LESSICO LATINO E ANALISI ELETTRONICA

NINO MARINONE
Università di Torino

L'uso del calcolatore elettronico per la redazione di indici e di concordanze ha avuto in questi ultimi anni diffusione notevole anche nell'ambito dei testi latini. Sono oggi disponibili indici di ogni genere: in ordine alfabetico diretto o inverso, indici di frequenza crescente o decrescente, indici distinti per settori o organizzati secondo qualsiasi altro criterio. Analogamente nelle concordanze le forme in ordine alfabetico accompagnate da un contesto di determinata estensione possono essere distinte per autore, opera, argomento, selezionate per tipo di linguaggio, segmentate per fasi diacroniche, con una molteplicità di risultatii che è in grado di sodisfare le più svariate esigenze del filologo[1]. Un servizio ancor più prezioso rende al filologo la ricerca di parole, che consente di reperire e di visualizzare (e quindi stampare immediatamente) le occorrenze di forme e sintagmi inseriti in contesti di qualsiasi estensione per un'opera, per un intero autore o per un gruppo di autori selezionabile a piacere[2].

Senonché por il latino (e ancor più per il greco) insorgono problemi di non facile soluzione. Infatti il calcolatore esige una formalizzazione implacabile e funziona su complessi rigidamente determinati. Il latino invece, anche solo considerando la lingua letteraria dalle origini al sec. V, presenta un'evoluzione diacronica che implica variazioni d'ogni genere non formalizzabili.

[1] Sia lecito recare ad esempio la concordanza delle opere dei grammatici latini raccolte in H. KEIL, *Grammatici Latini,* 7 vols., Leipzig 1857-1880 [rist. Hildesheim 1961], da me diretta e realizzata da Valeria Lomanto in collaborazione con l'Istituto di Linguistica Computazionale del Consiglio Nazionale delle Ricerche a Pisa: ammette la produzione di concordanze specifiche il cui totale teorico è rappresentato da un numero di 35 cifre. Cfr. N. MARINONE, V. LOMANTO, A. ZAMPOLLI, «Concordanza dei Grammatici Latini», *Suppl. al vol. 112 degli Atti dell'Accademia delle Scienze* II, Torino 1979; N. MARINONE, «Concordanze e indici dei grammatici latini tardoantichi e altomedievali», *La cultura in Italia fra tardo antico e alto medioevo*, Roma 1981, I, pp. 447-52; V. LOMANTO, «Grafia del tardo latino nell'elaborazione elettronica dei testi», *ibid.*, I, pp. 373-395; N. MARINONE, «A Concordance to Latin Grammarians», *Linguistica computazionale,* 1, 1981, pp. 127-129.

[2] A questo proposito non posso omettere di menzionare il monumentale *Thesaurus Linguae Graecae,* a cura di Th. Brunner (University of California, Irvine 1987-1988) registrato su CD-ROM contenente 60.123.000 parole, cioè l'intera letteratura greca da Omero al sec. VI. Se ne prevede l'estensione al greco bizantino fino al 1453, e la ricerca distinta per generi leterari e argomenti.

L'ordinamento stesso delle forme, che appare l'operazione meno complessa, non è scevro da problemi. Anzitutto la grafia. Il calcolatore considera autonoma qualsiasi unità graficamente diversa, anche se di valore identico. Ne consegue che vengono inesorabilmente distinti i casi di assimilazione e dissimiliazione, le parole che ammettono grafia unita e divisa, le varianti di scrittura (come ad es. la presenza o l'assenza di *h*) e così via. Quindi *adfero* è tutt'altro che *affero, quam ob rem* sono tre parole del tutto diverse da *quamobrem, exsisto* è nettamente distinto da *existo*.

Per ovviare a tale inconveniente è necessario raccogliere le varianti grafiche sotto un unico complesso di forme che viene assunto come termine di riferimento. Si procede cioè a quella che si può definire lemmatizzazione grafica, creando un lemma plurimo ma unitario detto «costellazione lemmatica», nel cui ambito nessuna forma è privilegiata. Ad es. *adfero* e *affero* diventano le due forme a cui il calcolatore riconduce unitariamente sia la forma con assimilazione sia quella senza. Lo studioso può partire indifferentemente dall'una o dall'altra ottenendo sempre lo stesso risultato di analisi, anzi può scegliere la grafia da privilegiare in relazione all'autore, al tipo di linguaggio, al periodo storico o a qualsiasi altra esigenza della propria ricerca.

