ACCADEMIA NAZIONALE DEI LINCEI
ANNO CCCLXV – 1968

QUADERNO N. 110

**PROBLEMI ATTUALI**
**DI SCIENZA E DI CULTURA**

ATTI DEL CONVEGNO SUL TEMA:

# L'automazione elettronica e le sue implicazioni scientifiche, tecniche e sociali

(Roma, 16–19 ottobre 1967)

(*ESTRATTO*)

ROMA
ACCADEMIA NAZIONALE DEI LINCEI
1968

ACCADEMIA NAZIONALE DEI LINCEI

Estratto dal **Quaderno N. 110** - Atti del Convegno sul tema: « *L'automazione elettronica e le sue implicazioni scientifiche, tecniche e sociali* »
(Roma, 16-19 ottobre 1967)

ALDO DURO e ANTONIO ZAMPOLLI

# ANALISI LESSICALI MEDIANTE ELABORATORI ELETTRONICI (*)

RIASSUNTO. — Il primo capitolo della relazione, dopo avere distinto e definito il diverso valore di espressioni come *analisi lessicale*, *analisi lessicologica*, *analisi lessicografica*, si sofferma in particolare su quest'ultima, cui è assegnato il compito di raccogliere materiale documentario per la costituzione dell'inventario di una lingua o per la compilazione di un grande vocabolario storico. A questo fine, l'analisi lessicografica ha bisogno non di liste di parole o indici di frequenza, ma di vere e proprie concordanze (*concordanze di forme* e *concordanze di lemmi*) e, soprattutto, di schede–contesto, cioè di schede contenenti citazioni di passi d'autore o comunque testimonianze della storia della lingua; è proprio la scheda–contesto l'obiettivo finale a cui deve tendere l'analisi compiuta mediante elaboratori elettronici. Viene poi delineato il passaggio dalle concordanze per forme alle concordanze per lemmi, e spiegato in che cosa consista l'opera di *lemmatizzazione* e quali siano le difficoltà che rendono oltremodo difficile l'opzione per una lemmatizzazione interamente automatica mediante calcolatore. La principale di queste difficoltà è costituita dagli *omografi*, la cui presenza nel lessico italiano è numerosa al punto da far ritenere poco meno che utopistica l'attuazione di un programma di lemmatizzazione automatica. Risultati positivi ha dato invece un esperimento di lemmatizzazione *semiautomatica*, che l'Accademia della Crusca ha condotto su alcuni capitoli dei *Promessi sposi* presso il Centro Nazionale Universitario di Calcolo Elettronico di Pisa: in ogni capitolo, il linguista ha il compito di lemmatizzare soltanto le forme nuove e gli omografi; le forme univoche, una volta che sono state incontrate e lemmatizzate, vengono via via acquisite dalla memoria del calcolatore, che sarà poi in grado di operare automaticamente la loro lemmatizzazione tutte le volte che le incontrerà.

Il secondo capitolo, redatto dal dott. Zampolli, esamina i problemi tecnici connessi con l'analisi elettronica di testi linguistici, illustrando i procedimenti attraverso cui si giunge dalla perforazione di un testo alle schede finali, e soffermandosi in particolare sulla fase più delicata delle elaborazioni elettroniche: la lemmatizzazione automatica.

## I. Problemi scientifici e considerazioni generali.

Il limite della fede che noi abbiamo nella macchina, e dell'aiuto che a lei chiediamo, è già implicitamente dichiarato nel titolo stesso di questa comunicazione, titolo che non abbiamo scelto noi di proposito e che tuttavia riflette esattamente il nostro pensiero: « analisi compiuta » non *da*, ma « *mediante* elaboratori elettronici ».

Crediamo infatti, ed è opinione maturatasi in tutti questi mesi di intensa sperimentazione, che ove si tratti non semplicemente di formare statistiche

(*) La relazione è distinta in due capitoli: il primo è dovuto ad A. Duro, il secondo ad A. Zampolli.

di parole in quanto segni grafici – e perciò meccanicamente riconoscibili da mezzi automatici – ma di operare analiticamente su parole significative le quali, pur nell'identità grafica, possono essere diverse per funzione e per contenuto semantico, come sono gran parte delle unità lessicali di cui è costituito il vocabolario di una lingua, la macchina può essere, senz'altro, un sussidio utilissimo, ma non è in grado di sostituirsi all'opera artigianale e intelligente dell'uomo.

Questo nostro atteggiamento di prudente riserva di fronte alle effettive o teoriche possibilità della macchina è condiviso da molti studiosi delle discipline umanistiche; altri invece spingono la loro fede molto più avanti. Il diverso atteggiamento si risolve, nella pratica, non in una totale accettazione o sconfessione dei sistemi automatici, ma in modi differenti di programmare il lavoro (1).

Per circoscrivere il discorso alle ricerche linguistiche e alla documentazione lessicale, sarebbe un grave anacronismo, crediamo, disconoscere gli immensi vantaggi che, anche soltanto dal punto di vista della celerità, dell'economia di costi, dell'esattezza dei risultati, gli spogli elettronici offrono in confronto ai metodi tradizionali di spoglio lessicografico: estensione dell'analisi a un numero di opere praticamente illimitato; elevata velocità di elaborazione e selezione; omogeneità dei metodi di ricerca e quindi dei risultati ottenuti; possibilità di compiere grande varietà di analisi d'ogni tipo, anche non programmate all'inizio, partendo sempre da un medesimo supporto d'informazione; ottenimento, in uscita, di schede–contesto ricche di dati a stampa e in perforazione.

Pretendere di elencare, non diciamo tutte, ma le più importanti fra le imprese attuate con l'aiuto dei mezzi elettronici in questo primo decennio della loro diffusione, nel campo dell'analisi letteraria, linguistica, stilistica, e persino metrica, significherebbe affrontare un'indagine non solo ardua ma che rischierebbe di non essere neppur lontanamente esaustiva. Ci limitiamo perciò a ricordare quelli che sono stati i momenti decisivi del cammino dell'automazione applicata all'analisi che a noi interessa, cioè all'analisi lessicale, o, più ristrettamente ancora, alla documentazione lessicografica.

In primo luogo, sia per ragioni di precedenza cronologica, sia per l'affinità dei metodi e dello scopo finale, sia infine per i legami di stretta collaborazione che l'Accademia della Crusca ha avuto con il Centro d'Automazione di Gallarate, va menzionato l'avvio dell'*Index Thomisticus* e la ben docu-

(1) Proprio per ciò che ha relazione con il nostro settore di studio, ecco come B. Quemada riassumeva il significato dei vari interventi avutisi al Colloquio internazionale di Besançon (giugno 1961) sulla meccanizzazione delle ricerche lessicologiche: « A la lumière des échanges de vues s'est trouvé confirmée la tendance qui divise, en apparence plus qu'en réalité, les utilisateurs des machines. Ainsi semblent s'opposer lexicologues ' classiques ' qui désirent bénéficier des moyens mécaniques pour accroître leurs possibilités de travail en suivant des normes d'exploitation ayant fait leurs preuves, et lexicologues ' modernes ' qui, eux, n'auraient jamais abordé la lexicologie sans l'existence des machines, et qui de ce fait songent à des applications très différentes des précédentes » (in *Cahiers de lexicologie*, n. 3, 1962, p. 3).

mentata notizia che il direttore dell'impresa, p. Roberto Busa, ne dava già nel 1951 nel volumetto *Sancti Thomae Aquinatis Hymnorum ritualium VARIA SPECIMINA CONCORDANTIARUM...*(2); a cui sono da aggiungere gli *Indicis Thomistici Specimina*, messi insieme e pubblicati sempre da p. Busa nel 1963, che confermano le previsioni da lui espresse dodici anni prima circa le grandi possibilità di sviluppo dei mezzi elettronici nell'àmbito dell'analisi linguistica.

