Procedura per l'elaborazione automatica

di ANTONIO ZAMPOLLI

Nella ricca gamma di applicazioni dell'elaborazione elettronica dei dati alle ricerche linguistiche e letterarie, l'automazione degli spogli lessicali ha un posto di grande rilievo.

Negli anni tra le due guerre mondiali erano state tentate, senza successo, diverse tecniche per la meccanizzazione, almeno parziale, delle attività di spoglio. Nel 1949 R. Busa S.J. mise a punto la prima procedura di spoglio con macchine elettrocontabili, con lo scopo di produrre gli indici dell'*opera omnia* di S. Tommaso d'Aquino.

Alcuni anni più tardi il suo esempio fu seguito da altri Centri specializzati sparsi un po' dovunque: in Francia, in Belgio, in Inghilterra, in Olanda. Restava tuttavia praticamente senza risposta, nonostante alcune procedure costosissime e poco felici di ripiego, l'esigenza di produrre concordanze di struttura adeguata.

Questo problema fu risolto quando, verso il principio degli anni '60, i calcolatori elettronici, già diffusi in campo commerciale e industriale, furono impiegati nelle procedure di spoglio dei testi. Il loro uso si diffuse molto rapidamente, ed oggi è disponibile presso molti Centri una ricca gamma di procedure e di programmi per produrre, con il calcolatore, indici di frequenza, rimari, indici inversi, concordanze, schede contestualizzate, ecc.

Il maggiore impulso è venuto senza dubbio dal fatto che i principali progetti di grandi dizionari storici nazionali hanno adottato o stanno adottando gli spogli elettronici. Gli spogli elettronici di testi si moltiplicano in tutto il mondo, e si vengono creando grandi 'biblioteche elettroniche', o 'banche di dati linguistici' che dir si voglia, tanto che è pienamente giustificato, come ha fatto J. Raben, parlare di "The death of the hand made concordance".[1]

[1] Per una rassegna dello sviluppo e dello stato attuale della lessicologia e della lessicografia assistite dai calcolatori, e per la relativa bibliografia, si veda il primo capitolo di A. ZAMPOLLI, *Problemi di linguistica applicata computazionale*, Pisa 1974.

Un panorama dei progetti in corso in Italia e in Europa è fornito dagli articoli raccolti in A. ZAMPOLLI, *Linguistica matematica e calcolatori*, Firenze 1973.

Tuttavia, almeno a mia conoscenza, non sono ancora stati messi a punto dei metodi adeguati per il trattamento delle varianti nelle concordanze e negli indici.[2] Esse vengono di solito completamente trascurate nello spoglio. Al più, si contrassegnano i luoghi che presentano qualche variante[3] in modo che l'utente, avvertito, possa ricorrere, se crede, al testo, per conoscere le varianti del contesto in questione.

Quando C. Segre propose all'Accademia della Crusca e al CNUCE il progetto di spoglio delle tre redazioni del *Furioso* fu subito chiaro che era giunto il momento di studiare e mettere a punto, grazie alla collaborazione tra filologi lessicografi e linguisti computazionali, una metodologia per il trattamento delle varianti.

Dalle prime discussioni, svoltesi nel Comitato Tecnico dell' Opera del Vocabolario dell'Accademia della Crusca, emersero proposte molto diverse tra loro, ma alla fine decidemmo di sperimentare la procedura che ora presentiamo in questa sede.

1. *Alcune note terminologiche*

Prima di descriverla con qualche dettaglio, mi sembra utile precisare alcuni termini usati comunemente nella pratica quotidiana degli spogli lessicali automatici: non mi propongo di dare delle definizioni rigorose, che in taluni casi non esistono, ma solo di indicarne intuitivamente il contenuto così come si è venuto affermando nell'uso, al fine di agevolare e semplificare la descrizione delle successive fasi delle nostre elaborazioni. Chiamiamo *parole* (o *occorrenze*) le successive unità grafiche di cui è costituito un testo: per il programma, l'unità grafica corrisponde a una o più lettere (o caratteri equivalenti) tra spazi o segni di interpunzione.[4] L'insieme delle indicazioni che individuano la posizione della parola nel testo costituiscono il suo *riferimento*: per esempio

[2] Diversa è la situazione degli studi per formalizzare e meccanizzare le operazioni per il calcolo dello *stemma codicum*; nell'ultimo ventennio sono state studiate alcune procedure che, se si prescinde dalle possibiltà di applicazione pratica, sono certamente apprezzabili sul piano teorico.

[3] Si vedano, per esempio, le concordanze dell'*opera omnia* di Seneca pubblicate a cura di R. Busa e A. Zampolli presso l'editore Georg Olms (Hildesheim 1975).

[4] I caratteri codificati che costituiscono il testo registrato su schede si dividono, per il programma, in due grandi categorie funzionali:

b) *lettere*, cioè codici che compongono le parole (per esempio: lettere dell'alfabeto, segni diacritici, apostrofo, ecc.).

b) *lettere*, cioè codici che compongono le parole (per esempio: lettere, segni diacritici, apostrofo, ecc.).

Una parola è definita come una serie ininterrotta di *lettere* tra due *separatori*.

i numeri di pagina e riga, oppure i numeri di canto, ottava, e verso.

Chiamiamo *record* una unità di trattamento registrata su un supporto accettabile dal calcolatore (nastro magnetico, dischi, ecc.), rappresentante l'elemento base di una certa elaborazione elettronica, e composta a sua volta da un certo numero di elementi di informazione (*campi*) tra loro collegati. Per esempio, possiamo pensare che un testo sia registrato come una serie di *record-parola*: ogni parola del testo è registrata su un record (cui d'ora in poi ci riferiremo con la sigla RP) il quale comprende tre campi distinti: il *campo-parola* che contiene una ed una sola parola del testo, il *campo-riferimento* che contiene il riferimento di questa parola, e il *campo-n.record* che contiene il numero progressivo del record nella serie ordinata dei record.

Chiamiamo *selezione* l'operazione con la quale si dispongono dei *record* in una data sequenza (crescente o decrescente) in base ad una certa *chiave* di ordinamento. Ad esempio, in una guida del telefono, gli abbonati sono ordinati per ordine alfabetico, e la chiave di ordinamento è il cognome-nome dell'abbonato. Ma, sempre in una guida del telefono, una diversa sequenza degli abbonati è quella dell'elenco stradale. Qui la chiave di ordinamento è la strada e il numero civico. Come si vede, la chiave può essere composta da più *campi*, collegati da una precisa *gerarchia*: all'interno della sequenza alfabetica per strada (campo n. 1), gli abbonati sono ulteriormente ordinati per numero civico (campo n. 2).[5]

In un testo sufficientemente lungo non tutte le parole sono diverse: alcune sono ripetute una o più volte. Chiamiamo *forme grafiche* le parole "diverse" presenti nel testo: l'elenco delle forme grafiche ci darà quindi tutte le parole diverse del testo esaminato; accanto a ciascuna forma grafica possiamo far scrivere il numero delle sue apparizioni nel testo, che chiamiamo *frequenza assoluta* della forma. Per esempio nella frase *chi non è con me è contro di me* le parole sono 9, ma le forme grafiche sono 7: *è* e *me* hanno frequenza 2, *chi*, *non*, *con*, *contro*, *di* hanno frequenza 1.

[5] Nella definizione di *record* e di *selezione* ho seguito da vicino le voci *record* e *ordinamento* del *Dizionario di Informatica* di A. CHANDOR (traduz. dall'inglese di G. RAPELLI), Bologna 1972. Ho usato il termine *selezione* anche per i processi di ordinamento con il calcolatore (mentre il *Dizionario* cit. riserva *selezione* solo all'ordinamento con macchine a schede) per rispettare l'uso prevalente tra gli utenti linguisti dei calcolatori.

Come vedremo, a una stessa forma grafica possono corrispondere unità linguistiche diverse: per esempio la forma grafica *danno* può essere sia la prima persona singolare presente indicativo di *dannare*, sia la terza persona plurale presente indicativo di *dare*, sia il sostantivo sinonimo di 'guasto', 'lesione'. Diciamo che la forma grafica *danno* è *omografa*, e corrisponde a tre *forme lessicali* distinte.

Chiamiamo *lemma* quella voce di base che rappresenta, in un dizionario, le varie forme di un testo. Per esempio, alle forme *dite* e *diremo* corrisponde il lemma *dire*; alle forme *va* e *andarono* il lemma *andare*, alle forme *bello*, *belle*, *belli* il lemma *bello*.

Chiamiamo *lemmatizzazione* o (come avrebbe preferito dire G. Devoto) *lemmazione* l'operazione di ricondurre ciascuna forma e le relative occorrenze al rispettivo lemma.[6]

Chiamiamo *record-parola lemmatizzato* (RPL) un RP al quale è stato aggiunto, in un apposito campo, il lemma della parola in esso contenuta.

La *frequenza assoluta del lemma* è il numero delle sue occorrenze nel testo, e corrisponde alla somma delle frequenze delle sue forme.

Nell'elenco dei *lemmi* (o delle *forme*) *per frequenza decrescente*, i lemmi (o le forme) sono elencati dal più frequente al meno frequente: lemmi (o forme) di frequenza uguale sono ordinati alfabeticamente entro la stessa *classe* di frequenza.

