PARTE QUARTA

RAPPORTI SULLE RICERCHE LESSICOGRAFICHE

IL LESSICO ITALIANO DELLE ORIGINI E L'INFORMATICA LINGUISTICA

D'Arco Silvio Avalle

I lavori per la compilazione del «Vocabolario storico della lingua italiana» (VSLI), promosso dall'Accademia della Crusca in collaborazione per la parte elettronica con il CNUCE di Pisa, hanno avuto inizio a Firenze il 1 ottobre 1965. Il progetto prevedeva uno spoglio completo dei testi più antichi sino alla fine del XIV secolo, ed uno spoglio selettivo, invece, dei testi posteriori sino all'epoca moderna. Principio fondamentale era di prendere in esame le opere di cui si avesse un'edizione a stampa.

Nel 1972, considerata la mole del *corpus*, l'Accademia decise di concentrare provvisoriamente i suoi sforzi sui testi dei primi secoli, dalle origini al 1375, in vista della pubblicazione di un vocabolario parziale, denominato «Tesoro della lingua italiana delle origini» (TLIO).

Alla fine del 1974 i materiali relativi ai secoli posteriori al XIV che erano stati spogliati elettronicamente, ammontavano già a 151 opere per un totale di 27.300 pagine. Di tali opere, 84, per un totale di 15.183 pagine, erano giunte a livello di lista-testo. Altre 37, per un totale di 9.010 pagine, erano giunte a livello di concordanza delle forme. Per altre 30 opere, infine, corrispondenti a 3.107 pagine, si avevano le concordanze dei lemmi: fase questa che precede immediatamente la compilazione della scheda-contesto.

Alle opere qui sopra elencate vanno aggiunte quelle spogliate con metodi manuali o xerografici. Alla fine del 1974 tali opere, sempre limitatamente ai secoli posteriori al XIV, ammontavano a

234, per un totale di 64.816 pagine. Da esse erano state ricavate 272.418 schede-contesto pronte ad essere immesse nell'Archivio dell'«Opera del Vocabolario».

In conclusione, sommando i dati delle opere spogliate elettronicamente e di quelle spogliate manualmente o xerograficamente, si ha un complesso di 385 opere per un totale di 93.116 pagine, ferme ai vari livelli della lista-testo, delle concordanze delle forme e delle concordanze dei lemmi per la sezione spogliata elettronicamente, ed alla fase terminale della scheda-contesto per la sezione spogliata manualmente o xerograficamente.

Gli spogli manuali e xerografici, finalizzati sostanzialmente alla redazione delle voci, avevano il vantaggio di procedere più speditamente di quelli elettronici. Essi però rappresentano una fase intermedia nella compilazione del VSLI, senza possibilità di utilizzazione immediata per chi si voglia documentare esaustivamente su questo o quel settore della lingua italiana. Gli addetti a questo tipo di spoglio scelgono infatti solo le forme lessicograficamente più interessanti, per cui il materiale sino ad ora raccolto con tale metodo, se può riservare sorprese gradevoli al ricercatore, non dà certezze di alcun genere, non dico, come è ovvio, sul piano statistico, ma neppure su quello, poniamo, sintattico e morfologico.

Col passare del tempo ed in considerazione del crescente interesse per una lessicografia di carattere statistico e matematico, gli spogli manuali e xerografici, di impostazione, tutto sommato, tradizionale, vennero progressivamente abbandonati in favore dello spoglio elettronico, forse meno produttivo in fatto di resa immediata, ma utilissimo per quel che riguarda l'esaustività dell'informazione, soprattutto a livello di concordanza delle forme. In questo campo l'«Opera del Vocabolario» ha raccolto un materiale vastissimo che si propone di mettere a disposizione degli studiosi

tramite un catalogo denominato «Lista delle concordanze delle forme delle opere posteriori al XIV secolo possedute dall'Opera del Vocabolario».

Sempre a proposito dell'attività dell'«Opera», andrà ricordato che per i testi del Duecento l'Accademia aveva pensato, in un primo tempo, ad una collaborazione con l'Istituto di Lingua e Letteratura Italiana dell'Università di Utrecht. Tale collaborazione è però venuta meno ad un certo punto per ragioni varie che è inutile qui elencare. In compenso il Centro di Lessicografia dell'Accademia ha acquisito, grazie alla generosità del defunto socio dott. Raffaele Mattioli, i materiali raccolti ai fini delle «Concordanze della lingua poetica italiana delle origini» (CLPIO) programmate dalla Casa Editrice R. Ricciardi di Milano, con l'intesa che le schede-contesto ricavate da tali «Concordanze» vengano immesse nell'Archivio dell'«Opera».

