R. BINDI P. ORSOLINI A. ZAMPOLLI

METODOLOGIE PER GLI SPOGLI LESSICALI AUTOMATICI

1. Cenni storici

1.2. Mentre l'esempio del lavoro pioneristico di R. Busa (1951) nel settore degli spogli elettronici (il suo primo esperimento risale al 1949 e nel 1953 egli fondava il CAAL - Centro per l' Automazione dell'Analisi Linguistica - di Gallarate) veniva seguito, a cavallo del 1960, da Centri di molti altri paesi europei, in Italia invece, nel periodo 1950-1965, oltre alla attività del CAAL, si registrano due soli progetti: la "omogeinizzazione degli indici del Frank e dello Spanke" (Roncaglia, Pacifico, 1961), e la tesi di Zampolli (1960), che costituisce il primo esempio di spoglio elettronico di un testo italiano (Veglia d'Armi di Diego Fabbri). Condotto a livello fonematico, morfologico e lessicale, si proponeva essenzialmente indagini di statistica linguistica (Busa et alii 1962, Tagliavini 1968, Zampolli 1968 A).

Per superare lo scetticismo nel nostro paese è stato decisivo l'esempio autorevole della Accademia della Crusca la quale iniziava la serie degli spogli per la costituzione dell'archivio lessicale, previsto per la redazione del Vocabolario storico, presso il CNUCE (Centro Nazionale Universitario di Calcolo Elettronico dell'Università degli Studi di Pisa, oggi Istituto del CNR), che, fondato

nel 1965, aveva celebrato la propria inaugurazione pubblicando gli Indici e le Concordanze della Divina Commedia (Tagliavini 1965). Nello stesso anno, anche il CAAL trasferiva presso il CNUCE le elaborazioni dell'Index Thomisticus (Busa, Zampolli 1968).

Nel 1967 l'Accademia Nazionale dei Lincei inseriva in un Convegno sul tema "L'Automazione elettronica e le sue implicazioni scientifiche, tecniche e sociali" una sezione specializzata (Duro, Zampolli 1968; Tagliavini 1968).

Da quel momento i progetti si moltiplicavano rapidamente (Zampolli 1968B), e nel 1968 erano così numerosi che A. Zampolli poteva costituire presso il CNUCE la Divisione Linguistica (DL).

Il convergere nello stesso luogo delle esperienze condotte fino a quel momento, a Gallarate Padova e Firenze, e la costituzione della DL concludono, per così dire, una prima fase che, nonostante la relativa modestia delle risorse a disposizione, ha delineato alcuni tratti salienti dello sviluppo successivo. Il confronto dei risultati e degli insegnamenti delle esperienze citate, orientava infatti, naturalmente, a studiare fin dall'inizio una metodologia e delle procedure generalizzate, improntate a criteri di generalità, anzichè alla ricerca, come per lo più accadeva in quegli anni, di una procedura ad hoc per il trattamento ottimale di un testo determinato.

Tale generalizzazione, presupponendo necessariamente una prima tipologizzazione dei fatti da rappresentare in 'machine readable form', apriva la strada a un tentativo di normalizzazione.

D'altra parte, la constatazione della complessità - spesso non evidente in prima istanza - degli aspetti organizzativi della interazione tra le analisi manuali e le operazioni automatiche negli spogli di grandi corpora di testi; alcuni esperimenti intesi a misurare le variazioni indotte da diversi procedimenti di lemmatizzazione nella 'performance' dei lemmatizzatori; il desiderio di favorire la coerenza nel lavoro di équipes; la rivalutazione, in quegli anni, del ruolo del lessico nei modelli linguistici; tutti questi fattori indussero a proporre (Zampolli 1968c)la costruzione per l'italiano di strumenti linguistico-computazionali, quale un dizionario di macchina (DMI) e dei parsers, intesi da un lato come la traduzione computazionale dei corrispettivi componenti del sistema linguistico, dall'altro applicabili per aumentare il grado di automazione degli spogli, con la conseguente riduzione dell'onere degli interventi manuali.

1.2. I compiti affidati alla DL erano quelli di "promouvoir et d'effectuer des recherches dans le domaine de la linguistique mathématique et computationelle, d'assurer les relations scientifiques et l'échange d'informations entre les

charcheurs italiens et étrangers, d'organiser et de coordonner l'activité didactique pour diffuser les expériences et les méthodologies et enfin de prêter assistance scientifique et techinique aux Instituts intéressés. A la différence de l'assistance fournie à d'autres catégories d'usagers, qui se limite en général à assurer les instruments et le temps de calcul, l'assistance réservée aux linguistes et aux humanistes comprenait également la consultation et la collaboration pour la planification scientifique et technique des projets, l'analyse et la mise au point des programmes nécessaires et l'exécution des élaborations", (Zampolli, 1973 A, p. XIII).

Le caratteristiche della collaborazione offerta dalla DL facilitarono indubbiamente la diffusione dei progetti di text-processing nel settore linguistico letterario e filologico. Ai ricercatori interessati ad automatizzare lo spoglio venivano infatti evitati la fatica e i costi della analisi e della programmazione delle procedure. Essi trovavano inoltre nelle metodologie e negli standards della DL uno schema di riferimento per la pianificazione dei dati e dei procedimenti, richiesta dall'automazione delle elaborazioni.

Come importante contropartita, la applicazione delle procedure della DL a progetti di nuovo tipo ne provocava

l'arricchimento con nuove funzioni e la estensione a nuove strutture di dati.

Per una descrizione delle procedure si veda Bortolini Tagliavini Zampolli (1971); per una tipologia dei progetti, si rinvia a Zampolli (1973A) e (1973B).

I progetti in questione, impiegando le procedure della DL, ne adottavano naturalmente gli standards per la rappresentazione dei testi in 'machine readable form'.

Questi testi hanno arricchito progressivamente l'archivio di materiali linguistici della DL.

In questo periodo iniziò anche la progettata costruzione del DMI (Zampolli 1968c)presso la cattedra di linguistica matematica dell'Università di Pisa, continuata poi presso la DL.

Nel 1976 iniziarono i primi esperimenti per la applicazione del DMI alla lemmatizzazione di testi, in particolare su un corpus di giornali (Crocetti, 1976) e sulla opera omnia di Montale.

Nel contempo veniva avviata la utilizzazione del DMI per ricerche sul sistema lessicale dell'italiano, in particolare a livello morfologico (Ruimy, Turrini (1979) in Zampolli et alii 1979) e semantico (Calzolari Moretti 1977, e Calzolari, Pecchia, Zampolli, in stampa).

Nel periodo in questione, vennero organizzate a Pisa

anche le prime 4 edizioni della Scuola Estiva Internazionale 'Mathematical and Computational Linguistics', nelle quali furono trattati i temi seguenti: introduzione alla linguistica computazionale e alla stilostatistica (1970; Zampolli 1973); Metodi elettronici in lessicologia e lessicografia (1972; articoli in "Cahiers de Lexicologie", nn. 28-31); Metodi per la descrizione formale della sintassi e della semantica delle lingue naturali (1974; Zampolli 1977); 'natural language understanding' (1976; in stampa).

La Scuola Estiva, radunando a Pisa studiosi di varie tendenze, ha contribuito, per giudizio unanime, allo sviluppo nel nostro paese non solo della linguistica matematica e computazionale, ma anche delle correnti più recenti della linguistica generale.

La DL ha ricevuto dalla Scuola Estiva lo stimolo per nuovi progetti. Nel campo della sintassi dell'italiano, è stata avviata la costruzione di un analizzatore fondato sull'ATN di Woods (1976), che si propone da un lato come strumento e occasione per ricerche teoriche nello ambito dei modelli sintattici, della psicolinguistica e della teoria della computabilità, dall'altro si propone come componente in sistemi applicativi per il trattamento automatico di frasi dell'italiano

Il progetto per la costruzione di un sistema di lemmatizzazione semiautomatica dello spagnolo, stimolato dalle richieste di collaborazione di ricercatori ispanoamericani presenti alla Scuola, che dovrebbe essere inserito in procedure di spoglio come aiuto per i lemmatizzatori, si propone anche di verificare limiti e possibilità di un analizzatore morfosintattico per la disambiguazione degli omografi attraverso il riconoscimento di cooccorrenze particolari anzichè attraverso la assegnazione di un indicatore sintagmatico all'intera frase (G.Cappelli et alii, 1979).

