ACCADEMIA NAZIONALE DEI LINCEI

ANNO CCCLXXII - 1975

CONTRIBUTI DEL
CENTRO LINCEO INTERDISCIPLINARE
DI SCIENZE MATEMATICHE E LORO APPLICAZIONI

N. 13

COLLOQUIO SUL TEMA:

LE TECNICHE DI CLASSIFICAZIONE E LORO APPLICAZIONE LINGUISTICA

(Firenze, 13 dicembre 1972)

(*ESTRATTO*)

ROMA
ACCADEMIA NAZIONALE DEI LINCEI
1975

ANTONIO ZAMPOLLI

L'ELABORAZIONE ELETTRONICA DEI DATI LINGUISTICI: STATO DELLE RICERCHE E PROSPETTIVE

Lo scopo di questa comunicazione non è quello di passare ordinatamente e sistematicamente in rassegna le diverse applicazioni dei calcolatori alla elaborazione dei dati linguistici. Non solo tale rassegna esigerebbe una classificazione di queste attività che per diversi motivi non sembra utile [1] ma andrebbe ben oltre il tempo concessoci. Preferisco invece discutere alcune ricerche, attraverso le quali è possibile, a mio avviso, illustrare non solo i risultati conseguiti negli ultimi anni, ma anche e soprattutto la natura dei problemi che oggi sono all'attenzione dei ricercatori e la situazione di stallo conseguente alla mancata soluzione di alcuni di essi. Le soluzioni di questi problemi costituiranno probabilmente le mete degli sforzi dei ricercatori nei prossimi anni e determineranno le linee di ricerca e le prospettive di sviluppo.

Mi sono basato soprattutto sulle mie esperienze di collaborazione ai numerosi progetti di spoglio condotti dai diversi utenti della Divisione Linguistica del CNUCE e sulle idee e sulle esperienze esposte dai rappresentanti delle mag-

(1) Una classificazione delle diverse applicazioni, dal momento che le tecniche per la elaborazione elettronica dei dati sono ormai impiegate praticamente in tutti i settori della ricerca linguistica, dovrebbe riflettere in realtà una classificazione delle ricerche linguistiche. Il panorama di queste ricerche oggi presenta, accanto a uno straordinario fiorire di attività, non poche contraddizioni, al punto che non sembra possibile tentarne una classificazione se non ponendosi nella prospettiva di una posizione teorica legata a una corrente o a una scuola ben determinata. E infatti gli autori che hanno fino ad oggi tentato una rassegna delle attività della linguistica computazionale (CL) hanno cercato invano, a mio giudizio, di sottrarsi a questo vincolo. Si può mostrare che anche coloro i quali – provenendo da discipline non linguistiche come la scienza dell'informazione, la logica, la matematica, l'ingegneria, la cibernetica – hanno adottato un criterio di classificazione dichiaratamente non linguistico, di fatti si sono mossi nel quadro di una teoria linguistica determinata. È il caso per esempio delle classificazioni operate in base alla « profondità » del trattamento dei dati e della complessità con cui opererebbero i diversi algoritmi, nelle quali la profondità viene misurata per confronto con i modelli generativisti di stampo semanticista, e il grado di utilità, di adeguatezza e di scientificità delle elaborazioni viene ritenuto direttamente proporzionale alla complessità degli algoritmi e all'impiego di strumenti matematici e formalizzanti. In realtà, molti cultori della CL, provenienti dalle discipline tecnico–scientifiche citate, hanno recato un grave danno alla credibilità scientifica della CL, perseguendo e valutando le applicazioni non per la loro utilità nei confronti della linguistica, delle discipline letterarie, della filologia, ecc., ma per la complessità dell'algoritmo e per il grado di formalizzazione che esso richiede. Per una bibliografia delle classificazioni proposte si veda Zampolli (1969 e 1970).

in altri Centri (Besançon, Leida, Cambridge, ecc.) che iniziavano la produzione di indici con sistemi UR a schede perforate [3].

Restava tuttavia praticamente senza risposta, nonostante alcune procedure costosissime e poco felici di ripiego, l'esigenza di produrre concordanze di struttura adeguata.

1.1.2. *Introduzione dei Calcolatori.*

Verso la fine degli anni '50 i calcolatori elettronici, già diffusi in campo commerciale e industriale da più di 10 anni in America e in Europa, affiancarono le macchine UR nelle elaborazioni per lo spoglio di testi.

L'adozione della nuova tecnologia non avvenne senza perplessità da parte dei lessicografi, e fu il principale argomento di discussione nei convegni di Tübingen (1960) e Besançon (1961). La scheda perforata, che può essere letta, maneggiata, e seguita costantemente di operazione in operazione attraverso la macchina UR, in fondo continuava ad essere vista come una normale scheda lessicale manuale, strumento tradizionale di generazioni di ricercatori.

L'uso dei calcolatori si diffuse in seguito, più rapidamente di quanto era stato previsto nei due congressi citati, ma gli sforzi dei lessicografi furono rivolti, in ultima analisi, a far eseguire al calcolatore le operazioni più pesanti della procedura tradizionale, senza modificarla sostanzialmente.

Il Congresso di Praga (1966) consacrò, per così dire, con accenti trionfalistici, l'avvento delle « machines dans la linguistique ». Il panorama dei centri attivi un pò ovunque in Europa, la varietà delle applicazioni e le dimensioni dei progetti suscitavano un grande entusiasmo. Era disponibile una ricca gamma di procedure e di programmi per produrre con il calcolatore indici di frequenza, rimari, indici inversi, concordanze, schede contestualizzate, ecc.

I principali progetti di dizionari storici avevano adottato o stavano adottando gli spogli elettronici, e in tutto il mondo si moltiplicavano gli spogli di testi, così che era pienamente giustificato parlare di *The death of the hand made concordance* (Raben, 1969) [4].

(3) Per questo periodo si possono consultare i primi numeri dei *Cahiers de Lexicologie* e De Tollenaere (1963).

(4) Come indicazioni bibliografiche generali per questo periodo si possono citare gli Atti di alcuni Congressi « Colloque International sur la Mécanisation des Recherches Lexicologiques », Besançon, 1961, in *Cahiers de Lexicologie*, 3 (1962); « Literary Data Processing Conference », Yorktown Heights (New York) (1964); « Le applicazioni dei Calcolatori Elettronici alle Scienze Morali e alla letteratura », Milano, 1962; « Les Machines dans la Linguistique », Praga, 1966; « L'Automazione Elettronica e le sue implicazioni scientifiche, tecniche e sociali », Accademia dei Lincei, 1968; « L'elaboration électronique en Lexicologie et en Lexicographie », Pisa, 1970; « Lexicon electronicum latinum », Pisa, 1968) e alcune riviste sorte nel frattempo per rispondere alle esigenze di diffondere informazioni in questo settore in così rapida espansione (oltre ai già citati *Cahiers de Lexicologie* del « Centre d'Etude du Vocabulaire Français » di Besançon, la *Revue* dell'« Organisation Internationale pour l'Etude des Langues Anciennes par Ordinateur » di Liegi, *The Finite String* della « Association for Computational Linguistics » degli USA, *Calculi* della « American Philological Association », *Computers and the Humanities* di New York, ITL dell'« Instituut voor Togepaste Linguistick» di Lovanio, ecc.).

1.1.2.1. *Alcuni esempi.*

Mi sembra opportuno riferire qui rapidamente alcuni lavori a titolo di esempio.

I progetti sono così numerosi che, nonostante le inchieste condotte per molte lingue, siamo ben lontani da un censimento soddisfacente. Poiché i progetti in corso presso il CNUCE coprono ampiamente le diverse zone di questo settore, mi sembra opportuno scegliere gli esempi tra di essi.

Archivi lessicali per la compilazione di grandi dizionari storici di una lingua.

Il progetto più significativo è senz'altro la costituzione degli archivi per il *Tesoro della Lingua Italiana* delle origini e per il *Vocabolario Storico della Lingua Italiana*, ad opera dell'Accademia della Crusca (5). Sono stati elaborati elettronicamente 30 milioni di occorrenze, il 40% delle quali sarà rappresentato nell'archivio. Negli ultimi 5 anni è stata pressoché completata la preparazione dell'input, ed ora le risorse umane e tecniche sono dedicate principalmente alla lemmatizzazione e alle scelte degli esempi. Le operazioni lessicologiche hanno raggiunto un notevole livello di formalizzazione, e formano una procedura che funge da modello per le elaborazioni degli altri progetti di spoglio, italiani e europei.

In questi lavori che affondano le loro radici nella ricca tradizione lessicologica e lessicografica europea, si riflette la tendenza moderna a creare grandi archivi ed inventari, che è stata determinata dalla diffusione dei calcolatori e dallo sviluppo della linguistica descrittiva. « The idea of creating lexicological archives is slowly making headway in each country. Slowly and cautiously without speaking of the naive myth of exhaustibility, these positive principles of cumulative work seem to be gaining the acceptance that they deserve (...) New methods of inventory–making have profoundly changed the conditions and the perspectives of lexicological text analysis » (B. Quemada, 1973, p. 425).

Gli esperimenti intrapresi su larga scala alcuni anni fa da centri nazionali, per es. a Besançon, Firenze, Nancy, Leida, si sono trasformati sotto un certo punto di vista in imprese di routine nelle quali lavorano grandi équipes di lessicografi.

Ma sembra chiaro che sta iniziando una nuova fase sperimentale: la quantità dei materiali raccolti – decine di milioni di citazioni – richiede che

(5) Cito come esempi di archivi lessicali di lingue con alfabeto non latino, preparati in Italia:

– il Progetto di un Dizionario Storico della Lingua Rumena del XVI secolo, i cui lavori, in collaborazione tra il CNUCE di Pisa e Bucarest, sono diretti da F. Dimitrescu (1973);

– il Lessico dell'ittita cuneiforme, diretto dal Prof. P. Meriggi (1973) di Pavia;

– l'archivio dell'assiro babilonese in collaborazione tra L'UCLA, il CNR e il CNUCE (Buccellati 1974).

il calcolatore venga adoperato non solo nell'assemblaggio degli archivi, ma anche come aiuto nella loro gestione. Avremo modo di discutere questa situazione più avanti.

Archivi lessicali per la redazione di Dizionari Storici relativi a discipline particolari.

A titolo di esempio cito il progetto per il « Vocabolario Storico della Lingua Giuridica » (Ciampi, 1973).

Questi progetti sono caratterizzati dalla esigenza di effettuare una forte selezione degli esempi, perché il lessico che deve essere messo in evidenza e accolto nell'archivio è un lessico nettamente specializzato: cioè il rapporto percentuale tra occorrenze in *input* e occorrenze che diventano *items* dell'archivio è molto basso. Perciò questi lavori hanno l'aspetto di esperienze interessanti sia metodologicamente sia economicamente, anche in relazione alla attuale richiesta di dizionari terminologici che proviene da numerosi settori scientifici e tecnici.

Lessici, indici e concordanze dell'opera omnia di un autore.

Merita rilevare che particolarmente numerosi sono i progetti relativi all'opera omnia di filosofi e pensatori. L'indice del lessico è considerato non solo come la chiave che permette di ricercare agevolmente i passaggi ove sono definite o esplicitate le idee, ma anche come uno strumento per chiarire e verificare l'interpretazione. Si pensi per esempio all'*Index Thomisticus* già citato, che dovrebbe iniziare entro l'anno la pubblicazione del primo dei 60 volumi previsti, e all'archivio dell'opera omnia di A. Rosmini, testè completato.

Spogli di opere diverse.

L'inventario più completo delle opere sottoposte a spoglio elettronico è probabilmente costituito dalle Sezioni « Directory of Scholars Active » e « Verbal Materials in Machine-Readable Form » che appaiono periodicamente in « Computers and the Humanities ». In Italia sono in corso di spoglio testi di diverse lingue (ventotto). Si tratta per lo più di studi stilistici, effettuati spesso per tesi di laurea su base lessicale e statistica, aventi come strumenti preferiti concordanze e indici. Le metodologie sono ben note, ma può essere interessante segnalare qualche esperienza di concordanze parzialmente diverse dalle forme tradizionali.

Le concordanze che chiamiamo *contrastive* si sono rivelate utili per confrontare un testo di base con un insieme di testi satelliti (per esempio un originale con le sue traduzioni in altre lingue, o con le sue successive redazioni). A ogni forma del testo originale si attribuiscono come contesti non solo il normale contesto, ma anche le unità contestuali corrispondenti nei testi satelliti. Analogamente, ogni forma del testo satellite riceve come contesto sia il proprio contesto normale, sia i contesti corrispondenti negli altri testi satelliti e nel

testo base [6]. Questo metodo ha permesso utili considerazioni critiche sul meccanismo della traduzione di opere letterarie. Le concordanze *sintattiche* hanno come esponente un classificatore o un insieme di classificatori dei tipi sintattici. Il contesto è costituito dalla proposizione che appartiene al tipo sintattico in esponente, accompagnata dalle proposizioni ad essa collegate da rapporti immediati di coordinazione, reggenza e subordinazione (cfr. De Mauro e Policarpi, 1971).

Le concordanze che chiamiamo *tematiche* ci sono state richieste per l'analisi delle strutture di testi poetici di tradizione orale. In queste concordanze l'esponente è costituito dalla sigla di un *tema* o di un *motivo*, e il contesto è l'intero brano corrispondente. Ovviamente sono prodotte anche tutte le possibili liste di correlazione tra lessico, temi, motivi, ecc. Il lavoro più avanzato in questo settore è l'analisi strutturale dei componimenti lirici di Pindaro e di Bacchilide.

Un particolare approfondimento metodologico richiedono gli spogli di testi condotti con interesse prevalentemente filologico. Il progetto di indici e concordanze di d'Arco Silvio Avalle (1973) sulla lingua poetica in Italia avanti la fine del XIII secolo è particolarmente interessante, perché presenta un ciclo completo di operazioni che, in interazione con il calcolatore, portano dalla trascrizione dei manoscritti, anteriori alla fine del XIII secolo, alla pubblicazione in fotocomposizione del testo riprodotto diplomaticamente e delle sue diverse analisi (indici, concordanze, ecc.).

Questo progetto è a mio avviso destinato a diventare un modello nel suo genere. Con una sola trascrizione del testo, e dunque con un solo controllo, si ottiene la stampa sia del testo critico sia degli indici e delle concordanze. Le correzioni eseguite con il metodo interattivo per mezzo del calcolatore sono più economiche e sicure delle correzioni di bozze. È dimostrato dalle esperienze di più ricercatori che gli indici e le concordanze sono strumenti molto efficaci per scoprire eventuali errori, sviste, incoerenze dell'editore di un testo. Le concordanze e gli indici che si producono ogni giorno al CNUCE mettono in evidenza errori e lapsus in edizioni critiche ritenute giustamente eccellenti e sicure. È dunque un grosso vantaggio poter pubblicare l'edizione di un manoscritto dopo averlo potuto analizzare con il calcolatore.

Un altro problema tipico è quelli del trattamento delle varianti. Oltre ad alcuni tentativi di automatizzare le fasi della *collatio* e della *recensio codicum*, che hanno condotto a risultati positivi per quanto concerne lo studio dei relativi algoritmi ma non sembrano ancora destinati ad applicazioni pratiche, va segnalato lo sforzo che alcuni filologi italiani stanno compiendo per definire dal punto di vista metodologico il trattamento delle varianti di un testo del quale si pubblicani indici, concordanze, ed elaborazioni statistiche. Si sono sperimentati metodi diversi. La soluzione più semplice prevede unica-

(6) Si veda. per esempio, la tesi di laurea di A. Penazzo (Parma, 1972) sulle diverse traduzioni italiane di alcuni brani del Faust di Goethe.

mente l'apposizione di un contrassegno alle parole per avvertire che hanno delle varianti che non vengono specificate. Soluzioni più complesse prevedono che tutte le informazioni dell'apparato critico vengano inserite nelle elaborazioni. In questo modo, le singole varianti e i rapporti tra le singole varianti sono presentati esplicitamente e perspicuamente nelle concordanze, negli indici, nelle statistiche. Il progetto più avanzato è lo spoglio delle tre redazioni dell'*Orlando Furioso* di L. Ariosto diretto da C. Segre [7].

Banche di dati.

L'archivio dell'*Atlante Linguistico Italiano* comprende circa 9 milioni di parole in trascrizione fonetica, raccolte con inchieste condotte in 1000 località italiane, in base a un questionario di circa 700 domande (cfr. Grassi 1973). Sono già state pubblicate alcune analisi di tale questionario (Ali). È in corso la trasposizione in 'machine readale–form' delle risposte dialettali. L'adozione delle tecniche elettroniche ha indotto a capovolgere l'ordine tradizionale delle operazioni, nel quale la messa in carta dei materiali precedeva gli indici e le analisi. Il piano attuale prevede invece la pubblicazione dello Atlante solo dopo la pubblicazione di indici regionali, di dizionari per località, di carte relative a fenomeni linguistici particolari, ecc. Si pensa già a un progetto di estensione dell'archivio con nuove inchieste, strutturate in una banca dinamica delle conoscenze dialettali. Si deve tener presente il rinnovamento dell'interesse per la dialettologia in Italia, motivato, tra l'altro, dall'urgenza di fissare, prima che vada perduto, il patrimonio culturale regionale, e dalla posizione privilegiata degli studi dialettali e regionali nel quadro delle ricerche sociolinguistiche, attualmente molto sentite nel contesto di mobilità sociale e geografica di gruppi e di individui.

1.1.2.2. *Procedure di spoglio.*

Esamianiamo a grandi linee lo schema di uno spoglio lessicale, così come viene condotto di solito presso la maggior parte dei Centri Specializzati. Esiste naturalmente una grande varietà di procedure, ma esse hanno in comune alcune funzioni, ordinate, di base, alle quali qui mi riferisco.

Fase 1. – *Preparazione dell'input.*

Il testo viene registrato su un supporto operabile dal calcolatore. Le operazioni fondamentali sono le seguenti. Il testo viene ricopiato, da un operatore, di solito su schede meccanografiche. Le schede, o il loro equivalente

(7) È stato messo a punto un sistema che permette di comunicare al calcolatore il testo delle due prime redazioni semplicemente segnalando le differenze tra queste e la terza edizione. L'algoritmo di elaborazione decide in base a regole formali, che considerano la natura delle varianti, quali varianti porre in esponente, quali riportare nel contesto, nonché il loro conteggio nel calcolo delle frequenze.

meccanografico, vengono listate, cioè viene stampato dal calcolatore il testo così come è stato copiato dall'operatore. Questa lista viene letta e confrontata, di solito dal ricercatore, con il testo originale, per ricercare e segnare gli inevitabili errori di ricopiatura. Questi errori vengono corretti per mezzo di schede. Il testo così corretto viene registrato in forma definitiva su un nastro magnetico o su di un supporto equivalente.

Fase 2. – *Elaborazioni relative alle forme grafiche.*

Al calcolatore viene di solito richiesto di elaborare una serie di indici e concordanze nei quali l'unità di elaborazione è la *forma grafica*. In questa fase cioè la parola è definita, per il calcolatore, come una sequenza di caratteri alfabetici (lettere e segni diacritici: accenti vari, dieresi, apostrofo, ecc.) tra due spazi. Il calcolatore considera due sequenze identiche di caratteri alfabetici come due parole uguali, e cioé come due occorrenze di una stessa forma grafica. Perciò tutte le sequenze *faccia* del testo sono considerate come occorrenze di una stessa forma, anche se hanno significati diversi nei diversi luoghi del testo. Per la stessa ragione saranno tenute distinte due sequenze come *amavo* e *amerò*. Gli indici più comunemente elaborati sono l'index locorum (elenco delle forme grafiche ciascuna seguita dalle indicazioni di tutti i luoghi del testo ove appare), gli indici alfabetici diretto e inverso delle forme e delle relative frequenze, l'indice delle forme in ordine di frequenza decrescente, diversi tipi di rimario, le concordanze delle forme, e, raramente, le schede lessicali.

Fase 3. – *Lemmatizzazione.*

Sulle liste delle concordanze per forma, il ricercatore analizza linguisticamente le varie forme. Normalmente questa analisi si limita alla *lemmatizzazione*; il ricercatore, cioè, scrive accanto a ciascuna forma il lemma cui la forma deve essere assegnata. Se la forma è omografa, il lemmatizzatore indica il lemma per ciascuna delle sue occorrenze, esaminando il contesto di ciascuna occorrenza. Un operatore perfora poi queste indicazioni su schede per mezzo delle quali il calcolatore aggiunge il lemma a ciascuna parola del testo registrato su nastro magnetico. Di solito, come si è detto, l'analisi non arriva a livelli più approfonditi. Per esempio non viene operata a livello morfologico, per distinguere forme di uno stesso lemma omografe sul piano morfologico (del tipo *dica*), nè a livello sintattico, per distinguere le diverse ' costruzioni ' o ' strutture ' nelle quali la parola è inserita, nè a livello sematico, per distinguere le diverse accezioni di una parola polisemica. Spesso, nel corso della lemmatizzazione, viene effettuata una scelta dei materiali lessicali; cioè vengono scelti i lemmi e le occorrenze ritenuti ' interessanti ': le altre parole vengono eliminate, e cioè ricevono un contrassegno che le esclude dalle elaborazioni successive (per esempio dalla produzione delle schede lessicali).

Fase 4. – *Elaborazione dei risultati dello spoglio.*

Si richiede al calcolatore l'elaborazione e la stampa degli stessi indici elencati alla fase 2, ma questa volta l'unità di elaborazione è il lemma.

Si pone a questo punto, di solito, il problema di mettere a disposizione di tutti gli studiosi questi risultati, pubblicando indici e concordanze dei lemmi. Farli ricomporre tipograficamente presenta l'inconveniente della correzione delle bozze, la cui gravità appare chiara se si considera che i risultati dello spoglio consistono in un numero di righe che è circa dieci volte quello del testo originale. Si preferisce riprodurli con metodi fotografici o xerografici, o ricorrere alla stampa per mezzo della fotocomposizione.

Le schede lessicali vengono prodotte dal calcolatore in due o tre serie complete, che vengono conservate in appositi archivi–schedario.

Tutta questa documentazione (indici, concordanze, e soprattutto le schede lessicali) è di solito alla base della redazione dei grandi dizionari storici, sincronici, ecc. È chiaro che quanto minore è il grado di analisi nel corso dello spoglio, tanto più oneroso e complesso sarà il compito del redattore. L'alternativa tra l'approfondire l'analisi e la scelta dei materiali in fase di spoglio, e il rimandarle alla fase di redazione, è oggi motivo di vivacissime discussioni tra lessicologi e lessicografi.

1.1.3. *La crisi attuale: Il problema centrale della lessicologia e della lessicografia assistite dai calcolatori.*

Verso il 1969–1970 cominciarono a levarsi le prime voci di insoddisfazione e di critica contro il generale ottimismo che, come abbiamo visto, caratterizzava, e caratterizza ancora oggi, l'ambiente degli utenti del calcolatore per gli spogli lessicali.

Ci si è resi conto che lo sviluppo delle applicazioni secondo le linee, per così dire, ormai ' classiche ' degli anni '50 e '60 è giunto a un punto di saturazione. Se si continua secondo la metodologia corrente o, in altre parole, secondo le attuali regole del gioco, non esistono concrete prospettive di sviluppo. Le tecniche e le metodologie in uso permettono di fare eseguire al calcolatore alcune delle operazioni tradizionalmente riservate al lessicografo; essenzialmente, quelle connesse alla trascrizione dei contesti per ciascuna occorrenza del testo e al loro ordinamento alfabetico. Ma per quanto le macchine diventino sempre più veloci e i programmi più sofisticati, i lessicografi non possono profittarne in proporzione, perché le metodologie attuali producono già molti più dati di quanti una équipe redazionale di dimensioni ragionevoli possa elaborarne lavorando secondo le procedure attuali. Ci si trova di fronte a una alternativa abbastanza chiara. Si accettano i limiti delle applicazioni odierne, si continuano a registrare *corpora* sempre più vasti di testi, a produrre archivi di concordanze o di schede lessicografiche senza alcuna previa analisi e classificazione linguistica. Tali elaborazioni vengono rinviate a una successiva fase di redazione, senza preoccuparsi per il momento degli enormi pro-

blemi che saranno posti dalla quantità dei materiali lessicali archiviati. Oppure si rinnovano le metodologie attuali. Si applicano i ritrovati di alcune discipline correlate, quali la linguistica computazionale, la statistica linguistica, l'*information retrieval*, ecc. Si sfruttano adeguatamente le nuove possibilità che la tecnologia dei calcolatori propone nel suo rapidissimo sviluppo; in particolare le tecniche conversazionali che permettono di usare il calcolatore come uno strumento redazionale attivo, in stretta interazione con il lessicografo (8).

1.2. Impiego dei calcolatori. Nuove tendenze e prospettive.

Esamianiamo ora le linee di ricerca secondo le quali si cercano di superare le difficoltà e i limiti dei metodi attuali.

1.2.1. *Preparazione dell'input.*

Le difficoltà sono sia di ordine tecnico (9) sia di ordine economico.

(8) In queste affermazioni ci trovammo d'accordo B. Quemada ed io nelle nostre comunicazioni al « Colloque sur l'élaboration électronique en lexicologie et lexicographie » da me organizzato a Pisa nell'estate del 1970 (Zampolli ed., 1973 a). Questa nostra presa di posizione fu confermata dalla « Tavola Rotonda sui grandi lessici storici » organizzata a Firenze presso l'Accademia della Crusca nella primavera del 1971. Essa propose di inserire nella 2ª Scuola Estiva Internazionale di Pisa, accanto al già previsto indirizzo dedicato alle grammatiche formali e al trattamento automatico di strutture sintattiche e semantiche, un nuovo indirizzo per studiare in che misura e in che forma il calcolatore possa assistere il lessicologo e il lessicografo anche nella fase di analisi dei dati lessicali e di redazione, oltre che nella loro raccolta. Lo svolgimento contemporaneo dei due indirizzi ha permesso di osservare concretamente quanti temi e problemi in comune abbiano lessicologia e lessicografia da un lato, e linguistica computazionale e molte correnti teoriche attuali della linguistica dall'altro. È risultato molto chiaro che lo studio del lessico si prospetta oggi anche per queste discipline come un tema centrale. I linguisti si rendono conto che è indispensabile affrontare il compito di applicare, in estensione, i sistemi di classificazione e le tecniche di analisi e di descrizione a sottoinsiemi lessicali sempre più vasti, e sono portati a rivalutare l'opera di raccolta dei dati, portata avanti fino ad oggi quasi esclusivamente dai lessicografi, e le loro esperienze. I lessicografi, dal canto loro, prendono coscienza del fatto che l'interesse dei linguisti teorici e dei linguisti computazionali, convergendo sul lessico, ha prodotto strumenti teorici e tecnici che rispondono ad esigenze fondamentali del compito lessicografico. Questo punto di vista è stato chiaramente espresso anche da L. Venezki e W. P. Lehman alla « International Conference on English Lexicography » del 1972 a New York (cfr. Zampolli, 1973 c) ed ha trovato il consenso di quanti, linguisti e lessicografi, hanno partecipato al « Deuxiéme Colloque International de Linguistique ed de Traduction » (Montréal 1972) cfr. Zampolli, 1973 d e al « Colloque sur l'analyse des corpus linguistiques: problémes et méthodes de l'indexation maximale » (Strasburgo, 1973). Un'ulteriore conferma di questo atteggiamento è costituito dal fatto che le applicazioni dei calcolatori allo spoglio di grandi corpora è stato posto per la prima volta in modo esplicito tra i temi della « 5 TH International Conference on Computational Linguistics » di Pisa (1973).

(9) È noto infatti che le tastiere delle macchine abitualmente disponibili per ricopiare il testo su supporti leggibili dal calcolatore sono così povere di caratteri da rendere necessaria indirizzo e gli eventuali nuovi caratteri di correzione; immettere le correzioni nel calcolatore;

L'operazione di ricopiare il testo è di per sè lunga e molto costosa. Il ciclo di controllo e di correzione, che secondo la nostra esperienza deve essere ripetuto in media 3 volte, triplica il tempo e il costo. Per preparare 3 pagine di circa 40 righe ciascuna, occorrono di solito circa 3 ore così ripartite: perforazione 1 ora; verifica 50′; lettura per controllo 30′; perforazione ed esecuzione delle correzioni 20′; due cicli ulterioi di controllo e di correzione 20′. Non vanno poi dimenticati i ' tempi morti ' per il passaggio dei materiali tra i diversi esecutori del ciclo (controllore, perforatore, calcolatore). I rimedi a questa situazione possono essere sia tecnici sia organizzativi.

L'evoluzione tecnologica sta per rendere obsolete le perforatrici di schede o di nastro: per una documentazione completa dei nuovi mezzi disponibili rinvio all'articolo di B. R. Schneider (1971), soprattutto per le possibilità offerte dai lettori ottici e dai terminali [10].

una codificazione difficile e gravosa per rappresentare la ricca varietà di caratteri presenti nei testi. Le complicazioni della codificazione aumentano il numero degli errori di battitura. Il ricercatore, nel tentativo di semplificarla, rinuncia spesso a rappresentare grafemi che non sembrano immediatamente indispensabili per le elaborazioni che egli ha progettato. Accade poi non di rado che il testo così impoverito non sia sufficente per altre ricerche o altri ricercatori, i quali debbono provvedere a riperforarlo. La scarsità di caratteri disponibili nelle normali stampanti, nuoce, naturalmente, anche alla rapidità e alla esattezza del *controllo a lettura*.

(10) Qualcuno ha detto che nel campo dei lettori ottici l'utopia è sempre dietro l'angolo. Dapprima si è pensato a un lettore ottico per un solo *font*; oggi esistono lettori ottici *multifont*, e già si lavora a un lettore ottico *omnifont*. Questo non esisteste ancora, ma è stato dimostrato che con tecniche appropriate è possibile aggiungere con poca spesa nuovi *fonts* a quelli leggibili da un certo sistema. Se si deve spogliare un corpus stampato per intero con un solo *font*, è consigliabile calcolare se convenga o no sostenere le spese della programmazione necessaria per aggiungere questo nuovo *font*. Comunque è sempre possibile battere il testo con una macchina da scrivere dotata di un *font* già accettato da un determinato sistema ottico, e di farlo poi leggere da questo. Sembra accertato, da esperienze condotte, che rispetto alla perforazione il costo scenda ai 2/5 e che gli errori siano meno di 1/5. Il grado di accuratezza in perforazione influisce poi sull'andamento dei controlli successivi, perché gli errori che sfuggono al controllo sono in proporzione alla quantità di errori presenti. Inoltre la registrazione del testo prodotta da un lettore ottico può essere una immagine della pagina stampata. Ogni carattere, ogni corpo, e la loro collocazione nella pagina sono registrati automaticamente. Ciò significa che è possibile rilevare e contrassegnare automaticamente le categorie di informazione indicate dal tipo di formato, come, per esempio, le diverse sezioni di un articolo di dizionario.

Un interessante sistema è quello dei terminali CTR (Cathode Ray–Tube Terminal). Si tratta di un terminale che scrive su uno schermo TV (anziché o oltreché su un foglio di carta), e tutto ciò che appare sullo schermo può essere trasmesso al calcolatore. Ci sono molti tipi, alcuni dei quali prevedono uno schermo opaco per non affaticare la vista e possiedono un insieme di caratteri molto ampio. Alcuni tipi sono anche programmabili, nel senso che l'insieme dei caratteri è prodotto dal *software*, e può essere moltiplicato a piacere. Il teleschermo è uno strumento nel quale il testo è perfettamente elastico; la cancellazione è istantanea e gli errori possono essere localizzati corretti e verificati con una sola operazione, anziché con le 6 operazioni che, nella procedura normale, sono: stampare il testo con un numero di ordine (indirizzo) univoco per ogni parola o riga; annotare l'indirizzo dell'errore e scrivere accanto la correzione relativa; perforare l'ordine opportuno (cambia in, cancella, inserisci) per ogni errore, il suo

Ancora di più è possibile fare sul piano organizzativo, nazionale e internazionale. In numerosi paesi, alcune case editrici hanno adottato il sistema di stampare con il *type–setting* o con la *fotocomposizione* e così producono una grande quantità di testi su nastro magnetico o di carta. In Italia stiamo già lavorando per assicurare la disponibilità di questi materiali per le ricerche linguistiche.

L'obiettivo più importante è però quello di creare degli organismi per coordinare a livello nazionale gli sforzi dei vari Centri e dei singoli ricercatori. È ovvio, innanzitutto, che occorre evitare i doppioni. Molti sarebbero sorpresi di sapere quante diverse versioni meccanografiche esistono di uno stesso testo, per es. Omero, la Bibbia, la Divina Commedia, e quanti programmi diversi operino lo stesso tipo di spoglio e di elaborazione sullo stesso tipo di macchine.

È necessario che i testi vengano registrati secondo gli stessi criteri tecnici e scientifici, così da poter essere facilmente scambiati tra i diversi ricercatori, e da poter costituire una grande biblioteca elettronica standardizzata, elaborabile con gli stessi programmi fondamentali. L'adozione di uno schema unico di registrazione assicura anche che vengano riprodotte tutte le informazioni presenti in un testo a livello grafemico, e quindi ne garantisce la utilizzabilità per le ricerche diverse. L'adozione di un tale *standard* permette ovvie economie di tempi e costi, e risponde a precise esigenze di ordine scientifico: i risultati dello spoglio di un testo dovrebbero prestarsi alla comparazione con gli altri testi dello stesso autore, della stessa epoca, della stessa scuola, dello stesso genere letterario, ecc.

1.2.2. « *Utility programs* ».

La standardizzazione degli archivi di testi è condizioni necessaria per la standardizzazione dei programmi.