Chiamiamo *contesto* di una parola un insieme variamente delimitato di parole che le sono contigue, o comunque associate, nel testo, scelte di solito fra quelle che la precedono o la seguono immediatamente.

Chiamiamo *record-parola contestualizzato* (RPC) un RP a cui è stato aggiunto, in un apposito campo, il contesto relativo alla parola registrata nel suo campo-parola, che d'ora in poi chiameremo *parola-esponente* del RPC.

Chiamiamo *contestualizzazione* di un testo l'operazione con la quale si ricopiano, su un nuovo nastro, tutti o alcuni RP di un nastro-testo, aggiungendo loro il contesto, e cioè trasformandoli in RPC.

La contestualizzazione è detta *integrale* se *tutti* i RP vengono ricopiati e trasformati in RPC; altrimenti è detta *selettiva*. Un

[6] Per le difficoltà che derivano al processo di lemmatizzazione dall'attuale stato del concetto di *parola* nelle teorie delle diverse scuole linguistiche, si veda U. Bortolini, C. Tagliavini, A. Zampolli, *Lessico di Frequenza della Lingua Italiana Contemporanea*, IBM Italia, 1971, pp. XLII sgg. (ripubblicato da Garzanti 1972).

esempio molto frequente di contestualizzazione selettiva è quello che considera parole da portare in esponente, e cioè da contestualizzare, solo le parole cosiddette piene o semantiche (sostantivi, aggettivi e avverbi qualificativi, verbi), mentre elimina le altre (le cosiddette parole vuote o sinsemantiche: articoli, pronomi, preposizioni, congiunzioni, ecc.).

Chiamiamo *concordanza* di un testo l'insieme *ordinato* di tutti i RPC generati dalla contestualizzazione. Se la chiave dell'ordinamento è l'ordine alfabetico delle parole-esponente parliamo di *concordanza delle forme*; parliamo invece di *concordanze dei lemmi* se la chiave è l'ordine alfabetico dei lemmi.

2. *Criteri per la contestualizzazione delle varianti*

I criteri di delimitazione del contesto vanno ormai uniformandosi presso i vari Centri che operano nel settore degli spogli.[7] Mancano invece dei metodi adeguati per l'inserimento delle varianti. Il metodo che abbiamo messo a punto per la elaborazione di un testo del quale esistono più versioni (per esempio tra-

[7] 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 è sempre al centro del suo contesto, ossia è sempre preceduta e seguìta 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

duzioni in lingue diverse, redazioni successive dell'autore, traduzioni diverse di un manoscritto, ecc.) si basa sulla composizione di *unità contestuali* (UC) particolari.

Sia T0 il testo originale e siano T1, T2... Tn le sue *n* diverse traduzioni (oppure sia T0 la redazione finale, e siano T1, T2... Tn le redazioni precedenti, ecc.). Supponiamo che, in linea di massima, sia possibile compiere le operazioni seguenti:

1. Segmentare T0 in una successione di *m* stringhe consecutive, cioè tali che ogni parola di T0 appartenga ad una e ad una sola stringa ben definita.

2. Numerare progressivamente queste stringhe da 1 ad *m*.

3. Segmentare T1 in una successione di *m* stringhe consecutive, tali che ciascuna di esse costituisca la traduzione (la redazione precedente, una diversa edizione, ecc.) di una e di una sola stringa di T0, in modo che, data una stringa di T0, sia possibile stabilire una corrispondenza biunivoca con una sola stringa di T1 e viceversa.[8]

4. Assegnare a ciascuna stringa di T1 lo stesso identico numero progressivo che individua la stringa di T0 della quale costituisce la traduzione (o la precedente redazione, ecc.).

5., 6..., 2 (n+1). Eseguire, per ciascuna delle altre versioni T2, T3... Tn, le operazioni 3 e 4.

Otteniamo così un insieme di $m \times n$ stringhe: T01, T02... T0m; T11, T12,..., T1m; Tn1, Tn2,... Tnm, dove il primo indice si riferisce al testo di provenienza, e il secondo è il numero progressivo della stringa. Se raggruppiamo assieme tutte le stringhe che hanno lo stesso numero progressivo, otterremo *m* gruppi, che chiamiamo unità contestuali (UC). Ciascuna unità contestuale contiene $n+1$ stringhe, e cioè una stringa di T0 e le stringhe che le corrispondono in ciascuna delle sue *n* versioni.

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.

b) 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.

[8] In pratica, non è sempre possibile suddividere To, T1..., Tn in un egual numero di stringhe corrispondenti. È da prevedere che ci siano brani aggiunti, soppressi, interamente rifatti. Essi verranno contrassegnati come "brani rifatti", e trattati a parte e cioè senza confronto tra le diverse versioni.

A questo punto possiamo procedere alla contestualizzazione. Se adottiamo il criterio esaustivo, creiamo un RPC per ciascuna parola di una UC,[9] e assegniamo a ciascuna di esse come contesto l'intera UC. In altri termini, tutte le parole, sia quelle del testo base, sia quelle delle sue diverse versioni, vengono portate in esponente, e ciascuna di esse riceve come contesto sia la stringa che la contiene, sia le stringhe che a questa corrispondono nelle altre versioni. Questo criterio è stato adottato con successo per elaborare un testo e le sue traduzioni,[10] ma appare decisamente ridondante per elaborare redazioni o edizioni diverse di un testo.

È chiaro che mentre nel primo caso le stringhe di una UC sono per definizione diverse tra di loro, nel secondo le stringhe di una UC sono uguali tra loro ogniqualvolta una stringa non ha subito modifiche da una versione all'altra. Inoltre, poiché, in questo caso, ogni parola ricorre identica in tutte le stringhe, si generano $n+1$ RPC identici per parola. Ciò accade anche per parole che rimangono invariate in stringhe che hanno subito una modifica parziale. Conviene quindi adottare un criterio selettivo che permetta di ridurre drasticamente il numero delle parole da portare in esponente e di semplificare le UC ridondanti.

La nostra procedura prevede che queste riduzioni avvengano sia automaticamente, in base a regole generali (*riduzione automatica*), sia applicando indicazioni fornite caso per caso dal filologo (*riduzione manuale*).

Darò alcuni esempi del tipo di regole che si potrebbero applicare riferendomi d'ora in avanti, per maggiore concretezza, alle tre redazioni del *Furioso*.

Per semplicità di esposizione, adotterò le seguenti convenzioni:

— lettere maiuscole tra parentesi indicano le tre diverse edizioni: (C), (A), (B)

— lettere maiuscole fuori parentesi indicano un verso: Cε (C), Aε (A), Bε (B). Quando nomino contemporaneamente le coppie C, A; (C, B); A, B, o la tripla C, A, B, intendo indicare rispettivamente due o tre versi che si corrispondono nelle tre redazioni, e cioè che hanno lo stesso riferimento;

[9] Se chiamiamo N0 il numero complessivo di parole di T0, N1 quello di T1..., Nn quello di Tn, il totale dei RPC sarà N=N0+N1...+Nn.

[10] Con questo sistema è possibile confrontare una parola dell'originale con tutte le sue traduzioni, e la traduzione di una parola con l'originale e con tutte le altre traduzioni. L'*optimum* sarebbe forse costituito da concordanze più analitiche, nelle quali ogni parola dell'originale fosse associata, in esponente, alla parola o al sintagma che la traduce. È facile convincersi che questa associazione richiede un lavoro preparatorio manuale estremamente lungo e faticoso, che in ogni caso sarebbe più agevole eseguire sulla documentazione fornita dalle concordanze qui descritte.

— lettere minuscole in corsivo indicano una parola di un verso: *c* ε C; *a* ε A; *b* ε B. Quando nomino le coppie (c, a), (c, b), (a, b) o la tripla (c, a, b), intendo indicare che *a* e/o *b* occupano, rispettivamente, in A e in B, la stessa posizione che *c* occupa in C. In questo caso dirò che le parole della coppia o della tripla sono, formalmente, *varianti reciproche*.[11] Quando scrivo $a=b$, $a\neq b$, A=C, A≠C, ecc. intendo dire che, dal punto di vista puramente grafico, due parole o due versi sono rispettivamente eguali (=) e diversi tra loro (≠).

La semplificazione di una unità contestuale può avvenire con la regola seguente:

Si confrontano tra di loro le coppie di versi:

C : A
C : B
A : B

Se i due versi di una coppia sono eguali tra loro, si cancella il secondo, avendo però cura di aggiungere alla UC l'indicazione della situazione iniziale. Poiché le situazioni possibili sono in tutto 5, proponiamo la seguente codificazione:

Situazione			*Codice*	*UC semplificata*
C=A;	C=B;	A=B	1	C
C=A;	C≠B;	A≠B	2	C, B
C≠A;	C=B;	A≠B	3	C, A
C≠A;	C≠B;	A=B	4	C, A
C≠A;	C≠B;	A≠B	5	C, A, B

Se portiamo in esponente tutte indistintamente le parole di una UC, per ogni parola che si trovi inalterata in due (o in tutti e tre i) versi della UC, generiamo una coppia (o una tripla) di RPC del tutto identici tra loro. Possiamo eliminare questa ridondanza con una regola del tutto analoga a quella precedente. Anche in

[11] Il fatto che una parola (o un gruppo di parole) *a* venga messo in relazione con un *c*, non implica necessariamente che il filologo ritenga che *c* sostituisca *a*, e cioè che *c* sia una variante di *a*, ma solo che, ai fini della ricostruzione di A, al posto di *c* deve essere inserita, in C, *a*.

