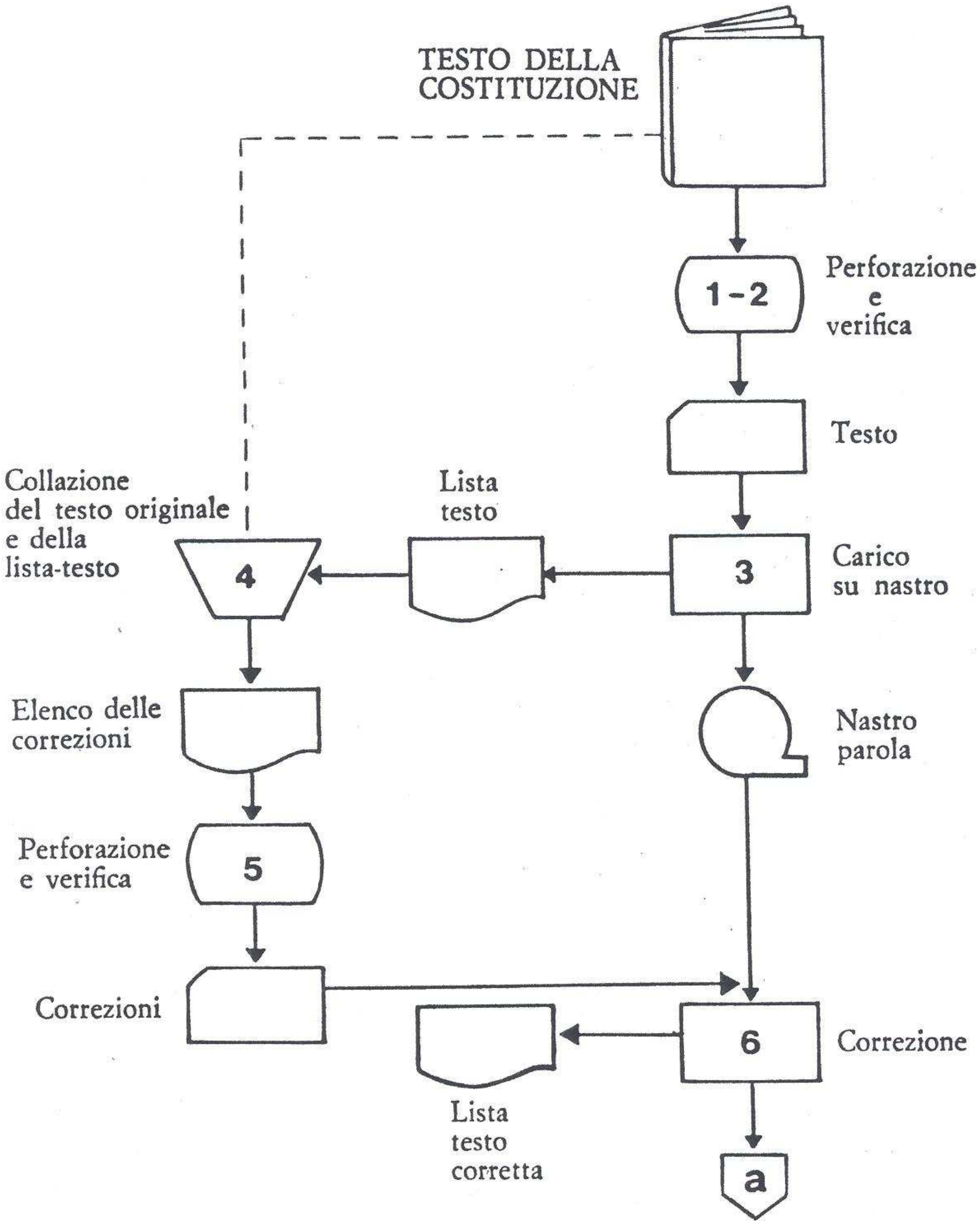
NOTA TECNICA

Per produrre le concordanze e gl'indici della Costituzione sono state eseguite una serie d'operazioni, il cui coordinamento logico e cronologico è rappresentato nel diagramma operativo (*flow-chart*) che si riporta a fianco di questa nota distribuito per comodità di consultazione in quattro diverse tavole, ma da considerarsi essenzialmente un diagramma unico. Ogni singola operazione è contrassegnata con un numero.

