STRUMENTI DI LESSICOGRAFIA LETTERARIA ITALIANA
VOLUME 20

GIUSEPPE SAVOCA

LESSICOGRAFIA LETTERARIA E METODO CONCORDANZIALE

LEO S. OLSCHKI EDITORE
MM

LESSICOGRAFIA COMPUTERIZZATA, CRITICA LETTERARIA E CONCORDANZE*

Il critico è uno studioso che, cercando di capire qualcosa dei testi letterari, è continuamente portato a rendersi conto di una realtà elementare, e cioè di avere a che fare con parole di cui per lo più egli non conosce la storia, la distribuzione e gli usi nelle varie opere di uno stesso autore o di un periodo. Egli oggi, nell'era informatica, se non lo è già, sarà presto costretto a prendere atto che ci sono a portata di mano, ma spesso misteriosamente inaccessibili, degli strumenti che, opportunamente addomesticati, possono rendergli utili servizi nello sforzo di accostarsi ai testi attraverso lo specifico della parola.

Da qui, credo, la necessità che egli si interessi alla lessicografia come scienza che, tra l'altro, si occupa della costituzione di vocabolari di opere e di autori, e all'informatica come scienza capace di far padroneggiare una massa di dati, di conoscenze (le parole sono dati) che altrimenti, con l'ausilio della sola memoria o dei repertori artigianali, è pressoché inimmaginabile riuscire a dominare.

Nell'ultimo romanzo di Renato Ghiotto, *I vetri* (Milano, Longanesi, 1987), si può cogliere una visione divoratrice della civiltà informatica, dominata da un grande, mostruoso e invisibile cervello a cui tutti sono asserviti. Il gran sacerdote di questa nuova e terribile religione è colui che, inaccessibile e misterioso, tiene le chiavi del calcolatore. Il protagonista del romanzo è solo uno delle migliaia di modesti servitori del mostro. Al più egli può lasciarsi andare a fantasie di onnipotenza, di fusione con il Grande Cervello che attestano in realtà la sua incapacità di uscire dall'asettica prigione tecnologica nella quale è chiuso.

La cosa interessante è che il narratore in prima persona si dichiara «impiegato della sezione narrativa», e si fa vedere in una cosiddetta «sala dei creativi», in cui «quasi tutti quelli del nostro gruppo fanno passare sul loro schermo brandelli di romanzo e si divertono a manipolarli, a invertire l'ordine delle pagine, a stravolgerli interpolando la frase di uno nella pagina di un altro, o anche semplicemente sosti-

---

* Testo pubblicato, con il titolo *Il letterato nell'era informatica*, in «Rivista IBM», n. 1, 1992, pp. 76-81.

tuendo puerilmente una parola, in modo che la protagonista baci il radiatore del suo amante invece che la bocca, o contando le "e" oppure i punti e virgola di un testo, senza peraltro usare la nozione finale in qualche modo; dubito che in queste operazioni ci sia alcunché di narrativo».

Si legge inoltre nel libro di Ghiotto una definizione del computer come «contenitore di conoscenze (e non di conoscenza)», su cui sarebbe il caso di meditare un poco. In termini appena più tecnici si direbbe che il computer tratta informazioni semplicissime, ridotte a due soli caratteri (lo zero e l'uno). E qualcuno forse si chiederà come sia possibile inserire nel cammino complesso che porta alla conoscenza critica uno strumento che lavora solo sui numeri o, al più, con parole che esso, per le sue elaborazioni, riduce a numeri.

Sospendo la risposta (ammesso che esista) a questa domanda, per prendere l'argomento un po' alla larga, e osservare che il rapporto tra lessicografia computerizzata e critica letteraria si può per analogia assimilare a quello che passa fra conoscenze e conoscenza nella definizione di Ghiotto. La critica è una forma di conoscenza basata su dati, su conoscenze particolari. Tutto ciò che accresce queste conoscenze puntuali, specifiche, si pone già di per sé nel circuito della conoscenza, e quindi della critica.

