NOTA TECNICA ALL'ESPERIMENTO

A. Descrizione dei tabulati

Nello schema premesso a ogni descrizione, le lettere minuscole indicano le zone o colonne, in ordine da sinistra verso destra, nelle quali sono distribuiti i dati del tabulato.

1. Testo

a - numero progressivo del componimento (o più esattamente del foglio che lo contiene) nella raccolta Barbi.
b - componimento.

E' la ristampa, (*) eseguita dal calcolatore, dei componenti prescelti per l'esperimento.

Essa è sembrata utile, nonostante siano allegate fotocopie dell'originale, sia perché di fatto costituisce una tappa effettiva del lavoro, che permette una accurata revisione della perforazione; sia per documentare come è stata eseguita la trascrizione; sia a riscontro di eventuali errori malauguratamente rimasti nonostante tutti i controlli; sia per facilitare il reperimento dei luoghi citati nei tabulati successivi.

---

* - Caratteri di stampa

Le stampatrici normalmente collegate agli elaboratori elettronici adoperati per ricerche matematiche, fisiche, commerciali, industriali, ecc. sono dotate di 48 (al massimo 60) diversi caratteri di stampa : le 26 lettere dell'alfabeto inglese maiuscolo, 10 cifre, 11 segni diversi, quali ad esempio : * + - % ecc.

La stampa di elaborazioni di testi naturali, richiede un numero molto maggiore di caratteri. Per questo motivo il CNUCE di Pisa disporrà in un prossimo futuro di una catena di 120 caratteri comprendente maiuscole, minuscole, punteggiatura, segni diacritici (accenti, dieresi, notazioni metriche, ecc.) che potranno essere stampati sopra e sotto le lettere.

Abbiamo dovuto usare per il nostro esperimento una stampatrice dotata di soli 48 segni, con lettere solo maiuscole, e nella quale un medesimo segno (') rappresenta sia l'accento sia l'apostrofo.

La catena CNUCE risponderà sicuramente a tutte le esigenze del lavoro definitivo.

2. Incipitario I : versi elencati secondo l'ordine alfabetico (*) della parola iniziale.

a - parola iniziale del verso.
b - verso.
c - riferimento (n. di componimento e n. di verso).

In caso di uguaglianza della parola iniziale, la sequenza è stabilita tenendo conto anche delle parole seguenti. Il calcolatore ha ordinato cioè i versi considerando come lettere anche gli spazi (che dividono le parole), gli apostrofi e gli accenti. Lo spazio precede l'apostrofo e l'accento, che a loro volta precedono le lettere.

Una possibile alternativa, di fatto applicata in alcuni incipitari, consiste nel considerare per l'ordinamento solo le lettere, trascurando spazi, accenti e apostrofi.

3. Incipitario II : versi iniziali di componimento elencati secondo l'ordine alfabetico della parola iniziale.

Vale tutto quanto detto per l'incipitario I con la differenza che qui sono compresi solamente i versi iniziali di ciascun componimento.

4. Concordanze delle forme

a - numero d'ordine della forma nella sequenza alfabetica.
b - forma.

---

* - Criteri dell'ordinamento alfabetico

A motivo del brevissimo tempo a disposizione, è stato necessario adoperare oltre a programmi scritti per questo esperimento, anche programmi predisposti per lavori precedenti, i quali prevedono particolari criteri di controllo della sequenza alfabetica. Come conseguenza, nei tabulati allegati, l'accento e l'apostrofo entrano nell'ordinamento alfabetico come una vera e propria lettera della parola, che precede nella sequenza tutte le altre.

Così 'fond, e 'l, 'n ecc. precedono a, abbada, abbiamo, ecc.

c - contesto : l'intero verso nel quale la forma occorre.
I tre asterischi sostituiscono la parola in esponente.
d - riferimento (numero di componimento e numero di verso).

Nei termini stessi della definizione di concordanza sono impliciti i principali problemi di struttura.

La concordanza è una lista delle occorrenze dei diversi elementi lessicali di un testo, ordinate in una sequenza conveniente e accompagnati ciascuna da una appropriata porzione del suo contesto.

