Dal lessico di frequenza al lessico di notorietà

Domenico Russo
Università degli Studi "G. D'Annunzio"
Dipartimento di Studi Comparativi
Viale Pindaro 42 – 65127 Pescara
russo@unich.it

Abstract
The statistic lexicography has produced a series of frequency dictionaries of the most important languages of the world. Nevertheless, frequency dictionaries are subjected to three conditions which weak their descriptive capacity: the small number of types of texts; the impossibility to detect the collocations and the complex lexemes, the impossibility to detect the available words.
It's also for these reasons that the speaker assigns a different place to a word instead of the place that the same word has in the frequency list. From that arises the attempt of a knowledge lexicon, able to reaorganize the frequency list extracts from the text, according to the speaker lexical competence.

1 Il LIN 2004

L'elaborazione del *Lessico italiano di notorietà* si iscrive nelle linee epistemologiche che definiscono due rilevantissime imprese lessicografiche, quelle sul vocabolario disponibile di Gougenheim[1] e quello delle verifiche di comprensione di De Mauro.[2]

Così come le liste di frequenza stabiliscono per ogni lemma il relativo indice di frequenza, d'uso e di rango, allo stesso modo il *Lessico italiano di notorietà* fornisce per ogni lemma il relativo indice di notorietà, di rango e di attrazione. Così come vocabolari fondamentali e di base presentano elementi lessicali che sfuggono, pur essendo funzionalmente rilevanti, alla quantificazione statistica, anche il *Lessico italiano di notorietà* include parole disponibili di cui stabilisce i relativi indici. In questo modo il LIN genera un lemmario di natura unitaria, indicizzato, replicabile, i cui indici dipendono in modo diretto dal tempo della rilevazione e dalla massa parlante intervistata.

Un lessico con le caratteristiche del LIN permette una pluralità di applicazioni e di sviluppi, ma è utile soprattutto in due casi. Il primo è relativo allo studio delle configurazioni lessicali di natura individuale, di gruppo o comunitarie. Il secondo è relativo allo studio delle evoluzioni lessicali, individuali, di gruppo o di comunità, relativamente a questa o quella pa-

[1] Cfr. Gougenheim 1956 e *Idem* 1964.
[2] Cfr. De Mauro 1980: 102-12 e *Appendice*: 147-70 per il primo lemmario; *Idem* 1994: 115-21 per la genesi; GDU *Introduzione*: *passim* e *Postfazione*: *passim*; *Idem* 2004: 144 per la fortuna editoriale.

rola, a questo o a quel gruppo o insieme di parole. Entrambi i settori citati, com'è noto, alimentano con dati e temi di riflessione la ricerca teorica di base e il versante applicativo delle linguistica (primo tra tutti l'educazione linguistica).

1.1 Il lemmario del LIN

Il lemmario del LIN è costituito da 5801 lemmi che comprendono 3812 (pari al 71% ca) dei 5356 lemmi del LIF e dai 1989 lemmi del vocabolario AD del GDU.[3] I lemmi LIF sono scelti in base alla divisione dell'intero lemmario in 6 intervalli di rango d'uso composti rispettivamente il primo di 500 lemmi, il secondo e il terzo di 750 e i restanti tre di circa 1000 lemmi ciascuno. I lemmi selezionati dal primo intervallo sono 497 (esclusi *essere* e *avere* ausiliari e *volta*), dagli altri intervalli sono stati prelevati i primi 663 lemmi di ognuno (con l'esclusione di *coda* dal terzo e di *serbare* dal quinto).

Il principio su cui si basa il test di notorietà usato per stabilire il LIN riprende e rimodella una procedura introdotta da De Mauro e dai suoi collaboratori nello studio del lemmario del VdB[4] che consiste, nell'essenziale, nel correlare in opportune unità di rilevazione lemmi di cui è noto il valore d'uso, nel nostro caso i lemmi LIF, e lemmi che non hanno valori di frequenza d'uso come sono appunto le parole disponibili, nel nostro caso il vocabolario AD del GDU.

Le unità di rilevazione del LIN sono 663[5] serie di dieci lemmi ciascuna. In ogni serie sono presenti 6 lemmi LIF, uno per ognuno dei 6 intervalli di valore d'uso stabiliti, 3 lemmi AD e un lemma il cui grado di notorietà e di frequenza d'uso è esterno sia al vocabolario AD che ai valori LIF.