L'insieme di relazioni internazionali stabilite anche in occasione della Scuola Estiva e di altri Convegni organizzati a Pisa (Calzolari-Zampolli 1977), favorì le attività del progetto 'Banca Internazionale di Dati Linguistici', il cui scopo principale è quello di produrre gli strumenti organizzativi e tecnici necessari per acquisire testi registrati in 'machine readable form' presso Centri di altri paesi, e inserirli nella nostra nastroteca, rendendoli così disponibili ai ricercatori italiani.

1.3. Lo sviluppo e l'arricchimento delle attività della DL diversificarono progressivamente e le sue esigenze organizzative e il tipo di risorse necessarie, rispetto al

resto del CNUCE, che nel frattempo aveva conosciuto una rapida espansione.

Verso il 1975-76 divenne evidente per gli organi direttivi del CNUCE che alla autonomia scientifica della DL era necessario far corrispondere l'autonomia organizzativa e amministrativa.

Alla fine di un lungo periodo di incertezza, nel giugno 1978, fu decisa la costituzione del Laboratorio di Linguistica Computazionale, ora denominato Istituto (ILC) afferente al Comitato 08.

Il processo di costituzione del nuovo organo, avviato in pratica nella primavera del '79, è stato lungo e faticoso, anche per il concomitante riordinamento generale del CNR e per il temporaneo blocco dei concorsi. Ora sembra prossimo il momento in cui l'Istituto potrà disporre delle risorse e delle strutture necessarie per la piena funzionalità. Il problema più importante è certamente quello del completamento delle lacune dell'organico.

Molto comunque è stato già fatto, sul piano dello studio e, parzialmente, della acquisizione delle attrezzature.

In particolare, è stato costituito il reparto di elaborazione, che ha la responsabilità della pianificazione e della esecuzione delle procedure richieste dai diversi progetti, soprattutto quelli in collaborazione con altri Istituti.

2. Procedure di spoglio

Le procedure in questione sono state messe a punto, in una prima versione per l'IBM 1401, da A. Zampolli per gli spogli dell'Index Thomisticus e dell'Opera del Vocabolario dell'Accademia della Crusca. Successivamente esse sono state implementate sui calcolatori del CNUCE (IBM 360/30 e, successivamente, IBM 370/158 e 370/168) dagli analisti-programmatori della DL. Il sistema è stato arricchito progressivamente sulla base delle esperienze condotte in collaborazione con i numerosi Istituti italiani e stranieri che lo hanno via via utilizzato.

2.1. La possibilità di applicare questo nostro sistema di text-processing a lingue, epoche e generi diversi presuppone uno standard per rappresentare in 'Machine readable form' la straordinaria varietà di 'grafemi' che possono apparire nei testi in linguaggio naturale. (Con il termine 'grafemi' indichiamo qui non solo i caratteri alfanumerici stampati nel testo, ma anche informazioni fornite dal testo in altra forma - spaziature, suddivisioni, ecc. o inserite nel testo in fase di preedizione: per es., contrassegni di parti non autentiche, di citazioni di altri autori, ecc.).

Questo standard consiste essenzialmente in un inventario di grafemi intesi nel senso predetto, e in una serie di tabelle che specificano la rappresentazione di questi grafemi nelle diverse fasi dello spoglio (su schede perforate, su nastro, in lemmatizzazione, in stampa). Si potrebbe dire che lo 'standard' funziona come un 'questionario' al quale ogni ricercatore che lo usi deve rispondere chiedendosi se un certo 'fenomeno' o 'grafema' sia presente o no nel testo che egli si accinge a spogliare, e che quindi lo guida passo passo alle analisi delle informazioni da rappresentare in input e alla definizione delle loro funzioni nelle elaborazioni successive.

La presenza di questo standard garantisce che un testo venga memorizzato in 'machine readable form' con tutte le informazioni pertinenti per le elaborazioni generalmente richieste dalla comunità degli studiosi. Il fatto che esso venga utilizzato in pratica da tutti i progetti di text-processing' nel nostro paese nel settore della ricerche linguistiche e filologiche, garantisce la reciproca scambiabilità dei testi registrati in 'machine readable form' da ricercatori diversi per scopi diversi, testi che possono confluire così a costituire corpora tecnicamente omogenei sempre più rappresentativi, controllabili, e strutturati lungo le diverse dimensioni (storiche, geografiche, di genere letterario, ecc.) di una lingua.

La 'biblioteca elettronica' conservata su nastro magnetico presso l'LC è tra le più estese oggi esitenti. Essa costituisce il primo passo verso la costituzione di una vera e propria banca di dati linguistici. Comprende oltre 5.000 testi, in più di 30 lingue, registrati con uniformità di criteri tecnici e scientifici, ed elaborabili perciò tutti con gli stessi programmi fondamentali che costituiscono le nostre procedure. Ogni qualvolta un nuovo programma viene scritto per una nuova elaborazione richiesta da un progetto specifico, esso è immediatamente applicabile a tutti i testi dell'archivio. Naturalmente, è necessario una operazione preliminare di preedizione, con la quale si decidono, di volta in volta, le modalità di applicazione dei nostri standards di rappresentazione ai testi da sottoporre a spoglio, e si introducono in essi eventuali informazioni complementari rilevanti per le elaborazioni successive (1).

2.2. Il diagramma operativo (flow-chart) che riportiamo, suddiviso per comodità di consultazione in 4 tavole diverse, alle pagine 42-45, rappresenta il flusso cronologico e le connessioni logiche delle diverse operazioni. Ogni singola operazione è contrassegnata con un numero. La flow-chart è riportata nell'appendice 1.

Op.- 1 - Per ottenere che il testo possa essere letto dal calcolatore, si suole riprodurlo su schede o su nastri per mezzo di una macchina perforatrice. Spesso ci si serve ancora di schede, nelle quali si riportano le parole, la punteggiatura, la divisione sia secondo il riferimento organico (versi, capitoli, canti, ecc.) sia per pagine e righe ('schede-testo')(2).

Op. 2 - Le schede-testo perforate vengono introdotte in una macchina detta 'verificatrice', sulla cui tastiera il testo viene nuovamente ribattuto, per opera di persona diversa da quella che ha eseguito la perforazione. Se v'è una discordanza la macchina si blocca e l'operatore controlla se è vouta a un errore proprio o a un errore di chi l'ha preceduto nel perforare la prima volta il testo.

Op. 3 - Il calcolatore 'legge' le schede, cioè traduce i fori di ciascuna di esse negli elementi corrispondenti del proprio linguaggio interno, scomponendo il testo nelle singole parole (definite come sequenze continue di lettere o altri segni equivalenti comprese fra due spazi o segni d'interpunzione), che nel nostro caso sono le unità elementari dell'elaborazione (3). Via via che

legge il testo e lo scompone, il calcolatore lo registra parola per parola su un nastro magnetico (il cosiddetto 'nastro-parola'), dando a ciascuna delle parole un numero progressivo che la individua in modo univoco. Contemporaneamente, per mezzo di una macchina stampatrice che gli è collegata, stampa la cosiddetta 'lista-testo', che non è poi altro che il testo originale, così come è stato perforato nelle schede, riprodotto con i caratteri di cui la stampante dispone.

Op.4 - La lista-testo prodotta dal calcolatore viene collazionata con il testo originale (4). Gli errori trovati vengono messi in evidenza in fogli particolari, con un modulo opportunamente predisposto, nei quali si riporta il numero progressivo che individua la parola o, più in generale, l'informazione errata.

Op.5 - L'operatore, ricopiando esattamente il modulo, perfora in forma corretta, su apposite schede di rettifica, le parole e le informazioni che in un primo tempo erano state sbagliate (5).