Op. 1. - Per ottenere che il testo possa essere letto dal calcolatore, si suole riprodurlo su schede o su nastri di carta per mezzo di una macchina perforatrice. Per la Costituzione ci siamo serviti di schede, sulle quali abbiamo riportato le parole, la punteggiatura, la divisione sia per articoli e commi, sia per pagine e righe (' schede-testo '). Nel trasportare su scheda mediante perforazione i caratteri della stampa abbiamo seguito le norme convenzionali proposte dalla Sezione linguistica del Centro nazionale universitario di calcolo elettronico (CNUCE), di cui, nelle elaborazioni elettroniche della Costituzione, abbiamo usato i programmi e le procedure non meno che gli elaboratori IBM 360/30 e IBM 7090 (1).

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 discor-

(1) 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 opportunamente 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.



LEGGENDA

SCHEDA

PERFORATRICE E VERIFICATRICE

NASTRO MAGNETICO

ELABORATORE

PROSPETTO DA STAMPATRICE

OPERAZIONE MANUALE

danza la macchina si blocca e l'operatore controlla se è dovuta 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 (2). 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 stampatrice dispone (3).

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.

(2) 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.

(3) Analogamente a quanto si è detto per le perforatrici, anche le stampatrici, essendo generalmente destinate ad applicazioni aziendali, dispongono di un numero limitato di caratteri e non offrono varietà di corpi o di serie tipografiche, anche se con vari accorgimenti è possibile rappresentare in modo univoco qualsiasi carattere del testo originale.

(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 (in quello della Costituzione sono 9.380) 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.

È 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

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 del CNUCE, 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 è stato adottato, di regola, un contesto limitato a un numero massimo di circa 120 caratteri, che equivale in media a 12-15 parole (8).

ricerche successive da ricercatori diversi. Per questo motivo tutti i testi elaborati elettronicamente per il *Vocabolario giuridico* seguono i criteri stabiliti dalla Sezione linguistica del CNUCE, 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 produrre 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 tende 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 dispendio di tempo: i contesti potenzialmente ambigui debbono almeno rivelarsi tali, così da costringere a ricorrere al testo per evitare false interpretazioni. Un terzo fattore da non trascurare, soprattutto nei progetti di maggiore estensione, è il tempo di calcolatore necessario a generare i contesti, ordinarli alfabeticamente e stamparli; questo tempo, com'è naturale, aumenta più o meno proporzionalmente alla lunghezza dei contesti.

(8) Le esperienze nostre, e degl'istituti che seguono le nostre procedure, hanno di regola avuto per oggetto le concordanze delle forme, sulle quali il ricercatore deve poter individuare e dividere le forme omografe, assegnare lemmi, e spesso anche scegliere quali contesti escludere dalle elaborazioni successive, e quali invece ritenere nella documentazione lessicografica a conclusione dello spoglio.

Spessissimo un contesto comprende l'intera frase, abbastanza spesso un periodo tra due punteggiature forti. Per l'italiano, il ricorso al testo risulta necessario, in media, meno di una volta su cento. Inoltre, questa lunghezza permette di stampare il contesto e il suo riferimento,

Il programma tiene conto di diversi fattori (per es. la punteggiatura, la fine o l'inizio di un capitolo, ecc.) nel delimitare il contesto (9). 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.

almeno quello topografico, su una sola riga di calcolatore, senza andare a capo, con un notevole risparmio di carta e di tempo nella stampa.

(9) I criteri di delimitazione del contesto vanno ormai uniformandosi presso i centri di lessicografia che utilizzano impianti meccanografici. Praticamente 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, nell'evidente proposito di sfruttare al massimo la velocità della macchina e di eliminare ogni intervento umano.

c) Il contesto è costituito da un'intera unità di riferimento: il verso, il paragrafo, il comma, ecc.

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.

