

ANTONIO ZAMPOLLI

L'elaboratore elettronico negli studi linguistici

L'elaboratore elettronico negli studi linguistici

Antonio Zampolli

Dall'inizio degli anni '50 la stampa ha spesso commentato in modo suggestivo l'accostamento elettronica-linguistica. Allora come oggi l'adozione di mezzi d'avanguardia non giustifica preoccupazione o esaltazione alcuna per l'incontro tecnica-umanesimo: il calcolatore si è rivelato ancora una volta un potente strumento al servizio del lavoro umano.

La statistica linguistica ha una storia ben più antica dell'avvento dei calcolatori. Nell'antichità i grammatici alessandrini contavano gli *hapax*, cioè le forme che appaiono una sola volta nel corso di un'opera. Nel Medioevo, vi fu una corrente di erudizione che promosse le concordanze della Bibbia, repertorio di tutte le parole dei testi biblici, in tutti i loro impieghi e nel loro contesto. Gruppi di monaci erano allora incaricati di fare quello che fanno oggi i calcolatori del CNUCE di Pisa. Molto citata, benché probabilmente apocrifa, la notizia che il cardinale Hugh di Saint-Cher († 1263) impiegò non meno di 500 monaci per compilare la prima autentica concordanza della Bibbia. Nel XVI e XVII secolo gli indici di certe edizioni latine avrebbero soddisfatto i nostri statistici più esigenti. Edizioni bellissime, dal punto di vista tipografico, come quelle francesi, ad *usum delphini*, elencavano, infine, tutti i luoghi, anche delle parole più frequenti, come la *et*. Quale fosse la fatica e l'animo dei pazienti compilatori non è difficile immaginare: ancora in epoca recente, al completamento del 1° volume del suo *Lexicum Plautinum* (Lipsia, 1924) Gonzales Lodge scrive: *Plus triginta anni celeriter praeterierunt postquam, adulescentia imperitiaque deceptus, consilium cepi lexici Plautini conficiendi* (« Più di trenta anni sono passati da quando, ingannato dall'inesperienza e dal giovanile entusiasmo, decisi di compilare un lessico di Plauto »). E dopo 9 anni, nel secondo volume, si congedava così: *Nunc Plaute, amice mi et inimice per totam fere vitam meam, vale* (« Ora Plauto, amico e nemico mio per quasi tutta la vita, addio »).

Il metodo di spoglio manuale di un testo consiste nel trascriverne su schede, una per una, tutte le parole

con il relativo contesto, e cioè con la frase cui appartengono; le schede vengono poi disposte a mano in ordine alfabetico. È chiaro che, se per ogni parola si cita, in media, una frase di 10 parole, il lavoro completo consiste nel trascrivere 10 volte l'intero testo.

L'annuncio, dato nel 1951, che impianti meccanografici venivano usati per l'indice della *Summa Theologiae* di S. Tomaso d'Aquino, creò una situazione completamente nuova nel campo della lessicografia. È a Padre Roberto Busa S.J. della Facoltà di filosofia dell'Aloisianum di Gallarate che viene universalmente riconosciuto il merito di avere per primo al mondo applicato impianti elettronici allo spoglio di testi. Ben presto seguirono altri progetti e si fondarono centri specializzati, dai nomi diversi e significativi: un elettrizzante spirito pionieristico anima le comunicazioni degli studiosi che, negli anni '50, danno notizia di questi primi esperimenti di applicazione dei calcolatori elettronici all'analisi linguistica. L'accostamento e la collaborazione dell'elettronica e della linguistica colpivano naturalmente l'immaginazione, tanto più che si inserivano in un contesto sociale e culturale impegnato a risolvere il temuto dualismo di tecnica e umanesimo.

Nell'« Almanacco letterario Bompiani » del 1962, alcuni illustri linguisti e pensatori italiani erano chiamati a commentare in questa prospettiva il binomio « linguistica-calcolatore elettronico ». Numerose iniziative moltiplicatesi e maturate negli anni '60 hanno verificato puntualmente le previsioni e i voti degli studiosi intervistati. Si potrebbe dire, parafrasando uno slogan fortunato, che le applicazioni linguistiche dei calcolatori sono già entrate nella seconda decade di

vita e nella seconda generazione. Il fenomeno è stato paragonato alla invenzione della stampa del XIV secolo: ma né allora né ora l'introduzione di tecniche d'avanguardia giustifica preoccupazione o esaltazione per la simbiosi o lo scontro della tecnica e dell'umanesimo. La spiegazione profonda dell'aiuto che il calcolatore può dare al linguista, con la sua capacità di trattare velocemente una imponente quantità di simboli, è insita nella natura stessa del linguaggio, come è concepito dalle moderne teorie della linguistica generale.

Le unità o segni linguistici di una lingua, concepita come sistema strutturato che assicura la comprensione nelle comunicazioni interpersonali, non sono raccolti né catalogati naturalmente in alcun luogo, ma sono per così dire depositati, in misura diversa, nella coscienza di tutti gli individui che parlano la stessa lingua, i quali li realizzano e li ripetono ogniqualvolta parlano o scrivono. Per studiare la lingua è necessario analizzare, segmentare e confrontare i diversi « testi » reperibili, parlati o scritti, dalla cui comparazione risulta da un lato ciò che è *normale* realizzazione delle possibilità del sistema, dall'altro ciò che invece è realizzazione personale e individuale, dunque fatto di *stile* e, in certe condizioni, di *poesia*. Forse la risposta migliore alla domanda « *a che cosa serve il calcolatore per un linguista?* », la forniscono esempi di applicazioni concrete scelte tra i numerosi progetti che istituti specializzati, con sede in diverse città d'Ita-

Frontespizio della prima edizione del Vocabolario della Crusca. L'opera, stampata nel 1612 e contenente 60.000 citazioni, costituì il modello dei grandi dizionari europei successivi.