Op.6 - Il calcolatore ricopia il nastro-parola correggendone gli errori e contemporaneamente stampa una nuova edizione della lista-testo.

Op.7 - Il calcolatore ricopia le parole del testo corretto, aggiungendo a ciascuna una porzione del contesto in cui si trova, delimitata secondo convenzioni che il ricercatore ha fissato scegliendo fra le possibilità che un programma flessibile di concordanze, quale il programma tipo dell'ILC, mette a sua disposizione. Questo programma permette di fissare di volta in volta il numero massimo di caratteri che possono comporre il contesto. Due esigenze contrastanti si fanno spesso sentire nella scelta di questo numero. Da un lato si sarebbe portati a richiedere un contesto molto ampio, per rendere l'esempio più 'significativo'(6). Da un altro lato ci si preoccupa di contenerlo entro limiti di lunghezza ragionevoli, sia per renderne più veloce la lettura sia per rendere più maneggevoli le concordanze (7). Sulla scorta di precedenti esperienze si adotta, di regola, un contesto limitato a un numero massimo di circa 120 caratteri, che equivale in media a 12-15 parole (8).

Il programma tiene conto di diversi fattori (per es. la punteggiatura, la fine o l'inizio di un capitolo, ecc.) nel delimitare il contesto. Per questo motivo la parola di cui si dà il contesto non è sempre al centro del rigo, ma la sua posizione è condizionata dalla

presenza, alla sua destra o alla sua sinistra, di tali fattori.

E' possibile anche specificare quali elementi del testo vadano 'contestualizzati' e quali no, ed elencare e classificare gli elementi che hanno la funzione di 'limiti' di contesto.

Op.8 - Le parole con i relativi contesti vengono ordinate alfabeticamente.

Op.9 - Il calcolatore stampa la lista delle concordanze per forma. Contemporaneamente ricopia, su un nastro, forme e contesti numerati progressivamente e produce per ogni forma una scheda che contiene la forma stessa e il numero progressivo corrispondente.

A questo punto ha luogo di solito la lemmatizzazione, la quale richiede una serie di interventi umani che, spezzando il ritmo delle elaborazioni interamente automatiche dello spoglio, aumenta considerevolmente il tempo, i costi, e i rischi di errore.

Molti ricercatori, in particolare anglosassoni e statunitensi, decidono di non lemmatizzare affatto i testi, e si limitano a produrre degli indici, delle concordanze, o delle schede-contesto nei quali gli esponenti non sono unità definite secondo criteri linguistici, ma semplici

forme grafiche, e cioè 'parole' come le riconosce abitualmente il calcolatore: sequenze di lettere tra due spazi o tra due separatori in genere. Indubbiamente questa semplificazione ha qualche vantaggio. La velocità del calcolatore nell'operare su simboli viene sfruttata completamente e si possono produrre rapidamente e a costi minori grandi quantità di spogli che, se diffusi, rendono innegabili servigi agli studiosi. Alle volte ci sono ragioni scientifiche che sconsigliano la lemmatizzazione, per esempio nel caso di testi in lingue o strati di lingua poco noti, nei quali molti lemmi non sarebbero attestati o addirittura neppure ricostruibili.

Tuttavia, molto spesso, in particolare negli spogli di grandi corpora per la redazione di ampi dizionari storici, una qualche analisi e classificazione dei materiali lessicali sembra indispensabile prima della conclusione degli spogli, per evitare che i redattori del dizionario restino sommersi dalla quantità di dati da scegliere e da ordinare. E' inevitabile chiedersi se il calcolatore, per aiutare effettivamente il lessicografo, non debba affiancarlo anche, e soprattutto, nella fase di classificazione dei dati lessicali che il calcolatore raccoglie in proporzioni non commesurabili alle possibilità umane di elaborazione.

Il cosiddetto dizionario di macchina, o lessico automatico (LA), fornisce una prima risposta a questa esigenza.

Descriveremo qui di seguito le operazioni così come vengono compiute nel caso non sia disponibile o non sia per qualche motivo consigliabile la adozione di un vocabolario-macchina o lessico automatico (LA).

Nel paragrafo 2.2. descriveremo invece rapidamente le operazioni della nostra procedura di lemmatizzazione semiautomatica con l'uso di un LA, osservando che le operazioni in tale paragrafo descritte sostituiscono le operazioni 10-12 della procedura la cui illustrazione continuiamo qui di seguito.

Op.10 - Le indicazioni dei lemmi e le eventuali indicazioni accessorie vengono scritte a mano sulla lista delle concordanze accanto a ciascuna forma.

Op.11 - Si perforano queste indicazioni sulle 'schede-forma' preparate dal calcolatore.

Op.12 - Il nastro delle concordanze per forma viene ricopiato con l'aggiunta, per ogni parola, del lemma perforato nella scheda relativa.

Op.13 - Si ordinano le parole per lemma, forma, riferimento.

Op.14 - Si stampano le concordanze per lemma.

Op.15 - Sulle concordanze lemmatizzate il ricercatore esegue alcuni controlli per accertare l'esattezza della lemmatizzazione.

Op. 16 - Dal nastro delle concordanze, già ordinate per lemma, si ricava un nastro contenente solo i lemmi in ordine alfabetico, con le relative frequenze; contemporaneamente si stampa l'elenco alfabetico dei lemmi.

Op. 17 - Il nastro dei lemmi viene riordinato secondo l'ordine decrescente delle frequenze.

Op. 18 - Si stampano i lemmi in ordine di frequenza.

Op. 19 - Dalle concordanze dei lemmi si stralciano e stampano elenchi di lemmi particolari (nomi propri, cose notevoli e altre classi di lemmi che il ricercatore abbia ritenuto opportuno di segnalare).

Op. 20 - Le parole vengono ridisposte in un nuovo nastro in ordine alfabetico di forma e, in casi di omografia, di lemma.

Op. 21 - Da questo nastro se ne ricava un altro che contiene l'elenco alfabetico delle forme lemmatizzate con le frequenze di ciascuna, e contemporaneamente si stampa l'elenco delle forme in ordine alfabetico.

Op. 22 - Viene generato un nuovo nastro nel quale le forme sono ordinate secondo l'ordine decrescente delle rispettive frequenze.

Op. 23 - Si ottiene l'elenco delle forme in ordine di frequenza.

2.2. Lemmatizzazione semiautomatica

Le operazioni qui descritte sostituiscono, come si è detto, le operazioni 10-12 della procedura precedentemente descritta.

1.2.3.1. Il lessico automatico

Per maggiore chiarezza riassumo rapidissimamente cosa si intende con i termini LA e consultazione di un LA. Mi riferirò costantemente, per brevità di esposizione, alla più semplice tra le diverse possibili strutture di un LA. Altrove si è descritta una organizzazione più complessa . Nella sua forma più semplice un LA consiste in una serie di forme grafiche registrate su nastro, su disco, o su altro supporto leggibile dal calcolatore. Queste forme sono ordinate alfabeticamente. Ogni forma è accompagnata da una serie di informazioni linguistiche di natura diversa a seconda dei diversi impieghi cui il LA è destinato. Chiamiamo analisi o funzione di una forma l'insieme delle informazioni che la accompagnano : per esempio se il LA è compilato come ausilio agli spogli lessicali, per ogni forma saranno dati il lemma cui la forma deve essere assegnata e, spesso, la clas-

sificazione grammaticale e morfologica del lemma e della forma. Se il LA è compilato per tradurre automaticamente da una lingua all'altra, saranno date anche la forma corrispondente nella lingua di uscita, alcune indicazioni per l'analisi sintattica e semantica della proposizione, ecc. Se il LA serve per statistiche fonematiche, etimologiche, sociolinguistiche, saranno dati anche la trascrizione fonematica della forma, la sua etimologia, i suoi registri d'uso, ecc.

