paragoniamo i prefissi allo stesso valore di N, $V(1)$ e $\mathscr{P}$ diventano equivalenti ($\mathscr{P}$ è $V(1)$ diviso per N: a parità di N le due misure ordinano i prefissi nella stessa maniera).
- Per ciascun data-set, consideriamo un valore "basso" e un valore "alto" di N (un terzo e tre quarti del valore massimo di N nel data-set).

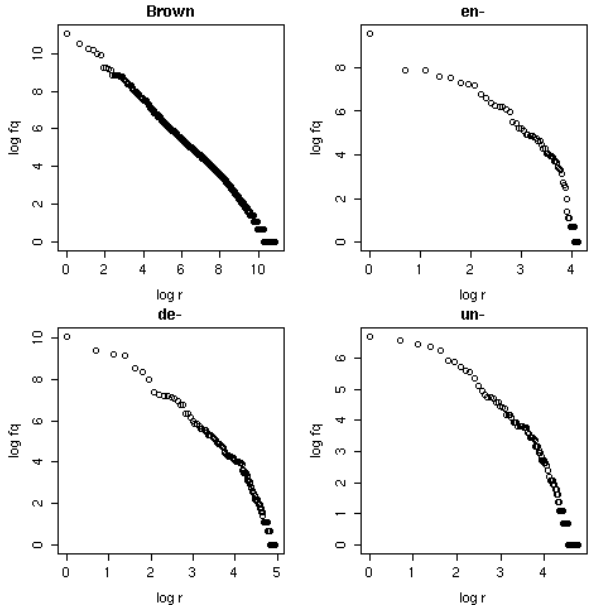

Figura 5: Curve rango/frequenza del Brown corpus e di tre classi di parole prefissate (lemmi verbali).

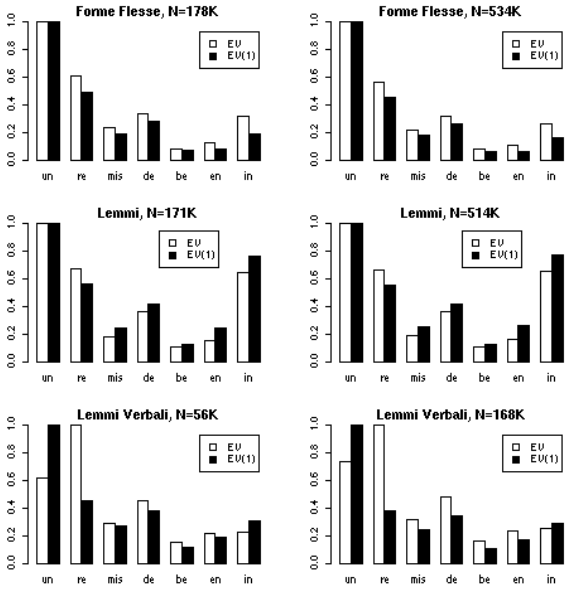

Figura 6: EV e $E[V(1)]$ per valori bassi e alti di N.

- Notiamo valori stabili per N alto e basso.
- Valori abbastanza stabili anche tra data-sets, tranne che per *in-*.
- Valori abbastanza ragionevoli, tranne che per *in-* (ma *in-* è comunque un caso a parte, vedi sopra).
- I livelli successivi di elaborazione (selezione di lemmi, selezione di lemmi verbali) non migliorano i risultati (anzi...)

6 Produttività, trasparenza semantica e contesto

- Prosecuzione di esperimenti (con risultati mediocri) di Baroni/Vegnaduzzo (2003).
- Una teoria sulla produttività: presenza di molte forme semanticamente trasparenti che contengono un affisso fa notare presenza di affisso e formulare generalizzazioni corrette in fase di acquisizione del linguaggio.
- Predizione: alta correlazione tra trasparenza semantica e produttività.
- Trasparenza semantica media di forme con prefisso come misura di produttività di prefisso.
- Misura basata su cause, non sintomi!
- (Ma cosa causa la trasparenza semantica?)

6.1 Misure automatiche della trasparenza semantica

- Misure di *somiglianza semantica* tra due parole in NLP (Manning/Schütze 1999).
- Trasparenza semantica: grado di somiglianza tra forma con affisso e base.
- Approccio contestuale alla semantica.
- Cruse (1986, p. 1):

> [T]he semantic properties of a lexical item are fully reflected in appropriate aspects of the relations it contracts with actual and potential contexts [...] [T]here are are good reasons for a principled limitation to linguistic contexts.

- Due interpretazioni pratiche e knowledge-poor di approccio contestuale:
  - Le parole semanticamente simili capitano in contesti simili.
  - Le parole semanticamente simili capitano una vicina all'altra.

