A. Z A M P O L L I

L A   S E Z I O N E   L I N G U I S T I C A

D E L   C.N.U.C.E.

COMUNICAZIONE LETTA AL COLLOQUE INTERNATIONAL SUR LA RECHERCHE COMPUTATIONELLE EN PHILOLOGIE, ORGANIZZATO A LIEGI DAL 'LABORATOIRE D'ANALYSE STATISTIQUE DES LANGUES ANCIENNES, NEI GIORNI 7-8 SETTEMBRE 1969.

---===ooOoo===---

1. In occasione della inaugurazione del C.N.U.C.E. nell'Ottobre 1965, veniva consegnata al Capo dello Stato, Giuseppe Saragat, una copia delle concordanze della Divina Commedia, ottenute elaborando il testo del poema con i sistemi IBM 7090 e 1401 /Tagliavini, 1965/.

Nell'estate del 1964, per il contributo del Consiglio Nazionale delle Ricerche, l'Accademia della Crusca ha potuto finalmente riprendere quell'attività lessicografica che aveva costituito il principale suo compito, e che un decreto ministeriale aveva troncato nel 1923, quando la quinta edizione del suo vocabolario era giunta, con l'undicesimo volume, appena alla fine della lettera O. La ripresa dei lavori, per la compilazione di un grande vocabolario storico della lingua italiana, fu annunziata dal nuovo presidente, il Prof. Giacomo Devoto, nella pubblica seduta del 31 Ottobre 1964 /Duro, 1968 p.245/. Si presentò subito in tutta la sua urgenza il problema di organizzare lo spoglio dei testi, per creare un grande Archivio della lingua italiana, da cui si dovrà compilare innanzitutto, un tesoro delle origini della lingua italiana, nel quale confluire gran parte del materiale documentario relativo al periodo delle origini, fino alla data convenzionale del 1375. Seguir poi il grande Vocabolario Storico che, assorbendo in sè, convenientemente diradato, il materiale del Tesoro, proseguirà a fare la storia della lingua italiana fino ai giorni nostri. Si trattava di mettere assieme decine e decine di milioni di schede, estratte da non meno di 20.000 volumi. Nel momento in cui si affacciava la domanda, essa aveva già pronta la risposta: ricorrere all'aiuto dei sistemi elettronici. In Italia c'era già una 'tradizione' o 'scuola' che dir si voglia, in tal senso. Fin dal 1949, P.Roberto Busa S.J. aveva avviato, a Gallarate, l'attività del C A A L (Centro per l'automazione dell'analisi letteraria) /Busa 1962/ con lo scopo principale di elaborare l'Index Thomisticus, e cioè gli indici e le concordanze della opera omnia di S.Tommaso d'Aquino, a voi ben noto.

Il Prof. Wisbey /Wisbey, 1965/ espone molto efficacemente l'inserimento delle tecniche di automazione negli spogli lessicografici, e attribuisce a P.Busa la priorità in queste applicazioni.

Dopo un decennio di attività sulle macchine UR, verso il 1960 il CAAL decise di trasferire le proprie elaborazioni sul computer (a quell'epoca, in Italia, era in piena espansione la

cosiddetta 2a generazione dei calcolatori). Ne fu dato l'annuncio al convegno di Tubinga organizzato dal CAAL alla fine del 1960 /Convegno 1960/ e particolarmente utili furono le discussioni tra sostenitori delle macchine UR e sostenitori dei computers che ricordo particolarmente animate al 'Colloque', organizzato nel maggio 1961 a Besançon dal Prof. Quemada.

L'Accademia della Crusca si rivolse dunque al CAAL, e sul finire del 1964 ottenemmo le prime concordanze e schede - contesto (1) di un gruppo di 20 sonetti del Fiore (poemetto attribuito a Dante Alighieri), e lo spoglio selettivo del Varmo di Ippolito Nievo: dico 'ottenemmo' perchè già da 4 anni lavoravo con P.Busa, che mi incaricò di collaborare con il Prof. A.Duro, direttore del Vocabolario della Crusca. Dopo il successo di quei primi esperimenti, l'Accademia della Crusca avviò, presso il CAAL, la perforazione dei primi 4 milioni di parole; iniziai lo studio della procedura per lo spoglio elettronico, e compilai una serie di programmi per l'ottenimento di concordanze e di schede-contesto lemmatizzate. I programmi da me studiati, come pure la procedura e la codificazione, ebbero necessariamente sin dal principio le caratteristiche proprie di un programma di utilità (utility program): si doveva infatti tener presente che lo spoglio avrebbe riguardato testi di 10 secoli, appartenenti a tutti i generi letterari, a diversi stati di lingua, e perfino a diversi sistemi linguistici, (per es. diversi dialetti). Soprattutto le edizioni sono diversissime, e redatte con una incredibile varietà e incoerenza di criteri. Basterà pensare, per esempio, a quante svariate forme può assumere il riferimento in una tale situazione, per avere un'idea dei problemi posti in fase di analisi e di programmazione.

Per l'esecuzione dei lavori, ci rivolgemmo nel 1966 al C.N.U.C.E. (2) di Pisa, organismo che, per la sua costituzione, ci offriva la disponibilità di macchine e l'organizzazione relativa in termini economici accessibili. Per una notevole coincidenza, anche le elaborazioni dell'Index Thomisticus furono trasferite a Pisa all'inizio del 1967, e così potei seguire presso il C.N.U.C.E. entrambi i progetti.

Ben presto seguirono numerosi altri progetti di ricerche linguistiche e particolare efficacia ebbe a questo scopo la seduta dedicata all'automazione in linguistica nel corso del Convegno sul tema 'L'automazione elettronica e le sue implicazioni scienti

fiche, tecniche e sociali, organizzato a Roma il 16-19 ottobre 1967 dall'Accademia Nazionale dei Lincei, nel corso del quale C.Tagliavini /Tagliavini 1968/, A.Duro ed io /Duro-Zampolli 1968/ presentammo lavori in corso al CNUCE.

2. I progetti nel settore umanistico divennero così numerosi, che nel Luglio del 1968 fu decisa, in seno al CNUCE, la costituzione di una sezione linguistica, la quale da un lato fornisce assistenza tecnica e scientifica agli Utenti, dall'altro promuove studi e ricerche originali volti a migliorare ulteriormente le applicazioni linguistiche dei calcolatori /Torrigiani 1968/.

La sezione, da me diretta, conta 12 collaboratori a tempo pieno, e inoltre alcuni laureati borsisti. Essa utilizza in parte il 7090 e il 1401 16K del CNUCE, ma in particolare le è riservato a pieno tempo un IBM 360/30, dotato di un lettore perforatore di schede, di 4 nastri, 2 dischi e 2 stampatrici, che entro il prossimo anno sarà affiancato da un più potente IBM 360/65. Una stampatrice veloce 1403 mod. 1 (1100 righe/min.) sta per essere dotata di una catena a 120 caratteri diversi per la stampa di testi in alfabeto latino. In questa catena, le combinazioni lettera + segno diacritico più frequenti nel corpus di testi in elaborazione al CNUCE, sono previste come caratteri a sè: avremo così già fisicamente predisposte in un unico carattere le combinazioni del tipo à ò ó é è ecc. La seconda stampatrice IBM 1404 (600/righe/min.) può stampare sia fogli normali, sia schede meccanografiche perforate. E' dotata infatti di un dispositivo di alimentazione e lettura di schede, sulle quali nel contempo si può stampare come su normali moduli di carta: la useremo per produrre quelle schede-contesto, che molti lessicografi /De Tollenaere, 1963/ giudicano strumento insostituibile del proprio lavoro (3).

Questa stampatrice sarà dotata di alcune catene speciali a 120 segni, per diversi alfabeti: per ora sono in ordine una catena per alfabeto latino, e una catena per alfabeto greco. Entrambe le catene sono concepite in modo da poter comporre assieme, in un unico carattere di stampa, segni appartenenti a 'slugs' distinti sulla catena. Ciò è possibile sopprimendo, per mezzo di uno speciale dispositivo comandato a programma, la spaziatura verticale automatica del carrello dopo un ciclo di stampa.

Queste apparecchiature dovrebbero risolvere il problema del

la stampa di elaborati di controllo e di tabulati definitivi a tira tura limitata. Tale ricchezza di caratteri è importante, come ben può valutare chi ha provato a controllare liste di testi contenenti 300-400 grafemi diversi, o più, rappresentati e codificati con i 48 caratteri delle stampatrici 'standard'; sarà sempre possibile ricorrere, per la pubblicazione, alla riproduzione fotografica in offset dei tabulati, come abbiamo già fatto /Tagliavini 1965, Zampolli 1967, Accademia della Crusca 1967/ e stiamo facendo per le concordanze non lemmatizzate dell'opera omnia di Seneca (4).

Tuttavia, la moderna tecnica della fotocomposizione presenta, ai fini della pubblicazione, vantaggi innegabili. D.G.Hays li ha chiaramente elencati al Seminar on Computational Linguistics di Bethseda /Hays 1966/ (8): essa unisce alle qualità estetiche della stampa e alla varietà pressochè illimitata di stili corpi e caratteri impiegabili nella stessa pagina, l'esattezza assoluta della composizione, mentre il type-setting, a causa degli errori non infrequenti nella caduta meccanica dei caratteri trascinati per forza di gravità, richiedeva la rilettura delle bozze. Questa necessità annullava uno dei vantaggi costitutivi della meccanografica: la certezza che se la prima registrazione dei dati è esatta, essa è mantenuta in tutte le trasmissioni e trasformazioni di questi dati nelle fasi successive della elaborazione.

A tutti noi è evidente come tale certezza sia indispensabile in molte elaborazioni linguistiche, ad es. nella pubblicazione di concordanze, nelle quali viene stampato, mediamente, un numero di righe 10 volte maggiori del numero di righe in input. Nel dicembre del 1968, abbiamo stampato, in collaborazione con l'Accademia della Crusca, un primo volumetto di indici e di concordanze di una novella di anonimo del 400, /Accademia della Crusca, 1968/ che è il primo esperimento nel suo genere in Italia e, a quanto sappiamo, in Europa (5). Alla fine dell'anno pubblicheremo indici, concordanze, rimario, ecc. del Canzoniere del Petrarca. Non disperiamo di avere in seguito una fotocompositrice presso di noi.

3. La nostra sezione, per decisione del Comitato direttivo del CNUCE /Zampolli, 1968a/, fornisce agli utenti umanisti, non solo l'uso delle macchine e l'organizzazione e la esecuzione delle elaborazioni, ma anche collaborazione scientifica e la stesura dei nuovi programmi, quando non siano sufficienti quelli già predisposti

per lo spoglio lessicografico e la elaborazione statistica di un testo. La serie di questi programmi, che abbiamo già sperimentato su più di 40 milioni di parole in 15 lingue diverse, funzionava finora sul 1401/16 K, ed è stata quasi interamente ritrascritta per il 360.