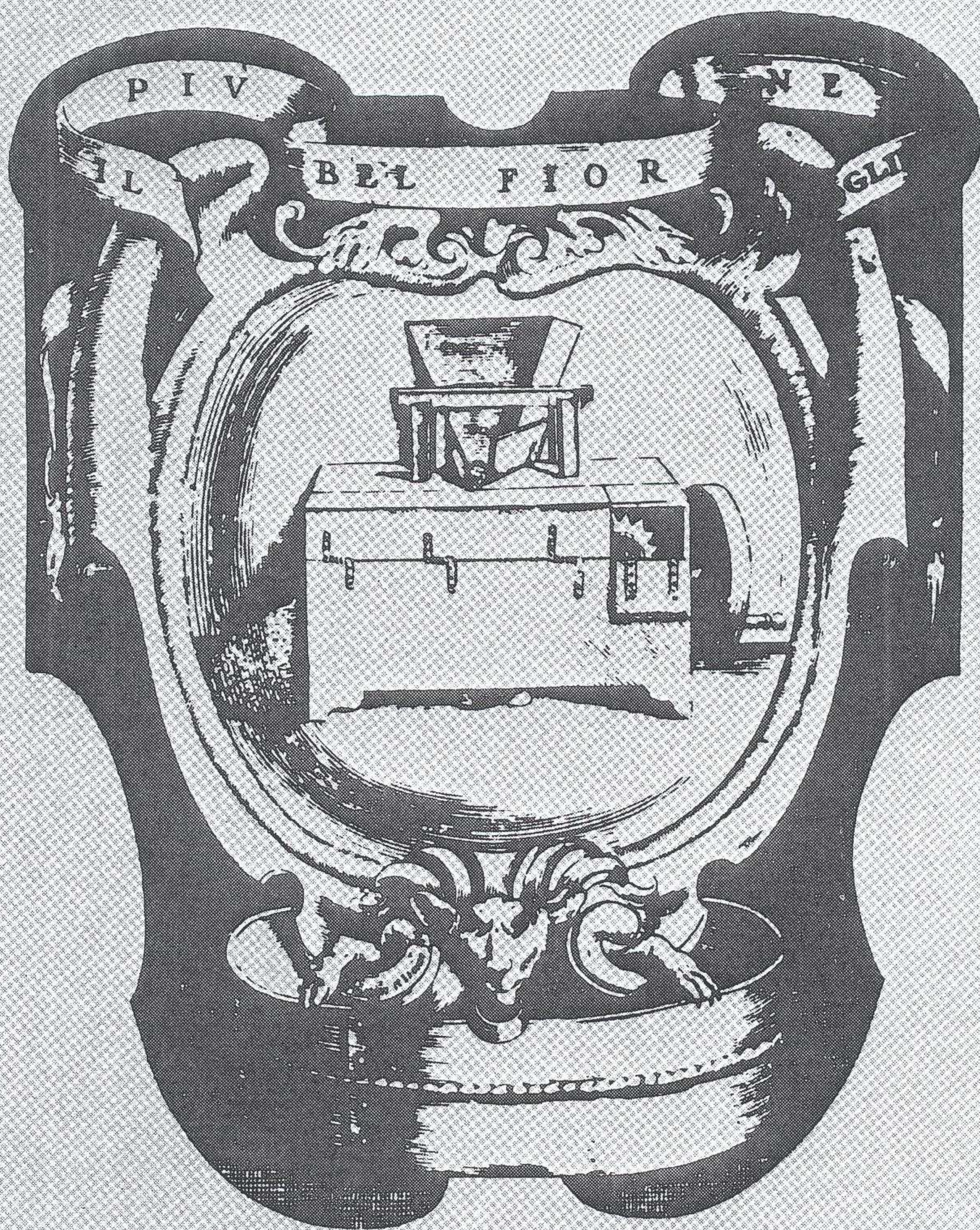
VOCABOLARIO DEGLI ACCADEMICI DELLA

CRVSCA, CON TRE INDICI DELLE VOCI.

locuzioni, e prouerbi Latini, e Greci, posti per entro l'Opera.

CON PRIVILEGIO DEL SOMMO PONTEFICE,
Del Re Cattolico, della Serenissima Repubblica di Venezia, e degli
altri Principi, e Potentati d'Italia,

E FAVOR D'ITALIA, DELLA MAESTA CESAREA,
Del Re Cristianissimo, e del Sereniss. Arciduca Alberto.



IN VENEZIA MDCXII

Appresso Giouanni Alberti

lia, conducono in collaborazione con il CNUCE di Pisa, utilizzando gli elaboratori elettronici IBM 1401 e IBM 7090.

METODI DI LAVORO

Prima di esaminarli a uno a uno, è bene descrivere brevemente il metodo di lavoro e i principali risultati che il calcolatore fornisce nel corso della elaborazione dei testi che gli vengono affidati. Come si vedrà, i progetti sono diversi, e per la natura dei testi analizzati e per lo scopo della ricerca. Tuttavia hanno in comune tutta la prima fase dell'elaborazione, nella quale si compie l'analisi lessicale, fino alla raccolta delle frequenze, delle concordanze e dei diversi tipi di indici nei quali vengono catalogate le parole sottoposte a indagine. Definiamo alcuni semplici termini: un testo è costituito di *parole*: per esempio ⁽¹⁾ nella terzina dantesca « Per me si va nella città dolente, / Per me si va ne l'eterno dolore, / Per me si va tra la perduta gente, » le *parole* sono 24 (8 per ciascun verso). Ma alcune di esse sono uguali: se non teniamo conto delle ripetizioni, troviamo che le *forme* o *parole diverse* della terzina sono in tutto 14 (per, me, si, va, ne, la, città, dolente, l', eterno, dolore, tra, perduta, gente): il numero di volte che ciascuna *forma* è ripetuta viene detto *frequenza assoluta* della forma nel testo. Pensiamo ora di dover cercare le forme in un qualsiasi vocabolario: non cercheremo certo, ad esempio, *va*, ma *andare*; e non troveremo *perduta*, ma solo *perdere* o *perduto*. Chiamiamo *lemmi* le forme di base *andare*, *perdere*, e cioè, in generale, le *parole-titolo* che troveremmo registrate, normalmente in neretto, nel vocabolario. Le forme ed i lemmi di un testo possono essere ordinati secondo diversi criteri: per esempio, in ordine alfabetico; o dal più frequente al meno frequente; oppure in *ordine alfabetico inverso*, per lo studio delle desinenze e delle rime ecc.

Se, dopo aver ordinato alfabeticamente i lemmi o le forme, riportiamo per ciascuno di essi tutte le frasi del libro in cui sono usati, con anche l'indicazione della pagina e della riga ove stanno scritti, otteniamo la lista delle *concordanze* (così chiamate perché raggruppano assieme tutte le frasi che *concordano* tra loro nel contenere una medesima parola). Se poi chiediamo al calcolatore di ampliare i brevi contesti delle concordanze in contesti più lunghi e di stamparli non più su fogli ma su schede meccanografiche, uno per scheda, otteniamo un risultato importantissimo del-

la documentazione lessicografica e cioè la *scheda-contesto*.

Il calcolatore « legge » il testo perforato su schede, lo registra su nastro magnetico, scomponendolo nelle unità elementari della elaborazione: parole, oppure grafemi, fonemi, sillabe, sintagmi ecc. A questo punto ha inizio il ciclo delle elaborazioni ad altissima velocità: per perforare mezzo milione di parole (circa 60.000 righe) occorrono quasi 350 ore di una esperta perforatrice; per conteggiare le frequenze con il calcolatore bastano 2 ore. La sproporzione tra velocità dell'elaboratore e velocità umana, nonché la possibilità di errori, anche di semplice disattenzione, hanno convinto i ricercatori a rivolgere i propri sforzi verso una maggiore automazione, per mezzo del cosiddetto « dizionario di macchina ». In altre parole, si chiede al calcolatore di fare come lo scolaro che non conosce una parola: egli la cerca in un vocabolario, dove ne troverà la descrizione.

Il calcolatore evidentemente non può lemmatizzare le parole non previste nel vocabolario e soprattutto non sa distinguere tra le parole omografe; il calcolatore non è in grado di decidere, per esempio, se la parola *danno* in un certo contesto sia il sostantivo maschile sinonimo di *rovina*, o non piuttosto forma del verbo *dare*, o addirittura il presente di *dannare*; né può distinguere fra *filare* di alberi e il verbo *filare*: tra gli aggettivi *militare*, *popolare*, *salutare* e i verbi omografi. Gli esempi per ogni lingua sono, sia pure in misura diversa, purtroppo molto numerosi.