Nonostante oggi si cominci a diffidare del facile ottimismo informatico, di un'Arcadia telematica che si rivela difficile da abitare, è innegabile che la rivoluzione informatica investe e sempre di più condizionerà tutti gli aspetti e i settori della vita sociale e culturale. Come accadde già al tempo della scoperta rivoluzionaria della stampa, oggi, dinanzi alla diffusione dei sistemi elettronici, nessuna attività intellettuale può fare a meno di confrontarsi con l'uso degli strumenti di calcolo. E nessuno studioso serio, al di là degli usi strumentali di macchine e tecnologie, potrà dire che, sul piano teorico e concettuale, la sua disciplina non abbia niente da spartire con l'informatica.

Informatica vuol dire trattamento automatico delle informazioni, e ogni scienza, ogni pensiero, ogni atto umano hanno a che fare sempre con informazioni che vengono elaborate, sintetizzate, trasformate, interpretate. Tutto ciò che è strutturato in un modello formalizzabile è per ciò stesso informatizzabile, cioè riducibile a oggetto di elaborazione informatica. È questo il caso di tutte le scienze fisiche; ma è anche il caso della critica letteraria? Sembrerebbe di no, stando almeno alla condizione attuale degli studi e alle prese di posizione di critici autorevoli (deducibili, però, più che altro *e silentio*).

Per quanto mi riguarda, credo che lo statuto della critica letteraria sia in movimento, e che essa, tramontata l'equivalenza tra critica e valutazione (estetica, ideologica, sociologica, ecc.), tenda ad acquisire il rigore e l'oggettività delle altre scienze. È obiettivamente su questa tendenza della critica a porsi come scienza che essa potrà incontrarsi proficuamente con la scienza del computer in forme e modi di cui ancora si intravede ben poco. Tuttavia, se riflettiamo per un attimo sul ruolo sempre

più decisivo che gli strumenti linguistici hanno nella formazione e nella elaborazione dei concetti critici, e se consideriamo il fatto che la linguistica al calcolatore ha già un suo proprio ambito metodologico e applicativo, ci verrà facile immaginare che, con la mediazione della linguistica, la critica potrà non essere più considerata estranea all'ambiente delle scienze informatiche.

Com'è a tutti noto esiste, da molto prima dell'avvento dei calcolatori, un settore di studi che si chiama linguistica matematica o statistica linguistica. Uno dei princìpi di base di questa disciplina, in verità in Italia (a parte lodevoli eccezioni) poco coltivata, riguarda il dato elementare che in un sistema linguistico ogni parola, accanto alla forma e al contenuto semantico, possiede, anche se i valori relativamente al codice linguistico italiano non sono noti, una frequenza assoluta e una determinata probabilità di ricorrenza. Il che significa, schematizzando molto, che se tutte le parole di questa mia nota venissero registrate e poi memorizzate al computer e sottoposte a una semplice elaborazione quantitativa, esse si potrebbero leggere, ad esempio, in una lista di frequenza nella quale tutti i vocaboli da me usati sarebbero accompagnati da un numero indicante quante volte ognuno di essi è stato da me adoperato.

È facilmente prevedibile (prescindendo dalle parole funzionali ad alta frequenza come articoli, congiunzioni, preposizioni, ecc.) che in uno scritto che si occupa dei rapporti tra linguistica, critica e informatica queste parole abbiano un'alta frequenza perché, ovviamente, l'argomento determina la scelta di alcuni termini, che sono poi le parole tema o anche le parole chiave di un determinato atto di linguaggio. Se poi l'elaborazione quantitativa si facesse su una zona più vasta delle attività linguistiche di un soggetto, come di qualunque testo o autore, si scoprirebbe che un piccolo numero di parole di base o di lemmi fondamentali ad alta frequenza costituisce la maggior parte di ogni discorso, come di ogni opera.

Secondo i rilievi di Pierre Guiraud, che ha studiato i caratteri statistici del vocabolario francese, le 100 parole più frequenti occuperebbero il 60% di qualunque testo scritto in francese, mentre le prime 1000 ne coprirebbero l'85%. Ognuno vede, ad esempio, quale vantaggio trarrebbero da simili dati coloro che volendo fare un corso rapido di una lingua straniera puntassero direttamente sul vocabolario fondamentale individuato con metodi quantitativi. Conoscendo 100 parole di una lingua straniera, e alcune regole di combinazione, potremmo farci capire almeno al 60% delle nostre esigenze. E non è poco!