Tutti noi abbiamo esperienza di questo tipo di elaborazioni; mi limiterò pertanto ad elencare i programmi principali, dopo una breve premessa sul loro carattere di generalità.

E' stato ripetuto tante volte da diventare un luogo comune: /Kay, 1965 e 1967/ che il problema della creazione dell'input è un freno, uno scoglio nelle elaborazioni linguistiche computazionali, per la varietà dei caratteri da rappresentare rispetto al numero di codici disponibili nelle normali perforatrici e per la grande quantità di dati che costituiscono di solito il corpus da elaborare; che, per questi motivi, è auspicabile un sistema standard di codifica, così da permettere l'acquisizione e lo scambio di testi tra ricercatori o Centri diversi: ci si sta avviando infatti a una situazione di fatto per cui è sempre più probabile che il ricercatore trovi già registrato per il calcolatore il testo che gli interessa; che la standardizzazione difficilmente può riguardare i sistemi di perforazione: infatti, per quanto un sistema possa essere ottimizzato nel senso, per es., di scegliere i codici meno 'costosi' per i caratteri più frequenti, si presenteranno sempre dei testi per il quale il sistema ottimale non è quello standard, a motivo, per restare nell'esempio scelto, di una diversa distribuzione di frequenza dei grafemi; che la standardizzazione va operata quindi per mezzo di una conversione da un codice di perforazione lasciato alla libera scelta del ricercatore, a un sistema di codifica standard per la registrazione del testo su nastro magnetico, o disco, ecc.

Noi siamo fondamentalmente d'accordo con tutte queste osservazioni, che vengono spesso formulate anche per la codifica in sede di stampa, anche se la nostra esperienza ci mostra che un sistema standard di perforazione può servire per una quantità di testi maggiore di quello che comunemente si afferma.

E' anche facile dimostrare che, purchè la perforazione rispetti alcuni pochi vincoli, è possibile compilare un programma di conversione generalizzato, adattabile all'input per mezzo di schede controllo:

3.1 Difatti il nostro programma di carico, il quale stampa una lista che riproduce fedelmente il testo e viene di solito letta per il controllo, è munito di una routine generalizzata di conversione; contemporaneamente trasferisce le schede-testo su nastro, nel quale le diverse informazioni contenute nelle schede sono assegnate a categorie funzionali distinte. La standardizzazione è resa possibile proprio da una classificazione dei grafemi e delle informazioni in genere presenti nei testi, operata assumendo come criterio distintivo la funzione, o, più precisamente, l'insieme di funzioni complesse che ciascun grafema o ciascuna informazione debbono esercitare nelle elaborazioni dell'intera procedura. Non potendo qui entrare in dettagli, rinvio a una monografia che uscirà presso il C.N.U.C.E. /Zampolli 1970/. Ogni elemento o unità del testo, occupa un record a sè, numerato progressivamente, e questo sistema facilita la fase di correzione che - e questo fatto è sottolineato invece meno spesso di quanto sarebbe necessario - costituisce uno scoglio e un freno, maggiore del costo della perforazione.

3.2. Il programma di correzione è basato sulla convinzione, frutto di una esperienza decennale, che è meglio correggere il nastro anzichè le schede, e che la prima registrazione del nastro deve realizzarsi avendo come unità di registrazione unità il più elementari e 'corte' possibili. Particolare attenzione va posta a quelle categorie di informazioni che riguardano uno 'spezzone' di testo, e a quelle che sono concatenate tra di loro, come per es. la numerazione delle righe.

Così, per esempio, non assegnamo subito a ogni parola il suo riferimento tipografico o organico, ma registriamo il riferimento solo dove esso cambia, come record a sè. Adoperiamo anche altri accorgimenti, frutto dell'esperienza e resi possibili dalla suddetta classificazione funzionale: per es. tutta una serie di controlli a programma, sul tipo di quelli descritti da Szanser /1969/, e abbiamo la possibilità di decidere di volta in volta (per mezzo di schede controllo in fase di carico o di particolari 'tipi-record' in correzione), la funzione di grafemi, per così dire, omografi (per es. i puntini, il puhto interrogativo, e esclamativo, ecc., in alcuni casi sono da considerare punteggiature 'forti', in altri punteggiature 'deboli', ai fini della costruzione dei contesti; oppure la presenza di un riferimento organico può interrompere oppure

no la numerazione progressiva delle righe del testo).

Il programma 'legge' da schede di correzione, distinte naturalmente in diversi tipi-inserimento, sostituzione, cancellazione ecc. - e modifica su nastro i records corrispondenti, richiamati per mezzo del numero progressivo che li distingue univocamente.

3.3 Altri programmi, dopo il sort del nastro parola, stampano le forme grafiche del testo in ordine alfabetico e in ordine di frequenza, compiendo contemporaneamente una serie di operazioni statistiche facilmente immaginabili (calcolo del rango, della sommatoria progressiva delle frequenze, ecc.).

3.4 Alcuni programmi sono predisposti per ricercare e stampare, o per ricercare e modificare parole o altri tipi di informazioni specificati di volta in volta per mezzo di schede. La lista delle forme, esaminata prima di proseguire con le elaborazioni, si è rivelata infatti uno strumento di controllo prezioso, soprattutto per la ricerca di eventuali incoerenze dell'editore: è questo un capitolo molto importante per molti dei progetti in corso al C.N.U.C.E., che prevedono, come s'è detto, lo spoglio di testi in edizioni critiche diversissime per concezione e per attendibilità.

I programmi in questione permettono di ricercare ed eventualmente di riparare simili inconvenienti.

3.5 Un programma genera i contesti, cioè aggiunge a tutte le parole del testo, o solo a categorie di parole formalmente individuate (ciò serve ad esempio per gli spogli selettivi), il 'contesto' nel quale appare. L'algoritmo tiene conto di diversi fattori (punteggiatura, segni filologici di apparato critico, livelli di riferimento, ecc.) nel delimitare il contesto: entro certi limiti, l'elenco e la gerarchia dei fattori possono essere variati per mezzo di schede controllo, come pure l'ampiezza del contesto.

3.6 Una volta generati e alfabetizzati i contesti, un programma stampa le concordanze delle forme grafiche. Nel caso si tratti di una lingua per la quale disponiamo di un vocabolario di macchina (DM), questo stesso programma contiene l'algoritmo di consultazione (6).

I contesti delle forme lemmatizzate possono essere stampati,o meno, a seconda che lo studioso voglia o no controllare la lemmatizzazione automatica. I contesti delle forme non lemmatizzate (perchè assenti nel DM o perchè omografi possibili) vengono sempre esaminati da persona competente.

3.7 Rinvio, per la descrizione dei nostri metodi e programmi di lemmatizzazione, alla comunicazione citata all'Accademia dei Lincei /Duro-Zampolli, 1968/.

3.8 Una serie di programmi provvede alla stampa delle diverse frequenze e concordanze per lemmi, delle schede contesto, di indici diversi, rimari, incipitari, ecc.

3.9 Una serie di programmi messi a punto da noi o adattatidalle biblioteche-programmi del C.N.U.C.E., della SHARE, ecc., esegue le diverse elaborazioni statistiche (dispersione, entropia, ecc.) sulle unità linguistiche analizzate dallo spoglio e sulle loro frequenze di occorrenza e di concatenazione.

3.10 Particolarmente elaborati sono i programmi che preparano i diversi tipi di risultati per la fotocomposizione: essi sono complicati da un lato da problemi come la giustificazione, l'hifenation, la distribuzione degli spazi, dall'altro dalla constatata frequenza con la quale gli utenti desiderano intervenire all'ultimo minuto prima della stampa definitiva, per modifiche di carattere generale o a luoghi particolari.

3.11 Si è già detto del programma che permette di realizzare i diversi ordini di selezione richiesti dalle lingue diversissime in elaborazione. Essenzialmente, essi incorporano una tabella che specifica per ogni grafema sia a quale dei diversi livelli ordinati gerarchicamente, normalmente presenti nell'ordinamento lessicografico di una lingua, il grafema vada assegnato, sia la sua posizione relativa all'interno di tale livello.

3.12 Una serie di programmi è poi messa a punto per ricerche di carattere fonematico. Rinvio per essi alle mie pubblicazioni /Zampolli, 1960,1968 ecc./, (cfr. anche /Silva, 1969/), e cito solamente le

principali funzioni che essi svolgono: un sistema di trascrizione automatica dall'alfabeto ortografico all'alfabeto fonematico (sperimentato per l'italiano); divisione in sillabe fonetiche e fonematiche; applicazione di regole di ridondanza e di contestualizzazione (in una prospettiva binaristica di tratti distintivi), ricerca di n-grammi e n-fonemi; ricerca di gruppi consonantici o vocalici; calcolo dell'entropia (di grado n) dei fonemi, delle sillabe,ecc.; statistiche di prosodemi o tratti soprasegmentali; statistiche sulle sequenze sillabiche e sulla struttura delle parole; ricerche sul rendimento funzionale, ecc.

3.13 Il carattere di generalità del nostro sistema di codifica e dei nostri programmi è dimostrato se non altro dal numero di lingue e di testi elaborati. Possiamo affermare, con un ragionevole grado di certezza, che se un ricercatore ci propone un testo da elaborare in una lingua qualsiasi, purchè trascrivibile con un alfabeto, l'esecuzione dello spoglio lessicografico, dell'analisi lessicale e statistica, e in genere delle elaborazioni che implicano ricerche di unità, di combinazioni di unità, e la produzione di indici e concordanze, è semplice routine, dalla codificazione del testo, alla pubblicazione dei risultati finali. Come vedremo anche nei punti successivi, la concentrazione di lavori che si è verificata al C.N.U.C.E. presenta notevoli vantaggi che derivano essenzialmente dall'adozione di uno stesso sistema di registrazione e di analisi, il quale rende interscambiabili testi diversissimi ed operabile univocamente una immensa 'biblioteca elettronica', e dal fatto che il nostro sistema, frutto di esperienze multiple e pluriennali, funziona come schema cui si riferisce il ricercatore non solo nel codificare, ma anche nello scegliere le informazioni da registrare per i testi che lo interessano. La presenza di questo ampio generale e ben formalizzato schema permette di esplicitare e non trascurare categorie di informazioni la cui presenza nel testo non è immediatamente evidente, e spesso induce il ricercatore ad introdurre alcune informazioni non immediatamente utilizzabili da parte sua, ma indispensabili per lo sfruttamento del testo da parte di altri ricercatori.

4. Lavori in corso

Gli Istituti Universitari o del CNR che si avvalgono dei pro-

grammi, degli impianti, o della collaborazione scientifica e tecnica del CNUCE sono oggi poco meno di 50, distribuiti in molte regioni italiane, e alcuni in altri Paesi europei. Mi sembra più utile in questa sede indicare i principali settori nei quali operano, anzichè elencare i progetti ad uno ad uno.