Dico *purtroppo*, perché almeno fino a che non si troveranno degli algoritmi sicuri per operare lo scioglimento della omografia, l'intervento dello studioso costituisce un collo di bottiglia che rallenta inevitabilmente la serie automatica delle elaborazioni. Il problema dell'omografia riflette una caratteristica di alcuni punti del sistema linguistico — l'ambiguità — che sembra contrastare con la sua funzione comunicativa. Appare probabile, addirittura vicino, il momento in cui sarà possibile togliere automaticamente l'ambiguità tra parole omografe che appartengono a categorie grammaticali diverse (pur su un altro piano, è quanto fanno, almeno parzialmente, alcuni programmi di traduzione automatica dal russo all'inglese, come quello in funzione al CETIS - Euratom a Ispra). In questa direzione sono particolarmente utili e stimolanti le ricerche di linguistica « algebrica » o « algoritmica », come la chiama G. Lepschy, della quale la grammatica trasformazionale è forse l'esempio più noto; questo tipo di approccio ai fenomeni linguistici è particolarmente adatto a formulare algoritmi di analisi grammaticale e sintattica eseguibili dai calcolatori. Ma dove l'ambiguità esiste solo sul piano semantico, qualsiasi previsione ottimistica pare azzardata.

PROGETTI DI ANALISI LINGUISTICA IN CORSO AL CNUCE DI PISA

L'Accademia della Crusca ha iniziato la compilazione di un ampio *Tesoro delle origini della lingua italiana* e di un non meno ampio *vocabolario storico della lingua* ⁽²⁾.

Il carattere scientifico di un vocabolario storico, basato su una larga documentazione lessicografica, si riflette nel gran numero di esempi tratti da opere letterarie, riportati per ogni parola a testimonianza della sua evoluzione e a chiarimento dei suoi significati.

La prima edizione della Crusca, stampata in un solo volume nel 1612, conteneva 60.000 citazioni e costituì il modello dei grandi dizionari europei successivi. La quinta impressione, iniziata nel 1863 si arrestò incompiuta, nel 1923, alla lettera O: contiene, in 11 volumi, circa 340.000 esempi.

Il progetto del nuovo grande vocabolario storico iniziato nel 1964 prevede la raccolta, sotto la guida del prof. Aldo Duro, di 100 milioni di schede, che dovrà essere eseguita in poco più di 10 anni, se si vogliono salvaguardare la continuità e la coerenza del metodo. Solo il calcolatore può setacciare parola per parola le opere più importanti e lessicalmente più ricche della nostra letteratura, in un tempo così breve.

Il risultato finale della elaborazione delle migliaia di testi sottoposti allo spoglio elettronico, è costituito dalle schede-contesto, cioè, come si è detto, da schede che, per ogni parola, riportano stampato un contesto sufficientemente lungo da utilizzare come citazione, come esempio per quella parola.

La documentazione lessicografica a disposizione del futuro redattore del vocabolario sarà la più ricca mai raccolta, e la consultazione sarà agevolata dai numerosi dati intermedi — lemmari, indici di frequenze, concordanze — che riassumono il materiale distribuito e ne costituiscono un prezioso indice statistico. Del resto, accanto all'archivio stampato su schede, rimarrà operabile l'archivio registrato su nastro magnetico per indagini settoriali supplementari ed approfondite.

Non dissimili, anche se più limitati per quanto riguarda sia la mole delle opere da analizzare sia il tipo di lessico da estrarre — giacché si tratta sempre di terminologie settoriali — sono gli scopi che si prefiggono altre imprese, tutte finanziate dal Consiglio Nazionale delle Ricerche, che hanno chiesto accesso ai complessi elettronici del CNUCE, con la consulenza tecnico-scientifica dell'Accademia della Crusca. Si tratta, in primo luogo, del *Vocabolario giuridico ita-*

(1) Per questo esempio e per la descrizione del lavoro viene seguita da vicino la relazione presentata dal prof. Aldo Duro e dal dott. A. Zampolli all'Accademia dei Lincei, in occasione del Convegno *L'automazione e le sue implicazioni scientifiche e sociali*, tenuto a Roma nei giorni 16-19 ottobre '67.

(2) Una rassegna completa dei lavori attualmente in corso presso il CNUCE di Pisa andrebbe troppo oltre lo spazio concessoci; la scelta dei temi o lo spazio ad essi dedicato dipende solo da comodità di esemplificazione.

liano che, con sede a Firenze e sotto la direzione del prof. Piero Fiorelli, attende alla raccolta del materiale documentario, estratto da opere comunque di interesse giuridico o storico-giuridico, per la compilazione di un *Vocabolario giuridico italiano* su basi storiche: codici, leggi, statuti, giurisprudenza, dottrina ecc.

Su opere interessanti la storia del pensiero, della scienza, della cultura, in quanto abbiano attinenza con i suoi obiettivi istituzionali, compie invece analisi, il *Lessico intellettuale europeo*, centro di ricerca dell'Istituto di filosofia dell'Università di Roma.

Queste analisi elettroniche sono interessanti anche sotto l'aspetto di esperimento economico, in quanto il lessico che nelle opere analizzate dovrà essere messo in evidenza e trasferito su schede è un lessico nettamente

specializzato, per cui il rapporto percentuale fra le schede dei risultati finali e le schede contenenti in perforazione la totalità delle parole di ciascuna opera analizzata è bassissimo.

Al contrario, si ha quasi totale coincidenza fra occorrenze e contesti utili nelle concordanze delle opere minori in volgare di Dante Alighieri (*Rime*, *Vita nova*, *Convivio*), portate già quasi a termine, che l'Istituto dell'Enciclopedia Italiana, ricorrendo alla competenza del prof. Aldo Duro per quanto concerne l'avvio delle varie fasi di lavoro e la lemmatizzazione, ha voluto ottenere elettronicamente ai fini di potersene valere per la compilazione dell'*Enciclopedia Dantesca* cui attende una particolare sezione dell'Istituto. Mentre, infatti, per il sommo pcema si dispone già di concordanze integrali e sicure (l'ultima in ordine di tempo è stata realizzata proprio dalla IBM Italia presso il CNUCE), mancavano ccncordanze aggiornate e integrali per le altre opere italiane di

Dante. La lacuna è stata così colmata, anche se per il momento le concordanze ottenute non sono state stampate e pubblicate, ma soltanto listate per uso interno dai redattori dell'*Enciclopedia Dantesca*.

Carattere di ricerca lessicografica riveste anche un lavoro sulla lingua indoeuropea più anticamente attestata, la lingua degli ittiti, decifrata e riconosciuta indoeuropea solo nel 1917, che conserva nella grammatica preziosi relitti arcaici. L'équipe del prof. Pietro Meriggi, del Centro Studi Egei del CNR presso l'Università di Pavia, si propone di pubblicarne un dizionario basato sui testi storici, utilizzando come documenti di base le schede-contesto prodotte dal calcolatore. Lo studioso deve innanzitutto trascrivere il testo originale con un sistema di codici comprensibili alla ragazza perforatrice, la quale non può certo decifrare ideogrammi e caratteri cuneiformi. La possibilità di usare il testo, una volta che sia stato registrato su schede, per tante elaborazioni quante se ne desiderano, ripaga lo sforzo richiesto allo studioso per operare la trascrizione. Mi sembra, questo delle lingue di difficile decrittazione, un chiaro esempio di un fatto difficilmente contestabile: che cioè la difficoltà di *pre-edizione dei testi* (ossia di preparazione alla perforazione) non deve essere un elemento per rinunciare alla elaborazione automatica, ma anzi un motivo di più per ricorrervi.