Non disponiamo, a mia conoscenza, di simili studi per l'italiano. Tuttavia, stando ai dati che io posso fornire servendomi delle concordanze pubblicate e da pubblicare a mia cura nella serie di «Strumenti di Lessicografia Letteraria Italiana», sono nella condizione di potere rilevare che quella di Guiraud non è una legge che valga immutabile anche per l'italiano letterario (e poetico in particolare). Infatti, i testi delle poesie di alcuni autori, che cito a scopo esemplificativo, danno valori costantemente più bassi del 60% ipotizzato da Guiraud per il francese. Le prime cento parole in Gozzano coprono il 50,80% del testo (20.090 su 39.557), in Corazzini il 56,22%

(10.084 su 17.936), in Sbarbaro il 57,36% (7.902 su 13.774), in Cardarelli il 54,02% (6.058 su 11.213), in Ungaretti il 51,66% (10.399 su 20.128), in Palazzeschi il 51,90% (28.551 su 55.008), in Montale il 54,71% (40.318 su 73.682), ecc.

Non avanzo ipotesi di comparazione di questi dati, ma credo sia il caso di osservare che non è possibile condurre alcuna analisi di linguistica testuale su base quantitativa se non si dispone di dati su autori e opere diversi che possano venire confrontati. Questo è un principio fondamentale. Non si può ritenere per nulla seria una critica che pretenda di analizzare un testo su base quantitativa se non si hanno punti di riferimento esterni.

Faccio un solo esempio, relativo al campo lessicale ruotante intorno a «morire» («morte», «mortale», ecc.), che in un poeta come Sergio Corazzini (un crepuscolare morto a ventun anni) supera la percentuale dell'uno per cento. Il che significa che su ogni cento parole presenti nelle sue poesie almeno una è attinta tra le parole con la radice del verbo «morire». Il dato isolato di per sé non è significativo di una poesia che canti, per così dire, l'essere per la morte. Ma se possiamo confrontare questo valore con gli altri di poeti cronologicamente vicini a Corazzini, come Pascoli, Palazzeschi, Gozzano, Cardarelli, Moretti, Govoni, Campana, D'Annunzio, ecc., ci accorgiamo che in questi il campo lessicale di «morire» presenta valori che vanno dallo 0,11% di Campana allo 0,48% di D'Annunzio (*Poema paradisiaco*). A questo punto la deduzione statistica e critica da trarre è che, presentando Corazzini un valore percentuale più che doppio rispetto a quello di altri poeti, in lui il tema del «morire» deve essere assolutamente centrale. L'analisi concreta del testo ci fornisce poi gli elementi per dare significato critico al peso statistico del «morire». Naturalmente questo è solo un esempio minimo tra i tanti che si potrebbero addurre. Va, in generale, osservato che per fare questi rilievi occorre avere colto il nesso tra lessicografia computerizzata e critica, mentre, a monte, è indispensabile poter disporre di realizzazioni moderne di questi strumenti antichi ma, ahimè!, abbastanza rari per la letteratura italiana che sono le concordanze.

Una concordanza all'altezza dei tempi, rispetto a quelle tradizionali, a mio modo di vedere deve servire non solo a documentare esaustivamente il lessico di un autore sul piano dei significanti (e quindi dei significati), ma anche deve contribuire all'arricchimento dei valori quantitativi espressi da un determinato *corpus* linguistico.

Non mi sembra assurdo prevedere che, alla distanza, l'uso del computer possa far cambiare fisionomia alla critica letteraria. Per il momento essa potrebbe servirsi delle concordanze (o meglio dedicarsi anche alle concordanze) in una duplice direzione che indicherei, la prima, come lettura paradigmatica dei testi, e come lettura propriamente informatica la seconda.

È ovvio che queste due direzioni di indagine si possono distinguere solo a scopo didattico in quanto esse sono in realtà convergenti. La premessa a questo tipo di lettura sta però nella disponibilità di strumenti, appunto le concordanze, che ubbidiscano ad alcuni requisiti di scientificità senza i quali la critica saprà sempre di vecchio

umanesimo (e dico ciò con tutto il rispetto non solo per la critica "classica", ma anche per i molti eccellenti studiosi che non riescono tuttora a confrontarsi con le nuove tecnologie di trattamento dei testi).