Così per es., nel caso che due parole consecutive (*c*1 e *c*2) si ritrovino identiche ma invertite in A, *a*1 sarà uguale a *c*2, e *a*2 uguale a *c*1. Tuttavia il codice di LINK — si veda più avanti nel testo — permette di segnalare che, in realtà, $c1=a2$ e $c2=a1$.

Nello schema generale della nostra procedura, c'è la possibilità, che non è stata sfruttata nella elaborazione del *Furioso*, di classificare le varianti, o almeno quelle tra di esse la cui natura è ben chiara. Per esempio: variante grafica, morfologica, lessicale, ecc. Questa classificazione permetterebbe evidentemente di elencare le varianti per categorie, e di fornire statistiche sulla frequenza dei diversi tipi.

questo caso sono possibili solo 5 diverse situazioni che codifichiamo e semplifichiamo così:

Situazione	*Codice di situazione*	*Parole esponente dopo la semplificazione*
c=a; c=b; a=b	1	c
c=a; c≠b; a≠b	2	c,b
c≠a; c=b; a≠b	3	c,a
c≠a; c≠b; a=b	4	c,a
c≠a; c≠b; a≠b	5	c,a,b

Sono possibili, naturalmente, regole ancora più selettive. Per esempio, si può richiedere che nella situazione del tipo

C≠A, ma con $c=a$

si porti in esponente solo *c*, con un contesto costituito solo dal verso C, e, naturalmente, con i due codici di situazione relativi. Questa semplificazione può essere giustificata nel caso che le differenze tra una coppia (o una tripla) di versi siano di poco conto, tanto che interessino solo per lo studio della parola che varia, ma non di quelle che rimangono inalterate.[12]

Naturalmente, se si introduce questa regola, il calcolatore la applica meccanicamente a tutte le situazioni del tipo ora descritto. Se invece si ritiene che in alcuni casi sia utile riportare per *c* sia C sia A, e in altri invece solo C, il filologo dovrà specificare di volta in volta quali versi riportare come contesto.

Per il *Furioso* è stata scelta appunto una procedura di riduzione semiautomatica che descriveremo più avanti.

3. *Lo spoglio delle tre redazioni del "Furioso"*

3.1. *Risultati prevedibili*

3.1.1. *Indice alfabetico dei lemmi.* — Si tratterà probabilmente di un indice "sinottico", nel quale i lemmi, ordinati alfabeticamente, saranno accompagnati dalle rispettive frequenze, assolute e percentuali, in (C) in (A) e in (B). Eventualmente, per risparmiare spazio, si potranno adottare accorgimenti grafici particolari;

[12] Per esempio modifiche nella forma dell'articolo possono presentare scarso interesse per parole che fanno parte di un altro gruppo nominale.

p. es. stampare una sola volta la frequenza se essa è identica nelle tre redazioni (o due volte se è la stessa in due di esse).

3.1.2. *Indice alfabetico delle forme.* — Si stamperà probabilmente in modo del tutto analogo.

3.1.3. *Indice dei lemmi per frequenza decrescente.* — La struttura di questo indice è più complessa, perché, a motivo del variare delle frequenze, i lemmi possono occupare ranghi diversi negli indici relativi alle tre redazioni. Comunque gli elementi saranno quelli ormai abituali: il lemma (in ordine di frequenza decrescente), il suo rango, un numero progressivo, la sua frequenza assoluta e percentuale, le sommatorie relative a queste due frequenze.[13]

Dovremo scegliere tra diverse soluzioni possibili; per es.:
— stampare tre indici separati, uno per ogni redazione;
— stampare i tre indici l'uno affiancato all'altro, su tre colonne, con un unico numero progressivo a margine;
— stampare i lemmi secondo l'ordine decrescente di (C), affiancati dai rispettivi rango, progressivo, frequenza e sommatoria in (C), nonché dal loro rango in (A) e in (B), ed eventualmente dalle variazioni di frequenza in (A) e in (B) rispetto a (C), formulate con espressioni del tipo ± x, dove x = frequenza di *c* meno frequenza di *a* (o di *b*).

3.1.4. *Indice delle forme per frequenza decrescente.* — Sceglieremo probabilmente una soluzione analoga a quella adottata per i lemmi.

3.1.5 *Indice inverso dei lemmi.* — Quest'indice è del tutto analogo a quello alfabetico (vedi 3.1.1.), salvo che i lemmi sa-

[13] Di solito organizziamo la lista per frequenza decrescente in questo modo. I lemmi vengono ordinati per frequenza decrescente e, a parità di frequenza, in ordine alfabetico. Ad ogni lemma vengono associati:

1. Un numero progressivo nell'ordine in questione.
2. Il rango del lemma nella lista. Il rango viene calcolato come segue: si assegna rango 1 al lemma più frequente, 2 al lemma che lo segue immediatamente, ecc. Quando *x* lemmi hanno frequenza eguale, a ciascuno di essi si assegna lo stesso rango, ottenuto con la formula

$$\frac{(n - 1) + (n + x)}{2}$$

dove *x*, come si è detto, è il numero dei lemmi che hanno la stessa frequenza, e *n* è il numero progressivo del primo di questi *x* lemmi.

3. La sua frequenza assoluta.
4. La sua frequenza percentuale.
5. La sommatoria delle frequenze assolute: cioè la somma della frequenza del lemma in questione e delle frequenze di tutti i lemmi che lo precedono.
6. La sommatoria delle frequenze percentuali.

Altrettanto dicasi per la lista delle forme.

ranno ordinati alfabeticamente a partire da destra anziché da sinistra.

3.1.6. *Indice inverso delle forme.* — Del tutto analogo al precedente.

3.1.7. *Indici statistici vari.* — La procedura adottata fornisce alcuni prospetti che potremo decidere se pubblicare o no solo dopo averli esaminati.

Per esempio, dopo avere calcolato per ciascun lemma (forma) un indice numerico che esprima l'entità del variare della sua frequenza nelle tre redazioni,[14] potremo ordinare i lemmi (le forme) secondo il valore decrescente di questo indice, e stampare solo i lemmi il cui indice supera una certa soglia. Oppure stampare tutte e sole le forme (lemmi) che hanno varianti nelle altre redazioni, facendo seguire ad ogni forma (lemma) le rispettive varianti.

3.1.8. *Concordanze dei lemmi.* — La loro struttura non è ancora definita completamente. Essa dipenderà, in misura maggiore degli altri indici, dalla tecnica scelta per la stampa del volume.[15]

I contesti saranno raggruppati sotto le rispettive forme, ordinate alfabeticamente all'interno dei rispettivi lemmi, anch'essi in ordine alfabetico.

Resta da decidere in quale ordine disporre i contesti all'interno di una forma: direttamente in ordine di testo, oppure con un altro criterio; per esempio prima i contesti la cui parola-esponente è presente in (C), o in (C)+(A), o (C)+(B), o (C)+(A)+(B), poi

[14] Le formule possibili sono più di una, e non posso qui illustrarle tutte. La più semplice, che indico a puro titolo di esempio, è probabilmente

$$\frac{fc - fa}{fc}$$

dove *fc* è la frequenza di una parola in (C) ed *fa* è la frequenza della stessa parola in (A). Analogamente per (B).

[15] Le tecniche in uso per la pubblicazione dei risultati degli spogli sono riconducibili a 4 tipi fondamentali:

1. I tabulati degli indici, delle concordanze, ecc., vengono affidati a una tipografia che li ricompone con i mezzi tradizionali.
2. Dai tabulati vengono ricavate, previo un opportuno rimpicciolimento, delle matrici per *l'offset*.
3. Il calcolatore è collegato con una fotocompositrice che registra gli indici, le concordanze, ecc., su un film che viene poi stampato, per lo più in *offset*.
4. Il calcolatore, collegato a un impianto di tipo COMP, registra gli indici, ecc., su *microphiches* che vengono distribuite a chi le richiede.

La soluzione 1 sembra fin d'ora da scartare, perché molto più costosa e soprattutto più onerosa delle altre, richiedendo la rilettura delle bozze. La soluzione 4, che probabilmente si affermerà nel prossimo futuro, sarebbe opportuna soprattutto in quei paesi nei quali le *microphiches* sono già un sistema standard di diffusione delle informazioni. Si dovrà quindi scegliere, probabilmente, tra le soluzioni 3 e 4. La fotocomposizione è molto più cara, ma, permettendo una varietà di caratteri e di corpi paragonabili a una normale *linotype*, consentirebbe tutta una serie di accorgimenti grafici per rendere più agevole la consultazione e più maneggevole il volume.

quelli delle parole presenti in (A) o in (A)+(B), poi quelli delle parole presenti solo in (B).

