L'ambiguità nel lessico di alta frequenza

di *Federica Casadei*

1
Sporcarsi le mani con i numeri

Non c'è dubbio che uno dei più rilevanti aspetti di continuità nella pur vastissima gamma degli interessi di De Mauro sia rappresentato dall'attenzione per la dimensione quantitativa dei fatti linguistici: «cifra caratteristica del suo profilo di studioso», scrive Serianni[1] all'indomani della scomparsa; «tratto demauriano per eccellenza», nel ricordo di Lorenzetti[2] di De Mauro lessicografo.

Del comprendere l'importanza dell'analisi quantitativa e statistica dei fenomeni linguistici, e in particolare di quelli lessicali, De Mauro è stato del resto assoluto pioniere: nel 1961, cioè in un'epoca in cui, come dirà egli stesso, «le rilevazioni elettroniche e, più in genere, la statistica linguistica erano […] viste o, meglio, intraviste come un oggetto misterioso e malfido»[3], firma per l'Enciclopedia Treccani la voce *Statistica linguistica*[4], nella quale sottolinea la portata non solo descrittiva ma anche teorica delle indagini linguistiche quantitative. Una convinzione, questa, che De Mauro ribadirà in molte occasioni (tra le quali citiamo almeno i due lavori *Quantità-qualità: un binomio indispensabile*[5] e *Informatica e linguistica*[6] e il volume *Parole e numeri* curato con Isabella Chiari[7]) e che darà i suoi frutti, per menzionare solo le tappe cruciali in ambito

1. L. Serianni, *Un linguista democratico*, in "Il Sole 24 Ore", 6 gennaio 2017.

2. L. Lorenzetti, *De Mauro: una nuova percezione del lessico*, in "La Lingua Italiana", 2017 (solo online, www.treccani.it/lingua_italiana/speciali/DeMauro/Lorenzetti.html).

3. T. De Mauro, *Fogli di un diario linguistico 1965-2015*, in "Nuovi Argomenti", 73, 2016, pp. 9-30: 9.

4. T. De Mauro, voce *Statistica linguistica*, in *Enciclopedia Treccani*, *Appendice III*, Istituto dell'Enciclopedia Italiana, Roma 1961.

5. Relazione presentata da De Mauro al Convegno "Statistica e scienze umane" tenutosi alla Sapienza Università di Roma nel 1992, pubblicata poi in T. De Mauro, *Capire le parole*, Laterza, Roma-Bari 1994, pp. 97-106.

6. Pubblicato originariamente in *Calcolatori e scienze umane*, a cura di E. Presutti, Etas Libri, Milano 1992, pp. 195-210, poi raccolto in De Mauro, *Capire le parole*, cit., pp. 107-18.

7. *Parole e numeri: analisi quantitative dei fatti di lingua*, a cura di T. De Mauro e I. Chiari, Aracne, Roma 2005.

lessicale, nella creazione del *Vocabolario di base della lingua italiana* (VdB)[8], poi del *Lessico di frequenza dell'italiano parlato*[9] e infine dello stesso *Grande Dizionario Italiano dell'Uso* (GRADIT)[10]. Il dizionario ha infatti tra i suoi più notevoli tratti innovativi l'indicazione della fascia d'uso nella quale si colloca ciascun lemma e ciascuna sua eventuale accezione; con questa scelta, che trasforma le categorie lessicostatistiche del VdB in marche d'uso, si chiude un cerchio e il VdB diviene «strumento nello strumento», per usare l'espressione di Bisconti[11].

In omaggio dunque all'insegnamento demauriano di «sporcarsi le mani con computi di natura statistica»[12] e di usare anche i numeri per l'analisi dei fatti lessicali, presento in questo contributo alcuni dati sull'ambiguità nel lessico italiano di alta frequenza, dai quali mi pare emergano interessanti spunti di riflessione su due grandi questioni che sono state al centro degli studi lessicologici di De Mauro: da un lato l'importanza della frequenza, per le ricadute che essa ha sulla forma e sul contenuto delle parole; dall'altro il tema della mancata relazione uno a uno tra forme e significati che si verifica quando a uno stesso significante corrispondano, per polisemia o per omonimia, significati diversi.