4.1 Documentazione lessicografica per la redazione di grandi dizionari storici di una lingua.

Esempi: Dizionario Storico della lingua italiana dell'Accademia della Crusca: Dizionario dei secoli XV-XVI dell'Accademia Rumena di fonetica e dialettologia di Bucarest; Dizionario dei testi storici della lingua Ittita, dell'Istituto di Glottologia di Pavia, per il Centro di Studi Egei del CNR. E' comune a imprese di questo tipo l'esigenza di elaborare testi tra loro diversissimi, per datazione, lingua, genere letterario, criteri di edizione. Da un lato dunque i programmi devono contenere algoritmi flessibili, adattabili alle differenti strutture e codificazioni dei testi; dall'altro, poichè i materiali devono alla fine confluire in un unico archivio, la loro presentazione e codificazione deve risultare il più omogenea possibile, sia per poterli rielaborare automaticamente come un unico corpus, sia per assicurare, a chi consulta attestazioni di una medesima voce riunite assieme dalle fonti più disparate, l'esatta e, quanto più possibile, immediata comprensione delle informazioni disponibili per ciascuna occorrenza.

Riferiamoci per alcuni esempi alle schede-contesto. In esse il lemma e il 'sottolemma' devono essere formulati in modo tale che la ricerca e il reperimento delle occorrenze nell'archivio finale non siano ostacolati da quell'insieme di alternative possibili che 'tormentano' chiunque si trovi a dover lemmatizzare un testo (7). Il numero e la complessità di tali alternative, che hanno le loro cause profonde nella situazione del concetto di parola nella linguistica contemporanea (8) sono notevolmente accresciuti dalla distribuzione dei testi in strati di lingua contigui e diversi, sia lungo la dimensione sincronica - stili, generi letterari, regionalismi, ecc. - sia lungo la dimensione diacronica - si pensi per es. al variare della lingua italiana dalle origini ai giorni nostri. Per questo motivo, l'utilità o addirittura la possibilità di un dizionario di macchina vengono sovente poste in

discussione /Duro, 1968/. Non è in questa sede che ci possiamo addentrare nella discussione: la mia personale opinione è che gli stessi motivi che rendono difficile la compilazione di un DM, ne esaltino l'utilità, come strumento che praticamente garantisce la coerenza di comportamento di 'équipes' di lemmatizzatori diversi, operanti per un certo numero di anni. E' però necessario, almeno in un primo tempo, concepire il DM non come un sostituto del lemmatizzatore, ma come uno strumento che registra tutte le decisioni e le conoscenze accumulate via via nel corso della lemmatizzazione, con il quale il lemmatizzatore può - per così dire - conversare. In definitiva, invece che ordinati per forme grafiche, i contesti verrebbero sottoposti all'esame del lemmatizzatore già ordinati per lemma, o comunque accompagnati dall'insieme di informazioni registrate nel DM a proposito di ciascuna forma grafica.

Il lemmatizzatore potrà controllare i risultati di questa prima lemmatizzazione automatica, intervenendo là dove sono necessarie correzioni o decisioni. Abbiamo calcolato che, pur con questo controllo, l'impiego di un DM consentirebbe, in media, un risparmio del 75% del tempo rispetto a una lemmatizzazione interamente artigianale. Naturalmente si potranno identificare nel corpus da spogliare delle 'tranches' per le quali è utile compilare un DM a parte, oppure, mantenendo solo un DM, introdurre, nella matrice delle codificazioni (9), indicazioni concernenti lo stile, l'epoca, ecc., che consentano di rendere di volta in volta operanti o inoperanti, in relazione alla 'tranche' da elaborare, alcuni elementi del DM.

Intendiamo per riferimento, quell'insieme di informazioni che individuano la posizione della parola nel testo stampato, e le dividiamo in due categorie principali: riferimento tipografico (volume, tomo, pagina, riga, ecc.) e riferimento organico (quando l'autore o l'editore hanno diviso il testo in parti, la parola è identificata per mezzo del nome o del numero della parte ove appare; per es. capitolo e riga, oppure canto ottava verso, fino alle suddivisioni complesse della filosofia scolastica: libro, lezione, capitolo, questione, articolo, paragrafo, ecc.). Noi chiediamo a tutti i nostri Utenti di riportare sempre il riferimento tipografico, e, se presente, quello organico. Nel caso di riferimento organico si trovano spesso situazioni disparate e, per così dire, fantasiose: per es. il riferimento deve essere formulato così: Titolo della dedica, Introduzione, rubrica par. 1, ecc.; in altri casi, il solo rife-

mento possibile è l'incipit della parte: per es. il primo verso di un sonetto. Si aggiungano altre complicazioni; per es. la numerazione delle righe nel riferimento tipografico non corrisponde alla numerazione delle righe del riferimento organico: il caso più semplice è quello di un verso che, in una tragedia, sia stampato su due o più righe, perchè viene pronunciato da due o più personaggi diversi.

Un altro esempio è formato dalle oscillazioni degli editori per quanto concerne l'apparato critico e in genere l'espressione a livello grafemico di uno stesso fenomeno: per es. l'integrazione di una lettera o di una parte di testo è espressa racchiudendo tra [ ] o tra ( ) o tra ⟨ ⟩, o ponendo in corsivo, o contrassegnando con puntino sottoscritto, ecc. le lettere integrate.

4.2. Documèntazione lessicografica per la redazione di dizionari storici di discipline particolari, o di particolari strati di lingua.

A titolo di esempio ricordo il progetto per la compilazione del 'Lessico Intellettuale Europeo' e quello per il 'Vocabolario Giuridico' /Fiorelli 1957 e 1966/:entrambi questi progetti seguono le tecniche e le metodologie messe a punto per l'Accademia della Crusca, in particolare si propongono come risultato finale l'ottenimento di schede-contesto che, almeno in parte, dovranno confluire anche nell'Archivio della Lingua Italiana da cui verrà tratto il vocabolario della Crusca. Questi progetti presentano pertanto le difficoltà esaminate al punto 4.1, con in più la complicazione di dover operare una maggiore selezione nello spoglio: a differenza infatti degli spogli per il vocabolario della Crusca, interessano qui soprattutto termini pertinenti lo strato o stile di lingua studiata che si trovano immersi nel lessico comune.

4.3. Documentazione lessicografica per la redazione di dizionari, lessici, dell'opera omnia di un autore.

Sempre a titolo di esempio, nomino gli spogli dell'opera omnia di Antonio Rosmini e di L.A.Seneca.

Mi sembra utile notare che in questa categoria rientrano soprattutto filosofi e pensatori: forse ciò è dovuto a una semplice coincidenza, ma forse è dovuto invece allo scopo che ci si propone spogliando le opere di un filosofo. L'indice del lessico viene con

siderato, come mi disse più volte P.Busa, come una chiave di ingresso al pensiero dell'autore, come un mezzo di ricerca dei luoghi ove vengono definite o esplicitate le idee, e lo studio della evoluzione e nell'uso della terminologia, è rivelatore della evoluzione del pensiero del filosofo; scopi, questi, che richiedono naturalmente lo spoglio dell'opera omnia dell'autore.

In questa categoria rientrerebbe naturalmente il progetto dell'Index Thomisticus, che svolge le sue elaborazioni al C.N.U.C.E. ancora in questi giorni. Ma P.Busa sta per trasferirsi per alcuni anni negli U.S.A.,ove concluderà l'Index Thomisticus. Egli lascerà però una copia dei materiali (nastri e schede) presso di noi.

4.4. Spogli di opere di singoli autori, per studi diversi (metrici, grammaticali, stilistici, tematici, ecc.) per lo più su base lessicale e statistica.

Ciascuno degli studi citati, richiederebbe un discorso a sè, che qui non è certamente il caso di fare, per ovvi motivi di spazio. Del resto, voi potete facilmente immaginarlo, grazie alla vostra personale esperienza: infatti sono qui presenti esperti di ciascun settore. Mi limito perciò a citare alcune lingue (greco classico, latino medioevale rinascimentale e moderno, spagnolo, francese, tedesco, sanscrito (10), ebraico, aramaico, paleoslavo, italiano dalle origini ai giorni nostri, ecc.) e alcuni autori (Pindaro, Bacchilide, Aristotile, S.Bernardo, Baumgarten, Kant, Goethe, Canovacci della Commedia dell'Arte, Fabbri, Machado, Gide, K.Marx, Saba, Gozzano, ecc.), per dare un'idea dell'estensione dei materiali registrati e dell'effettivo grado di 'utilità' o 'generalità' che dir si voglia dei nostri programmi e del nostro sistema di codifica. Particolarmente utili si rivelano a questo proposito la classificazione dei grafemi in classi o categorie funzionali, che è alla base del nostro sistema di elaborazione, e la flessibilità dei nostri programmi per l'ordinamento alfabetico, per cui è possibile elaborare contemporaneamente testi in lingue e alfabeti diversi.

4.5. Dialettologia

"It is possible now to draw diagrams - for example, maps - under the control of a computer (...) The possibility of combinig

map making with statistical operations in fascinatig, since it means reduced labor and therefore, probably, greater sophistication in future dialect surveys" (Hays 1969). W.N.Francis, J.Svartik,G.M. Rulin della Brown University, in una comunicazione /Francis 1969/al recente ICCL /ICCL/, affermano che la dialettologia, soprattutto nella sua fase interpretativa, è un campo della linguistica particolarmente aperto all'uso dei calcolatori. La dialettologia, tipicamente, ha a che fare con un grande corpus di dati, usualmente in forma di parole singole o di corte frasi, ed è interessata a ordinare e comparare elementi singoli su basi diverse: fonetiche, morfologiche, lessicali, geografiche, ecc. Il maggior ostacolo che ha fin qui allontanato dall'uso dei calcolatori nello studio dei dialetti, è il fatto che per molti dialetti i dati sono stati raccolti ed editi prima dell'epoca dei calcolatori. Il problema di preparare una grande quantità di dati, per lo più in trascrizione fonetica, per il calcolatore, è un compito formidabile. I risultati ottenuti, in modo relativamente agevole e con un enorme risparmio di ore-uomo rispetto ai metodi tradizionali, nell'ordinare e mettere in carta i materiali dialettali, sono senza dubbio tali da incoraggiare altri a investire tempo e denaro nel preparare i dati per il calcolatore piuttosto che nell'ordinare e fare la carta a mano. In particolare Francis et alii illustrano un esperimento di utilizzazione dei materiali dialettali registrati nel 4° volume del 'Survey of English Dialects' /Orton 1967/, per studiare le variazioni nella situazione delle fricative interdentali in 10 contee del Sud dell'Inghilterra. Dopo aver trascritto su schede, per ciascuna risposta, la parola inglese standard con l'equivalente dialettale in trascrizione fonetica e il numero codice della località, hanno chiesto al calcolatore elenchi di vario tipo, nei quali le informazioni registrate sono variamente correlate le uno alle altre. Ulteriori liste e conteggi statistici sono stati eseguiti per rilevare la diversa distribuzione dei fonemi fricativi in posizione iniziale rispetto ai contesti fonici seguenti: contemporaneamente, il plotter stampava una carta per ciascuna delle distribuzioni studiate. Recentemente ho saputo che progetti analoghi sono in corso in Germania, e negli U.S.A. all'Università del Wisconsin, sotto la direzione di G.Cassidy, si prepara il D.A.R.E. (Dictionary of American Regional English); interessante è anche il lavoro R.W.Shuy,

a proposito di un programma di 'automatic retrieval' per l'Atlante linguistico degli U.S.A. e del Canada /Shuy,1966/. Gordon R.Wood, ha illustrato all'ICCL una applicazione analoga, per la 'scoperta' e lo studio dei diversi tipi di americano regionale, lungo la costa dell'Atlantico. Egli utilizza tra l'altro liste di frequenza per identificare le parole che, scelte a preferenza di altre equivalenti, caratterizzano le preferenze tipiche di un sistema linguistico regionale, e riconosce l'utilità delle concordanze lessicali, grammaticali e sintattiche /Wood 1969/.