Altra tappa fondamentale, il Colloquio internazionale svoltosi a Strasburgo nel novembre 1957 nel quadro dei Colloqui internazionali del Centre National de la Recherche Scientifique (i cui atti furono poi raccolti nel volume *Lexicologie et lexicographie françaises et romanes* nel 1961), e in particolare l'intervento di B. Quemada su « La technique des inventaires mécanographiques », con la discussione che ne seguì. A due anni dal Colloquio di Strasburgo, il 1° numero dei *Cahiers de lexicologie*, del 1959, contiene già una particolareggiata esposizione, fatta dallo stesso Quemada, del programma di lavoro per l'impianto di uno schedario meccanografico che costituisse l'inventario generale del vocabolario francese.

Non meno importante fu il Colloquio internazionale di Besançon, del 1961, i cui atti sono pubblicati nel 3° fascicolo dei *Cahiers de lexicologie*.

Ai due Colloqui, di Strasburgo e di Besançon, e al fervore di ricerche che ne seguì, è dovuta l'istituzione in Francia del Centre de recherche pour un Trésor de la langue française, con sede a Nancy, che sotto la direzione di P. Imbs ha già quasi condotto a termine lo spoglio elettronico di circa 2500 opere per un complesso di oltre 300 milioni di citazioni.

Attività parallele o affini, che non possono essere passate sotto silenzio neppure in una così fuggevole e sommaria rassegna, si svolgono presso il « Centro per gli studi letterari e linguistici mediante calcolatore » dell'Università di Cambridge (che ha in programma l'elaborazione di testi letterari in lingue europee moderne e medievali per la produzione di indici e concordanze, anche a fini lessicografici); presso il « Laboratorio di analisi statistica delle lingue antiche » dell'Università di Liegi (per la produzione automatica di indici e concordanze, e per indagini morfologiche e sintattiche su opere di autori classici); presso il « Laboratorio meccanografico » dell'Istituto per la lingua cèca di Praga (che ha fra l'altro in programma, in collaborazione con l'Accademia cecoslovacca delle Scienze, il trattamento meccanografico dei lemmi contenuti nel Dizionario della lingua cèca letteraria); presso l'Accademia delle Scienze di Berlino Est; presso le Università di Göteborg in Svezia, di Amsterdam e Leida in Olanda, ecc. Un « Inventario linguistico dell'italiano delle Origini e del Duecento », con analisi grammaticale, sintattica e strutturale dei testi compiuta elettronicamente, è in corso di attuazione a cura dell'Istituto di lingua e letteratura italiane dell'Università di Utrecht. Molto più ampio il programma dell'Accademia della Crusca, che, con il finanziamento del

(2) Il titolo continua: *Primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate* (Milano, Fratelli Bocca, 1951).

Consiglio Nazionale delle Ricerche, sta attivamente operando per la costituzione di un immenso archivio della lingua italiana, ricco di oltre 100 milioni di citazioni su schede (da ottenere in parte con elaborazione elettronica, in parte con spogli artigianali), in vista della redazione di un grande Vocabolario storico della lingua italiana.

* * *

Per *analisi lessicale* (compiuta con mezzi elettronici) intendiamo una analisi che si proponga qualche cosa di più della semplice ricerca, a fini statistici, di parole, o elementi di parole, o gruppi di parole – si tratti di fonemi, grafemi, morfemi, sintagmi, lessemi o altro – un'analisi che, partendo da un testo dato, ne tragga almeno una lista di vocaboli, ciascuno dei quali sia però corredato di un contesto che lo presenti in tutta la complessità delle sue accezioni, del suo uso, della sua funzione, nel suo pieno valore insomma. Questa lista, stampata, è ciò che comunemente si chiama *concordanza* o *lista di concordanze*; essa può rappresentare sia un punto d'arrivo, sia una fase intermedia, dell'indagine linguistica; nell'uno e nell'altro caso, essa è diversamente utilizzabile, a seconda del programma con cui è stata prodotta. Se le concordanze vengono sfruttate per operare la cosiddetta « lemmatizzazione » delle forme flessive quali si trovano nel testo, se alle prime *concordanze per forma* si fanno seguire le *concordanze per lemmi* – ottenute anche queste meccanicamente – e se alla fine si chiede al calcolatore di ampliare i brevi contesti delle concordanze in contesti più lunghi, e di stamparli non più su fogli ma su schede, avremo così esaurito la serie di operazioni che, in termini d'automazione, costituiscono l'analisi lessicale [3]. Ulteriormente, sarà da distinguere un'*analisi lessicologica*, che ha fini e metodi di ricerca suoi propri, e che in genere ha bisogno di contesti brevi (in concordanze o su schede), dall'*analisi lessicografica*, che ha piuttosto carattere documentario, proponendosi come scopo la costituzione di un archivio o inventario della lingua, sia generale, sia circoscritto a periodi o ambienti particolari.

(3) Si rende qui necessario un chiarimento terminologico. Chiamiamo **parole** o **occorrenze** le unità grafiche di cui è costituito un testo (per la macchina, la parola è un segno o un gruppo di segni separato prima e dopo da uno spazio): nella terzina dantesca « Per me si va ne la città dolente, Per me si va ne l'etterno dolore, Per me si va tra la perduta gente » le parole o occorrenze sono 24 (8 per ciascun verso). Chiamiamo **forme** le unità lessicali, così come s'incontrano nel testo, nella forma flessa della declinazione nominale o della coniugazione verbale, o nella loro forma invariata se si tratta di vocaboli non soggetti a flessione: perciò *va* e *perduta* sono due forme, così come *per* e *tra*; nel computo statistico, peraltro, i tre *per*, nello stesso modo che i tre *me*, i tre *si*, i tre *va*, contano per una forma sola, sicché le forme della terzina citata sono in tutto non 24 ma 14 (*per*, *me*, *si*, *va*, *ne*, *la*, *città*, *dolente*; *l'*, *etterno*, *dolore*; *tra*, *perduta*, *gente*). Chiamiamo **lemma** la parola ricondotta alla forma di base, quella cioè con cui ciascuna unità lessicale viene usualmente registrata nel vocabolario: alla forma *va* corrisponde il lemma *andare*; alla forma *perduta* il lemma *perduto* o, secondo i casi, *perdere*. L'operazione di ricondurre ciascuna forma al rispettivo lemma è ciò che noi chiamiamo **lemmatizzazione** (o, come preferirebbe dire G. Devoto, *lemmazione*).

In vista di questo scopo, obiettivo finale della documentazione automatica è, per noi, la scheda–contesto (4). La somma delle indicazioni che questa deve portare in chiaro, la giusta lunghezza che deve avere il passo citato per consentire l'esatta interpretazione della parola nella sua denotazione e nelle sue connotazioni, il numero e la qualità delle perforazioni da praticare nella scheda per la classificazione e la selezione meccanica, sono altrettanti problemi dalla cui soluzione dipende il grado di utilità dello schedario in cui le schede confluiranno.

Che cos'è, in concreto, una scheda–contesto? È un cartoncino rettangolare, del formato internazionalmente usato per il trattamento meccanografico, sul quale è stato impresso dalla stampatrice elettronica un passo appartenente all'opera analizzata; il passo, che può essere lungo fino a una decina di righe, contiene ed esemplifica una determinata parola, che la scheda stessa reca in esponente sia nella forma flessa con cui compare nel testo (ammesso che si tratti di parola soggetta a flessione) sia nella forma lemmatizzata. Oltre alla citazione e agli esponenti, sono stampati anche alcuni riferimenti indispensabili: nome dell'autore, titolo dell'opera, esatta posizione della parola e del passo nel volume, data di composizione o di edizione dell'opera. Questi stessi riferimenti, insieme con il lemma e la eventuale forma flessionale, sono anche perforati, in tutto o in parte, sulla scheda, al fine sia di facilitare l'inserzione meccanica nello schedario secondo un criterio alfabetico e cronologico, sia di permettere successivi smistamenti o selezioni, sempre con mezzi automatici.