A seconda degli scopi cui il LA è destinato può variare sensibilmente anche il numero delle forme che lo compongono. Per esempio se il LA serve per tradurre testi scientifici relativi a una disciplina sepcifica, esso contiene di solito solo le voci più frequenti nella lingua comune e i termini tecnici della disciplina in questione. Se invece il LA serve per statistiche sull'intero sistema lessicale, l'insieme delle voci che compongono il LA è molto più ricco. Il LA si propone in tal caso come un sottoinsieme rappresentativo dell'intero lessico di una lingua, e al limite vorrebbe coincidere con esso.

E' importantissima, dal punto di vista applicativo, la distinzione tra forme univoche e forme omografe. Diremo univoca una forma alla quale, nel LA, corrispondono due o più analisi distinte. Per esempio se in un dato LA la analisi delle forme è rappresentata solo dal lemma, una for-

ma come l'italiano dica sarà, in tale LA, univoca: essa può appartenere infatti solo al lemma dire.
Sarà invece omografa nello stesso LA, la forma amo, che può appartenere sia al sostantivo amo sia al verbo amare.
Consideriamo invece un LA nel quale l'analisi assegnata a ciascuna forma comprenda, oltre al lemma, anche la classificazione morfologica della forma (genere, numero, grado per le forme nominali; modo, tempo, persona per i verbi, ecc.). In un tale LA sarà omografa anche la forma dica, cui corrisponderanno 4 distinte analisi: rispettivamente 3° persona singolare all'imperativo, 1ª, 2ª, e 3ª persona singolare del congiuntivo presente. Se invece le analisi di un LA comprendessero solo la trascrizione in alfabeto fonetico delle forme, in tale LA forme come amo e dica sarebbero univoche, perchè vanno in ogni caso trascritte rispettivamente come /amo/e/dika/, e sarebbero omografe solo le forme non omofone in italiano, come ancora (/ankóra/- /ánkora/) e pèsca (/pèska/-/péska/). Come esempio possiamo riferirci al Lessico Automatico Italiano (LAI) che stiamo predisponendo all'ILC sotto gli auspici del Comitato 08 del CNR.

Abbiamo registrato su nastro magnetico circa 150.000 lemmi, ottenuti dall'unione delle nomenclature dei principali dizionari. Ad ogni lemma sono state assegnate informazioni riguardanti l'etimologia, la funzione grammaticale, la

polisemia, la scomposizione in morfemi, gli eventuali registri di uso, le costruzioni sintattiche. Queste informazioni sono espresse in alcuni casi con un linguaggio completamente formalizzato (per esempio le costruzioni), mentre in altri casi la definizione è costituita da una parte formalizzata e da un'altra in linguaggio naturale, come nella esplicitazione delle polisemie.

Un algoritmo di flessione ha generato automaticamente circa 1.000.000 di forme, che sono registrate sia secondo l'ortografia corrente sia secondo l'alfabeto fonetico. Si veda in proposito Zampolli (1973 b).

1.2.3.2. Consultazione di un LA

L'algoritmo di consultazione di un LA così organizzato è estremamente semplice. Come si vede nell'organigramma di Fig. 1, in ingresso vengono posti il LA (1) e il nastro delle concordanze per forma (2).

L'algoritmo di consultazione legge una sola parola alla volta, dal nastro 2, e la confronta con le forme grafiche che compongono il LA. Si possono verificare 3 diverse condizioni.

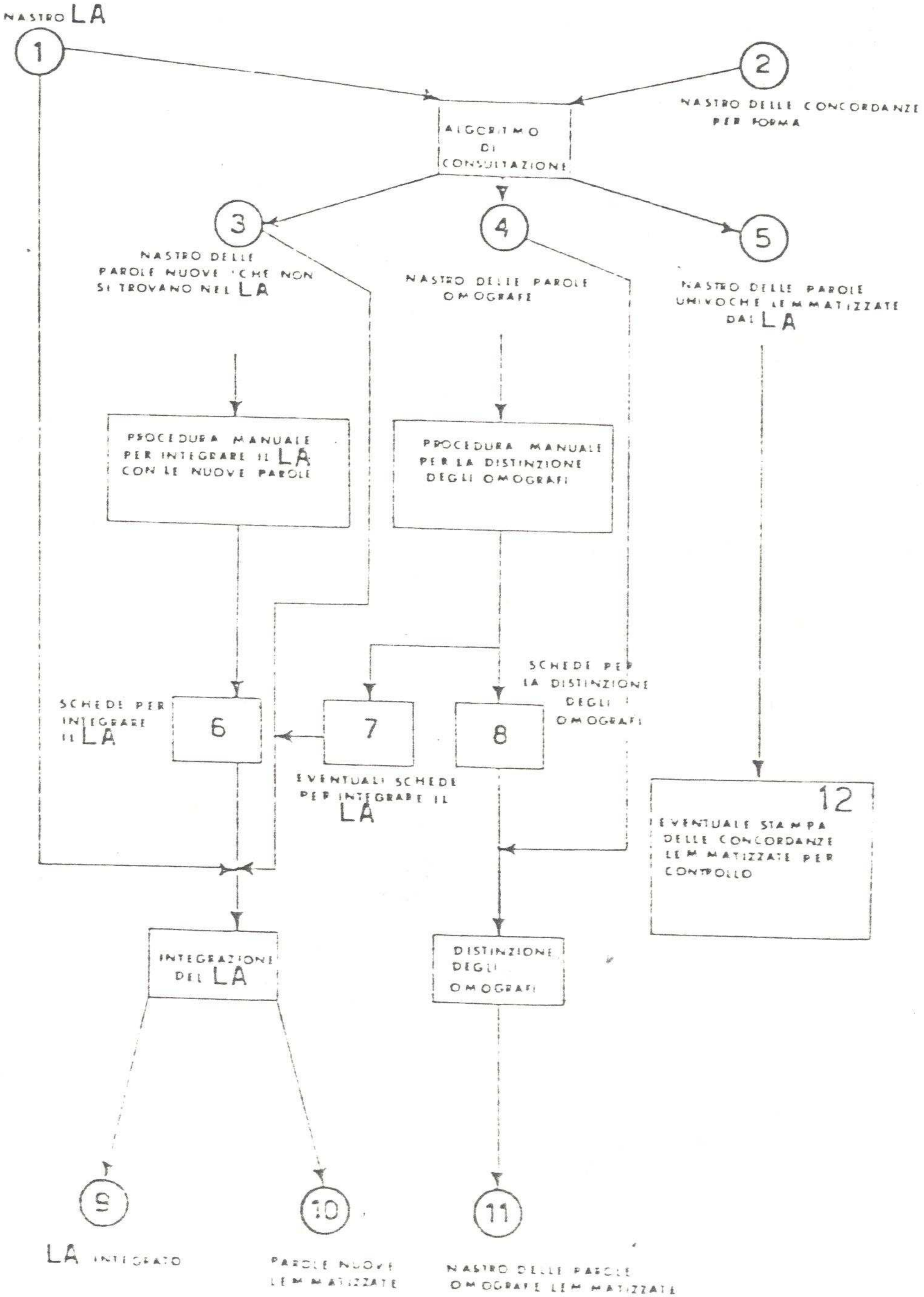
NASTRO LA
1
2
NASTRO DELLE CONCORDANZE PER FORMA
ALGORITMO DI CONSULTAZIONE
3
4
5
NASTRO DELLE PAROLE NUOVE CHE NON SI TROVANO NEL LA
NASTRO DELLE PAROLE OMOGRAFE
NASTRO DELLE PAROLE UNIVOCHE LEMMATIZZATE DAL LA
PROCEDURA MANUALE PER INTEGRARE IL LA CON LE NUOVE PAROLE
PROCEDURA MANUALE PER LA DISTINZIONE DEGLI OMOGRAFI
SCHEDE PER INTEGRARE IL LA
6
7
8
SCHEDE PER LA DISTINZIONE DEGLI OMOGRAFI
EVENTUALI SCHEDE PER INTEGRARE IL LA
12
EVENTUALE STAMPA DELLE CONCORDANZE LEMMATIZZATE PER CONTROLLO
INTEGRAZIONE DEL LA
DISTINZIONE DEGLI OMOGRAFI
9
10
11
LA INTEGRATO
PAROLE NUOVE LEMMATIZZATE
NASTRO DELLE PAROLE OMOGRAFE LEMMATIZZATE