Resta da decidere anche se le concordanze comprenderanno sia le parole piene che le parole vuote, o se, invece, comprenderanno solo le piene, mentre delle vuote sarà stampato solo l'*index locorum*.[16]

Un'altra decisione riguarda la rappresentazione delle varianti. Sembra probabile che stamperemo per esteso, l'uno sotto l'altro, i due o tre versi di una UC, come nella figura 3, ma non è escluso che, quand'è possibile, si stampi solo il verso di (C) inserendovi, tra parentesi, le varianti di (A) e di (B).

3.1.9. *Rimario.* — Se la parola in rima di C non ha varianti, nel rimario appare solo C, con l'indicazione, per mezzo del codice di situazione, della presenza di eventuali varianti relative alle altre parole del verso. Se alla parola *c* in posizione di rima in C corrisponde una variante *a* nella rima di A, si riporta la coppia di versi C, A sia sotto la rima di *c* che sotto quella di *a*. (Si procede in modo analogo per varianti in B, o in A e in B.)

Tutti i versi di una redazione che non hanno corrispondenza nelle altre, compaiono da soli sotto la propria rima, con accanto l'avvertimento, in codice, che si tratta di versi rifatti.

L'ordinamento terrà conto dei seguenti campi: 1) rima; 2) parola in rima, parola precedente la rima, e così via fino alla parola iniziale; 3) ordine di testo.

3.2. *Procedura operativa di spoglio*

Per produrre le concordanze, gli indici e il rimario delle tre redazioni del *Furioso*, deve essere eseguita una serie di operazioni, il cui coordinamento logico e cronologico è rappresentato nel diagramma operativo (*flow-chart*), che si riporta alla fine del saggio, distribuito per comodità di consultazione in 5 diverse tavole, ma da considerarsi essenzialmente un diagramma unico. Ogni singola operazione è contrassegnata con un numero.

[16] Com'è noto, l'*index locorum* consiste nell'elencare, sotto il lemma (o la forma), tutti i riferimenti dei luoghi ove esso (essa) occorre. È noto anche che le parole vuote coprono circa il 50% di qualsiasi testo, e che le prime 100 parole più frequenti ne coprono circa il 60%. Il trattamento che viene riservato alle parole vuote e alle parole piene di altissima frequenza è il risultato di due esigenze contrapposte: da un lato si cerca di ridurre le dimensioni del volume, dall'altro si vuole fornire una documentazione completa e non selettiva. L'*index locorum* è la forma più semplice di compromesso. Sono state proposte anche soluzioni più elaborate, che consistono nel contestualizzare in modo diverso categorie diverse di parole: per le preposizioni riportare le 3 parole immediatamente seguenti; per le congiunzioni la tripla di parole di cui la congiunzione è al centro, ecc.

Possiamo raggruppare le operazioni in 3 fasi principali:

a) op. 1 - 6. Si trascrive, in forma leggibile dal calcolatore, il testo base: nel nostro caso (C).

b) op. 7 - 13. Si segnalano le differenze tra il testo base e ciascuna delle altre versioni. Il calcolatore, in base a queste segnalazioni, modifica il testo base, ricostruendo così le altre versioni ad una ad una.

c) op. 14 - 34. Il calcolatore produce concordanze ed indici 'sinottici', elaborando contemporaneamente il testo base e le sue versioni.

Le operazioni 1 - 6 non differiscono da quelle eseguite nella normale procedura di spoglio in uso presso gli Utenti della Divisione Linguistica del CNUCE.[17]

Le operazioni 7 - 13 sono invece specifiche della elaborazione di versioni diverse di uno stesso testo. Esse sono state studiate in stretta collaborazione con C. Segre per la elaborazione del *Furioso*, e sono state programmate in modo da poter servire anche ad altri Utenti.[18]

Le operazioni 14 - 34 seguono da vicino la normale procedura di spoglio, ma hanno richiesto una serie di modifiche ai program-

[17] Nel 1968 la Direzione del CNUCE (allora Istituto dell'Università di Pisa, e oggi Istituto del CNR decise di istituire una Divisione Linguistica con lo scopo di promuovere e di eseguire delle ricerche nel settore della linguistica matematica e computazionale, di assicurare le relazioni scientifiche e gli scambi di informazioni tra i ricercatori italiani e stranieri, di organizzare e coordinare l'attività didattica per diffondere le metodologie e le esperienze, e infine di prestare assistenza scientifica e tecnica agli Istituti utenti. A differenza dell'assistenza fornita ad altre categorie di utenti, che si limita in genere ad assicurare gli strumenti e il tempo di calcolo, l'assistenza riservata ai linguisti e agli umanisti comprende anche la consulenza e la collaborazione sul piano della definizione scientifica e tecnica dei progetti e l'analisi, la stesura e la messa a punto dei programmi necessari per le elaborazioni (cfr. A. Faedo, *Discorso di apertura* in A. Zampolli, *Linguistica matematica e calcolatori*, Firenze 1973, pp. VII-IX). Questa decisione ha contribuito in maniera determinante allo sviluppo qualitativo e quantitativo delle ricerche e delle applicazioni in Italia, testimoniato dal numero (oltre 50) degli Istituti Universitari e di Ricerca utenti della Divisione Linguistica del CNUCE, e dalla varietà dei loro progetti, che vanno dalla lessicografia alla filologia, alla dialettologia, alla linguistica storica, alla psicolinguistica, alla demologia, alla linguistica matematica, alla linguistica teorica, ecc. Sono stati registrati per il calcolatore più di 5.000 testi, in oltre 20 lingue, per un totale complessivo di oltre 80.000.000 di parole. I vantaggi scientifici e economici offerti dalla riunione di tanti progetti sono incommensurabili.

Si pensi per es. al fatto che l'adozione di uno stesso sistema generale di registrazione e di elaborazione permette di utilizzare dei programmi già sperimentati e dà la possibilità ad ogni ricercatore di utilizzare i testi o i programmi elaborati per altri progetti.

In pratica, un ricercatore può condurre le sue ricerche in una grande biblioteca elettronica, anziché sul solo testo che egli stesso ha registrato. La presenza di un insieme di *corpora* così vasti, registrati e analizzati con uniformità di criteri scientifici e tecnici, assicura la comparabilità tra i dati statistici rilevati nei vari testi, e permette di verificare, finalmente, le ipotesi e i modelli statistici formulati fino a qui su una quantità insufficiente di dati. Un esempio di quanto ho detto è fornito anche dallo spoglio del *Furioso*. Infatti abbiamo utilizzato il testo della redazione (C), che era già stato perforato e corretto dall'Accademia della Crusca per l'Opera del Vocabolario, ed abbiamo realizzato così un notevole risparmio di tempo e di spesa.

[18] I programmi che abbiamo appena approntati per l'esperimento qui descritto sono stati applicati per lo spoglio di un testo latino di cui esistono varie redazioni.

mi esistenti per adattarli alla presentazione 'sinottica' dei materiali provenienti dalle diverse versioni.

Op. 1. — Per ottenere che il testo possa essere letto dal calcolatore, si suole riprodurlo su schede, su nastri di carta o magnetici, per mezzo di una macchina perforatrice o di un terminale. Per il *Furioso*, come si è detto, ci siamo serviti delle schede del testo (C) predisposte dall'Accademia della Crusca per l'opera del Vocabolario, sulle quali erano riportate le parole, la punteggiatura, la divisione sia per canto, ottava e verso, sia per pagine e righe ('schede-testo'). Nel trasportare su scheda mediante perforazione i caratteri del testo stampato sono state seguite le norme convenzionali proposte dalla Divisione Linguistica del CNUCE, di cui nelle elaborazioni elettroniche del *Furioso* si sono usati i programmi e le procedure non meno che gli elaboratori IBM 360/67 e IBM 370/158.[19]

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 essa è dovuta a un errore proprio o 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.[20] Via via che legge

[19] Di norma il numero dei caratteri diversi da rappresentare oltrepassa il centinaio (comprendendo nel conto lettere dell'alfabeto, cifre numeriche, segni diacritici e interpunzioni; il 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 del *Furioso*, 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 biunivoca tra i 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.

[20] In altre ricerche tali unità sono costituite dai grafemi, dai fonemi, dalle sillabe, dai morfemi, dai sintagmi, o da altro ancora.

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.[21]

Op. 4. — La lista-testo prodotta dal calcolatore viene collazionata con il testo originale.[22] Gli errori trovati vengono messi in evidenza in fogli particolari, con un modulo opportunamente predisposto, nel quale 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 errate.[23]

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

Op. 7. — Il calcolatore stampa il testo [nel nostro caso, come

[21] 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. Per ovviare almeno in parte a questi inconvenienti, il CNUCE ha dotato le proprie stampatrici di catene speciali, che sono state costruite appositamente dalla IBM, le quali contengono maiuscole, minuscole, interpunzioni, e i segni diacritici più frequenti.

[22] 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 solito 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.