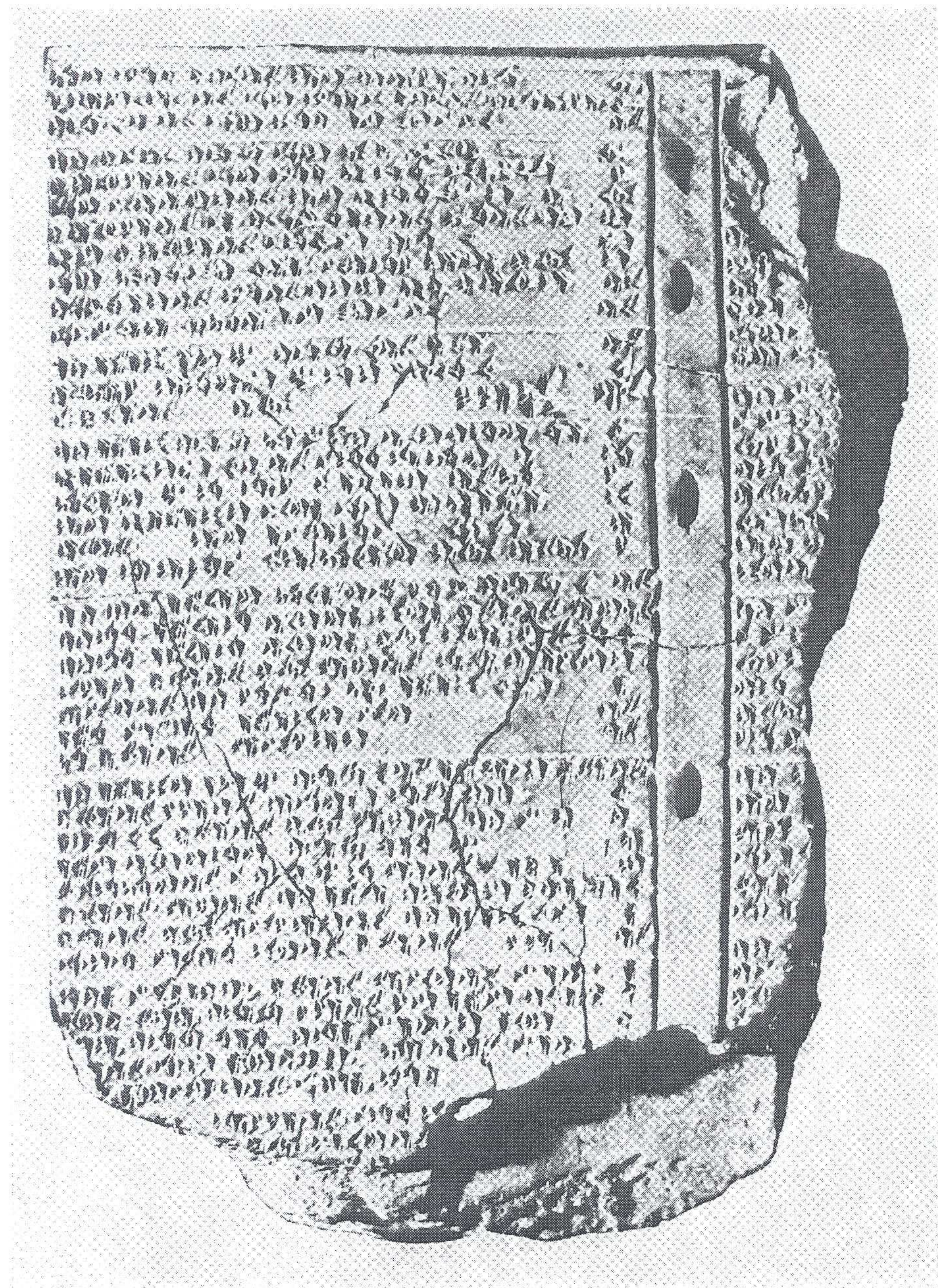
* * *

Di estrema importanza sono anche le elaborazioni, in corso a Pisa, di testi di singoli autori, per studiarne l'evoluzione dello stile e del pensiero. I professori Pietro Ferrarino e Alfonso Traina, dell'Università di Padova, curano le concordanze dell'opera omnia di L. Anneo Seneca; il gruppo del prof. F. M. Sciacca, dell'Università di Genova, ha iniziato l'analisi delle opere di A. Rosmini per la compilazione di un lessico rosminiano.

La dottoressa Carla Colombo ha avviato per conto del prof. Carlo Bo, dell'Università Bocconi di Milano, la concordanza delle opere di André Gide, iniziando dal *Journal*. Questo diario è la testimonianza quotidiana dell'evoluzione spirituale che portò Gide, attraverso disparate esperienze di vita, a vivere atteggiamenti esistenziali e convinzioni filosofiche di grande diversità. Per questo il lessico, particolarmente quello attinente in qualche modo al pensiero religioso, verrà studiato soprattutto chiedendo al calcolatore di raggrupparlo per *tranches* cronologiche, e le stesse concordanze recheranno accanto, anziché la pagina e la linea come di consueto, l'anno e la data: il critico cercherà così di penetrare i diversi successivi significati attribuiti alle parole-« chiave » dell'autore.

I procedimenti elettronici si possono applicare con successo a una ricerca di tipo nuovo, che riguarda la *morfo-*

Iscrizioni cuneiformi in lingua ittita. Questa antichissima lingua indoeuropea è stata decifrata solo nel 1917.



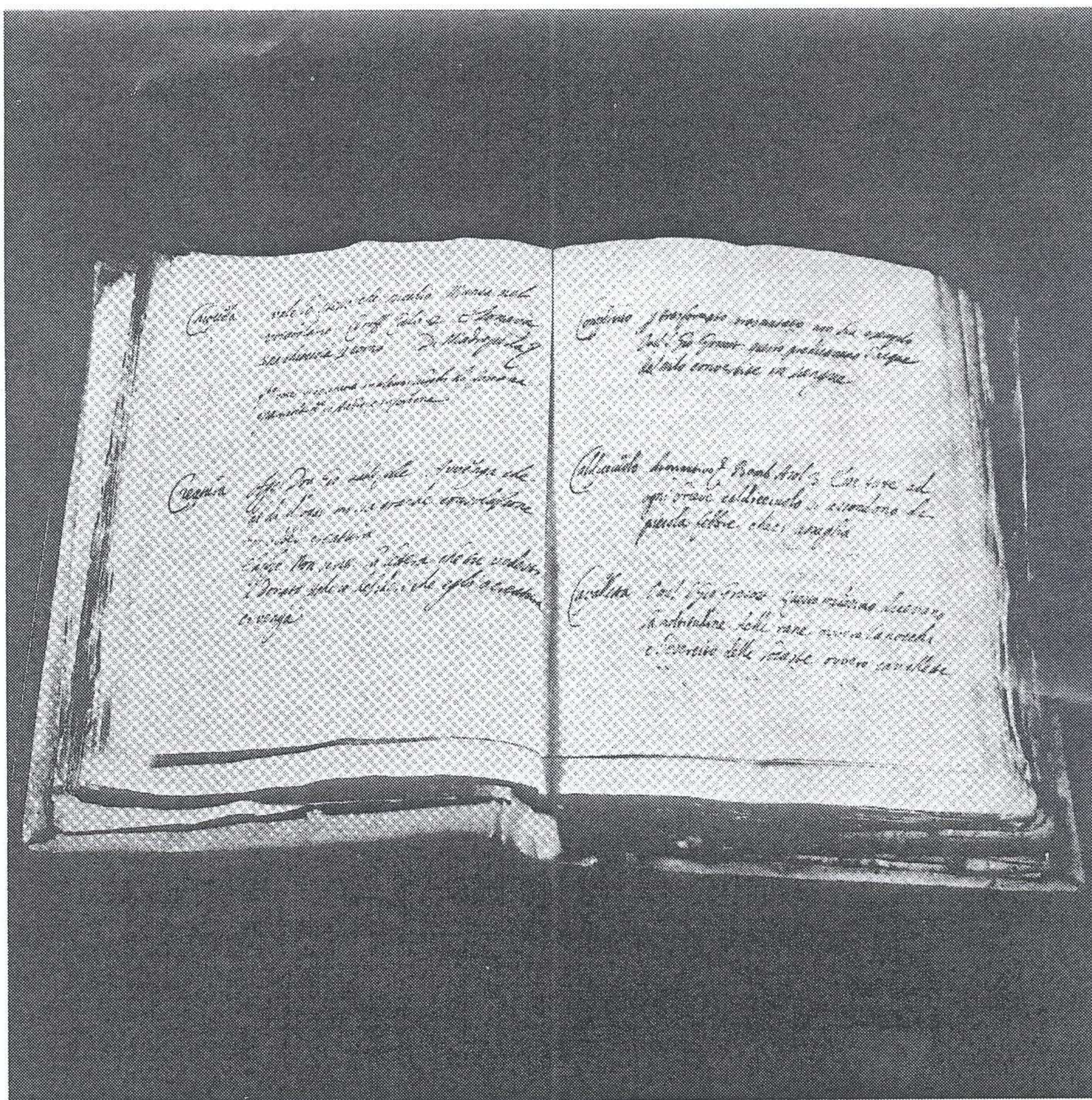
logia strutturale della poesia corale greca antica (Pindaro, Bacchilide e gli altri poeti corali tramandati frammentariamente). Questa ricerca che conta sulla inesorabilità del calcolatore per rendere obiettiva in massimo grado l'analisi degli elementi che caratterizzano formalmente la struttura poetica dei testi, è condotta dal prof. Carlo O. Pavese ed è sostenuta dal Centro di Studi per la lirica greca dell'Università di Urbino.