Proposito di questa comunicazione è di illustrare brevemente i procedimenti attraverso i quali si giunge, dalla fase di concordanze non lemmatizzate, alla scheda–contesto. Prima però desideriamo liberare il terreno da alcune reali o possibili obiezioni. Questa, per esempio: se sia davvero utile, in tutti i casi e per qualsiasi tipo di spoglio, ricorrere ai complessi elettronici, e non convenga invece mantenere i vecchi sistemi artigianali almeno per quelle opere da cui può essere estratto scarso materiale documentario. L'obiezione è giusta. Dal punto di vista dei costi di produzione, è economico ricorrere all'analisi automatica soltanto per opere da sottoporre a spoglio integrale o fittissimo, per le quali il rapporto fra il totale delle schede finali ottenute (schede–contesto) e il totale delle parole che costituiscono il testo analizzato (occorrenze, o parole *input*) sia non inferiore al 10%. Da una serie di calcoli, fatti sulla base di un volume di 300 pagine, suppergiù equivalenti a 10.000 righe e a 100.000 parole, si ricavano le seguenti cifre indicative dei costi (5):

(4) Dicendo « per noi » intendiamo riferirci all'Accademia della Crusca, la quale, con l'assistenza tecnica del Centro di Gallarate e del dott. Zampolli, è stata il primo e forse il solo centro di ricerca a proporsi come fine essenziale delle proprie elaborazioni elettroniche la scheda–contesto, e a realizzarla in maniera pienamente soddisfacente.

(5) Per calcolare le spese di utenza del calcolatore, si è tenuto conto dei costi unitari addebitati dal Centro Nazionale Universitario di Calcolo Elettronico (CNUCE) di Pisa, notevolmente inferiori a quelli che le varie società praticano ai clienti ordinari.

* *A. Duro e A. Zampolli.*

*a*) In uno spoglio elettronico integrale, che dia un *output* di 50.000 schede–contesto, ossia utilizzi in media una parola su due (il che significa la totalità delle occorrenze lessicali, e una bassa percentuale delle parole grammaticali), ogni scheda–contesto viene a costare, tra macchina, materiali, e lavoro umano specializzato, lire 28.

*b*) In uno spoglio elettronico fittissimo, che dia come risultato 25.000 schede–contesto, con un rapporto tra *output* e *input* di 1 a 4, il costo di ciascuna scheda finale è di lire 42.

*c*) Per spogli selettivi più radi, si può salire fino a 160 lire a scheda per una selezione che frutti il 5% di schede finali rispetto all'*input*, cioè utilizzi una parola ogni due righe di testo.

Di fronte a queste cifre che, come s'è visto, variano in rapporto inverso alla fittezza dello spoglio, i costi delle schede ottenute manualmente o col sussidio di apparecchiature fotomeccaniche sono fissi: 200 lire è il costo medio di una scheda trascritta manualmente; non meno di 120 lire l'una costano le schede ottenute con tecniche fotoxerografiche. Una comparazione fra i tre diversi sistemi di spoglio (tutti e tre in atto presso l'Accademia della Crusca) è quindi possibile solo quando il rapporto tra le parole utilizzate per la schedatura e gli scarti si aggiri su valori uguali o inferiori al 5%. Per valori più alti, il costo degli spogli compiuti con elaboratori elettronici è nettamente inferiore.

Il confronto non deve però essere fatto soltanto sul piano economico. Non si può ignorare il fatto che, dalla elaborazione automatica dei testi, si ottengono, oltre alle schede di citazione, parecchi altri risultati intermedi che, anche se non siano ritenuti direttamente utili ai fini della documentazione lessicografica, possono essere messi a disposizione di altri studiosi, interessati a vari tipi di analisi, linguistica o stilistica. Tra questi risultati intermedi, i più importanti sono gli indici di frequenza e le concordanze. Queste ultime, in particolar modo, offrono la possibilità di spingere l'indagine a fondo, a livelli che in nessun altro modo potrebbero essere raggiunti; infatti, la presentazione simultanea, su una medesima lista, di tutti gli esempi di una stessa parola, consente di vedere, quasi con un sol colpo d'occhio e nel vivo dei suoi contesti, tutte le somiglianze, le affinità, o al contrario le diversità d'uso della parola, sotto l'aspetto semantico, morfologico, sintattico, stilistico, ecc.

A titolo sperimentale, abbiamo voluto sottoporre *I Malavoglia* del Verga a uno spoglio selettivo, eseguito contemporaneamente con due tecniche diverse: con il cosiddetto metodo xerografico (che è soltanto un perfezionamento del metodo tradizionale della schedatura a mano) e con il sistema elettronico. Non riporteremo tutti i dati comparativi dei risultati dei due differenti tipi di spoglio, che sono peraltro molto istruttivi, e che potranno fornire materia per una pubblicazione a sé; citeremo solamente alcune tra le più interessanti osservazioni di carattere linguistico rese possibili anche da una semplice e frettolosa lettura della lista delle concordanze stampate dal calcolatore.

L'aggettivo *lustro*, che nel romanzo ricorre sette volte, e sempre al plurale, in sei casi è attributo di *occhi*, una di *capelli*; l'aggettivo *lucido* non s'incontra

mai. *Mortale*, che ha sei ricorrenze, si trova soltanto in compagnia di *peccato*. Il verbo *osare*, nelle venti sue comparizioni, è sempre preceduto da negazione. Il verbo *ingannare*, che ha sei ricorrenze, si trova cinque volte nell'espressione *ingannare il tempo*, una nell'espressione *ingannare la noia*. La *Sicilia* non è mai menzionata nel romanzo; *siciliano* compare una volta sola.

Uno dei problemi più ardui negli spogli integrali è costituito dalla distinzione tra parole lessicali e parole grammaticali (o, secondo altre terminologie, parole piene e parole vuote, parole autosemantiche e parole sinsemantiche), soprattutto in merito al trattamento da riservare alle seconde – cioè agli articoli, preposizioni, congiunzioni, pronomi, avverbi non modali – che, come è ormai noto, costituiscono all'incirca il 50% del totale delle occorrenze di un qualsiasi testo. È ovvio che, anche trattandosi di un autore dell'importanza di Dante, sarebbe assurdo voler riportare in un vocabolario storico, per ampio che sia, tutti gli esempi dell'articolo *la* (che nella *Divina Commedia* assommano a 2824), della congiunzione *e* (3884), del pronome e congiunzione *che* (4511), e così via. Poco economico, e in realtà scarsamente redditizio, sarebbe anche proporsi di scegliere con criteri rigorosamente lessicologici o lessicografici gli esempi più utili e belli fra le migliaia che le concordanze registrano per ogni forma delle particelle di altissima frequenza. Si può immaginare di fare una scelta, che sia più che casuale, fra le 9435 ricorrenze della congiunzione *e*, o fra le 4188 ricorrenze della preposizione *a*, dei *Reali di Francia* di Andrea da Barberino (citiamo quest'opera perché ne abbiamo sott'occhio le concordanze)? Una scelta di questo genere si potrà fare tutt'al più su cinque o sei opere, la *Divina Commedia*, per esempio, o il *Canzoniere* del Petrarca, o i *Canti* del Leopardi, o sui primissimi documenti del nostro volgare; ma per opere di minor valore linguistico e letterario è inevitabile che ci si debba affidare al caso, ossia a una selezione automatica eseguita dallo stesso calcolatore, che scarti un'alta percentuale degli esempi di tutte le particelle grammaticali, lasciando all'analisi del linguista una percentuale minima: un esempio su 10, su 50, su 100, secondo che l'indice di frequenza sia per ciascuna particella più o meno alto.

Occorre insistere, peraltro, sul fatto che questa fortissima riduzione (o addirittura, per autori e testi di minore rilievo, l'esclusione) delle ricorrenze delle parole grammaticali riguarda soprattutto la fase finale, quella delle schede–contesto, e quindi il momento più strettamente lessicografico dell'analisi. Per chi invece compie ricerche di tipo statistico, o in genere lessicologico, può essere utile avere l'elenco completo di tutte le ricorrenze delle parole grammaticali anche meno significative [6].