Fig. 1.

a) La parola è identica a una forma che nel LA figura come univoca

Il programma la lemmatizza automaticamente, cioè la ricopia sul nastro di uscita (5), accompagnata, oltre che dal contesto e dal riferimento, dalla analisi che nel LA è associata alla forma in questione.

b) Nel LA non compare nessuna forma identica alla parola cercata

Il programma la scrive con i relativi contesti e riferimenti sul nastro di uscita (3) che, verrà adoperato per stampare le concordanze di tutte le parole nuove, e cioè presenti nel testo ma assenti nel LA. I lemmi (e le altre eventuali informazioni) vengono scritti a mano in questo elenco e di qui vengono perforati su schede (6) che servono sia per lemmatizzare le parole nuove (nastro IO) sia per inserire le nuove voci nel LA (LA integrato con forme nuove, nastro 9).

c) La parola è identica a una forma che nel LA figura come omografa

Salvo quanto si dirà in seguito a proposito delle possibilità di distinguere automaticamente gli omografi, la loro lemmatizzazione deve essere fatta a mano. Tutte le parole omo-

grafe possibili del testo vengono ricopiate sul nastro di uscita (4), accompagnate ciascuna dal proprio contesto e riferimento, nonchè dalle due o più analisi proposte dal LA. Il nastro 4 servirà per stampare le concordanze delle forme omografe. Accanto ad ogni forma il calcolatore stamperà anche le analisi possibili. Il linguista dovrà leggere i contesti, per assegnare ciascuna occorrenza della forma all'una o all'altra delle analisi possibili. L'analisi assegnata verrà trasferita, per mezzo di schede (8) su nastro (11: contiene le parole omografe del testo, lemmatizzate). Se l'esame dei contesti di una forma rivela una analisi non compresa tra quelle già previste per tale forma del LA, tale analisi deve essere aggiunta al LA, per mezzo di schede (7).

3. Risultati principali della procedura

Riportiamo qui, a titolo di 'specimina', alcuni esempi dei risultati tratti da lavori in corso, o già pubblicati, presso l'ILC.

E' necessario premettere, a questo proposito, una breve precisazione relativa ai sistemi più adoperati per la diffusione dei risultati dello spoglio.

Essi sono:

a) Riproduzione fotostatica, o in offset, dei tabulati direttamente prodotti dalla stampante on-line del calcolatore. Eventualmente, prima della riproduzione, i tabulati vengono ridotti fotograficamente o fotostaticamente.

Gli specimina n°10 e n°14 riportano esempi tratti da pubblicazioni stampate con questo metodo, che è il più rapido ed economico.

b) <u>Ricomposizione tipografica dei tabulati</u>

E' questo indubbiamente il metodo più lungo e costoso: il suo uso è giustificabile solo in casi del tutto particolari.

c) <u>Fotocomposizione comandata dal calcolatore</u>

Le fotocompositrici, come é noto, sono apparecchiature che possono ricevere i dati dal calcolatore direttamente on-line, oppure off-line, per mezzo di supporti diversi (dischetti, nastri, ecc.) e li incidono su un film, grazie ad opportuni comandi intercalati ai dati stessi, i quali specificano il set di caratteri, il formato, il corpo, le spaziature, ecc., della pagina da comporre. La fotocompositrice può operare in base a meccanismi diversi: una sorgente di luce emette un raggio che viene

filtrato da un disco o da una matrice trasparente nei quali sono incisi o disegnati i diversi caratteri. Oppure, nelle più veloci e perfezionate, viene emesso un raggio di luce che disegna sulla pellicola il carat tere seguendo il principio dello 'scanning' televisivo. In questo caso i caratteri sono programmabili nella me moria del calcolatore sotto forma di matrici reticola ri nelle quali sono specificati i quadratini da 'annerire' nel film.

I costi sono in genere più elevati che nel sistema a), tuttavia il risultato, dal punto di vista grafico è decisamente superiore, non di rado superiore a quello di una monotype.

Anche la programmazione è molto complessa. L'ILC sta predisponendo perciò una serie di programmi generalizzati da mettere a disposizione degli utenti delle proprie procedure.

Gli esempi qui riportati sono riprodotti dalle Concordanze del Canzoniere del Petrarca. (specimina n.11).

d) Microfiches prodotte dal calcolatore

Il calcolatore 'stampa' i risultati dello spoglio su microfiches, mediante un dispositivo speciale ad esso collegato. Alcuni tipi di microfiches presentano la possibilità di essere 'riletti' dal calcolatore.

Questo metodo presenta ovvi vantaggi per la manegevolezza e il costo ridotto (una volta che sia stato ammortizzato il costo iniziale delle apparecchiature). Si potrebbe addirittura pensare a un servizio che produca le microfiches di tutte le parti degli indici e delle concordanze su richiesta.

L'ICL sta esamindando per ora la fattibilità di questo metodo.

Ciò premesso, elenchiamo qui di seguito i principali tipi di risultati, prodotti dalla procedura descritta, dei quali si allegano nell'appendice 2 i relativi 'specimina'.

1. Preedizione (preparazione del testo)
2. Lista (corretta del testo
3. Forme in ordine alfabetico
4. Forme in ordine di frequenza decrescente
5. Index locorum delle forme
6. Indice inverso delle forme
7. Incipitario
8. Rimario
9. Concordanze delle forme: contesti ritagliati con la procedira standard dell'ICL
10. Concordanze delle forme
11. Concordanze dei lemmi: contesto stabilito con interventi dello studioso

12. Concordanze delle forme: parola contestualizzata sempre nel centro del proprio contesto
13. Concordanze contrastive dei lemmi
14. Concordanze delle forme con annotazioni manuali dei lemmi
15. Concordanze dei lemmi
16. Concordanze dei lemmi con sottolemmi
17. Lemmi e Forme in ordine alfabetico
18. Lemmi in ordine di frequenza decrescente
19. Concordanze delle Forme con proposta di lemmatizzazione per mezzo del Dizionario Macchina dello Italiano
20. Concordanze delle Forme con proposte di lemmatizzazione per mezzo del Dizionario Macchina dello Italiano e integrazioni manuali del ricercatore
21. Stampa dei contesti su schede meccanografiche

N O T E

(1) Le aggiunte introdotte con la preedizione possono servire a rendere esplicite, chiare, ed univoche per la perforazione (e i controlli successivi) informazione già presenti nel testo (riferimenti, grafemi funzionalmente 'ambigui' - per es. - trattini e apostrofi nei confronti della suddivisione delle parole, ecc.), oppure ad introdurre informazioni nuove. Tipico esempio è quello dei contrassegni apposti per tener distinte, nelle elaborazioni successive, le citazioni alla lettera, nomi propri, cose notevoli, ecc. La fase di preedizione è più complessa nel caso in cui si vogliano ottenere risultati diversi che quelli comunemente prodotti negli spogli lessicografici. Si considerino due esempi: il rimario e le concordanze contrastive. Per produrre un rimario è necessario che il calcolatore possa ordinare i versi di un testo sulla rima, intesa come serie di fenemi compresi tra l'ultimo accento tonico del verso e la fine del verso stesso. La complessità delle operazioni preliminari richieste cresce pertanto proporzionalmente alla distanza tra il sistema grafico e il sistema fonologico di una lingua. Per l'italiano l'ILC consiglia di solito un sistema che sfrutta non solo la stretta corrispondenza della nostra lingua tra grafia e fonetico, ma anche il fatto che le parole piane e tronca costituiscono, nel loro insieme, la parte di gran lunga più frequente in un testo. Chi prepara il testo segue la seguente regola: contrassegna con un puntino sottoscritto la vocale iniziale di rima in tutti e solo i versi nei quali essa non coincide con il penultimo grafema vocalico o con la vocale finale di parola tronca, accentata graficamente. Per esempio non contrassegnerà nulla se il verso finisce in parole come dire, amare, città, consultazione, grido, luogo, pietà, ecc. Dovranno essere invece contrassegnate parole come me devono, specifica, figlio, ecc. L'algoritmo che costruisce il record su cui operare l'ordinamento per il rimario aggiungerà perciò il record - verso la zona 'rima', riportando in essa la stringa finale di lettere del verso a partire dalla vocale col puntino sottoscritto, oppure, se tale puntino manca nel verso, la stringa finale che inizia col penultimo grafema vocalico del verso stesso. L'esempio del

./.