* * *

Anche nel campo della poesia di tradizione orale si è eseguito, per il Gruppo italiano di Studi Demologici e il Comitato per la raccolta Barbi (cui collaborano, tra gli altri, il prof. Vittorio Santoli, il prof. A. Mario Cirese, Paola Raichich e Salvatore Barone) un esperimento che ha dimostrato la possibilità e la convenienza di uso degli elaboratori. Gli ottimi risultati conseguiti hanno dato il via a uno spoglio sistematico delle tradizioni popolari italiane, iniziando dalle numerose raccolte di canti popolari. Accanto alla poesia d'arte esiste una larga varietà di testi, tramandati per lo più oralmente e in qualche caso appoggiati dalle stampe dei girovaghi e cantastorie, che costituiscono la parte più importante, e a volte unica, della cultura e dell'espressione poetica degli strati popolari di una nazione.

Sono proverbi, sconsigli, formule magiche, giochi, indovinelli, massime, fiabe, canzoni che hanno una propria storia, a volte secolare, e un proprio modo particolare di tradizione e di diffusione. Delle numerosissime lezioni la critica non fa opera di « eliminazione » ma conserva anche le varianti: non si cerca cioè la ricostruzione di un archetipo, ma la rappresentazione più estesa possibile della tradizione di un componimento. Si può dire che l'unità non è rappresentata dal singolo componimento; i versi o gruppi di versi che lo costituiscono hanno ciascuno una vita più o meno indipendente nello spazio e nel tempo, così che versi originariamente appartenenti a un canto sono inseriti in canti differenti tra loro e di regioni lontanissime. Il riscontro delle analogie, che richiedeva fino a oggi una lunga schedatura manuale ed era, in pratica, affidato alla memoria dello studioso, è enormemente facilitato dalle concordanze. Spingendo oltre l'automazione delle analisi, il calcolatore potrà segnalare accoppiamenti tipici e frequenti di parole, aiutando in questo modo a rivelare la presenza di formule, moduli e schemi ritmici, che costituiscono il canovaccio sul quale si costruiscono le improvvisazioni, rielaborazioni e innovazioni individuali.

Un altro problema, già parzialmente affrontato con il calcolatore, consiste nell'esaminare le parole in posizione di rima per integrare la descrizione tradizionale degli schemi metrici popolari: accanto alle forme « canoni-



che » della rima, della consonanza e della assonanza, si intuiscono un tessuto più sottile e una varietà di rispondenze foniche, accettate dalla coscienza popolare, che sono ancora da enucleare e inventariare.

È agli inizi, sempre per il Gruppo italiano di Studi Demologici, la costituzione di un immenso schedario in vista della compilazione di una enciclopedia delle tradizioni popolari, con materiale che verrà estratto da precedenti enciclopedie e da studi monografici, combinando assieme le tecniche della schedatura lessicografica e della documentazione automatica.

* * *

Le ricerche statistiche sulla lingua italiana contemporanea dirette dal prof. Carlo Tagliavini, dell'Università di Padova, ci conducono invece al campo della fonetica⁽³⁾, disciplina particolarmente adatta all'applicazione di tecniche statistiche basate sui calcolatori. Nel 1948 il celebre linguista francese Marcel Cohen riponeva grandi speranze per il progredire di questi studi nell'uso di « machines à calculer »; oggi le principali scuole linguistiche ne raccomandano l'impiego per integrare con dati quantitativi la conoscenza del « rendimento funzionale », cioè del valore significativo di ciascun fonema.

G. K. Zipf, il celebre comportamentista americano, ha formulato, sulla

Gli esempi che dovevano servire alla compilazione o al rifacimento di un Vocabolario della Crusca venivano talora ricopiati l'uno di seguito all'altro, in ordine approssimativamente alfabetico, su grandi quaderni.