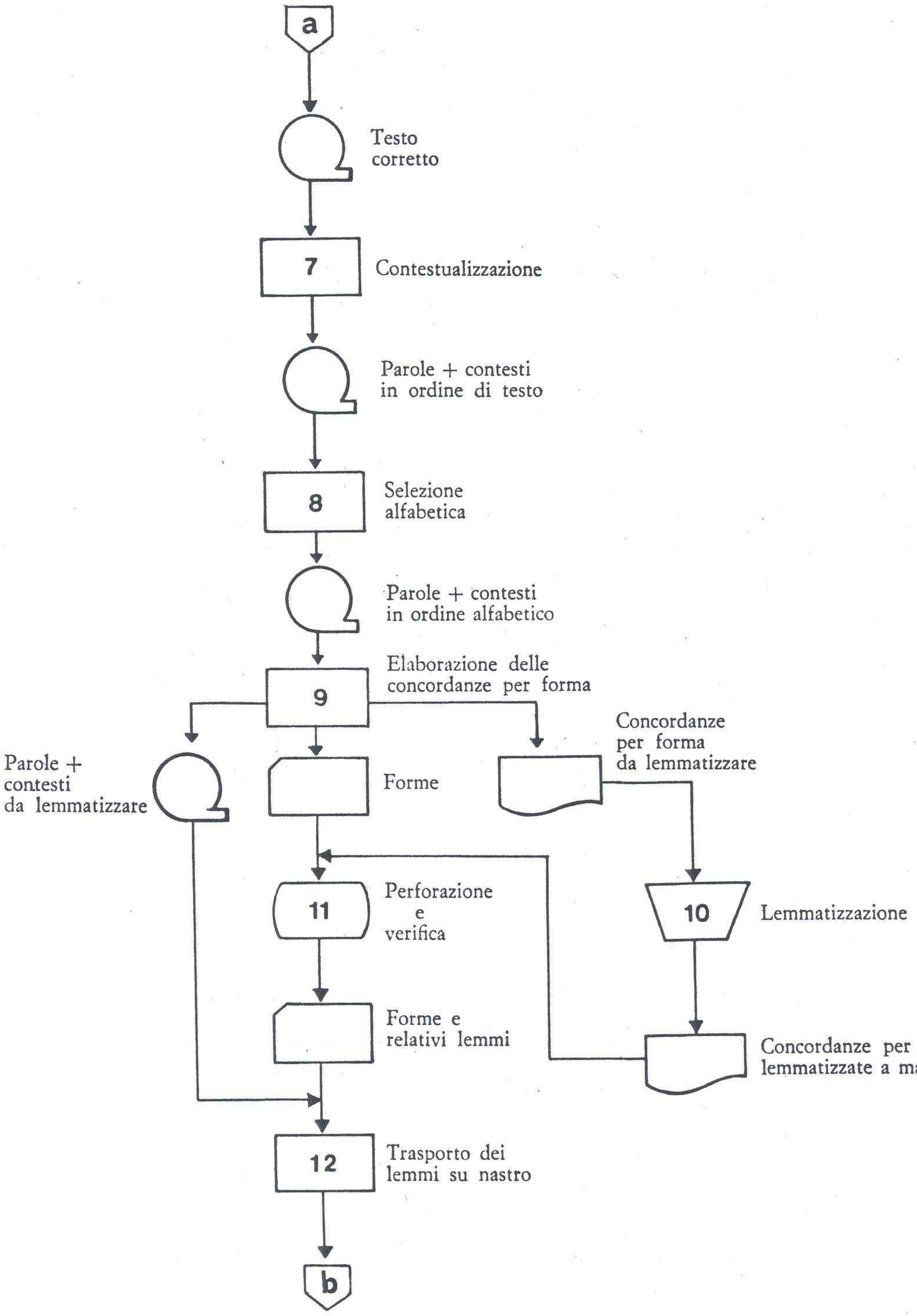
e) Il contesto è costituito sempre e solo da tutte le parole comprese tra due segni di punteggiatura. I tipi *c*, *d*, *e* hanno in comune una caratteristica ben precisa: il testo è segmentato in 'sintagmi' successivi, e tutte le parole di un sintagma hanno l'intero sintagma per contesto, cioè hanno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


a
Testo corretto
7
Contestualizzazione
Parole + contesti in ordine di testo
8
Selezione alfabetica
Parole + contesti in ordine alfabetico
9
Elaborazione delle concordanze per forma
Parole + contesti da lemmatizzare
Forme
Concordanze per forma da lemmatizzare
11
Perforazione e verifica
10
Lemmatizzazione
Forme e relativi lemmi
Concordanze per forma lemmatizzate a mano
12
Trasporto dei lemmi su nastro
b

È 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 ([10]).