L'unità lessicale sotto la quale raggruppare le occorrenze deve essere la forma oppure il lemma? Se si sceglie la forma, essa deve essere definita semplicemente come individualità grafica, o come unità grafico-semantica, con conseguente distinzione delle forme omografe? (Se si sceglie il lemma, la sua definizione è ancora più complessa).

In quale ordine le diverse occorrenze devono essere elencate sotto il relativo esponente? Ad esempio, se l'esponente è il lemma, le occorrenze devono essere raggruppate per forma (ad es. sotto il lemma avere tutti gli abbiamo, poi gli avete, poi gli avevamo, ecc.); oppure direttamente in ordine di apparizione nel testo; oppure, nel nostro caso, in ordine di incipitario?

I criteri di delimitazione del contesto attualmente in uso nei Centri specializzati possono essere raggruppati nei tipi seguenti :

1. la parola esponente è sempre al centro del suo contesto: ossia è sempre preceduta e seguita da un egual numero di battute o di parole;
2. il contesto è scelto in base alla natura della parola: per le parole grammaticali è un trinomio, per le preposizioni sono prese solo le due parole seguenti, ecc.;
3. il contesto è costituito da una intera unità di riferimento: il verso, il paragrafo, il versetto, il comma, ecc.;
4. i limiti del contesto sono segnati in fase di preedizione: il testo viene suddiviso in pericopi, ciascuna delle quali funge da contesto per tutte le parole che la compongono;
5. il contesto è regolato tenendo conto di determinati segni quali l'interpunzione, il cambio di riferimento, ecc.

In questo esperimento si è deciso : 1) di porre come esponente la forma grafica, cioè senza lemmatizzare e senza separare i diversi significati delle forme omografe; 2) di elencare le occorrenze di ciascuna forma in ordine di componimento e verso; 3) di assumere il verso come unità di contesto.

E' allo studio se e come, nel lavoro sulla intera raccolta, operare la lemmatizzazione delle concordanze, e in particolare la separazione degli omografi.

L'insieme di operazioni comunemente raggruppate sotto il termine "lemmatizzazione" richiede una serie di interventi umani che spezzano il ritmo

Per ciò gli sforzi dei ricercatori tendono a ridurre sempre più, per mezzo del cosiddetto "dizionario di macchina", la necessità di interventi umami. Sarebbe troppo lungo discutere qui la struttura e il funzionamento del "dizionario di macchina" per una lemmatizzazione semiautomatica, certamente possibile anche per i canti popolari; rinvio per questo argomento alla comunicazione : Analisi lessicali mediante elaboratori elettronici, A. Duro e A. Zampolli. (Convegno sul tema : L'automazione elettronica e le sue implicazioni scientifiche, tecniche, sociali, Accademia dei Lincei, Roma, Ottobre 1967). Mi sembra però necessario rilevare che i canti popolari pongono problemi affatto particolari per la loro provenienza da sistemi linguistici dialettali o comunque regionali diversi .

Un'ultima osservazione : se un verso contiene in media 6 parole, e viene ripetuto come contesto per ciascuna di esse, una concordanza completa è per lo meno 6 volte più estesa, come numero di righe, del testo originale su cui è basata .

I Centri che pubblicano concordanze, mettono in atto diversi sistemi per ridurne la mole, sia per diminuire le spese di edizione, sia per rendere più agevole la consultazione . Il metodo più semplice, e che verisimilmente potrà essere impiegato anche da noi , consiste nel comunicare al calcolatore una lista di parole da escludere, del tutto, o secondo percentuali prestabilite, dalle concordanze . Questa tecnica si basa su una caratteristica ormai comprovata del linguaggio naturale . Le parole "grammaticali" o "vuote" (pronomi, congiunzioni, preposizioni, ecc) rappresentano il 50% di un testo e, inoltre, le 50 parole più frequenti, quasi tutte parole "vuote" ne rappresentano da sole il 40-50% (Si veda a questo proposito il tabulato n. 6) . Di solito, in concordanze con finalità lessicografiche, vengono escluse o decimate appunto le parole grammaticali più frequenti . La completezza della documentazione lessicale viene affidata al nastro magnetico che contiene le concordanze complete che possono essere stampate per singole parole richieste di volta in volta dai singoli studiosi .