Senza avere notizie delle ricerche citate, e indipendentemente quindi dal loro esempio, in Italia da circa un anno applichiamo il calcolatore alla dialettologia. In collaborazione con il Prof. C. Grassi di Torino, abbiamo compilato concordanze, per così dire, contrastive, di diverse versioni di un unico testo dialettale biellese, raccolte in epoche diverse, per studiare i mutamenti causati dall'evoluzione diacronica.

Abbiamo quindi affrontato il problema di redigere gli indici e le carte dell'ALI (Atlante Linguistico Italiano) /Grassi-Terracini, 1967/.

Lo studio scientifico della dialettologia, inaugurato da G.I. Ascoli sul finire dell'800, portò al sorgere della geografia linguistica e all'esame sincronico di una unità linguistica; questo indirizzo, contrastando le idee dei neo grammatici sulla rigidità delle leggi della evoluzione fonetica, ebbe notevole peso nello sviluppo teorico della linguistica generale. Tutta una serie di inchieste dialettali sfociò nella compilazione di atlanti linguistici regionali e nazionali, come l'ALF (atlante linguistico francese) dell'abate J. Gillieron (1902-1912) e l'AIS (atlante italiano svizzero) di K.Jaberg e J.Jud (1928-1940; 1960 gli indici) (11).

I progetti di un atlante dialettologico italiano, sostenuti a più riprese (1909; 1914; 1921) da illustri linguisti (D'Ovidio, Goidanich, Parodi, Salvioni, Bartoli, Bertoni, ecc.) si concretarono infine negli anni del primo dopoguerra, per merito della società Filologica Friulana "G.I.Ascoli" nel grandioso piano dell'ALI (atlante linguistico italiano) (11), le cui inchieste, condotte in massima parte da U.Pellis in circa 1000 punti della penisola e delle isole, su un questionario di oltre 7000 domande, si conclusero nell'autunno del 1965 sotto la direzione del compianto B.Terracini. Nel laboratorio dialettologico dell'Università di Torino, sono oggi rac

colti circa 3 milioni e mezzo di risposte per un totale, approssi mativamente calcolato, di 9 milioni di parole.

Da questo archivio che, per ricchezza di lessico e fittezza di puhti investigati, non ha uguali si debbono ora estrarre le parole da riportare nelle previste carte dialettali (circa 3000) e, compito più arduo, ricavarne degli indici che ne rappresentino la chiave di accesso. Nei primi contatti con la IBM e altre società, le difficoltà di ordine tecnico (es. ricchezza di segni del l'alfabeto fonetico) e scientifico (es. mancanza di strutture e corrispondenze formali tra domanda italiana e risposta dialettale) sembrarono in un primo tempo sconsigliare l'impiego di procedimenti automatici.

Si può dimostrare che:

- la nostra tecnica di codificazione fonetica risolve i problemi di input così come la catena di stampa quelli di output;
- è utile rovesciare l'ordine delle fasi della procedura tradizionale (che vorrebbe prima la messa in carta dei materiali, e poi il loro ordinamento in un indice di accesso alle carte) in questa successione:

a) registrare meccanograficamente tutte le parole
b) ove possibile 'tipicizzarle', cioè ricondurle a un unico 'tipo' italiano o toscano equivalente
c) elaborare l'archivio così ottenuto secondo i parametri suggeriti dalle diverse categorie di informazione registrate (12)
d) pubblicare, prima ancora delle carte, gli indici alfabetici regionali, grammaticali, fonetici, tavole dialettali comparative e altre esaurienti documentazioni del materiale raccolto
e) scegliere sulla base degli indici le parole da riportare in carta con il vantaggio metodologico di dominare, per loro mezzo, la massa dei dati (per es. di reperire parole dialettali ottenute, per così dire, casualmente, in risposta a domande non direttamente intese a provocarle ecc.)
f) porre in input la registrazione magnetica delle parole scelte e comporre automaticamente in fotocompositrice la carta (ciò, ho appurato, è possibile individuando, su un sistema di coordinate cartesiane, i punti-località e lo spazio disponibile per stamparvi accanto la risposta).

Oltre all'economia di tempo e ai vantaggi metodologici, alla fine si otterrà un archivio dei dialetti italiani operabile automa

ticamente, le cui voci, già si pensa, potranno essere lo 'heading' sotto cui raccoglierenuovi dati e monografie di una enciclopedia elettronica delle conoscenze dialettali, regolarmente aggiornabile.

Va tenuto presente, a questo proposito, il rinnovato interesse per la dialettologia testimoniato da numerosi convegni di questo e dello scorso anno, motivato dalla posizione, per così dire, privilegiata della dialettologia nel quadro degli studi dei rapporti tra linguaggio e società, e soprattutto nel contesto nazionale attuale di mobilità sociale o geografica di individui e gruppi: sono pertanto in corso o in progetto numerose iniziative e inchieste che produrranno un ricco materiale documentario.

Sul piano diacronico, i dialetti per le peculiari caratteristiche di reazione del linguaggio ai fatti sociali, conservano l'unica traccia di precedenti diverse strutturazioni di comunità ecohomiche e sociali (ad es. montane e rurali). Sul piano sincronico rivelano e testimoniano importanti fenomeni di evoluzione e di assestamento: basta pensare, per esempio, alla diversa configurazione dei rapporti, sul piano dei modelli di prestigio nel comportamento, tra dialetti e lingua regionale o nazionale, e tra dialetti di gruppi sociali diversi. A questo si aggiungano le preoccupazioni per l'avvenire e la conoscenza dei dialetti che ha avuto echi in recenti disposizioni ministeriali per la scuola.

In questo contesto è in corso un progetto di raccolta documentaria dialettale per tutte le regioni, per la cui realizzazione i promotori attendono dalla automazione un importante contributo.

4.6. Elaborazioni di grandi files di dati linguistici.

S.Lamb, chiamerebbe probabilmente così questo settore (Processing of linguistic Files) nel celebre articolo nel quale classifica le attività della Linguistica Computazionale /Lamb 1965/.

Un ottimo esempio è costituito dal progetto del Gruppo Italiano di Studi Demologici e del Comitato per la 'Raccolta Barbi', nel campo della poesia di tradizione orale (13).

Esiste una tradizione di rigore filologico applicata allo studio della letteratura popolare in quanto diversa dalla poesia d'arte. La poesia popolare italiana presenta una larga varietà di generi e tipi, diffusi oralmente e appoggiati in qualche caso dalle stampe di girovaghi o cantastorie: si tratta di proverbi scongiuri indo-

vinelli, fiabe e canzoni che hanno una propria storia a volte secolare e un modo particolare di tradizione che impongono metodi di ricerca adeguati. Delle numerosissime varianti la critica non fa opera di 'eliminazione', ma conserva anche le varianti che presentano segni chiari di modifiche, o innovazioni felici e deterioranti, rispetto a condizioni che si ritengono più antiche. Non si cerca cioè la ricostruzione di un archetipo, ma la ricostruzione più estesa possibile della tradizione di un componimento. Si può dire che l'unità non è rappresentata dal singolo componimento, ma che i versi o gruppi di versi che lo costituiscono hanno ciascuno una vita più o meno indipendente nello spazio e nel tempo. Si trovano canti con una tradizione molto compatta ma senza che vi sia un particolare comune a tutte le versioni: l'area di diffusione di una variante è magari diversa da quella della variante dei versi immediatamente successivi nello stesso canto, e si trovano versi, originariamente appartenenti a un canto, inseriti in canti differenti tra loro o di regioni lontanissime. E' evidente come il riscontro delle analogie, che era in pratica affidato alla memoria dello studioso, o a una lunga schedatura manuale, sia enormemente facilitato dalle concordanze, che mettono accanto e perciò permettono di evidenziare versi coincidenti o parzialmente simili, provenienti da tutte le raccolte spogliate. Un'altra direzione di studio è la ricerca di accoppiamenti tipici e frequenti di parole, che aiuti a rilevare la presenza di formule moduli e schemi ritmici, che costituiscono il canovaccio sul quale si costruiscono le improvvisazioni, rielaborazioni e innovazioni individuali. Un altro problema, già parzialmente affrontato con il calcolatore, consiste nell'esaminare le parole in posizione di rima per integrare la descrizione tradizionale degli schemi metrici popolari: accanto alle forme 'canoniche' della rima della consonanza e della assonanza, si intuisce un tessuto più sottile e una varietà di rispondenze foniche accettate dalla coscienza popolare, che sono in parte da enunciare e da inventariare (14). Lo strumento classico è il rimario, che noi abbiamo sviluppato in forme diverse (15).

In questa direzione, è ancora tutta da sviluppare la metodologia della trascrizione fonetica delle parole in rima: è nostra convinzione infatti che per rimari e dizionari inversi la trascrizione fonetica sia in molti casi indispensabile; del resto A.Juilland ha fornito un primo esempio con il suo indice inverso della lingua fran

cese /Juilland 1965a/. I lessici automatici, per la varia provenienza geografica dei testi popolari raccolti, potranno essere utilizzati parzialmente come documentazione utile per gli studi sull'italiano regionale, dei quali i linguisti lamentano la rarità.

Il progetto della raccolta Barbi fornisce, a mio avviso, anche un notevole esempio di applicazione nel settore, a quanto mi consta nuovo all'uso dei calcolatori, delle edizioni. La raccolta Barbi, come si è detto, esiste al presente come una collezione di manoscritti, che attende di essere edita. Essa viene ora perforata direttamente su schede e il testo è corredato da tutto quell'insieme di comandi che ne permetteranno l'edizione per mezzo della fotocomposizione. Con queste stesse schede, ancor prima della pubblicazione del testo, si otterranno indici, incipitari, rimari, concordanze: e dunque, non solo con una unica battitura si ottengono due risultati (a. la composizione per la edizione, che elimina l'intervento del tipografo, b. le elaborazioni lessicografiche e linguistiche), ma la redazione stessa del testo può essere pubblicata dopo essere passata al vaglio critico dello spoglio. Nella quotidiana esperienza dell'Accademia della Crusca, il semplice esame della lista delle forme e delle relative frequenze fa scoprire refusi o errori anche nelle edizioni critiche giustamente più celebrate e stimate.

