932/2

SLI
SOCIETÀ DI LINGUISTICA ITALIANA

L'INSEGNAMENTO DELL'ITALIANO IN ITALIA E ALL'ESTERO

ATTI DEL QUARTO CONVEGNO INTERNAZIONALE DI STUDI
Roma, 1-2 giugno 1970

a cura di
MARIO MEDICI e RAFFAELE SIMONE

Volume secondo

BULZONI ROMA 1971

UMBERTA BORTOLINI
ANTONIO ZAMPOLLI
Padova

Lessico di frequenza della lingua italiana contemporanea: prospettive metodologiche

Per iniziativa dell'IBM Italia e con la direzione di Carlo Tagliavini, è in preparazione presso l'Istituto di glottologia e fonetica dell'università di Padova un lessico di frequenza della lingua italiana contemporanea. Basato su uno spoglio di 500.000 parole, esso fornirà una prima raccolta di circa 5.000 lemmi, che pensiamo abbia non solo un notevole interesse pratico per l'insegnamento dell'italiano, ma possa anche essere un utile strumento di lavoro per lo studio del lessico contemporaneo. Esporremo qui, brevemente, alcuni problemi teorico-metodologici connessi con l'elaborazione di tale opera, e innanzi tutto alcuni problemi legati al concetto di f r e q u e n z a di una parola in una lingua e al m e t o d o per rilevarla a partire da uno spoglio di testi.

I testi sottoposti a spoglio si configurano, nelle intenzioni dei compilatori dei dizionari di frequenza, come campioni della lingua intesa quale universo statistico e di conseguenza pongono i problemi della campionatura per i quali la statistica ha elaborato tutta una serie di metodologie e di tecniche (M. Boldrini, *Statistica* 1960, p. 181 ss.).

Da un lato occorre delimitare l'« universo », dall'altro assicurarsi della rappresentatività del campione. Se si conoscono bene le caratteristiche dell'universo, si può fabbricare un campione in cui ciascuna di queste caratteristiche sia distribuita secondo le stesse proporzioni, con una sottile stratificazione (C. Muller, *Initiation* 1968, p. 15). Che non sia questo il caso della lingua, è facilmente intuibile.

Le frequenze si osservano in un testo, cioè in una realizzazione del sistema linguistico da parte di un individuo, realizzazione determinata da un lato dalle sue caratteristiche personali, dall'altro da quelle della comunità cui appartiene, caratterizzata sul piano sociale, geografico, storico, ecc. Se il testo è un testo letterario, nella sua produzione influiscono anche le regole e le peculiarità del genere letterario e, in ogni caso, la situazione in cui viene prodotto. I lessicografi hanno sempre parlato di un linguaggio tecnico e di

uno poetico, di uno popolare e di uno dotto, ecc. È legittimo chiederci se attualmente possediamo, della struttura dell'universo linguistico, una conoscenza sufficiente a decidere le dimensioni e la composizione di un campione rappresentativo, e addirittura, a stabilire se e in quali limiti il rapporto fra testo e lingua sia legittimamente equiparabile, con rigore metodologico, al rapporto campione-universo.

Vi sono almeno due maniere diverse di guardare alla lingua come ad un universo statistico.

Secondo una prima concezione, le unità del sistema linguistico sarebbero caratterizzate, oltre che dai tratti qualitativi emergenti dalle opposizioni e dalle relazioni che formano la struttura del sistema stesso, anche dalle loro rispettive probabilità di uso. Queste probabilità non sono direttamente osservabili, come non lo è del resto il sistema; esse però si tradurrebbero nel fatto (dato per certo) che le unità linguistiche ricorrerebbero nei testi, parlati e scritti, con frequenza relativamente stabili. In questa prospettiva, le frequenze osservabili nei testi vengono assunte come approssimazioni delle probabilità non osservabili del sistema.

Tale concezione ha ricevuto una formulazione teorica esplicita da P. Guiraud, anche se essa già traspariva da studi precedenti, e fu accettata e sviluppata in primo luogo da G. Herdan e poi da altri cultori della materia, i quali in essa riconoscono il fondamento teorico sul quale la statistica linguistica pone la propria autonomia come scienza.