5. Forme in ordine alfabetico e relative frequenze .

a - numero d'ordine della forma nell'elenco di tutte le forme in sequenza alfabetica ,
b - forma ,
c - sua frequenza complessiva nell'insieme dei componimenti .

Naturalmente non essendosi operata la distinzione degli omografi, le forme sono intese in senso grafico : così, per es., la frequenza della forma che (102) comprende sia le occorrenze del che pronome sia quelle del che congiunzione .

A lemmatizzazione avvenuta, sarebbe molto semplice ristampare questo elenco con frequenze distinte per le forme omografe .

Nelle liste di frequenza in sede di lavoro definitivo, saranno da valutare le consequenze delle ripetizioni di parole dovute alla presenza di numerose lezioni di uno stesso componimento, con un numero di varianti anche minimo, come è caratteristico dei testi di tradizione popolare .

Analogo problema pongono le ripetizioni in uno stesso componimento di versi identici non legate solo a uno schema musicale .

6. Forme elencate in ordine di frequenza decrescente .

a - rango : numero progressivo di ciascuna frequenza nell'ordine decrescente .
b - frequenza .
c - forma .
d - numero d'ordine della forma nella sequenza alfabetica normale .
e - sommatoria progressiva delle frequenze .

Ogni foglio contiene 50 forme ; in calce è stampata la somma delle loro frequenze.

A parità di frequenza le forme sono in ordine alfabetico inverso (dalla z alla a , anzichè dalla a alla z); ciò solamente, perchè per risparmio di tempo, abbiamo usato un programma così predisposto per un lavoro precedente .

Nel lavoro definitivo potrà essere interessante aggiungere altre liste di frequenza parziale, per così dire specializzate : ad es. distinte per regione, o per tipo di componimento o per posizione del verso (iniziale, finale), ecc.

Per quanti si occupano di descrivere la composizione quantitativa del linguaggio e vi ricercano leggi statistiche, questa lista è certamente la più utile . Essa fornisce infatti i dati numerici necessari per l'applicazione delle formule proposte dai diversi autori (Zipf, Guiraud, Herdan, ecc.) : lunghezza del testo, estensione del vocabolario, rapporto rango-frequenza ecc . Ma essa è utile a fini pratici per una migliore organizzazione del lavoro, per decidere cioè un diverso trattamento per parole di frequenza relativa altissima .

Per inciso osservo che, già da questo esperimento, le leggi fondamentali della distribuzione delle frequenze formulate su testi letterari, risultano valide, in prima approssimazione, anche per questi canti .

Le parole dei 100 componimenti, pari a 521 versi, sono 3745 ; le forme grafiche diverse corrispondenti sono 987, distribuite in 41 tipi o ranghi di frequenza .

La forma più frequente (E) che rappresenta lo 0,10 % delle 973 forme, ha una frequenza assoluta di 173 occorrenze, che rappresentano il 4,62 % delle 3745 occorrenze complessive . Rispettivamente per le 5, 10, 50, 100 forme

più frequenti la situazione è la seguente :

	% sulle 987 forme	frequenza assoluta	frequenza % sulle 3745 occorrenze
1	0,10	173	4,62
5	0,51	511	13,64
10	1,01	806	21,52
50	5,07	1739	46,44
100	10,13	2214	59,12

7. Rimario I : versi elencati secondo l'ordine alfabetico della parola finale .

a - rima .
b - parola in rima .
c - verso .
d - riferimento (n. di componimento e n. di verso) .

Nello stabilire l'ordine si sono seguiti i seguenti criteri :

1. ordine alfabetico delle rime (e cioè della terminazione dei versi a partire dall'ultima vocale accentata) .
2. all'interno della stessa rima, ordine alfabetico della parola che contiene la rima .
3. nel caso in cui si abbia identità non solo della rima, ma anche della parola che la contiene, si è tenuto conto dell'ordine del verso nell'incipitario . (Come alternativa si potrebbe proporre l'ordine di numero di componimento) .

8. Rimario II

Elenco delle "rime perfette" delle "assonanze regolari" delle parole comunque collocate in reciproca relazione di "proposta e risposta" di rima .

a-b parole in reciproca relazione di "proposta e risposta" di rima .
c - riferimento (n. di componimento) .
d - l'eventuale asterisco segnala coppie di parole che pure essendo secondo lo schema metrico in posizione di "proposta e risposta" di rima, non sono tra loro né in rima né in assonanza regolare .