6.1.1 Similarità di contesto

- Coseno (correlazione) di vettori normalizzati che rappresentano frequenza di co-occorrenza con altre parole all'interno di una certa finestra:

$$\cos(\overrightarrow{x}, \overrightarrow{y}) = \overrightarrow{x} \cdot \overrightarrow{y} = \sum_{i=1}^{n} x_i y_i$$

- Nel nostro caso:
  - Targets: coppie forma prefissata/base.
  - Contesti: content words.
  - Finestra: frase.

6.1.2 Co-occorrenza

- Misurata tramite la *Mutual Information* (MI):

$$MI(w_1, w_2) = \log \frac{Pr(w_1, w_2)}{Pr(w_1)Pr(w_2)}$$

- Nel nostro caso:
  - Targets: coppie forma prefissata/base.
  - Finestra: documento.
  - Esperimento anche con frequenze di occorrenza e co-occorrenza prese da Google (per via di enorme mole di dati che contiene, la rete si è rivelata corpus ideale in vari problemi relati – e.g., Turney 2001; Baroni/Vegnaduzzo 2004).

6.1.3 Calcolo delle medie

- Risultati poco convincenti se si prendono in considerazione *tutte* le forme di una classe.
- Il fatto che un sotto-insieme di parole con un certo prefisso abbiano trasparenza semantica alta è più significativo del fatto che altre parole della stessa classe siano opache.
- E.g., *re-* ha sia molte forme trasparenti che molte forme opache.
- Soluzione: calcola media di coseno/MI prendendo in considerazione solo gli n valori più alti in ciascuna classe.
- Nei miei esperimenti, $n = 20$.
- Valore arbitrario: servirebbe approccio più ragionato a questo ed altri parametri.
- A differenza di approcci classici, hapax legomena non giocano ruolo centrale: misure più robuste?

6.2 Risultati

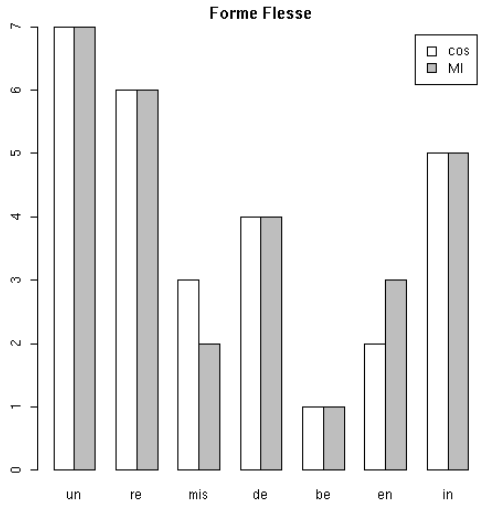

Figura 7: Misure di produttività basate su contesto e co-occorrenza, forme flesse.

- Nelle figure 7, 8 e 9, risultati riportati come ranghi – entità di differenze non significativa.
- I risultati migliori di nuovo con i dati meno elaborati (forme flesse).
- Coseno meglio di MI.
- *in-*, come al solito, più produttivo del previsto.
- Google, a differenza che in altre applicazioni, non è un toccasana.

7 Conclusioni

- Work, ovviamente, in progress.
- Gli stessi metodi vanno testati su altri data-sets.
- Questione di elaborazione/pulizia automatica dei dati rimane aperta.
  - In particolare, sarebbe utile poter distinguere tra funzioni diverse di prefissi (e.g., *re-*, *in-*): si potrebbero applicare metodi usati in Word Sense Disambiguation?
  - Lemmatizzazione/restrizione a una classe morfosintattica non sembrano esser d'aiuto – anzi, forse sono dannosi.

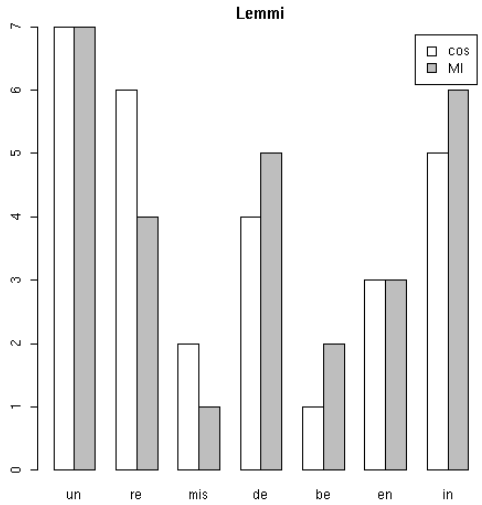

Figura 8: Misure di produttività basate su contesto e co-occorrenza, lemmi.