2
Polisemia e omonimia nel vocabolario di base italiano

2.1. La polisemia nel lessico generale e di alta frequenza

Pur essendo disponibili poche cifre precise sulla quantità di lessemi monosemici e polisemici registrati nelle grandi fonti lessicografiche (pressoché nessuna delle quali fornisce questo dato), non c'è dubbio che i primi siano la maggioranza. Per l'inglese, la versione 3.0 di WordNet contiene 155.287 parole, l'83% delle quali risultano monosemiche; la quota di polisemia si attesta quindi al 17%, come già nelle precedenti versioni del database (Miller riferiva infatti che «approximately 17% of the words in WordNet are polysemous»[13]). Per l'i-

8. Il *Vocabolario di base della lingua italiana* compare nella prima versione come appendice al volume di T. De Mauro, *Guida all'uso delle parole*, Editori Riuniti, Roma 1980, pp. 149-83. Nel 2016 ne è stata pubblicata una nuova versione online sul sito di *Internazionale*, www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana. Se non diversamente specificato, i dati che presento in questo lavoro fanno riferimento alla versione del 1980.

9. T. De Mauro, F. Mancini, M. Vedovelli, M. Voghera, *Lessico di frequenza dell'italiano parlato*, Etas Libri, Milano 1993.

10. *Grande Dizionario Italiano dell'Uso*, ideato e dir. da T. De Mauro, 6 voll., UTET, Torino 1999 (II ed. 8 voll. con CD-ROM, 2007).

11. V. Bisconti, *La svolta lessicografica di Tullio De Mauro e i dizionari italiani contemporanei*, in "Chroniques Italiennes", 23, 2012, 2, pp. 1-26: 15.

12. De Mauro, *Capire le parole*, cit., p. 106.

13. G. A. Miller, *WordNet: A Lexical Database for English*, in "Communications of the ACM", 38, 1995, 11, pp. 39-41: 40.

taliano, dei circa 260.000 lessemi registrati nel GRADIT quelli polisemici sono oltre 50.000, cioè il 19%[14].

Minoritaria dunque nel lessico generale, costituito in larghissima parte da lessemi monosemici, la polisemia domina invece il lessico di alta frequenza. Che la frequenza d'uso di una parola correli con la sua polisemia è noto alla statistica linguistica a partire dagli studi di Zipf[15], il cui principio di versatilità economica afferma che le parole più frequenti, essendo semanticamente più generiche, possono assumere un numero maggiore di significati rispetto alle parole di minor frequenza. La correlazione tra frequenza e polisemia individuata da Zipf, confermata poi nei decenni successivi da molti studi, è tale che il grado di polisemia di un lessema può essere utilizzato come indicatore della sua frequenza; così avviene ad esempio in WordNet, dove in mancanza di dati affidabili sulla frequenza dei lessemi registrati quest'ultima viene assegnata in base alla loro polisemia («WordNet uses polysemy as an index of familiarity»[16]).

Varie analisi effettuate, in lingue diverse, su campioni di parole di alto uso mostrano che in esse il numero di accezioni è molto superiore alla media. Per l'italiano il VdB non dà informazioni sulla polisemia dei lessemi registrati, ma è possibile ricavarle servendosi del GRADIT, poiché questo, come si è detto, marca ogni lessema e ogni sua accezione secondo la fascia d'uso[17]. L'analisi che ho condotto[18] utilizzando appunto il GRADIT mostra che la quasi totalità del VdB è costituita da lessemi polisemici: come si vede dai dati riportati nella tabella 1, la polisemia investe circa il 90% del VdB nel suo complesso e il 96% del vocabolario fondamentale; e in realtà il conteggio sottostima la quota di polisemia nel VdB, per ragioni legate ai criteri di lemmatizzazione e definizione adottati dal GRADIT[19].