Ci sono oggi per esempio numerosissimi programmi di contestualizza-

indirizzo e gli eventuali nuovi caratteri di correzione; immettere le correzioni nel calcolatore; stampare le righe corrette; controllare che le correzioni siano state apportate nel luogo e nel modo voluti. Con il teleschermo invece l'utente semplicemente muove con un tasto il cursore, un contrassegno che indica e determina il carattere sul quale si può operare in un dato momento. L'utente può, operando con una tastiera, cancellare, sostituire, inserire determinati caratteri, parole, o intere frasi. Se qualcosa viene cancellato, lo spazio vuoto viene riempito automaticamente dal testo che si muove verso sinistra; se qualcosa viene aggiunto, il testo si sposta proporzionalmente verso destra. Alcuni terminali hanno anche la capacità di *formatting* che semplifica la codificazione nelle applicazioni (per esempio quelle bibliografiche o nelle predisposizioni di lessici) in cui una stessa struttura si ripresenta ripetutamente. Così, per esempio, la struttura appare sullo schermo come un modulo che deve essere riempito: autore, titolo, editore, città, data. L'operatore riempie gli spazi appositi. Ogni elemento è così classificato automaticamente in virtù della sua posizione sullo schermo prima di entrare in memoria. Un altro vantaggio è dato dal fatto che nel caso della battitura del testo o delle correzioni il calcolatore può integrare l'operatore applicando un programma che esercita dei controlli di correttezza formale: per esempio verifica se sono rispettate certe regole di fonotassi o consulta un lessico automatico. (cfr. Szanser 1969).

zione, cioè programmi che per ogni parola del testo (o per alcune categorie di parole del testo) producono un *record* costituito da almeno tre elementi: la parola, il suo riferimento al testo, il suo contesto immediato. Evidentemente, le differenze tra i diversi programmi consistono nell'algoritmo che ritaglia il contesto. Dopo aver esaminato i diversi tipi di algoritmi in uso presso i centri specializzati nello spoglio di testi, la *Divisione Linguistica del CNUCE* ha messo a punto un programma di contestualizzazione al quale l'utente può chiedere, con poche *control–cards*, di eseguire l'uno o l'altro tipo di algoritmo (11). Questo lavoro di generalizzazione è stato fatto per tutte le fasi dello spoglio lessicale, dello spoglio fonetico, di alcune elaborazioni statistiche. Le esperienze condotte su circa 2.000 testi per circa 50 milioni di parole in 28 lingue diverse, ci consentono di affermare che le nostre procedure sono utilizzabili, grazie a poche schede–controllo, per elaborare qualsiasi testo in qualsiasi lingua, o almeno tutti i testi che possono essere ricondotti a una scrittura alfabetica.

1.2.3. *Lemmatizzazione semiautomatica.*

L'insieme delle operazioni comunemente raggruppate sotto il termine *lemmatizzazione* richiede una serie di interventi umani che, rompendo il ritmo delle elaborazioni interamente automatiche dello spoglio, aumenta considerevolmente il tempo, il costo, e i rischi di errore.

Molti ricercatori, in particolare anglosassoni e statunitensi, decidono di non lemmatizzare affatto i testi, e si limitano a produrre degli indici, delle concordanze, o delle schede–contesto nei quali gli esponenti non sono unità definite secondo criteri linguistici, ma semplici *forme grafiche*, e cioè 'parole' come le riconosce abitualmente il calcolatore: sequenze di lettere tra due spazi o tra due separatori in genere. Indubbiamente questa semplificazione ha qualche vantaggio. La velocità del calcolatore nell'operare su simboli viene sfruttata completamente e si possono produrre rapidamente e a costi minori grandi quantità di spogli che, se diffusi, rendono innegabili servigi agli studiosi. Alle volte ci sono ragioni scientifiche che sconsigliano la lemmatizzazione, per esempio nel caso di testi in lingue o strati di lingua poco noti, nei quali molti lemmi non sarebbero attestati o addirittura neppure ricostruibili.

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

(11) Si veda a questo proposito Zampolli (1972).

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

1.2.3.1. *Il lessico automatico.*

Per maggiore chiarezza riassumo rapidissimamente cosa si intende con i termini LA e consultazione di un LA. Mi riferirò costantemente, per brevità di esposizione, alla più semplice tra le diverse possibili strutture di un LA In nota è descritta una organizzazione più complessa (12). Nella sua forma più semplice un LA consiste in una serie di *forme grafiche* registrate su nastro, su disco, o su altro supporto leggibile dal calcolatore. Queste forme sono ordinate alfabeticamente. Ogni forma è accompagnata da una serie di informazioni linguistiche di natura diversa a seconda dei diversi impieghi cui il LA è destinato. Chiamiamo *analisi* o *funzione* di una forma l'insieme delle informazioni che la accompagnano: per esempio se il LA è compilato come ausilio agli spogli lessicali, per ogni forma saranno dati il lemma cui la forma deve essere assegnata e, spesso, la classificazione grammaticale e morfologica del lemma e della forma. Se il LA è compilato per tradurre automaticamente da una lingua all'altra, saranno date anche la forma corrispondente nella lingua di uscita, alcune indicazioni per l'analisi sintattica e semantica della proposizione, ecc. Se il LA serve per statistiche fonematiche, etimologiche, sociolinguistiche, saranno dati anche la trascrizione fonematica della forma, la sua etimologia, i suoi registri d'uso, ecc.

(12) Un metodo di ricerca più complesso, usato soprattutto come input al *parser*, è quello della segmentazione. Nel LA sono registrate non tutte le forme grafiche di un lemma, ma solo il suo tema (o i suoi diversi temi, se necessario); il termine *tema* non deve essere inteso nel senso tecnico della linguistica, ma nel senso operativo di sequenza di grafemi che resta invariabile nel corso di tutta (o parte) la flessione. Il tema è accompagnato da un codice che rinvia a una tabella ove sono enumerate tutte le desinenze possibili, nel sistema morfologico considerato, per la categoria alla quale il tema appartiene. Parole invariabili e forme irregolari sono registrate nel LA come termini indipendenti. Se la parola da lemmatizzare non coincide con uno di questi termini, il programma cerca di segmentarla in *tema* + *desinenza*, in modo che il tema sia presente nel LA e che la desinenza faccia parte di quelle indicate nella tabella associata a tale tema. Naturalmente quando il programma trova più di una segmentazione corretta, tratterrà la forma come omografa. I sistemi di segmentazione variano da algoritmi interessanti da un punto di vista puramente informatico ad algoritmi che incorporano una teoria linguistica morfologica specifica, e funzionano come interpreti di regole morfologiche e morfofonologiche. Per una discussione di questi algoritmi, particolarmente interessanti quando trattano oltre ai fenomeni di flessione, elisione e troncamento, anche quelli della derivazione mediante affissi e della composizione, si veda Kay (1974); il sistema di consultazione per segmentazione é indubbiamente più interessante dal punto di vista algoritmico, ma non è detto che sia sempre più economico della ricerca per forme. Per confrontare correttamente il rendimento dei due metodi, occorre tener conto dell'intera procedura di spoglio nella quale la consultazione è inserita, e delle dimensioni del testo da lemmatizzare. La ricerca per forme e non per temi conviene quando la procedura prevede in ogni caso la alfabetizzazione del testo e il suo rendimento cresce proporzionalmente al numero complessivo delle occorrenze. Il LAI (Lessico Automatico Italiano) esiste nelle due versioni, e cioè come inventario di temi (150.000 circa) e di forme (1.000.000 circa).

A seconda degli scopi cui il LA è destinato può variare sensibilmente anche il numero delle forme che lo compongono. Per esempio se il LA serve per tradurre testi scientifici relativi a una disciplina specifica, esso contiene di solito solo le voci più frequenti nella lingua comune e i termini tecnici della disciplina in questione. Se invece il LA serve per statistiche sull'intero sistema lessicale, l'insieme delle voci che compongono il LA è molto più ricco. Il LA si propone in tal caso come un sottoinsieme rappresentativo dell'intero lassico di una lingua, e al limite vorrebbe coincidere con esso (13).

È importantissima, dal punto di vista applicativo, la distinzione tra forme univoche e forme omografe. Diremo *univoca* una forma alla quale, nel LA, corrispondono due o più analisi distinte. Per esempio se in un dato LA la analisi delle forme è rappresentata solo dal lemma, una forma come l'italiano *dica* sarà, in tale LA, univoca: essa può appartenere infatti solo al lemma *dire*. Sarà invece omografa nello stesso LA, la forma *amo*, che può appartenere sia al sostantivo *amo* sia al verbo *amare*. Consideriamo invece un LA nel quale l'analisi assegnata a ciscuna forma comprenda, oltre al lemma, anche la classificazione morfologica della forma (genere, numero, grado per le forme nominali; modo, tempo; persona per i verbi, ecc.). In un tale LA sarà omografa anche la forma *dica*, cui corrisponderanno 4 distinte analisi: rispettivamente 3ª persona singolare dell'imperativo, 1ª, 2ª, e 3ª persona singolare del congiuntivo presente. Se invece le analisi di un LA comprendessero solo la trascrizione in alfabeto fonetico delle forme, in tale LA forme come *amo* e *dica* sarebbero univoche, perché vanno in ogni caso trascritte rispettivamente come /ámo/ e/ díka/, e sarebbero omografe solo le forme non omofone in italiano, come *ancora* (/ánkora/ – /ankóra/) e *pèsca* (/pèska/– /péska/). Come esempio possiamo riferirci al Lessico Automatico Italiano (LAI) che stiamo predisponendo al CNUCE sotto gli auspici del Comitato 08 del CNR.

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

(13) Si pone il problema se il lessico di una lingua debba o no considerarsi finito. Spesso la risposta è diversa a seconda che si tratti di morfemi o di parole. Si veda per esempio Bloomfield (1933, cap. 6) Hjemslev (1953, cap. 12 e 14) Spang–Hanssen (1967) Wagner (1967, p. 17) Ch. Muller (1968, p. 134). L. Guilbert (1967, pp. 116 e segg.) e Rey–Debove (1970, pp. 3 e segg.) hanno discusso il problema dal punto di vista generativista. Si cfr. anche Zampolli (1973 b, p. 165) Un LA può essere compilato con procedimenti diversi. Per una lingua il cui lessico è ben noto, si registrerà dapprima un lemmario ottenuto, di solito, dall'unione delle nomenclature dei principali dizionari esistenti. Si applicherà poi a questo lemmario un algoritmo di flessione capace di generare tutte le forme possibili secondo il sistema linguistico dato. Se invece si deve lemmatizzare un corpus per la cui lingua non esiste ancora un dizionario attendibile, si procede per gradi. Le forme del primo testo del corpus saranno lemmatizzate a mano. Il secondo testo sarà lemmatizzato adoperando un LA contenente solo le forme trovate nel primo: il LA lemmatizzerà così automaticamente solo le forme del secondo testo che comparivano anche nel primo, mentre quelle che sono nuove rispetto al primo saranno lemmatizzate a mano, e poi saranno inserite nel LA. Così il LA adoperato per lemmatizzare il terzo testo sarà costituito dall'unione delle forme dei due testi precedenti, e così via.

cale, la polisemia, la scomposizione in morfemi, gli eventuali registri di uso, le costruzioni sintattiche. Queste informazioni sono espresse in alcuni casi con un linguaggio completamente formalizzato (per esempio le costruzioni), mentre in altri casi la definizione è costituita da una parte formalizzata e da un'altra in linguaggio naturale, come nella esplicitazione delle polisemie. Un algoritmo di flessione ha generato automaticamente circa 1.000.000 di forme, che sono registrate sia secondo l'ortografia corrente sia secondo l'alfabeto fonetico. Si veda in proposito Zampolli (1973 b).

1.2.3.2. *Consultazione di un LA.*

L'algoritmo di consultazione di un LA così organizzato è estremamente semplice. Come si vede nell'organigramma di Fig. 1, in ingresso vengono posti il LA (1) e il nastro delle concordanze per forma (2) [14].

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

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

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

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

Il programma la scrive con i relativi contesti e riferimenti sul nastro di uscita (3) che, verrà adoperato per stampare le concordanze di tutte le *parole nuove*, e cioè presenti nel testo ma assenti nel LA. I lemmi (e le altre eventuali informazioni) vengono scritti a mano in questo elenco e di qui vengono perforati su schede (6) che servono sia per lemmatizzare le parole nuove (nastro 10) sia per inserire le nuove voci nel LA (LA integrato con forme nuove, nastro 9).

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

Salvo quanto dirò in seguito a proposito delle possibilità di distinguere automaticamente gli omografi, la loro lemmatizzazione deve essere fatta a mano. Tutte le parole omografe possibili [15] del testo vengono ricopiate sul

(14) Come si ricorderà, tale nastro contiene le parole del testo ordinate alfabeticamente, ciascuna registrata in un proprio record e accompagnata da riferimento e contesto.

(15) Nella terminologia degli spogli è invalsa, forse impropriamente, l'abitudine di distinguere tra omografia *possibile* nel sistema e omografia *attuale* in un corpus. Il primo termine indica forme grafiche che, secondo i criteri di lemmatizzazione prescelti, possono appartenere a più lemmi nel sistema linguistico cui il corpus viene riferito. Il secondo termine designa forme che, oltreché possibili nel sistema, sono anche di fatto presenti nel corpus come realizzazioni di almeno due lemmi diversi. Naturalmente, il fatto che una forma omografa possibile non sia omografa attuale in un corpus può essere constatato solo esaminandone tutte le occorrenze, e si deve indicare con oppurtune annotazioni all'utente dello spoglio quale, tra le diverse possibilità, è quella di fatto realizzata.

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

1.2.3.3. *Obiezioni all'impiego di un LA per la lemmatizzazione.*

Com'è noto, i primi LA furono messi a punto per la traduzione automatica dal russo all'inglese e viceversa, a partire dal 1946; agli inizi, ci fu anzi chi ritenne che un buon LA fosse condizione necessaria e sufficiente per tradurre automaticamente.

I primi testi di linguistica computazionale dedicano grande spazio ai sistemi per compilare e per consultare un LA (16). Tutti questi studi hanno prodotto sistemi molto noti e affinati per la gestione e la consultazione di un LA.

Non deve però stupire se, nonostante questo, pongo l'impiego dei LA tra i possibili sviluppi della lessicografia assistita dai calcolatori. In realtà sono pochissimi le imprese lessicografiche e in genere gli autori di spogli lessicali e statistici che abbiano adottato un LA per rendere automatica o almeno semiautomatica la lemmatizzazione.

Molti obiettano che la lemmatizzazione per mezzo di un LA secondo la procedura abituale espone al rischio di gravi errori. Il più grave sarebbe quello che una forma di fatto omografa nel sistema linguistico cui il testo da lemmatizzare appartiene, sia stata invece inserita come univoca nel LA, per errore o per insufficente conoscenza della lingua. In questo caso il calcolatore lemmatizzerebbe direttamente come univoche le occorrenze di tale forma nel testo, senza sottometterle all'esame del linguista. Più lo strato di lingua studiato si estende diacronicamente, più si arretra nel tempo verso sistemi che non sono completamente presenti alla competenza di chi compila il LA, più questo rischio è grave (17).

(16) Si vedano per esempio gli articoli il Locke e Booth (1955), e, per una bibliografia di questo periodo, Mounin (1964).

(17) Una forma può essere omessa nel LA o perché è stato omesso il lemma corrispondente, o perché, pur essendo presente il lemma, l'algoritmo di flessione non prevede regole capaci di generare tale forma. In questo caso, non del tutto improbabile quando si ha che fare con stati di lingua non del tutto consolidati o che includono un grande numero di varietà regionali e simili (si pensi per esempio alle varianti della lingua italiana delle origini), conviene, se possibile, adottare la consultazione per segmentazione, che permette di arricchire e complicare le regole del componente morfologico e morfonologico, aumentando così il numero delle forme potenzialmente generabili dal sistema.

Tuttavia l'ostacolo è superabile se, anziché considerare definitiva la lemmatizzazione delle forme univoche secondo il LA, se ne stampano i contesti lemmatizzati affinché il linguista controlli i lemmi attribuiti dal LA (cfr. n. 12 in Fig. 1).

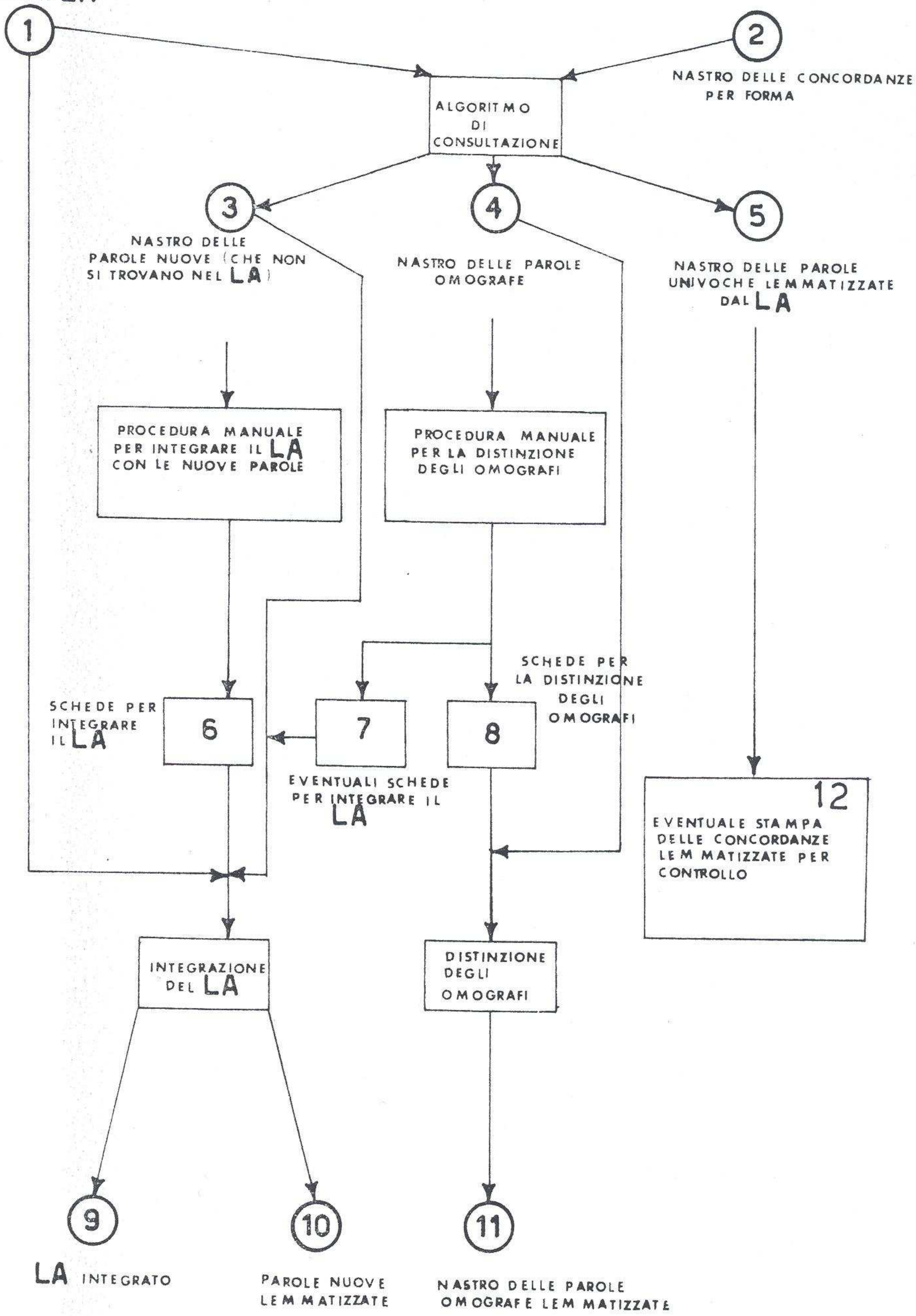

Fig. 1.

Appositi esperimenti hanno mostrato che anche in questo casi si risparmierebbe sempre un tempo notevole (il 70% circa) rispetto alla procedura di lemmatizzazione manuale.

Un'altra obiezione è quella di coloro i quali affermano che si dovrebbe creare un LA diverso per ciascuno spoglio, perché a seconda della natura del testo o degli scopi, l'autore dello spoglio può adottare criteri linguistici diversi nella lemmatizzazione. Divergenze di questo tipo certamente sono giustificabili in alcuni casi anche se, a mio avviso, è evidentemente auspicabile che i criteri di lemmatizzazione siano comuni per il maggior numero possibile di testi: si pensi per esempio alla necessità di poter comparare tra loro i dati sulle frequenze lessicali di testi spogliati da ricercatori diversi.

Tuttavia, non è vero che un LA debba necessariamente produrre un solo tipo di lemmatizzazione. Entro certi limiti, esso permette diversi tipi di lemmatizzazione, e ne assicura nel contempo la comparabilità.

Evitando il gran numero di trascrizioni necessarie nella procedura artigianale, l'uso di un LA diminuisce notevolmente il rischio di errori casuali. Nello stesso tempo funge da sistema di riferimento che assicura coerenza di comportamento nei casi nei quali la presenza di alternative sistematiche richiede che la formulazione del lemma avvenga secondo regole prefissate. Il LA funziona come registrazione delle decisioni prese e, potendo essere stampato rapidissimamente, o consultato tramite video, fornisce a ogni istante un quadro completo delle forme esaminate e dei trattamenti adottati. Accanto a questa funzione normalizzatrice, che è tanto più importante quanto più è esteso il corpus sottoposto a spoglio e quanto più numerosi e distribuiti nel tempo sono i lemmatizzatori, il LA esercita un insieme di funzioni collaterali; si pensi per es. alla possibilità, suggerita tra gli altri da C. A. Mastrelli per l'Accademia della Crusca, di retrodatare automaticamente le forme nel corpus scelto.

Un'altra obiezione riguarda il costo e il tempo richiesti dalla compilazione di un LA. Per es. per completare la parte del LAI che serve per la lemmatizzazione di testi italiani contemporanei, hanno lavorato 7 persone per 2 anni. Si deve però considerare che questo tempo equivale al tempo che il LAI permetterà di risparmiare nella lemmatizzazione di 2 milioni di parole, appena un ventesimo delle parole presenti nella biblioteca elettronica del CNUCE.

Occorre soprattutto considerare che il LA, e quindi il tempo speso per compilarlo, sono utili non solo per la lemmatizzazione, ma anche per numerose ricerche linguistiche, alcune delle quali sono premesse necessarie per lo sviluppo stesso della lessicologia e della lessicografia, sia in quanto discipline autonome, sia in quanto assistite dal calcolatore. Accanto a statistiche sulle unità dei diversi componenti che costituiscono il sistema linguistico italiano [18],

(18) Si vedano, per esempio a proposito delle statistiche fonologiche, Troubeckoy (1939 cap. 7) Martinet (1960, n. 3.38), Hockett (1955 e 1966) Kramsky (1964), Tagliavini (1968) Muljacic (1969, p. 306) Zampolli (1968 e 1973 b), Saramandu (1966), Guiraud (1954 e 1967), Dubois (1964).

il LAI permette la costruzione di modelli dei meccanismi operanti nella organizzazione del lessico [19] e di una banca di dati linguistici [20].

1.2.3.5. *Distinzione automatica degli omografi.*

Risulta evidente dalla procedura esposta al n. 1.2.3.2. che per completare la automazione della lemmatizzazione occorrerebbe poter distinguere automaticamente gli omografi. Questo problema è stato ed è affrontato da tempo nei progetti di traduzione automatica, mentre conosco solo pochi tentativi da parte di imprese lessicografiche. A Nancy l'èquipe del *Trèsor* aveva iniziato a studiare degli algoritmi per gli omografi di altissima frequenza, e lo stesso stiamo facendo a Pisa. Analoghi tentativi sono in corso presso la 'Accademia Reale Spagnola' (Del Campo *et al.*, 1974).

Per l'omografia funzionale (del tipo sostantivo–verbo: *faccia*), si propone un *parser* sintattico.

Il principio è evidente. Il LA associa a un omografo di questo tipo, non una sola analisi grammaticale, ma più analisi distinte, una per ciascuna diversa funzione sintattica che l'omografo potrebbe esercitare. Se il *parser* ha successo e assegna un solo indicatore sintagmatico alla frase, allora il calcolatore potrà risolvere automaticamente l'omografia scegliendo la descrizione grammaticale dell'omografo che ha permesso al *parser* di avere successo.

Per l'omografia di tipo radicale, del tipo *mozzo* (della ruota e della nave), che non comporta diverse analisi grammaticali, si proponeva fino al 1966–67 un *parser* a livello semantico. Si pensava per esempio di classificare tutte le voci presenti nel LA secondo categorie o comunque componenti semantiche, e di formulare delle regole specificanti la possibilità o impossibilità di relazione, di selezione, di coocorrenza tra le diverse categorie semantiche, almeno in certe posizioni della struttura sintattica, con il proposito, più o meno esplicito, di costruire un modello semantico globale, un reticolo di rapporti collegante, al limite, tutte le parole del lessico.

Oggi la situazione è cambiata e si tende a non separare i due momenti del *parser*. Lo stato delle procedure di analisi sintattica e semantica è descritto in Zampolli (1974 e), al quale qui rinvio. Qui basterà osservare che i sistemi più recenti citati in tale rassegna, se da un lato riescono a trattare una grande varietà di strutture sintattiche complesse, dall'altro sono limitati di solito a un sottoinsieme lessicale molto ristretto, e il loro funzionamento presuppone che gli enunciati che vengono sottoposti ad analisi si riferiscano a un argomento preciso noto a priori, a una parte di realtà ben delimitata, la cui conoscenza è data al sistema come una rappresentazione formale esplicita.

Le occasioni di incontro tra lessicografi e lessicologi, impegnati nello spoglio di vasti corpora di testi, e linguisti computazionali impegnati nello studio dei cosidetti sistemi integrati di analisi linguistica sono mancate anche nel pas-

(19) Si veda per esempio la ricerca sui suffissi in Zampolli (1973 e).
(20) Si veda più avanti il capitolo sulla banca dei dati.

sato più recente. Negli ultimi tempi si è cercato deliberatamente di creare delle occasioni di incontro e di discussione, da parte di ricercatori che, come lo scrivente, avendo esperienza in entrambi i campi, sono convinti della possibilità e necessità di scambi metodologici.

Nel corso della Scuola Estiva 1972 e della ICCL 1973 già citate, gli autori dei sistemi di *parser* più recenti, interrogati da lessicografi e lessicologi sulla possibilità di estendere i loro modelli fino a includere sottoinsiemi sempre più vasti, e al limite tutta la lingua di uso comune, hanno espresso giudizi contrastanti sulla possibilità teorica di tale estensione, ma hanno tutti concordato sulla impossibilità pratica di realizzarla concretamente, almeno per numerosi decenni. Ciò perché, a differenza dei sistemi di *language understanding*, che si occupano di sottoinsiemi di lingua naturale estremamente ridotti, e per lo più in certa misura già formalizzati, i lessicografi sottopongono a spoglio corpora di testi distribuiti ampiamente sia diacronicamente sia sincronicamente (regione, argomento, registro d'uso, genere letterario, ecc.). Si deve quindi concludere che la soluzione automatica di tutte le omografie di un testo è almeno oggi una meta utopistica. Ma non si deve escludere né rinunciare ad automatizzare in parte la soluzione dell'omografia. Anche se, come vedremo, è raro che venga studiata o almeno resa nota l'efficienza dei sistemi di *parser*, possediamo comunque alcuni dati indicativi, secondo i quali su testi di lingua inglese, scelti senza particolari restrizioni, sarebbe possibile risolvere automaticamente, incorporando alcuni dei ritrovati più recenti dei sistemi integrati, il 60% delle omografie. Con algoritmi ad hoc, che, pur incorporando tali ritrovati, non mirino a riconoscere la struttura completa della frase, ma si propongano di risolvere sopratuttto l'omografia funzionale la quale, come è noto, è generata per la maggior parte da poche parole di altissima frequenza, sembrerebbe possibile raggiungere e in talune lingue sorpassare l'80%. Risultato apprezzabilissimo perché, a differenza dei sistemi integrati ove importa il processo di analisi per se stesso, negli spogli importa essenzialmente ridurre il lavoro umano.

Probabilmente i primi passi concreti dovrebbero essere mossi verso una interazione uomo–macchina nello sciolgimento della omografia, in attesa di raggiungere uno ' stato dell'arte ' nel quale il programma dovrebbe eseguire il parsing delle frasi per le quali è adeguato, e dovrebbe richiedere la collaborazione del linguista per quelle che sono al di fuori delle sue capacità. Questo colloquio uomo–macchina richiede l'uso di un terminale video. Alcune dimostrazioni di questo colloquio le abbiano organizzate con B. Quemada per i partecipanti alla Scuola Estiva nel 1972.

Sullo schermo appare la forma da analizzare. Accanto ad essa appaiono le diverse proposte di analisi fornite dal LA e dal *parser*, numerate progressivamente.

Al di sotto di queste informazioni, il ricercatore può far apparire, uno dopo l'altro, i contesti della forma. Se il *parser* è riuscito ad analizzare alcune occorrenze, il lemmatizzatore controlla l'esattezza dell'analisi scelta. Altrimenti l'opera egli stesso, per mezzo della tastiera o del *light–pen* scegliendo

tra le analisi proposte dal LA. Nel caso che nessuna di queste convenga ad una certa occorrenza, egli aggiungerà la nuova analisi a quelle già esistenti nel LA. Evidentemente quest'ultimo caso è equivalente, dal punto di vista della procedura, al caso di un testo da analizzare per la cui lingua non esiste un LA: mano a mano che l'analisi procede, si forma un LA che viene arricchito dagli spogli successivi.

1.2.3.6. *Selezione dei materiali lessicali.*

Il tema della selezione domina il processo di compilazione dei dizionari e in particolare dei dizionari storici (21).

Per incominciare, i compilatori devono scegliere un corpus di testi, che possa essere considerato rappresentativo, ai fini del dizionario, del corpus totale dei testi teoricamente disponibili nella lingua in questione (22). Non si può negare l'opportunità di una sorta di *feedback*. Mano a mano che lo spoglio, procedendo, fornisce nuove informazioni sulla struttura del campione in esame, dovrebbe essere possibile modificare il corpus stesso, per esempio riducendo il numero dei testi relativi a un sottoinsieme di lingua o a classi di fatti linguistici sufficentemente documentati; oppure, viceversa, introducendo testi che si suppone contengano fenomeni non ancora apparsi, ecc. Questo tipo di *feedback*, fino ad oggi pochissimo realizzato, richiede una organizzazione particolare dello spoglio che consenta di gestire un archivio costantemente aggiornato e rapidamente accessibile in tutte le sue parti. Le caratteristiche di un siffatto archivio sono quelle proprie della cosiddetta *banque de mots*, di cui parleremo al numero seguente, organizzata attraverso un LA.

(21) Il dizionario storico tradizionale, del tipo rappresentato dall'*Oxford English Dictionary*, dalle diverse edizioni del *Vocabolario della Crusca* e da altri dizionari simili per molte lingue nazionali, può essere considerato essenzialmente come l'insieme ordinato di numerose citazioni selezionate da un vasto corpus di testi relativi alla lingua e al periodo abbracciati dal dizionario. Le citazioni che accompagnano ogni lemma sono riunite dal redattore per mostrare le forme, gli usi e le collocazioni di ciascuna parola e le distribuzioni di ciascuna di queste caratteristiche nel tempo, nello spazio, nel livelli di stile, ecc. « From this point of view the definitions serve merely as sign–posts or markers to the group of citations which follow them, and the primary function of the dictionary is to present the selection of citations and only secondarily to provide a list of definitions or ' meanings ' ». (J. A. Aitken, 1972). Di conseguenza dizionari di questo tipo, che presentano raccolte di parecchie centinaia di migliaia di esempi, sono principalmente il frutto di un procedimento di selezioni successive.

(22) Per esempio: J. A. Aitken ha calcolato per che il DOST (Dictionary of the Older Scottish Tongue) il totale delle parole nei testi disponibili di scozzese antico sia dell'ordine di 1.000 milioni. In pratica solo un quinto di queste, circa 200 milioni, (pari a 3.000 volumi) possono essere incluse nel corpus. La scelta del corpus campione avviene di solito sulla base di giudizi sull'importanza e sulla rappresentività dei testi, o in seguito a fattori contingenti, per esempio la disponibilità di edizioni critiche. Sono molto rari i casi nei quali è possibile spogliare l'intero corpus di testi disponibili: si tratta per lo più di sottoinsiemi di lingua ben delineati o di lingue morte delle quali sono sopravvissuti relativamente pochi testi.

Il secondo processo di selezione consiste nello sceglicre da corpus gli esempi che devono confluire nell'archivio.

I lessicografi sono d'accordo, in genere, con il calcolo di Aitken secondo il quale un esperto lessicografo può esaminare, redigendo un articolo di dizionario, circa 10.000 schede contesto all'anno. A questa velocità elaborare una documentazione composta da 10.000.000 di citazioni richiederebbe 100 collaboratori per 10 anni. In pratica gli archivi dei grandi dizionari storici basati su spogli manuali negli ultimi 150 anni comprendevano da un mezzo milione a 10 milioni di occorrenze, e ciononostante i preventivi di tempi sono stati quasi sempre superati di molte decine di anni. L'impiego del calcolatore, che permette di raccogliere in un tempo relativamente breve un numero di esempi molto superiore, rende, se possibile, più drammatico il problema di scegliere le occorrenze particolarmente « interessanti » da inserire nell'archivio [23].

L'Accademia della Crusca e il CNUCE hanno messo a punto una procedura per la composizione di un archivio lessicale che è, a tutt'oggi, da un punto di vista globale, tra le più economiche e perfezionate [24]. Lo spoglio elettronico prevede la registrazione integrale del testo in *machine readable form* e la produzione delle concordanze di tutte le forme del testo. Una equipe di lemmatizzatori legge le concordanze per forma lemmatizzandole, e nel contempo contrassegnando le occorrenze ritenute degne di entrare nell'archivio. Alcuni esperimenti per automatizzare almeno in parte questa selezione sono stati compiuti al C.N.U.C.E. e presso il T.L.F. di Nancy. Essi si fondano soprattutto su metodi statistici e i risultati ottenuti sono ancora molto discutibili. Nel corso della Scuola Estiva di Pisa del 1972 si è ampiamente discusso dei mezzi atti ad abbreviare queste operazioni di scelta e a fondarle, se possibile, su criteri non esclusivamente dipendenti dal giudizio individuale del lemmatizzatore. L'idea di base prevede l'interazione del lessicografo con il programma, per mezzo di uno schermo video. Il calcolatore assisterebbe attivamente il lessicografo: nel procedimento di lemmatizzazione, mediante la consultazione di un dizionario di macchina; nel processo di selezione, mediante il riconoscimento automatico nel testo di alcune strutture descritte in precedenza come fattori rilevanti per l'accettazione o la esclusione degli esempi, e mediante il continuo controllo statistico dell'accumularsi degli elementi scelti.

Il terzo processo di selezione appartiene alla fase di redazione vera e propria del dizionario. I redattori devono analizzare i materiali raccolti nello archivio, scegliere gli esempi più rappresentativi ed organizzarli opportuna-

(23) La forma tradizionale dell'archivio é quella di un insieme di schede–contesto; ad ogni parola scelta nei testi spogliati viene intestata una scheda, nella quale si ricopia anche il riferimento al luogo del testo da cui proviene e il contesto nel quale appare. Le schede vengono ordinate, di solito, secondo l'ordine alfabetico delle intestazioni (forme, lemmi o lessemi) e, all'interno di una stessa intestazione, in ordine cronologico, di autore, di opera, ecc.