Come si è detto, nel 1972 avevano inizio i lavori per il «Tesoro» (TLIO), lavori, per altro, già preventivati sin dal 1969 (cfr. SFI, XXVII, 1969, p. 258). Nonostante le più limitate prospettive, l'impresa non poteva dirsi di poco conto, sia per la mole, sia per la natura dei materiali da spogliare. Infatti ad un calcolo approssimativo, le opere rientranti nel canone comprendono circa diciotto milioni di occorrenze. La cifra, di tutto rispetto, soprattutto se si tien conto del numero ridotto dei collaboratori addetti alle operazioni di spoglio, è tanto più preoccupante, quando si pensi alla qualità di quelle opere e, in particolar modo, delle loro edizioni. I testi delle origini vogliono dire un'altra lingua, nel complesso differentissima dall'italiano standard, la cui lettura e lemmatizzazione coinvolgono spesso responsabilità e competenze particolari. Quanto poi alla qualità delle edizioni, tali responsabilità aumentano proporzionalmente col diminuire dell'attenzione e

della abilità dei singoli curatori. Di qui la necessità di tutta una serie di controlli e di verifiche al fine di ridurre al minimo i guasti di una editoria non sempre precisa o, comunque, tecnicamente superata.

Fra i vari problemi connessi con lo spoglio del materiale delle origini, il più importante, a mio avviso, è quello riguardante la sua definizione linguistica e culturale. Ricorderò, a questo proposito, che l'«Opera del Vocabolario» è da tempo in possesso di una massa ingente di dati bibliografici, tuttora in via di arricchimento, relativi alle edizioni a stampa dei testi appartenenti all'epoca prevista dal TLIO. Tale raccolta, unica nel suo genere, ma dispersa fra i vari Uffici dell'«Opera», è stata sottoposta ultimamente ad un'attenta revisione e ad una rielaborazione unitaria, al fine di seguire più da vicino il progredire dei lavori del TLIO e di rendersi, nello stesso tempo, più esatto conto della consistenza del *corpus*. I criteri adottati al riguardo sono stati di ridurre le schede bibliografiche ad un modello standard che ne renda più agevole la consultazione e, nello stesso tempo, di arricchirle di un certo numero di informazioni tanto dal punto di vista bibliografico, quanto da quello storico, letterario e linguistico.

Dopo varie prove, si è deciso di dare alle schede la seguente struttura (si veda l'allegato n. 1).

1. In alto a sinistra è indicata la data o, in mancanza di meglio, l'epoca approssimativa cui risale l'opera. Per convenzione tale data può essere rappresentata con il *terminus ante quem* (B), con l'anno preciso (C), con l'anno preciso seguito da *circa* (D), o, ancora, da un punto interrogativo (E), il *terminus post quem* (F), con l'indicazione di due date precise, quando se ne conoscano la data di inizio e di fine (G), con l'indicazione di due date precise separate da una barra obliqua, quando si vogliono indicare i termini cronologici entro cui oscilla l'anno di composizione dell'ope-

ra (H), con un'espressione più vaga, ad esempio sec. XIII, prima metà del XIV secolo, e così via (I), ed, infine, per le opere scritte alla fine o all'inizio di un secolo, con l'indicazione rispettivamente *exeunte* ed *ineunte* (L). Tali indicazioni, opportunamente codificate, come risulta dalla sigla posta in alto a destra delle singole schede, permettono al calcolatore di ordinare le opere cronologicamente e, nel caso di aggiunta di nuove opere, di inserirle al posto che loro compete. In caso di opere la cui datazione coincida, l'ordine adottato è quello alfabetico del nome dell'autore e, per le adespote, del titolo o dell'incipit.

2. Seguono i dati bibliografici dell'opera, che comprendono:

2.1. Il nome ed il cognome o, comunque, gli altri dati onomastici dell'autore; fanno eccezione Petrarca, Boccaccio, Pucci, Sacchetti, per cui si è preferito l'ordine inverso; nel caso che l'opera sia adespota, comparirà il titolo o l'incipit.