Un ulteriore passo in avanti sarà forse reso possibile dai lettori ottici, già oggi disponibili sul mercato, che 'leggono' testi scritti con certi tipi di macchine da scrivere. L'editore o il curatore potranno battere il testo come d'abitudine in forma di dattiloscritto, che verrà assunto elettronicamente senza passare per il tramite della perforazione.

E' agli inizi, sempre per il gruppo italiano di studi demologici, la costituzione di un immenso schedario in vista della compilazione di una enciclopedia delle tradizioni popolari, per il quale si combineranno assieme le tecniche della schedatura lessicografica e della 'information retrieval'.

4.7. Statistica linguistica

L'attenzione dei linguisti sulla costanza della frequenza dei grafemi e dei fonemi è stata attirata per la prima volta dai lavori di stenografi e crittografi sulla frequenza delle lettere.

Gli spogli dei pedagoghi hanno attirato l'attenzione sul fatto

che un piccolo numero di parole spesso ripetute costituisce la maggior parte di un testo /Guiraud, 1959, p.44-50/.

Gli studi sulle regolarità messe in evidenza da questi dati, condussero alla formulazione dei primi modelli statistici tra i quali ebbero particolare risonanza quelli proposti da G.K.Zipf dopo la prima guerra mondiale /Zipf, 1937/.

I linguisti cominciarono a prendere coscienza delle possibilità di analisi fondate su un metodo rigoroso e particolare attenzione fu rivolta alla statistica linguistica dalla scuola di Praga. Questo interesse si accentuò nel secondo dopoguerra, come risulta tra l'altro dalle comunicazioni al VI° Congresso dei Linguisti del 1948, di linguisti di formazione tradizionale, come il grande semitista Marcel Cohen, I Whathomugh e M.Durand. Press'a poco in quegli anni, D.W.Reed /1949, pp.235-236/ in un articolo su Word del 1949 anticipa chiaramente il problema della frequenza come carattere costitutivo di una forma linguistica e insiste sul rapporto tra la frequenza in testi particolari e la probabilità in lingua.

P.Guiraud ha dato una formulazione teorica a queste affermazioni. La lingua potrebbe essere intesa come un ordine sistematico di engrammi, ciascuno presente con la sua frequenza, la quale costituisce un attributo oggettivo quanto la forma o il significato /Guiraud 1959 p. 16 e p.25/.

Egli ha anche formulato le basi teoriche della possibilità di analisi della "langue" e della "parole" equiparate a "codice" e "messaggio", con i concetti e le tecniche della teoria dell'informazione. Molti aspetti del funzionamento della "langue" posti dal Saussure a fondamento della sua dottrina, quali "la différentialité, la segmentation, la linéarité, l'arbitraire du signe linguistique... sont.... des caractères généraux et pré-linguistiques communs à toute une catégorie de signes informationnels" /Guiraud 1954, p.119/.

L.Heilmann riassume così: "Se, dunque, oggi vogliamo tener presente 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, attuantesi in un processo individuale fisiopsichico sottoposto al determinismo statistico" /1962b p.137/.

Un ulteriore sviluppo di questi principi, è costituito dalla equiparazione, proposta da Herdan, tra l'antinomia saussuriano "lan-

gue-parole" da una parte e il rapporto "popolazione statistica-campione" dall'altra. La 'langue' equivale alla "totality of engrams in the brains members of the speech community with their probabilities of occurence" e la 'parole' "for random samples for it" /Herdan,1960, p.34/.

Queste premesse, secondo alcuni, costituirebbero le basi teoriche sulle quali la statistica linguistica si propone come disciplina indipendente /Roceric-Alexandrescu, 1968, p.18/.

Queste premesse non hanno però trovato, almeno fino ad oggi, un consenso generale. Alcuni autori, e non solo linguisti ma anche di formazione statistica, hanno rilevato una eccessiva semplificazione nelle affermazioni di Herdan e del primo Guirand, ed hanno proposto, come vedremo più avanti, una maggiore cautela, consigliando di utilizzare e, se il caso, attendere, il risultato degli spogli di testi che lo sviluppo della LC rende sempfe più numerosi e attendibili.
Già nel 1964 al Convegno di Strasburgo Statistique et analyse linguistique, R.Moreau delineava esplicitamente questa situazione della linguistica: "... Les premiers pas de la statistique appliquée à la linguistique ont précisement consisté à admettre des règles de 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 supp sait 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. Le (...) CREDIF (...) a fait une première brèche dans cette croyance en introduisant sa notion de mots disponibles. M.me Hirschberg et Ch. Muller ont montré que, sauf cas particulier, cette règle de jeu n'était pas vérifiée" /Moreau, 1964, p.130/.

Un'atmosfera di meditata prudenza caratterizzò del resto molti degli interventi al Congresso. Nel volume degli atti apparso due anni dopo, l'introduzione a firma di Ch. Muller e B.Pottier ne dà testimonianza: "...Comme dans toute évolution des méthodes de recherche, après un periode d'expansion, voir d'excès, il vient un moment d'équilibre. Les linguistes savent à present qu'il faut tenir compte des critères quantitatives; ils ignorent dans quelle mesure il faut les considerer comme pertinents, en fonction des champs spécifiques d'études" /Statistique, 1966, p.1/. Nelle conclusioni, che ricordo approvate dall'assemblea dei congressisti, e riportate a chiusura degli atti (pp. 133-134) è rilevante l'invito a un maggior rigore

metodologico nella raccolta dei dati ai quali si vogliono applicare le tecniche statistiche, e l'auspicio di una partecipazione attiva dei linguisti e dei filologi tradizionali come garanzia di rigore nella definizione delle unità linguistiche e di utilità nella scelta dei problemi da sottoporre al trattamento statistico. A questo scopo si auspica che le nozioni di base prendano posto nella formazione dei linguisti. Queste raccomandazioni non sono rimaste inascoltate, come dimostrano la pubblicazione di alcuni manuali dedicati ai linguisti (per es. /Muller, 1968/), l'introduzione a livello accademico, nei dipartimenti di linguistica di numerose università europee e americane, di corsi riguardanti diversi settori della linguistica matematica, e soprattutto la revisione critica e programmatica di alcune convinzioni di base della statistica linguistica, l'ideazione e la verifica di nuove tecniche per lo spoglio, e la moltiplicazione degli spogli per fornire i dati di base alle operazioni di induzione e verifica delle ipotesi della statistica linguistica. I materiali linguistici registrati e analizzati presso il C.N.U.C.E. e i centri analoghi, costituiscono una fonte preziosa e di importanza decisiva. Ne abbiamo già iniziato lo sfruttamento in questo senso, e per l'italiano condurremo delle analisi in modo sistematico a diversi livelli. Uscirà entro l'anno un mio studio sulla fonematica, è agli inizi un progetto per la sintassi, è molto avanzato uno studio a livello lessicale, che entro la fine dell'anno si concreterà nella pubblicazione di un Dizionario di frequenza dell'italiano. /Bortolini et alii/. Esso seguirà da vicino l'esempio della bella collezione dell'editore Mouton, dedicata ai Dizionari di frequenza delle lingue romanze scritte, in corso di realizzazione sotto la direzione di Alphonse Juilland all'Università di Stanford. Sono già apparsi il volume dedicato allo spagnolo (1964) e al rumeno (1965).

Il motivo principale della scelta del modello proposto dallo Juilland, è che così contiamo di rendere confrontabili i nostri risultati con i suoi. Ch.Muller con il suo concetto di lessico opposto a quello di vocabolario, con la distinzione tra lessico di un idioma, lessico di una lingua, lessico di un gruppo umano, lessico di un individuo del gruppo, e soprattutto lessico di situazione; R.Moreau con la distinzione tra parole ergodiche nella lingua e parole ergodiche in un centro di interesse; i lavori del CREDIF con la distinzione tra lessico disponibile e lessico di frequenza, han-

no chiaramente illustrato, anche su un piano teorico, le difficoltà insite nel presupposto stesso delle operazioni che conducono alla compilazione di un vocabolario fondamentale.

L'intuizione statistica che sta alla base della compilazione dei dizionari di frequenza, è stata chiaramente espressa da Ch.Muller, "La Fréquence (F) d'une unité lexicale, mesurée dans un corpus assez étendue et judicieusement composé, est considérée très généralement comme une bonne estimation de sa probabilité d'emploi: d'une observation dirigée sur un fait de discours, on induit un fait de langue, inobservable par définition; c'est par le même processus qu'un échantillon bien composé fournit une estimation satisfaisante sur les caractères d'une population trop vaste pour être soumise toute entière à des tests" /Muller 1965, p.33/. Ma, osserva sempre Muller, "on peut s'interroger (...) sur la légitimité de l'operation,et sur l'existence d'une fréquence 'en langue'„.

Quale che sia la definizione della lingua come universo statistico, i testi sottoposti a spoglio si configurano nelle intenzioni dei compilatori dei dizionari di frequenza, come campioni di questo universo, e di conseguenza si pongono i problemi della campionatura per i quali la statistica ha elaborato tutta una serie di metodologie e di tecniche /Boldrini 1960, p.181 e s./.

Da un lato occorre delimitare l'universo, dall'altro assicurarsi della rappresentatività del campione. Se si conoscono bene le caratteristiche dell'universo, si può fabbricare un campione in cui ciascuna di queste caratteristiche sia distribuita secondo le stesse proporzioni, secondo una stratificazione sottile.

Gli studiosi del problema consigliano di aumentare quanto più è possibile il numero degli strati: "Stratifiez à outrance", come dice Moreau.

D'altra parte la linguistica quantitativa ha già ottenuto alcuni importanti risultati, soprattutto per quanto riguarda la applicazione e l'adattamento dei metodi classici della statistica ai problemi della campionatura lessicale /Muller, 1964, pp.133-134/.

Penso che gli studi tradizionali di stile e di lessicografia possano fornire i dati di partenza per una primissima stratificazione a priori della lingua, o meglio i dati per identificare, a livello di ipotesi di lavoro, alcuni sottoinsiemi sufficientemente definiti e egualmente caratterizzati rispetto a un parametro comune. Ai risultati degli spogli si applicheranno poi le tecniche statistiche

per verificare, da un lato, l'omogeneità delle strutture quantitative dei sottogruppi, dall'altro, per identificare le eventuali parole tematiche e ergodiche.

La determinazione degli strati di lingua potrà derivare da un lavoro accurato e documentato di induzioni da spogli sempre più numerosi, sempre che, naturalmente, questo stesso lavoro provi l'esistenza di strati omogenei e definibili di lingua sul piano delle strutture quantitative.