(24) Essa prevede tre tipi di spoglio, distinti dalla densità: lo spoglio fitto viene realizzato con mezzi elettronici; lo spoglio medio con procedure xerografiche; lo spoglio rado con la tradizionale schedatura a mano.

mente per illustrare ciascuna accezione, la sua distribuzione ed evoluzione nel tempo, nello spazio, ecc., scrivere la definizione e alla fine seguire fino alla stampa il risultato di tutte le operazioni. Si può immaginare un procedimento di redazione assistita dal calcolatore. Il LA fornirebbe innanzitutto un primo schema della struttura dell'articolo, eventualmente modificabile in base alle nuove evidenze emergenti dall'esame dell'archivio.

Il redattore potrebbe far comparire in sequenza sul video gli esempi, e, per mezzo di semplici comandi sulla tastiera o con il *light–pen*, eseguire in una stessa fase le seguenti operazioni:

indicare con dei numeri il raggruppamento degli esempi in classi;

chiedere al calcolatore di ripresentare gli esempi raggruppati secondo le classi prescelte;

indicare gli esempi che devono confluire nel dizionario, eventualmente tagliando nel modo più opportuno quelli troppo lunghi;

inserire in testa ad ogni gruppo la definizione o, più generalmente, i classificatori pertinenti;

inviare i materiali così elaborati (definizioni e esempi) ad un processo automatico di stampa, per es. la fotocomposizione.

Uno schema di questo tipo è stato discusso a Pisa (Aitken, 1972; Bailey, 1972; Quemada, 1972; Zampolli, 1974 b) e la Sezione Linguistica del CNUCE sta sviluppando una procedura interattiva che completa un primo esperimento presentato nel 1972 alla Scuola Estiva [25].

Si può giustamente obiettare a questo schema che non sarebbe possibile operare, sullo schermo, l'esame sinottico dei diversi esempi, cosa che il lessicografo è abituato a fare « sparpagliando » le schede–contesto sulla scrivania. Alcuni hanno proposto di rimediare mettendo due video in parallelo, su uno dei quali sarebbe fisso il contesto in esame, mentre sull'altro verrebbero fatti scorrere dal ricercatore gli altri contesti della stessa forma. Altri hanno proposto di ricorrere a tecniche particolari, che combinano assieme un terminale video e le microfiches. Sembra certo che occorrerà ancora un serio e forse lungo lavoro di sperimentazione in stretta collaborazione tra linguisti computazionali e lessicografi. In ogni caso questa è certamente una delle più interessanti prospettive di sviluppo, che si inserisce naturalmente nel progetto più generale di una *banque de mots*.

(25) Il ricercatore può richiedere al calcolatore di raggruppare immediatamente i contesti che hanno ricevuto la stessa analisi, e di far scorrere ciascun gruppo sullo schermo, preceduto dalla analisi e da altri eventuali commenti da lui inseriti in precedenza. Nello stesso tempo è possibile operare, per es. con l'opportuno uso del *light–pen*, una scelta dei contesti da inviare all'articolo del dizionario, ed eventualmente « tagliare » i contesti troppo lunghi nel modo più opportuno. Basterà infatti per esempio appoggiare il *light–pen* sopra il primo carattere di una parola e sotto l'ultimo carattere di un'altra per comandare al calcolatore di eliminare dal contesto tutte le parole comprese tra i due punti segnati, sostituendole per esempio con dei puntini tra parentesi.

1.3. La banque de mots.

Alla Scuola di Pisa si sono delineate due diverse interpretazioni del termine *banque de mots* o archivio lessicale.

Da un lato alcuni, come Bahr (1972) l'hanno inteso come una specie di LA, nel quale ogni parola è accompagnata da tutte le informazioni fonologiche, morfologiche, sintattiche, semantiche note. Quemada ed io l'intendevamo come un vero e proprio archivio di parole estratte da corpus di testi analizzati, elaborati, e non stampati, ma conservati per il calcolatore e messi a disposizione dei linguisti e dei lessicografi mediante le moderne tecniche conversazionali di interazione uomo–macchina.

Penso che una *banque de mots* o archivio lessicale debba contenere entrambi questi elementi, il LA e i testi, e che, poiché l'archivio è pensato come dinamico, sia il LA sia il corpus debbano avere caratteristiche dinamiche.

Il LA può costituire la registrazione ove si depositano le conoscenze acquisite da linguisti e lessicografi sul lessico e sulle sue relazioni funzionali con gli altri componenti del sistema linguistico. Il LA fornisce così un inventario delle unità del sistema lessicale e delle loro proprietà, le quali fungono da classificatori nei confronti delle unità del testo. L'unità, linguisticamente definita e classificata del LA, è la chiave attraverso la quale normalmente si accede alle unità corrispondenti del corpus.

L'indexamento del corpus richiede il riconoscimento e la classificazione di tali unità nel quadro di una teoria linguistica. Le teorie linguistiche, soprattutto oggi, sembrano mutare piuttosto rapidamente. È inevitabile che chi spoglia corpus di notevoli dimensioni si ponga il problema se, in queste condizioni, i risultati del proprio lavoro siano così legati a una teoria linguistica specifica, da risultare scarsamente utilizzabili o addirittura non interpretabili in una teoria diversa.

La risposta a questo interrogativo, che si pone in maniera tanto più drammatica quanto più ampio è il corpus e quindi proporzionalmente più costoso lo spoglio, non è facile, come si è potuto constatare nel corso delle vivaci e talora aspre discussioni su questo argomento che hanno avuto luogo al *Colloque sur l'indexation maximale* di Strasburgo (1972).

In ogni caso essa non può consistere in un *si* o in un *no* netti, ma deve invece prendere la forma di procedure cautelative che assicurino, nei limiti del possibile, la dinamicità dell'analisi linguistica del corpus.

Tra l'atteggiamento estremo di chi sostiene che l'analisi di un corpus dipende completamente da una teoria e non ha senso se non è spinta fino al livello di dettaglio massimo previsto dalla teoria, e l'altro estremo di chi, dalle stesse considerazioni, conclude che l'indexamento deve arrestarsi alle unità definite al livello grafemico, lasciando all'utente degli indici ogni analisi, si pongono a nostro avviso dei procedimenti intermedi. Questi processi sembrano poggiare su una convinzione, talora non esplicitamente asserita, che oltre alle unità puramente grafemiche sia possibile individuare delle unità

le quali, pur essendo definite con criteri linguistici, sono, per così dire, ' neutre ' rispetto alla maggior parte delle teorie linguistiche e cioè indicate e delimitate univocamente da esse, anche se il loro status e la loro classificazione possono variare da teoria a teoria. Si tratterebbe, in altre parole, di unità minime le quali possono o coincidere con le unità della teoria, oppure possono essere raggruppate in una unità che comprende tutte le loro occorrenze, la quale coincide con una unità della teoria.

In tal caso, il LA sarebbe concepito come un inventario di tali unità minime, e l'analisi consisterebbe nel riconoscere queste unità nel corpus. In altre parole, l'indexamento consisterebbe nel collegare con una unità del LA tutte le sue occorrenze nel corpus. Molti degli spostamenti indotti nello stato, nelle definizioni e nelle reciproche relazioni delle unità della teoria potrebbero essere registrati nel LA senza alterare l'inventario delle sue unità, ma semplicemente modificandone le descrizioni. Il corpus risulterebbe automaticamente analizzato secondo la nuova forma della teoria, senza che sia necessario intervenire su di esso.

La validità di questa soluzione può essere discussa sotto molti punti di vista. Il punto cruciale consiste naturalmente nel dimostrare che esistono le unità di base che possono essere trattate da una teoria o da un'altra semplicemente cambiando la loro descrizione. Non ho lo spazio per discutere questo argomento. Tuttavia posso affermare che la nostra esperienza di spogli conferma che è possibile costruire un inventario soddisfacente di tali unità, o almeno un inventario nel quale il margine di errore dovuto all'adattamento a diversi sistemi teorici sia compensato dalla possibilità di estendere la analisi a corpus di dimensioni altrimenti proibitive [26].

Naturalmente il LA deve essere aperto alle modifiche richieste da nuove teorie (cfr. Lehman, 1971), e può così servire per confrontare la loro efficacia descrittiva.

Ma il LA deve poter essere integrato o modificato anche in base a dati emergenti dall'analisi del corpus.

Una delle informazioni che vengono dal corpus al LA è quella della frequenza. Questo darà modo di colmare progressivamente la lacuna giustamente lamentata da J. Rey–Debove (1969, p. 372). Ad ogni parola, ad ogni sua accezione, alle costruzioni grammaticali cui partecipa, ecc. la banca delle parole dovrebbe permettere di assegnare una frequenza di uso, distinta nei vari sottoinsiemi di testi identificabili sull'asse diacronico e sincronico.

Naturalmente, per poter organizzare il corpus secondo la struttura di accezioni, di relazioni, di proprietà proposte dal LA, è necessario analizzarlo, compito tanto più improbo quanto più l'analisi è dettagliata e il corpus è esteso.

Ritorna qui il problema di come ridurre la sproporzione tra i dati prodotti dagli spogli e possibilità elaborative dell'uomo. Possiamo solo elencare alcuni elementi preparatori per una risposta provvisoria.

(26) Queste affermazioni sono discusse in Zampolli (1974 c) e Zampolli (1974 d).

È necessario in primo luogo coordinare gli sforzi di quanti operano su una stessa lingua, e creare delle banche di dati e di programmi a livello nazionale o internazionale. Queste banche devono essere centralizzate, ma devono essere inserite in una rete di terminali, o di calcolatori, là dove questa esiste. Il centro nazionale che gestisce la banca deve essere dotato di *hardware* e *software* specificamente designati per esigenze linguistiche, ivi compresi sistemi di stampa che permettano delle edizioni soddisfacenti.

L'analisi dei dati deve essere decentrata, e cioè compiuta da piccole èquipes di ricercatori o addirittura da ricercatori isolati, dotati di strumenti modesti, ma aventi accesso diretto alla banca dalla quale estraggono i testi da analizzare e nella quale li ' ridepositano ' analizzati. Le moderne tecnologie dei terminali, del *time–sharing*, dell'uso interattivo permettono una simile cooperazione, mentre l'uso di un LA e l'insieme di criteri di analisi formulati esplicitamente che esso presuppone assicurano un livello accettabile di coerenza tra i diversi esecutori delle analisi. La presenza di un LA funziona come la registrazione di tutte le soluzioni già trovate, e suggerisce uno schema comune di analisi. Dice quali usi, quali costruzioni sono stati trovati, con quale frequenza. Permette un primo livello di automazione dell'analisi: permette una lemmatizzazione semiautomatica e, associata ad un *parser* sintattico, permette il riconoscimento di alcune strutture [27]. Mette in evidenza le lacune del corpus, e suggerisce quando occorre ricorrere alle competenze di informatori [28].

Questo dialogo a 3, LA–corpus–ricercatore, sembra il solo sistema per poter passare da un archivio statico, stampato in una marea di schede o di concordanze, a una moderna banca di dati lessicali che metta a frutto le risorse offerte dalla tecnologia moderna. Di essa verrebbero stampate (di preferenza automaticamente con sistemi a microphiches collegati al calcolatore) solo alcune parti su richiesta dei ricercatori, mentre, nella maggior parte dei casi, essa verrebbe consultata per mezzo di terminali. Per tutto quanto si è detto è importante evidentemente che si abbia una *banque de mots* centralizzata per ogni lingua.

In questa direzione è svolta la attività della Divisione Linguistica del CNUCE di Pisa. Praticamente tutti gli spogli automatici che si eseguono in Italia sono condotti dai diversi Istituti italiani presso la Divisione Linguistica del CNUCE, con unità di criteri tecnici e scientifici. Più di 2000 opere in corso di spoglio costituiscono il primo nucleo sul quale, una volta pronto

(27) Il riconoscimento di alcune strutture privilegiate può richiedere un parser relativamente semplice. Prendiamo l'esempio delle diverse costruzioni di un verbo. Se nel LA esse sono indicate, è possibile far raggruppare tutti i contesti nei quali compaiono degli elementi formali che indichino le diverse costruzioni: per es. le preposizioni. Naturalmente resterebbero degli esempi non analizzati, da sottoporre alla attenzione dei ricercatori.

(28) Il calcolatore, oltre a fornire la documentazione per mezzo degli spogli, è utilissimo anche per le sue capacità di sintetizzare rapidamente delle strutture linguistiche *ad hoc*, la cui « grammaticalità » può essere sottomessa al giudizio del parlante, al fine di esaminare esaurientemente tutte le possibili costruzioni e classificazioni di un lemma.

il LAI, potremo cercare di costruire la *banque de mots* della lingua italiana. Naturalmente la completa interattività di tale archivio è solo una prospettiva del futuro, ma di un futuro forse più vicino di quanto si creda, probabilmente nell'arco di 10–20 anni. Quello che importa, è eseguire *ora* gli spogli in modo da non precludere e neppure ostacolare la realizzazione di questa prospettiva, che fra tutte le mete canoniche della CL sembra la più sicuramente realizzabile. Anche in questo senso è preziosa la preveggenza con la quale il CNR e in particolare A. Faedo (1973) hanno assicurato al mondo scientifico italiano un Centro che, come la Divisione Linguistica del CNUCE, sta creando le premesse di una banca di dati linguistici.

2. STATISTICA LINGUISTICA

Fino agli inizi, si sono delineati due filoni principali negli studi statistici. Il filone che potremmo chiamare filologico–stilistico si propone di scoprire e descrivere caratteristiche quantitative individuali, distintive e misurabili, per adoperarle in una definizione 'oggettiva' dello stile di un autore, o per risolvere problemi filologici di cronologia, di attribuzione, ecc. (29). Il filone che potremmo chiamare linguistico si propone invece la costruzione di modelli quantitativi generali capaci di descrivere le regolarità statistiche osservate nelle distribuzioni di frequenza delle unità linguistiche, e di spiegarne eventualmente le cause (30).

I due filoni si sono sviluppati spesso secondo linee indipendenti. Dato l'argomento della presente comunicazione, ci interesseremo qui al secondo

(29) Per una rassegna storica si veda Bailey, *Statistics and style: a historical survey* (1969); per la bibliografia, Guiraud, Whatmough (1954) e Bailey e Dolezel (1968).

(30) Per questa distinzione si veda anche Plath (1962), il quale distingue i modelli statistici *macrolinguistici* (come esempio dei quali egli discute la legge di Zipf e le diverse interpretazioni proposte) dalle statistiche per lo studio dello stile e dell'attribuzione. Si veda anche Zampolli (1969), p. 9.

Una distinzione analoga si trova in Petöfi (1969). Egli parla di analisi quantitativa « of language » e « of texts ». Nel primo caso si definisce la frequenza di tutti gli elementi che costituiscono la struttura o di tutte le classi concernenti il linguaggio dato, e si ponderano gli elementi della struttura con questi valori (p. 10); in altri termini si studiano le frequenze caratteristiche di un intero linguaggio (p. 11). Nel secondo caso, si investigano le caratteristiche di testi individuali o quelle del linguaggio poetico, scientifico, ecc. (*ivi*). Egli ritiene anzi che le ricerche nel settore della cosiddetta « text–theorie », che sono appena agli inizi, siano state fino ad oggi principalmente statistiche (*ivi*).

Nel secondo filone sono particolarmente noti, anche se non poco discussi (cfr. Levison *et alii*, 1968) gli sforzi di stabilire dei tests di « authorship ». Essi si propongono di scoprire « some habitual occurrence common to all writers of the class being studied (...) which each individual uses in his own way or at his own level. It must be shown that (1)–within any single work of an author and between works of the same author, the variations in the occurrence are such that chance alone will readily explain, and (2)–that between the works of different authors the variations often exceed anything that chance can explain ». (Michaelson e Morton, 1972, p. 27). Si veda a questo proposito anche A. Kasher (1972).

filone, ma dovrebbe apparire chiara, in conclusione, la necessità di ricondurre i due tipi di ricerca a un unico modello esplicativo.

2.1. Gli inizi.

È noto che i primi studi sulla frequenza delle unità linguistiche nei testi non sono state condotte, salvo poche eccezioni, da linguisti, ma da ricercatori provenienti da altre discipline, i quali si proponevano di fondare applicazioni pratiche di vario genere sulla conoscenza delle strutture e delle caratteristiche quantitative dei fatti linguistici.

Così i crittografi hanno compilato delle statistiche sulle frequenze dei grafemi o di gruppi di grafemi nei testi, con l'intento di adoperarle nello stabilire o nel decifrare dei codici. Audiologi e foniatri hanno indagato la frequenza dei ' fonemi ', allo scopo di creare dei tests delle capacità uditive o di studiare le cause e le conseguenze di alcune fonopatie. Stenografi e costruttori di macchine da scrivere hanno compilato statistiche sulla frequenza di grafemi, gruppi di grafemi, sillabe e parole, per perfezionare i propri sistemi di trascrizione privilegiando opportunamente le unità più frequenti. Un pò più tardi, anche gli insegnanti di lingue straniere posero mano a spogli statistici di testi per determinare gli elementi di maggiore frequenza, soprattutto sul piano lessicale, in modo da privilegiarli nell'insegnamento (31).

In questi primi studi, i ricercatori non si riferivano quasi mai a un modello esplicito (32). Tuttavia è possibile riconoscere la giustificazione, la ragione stessa di essere di queste ricerche nella convinzione, implicita, che ciascuna unità linguistica sia caratterizzata da una propria frequenza (o almeno dalla appartenenza a una classe di frequenze più o meno precisamente delimitata) e che questa frequenza sia relativamente stabile nei diversi testi. Se ci si sobbarcava allo spoglio di un corpus (33) per stabilire in esso le frequenze dei diversi grafemi o gruppi di grafemi sulle quali fondare l'ottimizzazione di un sistema stenografico, si riteneva, evidentemente, che nei testi da trascrivere stenograficamente in futuro si sarebbe incontrata una distribuzione di frequenza molto simile.

Però solo nell'immediato secondo dopoguerra si verificano i primi tentativi di dare una formulazione teorica esplicita a queste convinzioni, anche

(31) Si vedano per es. per la crittografia Kerckhoffs (1883) e Valero (1896); per la audiologia Bocca e Pellegrini (1950) e la relativa bibliografia; per la stenografia Kaeding (1897–98), Aliprandi (1940) e Estoup (1912). Per i dizionari di frequenza, l'introduzione a Bortolini, Tagliavini, Zampolli, (1971). In generale è ben documentato Guiraud (1964 b).

(32) Questa assenza di preoccupazioni teoriche va imputata non solo alla mancanza di una preparazione linguistica ' professionale ', ma anche alla difficoltà oggettiva di incorporare nelle teorie linguistiche del tempo i fatti quantitativi.

(33) Si pensi che il lavoro di Kaeding (1988), si basa sullo spoglio manuale di circa 11.000.000 di parole con lo scopo di stabilire le frequenze dei singoli grafemi, delle singole sillabe, e, infine, delle parole ad uso degli stenografi del sistema Stolze–Schrey.

se nell'intervallo tra le due guerre i linguisti avevano iniziato ad occuparsi dei problemi metodologici connessi allo studio delle frequenze dei fatti linguistici, anticipando in molti casi problemi ancora oggi irrisolti. Esamianimo brevemente a titolo di esempio i contributi portati da N. Trubeckoj e da G. K. Zipf.

N. Trubeckoj (1939) dedicò un intero capitolo dei suoi *Grundzüge der Phonologie* alla statistica fonematica, la quale aveva attirato l'attenzione di altri studiosi della Scuola di Praga, in primo luogo V. Mathesius (1929, 1931) e B. Trnka (1935). Trubeckoj ha posto chiaramente alcuni problemi metodologici.

(1) Egli ha affermato che prima di ogni ricerca statistica sulle frequenze degli elementi fonologici di una lingua, è necessario costruire un modello che consenta il « conto preciso delle possibilità teoriche (tenendo conto di tutte le regole di neutralizzazione e di combinazione (p. 296), perché « le cifre assolute della effettiva frequenza dei fonemi hanno una importanza soltanto secondaria » (p. 296). Bisogna invece metterle in rapporto « col numero della frequenza che dovremmo avere in teoria in base alle regole di combinazione » (p. 288). In alcuni casi la costruzione del modello è semplice e il calcolo delle possibilità teoriche è facile. Trubeckoj fornisce un esempio di calcolo per le parole di due fonemi in tedesco (34). In altri casi il calcolo è più complicato, ma cionondimeno « non ci si deve lasciare spaventare dalle difficoltà tecniche di un tale calcolo. Infatti, solo se confrontate colle cifre di frequenza possibili, ricavate in base a questi calcoli, le cifre di frequenza effettive prendono valore, in quanto esse indicano se un fonema nella lingua considerata è usato molto o poco ». (p. 298). Bisogna riconoscere che questo suggerimento di Trubeckoj è stato seguito assai raramente, e non solo nelle statistiche a livello fonologico, ma anche agli altri livelli. La ragione va ricercata soprattutto nelle « difficoltà tecniche » da lui previste. Queste difficoltà sono talora di natura prevalentemente operativa: a livello fonologico la costruzione del modello completo non presenta difficoltà teoriche, ma il calcolo delle possibilità, praticamente ineseguibile a mano, è un tipico lavoro da eseguire mediante un programma di calcolatore che incorpori in forma di diagramma a stati finiti (35) le regole fonotattiche e morfofonemiche di una lingua. Ad altri livelli, per esempio a livello sintattico, le difficoltà sono in primo luogo di natura teorica,

(34) « In tedesco possono stare all'inizio di parola 18 consonanti (....) e in fine 14 (....), mentre tutti i fonemi vocalici accentati (cioé 10, se non si distinguono *ä* ed *e*) possono stare tanto in posizione iniziale che finale; non sono ammessi i gruppi *i* + *i*, *au* + *r*, *au* + *y*, *eü* + *r*, *eü* + *y*, *ai* + *r*, *ai* + *y*, *o* + *y*, *ö* + *y*. Quindi in tedesco in teoria sono possibili (18 × 10) — 1 = 179 parole del tipo « consonante + vocale » e (14 × 10) — 8 = 132 parole del tipo « vocale + consonante », senza considerare le differenze di taglio sillabico). In realtà il tipo « consonante + vocale » è rappresentato in tedesco da 57 parole (....) e il tipo « vocale + consonante » da 37 parole (....). (p. 299).

(35) Le ragioni per le quali preferisco un sistema a stati finiti (eventualmente aumentato in un ATN) anziché CF e CS, sono non di natura teorica, ma programmativa.

ma è certamente possibile operare con modelli parziali corrispondenti a sotto-insiemi limitati di una lingua.

(2) Trubeckoj ha affermato chiaramente in base a rigorosi argomenti teorici, la complementarità della statistica sui testi e sul lessico. « Per il primo scopo bisogna esaminare statisticamente dei testi continuati, per il secondo dei vocabolari » (288). Per quanto riguarda la fonologia, la statistica « da una parte deve mostrare con quale frequenza un elemento fonologico della lingua studiata (fonema, nesso fonematico, tipi di parola o morfema) appare nel discorso, dall'altra quale è il carico o aggravio funzionale di questo elemento o di una data opposizione fonologica » (288). Anche a questo proposito bisogna riconoscere che, mentre non mancano buoni lavori di statistica sui testi a livello fonematico, lessicale, e, entro certi limiti, anche sintattico, mancano quasi completamente le statistiche sul vocabolario e, in termini generali, sugli elementi che costituiscono un sistema linguistico a vari livelli. Anche se non mancano difficoltà teoriche, le difficoltà che fondamentalmente hanno impedito la realizzazione del programma tracciato da Trubeckoj sono operative. In primo luogo mentre i testi sono, per così dire ' oggetti dati ', gli inventari di unità o funzioni del sistema devono essere costruiti dal linguista e il lavoro spesso è lungo e faticoso. Si pensi per esempio all'inventario di tutte le forme fonetiche di una lingua, necessario per il calcolo delle coppie minime, a sua volta indispensabile per « fissare in cifre per ciascuna lingua anche il grado di utilizzazione distintiva (' carico funzionale ') delle singole opposizioni fonologiche e il carico medio generale dei fonemi » (200). Una volta poi che questi inventari siano compilati, le loro stesse dimensioni ne impediscono lo spoglio manuale. Mentre è possibile limitare le dimensioni del corpus accettando un rischio di errore che può essere stimato con le opportune tecniche statistiche relative alla campionatura [36], non ha praticamente senso, se non in casi particolari, il campionamento di un inventario, per esempio di un vocabolario. Abbiamo già visto in 1.2.3.2 come un lessico automatico possa essere lo strumento adatto per risolvere i problemi operativi di queste ricerche [37].

(36) Cfr. Muller (1973), cap. 13.

(37) Si veda Zampolli (1973b), pp. 168–171, in nota, per una descrizione del lavoro in corso al CNUCE al fine di produrre una statistica della frequenza delle coppie minime nel lessico italiano. Muljacic (1969, p. 307) afferma; « Per ora è praticamente impossibile trovare tutte le coppie minime in cui possono ricorrere tutti i fonemi di una lingua, soprattutto i fonemi frequenti. Ciò sarà possibile quando i dizionari conterranno tutte le forme di tutti i paradigmi delle parti del discorso variabili in trascrizione fonologica ». Anche ammettendo che un tale dizionario esista, la scoperta delle coppie minime è praticamente impossibile senza un calcolatore. P. Torrigiani ha sviluppato al CNUCE un algoritmo che, dato un elenco di forme, scopre elenca e conteggia tutte le coppie minime. L'algoritmo raggruppa innanzitutto le forme (nel caso specifico le circa 1.200.000 forme del LAI, trascritte fonematicamente) per lunghezza. Sia n una lunghezza (ove $1 \leq n \leq 30$ per l'italiano) e m il numero di forme di lunghezza n. Le operazioni (da un punto di vista logico, non operativo) possono essere descritte così. Si deve considerare una prima serie di m forme di lunghezza $n-1$, ottenute a partire dalle forme originarie cancellando il primo fonema. Dopo averle disposte in ordine alfabetico, basterà

È augurabile che lo sviluppo della CL permetta finalmente di svolgere il programma così opportunamente delineato da Trubeckoj. Il LAI che stiamo mettendo a punto con il contributo del CNR offre la possibilità, per la prima volta, di attuarlo.

(3) Per quanto riguarda le frequenze delle unità linguistiche, neanche Trubeckoj discute esplicitamente il loro stato teorico. Egli pone chiaramente il problema di distinguere « quali fenomeni fonologici sono determinati dallo stile, e quali invece ne sono indipendenti » (291). Secondo i suoi dati, almeno in tedesco, « la lunghezza delle unità significative (parole o morfemi) e la frequenza dei nessi consonantici sono condizionate dallo stile. Invece la fraquenza dei singoli fonemi sembra abbastanza indipendente dallo stile del testo (*ivi*) ». Tuttavia egli non cerca di spiegare la stabilità delle frequenze e le loro differenze: « Finora le lingue esaminate statisticamente riguardo alle frequenze dei fonemi sono troppo poche. Perciò dovrebbero essere ancora premature le interpretazioni dei dati statistici e i giudizi d'indole generale su questo problema (292).

Diverso è l'atteggiamento di G. K. Zipf (1935, 1949). Il suo nome è legato soprattutto alla cosiddetta 'legge di Zipf'. Ma, anche senza considerare il tabulare le sequenze uguali per trovare tutte le coppie di forme che differiscono solo per il primo fonema. Cancellando invece il secondo fonema delle forme originarie, ordinando alfabeticamente le *m* sequenze (di lunghezza sempre *n*—1) che ne risultano, e tabulando le sequenze uguali, si trovano tutte le coppie che differiscono solo per il secondo fonema. Si procede così fino ad elaborare le *m* sequenze di lunghezza *n*—1 nelle quali è stato cancellato l'ultimo fonema. (Naturalmente non si considerano coppie minime gli omofoni, e cioè le sequenze di lunghezza *n*—1 uguali se anche le *m* sequenze originali corrispondenti sono uguali). Questo processo è ulteriormente complicato dalla necessità di suddividere il totale di coppie minime scoperte per ogni opposizione in totali parziali di diversa natura: posizione nella parola, lunghezza della parola, categorie grammaticali delle parole della coppia, loro 'registro' e frequenza di uso, ecc. Ma l'operazione più complessa sta nel tenere distinte le opposizioni che avvengono tra i diversi tipi di morfemi. Consideriamo per esempio due verbi regolari come *arare* e *amare*. È chiaro che ciascuna delle loro flessioni produrrà una coppia minima per l'opposizione /*r*/—/*m*/. Ho deciso pertanto di contare tutte le coppie di questo tipo a parte. I tipi tenuti distinti e i relativi criteri formali di distinzione delle coppie sono per ora:

(*a*) flessioni di uno stesso lemma: *aro*: *ari*;

(*b*) flessioni corrispondenti di lemmi appartenenti alla stessa parte del discorso; in questo caso si operano due conteggi: nel primo si contano tutte le coppie di questo tipo, nel secondo si contano solo le serie di coppie (per esempio la serie del tipo *amo*: *aro*, *ari*: *ami*, *ara*: *ama*, ecc., incrementa solo di 1 questo secondo totale);

(*c*) flessioni di lemmi che, pur appartenendo a categorie grammaticali o comunque a paradigmi diversi, costituiscono una serie che si ripete regolarmente: per es. *corto*, *corti*, *corta* e *conto*, *conti*, *conta* (vb.). Il trattamento é analogo al caso (*b*);

(*d*) forme di categoria grammaticale diversa e tema uguale, nelle quali l'opposizione è dovuta alla desinenza: conte (sost.) – conto (vb.).

Nel suo insieme, il lavoro del conteggio delle coppie minime, che richiede un tempo di calcolatore molto alto, eseguito a mano, per esempio sul milione di forme del LAI, richiederebbe qualcosa come 500 anni–uomo.

fatto che studi successivi hanno svuotato questa legge di buona parte del suo significato [38] l'apporto principale di Zipf allo sviluppo della statistica linguistica va ricercato a mio avviso nel fatto che egli ha tentato di spiegare le regolarità statistiche emergenti dai testi nel quadro di una teoria dell'atto di comunicazione. Abbiamo osservato (Zampolli, 1974 e) che anche oggi possediamo a questo proposito solo qualche tentativo appena abozzato, e lo sviluppo degli studi statistici ha mostrato che Trubeckoj aveva ragione nell'affermare che i dati statistici disponibili nel periodo tra le due guerre mondiali non erano sufficienti per interpretazioni di carattere generale come quelle tentate da Zipf.

Tuttavia resta il fatto che Zipf ha posto e tentato di risolvere problemi che oggi, mutatis mutandis, sono stati risollevati da più parti, e in primo luogo da quei trasformazionalisti che tentano di ricuperare la statistica linguistica nel quadro dello studio della performance, dopo la condanna o almeno il disinteresse mostrato nei primi tempi dalla scuola generativo–trasformazionale [39]. Mutatis mutandis e fatte le debite proporzioni, si potrebbe provare

(38) Come è noto, la legge di Zipf è stata spiegata da alcuni come una conseguenza necessaria della legge delle probabilità. Miller (1965, p. VII) la spiega paradossalmente così. Supponiamo di acquistare una dozzina di scimmie e di incatenarle a una macchina da scrivere fintanto che abbiano prodotto una sequenza sufficientemente lunga di caratteri. Supponiamo poi di definire una ' parola ' come una sequenza ininterrotta di lettere tra due spazi. Se contiamo le occorrenze di queste ' parole ', e le ordiniamo per frequenza decrescente, troviamo che le loro frequenze rispondono perfettamente alla legge di Zipf. Il punto cruciale sta nel fatto che si assume che gli spazi siano distribuiti a caso lungo il testo, e perciò ci saranno necessariamente più occorrenze di ' parole ' corte che non di ' parole ' lunghe. Si aggiunga poi l'osservazione che la varietà di parole diverse disponibili cresce esponenzialmente con la loro lunghezza, e il fenomeno descritto da Zipf diventa inevitabile: poche parole corte saranno usate un gran numero di volte, mentre un gran numero di parole lunghe non sarà usato se non di rado o mai. La spiegazione matematica di questo fatto è stata operata da B. Mandelbrot (1954).

(39) Nel quadro della scuola generativo–trasformazionale, l'aspetto quantitativo e statistico è stato considerato da almeno tre punti di vista.

(*a*) Come componente di processi markoviani a stati finiti. « Per completare questo modello (una macchina a stati finiti) assegneremo una certa probabilità da ogni passaggio da uno stato all'altro. Possiamo allora calcolare il grado di ' incertezza ' associato ad ogni stato e definire il ' contenuto di informazione ' della lingua come l'incertezza media, ponderata dalla probabilità di trovarsi negli stati associati. Dato, però, che stiamo studiando la struttura grammaticale del linguaggio e non quella statistica, non ci occuperemo di questa generalizzazione ». Chomsky, 1957, p. 25 della trad. it.

Ovviamente, poiché questi processi vengono considerati inadeguati per costruire una grammatica di una lingua naturale, cade ogni interesse anche per il loro complemento statistico. Questo fatto è esplicito in Halle (1958; p. 22). Nel recensire Herdan (1956), egli identifica, a mio avviso correttamente, la proporzione ' langue–parole: popolazione–campione ' di Herdan con lo ' schema d'urna ', proposto per esempio da Yule (1944) (« A careful reading reveals that this (la concezione di Herdan) differs from Yule's ' treasure chest of tickets ' only in its abstruse language » p. 24), ma fa rientrare lo ' schema d'urna ' nel novero dei modelli a stati finiti che Chomsky dichiara, com'egli ricorda, inadeguati.

Questa inclusione non sembra però del tutto corretta. Halle sostiene che un modello probabilistico è ' fundamentally a negation of syntax, for it does not distinguish a normal English book from one in which the identical set of words occurs in inverse order, or one in

a tradurre con termini attuali alcune esigenze metodologiche espresse da Zipf riassumendole nell'affermazione che il comportamento statistico delle unità linguistiche, e cioè il modo e la misura del loro occorrere nei testi, può essere

which the same words are scrambled up in any one of a multitude of conceivable ways ' (p. 22). Si potrebbe obiettare che un modello d'urna *non* si propone di generare la struttura sintattica di un testo, come del resto risponde ad Halle lo stesso Herdan (1964 b, p. 62).