2.2. Il titolo o l'incipit dell'opera

2.3. Le indicazioni relative all'edizione moderna scelta dall'«Opera del Vocabolario» (talune di queste scelte potranno essere modificate nel corso del tempo), vale a dire, nell'ordine:

2.3.1. il cognome e il nome del curatore

2.3.2. il titolo del volume complessivo, nel caso che l'opera sia contenuta in una miscellanea, una antologia, e così via

2.3.3. l'editore

2.3.4. il luogo di edizione

2.3.5. la data di edizione

2.3.6. ed, eventualmente, le pagine dove l'opera è stampata.

Per le opere edite in pubblicazioni periodiche, si indicano:

2.3.7. la sigla della rivista

2.3.8. il numero del volume, della serie, ecc.

2.3.9. l'anno di pubblicazione, e, infine

2.3.10. le pagine

3. Chiudono la scheda dodici diversi tipi di informazione relativi a:

3.1. il genere di spoglio (T.S.) effettuato, ora costantemente elettronico (E).

3.2. il numero delle pagine (N. PAG.) occupate dal testo nell'edizione.

3.3. la fase di elaborazione cui il testo è pervenuto (FASE). Si è tenuto conto delle fasi più importanti, sei in tutto, ognuna delle quali è contrassegnata da una sigla. Laddove tale sigla manca, significa che il testo è ancora giacente presso l'Ufficio Filologico.

3.4. il numero delle occorrenze (OCCORRENZE), approssimativo e non ancora calcolato per le opere tuttora giacenti presso l'Ufficio Filologico.

3.5. le schede-contesto ottenute (SKS).

3.6. il tipo di controllo, quando effettuato su ms., su microfilm, ecc., parziale o totale, e così via (CONTROLLO).

3.7. il movimento (MOV), che viene contrassegnato in perforazione con un asterisco ogni qual volta l'opera subisca un avanzamento di fase (cfr. 3.3.). Tale informazione servirà ad automatizzare i rendiconti semestrali sul lavoro compiuto nell'ambito dell'«Opera».

3.8. la localizzazione geografica dell'opera ricavata dai dati linguistici. Nel caso di opere trascritte da amanuense alloglotto e adattate al suo modello linguistico, si dà, quando posssibile, una doppia indicazione relativa alle due zone di provenienza rispettivamente dell'autore e del trascrittore. La localizzazione comporta un reticolo abbastanza fitto, comprendente ben centodieci zone linguistiche.

3.9. l'indicazione se si tratta di opera originale, oppure di volgarizzamento, o, ancora, se è mista.

3.10. l'indicazione se l'opera è in versi, in prosa o mista di prosa e di versi.

3.11. l'indicazione della categoria o genere letterario cui l'opera appartiene. Anche in questo caso si è stati abbastanza analitici, come risulta dal siglario che prevede ben ventuno categorie.

3.12. In fondo a sinistra delle singole schede è infine indicato in codice la denominazione del nastro magnetico contenente l'opera schedata.

La raccolta di queste indicazioni è costata uno sforzo notevole ed un'attenzione particolare. Tutto il materiale assomma a circa 1.800 schede, già tutte perforate ed in fase di correzione, oltre ai rinvii dell'autore all'opera e viceversa. I dati tuttora mancanti (nella fattispecie il computo approssimativo delle occorrenze per le opere tuttora giacenti presso l'Ufficio Filologico oppure ferme alle prime fasi dell'elaborazione elettronica) sono in via di completamento. A conclusione di questi lavori le schede verranno pubblicate in un'opera a parte intitolata «Le edizioni a stampa dei testi italiani ad uso del TLIO» (ES).

Le informazioni raccolte nell'ambito di questo progetto hanno un'importanza tutta particolare, sia sotto il profilo linguistico e letterario, sia per quel che riguarda i bisogni dell'utenza (cfr. 3.1., 3.3., 3.5., e 3.6.) e le esigenze dell'amministrazione (cfr. 3.7. e 3.12.).