[23] In media, per perforare un testo di circa 10.000 parole (nel *Furioso* sono 120.000 circa), 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; 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. La concentrazione dei progetti di spoglio e in genere di elaborazione di testi in lingua naturale presso il CNUCE ha garantito la adozione di una norma comune di registrazione e di analisi dei testi. Si è così creata in Italia una situazione di fatto che anche i paesi tecnologicamente più progrediti ci invidiano, e che viene presa a modello dalle organizzazioni internazionali del settore.

si è detto, la redazione (C)], una parola per riga, in un prospetto predisposto in modo da facilitare al filologo la trascrizione delle varianti e l'eventuale indicazione dei versi da assegnare come contesto alle singole parole.

Come si vede dall'esempio riportato nella figura 1, il prospetto è diviso in colonne e in righe delimitate da trattini.

Ogni riga si riferisce a una parola di C, che il calcolatore stampa nella colonna 3, preceduta dal proprio numero record (colonna 1) e tipo record (colonna 2).[24]

Il filologo deve riempire le altre colonne, le quali sono così suddivise:

le colonne 4 — 5 si riferiscono a (C); le colonne 7 — 13 e 21 ad (A), le colonne 14 — 20 e 22 a (B).

Op. 8. — Il filologo trascrive nella colonna 9 la eventuale variante *a* di *c*, preceduta dal tipo record relativo (col. 8), e nella colonna 16 la eventuale variante *b*, anch'essa preceduta dal rispettivo tipo record (Col. 15).[25] Egli può evitare di trascrivere *b*, se è uguale ad *a*, ma in tal caso deve apporre una croce nella colonna 4, la quale è riservata appunto ad indicare che *a* è uguale a *b*. Se *a* è composta da più parole esse devono essere scritte una sotto l'altra (col. 9) e numerate progressivamente (col. 7). Se la variante di *c* in A non è una parola (o un gruppo di parole), ma consiste nella soppressione di *c*, le colonne 8 e 9 resteranno vuote, e nella colonna 7 dovrà essere scritto uno *zero*. In modo del tutto analogo ci si comporta per *b*, la colonna 14 avendo la stessa funzione della colonna 7 per *a*. Nei casi di versi interamente rifatti in (C) rispetto ad (A),[26] si scrive il verso A una parola per riga, nelle colonne 7, 8 e 9, con l'apposizione di una croce in colonna 21, destinata appunto a contraddistinguere i rifacimenti.

Allo stesso modo, se C è rifatto interamente rispetto a B, si scrive B nelle colonne 14 - 16 e si appone una croce a colonna 22. La colonna 10 viene utilizzata quando una qualsiasi parola di (C) appare identica in (A), ma spostata all'interno dello stesso verso

[24] In effetti, una riga può essere occupata, oltre che da una parola, da un'altra unità di informazione, come punteggiatura, riferimento, segno di fine riga, ecc.
Il tipo record distingue appunto diverse categorie di informazioni. Per es., A punteggiatura, R riferimento, P numero di pagina, 'blank' parola, ecc.

[25] Seguo anche qui la notazione secondo la quale le minuscole in corsivo rappresentano una parola, le maiuscole un verso, e le maiuscole tra parentesi il testo di una redazione.

[26] Il fatto che C e A (o B) siano scritti l'uno accanto all'altro non serve a metterli in relazione tra loro, ma permette solo di ricostruire il testo di A (o di B) nella successione voluta.

o in un verso attiguo. Questa parola viene trattata, in A, in tutto e per tutto come una comune variante, ma la si collega alla corrispondente parola di C, scrivendole accanto il numero record di quest'ultima a colonna 10. Vedremo più avanti la funzione di questa informazione (LINK) nella procedura. Analogo impiego rispetto a (B) ha la colonna 17.

Fino a questo punto l'intervento del filologo ha il solo scopo di inserire le varianti di (A) e di (B) rispetto a (C), cioè di comunicare al calcolatore tutte le informazioni necessarie per ricostruire, prendendo come base (C), le redazioni (A) e (B): questo intervento è necessario sia nella procedura automatica di riduzione, sia in quella manuale. Le colonne 5 - 6, 11 - 13, e 18 - 20 invece devono essere compilate solo se, come nel caso del *Furioso*, viene scelta la seconda procedura. Poiché nel contesto di *c* comparirà in ogni caso automaticamente C, il filologo deve specificare solo se richiedere anche A (croce a col. 5) e B (croce a col. 6). I versi richiesti nel contesto di *a* devono invece essere indicati tutti dal filologo: una croce apposta nelle colonne 11, 12, 13 indica che vengono richiesti rispettivamente C, A, B. Analoga funzione hanno, per *b* le colonne 18, 19, 20.

Poiché, come si è detto, le parole di un verso rifatto ricevono automaticamente questo solo verso come contesto, non occorrerà compilare le colonne 5, 6, 11-13, 18-20.

Op. 9. — Per ogni riga nella quale il filologo è intervenuto, l'operatore perfora su una apposita 'scheda-variante' (o introduce da terminale) le informazioni annotate nel prospetto e il numero record che contraddistingue univocamente la riga.

Op. 10. — Il calcolatore legge le schede varianti e il nastro-testo, e ricostruisce in memoria, una dopo l'altra, le triple di versi C, A, B.

Se è stata scelta la procedura automatica, il programma semplifica le triple e produce un RPC per ogni parola indicata dalle regole di contestualizzazione.[27] Se invece è stata scelta la procedura manuale, il programma genera un RPC per ogni parola di (C), associando un contesto che contiene certamente C, e A e/o B solo su richiesta del filologo. Per le parole di (A) e di (B), invece, come si è detto, porta in contesto solo i versi da esso esplicitamente ri-

[27] Il programma di contestualizzazione non è rigido, ma permette all'utente la scelta tra alcune alternative: innanzitutto tra procedura manuale e procedura automatica; in secondo luogo, all'interno di quest'ultima, tra diversi criteri di riduzione delle ridondanze.

chiesti. In ogni caso, al termine di questa operazione, il RPC contiene:
— la parola esponente,
— il suo numero record,
— una sigla che indica la redazione da cui proviene,
— il suo riferimento,
— il codice che indica la situazione iniziale della tripla prima della semplificazione,
— il codice che indica la situazione della parola e delle sue varianti,
— l'eventuale codice che indica "verso completamente rifatto",
— l'UC: cioè uno, due, o tre versi, ciascuno accompagnato dalla sigla che indica la redazione da cui proviene e dall'eventuale LINK.

Contemporaneamente, il calcolatore stampa una lista 'testo +variante' che riproduce, in pratica, la struttura e le informazioni presenti nel prospetto compilato con la op. 7.

In più, dopo l'ultima parola di ogni verso, stampa la relativa tripla di versi, così come li ha ricostruiti, e inoltre sottoscrive, ad ogni parola che sarà portata in esponente, la sigla del verso o dei versi (C, A, B) che ne costituiranno il contesto.

Op. 11. — La lista 'testo+varianti' viene esaminata dal filologo. Se ci sono errori, si correggono le schede varianti e si ripetono le operazioni 9, 10, e 11 limitatamente alle unità errate.

Op. 12. — I RPC vengono ordinati secondo i seguenti campi
1) Ordine alfabetico della parola-esponente
2) Ordine crescente del LINK
3) Ordine crescente del numero record.

Op. 13. — Il calcolatore stampa le concordanze per forma; contemporaneamente ricopia, su un nastro, forma e contesti numerati progressivamente e produce per ogni forma una scheda che contiene la forma stessa e il numero progressivo corrispondente. La figura 2 riproduce un primo esempio di concordanza per forme, in versione largamente provvisoria. Il contesto è stampato nella colonna 5. Se esso consiste di più versi, il verso che contiene la parola esponente è stampato per primo con accanto l'indicazione, in codice, della redazione da cui proviene (col. 3), delle situazioni della tripla (col. 1) e della parola esponente nella tripla (col. 2) prima delle semplificazioni. Le colonne 1 e 2 sono invece vuote in corrispondenza al secondo e al terzo verso della tripla, e ciò contribuisce a mettere in evidenza le coppie e le triple di versi che formano una UC unica. Le stesse colonne contengono invece un asterisco se si tratta di verso rifatto. La sigla a colon-

na 3 indica da quale redazione proviene il verso. La colonna 4 è normalmente vuota; contiene un asterisco solo se la parola esponente in questione è legata da un LINK alla parola esponente precedente.[28] La colonna 6 contiene, in corrispondenza al primo verso di una UC, il riferimento (canto, ottava, verso; pagina e riga).

Op. 14. — Le indicazioni dei lemmi e le eventuali indicazioni accessorie vengono scritte a mano sulla lista delle concordanze accanto a ciascuna forma.[29]

Op. 15. — Si perforano queste indicazioni sulle 'schede-forma' preparate dal calcolatore (o si introducono da terminale).

Op. 16. — Il nastro delle concordanze per forma viene ricopiato con l'aggiunta, per ogni parola, del lemma perforato nella scheda relativa.

Op. 17. — Si ordinano le parole per lemma, forma, riferimento.

Op. 18. — Si stampano le concordanze per lemma.

Op 19. — Sulle concordanze lemmatizzate il ricercatore esegue alcuni controlli per accertare l'esattezza della lemmatizzazione.

Op. 20. — 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. 21. — Il nastro dei lemmi viene riordinato secondo l'ordine decrescente delle frequenze.

Op. 22. — Si stampano i lemmi in ordine di frequenza.