(*b*) Come criterio di grammaticalità. « La nozione di frase grammaticale non può essere in nessun modo identificata con la nozione di ' alto grado di approssimazione statistica allo inglese ' » (Chomsky 1957, p. 16 della trad. it.) « Une phrase grammaticale n'est pas la même chose qu'un énoncé très fréquent dans le corpus, dans le discours (...) En tout cas, on ne voit pas comment on pourrait établir de relation significative entre la plus ou moins grande probabilité d'occurrence d'une phrases dans le corpus et sa grammaticalité » (Ruwet, 1968, p. 39).

Le ragioni cruciali di questo rifiuto della frequenza come criterio di grammaticalità sembrano essere due: (1) La frequenza si misura in un corpus: ora la presenza in un corpus non garantisce la grammaticalità, perché un corpus può contenere degli ' errori ' di vario genere, nè la assenza la esclude, perché il corpus é finito mentre l'insieme delle frasi di un linguaggio è infinito. In modo generale, ogni specie di frasi ben formate occorrono assai raramente in un corpus, per ragioni che non hanno nulla a che vedere con la grammaticalità: sono troppo lunghe per essere comprese facilmente, sono false, sono banali, ecc. (2) A rigore, non ci sarebbe alcun senso a parlare di probabilità maggiori o minori di una frase. Se si spoglia un corpus abbastanza ampio, la frequenza di ciascuna delle frasi che vi figurano è estremamente bassa, e più crescono le dimensioni del corpus, più questa frequenza tende verso zero. Inoltre, se si ammette che il numero delle frasi di una lingua è infinito, il problema della frequenza relativa delle frasi individuali cadrebbe: la frequenza sarebbe zero. (Ruwet 1968, p. 40, Chomsky 1965, p. 228 n. 5, 1966, p. 20). Questo discorso sarebbe valido anche per la nozione di ' probabilità di un tipo di frase ' (Chomsky 1975, p. 228 n. 5, e 1966, p. 20), perché anche i tipi di frase, intesi come ' sequenze definite di classi di parole o di morfemi ', sono in numero eccezionalmente elevato. Tuttavia, non si può negare che sia possibile stabilire le frequenze relative, in un corpus, di tipi di frase specifici, generati dalla grammatica, e di cercare una spiegazione alle variazioni di frequenza constatate. In questo caso, la ricerca – secondo Ruwet 1968, p. 374 – non concerne la grammaticalità, ma la accettabilità delle frasi, e quindi l'uso che i parlanti concretamente fanno dei meccanismi del linguaggio. (Tuttavia Chomsky non sembra accettare questo punto di vista – cfr. 1965, p. 228, n. 5).

(*c*) Come è noto, lo studio della *performance* è appena agli inizi nella scuola generativo trasformazionale. Tuttavia in esso sembra si cerchi di ricuperare l'aspetto quantitativo e statistico del linguaggio. Chomsky, che pure si è interessato con molta acutezza dello stato attuale della statistica linguistica (e proprio a proposito di modelli di esecuzione; cfr. Chomsky e Miller, 1963, pp. 291 e segg., e pp. 334–344 dell'ed. it.), ha concluso che « Although statistical properties of language and explanatory models for observed uniformities are certainly worth studying, it seems that such investigations have not yet reached the point where they make a significant contribution to the understanding of linguistic processes » (Chomsky 1956, p. 105, e vedi anche Chomsky e Miller 1963, p. 343 della trad. it.). Egli tuttavia dichiara esplicitamente più volte che questa situazione non è una conseguenza necessaria della natura della statistica linguistica, e non esclude possibili contributi futuri (Chomsky 1957, p. 19 della trad. it., (d); 1963, p. 243–344 della trad. it.). Egli critica anche l'opinione di alcuni che « Le frasi più accettabili sono quelle che hanno maggiore probabilità di essere prodotte,.. » (Chomsky 1965, p. 51 della trad. it.). Ma negli ultimi tempi non pochi trasformazionalisti hanno ripreso questa opinione. Ricerche psicolinguistiche cercano di determinare la complessità relativa di certe strutture, o le cosiddette regole del *riordinamento stilistico*, alcune delle quali « non cambiano il contenuto in modo significativo e hanno soprattutto l'effetto di convertire

interpretato e spiegato solo sviluppando lo studio della performance nel quadro globale di una teoria della comunicazione linguistica.

Lo scopo primario esplicitamente dichiarato da Zipf è quello di osservare, misurare, e per quanto possibile definire, formulando ' leggi tentative ', « the underling forces which impel and direct linguistic expression » (1935, p. 3).

L'esame dei dati quantitativi forniti dagli spogli cui aveva accesso ha convinto Zipf della esistenza di straordinarie regolarità statistiche presenti nei testi a tutti i livelli, fonologico, morfologico, lessicale, sintattico. Queste regolarità possono essere espresse e misurate da formule matematiche. Di contro a queste regolarità nelle frequenze d'uso delle diverse unità linguistiche, sta il fatto che quando parliamo « we select words according to their meanings, and according to the ideas and feelings wich we wish to convey; both content and direction of our speech are dictated almost solely by exigencies of meaning and emotion », e certamente non dallo scopo « of preserving or restoring any imaginable condition of equilibrium in the resultant frequency-distribution of the elements of our speech » (1935, p. XIII).

Dobbiamo dunque concludere che le regolarità statistiche che osserviamo nei testi sono una conseguenza dei meccanismi generali che presiedono alla produzione degli enunciati e degli elementi che entrano in gioco in questo processo e devono essere considerate come manifestazioni « of the forces of the mind in the process of functioning (*ivi*) ». Allo stesso modo i fisici « investigano le forze intangibili della gravitazione osservando le loro influenze sugli oggetti percepibili ».

È principalmente la funzione di una unità linguistica nel processo globale del comunicare che determina la sua frequenza, e la frequenza può modi-

una sequenza in un'altra equivalente con minore autoinclusione » (Chomsky 1965 p. 23 della trad. it., n. 1).

Il concetto di *complessità* è cruciale in questi studi e in quelli di stilistica letteraria trasformazionale. Questi ultimi sono basati sull'ipotesi che le trasformazioni che un autore usa siano una parte significativa di ciò che il lettore percepisce come stile letterario. Questa ipotesi parte dalla nozione che i parlanti e gli scrittori possono usare vari e diversi modelli (sentences types) per esprimere fondamentalmente lo stesso contenuto.

Un autore, teoricamente, ha un'opzione in ogni situazione linguistica: per esempio può usare o no la trasformazione passiva. In questo senso, la grammatica generativa trasformazionale ha il potere di differenziare stili diversi. Essa formalizzarebbe la nozione della complessità delle frasi prodotte individuando le trasformazioni in esse applicate. In questa direzione sono stati condotti studi statistici per misurare l'uso dei diversi tipi di trasformazione in autori diversi. (cfr. Ohmann 1964, Thorne 1965, Hayes 1969). I limiti di queste ricerche sono diversi. Lo stile esiste a tutti i livelli linguistici, non solo a quello sintattico. Le grammatiche trasformazionali attualmente esistenti non sono certo così estese da permettere l'analisi di un intero testo. Ma soprattutto manca quasi completamente un modello della perfomance, e « Transformational grammars disavowed any claim that the rules specified in their analyses have any exact representation in whatever mental operations produce sentences in real life ». Per questo, recentissimamente, alcuni hanno proposto la costruzione di una « ' performance grammar ', i.e., a grammar consisting of a set of rules that can be shown to govern the actual verbal behaviour of the users of a given language » (Carroll, 1973, p. 2).

ficare la forma. Occorre quindi – conclude Zipf – studiare statisticamente le relazioni tra la funzione e la frequenza, e tra la frequenza e la forma [40].

Il problema cruciale diventa allora la spiegazione della frequenza delle unità linguistiche, e in particolare, nel pensiero di Zipf, delle parole. Fondamentalmente, egli concepisce un concetto come « a bundle of semantic features » (Miller, 1965, p. IX) che egli chiama ' genes of meaning '. La frequenza di una parola [41] riflette la frequenza, nella vita dei parlanti, del concetto che essa esprime o di alcuni dei suoi tratti. Commenta G. A. Miller (*ivi*): « It is even possible that statistical analysis may yet prove relevant if we can learn how to look behind those surface symptoms called ' words ' to the deeper cognitive processes that underlie them ».

Non vale la pena esaminare le forze psico–biologiche ipotizzate da Zipf. Ciò che importa qui è sottolineare la sua concezione della frequenza di uso come risultato di un processo dinamico (per paragonarla in seguito alla concezione più « statica » di G. Herdan), nel quale intervengono a un tempo fattori extralinguistici e funzioni e caratteristiche del sistema linguistico.

2.2. La lingua come universo statistico.

Come abbiamo detto, è nel secondo dopoguerra che si verificano i primi tentativi di dare una formulazione esplicita ai presupposti teorici sottostanti le operazioni con le quali, come abbiamo visto, si estrapolavano, dalle frequenze di occorrenza delle unità linguistiche nel corpus spogliato, le frequenze caratteristiche di queste unità, le quali autorizzavano l'attesa che in altri testi esse sarebbero riapparse con la stessa frequenza.

Si delinearono ben presto due tendenze principali. Secondo una prima concezione, le unità del sistema linguistico sarebbero caratterizzate, oltre che dai tratti qualitativi emergenti dalle opposizioni e dalle relazioni che formano la struttura del sistema stesso, anche dalle loro rispettive probabilità

(40) Zipf spiega il fatto che in lingue, anche storicamente indipendenti tra loro, le aspirate e le occlusive sorde sono più frequenti delle corrispondenti sonore, come conseguenza della tendenza dei sistemi fonologici a stabilire e a mantenere un equilibrio tra la frequenza di uso di un fonema e la sua complessità valutata come numero di tratti articolatori che lo costituiscono. (Secondo Trubeckoj, che si riferisce anche alla recensione del libro di Zipf di Cohen (1935), nella sua formulazione fonologica questa teoria suonerebbe così: « fra i due termini di una opposizione privativa il termine non–marcato è più frequente, nel discorso continuato, di quello marcato ». Si veda appunto Trubeckoj (1939, cap. 7, n. 3) per una valutazione della teoria, e Guiraud (1960, cap. IX e X) per una sua applicazione allo studio dei cambiamenti fonetici). Un esempio di relazione inversa si avrebbe, secondo Zipf, tra frequenza e lunghezza delle parole. In genere, salvo eccezioni, le parole più corte sono le più frequenti, e quando una parola lunga diventa di uso troppo frequente scatterebbero dei meccanismi equilibratori che tendono ad accorciarla.

(41) Zipf (1935, cap. V, parte seconda) distingue tra *articulatory words* e *conceptual words*. Il discorso è diverso per le due categorie.

d'uso. Queste probabilità non sono direttamente osservabili, come non lo è, del resto, il sistema; esse però si tradurrebbero nel fatto – *dato per certo* – che le unità linguistiche ricorrerebbero nei testi, parlati e scritti, con frequenze relativamente stabili. In questa prospettiva le frequenze osservabili nei testi vengono assunte come *approssimazioni* delle probabilità non osservabili del sistema. Questa concezione traspare per esempio in un articolo in « Word » del 1949, nel quale D. W. Reed anticipa chiaramente il concetto di frequenza come carattere costitutivo di una forma linguistica e insiste sul rapporto tra la frequenza in un corpus e la *probabilità* in lingua (Reed, 1949, pp. 235–236). Ma essa ha ricevuto una formulazione teorica esplicita da P. Guiraud (1954), il quale ha formulato le basi teoriche della possibilità di analizzare la « langue » e la « parole » con i concetti e le tecniche della teoria dell'informazione, equiparandole a « codice » e « messaggio ». Egli ha affermato esplicitamente che la « langue » può essere intesa come un ordine sistematico di engrammi, ciascuno corredato da una specifica probabilità, la quale costituisce un attributo oggettivo come la forma o il significato (Guiraud 1959, p. 16 e p. 23). Con ciò stesso si pensava di spiegare, sul piano teorico, il fatto che, come aveva già affermato Zipf, il variare delle frequenze d'uso di una unità per motivi extralinguistici può causare un mutamento nella sua forma e nella sua funzione. (Guiraud, 1959, cap. X).

Ma la sua formulazione più chiara e più decisa si deve a G. Herdan. Egli propone l'equiparazione tra la dicotomia saussuriana « langue–parole » da una parte e il rapporto « popolazione statistica–campione » dall'altra. La « langue » equivale alla « totality of engrams in the brains of members of the speech community together with their probabilities of occurrence » e la « parole » « for random samples from it » (Herdan, 1960, p. 34). « Language as chance refers to the langue–parole dichotomy in its interpretation as that between statistical universe and sample, whereas language as choice refers to the significant–signifié dichotomy in its interpretation as being subject to the law of duality » (Herdan, 1966, p. 7), e più precisamente « what 'la langue' comprises are not only engrams as lexical forms, but these engrams plus their respective probabilities of occurrence... This seemingly small modification of de Saussure's conception has the important consequence that 'la langue' now has the essential characteristic of a statistical population... 'La parole' then appears as it ought to according to its meaning, as a term for statistical samples withdrawn from 'la langue' as the population » (Herdan, 1966, p. 28). Questa formulazione, che prometteva di fornire delle utili aperture per lo studio delle « modalités de la transition de la langue à la parole et de celle de la parole à la langue » (Heger, 1969, p. 54), e cioè del processo di attualizzazione della virtualità del sistema negli atti di comunicazione sul piano della « parole », è stata accolta con favore in un primo tempo soprattutto dai linguisti europei. L. Heilmann riassumeva con la consueta chiarezza e efficacia così: « Se, dunque, oggi vogliamo tener presenti tutti i fattori che si riconoscono nel funzionamento della lingua, noi dobbiamo considerare quest'ultima come un sistema qualitativo e quantitativo di segni arbitrari doppiamente articolati

(sul piano del significante e su quello del significato), caratterizzati da forma, contenuto e frequenza, attuantisi in un processo individuale fisiopsichico sottoposto al determinismo statistico ». (Heilmann, 1963, p. 63). Il modello di Herdan inseriva gli studi sugli aspetti quantitativi del linguaggio direttamente al centro della teoria linguistica strutturalistica di tipo europeo, e anche recentemente molti cultori della statistica linguistica riconoscono nel modello di Herdan il fondamento teorico della statistica linguistica come scienza (42) e non si può negare che esso presenti un fascino ' teorico ' del tutto particolare, soprattutto a ragione della sua essenzialità e semplicità.

Queste premesse non hanno però trovato, almeno fino ad oggi, un consenso generale.

Il punto cruciale e determinante non sta, come alcuni hanno invece affermato, nel presunto carattere « statico e tassonomico » della ' langue ' saussuriana opposta all'« aspetto creatore » della competenza chomschiana (43), ma nel fatto che gli spogli di testi moltiplicatisi grazie alla diffusione dei calcolatori hanno mostrato e continuano a mostrare che la frequenza delle unità linguistiche, salvo casi eccezionali, non è stabile. Già gli autori del « *Français fondamental* introducendo la nozione di disponibilità scossero per primi questa concezione e R. Moreau, al Convegno di Strasburgo del 1964 sul tema ' Statistique et analyse linguistique ', delineava esplicitamente questa situazione affermando: « les premiers pas de la statistique appliquée à la linguistique ont précisément consisté à admettre des régles du jeu qui soient simples. C'est ainsi qu'on a énoncé (voir par exemple Guiraud) que la fréquence des mots était constante dans la langue, ce qui supposait donc que l'on pouvait assimiler le choix d'un mot au tirage d'une boule dans une urne dont la composition reste inchangée au cours du temps » (Moreau, 1966, p. 130).

E. Heger ritiene indispensabile modificare il modello di Herdan, introducendo tra la langue e la parole un terzo livello, la parole, nel quale solo sarebbe possibile la quantificazione (44). Contro la soluzione di Herdan « à première vue tout à fait convaincante, il faut faire remarquer d'abord que les probabilités d'occurrence qu'elle postule et qu'elle attribue aux unités de langue, n'existent pas indépendamment de facteurs thématiques et stylisti-

(42) Ca si alte ramuri ale lingvisticii moderne, linguistica statistică – în calitate de disciplină consistuită – se reclamă ca avînd bazele teoretice în conceptia despre limbă a lui F. de Saussure « (Roceric–Alexandrescu, 1968, p. 18).

(43) L'argomento è stato discusso molte volte e sarebbe inoltre troppo lungo riprenderlo qui. Si confrontino per es. Ruwet (1968, pp. 18, 50, 51, 390–391) e Heger (1969, pp. 51 e segg.).

(44) Per distinguere le unità negli studi di statistica linguistica in generale Heger (1969) propone i termini *linguème*, *type* e *occurence*. I *linguèmes* apparterebbero alla *langue*, le *occorrenze* alla *parole*, e i *tipi* alla Σ-*parole*. La Σ–*parole* é, secondo Heger, un piano, intermedio tra la *langue* e la *parole*, le cui unità sono quantificabili, mentre non lo sono le unità della *langue*, virtuali, né quelle della *parole* che sono individuali e « liées à leur hic et nunc respectif » (p. 57) Heger considera la *langue* come un sistema di unità in numero limitato e fissato. Ad essi corrispondono, nello stesso numero, *i tipi* della Σ–*parole*, mentre nella *parole* il numero delle occorrenze è illimitato.

ques, c'est–à–dire de facteurs étrangers au système qu'est la langue ». Heger (1969, pp. 56–57) M.me Hirscheberg et Ch. Muller (45) hanno dimostrato che la stabilità della frequenza lessicale è un mito: le frequenze delle parole, salvo casi eccezionali, non sono stabili, ma variano, in dipendenza dallo stile e dal

(45) Sono interessanti a questo proposito le considerazioni di Ch. Muller sui rapporti tra lessico (lexique) e vocabolario (vocabulaire): egli riserva il primo di questi termini alla « langue », il secondo al « discours ». In questa terminologia, le unità che compongono il lessico sono « lessemi » (lexémes); quando queste unità virtuali sono attualizzate nel discorso, ciascuna di esse è un vocabolo (vocable), cui corrisponde un certo numero di occorrenze nel testo. Si definisce dunque come vocabolario di un testo l'insieme di vocaboli presenti, ciascuno con una propria frequenza, nel testo in esame. Il vocabolario di un testo presuppone l'esistenza di un lessico, che trascende il testo, del quale il vocabolario non è che un campione. Si può concepire una serie di insiemi lessicali, ciascuno dei quali contiene il successivo: il lessico di un idioma, nel senso estensivo, entro certi limiti cronologici; il lessico di una lingua, considerato sincronicamente; il lessico di un gruppo umano limitato appartenente alla comunità linguistica, il lessico di un individuo di questo gruppo; il lessico dell'individuo in una situazione definita stilisticamente e tematicamente.

La nozione di lessico di situazione (lexique de situation) ingloba due tipi di elementi: i primi sono d'ordine stilistico legati all'interlocutore e all'effetto che il parlante vuole produrre; i secondi sono di ordine tematico, legati a ciò che il parlante vuol comunicare, al contenuto del messaggio. La situazione provoca, all'interno del lessico individuale, « un déplacement des probabilités d'emploi »; certe unità sono escluse dal discorso, e cadono a probabilità zero. Si può ammettere cioè che, al momento della redazione di un testo o della concezione di un enunciato, l'autore abbia un certo numero di unità del suo lessico « in gioco », mentre le altre sono escluse per motivi stilistici e tematici. Tra le unità lessicali di una lingua, o lessemi, e gli insiemi citati si potrebbero stabilire delle relazioni di inclusione o di esclusione. Se queste relazioni potessero essere stabilite con rigore, sarebbe possibile anche la quantificazione del lessico, mentre è possibile quantificare con rigore solo il vocabolario. La relazione di inclusione può essere stabilita di solito attraverso l'inventario del vocabolario; ma la relazione di esclusione può essere affermata solo per ragioni diacroniche, dunque in un piccolo numero di casi (Muller, 1968, pp. 136–140).

Si è potuto sperare, continua Muller, che la frequenza delle unità lessicali fosse stabile presso l'individuo, variabile nel gruppo; perciò si attribuivano alla statistica lessicale straordinarie possibilità di risolvere problemi filologici di attribuzione. Egli però teme che questa stabilità sia eccezionale. Ne deriva che, se gli elementi quantitativi semplici che compongono la struttura del vocabolario di un testo sono facilmente riconoscibili per mezzo dello spoglio (numero dei vocaboli del testo, frequenza di ciascuno, classi di frequenza e rapporto tra loro), non è invece possibile trasportare questa nozione di struttura nell'ambito del lessico, poiché la probabilità di un lessema nel lessico di un individuo può essere conosciuta o stimata solo se la frequenza è stabile (tenuto conto delle variazioni aleatorie di cui conosciamo la natura) in tutti i discorsi di questo individuo indipendentemente dalla situazione che è, invece, determinante sia sul piano stilistico sia su quello tematico. Quanto alla probabilità di un lessema nel lessico di un gruppo umano, si dovrebbe supporre che la sua frequenza sia costante presso tutti gli individui del gruppo, e Muller giudica questa possibilità ancora meno verisimile.

Ciononostante egli persiste ad ammettere una struttura quantitativa nel lessico, che si definisce con delle probabilità associate ai lessemi, ma non le riconosce « quelque réalité que dans un lexique de situation » (Muller, 1968, p. 141). Per il lessico di un individuo, o di un gruppo, bisognerebbe fare una media di tutte le situazioni di discorso nel quale l'individuo o gli individui del gruppo potrebbero trovarsi, ponderando ciascuna con la propria probabilità. Questo compito gli sembra meno utopistico nel caso di un gruppo, nel quale il grande numero dei membri diminuisce la difficoltà di ponderare la situazione.

tema. Anche i dati che stanno emergendo dall'esame del « Lessico Italiano di Frequenza » (Bortolini, Tagliavini e Zampolli, 1971: d'ora in poi lo chiamerò *a* e dal « Frequency Dictionary of Italian Words » (Juilland e Traversa, 1973: d'ora in poi lo chiamerò *b*, confermano le osservazioni e le affermazioni degli autori ora citati. Per comodità di esposizione, ho posposto la illustrazione di questi dati al paragrafo successivo. Essi concordano in parte con una seconda concezione della lingua come universo statistico, che è sempre stata presente come alternativa a quella di Guiraud e di Herdan.

In questa seconda concezione, il concetto della probabilità come caratteristica intrinseca dell'unità nel sistema non è espresso. L'universo statistico è definito piuttosto come l'insieme di tutti i testi (parlati e scritti) prodotti in un certo periodo di tempo. I testi del corpus sono dunque considerati come campioni estratti da un insieme più esteso di testi, che costituisce la popolazione statistica. La relazione tra testi campione e popolazione è complicata innanzitutto dal fatto che « A satisfactory delimitation of the text population (or populations) is a complicated task because for all practical purposes text populations are open set (uncountable) » (Dolezel, 1969, p. 22). Ma anche ammettendo che l'insieme sia definibile esattamente, una gran parte di questo insieme sfugge a ogni misura: per es. le lettere, le conversazioni private, ecc. All'interno di questa popolazione statistica, i testi devono essere raggruppati in classi o strati, a seconda dei fattori che ne condizionano la produzione, quali lo stile e la situazione (in senso lato) e che sembrerebbero essere responsabili delle differenze riscontrate nelle frequenze delle unità linguistiche. Un corpus potrà essere considerato rappresentativo di una classe omogenea di testi, se esso è composto da testi tratti tutti da questa classe. Se si vuole invece un corpus rappresentativo della popolazione nel suo complesso, la sua composizione dovrà riprodurre la stratificazione della popolazione, e cioè contenere non solo tutte le categorie della popolazione, ma anche nella stessa proporzione della popolazione. Sono evidenti le difficoltà connesse alla definizione non solo qualitativa ma anche quantitativa delle varie classi o sottoinsiemi di testi nella popolazione. Basta pensare alle a volte incredibili diversità di composizione del corpus assunto come campione dagli autori di dizionari di frequenza, e alla soggettività e arbitrarietà delle loro scelte nonostante la serietà degli sforzi volti a riprodurre nella composizione del corpus campione la complessità di strati della lingua nel suo insieme (46).

Un'altro problema, strettamente connesso al precedente, è quello di porre in relazione le classi di testi con il variare o con la stabilità delle diverse carat-

(46) Gìi autori dei dizionari di frequenza hanno quasi sempre composto il corpus da spogliare con brani scelti da testi diversi, definiti da una differenza di origine (autore o soggetto parlante) o altro (genere letterario, ambiente sociale, situazione concreta, ecc.). Non solo le dimensioni complessive del campione scelto a rappresentare una determinata lingua sono, come abbiamo visto, notevolmente diverse, ma la scelta stessa dei testi risponde a criteri differenti. Molti sforzi sono volti a riprodurre nella composizione del corpus campione la complessità di strati e di stili della lingua nel suo insieme. Si hanno classificazioni e suddivisioni particolarmente complesse.

teristiche quantitative. R. Moreau, che può essere considerato un esponente autorevole della concezione di cui stiamo parlando, distingue tre classi di parole: parole di *classe ouverte*, parole di *classe fermée*, e parole per le quali non ha senso parlare di frequenza nella popolazione. Queste ultime sarebbero approssimativamente quelle che il CREDIF, nella propria terminologia, chiama *mots disponibles*, e per le quali ha preferito non indicare la frequenza nel *Français fondamental*. Si deve constatare infatti che per molte parole non è possibile delimitare una qualche classe di testi nella quale esse ricorrano con frequenza stabile.

Le parole di *classe fermèe* sarebbero tutte le parole ' atematiche ', quelle cioè che servono ad « esprimerci a proposito delle cose piuttosto che a esprimere le cose stesse ». Si collocherebbero in questa categoria un certo numero di aggettivi e di voci correnti, di parole grammaticali, alcuni nomi assai generali: « termini più o meno comuni a tutti i soggetti e a tutte le situazioni ». Poiché il loro impiego non sembra variare sensibilmente nei diversi centri di interesse, la stima della loro frequenza a partire da testi campione non porrebbe problemi.

Diverso è il caso delle parole di *classe ouverte* o ' tematiche ' che, al contrario, presentano oscillazioni di frequenza da un testo all'altro, e spesso tra brani di uno stesso testo.

Per valutare la loro frequenza, il solo modo operativo conveniente consisterebbe nello stratificare a priori la lingua, nel delimitare dei centri d'interesse, all'interno dei quali le parole siano ' tematiche '.

L'idea di Moreau è appunto che le parole di uso non stabile nella lingua possano, invece, esserlo in testi con un tema, un centro di interesse comune. Egli propone perciò di stratificare una volta per tutte la lingua di una data epoca in un certo numero di centri d'interesse, che potranno essere anche molto numerosi, del tipo, per esempio: matematica, fisica, chimica, ecc. Allo interno di ciascuno strato sarà determinata la frequenza delle parole tematiche [47].

(47) Per esempio, nel centro di interesse « algebra elementare », parole tematiche come ' eguaglianza, diseguaglianza, equazione, incognita, variabile, funzione ' saranno frequenti e stabili. I risultati ottenuti sui testi accessibili alla campionatura potranno essere estesi agli altri testi, alla parte inesplorata del centro d'interesse. Come ha senso classificare per frequenza le parole generalmente stabili nella lingua, così potranno essere messe in ordine di frequenza, all'interno del loro centro d'interesse, le parole tematiche. Ma vi sono notevoli difficoltà relative alla determinazione dei centri d'interesse. In primo luogo ci si chiede se non si debbano cercare per ogni parola, anche di frequenza debole, i centri d'interesse cui si riferisce, generando così alla fine una straordinaria moltiplicazione degli strati. O se non convenga, invece, definire a priori, nella lingua, un numero ragionevole di centri d'interesse e cercare poi, all'interno di ciscuno, solo la frequenza di alcune parole. Per Moreau, la risposta sta nella constatazione che, per alcune parole, è impossibile delimitare un qualsiasi centro d'interesse nel quale esse ricorrano con frequenza stabile, cioè, secondo la sua terminalogia, siano « ergodiche »; sarebbero approssimativamente quelle che, il CREDIF, nella propria terminologia, chiama *mots disponibles*. Definire una frequenza per le parole disponibili non ha senso, dice Moreau: infatti il CREDIF ha preferito non indicarne la frequenza nel francese fondamentale.

Occorre innanzitutto osservare che questo modello lessicale non è del tutto rispondente ai dati dei nostri spogli, o che almeno si dovrebbe precisare entro quali limiti di fluttuazione si considera ' stabile ' la frequenza di una parola. Nei due dizionari di frequenza dell'italiano, già citati, si sono riscontrate variazioni di frequenza statisticamente significative anche all'interno delle parole che Moreau colloca tra quelle di *classe fermèe*. È significativo poi il fatto che le variazioni nella frequenza delle *parole di relazione* tra i sottoinsiemi di cui essi sono composti appaiano strettamente connesse alla natura del sottoinsieme, dal momento che hanno lo stesso andamento per la maggior parte delle unità appartenenti alla medesima categoria grammaticale (48).

Non ci poniamo qui il problema, che meriterebbe una discussione troppo ampia, se abbia o no senso, nella concezione che stiamo esaminando, parlare di frequenza di uso di una unità nella popolazione nel suo complesso anche quando la sua frequenza varia sensibilmente tra le diverse classi di testi.

Piuttosto vale la pena dichiarare esplicitamente le lacune, peraltro già accennate, che impediscono di considerare soddisfacente la concezione ora discussa. Innanzitutto, manca quasi completamente un collegamento esplicito tra la lingua come sistema e la lingua–popolazione statistica intesa come insieme di testi. In secondo luogo, e, probabilmente non indipendentemente da questa assenza, mancano dei criteri per dividere i testi in classi e manca un inventario dei fatti linguistici il cui comportamento statistico può essere descritto in relazione alle classi di testi.

Sembra naturale, a questo punto, di fronte alla constatazione che oltre alle stabilità si devono trattare anche le variazioni tra classi di testi, prendere in considerazione le idee e i dati del filone ' filologico–stilistico ', che fin dagli inizi si è concentrato sulle differenze tra testi più che sulle regolarità generali in una lingua. Questi studi possono fornire i dati di partenza per una primissima stratificazione a priori o, meglio, i dati per identificare, per lo meno come ipotesi di lavoro, alcuni sottoinsiemi sufficientemente definiti ed egualmente caratterizzati rispetto a uno o più parametri comuni. Inoltre essi hanno

(48) Per esempio in *a*:

	Cinema	Teatro	Romanzi	Giornali	Suss.
Il (art.)	5930	7730	10892	13453	16747
Di (prep.)	3021	3434	5132	7044	6770
Da (prep.)	692	729	1070	1249	1400
Fra (prep.)	37	42	64	108	113
In (prep.)	1014	1216	2020	2420	2181
Tra (prep.)	40	48	84	95	122

radunato una serie di dati sulle caratteristiche che possono essere considerate «as a statistic trait in the patterning of style» (Bailey, 1969, p. 221), che possono cioè essere trattate « as ' elementary quanta ' –as ' distinctive features of style – » (Doležel, 1969, p. 21). Tuttavia manca di solito qualsiasi riferimento a una teoria che offra « a description of the orgins and underlying factors responsible for the variety of styles » (Doležel, 1969, p. 11). La mancanza di una simile teoria è in parte responsabile di quelle incertezze e contraddizioni, derivanti soprattutto dalla constatazione « of the unstable statistical nature of style chai acteristics », che hanno caratterizzato verso la metà degli anni '60, in contrasto con l'entusiasmo degli anni '50, lo stato d'animo dei ricercatori che ha trovato una efficace espressione nel « Convegno ' Statistique et analyse linguistique ' dell'aprile 1964 a Strasburgo (49).

Si è ancora, dunque, alla ricerca di un modello quantitativo comune al filone stilistico filologico e al filone della comunicazione linguistica, il quale caratterizzi esplicitamente le variabili linguistiche ed extralinguistiche che influenzano la produzione dei testi e indichi come individuare i fatti linguistici determinati quantitativamente da questi fattori. Un tale modello, come si è detto, non esiste, anche se alcuni ricercatori hanno portato recentemente alcuni contributi relativi alle caratteristiche che dovrebbe avere, alle esigenze cui dovrebbe rispondere, e ai fatti che dovrebbe spiegare.

Riferisco qui lo schema proposto tra gli altri da Doležel (1969) sul versante stilistico e da Rosengren (1971) sul versante linguistico. Esso si pone nello ambito della teoria generativo–trasformazionale e in particolare presuppone una teoria della ' performance '. Esso ' soffre ' dunque dello stato estremamente lacunoso di questa teoria. Non si va al di là di semplici affermazioni di questo tipo: « I am convinced that contemporary linguistics provides a suitable general background for explaining stylistic variety in the distinction of linguistic competence and linguistic performance. Style–forming processes are to be considered a fundamental component of linguistic performance: style originates in the processes of producing (encoding) a text (a discourse) by a certain speaker. Therefore, stylistic theory is a substantial part of the theory of linguistic performance » (Doležel, p. 11). Per il resto, la teoria della ' performance ' viene ' messa tra parentesi ', si assume cioè che nella produzione di un testo esistano dei meccanismi di ' performance ', i quali essenzialmente presiedono alla scelta e alla applicazione delle regole fornite dalla ' competence ', e che questi meccanismi incorporino i processi di formazione dello stile. Le frequenze riscontrate nei testi sono determinate globalmente dai meccanismi di produzione del testo. I meccanismi che operano in modo deterministico sono responsabili delle frequenze stabili nella lingua (ammesso che ci siano: il che equivale a dire che queste frequenze sono determinate dal funzionamento interno della lingua), mentre i meccanismi che implicano una scelta sono responsabili del variare delle frequenze, e quindi devono

(49) Espressioni che sono state poi notevolmente mitigate negli Atti.

essere considerati come componenti degli « style–forming processes ». Ma non si dice null'altro su questi meccanismi, dei quali praticamente ci si limita ad affermare l'esistenza. Nulla viene detto poi sugli elementi che mettono in moto questi meccanismi o, per così dire, sul loro input: il contesto dell'atto di comunicazione; evidentemente una tale specificazione esigerebbe una teoria della comunicazione linguistica.

Possiamo quindi trascurare, date tutte queste indeterminatezze, l'aspetto che collega lo schema di cui stiamo parlando con la teoria generativo–trasformazionale, e considerare il programma di lavoro che esso propone come una illustrazione della natura e della complessità dei problemi da risolvere.