Nello stabilire l'ordine delle coppie di parole si sono seguiti i seguenti criteri :

1. ordine alfabetico della prima parola della coppia .
2. in corrispondenza a una stessa parola, sono collocate in coda alla altre le coppie contrassegnate da asterisco .
3. ordine alfabetico della seconda parola della coppia .
4. a parità di entrambe le parole, si è tenuto conto dell'ordine di componimento .

La coppia di parole in relazione viene ripetuta nell'ordine predetto, per entrambe le parole che la compongono .

Il numero di coppie così generate è dato dalla formula $\underline{n(n-1)}$, dove $\underline{n}$ è il numero delle parole in relazione , nello stesso componimento, secondo una determinata rima .

Per es., se lo schema del componimento è $a\ b\ a^1\ b^1\ a^2\ b^2\ c\ c^1$, nell'elenco appaiono le seguenti coppie :

$a\ a^1, a\ a^2, a^1\ a, a^1\ a^2, a^2\ a, a^2\ a^1$. $\quad 3(3 - 1) = 6$

$b\ b^1, b\ b^2, b^1\ b, b^1\ b^2, b^2\ b, b^2\ b^1$. $\quad 3(3 - 1) = 6$

$c\ c^1, c^1\ c$. $\quad 2(2 - 1) = 2$

Oltre alle parole finali collocate in reciproca relazione di "proposta e risposta" di rima, nei testi scelti compaiono parole finali di verso, per le quali non è stata determinata nessuna corrispondenza con altre , o perchè non era immediatamente determinabile lo schema metrico del componimento (cfr. per es. comp. n. 0170 versi 5-8) ; oppure perchè,essendo inserite in un preciso e ben noto schema metrico, proprio per la natura di questo, non si presentavano in rapporto di rima perfetta o assonanza regolare, ma di "consonanza atona", come è il caso del secondo verso degli stornelli .

Di queste parole non è stato stampato alcun elenco in attesa delle ulteriori decisioni per il lavoro definitivo.

Rimandiamo alla relazione generale, sia per la esposizione delle ragioni che hanno indotto ad aggiungere questo secondo rimario al primo (n. 6) di tipo tradizionale, sia per i complessi problemi connessi alla elaborazione degli indici metrici, la quale rappresenta uno degli aspetti più nuovi e interessanti del lavoro .

B. Fasi dell'esperimento

1. Preedizione del testo

L'operazione di preedizione ha la funzione di introdurre informazioni non presenti a livello grafemico nel testo stampato : ad es. distinguere brani interpolati da omettere, segnalare luoghi di lettura incerta ecc.

Per i componimenti della raccolta Barbi, il lavoro di preedizione presenta particolare importanza ma buona parte delle informazioni da aggiungere (quali ad esempio località e data di raccolta, dati sull'informatore ecc.) non si riferisce direttamente al testo.

Di conseguenza è possibile svolgere il lavoro di perforazione dell'input seguendo due linee distinte e reciprocamente indipendenti : si portano su schede da un lato il testo dei componimenti, dall'altro tutte le informazioni relative a ciascuno di essi. Sarà compito del calcolatore associare testo e informazioni sulla base di un comune numero di codice.

In questo esperimento ci siamo occupati solo del testo. Gli elementi da inserire a livello grafemico sono stati solo due : si è associata ad ogni verso una lettera, per descrivere nel modo tradizionale (ad es. A, B, A, B, C, C) lo schema metrico del canto; - è stata segnata la posizione dell'accento tonico sulla parola in posizione di rima.

Il lavoro di preedizione dei 100 componimenti scelti è stato svolto dalla Dr. ssa Paola Tabet Raicich e dal Sig. Salvatore Barone.

2. Perforazione

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

I componimenti perforati per questo nostro esperimento non presentano difficoltà particolari di codificazione : richiedono infatti solo 30 caratteri, tra lettere (senza distinzione tra maiuscole e minuscole) accenti e apostrofo.

Di contro, la natura e le condizioni dei manoscritti che devono essere trascritti pongono difficoltà tali che non consentono di affidare la perforazione ai normali centri di servizio operativo, e consigliano sia eseguita da persone specializzate, in grado di decifrare e comprendere i testi. La perforazione relativa all'esperimento è stata eseguita dalle stesse persone che si sono occupate della preedizione.