[28] Il codice LINK ha una duplice funzione. Per quanto concerne gli indici di frequenza, permette di considerare a parte le varianti che consistono non nel modificare ma semplicemente nello spostare la posizione della parola.

Per quanto riguarda le concordanze, esso si traduce in un contrassegno il quale avverte il lettore che la parola esponente *a* (o *b*) in questione è collegata alla parola *c* che la precede immediatamente nelle concordanze, rispetto alla quale è appunto cambiata semplicemente di posizione.

[29] Non abbiamo ancora deciso se impiegare o no nella lemmatizzazione il *Lessico Automatico dell'Italiano* approntato dalla Divisione Linguistica del CNUCE con il patrocinio del Comitato 08 del CNR. Il *Lessico Automatico*, che è costituito dalla registrazione su nastro magnetico di circa un milione di forme corrispondenti a circa 150.000 lemmi, è stato predisposto principalmente per consentire la lemmatizzazione semiautomatica dell'italiano contemporaneo. Un giudizio sull'opportunità di impiego deve basarsi su dati che solo un esperimento può fornire: e cioè quale sia la percentuale delle parole del *Furioso* presenti nel *Lessico Automatico*, e soprattutto in che misura il sistema di omografie presenti nella lingua del *Furioso* sia diverso da quello dell'italiano contemporaneo.

Per maggiori informazioni sul *Lessico Automatico* si veda: A. Duro - A. Zampolli, *Analisi Lessicali mediante elaborazioni elettroniche*, in *Atti del Convegno sul tema: L'automazione elettronica e le sue implicazioni scientifiche, tecniche e sociali* (Accademia Nazionale dei Lincei, 16-19 ottobre 1967), Roma 1968, pp. 132-139; A. Zampolli, *Linguistica matematica e calcolatori*, Firenze 1973, pp. 133-199; A. Zampolli, *Problemi di linguistica applicata computazionale*, Pisa 1974, pp. 15 sgg. (in queste pagine sono discusse anche le possibilità di distinguere automaticamente i diversi significati degli omografi).

Op. 23. — Dal nastro delle forme per frequenza si calcolano e si stampano progetti statistici particolari (vedi 3.1.7).

Op. 24. — Le parole vengono ridisposte in un nuovo nastro in ordine alfabetico di forma e, in casi di omografia, di lemma.

Op. 25. — 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. 26. — Viene generato un nuovo nastro nel quale le forme sono ordinate secondo l'ordine decrescente delle rispettive frequenze.

Op. 27. — Si ottiene l'elenco delle forme in ordine di frequenza.

Op. 28. — Dal nastro parole + contesti si ricopiano, su un nuovo nastro, solo le parole in posizione di rima.

Op. 29. — Si ordinano queste parole per rima,[30] parola finale, parola penultima ... parola iniziale, riferimento.

Op. 30. — Si stampa il rimario.

A seconda della tecnologia di stampa che verrà scelta, possono essere, o no, richieste altre operazioni: per esempio, se useremo

[30] In altri termini, la chiave di questo ordinamento consiste nei campi seguenti:

— rima
— ultima parola del verso
— penultima parola del verso
. . . . . .
— prima parola del verso
— sigla della redazione cui il verso appartiene
— riferimento

La rima è definita, normalmente, come la sequenza di fonemi compresa tra l'ultima vocale tonica e la fine del verso.

In alcune lingue, la sola rappresentazione possibile della rima è costituita dalla sua trascrizione fonematica; in altre ci si può servire della normale ortografia.

Nel primo caso, la trascrizione fonematica può essere operata a mano oppure consultando, qualora sia disponibile, un *lessico automatico* che contenga la trascrizione fonematica di ogni forma, o almeno le informazioni fonetiche necessarie per applicare un algoritmo di trascrizione automatica (per es. queste informazioni, per l'italiano, sono: la posizione dell'accento, l'apertura di *e*, *o*, la sonorità di *s*, *z*, la semivocalità, la pronuncia non palatalizzata dei nessi *gl*, *gn*).

Nel secondo caso, basta contrassegnare l'ultima vocale tonica del verso.

Per l'italiano contemporaneo il calcolatore trova questa informazione nel *Lessico Automatico* citato.

Per testi di epoche precedenti, adottiamo una procedura semiautomatica basata sulla constatazione che le parole piane sono le più frequenti in italiano. Essa prevede che l'utente contrassegni (ad es. con un puntino sottoscritto nel testo o nella lista testo — nel caso del *Furioso* nel prospetto per l'inserimento delle varianti) l'ultima vocale tonica del verso solo quando non coincide con il penultimo grafema vocalico. (Naturalmente la presenza di un eventuale accento grafico, come nelle parole tronche, rende superfluo questo contrassegno.)

Il calcolatore, in assenza di contrassegni o di accenti grafici, calcolerà la rima a partire dal penultimo grafema vocalico.

una fotocompositrice, occorrerà trasformare i nastri che contengono i risultati finali inserendo gli opportuni comandi per la giustificazione della riga, della pagina, per la gestione dei diversi corpi, ecc.

Ritengo doveroso ringraziare il dottor E. Picchi della Divisione Linguistica del CNUCE, il quale non solo ha provveduto alla analisi della procedura e alla predisposizione del software relativo, ma ha anche contribuito alla fase di studio generale dello schema di lavoro qui esposto.

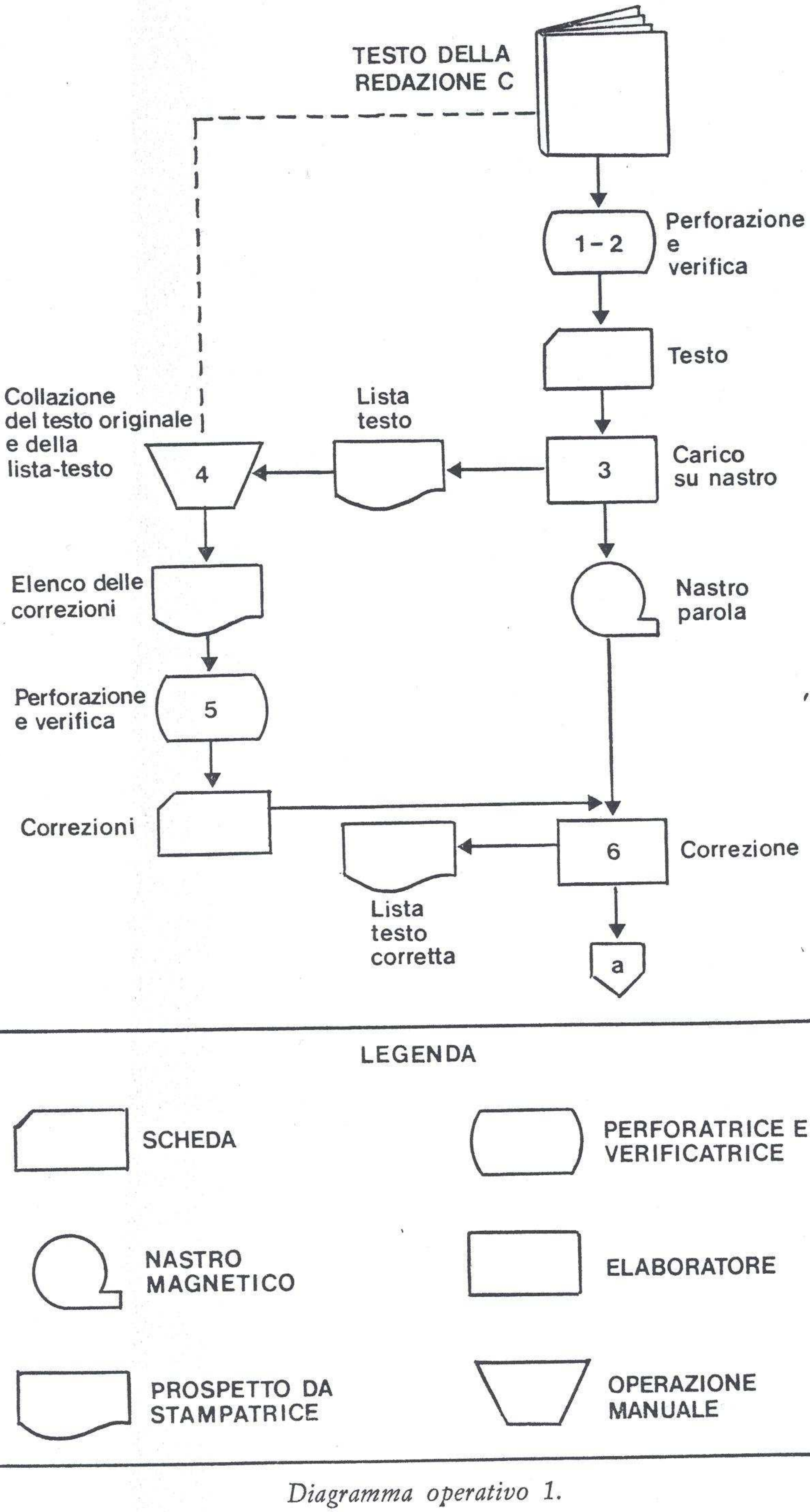

Diagramma operativo 1.