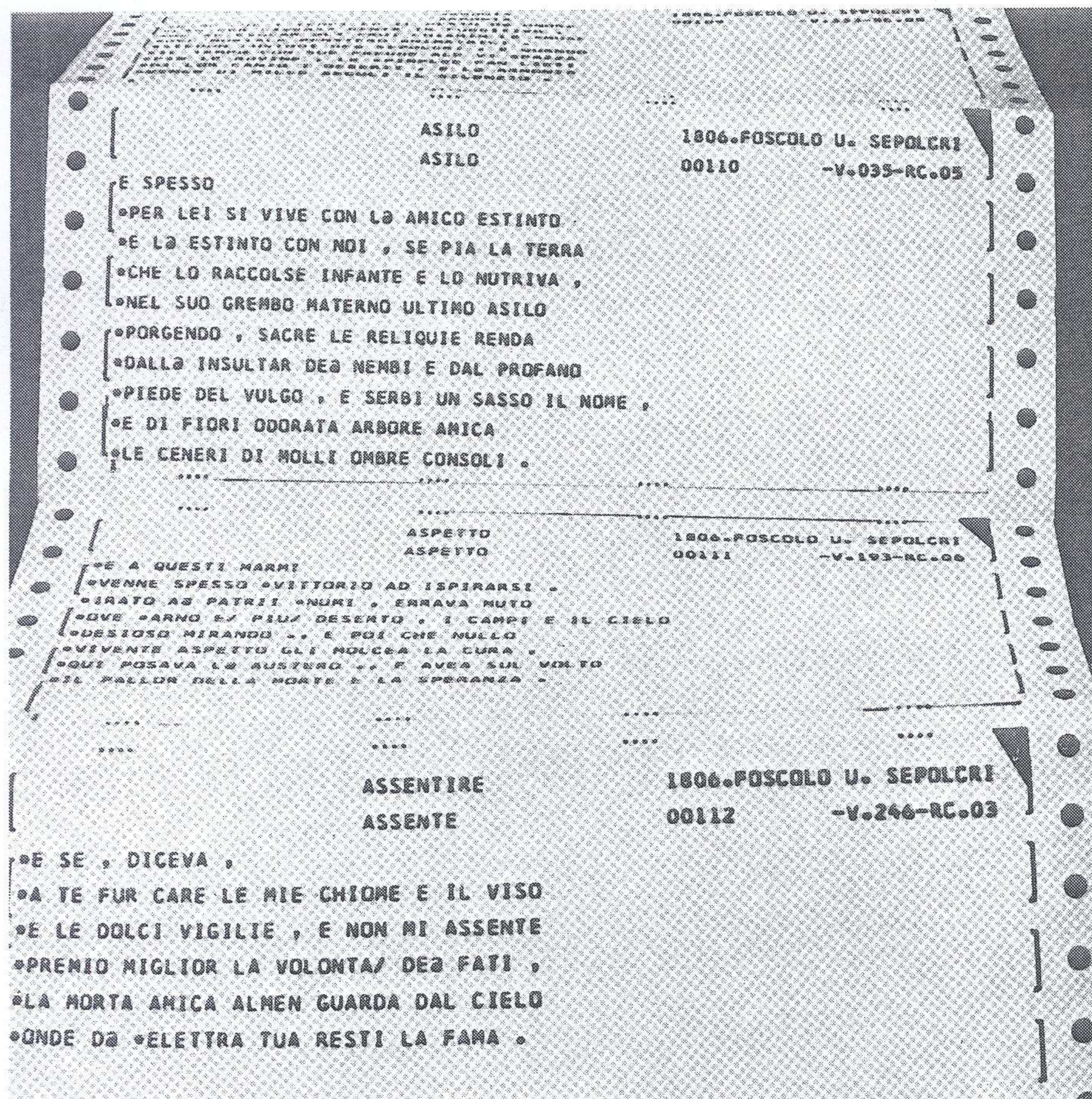
base di un ampio panorama delle frequenze dei fonemi nelle più disparate lingue del nostro pianeta, una suggestiva ipotesi sul meccanismo dell'evoluzione storica delle lingue, per la quale, ad esempio, il latino si è differenziato nelle lingue romanze contemporanee (italiano, ladino, sardo, francese, romeno, spagnolo, portoghese, catalano ecc.)

Conoscere la frequenza dei fonemi è utile anche ad altre discipline, per esempio alla scienza delle telecomunicazioni che deve adattare i mezzi tecnici di trasmissione alle caratteristiche della voce umana, o nello studio delle malattie dell'udito.

Le capacità di inserimento sociale di un paziente che soffre di ipoacusia, può essere valutata meglio con un test di percezione uditiva formulato con combinazioni di suoni o parole che rispecchiano la composizione statistica del linguaggio comune, piuttosto che con materiale acustico scelto a caso.

Il calcolatore, nel rilevare ed elaborare le frequenze dei fonemi, esplica in massimo grado le sue capacità statistiche. Purtroppo l'ortografia italiana è ben lontana dal rappresentare fedelmente la pronuncia reale; si pensi per esempio alla *h* di parole come *ho*, *hanno* ecc. cui non corrisponde alcun suono, o al trigramma

(3) Non è qui possibile per semplicità di esposizione adoperare coerentemente la terminologia scientifica che distingue fonetica, fonemica, fonologia.



Dall'elaborazione elettronica hanno origine le schede-contesto che riproducono la parola analizzata, la sua forma lemmatizzata e, per esteso, il contesto in cui la parola si trova.

sci che nella parola *sciabola* rappresenta un unico fonema.

Prima di chiedere al calcolatore di scatenare ad altissima velocità la sua abilità statistica, occorre trascrivere il testo nell'alfabeto fonemico, nel quale ad ogni fonema corrisponde un solo simbolo e viceversa. La trascrizione lunga e faticosa è fonte di innumerevoli errori, come ho potuto sperimentare personalmente quando, ai tempi della mia tesi di laurea sulla fonetica dell'italiano, non avevo ancora deciso di ricorrere all'aiuto del calcolatore e trascrivevo il testo lettera per lettera con penna e carta quadrettata. Fortunatamente un attento esame dei rapporti fra sistema grafemico e sistema fonemico mi ha permesso di formulare un algoritmo per automatizzare anche questa fase preliminare della ricerca che ora con il mio maestro, prof. Carlo Tagliavini, e con la sua collaboratrice dott.ssa Umberta Bortolini, applico allo studio fonologico del vocabolario italiano.

* * *

Lo studio del linguaggio con i calcolatori, oltre alla linguistica, interessa oggi discipline diverse; psicologia, antropologia, sociologia, archeologia, documentazione automatica, traduzione automatica, composizione tipografica, logica, teorie dell'informazione, dattilografia, crittografia, fol-

lore, foniatria, logopedia, pedagogia ecc.

Non è possibile passarle tutte in rassegna e devo limitarmi a un solo esempio preso dalle applicazioni psichiatriche.

L'interesse degli psicologi e degli psichiatri per il linguaggio è legato al fatto che l'espressione linguistica riveste una grande importanza nella indagine della intera personalità.

La scoperta di differenze linguistiche tra diverse categorie nosologiche potrebbe essere utilizzata e per la diagnosi e per una maggiore comprensione della malattia mentale.

Lo studio del linguaggio intrapreso nella Clinica di psichiatria della Università di Pisa dall'équipe guidata dal prof. Giannini riguarda gli adulti psicotici (maniaci, schizofrenici) e nevrotici e i bambini intellettivamente ipodotati.

L'analisi morfologico-lessicale del linguaggio maniacale ha per ora dimostrato eccezionale l'uso del *futuro*, raro il *presente*, più frequente il *passato*. Ciò pone in dubbio le reali capacità di progettazione ed infuturazione del maniaco, ravvicinando il mondo maniacale a quello del depresso, ancorato al passato.

Il decremento degli aggettivi sembra testimoniare la diminuita attitudine a qualificare o a differenziare le caratteristiche della realtà.

Lo studio del linguaggio di bambini intellettivamente ipodotati è stato fatto in correlazione con lo studio di bambini di normale dotazione mentale, provenienti dallo stesso am-

biente socio-economico e culturale. I primi risultati mostrano che la quantità delle parole diverse in rapporto al numero totale delle parole pronunciate (come dire la ricchezza del vocabolario impiegato) è inferiore nel gruppo degli insufficienti mentali. Si è osservata anche la diminuzione degli aggettivi, in particolare la percentuale di aggettivi diversi. Nei normali la percentuale degli aggettivi sul totale delle parole è risultata del 6,34 %, nei subnormali del 4,61 %. Tuttavia il rapporto tra frequenza dei verbi e frequenza dei sostantivi matura allo stesso modo nello sviluppo dell'insufficiente mentale e del bambino normale. Nel subnormale la difficoltà sta dunque nell'entrare in possesso del patrimonio linguistico, ma quando ciò è possibile il linguaggio verrebbe acquisito rispettando certe strutture che sono costanti e peculiari del linguaggio umano.