Por la stessa ragione per cui risultano distinte le varianti grafiche avviene esattamente il contrario in campo morfologico per le omografie. Ad es. *uis* sostantivo è accomunato a *uis* forma verbale, l'ablativo *tristi* è tutt'uno con la forma sincopata del perfetto *triui,* e così pure per altri innumerevoli casi[3]. In sostanza è sempre riservata all'intelligenza umana la capacità di interpretare un testo in modo corretto.

Considerando il problema da una diversa angolazione, si affaccia la possibilità di introdurre nell'analisi una lemmatizzazione morfologica che agisca insieme con quella grafica. Vale a dire: il sistema riconduce ogni forma ad un lemma (o ad una costellazione lemmatica, nel senso delineato poc'anzi), che corrisponde alla voce registrata in un dizionario latino. Ma ecco subito un ulteriore problema: quale dizionario? Di norma si giudica un dizionario sia dalla quantità delle entrate in rapporto alla fascia evolutiva della lingua che esso contempla sia soprattutto dal metodo con cui sono compilate le singole voci distinguendo le accezioni e illustrandole con esempi idonei forniti di riferimenti esatti. Nel nostro caso, poiché il calcolatore non accede ai valori semantici, è di gran lunga prevalente la valutazione quantitativa del patrimonio lessicale. Quindi è opportuno assumere a fondamento la lista che contiene il maggior numero di vocaboli, e non il dizionario migliore nel senso tradizionale, poiché in realtà quello che occorre per l'analisi computazionale è un semplice lemmario. L'asserzione può sembrare paradossale, ma esprime in modo conciso la situazione.

Escludendo il *Thesaurus Linguae Latinae* perché non ancora finito, si ritiene che il *Lexicon Totius Latinitatis* di Egidio Forcellini[4] sia il più com-

[3] Tra i linguisti computazionali si parla di «disambiguare» le forme omografe estendendo l'analisi al contesto prossimo al termine in esame. Sono interventi di qualche efficacia per le lingue moderne, ma certo non risolutivi per il latino, il cui sistema è assai più articolato.

[4] Nella quarta edizione curata da F.Corradini e G. Perin, Padova 1864-1887 [rist. Bologna 1965].

pleto fra i dizionari latini, e come tale è stato adottato in varie elaborazioni computazionali. Senonché tale convinzione è fondata sul complesso dei 92.052 lemmi presenti nel Forcelllini, in cui sono incluse anche le 29.624 voci dell'onomastico, che per molteplici ragioni non servono al nostro scopo; inoltre i frequenti rinvii incrementano sì il numero dei lemmi ma non certo la quantità e la qualità dell'informazione, che è anche inficiata dalla scarsa attendibilità delle edizioni adottate. Da un preciso controllo il lemmario del Gradenwitz[5], che elenca 52.687 voci con esclusione totale dei nomi propri, offre una maggior ricchezza quantitativa; quindi nel nostro caso è senz'altro preferibile. Però, così facendo, il problema è solo impostato, non risolto.

Infatti il calcolatore ha quell'altra caratteristica (o pregio o difetto, secondo i punti di vista) cui accennavo prima: per produrre risultati validi richiede un'impostazione dei dati rigorosamente formalizzata. Ma nessun dizionario latino è così strutturato. Un studio comparativo di Valeria Lomanto[6] ha messo in rilievo le divergenze nell'impostazione dei lemmi tra il Forcellini, il Georges (con i supplementi del Gradenwitz) e il *Thesaurus*. La stessa parola subisce spesso trattamenti diversi: ad es. un participio con valore nominale è annoverato dall'uno come lemma autonomo, dall'altro come sottolemma del verbo, dall'altro ancora all'interno del verbo stesso; mentre un altro participio che ammette lo stesso valore è trattato in modo diverso. Ciò è giustificato da motivi che la competenza di ciascun autore della voce ha ritenuto validi; ma il calcolatore non li capisce.

La sola via d'uscita è la redazione di un nuovo repertorio lessicale che tenga conto di questa particolare esigenza. Esso è stato realizzato in questi ultimi anni nell'ambito dal progetto di ricerca «lessicografia latina» finanziato dal Ministero della Pubblica Istruzione e da me coordinato, a cui collaborano vari docenti delle università di Bologna, Firenze, Genova, Milano, Pisa e Torino. Adottando particolari criteri di catalogazione sono state compilate circa 57.000 schede lessicali utilizzando, oltre al Gradenwitz, il Georges[7] e il recente dizionario di Oxford[8]. La registrazione dei dati, eseguita su una griglia di 80 colonne, è fondata sulla segmentazione delle forme distribuita in sei campi utilizzati nel modo seguente.

