UN CORPUS DELL'ITALIANO SCRITTO CONTEMPORANEO

DALLA PARTE DEL RICEVENTE

***Alessandro Laudanna*^, *Anna M. Thornton**^,**
***Giorgina Brown*^, *Cristina Burani*^, *Lucia Marconi*°**

^ Istituto di Psicologia, CNR, Viale Marx 15, 00137, Roma
* Università degli Studi, Via Camponeschi 2, 67100, L'Aquila
° Istituto per i Circuiti Elettronici, CNR, Via De Marini 6, 16149, Genova

In S. Bolasco, L. Lebart e A. Salem (a cura di), *III Giornate internazionali di Analisi Statistica dei Dati Testuali*. Volume I, pp.103-109. Roma: Cisu

alaudanna@unisa.it
alessandro.laudanna@istc.cnr.it

UN CORPUS DELL'ITALIANO SCRITTO CONTEMPORANEO DALLA PARTE DEL RICEVENTE

Alessandro Laudanna^, Anna M. Thornton*^,
Giorgina Brown^, Cristina Burani^, Lucia Marconi°

^ Istituto di Psicologia, CNR, Viale Marx 15, 00137, Roma
* Università degli Studi, Via Camponeschi 2, 67100, L'Aquila
° Istituto per i Circuiti Elettronici, CNR, Via De Marini 6, 16149, Genova

In this paper we describe the criteria we adopted for the selection of a corpus composed of 3,000,000 words from Italian contemporary written texts. The corpus will give rise to a frequency dictionary, which should have two main characteristics: i) representativeness of the Italian texts which are actually read, rather than of all possible written texts, ii) usefulness for psycholinguistic research.

KEY WORDS: *Corpora, Frequency, Frequency Dictionary, Psycholinguistics*

1. Il ruolo della frequenza nell'uso del lessico

La frequenza di occorrenza è di centrale importanza in tutti i compiti che richiedono il riconoscimento e la comprensione di parole scritte da parte di persone adulte, sia normali che con disturbi acquisiti del linguaggio. La frequenza è uno dei fattori più importanti nell'influenzare il tempo e l'accuratezza della decisione lessicale[1], il tempo di pronuncia, la durata delle fissazioni oculari. Influenza inoltre le prestazioni di lettura, scrittura, comprensione, produzione ecc. e i processi di acquisizione e sviluppo del linguaggio (per una rassegna si veda Colombo, 1993). Come viene argomentato più in dettaglio in Thornton, Burani e Laudanna (in questo volume), anche la frequenza dei costituenti morfologici della parola, in particolare della radice, ha effetti sul *processing* lessicale. Sono state fornite molte interpretazioni dell'effetto di frequenza, a volte anche sensibilmente diverse tra loro, ma si può sostenere che l'aspetto comune alle varie interpretazioni è che la forza o la disponibilità della rappresentazione di una parola nel lessico mentale è funzione del numero di volte che quella certa parola è stata incontrata. Ciò fa sì che la rappresentazione in memoria delle parole ad alta frequenza sia più rapidamente

[1] La decisione lessicale è un compito sperimentale di vasta utilizzazione nella psicolinguistica, e consiste nel far decidere ad un soggetto se una certa sequenza di lettere o suoni corrisponda o meno ad una parola della propria lingua.

attivabile e più agevolmente discriminabile da quella di altre parole simili dal punto di vista della forma fonica o grafica.

La frequenza è stata oggetto anche di studi di tipo quantitativo o semi-quantitativo: si è trovato che essa spiega all'incirca il 50% della varianza osservata nei compiti sperimentali (Whaley, 1978), che i tempi di riconoscimento di una parola sono una funzione inversa della frequenza su base logaritmica della parola stessa (Scarborough, Cortese e Scarborough, 1977), che i giudizi soggettivi di frequenza dei parlanti sono un migliore predittore delle prestazioni sperimentali rispetto ai valori di frequenza oggettivi riportati nei dizionari di frequenza (Gordon, 1985).

Per determinare con relativa esattezza la frequenza reale di una parola in una data lingua si ricorre solitamente a dei valori di frequenza forniti da dizionari basati su campioni di lingua, il più delle volte ricavati da testi scritti. Perché la frequenza del campione si approssimi nel modo più fedele possibile alla frequenza reale delle parole nella lingua di riferimento è necessario che il campione risponda a criteri di rappresentatività. Quali siano i criteri di rappresentatività auspicabili dal punto di vista della ricerca sperimentale sul lessico è l'oggetto del prossimo paragrafo.