Queste premesse non hanno però trovato almeno fino ad oggi, un consenso generale. Già gli autori del *Français fondamental*, introducendo la nozione di 'disponibilità' scossero per primi questa concezione e R. Moreau al Convegno di Strasburgo del 1964 (*Statistique et analyse linguistique*) delineava esplicitamente questa situazione affermando: « les premiers pas de la statistique appliquée à la linguistique ont précisément consisté à admettre des règles du jeu qui soient simples. C'est ainsi qu'on a énoncé (voir par exemple Guiraud) que la fréquence des mots était constante dans la langue, ce qui supposait donc que l'on pouvait assimiler le choix d'un mot au tirage d'un boule dans une urne dont la composition reste inchangée au cours du temps » (R. Moreau, *Intervention* 1966, p. 30). Anche la Hirscheber e C. Muller hanno dimostrato che questa regolarità non si verifica: le frequenze delle parole, salvo casi eccezionali, non sono stabili, ma variano, in dipendenza dallo stile e dal tema da testo a testo, perfino « pour les mots de relation, qui sont les éléments les moins thématiques du lexique, car beaucoup d'entre eux, la plupart même, subissent visiblement l'effet des situations stylistiques » (Muller, *Initiation* 1968, p. 141).

Nella seconda concezione, il concetto della probabilità come caratteristica intrinseca dell'unità del sistema linguistico, non è espresso; l'universo statistico è definito piuttosto come l'insieme di tutti i testi (parlati e scritti) prodotti in un certo periodo di tempo. Anche ammettendo che l'insieme sia definibile esattamente, una gran parte di questo insieme sfugge ad ogni misura: per esempio le lettere, le conversazioni private, ecc.

Moreau distingue due categorie di parole: parole di *classe ouverte* e parole di *classe fermée*.

Le parole di *classe fermée* sarebbero tutte le parole atematiche, quelle cioè che servono ad « esprimerci a proposito delle cose piuttosto che ad esprimere le cose stesse ». Si collocherebbero in questa categoria un certo numero di aggettivi e di voci correnti, le parole grammaticali, alcuni nomi assai generali: « termini più o meno comuni a tutti i soggetti e a tutte le situazioni ». Poiché il loro impiego non varia nei diversi centri di interesse, la stima della loro frequenza a partire da testi campione non porrebbe problemi.

Diverso è il caso delle parole di *classe ouverte* o tematiche che, al contrario, presentano oscillazioni di frequenza da un testo all'altro, e spesso tra brani di uno stesso testo.

Per valutare la loro frequenza, il solo modo operativo conveniente consisterebbe nello stratificare *a priori* la lingua, nel delimitare dei centri d'interesse, all'interno dei quali le parole siano tematiche. L'idea di Moreau è appunto che parole di uso non stabile nella lingua possano invece esserlo in testi con un tema, un centro d'interesse comune. Egli propone perciò di stratificare una volta per tutte la lingua di una data epoca in un certo numero di centri d'interesse, che potranno essere anche molto numerosi, del tipo, per esempio : matematica, fisica, chimica, ecc. All'interno di ciascuno strato sarà determinata la frequenza delle parole tematiche in questo insieme.

I dati da noi ottenuti negli spogli per il lessico di frequenza dell'italiano, sembrano confermare questa seconda concezione; attestano infatti l'esistenza di variazioni nella frequenza delle parole di relazione tra i cinque sottoinsiemi da noi esaminati (cioè: cinema, teatro, romanzi, giornali, sussidiari), variazioni che appaiono strettamente connesse alla natura del sottoinsieme, dal momento che hanno lo stesso andamento per la maggior parte delle unità appartenenti alla medesima categoria grammaticale. Se consi-

deriamo, infatti, i seguenti gruppi di parametri, per i cinque sottoinsiemi di 100.000 occorrenze esatte ciascuno:

A numero dei lemmi
» delle forme
B frequenza del 1° lemma
» del 50 »
» del 100 »
C numero dei lemmi, forme, occorrenze di: ARTICOLI, SOSTANTIVI, AGGETTIVI, PREPOSIZIONI
D numero dei lemmi, forme, occorrenze di: ESCLAMAZIONI, PRONOMI, VERBI, AVVERBI, CONGIUNZIONI,

vediamo che la sequenza è in ordine crescente:

— CINEMA TEATRO ROMANZI SUSSIDIARI GIORNALI, per i parametri A e C e in ordine decrescente:
+ CINEMA TEATRO ROMANZI GIORNALI SUSSIDIARI, per i parametri B e D.