Prima di impegnare i metodi statistici in nostro possesso nello studio dell'universo linguistico, si dovrà apprendere di più sperimentalmente e induttivamente sulle strutture quantitative di sottoinsiemi linguistici sicuramente definibili; la scuola francese degli ultimi anni, è ritornata per così dire un passo indietro rispetto alle generalizzazioni dei primi tempi (Moreau, 1964, p. 130/; esercita e verifica le tecniche statistiche note, in ambiti ben delimitati, e da questi cerca di estrarre regole strutturali valide limitatamente all'insieme osservato, confrontando poi tra loro parti distinte di una stessa opera e l'opera nel suo complesso con queste parti; oppure confronta opere di uno stesso autore, o opere di due autori diversi, evitando, nel confronto, la mediazione della "norma", nel senso di Guiraud /Muller, 1964, pp.5-6/. Questo atteggiamento è del resto favorito dallo sviluppo della lessicografia automatica in Francia, ove il numero degli spogli disponibili si accresce giorno per giorno grazie all'attività dei Centri di Besançon e Nancy.

Questo atteggiamento sembra estremamente ragionevole.

I testi dei diversi settori, registrati su nastro magnetico al CNUCE e sottoposti a spoglio lessicografico, offrono alle ricerche di statistica linguistica un corpus straordinariamente vasto (circa 50 milioni di parole) che copre un gran numero di lingue e di epoche storiche, (ittita, sanscrito, greco antico, latino della classicità, ebraico, aramaico, nabateo, latino medioevale rinascimentale e scientifico, italiano dalle origini ai giorni nostri scritto e parlato, francese, tedesco, spagnolo, paleoslavo, rumeno, diversi dialetti italiani, ecc.). Le unità linguistiche sono classificate e registrate secondo criteri linguistici e tecnici uniformi, tali da permettere un rapido conteggio interamente automatico in qualsiasi settore del corpus, con la immediata raccolta dei dati quantitativi richiesti dalle elaborazioni statistiche.

Questa coerenza e omogeneità, sono indubbiamente uno dei principali vantaggi del concentrarsi delle elaborazioni in grandi 'biblioteche elettroniche' nazionali, e sono uno dei motivi più convincenti per intensificare gli scambi di informazione e i tentativi di coordinamento tra nazioni diverse, che, come il Prof. Delatte ha detto esplicitamente, è uno degli obbiettivi di questo convegno.

4.8. Elaborazioni per altre discipline umanistiche

Mentre l'interesse dei linguisti per il linguaggio naturale è dato per definizione, l'interesse di altre discipline deriva da una attenzione non al linguaggio per se ma alla funzione del linguaggio come veicolo primario per la trasmissione delle informazioni nella società umana. Che si accetti o no l'idea della cosiddetta 'esplosione dell'informazione', è un dato di fatto che molte organizzazioni impegnate nell'elaborazione di informazioni su larga scala si rivolgono alla automazione, e mirano a costruire dei sistemi per analizzare il contenuto di testi in linguaggio naturale. Idealmente la content-analysis consiste nel determinare i concetti presenti nei testi e le relazioni tra questi concetti. I concetti e le relazioni che vengono identificati sono tradotti in un insieme di frasi canoniche che rappresentano il contenuto del documento: nel caso di un sistema 'fact retrieval or question-answering', queste frasi rappresentano, per così dire, la conoscenza del sistema, e servono come base per generare risposte pratiche a richieste specifiche. Naturalmente, le procedure di fatto esistenti di 'content analysis', sono solo una lontana approssimazione di questo ideale: è chiaro che analizzare il contenuto di un testo in linguaggio naturale implica un elevato ordine di formalizzazione delle conoscenze sintattiche e semantiche, oggetto principale di attenzione da parte di molte correnti linguistiche moderne. Non è possibile analizzare in questa sede le difficoltà connesse a questo problema, il che equivarrebbe a esaminare le teorie delle numerose correnti che caratterizzano il recente risorto sviluppo della semantica. L'articolo di Todorov nel primo numero di Langages costituisce una chiara sintesi delle diverse scuole, fino all'anno della sua pubblicazione. E' noto che lo sviluppo dei fermenti semantici contenuti nelle idee di Chomsky, ha generato un ulteriore rigoglio di studi, per una sintesi dei quali rinvio alle comunicazioni di Binnik /1969/ e di Schwarz

/1969/ alla recente Second International Confererence on computational Linguistics di Stoccolma /ICCL/ e alla raccolta di saggi curata da E.Bach e R.T.Harms./1968/. In Italia, il problema dei rapporti tra sintassi e semantica non è certo nuovo, ricordo uno per tutti, il saggio "Sintaxe et sémantique" di G.Devoto /1953/. Oltre al gruppo già citato di Parisi, si hanno altri studi in questo settore, per es. a Bologna. /Calboli,1969/ /Colombo,1969/.Quello che mi preme rilevare, è come studi di questo tipo conducano a un ricupero della descrizione del lessico e dei suoi tratti,richiedendo così anche sul piano della linguistica quelle compilazioni di dizionari almeno tentativamente formalizzati che erano già alla base delle procedure computazionali per il trattamento automatico dell'informazione contenuta in testi o messaggi in lingua naturale (traduzione automatica, documentazione automatica, computer assisted instruction, ecc.).

A conferma di questo fatto potrei citare molti autori: scelgo, tra tutti, le affermazioni di H.H.Josselson (che, come è noto, da molto tempo lavora alla compilazione di dizionari computazionali del russo /Josselson 1966/ nella comunicazione "The Lexicon: A System of Matrices of Lexical Units and Their Properties", presentata all'ICCL 1969, ove egli si riferisce esplicitamente all'opera di Ch. Fillmore /1966,1968,/ecc.: che ha operato "La trascrizione da una codificazione puramente sintattica (...) a una codificazione semantica (...) con i suoi casi grammaticali" /Josselson, 1969, p.11/. I problemi che generalmente vengono riferiti alla semantica, aveva o avuto una gran parte nel determinare la crisi della traduzione automatica e dell'information retrieval - almeno nei suoi approcci di tipo linguistico - nei primi anni del 1960. /Kuno, 1966/. Oggi, non indipendentemente forse dal risveglio degli studi della cosiddetta 'sematica generativa' negli USA dopo la stasi beavorista e distribuzionalista, anche la linguistica applicata ritorna ad affrontare il problema della semantica in un prospettiva nuova (certo molto più ambiziosa e forse utopistica, per le difficoltà teoriche di una formalizzazione della descrizione semantica /De Mauro, 1968/) descritta criticamente da D.G.Hays /1969/: lo scopo è, al limite, quello di riprodurre globalmente l'attività di un 'human information processor'; si vedano a questo proposito i saggi editi da M.Minsky con il titolo significativo Semantic Information Processing, 1968.

L.Karttunen, nella sua comunicazione all'ICCL, considera un meccanismo concepito per leggere un testo in una lingue naturale, per 'interpretarlo', per poi memorizzare il contenuto in qualche modo, ad esempio con lo scopo di rispondere a delle domande su questo contenuto. Egli analizza alcuni requisiti fondamentali che la macchina dovrebbe avere. Essenzialmente deve essere capace di costruire un 'file' che consiste nelle registrazioni di tutte le unità individue menzionate (eventi, oggetti, persone, ecc.), e per ogni 'individuo' dovrebbe essere registrato tutto quello che di esso è stato detto: deve riconoscere quando è menzionata una nuova unità nel testo e memorizzarla con la caratterizzazione per futuri riferimenti.

Karttunen ammette che, almeno per ora, un simile 'interprete' non è realizzabile, ma afferma che questo non deve scoraggiarci dallo studiare in astratto quali capacità dovrebbe possedere; esamina così per es. in che condizioni l'apparizione di un sintagma nominale indefinito implichi l'esistenza di una qualche unità o individualità specifica.

4.8.1. In assenza di un modello realmente funzionante di analisi del contenuto, non mancano però tecniche e procedure che sono di grande aiuto nell'analisi di documenti. Mi riferisco alla comunicazione di L.Fossier; che abbiamo sentito oggi, sullo studio de l'analisi del contenuto di fonti diplomatiche medioevali. /Fossier/.

Anche al C.N.U.C.E. abbiamo in corso progetti analoghi, i nostri, come il suo, ancora in fase sperimentale. Cito a titolo di esempio il progetto dell'Istituto di Storia medioevale e moderna, Paleografia e Diplomatica dell'Università di Pisa, che si propone di analizzare circa 1500 Carte Lombarde dal 774 al 1100 (Atti notarili, documenti pubblici e privati, placiti, diplomi, bolle, ecc.), per ottenere indici di nomi di luogo, di enti ecclesiastici, di nomi di persone, di cose notevoli dal punto di vista storico-economico e storico-giuridico, e la elaborazione di tutti i rapporti possibili tra i suddetti elementi, al fine soprattutto di identificare persone che appaiono più volte nei documenti, gruppi familiari, parentele e la "microtoponimia" delle singole località. Non è il caso di ripetere quanto ha detto L.Fossier: indici lessicali e concordanze sono già un notevole aiuto - lo abbiamo sperimentato - per il reperimento delle informazioni; inoltre

forniscono nel contempo una documentazione linguistica interessante, sia per la lingua latina medioevale, sia per i molti elementi romanzi che così si possono reperire e documentare. La fase di preedizione dei testi ha in questo progetto notevole importanza. La divisione del testo in 'pericopi' secondo la tradizione diplomatica, conferisce ai contesti stampati nelle concordanze caratteristiche che li rendono particolarmente adatti agli studi diplomatici, i quali del resto ricevono utili documentazioni statistiche per es. sui sistemi di datazione, sulla natura giuridica del documento, su formule, ecc. Queste statistiche sono possibili perchè, appunto in fase di preedizione, si riempie per ciascun documento una 'matrice', in cui appaiono codificate molte informazioni sul documento e sul suo contesto. Stiamo poi mettendo a punto un sistema di "descrittori", inseriti nel testo, per individuare le unità referenziali, e un sistema per correlarle tra di loro.

4.8.2. L'Istituto di Archeologia dell'Università di Pisa, ha progettato una elaborazione sul catalogo della ceramica greca di D.Beazley, come pr no nucleo di archivio-schedario gestibile automaticamente della ceramica greca, da aggiornare regolarmente. Ci si propone di correlare statisticamente le diverse informazioni registrate per ciascun oggetto.

Le applicazioni della elaborazione automatica all'Archeologia hanno avuto un notevole sviluppo. Sono ben note le classificazioni di queste attività proposte da Gardin /1965/, e un quadro aggiornato della situazione è fornito dai Preprints per il 'Colloque international sur l'emploi des calculateurs en archeologie: problèmes sémiologiques et mathématiques' organizzato dal CNRS francese a Marsiglia lo scorso aprile.

4.8.3. In una categoria del tutto particolare dobbiamo collocare le applicazioni del calcolatore nel settore della musicologia e della musica.