2. Rappresentatività del corpus

Quando in un contesto psicolinguistico si parla della frequenza di parola, ci si riferisce ad una ideale frequenza media, che faccia astrazione da particolari idiosincrasie, frequenze soggettive, sovraesposizioni di talune categorie di persone a lessici specialistici, etc. E' perciò necessario che un lessico di frequenza centrato sul ricevente non sia basato su tipi testuali stabiliti a priori o campionati in maniera distorta, ma su criteri di selezione del corpus che permettano di far emergere per quanto possibile la frequenza media di ricezione. Come argomenteremo più diffusamente tra breve, un tale lessico di frequenza dovrebbe essere rappresentativo dell'italiano effettivamente letto (piuttosto che di quello più frequentemente prodotto) e dovrebbe essere ricavato da testi a larga diffusione, riproducendone l'incidenza proporzionale. In accordo con i criteri appena esposti, uno degli aspetti che dovrebbe caratterizzare la selezione del corpus sul quale si basa un lessico di frequenza rappresentativo dell'italiano scritto contemporaneo è il fatto che tale selezione venga effettuata sulla base di una indagine reale e non di scelte compiute a priori. Tali scelte spesso portano come conseguenza a favorire dal punto di vista delle occorrenze certi tipi testuali, sfavorendone altri.

Dal punto di vista della rappresentatività dei testi scritti effettivamente fruiti, gli strumenti attualmente esistenti per l'italiano sono insoddisfacenti. Intanto, nessuno di essi si basa su di un corpus di italiano scritto contemporaneo di dimensioni rilevanti: alcuni tra i più usati lessici di frequenza dell'italiano si basano infatti su corpora di sole 500.000 occorrenze di parole; per di più, non sono stati comunque pubblicati i dati sulle parole ricorrenti nelle fasce di frequenza più basse. Inoltre, nessuno degli attuali dizionari di frequenza per

l'italiano scritto è rappresentativo in modo bilanciato di diversi usi scritti della lingua. Per esemplificare questa affermazione, nella Tabella 1 sono riportati i sottoinsiemi testuali di cui è composto il *Lessico di frequenza della lingua italiana contemporanea* (LIF: Bortolini, Tagliavini e Zampolli, 1971), la loro incidenza percentuale nel LIF stesso, e la loro incidenza percentuale quale emerge da una recente indagine dell'ISTAT (ISTAT, 1993) sulle letture degli italiani.

Tabella 1
Incidenza dei diversi tipi di testi nel LIF e nell'indagine ISTAT sulle letture degli italiani (ISTAT, 1993)

	LIF	Indagine ISTAT
TEATRO	20%	0.3%
ROMANZI	20%	8.7%
SCENEGG.CINEMAT.	20%	0
PERIODICI*	20%	83%
SUSSIDIARI	20%	dato mancante

* Nel LIF sotto la voce periodici sono compresi anche i quotidiani. Per uniformare il dato, anche la percentuale dell'indagine ISTAT è stata calcolata accorpando quotidiani e periodici.

Seguendo i criteri di rappresentatività sopra delineati, abbiamo costituito un corpus sul quale stiamo costruendo una banca dati lessicale dell'italiano scritto contemporaneo. Al progetto partecipano la Scuola Normale Superiore di Pisa, l'Istituto per i Circuiti Elettronici del CNR di Genova e l'Istituto di Psicologia del CNR di Roma. Uno dei criteri guida a cui si ispira il corpus è, come già detto, far sì che esso permetta di far emergere la frequenza media di ricezione. Esso vuole dunque risultare rappresentativo soprattutto dell'italiano effettivamente letto, piuttosto che di quello più frequentemente prodotto. Da questa scelta consegue che non debbano essere necessariamente presi in considerazione tutti i possibili usi scritti dell'italiano, ma soltanto i tipi di testi che si presuppone siano più letti. Questa scelta porta inoltre ad escludere dal corpus una serie di tipi di testi (quali codici e leggi, verbali, rapporti tecnici, saggi scientifici pubblicati su stampa specialistica) che si presume vengano letti da un numero molto scarso di persone, benché rappresentino una percentuale notevole di ciò che viene scritto. Questa scelta si giustifica anche perché uno degli scopi dichiarati del corpus che abbiamo costituito, e del lessico di frequenza che deve scaturirne, è quello di avvicinarsi al lessico mentale di un

parlante di media cultura. La differenza tra testi ad alta incidenza nella scrittura e testi ad alta incidenza nella lettura può essere definita forse più chiaramente in termini operazionali come un rapporto quantitativo tra chi produce il testo e chi lo legge: in casi come quelli delle sentenze di tribunale o degli articoli scientifici di tipo specialistico il quoziente dato da questo rapporto è molto più alto che in casi come quelli dell'articolo di quotidiano o di rotocalco.