(6) « Puisque je parle de ces Index, il est peut–être regrettable que nous n'ayons pas indexé tous les mots–outils; nous ne l'avons pas fait pour des raisons matérielles: cela aurait doublé le prix de vente. Je voudrais vous donner un exemple précis: il y a quelques jours, j'analysais l'emploi de l'article dans la Chanson de Roland; je possédais un Index complet, je ne m'en suis pas servi parce que j'ai plus vite fait de relire le texte et de rechercher, moi–même, tous les articles plutôt que d'aller à l'index. Mais il est possible que, dans un dévelop-

* * *

Per spiegare in modo più concreto che cosa siano e in che differiscano le concordanze di forme e le concordanze di lemmi, possiamo dire, con un esempio, che, nelle concordanze per forme, il verbo *avere* comparirà in esponente con questo aspetto d'infinito solamente in testa ad una serie di esempi che contengano effettivamente il verbo nella forma dell'infinito. Quei passi, invece, nei quali il verbo è presente con un'altra forma della sua coniugazione, si troveranno via via dislocati sotto gli esponenti *ho*, *hai*, *ha*, *abbiamo*, *avevo*, *ebbi*, *avessi*, *avendo*, *avuto*, ecc., e quindi a grande distanza l'uno dall'altro, sotto lettere dell'alfabeto diverse.

Nelle concordanze per lemmi, al contrario, tutti gli esempi saranno riuniti sotto l'unico esponente *avere*, ordinati secondo la successione alfabetica o paradigmatica delle forme della coniugazione cui ciascun gruppo di esempi appartiene.

In che modo si compie questo raggruppamento delle forme verbali, e quello di altre forme declinabili (come i sostantivi e gli aggettivi) che devono essere riunite sotto la forma del singolare maschile?

Le soluzioni sono diverse, ma le vie sono praticamente due: o attraverso una serie di operazioni automatiche (di cui diremo tra poco); o attraverso una lemmatizzazione artigianale, fatta dallo specialista, che pazientemente scorre, esempio per esempio, tutte le forme registrate nelle liste di concordanze, ne analizza il valore morfo-semantico, e, se si tratta di forme flesse, assegna a ciascuna la rispettiva forma di base: a un *vadano* attribuisce il lemma *andare*, a un *azzurre* il lemma *azzurro*, a una preposizione articolata *degli* la forma della preposizione semplice *di*. L'operazione scorre piana e veloce finché s'incontrano soltanto forme univoche, che possono cioè appartenere a un lemma unico, il quale a sua volta non consuoni con altre parole, di etimologia e significato diversi, o, come altrimenti si dice, non sia *omografo* di altri vocaboli. Di fronte a casi di omografia, invece, l'opera di lemmatizzazione è meno rapida ed è irta di pericoli e di difficoltà.

Un *ridesse* riportato direttamente a *ridere* non terrebbe conto che *ridesse* può essere anche congiuntivo di *ridare*. Dire che *luci* è il plurale di *luce* è affermazione incauta, se non si è prima accertato, con una attenta lettura del contesto, che non sia invece una forma del verbo *lùcere*. Trovando in tutta una sequenza di esempi la parola *danno*, sempre preceduta da *il*, non si sia troppo frettolosi a classificarla come sostantivo maschile; potrebbe essere invece, specie in testi antichi o poetici, forma del verbo *dare* (per es.: *essi il danno per certo*), o addirittura presente di *dannare* (per es.: *io il danno a questa pena*). Se poi dalle forme passiamo ai lemmi, vediamo che non c'è diversità grafica, e perciò non è possibile una distinzione automatica, tra

pement ultérieur de la science, il soit utile d'avoir, dans un Index, tous les mots-outils ». Così P. Guiraud, in un suo intervento al Colloquio di Strasburgo (in *Lexicologie et lexicographie* cit., p. 67).

parole come *miglio* (pianta e seme) e *miglio* (misura di distanza); fra *trattore* persona e *trattore* veicolo; fra *lustro* e *parco* aggettivi e *lustro* e *parco* sostantivi; tra un *filare* d'alberi e il verbo *filare*; tra gli aggettivi *militare*, *popolare*, *salutare* e i verbi omografi. E se nella pronuncia *pànico* e *nòcciolo* hanno suono differente rispetto a *panìco* e *nocciòlo*, nella scrittura corrente, dove l'accento non è segnato, le due parole non si distinguono, così come non si distinguono *pésca* e *pèsca*, o *cólto* e *còlto*.

L'esistenza di omografi, che la lingua italiana possiede in grandissimo numero, e la frequenza delle varianti grafiche, delle quali pure la nostra lingua ha grande ricchezza, specialmente in dimensione diacronica, sono tra gli ostacoli più seri che si oppongono all'attuazione di un programma di lemmatizzazione automatica, cioè di una connessione forma–lemma compiuta dal calcolatore stesso. Né meno gravi sono le difficoltà soggettive, procedenti dalla relatività stessa del concetto di omografia, specialmente per ciò che riguarda gli usi sostantivati di alcune parti del discorso (aggettivi, participi, infiniti, ecc.) e i molti casi in cui non c'è separazione chiara tra omografia e polisemia.

La lemmatizzazione automatica non è tuttavia un sogno illusorio di pochi patiti della macchina elettronica. Operando sulla lingua latina, il p. Busa ha potuto effettuarla con ottimi risultati per i dieci milioni e mezzo di parole dell'*Index Thomisticus*. Sempre su testi latini, ma di età classica, è allo studio, e in gran parte ormai perfezionato, un programma di analisi e lemmatizzazione automatica presso il Laboratorio di Liegi. Un « dizionario di macchina », o dizionario di forme flesse, per l'attribuzione automatica di 400.000 forme verbali ai rispettivi lemmi è stato programmato ed è utilizzato a Nancy nella raccolta del materiale documentario per il *Trésor de la langue française*.

Ma né il latino né il francese posseggono una così numerosa serie di voci omografe quante ne ha l'italiano, e per di più il nostro àmbito di ricerca si estende dalle origini ai giorni nostri, per ben otto secoli, e sarebbe materialmente impossibile prevedere tutte le forme della flessione verbale e nominale, e tutte le varianti, che le parole della nostra lingua hanno avuto via via nei secoli.

È soprattutto questo il motivo per cui abbiamo sempre sostenuto l'opportunità di compiere la lemmatizzazione con il metodo tradizionale, intervenendo cioè *manualmente* sulle liste di concordanze per riportare le forme flesse ai rispettivi lemmi.

Tuttavia, per non rinunciare del tutto, con aprioristico scetticismo, a un programma di lemmatizzazione meccanica, ci siamo alla fine decisi, proprio in vista di questo Convegno, a tentare un esperimento di lemmatizzazione che chiameremo « semiautomatica », portato a termine in questi ultimi mesi con un programma scritto apposta dal dott. Zampolli, del Centro Studi della IBM–Italia di Pisa, che di tutte le ricerche che l'Accademia della Crusca compie con elaboratori elettronici è l'intelligente e infaticabile programmatore e coadiutore.

L'esperimento ha avuto per oggetto *I promessi sposi*, nell'edizione 1840–42, e per strettezza di tempo si è dovuto limitarlo ai soli primi dieci capitoli. Il programma, del resto molto semplice, può essere esposto in poche parole. Dopo aver trasferito il romanzo su schede perforate, e da queste su nastro magnetico, si è condotta l'elaborazione elettronica del materiale fino alla fase delle concordanze per forma; ma, per questa fase, si è proceduto in modo diverso dal solito: si sono stampate su lista le concordanze non dell'intera opera, ma di un capitolo per volta. A mano a mano che veniva lemmatizzato manualmente un capitolo, il calcolatore immagazzinava nella propria memoria le singole forme insieme con il lemma a ciascuna assegnato, e se ne serviva poi per una lemmatizzazione automatica delle forme uguali che incontrava nei capitoli successivi, presentando alla lemmatizzazione dello specialista soltanto le forme nuove. Si è proceduto con questo sistema fino al decimo capitolo, avendo peraltro cura di escludere dalla lemmatizzazione le forme omografe. Queste sono state via via accantonate, e alla fine del decimo capitolo il calcolatore ne ha stampato una lista completa, perché ne fosse fatta un'unica oculata lemmatizzazione. A differenza, infatti, delle voci univoche, le quali possono essere ricondotte senza esitazioni alla loro unica forma di base (*vedevano* non può essere altro che l'imperfetto del verbo *vedere*), le forme omografe possono essere attribuite a più lemmi; più esattamente, ciascuna di esse può appartenere a un solo lemma, da scegliersi però in un gruppo più o meno largo. Per esempio, trovando un certo numero di passi conte nenti tutti la medesima forma *legge*, si sono dovuti distinguere quelli in cui *legge* è sostantivo femminile (*légge*) e quelli in cui è voce del verbo *leggere* (*lègge*).