Per quanto concerne la prima, i progetti in corso al C.N.U.C.E. sono paragonabili agli studi statistici sullo stile in poesia e nella letteratura in generale: con la differenza, ovvia, che la formalizzazione dei dati è precostituita, cosicchè la ricerca degli 'stilemi' può venire svolta più facilmente e con rigore matematico, soprattutto senza l'ostacolo dei problemi connessi alla semantica. I

progetti in corso di propongono lo studio dello stile di partiture musicali del '500 italiano. Per il sistema di trascrizione musicale un prezioso aiuto ci è venuto dalla rivista Computer and the Humanities di New York. /Erickson, 1968/.

Con l'occasione degli studi musicologici, siamo entrati in contatto con il laboratorio Fonologico del Conservatorio di Firenze, che ha in corso l'elaborazione di programmi per usare il calcolatore come esecutore di musica, che arricchisce la già cospicua serie degli strumenti elettronici oggi in uso nei laboratori musicali. Il programma preparato elabora dati relativi alla emissione di 30.000 frequenze udibili la cui generazione avviene in una sezione dell'unità centrale del 7090 IBM. Il programma, oltre a permettere la realizzazione di una quantità elevatissima di intervalli e ritmi inediti e preclusi a qualsiasi altro strumento musicale, accetta una serie di istruzioni riguardanti l'immediata modifica dei valori di frequenza e di tempo di tutti o parte dei suoni di un qualsiasi testo musicale precedentemente eseguito o anche semplicemente memorizzato nelle memorie centrali e periferiche.

E' possibile, infine, modificare il 'temperamento' di ogni testo dato con una sola istruzione che determina l'entità, pressochè illimitata, della variazione sia in aumento che in diminuzione, del rapporto intervallare originario. /Grossi,1969/.

4.9 Psicologia, Psichiatria, Pedagogia

4.9.1. La storia e il recente sviluppo della psicolinguistica sono ben noti, e altrettanto sono noti i punti di incontro tra la psicologia e la linguistica /Titone, 1964, pp.49 e segg./, /Tagliavini, 1966, vol. I, cap.IV/. Non mi riferisco solo agli argomenti comuni come l'apprendimento del linguaggio, i disturbi del linguaggio, l'apprendimento di una seconda lingua, il bilinguismo, ecc. /Soporta, 1961/ Lo studio psicologico del comportamento linguistico ha ricevuto un nuovo impulso dall'influenza chomskiana. /Schlesinger,1967/: "I linguisti hanno specificato i principi del generare meccanicamente le frasi grammaticali, ed è stato suggerito che principi simili possono essere operati nel parlante umano. E' compito della ricerca psicologica esplorare questi suggerimenti, e la loro fecondità è attestata dal considerevole numero di studi psicolinguistici, pubblicati negli ultimi anni, e concernenti le variabili sintattiche, in contrasto con la quasi totale as-

senza di tali studi prima della pubblicazione delle Syntactic Structures di Chomsky". (p.397) (16). (E' noto che Chomsky ha collaborato strettamente con gli psicologi, per es. Miller(/Chomsky, 1958/, ecc.). Le interazioni tra la psicologia e la linguistica, e soprattutto alcuni presupposti della scuola Chomskiana, per es., che l'uomo nasca con dei meccanismi innati assai particolari i quali giuocano un ruolo considerevole nelle acquisizione del linguaggio, ripropongono il tema di chiarificare i rapporti e di tracciare i confini tra linguistica e psicologia, già molto dibattuto nella scuola linguistica americana, attorno ai tempi della prima guerra mondiale, si parlava di mentalismo, psicologismo, ecc., termini che oggi ritornano a proposito di Chomsky) e ridotto a giuste proporzioni da P.A.Gemelli e G.Zunini nella loro Introduzione alla Psicologia, p.258 segg. n.Ruwet /1968,p.37/ afferma che "Se la linguistica vuole essere una scienza bisogna pure che essa si ponga la questione dei propri rapporti con la psicologia", e segnala che "lo sviluppo della grammatica generale è strettamente legato a quello delle nuove concezioni nel dominio psicologico, e all'abbandono di una psicologia strettamente beaviorista". Chomsky e Halle hanno mostrato che un certo numero di principi teorici difesi dagli strutturalisti non potrebbero spiegarsi che a partire da una certa concezione della percezione. "Sembra inevitabile - dice ancora Ruwet - che una teoria linguistica si fondi, esplicitamente o no, su dei presupposti psicologici". Del resto, pur con tutte le limitazioni e le distinzioni tra modello e realtà che il modello vuole rispecchiare /I.I.Revzin, cap.I/, Chomsky afferma chiaramente che i suoi saggi "affrontano lo studio del linguaggio quale branca della psicologia umana teorica. Il loro scopo consiste nel presentare e chiarire le capacità mentali che permettono a un essere umano di imparare a usare una lingua" /Chomsky,1969,p.31/.

In Italia, una equipe del Centro di Psicologia del CNR a Roma, situa i propri studi in questo contesto /Parisi,1968/. In particolare, in accordo con i più recenti sviluppi cosiddetti 'postchomskiani', dirige le ricerche verso la descrizione formalizzata di alcuni settori di lessico /Parisi,1969/. Queste ricerche utilizzano il materiale documentario, gli esempi, prodotti dai nostri spogli: ciò, chiaramente, non sorprende. E' vero che nella concezione generativa-trasformazionale della grammatica, il corpus ha un valore diverso da quello che gli attribuiscono altre scuole linguistiche,

per es. i distribuzionalisti, o i grammatici tradizionali, pratici, che prendono i loro esempi presso gli scrittori.

Essenzialmente, le scuole generative osservano che il corpus, per quanto vasto, è, per definizione, finito, mentre "i linguisti hanno generalmente ammesso che una grammatica deve essere capace di predire, a partire da osservazioni in numero necessariamente limitato, un numero indefinito di frasi che non figurano in questi corpus, e che tuttavia se venissero emesse, sarebbero considerate dai parlamenti come facenti parte della lingua" (Ruwet, 1968, p.36). Pertanto la presenza o meno in un corpus non può essere assunta come criterio di 'grammaticalità' di una frase. D'altra parte, un corpus può comprendere errori d'attenzione, lapsus, parole incomplete,che i parlanti 'rifiuterebbero', correggerebbero. Tuttavia "è la 'esecuzione' che fornisce i dati di osservazione, corpus di ogni tipo, scritti o orali (conversazioni registrate, interviste, drammi, articoli di giornali, testi letterali, ecc.), che permettono di affrontare lo studio della 'competenza'" /Ruwet,1968,p.18/.

4.9.2. Ma ci sono altri motivi per i quali la sezione linguistica del CNUCE collabora con psichiatri, psicologi, pedagoghi.

Il calcolatore viene usato in questi settori per motivi diversi: R.Moreau, nella presentazione del Simposium 'Computers in psychological Research' del 1966 dice: "L'homme a la 'competence', c'est-à-dire la capacite de decider et d'agir en toute connaissance de cause, mais son systemè nerveux est souvent insuffisant pour traiter des problèmes trop compliques. Il n'a pas la 'performance' . C'est ainsi que nous avons la competence pour multiplier deux nombres ayant chacun cent mille chiffres, mais nous n'arriverons jamais au bout de cette tache. L'ordinateur, lui, a la performance. Si on lui dit comment executer le travail, donc si on lui fournit la competence, il menera à bien le calcul en quelques minutes (...) ce symposium conduit à penser que l'ordinateur en devenant capable de simuler des phenomènes humaines a dejà acquis une partie de la competence. Serait-ce là le premier pas vers l'acquisition par la machine de ce que nous appelons l'intelligence?"/ Moreau,1966, p.VIII/. Anche senza far riferimento agli studi che vengono raggruppati sotto l'etichetta comune di 'intelligenza artificiale', dei quali abbiamo in Italia un notissimo esempio nella scuola di S.Ceccato /1966/, sono ben note le ricerche per la creazione di modelli con-

cettuali e strumenti matematici che "aiutino a studiare la stupefacente complessità dei cervelli animali" /Braitenberg, 1968/; anche di queste abbiamo in Italia degli esempi, che sono stati illustrati nella seconda seduta del già citato convegno dell'Accademia Nazionale dei Lincei sull'Automazione.

Le discipline in questione traggono poi dall'uso del calcolatore molti dei vantaggi che il calcolatore offre alle scienze mediche in generale /Ledley,1956/ /Maccacaro,1968/: la gestione di archivi di dati sufficientemente compatti e di immediata consultazione, l'applicazione di metodologie statistiche per riconoscere nei fenomeni delle sistematicità che possono essere completamente mascherate da fattori pertubativi casuali, ecc. /Ledley, 1967/.

Presso il C.N.U.C.E. sono in corso numerosi progetti di questo tipo.

4.9.3. A noi interessa qui citare le ricerche che hanno come oggetto il linguaggio in quanto comportamento privilegiato dell'uomo; la premessa teorica di questi studi è che 'la struttura del linguaggio, e non solo i suoi aspetti tematici e contenutistici, sia in qualche maniera espressione delle caratteristiche psicologiche di chi parla' /Sarteschi et alii, 1968, p.1/ (17).

L'Istituto di Psichiatria dell'Università di Pisa, in collaborazione con il C.N.U.C.E., ha già presentato i primi risultati di alcune ricerche.

Una di queste aveva lo scopo di accertare, attraverso lo studio delle corrispondenze fra espressione linguistica e caratteristiche della personalità, se la struttura morfologica del linguaggio nella risposta ai tests proiettivi potesse fornire elementi integranti le informazioni ottenute per altre vie, per es. con la tradizionale siglatura. Abbiamo scelto lo psicodiagnostico di Rorschach in quanto le dieci tavole costituiscono uno stimolo altamente standardizzato e, nel contempo, per la indefinitezza delle rappresentazioni, scarsamente predeterminanti l'espressione verbale del soggetto. E' noto infatti come le espressioni linguistiche, impiegate di fronte a ogni tavola, differiscono spesso notevolmente anche quando le risposte ricevono nello psicogramma la stessa siglatura. La ricerca è stata condotta su pazienti psichiatrici (schizofrenici e neurotici) e rispettivi familiari, compiendo uno studio statistico delle categorie grammaticali tradizionali. La e

laborazione dei materiali linguistici ha richiesto due fasi distinte: la prima per lo spoglio delle risposte, che è stata condotta con le procedure in uso al C.N.U.C.E. per gli spogli lessicografici, la seconda per l'analisi statistica dei dati quantitativi forniti dallo spoglio.

Tra i risultati conseguiti, va sottolineata la stretta analogia dei profili grammaticali dei quattro sottogruppi, (schizofrenici, loro familiari; neurotici, loro familiari), che testimonierebbe la "esistenza di una rigida struttura morfologica del discorso, scarsamente modificabile anche da parte di processi psicopatologici".