3. Determinazione dell'ampiezza del corpus e sua ripartizione

Nella selezione del corpus, ci siamo serviti dei risultati di una recente indagine ISTAT sulle letture degli italiani, basata su di una popolazione di individui di età superiore agli 11 anni, componenti di 24.000 famiglie (ISTAT, 1993). Nell'indagine sono riportate le preferenze di lettura del campione, suddivise per tipo di pubblicazione: quotidiani, periodici (settimanali e mensili) e libri. L'incidenza quantitativa dei tre tipi è stata da noi ponderata per i seguenti fattori: numero di fruitori per ciascun tipo, numero di unità lette per ciascun tipo, numero di occorrenze mediamente contenute in ciascun tipo. Le incidenze quantitative così trasformate sono state infine rapportate al numero di occorrenze che intendevamo raggiungere (3 milioni). La dimensione di 3 milioni di occorrenze ci è apparsa come la dimensione minima capace di garantire al contempo l'inclusione nel corpus di un numero sufficientemente alto di lemmi a bassa frequenza e la rappresentatività del campione. Inoltre il corpus da cui ricavare il lessico di frequenza deve contenere nelle nostre intenzioni un numero di occorrenze sufficiente per permettere la registrazione di unità lessicali anche di una fascia di frequenza medio-bassa, in misura sensibilmente superiore a quella degli altri lessici di frequenza disponibili per l'italiano. Molte indagini linguistiche e psicolinguistiche richiedono infatti di considerare l'interazione di diverse variabili lessicali e sublessicali, soprattutto nella fascia di frequenza medio-bassa: la scelta dei lemmi appartenenti a tale fascia dovrebbe poter spaziare in un insieme non troppo limitato di parole. Allo stesso tempo, la dimensione di 3 milioni di occorrenze del corpus permette il trattamento in tempi ragionevoli del materiale lessicale selezionato.

La suddivisione dei 3 milioni di occorrenze per tipo di pubblicazione ha dato luogo alla seguente ripartizione: 1.500.000 occorrenze da quotidiani, 1.000.000 di occorrenze da periodici, 500.000 occorrenze da libri. Una simile modalità di ripartizione del corpus, sempre basata sui dati dell'indagine ISTAT citata, è stata seguita anche per le partizioni interne a ciascun tipo (genere di libri e riviste, argomento all'interno dei quotidiani). Anche la scelta degli specifici quotidiani, periodici e libri è stata operata in base a vari dati sulla loro lettura, laddove disponibili, o sulla loro diffusione. Nel paragrafo seguente analizzeremo come si è arrivati alla ulteriore ripartizione quantitativa dei sottosettori del corpus.

4. Ripartizione quantitativa dei sottosettori del corpus

Secondo i dati dell'indagine ISTAT, in Italia 30 milioni di persone leggono abitualmente quotidiani, 28 milioni leggono periodici e 18 milioni leggono libri. Per rispettare tale ripartizione, si dovrebbe prelevare quindi il 39% delle occorrenze da quotidiani, il 37% da periodici e il 24% da libri. Questa ripartizione non terrebbe però conto della incidenza relativa dei tre tipi di pubblicazione nella lettura, non terrebbe cioè conto del fatto che, ad esempio, la media di libri letti in un anno, da chi legge libri, è di 6,33, mentre la media di quotidiani letti in un anno è di 255. Questo tipo di incidenza è stato dunque il primo fattore da noi utilizzato per ponderare i dati iniziali.

In secondo luogo abbiamo dovuto tener conto del fatto che il numero medio di occorrenze di parole varia a seconda del tipo di pubblicazione. Sulla base di un campionamento non sistematico da noi effettuato, abbiamo stimato il numero medio di occorrenze per unità pubblicata (rispettivamente 70.000 per un quotidiano, 35.000 per un periodico e 80.000 per un libro), al fine di ponderare ulteriormente il numero di occorrenze per tipo di pubblicazione.

E' infine lecito ipotizzare che, a parità di altre condizioni, un lettore non legga in media la stessa proporzione relativa di testo scritto di un quotidiano, di un periodico e di un libro. Considerata anche questa terza restrizione sulla determinazione dei sottosettori, la proporzione risultante diventava: 57% di occorrenze da quotidiani, 33% da periodici e 10% da libri.