Quali sono stati i vantaggi, e quali il risultato e il significato di questo esperimento?

Innanzi tutto, si è sostituito parzialmente all'opera e alla mente dell'uomo la memoria del calcolatore, il quale si è dimostrato capace di riconoscere le forme già incontrate in precedenza, e di riportarle automaticamente al lemma ch'era stato loro assegnato una volta per tutte.

Così, mentre nel 1° capitolo tutte le 2243 forme univoche sono state lemmatizzate manualmente, nel 2° capitolo, su un totale di 4460 occorrenze e 1514 forme, il calcolatore ha saputo riconoscere e lemmatizzare 3309 occorrenze per 601 forme e 517 lemmi, lasciando al linguista il compito di assegnare il lemma alle sole 913 forme nuove. Nel 3° capitolo si hanno 4889 occorrenze, 772 forme nuove, 762 forme lemmatizzate dal calcolatore. I dati relativi a questi e altri capitoli risultano chiaramente dalla tabella comparativa stampata nella pagina seguente.

Le forme omografe, che, come già s'è detto, sono state lemmatizzate con un'unica operazione alla fine (la distinzione è infatti più facile quando si hanno presenti e comparabili tutti quanti i contesti in cui ciascuna parola si trova usata), sono risultate essere: 523 nel 1° capitolo, 413 nel secondo, ecc. (vedi colonna 8 della tabella).

*Statistiche relative ai primi 10 capitoli dei Promessi sposi.*

| Capitolo | Occorrenze | | | | Forme | | | | Lemmi | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1. Totale | 2. vecchie | 3. nuove | 4. omografe | 5. Totale | 6. vecchie | 7. nuove | 8. omografe | 9. Totale | 10. vecchi | 11. nuovi |
| 1. | 7603 | — | [7603] | 1200 | 2766 | — | [2766] | 523 | 1622 | — | [1622] |
| 2. | 4460 | 3309 | 1151 | 917 | 1514 | 601 | 913 (20,47 %) | 413 | 1050 | 517 | 533 (11,95 %) |
| 3. | 4889 | 3873 | 1016 | 1075 | 1534 | 762 | 772 (15,79 %) | 404 | 1037 | 641 | 396 ( 8,09 %) |
| 4. | 5375 | 4303 | 1072 | 1092 | 1827 | 875 | 911 (16,94 %) | 502 | 1330 | 798 | 532 ( 9,89 %) |
| 5. | 5155 | 4080 | 1075 | 1102 | 1702 | 978 | 724 (14,04 %) | 445 | 1181 | 801 | 380 ( 7,37 %) |
| 6. | 4891 | 4266 | 625 | 1012 | 1556 | 987 | 569 (11,63 %) | 417 | 1070 | 821 | 249 ( 5,09 %) |
| 7. | 6424 | 5622 | 802 | 1415 | 2018 | 1226 | 792 (12,32 %) | 480 | 1342 | 1020 | 322 ( 5,01 %) |
| 8. | 7499 | 6587 | 912 | 1770 | 2211 | 1364 | 847 (11,29 %) | 577 | 1468 | 1056 | 412 ( 5,49 %) |
| 9. | 7253 | 6352 | 901 | 1663 | 2189 | 1373 | 816 (11,25 %) | 562 | 1528 | 1143 | 385 ( 5,30 %) |
| 10. | 7865 | 7015 | 850 | 1918 | 2275 | 1538 | 737 ( 9,37 %) | 576 | 1576 | 1252 | 324 ( 4,11 %) |

*Nota.* – I numeri in parentesi, alle colonne 7 e 11, esprimono il rapporto percentuale delle forme nuove e dei lemmi nuovi di ciascun capitolo rispetto al totale delle occorrenze del capitolo stesso (indicato nella colonna 1).

Nell'insieme, su un totale di 61.414 parole presenti nei primi dieci capitoli del romanzo, le occorrenze di forme omografe sono risultate essere 13.164.

Per ciò che riguarda il risparmio di tempo, e quindi di energie, nel lavoro umano, esso apparirà evidente quando si dica che, mentre la lemmatizzazione manuale del primo capitolo ha richiesto otto ore, il secondo ne ha richiesto soltanto quattro, e queste si sono successivamente ridotte a una media di due–tre ore per capitolo. Anche sotto questo aspetto, dunque, il risultato dell'esperimento deve ritenersi positivo.

Sul piano dell'analisi linguistica, il rilievo più importante riguarda il rapporto fra il totale delle parole presenti in ciascun capitolo (vedi colonna 1 della tabella), le forme nuove, che in esso compaiono per la prima volta (colonna 7), e le forme già incontrate nei capitoli precedenti (colonna 6); e, passando dalle forme ai lemmi, il rapporto fra lemmi nuovi e lemmi già incontrati precedentemente (vedi, per questi, le colonne 9, 10, 11 della tabella). Questi dati comparativi consentono una prima importante osservazione: mentre il numero delle forme nuove varia notevolmente di capitolo in capitolo pur mantenendosi pressappoco costante entro tali limiti di variazione, il numero dei lemmi nuovi tende invece a decrescere, e continuerebbe a decrescere ancora e ad assottigliarsi, quando l'analisi fosse estesa ai capitoli successivi al decimo. La statistica risulta anche più persuasiva se i dati numerici relativi alle forme nuove e ai lemmi nuovi si traducono in misura percentuale rispetto al totale delle occorrenze di ciascun capitolo (come appunto abbiamo fatto nelle colonne 7 e 11).

* * *

Per quanto piena e incondizionata possa essere la nostra fiducia nelle possibilità degli elaboratori elettronici, e nell'ampiezza delle operazioni che essi possono compiere, è difficile non essere alquanto scettici quando si tratti di analisi s e m a n t i c a, la quale appartiene al dominio dei concetti più che delle forme, e richiede perciò raziocinio, intelligenza, sensibilità, doti di cui la macchina ovviamente non è dotata.

Per la distinzione degli omografi, bisogna necessariamente passare attraverso l'analisi semantica, ed è per questo che noi, sostenitori dell'automazione fino alla fase delle concordanze di forme, ci sentiamo meno sicuri a sostenerne l'uso anche per le operazioni di lemmatizzazione, che preferiamo continuare a compiere, almeno in parte, con i metodi tradizionali.

Si deve tuttavia riconoscere che certe analisi semantiche, e quindi certe distinzioni di omografi, possono essere effettuate anche mediante il calcolatore, quando all'individuazione di un significato sia possibile giungere attraverso l'analisi di unità grafiche, cioè di associazioni di parole e di connessioni sintattiche o grammaticali. La macchina, per esempio, è in grado di riconoscere il sostantivo plurale *dèi* e distinguerlo dalla preposizione articolata *déi* (nei casi, s'intende, in cui non siano graficamente distinti dall'accento timbrico, che sarebbe già un dato formale più che sufficiente per la discriminazione

automatica delle due funzioni), tutte le volte che lo incontra preceduto da *gli*, *agli*, *degli*, e simili, una volta che questa istruzione le sia stata impartita.

La macchina sarà anche capace di cogliere nelle espressioni *sei uomini*, *sei giorni*, *sei belle ciliegie* la presenza del numerale 6, e invece nei nessi *sei stanco*, *sei un brav'uomo* la presenza del verso *essere*, una volta che le sia stato insegnato che il numerale *sei* è normalmente seguito da nomi o aggettivi plurali, e al contrario il verbo *sei* da nomi, pronomi o aggettivi singolari. Ma se la sua capacità di distinguere il plurale dal singolare fosse fondato sulla natura della vocale finale, incontrando *sei radio* o *sei auto* ci darebbe risposta sbagliata, interpretando *radio* e *auto*, per la loro terminazione in *-o*, come singolari e qualificando di conseguenza *sei* come verbo.