Agli intervistati si chiede di attribuire un punteggio da 1 a 10 alle parole in base alla percezione di notorietà soggettiva che ogni lemma suscita alla loro attenzione metalinguistica.[6] L'attribuzione dei punteggi deve avvenire in cinque turni di assegnazione successivi, in ognuno dei quali l'intervistato deve selezionare una coppia di parole composta dalla parola più nota e da quella meno nota tra quelle in esame, attribuendo il punteggio maggiore alla parola nota e il minore alla meno nota.[7]

2 L'indice di notorietà LIN

2.1 I gradi di notorietà delle parole

Una rilevazione sulla notorietà delle unità lessicali comprese in una lista chiusa condotta usando valori numerici decimali permette di ordinare la serie dei lemmi in una graduatoria

[3] Estratti dalla prima edizione, 1999.

[4] Cfr. in particolare Risso 1986 e relativa tesi di laurea.

[5] Cfr. Russo 2004.

[6] La nozione di «notorietà» adottata dal LIN è la sommatoria della varietà dei fatti che determinano un parlante a sostenere che conosce o non conosce e quanto nell'uno e nell'altro caso una parola e si basa sulla doppia assunzione che la dizione *conoscere una parola* coincida con l'intero insieme dei possibili usi linguistici e che, se interrogato in proposito, un parlante è sempre in grado di pronunciarsi in merito.

[7] Cfr., per i dettagli e le ragioni delle scelte citate, Russo 2004 a:3-8 e 2005 b *passim*.

anch'essa decimale calcolando in modo semplice ed efficace la media dei punteggi raccolti da ogni unità. Tuttavia, una scala di gradi di notorietà decimale non è congruente con una buona rappresentazione dei fenomeni lessicali. Non lo è perché un andamento decimale è un andamento lineare mentre l'andamento dei dati lessicali segue andamenti curvilinei, di natura iperbolica nel caso dei dati lessicali tratti da corpus, di natura parabolica nel caso dei dati lessicali tratti da parlanti. In altri termini, la scansione decimale dei gradi di notorietà, costruendo intervalli di raggruppamento identici, mentre aiuta a costruire una graduatoria di notorietà dei lemmi in lista, si fa però sfuggire i dati relativi al diverso rilievo sistemico, che invece andrebbe calcolato e messo in evidenza, dei lemmi che si raccolgono alle varie altezze.

Il LIN cerca di tener conto di quest'esigenza omologando gli intervalli di costituzione dei gradi di notorietà all'andamento delle frequenze d'occorrenza dei lemmi nei testi. Quest'andamento, come è noto, conosce percentuali altissime alle prime centinaia di unità lessicali, per degradare poi in relazione agli insiemi successivi. Considerando infatti sezioni uguali di lemmario, ad esempio di 500 unità come si fa nel LIF, si vede che il numero delle occorrenze che si raccolgono nella prima sezione rappresentano più dell'80,652% del totale, mentre, a percentuali sempre più basse, le occorrenze che si raccolgono nelle sezioni successive degradano rapidamente (6,677%, 3,799%, 2, 564, ecc).

Dal punto di vista dell'interpretazione sistemica dei dati queste percentuali dicono che i primi lemmi della lista svolgono all'interno del sistema linguistico un ruolo enormemente più rilevante rispetto alle unità comprese negli intervalli sucessivi.

In questo lavoro si assume che, così come nei testi, anche nella competenza lessicale dei parlanti si determini una situazione tale per cui poche unità lessicali di massima notorietà rinviino a un alto grado di salienza funzionale, grado che si distacca sensibilmente da quello di tutte le altre unità lessicali di notorietà via via inferiore.