CONCLUSIONE

Il calcolatore, non sembra dubbio, si avvia a diventare uno strumento *normale* delle ricerche linguistiche: quando ciò si verificherà, la *computational linguistics* (il neologismo anglosassone è malamente traducibile) entrerà nella sua terza generazione. A Pisa ci prepariamo a questo, tanto è vero che si è ormai rivelata necessaria una intera équipe per dirigere e svolgere i compiti per i quali un anno fa bastava una sola persona. Operiamo sull'*hardware*, cioè sulla composizione fisica del calcolatore, per esempio ampliando, con catene supplementari interscambiabili di stampa, la gamma degli alfabeti e dei caratteri disponibili. Soprattutto, riteniamo nostro compito consolidare la metodologia e costruire quello che i miei colleghi ingegneri chiamano, con curioso termine gergale, *software*: cioè l'insieme di programmi e procedure che completano il calcolatore e ne fanno qualcosa di più di un complicato inerte castello di fili e transistor. Le tecniche di codificazione e i programmi di *utilità* (cioè adattabili a lavori diversi), frutto della preziosa collaborazione dei numerosi istituti ricordati, sono a disposizione dei ricercatori che si rivolgono al CNUCE per l'automazione dei loro progetti. La standardizzazione dei metodi risparmia al linguista il tempo necessario a organizzare il dialogo tecnico con la macchina: la standardizzazione dei criteri di registrazione assicura la interscambiabilità dei testi disponibili su nastro magnetico: un testo, una volta schedato, è usufruibile per ricerche successive di studiosi diversi anche non previste dal progetto iniziale. Attualmente il materiale linguistico contenuto nei nostri archivi magnetici può essere valutato in circa 30 milioni di parole, ed è in rapido accrescimento: si tratta, come affermano gli esperti anche stranieri, di una tra le più importanti e moderne biblioteche elettroniche del mondo.

4. La pré-édition

4.1. Elle consiste pour nous, à lire le texte mot par mot et à y ajouter à la main les signes ou les symboles qui devront être perforés avec les lettres et la ponctuation.

4.2. Elle comprend:

- a) corriger les erreurs d'impression du texte
- b) préciser la référence chaque fois qu'elle change
- c) caractériser certains symboles et certaines situations graphiques: par exemple, marquer le point qui n'est pas un point final, mais un point d'abréviation d'un mot, ou bien qui représente l'un et l'autre; marquer le tiret qui n'est pas un signe de ponctuation, mais un trait-d'union etc....
- d) spécifier certains types de phrases ou certains types de mots que l'on veut mettre en évidence. Nous appelons »spécificatifs« ces signes spéciaux et, jusqu'à maintenant, nous en avons trois systèmes:
 - dans l'Index Thomisticus, pour distinguer les phrases principalement selon »l'autenzia«, c'est-à-dire selon l'attribution de paternité: citations littérales, citations selon le sens, références à des titres d'œuvres, etc...
 - dans les textes de Qumran, principalement selon l'état paléographique: lecture incertaine ou alternative, particularité due au scribe, mot effacé etc...
 - pour les concordances de Sénèque, également pour le rapport entre les mots du texte et leurs variantes dans l'apparat, afin que variantes et mots du texte figurent ensuite, et naturellement en corrélation, dans les concordances finales
- e) distinguer les »formules« des »mots«: c'est-à-dire les symboles et les expressions, par exemple arithmétiques ou géométriques, qui ne sont pas des »mots«: ainsi MI et DII qui étaient les nombres romains 1001 et 502, et non le vocatif de *meus* ni le nominatif pluriel de *deus*, ou bien AB qui signifiait le segment *AB* et non la préposition *ab*.

5. La perforation

5.1. Bien que nous continuions à perforer sur cartes, nous aussi, sommes convaincus qu'il est préférable pour ce genre de travaux de perforer sur bandes de papier perforé.

5.2. Mais notre préoccupation essentielle est de perforer avec le plus grand soin: comment réduire les erreurs à une quantité minime et négligeable, puisque sur une grande quantité l'absence totale et complète d'erreurs est une limite aussi irréalisable que souhaitable?

Je ferai sur ce sujet un exposé à part.

Je me bornerai à dire que, une fois le texte perforé, nous le vérifions d'abord à la machine, puis nous l'imprimons (avec l'ordinateur) deux fois de suite, et le confrontons chaque fois avec le document original, en le lisant mot par mot.

- 5.3. La transcription de toutes les œuvres de Saint Thomas sur des cartes perforées à reporter ensuite sur bande magnétique, a nécessité leur lecture au moins six fois mot par mot: deux fois pour la pré-édition, une fois pour la perforation, une fois pour la vérification, deux fois pour le contrôle par la lecture. Lire six fois 1.700.000 lignes de texte — puisque tel en est le nombre dans l'œuvre de Saint Thomas — équivaut donc à lire une fois 10 millions de lignes.

6. La lemmatisation

- 6.1. La lemmatisation, comme je disais, consiste à reconnaître et à codifier ce qui, en elle, est l'unité graphique et en même temps significative.

- 6.2. La lemmatisation peut être au niveau du mot ou au niveau de la forme. Pour rester dans les limites de la terminologie Aristotelicienne-Thomiste, je dirai que la lemmatisation au niveau-forme considère les mots comme »termes« de la résolution d'une proposition dans ses éléments.

Par contre la lemmatisation au niveau du mot considère des mêmes termes la »suppositio« c'est-à-dire la fonction représentative signifiante de ce terme précisément dans cette proposition.

La lemmatisation au niveau du mot ne requiert pas un »dictionnaire de machine« contrairement à la lemmatisation au niveau de la forme.

- 6.3. Il est évident que la lemmatisation au niveau du mot demande beaucoup plus de temps que la lemmatisation au niveau de la forme. C'est pourquoi l'analyse linguistique de textes de grande étendue doit être faite automatiquement grâce au dictionnaire de machine. Ou peut peut-être en conclure que la lemmatisation au niveau du mot peut-être prise en considération seulement comme une des méthodes pour parvenir à constituer un dictionnaire de machine suffisant. Je me permets de proposer qu'un des points de discussion de ce colloque soit la comparaison entre ces deux lemmatisations.

Pour ce qui nous concerne, je renvoie à mon exposé sur notre dictionnaire latin de machine, qui contient, en ce moment, 80.000 formes de mots latins, et sur les problèmes, dont il nous impose la solution.

- 6.4. Je propose aussi que soit inclus dans les points à discuter la comparaison entre concordances non lemmatisées et concordances lemmatisées.

7. Le choix automatique du contexte

- 7.1. Dès le début, une des principales objections contre la compilation

automatique des concordances fut la suivante: l'ordinateur ne peut pas choisir intelligemment les mots du contexte, car ce travail requiert l'habileté spécifique de l'homme qui compile; et, par conséquent, les contextes découpés automatiquement sur la base de schémas fixes ne sont vraisemblablement pas suffisants.