Potremmo rappresentare la situazione così:

CINEMA-TEATRO	ROMANZI	GIORNALI SUSSIDIARI
I	II	III

I gruppi I e III si dispongono simmetricamente rispetto al gruppo II, che, almeno per i parametri considerati, rappresenta valori pressoché coincidenti con il valore medio dei cinque sottoinsiemi.

A quale motivo attribuire queste regolarità? Significano forse che i denominatori comuni, secondo i quali abbiamo raggruppato i testi sono rilevanti rispetto ai parametri citati? Ci proponiamo di sviluppare in seguito l'esplicazione di queste osservazioni, che promettono risultati interessanti anche al di fuori dello studio della struttura quantitativa e del rapporto campione-universo.

La determinazione dei centri d'interesse avrà evidentemente qualcosa di arbitrario e la varietà di strutture e partizioni che si riscontrano nella campionatura dei dizionari di frequenza lo conferma. Gli studiosi del problema

consigliano di aumentare quanto più è possibile il numero degli strati: « Stratifiez à outrance », raccomanda Moreau (*Au sujet de* 1962, p. 157).

D'altra parte la linguistica quantitativa ha già ottenuto alcuni importanti risultati, soprattutto per quanto riguarda l'applicazione e l'adattamento dei metodi classici della statistica ai problemi della campionatura lessicale.

Riteniamo che gli studi tradizionali di stilistica e di lessicografia possano fornire i dati di partenza per una primissima stratificazione *a priori* della lingua, o meglio i dati per identificare, per lo meno come ipotesi di lavoro, alcuni sottoinsiemi sufficientemente definiti ed egualmente caratterizzati rispetto a un parametro comune. Ai risultati degli spogli si applicheranno poi le tecniche statistiche per verificare, da un lato, l'omogeneità delle strutture quantitative dei sottogruppi, dall'altro, per identificare le eventuali parole tematiche.

La determinazione degli strati di lingua potrà derivare da un lavoro accurato e documentato di induzione da spogli sempre più numerosi, ammesso che, naturalmente, questo stesso lavoro provi l'esistenza di strati omogenei e definibili di lingua sul piano delle strutture quantitative di sottoinsiemi linguistici sicuramente definibili. Gli studi più recenti sono ritornati, per così dire, un passo indietro rispetto alle generalizzazioni dei primi tempi. Questo atteggiamento è del resto favorito dallo sviluppo della lessicografia automatica in molti Paesi, dove il numero degli spogli disponibili si accresce giorno per giorno grazie all'attività di centri specializzati.

Il procedimento di stratificare il *corpus*, suggerisce immediatamente l'idea di « completare la nozione di frequenza con quella della stabilità della frequenza, o, se si preferisce, di correggere la frequenza con il modo nel quale si distribuisce nel *corpus* »: è inevitabile che la frequenza globale di un elemento sia soggetta a degli accidenti che porterebbero a sopravvalutarla, per esempio nel caso che una situazione imprevedibile accumuli in una parte del *corpus* un numero elevato di occorrenze di questo elemento (Muller, *Fréquence* 1965, p. 34). Sono stati proposti vari tipi di tecniche per tener conto di questo fatto (A. Juilland 1965, p. XLV): una tecnica semplice, che consiste nell'assumere come indice dell'uso di un'unità lessicale il numero dei sottoinsiemi nel quale occorre (che chiameremo R: se il numero del sottoinsieme è N, R varietà da 1 a N); una varietà di tecniche più complesse che producono diverse formule per misurare la distribuzione della frequenza nei sottoinsiemi.

A. Juilland e E. Chang Rodriguez (1964, pp. LII-LIV) hanno introdotto una misura della ripartizione, che chiamano d i s p e r s i o n e (ingl. *disper-*

sion), nel loro *Frequency dictionary of Spanish words*, basato sullo spoglio di 500.000 occorrenze circa, ripartite in cinque sottoinsiemi di eguali dimensioni, cioè di 100.000 occorrenze ciascuno [1]: l'hanno applicata anche nel volume dedicato al rumeno, e ne preannunciano l'uso nei volumi successivi.