Tenendo conto di quanto detto sopra, va subito osservato che per quanto riguarda gli intervalli estremi della scala di notorietà le caratteristiche di una rilevazione a banda larga come quella da cui derivano i dati del LIN sono tali che non permettono in alcun modo di cogliere il punto di discontinuità tra il primo intervallo e i restanti del lemmario. Tuttavia è possibile reintrodurre questo punto di discontinuità nella lettura dei dati se si tiene conto da una parte della consistenza quantitativa in termini di lemmi dei due intervalli estremi, che copre il 20% del totale dei lemmi utilizzati nella rilevazione, e dall'altra degli indici di attrazione relativa dei primi due turni di assegnazione. In questo modo si assume che nel giro di due turni tutti i lemmi degli intervalli 1 e 7 previsti dalla rilevazione, cioè teoricamente quelli più e quelli meno noti e dunque più forti come attrattori, si vedano assegnato il loro coefficiente. Ciò corrisponde a resecare nella serie delle medie risultanti dalla rilevazione il 20% dei valori estremi per entrambi i segmenti. Tenuto conto del fatto che le medie si dispongono in una scala di 901 unità di valore e che questa è la stessa scala che si ottiene con la normalizzazione prevista dal calcolo dell'indice di notorietà elaborato in questo lavoro, la nostra lettura considera come segmenti estremi tutti i valori di media compresi tra 10.00 e 8.00 per il primo intervallo e tra 2.99 a 1.00 per il settimo intervallo.

Per quanto riguarda gli intervalli intermedi, il criterio che regola l'omologazione degli intervalli successivi al primo (e precedenti l'ultimo) prende in considerazione la densità nume-

rica media dei lemmi che nei raggruppamenti LIF risultano equifunzionali perché di identico valore d'uso. Infatti, al decrescere dei valori del delta tra i valori d'uso minimo e massimo di ogni intervallo considerato aumenta il numero dei lemmi che mediamente entra con lo stesso valore in un punto di valore d'uso. Ciò dato, poiché i valori dei lemmi per punto di valore d'uso sono inversamente proporzionali ai valori delle percentuali dei delta sul loro totale, gli stessi valori sono direttamente proporzionali agl'inversi di queste percentuali normalizzate a 100, il che permette di dividere la serie delle medie ottenute dalla rilevazione o, come si fa qui, la serie dei valori decimali dell'intervallo costitutivi dell'indice di notorietà in intervalli omologhi a quelli presentati dai valori standard.

2.2 Il calcolo sulla frequenza delle attribuzioni

Come che sia dell'omologazione proposta più su, resta il fatto che il ricorso alla media aritmetica come principio di determinazione del rango di notorietà di un'unità lessicale e la scansione decimale dei vari gradi permette di ottenere eccellenti dati conoscitivi d'insieme.

Nello stesso tempo, tuttavia, proprio in virtù del suo stesso potere ermeneutico, la media aritmetica porta a porre domande ulteriori. Spinge in particolare a chiedersi anzitutto in che modo si comporta la varietà dei parlanti che originano il valore di notorietà di una singola unità lessicale, o, ancora, quale sia la varietà dei comportamenti dei parlanti che originano gruppi di unità lessicali che hanno lo stesso valore di notorietà.

Domande di questo tipo non hanno modo di essere poste quando si stabilisce il valore d'uso delle parole in base a un valore di dispersione calcolato su un insieme ridotto di tipi testuali, come avviene di necessità in molte liste di frequenza. Emergono invece più facilmente dove, come nel caso della notorietà, la varietà dei tipi testuali si trova sostituita dalla varietà dei soggetti parlanti che compongono la popolazione testata.

In questo secondo caso il calcolo della media è tale che il risultato finale sottrae alla lettura proprio ciò che invece sarebbe interessante sapere. In altri termini, un indice di notorietà puramente medio presenterebbe inconvenienti in parte omologhi a quelli che presenta un'indice di valore d'uso delle parole che non prevede la correzione portata dal valore della dispersione. E' dunque necessario elaborare un indice di notorietà che contenga al suo interno il procedimento di calcolo della media aritmetica, ma sappia, allo stesso tempo, tener conto dell'andamento ogni volta particolare dei valori che costituiscono la media stessa.

L'indice di notorietà LIN tiene conto del diverso comportamento dei soggetti parlanti che costituiscono la popolazione di rilevazione prendendo in considerazione anzitutto la frequenza con cui i diversi punteggi vengono attribuiti; indicizza, in secondo luogo, i valori delle frequenze dei punteggi in modo da assumere all'interno del calcolo il peso funzionale dei diversi punteggi e normalizza infine il risultato in modo da ottenere una scala decimale. L'indice di notorietà LIN elaborato insieme al collega Sergio De Grossi risponde cioè alla formula:

$$N = \frac{\sum_{1}^{10} f_i p_i \ln(p_i)}{P \ln(10)}$$

dove N è il valore di notorietà che risulta dalla sommatoria da 1 a 10 normalizzata sui parlanti P dei prodotti delle frequenze dei punteggi *fi* per i punteggi *pi* per il loro logaritmo.