./. le concordanze contrastive è invece più complesso. Con questo termine indichiamo delle concordanze eseguite contemporaneamente su più 'versioni' di uno stesso testo, quali redazioni successive, manoscritti diversi, o traduzioni in altre lingue. Per ogni parola di ciascuna 'versione' le concordanze riportano il contesto della versione da essa cui proviene, nonchè i contesti corrispondenti nelle altre 'versioni'. La corrispondenza viene stabilita in fase di preedizione, nel modo seguente. Il preparatore suddivide dapprima il testo base in 'pericopi' o 'unità contestuali' che numera progressivamente. Ripete poi l'operazione sulle altre versioni, avendo cura di assegnare, in tutte le versioni, numeri progressivi eguale a tutte le 'pericopi' corrispondenti.

(2) Di norma il numero dei caratteri diversi da rappresentare oltrepassa il centinaio (comprendendo nel conto lettere dell'alfabeto, cifre numeriche, segni diacritici e d'interpunzione; e tutto in chiaro e in nero, in tondo e in corsivo, in maiuscole e in minuscole, ecc.), mentre le macchine perforatrici oggi in uso permettono di distinguere, per mezzo delle opportune combinazioni di due o tre fori, non più di 48 o, al massimo, 64 segni diversi. Per questo motivo, nell'elaborare il testo della Costituzione si è dovuto far ricorso, come si sarebbe fatto con qualsiasi altro testo non numerico, a una codificazione complessa che permettesse di stabilire una corrispondenza univoca tra i singoli fori della scheda e i singoli caratteri da rappresentare. Si sono dovuti adoperare a questo scopo combinandoli opportunatamente tra loro, due dei metodi più comuni. Il primo consiste nel rappresentare un dato carattere del testo con una sequenza di due o più perforazioni. L'altro metodo consiste nell'adottare l'equivalente funzionale di quello che nella macchina da scrivere è il tasto delle maiuscole: tra i 64 codici disponibili se ne scelgono alcuni con funzione di 'chiave'; ciascuno dei codici restanti assume un significato diverso a seconda dell'ultima chiave precedente. Com'è chiaro, il numero complessivo dei caratteri che si possono rappresentare grazie a quest'accorgimento è dato dal prodotto del numero dei caratteri usati come chiave per il numero di quelli rimanenti.

3) In altre ricerche, anch'esse in senso largo linguistiche, tali unità sono costituite dai grafemi, dai fonemi, dalle sillabe, dai morfemi, dai sintagmi, o da altro ancora.

4) Si è rivelato molto utile, per una prima revisione di ciò che è stato perforato, anche l'esame accurato degli elenchi delle forme e delle relative frequenze. Esso permette tra l'altro d'identificare rapidamente forme 'impossibili' nella lingua o nel testo considerati, che, essendo dovute per lo più ad errori di perforazione, s'incontrano di regola non più d'una volta. Lo stesso esame permette pure di rilevare incoerenze e discordanze nella grafia di forme particolari che presentano la possibilità di varianti grafiche. Questo fatto è importante nella perforazione di testi che abbiano un'edizione critica, di cui si voglia verificare il grado di accuratezza.

5) In media, per perforare un testo di circa 10.000 parole occorrono circa otto ore di una dattilografa-perforatrice, e altrettante ne sono necessarie per verificare il testo perforato e correggere gli errori.
La lettura della lista-testo, che equivale approssimativamente a una correzione di bozze di stampa, richiede altre cinque o sei ore. Di fronte a questi tempi 'lunghi' stanno quelli 'brevi' necessari per le fasi automatiche dello spoglio: le diverse liste di frequenza e le concordanze per forma si ottengono in non più di mezz'ora, a cui si deve aggiungere un quarto d'ora per la stampa dei risultati.
E' logico quindi che molti sforzi siano riservati a migliorare e alleggerire la fase di preparazione del testo da immettere nel calcolatore, ma è importante soprattutto registrare i testi con criteri scientifici e tecnici uniformi, così che possano essere utilizzati per elaborazioni e ricerche successive da ricercatori diversi. Per questo motivo tutti i testi elaborati elettronicamente per il Vocabolario giuridico seguono i criteri stabiliti dalla Sezione linguistica, oggi ILC ., presso la quale sono oggi in corso di elaborazione elettronica più di cinquanta milioni di parole, in ventuno diverse lingue.

(1) Le aggiunte introdotte con la preedizione possono servire a rendere esplicite, chiare, ed univoche per la perforazione (e i controlli successivi) informazione già presenti nel testo (riferimenti, grafemi funzionalmente 'ambigui' - per es. - trattini e apostrofi nei confronti della suddivisione delle parole, ecc.), oppure ad introdurre informazioni nuove. Tipico esempio è quello dei contrassegni apposti per tener distinte, nelle elaborazioni successive, le citazioni alla lettera, nomi propri, cose notevoli, ecc. La fase di preedizione è più complessa nel caso in cui si vogliano ottenere risultati diversi che quelli comunemente prodotti negli spogli lessicografici. Si considerino due esempi: il rimario e le concordanze contrastive. Per produrre un rimario è necessario che il calcolatore possa ordinare i versi di un testo sulla rima, intesa come serie di fenemi compresi tra l'ultimo accento tonico del verso e la fine del verso stesso. La complessità delle operazioni preliminari richieste cresce pertanto proporzionalmente alla distanza tra il sistema grafico e il sistema fonologico di una lingua. Per l'italiano l'ILC consiglia di solito un sistema che sfrutta non solo la stretta corrispondenza della nostra lingua tra grafia e fonetico, ma anche il fatto che le parole piane e tronca costituiscono, nel loro insieme, la parte di gran lunga più frequente in un testo. Chi prepara il testo segue la seguente regola: contrassegna con un puntino sottoscritto la vocale iniziale di rima in tutti e solo i versi nei quali essa non coincide con il penultimo grafema vocalico o con la vocale finale di parola tronca, accentata graficamente. Per esempio non contrassegnerà nulla se il verso finisce in parole come dire, amare, città, consultazione, grido, luogo, pietà, ecc. Dovranno essere invece contrassegnate parole come me devono, specifica, figlio, ecc. L'algoritmo che costruisce il record su cui operare l'ordinamento per il rimario aggiungerà perciò il record - verso la zona 'rima', riportando in essa la stringa finale di lettere del verso a partire dalla vocale col puntino sottoscritto, oppure, se tale puntino manca nel verso, la stringa finale che inizia col penultimo grafema vocalico del verso stesso. L'esempio del

./.

./. le concordanze contrastive è invece più complesso. Con questo termine indichiamo delle concordanze eseguite contemporaneamente su più 'versioni' di uno stesso testo, quali redazioni successive, mano scritti diversi, o traduzioni in altre lingue. Per ogni parola di ciascuna 'versione' le concordanze riportano il contesto della versione da essa cui proviene, nonchè i contesti corrispondenti nelle altre 'versioni'. La corrispondenza viene stabilita in fase di preedizione, nel modo seguente. Il preparatore suddivide dapprima il testo base in 'pericopi' o 'unità contestuali' che numera progressivamente. Ripete poi l'operazione sulle altre versioni, avendo cura di assegnare, in tutte le versioni, numeri progressivi eguale a tutte le 'pericopi' corrispondenti.