Anche per la raccolta Barbi, si è adottato il sistema messo a punto per l'Accademia della Crusca e per gli altri progetti linguistici in corso [illegible] di Pisa. Ciò per rispondere alla esi-

3. Verifica

Per elaborazioni condotte su una grande quantità di parole, alla perforazione si fa normalmente seguire una operazione di verifica del testo. Le schede già perforate vengono alimentate in una macchina verificatrice, sulla quale l'operatore ribatte il testo una seconda volta : la macchina segnala le eventuali discordanze tra la prima e la seconda battuta.

Questa operazione, che elimina di solito più del 95% degli errori presenti, non è stata eseguita per questo esperimento ma deve essere prevista in sede di lavoro effettivo.

4. Elaborazioni elettroniche

Le elaborazioni sono state eseguite sui calcolatori IBM 1401 16K e IBM 7090 del Centro Nazionale Universitario di Calcolo Elettronico di Pisa.

A partire dalle schede contenenti i versi perforati, le elaborazioni si sono svolte del tutto automaticamente, senza nessun intervento umano. Sono stati usati complessivamente 9 programmi diversi, 4 dei quali già predisposti per lavori precedenti e 5 scritti appositamente.

Dato il carattere di primo esperimento, si è volutamente semplificata la elaborazione, e i risultati ottenuti devono essere considerati puramente indicativi.

Il calcolatore "legge" le schede perforate e registra il testo su due nastri magnetici diversi :

- il nastro testo che contiene i singoli versi registrati ciascuno come unità elementare di elaborazione, accompagnati dal numero di contesto e di verso nel contesto.
- il nastro parola nel quale le unità elementari di registrazione sono le parole, intese come serie consecutiva di lettere tra due spazi. Per ciascuna parola il calcolatore ripete le informazioni che la riguardano, anche se nel testo stampato e perforato occorrono una sola volta, quali numero di pagina, di verso ecc.

Dal primo nastro, mediante opportuni ordinamenti, i versi vengono ricopiati su altri nastri e stampati nella sequenza voluta dall'incipitario e dal rimario. Dal secondo nastro mediante opportuna selezione le parole vengono disposte in sequenza alfabetica, generando un nuovo nastro con la ricapitolazione delle forme, dal quale si ristampano i due tabulati delle frequenze. Successivamente le parole vengono ricopiate, corredate di un contesto, su un terzo nastro, dal quale, previo ordinamento alfabetico, si stampano le concordanze.

C. Conclusione

L'esperimento è stato a nostro avviso molto utile. Ha posto in luce alcuni momenti fondamentali della elaborazione nei quali si devono operare delle scelte tra le diverse possibili alternative che si sono presentate (per esempio quali informazioni introdurre in preedizione, soprattutto per la elaborazione delle strutture metriche; con quali criteri stabilire la sequenza dei vari elenchi di versi (incipitari e rimari); come alleggerire le concordanze e se lemmatizzarle o no; ecc.). Nel contempo ha fornito nuove informazioni, di carattere propriamente sperimentale, utilizzabili sia nelle scelte iniziali in sede di programmazione, sia per la organizzazione pratica del lavoro (sequenza delle operazioni, organizzazione della perforazione, tempi di lavoro).

D'altro lato ha confermato che la elaborazione elettronica, sottraendo lo studioso alla fatica delle operazioni di spoglio, di trascrizione, di ordinamento, gli permette di richiedere un maggior numero di risultati e di immaginare e organizzare una documentazione esaustiva, in forme nuove e più ricche rispetto a quelle tradizionali.

La perforazione del testo ha richiesto complessivamente 3 ore; ma i tempi macchina della elaborazione elettronica per l'intero esperimento sono stati i seguenti :

Calcolatore IBM 1401	18 minuti primi
Calcolatore IBM 7090	45 minuti secondi

Antonio Zampolli

Centro Studi IBM

Avvertenza : I tabulati ottenuti con l'unità IBM 1403/2 sono stati ridotti all'80% delle dimensioni originali e ne è stato eseguito il fototrasporto su matrici stampate poi in offset.