Concordanze per forma o lemmatizzate?

Giovanni Nencioni

L'iniziativa, nuova nella storia dell'Accademia della Crusca, di applicarsi alla lessicografia e alla lessicologia delle lingue settoriali, in particolare della nomenclatura tecnica, ha trovato nell'accordo di collaborazione con la Scuola Normale Superiore di Pisa, e precisamente col suo Centro di ricerche informatiche per i beni culturali, un punto di riferimento e di confronto essenziale. Quel Centro è un laboratorio la cui ricca attrezzatura tecnologica è governata da tecnici che conoscono non solo le risorse dell'informatica, ma i problemi per i quali le discipline scientifiche o umanistiche ricorrono ad essa. E non si può dire di quanto conforto sia al linguista o filologo o letterato tecnicamente inesperto vedersi chiarire le possibilità e i limiti della "macchina" da chi non ignora le sue necessità e prevede le sue richieste. Gli si escludono possibilità da lui ingiustamente attribuite all'automa e sottratte alla propria mente, ma gli si offrono strumenti per operazioni cronologicamente o mentalmente impervie o per connessioni insospettate. La sua alacrità mentale viene accelerata ed esaltata, ma anche ridefinita, come quella dello scienziato che si sente integrato ma non diminuito dalla strumentaria del laboratorio. Vale a dire che, nel caso concreto, lo studioso ormai consapevole di quel che può chiedere alla propria mente e allo strumento, ha la scelta di ripartire i compiti motivando razionalmente i termini della cooperazione.

Porto come esempio il caso concreto occorso a me, consulente lessicografico della Crusca, nell'avviare la collaborazione col Centro di ricerche informatiche della Scuola Normale. Si trattava d'impostare le concordanze di una vasta opera, *Le vite de' più eccellenti pittori, scultori e architettori* di Giorgio Vasari, un capolavoro a doppio titolo, come creatore di un nuovo genere, la moderna storiografia artistica, e come testo letterario. Consigliavano l'impresa la scarsa presenza del lessico vasariano nei dizionari della lingua comune e il legittimo desiderio di mettere finalmente a disposizione degli studiosi lo spoglio completo e preciso di un universo terminologico e concettuale concernente le arti.

Erano possibili, a detta dei tecnici informatici, due vie operative, sfocianti in due soluzioni diverse: attuare le *concordanze per forma,* cioè collocando in

esclusivo e primario ordine alfabetico le singole parole (con termine informatico *occorrenze*) uscite dallo spoglio automatico del testo, ognuna contestualizzata al centro di un segmento tagliato automaticamente, oppure attuare le *concordanze lemmatizzate*, cioè collocare le parole in un ordine alfabetico secondario, dopo aver sussunto le varianti flessionali di ognuna sotto la forma base del paradigma posta come lemma, e magari, per colmo di perfezione, tagliare i segmenti di contestualizzazione in modo che conservassero una struttura sintattica e semantica compiuta. La soluzione delle concordanze lemmatizzate era certo quella preferibile, anche perché corrispondente alla regola seguita dai normali dizionari, potendosi ovviare alla difficoltà di reperire la forma base del paradigma di forme eteroclite mediante rinvii. Ma sarebbe occorso inserire nel processo automatico una lunga fase di lavoro manuale, cioè il lavoro stesso della lemmatizzazione, perché una lemmatizzazione automatica è, a tutt'oggi, eseguibile solo se la memoria informatica possiede un dizionario di riferimento, che non esiste per la nostra lingua antica e, anche nel caso di un'operazione eseguita sulla lingua moderna, per la quale il dizionario di riferimento esiste, abbisogna di un accurato lavoro di revisione. La lemmatizzazione della lingua antica, poi, solleva difficoltà particolari, quali la supposizione del lemma base quando esso non sia attestato, o la normalizzazione del lemma base quando il testo fornisca solo forme eteroclite rispetto a quella divenuta normale. La soluzione delle concordanze per forma si presentava, invece, spedita, perché eseguibile in tutto e per tutto automaticamente, fondata com'è sulla forma grafica della parola e quindi distinguente fino alle varianti prodotte, a parità di segni fonetici, dai segni diacritici introdotti dall'editore. Le concordanze per forma coincidono così con l'indice di frequenza, che segue un ordine alfabetico primario ed è privo di rinvii; l'inserimento di questi introdurrebbe infatti nel processo automatico delle occorrenze per forma una fase manuale, perché implica un esame di confronto e di riconnessione tra forme diverse, che è l'operazione fondamentale della lemmatizzazione.

Si doveva dunque, per le concordanze vasariane, scegliere tra le due soluzioni: a lungo o a breve termine. Abbiamo scelto quella delle concordanze per forma, cioè della totale automazione, quindi a breve termine. Era la via - ci siamo obiettati umanisticamente - del macchinismo bruto e cieco, della inintelligenza, della indistinzione; ma è poi prevalso un ragionamento di genere pragmatico, che possiamo riassumere così. L'informatica è un leviatano benefico e terrifico a un tempo. Ha aperto all'uomo un gorgo di strumentalità cognitiva, la cui voracità urge di essere saziata, tanto nel campo scientifico e tecnologico che umanistico. Mentre, prima della comparsa del mostro, il linguista o filologo o critico letterario includeva nella propria ricerca una fase preparatoria, non essendo i dizionari a sua disposizione strumenti obiettivamente sufficienti a una indagine rigorosa ma dovendo di volta in volta essere verificati e completati dallo studioso, oggi questo - poniamo il critico letterario o lo storico della lingua - sente con struggimento la mancanza di una biblioteca di concordanze di tutti i testi della letteratura italiana, pensando

quali risultati potrebbe trarne rapidamente, interrogandole con sapienti *thesauri*, cioè risparmiandosi i faticosi e parziali lavori di scavo o di controllo cui era costretto alcuni decenni fa. Né gli passa più per il capo il pericolo di affogare sotto una valanga di schede, come accadde realmente - stando alla testimonianza di Anatole France - al dottissimo Fulgence Tapir che aveva schedato ciò che concerneva le arti di tutto il mondo, ivi compresa l'arte dei Pinguini; perché le banche dati informatiche sono immediatamente e fulmineamente accessibili a distanza, il che rende ancora più struggente il senso della loro rarità e povertà. Il male è che il leviatano non può saziarsi da solo, ma con l'aiuto di lessicografi speciali, i quali devono rendersi conto di essere, a differenza degli autonomi lessicografi preinformatici, in un rapporto diretto e complementare con studiosi che aspettano il risultato del loro lavoro, ponte necessario tra i testi e moderne analisi che esigono dati non approssimativi e parziali, ma esatti e totali, statisticamente apprezzabili.

Rendendoci conto di ciò e sentendo, di conseguenza, il nostro rapporto di solidarietà, oltre che di complementarità, con quegli studiosi, abbiamo pensato che era nostro primo dovere gettare nelle fauci del leviatano il nostro piccolo ma sollecitamente utile tributo; cioè procurare al più presto uno strumento meno elaborato e raffinato, ma perciò più oggettivo, e meno oblativo per il consultatore comune, ma bene dominabile dagli studiosi cui principalmente è rivolto.

Ci pare opportuno render ragione qui di una decisione meditata, perché il campo in cui lavoriamo è mosso da orientamenti ed esperienze diversi, prodotti dai progressi della tecnologia informatica e degli studi umanistici collegati con essa; orientamenti ed esperienze che è utile siano conosciuti, confrontati e discussi.