14. Il dato è riferito da De Mauro nella *Postfazione* al *Grande Dizionario Italiano dell'Uso*, cit., vol. 6, pp. 1163-83: 78.

15. Cfr. G. K. Zipf, *The Psycho-Biology of Language*, Routledge & Sons, London 1936; Id., *The Meaning-Frequency Relationship of Words*, in "Journal of General Psychology", 33, 1945, pp. 251-6; Id., *Human Behaviour and the Principle of Least Effort. An Introduction to Human Ecology*, Addison-Wesley Press, Cambridge 1949.

16. R. I. Tengi, *Design and Implementation of the WordNet Lexical Database and Searching Software*, in *WordNet: An Electronic Lexical Database*, ed. by C. Fellbaum, The MIT Press, Cambridge 1998, pp. 105-27: 13.

17. Le marche usate dal GRADIT, e che uso anche qui, per indicare le tre fasce del VdB sono: FO per il vocabolario fondamentale (i lessemi in assoluto più frequenti in italiano), AU per il vocabolario di alto uso (i lessemi di alta, seppur minore, frequenza), AD per il vocabolario di alta disponibilità (lessemi relativamente infrequenti nello scritto e nel parlato ma noti ai parlanti perché legati a oggetti e azioni di grande rilevanza nella vita quotidiana). Tutti i dati relativi al GRADIT che presento in questo lavoro si riferiscono alla versione elettronica dell'edizione del 2007.

18. Cfr. F. Casadei, *La polisemia nel vocabolario di base dell'italiano*, in "Lingue e Linguaggi", 12, 2014, pp. 35-52.

19. Questo aspetto è illustrato più chiaramente in Casadei, *La polisemia nel vocabolario di base*, cit., pp. 40-2.

Tabella 1
Polisemia nei lessemi del VdB registrati nel GRADIT

	lessemi	di cui polisemici
FO	2.077	2.001 (96%)
AU	2.663	2.440 (92%)
AD	1.988	1.562 (79%)
tot VdB	6.728	6.003 (89%)

2.2. L'omonimia nel lessico generale e di alta frequenza

Più complesso il quadro relativo all'omonimia, forse l'unico grande fenomeno lessicale per cui non è stata formulata nessuna legge quantitativa e che presenta molti aspetti irrisolti nella definizione e grande variabilità nel trattamento lessicografico[20]. Sappiamo dell'enorme diffusione dell'omonimia nei testi grazie ai dati relativi all'etichettatura dei *corpora* linguistici: come scrive De Mauro[21]

> dobbiamo all'esperienza di applicazione ai testi di programmi di lemmatizzazione automatica una stima della grande incidenza percentuale degli omonimi relativi nei testi delle lingue più diverse. Oggi sappiamo che nelle lingue europee circa la metà delle forme di parole dei testi sono plurietichettabili, riconducibili cioè a più di un lessema.

Più difficile, invece, valutare l'incidenza dell'omonimia nel lessico, perché i dizionari (inclusi i repertori di soli omonimi) differiscono molto nei criteri di lemmatizzazione degli omonimi e dunque nella quantità e nel tipo di omonimi che registrano.

Nel GRADIT, che come ogni dizionario generale registra solo le omonimie (anzi propriamente le omografie) nella forma di citazione – dunque solo casi come *ara* 'altare' e *ara* 'unità di misura' ma non *ara* voce del verbo *arare* – l'omonimia coinvolge 14.547 lessemi monorematici, pari al 6% del totale. Considerando invece anche le omonimie che si verificano in forme diverse da quella di citazione, come nel repertorio di omonimi HOMO che ho realizzato[22], i lessemi coinvolti nell'omonimia risultano essere circa il 14% dei monorematici del GRADIT; una percentuale, quindi, più che doppia rispetto a quella ottenuta consi-

20. Per una discussione dei problemi definitori e di trattamento lessicografico degli omonimi rinvio a F. Casadei, *L'omonimia nel lessico italiano*, in "Studi di Lessicografia Italiana", 33, 2016, pp. 187-228: 187-93.