Come già aveva affermato W. Winter (50) in una comunicazione al IX Convegno Internazionale dei linguisti, anche secondo L. Doležel e I. Rosengren il processo centrale nella produzione di un testo per quanto riguarda lo stile è un processo di scelta. Grosso modo il processo selettivo può essere delineato così: a ogni passo della produzione del testo, il parlante sceglie un « modo di espressione » da un insieme A di alternative offerte dalla sua competenza inguistica: A = (*a*, *b*. . .*n*). Nel caso più semplice, un insieme di alternative consiste di soli due membri; per esempio A = (costruzione attiva, costruzione passiva). Nel produrre una proposizione, il parlante decide se usare la forma attiva o quella passiva. Questa scelta è ripetuta ogni qualvolta l'insieme è presentato al parlante, nel nostro esempio alla produzione di ogni proposizione. In altri casi, naturalmente, l'operazione di scelta non avviene ad ogni proposizione; per esempio la scelta tra un insieme di sinomini avviene solo quando si deve esprimere il ' contenuto ' corrispondente a tale serie di sinonimi.

In ogni caso, la ripetizione della operazione di scelta ha come conseguenza la distribuzione probalistica degli elementi dell'insieme A: P(A)=P(a), P(*b*) · · · P(*n*). A ognuna delle alternative corrisponde quindi una certa frequenza di occorrenza nella popolazione di testi prodotti.

(50) ' From a linguist's point of view, references to style as commonly found in literary analyses of texts are often too intuitive and impressionistic to carry much weight '. La valutazione statistica sembra a Winter un mezzo per tentare di scoprire le proprietà distinte di certi gruppi di testi. ' A style may be said to be characterized by a pattern of recurrent selections from the inventory of optional features of a language. Various types of selections can be found: complete esclusion of optional element, obligatory inclusion of a feature optional elsewhere, varying degrees of inclusion of a specific variant without complete elimination of competing features. In these properties, styles agree with dialects (. . .) The identification of different styles can be undertaken in much the same way as the identification of different dialects of a language: a boundary between two styles can be established on the basis of the isoglosses '. (Winter, 1969, p. 3). Anche se la caratterizzazione completa di uno stile presupporrebbe l'individuazione di tutte le proprietà distintive di una intera base di enunciati, non è necessario attendere una descrizione esaustiva; si può iniziare proponendo una identificazione sulla base di alcune « isoglosse » caratteristiche. Idealmente l'identificazione di stili contrastanti dovrebbe essere fatta senza ricorrere alla intuizione o alla tradizione. In pratica, – suggerisce Winter – si comincerà assumendo come ipotesi di lavoro le suddivisioni tradizionali e verificandole.

Si assume che il processo di scelta sia controllato da fattori di natura pragmatica. Ogni parlante compone il testo « in accordance with his individual idea of efficiency, on the one hand, and in accordance with the requirements of the context (in the broadest sense) on the other. Those qualities of the speaker that are relevant for the selection processes (verbal preferences, mental type, stylistic skills, etc.) represent the *subjective factor* of selection – Context Q, can be designated as the *objective–factor* of selection; it is independent of the speaker, though it exercise its influence through him – Context is represented by such factors of verbal communication as the form of language (written or oral), the form of discourse (dialogue or monologue), the genre, the function of the text, and so on » (p. 13).

I fattori pragmatici, soggettivi e oggettivi, sono sempre presenti quando un testo viene prodotto. Tuttavia il modello riconosce che l'impatto di questi fattori nei processi di formazione dello stile varia. Teoricamente esistono tre possibilità fondamentali.

(1), Un parlante (X) quando sceglie tra le possibili alternative, è controllato solo dalle proprie preferenze e abitudini personali; l'influenza del contesto è eliminata. Questo caso sembra piuttosto eccezzionale. Empiricamente, questo tipo potrebbe essere rappresentato « by a poet who imposes his highly personal style on all texts he produces, regardless of differences in genres or forms »,

(2) Il parlante X sopprime completamente le preferenze personali ed è interamente soggetto all'influenza del contesto quando sceglie nell'insieme di alternative. Esempio pratico: i documenti burocratici.

(3) Il parlante X adatta le proprie scelte al contesto e allo stesso tempo conserva alcune caratteristiche individuali che lo distinguono dagli altri parlanti. È questo probabilmente il caso del parlante–comune.

Nel modello, le due classi di fattori (soggettivi e oggettivi) sono considerate come due forze in competizione tra loro: uno stesso parlante può comportarsi di volta in volta nel modo (1), (2) e (3), anche se (3) può essere considerato come il comportamento prevalente. I processi selettivi dovrebbero spiegare l'origine e la varietà degli stili. Non possediamo una teoria della produzione dell'atto di parola che incorpori questi processi e le variabili che operano su di essi in input. Questa constatazione è vera anche per la teoria generativo–trasformazionale della performance. Lo studio delle variabili che caratterizzano le differenze tra i vari stili è ancora quasi tutto da fare. Queste caratteristiche debbono essere individuate e misurate ai vari livelli: fonematico, morfologico, lessicale, sintattico, ecc. Alcune di queste misure sono state già studiate, ma di solito al di fuori di uno schema generale.

Possiamo provare a vedere come dovrebbe essere organizzato questo studio. Innanzitutto si deve stabilire una prima tentativa classificazione dei testi basata sulla esistenza di fattori pragmatici di produzione. I fattori sogget-

tivi e oggettivi devono essere determinati empiricamente, prima di conoscere la struttura statistica del corpus. Dati i testi del corpus TCL, essi saranno classificati secondo due dimensioni: (*a*) secondo i fattori soggettivi: la classificazione produce le classi T (Xi), cioè i testi T della classe T (Xi) sono prodotti tutti da uno stesso parlante Xi; (*b*) secondo i fattori oggettivi; la classificazione produce le classi T (Qi), cioè i testi della classe T (Qi) sono prodotti tutti nel contesto Qi. Le classi T (Xi, Qi) rappresenteranno quindi insiemi di testi prodotti da un certo scrittore in un certo contesto. Ogni classe di testi, così definita, può essere interpretata come una ' popolazione ' ed essere usata per determinare la proprietà statistiche caratteristiche dei testi.

Esaminiamo i diversi casi possibili:

(1) Supponiamo che alcune caratteristiche risultino stabili statisticamente nell'insieme T (L). Ciò significa, naturalmente, che queste caratteristiche non hanno capacità discriminante per lo studio degli stili nell'insieme dei testi T (L). Possiamo dire che queste caratteristiche sono di natura sopra–stilistica. Se è vero ciò che molti ritengono, (si ricordino le affermazioni di Trubecrkoy), la distribuzione dei fonemi e dei grafemi ne sarebbe un esempio.

(2) Supponiamo che i dati mostrino che alcune caratteristiche non sono statisticamente stabili in un testo T o in una classe T (Xi, Qi). Questo problema è stato molto discusso. Possiamo cercare quali fattori occasionali, transienti determinino la fluttuazione. Salvo quanto diremo in 3, potremo ritenere queste caratteristiche come sub–stilistiche. Tuttavia l'esistenza di queste caratteristiche è un punto interessante per la statistica linguistica, che deve creare un modello per ' spiegare ' anche queste caratteristiche.

(3) I valori di una caratteristica sono statisticamente stabili nei testi di un certo parlante, indipendentemente dal ' contesto ', mentre variano da parlante a parlante. Tali caratteristiche possono essere assunte come espressione dello stile individuale, cioè prodotte da fattori soggettivi: (S–C).

(4) Una caratteristica è omogenea nel corpus T (Qi), e cioè varia da un contesto all'altro, ma all'interno del contesto non varia per i singoli parlanti. In questo caso la caratteristica può essere assunta come risultato di fattori oggettivi (*O–C*).

(5) Se una S–C fluttua entro intervalli diversi in contesti diversi, possiamo ritenere che essa sia controllata da entrambi i tipi di fattori: La chiameremo allora soggettiva–oggettiva (SO–C).

Le caratteristiche così specificate « can be treated as ' elementary quanta ' –as ' distinctive features, of style–that are practically unlimited in their combinatorial possibilities » (Doležel, 1969, p. 21).

Questo schema non deve indurre a credere che una caratteristica classificata in un certo modo (*S–C*, *O–C*, *SO–C*, ecc.) in un corpus, debba risultare

appartenente alla stessa classe in qualsiasi altro corpus. La natura statistica specifica di una caratteristica stilistica non è determinata dalla sua natura linguistica, ma dai fattori che la controllano. Poiché è possibile che fattori diversi si alternino al controllo di una data caratteristica, è possibile che « the statistical nature of a linguistically identical characteristic is subject to change in various texts (or text clas) » (*ivi*, p. 22). Questo punto è estremamente interessante ma naturalmente complica molto i procedimenti di scoperta delle caratteristiche, della loro natura, dei fattori che le controllano.

Accanto allo sviluppo di schermi teorici e di modelli provvisori del tipo ora illustrato, che possono servire come base per un piano di lavoro, è necessario accumulare gli spogli e la elaborazione dei dati relativi per estendere, per così dire, sperimentalmente, l'inventario delle caratteristiche e scoprire il loro comportamento in sottoinsiemi diversi di testi. Come si può facilmente vedere, questo modello è più che altro il frutto di un esercizio combinatorio su alcuni elementi la cui esistenza è postulata e non provata. Metodologicamente, si può sempre assumere questo modello, o un altro modello equivalente, come ipotesi di lavoro e poi provare a verificarlo sperimentalmente. È essenziale, a mio avviso, constatare che: i modelli precedenti sono già stati falsificati; non disponiamo di una teoria adeguata o addirittura, forse, non possediamo alcuna teoria dei processi che determinano la scelta tra le possibilità del sistema nell'atto di comunicazione; le attuali proposte di modelli sono poco più che semplici espressioni della esigenza di costruire questa teoria, e di accumulare dati sperimentali sul grado di stabilità e sulla distribuzione delle frequenze del maggior numero di fatti linguistici possibili; nella situazione attuale manca un modello sufficentemente elaborato che dia alla statistica linguistica una base teorica adeguata. Per uscire da questa situazione, occorre tenere conto delle poche certezze finora acquisite, che sembrano indicare l'esistenza di sottoinsiemi relativamente omogenei di testi, e assumere questa come ipotesi di lavoro. Tra le operazioni da compiere alcune sono particolarmente urgenti. Esaminare il comportamento statistico del maggior numero possibile di caratteristiche a tutti i livelli (e non solo quelle studiate tradizionalmente, quali la lunghezza delle parole, la lunghezza delle frasi, le frequenze lessicali; naturalnente queste erano state scelte soprattutto per la semplicità della procedura di spoglio: altre caratteristiche, per esempio i ' patterns ' sintattici e le scelte tra sinonimi implicano definizioni linguistiche e procedure di spoglio molto più sofisticate e laboriose; in questo la CL è chiamata direttamente in causa). Procedere euristicamente, cercando di raggruppare i testi in classi omogenee in base alle somiglianze nella distribuzione di frequenza delle categorie studiate. Confrontare le classi di testi così ottenute a posteriori con le classi di testi determinate a priori, raggruppando i testi secondo la presunta omogeneità dei fattori di produzione, dove ' omogeneità ' significa che uno o più fattori noti sono costanti.

2.3. Alcuni dati.

Come preannunciato al paragrafo precedente, riporterò qui alcuni dati ricavati dallo studio, peraltro appena iniziato, di due dizionari di frequenza dell'italiano scritto pubblicati di recente: il *Lessico di Frequenza della Lingua Italiana Contemporanea* di U. Bortolini, C. Tagliavini e A. Zampolli (1971; d'ora in poi indicato come *a* e il *Frequency Dictionary of Italian Words* di A. Juilland e A. Traversa (1973, d'ora in poi indicato come *b*.

Il mio scopo non è quello di confrontare i due dizionari, anche se, forse per la prima volta, ci troviamo a possedere per una stessa lingua due dizionari il cui confronto, pur entro certi limiti, potrebbe avere qualche significato.

Sta di fatto che i dizionari di frequenza compilati fino ad oggi per la stessa lingua sono di solito molto diversi tra loro. Poiché essi sono compilati con metodi diversi, non è possibile decidere se le differenze siano da attribuire semplicemente ai metodi di compilazione, oppure alle popolazioni statistiche che i dizionari di frequenza si propongono di rappresentare (51).

Le differenze metodologiche di solito riguardano:

a) *I criteri di composizione del corpus* sul cui spoglio è fondato il dizionario. I criteri sono: (1) quantitativi: dimensioni globali del corpus; numero dei sottoinsiemi nei quali il corpus è stratificato e loro porporzioni relative; numero dei testi diversi in ciascun sottoinsieme; quantità e distribuzione di parole o di frasi prese da ciascun testo; (2) qualitativi: scelta e definizione dei sottoinsiemi; scelta dei testi nei sottoinsiemi (52).

b) La definizione e la individuazione delle unità lessicali: in altre parole, poiché di solito le unità di lavoro sono i lemmi, i criteri di lemmatizzazione (53).

(51) Secondo alcuni queste differenze invalidano il concetto stesso di dizionario di frequenza. Il loro discorso, ridotto all'osso, è il seguente: in che senso e in che misura si può ritenere che un dizionario di frequenza sia rappresentativo di una lingua, se ogni nuovo dizionario fornisce parole e frequenze diverse dai precedenti, e che affidamento possono dare tali dizionari per le applicazioni cui sono destinati? Prima di interpretare le differenze tra i dizionari di frequenza, è necessario accertare quali di esse siano dovute alle tecniche di compilazione e quali invece al corpus spogliato. Queste ultime possono essere adoperare per verificare i limiti del ' modello d'urna ' e della interpretazione di Herdan.

(52) La nota 15 di *a* (pag. XV e segg.) e Zampolli (1969, p. 73 e segg.) forniscono alcuni esempi delle « a volte incredibili diversità di composizione e di dimensioni del corpus assunto come campione » e della soggettività che caratterizza la scelta dei testi. Molto poco possiamo aggiungere ai consigli già citati di Trubeckoj « Per prudenza però, è bene (...) ricercare la neutralizzazione dei generi stilistici. La cosa migliore è la notazione di varie conversazioni o prendere dei giornali nei quali siano rappresentati diversi generi stilistici (articoli di fondo politici, telegrammi, articoli semiscientifici, comunicazioni ufficiali, notiziari sportivi, notiziari aziendali, romanzi d'appendice, ecc.) (Trubeckoj, 1949, p. 292). Questa situazione è una prova indiretta della inadeguatezza dello schema d'urna e della mancanza di un modello della lingua come ' popolazione statistica stratificata ' (cfr. Muller, 1968, p. 15) che fornisca i criteri per la rappresentitività di un corpus.

(53) Si confrontino a questo proposito Engwall (1974, capitolo 3), e Lyne (1972).

c) Le formule che determinano la scelta e l'ordinamento delle unità lessicali (d'ora in poi ci riferiamo direttamente a queste come a dei lemmi) da inserire nel dizionario [54].

(54) Rassegne dei diversi metodi sono contenute in Juilland e Chang Rodriguez (1964), Zampolli (1969, pp. 75 e segg.), Bortolini, Tagliavini, Zampolli 1971, XXI e segg.) Engwall (1974, capitolo 5). Nei primi dizionari di frequenza *la frequenza* (F) fu l'unico criterio di scelta e di classificazione delle parole, per es. negli studi di Kaeding (1897–98) e di Henmon (1924). In seguito si è preso in considerazione anche il numero dei testi nei quali una parola è rappresentata, e cioè la *ripartizione* o *rango* (R). L'introduzione della ripartizione è fondata sulla convinzione che sia necessario « compléter la notion de fréquence par celle de stabilité de la frequence, ou, si l'on préfère, de corriger la fréquence par la façon dont celle–ci se distribue dans le corpus » (Muller, 1965, p. 34). Così per esempio, Wander Beke (1929), Kenniston (1933) Josselson (1953), Garcia–Hoz (1953), *il Français Fondamental* del CREDIF (1965) adoperano in varie combinazioni F e R. Juilland (1964) ha introdotto un metodo di calcolo che assegna ad ogni lemma del corpus un indice che egli chiama U (*usage*), il quale combina assieme alla frequenza, anziché la ripartizione, la dispersione (D). L'indice D vuole rendere conto del modo più o meno uniforme nel quale la frequenza complessiva di un lemma è la ripartita tra i 5 sottoinsiemi nei quali Juilland ha diviso il suo corpus. Nella formulazione datane da Muller (1965) il procedimento richiede che si calcoli anzitutto

$$S = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (f_i - \bar{f})^2}$$

dove f_i è la frequenza del lemma nell'iesimo sottoinsieme, n è il numero dei sottoinsiemi (nel nostro caso 5) e $\bar{f}$ è la frequenza media, cioè F/n.

Si definisce coefficiente di variazione il rapporto

$$V = \frac{S}{\bar{f}}$$

che, a motivo del denominatore, è manifestamente indipendente dalla frequenza. Questo coefficiente di variazione è nullo quando

$$f_1 = f_2 = f_3 = f_4 = f_5 = \bar{f}$$

infatti in tal caso è nullo anche S.

Il suo valore è invece massimo quando il lemma appare in uno solo dei sottoinsiemi; in questo caso si ha

$$V = \sqrt{n - 1}.$$

Nel nostro corpus, diviso in 5 sottoinsiemi, il valore massimo di V è 2. Dividendo V, il coefficente di variazione, per $\sqrt{n - 1}$ si ottiene un indice che varia da 0, ripartizione perfettamente uniforme, a 1, quando il lemma appare in un solo sottoinsieme.

Il rapporto

$$\frac{V}{\sqrt{n - 1}}$$

varia dunque tra 0 e 1, crescendo col diminuire dell'uniformità della ripartizione. Introducendo il complemento all'unità, per invertire il senso della variazione, si ottiene la formula (da noi adottata e corrispondente a quella di Juilland e Rodriguez)

$$D = 1 - \frac{V}{\sqrt{n - 1}}$$

Sarebbe evidentemente interessante far variare una sola di queste caratteristiche e studiare le conseguenze sulla composizione del dizionario, In particolare, nel quadro problematico discusso al paragrafo precedente, interesserebbe soprattutto far variare la composizione qualitativa del corpus.

Osserviamo che *a* e *b* hanno alcune importanti caratteristiche metodologiche in comune.

(*a*) I criteri quantitativi di composizione del corpus sono, a livello generale, i medesimi. Sia *a* sia *b* infatti sono basati su un insieme di 500.000 occorrenze ripartite in 5 sottoinsiemi di 100.000 occorrenze ciascuno.

(*b*) I criteri di lemmatizzazione sono stati sensibilmente diversi nei due dizionari, tanto che un confronto diretto tra i rispettivi lemmari sarebbe del tutto privo di senso. Tuttavia è stato possibile compiere una serie di interventi che, eliminando quasi del tutto le conseguenze dei diversi criteri di lemmatizzazione, hanno permesso di ridurre le differenze entro limiti probabilmente accettabili ai fini del confronto [55].

che qualifica la ripartizione della frequenza di ogni lemma del corpus per mezzo di un *coefficente di dispersione D* che varia tra 0 (frequenza accumulata in un solo sottoinsieme) e 1 (ripartizione perfettamente uniforme della frequenza del lemma negli *n* sottoinsiemi).

Per esemplificare il comportamento di questo indice, calcoliamo per i 7 lemmi seguenti, che rappresentano tutte le diverse possibili ripartizioni per F = 5 e *n* = 5. (R rappresenta il numero dei sottoinsiemi nei quali il lemma appare).

Lemma	F	f1	f2	f3	f4	f5	R	D
L 1	5	1	1	1	1	1	5	1,000
L 2	5	2	1	1	1	0	4	0,685
L 3	5	2	2	1	0	0	3	0,577
L 4	5	3	1	1	0	0	3	0,450
L 5	5	3	2	0	0	0	2	0,368
L 6	5	4	1	0	0	0	2	0,225
L 7	5	5	0	0	0	0	0	0,000

L'indice U è calcolato con la formula:

$$U = DF$$

U è uguale ad F quando la parola è ripartita uniformemente nel corpus, perché in tal caso, D = 1. Invece U = 0 quando le occorrenze sono accumulate tutte in un solo sottoinsieme perché in tal caso D = 0. Per lemmi di frequenza complessiva uguale, U è tanto più vicino ad F quanto più la ripartizione è uniforme e decresce proporzionalmente al decrescere di D.

(55) In alcuni casi questi interventi sono stati abbastanza semplici. Per esempio, mentre in *a* tutte le occorrenze dell'articolo determinativo, comprese quelle nelle cosiddette preposizioni articolate, sono state riunite in un unico lemma *il*, che presenta frequenza complessiva

(*c*) Le formule statistiche per la scelta e l'ordinamento dei lemmi sono state le medesime in entrambi i dizionari. A ciascun lemma del corpus è stato assegnato, sia in *a* sia in *b*, il valore U (*uso*, inglese *usage*) con una formula che tiene conto della frequenza globale del lemma nel corpus e della uniformità della sua ripartizione nei 5 sottoinsiemi. Sia in *a* sia in *b* sono stati accolti i (circa) 5.000 lemmi con maggior valore di U (56).

54752 ed occupa il rango 1, in *b* appaiono 3 lemmi distinti, *il*, *la*, *lo*, con frequenze rispettive 24333, 23663 e 6821, e ranghi, rispettivamente, 2, 3, e 9. Delle due soluzioni che permetterebbero un confronto:

(1) suddividere la frequenza di *il* in *a* per creare tre lemmi distinti, *il*, *la*, *lo*,

(2) riunire in *b* la frequenza dei tre lemmi *il*, *la*, *lo*, in un solo lemma *il*,

la prima è impossibile, perché in *a* appare, sotto *il*, la forma *—ll'* che comprende le occorrenze nelle preposizioni articolate sia dall'articolo maschile sia dell'articolo femminile. (Si noti che, con una evidente incoerenza, *l'* in *a* è stata invece distinta in *masch.* e *femm.*). Si è quindi operato l'intervento (2). Così facendo l'articolo determinativo è salito anche in *b* al primo rango, con frequenza 54817. Se calcoliamo lo scarto ridotto tra le due frequenze, 54752 e 54817, troviamo

$$Z = 0,1$$

il quale indicherebbe che la differenza tra le due frequenze ha più di 9 probabilità su 10 di essere ottenuta per puro effetto del caso.

Purtroppo in molti casi la soluzione non è stata così semplice, ed ha richiesto un lungo lavoro da ' detective ' sui lemmari dei due dizionari, perché spesso i criteri di lemmatizzazione dichiarati nelle introduzioni non sono sufficientemente espliciti o addirittura sono stati contraddetti nella applicazione. Del resto questa è una situazione abbastanza comune negli spogli (cfr. Engwall, 1974, cap. 6; Köster, 1970, p. 291; Lyne, 1972 e 1973), cui solo l'uso di un lessico automatico può porre rimedio. Una descrizione dettagliata degli interventi compiuti sarebbe troppo lunga, e rimando pertanto a Zampolli (1974 D). Mi limito ad accennare a una classe di interventi cui mi riferirò nel seguito dell'esposizione. Come è noto, uno dei punti più controversi nella lemmatizzazione, non solo di testi italiani ma anche di molte altre lingue, è costituito dalle « parole che possono assumere un altro valore sintattico per il passaggio di categoria grammaticale » (*a*, LV), e cioè di quelle parole che *b* chiama « syntactic homographs », le quali « are identical occurrences of members of different word classes: the same shapes, often with the same lexical meaning, are distinguished by sentence function » (*b*, XXVI). Si pensi ad esempio all'uso sostantivato di infiniti, participi passati e presenti, all'uso aggettivale di participi passati e presenti, agli aggettivi sostantivati e ai sostantivi in funzione aggettivale, agli aggettivi usati avverbialmente, all'uso preposizionale e avverbiale di certe parole, ecc. Mentre *b* ha in genere distinto accuratamente questi « omografi sintattici », *a* li ha lasciati uniti sotto uno stesso lemma, cui ha assegnato però entrambe le categorie grammaticali. Così, per es., troviamo che in *a* il lemma *su* è accompagnato dalla doppia sigla AV PZ, il lemma *questo* dalla sigla AG PR. I participi, invece, sono stati per lo più lemmatizzati sotto i verbi corrispondenti. In questi casi, molto numerosi, siamo intervenuti riunendo anche in *b* quegli omografi che *a* non aveva distinto. Questo fatto spiega perché nelle tabelle relative alle percentuali delle categorie grammaticali che seguono, (relative sia ad *a* sia a *b*) compaiono le frequenze di coppie di categorie, come *avv.* + *prep.*, *agg.*+ *pron.*, ecc. Questi interventi hanno anche ridotto il numero di lemmi in *b* da 5014 a 4505. Rinvio a Zampolli (1974 D) per dettagli su questo punto e sulle conseguenze nella applicazione delle formule statistiche ai due lemmari, delle quali parlerò tra poco.

(56) Come abbiamo detto, la formula per il calcolo di U proposta da Juilland, usata anche da Allen (1970 e 1971), è stata sottoposta a diverse critiche: Muller (1965), Carroll (1970)

Le differenze riguardano alcuni criteri di composizione del corpus, sia di tipo qualitativo, sia di tipo quantitativo.

– *Limiti cronologici*. I testi di *a* sono tutti posteriori alla seconda guerra mondiale; sono compresi infatti tra il 1947 e il 1968. I testi di *b* appartengono invece al periodo tra le due guerre: il primo testo è del 1920, il più recente del 1940.

– *Sottoinsiemi*. Tre sottoinsiemi di *a* possono essere messi in corrispondenza con tre sottoinsiemi di *b*, almeno secondo le etichette loro assegnate dagli autori: ' teatro ' di *a*, indicato d'ora in poi con la sigla Ta, e ' dramatic Literature: Plays ' di *b*, sigla Tb; ' romanzi di *a* sigla Ra, e ' Fictional Literature: ' Novels and short Stories ' di *b* sigla Rb; giornali di *a* sigla Ga, e ' periodical Literature: Newspapers and Magazines ' di *b* sigla Gb [57]. Gli altri insiemi

e Rosengren (1971). Questi ultimi due hanno proposto, in sostituzione di U, due nuovi indici di uso, rispettivamente UM e KF, che sarebbe troppo lungo illustrare qui. Tuttavia abbiamo preferito usare U nel compilare *a* perché ci sembrava « importante nello stato attuale delle ricerche di statistica linguistica assicurare la compatibilità tra i risultati di ricerche diverse » (*a*, XXIX), e in particolare con *b*, già annunciato, nella collana « The romance languages and their structures » diretta da Juilland, quando mettemmo mano alla compilazione di *a*. (A questo proposito è interessante riferire un esperimento di Engwall (1974). Egli ha calcolato i valori di F, R, U, UM, KF per i lemmi di un corpus di 50.000 occorenze tratte da 25 romani, francesi. Egli ha poi ordinato i lemmi in ordine decrescente secondo ciascuno dei 5 metodi. Per costatare il grado di concordanza tra di essi, ha usato l'indice di Spearmann (vedere più avanti nota). I valori trovati indicano una correlazione molto forte tra U, UM e KF. Questo è un'argomento che, sia pure a posteriori, ci tranquillizza sulla opportunità di aver rinunciato a perfezionamenti di U per favorire la comparabilità di *a* e di *b*. Ci siamo però discostati per quanto riguarda la scelta del limite inferiore di accettazione. Nel compilare *b* Juilland e Traversa si sono proposti di accogliere 5.000 lemmi, ed hanno scelto come « cut–off point the coefficient which fell closest to 5.000 ». In *b* « this boundary happens to fall at the coefficient of 2,39, wich gives us 5.014 words » (Juilland e Traversa, 1973, XLVI). In *a* il limite immediatamente successivo a 5.000 cadrebbe al coefficente 2,00, in corrispondenza del lemma 5001. Ma « abbiamo osservato che, giungendo fino al lemma 5356, l'ultimo dei lemmi che hanno U = 1,78, si includono tutti i lemmi con R = 3. Abbiamo pertanto fissato ad U = 1,78 la soglia inferiore entro la quale accogliere i lemmi » in *a*. Si noti che 1,78 è il valore di U per F = 3 e R = 3. « Così facendo abbiamo aggiunto ai 5001 lemmi solo 95 lemmi con R < 3 e 260 lemmi con R = 3. Poiché tra il lemma 1 e il lemma 5001 i lemmi con R<3 sono 331 » *a* « risulta costituito da 426 lemmi con R < 3 e da 4930 con R = 3 ». (*a*, XXX). Occorre qui aggiungere un'altra osservazione. In *a* il coefficente 2,39 corrisponde al lemma 4352; 662 lemmi meno che in *b*. Questo fatto si spiega considerando la diversa struttura quantitativa lessicale del cinema, presente in *a*, rispetto a quello della prosa tecnico–scientifica presente in *b*, ma soprattutto con la diversità dei criteri di lemmatizzazione, più analitici in *b*. Infatti, dopo le operazioni di unificazione descritte alla nota precedente, alla soglia di 2,39 corrispondevano 4334 lemmi in *a* e 4503 in *b*.

(57) Questa corrispondenza è stata affermata in un colloquio svoltosi tra Juilland, Tagliavini, Zampolli nel 1971. Naturalmente essa va accolta con beneficio di inventario, ed è condizionata dalla soggettività della definizione dei sottoinsiemi, la quale, come si è detto, attende per principio una verifica ed eventualmente una correzione a posteriori sulla base dei risultati dell'analisi. Tuttavia i dati del nostro spoglio sembrano confermare, almeno in prima istanza, la corrispondenza sopraddetta.

hanno invece ' etichette ' indipendenti: ' cinema ' (sigla C) e ' sussidiari ' (sigla S) in *a*, ' Essaystic Literature: Essays ' (sigla E) e ' Technical Literature: Technical and Scientific Writings ' (sigla TS) in *b*.

– *Composizione dei sottoinsiemi*. *a* e *b* hanno seguito criteri diversi nel comporre i sottoinsiemi. In primo luogo la composizione di *a* è meno varia, più omogenea. In *a* infatti troviamo: Ta, 10 testi e 10 autori; Ra, 10 testi e 10 autori; Ga, 5 testate; C, 8 copioni (uno dei quali in 4 episodi di autori diversi) e 9 autori; S, 3 testi (però ogni sussidiario era suddiviso in sottosezioni relative a materie diverse). In *b* troviamo invece la seguente composizione: Tb, 60 testi e 27 autori; Rb 55 testi e 44 autori; Gb, 32 testate; E, 50 testi e 42 autori; i testi TS sono presi da 33 pubblicazioni diverse. Inoltre, mentre in *b* sono state estratte a sorte dai testi delle frasi isolate (in media una frase per pagina) (58), in modo da raggiungere 100.000 parole per sottoinsieme, in *a* da ogni testo sono stati presi dei brani continui. Nel cinema è stato preso l'intero testo, nel teatro e nei romanzi da ciascuno dei 10 testi è stato preso un blocco di 10.000 parole. Dai giornali sono stati presi articoli interi dalle diverse sezioni; dai sussidiari sono stati presi dei blocchi da ogni materia.

La diversità di composizione del corpus di *a* e di *b* risulta dunque da almeno tre fattori diversi. Ne segue che le differenze risultanti da un eventuale confronto tra *a* e *b* possono essere correlate, in linea di principio, a uno o all'altro di questi fattori, o a loro combinazioni. Non ci troviamo cioè nella situazione corretta per un confronto, che sarebbe evidentemente quella di mantenere costanti tutti i fattori salvo uno: in particolare, per quello che in questi discorso ci interessa, le ' etichette ' dei sottoinsiemi. Nulla può essere fatto, ora, per ridurre questa pluralità di fattori.

Una volta chiarito che non ci proponiamo di confrontare *a* e *b* alla ricerca di spiegazioni delle loro eventuali differenze, esaminiamo alcuni dati sulla loro struttura quantitativa, approfittando delle somiglianze metodologiche nella loro elaborazione.

Dall'esame dei 5 sottoinsiemi di *a* provengono alcuni dati che possono servire come esempio dei dati di fatto che, accumulandosi col diffondersi degli spogli elettronici, depongono a favore di un modello ' stratificato ' della lingua come popolazione statistica.

Applichiamo il calcolo del X^2 alle frequenze delle categorie grammaticali per tutte le 10 coppie che possono essere formate con i 5 sottoinsiemi di *a* (59). I valori ottenuti (v. Tabella 1) sono tutti molto alti, e l'ipotesi di una differenza

(58) Per maggiori dettagli si veda Juilland e Chang Rodriguez (1964, XXVI, XXVIII).

(59) Come è noto il test del X^2 (o test di Pearson) « est d'une application constante en statistique linguistique » (Muller, 1973, p. 116) nella quale viene usato non solo per verificare « l'hypothèse nulle » (*ivi*, p. 125) o come misura di correlazione (*ivi*, p. 145), ma anche per mettere in rilievo fatti interessanti di altro tipo (*ivi*, p. 125).

Nel nostro caso, lo abbiamo adoperato come indice di somiglianza tra le distribuzioni di frequenza delle categorie grammaticali di una coppia di insiemi.

significativa nella ripartizione delle categorie grammaticali tra i sottoinsiemi può essere senz'altro mantenuta (60). Per es. il valore che si trova all'incrocio tra la prima riga e la seconda colonna rappresenta il valore di X^2 calcolato per la ripartizione delle categorie grammaticali nel cinema e nel teatro.

Se consideriamo ora la riga del cinema, vediamo che i valori di X^2 crescono da sinistra verso destra. In altre parole la differenza misurata dal χ^2 tra la distribuzione delle categorie grammaticali nel cinema e negli altri sottoinsiemi cresce con l'ordine seguente: C, Ta, Ra, Ga, S. Consideriamo ora l'ultima riga, quella dei sussidiari: la differenza tra i sussidiari e gli altri sottoinsiemi decresce. Si delinea quindi un possibile ordinamento di sottoinsiemi C, Ta, Ra, Ga, S, regolarmente confermato dalle altre tre righe (61).

La formula del X^2 come è noto, è

$$X^2 = \Sigma \frac{(o - t)^2}{t}$$

dove t è un valore teorico e o il valore osservato.

Nel calcolo di X^2 sulle frequenze delle categorie grammaticali in 2 sottoinsiemi (x e y), se oxa è la frequenza della categoria (a) nel sottinsieme x, oya è la frequenza della categoria (a) nel sottinsieme y, Tx è il totale delle occorrenze in x, Ty è il totale delle occorrenze in y, allora, per esempio, al valore osservato oxa corrisponde il valore teorico

$$\frac{(oxa + oya)\,Tx}{Tx + Ty}.$$

Il numero dei valori osservati è nel nostro caso 36, e cioè uguale al numero delle categorie grammaticali (18) moltiplicato 2.