Sarebbe troppo lungo esporre e discutere la formula con la quale viene calcolato l'indice di dispersione e il modo per collegare quest'indice alla frequenza (per cui ottengono un coefficiente che chiamano *usage* e che noi chiameremo u s o), tanto più che sono ampiamente illustrati nei lavori di Juilland a cui rimandiamo (*Frequency dictionary of Spanish words*, pp. LXII–LXIV e LXIV–LXIX).

La formula ha l'effetto di correggere la frequenza con la dispersione piuttosto che viceversa, anche se si potrebbe forse dire che i due fattori sono coordinati. Basterà qui dire che l'indice di dispersione varia da 0 a 1: ha il valore 0 quando tutte le occorrenze sono concentrate in un solo sottoinsieme, assume il valore 1 quando le occorrenze sono uniformemente ripartite nei sottoinsiemi; tra questi due valori l'indice di dispersione cresce proporzionalmente all'uniformità della distribuzione. L'Uso è uguale alla Frequenza quando la parola è ripartita uniformemente nel *corpus*, poiché in tal caso D è uguale ad 1; U = 0. quando D = 0. A Frequenza uguale, U è tanto più vicino a F quanto più la ripartizione è perfetta. U decresce proporzionalmente al decrescere di D. Per altri commenti rinviamo alla citata recensione di Muller.

La significatività dell'indice di dispersione e quindi del coefficiente di Uso può essere accresciuta aumentando il numero dei sottoinsiemi.

La difficoltà di tale aumento risiede nel dover specializzare maggiormente il campione, cioè nell'immaginare un numero maggiore di strati o sottoinsiemi e nel reperire i testi relativi. La elaborazione e il calcolo non ne sarebbero

[1] La coincidenza di questi dati di Juilland con i nostri non è casuale ma voluta e ciò soprattutto allo scopo di avere più agevoli comparazioni non solo con i dizionari di frequenza redatti per lo spagnolo e il rumeno, lingue geneticamente affini all'italiano, ma soprattutto in seguito con quello italiano che rientra nel programma di Juilland e che, data la diversità del *corpus* campione, non sarà certo una ripetizione del nostro lessico. Per questa stessa ragione non abbiamo ritenuto opportuno accogliere nessuna delle modifiche proposte da S. Allen (*Vocabulary data processing*, 1969), da J. B. Carroll (*An alternative to Juilland's usage coefficient...*, 1970) e da J. Mistik (*Frekuencia slov v slovencine...*, 1969) alla formula di Juilland e di Chang Rodriguez.

però minimamente appesantiti o complicati usando il calcolatore; infatti, né la fase di lemmatizzazione né la fase di applicazione comporterebbero maggiori difficoltà di programmazione o di esecuzione.

Il numero di lemmi scelti tra quelli risultanti dallo spoglio per essere pubblicati, è di solito, nei dizionari sinora apparsi, fissato con criteri estrinseci, quali le dimensioni che si vogliono dare al volume o problemi di tipo editoriale. Non esponiamo qui, per limiti di tempo, le ragioni del procedimento che abbiamo adottato: basti dire che abbiamo preso tutti i lemmi (5.356) con U maggiore o uguale a 1,78, che è il valore massimo di U raggiungibile da un lemma con tre occorrenze presenti in tre sottoinsiemi.

Riferiremo invece brevemente su alcuni problemi di lemmatizzazione, operazione imprescindibilmente legata all'intervento umano; ove si tratti infatti, non semplicemente di formare statistiche di parole come unità grafiche e perciò meccanicamente riconoscibili da mezzi automatici, ma di operare analiticamente su parole significative, che, pur nell'identità grafica, possono essere diverse per funzione e per contenuto semantico, la macchina può agevolare ma non sostituire l'opera intelligente dell'uomo [2]. Il calcolatore, in effetti, ci ha fornito una lista di forme basata sull'identità grafica delle singole occorrenze, ma non sempre l'unità grafica coincide con l'unità lessicale: vi sono alcuni casi, infatti, in cui più unità grafiche possono essere considerate una sola parola o per ragioni morfosintattiche, come i tempi composti dei verbi, o per ragioni lessicali, come le locuzioni verbali, avverbiali, ecc. Vi sono viceversa, alcuni casi in cui un'unità grafica rappresenta più unità lessicali come ad esempio le preposizioni articolate o le forme omografe.