21. T. De Mauro, voce *Basi di conoscenze e banche dati lessicali XXI secolo*, in *Enciclopedia Treccani* online, 2009 (www.treccani.it/enciclopedia/basi-di-conoscenze-e-banche-dati-lessicali_XXI_Secolo).

22. HOMO è un repertorio di omonimi italiani che include le omonimie che coinvolgono le forme di tutti i lessemi monorematici presenti nel GRADIT, per un totale 112.344 omonimi riconducibili a 35.557 lessemi. I criteri di costruzione di HOMO e i dati che ne emergono sono esposti in Casadei, *L'omonimia nel lessico italiano*, cit.

derando solo le forme di citazione, ma che comunque rappresenta una quota minima del lessico complessivo.

Per quanto riguarda il lessico di alta frequenza, nel nuovo VdB del 2016 i lessemi che hanno un esponente – cioè che hanno, nella forma di citazione, almeno un omonimo – sono quasi il 14% (997 su 7.248), una percentuale più che doppia, quindi, rispetto al 6% del lessico generale. E cifre ben più significative emergono se si considerano anche le omonimie in forme diverse da quella di citazione: dai dati di HOMO risulta non solo che i lessemi del VdB danno luogo da soli all'11% di tutti gli omonimi, ma soprattutto che il VdB, in proporzione al numero di lessemi che lo costituiscono, è la fascia lessicale più "omonimogena". Infatti è coinvolto in omonimie il 55% dei lessemi del VdB e ben il 64% di quelli del vocabolario fondamentale (si veda tabella 2), contro il 24% dei lessemi di uso comune, il 15-17% di quelli di basso uso e obsoleti, il 10% dei tecnico-specialistici[23].

Tabella 2
Omonimia nei lessemi del VdB registrati nel GRADIT

	lessemi	con almeno un omonimo in una delle forme
FO	2.077	1.324 (64%)
AU	2.663	1.428 (54%)
AD	1.988	978 (49%)
tot VdB	6.728	3.730 (55%)

2.3. Ambiguità e frequenza

I dati sin qui esposti mostrano che, si tratti di polisemia o di omonimia, il VdB è la fascia lessicale nella quale si concentra la maggiore quantità di ambiguità, con un forte scarto rispetto alle fasce di minor frequenza. E questa correlazione tra frequenza e ambiguità si manifesta anche all'interno del VdB stesso, poiché la quantità di lessemi ambigui per polisemia o per omonimia è massima nel vocabolario fondamentale e decresce progressivamente nel vocabolario di alto uso e in quello di alta disponibilità. Mettendo insieme i dati relativi alla polisemia e quelli relativi all'omonimia (si veda tabella 3), ne risulta che più della metà del VdB nel suo complesso – e oltre il 60% del vocabolario fondamentale – è costituita da lessemi che sono al tempo stesso polisemici e coinvolti in omonimie:

23. I conteggi precisi sulla quantità di omonimi nelle varie fasce d'uso si trovano in F. Casadei, *Frequenza, lunghezza e omonimia: un'analisi degli omonimi nel vocabolario di base italiano*, in "Lingue e Linguaggi", 19, 2016, pp. 61-75: 67.

Tabella 3
Lessemi del VdB sia polisemici che coinvolti in omonimie

	lessemi	con polisemia e omonimia
FO	2.077	1.309 (63%)
AU	2.663	1.388 (52%)
AD	1.988	916 (46%)
tot VdB	6.728	3.613 (54%)

Emerge inoltre una notevole ambiguità interna alla fascia lessicale basico-comune: come mostra la tabella 4, il 93% dei lessemi VdB polisemici ha più accezioni che ricadono nel VdB o al massimo nel vocabolario comune (CO), e quasi il 60% dei lessemi con omonimi trova un omonimo nel VdB o nel vocabolario comune; in entrambi i casi ciò avviene in misura maggiore nella fascia di massima frequenza per decrescere progressivamente nelle altre.