(2) Di norma il numero dei caratteri diversi da rappresentare oltrepassa il centinaio (comprendendo nel conto lettere dell'alfabeto, cifre numeriche, segni diacritici e d'interpunzione; e tutto in chiaro e in nero, in tondo e in corsivo, in maiuscole e in minuscole, ecc.), mentre le macchine perforatrici oggi in uso permettono di distinguere, per mezzo delle opportune combinazioni di due o tre fori, non più di 48 o, al massimo, 64 segni diversi. Per questo motivo, nell'elaborare il testo della Costituzione si è dovuto far ricorso, come si sarebbe fatto con qualsiasi altro testo non numerico, a una codificazione complessa che permettesse di stabilire una corrispondenza univoca tra i singoli fori della scheda e i singoli caratteri da rappresentare. Si sono dovuti adoperare a questo scopo combinandoli opportunatamente tra loro, due dei metodi più comuni. Il primo consiste nel rappresentare un dato carattere del testo con una sequenza di due o più perforazioni. L'altro metodo consiste nell'adottare l'equivalente funzionale di quello che nella macchina da scrivere è il tasto delle maiuscole: tra i 64 codici disponibili se ne scelgono alcuni con funzione di 'chiave'; ciascuno dei codici restanti assume un significato diverso a seconda dell'ultima chiave precedente. Com'è chiaro, il numero complessivo dei caratteri che si possono rappresentare grazie a quest'accorgimento è dato dal prodotto del numero dei caratteri usati come chiave per il numero di quelli rimanenti.

3) In altre ricerche, anch'esse in senso largo linguistiche, tali unità sono costituite dai grafemi, dai fonemi, dalle sillabe, dai morfemi, dai sintagmi,o da altro ancora.

4) Si è rivelato molto utile, per una prima revisione di ciò che è stato perforato, anche l'esame accurato degli elenchi delle forme e delle relative frequenze. Esso permette tra l'altro d'identificare rapidamente forme 'impossibili' nella lingua o nel testo considerati, che, essendo dovute per lo più ad errori di perforazione, s'incontrano di regola non più d'una volta. Lo stesso esame permette pure di rilevare incoerenze e discordanze nella grafia di forme particolari che presentano la possibilità di varianti grafiche. Questo fatto è importante nella perforazione di testi che abbiano un'edizione critica, di cui si voglia verificare il grado di accuratezza.

5) In media,per perforare un testo di circa 10.000 parole occorrono circa otto ore di una dattilografa-perforatrice, e altrettante ne sono necessarie per verificare il testo perforato e correggere gli errori.
La lettura della lista-testo, che equivale approssimativamente a una correzione di bozze di stampa, richiede altre cinque o sei ore. Di fronte a questi tempi 'lunghi' stanno quelli 'brevi' necessari per le fasi automatiche dello spoglio: le diverse liste di frequenza e le concordanze per forma si ottengono in non più di mezz'ora, a cui si deve aggiungere un quarto d'ora per la stampa dei risultati.
E' logico quindi che molti sforzi siano riservati a migliorare e alleggerire la fase di preparazione del testo da immettere nel calcolatore, ma è importante soprattutto registrare i testi con criteri scientifici e tecnici uniformi, così che possano essere utilizzati per elaborazioni e ricerche successive da ricercatori diversi. Per questo motivo tutti i testi elaborati elettronicamente per il Vocabolario giuridico seguono i criteri stabiliti dalla Sezione linguistica, oggi ILC ., presso la quale sono oggi in corso di elaborazione elettronica più di cinquanta milioni di parole, in ventuno diverse lingue.

(6) Non è qui il caso di porre in termini scientifici il problema di cosa si debba intendere per 'contesto' di una parola, problema che facilmente conduce in zone di confine fra linguistica,logica, psicologia, ecc.: proprio per evitarlo abbiamo adoperato il termine 'significativo' invece di termini più tecnici e appropriati. Sul piano pratico la 'sufficienza' o la 'insufficienza' di un contesto vanno giudicate in rapporto all'impiego che se ne vuole fare.

(7) La facilità di consultazione è al centro delle preoccupazioni della maggior parte dei compilatori che si servono dei mezzi elettronici per produr re concordanze, i quali per lo più si propongono di dare al futuro utente non uno strumento che escluda sempre il ricorso al testo, ma piuttosto uno strumento per riconoscere, prima di ricorrere al testo, le occorrenze 'interessanti' ai fini della ricerca. Con il calcolatore, infatti, si ten de a produrre non più le concordanze di una sola opera, ma le concordanze dell'opera omnia di un autore o di un intero periodo storico; lo spoglio è per lo più integrale, e in ogni caso molto più fitto che nelle concordanze tradizionali; ne segue che sono più numerosi e vari i potenziali settori d'interesse degli utenti (grammaticale, lessicografico, documentario, ecc.). Il contrasto tra l'esigenza di ampliare e l'esigenza di ridurre il contesto è bene esemplificato nella procedura di lemmatizzazione manuale. Il tempo di 'lettura' delle concordanze è più o meno proporzionale alla lunghezza del contesto, soprattutto se la parola esponente non è sempre rigidamente allineata al centro di questo. D'altra parte il ricorrere al testo per completare un contesto insufficiente richiede un grosso dispen dio di tempo; i contesti potenzialmente ambigui . ebbono almeno rivelarsi tali, così da costringere a ricorrere al testo per evitare false interpretazio ni. Un terzo fattore da non trascurare, soprattutto nei progetti di maggiore estensione, è il tempo di calcolatore necessario a generare i contesti, ordinar li alfabeticamente e stamparli ; questo tempo, come è naturale, aumenta più o meno proporzionalmente alla lunghezza dei contesti.

(8) L'algoritmo di contestualizzazione può essere portato come esempio di cosa intendiamo per 'flessibilità' dei programmi che compongono le nostre procedure. L'utente deve infatti poter compiere delle scelte sia per quanto riguarda la organizzazione e la forma dei risultati, sia per quanto riguarda le regole algoritmo-linguistiche applicate per produrli. L'utente che vuole produrre una concordanza può:
1) specificare quali unità linguistiche porre in esponente (lemmi; lemmi e forme; forme lessicali; forme grafiche; sintagmi; strutture sintattiche; codici semantici o tematici; ecc.);
2) specificare gli elementi che costituiscono l'esponente (per es. nel caso del lemma parola-lemma, codici morfosintattici, codici di omografia, eventuali commenti e rinvii) e le modalità di formulazione (naturalmente entro un certo limite di caratteri);
3) scegliere tra spoglio integrale e selettivo: portare in esponente tutte le unità del livello prescelto, o solo alcune (per es. vengono spesso omesse le parole vuote o grammaticali); per ciascun esponente, riportare i contesti di alcune occorrenze opportunamente prescelte, tralasciando le altre o riportandone solo i riferimenti;
4)decidere l'ordinamento dei diversi elementi della concordanza, sia per quanto concerne la sequenza degli esponenti (per es., se disporre i lemmi e le forme sotto i relativi lemmi in una unica serie alfabetica, o divisi per categoria grammaticale, o per lingua, ecc.), sia per quanto riguarda l'ordine dei contesti sotto i relativi esponenti (in ordine di testo, in ordine cronologico, ordinati secondo le parole che seguono e/o precedono nel contesto, ecc.);
5) scegliere le regole per il 'taglio' dei contesti, in un insieme di regole disponibili, che abbiamo messo a punto sulla base di un inventario dei criteri di contestualizzazione più frequentemente adoperati . Questi criteri possono essere raggruppati nei tipi seguenti:
a) Nei sistemi di kwic-index, messi a punto soprattutto per applicazioni documentarie, tutte le parole dei titoli che compongono la lista bibliografica da elaborare, o, più spesso, le sole parole 'lessicali' che vi compaiono, vengono elencate in ordine alfabetico, e ricevono come contesti i titoli, o parti di titolo, nei quali occorrono. I programmi di kwic-index oggi più diffusi non sono adeguati per compilare concordanze a scopo lessicografico, per diversi motivi.

./.