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.

Op. 10. - Le indicazioni dei lemmi e le eventuali indicazioni accessorie vengono scritte a mano sulla lista delle concordanze accanto a ciascuna forma ([11]).

([10]) Nel nostro programma gli elementi del testo devono essere classificati in:

— elementi che devono avere un proprio contesto e devono essere presenti nei contesti degli altri (per es. le parole ' lessicali ');

— elementi che devono entrare a far parte dei contesti degli altri, ma non ricevono contesto proprio (per es. i segni di punteggiatura);

— 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);

— 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).

Degli elementi del testo si deve anche fare una seconda distinzione in due gruppi:

— elementi che hanno funzione di ' delimitare ' il contesto (per es. un segno di punteggiatura, o la fine di un capitolo);

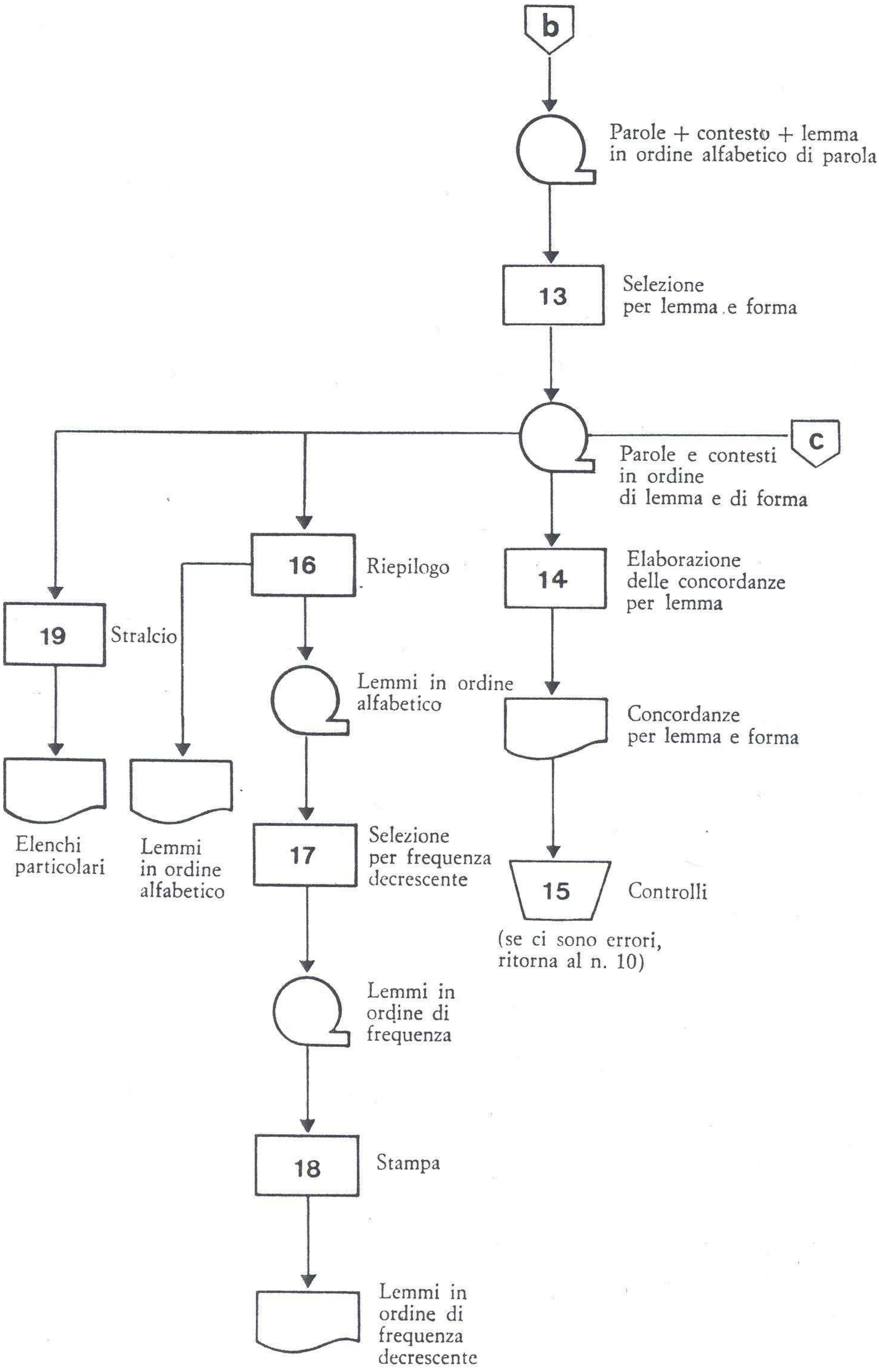
— elementi che non hanno funzione di delimitare il contesto (per es. una parola).