Tabella 4
Ambiguità interna alla fascia lessicale basico-comune

	lessemi VdB con più accezioni VdB/CO	lessemi VdB con omonimi VdB/CO
FO	96%	61%
AU	93%	60%
AD	89%	51%
tot VdB	93%	58%

Nel caso della polisemia questo dato si può spiegare con il *feedback* tra frequenza e polisemia, per cui da un lato la frequenza d'uso è fattore di innesco per lo sviluppo della polisemia, dall'altro l'elevata polisemia di un lessema ne favorisce la frequenza accrescendo la quantità di contesti in cui può essere usato[24]. Meno ovvio, invece, quale possa essere la spiegazione nel caso dell'omonimia: si potrebbe supporre che abbia un ruolo la lunghezza delle forme (dato che più è corta una parola più è probabile che trovi un omonimo, e dato che le parole più frequenti sono più corte, è più probabile che una parola trovi un omonimo nel lessico di alta frequenza); ma così non è, poiché la lunghezza media delle forme VdB che hanno omonimi anch'essi VdB/CO risulta uguale a quella delle forme VdB i cui omonimi sono esterni al VdB/CO.

24. Cfr. i recenti lavori di G. Fenk-Oczlon, A. Fenk, *The Association between Word Frequency and Polysemy: A Chicken and Egg Problem?*, in *Proceedings of the XIIth International Conference «Cognitive Modeling in Linguistics»*, ed. by V. Solovyev, V. Polyakov, Kazan State University Press, Kazan 2010, pp. 167-70 e G. Fenk-Oczlon, A. Fenk, *Frequency Effects on the Emergence of Polysemy and Homophony*, in "International Journal of Information Technologies and Knowledge", 4, 2010, 2, pp. 103-9.

Quest'ultimo dato, peraltro, è coerente con quanto emerge dall'analisi generale dell'omonimia, cioè che frequenza e lunghezza agiscono come variabili indipendenti: a parità di frequenza le forme più brevi hanno più omonimi, ma a parità di lunghezza le forme più frequenti hanno più omonimi[25]. Ad esempio tra i lessemi bisillabi – dunque tutti di pari lunghezza – che il GRADIT marca rispettivamente come appartenenti al vocabolario fondamentale, come obsoleti e come tecnico-specialistici, i primi hanno un omonimo nel 73% dei casi, i secondi nel 50%, i terzi nel 33%; e anche all'interno del VdB, dove i lessemi di alto uso e quelli di alta disponibilità hanno identica lunghezza media, la quota di omonimie è maggiore nei primi che nei secondi.

3
L'ambiguità come proprietà desiderabile del codice linguistico

Anche tra i fautori delle indagini quantitative e statistiche è oggetto di dibattito se la frequenza sia di per sé un fattore esplicativo dei fatti linguistici o se piuttosto non sia, come ritiene Greenberg, un sintomo che a sua volta deve essere spiegato («frequency is itself but a symptom and the consistent relative frequency relations [...] are themselves in need of explanation»[26]). Pur senza dirimere questo dubbio, i dati sopra esaminati indicano certamente che frequenza e ambiguità sono connesse in modo non casuale, e che esiste un "effetto frequenza" nello sviluppo sia della polisemia che dell'omonimia. Fatto, quest'ultimo, particolarmente significativo, posto che l'omonimia è tradizionalmente ritenuta un fenomeno del tutto accidentale e privo, a differenza della polisemia, di qualunque valore semiotico (nelle parole di Ullmann, «homonymy is not necessarily an unrestricted universal. [...] one could easily imagine an idiom without any homonyms; it would be, in fact, a more efficient medium»[27]); ci aspetteremmo, cioè, che solo la lunghezza delle forme, e non anche la loro frequenza indipendentemente dalla lunghezza, correlasse con l'esistenza di omonimie.