A un calcolatore l'analisi sintattica è sufficiente per stabilire i due diversi valori degli omografi *desti* (plurale dell'aggettivo *desto*, voce del verbo *destare*, e passato remoto del verbo *dare*) in due frasi come «i bambini *sono desti*» e «il denaro che *desti a lui*»; ma in espressioni quali *non desti*, *mi desti*, *li desti*, *appena desti* dovrà sospendere il giudizio, in quanto sono tutte espressioni polivalenti, non univoche.

Per un'alta percentuale di ricòrrenze, la presenza e la posizione dell'articolo *il* o *le* e della preposizione *di* possono essere elementi sufficienti a far decidere al calcolatore con criteri formali di scelta se *sale* sia un sostantivo femminile plurale (*le sale del palazzo*), o un sostantivo maschile singolare (*versare il sale*; *sale di magnesio*), o una voce del verbo *salire* (*sale il monte*, *sale le scale*). Ma quale confusione di attribuzioni farà incontrando espressioni come *sale di corsa*, *le sale la mosca al naso* e altre frasi trabocchetto?

Molte analisi formali potrebbero essere insegnate alla macchina, anche assai più complicate e ardue di quelle che abbiamo ora esposto. Ma un programma che si proponesse di prevedere e risolvere in analisi formali tutte le possibili omografie e polisemie di una lingua, esigerebbe un impegno di spese, di tempo, e di energie intellettuali enormemente più grande di quello richiesto dalla lemmatizzazione manuale, sia pure ripetuta opera per opera. Questo, almeno per ciò che concerne analisi e ricerche aventi per fine la documentazione lessicografica. Nell'àmbito della traduzione automatica può darsi che le prospettive e i risultati già raggiunti siano diversi, anche perché le lingue prese in considerazione per la traduzione automatica sono, di solito, lingue circoscritte a un'epoca determinata, l'attuale, costituenti quindi un sistema ben definito.

Meno pessimistiche sono le deduzioni che possiamo trarre dall'esperimento di lemmatizzazione semiautomatica condotto sui *Promessi sposi*. Esclusi gli omografi, l'intervento del calcolatore sulle forme univoche può essere di effettivo e notevole aiuto, consentendo un fortissimo risparmio di tempo, di denaro e di fatica. A una condizione, tuttavia: che la lingua sulla quale l'analisi viene operata automaticamente sia, se non la lingua di un singolo autore o addirittura di una determinata singola opera (come nel caso dei *Promessi sposi*), una lingua appartenente a una sezione temporale e areale il più possibile limitata, una lingua cioè sincronica e sintopica. Altrimenti, quando anche si

fossero accuratamente separate tutte le forme univoche dalle forme omografe, per consentire al calcolatore di intervenire automaticamente nella lemmatizzazione delle prime trascurando le seconde, nessuno potrà mai dirsi sicuro che, nella gran massa di testi da analizzare, appartenenti a una lingua che conosce un'evoluzione diacronica di otto secoli e una dimensione diatopica estesa a tutte le regioni d'Italia, non s'incontrino forme arcaiche, poetiche, dialettali, e varianti individuali o arbitrarie o comunque non prevedibili, che facciano sorgere l'omografia anche là dove non era sospettata, e venga così messa in serio pericolo la solidità di una struttura tanto minuziosamente e laboriosamente congegnata.

**II. Problemi tecnici (con speciale riferimento al «dizionario di macchina»).**

Per ottenere che il testo sia leggibile dal calcolatore è necessario riprodurlo su schede o su nastro perforato servendosi di una macchina perforatrice.

Solitamente vengono riportati, oltre le parole e la punteggiatura, gli eventuali segni propri dell'edizione critica, informazioni sulla struttura grafica (divisione in pagine e in righe; maiuscolo, minuscolo, corsivo, ecc.), la suddivisione del testo nelle parti (capitoli, strofe, paragrafi, ecc.) utilizzabili come riferimento; spesso vengono aggiunte informazioni non presenti a livello grafemico nel testo stampato ma inserite nella fase cosiddetta di «preedizione»: ad esempio frasi che sono citazioni, o brani interpolati da omettere.

Di norma il numero di caratteri diversi da rappresentare oltrepassa il centinaio, mentre le macchine perforatrici in uso oggi offrono solo 48 o 64 segni diversi; si deve pertanto mettere in atto una particolare tecnica di codificazione. Le soluzioni più comuni consistono nel rappresentare un carattere stampato con una sequenza di due o più perforazioni, oppure nell'adottare l'equivalente funzionale del tasto delle maiuscole nella macchina da scrivere.

Tra i 64 codici disponibili se ne scelgono alcuni con funzione di « chiave », senza significato proprio. Ciascuno dei codici restanti assume, quando viene letto dal calcolatore, un significato diverso a seconda dell'ultima chiave precedente. È chiaro che se le chiavi sono $n$, i caratteri rappresentabili complessivamente sono $n(64-n)$.

L'operazione di ricopiatura del testo su schede, con tutti i controlli inevitabili, è costosa e lunga a confronto con le operazioni successive del calcolatore. Di qui l'esigenza (sottolineata anche di recente al Congresso di Grenoble, 23–25 Agosto 1967 « International Conference on computational Linguistics ») che i diversi Centri operanti in questo settore adottino un metodo comune nel perforare i testi, in modo che l'insieme di perforazioni praticate secondo un sistema standard garantisca l'utilizzazione del testo perforato, o almeno del nastro magnetico corrispondente, per altre ricerche, anche non previste nel programma iniziale. In questo senso già si opera al Centro Studi IBM

di Pisa, dove il sistema di perforazione messo a punto per l'Accademia della Crusca viene normalmente impiegato per gli altri progetti linguistici in corso.

Il problema della sperequazione tra segni disponibili e segni da rappresentare si pone anche per la stampa dei risultati delle elaborazioni.

Al Centro Nazionale Universitario di Calcolo Elettronico di Pisa lo si sta risolvendo con il dotare la macchina stampatrice veloce IBM di catene di stampa fornite di un massimo di 120 caratteri e intercambiabili facilmente con altre catene per lingue scritte con alfabeto non latino. Nel frattempo, si codifica anche in sede di stampa, con delle « chiavi » sottoscritte ai caratteri usati non univocamente.

Il calcolatore « legge » le schede o il nastro perforato e registra il testo su nastro magnetico, scomponendolo nelle diverse unità elementari di elaborazione (nel nostro caso le parole; in altri lavori i grafemi, i fonemi, le sillabe, i periodi, i sintagmi, ecc.) e ripete per ciascuna unità informazioni che la riguardano, anche se nel testo stampato e perforato occorrono una sola volta, quali numero di pagina e di riga, riferimento, qualifiche dell'apparato critico, ecc. A questo punto può avere inizio il ciclo delle elaborazioni vere e proprie la cui velocità è molto superiore rispetto alle operazioni precedenti: per perforare un testo di mezzo milione di parole (circa 60.000 righe) occorrono 350 ore–uomo circa; per conteggiare le frequenze con il calcolatore bastano due ore.

I problemi posti dalle elaborazioni sono in primo luogo problemi di tecnica di programmazione (indici alfabetici diretti e inversi, rimari, frequenze varie, ecc.), che tuttavia non possono prescindere da considerazioni di carattere più strettamente linguistico. Questo si verifica già nella costruzione delle concordanze.

I criteri di delimitazione del contesto vanno ormai uniformandosi presso i vari Centri e praticamente possono essere raggruppati nei tipi seguenti:

1. la parola esponente è sempre al centro del suo contesto: ossia, è sempre preceduta e seguita da un egual numero di battute o di parole;

2. il contesto è scelto in base alla natura della parola: per le parole grammaticali è un trinomio, per le preposizioni sono prese solo le due parole seguenti, ecc.;

3. il contesto è costituito da una intera unità di riferimento: il verso, il paragrafo, il versetto, il comma, ecc.;

4. i limiti del contesto sono segnati in fase di preedizione: in altre parole il testo viene suddiviso in pericopi, ciascuna delle quali funge da contesto per tutte le parole che la compongono;

5. il contesto è regolato tenendo conto di determinati segni quali l'interpunzione, il cambio di riferimento, ecc.