In questo modo l'indice di notorietà, che come si vede consiste in definitiva nel calcolo di una media aritmetica 'arricchita', risulta sensibile a variazioni anche molto piccole di dispersione dei punteggi raggiungendo un potere di discriminazione dei valori per lemma praticamente pari all'unità, oltre a consentire livelli d'osservazione e considerazione pari al numero di cifre decimali che si ha interesse di assegnare come criterio di sezionamento e raggruppamento delle unità lessicali comprese nella lista.

Bibliografia

A. Dizionari

GDU, *Grande dizionario italiano dell'uso*, ideato e diretto da Tullio De Mauro, Torino, Utet, 1999.

Gougenheim, G., Michéa, R., Rivenc, P., Sauvageot, A. (1956), *L'élaboration du français élémentaire. Etude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris, Didier.

Gougenheim G., Michéa R., Rivenc P., Sauvageot A. (1964), *L'élaboration du français fondamental (1er degré). Etude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris, Didier.

LIF, *Lessico di frequenza della lingua italiana contemporanea*, Bortolini U., Tagliavini C., Zampolli A., Milano IBM-Italia, 1971; Garzanti, 1972.

Migliorini, B. (1943), *Der Grundlegende Wortschatz des Italienischen*, Marburg, Elwert.

Russo, D. (2004b), *LIN. Lessico Italiano di Notorietà 2004. Prototipo*, Roma, Aracne.

Russo, D. (2005a), *LIN. Lessico Italiano di Notorietà 2004. Dal lessico dei bambini a quello dei nonni. Indagine sulla configurazione della competenza lessicale delle diverse età*, Roma, Aracne.

Russo, D. (2005c), *LIN. Lessico Italiano di Notorietà 2004. Due indagini campione*, Roma, Aracne.

Russo, D. (2005d), *LIN. Lessico Italiano di Notorietà 2004. Il lessico di alta disponibilità*, Roma, Aracne.

B. Altra Letteratura

Burani, C., Thornton, A. M. (1993), 'Strumenti per la ricerca psicolonguistica: lessici di frequenza della lingua italiana'. *Giornale Italiano di Psicologia*, 20, pp. 495-506.

Chiari, I., De Mauro, T., (a cura di) (2005), *Parole e numeri*, Roma, Aracne.

De Mauro, T. (1980), *Guida all'uso delle parole*, Roma, Editori Riuniti.

De Mauro, T. (1994), *Com'è nato il Vocabolario di Base*, in Thornton, Iacobini, Burani (1994), pp. 115-21.

De Mauro, T. (1999), *Introduzione* e *Postfazione* al GDU: VII-XLII e 1163-83.

De Mauro, T. (2004), *La cultura degli italiani*, a c. di Francesco Erbani, Roma-Bari, Laterza.

Genuini, S., Vedovelli, M. (a c. di) (1983), *Teoria e pratica del Glotto-kit. Una carta d'identità per l'educazione linguistica*, Milano, Angeli.

Risso, C. (1986), *La disponibilità lessicale: nozione teorica e dati sperimentali.* «Linguaggi», III (1986) 1-2, pp. 6-13.

Russo, D. (2004a), *Test di conoscenza delle parole. Schede di rilevazione per calcolare il grado di notorietà delle parole più usate in italiano*, Roma, Aracne.

Russo, D. (2005b.), 'La rilevazione dei gradi di notorietà dei lemmi del Vocabolario di Alta Disponibilità', in Chiari/De Mauro 2005, pp. 233-46.

Sgroi, S. C. (1981), 'I lessici fondamentali e di frequenza della lingua italiana', *Quaderni di semantica* 2, 2, pp. 281-95.

Thornton, A. M., Iacobini, C., Burani, C. (1994), DBVDB. *Una base di dati sul Vocabolario di Base della lingua italiana*, Roma, Bulzoni.