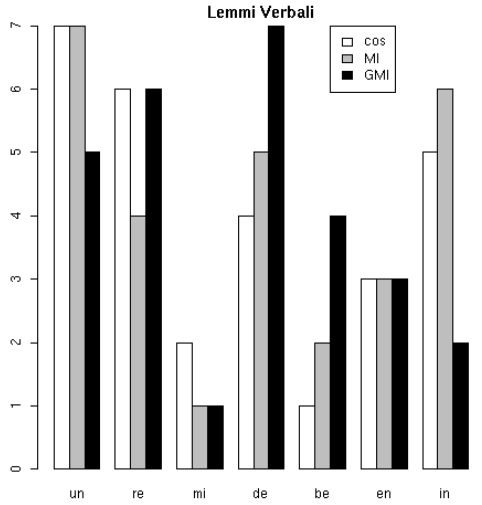

Figura 9: Misure di produttività basate su contesto e co-occorrenza, lemmi verbali.

- Calcolare valori attesi di V e $V(1)$ (sia pur assumendo modello statistico di dubbia validità) sembra essere un metodo efficace per evitare problemi dovuti a dipendenza di misure da N.
- Sarebbe interessante paragone diretto con metodo Gaeta/Ricca.
- Tra i metodi basati su somiglianza semantica, metodo basato su similarità di contesto sembra migliore: si potrebbero provare altre misure oltre a coseno.
- Validità di assunzione alla base di approccio (coseno/MI misurano trasparenza semantica) va testata.
- Un problema al cuore dello studio della produttività: se misure vogliono essere metodo oggettivo per quantificare produttività, nel caso in cui siano in disaccordo con intuizioni di morfologi, dobbiamo fidarci delle misure o dei morfologi?
- Convergere di misure diverse agli stessi valori sarebbe criterio di valutazione più convincente (anche: utilizzo in applicazioni concrete?)
- Convergenza di misure di Baayen e misure basate sul contesto agli stessi valori può essere indizio che siamo sulla strada giusta (ma bisogna accertarsi che non ci siano dipendenze nascoste tra i due gruppi di misure).

Riferimenti bibliografici

Aston, Guy/Burnard, Lou (1998). *The BNC Handbook.* Edinburgh: Edinburgh University Press.

Baayen, Harald (1991), Quantitative aspects of morphological productivity. In: *Yearbook of Morphology 1991*, 109-149.

Baayen, Harald (1996), The effects of lexical specialization on the growth curve of the vocabulary. In: *Computational Linguistics* 22, 455-480.

Baayen, Harald (2001), *Word frequency distributions*. Dordrecht: Kluwer.

Baayen, Harald/Lieber, Rochelle (1991), Productivity and English derivation: a corpus-based study. In: *Linguistics* 29, 801-843.

Baroni, Marco (2003), Distribution-driven morpheme discovery: A computational/experimental study. In: *Yearbook of Morphology 2003*, 213-248.

Baroni, Marco/Bernardini, Silvia/Comastri, Federica/Piccioni, Lorenzo/Volpi, Alessandra/Aston, Guy/Mazzoleni, Marco (2004), Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. In: *LREC 2004*.

Baroni, Marco/Vegnaduzzo, Stefano (2003), Assessing morphological productivity via automated measures of semantic transparency. Presentazione orale al Workshop on Explaining Productivity di *DGFS 2003*.

Baroni, Marco/Vegnaduzzo, Stefano (2004), Identifying subjective adjectives through web-based mutual information. *KONVENS 2004*.

Cruse, Alan (1986), *Lexical semantics*. Cambridge: Cambridge University Press.

Evert, Stefan (2004), A simple LNRE model for random character sequences. In: *JADT 2004*.

Evert, Stefan (2004), *The statistics of word cooccurrences: Word pairs and collocations*. PhD thesis, University of Stuttgart/IMS.

Evert, Stefan/Lüdeling (2001), Measuring morphological productivity: Is automatic preprocessing sufficient? In: *Corpus Linguistics 2001*, 167-175.

Gaeta, Livio/Ricca, Davide (2003), Corpora testuali e produttivit morfologica: i nomi d'azione italiani nelle annate della Stampa. In: *Parallela IX*.

Kageura, Kyo (1998), A Statistical Analysis of Morphemes in Japanese Terminology. In: *COLING-ACL 98*, 638-645.

Kučera, Henry/Francis, Nelson (1967), *Computational analysis of present-say American English*. Providence: Brown University Press.

Li, Wentian (2002), Zipf's Law everywhere. In: *Glottometrics* 5, 14-21.

Mandelbrot, Benoit. 1953. An informational theory of the statistical structure of languages, in Willis Jackson (ed.) *Communication Theory*. London: Butterworth: 486-502.

Manning, ChristopherSchütze/Hinrich (1999), *Foundations of statistical natural language processing*. Cambridge MA: MIT Press.

Plag, Ingo (2004), Productivity. In corso di stampa in: Keith Brown (a cura di) *Encyclopedia of language and linguistics, 2nd edition*. Oxford: Elsevier.

Turney, Peter (2001), Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *ECML 2001*, 491-502.

Zipf, George Kingsley (1949), *Human behavior and the principle of least effort*. Cambridge MA: Addison-Wesley.

Zipf, George Kingsley (1965), *The psycho-biology of language*. Cambridge MA: MIT Press.