/. b) 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. Il programma è di semplice stesura e la composizione dei contesti velocissima; spesso alla scelta di questo metodo si accompagna la decisione di non lemmatizzare, nello evidente proposito di sfruttare al massimo la velocità della macchina e di eliminare ogni intervento umano, ed i contesti vengono ordinati, sotto le rispettive forme, secondo l'ordine alfabetico delle parole che seguono (e/o precedono) la parola esponente nel suo contesto.

c) Il contesto è costituito da un'intera unità di riferimento: il verso, il paragrafo, il versetto, il comma, ecc. Ovviamente il contesto è tanto più 'significativo' quanto più l'unità di riferimento è correlata a una unità di natura linguistica; a tale correlazione è inversamente proporzionale il rischio che il contesto sia inadeguato, soprattutto se la parola esponente occorre ai limiti del contesto stesso.

d) I limiti di contesto sono segnati in fase di 'preedizione': il testo viene suddiviso in 'pericopi' per mezzo di contrassegni che vengono perforati e conservati nelle successive elaborazioni: ciascuna pericope funge da contesto per tutte le parole che la compongono.

e) Il contesto è costituito sempre e solo da tutte le parole comprese tra due segni di punteggiatura. I tipi c, d, e, hanno in comune una caratteristica ben precisa: il testo è segmentato in 'sintagmi' successivi e tutte le parole di un sintagma hanno l'intero sintagma per contesto, cioè hanno il medesimo contesto.

f) Il contesto è scelto in base alla natura della parola: per le parole grammaticali è spesso un 'trinomio' del quale la parola grammaticale è al centro, per le preposizioni sono prese per lo più le due parole successive, ecc. Il presupposto è, evidentemente, che le parole di cui si deve costruire il contesto siano già, in qualche modo, classificate; quindi questo metodo è usato spesso per concordanze di lemmi o comunque nella parte conclusiva dello spoglio (e non come strumento di lavoro, per es., nella fase di lemmatizzazione). Sono interessanti a questo proposito le possibilità di automatizzare almeno in parte l'analisi sintattica, scegliendo, per ogni parola, quella parte terminale della struttura che si giudica interessante come contesto per la categoria grammaticale a cui la parola appartiene.

./.

./. g) Il calcolatore fornisce, con un metodo qualsiasi (per lo più di tipo d), un primo contesto spesso sovrabbondante, che viene poi ridotto alle dimensioni richieste per mezzo di schede con cui si comunicano al calcolatore le parole che lo studioso, dopo un accurato esame, ha deciso di eliminare per snellire il contesto.
h) Il contesto è regolato tenendo conto di determinati segni quali l'interpunzione, il cambio di riferimento, ecc. Come nel tipo b, il contesto viene costruito per ogni parola, cosicchè varia da una parola a quella successiva, ma, a differenza del tipo b e a somiglianza invece dei tipi c,d,e, è regolato sulla presenza di elementi ben definiti, cosicchè la parola può trovarsi collocata diversamente nel contesto: verso l'inizio, verso il centro, verso la fine, a seconda dei casi.
L'algoritmo del nostro programma è potenzialmente in grado di generare contesti secondo tutti i tipi, anche se non è stato sufficientemente sperimentato il tipo f, dal momento che usiamo le concordanze in fase di lemmatizzazione prima di qualsiasi analisi.
E' possibile anche specificare quali elementi del testo vadano 'contestualizzati' e quali no, ed elencare e classificare gli elementi che hanno la funzione di 'limiti' di contesto.
Nel nostro programma gli elementi del testo devono inoltre essere classificati in :
a) elementi che devono avere un proprio contesto e devono essere presenti nei contesti degli altri (per es. le parole 'lessicali');
b) elementi che devono entrare a far parte dei contesti degli altri, ma non ricevono contesto proprio (per es. i segni di punteggiatura);
c) elementi che non devono né entrare a far parte dei contesti né ricevere contesto proprio; di solito si tratta di 'codici' introdotti nel testo per compiere alcune operazioni del programma (per es. i segni di divisione in pagine e in righe);
d) elementi che devono ricevere contesto proprio, ma non entrare nei contesti altrui (per es. le varianti, qualora se ne voglia tener conto, nelle elaborazioni di un testo fornito d'apparato critico).

B I B L I O G R A F I A

Accademia Nazionale dei Lincei (1968), Atti del Convegno sul tema: L'Automazione elettronica e le sue implicazioni scientifiche, tecniche e sociali (Accademia dei Lincei, Roma, 1967), Roma.

U. Bortolini, C. Tagliavini, A. Zampolli (1971), Lessico di frequenza della lingua italiana contemporanea, IBM Italia.

R. Busa (1951), Sancti Thomae Aquinatis Hymnorum Ritualium. Varia Specimina Concordantiarum. Primo saggio di Indici di parole automaticamente composti e stampati da macchine IBM a schede perforate, Milano.

R. Busa, C. Croatto-Martinolli, L. Croatto, C. Tagliavini, A. Zampolli (1962), Una ricerca statistica sulla composizione fonologica della lingua italiana parlata eseguita con un sistema IBM a schede perforate, in International association of logopedics and phoniatrics (1962), Proceding of the XIIth International Speech and Voice Therapy Conference, Padova, pp. 542-562.

R. Busa, A . Zampolli (1968), Centre pour l'Automation de l'Analyse Linguistique (C.A.A.L.), Gallarate, in Les machines dans la linguistique, Prague, pp. 25- 34.

N. Calzolari, L. Moretti (1976), A Method for a Normalization and a possible Algorithmic Treatment of Definitions in the Italian Dictionary, in Preprints del Coling 76.

N. Calzolari, C.Pecchia, A.Zampolli, (1980) Working on Italian Machine dictionari: a Semantic Approach, in N. Calzolari, A.Zampolli (1977) vol. II.

N. Calzolari, A.Zampolli, (1977) Computational an Mathematical Linguistics, Firenze vol. I (1977) vol.II (in stampa).

A. Cappelli, G.Ferrari, L.Moretti, I.Prodanoff, O.Stock (1980) Analisi di un testo italiano con un analizzatore su A.T.N. in pubblicazione su "Informatica" 1980.

G. Cappelli, N.Catarsi, D. Ratti, A.Saba (1979) Análisis automática de textos en lengua española: métodos y aplicaciones, Pisa.

C. Crocetti, (1976), Criteri di analisi del testo per una statistica del lessico dei quotidiani, Tesi di Laurea dell'Università di Pisa.

A. Roncaglia, M.Pacifico (1961), Un esperimento di filologia elettronica: l'omogeneizzazione degli Indici del Franck e dello Spanke, in Almanacco Letterario Bompiani,pp.135-142.

C. Tagliavini (1965), Concordanze della "Divina Commedia" Pisa (Ci si riferisce all'introduzione dell'edizione IBM Italia).

C. Tagliavini (1968) Applicazioni dei calcolatori elettronici all'analisi e alla statistica linguistica, in Accademia Nazionale dei Lincei (1968) pp. 111-118.

W. Woods (1976), Lunar Rocks in Natural English: Explorations in Natural Language Question Answaring, in Zampolli, (1976B).

A. Zampolli (1960), Studi di statistica linguistica eseguiti con impianti IBM, (Tesi di Laurea dattiloscritta),Padova.

A. Zampolli (1968A), Recheche statistique sur la composition phonologique de la langue italienne executée avec un systéme IBM in Les Machines dans la Linguistique, Prague, pp. 25-34.

A. Zampolli (1968B), L'elaboratore elettronico negli studi linguistici in "Rivista IBM", n. 2, pp. 14-19.

A. Zampolli (1968C), Projet d'un dictionnaire italien de machine, Intervention, in Busa (1968) pp. 109-126.

A. Zampolli (a c. di) (1973A), Linguistica Matematica e Calcolatori. Atti del Convegno e della prima Scuola Internazionale, Firenze.

A. Zampolli (1973B), La Section Linguistique du CNUCE , in Zampolli (1973A), pp. 133-199.

A. Zampolli, N. Calzolari, A. Cappelli, G. Ferrari, L.Moretti. L. Pecchia, E. Picchi, I. Prodanoff, N. Ruimy, G. Turrini, Il Dizionario di Macchina dell'Italiano, in Atti del Convegno: "Logiche, Calcoli, Formalizzazioni e Lingue Storiconaturali, Catania 17-19 Settembre 1976, (in stampa).