(60) I valori del X^2 sono tutti estremamente significativi, e dunque sembra di poter concludere che nella distribuzione delle frequenze delle categorie grammaticali sono presenti delle deviazioni che richiedono delle spiegazioni diverse dal puro caso.

(61) Riportiamo qui sotto, per ogni sottoinsieme, i due sottoinsiemi che, in ordine decrescente, risultano 'più' simili in base al X^2 sulle categorie grammaticali:

C : Ta, Ra
Ta : Ca, Ra
Ra : Ga, Ta
Ga : Sa, Ra
Sa : Ga, Ra

La osservazione più immediata, confermata dai dati che esporremmo in seguito, è che siano individuabili due blocchi, da un lato C e Ta, dall'altro S e Ga, tra i quali si situa Ra, che costituisce per così dire un intermediario tra essi.

Se consideriamo l'ordine di 'somiglianza decrescente' (e cioé X^2 crescente) tra le coppie:

a) (1) C—Ta
(2) Sa—Ga
b) (3) Ra—Ga
(4) Ra—Ta
(5) Ra—S
(6) Ra—C
c) (7) Ga—Ta
(8) S—Ta
(9) Ga—C
(10) S—C

TABELLA 1.

Valori di X^2 *per la ripartizione delle frequenze delle categorie grammaticali per i 5 sottoinsiemi di a.*

	C	Ta	Ra	Ga	S
C	—	613,993	6310,065	14676,763	18179,802
Ta	613,993	—	3572,389	10369,345	13042,016
Ra	6310,065	3572,389	—	2665,428	4183,161
Ga	14676,763	10369,345	2665,428	—	1143,189
S	18179,802	13042,016	4183,161	1143,189	—

Se applichiamo il calcolo del X^2 alle frequenze dei lemmi presenti in ogni coppia di sottoinsiemi (62) (Tabella 2) (63), troviamo pienamente confermato l'ordine in questione (64). La Tabella 2 riporta, per ogni coppia di sottoinsiemi, i valori di

$$Z = \sqrt{2\,X^2} - \sqrt{2v - 1}$$

dove v sono i gradi di libertà. Questi valori sono comparabili per ogni coppia perché indipendenti da v, diverso per ogni coppia.

vediamo che ai primi due posti stanno le coppie che costituiscono i due blocchi, seguite (3—6) dalle coppie costituite da Ra con gli elementi dei 2 blocchi, seguite infine (7—10) dalle coppie costituite da elementi di blocchi diversi.

(62) Cioè nel calcolo di X^2 per una data coppia sono stati inclusi solo i lemmi presenti in entrambi i sottoinsiemi della coppia.

(63) Le tabelle del X^2 forniscono di solito i dati fino a 30 o al massimo 100 gradi di libertà. Nel caso in questione i gradi di libertà sono sull'ordine dei 4000–4500, e variano da coppia a coppia.

Per questo motivo nella Tabella 2 (come anche nelle Tabelle 8 e 14) non abbiamo riportato il valore di X^2 ma un valore Z calcolato a partire dal X^2 con la formula seguente (dove v sono i gradi di libertà):

$$Z = \sqrt{2\,X^2} - \sqrt{2v - 1}$$

che ha il vantaggio di essere indipendente dai gradi di libertà. Anche in questo caso si deve notare che i valori di Z sono sempre estremamente significativi.

(64) Mentre le due coppie ' più ' simili sono le stesse che per le categorie grammaticali (vedi nota 61 in principio), l'ordine decrescente di somiglianza tra le coppie è parzialmente diverso:

C—Ta , Ra—Ta , Ga—Ra , Ra—C , S—Ga , S—Ra , Ga—Ta , Ga—C , S—Ta , S—C

Essenzialmente, mentre resta confermata la somiglianza del blocco C—Ta, si ha un allontanamento di S da tutti gli altri sottoinsiemi, pur restando più simile a Ga.

TABELLA 2.

Valori di Z per l'insieme delle frequenze dei lemmi comuni ad ogni coppia di sottoinsiemi di a.

C	Ta	Ra	Ga	S
C	—	43,130	133,517	202,442	252,861
Ta	43,130	—	103,673	169,561	233,501
Ra	133,517	103,673	—	121,034	165,907
Ga	202,442	169,561	121,034	—	136,851
S	252,861	223,501	165,907	136,851	—

L'indice di correlazione tra i ranghi dei lemmi dei diversi sottoinsiemi segue lo stesso ordine (65) (v. Tabella 3).

(65) Come è noto, l'indice ρ di Spearman misura la correlazione tra i ranghi assegnati da due ordinamenti diversi alla stessa serie di dati. Nel nostro caso l'indice è stato calcolato per le 10 possibili coppie di sottoinsiemi di *a*. Data una coppia di sottoinsiemi x e y, il rango in x é stato assegnato nel modo seguente. I lemmi presenti in almeno uno dei due sottonsiemi della coppia (cioè presenti sia in x sia in y, presenti solo in x, presenti solo in y) sono stati ordinati secondo l'ordine delle frequenze in x. Il rango di un lemma in x è il suo numero progressivo in quest'ordine. Naturalmente, come d'uso, ai lemmi che hanno la stessa frequenza è stata attribuita come rango la media dei loro numeri progressivi. I lemmi con frequenza zero in x, e cioè presenti solo in y, si trovano alla fine di quest'ordinamento, ed hanno ricevuto, come rango in x, il rango immediatamente successivo a quello assegnato al lemma meno frequente in x. In modo analogo si è proceduto per assegnare il rango in y. (Il trattamento dei lemmi con frequenza zero mi è stato suggerito da Ch. Muller).

La formula per il calcolo di ρ per una coppia di sottionsiemi x e y è:

$$\rho = 1 - \frac{6\,\Sigma d^2}{n\,(n^2 - 1)}$$

dove d è la differenza tra il rango di un lemma in x e il suo rango in y, ed n è il numero dei lemmi presenti in almeno uno dei due sottoinsiemi della coppia. Come è noto, ρ varia da —1 a +1 con il significato seguente:

- sarà di +1 se c'è identità perfetta tra i due ordinamenti,
- sarà di —1 se i due ordinamenti sono esattamente inversi,
- sarà tanto più vicino a zero quanto più la dipendenza tra i due ordinamenti è debole.

Quando n è molto grande, la significatività può essere verificata (approssimativamente) calcolando

$$t = \rho \sqrt{\frac{n - 2}{1 - \rho^2}}$$

Tutti i valori sono risultati positivi e largamente significativi. Le Tabelle 3, 9 e 15 riportano i valori di t, che essendo indipendenti da n, sono comparabili tra loro.

Tabella 3.

Valori dell'indice di correlazione tra i ranghi dei lemmi nelle coppie di sottoinsiemi di a.

	C	Ta	Ra	Ga	S
C	—	0,678	0,592	0,491	0,434
Ta	0,678	—	0,594	0,509	0,425
Ra	0,592	0,594	—	0,440	0,437
Ga	0,491	0,509	0,440	—	0,476
S	0,434	0,425	0,437	0,476	—

Tabella 4.

Frequenze percentuali delle categorie grammaticali nei sottoinsiemi di a.

	C	Ta	Ra	Ga	S	Ta+Ra+Ga	Tot.
Sost.	13,376	14,848	16,729	18,929	20,440	16,804	16,820
Agg.	4,277	4,701	4,958	6,701	7,070	5,437	5,522
Art.	8,208	10,159	13,794	16,804	19,286	13,535	13,580
Prep. . . .	8,624	9,443	14,048	16,735	15,374	13,354	12,786
Verbi . . .	24,038	21,848	17,713	14,763	14,911	18,162	18,722
Pron. . . .	11,627	11,435	8,502	5,161	5,419	8,414	8,481
Avv. . . .	9,987	8,711	6,454	4,869	3,221	6,707	6,691
Cong. . . .	5,989	5,553	6,428	5,159	5,338	5,728	5,698
Num. . . .	0,811	0,817	0,631	0,998	0,490	0,813	0,749
Int.	0,778	0,773	0,131	0,020	0,017	0,315	0,350
Ausil. . . .	2,980	2,645	2,348	2,755	1,535	2,581	2,457
Avv.+Prep. .	0,983	1,054	1,565	1,310	1,234	1,308	1,227
Int.+Sost. .	0,007	0,018	0,008	0,001	0,002	0,009	0,007
Agg.+Pron. .	3,026	3,195	2,966	2,736	2,950	2,969	2,977
Agg.+Avv. .	2,726	2,468	2,052	2,097	1,765	2,208	2,227
Int.+Avv. .	0,436	0,312	0,130	0,078	0,059	0,175	0,205
Pron.+Avv. .	1,438	1,302	0,839	0,471	0,436	0,877	0,905
Avv.+Cong.	0,694	0,708	0,702	0,412	0,453	0,610	0,596

Come già per X^2, adopero quest'indice non per stabilire se esista o no correlazione tra i ranghi di due sottoinsiemi, ma come misura della loro maggiore o minore prossimità. L'analisi dei dati conferma largamente quanto detto nella nota 64.

TABELLA 4a.

Frequenze percentuali delle categorie grammaticali nei sottoinsiemi di a prima della fusione dell'omografia funzionale.

	C	Ta	Ra	Ga	S	Tot
Sost. . .	14,67	16,31	18,71	21,86	23,05	18,92
Agg.. . .	10,23	10,89	11,06	12,75	13,06	11,60
Art. . . .	8,12	10,02	13,36	15,88	18,36	13,14
Prep. . .	8,57	9,13	13,49	16,06	14,87	14,42
Verbi . .	26,72	24,19	20,17	17,22	16,19	20,90
Pron. . .	12,00	11,82	8,62	5,10	5,24	8,56
Avv. . .	12,69	11,22	8,23	6,06	4,24	8,49
Cong. . .	5,87	5,36	6,16	5,02	4,97	5,48
Int. . .	1,13	1,06	0,20	0,065	0,02	0,49

TABELLA 4b.

Frequenze percentuali delle categorie grammaticali nell'intero corpus di a.

	C	Ta	Ra	Ga	S	Ta+Ra+Ga	Tot
Sost.	14,504	16,547	18,392	21,606	22,792	18,711	18,705
Agg.	4,594	5,178	6,010	7,555	8,170	6,244	6,303
Art.	7,913	9,711	12,902	15,400	17,762	12,745	12,668
Prep. . . .	8,441	9,293	12,437	16,021	14,837	12,945	12,437
Verbi . . .	23,713	21,666	17,847	14,655	14,707	18,060	18,511
Pron. . . .	11,282	11,030	8,028	4,791	5,036	7,954	8,030
Avv. . . .	9,385	8,273	6,257	4,812	3,073	6,449	6,355
Cong. . . .	6,053	6,004	5,324	4,750	4,934	5,378	5,413
Num. . . .	0,772	0,913	0,916	0,966	1,227	0,987	0,916
Int.	0,917	0,935	0,185	0,037	0,021	0,386	0,418
Aus.	2,872	2,528	2,196	2,525	1,414	2,416	2,305
Avv.+Prep..	0,799	0,875	1,275	1,133	1,130	1,095	1,043
Int.+Sost. .	0,079	0,044	0,013	0,001	0,002	0,019	0,028
Agg.+Pron..	2,909	3,048	2,769	2,507	2,716	2,775	2,790
Agg.+Avv. .	2,619	2,347	1,911	1,913	1,624	2,057	2,082
Int.+Avv. .	0,447	0,326	0,153	0,073	0,056	0,184	0,211
Pron.+Avv..	1,386	1,245	0,784	0,432	0,402	0,821	0,849
Avv.+Cong.	0,985	0,919	0,863	0,520	0,560	0,768	0,769

Tabella 5.

Classi di frequenza in a.

Classe		Frequenza Percentuale				
		C	Ta	Ra	S	Ga
1	1–500	88,764	84,451	82,555	81,397	78,747
2	501–1000	5,272	6,135	6,763	8,240	8,217
3	1001–1500	2,532	3,184	3,728	4,340	4,845
4	1501–2000	1,481	1,956	2,466	2,547	3,080
5	2001–2500	0,979	1,292	1,713	1,609	2,015

Classe		Frequenza media				
		C	Ta	Ra	S	Ga
1	1–500	175,10	168,21	159,69	154,79	148,03
2	501–1000	10,40	12,07	13,48	15,67	15,44
3	1001–1500	51,96	62,26	7,21	8,25	9,10
4	1501–2000	29,22	3,85	4,77	4,84	5,79
5	2001–2500	1,93	2,54	3,31	3,06	3,78

Tabella 6.

Metà, Media, Mediana in a.

		Metà		Media		Mediana	
		% sul Totale delle occorrenze	Frequenza media del lemma	% sul Totale delle occorrenze	Frequenza media del lemma	% sul Totale delle occorrenze	Frequenza media del lemma
I Parte	C	97,339	55,498	85,475	242,967	0,838	28,450
	Ta	96,701	47,781	83,044	209,589	0,753	24,700
	Ra	95,967	43,990	81,186,	179,688	0,592	22,920
	S	96,106	48,049	79,923	171,948	0,630	24,990
	Ga	95,409	42,421	77,953	157,234	0,567	22,230
II Parte	C	2,661	1,517	14,525	4,603	99,161	28,510
	Ta	3,299	1,630	16,956	4,663	99,247	24,700
	Ra	4,239	1,848	18,814	4,811	99,408	22,920
	S	3,894	1,946	20,077	5,678	99,370	24,990
	Ga	4,591	2,041	22,047	5,509	99,433	22,230

La probabilità che i valori citati si dispongano sempre in quest'ordine per puro caso è del tutto irrilevante. Ci si deve dunque chiedere se questo ordinamento dei 5 sottoinsiemi abbia un significato.

Proviamo per esempio ad analizzare il dato globale relativo alle categorie grammaticali. Nella Tabella 4 riportiamo le frequenze percentuali di ogni categoria grammaticale nei 5 sottoinsiemi, disposti secondo l'ordine sopra specificato.

Rispetto a quest'ordine le categorie grammaticali si dividono in tre gruppi principali:

(*a*) le frequenze di sostantivi, articoli, preposizioni, aggettivi crescono secondo quest'ordine;

(*b*) decrescono secondo quest'ordine le frequenze di verbi, pronomi, avverbi.

(*c*) non seguono quest'ordine le congiunzioni e i numerali [66].

(66) A proposito delle percentuali riportate nella Tabella 4, si impongono alcune considerazioni preliminari.

1) Nella Tabella 4, oltre alle 11 categorie grammaticali (nell'ordine: sostantivi, aggettivi, articoli, preposizioni, verbi, pronomi, avverbi, congiunzioni, numerali, interiezioni, ausiliari) appaiono 7 coppie: avverbi + preposizioni, interiezioni + sostantivi, aggettivi + pronomi, aggettivi + avverbi, interiezioni + avverbi, pronomi + avverbi, avverbi + congiunzioni. La nota 55 spiega le ragioni di questi accoppiamenti, che vanno interpretati così: per esempio 0,083 è la frequenza percentuale, nel sottoinsieme C, che si ottiene sommando tra loro le frequenze di tutti i lemmi per i quali non è stato distinto l'uso avverbiale dall'uso preposizionale. Si noterà che le categorie maggiormente intaccate sono avverbi, pronomi, aggettivi e preposizioni. Praticamente, la percentuale del 9,987% di avverbi in C è inferiore alla percentuale effettiva, perché essa non include le occorrenze degli avverbi conteggiate nelle coppie avv. + prep., + agg., + int., + pro., + cong. Tuttavia si può dimostrare che questo fatto non invalida le considerazioni sull'ordinamento dei sottoinsiemi. Poiché la dimostrazione completa sarebbe troppo lunga, si suggerisce di confrontare la Tabella 4 con la Tabella 4 a, nella quale sono riportate le frequenze percentuali delle categorie grammaticali ricavate in una fase intermedia della elaborazione di *a*, nella quale non erano ancora stati fusi diversi usi grammaticali dei lemmi in questione.

2) Le percentuali riportate in 4 e in 4 a, come anche in 12 e 12 a, si riferiscono allo insieme delle occorrenze corrispondenti ai soli lemmi accettati in *a* e in *b*, e non ai due rispettivi corpora. La Tabella 4 b riporta le percentuali sull'intero corpus di *a*. Come si vede, le variazioni tra le Tabelle 4 a e 4 b sono minime, e consistono, come era logico attendersi, in un lieve aumento di sostantivi (poco meno del 2%) aggettivi (poco meno dell'1%) interiezioni (0,2% circa), e cioè delle categorie cui, per la maggior parte, appartengono i lemmi che non hanno raggiunto il valore di U necessario per essere accolti in *a*. Tuttavia si può agevolmente constatare che queste piccole variazioni non influiscono in alcun modo sulle considerazioni relative all'ordinamento dei sottoinsiemi. Anche se le occorrenze del corpus non rappresentate in *b* (circa 80.000) sono più numerose di quelle escluse da *a* (circa 20.000) pare lecito supporre che questa considerazione valga anche per *b*. Possiamo ora analizzare più in dettaglio l'ordinamento dei sottoinsiemi (C Ta Ra Ga S) secondo il quale, come si è detto:

(1) crescono le percentuali di sostantivi, aggettivi articoli, preposizioni;
(2) decrescono le percentuali di verbi, pronomi, avverbi, interiezioni, ausiliari.

(È da rilevare che delle due categorie che non seguono quest'ordine, numerali e congiunzioni, quest'ultima può essere scomposta in due sottocategorie: le congiunzioni di coordinazione

L'ordine C, Ta, Ra, S, Ga, è sostanzialmente confermato dai dati relativi alle classi di frequenza (Tabella 5) [67], alla metà, alla media, e alla mediana dei 5 sottoinsiemi (Tabella 6) [68].

che si comportano come 1, e quelle di subordinazione che si comportano come 2). Si notano 4 eccezioni su 45 caselle. 3 di esse riguardano l'ordine di Ga e S, che risulta invertito per preposizioni, verbi, pronomi. La quarta eccezione riguarda la frequenza degli ausiliari nei giornali, che si situa tra C e Ta. Per quanto riguarda le 7 coppie, due di esse sono costituite da categorie di tipo 2 (int. + avv., pro + avv.), e il loro comportamento è di tipo 2; quattro sono ' miste ': due (agg. + pron. e agg. + avv.) si comportano come il tipo 2, le altre due sono irregolari, come pure la settima coppia (pron. + avv.). Osserviamo inoltre che tutte le eccezioni cadono se consideriamo i tre blocchi cui accennammo nella nota 61.

Possiamo chiederci quali fatti linguistici corrispondano a queste regolarità. Regolarità di questo tipo erano state rilevate in passato, sia pure in forma più limitata. Si pensi per esempio al quoziente VAR (Verb–Adjective Ratio = numero di verbi/numero aggettivi), introdotto da Busemann (1925) e studiato da Boder (1940), Schlismann (1948) Miller (1951). Busemann correlò il VAR con la stabilità emotiva. Anche Boder sostenne che questo quoziente riflette caratteristiche individuali. Entrambi riconobbero che il VAR è più alto nel parlato che nello scritto, e Busemann spiegò questo fatto sostenendo che quanto maggiore è il tempo di cui disponiamo per produrre simboli, tanto più numerose sono le qualificazioni che aggiungiamo. Ma già Boder e soprattutto Miller riconobbero che « Il più importante fattore che influenzi il rapporto verbo–aggettivo è il genere di testo. Una stessa persona fornisce dei rapporti diversi a seconda del compito di fronte a cui si trova » (Miller, 1972, p. 183) « Ad esempio, nella sceneggiatura di un dramma il rapporto sale a 9 verbi per aggettivo, mentre in scritti scientifici scende a 1,3. Gli statuti legali, che trattano delle azioni di esseri umani, contengono 5 verbi per aggettivo. È interessante osservare che nelle tesi di laurea il rapporto è di 1,5 verbi per aggettivo, ma al momento in cui l'allievo arriva alla tesi di dottorato è riuscito a ridurre questo rapporto a 1,1 » *ivi*). Antosch (1968) ha studiato « la diagnosi dello stile letterario con il quoziente verbo–aggettivo ». Secondo lui il problema della differenza tra parlato e scritto implica la questione della posizione del dramma, che egli ritiene « intermedia » tra i due. I dati in suo possesso dimostrerebbero che VAR è più alto negli scritti drammatici che negli altri, negli scritti dialettali che nel linguaggio standard, nella prosa che nella poesia. Importante sarebbe il ' modo di presentazione ' (VAR decrescerebbe dal dialogo – più vicino al parlato – al monologo, alla prosa narrativa e discorsiva: e infatti VAR cambierebbe nei romanzi e nei drammi da parte a parte a secondo del modo di presentazione e del ' tema '. Con questo ultimo egli intende in realtà distinzioni del tipo favole popolari, favole letterarie, romanzi, prosa critica o scientifica. VAR, cioè, non si limiterebbe a distinguere testi drammatici e non drammatici, ma all'interno di questi ultimi differenzierebbe alcune diverse sottoclassi. In particolare la narrativa si distinguerebbe dalla prosa scientifica e descrittiva). VAR decresce nei sottoinsiemi di *a* secondo l'' ordine solito ':

$$C = (2\,,\,6)\,;\ Ta = (2\,,\,2)\,;\ Ra = (1\,,\,9)\,;\ Ga = (1\,,\,3)\,;\ S = (0\,,\,9)\,.$$

Secondo i criteri di Antosch, quest'ordine corrisponderebbe a un graduale allontanamento dal parlato. Altri elementi formali da lui indicati concorderebbero con l'ipotesi dei 3 blocchi : dialogo (C , Ta) dialogo e narrativa (Ra) descrizione (Ga e S). Sull'ordinamento C, Ta potrebbe influire il carattere dialettale del cinema e la maggior quantità di monologhi nel teatro. Sull'ordinamento Ga , S influirebbe la compresenza, in Ga, di elementi narrativi e descrittivi. A noi non interessa qui trovare elementi formali di questo tipo. Siamo convinti che le interpretazioni di Antosch si fondino su una eccessiva semplificazione dei fatti. Bisognerebbe innanzitutto analizzare all'interno dei sottoinsiemi i singoli testi e le loro parti. Ci interessa invece rilevare la presenza dei due gruppi di categorie grammaticali, del cui comportamento VAR è forse solo una conseguenza particolare, e osservare come questo fatto concordi

Le note 67–68 si trovano a pp. 85–86.

Ripetiamo l'esame degli stessi dati per *b*. Se consideriamo solo i tre sottoinsiemi con etichette comuni a quelle di *a* (teatro, romanzo, giornali) troviamo

globalmente, con gli altri dati, nell'esigere un modello quantitativo di lingua variamente stratificato, in primo luogo secondo il genere. Siamo invece ancora molto lontani dallo spiegare con certezza il comportamento dei due gruppi in questione. In un certo senso essi sono in rapporto con la struttura sintattica della proposizione e, in certi limiti, del periodo. Gli studiosi cui ho sottoposto i dati hanno avanzato spiegazioni diverse. Molti ritengono si tratti di una opposizione tra stile nominale e stile verbale. Alcuni insistono sulla lunghezza media, approssimativamente, delle proposizioni: poiché in media ogni proposizione ha un verbo, più si abbassa la lunghezza media delle proposizioni maggiore diviene la frequenza percentuale dei verbi a scapito delle altre categorie. Altri invece insistono su diverso grado di presenza del dialogo nei sottoinsiemi di *a*. Maggiore è la presenza del dialogo, più cresce la percentuale dei pronomi rispetto agli aggettivi. Crescendo il numero dei pronomi, il gruppo nominale si accorcerebbe, perché spesso la presenza del pronome esclude quella di articoli, aggettivi, e preposizioni. Occorre andare con i piedi di piombo e prendere spunto dalle indicazioni statistiche per analisi dettagliate dei contesti. (Per esempio le ragioni che determinano il variare delle percentuali delle interiezioni e forse anche degli ausiliari potrebbero essere diverse dalle ragioni delle altre variazioni). Lo ' Standard Corpus of Present–Day Edited American English ', raccolto alla Brown University nel 1963–64 sotto la direzione di W. Nelson Francis, può fornire, a motivo delle varietà di composizione, dei dati interessanti per il problema che stiamo trattando. Il corpus comprende 500 campioni di circa 2000 parole ciascuno, ripartiti nelle seguenti categorie o ' generi ':

A Press: reportage; B Editorial; C Rewiews; D Religion; E Skills and Hobbies; F Popular Lore; G Belles Lettres, Biography, ecc.; H Miscellaneous; J Learned and Scientific Writings; K Fiction: general; L Fiction: Mystery; M: Fiction Science; N Fiction: Adventure and Western; P Fiction: Romance and Love Story; R Humor. Avendo ottenuto dagli autori il nastro magnetico più recente nel quale le parole sono lemmatizzate e classificate grammaticalmente, mi riprometto di rianalizzarlo dal punto di vista che sto esponendo. Un primo rapido sguardo alle parole di maggiore frequenza e alle categorie grammaticali indicherebbe la possibilità di disporre i sottoinsiemi secondo un ordine con caratteristiche analoghe a quelle riscontrate nei sottoinsiemi di *a* e di *b*. In prima approssimazione si possono identificare 3 gruppi di sottoinsiemi: narrativa, giornali, scritti tecnici – saggistica. Le categorie grammaticali si dispongono in quest'ordine. M. L. Marckworth e L. M. Bell (1967) hanno studiato la distribuzione della lunghezza delle frasi nei 16 sottoinsiemi, con lo scopo dichiarato di ' determine whether sentence length is a significant parameter with quantitative description of writing style in the various literary prose genres of the corpus; i.e. does the genre impose some sort of constraint in this matter on the individual practitioner? ' Anche nel loro caso, il X_2 tra i diversi sottoinsiemi relativi alle frequenze delle lunghezze delle frasi è sempre significativo, e conforta l'ipotesi che la lunghezza delle frasi sia effettivamente una variabile dello stile dei generi. Dopo aver ordinato i generi secondo la lunghezza media delle frasi, è risultata immediatamente evidente una prima fondamentale suddivisione del corpu in due categorie: (I) informative prose (II) imaginative prose. È risultata evidente una correlazione tra una lunghezza media bassa e un'alta percentuale di materiale in forma di citazione. Nella categoria (I) questo materiale è costituito prevalentemente da citazioni di altre fonti; nella categoria (II) dalla riproduzione di dialoghi. Tutti i dati portano alla conclusione che ' a high average dialogue content is, in general, correlated with a low mean sentence lenght '. Gli autori individuano 3 elementi ' qualitativi ' che possono essere messi in relazione con i dati « quantitativi », ec he contribuiscono ad elevare la lunghezza media. (a) Il genere è tale che l'autore e l'ipotetico pubblico sono coscientemente interessati alla eleganza stilistica delle esposizioni. Questo fattore opererebbe per es. in ' belles lettres ' (G) e in ' Press–Reviews (C). (b) Il genere impone all'autore di ' make highly precise statements that exclude a set

confermato per tutti i parametri esaminati l'ordinamento Tb, Rb, Gb, in tutto simile all'ordinamento Ta, Ra, Ga.

of further possibilities; i.e., the statements are likely to be extensively modified for the purposes of exactness'. Per es. in H ('technical governmental legal statements') e J ('Learned and Scientific Writings'). (c) 'The genre is one in which a specialized set of concepts is assumed to be available in the audience'. Questo fattore sarebbe evidente in J e in D. I dati dimostrano che 'such paralinguistic elements do have a significant effect on authors, at least in terms of the specific quantitative measure under discussion'. Gli autori concludono che « The chief value of this experiment, over and above the determination of the factual results above, has been to demonstrate again that, using a text of sufficient size to minimize the idiosyncrasies of individual authors, strictly quantitative measurements can be made of what has heretofore been the preserve of the aesthetic elements of language style ».

Sarebbe interessante, evidentemente, confrontare i dati delle Tabelle 4, 4a, 4b, e 12 con una statistica sull'italiano parlato.

In attesa che si concludano alcune tesi in corso presso l'Istituto di glottologia di Pisa, gli unici dati in mio possesso sul parlato sono quelli pubblicati in P. Castrogiovanni (1973) e P. Castrogiovanni *et alii* (1969). La statistica è stata condotta su un corpus di risposte al test di Rorschach registrate presso la Clinica Psichiatrica dell'Università di Pisa. Il test di Rorschach è stato somministrato a 32 pazienti neurotici e schizofrenici e a un gruppo di controllo costituito da 87 loro familiari, metà uomini e metà donne, con età media di 44 anni, e livello di scolarità limitato nell'80% dei casi alle scuole elementari. Le frequenze percentuali delle categorie grammaticali nelle circa 20.000 parole raccolte per il gruppo dei familiari sono le seguenti: sostantivi 13,07; aggettivi 6,1; articoli 8,0; preposizioni 7,2; verbi 20,5; pronomi 16,0; avverbi 16,5; congiunzioni 6,6; interiezioni 3,7. Da un confronto con le Tabelle 4, 4a e 4b, risulta che questa distribuzione si situerebbe immediatamente a sinistra del cinema. Infatti le frequenze delle categorie del gruppo (1) sono minori, e quelle del gruppo (2) maggiori rispetto al cinema. Fanno eccezione i verbi, la cui frequenza è inferiore a quella del teatro. Questa situazione si presta a commenti diversi, che è tuttavia prematuro valutare. Le risposte al test consistono soprattutto in descrizioni di che cosa una tavola potrebbe raffigurare. Non si tratta cioè di un vero e proprio dialogo. Come abbiamo detto in precedenza, non sappiamo quali siano i tratti che causano le differenze nella distribuzione morfologica tra i diversi sottoinsiemi. Non possiamo cioè assumere una opposizione parlato: scritto. Per questo sarebbe necessario spogliare un corpus 'parlato' stratificato, e verificare il comportamento di ciascuno dei suoi sottoinsiemi, per vedere se hanno dei tratti comuni diversi dai sottoinsiemi dello 'scritto'.

Gli unici dati di cui dispongo provengono dalla ricerca in questione sulle risposte al Rorschach. Il « profilo grammaticale » (Castrogiovanni, 1973, p. 320) degli schizofrenici si differenzia infatti da quello dei neurotici e dei normali per una maggiore utilizzazione delle categorie (1) e una minore utilizzazione del gruppo (2). Non si può naturalmente considerare questa situazione alla stregua della 'stratificazione in generi', ma è interessante che la stessa bipolarità emerga anche dal confronto tra due categorie di malati mentali.

Sulla relazione tra genere e frequenza di tratti linguistici si vedano anche M. Levison *et alii* (1968), Brandwood (1969) Dyer (1973, p. 329). In sostanza, essi dimostrano che le frequenze di alcune caratteristiche, quali la distribuzione di lunghezza delle frasi e l'uso di certe parole di relazione, variano di meno tra opere di autori diversi, ma dello stesso genere, che tra opere dello stesso autore ma di genere diverso. In particolare essi insistono sulla differenza tra opere che riproducono il dialogo e le altre. Si vedano anche Michaelson e Morton (1972) p. 37. Kollmann (1973) ha studiato le variazioni tra generi diversi, in latino, delle parole di maggiore frequenza.

(67) Le classi di frequenza sono delimitate elencando i lemmi in ordine decrescente di frequenza e dividendoli a gruppi di 500. Per le prime 5 classi di frequenza (abbiamo omesso

Qualche oscillazione troviamo invece se consideriamo tutti e 5 i sottoinsiemi. L'ordine infatti è: Tb, Rb, E, TS, Gb per quanto riguarda le categorie grammaticali (Tabelle 7 e 10) (69); Tb, Rb, E, Gb, TS, per l'indice Z e l'indice di correlazione per lemmi (Tabelle 8 e 9); Tb, TS, Rb, E, Gb per i dati relativi alle classi di frequenza, alla metà, alla media e alle mediana (Tabelle 11 e 12, con qualche eccezione per la posizione di TS) (70).

Salta agli occhi che, se si considerano assieme *a* e *b*, Ra e Rb costituiscono una transizione tra due gruppi caratterizzati piuttosto nettamente: C, Ta, Tb da un lato, Ga, S, E, TS, Gb dall'altro.

Proviamo a considerare in dettaglio l'unione dei 10 sottoinsiemi.

L'ordinamento per X^2 sulle categorie grammaticali è rappresentato dalla sequenza C, Ta, Tb, Ra, Rb, Ga, E, S, TS, Gb. (Tabella 13).

Ci sono solo 5 inversioni, tutte concentrate l'una accanto all'altra. Esse sono infatti: nella riga Rb, tra E e Ga; nella riga di Ga, tra E, S, TS, Gb; nella riga di E tra S e TS.

Se esaminiamo l'ordinamento delle frequenze percentuali delle categorie grammaticali abbiamo una conferma dell'ordine nella sequenza da C a Ga,

le classi successive solo per motivi di spazio) la Tabella 5 riporta la frequenza media dei lemmi della classe e la loro frequenza percentuale rispetto al totale delle occorrenze nel rispettivo sottoinsieme. Come si vede, rispetto all'ordine C, Ta, Ra, S, Ga, decrescono la frequenza media e percentuale della prima classe, mentre crescono quelle delle classi successive (e anche quelle delle classi non riportate in tabella); la sola inversione è tra S e Ga nella 2ª Classe). Nel complesso, dunque, il blocco 'drammatico' presenta una maggiore concentrazione di occorrenze in corrispondenza ai lemmi più frequenti, e, di contro, una minor ricchezza di vocabolario.

(68) Siano gli x lemmi di un sottoinsieme ordinati per frequenza decrescente, e sia y il totale delle occorrenze del sottoinsieme.

– La prima parte della metà si riferisce ai primi $x/2$ lemmi.

– La prima parte della media si riferisce ai lemmi con frequenza superiore a y/x

– La prima parte della mediana si riferisce ai lemmi che corrispondono alle prime $y/2$ occorrenze

– La seconda parte di metà, media e mediana si riferiscono rispettivamente agli altri $x/2$ lemmi, ai lemmi con frequenza inferiore a y/x, ai lemmi corrispondenti alle altre $y/2$ occorrenze.

Come si vede, la frequenza percentuale sul totale delle occorrenze e la frequenza media dei lemmi decrescono secondo l'ordine C, Ta, Ra, S, Ga nelle prime metà, e crescono nelle seconde metà secondo lo stesso ordine.

Le inversioni riguardano Ra ed S (percentuale e frequenza media di metà e mediana nella prima parte e della mediana nella seconda parte; percentuale della media nella seconda parte) e S e Ga (frequenze media nella seconda parte). Questo fatto potrebbe essere interpretato come indicazione del carattere non del tutto ben definito del sottoinsieme 'sussidiari', che è un misto di diversi 'modi di esposizione': descrizione, narrativa, esercizi, ecc.