1. **Riferimenti:** lettera iniziale dell'area lessicale cui appartiene il lemma, individuata mediante N(ominale), P(ronominale), V(erbale), I(nvariabile); seguita da cinque cifre per la numerazione progressiva all'interno di ciascuna area.

2. **Indicatori**

a) *per il lemma:* la lettera V segnala che il lemma deve essere inserito in una costellazione lemmatica, e viceversa Y impedisce tale inserimento;

b) *per segmenti iniziali:* lettere alfabetiche indicano la presenza di alterazioni nei prefissi variabili (ad es. i preverbi) allo scopo di ottenere la lemmatizzazione grafica, e analogamente cifre per i prefissi pronominali;

[5] O. GRADENWITZ, *Laterculi uocum Latinarum*, Leipzig 1904 [rist. Hildesheim 1966].

[6] V. LOMANTO, «Lessici latini e lessicografia automatica», *Memorie dell'Accademia delle Scienze* II, Torino 1980, V, 4, pp. 111-270.

[7] Nell'ultima edizione: K. E. e H. GEORGES, *Ausführliches Lateinisch-Deutsches Handwörterbuch*, Hannover - Leipzig 1913-1918 [rist. Basel 1951].

[8] *Oxford Latin Dictionary*, by P.G.W. Glare, Oxford 1982.

c) *per segmenti mediani:* il segno + produce automaticamente il tema del perfectum, del supino e dei participi perfetto e futuro nei verbi regolari, e viceversa il segno — impedisce la formazione di tutte le forme nominali in qualsiasi verbo;

d) *per segmenti postfinali:* segnalano la presenza dei suffissi pronominali, delle enclitiche e di qualsiasi altro elemento che si unisca alla forma già in sé completa.

3. **Segmento base:** rappresenta la parte invariabile del lemma nella flessione (che non necessariamente si identifica con la radice o con il tema o con il lessema).

4. **Segmento finale:** codici alfanumerici indicano il tipo o i tipi de flessione e gli elementi finali (che non necessariamente si identificano con la terminazione o la desinenza), a cui deve essere unito il segmento base con o senza l'intervento di segmenti mediani. Nell'ambito delle quattro aree principali sono distinte 20 specie di declinazioni con inclusione automatica di segmenti mediani per ottenere l'analisi di comparativi e superlativi nonché di segmenti iniziali e postfinali per l'analisi delle forme pronominali; 36 specie di coniugazioni dell'infectum con inclusione di segmenti mediani per ottenere l'analisi automatica del participio presente, del gerundio e del gerundivo, ed anche con la separazione dei tempi, dei modi e delle diatesi per evitare false analisi in presenza di forme anomale; 4 specie di coniugazioni del perfectum per ammettere anche l'analisi delle forme sincopate; infine gli avverbi nominali e pronominali collegati al lemma da cui derivano.

5. **Lemma:** è prodotto automaticamente (tranne che nelle formazioni abnormi) il lemma o la costellazione lemmatica a cui si ricollega la forma analizzata, aggiungendo per comparativi e superlativi la forma dell'aggettivo o dell'avverbio o del participio da cui derivano, per il perfectum e i nomi verbali la forma del presente a cui sono ascritti; analogamente per la compresenza di due o più alternative nella flessione. Infine la sigla CP [comune/proprio] segnala che il lemma è omografo di un nome proprio.

6. **Classificazione:** è indicata automaticamente la categoria morfologica a cui appartiene ciascun lemma prodotto dall'analisi.

Il sistema funziona mediante l'applicazione di una complessa procedura di spoglio, che è stata realizzata da Andrea Bozzi e Giueppe Cappelli presso l'Istituto di Linguistica Computazionale di Pisa[9]. I primi esperimenti di questa analisi morfologica automatica del latino hanno già dato esito positivo.

A questo punto ci troviamo nella situazione di un agricoltore che ha acquistato una moderna mietitrebbia: la macchina è in grado di fare in breve tempo la mietitura di un vasto raccolto; peccato che il contadino possieda soltanto un fazzoletto di terra. Voglio dire che manca il materiale da sottoporre alla lemmatizzazione automatica. Infatti non tutti i testi latini finora registrati su supporto magnetico sono stati redatti con criteri idonei a tale procedura. Per la maggior parte degli autori occorre procedere ex novo, impegnandosi a trascrivere tutte la edizioni critiche più affidabili.

[9] Cfr. A. Bozzi, «Progetto di organizzazione di un vasto repertorio lessicale automatico della lingua latina», *Maia* 32, 1982, pp. 167-172; N. Marinone, «A Project for a Latin Lexical Data Base», *Linguistica Computazionale* 3, 1983, pp. 175-178.