Nel nostro algoritmo, i limiti possono essere distribuiti in classi, fino a un massimo di 9. Gli elementi della classe 9 sono ' limiti invalicabili ' di contesto: cioè, se nel costruire la parte destra di un contesto si incontra un limite di classe 9, il contesto non procederà in alcun modo più oltre verso destra, qualunque sia lo spazio ancora a disposizione. Si definiscono spesso limiti di classe 9 i contrassegni di un cambiamento di capitolo: evidentemente perché non si giudicano utili al contesto della parola finale di un capitolo le parole iniziali del capitolo successivo, e viceversa.

Gli elementi delle altre classi, dalla 1 alla 8, sono ' limiti valicabili ' di contesto. Se nel costruire la parte destra di un contesto s'incontra uno di questi limiti, il contesto potrà scavalcarlo e riprocedere verso destra solo dopo aver trovato un limite di classe eguale o maggiore nel procedere verso la direzione opposta. I vari segni di punteggiatura possono essere attribuiti a classi diverse, secondo la loro forza: per es. il punto fermo, l'interrogativo e l'esclamativo alla classe 3; le parentesi, le virgolette, i due punti, il punto e virgola alla classe 2; la lineetta e la virgola alla classe 1, ecc.

Nelle concordanze della Costituzione abbiamo considerato come invalicabili i limiti di contesto stabiliti in fase di preedizione (cfr. n. 9 *d*), come valicabili (senza ulteriore suddivisione di classi) quelli costituiti dai vari segni di punteggiatura.

([11]) Le operazioni di lemmatizzazione potranno essere automatizzate, almeno parzialmente, quando sarà disponibile il cosiddetto dizionario di macchina dell'italiano contemporaneo, in preparazione dal CNUCE con il contributo del C.N.R. Il dizionario di macchina sarà costituito da un elenco di forme registrate su nastro magnetico, ciascuna accompagnata dal lemma o dai lemmi cui può appartenere. Tale elenco di forme si ottiene generando automaticamente tutte le possibili flessioni di circa 120.000 lemmi della lingua italiana (cfr. Aldo Duro e Antonio Zampolli, *Analisi lessicali*, in *Atti del convegno sul tema: «L'automazione elettronica e le sue implicazioni scientifiche, tecniche e sociali»*, Roma, Accademia nazionale dei Lincei, 1968, pp. 119-139).


b
Parole + contesto + lemma
in ordine alfabetico di parola
13
Selezione
per lemma e forma
C
Parole e contesti
in ordine
di lemma e di forma
16
Riepilogo
14
Elaborazione
delle concordanze
per lemma
19
Stralcio
Lemmi in ordine
alfabetico
Concordanze
per lemma e forma
Elenchi
particolari
Lemmi
in ordine
alfabetico
17
Selezione
per frequenza
decrescente
15
Controlli
(se ci sono errori,
ritorna al n. 10)
Lemmi in
ordine di
frequenza
18
Stampa
Lemmi in
ordine di
frequenza
decrescente

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.

Antonio Zampolli