Non ha molto senso presentare oggi come uscite dal computer concordanze parziali e senza dati numerici. La parola «computer» viene da «computare», che significa calcolare. Ma non facendogli eseguire sull'opera che si va a concordare alcune semplici operazioni aritmetiche di cui esso è capace, il calcolatore, rispetto a una macchina da scrivere, offre in più il modesto servizio di ordinare alfabeticamente le parole.

Tornerò per concludere su quella che ho definito lettura paradigmatica, mentre per la lettura informatica è giocoforza confessare che ne sappiamo ancora troppo poco. Si può formulare qualche ipotesi, nel senso che tutto ciò che in un'opera è quantificabile e formalizzabile ha un suo rilievo e un suo marchio d'autore. Si potrebbe immaginare che, grazie al computer, noi potremmo avere una scheda per così dire informativa (oltre che informatica), la quale di ogni opera e di ogni autore ci desse dei dati relativi all'estensione del *corpus*, alle frequenze, alla distribuzione delle categorie grammaticali, ai nessi sintattici, alle ripetizioni, ecc.

Per parlare un po' all'ingrosso, direi che non è, ad esempio, senza importanza e senza significato conoscere l'estensione del lessico di un autore. Non è nemmeno senza significato, ad esempio, potere osservare che di opera in opera Montale presenta un tasso si accrescimento del lessico tra il 25 e il 35%: il che significa che in ogni nuova opera egli usa una percentuale, oscillante tra il 25 e il 35%, di parole sconosciute alla produzione poetica precedente. Sapere, ad esempio, che la lunghezza del testo delle poesie di Montale, che comprendono [prima del *Diario postumo*] 73.682 occorrenze di parola, è circa tre volte e mezza più ampia di quella di tutte le poesie di Ungaretti non ci dice molto sulla grandezza e sul valore dei due poeti; ma certo ci dà un elemento descrittivo, identificativo, che ha un suo interesse. Conoscere, ad esempio, che Petrarca nel *Canzoniere*, con un lessico di poco superiore per estensione alla metà di quello della *Divina Commedia*, usa in media una parola 17,43 volte, mentre Dante 13,57 volte, e ciò contro le aspettative statistiche, ci induce a considerare un elemento di "apprendibilità" quantitativa della lingua petrarchesca che, tra altre ragioni, contribuisce a spiegare la maggiore fortuna di Petrarca nel corso della storia poetica italiana. Non sarebbe senza importanza formalizzare, ad esempio, la metrica, la lunghezza dei versi, ecc.

Nella lettura paradigmatica vedrei un coronamento per quanto possibile scientifico, cioè fondato su dati e su elaborazioni al computer, della critica tradizionale su base testuale e linguistica. Lo strumento principe per questa auspicabile svolta nel lavoro critico a me sembra sia offerto dalle concordanze del lessico dell'autore che il critico intende studiare.

Anche se non dichiara i significati dei lemmi individuati ed esposti come entrate lessicografiche, una concordanza esaustiva è un dizionario speciale che nei confronti

dei normali vocabolari generali ha il grande vantaggio di essere completo e definitivo, rispetto alla "pronunzia" e al "possesso" che di un certo patrimonio di parole ha esercitato l'autore concordato.

Come i dizionari, anche le concordanze hanno un loro particolare fascino e una loro "poesia", ma esse, è bene ribadirlo, non sono la poesia di cui danno, con le parole (anche con tutte le parole di un testo), solo un'immagine. Per dirla con Montale, «le ombre che si nascondono / tra le parole», il «buio / delle parole», le "discordanze" tra segni e senso, tra parole dette e parole taciute, tra la scrittura e la sua ombra si sottraggono alla presa di qualunque dizionario. Quello che sempre sfugge e però sempre si cerca, con una concordanza come con un saggio critico, è proprio la poesia. Al limite una concordanza, non dichiarando i significati, è la forma meno arbitraria di vocabolario di un testo e insieme la forma di interpretazione più rispettosa del significato e del mistero della sua poesia.