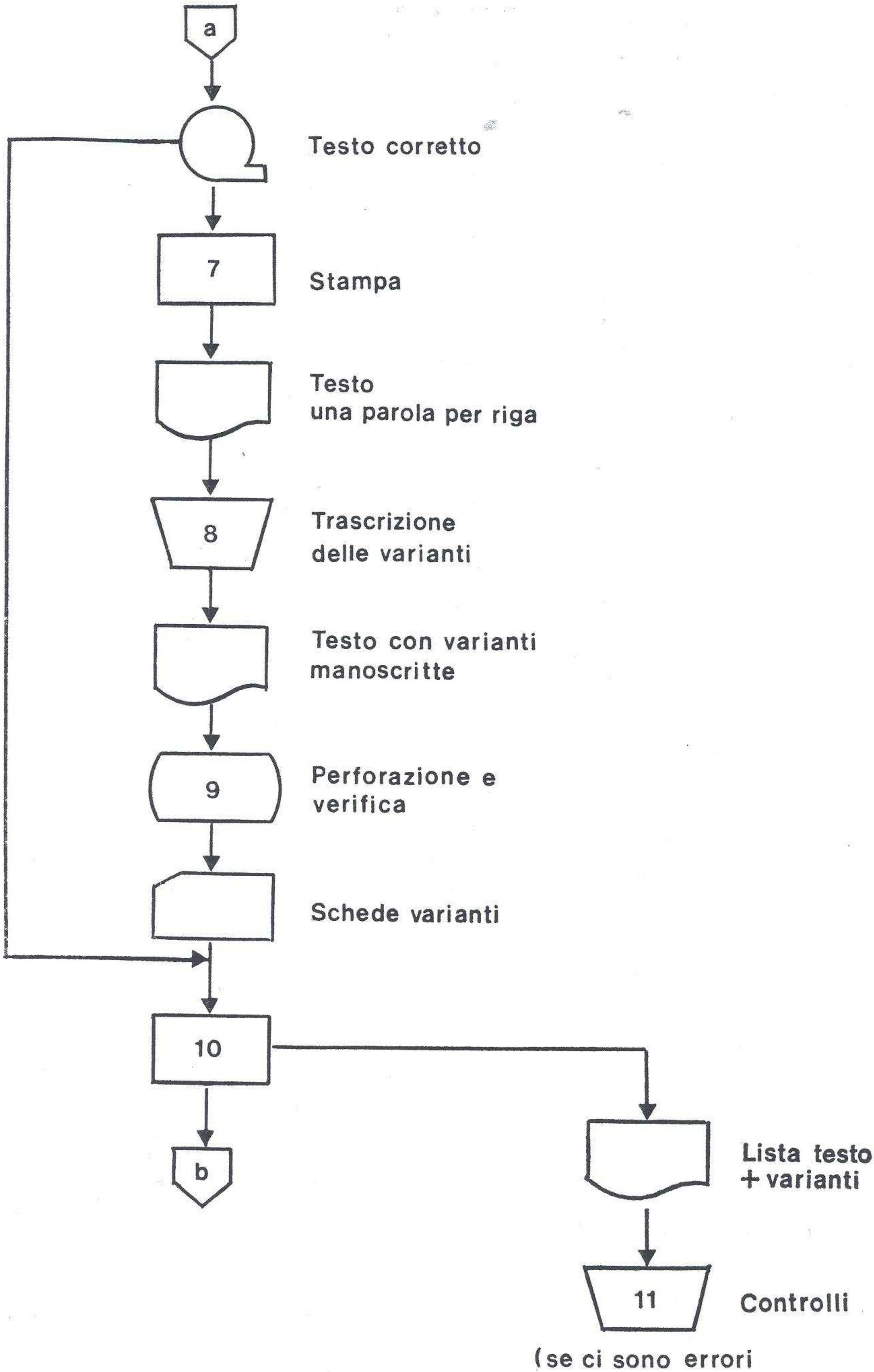

Diagramma operativo 2.

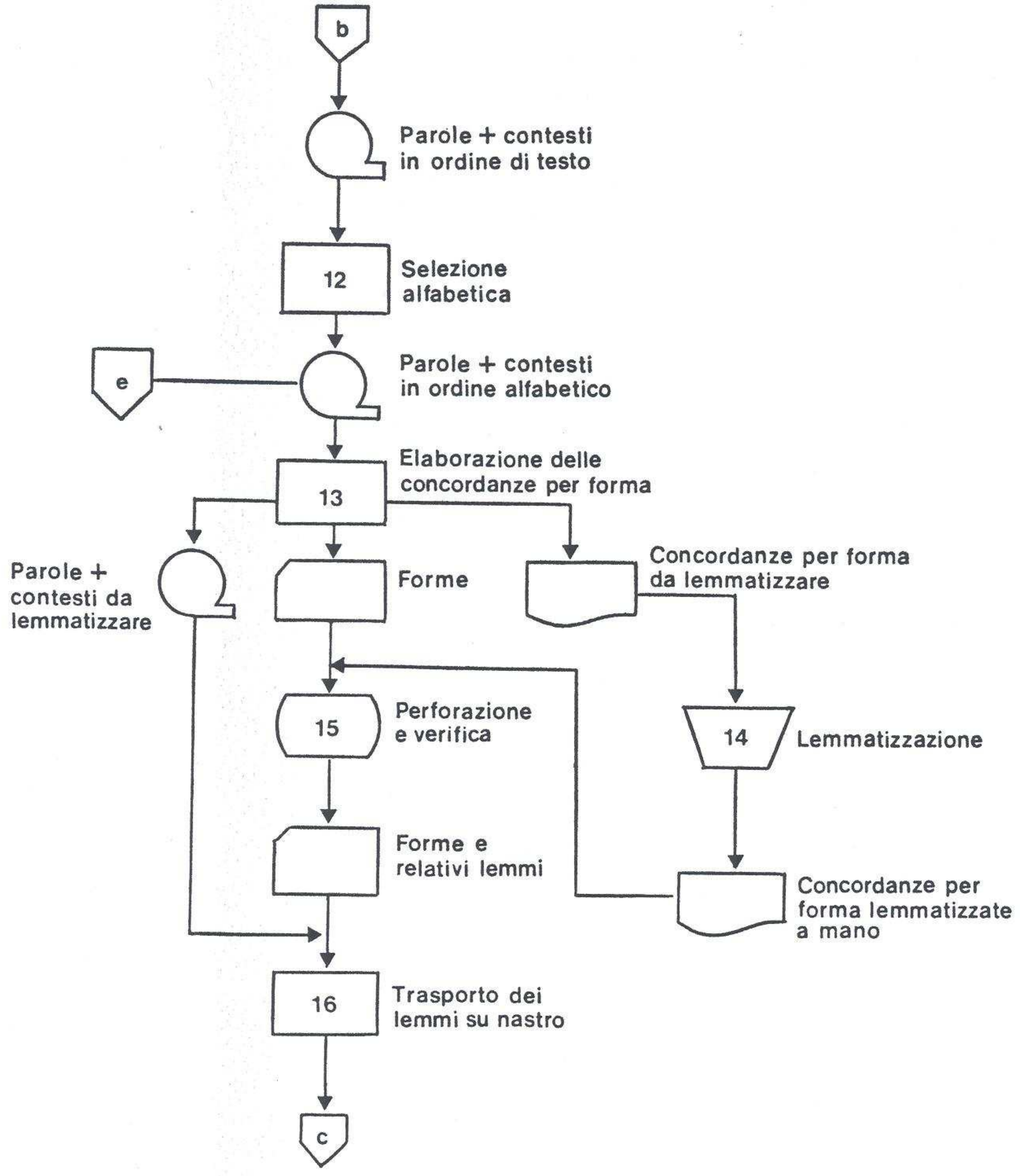

Diagramma operativo 3.