Subito si affaccia un'obiezione: il risultato merita così gravoso dispendio di energie e di danaro? tanto lavoro per sentirsi alla fine rispondere dal computer che *patrem* è una forma di *pater* della terza declinazione, che *inermis* può ricollegarsi sia ad *inermis* sia ad *inermus*, che *arator* oltre che sostantivo è anche imperativo futuro di *arare*, che *suaue* oltre che come neutro dell'aggettivo vale anche como *uel sua,* che *quinque* oltre che essere un numerale rappresenta anche l'analisi di *et quin* (che è un hapax plautino), e così via, come in un divertente giochetto elettronico.

Non è certo questo il traguardo finale. Penso piuttosto alla ricerca di parola. Attualmente l'interrogazione è circoscritta a una o più forme univoche; oppure, se chiedo il segmento inviariabile di una forma, la risposta è estesa a tutte le parole che contengono tale elemento. Chiarisco con una prova ipotetica. Imposto la ricerca sul settore desiderato (cioè un gruppo di autori o addirittura l'intero patrimonio letterario), digito sulla tastiera *rosarum:* leggerò sullo schermo tutte le occorrenze di tale genitivo, ciascuna inserita in un contesto e fornita dei riferimenti. Altro tipo di approccio: digito *ros*. Ottengo tutte le forme inizianti con questo segmento, ma non solo quelle attinenti al sostantivo *rosa:* compariranno anche le forme di *ros, rosaceus, rosarius, rosatus, roscidus* e *rosidus, rosetum, roseus, rosio, rosor, rosmarinum, rostellum, rostralis, rostrans, rostratus, rostrum, rosulentus,* il perfetto *rosi* e infine *rosus* sia participio sia sostantivo. Un risultato ineccepibile ma praticamente inutilizzabile se il campo dell'indagine è di estensione notevole, come capita in genere quando si ricorre al calcolatore; giacché per l'analisi di uno o due autori è più comodo far ricorso ai vari lessici speciali già esistenti. Inutile aggiungere quanto si complichi la faccenda nel caso di un verbo irregolare. In sostanza il sistema è pienamente efficiente solo quando si tratta di termini raramente usati; e questo è un inconveniente serio, a cui peraltro credo sia possibile porre rimedio, applicando alla ricerca di parola il sistema di lemmatizzazione automatica che abbiamo realizzato.

Si forma un lemmario che contiene l'elenco delle parole latine disposte in ordine alfabetico, ciascuna accompagnata dal codice alfanumerico che compare come riferimento già registrato per la lemmatizzazione. Si procede esattamente come in un elenco telefonico. Per ottenere una voce si compone il prefisso selettivo, che è costituito dalla lettera iniziale dell'area lessicale (cioè N per nominale, V per verbale, e così via). Si fa seguire il numero assegnato alla voce desiderata, che è di sole cinque cifre. Quindi, nel caso che si voglia limitare la ricerca (ad es. per un verbo alle sole forme del perfectum), si digita l'indicatore della sezione specifica. Ecco un esempio. Supposto che il verbo *fero* abbia il numero 13471, digito sulla tastiera V13471; otterrò tutte le occorrenze del verbo negli autori che intendo esaminare; se faccio seguire P avrò soltanto le forme del perfectum, se IP quelle dell'infectum e del perfectum insieme ma non quelle nominali, e così via. Analogamente per un sostantivo a duplice declinazione, como ad es. *barbaria/-es:* facendo seguire 1 al suo numero otterrò solo le forme della prima declinazione, facendo seguire 5 solo quelle della quinta declinazione.

Si obietta: perché non digitare direttamente il lemma desiderato? In realtà questa, che può sembrare una semplificazione in quanto elimina la necessità di consultare il lemmario, si traduce in una difficoltà per l'utente,

che non conosce i criteri con cui sono state registrate le voci latine nella ricerca della maggior formalizzazione possibile.

Questo progetto è in gran parte realizzato. Restano solo particolari settori da sistemare in modo definitivo per giungere ad una piena acribia filologica; ma soprattutto sono necessari testi latini registrati su supporto magnetico in modo adeguato al sistema di lemmatizzazione.

Non sussistono ostacoli tecnici. Piuttosto una remora notevole è costituita dai filologi classici, perché nei confronti della ricerca tecnologica rappresentano un gruppo di utenti molto scettico, inoltre scarso di numero e ancor più scarso di disponibilità finanziarie.