L'insieme di operazioni comunemente raggruppate sotto il termine « lemmatizzazione » richiede una serie di interventi umani che spezzano il ritmo delle elaborazioni interamente automatiche con le quali dal testo perforato si producono i prospetti statistici, le concordanze, le schede–contesto e in generale la stampa dei risultati finali.

In una procedura di lemmatizzazione interamente artigianale il lemma va scritto dapprima accanto a ciascuna delle parole del testo disposte alfabeticamente, e successivamente va perforato su schede, unitamente a un numero di codice per mezzo del quale il calcolatore associa il lemma a ciascuna delle occorrenze nel nastro–testo.

La sproporzione esistente tra velocità dell'elaboratore e tempi di intervento umano rende particolarmente delicato il compito di coordinare le fasi del lavoro, per sfruttare al massimo le disponibilità della macchina. Un esempio: nel nostro esperimento sui *Promessi Sposi*, la lemmatizzazione artigianale del primo capitolo ha richiesto otto ore, mentre l'intero ciclo delle operazioni elettroniche su di esso è durato circa 20 minuti. Tutto ciò spiega come gli sforzi dei ricercatori tendano a ridurre sempre più, per mezzo del cosidetto « dizionario di macchina », la necessità di interventi umani. A raggiungere questo obiettivo sono interessati non solo i lessicografi ma anche quanti, per studi di stilistica o di linguistica quantitativa, devono operare su lemmi e non sulle forme intese a livello grafemico, e quanti operano nei diversi settori della linguistica applicata (traduzione meccanica, linguistica automatica).

In termini generali, la struttura di un dizionario di macchina e il meccanismo della consultazione non sono dissimili da quelli di un normale dizionario stampato. Il dizionario di macchina è costituito da una serie di articoli (1) registrati nella memoria centrale o nelle memorie ausiliarie (nastri o dischi) del calcolatore, ciascuno dei quali si compone di due parti: un termine di ingresso (*indice*), che serve per la ricerca, e una serie di informazioni (*funzione*) su questo termine. Dall'altro lato si ha una parola, estratta da un contesto, che deve essere « ricercata » nel vocabolario; le informazioni reperite devono essere associate alla parola quando viene reinserita nel suo contesto.

Che cosa siano le funzioni, risulterà chiaro dai seguenti esempi. Nella traduzione automatica, le informazioni consistono principalmente in:

– un codice sintattico–semantico che specifica le proprietà distribuzionali e semantiche da utilizzare per collegare secondo schemi strutturali definiti la parola in questione con quelle che la attorniano;

– un insieme di istruzioni (o indirizzi di istruzioni) per la soluzione (tentativa) degli eventuali omografi;

– la parola (o il gruppo di parole) che traduce nella lingua di uscita la parola in questione.

Nelle applicazioni lessicografiche, la funzione è costituita principalmente dal lemma (inteso come forma di base cui la parola investigata va assegnata nei conteggi e soprattutto negli ordinamenti alfabetici delle operazioni finali) e da una serie di codici che classificano morfologicamente la parola, oppure la qualificano per un particolare trattamento nelle elaborazioni successive.

(1) O unità lessicali, traducendo in italiano il francese *unité lexicale*; in inglese, *dictionary entrie*.

Le tecniche di consultazione del dizionario di macchina sono riconducibili a due tipi fondamentali:

– il primo, usato soprattutto in sede di traduzione e documentazione automatica, ma anche, recentemente, in lessicografia (ad esempio a Liegi), prevede che le parole da lemmatizzare vengano ricercate nel dizionario di macchina rimanendo, sia in *input* sia in *output*, in ordine di testo;

– il secondo esige invece che le parole del testo siano previamente ordinate alfabeticamente, così come sono ordinate alfabeticamente le forme elencate nel dizionario di macchina.

È evidente che, mentre nel secondo caso la ricerca si riduce a un confronto tra due serie di parole disposte nello stesso ordine, nel primo caso si deve ricorrere a una tecnica più complessa non esistendo alcun rapporto tra l'ordine delle parole nel testo e la successione degli articoli nel dizionario. Per facilitare e velocizzare la ricerca, si mira a comprimere le dimensioni del dizionario, in modo che possa essere contenuto per intero nella memoria centrale, o registrato su memorie periferiche ad accesso rapido casuale (ramac, dispac).

Il metodo più generalmente seguito è quello di registrare non tutte le forme possibili a livello grafemico, ma solo le parole « invariabili » e i temi (2) di quelle « flessibili ». Ogni tema è accompagnato da un codice di « paradigma », e per ogni tipo di paradigma possibile sono registrate tutte le desinenze in forma di tabella.

In programmi molto avanzati, la distinzione non è solo tra tema e desinenza, ma addirittura tra « forma base » e « affissi » (prefissi e suffissi) cosicché la riduzione del numero delle voci di vocabolario è ancora maggiore.

Alcuni dizionari impiegati nella traduzione automatica coprono con circa 20.000 temi alcune centinaia di migliaia di forme grafiche. Ogni parola del testo da lemmatizzare viene innanzitutto « segmentata » dal calcolatore il quale procede tentativamente per scomposizioni successive, fino a che perviene a isolare un tema e una desinenza che oltre a essere previsti dal dizionario, siano qualificati con codici compatibili tra loro. Se il calcolatore esaurisce tutte le possibili scomposizioni senza trovarne una corretta, la parola viene segnalata e l'operatore umano può intervenire o immediatamente a *consolle* o in un secondo tempo per mezzo di schede perforate.

Nella prospettiva invece di una ricerca operata con le parole in ordine alfabetico, non è necessario che il dizionario sia accessibile contemporaneamente in tutte le sue parti; può essere registrato su nastro magnetico e letto in memoria centrale un poco per volta, a mano a mano che avanza il nastro delle parole. Il risparmio di spazio non costituisce più un problema, e le diverse forme flessive possono essere registrate integralmente, tutte.

(2) Il concetto di *tema* è naturalmente adattato alle possibilità operative di riconoscimento da parte della macchina e non coincide perciò esattamente con il valore scientifico del termine.

Il programma da noi usato per il citato esperimento sui *Promessi Sposi* è appunto di questo tipo. Esso richiede in *input* due nastri:

1. il *nastro–dizionario*, contenente
   *a*) le forme certamente univoche, con il relativo lemma,
   *b*) le forme omografe possibili senza lemma;

2. il *nastro delle parole* in ordine alfabetico, ciascuna accompagnata da riferimento e dal relativo contesto.

Ogni parola viene confrontata con il dizionario e, a seconda del risultato del confronto, viene smistata su uno dei 3 nastri di *output* che sono:

1. *Nastro delle parole « nuove »*, cioè non ancora previste nel dizionario. La loro lemmatizzazione viene sospesa. Il calcolatore stampa tutte le occorrenze delle forme nuove, con i relativi contesti, e predispone contemporaneamente le schede cui verrà apposto artigianalmente il lemma. Queste schede, in una fase successiva, permettono di associare il lemma alle parole momentaneamente accantonate, e nello stesso tempo aggiornano il dizionario, in modo che se nella prossima elaborazione si riscontreranno ancora queste forme, esse potranno essere lemmatizzate automaticamente.

2. *Nastro delle parole trovate, univoche.*
La parola è prevista come univoca; le viene associato immediatamente l'unico lemma possibile, e la fase di lemmatizzazione per essa è chiusa.

3. *Nastro delle parole trovate, omografe.*
La parola è prevista, ma il calcolatore non è istruito a scegliere tra i due o più lemmi diversi cui può essere assegnata. La sua lemmatizzazione viene rinviata a una fase successiva, nella quale si stampano i contesti in base ai quali il lemmatizzatore decide a quale lemma attribuire ogni singola occorrenza.

Nel corso dell'esperimento citato, come si è detto, per le forme omografe non era stato previsto nel vocabolario nessun lemma. Sarebbe però possibile, con un programma diverso, inserire nel vocabolario per ogni forma omografa tutti i lemmi prevedibili. Il calcolatore li stamperebbe, allora, ciascuno contraddistinto da un numero progressivo, in capo ai contesti da esaminare; il lemmatizzatore si limiterebbe a scrivere a lato di ciascun contesto il numero codice del relativo lemma.