La separazione delle forme omografe risalenti a lemmi diversi, che la lingua italiana possiede in gran numero, pone senz'altro i maggiori problemi al lemmatizzatore, poiché oltre a difficoltà puramente pratiche, esige delle decisioni di principio che spesso non sono né semplici né facili. Possiamo innanzi tutto decidere che due forme omografe rappresentano lemmi distinti sia quando vi riconosciamo un significato diverso (omonimi) sia quando pur avendo lo stesso significato le due forme esplicano funzioni morfosintattiche diverse.

[2] Purtroppo manca ancora per l'italiano un d i z i o n a r i o d i m a c c h i n a, che avrebbe notevolmente alleggerito l'opera di lemmatizzazione; esso rientra però nelle nostre ricerche e ci auguriamo di poterlo realizzare quanto prima.

Il primo è il caso dell'omonimia e nella delimitazione e classificazione delle unità sincroniche, è uno degli aspetti più importanti e difficili. La causa più comune dell'omonimia è l'evoluzione fonetica convergente: per influsso di regolari cambiamenti fonetici, due o più parole che avevano un tempo forme diverse, vengono a coincidere nella lingua parlata e in quella scritta, per esempio: *fiera* [bestia], *fiera* [mercato] (omofoni e omografi), e talvolta solo in quella scritta (omografi ma non omofoni), per esempio: *pèsca*, *pésca*; *àncora*, *ancóra*; *balia*, *balia*. L'omonimia può determinarsi anche in seguito ad uno sviluppo divergente del senso: infatti, quando due o più significati della stessa parola si distanziano a tal punto che non esiste più nella coscienza del parlante il senso dell'unica origine, la polisemia lascia il posto all'omonimia e l'unità della parola è distrutta.

Nella lingua comune vi sono molti omonimi secondari di questo tipo e solo un esperto in etimologia sarebbe in grado di connettere *penna* da scrivere con *penna* di uccello, *collo* [bagaglio] con *collo* [parte del corpo].

Non ci soffermeremo qui su casi di omografi ma non omofoni poiché l'aggiunta di un accento risolve facilmente tutti questi casi in cui l'omografia è causata solo da una deficienza del nostro sistema grafico, né agli omonimi parziali, cioè appartenenti a classi diverse di parole per esempio: (la) *faccia* sostantivo femminile e *faccia* congiuntivo presente del verbo *fare*, che non presentano difficoltà di separazione né ambiguità nella classificazione, poiché il criterio morfosintattico è evidente e i contesti ambigui sono poco probabili.

Restano i cosiddetti omonimi effettivi in cui una stessa forma esprime due o più significati senza che la funzione morfosintattica ne sia modificata come ad esempio: *collo* [bagaglio] e *collo* [parte del corpo] o *portiera* [sportello] e *portiera* [portinaia]. Per alcuni di questi casi è veramente difficile dire dove finisce la polisemia e dove comincia l'omonimia, poiché come ha detto Bloomfield non possiamo misurare « il grado di prossimità dei significati », solo in una prospettiva diacronica, il contrasto tra omonimi e unità a significato variabile (polisemia) è netto.

Se il numero dei casi imbarazzanti di omonimia è fortunatamente limitato, il numero, invece, delle parole che hanno assunto un altro valore sintattico per passaggio di categoria grammaticale si può dire infinito e, mentre in alcuni casi è abbastanza semplice stabilire il limite del passaggio poiché è chiaramente definibile, in molti altri è praticamente impossibile determinarlo. Così mentre l'uso sostantivato di infiniti, participi passati e presenti è facilmente delimitabile, per l'uso aggettivale del participio presente e soprattutto del participio passato, è espressamente difficile fissare i limiti e talvolta quasi impossibile misurare il grado del loro cambiamento di categoria, poiché già all'inizio essi hanno qualcosa dell'aggettivo e niente nel loro comportamento viene modificato dal nuovo uso né il senso deve necessariamente cambiare. Anche gli aggettivi sostantivati e i sostantivi in funzione aggettivale presentano dei casi in cui è veramente impossibile determinare il limite del passaggio di categoria, poiché sono passaggi che avvengono spesso e con la più grande facilità e di solito senza alcun cambiamento di significato. In questi casi la separazione è difficile e chi opera queste distinzioni potrebbe essere portato ad adottare delle decisioni diverse per i diversi casi citati.