Come ultimo accorgimento, abbiamo tenuto presente che nell'indagine ISTAT non vengono considerati i libri letti per motivi scolastici o professionali e, in accordo anche con una esigenza di semplicità nella ripartizione abbiamo aggiustato le proporzioni nel modo già riportato: 50% di occorrenze, pari a 1.500.000, da quotidiani; 33.3% di occorrenze, pari a 1.000.000, da periodici; 16.7%, di occorrenze, pari a 500.000, da libri.

Per i periodici, dato che i settimanali vengono letti tre volte più dei mensili, abbiamo ponderato tutti i valori relativi secondo un rapporto di 3 a 1. Per alcuni generi di periodici (periodici radiotelevisivi, periodici di cronaca mondana, periodici per bambini e ragazzi, fotoromanzi), che contengono meno parole rispetto a periodici di altro tipo, vi è stata una correzione ulteriore.

La sottoripartizione per tipo di pubblicazione è stata poi incrociata con una ripartizione di genere per libri e riviste, e con una ripartizione di argomenti per i quotidiani. Anche per questo secondo tipo di ripartizione, abbiamo fatto riferimento ai dati ISTAT.

Un'ulteriore scelta ha riguardato gli specifici libri, periodici e quotidiani da inserire nel corpus. La selezione è stata operata anche in questo caso sulla base dei dati relativi alla diffusione ed alla lettura. Per i periodici e i quotidiani sono stati utilizzati dati Audipress sulla lettura relativi al secondo semestre del 1993 ed al primo del 1994, dati Accertamento Diffusione Stampa per il periodo 1.9.92-31.8.93 e dati pubblicati dalla FIEG in *La stampa in Italia* (1990-92). Per i libri si è fatto riferimento alle classifiche dei libri più venduti nel 1992, 1993 e 1994, pubblicate nell'inserto *Tuttilibri* di *La Stampa* e realizzate da Adhoc-Gpf & Associati, e ai dati ISTAT sulla produzione editoriale del 1993, pubblicati su

La Rivisteria. I testi inclusi nella banca dati sono quindi quelli più letti nel periodo di tre anni 1992-1994.

Individuate le fonti, l'ultima decisione è stata quella relativa a quali parti di esse inserire nel corpus e quale ampiezza attribuire a tali parti. Per i periodici, l'unità di riferimento per il campionamento è stata la pagina. Si è cercato di mantenere costante la proporzione tra parti iniziali, centrali e finali degli articoli, nel caso in cui non fossero stati inclusi per intero nel corpus, e di mantenere per ogni periodico la proporzione tra i tipi di articoli o argomenti inclusi. In media ogni unità è composta da 600 occorrenze.

Per i libri, sono state incluse nel corpus mediamente 5.000 occorrenze da ogni libro per quanto riguarda narrativa, saggistica, fantascienza, libri rosa e gialli; 2.500 per scienze sociali ed umane, scienze naturali ed esatte, arte e teatro; 1.500 per hobby, viaggi e bambini/ragazzi. Il campionamento in questo caso è stato di tipo casuale sistematico: per ogni libro, sono state prese due pagine a intervalli regolari dopo aver scelto la prima pagina in maniera casuale. La frammentazione dei testi è giustificata dall'esigenza di ridurre il divario tra frequenza assoluta e indice d'uso, soprattutto diminuendo la probabilità di includere nel corpus sacche di ripetizione di lemmi a bassa frequenza.

Per i quotidiani, l'unità di riferimento è l'articolo, spezzato solo nel caso sia più lungo di 1.000 occorrenze. Sono state selezionate in media 41.667 occorrenze per ogni mese dal 1992 al 1994. Anche in questo caso, la dispersione temporale dei testi selezionati ha lo scopo principale di evitare, per quanto possibile, distorsioni sistematiche, dovute in questo caso a particolari scelte lessicali condizionate dal contesto temporale in cui il testo è stato prodotto.

RIFERIMENTI BIBLIOGRAFICI

Bortolini, U., Tagliavini, C., Zampolli, A. (1971). *Lessico di frequenza della lingua italiana contemporanea*. Milano: Garzanti.

Colombo, L. (1993). Locus o loci dell'effetto frequenza? In A. Laudanna, C. Burani (a cura di), *Il lessico: processi e rappresentazioni*. Roma: La Nuova Italia Scientifica.

Gordon, B. (1985). Subjective frequency and the lexical decision latency function: Implications for mechanisms of lexical access. *Journal of Memory and Language*, 24, 631-645.

Istituto Nazionale di Statistica (1993). *Indagine multiscopo sulle famiglie Anni 1987-1991. Vol. 7: Letture, Mass Media e Linguaggio*. Roma: ISTAT.

Scarborough, D.L., Cortese, C., Scarborough, H.S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, 3, 1-17.

Whaley, C.P. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior*, 17, 143-154.