(69) Per quanto riguarda le frequenze percentuali delle categorie grammaticali, valgono gli stessi raggruppamenti osservati in *a*. Si notino però i fatti seguenti. Un massimo in E per gli aggettivi; 3 inversioni tra Ts e Gb (preposizioni, verbi e pronomi: si ricordi l'analogo fenomeno rilevato per *a* tra S e Ga) e inoltre, analogamente ad *a*, la frequenza degli ausiliari si situa oltre quelle di Tb e di Rb. I numerali, come in *a*, sono irregolari, mentre le congiunzioni decrescono, come fanno in *a* le congiunzioni di subordinazione.

(70) Anche in questo caso Tb presenta della analogie con S.

TABELLA 7.

Valori di X^2 per la ripartizione delle frequenze delle categorie grammaticali per i sottoinsiemi di a.

	Tb	Rb	E	Ts	Gb
Tb	—	5762,045	9183,548	12067,696	12786,394
Rb	5762,045	—	848,290	1868,339	2284,064
E	9183,548	848,290	—	725,107	1229,993
Ts	12067,696	1868,339	725,107	—	596,42
Gb	12786,394	2284,064	1229,993	596,42	—

TABELLA 8.

Valori di Z per l'insieme delle frequenze dei lemmi comuni ad ogni coppia di sottoinsiemi di b.

	Tb	Rb	E	Gb	Ts
Tb	—	121,071	151,050	193,176	226,817
Rb	121,071	—	90,879	143,737	176,524
E	151,050	90,879	—	78,252	128,986
Gb	193,176	143,737	78,252	—	113,152
Ts	226,817	176,524	128,986	113,152	—

TABELLA 9.

Valori dell'indice di correlazione tra i ranghi dei lemmi nelle coppie di sottoinsiemi di b.

	Tb	Rb	E	Gb	Ts
Tb	—	0,629	0,541	0,408	0,326
Rb	0,629	—	0,522	0,319	0,305
E	0,541	0,522	—	0,522	0,462
Gb	0,408	0,319	0,552	—	0,533
Ts	0,326	0,305	0,462	0,533	—

Tabella 10.

Frequenze percentuali delle categorie grammaticali nei sottoinsiemi di b.

	Tb	Rb	E	Ts	Gb	Tb+Rb+Gb	Tot
Sost.	13,955	18,199	18,365	18,625	20,412	17,575	17,954
Agg.	4,435	6,269	8,088	7,560	7,957	6,250	6,888
Art.	10,216	16,030	16,912	17,988	18,357	14,935	15,942
Prep. . . .	10,541	15,458	17,149	20,165	18,396	14,864	16,347
Verbi . . .	21,383	16,050	12,903	12,732	12,815	16,678	15,120
Pron. . . .	11,368	7,045	5,746	4,266	4,289	7,508	6,516
Avv. . . .	8,136	4,457	4,149	4,123	3,471	5,316	4,844
Cong. . . .	7,168	6,273	6,953	5,998	5,278	6,221	6,332
Num. . . .	0,391	0,481	0,390	0,539	0,404	0,425	0,440
Int.	0,499	0,076	0,030	0,003	0,011	0,191	0,121
Aus.	2,555	1,603	1,411	1,422	2,687	2,283	1,936
Avv.+Prep.	0,928	1,771	1,266	1,314	1,090	1,264	1,274
Int.+Sost. .	0,003	0,001	0,004	0,0	0,0	0,001	0,001
Agg.+Pron.	3,752	2,831	3,136	2,929	2,588	3,047	3,044
Agg.+Avv.	2,464	1,988	2,206	1,549	1,518	1,982	1,945
Int.+Avv. .	0,271	0,084	0,074	0,067	0,073	0,141	0,113
Pron.+Avv..	0,912	0,425	0.434	0,195	0,226	0,515	0,437
Avv.+Cong.	1,025	0,960	0,782	0,526	0,437	0,802	0,745

Tabella 11.

Classi di frequenza in b.

Classe		Frequenza Percentuale				
		Tb	Ts	Rb	E	Gb
1	1–500	87,102	83,739	82,863	80,066	78,582
2	501–1000	5,968	8,320	7,041	7,337	8,636
3	1001–1500	3,058	4,037	3,872	4,334	4,916
4	1501–2000	1,751	2,091	2,491	2,904	3,105
5	2001–2500	1,156	1,141	1,665	1,434	2,058

Classe		Frequenza Media				
		Tb	Ts	Rb	E	Gb
1	1–500	136,50	130,54	132,86	133,87	129,49
2	501–1000	9,35	12,97	11,29	12,26	14,23
3	1001–1500	4,74	6,29	6,21	7,26	8,10
4	1501–2000	2,74	3,26	3,99	4,85	5,11
5	2001–2500	1,81	1,77	1,94	3,41	3,39

mentre abbiamo numerose inversioni tra E, S, TS, mentre Gb si situa regolarmente alla fine della sequenza. Se sostituiamo ai valori di E, S, TS la loro media, abbiamo un ordinamento pressoché totale, con solo 3 inversioni: nei sostantivi e negli aggettivi tra Ta e Tb, e negli avverbi tra Rb e Ga. (Tabella 16) [71].

TABELLA 12.

Metà, Media, Mediana in b.

		METÀ		MEDIA		MEDIANA	
		% sul Totale delle occorrenze	Frequenza media del lemma	% sul Totale delle occorrenze	Frequenza media del lemma	% sul Totale delle occorrenze	Frequenza media del lemma
I Parte	Tb	96,633	46,143	82,545	211,379	0,859	24,06
	Ts	96,158	49,570	80,191	166,666	0,628	25,77
	Rb	95,666	41,086	80,162	165,760	0,815	21,48
	E	95,035	38,295	78,064	158,788	0,530	20,15
	Gb	95,005	39,991	77,023	145,222	0,510	21,04
II Parte	Tb	3,365	1,620	17,455	4,636	99,141	24,06
	Ts	3,842	1,820	19,809	5,828	99,372	25,77
	Rb	4,334	1,861	19,838	4,755	99,385	21,48
	E	4,965	2,013	21,936	4,907	99,470	20,15
	Gb	4,995	2,089	22,977	5,443	99,490	21,04

Se passiamo a Z calcolato per i lemmi troviamo la sequenza

C, Ta, Tb, Ra, Rb, Ga, E, Gb, S, TS,

la quale presenta 7 inversioni che riguardano le solite categorie: due sono tra Ga e E, due tra E e Gb, tre tra S e TS (Tabella 14).

(71) Possiamo fare alcune osservazioni sulla Tabella 16. Innanzitutto, nelle coppie di sottoinsiemi di *a* e *b* aventi la stessa etichetta (T, R, G), il sottoinsieme di *a* si situa a sinistra di quello di *b*. Se accettassimo l'ipotesi che deriva dalle premesse di Antosch, che l'ordine in questione riflette un progressivo allontanamento dalle caratteriche del parlato, potremmo avanzare l'interpretazione che, col passare del tempo, le caratteristiche dei tre sottoinsiemi si siano modificate avvicinandosi a quelle del parlato. Questo fatto è rilevante soprattuto per i giornali: mentre Ga si situa immediatamente a destra dei romanzi e a sinistra di E, S, TS, Gb si situa a destra di questi ultimi. Inoltre, mentre l'ordinamento di C,T, R è costante e rigido, E, S, TS presentano non poche inversioni. Nella prospettiva di Antosh, questo fatto potrebbe essere interpretato come un indice che nei tre sottoinsiemi operano delle tendenze contraddittorie, o tendenze simili ma con diversa intensità. Ancora più semplicemente, si può rilevare che la definizione a priori di E, S, TS non è del tutto chiara.

TABELLA 13.

Valori di X^2 *per la ripartizione delle frequenze delle categorie grammaticali per i sottoinsiemi di a e di b.*

	C	Ta	Tb	Ra	Rb	Ga	E	S	Ts	Gb
C	—	613,993	1244,852	6310,065	11145,811	14676,763	16109,802	18179,802	19319,083	20006,249
Ta	613,993	—	635,242	3572,389	7425,966	10369,345	11644,866	13042,016	14567,272	15044,349
Tb	1244,852	635,242	—	2957,730	5762,045	8968,699	9183,548	11003,662	12067,696	12786,394
Ra	6310,065	3572,389	2957,730	—	1234,573	2665,428	3400,450	4183,161	5064,661	5530,797
Rb	11145,811	7425,966	5762,045	1234,573	—	1198,743	848,290	1180,391	1868,339	2284,064
Ga	14676,763	10369,345	8968,698	2665,428	1198,743	—	1295,715	1143,189	1319,847	1064,314
E	16109,802	11644,866	9183,548	3400,450	848,290	1295,715	—	964,967	725,107	1299,933
S	18179,802	13042,016	11003,662	4183,161	1180,391	1143,189	964,967	—	1188,65	966,931
Ts	19319,083	14567,272	12067,696	5064,661	1868,339	1319,847	725,107	1188,65	—	596,462
Gb	20006,249	15044,349	12786,394	5530,797	2284,064	1064,314	1229,933	966,931	596,462	—

TABELLA 14.

Valori di Z per l'insieme delle frequenze dei lemmi comuni ad ogni coppia di sottoinsiemi di a e di b.

	C	Ta	Tb	Ra	Rb	Ga	E	Gb	S	Ts
C	—	43,130	71,779	133,517	179,629	202,442	211,129	241,776	252,861	269,396
Ta	43,130	—	54,921	103,673	149,461	169,561	176,783	211,715	223,501	241,482
Tb	71,779	54,921	—	96,790	121,071	163,471	151,050	193,176	200,047	226,817
Ra	133,517	103,673	96,790	—	63,101	121,034	114,623	161,165	165,907	189,795
Rb	179,629	149,461	121,071	63,101	—	126,051	90,879	143,737	142,822	176,524
Ga	202,442	169,561	163,471	121,034	126,061	—	88,120	84,555	136,851	131,174
E	211,129	176,783	151,050	114,623	90,879	88,120	—	78,252	109,401	128,986
Gb	241,776	211,713	193,176	161,165	143,737	84,555	78,252	—	121,944	113,152
S	252,861	223,501	200,041	165,907	142,822	136,851	109,401	121,944	—	158,842
Ts	269,396	241,482	226,817	189,795	176,524	131,174	128,986	113,152	158,842	—

TABELLA 15.

Valori dell'indice di correlazione tra i ranghi dei lemmi nelle coppie di sottoinsiemi di a e di b.

	C	Ta	Tb	Ra	Rb	E	Ga	Gb	S	Ts
C	—	0,687	0,638	0,592	0,575	0,516	0,491	0,479	0,434	0,420
Ta	0,678	—	0,640	0,594	0,593	0,549	0,509	0,499	0,425	0,436
Tb	0,638	0,640	—	0,634	0,629	0,541	0,533	0,408	0,534	0,326
Ra	0,592	0,594	0,634	—	0,678	0,570	0,440	0,482	0,437	0,441
Rb	0,575	0,593	0,629	0,678	—	0,522	0,505	0,319	0,541	0,305
E	0,516	0,549	0,541	0,570	0,522	—	0,592	0,552	0,572	0,462
Ga	0,491	0,509	0,533	0,440	0,505	0,592	—	0,631	0,476	0,547
Gb	0,479	0,499	0,408	0,482	0,319	0,552	0,631	—	0,546	0,533
S	0,434	0,425	0,534	0,437	0,541	0,572	0,476	0.546	—	0,487
Ts	0,420	0,436	0,326	0,441	0,305	0,462	0,547	0,533	0,487	—

TABELLA 16.

Frequenze percentuali delle categorie grammaticali nei sottoinsiemi di a e di b.

	C	Ta	Tb	Ra	Rb	Ga	S+E+Ts	Gb	(T+R+G)*a*	(T+R+G)*b*	Tot *a*	Tot *b*
Sost.	13,376	14,848	13,955	16,729	18,199	18,929	19,143	20,412	16,804	17,575	16,820	17,954
Agg.	4,277	4,701	4,435	4,958	6,269	6,701	7,573	7,957	5,437	6,250	5,522	6,888
Art.	8,298	10,159	10,216	13,794	16,030	16,804	18,062	18,357	13,535	14,935	13,580	15,942
Prep.	8,624	9,443	10,541	14,048	15,458	16,753	17,563	18,396	13,354	14,864	12,786	16,347
Verbi	24,038	21,848	21,383	17,713	16,050	14,763	13,515	12,815	18,162	16,678	18,722	15,120
Pron.	11,627	11,435	11,368	8,502	7,045	5,161	5,144	4,289	8,414	7,508	8,481	6,516
Avv.	9,987	8,711	8,136	6,454	4,457	4,869	3,831	3,471	6,707	5,316	6,691	4,844
Cong.	5,989	5,553	7,168	6,428	6,273	5,159	6,096	5,278	5,728	6,221	5,698	6,332
Num.	0,811	0,817	0,391	0,631	0,481	0,998	0,473	0,404	0,813	0,425	0,749	0,440
Inter.	0,773	0,778	0,499	0,131	0,076	0,020	0,017	0,011	0,315	0,191	0,350	0,121
Aus.	2,980	2,645	2,555	2,348	1,603	2,755	1,456	2,687	2,581	2,283	2,457	1,936
Avv.+Prep	0,983	1,054	0,928	1,565	1,771	1,310	1,271	1,090	1,308	1,264	1,227	1,274
Int. + Sost.	0,007	0,018	0,003	0,008	0,001	0,001	0,002	0,0	0,009	0,001	0,007	0,001
Agg.+Pron.	3,026	3,195	3,752	2,966	2,831	2,736	3,005	2,588	2,969	3,047	2,977	3,044
Agg.+Avv.	2,726	2,468	2,464	2,052	1,988	2,097	1,840	1,518	2,208	1,982	2,227	1,945
Int.+Avv.	0,436	0,312	0,271	0,130	0,084	0,078	0,066	0,073	0,175	0,141	0,205	0,113
Pron.+Avv.	1,438	1,302	0,912	0,839	0,425	0,471	0,355	0,226	0,877	0,515	0,905	0,437
Avv.+Cong.	0,694	0,708	1,025	0,702	0,960	0,412	0,587	0,437	0,610	0,802	0,596	0,745

L'indice di correlazione tra i ranghi (Tabella 15) mostra l'ordine C, Ta, Tb, Ra, Rb, E, Ga, Gb, S, TS, con qualche inversione per gli ultimi 5 insiemi di destra.

Questi dati, anche secondo il giudizio di alcuni illustri cultori di statistica linguistica [72], sembrano convalidare l'orientamento attuale a dirigere

(72) Tra i diversi dati disponibili ne ho riportati solo alcuni per esemplificare i fatti dei quali un modello linguistico quantitativo dovrebbe tenere conto.
La intera serie dei dati sarà pubblicata in Zampolli (1974 d).

Sono particolarmente interessanti, per esempio, i risultati del confronto tra i singoli lemmi di *a* e di *b* e dei relativi sottoinsiemi. Il confronto è stato fatto considerando 14 lemmari: *a*, *b*, Ta + Ra + Ga, Tb + Rb + Gb, C, Ta, Ra, Ga, S, Tb, Rb, E, Ts, Gb, i quali sono stati confrontati a due per volta. Data una coppia di lenmari (x , y), sono stati tabulati i lemmi presenti in x e assenti in y e viceversa. Chiamiamo Lx il numero di lemmi presenti in x e assenti in y, e Fx il totale delle loro occorrenze. Chiamiamo Ly il numero di lemmi presenti in y e assenti in x e Fy il totale delle loro occorrenze. Chiamiamo Lx% la percentuale di Lx sul totale dei lemmi di x, Fx% la percentuale di Fy sul totale delle occorrenze di x, e analogamente Ly% e Fy%. Chiamiamo inoltre LC il numero dei lemmi comuni, FC il totale delle occorrenze comuni, e LC% e FC% le rispettive percentuali. Per tutte le coppie di sottoinsiemi di *a*, Lx, Fx, Ly, Fy, Lx%, Fx%, Ly%, Fy% crescono secondo il solito ordine C, Ta, Ra, Ga, S, mentre decrescono secondo lo stesso ordine LC, FC, LC%, FC%. Lo stesso avviene in *b* secondo l'ordine Tb, Rb, E, Gb, Ts. Se consideriamo invece le coppie formate dall'unione dei 10 sottoinsiemi, l'ordine C, Ta, Tb, Ra, Rb, ecc. non è rispettato, perché i lemmi comuni tra due sottoinsiemi dello stesso dizionario sono quasi sempre più numerosi dei lemmi comuni a due sottoinsiemi provenienti da dizionari diversi. Questo fatto non sorprende perché nella scelta dei lemmi da includere in un dizionario è intervenuta la dispersione D, che, come si è detto, tiene conto della ripartizione nei 5 sottoinsiemi del corpus.

Però se si considerano un sottoinsieme x di un dizionario e i 5 sottoinsiemi dell'altro, i valori Lx, Fx, Ly, ecc. rispettano l'ordinamento rispetto ad x. Per esempio, rispetto a C, l'ordine è Tb, Rb, E, Gb, Ts. Se consideriamo la distribuzione per categorie grammaticali di lemmi e occorrenze di *a* assenti in *b* e viceversa, troviamo che il contributo maggiore è dato da sostantivi, verbi, aggettivi e avverbi, nell'ordine. Abbiamo analizzato anche i singoli lessemi dei due lemmari. Per esempio abbiamo ordinato tutti i lemmi comuni ai due sottoinsiemi di una coppia secondo il valore decrescente di

$$* \mathrm{Sa}$$

dove, se

Oa è la frequenza del lemma in un sottoinsieme di *a*
Ob è la frequenza del lemma in un sottoinsieme di *b*
t è la frequenza teorica del lemma nella coppia:

$$t = \frac{\mathrm{Oa} + \mathrm{Ob}}{2}$$

allora

$$\mathrm{Sa} = \frac{(\mathrm{Oa} - t)^2}{t}$$

e l'asterisco è il segno (positivo o negativo) della differenza

$$(\mathrm{Oa} - t)$$

Si noti che, se

$$\mathrm{Sb} = \frac{(\mathrm{Ob} - t)^2}{t}$$

le ricerche verso la scoperta e la definizione di caratteristiche misurabili che permettano di costruire, su basi induttive, un modello stratificato della lingua come popolazione statistica.

allora

$$Sa = Sb$$

ma

$$*Sa \text{ è di segno opposto a } *Sb.$$

Pertanto l'ordinamento secondo *Sa è esattamente l'inverso dell'ordinamento secondo * Sb.

| SA | (che è uguale ad | SB | ed è uguale a zero se Oa = Ob = t viene usato come misura del grado di ' differenza ' tra le frequenze nei due sottoinsiemi.

Questo ordinamento permette alcune considerazioni estremamente interessanti, per le quali, non avendo il tempo di esporle qui, rimando a Zampolli (1974 d).

Desidero ringraziare Ch. Muller per aver verificato la parte statistica della ricerca.

Bibliografia

Abraham S., Kiefer F. (1965) – *An algorithmic definition of the morpheme*, in « Statistical Methods in Linguistics », IV, 4–9.

Accademia della Crusca (1967) – *Norme per la schedatura lessicografica*, Firenze.

Accademia della Crusca (1967) – *Concordanze degli Inni Sacri di A. Manzoni*, Firenze.

Accademia della Crusca (1968) – *Novella del Grasso Legnaiuolo. Testo. Frequenze. Concordanze*, Firenze.

Accademia della Crusca (1973) – *Tavola rotonda sui grandi lessici storici*, Firenze.

Aitken (1972) – *Historical dictionaires, word frequency distributions, and the computer*, in A. Zanpolli (1974 b).

Akademija Nauk SSSR (1968) – Naučnyj Sovet po Kibernetike, Sekcija Semiotiki, *Statistika reči*, Leningrad, e traduzione italiana: Accademia delle Scienze dell'URSS. Consiglio scientifico per la Cibernetica, Sezione di Semiotica, *Statistica linguistica*, con l'aggiunta di due appendici. Traduzione dal russo di M. Cârstea, Bologna, 1971.

Akhmanova O. S. (1969) – *Linguistics and the quantitative approach*, in *Actes du X.e Congrès international des linguistes*, Bucarest.

Akhmanova O. S., Mel'chuk I. A., Frumkina R. M., Paducheva E. V. (1963) – *Exact methods in linguistic research*, translated from Russian by D. G. Hays and D. V. Mohr, Santa Monica (California).

Aliprandi G. (1940) – *Frequenze dattilografiche*, in « Bollettino dell'Accademia Italiana di studi stenografici di Padova », XV.

Allen J. (1973) – *Speech Synthesis from Unrestricted Text*, Report, Research Laboratory of Electronics, MIT, Cambridge (Mass.).

Allen S. (1971) – *Frequency Dictionary of Present–Day Swedish*, Stockholm.

Antosch F. (1969) – *The Diagnosis of Literary Style with the Verb–Adjective Ratio*, in Doležel, Bailey (1969), 57–65.

Apostel L., Mandelbrot B., Morf A. (1957) – *Logique, langage, et théorie de l'information*, Paris.

Astrahan M. (1970) – *Speech Analysis by clustering or the hyperphoneme method*, AI Memo 124, Stanford (Calif.).

Atal B. S. (1971) – *Speech Analysis and Synthesis by Linear Prediction of the Speech Wave*, in IEEE Trans. on Computers.

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

D'Arco Silvio Avalle (1973) – *Projet pour une liste des concordances de la langue poétique en Italie avant la fin du XIII.e siècle*, in A. Zampolli (1973 a), 19–27.

Bach E. (1968) – *Nouns and Noun Phrases*, in Bach E., Harms R. T. (1968), 91–122.

Bach E., Harms R. T. (1968) (eds.) – *Universals in Linguistic Theory*. Holt, Rinehart and Winston, Inc.

Bailey R. W. (1969) – *Statistics and Style: A Historical Survey*, in Doležel, Bailey (1969), 217–236.

Bailey R. W. (1972) – *Interactive Lexicography*, in Zampolli (1974 b).

Bailey R. W., Dolezel L. (1968) – *An annotated bibliography of statistical stylistic*, Ann Arbor (Mich.).

Bahr J. (1972) – *The lexical Data Bank*, in A. Zampolli (1974 b).

Baronofsky S. (1970) – *Some heuristics for automatic detection and resolution of anaphora in discourse*, MA Thesis, Austin (Texas).

Bartoletti Colombo A. M. (1973) – *Per un vocabolario delle Costituzioni di Giustiniano. Premessa e saggio d'elaborazione*, Firenze.

BATORI I. (1974) – *Working with the Interactive Version of the T.G.T. System of Joyce Friedman*, in A. Zampolli (1974 a).

BELARDI W. (1959–60) – *Elementi di Fonologia generale*, Roma.

BELEVITCH V. – *Langage des machines et langage humain*, Bruxelles.

BLOOMFIELD L. (1933) – *Language*, New York.

BOBROW D. G., FRASER I. B. (1969) – *An Augmented State Transition Network Analysis Procedure*, in L. M. Norton, D. Walker (eds.), *Proceedings of the International Joint Conference on Artificial Intelligence*, Bedford (Mass.), 557–567.

BOBROW D. G., KLATT D. H. (1968) – *A Limited Speech Recognition System*, in Proc. AFIPS Fall Joint Computer Conference, Thompson, Washington, D. C., *33*, 305–318.

BOCCA E., PELLEGRINI A. (1950) – *Studio statistico sulla composizione fonetica della lingua italiana e sua applicazione pratica all'audiometria con la parola*, in « Archivio italiano di Otologia, Rinologia e Laringologia », LVI, suppl. n. 5, 116–141.

BODER D. P. (1940) – *The Adjective–Verb quotient; a contribution to the psychology of language*, in « Psych. Rec. », 3.

BOISVERT S., DUGAS A., BELANGER D. (1974) – *Obling: a Tester for Transformational Grammars*, in A. Zampolli (1974 a).

BOLDRINI M. (1958) – *Statistica, teoria e metodi*, Milano.

BOLELLI T. (1971) – *Linguistica Generale, Strutturalismo, Linguistica Storica*, Pisa.

BOLINGER D. (1965) – *The atomization of meaning*, in « Language », *41*, 4.

BOLOCAN G. M. (1961) – *Unele caracteristici ale stilului publicistic al limbii romane literare.* in « Studi si cercetări lingvistice », XII, 35–71.

BOOTH A. D., CLEAVE J. P., BRANDWOOD L. (1958) – *Mechanical Resolutions of Linguistic Problems*, London.

BORKO H. (1967) (Ed.) – *Automated Language Processing*, New York.

BORTOLINI U., TAGLIAVINI C., ZAMPOLLI A. (1971) – *Lessico di frequenza della lingua italiana contemporanea*, IBM Italia.

BOTHA R. P. (1968) – *The function of the lexicon in transformational generative grammar*, The Hague.

BRANDWOOD L. (1969) – *Plato's Seventh Letter*, in « Revue », IV, 1–25.

BROAD D. J. (1972) – *Basic directions in automatic speech recognition*, in « Int. Journal of man–machine studies », *4*, 105–118.

BUCCELLATI G. (1974) – *The Old Babylonian Linguistic Analysis Project: Goals, Procedures and First Results*, in A. Zampolli (1974 a).

BUCHANAN M. A. (1972) – *A graded Spanish word book*, Toronto (Ont.).

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

BUSA R. (1968) – *De lexico electronico latino*, Pisa.

BUSA R., ZAMPOLLI A. (1968) – *Centre pour l'Automation de L'Analyse Linguistique (C.A.A.L.), Gallarate*, in *Les machines dans la linguistique*, Prague, 25–34.

BUSEMANN A. (1925) – *Die Sprache der Jugende als Ausdruck der Entwicklungsrhythmik*, Jena.

BUSEMANN A. (1948) – *Stil und Charakter: Untersuchungen zur Psycologie der individuellen Redeform*, Meisenheim–Glan.

CAMILLI A. (1921) – *An Italian Phonetic Reader*, London.

CARROLL J. B. (1970) – *An alternative to Juilland's usage coefficient for lexical frequencies, and a proposal for a standard frequency index (SFI)*, in « Computer studies in the humanities and verbal behavior », III, *2*, 61–65.

CARROL J. B. (1973) – *Towards a Performance Grammar for Core Sentences in Spoken and Written English*, Princeton.

CASARES J. (1941) – *Nuevo concepto del Diccionario de la lengua y otros problemas de lexicografía y gramática*, Madrid.

CASARES J. (1950) – *Introducción a la lexicografía moderna*, Madrid.

CASTROGIOVANNI P., CERRI S. A., MAFFEI G., PASQUINUCCI P. J., TORRIGIANI G., ZAMPOLLI A. (1968) – *Analisi Linguistica delle risposte al test di Rorschach di schizofrenici e neurotici e dei rispettivi familiari*, in « Neopsichiatria », XXX, IV (1968), 810–837.

CASTROGIOVANNI P., TELARA A. (1973) – *Primi risultati di un'analisi statistica morfologica e lessicale delle risposte al test di Rorschach nelle prospettive di uno studio dei rapporti tra psicopatalogia e linguaggio*, in Zampolli (1973 a), 307–323.

CELCE–MURCIA M. (1972 a) – *Paradigms for Sentence Recognition*, Los Angeles (prepr.).

CELCE–MURCIA M. (1972 b) – *English Comparatives*, Ph. D. Thesis, UCLA, Los Angeles.

CHAFE W. L. (1970) – *Meaning and the Structure of Language*, Chicago.

CHOMSKY N. (1957) – *Syntactic Structures*, The Hague, trad. it. Bari, 1970.

CHOMSKY N. (1958) – Recensione a Belevitch (1956).

CHOMSKY N., MILLER G. A. (1963) – *Finitary Models of Language Users*, in R. D. Luce, R. R. Bush, E. Galanter, *Handbook of Mathematical Psychology*, New York. Trad.it in Chomsky (1969), *L'analisi formale del linguaggio*, Torino.

CHOMSKY N. (1965) – *Aspects of the Theory of Syntax*, Cambridge (Mass.).

CHOMSKY N. (1966) – *Topics in the Theory of Generative Grammar*, The Hague.

CHOMSKY N. (1971) – *Deep Structure, surface structure, and semantic interpretation*, in D. D. Steinberg, L. A. Jakobovits (1971), 183–216.

CIAMPI C. (1973) – *Les projets de recherche automatique des informations juridiques dans l'Institut pour la documentation juridique du Conseil National des Recherches*, in Zampolli (1973 a), 249–268.

CIRESE A. M. – *Inventaires et répertoires lexicaux, formulaires et métriques des chants populaires italiens*, in Zampolli (1973 a), 209–239.

COHEN M. (1935) – Recensione a ZIPF (1935) , in « BSL », XXXVI, 8.

COHEN M. (1967) – *Sur l'histoire de la statistique en linguistique*, in « Etudes de linguistique appliquée », 5, 3–8.

COLBY K. M., SCHANK R. (1973) (Eds.) – *Computer models of thought and language*, San Francisco (California).

COLES L. S. (1969) – *Talking with a robot in English*, in D. E. Walker, L. M. Norton (1969) (eds.), *Proceedings of the International Joint Conference on Artificial Intelligence*, Boston.

COLES S. L. (1971) – *Techniques for Information Retrieval Using an inferential Question–Answering System with natural Language Input*, Technical Note, ARI, Menlo Park (Calif.).

COLLINS A., QUILLIAN R. M. (1969) – *Retrieval Time from Semantic Memory*, in « J. Verb. Learn. and Verb. Behav. », 8, 240–47.

COLMERAUER A., KITTREDGE R., STEWART G. (1970) – *Résultats Préliminaires. Project de Traduction Automatique*, Université de Montreal.

CONWAY M. E. (1963) – *Design of a Separable Transition–Diagram Compiler*, in « CACM », VI, 7, 396–408.

CORBEIL J. (1971) – *Les Structures syntaxiques du français moderne*, Paris.

DENOOZ J. (1968) – *Programmes d'ordinateur pour la détermination et le traitement statistique des fréquences de phonèmes*, in « Revue », III, 41–53.

DEL CAMPO J., GONZALES I., MOLINA M. T., MARCOS F. (1974) – *The Automatic Syntactic Analysis as an Aid in Dictionary Making*, in Zampolli (1974 a).

DE MAURO T. (1970 A) – *Storia linguistica dell'Italia unita*, Nuova edizione riveduta, aggiornata e ampliata, Bari.

DE MAURO T. (1970 b) – *Introduzione alla Semantica*, Bari.

DE MAURO T., POLICARPI G. (1971) – *Ricerche sulla struttura del periodo italiano*, in SLI, *L'insegnamento dell'italiano in Italia e all'estero*, Roma, 583–694.

Dictionnaire de fréquences. Vocabulaire littéraire des XIX et XX.e siècles, Préf. par P. Imbs Paris, 1971.

DIMITRESCU F. (1973) – *Projet d'un dictionnaire de la langue roumaine du XVI.e siècle*, in Zampolli (1973 a), 41–48.

DI PIETRO R. J. (1965) – *The Phonemic Status of Juncture in Italian*, in *Proceedings of the Fifth International Congress of Phonetic Sciences*, Basel–New York, 261–263.

DI SPARTI A. (1973) – *Analisi sintattica automatica delle lingue naturali*, Palermo.

DOLEŽEL L. (1969) – *A Framework for the statistical Analysis of Style*, in Doležel, Bailey (1969), 10–25.

DOLEŽEL L., BAILEY R. W. (1969) (Eds.) – *Statistics and Style*, New York.

DOP (1969) – *Dizionario d'ortografia e di pronunzia*, Redatto da B. Migliorini, C. Tagliavini, P. Fiorelli, Torino.

DUBOIS J. (1962) – *Étude sur la dérivation suffixale en français moderne et contemporain*, Paris.

DUBOIS J. (1966) – *Utilisation des statistiques lexicographiques pour l'étude structurale du lexique*, in *Statistique et analyse linguistique*, Paris, 95–98.

DUBOIS C. e J. (1971) – *Introduction à la lexicographie: le dictionnaire*, Paris.

DURO A., ZAMPOLLI A. (1968) – *Analisi lessicali mediante elaboratori elettronici*, in *Atti del Convegno sul Tema: L'elaborazione elettronica...* (1968).

DURO A. (1973) – *Elaborations électroniques de textes effectuées par l'Accademia della Crusca, pour la préparation du Dictionnaire Historique de la langue italienne*, in Zampolli (1973 a), 53–75.

DYER R. (1973) – *The measurement of individual style*, in Zampolli (1973 a), 325–348.

EARLEY J. (1970) – *An efficient Context–free parsing algorithm*, in «CACM», XIII, 2, 94–102.

EBELING C. L. (1960) – *Linguistic units*, 's–Gravenhage.

ENGWALL G. (1972) – *A. Juilland, B. Brodin, C. Davidovitch. Frequency dictionary of French words* (recensione), in «Studia Neophilologica», *44*, 2, 445–462.

ENGWALL G. (1974) – *Fréquence et distribution*, Stockholm.

ESTOUP J. B. (1912–1916) – *Gammes sténographiques*. Recueil de textes choisis pour l'acquisition méthodique de la vitesse, précédé d'une introduction, Paris (la prima edizione è del 1907).

FAEDO A. (1973) – *Discorso di Apertura*, in Zampolli (1973 a), VII–IX.

FEIGENBAUM E. A. (1969) – *Artificial Intelligence: Themes in the second decade*, in *Proceedings of the International Conference on Information Processing*, New York.

FILLMORE C. J. (1968) – *The Case for Case*, in E. Each, R. T. Harms (1968), 1–88.

FRANCIS W. N., RUBIN G. M., SVARTIVIK J. (1969) – *A method of computer–produced graphical representation of dialectal variation in initial fricatives in Southern British English*, Preprint for the 3th ICCL.

FRIEDMAN J. (1969) – *Applications of a Computer System for Transformational Grammar*, Preprint for the 3th ICCL.

FRY D. B. (1973) – *Speech and Language*, preprint, Padova.

GARCIA HOZ V. (1953) – *Vocabulario usual, común y fundamental*, Madrid.

GARDIN J. (1965) – *A Typology of Computer Uses in Anthropology*, in D. Hymes (ed.), *The Use of Computers in Anthropology*, The Hague, 105–117.

GARVIN P. L. (1962) – *Computer Participation in Linguistic Research*, in «Language», XXXVIII, 4, 385–389.

GODEL R. (1948) – *Homonymie et identité*, in «Cahiers F. de Saussure», VII, 5–15.

GOLDMAN N. M. (1974) – *Sentence Paraphrasing from a Conceptual Base*, in A. Zampolli (1974 a).

GOLOPENTIA–ERETESCU S. (1966) – *La structure phonologique des monosyllabes roumains*, in «Cahiers de Linguistique théorique et appliquée», III, 59–67.