Ma perché questa concentrazione di ambiguità nel lessico di alta frequenza, se l'ambiguità è un ostacolo al buon funzionamento del codice linguistico – una patologia linguistica, nella classica definizione di Gilliéron?[28]

Zipf ritiene che un certo tasso di ambiguità lessicale rappresenti un accettabile compromesso tra l'esigenza di minimo sforzo del parlante (usare meno parole possibili e, idealmente, solo una che copra ogni possibile significato) e quella del ricevente (avere una forma diversa per ciascun significato)[29]. Alcuni studi

25. Cfr. Casadei, *Frequenza, lunghezza e omonimia*, cit., pp. 72-3.

26. J. H. Greenberg, *Language Universals: With Special Reference to Feature Hierarchies*, De Gruyter, Berlin-New York 1966, p. 70.

27. S. Ullmann, *Semantic Universals*, in *Universals of Language*, ed. by J. H. Greenberg, The MIT Press, Cambridge 1966, pp. 172-207: 197.

28. J. Gilliéron, *Pathologie et thérapeutique verbale*, Champion, Paris 1921.

29. Cfr. Zipf, *Human Behaviour and the Principle of Least Effort*, cit.

recenti[30] si spingono oltre, suggerendo che la pervasività dell'ambiguità lessicale dipenda dal fatto che quest'ultima è una proprietà *desiderabile* delle lingue, perché contribuisce ad aumentarne l'efficienza in più modi: da un lato consente il riutilizzo di forme la cui produzione e comprensione sono più facili (perché più familiari, più brevi, fonotatticamente più semplici, quali sono appunto quelle di maggior frequenza), dall'altro evita l'eccessiva ridondanza e la trasmissione di informazioni inutili che si avrebbero usando forme non ambigue anche laddove il parlante sia in grado di risolvere l'ambiguità servendosi di informazioni cotestuali e contestuali – cioè, come sappiamo, nella stragrande maggioranza dei casi. Capovolgendo il punto di vista che ritiene l'ambiguità un intralcio alla comunicazione, si sostiene che anzi l'ambiguità è condizione necessaria perché la comunicazione abbia successo e che ogni sistema comunicativo efficiente deve essere ambiguo, posto che il contesto fornisca dati sufficienti per la disambiguazione.

L'ipotesi che la mancata corrispondenza uno a uno tra forme e significati risponda a un principio semiotico volto a ottimizzare la capacità del codice linguistico non appare inedita a chi ha familiarità con quel che De Mauro ha teorizzato sulla natura del linguaggio verbale e dei processi di comprensione; cioè che l'esistenza di vari tipi di ambiguità nella relazione tra forme e significati – tra i quali l'omonimia, che De Mauro include con buona pace di Ullmann tra gli universali linguistici[31] – attesta non un errore di progettazione delle lingue, ma semmai che «le lingue sono costruite in modo da presupporre il continuo intervento dell'attività di disambiguazione»[32]. Delle tante che De Mauro ci ha lasciato, questa affermazione della natura semiotica dell'ambiguità lessicale, del suo essere tratto costitutivo e positivo delle lingue verbali, appare senz'altro una tra le più attuali.

30. Tra i lavori più rilevanti: S. T. Piantadosi, H. Tily, E. Gibson, *Word Lengths Are Optimized for Efficient Communication*, in "Proceedings of the National Academy of Sciences", 108, 2011, 9, pp. 3526-9; S. T. Piantadosi, H. Tily, E. Gibson E., *The Communicative Function of Ambiguity in Language*, in "Cognition", 122, 2012, pp. 280-91; J. Fortuny, B. Corominas-Murtra, *On the Origin of Ambiguity in Efficient Communication*, in "Journal of Logic, Language and Information", 22, 2013, 3, pp. 249-67; T. Wasow, *Ambiguity Avoidance is Overrated*, in *Ambiguity: Language and Communication*, ed. by S. Winkler, De Gruyter, Berlin-New York 2015, pp. 29-47; R. V. Solé, L. Seoane, *Ambiguity in Language Networks*, in "The Linguistic Review", 32, 2015, 1, pp. 5-36.

31. T. De Mauro, *Minisemantica dei linguaggi non verbali e delle lingue*, Laterza, Roma-Bari 1982, p. 94.

32. De Mauro, *Capire le parole*, cit., p. 26.