D'altra parte uno dei principi basilari della statistica è di operare su dati ben definiti: è ovvio, infatti, che da definizioni o convenzioni diverse si otterranno risultati numericamente ben differenti. Ci si è quindi prospettata la necessità di stabilire una norma che sottraesse tali decisioni all'arbitrarietà del lemmatizzatore e all'ispirazione del momento. L'analisi linguistica, però, non dà mai delle classificazioni nette, lascia sempre delle zone di indeterminatezza, nel *continuum* della lingua è, infatti, ben difficile poter tracciare dei limiti netti e dare delle classificazioni precise. Questi motivi ci hanno spinto verso la soluzione, del resto suggerita da precedenti esperienze, che è quella di seguire, o per lo meno di allontanarsene nel minor numero possibile di casi, la norma tradizionale dei dizionari. Nella convinzione che le lacune e le inesattezze della norma lessicografica sarebbero state compensate dalla sua comodità e relativa stabilità e che la stessa definizione di parola, che sembra di poter ricavare dalla pratica dei dizionari, pur confusa e istintiva, ha largamente contribuito a fissare i limiti dell'unità di parola nello spirito dei non specialisti. Tra i molti dizionari della lingua italiana abbiamo dato la preferenza a quello di B. Migliorini che, per l'autorità dell'autore, insigne linguista e storico della lingua, ci è sembrato offrire le maggiori garanzie.

Le operazioni di spoglio dei testi e di calcolo sono state eseguite con gli elaboratori del Centro Nazionale Universitario di Calcolo Elettronico (CNUCE) di Pisa. Sono stati usati i programmi generalizzati messi a punto dalla sezione linguistica del CNUCE che vengono adoperati per numerosi altri progetti. Ci preme sottolineare l'importanza del fatto che si sta costituendo una grande 'biblioteca elettronica' di dati linguistici e di testi, registrati con uniformità di criteri, e quindi interscambiabili e operabili univocamente.

Questa 'biblioteca elettronica', nella quale le unità linguistiche sono registrate secondo criteri linguistici e tecnici uniformi, permettono rapidi conteggi interamente automatici in qualsiasi settore del *corpus*, con la immediata raccolta dei dati quantitativi richiesti dalle elaborazioni statistiche. Sono così possibili l'applicazione e la verifica delle teorie e delle tecniche della statistica linguistica a un gran numero di dati e di testi, che, secondo tutti gli studiosi, sono condizioni essenziali allo sviluppo della statistica linguistica come scienza.

Un *corpus* siffatto permette finalmente lo svolgimento di quel procedimento induttivo di ricerca delle caratteristiche quantitative di strati e sottoinsiemi di lingua, che, come abbiamo detto, è problema centrale nella compilazione dei lessici di frequenza, e al quale la nostra ricerca ha voluto dare un primo contributo per la lingua italiana.

RIFERIMENTI BIBLIOGRAFICI

L. Bloomfield, *Language*, Londra 1935.
M. Boldrini, *Statistica: teoria e metodi*, Milano 1960.
P. Guiraud, *Les caractères statistiques du vocabulaire*, Parigi 1954.
P. Guiraud, *Problèmes et méthodes de la statistique linguistique*, Parigi 1960.
G. Herdan, *Type–token mathematics*, L'Aia 1960.
A. Juilland – E. Chang Rodriguez, *Frequency dictionary of Spanish words*, L'Aia 1964.
A. Juilland – P. M. H. Edwards – I. Juilland, *Frequency dictionary of Rumanian words*, L'Aia 1965.
R. Moreau, *Au sujet de l'utilisation de la notion de fréquence en linguistique*, « Cahiers de lexicologie », 3 (1962), pp. 140–158.
R. Moreau, *Intervention de M. R. Moreau*, « Statistique et analyse linguistique », Parigi 1966, pp. 125–132.
C. Muller, *Fréquence, dispersion et usage: à propos des dictionnaires des fréquence*, « Cahiers de lexicologie », 6 (1965), pp. 32–42.
C. Muller, *Initiation à la statistique linguistique*, Parigi 1968.