GOUGENHEIM G., MICHEA R., RIVENC P., SAUVAGEOT A. (1964) – *L'élaboration du français fondamental*, Paris.

GOUGENHEIM G., RIVENC P., M.me HASSAN (1964) – *Le français fondamental*, in «Tendances nouvelles en matière de recherche linguistique», Strasbourg, 53–72.

GRASSI C. (1973) – *Perspectives de l'emploi de l'élaborateur électronique en géographie linguistique et en dialectologie*, in Zampolli (1973 a), 233–239.

GRICE H. P. (1968) – *Utterer's meaning, sentence–meaning and word meaning*, in « Foundations of Language », IV, 225–242.
GRICE H. P. (1957) – *Meaning*, in « Philosophical Review », LXVI, 377–388.
GRIFFITHS T., PETRICK S. (1965) – *On the relative Efficiencies of Context–free Grammar Recognizers*, in « CACM », VIII, 5, 289–300.
GROSS M. (1972) – *Mathematical Models in Linguistics*, Prentice–Hall.
GUILBERT L. (1967) – *Le dictionnaire du français contemporaine*, in « Cahiers de Lexicologie », X, 1, 115–119.
GUIRAUD P. (1954 a) – *Language et communication. Le substrat informationnel de la sémantisation*, in « Bulletin de la Societé de Linguistique de Paris », L, 119–133.
GUIRAUD P. (1954 b) – *Les caractères statistiques du vocabulaire*, Paris.
GUIRAUD P. (1960) – *Problèmes et méthodes de la statistique linguistique*, Paris.
GUIRAUD P. (1967) – *Structures étymologiques du Lexique Français*, Paris.
GUIRAUD P., WHATMOUGH J. (1954) – *Bibliographie de la statistique linguistique*, Utrecht.
HALLE M., Recensione a HERDAN (1956) – in « Kratylos », III, I, 20–28.
HALLIDAY M. A. K. (1959) – Recensione a HERDAN (1956), in « Archivium Linguisticum », XI, 2, pp. 148–154.
HALLIDAY M. A. K. (1966) – *The English verbal group: A specimen of a manual of analysis*, London.
HALLIDAY M. A. K. (1970) – *Functional diversity in language as seen from a consideration of modality and mood in English*, in « Foundations of Language », VI, 322–361.
HARKIN D. (1957) – *The history of word counts*, in « Babel », III, 113–124.
HARRIS Z. S. (1957) – *Co–occurrence and Transformations in Linguistic Structure*, in « Language », XXXIII, 283–340.
HARRIS Z. S. (1959) – *Computable Syntactic Analysis: The 1959 Computer Science Analyzer*, in « TDAP » 15.
HARRIS Z. S. (1961) – *String Analysis of Sentence Structure*, The Hague.
HARRIS Z. S. (1968) – *Mathematical Structures of Language*, New York.
HARRIS Z. S. (1970) – *Two Systems of Grammar: Report and Paraphrase*, in Z. S. Harris, *Papers in Structural and Transformational Linguistics*, Dordrecht.
HAYES C. W. (1968) – *A Transformational–Generative Approach to Style: Samuel Johnson and Edward Gibbon*, in « Language and Style », I.
HAYES C. W. (1969) – *A Study in Prose Styles: Edward Gibbon and Ernest Hemingway*, in L. Doležel, R. W. Bailey (Eds.) (1969), 80–94.
HAYS D. G. (Ed.) (1966) – *Readings in Automatic Language Processing*, New York.
HAYS D. G. (1967) – *Introduction to Computational Linguistics*, New York.
HAYS D. G. (1969) – *Computational Linguistics: Introduction*, in *Encyclopaedia of Linguistics, Information and Control*, Pergamon Press, 49–51.
HEGER K. (1969) – *La sémantique et la dichotomie de langue et parole. Nouvelles contributions à la discussion sur les bases théoriques de la sémasiologie et de la onomasiologie*, in « Travaux de Linguistique et Littérature », VII, 47–111.
HEIDORN G. E. (1972) – *Natural Language Inputs to a Simulation Programming System*, Naval Postgraduate School (NPS–55 MD 72101 A), Montery (Calif.).
HEILMANN L. (1962–1963) – *Considerazioni statistico–matematiche e contenuto semantico*, in « Quaderni dell'Istituto di Glottologia » (Università di Bologna), VII, 34–45.
HENISZ–DOSTERT B., THOMPSON F. B. (1974) – *The REL system and REL English*, in Zampolli (1974 a).
HENNON V. A. C. (1924) – *A French word book based on a count of 4000.00 running words*, Madison.
HERDAN G. (1956) – *Language as Choice and Chance*, Groningen.
HERDAN G. (1960) – *Type–token mathematics. A textbook of mathematical linguistics*, The Hague.
HERDAN G. (1964 a) – *Quantitative Linguistics*, London.

HERDAN G. (1964 b) – *Quantitative Linguistics or Generative Grammar*, in « Linguistics », 4, 56–65.

HERDAN G. (1966) – *The advanced Theory of Language as Choice and Chance*, Berlin–Heidelberg–New York.

HEWITT C. (1971) – *Procedural embedding of Knowledge in PLANNER*, in *Proc. of the Second Joint Conference on Artificial Intelligence*, London, 167–182.

HJELMSLEV L. (1953) – *Prolegomena to a Theory of Language*, Bloomington (Ind.).

HJELMSLEV L. (1957) – *Pour une sémantique structurale*, in « Travaux du Cercle Linguistique de Copenhague », XII, 96–112.

HOCHETT C. F. (1955) – *A manual of Phonology*, Bloomington (Ind.).

HOCKETT C. F. (1966) – *The Quantification of Functional Load: A Linguistic Problem*, Santa Monica (California).

HOCKETT C. F. (1967) – *Language, Mathematics, and Linguistics*, The Hague.

HOUSEHOLDER F. W., SOPORTA S. (1967) – *Problems in Lexicography*, Bloomington (Ind.).

HOWE W. G., KRULEE G. K. (1971) – *A logic of English Questions*, RC 3490, IBM, Yorktown Heights.

JACKENDOFF R. S. (1972) – *Semantic Interpretation in Generative Grammar*, Cambridge (Mass.).

JONES L. V., WEPMAN J. M. (1966) – *A Spoken Word Count*, Chicago.

JOSSELSON H. H. (1953) – *The russian word count and frequency analysis*, Detroit.

JOSSELSON H. H. (1953) – *The Lexicon: A System of Matrices of Lexical Units and their Properties*, Preprint for the 3th ICCL.

JUILLAND A., CHANG–RODRIGUEZ E. (1964) – *Frequency Dictionary of Spanish Words*, The Hague.

JUILLAND A., EDWARDS P. M. H., JUILLAND I. (1965) – *Frequency Dictionary of Rumanian Words*, The Hague.

JUILLAND A., BRODIN D., DAVIDOVITCH C. (1970) – *Frequency Dictionary of French Words*, The Hague.

JUILLAND A., TRAVERSA V. (1973) – *Frequency Dictionary of Italian Words*, The Hague.

KAEDING E. W. (1898) – *Häufigkeitswörterbuch der deutschen Sprache*, Steglitz.

KAPLAN R. M. (1970) – *The MIND–System: A Grammar–Rule Language*, Rand Corporation, Santa Monica (Calif.) (RM–6265/1–PR).

KAPLAN R. M. (1971) – *Augmented Transition Networks as Psychological Models of Sentence Comprehension*, in *Second International Joint Conference on Artificial Intelligence*, London, 429–443.

KATZ J. J. (1972) – *Semantic Theory*, Harper and Row, Publishers.

KATZ J. J., FODOR L. A. (1963) – *The structure of a semantic theory*, in « Language », XXXIX, 2, 170–210.

KATZ J. J., POSTAL P. M. (1964) – *An Integrated Theory of Linguistic Descriptions*, Cambridge (Mass.).

KAY M. (1967 a), *Standards for Encoding Data in a Natural Language*, in « Computers and the Humanities », I, 5, 170–177.

KAY M. (1967 b) – *Experiments with a Powerful Parser*, RM–5452 PR, Rand Corporation, Santa Monica (Calif.).

KAY M. (1974) – *Automatic Grammatical analysis*, in Zampolli (1974 b).

KASMER A. (1972) – *The Book of Isaiah: Characterization of Authors by Morphological Data Processing*, in « Revue », III, 1–62.

KENNISTON H. (1933) – *A basic list of spanish words and idioms*, Chicago.

KERCKHOFFS A. (1883) – *La cryptographie militaire*, Paris.

KIEFER E. (1964) – *Some aspects of mathematical models in linguistics*, in « Statistical Methods in Linguistics », III, 8–26.

KIEFER F. (1968) – *Mathematical linguistics in Eastern Europe*, New York.

KIMBALL J. (1973) – *Formal Theory of Grammar*, Prentice–Hall.

KNEASE T. M. (1933) – *An Italian Word List from Literary Sources*, Toronto.

KNUTH D. E. (1968) – *Semantics of Context–free Languages*, in «Mathematical Systems Theory», II, *2*, 127–145.

KÖSTER P. (1970) – *Words and numbers. A quantitative approach to Swift and some understrappers*, in «Computers and the humanities», IV, 5, 289–304.

KOLLMANN E, D. (1973) – *Word Frequencies in Latin Literature*, in «Revue», IV, 1–18.

KRÁMSKÝ J. (1964) – *A Quantitative Phonemic Analysis of Italian Mono–, Di–, and Trisyllabic Words*, in «Travaux Linguistiques de Prague», I, 129–143.

KRÁMSKÝ J. (1969) – *The word as a linguistic unit*, The Hague.

KUČERA H., FRANCIS W. N. (1967) – *Computational Analysis of Present–Day American English*, Providence (Rhode Island).

KUČERA E., MONROE G. K. (1968) – *A Comparative Quantitative Phonology of Russian, Czech and German*, New York.

KUNO S. (1963) – *The Current Grammar for the Multiple–Path English Analyzer*, in *Mathematical Linguistics and Automatic Translation*, Report NSF 8, Cambridge (Mass.).

KUNO S. (1965 a) – *The Predictive Analyzer and a Path Elimination Technique*, in «CACM», VIII, 7, 453–462.

KUNO S. (1965 b) – *A System for Transformational Analysis*, in *Mathematical Linguistics and Automatic Translation*, Report NSF 15, Cambridge (Mass.).

KUNO S. (1967 a) – *A Context–Sensitive Recognition Procedure*, in *Mathematical Linguistics and Automatic Translation*, Report NSF 18, Cambridge (Mass.).

KUNO S. (1967 b) – *Computer Analysis of Natural Languages*, in *Proceedings of Symposium in Applied Mathematics*, vol. XIX, Providence, 52–110.

KUNO S., OETTINGER A. G. (1963) – *Multiple–Path Syntactic Analyzer*, in Popplewell (ed.), *Information Processing 1962*, Amsterdam, 306–312.

LAKOFF G. (1971 a) – *On Generative Semantics*, in D. D. Steinberg, L. A. Jakobovits (1971), 232–296.

LAKOFF G. (1971 b) – *Presupposition and relative well–formedness*, in D. D. Steinberg, L. A. Jakobovits (1971), 329–340.

LAMB S. M. (1961) – *The digital Computer as an Aid in Linguistics*, in «Language», XXXVII, 3, 382–412.

LAMB S. M. (1965) – *Linguistic Data Processing*, in D. Hymes (ed.), *The Use of Computers in Anthropology*, The Hague, 159–188.

LAMB S. M. (1966) – *Outline of Stratificational Grammar*, Washington, D. C.

Le applicazioni dei Calcolatori alle Scienze morali e alla Letteratura (1962), Milano.

LEHMANN W. P. (1973) – *On the Design of a Central Archive for Lexicography in English*, in McDavid, Duckert (1973), 312–317.

LEPSCHY G. C. (1964) – *Note sulla fonematica italiana*, in «Italia Dialettale», XXVII, 53–67.

LEPSCHY G. C. (1966) – *La linguistica strutturale*, Torino.

Les Machines dans la linguistique (1968), Prague.

LEVISON M., MORTON A. Q., WINSPEAR A. D. (1968) – *The seventh Letter of Plato*, in «Mind», LXXVII, n. 307, 309–325.

Lexicologie et Lexicographie françaises et romanes (1961), CNRS.

Literary Data Processing (1964), Yorktown Heights (N. Y.).

LOCKE W. N., BOOTH A. D. (1955) (eds.) – *Machine Translation of Languages*, New York and London.

LYNE A. A. (1972) – *The problem of non–comparability of word–frequency counts. A proposal for a standard decision–procedure to be followed in lemmatizing French*, Comunicazione presentata al III Congresso Internazionale AILA.

LYNE A. A. (1973) – *L'élaboration des listes de fréquence*, in «Cahiers de lexicologie», II, 23, 83–108.

LYONS J., WALES R. J. (1966) (eds.) – *Psycholinguistics Papers*, Edinburgh.

LYONS J. (1968) – *Introduction to Theoretical Linguistics*, Cambridge. Trad. it., *Introduzione alla Linguistica Teorica*, Bari, 1971.

MACLAY H. (1971) – *Overview*, in D. D. Steinberg, L. A. Jakobovits (1971), 157–182.
MANDELBROT B. (1954) – *Structure formelle des textes et communication*, Deux études, in « Word », X, 1, 1–27.
MAGNO CALDOGNETTO (1973) – *Scopi, metodi e applicazioni della fonetica sperimentale*, in «Miscellanea II », Udine.
MARCKWORT M. L., BELL L. M. (1967) – *Sentence–Lenght Distribution in the Corpus*, in H. Kučera, W. Nelson Francis (1967).
MARKOV A. A. (1961) – *Theory of Algorithms*, IMEI, Accademia delle Scienze dell'URSS vol. 42. (Tradotto dell'Israel Program for Scientific Translations).
MARTINET A. (1960) – *Eléments de linguistique générale*, Paris.
MATORÈ G. (1953) – *La méthode in lexicologie*, Paris.
MAXWELL A. E. (1967) – *Analysing Qualitative Data*, London.
MCCAWLEY J. D. (1968) – *The Role the Semantics in a Grammar*, in E. Bach, R. T. Harms (1968), 125–169.
MCDAVID R. I., DUCKERT A. R. (1973) (eds.) – *Lexicography in English*, New York.
MELBY A. K. (1974) – *Junction Grammar and Machine Assisted Translation*, in Zampoll (1974 a).
MENARD N. (1972) – *Mesure de la richesse lexicale. Méthodologie et vérifications expérimentales* (Thèse), Strasbourg (datt.).
MERIGGI P. (1973) – *Un lexique de l'hittite cuneiforme*, in Zampolli (1973 a), 111–113.
MICHAELSON S., MORTON A. Q. (1972) – *The Spaces in between: A multiple test of authorship for greek writers*, in « Revue », I, 23–77.
MICHÉA R. (1964) – *Les vocabulaires fondamentaux*, in « Recherches et techniques nouvelles au service de l'enseignement des langues vivantes », Strasbourg, 21–38.
MICHÉA R. (1967) – *La relation rang–fréquence et la structure statistique de la langue parlée*, in « Bulletin de la Societé de Linguistique de Paris », LXII, 9–14.
MICHÉA R. (1973) – *Notes et réflexions sur la distribution des fréquences lexicales*, in « Cahiers de Lexicologie », XXII, 1, 69–95,
MICHIE D. (1968) (ed.) – *Machine Intelligence 3*, New York.
MIGLIORINI B. (1941) – *La lingua del 900*, Firenze.
MIGLIORINI B. (1951) – *Che cos'é un vocabolario?*, Firenze.
MIGLIORINI B. (1960) – *Storia della lingua italiana*, Firenze.
MIGLIORINI B. (1963) – *La lingua contemporanea*, Firenze.
MIGLIORINI B., BALDELLI I. (1964) – *Breve storia della lingua italiana*, Firenze.
MILLER G. A. (1951) – *Language and Communication*, New York. Trad. it. *Linguaggio e Comunicazione*, Firenze, 1972.
MILLER G. A., HEISE G. A., LICHTEN W. (1951) – *The Intelligibility of Speech as a Function of the Context of Speech Materials*, in « J. Experimental Psychology », *41*, 329–355.
MILLER G. A., ISARD S. – *Some Perceptual Consequences of Linguistic Rules*, in « J. Verb. Learn. and Verb. Behavior », *2*, 217–228.
MINSKY N. (1968) (ed.) – *Semantic Information Processing*, Cambridge (Mass.).
MISTRIK J. (1969) – *Frekvencia slov v Slovenčine*, Bratislava.
MOLINA E. (1916) – *Le gamme stenografiche*, in « Bollettino stenografico italiano », XV, 4–8; 13–19; 55–64.
MOREAU R. (1962) – *Au suijet de l'utilisation de la notion de fréquence en linguistique*, in «Cahiers de Lexicologie », III, 140–158.
MOREAU R. (1966) – *Intervention de M.R. Moreau*, in *Statistique et analyse linguistique*, Paris 125–132.
MOUNIN G. (1964) – *La machine à traduire*, The Hague.
MOYNE J. A. (1972) – *Some Grammars and Recognizers for Formal and Natural Languages*, in Julius T. Tou, (Ed.), *Advances in Information System Science*, vol. 5, New York.
MULJAČIĆ Ž. (1969) – *Fonologia generale e fonologia della lingua italiana*, Bologna.

MULLER Ch. (1963) – *Le mot, unité de texte et unité de lexique en statistique lexicologique*, in « Travaux de Linguistique et de Littérature », 155–173.

MULLER Ch. (1965) – *Fréquence, dispersion et usage. À propos des dictionnaires de fréquence*, in « Cahiers de Lexicologie », 7, 2, 32–42.

MULLER Ch. (1967) – *Étude de statistique lexicale: Le vocabulaire du théâtre de Corneille*, Paris.

MULLER Ch. (1968) – *Initiation à la statistique linguistique*, Paris.

MULLER Ch. (1973 a) – *Éléments de statistique linguistique*, in Zampolli (1973 a), 348–378.

MULLER Ch. (1973 b) – *Initiation aux méthodes de la statistique linguistique*, Paris.

NEWELL A., BARNETT J., FORGIE J. W., GREEN C., KLATT D., LICKLIDER J. C. R., MUNSON J., REDDY D. R., WOODS W. A. (1973) – *Speech Understanding Systems*, Amsterdam.

NEWELL A., SIMON H. A. (1971) – *Human Problem Solving*, Englewood Cliffs (N. J.).

NOVAK L. A. (1963) – *Nekotorye vosprosy lingvostatistiki i častotnye slovari*. (Alcuni problemi di statistica linguistica e i vocabolari di frequenza), in « Ministerstvo Prosveščenija MSSR, Belčkij gosudarstvennyj Pedagogičeskij Institut im. A. Russo », Učenye Zapiski, vyp. 6, 22–76 (e traduzione italiana nella II appendice del volume *Statistica linguistica*, Bologna, 1971, 377–440).

OHMANN R. (1964) – *Generative Grammars and the Concept of Literary Style*, in « Word », XX, 423–439.

PACAK M. (1967) – *Homographs: Classification and Identification*, in « Études de linguistique appliquée », V, 88–105.

PALME J. (1971) – *A natural Language Parsing Program for Question Answering*, FOAP Rapport C 8268–11 (64), Stockholm.

PANCONCELLI CALZIA G. (1911) – *Italiano*, Leipzig–Berlin.

PAPP F. (1966) – *Mathematical Linguistics in the Soviet Union*, The Hague.

PETRICK S. R. (1965) – *A Recognition Procedure for Transformational Grammars*, unpubl. Doct. Diss., MIT, Cambrigde (Mass.).

PETRICK S. R. (1972) – *Computer–oriented grammars and parsing*, in Center For Applied Linguistics, 22–37.

PETRICK S. R. (1974) – *Semantic Interpretation in the REQUEST System*, in Zampolli (1974 a).

PETÖFI J. S. (1969) – *On the complex analysis of languages as synchronic systems*, in « Revue », I, 1–18.

PIKE K. L. (1967) – *Language in Relation to a Unified Theory of the Structure of Human Behavior*, The Hague.

PLATH W. (1962) – *Mathematical Linguistics*, in *Trends in European and American Linguistics*, 1930–1960, Utrecht.

PLATH W. J. (1974) – *Transformational Grammar and Transformational Parsing in the REQUEST System*, in Zampolli (1974 a).

QUEMADA B. (1972) – *Lexicology and Lexicography*, in T. A. Sebeok (ed.), *Current Trends in Linguistics*, 9, 395–476.

QUEMADA B. (1973) – *Les travaux du centre d'étude du français moderne et contemporain*, in Zampolli (1973 a), 121–32.

QUILLIAN M. R. (1968) – *Semantic Memory*, in M. Minsky (1968), 216–270.

RABEN J. (1969) – *The death of hand made concordance*, New York.

REED D. W. (1949) – *A statistical Approach to quantitative Linguistic Analysis*, in « Word », V, 235–247.

REGULA M., JERNEI J. (1965) – *Grammatica italiana descrittiva*, Berna.

REVZIN I. I. (1968) – *Les modelès linguistiques* (Modeli jazyka), Traduit et adapté par Y. Gentilhomme, Paris.

REY–DEBOVE J. (1970) – *Le domaine du dictionnaire*, in « Langages », 19, 3–34.

REY–DEBOVE J. (1971) – *Limites des applications de la linguistique à la lexicographie*, in G. E. Perren, J. L. M. Trim (eds.), *Applications of Linguistics*, Cambridge, 369–375.

REY–DEBOVE J. (1971) – *Limites des applications de la linguistique à la lexicographie*, in G. E. Perre, J. L. M. Trim (eds.), *Applications of Linguistics*, Cambridge, 369–375.

RIESBECK C. (1973) – *Computer Analysis of Natural Language in Context*, Ph. D. Thesis, Stanford (Calif.).

RÍHA A., MACHOVÁ S. (1974) – *Computer Testing of Generative Grammar*, in A. Zampolli (1974a).

ROBINSON I. J. (1962) – *Preliminary Codes and Rules for the Automatic Parsing of English*, RM–3339–PR, Rand Corp., Santa Monica (Calif.).

ROBINSON J. J., MARKS S. (1965) – *PARSE: A System for automatic Analysis of English Text*, RM–4564–PR, Rand Corp., Santa Monica (Calif.).

ROCERIC–ALEXANDRESCU A. (1968) – *Fonostatistica limbii române*, Bucarest.

RODRIGUEZ BOU I. (1952) – *Recuento de vocabulario español*, Rio Piedras (Puerto Rico).

ROMEO (1967) – *On the Phonemic Status of the So–called « Geminates » in Italian*, in « Linguistics », XXIX, 105–116.

ROSENGREN I. – *The quantitative concept of language and its relation to the structure of frequency dictionaries*, in « Études de linguistique appliquée », I, 1, 103–127.

ROSETTI A. (1947) – *Le Mot. Esquisse d'une Théorie générale*, Copenhague–Bucarest.

ROSIELLO L. (1965) – *Struttura, uso e funzioni della lingua*, Firenze.

RUBENSTEIN M. (1966) – *Directions in Semantic Research*, in *Seminar on Computational Linguistics*, Bethesda, 97–104.

RUMELHART D. E., LINDSAY P. H., NORMAN D. A. (1972) – *A process model for long term memory*, in Tulving, Donaldson, *Organization and Memory*, New York.

RUWET N. (1964) – *La linguistique générale aujourd'hui*, in « Archives européennes de Sociologie », V, 277–310.

RUWET N. (1968) – *Introduction à la grammaire générative*, Paris.

SAGER N. (1972) – *The Sublanguage Method in String Grammars*, in R. W. Ewton Jr., J. Ornstein (eds.), *Studies in Language and Linguistics* (1970–71), El Paso (Texas).

SALKOFF M. (1973) – *Analyse syntaxique automatique utilisante une grammaire en chaîne (String grammar)*, in Zampolli (1973 a), 499–517.

SALTARELLI M. (1970) – *A Phonology of Italian in a Generative Grammar*, The Hague.

SANDEWALL E. (1972) – *A first–order Calculus for expressing Conceptual Information*, Uppsala.

SARAMANDU M. (1966) – *Sur le rendement fonctionnel des types de structures phonématiques en roumain*, in « Cahiers de linguistique théorique et appliquée », III, 147–160.

SCHANK R. C., RIEGER C. J. (1974) – *Inference and Conceptual Memory*, in Zampolli (1974 a).

SCHLISMANN A. (1948) – *Sprach–und Stilanalyse mit einem vereinfachten Aktionsquotienten*, in « Wiener Zeitschrift für Philosophie, Psychologie, und Pädagogik », II, 2, 42 e segg.

SCHNEIDER R. B. (1971) – *The Production of Machine Readable Texts: Some of the Variables*, in « Computers and the Humanities », VI, 3, 39–47.

SEDELOW S. Y., SEDELOW W. A. jr. (1972) – *Language Research and The Computer*, The University of Kansas, Lawrence (Kansas).

SEILER M. (1964) – *On defining the word*, in « Proceedings of the ninth International Congress of Linguists », The Hague, 767–770.

SHANNON C. E. (1948) – *A mathematical theory of communication*, in « Bell System Technical Journal », XXVII, 379–423, 623–656.

SILVA G. (1969) – *Phontrns: An Automatic Ortographic–to–Phonetic Conversion System for French*, in « Computers and the Humanities », III, 5, 257–265.

SIMMONS R. (1970) – *Natural Language Question Answering Systems: 1969*, in « CACM », XIII, 1, 15–30.

SIMMONS R., BRUCE B. (1971) – *Some Relations between Predicate Calculus and Semantic Net Representations of Discourse*, in « Procedings of 2nd International Conference on Artificial Intelligence », London, 524–530.

SIMMONS R. (1973) – *Mapping English Strings into Meanings*, Techn. Report NL 10, Dept. of Computer Sciences, University of Texas, Austin (Texas).

SMITH R. N. (1972) – *Interactive Lexicon Updating*, in « Computers and the Humanities », VI, 3, 137–145.

SPANG-HANSSEN H. (1956) – *The study of Gaps between Repetitions*, in M. Halle, *et alii* (eds.), *For Roman Jakobson*, The Hague, 492–502.

SPANG-HANSSEN H. (1963) – *Sentence Lenght and Statistical Linguistics*, in *Structures and Quanta: Three Essays on Linguistic Description*, New York, 58–72.

SPANG-HANSSEN H. (1967) – *Fini et infini dans le vocabulaire*, in « Langages », 6, 100–105.

STACHOVITZ R. A. (1974) – *Beyond the Feasibility Study: Syntax and Semantics*, in Zampolli (1974 a).

Statistique et analyse linguistique (1966), Paris.

VON STCEHOV A. (1973) – *Syntactic analysis of Italian*, in Zampolli (1973 a), 541–562.

STEINBERG D. D., JAKOBOVITS L. A. (1971) (eds.) – *Semantics. An Interdisciplinary Reader in Pholosophy, Linguistics and Psychology*, Cambridge.

ŠTEINFELDT E. (1969) – *Dictionnaire des fréquences des mots dans la langue russe moderne.* Les 2500 mots les plus usuelles à l'usage des professeurs de Russe, Moscou.

STINDLOVA J. (1967) – *Le dictionnaire de la langue tchèque lettéraire: enregistrement des données sur cartes mécanographiques*, in « Cahiers de Lexicologie », X, 103–113.

Structure of Language and its Mathematical Aspects (1961), AMS, Providence (R.I.).

SZANSER A. J. (1969) – *Automatic error–correction in natural languages*, Preprint to the 3th ICCL.

TAGLIAVINI C. (1965) – *Concordanze della « Divina Commedia »*, Pisa (Ci si riferisce all'introduzione dell'edizione IBM Italia).

TAGLIAVINI C. (1968) – *Applicazioni dei calcolatori elettronici all'analisi e alla statistica linguistica*, in *Atti del Convegno sul tema: L'automazione elettronica e le sue applicazioni scientifiche, tecniche e sociali* (Accademia Nazionale dei Lincei, Roma, 1967), Roma, 111–118.

TESNIÈRE L. (1959) – *Éléments de syntaxe structurale*, Paris.

THORNDIKE E. L. (1932) – *A teacher's word book of the twenty thousand words found most frequently and widely in general reading for children and young people*, New York.

THORNDIKE E. L., LORGE I. (1944) – *The teacher's word book of 30.000 words*, New York.

THORNE J. P. (1965) – *Stylistics and Generative Grammars*, in « Journal of Linguistics », I, 49–59.

THORNE J. P., BRATLEY P., DEWAR H. (1968) – *The Syntactic Analysis of English by Machine*, in Michie (1968) (ed.), 231–310.

TOGEBY K. (1949) – *Qu'est–ce qu'un mot?*, in « Travaux du Cercle Linguistique de Copenhague », V, 97–112.

DE TOLLEANERE F. (1963) – *Nieuwe wegen in de lexicologie*, Amsterdam.

DE TOLLEANERE F. (1973) – *Travaux de l'Institut de Lexicologie Néerlandaise*, in Zampolli (1973 a), 29–39.

TORRIGIANI G. (1968) – *Problemi metodologici dell'analisi linguistica mediante elaboratori elettronici presso il CNUCE*, Pisa.

TRNKA B. (1966) – *A phonological analysis of present–day standard english*, Tokyo.

TRUBECKOI N. S. (1939) – *Grundzüge der Phonologie*, in « Travaux du Cercle Linguistique de Prague », VII. Edizione italiana a cura di G. M. PORRU, *Fondamenti di Fonologia*, Torino (1971).

VALERIO C. (1893–1896) – *De la Cryptographie*, Paris.

VANDER BEKE G. E. (1929) – *French word book*, New York.

VEILLON G. (1973) – *Les Problèmes d'analyse en traduction automatique*, in Zampolli (1973 a) 563–590.

VENEZKY R. L. (1973 a) – *Computational Aids to Dictionary Compilation*, Toronto.

VENEZKY R. L. (1973 b) – *Computer applications in lexicography*, New York.

VINAY J. P. (1968) – *Enseignement et apprentissage d'une langue seconde*, in « Le Langage », (Encicl. de la Pléiade, dir. par A. Martinet), Paris, 685–725.

WAGNER R. L. (1967) – *Les vocabulaires français*, Paris.

WALKER D. (1974) – *Speech Understanding, Computational Linguistics and Artificial Intelligence*, in Zampolli (1974 a).

WATTS A. F. (1946) – *The language and mental development of children*, London.
WILKS Y., HERSOVITZ A. (1974) – *An Intelligent Analyzer and Generator for Natural Language*, in Zampolli (1974 a).
WINOGRAD T. (1971) – *Understanding Natural Language*, New York.
WINTER W. (1969) – *Styles as Dialects*, in Doležel, Bailey (1969), 3–9.
WISBEY R. (1965) – *Computers and Lexicography*, in D. Hymes (1965) (ed.), *The Use of Computers in Anthropology*, The Hague, 215–234.
WOODS W. A. (1967) – *Context–Sensitive Recognition*, in *Mathematical Linguistics and Automatic Translation*, Report MSF 18, Cambridge (Mass.).
WOODS W. A. (1970 a) – *Context–Sensitive Parsing*, in «CACM», XIII, 7, 437–445.
WOODS W. A. (1970 b) – *Transition Network Grammars for Natural Language Analysis*, in «CACM», XIII, 10, 591–606.
WOODS W. A. (1974) – *Meaning and Machines*, in Zampolli (1974 a).
YULE G. U. (1938) – *On Sentence–Length as a Statistical Characteristic of Style in Prose, with Application to Two Cases of Disputed Autorship*, in «Biometrika», XXX, 363–390.
YULE G. U. (1944) – *The statistical study of literary vocabulary*, Cambridge.
ZAMPOLLI A. (1960) – *Studi di statistica linguistica eseguiti con impianti IBM* (Tesi di laurea dattiloscritta), Padova.
ZAMPOLLI A. (1967) – *Nota Tecnica*, in *Raccolta Barbi di Canti Popolari Italiani, Esperimento di Elaborazione Elettronica* E1/RB, Pisa.
ZAMPOLLI A. (1968 a) – *L'elaboratore elettronico negli studi linguistici*, in «Rivista IBM», 2, 14–19.
ZAMPOLLI A. (1968 b) – *Recherche statistique sur la composition phonologique de la langue italienne exécutée avec un système IBM*, in *Machines dans la Linguistique*, Prague, 25–34.
ZAMPOLLI A. (1969) – *Due conversazioni sullo stato attuale della linguistica computazionale*, Pisa.
ZAMPOLLI A. (1970) – *Progetti e Metodi della Sezione Linguistica del CNUCE*, in «Revue», III, 39–83.
ZAMPOLLI A. (1971) – *Nota Tecnica*, in A. M. Bartoletti Colombo, *La Costituzione della Repubblica Italiana, Testo, Concordanze, Indici*, Firenze, pp. XVII–XXVI.
ZAMPOLLI A. (1973 a) – *Linguistica Matematica e Calcolatori. Atti del Convegno e della Prima Scuola Internazionale*, Firenze, 1973.
ZAMPOLLI A. (1973 b) – *La Section Linguistique du CNUCE*, in Zampolli (1973 b), 133–199.
ZAMPOLLI A. (1973 c) – *L'automatisation de la recherche lexicologique: état actuel et tendances nouvelles*, in «META», XVIII, 1–2, 101–136.
ZAMPOLLI A. (1973 d) – *Congrès International de Lexicographie Anglaise*, in «Cahiers de Lexicologie», 22, 1, 113–126.
ZAMPOLLI A. (1973 e) – *Humanities Computing in Italy*, in «Computers and the Humanities», 7, 6, 343–360.
ZAMPOLLI A. (1974 a) (ed.) – *Computational and Mathematical Linguistics. Proceedings of the 5 th ICCL*, Firenze, in stampa.
ZAMPOLLI A. (1974 b) (ed.) – *Logic, Semantics, Lexicology, Computers*, Firenze (in stampa).
ZAMPOLLI A. (1974 c) – *Fifth International Conference on Computational Linguistics*, in «ALLC Bulletin», 2, 2, 52–57.
ZAMPOLLI A. (1974 d) – *Description quantitative des systèmes linguistiques: vérifications expérimentales de quelques modèles récents*, in Zampolli (1974 a).
ZAMPOLLI A. (1974 e) – *Problemi di linguistica applicata computazionale*, Pisa.
ZGUSTA A. (1971) – *Manual of Lexicography*, The Hague.
ZIPF G. K. (1935) – *The Psycho–biology of Language: an introduction to dynamic phylology*, Cambridge (Mass.). (Citato nell'ediz. 1968).