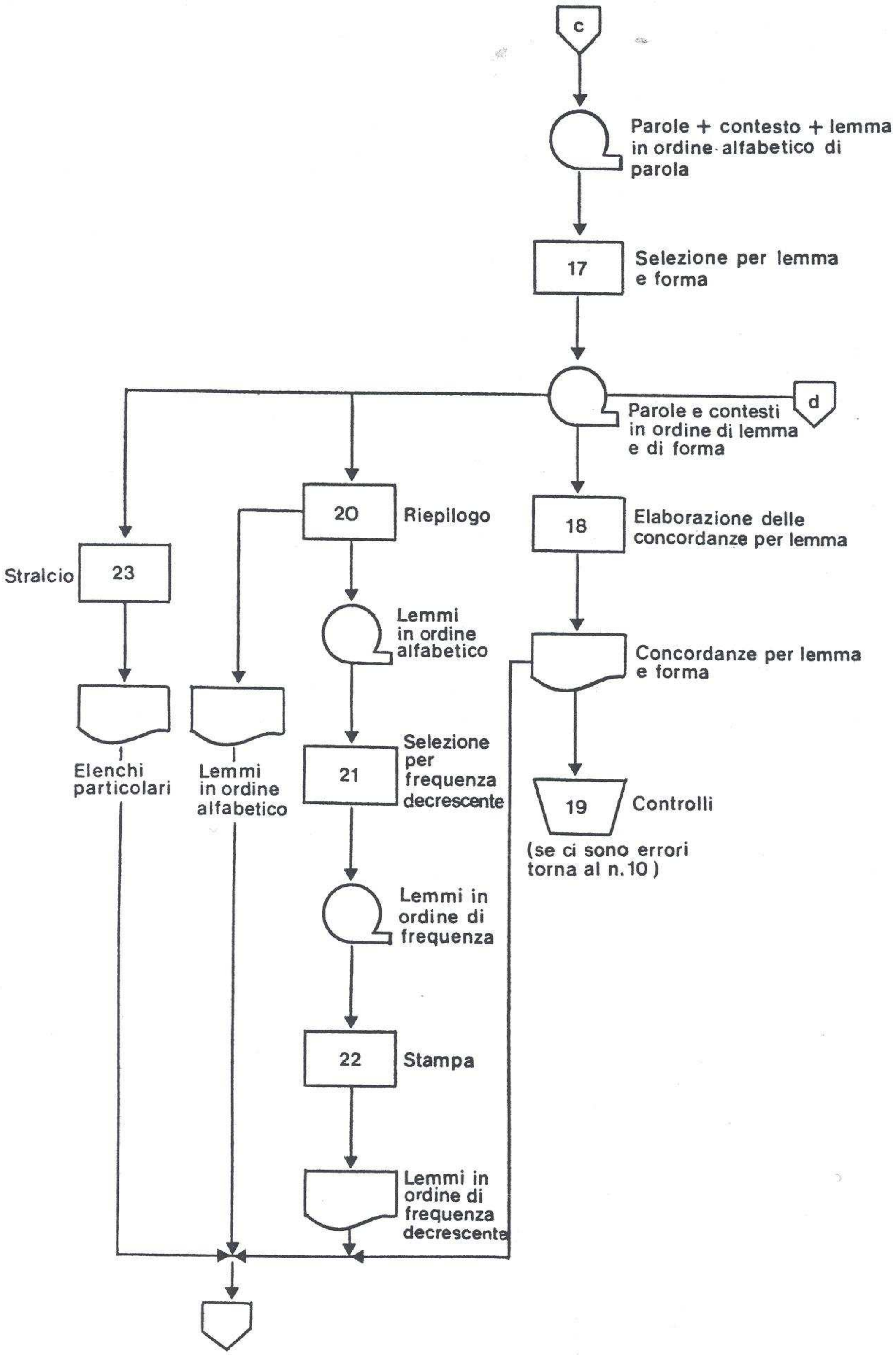

Diagramma operativo 4.

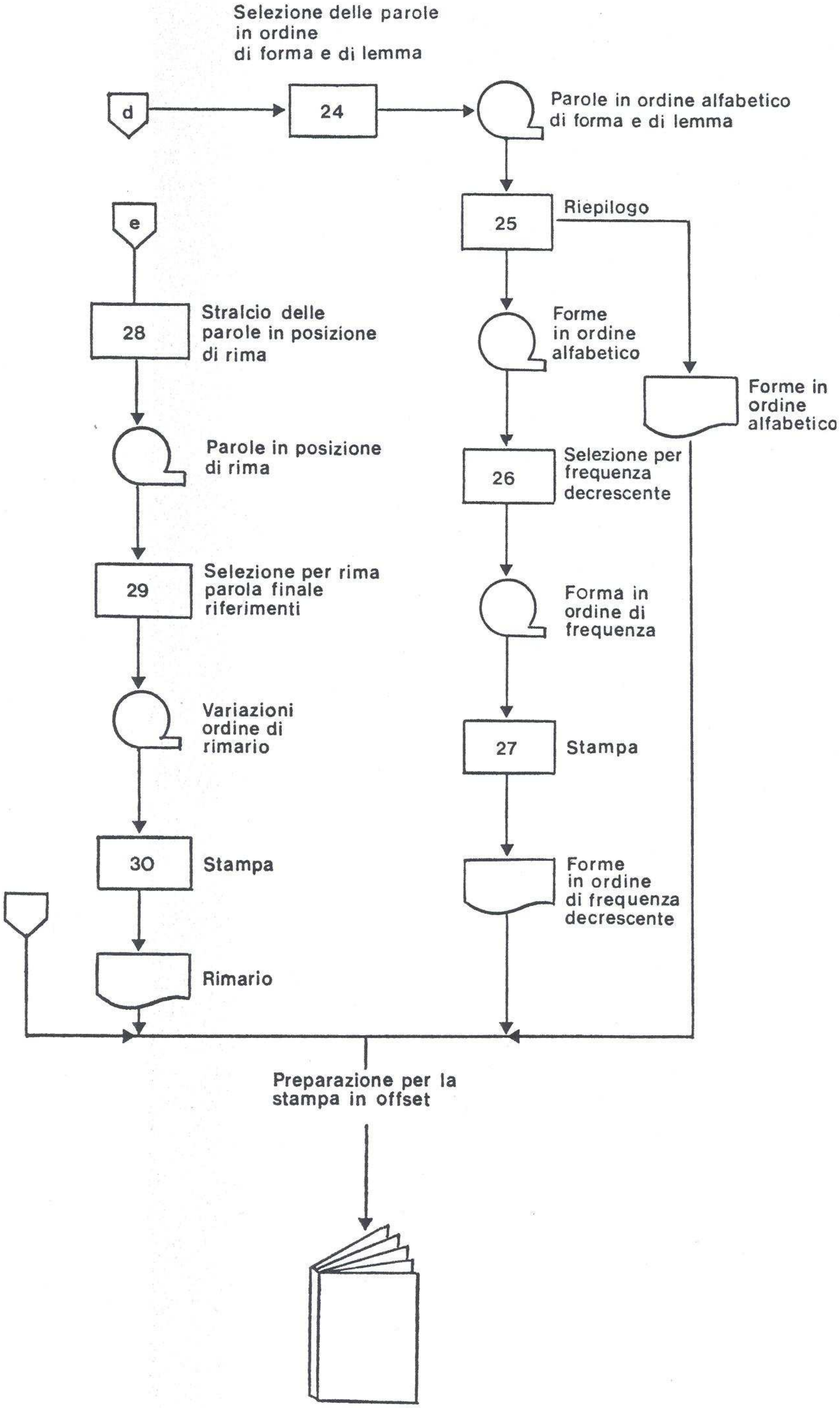

Diagramma operativo 5.

1 N. REC.	2	3 — C —	4 5 6 7 8 = A B E	9 — A —	10	11 12 13 14 15	16 — B —	17	18 19 20 21
019645		tanto			LINK	A C A B E		LINK	B C A B
019646		possente							
019647	1	.							
019648	R	R							
019649	Z	1 , , 0027:							
019650		Disegnando	X 1 2 3	LA DONNA		X X X X X X			
019651		levargli	X X 1 2	CHE GLI		X X X X X X	LEVARLI		X X X
019652		ella	X 1 2	VUOL TOGLIER		X X X X X X			
019653		la							
019654		testa							
019655	1	,							
019656	R	R							

Figura 1. — Prospetto per la segnalazione delle varianti

1	2	3	4	5	6	7
422	dolente					
	1	1	CAB		Il re, dolente per Ginevra bella	4, 60, 1 = PAG. 77, 01.
423	dolse					
	3	1	CB		e furon di lor molte a chi ne dolse;	4, 39, 7 = PAG. 71, 31.
424	domande					
	5	4	C		Pur ch'uscir di là su non si domande,	4, 32, 1 = PAG. 70, 01.
			A		Pur ch'uscir di là su non se dimande,	
			B		Pur che uscir di là su non se dimande,	
425	domandò					
	5	5	C		fêro a Rinaldo, il qual domandò loro	4, 55, 2 = PAG. 75, 26.
			A		fêro a Rinaldo, il quale intrò con loro	
			B		fêro a Rinaldo, il qual dimandò loro	
426	donna					
	4	4	C		Disse la donna: — O gloriosa Madre,	4, 03, 6 = PAG. 64, 22.
			AB		— O re del cielo, o gloriosa Madre,	
	4	4	AB		: disse fra sé la donna — che fia questo? —	4, 03, 7 = PAG. 64, 23.
			C		o Re del ciel, che cosa sarà questa? —	
	4	1	C		Vede la donna un'alta maraviglia,	4, 04, 5 = PAG. 63, 05.
	1	1	CAB		La donna il tutto ascolta, e le re giova,	4, 08, 1 = PAG. 64, 01.

Figura 2. — Concordanze per forma

LEGENDA

1 Progressivo della forma.
2 Codice di situazione della tripla di versi:
  1 C=A =B
  2 (C=A)=B
  3 (C=B)=A
  4 C=(A=B)
  5 C=A=B
3 Codice di situazione della parola nella tripla:
  1 c=a =b
  2 (c=a)=b
  3 (c=b)=a
  4 c=(a=b)
  5 c=a=b
4 Sigla della redazione (o delle redazioni) da cui proviene il verso. Se l'UC consiste in una coppia o una tripla di versi, la parola esponente proviene dal primo di essi.
5 Eventuale contrassegno di LINK.
6 Contesto.
7 Riferimento (Canto, ottava, verso, pagina, riga).