Si può facilmente dedurre da quanto è stato detto che il vero rischio del metodo sta nella possibilità di inserire nel dizionario come univoca una forma che invece può risultare di fatto omografa. Il calcolatore in tal caso lemmatizzerebbe senz'altro le parole corrispondenti come univoche, senza proporle all'esame del linguista.

Questo rischio, come è già stato sopra accennato, è tanto maggiore quanto più si estende diacronicamente lo strato di lingua esaminato e quanto più si

torna indietro nel tempo verso sistemi non perfettamente presenti alla coscienza linguistica di chi compone il dizionario di macchina.

Si pensa però di superare questo ostacolo, almeno fino a che l'esperienza d'uso del dizionario non ne avrà comprovata la completezza, col considerare provvisoria la lemmatizzazione operata per suo tramite. Basterà stampare anche il nastro delle forme univoche dotandole dei relativi contesti per consentire il controllo dei lemmi assegnati automaticamente. Si può facilmente calcolare che, pur con questa fase di controllo, il risparmio di tempo rispetto a una procedura interamente artigianale sarebbe notevolissimo, del 75% circa.

A favore dell'impiego del dizionario di macchina sta anche la considerazione che esso, evitando gran numero delle scrizioni di lemmi necessarie nella lemmatizzazione artigianale, diminuisce in maniera notevole il rischio di errori casuali di trascrizione, e nel contempo garantisce una maggiore sistematicità e coerenza nei casi, più numerosi di quanto si possa pensare, in cui la formulazione del lemma richiederebbe una discussione o il ricorso a un corpo di regole prefissate; il dizionario funziona come registrazione delle decisioni prese e, potendo essere stampato facilmente, fornisce in ogni momento il quadro completo delle forme esaminate e del trattamento scelto.

Naturalmente l'economia di tempo realizzata con l'uso di un dizionario di macchina è in relazione al numero delle forme univoche incluse. La percentuale di forme omografe possibili dipende dal grado di « profondità » al quale si vuole spingere l'analisi e la distinzione delle omografie e della polisemia. Del resto, non sembra che le categorie tradizionali, allo stato attuale delle conoscenze e soprattutto quando si applichino a grandi quantità di esempi, permettano una risposta coerente e integrale al problema della individuazione della unità lessicale. Pertanto entrano in gioco considerazioni di ordine economico e organizzativo: si tratta di trovare il giusto punto di equilibrio tra le analisi che si devono eseguire nella prima fase di raccolta e di organizzazione automatica della documentazione (nel nostro caso la costituzione dell'Archivio della lingua) e le analisi demandate agli utenti successivi.

Il problema se e come sia possibile operare automaticamente lo scioglimento dell'omografia si è posto, soprattutto sotto l'influsso della linguistica applicata, con soluzioni diverse, in genere fondate sul riconoscimento della struttura sintattica della frase, operato secondo le regole grammaticali proposte dalla scuola strutturalistica (grafi di Tesnière) e dalle diverse grammatiche generative e trasformazionali.

Per le parole omografe che appartengono alla stessa parte del discorso, si ricorre anche a una classificazione semantica del lessico, nella quale sono specificate le possibilità di relazione e di compatibilità tra le diverse categorie.

A un livello più generale si fa intervenire il macrocontesto o, in altre parole, il tema generale del testo, facendo ricorso ai microglossari: quando una parola può avere, oltre ad altre accezioni, un senso tecnico specifico, è molto probabile che sia quest'ultimo a ricorrere in un testo specializzato (per esempio, *fattore* in un libro di matematica molto probabilmente non significherà l'amministratore di una fattoria).

Da più parti si avverte però che i metodi più economici e più sicuri sono quelli che cercano di rilevare il valore semantico e morfosintattico di una parola dal suo contesto immediato, con regole stabilite individualmente per ogni forma omografa, come è stato appunto esemplificato nel capitolo precedente

Il problema della soluzione automatica della omografia e della polisemia è solo parzialmente comune alla lessicografia e alla traduzione automatica. Per esempio, è inutile distinguere nella lingua di partenza i segmenti la cui polisemia è omogenea nella lingua di arrivo, mentre nella lessicografia, in cui punto di arrivo sono le diverse unità lessicali, questa economia non è possibile. Soprattutto, vi è, per così dire, una precedenza, una dipendenza tra i due settori: la linguistica applicata attende dalla lessicografia e dalla linguistica non solo l'inventario completo delle omografie, ma anche la descrizione delle differenze distribuzionali che, formalizzate, dovrebbero permettere l'automazione delle operazioni di discriminazione.

In effetti bisogna riconoscere che gli inventari e le descrizioni esistenti sono incompleti e parziali, e che appunto la raccolta e l'analisi di concordanze stabilite su una grande quantità di testi è la prima tappa indispensabile per il completamento delle conoscenze in questo campo. Ciò non toglie che, sulla base delle conoscenze già acquisite, anche nelle elaborazioni lessicografiche si studi la possibilità di automatizzare l'analisi degli omografi.

Le tecniche per la generazione di un vocabolario di forme attualmente in uso nei Centri lessicografici di cui abbiamo conoscenza si possono ricondurre a due tipi principali.

Alcuni procedono, per così dire, induttivamente: iniziano senza un vocabolario di macchina l'elaborazione della prima opera, e ne lemmatizzano artigianalmente tutte le forme, utilizzate per costituire un primo vocabolario che viene via via arricchito con le « forme nuove » delle opere successive. È appunto questo il metodo da noi seguito nell'esperimento sui primi capitoli dei *Promessi Sposi*: le forme del primo capitolo, lemmatizzate a mano, sono state impiegate per la lemmatizzazione automatica del secondo, e così via. Altri iniziano invece con il vocabolario di macchina già approntato, ottenuto operando tutte le possibili flessioni di un lemmario assunto come base di partenza.

Il lemmario può essere desunto da un solo dizionario, o dalla fusione di più dizionari o con procedimenti ancora più complessi, che garantiscano maggiore completezza. In questa fase preparatoria il calcolatore può essere di aiuto non solo per il confronto e la fusione di dizionari diversi, ma anche per flettere i lemmi automaticamente anziché a mano. Questa operazione che è, per così dire, l'inverso di quella di segmentazione esaminata precedentemente, richiede che al calcolatore vengano comunicate le tabelle con le desinenze dei vari paradigmi, e che i lemmi da flettere siano prima ridotti ciascuno a un tema (o, nel caso di lemmi irregolari, a due o più temi) accompagnato da un codice di rinvio a una determinata tabella. Il calcolatore genera

tutte le flessioni possibili semplicemente giustapponendo al tema, una alla volta, le desinenze previste dalla tabella.

Con tale procedimento si otterrebbe anche un notevole aiuto per la ricerca della omografia possibile. Sarebbe sufficiente infatti disporre col calcolatore le forme in stretto ordine alfabetico e far segnalare le forme ripetute almeno una volta.

Probabilmente i dati e le esperienze fino a qui raccolte non permettono di decidere quali delle due tecniche sia preferibile per la generazione di un dizionario di macchina: la flessione automatica di un lemmario richiede un lungo lavoro di preedizione e di studio, e con ogni probabilità si includono forme che lo appesantiscono e che non si incontreranno di fatto nelle opere spogliate. D'altro lato, più è completo il dizionario, più si riduce il numero delle forme nuove, e quindi la spezzatura del ritmo elaborativo, mentre maggiore è la garanzia di organicità coerenza e sistematicità nella lemmatizzazione. Grande rilievo ha poi il fatto che si operi sincronicamente su una sezione cronologica di lingua ben delimitata, oppure diacronicamente su un lungo periodo.

Nonostante le difficoltà oggettive, ci sembra di poter concludere, alla luce delle diverse esperienze compiute, che è indispensabile continuare le ricerche per l'automazione almeno parziale della lemmatizzazione, se si vuole trarre dall'uso della macchina un aiuto veramente efficace alla fatica lessicografica.