- 7.2. Je tiens à remarquer que, dans une concordance, on peut faire l'étude d'une forme ou d'un lemme sur des plans différents: par exemple le but peut-être simplement de reconnaître morphologiquement ou de définir lexicalement une forme ou un lemme; il peut être aussi la recherche doctrinale des concepts dont ce mot est le soutien.
- 7.3. En tout cas, je soutiens que, pour un mot qui n'est pas un apax, des contextes d'environ 100 positions, c'est-à-dire de dix-douze mots, sont pratiquement et dans leur ensemble toujours suffisants pour une définition lexicale de ce mot, si l'ordinateur a été programmé pour reconnaître les limites des phrases en fonction de la hiérarchie des signes de ponctuation ou des autres situations graphiques.
- 7.4. Nos programmes, en effet, organisent la délimitation du contexte pour chaque mot de la façon suivante:
 - l'ordinateur cherche jusqu'à 50 positions — plus ou moins, selon les désirs — avant le mot et 50 positions après;
 - dans le cas où la cinquantième position coupe un mot, il cherche à quelle extrémité on peut inclure un mot entier, si l'on ajoute les positions gagnées à l'extrémité opposée en renonçant au mot que la cinquantième position couperait en deux;
 - en outre, si, avant la cinquantième position, il rencontre un point final ou un autre signe équivalent, il ne va pas au-delà dans ce sens, mais ajoute à l'extrémité opposé les positions qui n'ont pas été utilisées à la première extrémité;
 - mais si le mot se trouve à l'intérieur d'une citation littérale, l'ordinateur franchit le point final, sans cependant, aller au-delà de la citation, lorsque celle-ci est suffisamment longue pour fournir un contexte entier;
 - par contre, dans le cas où le mot se trouve entre deux points rapprochés, l'ordinateur franchit le point final ou de droite ou de gauche, si la présence ou l'absence de signes spécifiques, qu'a le mot, continue seulement à droite ou seulement à gauche; par contre il franchit le point toujours à droit, si la situation est la même des deux côtés, etc....

Nous avons avec nous des exemples de concordances obtenus avec ce programme, afin que la personne intéressée puisse les examiner.

- 7.5. Enfin la cohérence et la constance de son caractère systématique plaident en faveur de cette délimitation automatique des contextes: on offre

ainsi une documentation parfaitement objective et entièrement à l'abri des fluctuations et de la subjectivité inévitable du choix de l'homme: quand, par exemple, on employait vingt étudiants pour choisir les contextes pour la concordance du même auteur, comment pouvait-on garantir que leurs critères n'aient pas été différents et n'aient pas subi de variations?

8. La structure de nos indices et concordances finals

8.1. Les résultats de nos travaux, pour chacune des deux parties (œuvres authentiques et apocryphes), seront publiés, répartis en deux grosses sections: la section-indices et la section-concordances.

La section-indices contient la documentation relative aux formes et aux lemmes: elle est donc récapitulative.

La section-concordances contient la documentation relative à tous les mots: elle est donc analytique.

8.2. Avant tout on présentera les indices de chaque œuvre particulière (le Corpus Thomisticum en contient 126). Ce sont:

8.2.1. Le *laterculum formarum* (Fig. 53,54; p. 264—5) qui donnera, par ordre alphabétique:

— toutes les formes rencontrées dans l'œuvre, lemmatisées et codifiées morphologiquement;

— toutes leurs homographes possibles selon Forcellini;

— les fréquences absolues de chacune, soit totale, soit détaillée par signes spécifiques;

— les fréquences proportionnelles;

8.2.2. Le *conspectus lemmatum* (Fig. 58; p. 269) qui donnera également et de la même façon, en les divisant en 3 groupes, pour tous les lemmes rappelés par les mots de l'œuvre:

— les lemmes dont on aura rencontré au moins une présence effective;

— les lemmes représentés seulement par des formes homographes dont la valeur est absente de tout contexte examiné;

— les lemmes représentés seulement par des formes homographes sans occurrences, mais dont on n'aura pas même examiné la présence éventuelle, parce que très peu probable;

8.2.3. Le *lemmatum formae* qui présente, pour chaque lemme, la liste des formes qui le rappelle dans l'œuvre; chacune de ces formes est suivie seulement de la fréquence absolue.

8.3. Pour l'ensemble de toutes les œuvres les indices seront divisés en différents groupes:

8.3.1. Tout d'abord des listes analogues à celle du 8.2 qui représenteront un panorama complet de la terminologie thomiste.

8.3.2. Puis la liste des formes et des lemmes selon les fréquences; des formes par désinences (c'est-à-dire un »index a contrario«) et par codes mor-

phologiques; et la liste de l'homographie des formes et des désinences.

8.3.3. Enfin deux *tabulae vocum*, une pour les formes, une pour les lemmes; à côté de la colonne réservée aux formes et aux lemmes, on y trouvera autant de colonnes que d'œuvres comprises dans cette partie du »Corpus Thomisticum«; et dans chaque colonne on trouvera le chiffre total des présences de cette forme ou de ce lemme pour chaque œuvre, ainsi que son pourcentage par rapport au total des mots de l'œuvre. Ceci permettra de voir, d'un seul coup d'œil, la répartition des présences de chaque mot dans la succession des différentes œuvres.

8.4. La section concordances contiendra pour chaque forme de mots au moins l'indication de tous les endroits où elle se trouve de la première œuvre à la dernière.

Nous aurons donc une concordance unique pour les œuvres complètes, et non autant de cycles et de séries de concordances qu'il y a d'œuvres: nous ne donnerons pas même pour les œuvres plus importantes une concordance à part.

Ainsi la forme *connaturaliter*, par exemple, sera suivie de tous les passages qui la contiennent, d'abord dans les trois Sommes (in Sent., c. Gentiles, Theologiae), puis dans les Quaest. Disput., puis dans les Commentaires (aristotéliens, autres et bibliques), enfin dans les Opuscles.

8.4.1. La section concordance sera divisée en cinq groupes de telle sorte que les mots énumérés dans l'un ne se retrouvent dans aucun des autres.

8.4.2. Les mots se trouvant dans des phrases qui sont des références (par ex. ut dicit Philosophus in 5 Metaph.) seront d'abord réunis dans un index locorum, qui renverra à la seconde partie de ce premier groupe; cette seconde partie renferme, dans un ordre naturellement discontinu, toutes les lignes où se trouvent des références de ce type.

Cette documentation servira à celui qui voudrait étudier quels auteurs et quelles œuvres cita Saint Thomas, et avec quelle phraséologie.

8.4.3. Viendront ensuite, par ordre alphabétique et accompagnées de l'index locorum, toutes les phrases où l'A. se cite lui-même (par exemple: ut supra habitum est, ut infra dicetur).

8.4.4. Puis tous les débuts de période que l'A. cite en référence aux diverses parties du texte commenté, comme »incipits«.

8.4.5. Ces trois premières subdivisions de la section concordances ne dépasseront pas, dans la mesure où nous pouvons prévoir, 6 ou 7 pour cent de l'ensemble des mots.

8.4.6. Le groupe suivant par contre contiendra environ 60 ‰ des mots restants. Comme je le montrerai dans un exposé à part, un petit nombre de formes — 600, c'est-à-dire bien moins d'un pour cent — ont des

fréquences si élevées qu'elles totalisent 60 % des mots de l'œuvre. Le souci d'offrir une documentation à la fois intégrale et maniable, nous a menés à réunir ces mots en petits groupes, que nous appelons syntagmes, qui sont présentés par ordre alphabétique et qui sont suivis de l'index locorum, dont j'aurai l'occasion de parler plus en détail.

- 8.4.7. Le reste des formes, plus de 99 %, qui cependant représentent seulement un peu moins de 40 % du texte, sera présenté selon la manière traditionnelle des concordances.

Tous les contextes (d'environ 100 caractères, c'est-à-dire de 13 à 15 mots, découpés automatiquement par l'ordinateur, comme nous avons dit plus haut) qui contiennent la même forme seront réunis dans l'ordre du texte. Les formes seront ordonnées à l'intérieur de leur lemme, selon la séquence des codes morphologiques.

Nous aurons donc une concordance sur forme et non pas directement sur lemme. Nous mettrons en évidence, après la référence de chaque contexte, le code spécifique que le mot en question possédait; de cette manière, si cette phrase fait partie d'une citation littérale ou d'une citation de sens, l'utilisateur en sera informé.