Di interesse linguistico-letterario sono le informazioni relative alla data di composizione (1.), il numero delle occorrenze (3.4.), il tipo di lingua impiegata (3.8.), la qualità dell'opera, se originale o volgarizzamento (3.9.), la struttura formale, prosa o poesia (3.10.), e, infine, la categoria o genere letterario (3.11.). Com'è noto, l'apprezzamento lessicografico delle singole voci dipende in larga misura, oltre che dal contesto, dall'incidenza che i dati qui sopra elencati hanno sul loro valore; dati tanto più importanti,

quando si pensi ai condizionamenti linguistici derivanti nell'epoca medioevale alle singole opere dell'assetto formale ad esse assegnato dai loro autori. Ora, la somma di questi dati permette di porre al calcolatore domande di vario genere a seconda delle combinazioni in cui essi possono essere disposti: come, ad esempio, nel caso che si vogliano enucleare le voci (testi) di una certa zona linguistica per un periodo determinato e, nel caso, nell'ambito di una certa categoria o genere letterario, oppure calcolare i rapporti percentuali fra opere originali e volgarizzamenti in una certa zona ed in una certa epoca, e così via.

Ed ecco un elenco, puramente orientativo, delle combinazioni più importanti, tenendo come punto di riferimento fisso il numero delle occorrenze (3.4.).

I. graduatoria delle zone linguistiche (combinazione 3.4. X 3.8.)

II. graduatoria delle categorie o generi letterari (combinazione 3.4. X 3.11.)

III. graduatoria delle zone linguistiche nell'ambito delle singole categorie o generi letterari (combinazione 3.4. X 3.8. X 3.11.)

IV. rapporti percentuali fra prosa e poesia (combinazione 3.4. X 3.10.)

V. graduatoria delle zone linguistiche nell'ambito della prosa e della poesia (combinazione 3.4. X 3.8. X 3.10.)

VI. rapporti percentuali di opere originali e volgarizzamenti (combinazione 3.4. X 3.9.)

VII. graduatoria delle zone linguistiche nell'ambito delle opere originali e dei volgarizzamenti (combinazione 3.4. X 3.8. X 3.9.)

VIII. produzione totale per sezioni cronologiche determinate (combinazione 3.4. X 1.)

IX. graduatoria delle zone linguistiche nell'ambito di sezioni cronologiche determinate (combinazione 3.4. X 3.8. X 1.)

X. graduatoria delle categorie e dei generi letterari nell'ambito di sezioni cronologiche determinate (combinazione 3.4. X 3.11. X 1.)

XI. rapporti percentuali di prosa e poesia nell'ambito di sezioni cronologiche determinate (combinazione 3.4. X 3.10. X 1.)

XII. rapporti percentuali di opere originali e di volgarizzamenti nell'ambito di sezioni cronologiche determinate (combinazione 3.4. X 3.9. X 1.)

XIII. graduatoria delle zone linguistiche nell'ambito delle singole categorie o generi letterari suddivisi in sezioni cronologiche determinate (combinazione 3.4. X 3.8. X 3.11. X 1.)

XIV. graduatoria delle zone linguistiche nell'ambito della prosa e della poesia suddivisa in sezioni cronologiche determinate (combinazione 3.4. X 3.8. X 3. 10. X 1.)

XV. graduatoria delle zone linguistiche nell'ambito delle opere originali e dei volgarizzamenti suddivisi in sezioni cronologiche determinate (combinazione 3.4. X 3.8. X 3.9. X 1.) e così via.

Visto che i dati del repertorio (ES) sono praticamente esaustivi, almeno al limite delle informazioni raccolte sino ad ora (la distribuzione delle opere andate perdute nel corso dei tempi, o rimaste inedite, o sfuggite alla nostra attenzione è troppo «dispersa» perché ne venga modificato sostanzialmente il quadro generale), la lingua e la letteratura italiana dei primi secoli potranno d'ora in poi essere studiate su un piano rigorosamente statistico anche dal punto di vista della loro strutturazione socio-culturale, con notevole beneficio per l'articolazione interna del «Tesoro».

Lo stesso schema, sia pure con alcune semplificazioni, è stato adottato per la scheda-contesto. Essa presenta le seguenti caratteristiche.

In alto al centro abbiamo la forma, così come si trova nel testo. In alto a sinistra la voce o lemma cui la forma va riportata: ad esempio l'infinito per i verbi, il singolare per i sostantivi, l'espressione completa delle congiunzioni subordinanti composte per i loro singoli elementi, quando essi siano separati nella stampa, ecc. Come risulta dai fascicoli (cfr. allegato n. 2), il lemma è connotato grammaticalmente — ad esempio V. — verbo, S.M. — sostantivo maschile — o, addirittura, semanticamente — ad esempio *mondo (universo)*. Tali indicazioni servono a risolvere il problema dello smistamento automatico degli omografi a livello di lemma. Nel nostro caso *salutare V.* si distinguerà, in absentia, da *salutare AGG.*, oppure *mondo (universo) S.M.* — ma è possibile anche *mondo S.M.* senza connotazione semantica, struttura questa che potrebbe essere impiegata nei casi in cui la parola significhi appunto «universo» — si distinguerà, sempre in absentia, da *mondo (persona pura) S.M.*

Il problema degli omografi appartenenti ad una stessa categoria grammaticale, ma di significato diverso, è dei più spinosi, dato che non è sempre facile fare previsioni al riguardo. Tanto più in un *corpus* come quello del TLIO, che comprende testi di varia origine dialettale. A questo fine si è costituito un «Elenco degli omografi della lingua delle origini appartenenti ad una stessa categoria grammaticale» (E.1.) costituito sino a tutt'oggi da circa 3.180 voci suddivise in 1.060 schede. Tale elenco è costantemente aggiornato ogni qual volta i collaboratori rilevano nuovi casi di omografia. Quando i singoli lemmi omografi comportano ciascuno più di un significato (polisemia), se ne è adottato convenzionalmente uno, senza con questo che la parola lemmatizzata abbia proprio quel significato. Così ad esempio per *ruga (grinza, strada)* dal lat. RUGA — distinto da *ruga (bruco)* dal lat. ERUCA — si è adottato il lemma *ruga (grinza) S.F.*, anche se la forma lemmatizzata significa «strada».

L' «Elenco» (E.1.) è completato da quello degli omografi per la cui distinzione è sufficiente la connotazione grammaticale (E.2.) e che assommano a circa 1.386 voci suddivise in 462 schede.

L'«Elenco generale degli omografi» (= E.1. + E.2.), in tutto circa 4.566 voci suddivise in 1.522 schede, ha sostanzialmente un'interesse pratico. Esso serve ad orientare i collaboratori nel caso che sia difficile distinguere fra polisemia ed omografia. Al riguardo si è deciso che tutto ciò che non rientra nell' «Elenco generale degli omografi» va trattato alla stregua delle voci polisemiche.

Le indicazioni che seguono riguardano nell'ordine:

1. la data dell'opera contenente la forma posta in esponente
2. gli estremi bibliografici dell'opera
3. la zona linguistica
4. la categoria e genere letterario
5. il riferimento topografico (pagina e riga dell'edizione a stampa da cui si è ricavata la forma) e in più, eventualmente, il riferimento organico (capitolo, paragrafo, libro, ecc.)
6. il numero della scheda-contesto
7. il contesto, secondo il taglio (I) previsto dall'Ufficio Lessicografico, oppure (II) stabilito automaticamente dal programma, per un massimo di 756 battute distribuite in 12 righe di 63 battute l'una.

Inutile sottolineare l'interesse di tale sistema di dati. Una volta che l'Archivio dell' «Opera del Vocabolario» sarà completo, l'utente potrà richiedere informazioni di vario genere sulla distribuzione cronologica, linguistica e letteraria del tesoro lessicografico italiano delle origini. Ad esempio l'elenco completo delle voci (lemmi) di una certa zona linguistica risalenti ad una certa epoca e per un certo genere letterario, e così via.

Dato che la lemmatizzazione, come è ovvio, comporta uno sfoltimento massiccio delle forme (si è calcolato un abbattimento medio del 65 per cento) non sarà possibile procedere a computi statistici sulla consistenza (frequenze medie, e così via) del lessico del TLIO nelle sue singole forme. Questo però sarà possibile per il *corpus* Avalle (CLPIO) comprendente circa 600.000 occorrenze ricavate dalla trascrizione in edizione critica di ***tutti*** i mss. databili per ragioni paleografiche, storiche, ecc., entro i limiti del Duecento. Dato che le 600.000 occorrenze verranno ***tutte*** lemmatizzate senza abbattimenti di sorta, è ovvio che per quel che riguarda questo settore avremo una raccolta completa di ***tutti*** i dati relativi alla lingua poetica italiana delle origini, basata sui documenti autentici dell'epoca. Lo stesso dicasi per l'altro *corpus*, elaborato da A. Castellani con gli stessi criteri, della prosa italiana delle origini (sino al 1275), che completerà in tal modo la documentazione relativa al periodo più antico.