Nonostante ciò, il minor numero di parole e, specialmente, la più alta percentuale di sostantivi e articoli impiegati dagli schizofrenici, permettono agli autori di prospettare l'esistenza di alcuni aspetti del linguaggio, in stretto rapporto conn le caratteristiche della personalità e del comportamento, caratteristici dei pazienti schizofrenici. La scoperta di differenze linguistiche tra diverse categorie nosologiche potrebbe essere utilizzata e per la diagnosi e per una maggior comprensione della malattia mentale.

Un altro studio del linguaggio intrapreso nella clinica di psichiatria della Università di Pisa, riguarda gli adulti psicotici (maniaci, schizofrenici) e nevrotici, e i bambini intellettivamente ipodotati. L'Analisi morfologico-lessicale del linguaggio maniacale ha per ora dimostrato eccezionale l'uso del futuro, raro il presente, più frequente il passato. Ciò pone in dubbio le reali capacità di progettazione e di infuturazione del maniaco, ravvicinando il mondo maniacale a quello del depresso, ancorato al passato.

Il decremento degli aggettivi sembra testimoniare la diminuita attitudine a qualificare o differenziare le caratteristiche della realtà.

Lo studio del linguaggio di bambini intellettivamente ipodotati è stato fatto in correlazione con lo studio di bambini di normale dotazione mentale, provenienti dallo stesso ambiente socio-economico e culturale (18).

I primi risultati mostrano che la quantità delle parole diverse in rapporto al numero totale delle parole pronunciate (come dire la ricchezza del vocabolario impiegato) è inferiore nel gruppo degli insufficienti mentali. Si è osservata anche la diminuzione degli aggettivi, in particolare la percentuale di aggettivi diver

si (19). Nei normali la percentuale degli aggettivi sul totale delle parole è risultata del 6,34%, nei subnormali del 4,61%. Tuttavia il rapporto tra frequenza dei verbi e frequenza dei sostantivi matura allo stesso modo nello sviluppo dell'insufficiente mentale e del bambino normale. Nel subnormale la difficoltà sta dunque nell'entrare in possesso del patrimonio linguistico, ma quando ciò è possibile il linguaggio verrebbe acquisito rispettando certe strutture che sono costanti e peculiari del linguaggio umano.

4.10. Dizionario di macchina della lingua italiana

Ho già esposto altrove i motivi che ci hanno indotti ad intraprendere la compilazione di un DM dell'Italiano /Zampolli, 1968 e 1969/: oltre che automatizzare e facilitare le operazioni di lemmazione in procedure di spoglio, un DM permette la completa automatizzazione della trascrizione fonematica, fornisce i dati per tutta una serie di statistiche (fonematica, etimologica, stilistica, rendimento dei suffissi e dei sistemi di derivazione, ecc.) sul vocabolario che sembrano essere il necessario complemento alla statistica sui testi, ed è uno strumento utilissimo per assicurare coerenza e comparabilità a spogli e statistiche eseguiti da ricercatori diversi in tempi diversi su testi diversi.

Rinvio appunto per questi aspetti agli scritti citati, e voglio solo aggiungere un'ulteriore aspetto del DM, quello di strumento per organizzare e gestire un archivio delle conoscenze e dei dati relativi a una lingua, che è venuto assumendo negli ultimi tempi sempre maggiore importanza, sia per i motivi connessi con i più recenti sviluppi teorici della linguistica generale e computazionale, che ho brevemente riferiti al punto 8.1., sia per motivi di ordine pratico e, per così dire, applicativo, el quadro di alcune iniziative per il coordinamento dell'insegnamento della lingua italiana, che prevedono il costituirsi di un'ampia documentazione, il DM può avere una utile funzione organizzativa in fase di compilazione e di reperimento in fase di consultazione. Siamo confortati, in questa nostra convinzione, da iniziative analoghe, come quelle a carattere lessicografico tradizionale per la lingua ceca /Stindlova 1967/, e per la lingua ungherese /Papp 1965/, e quelle ambientate nel contesto teorico del

la linguistica americana contemporanea di H.H.Jasselson per il russo e di M.Gross /1969/ per il francese. Soprattutto quest'ultimo progetto ci sembra presentare notevole affinità con il nostro, e una collaborazione si prospetta molto utile. Molto utile sarà anche la collaborazione con il Centro di studi per l'Italiano di Utrecht del Prof. M.Alinei (20), che sta ora dedicando molta attenzione alla descrizione, in una prospettiva generativo-trasformazionale, di larghi settori di lessico. Attualmente, l'Università di Pisa ci ha concesso 3 borsisti e un ricercatore per il progetto del DM: abbiamo già pronto l'algoritmo di flessione del lemmario sia a livello fonematico sia a livello morfologico, e stiamo eseguendo la preedizione del dizionario prescelto, il Migliorini-Capuccini, che verrà integrato mediante il confronto con altri dizionari (Dizionario Enciclopedico Italiano, Devoto-Oli, Garzanti, ecc.).

5. Il CNUCE ha tra i propri compiti istituzionali lo svolgimento di una intensa attività didattica, che copre i più disparati settori delle applicazioni scientifiche dei calcolatori: medicina, statistica, metodi di simulazione, linguaggi di programmazione, analisi numerica, ecc. Per la frequenza a questi corsi, che di norma durano qualche settimana, sono messe a disposizione numerose borse di studio, aperte anche agli stranieri, e i frequentatori hanno accesso ai sistemi di elaborazione durante il periodo di istruzione.

L'attività didattica del CNUCE riceverà senza dubbio un notevole incremento dal corso di laurea in Scienza dell'Informazione, che avrà inizio con l'anno accademico 1969-1970, e che è il primo del genere organizzato in Italia: tra le materie di esame è previsto anche un insegnamento denominato 'Elaborazione elettronica di testi letterari'.

5.1. L'attività didattica della nostra sezione si svolge a tre livelli:

a) I ricercatori che iniziano un loro progetto presso di noi, vengono innanzi tutto introdotti ai metodi e ai problemi dell'automazione applicata alla linguistica, per mezzo di riunioni e del lavoro in comune; spesso gli Istituti utenti specializzano una o più persone che seguono la ricerca presso di noi.

b) Svolgo, da due anni nell'ambito del C.N.U.C.E., un corso di applicazioni linguistiche dei calcolatori. I partecipanti provengono da diverse Università, e sono eterogenei per formazione

(lettere, lingue, legge, ecc.) e per qualificazione (professo ri, assistenti, studenti, documentaristi, ecc.).

Dalle discussioni collettive a conclusione del corso, è risultato che, dopo una parte introduttiva comune, sarebbe auspicabile una ulteriore specializzazione, in relazione sia alla diversa formazione accademica dei partecipanti, sia ai loro concreti diversi interessi.

Per il nuovo programma sono in contatto con alcune università statunitensi e tedesche, ove si tengono da qualche anno corsi analoghi nelle facoltà di lettere o in quelle di informatica.

c) Stiamo organizzando una scuola estiva internazionale sull'esempio recente (estate 1969) di alcune università USA, che hanno messo a nostra disposizione le loro esperienze. Il programma prevede una parte introduttiva (un linguaggio di programmazione e nozioni generali sull'elaborazione di testi naturali) e corsi particolari di statistica linguistica, di stilistica, di analisi sintattica automatica.

6. Voglio esprimere la mia riconoscenza al Prof. Delatte e ai suoi Collaboratori, che ci hanno permesso di incontrarci e di scambiare le nostre esperienze, e soprattutto per uno dei temi propostoci; la necessità di organizzare lo scambio e la diffusione delle informazioni, dei testi registrati, dei programmi, ecc. Credo che questa necessità sia emersa da alcuni punti della mia comunicazione, ed è stata proprio qui, in questo congresso, confermata dalla constatazione di quanti lavori vengono svolti e sono sperimentati indipendentemente da più d'uno tra noi. Per parte mia, quasi tutte le comunicazioni udite mi hanno richiamato progetti in corso al CNUCE, e mi hanno fatto pensare a quanta energia, e perchè no, quanto tempo e denaro si sarebbe potuto risparmiare con una migliore informazione reciproca: l'impegno reciproco in questo senso mi sembra rivestire le caratteristiche di un preciso dovere morale, se si considerano il 'costo' e lo 'sforzo' richiesto dai nostri lavori e dalle nostre ricerche, che spesso sono sostenute dal pubblico denaro, e che altrettanto spesso forniscono un 'servizio' di comune utilità a quanti sono impegnati nello studio del linguaggio, facoltà privilegiata dell'uomo.

Il CNUCE di Pisa, che rappresenta oggi un esempio riteniamo

fortunato di collaborazione tra Enti diversi e che ha realizzato, almeno nell'ambito nazionale, l'unificazione dei metodi e delle procedure e lo scambio delle informazioni, vi assicura per mio tramite il proprio impegno e la propria collaborazione.

N O T E

1) - Intendiamo con questo termine una scheda meccanografica, la quale reca stampati i dati presenti nella scheda lessicografica tradizionale (lemma o esponente, autore, opera, riferimento, data e un contesto che chiamiamo 'lungo' o 'macrocontesto' in opposizione al contesto 'corto' o 'microcontesto' delle concordanze) e perforati i dati essenziali che permettano di operare automaticamente estrazioni, inserzioni, riordinamenti dello schedario (lemma, sigle di autore, opera e data).
Abbiamo messo a punto un sistema per condensare le perforazioni nel breve spazio ad esse riservato nella scheda. Le righe del macrocontesto (al massimo 13), se stampate dal calcolatore con un intervallo di 1/8 anzichè di 1/6 di pollice, cadono esattamente nelle interlinee libere tra le perforazioni; dunque, per le schede contesto prodotte dal calcolatore, sarebbero teoricamente a disposizione per la perforazione tutte e ottanta le colonne. E' necessario però condensare le informazioni perforate in poche colonne, perchè alle schede prodotte dal calcolatore andranno fuse schede ottenute a mano o con procedimenti di spoglio xerografico, nelle quali le righe stampate si sovrapporrebbero alle perforazioni. /Duro,1968a/.

2) - "Nel maggio del 1964 l'IBM Italia metteva a disposizione della Università italiana un Sistema elettronico del tipo IBM 7090 e, su indicazione del Ministro della Pubblica Istruzione, veniva designata l'Università di Pisa quale sede del calcolatore. Nasceva in tal modo, con una convenzione tra l'Università stessa e la IBM Italia, il Centro Nazionale Universitario di Calcolo Elettronico, con sede presso l'Università degli Studi di Pisa (...).
La gestione e l'amministrazione del Centro sono affidate ad un apposito Comitato Direttivo del quale fanno parte: due professori ordinari della Facoltà di Scienze Matematiche Fisiche e Naturali dell'Università, designati dalla Facoltà stessa; due rappresentanti dell'IBM Italia, designati dalla Direzione Generale; e il Direttore del Centro che presiede il Comitato, nominato dall'Università in accordo con l'IBM Italia (...).
All'iniziativa del Centro possono partecipare tutte le Università italiane e gli Istituti di ricerca. Il tempo di impiego del Sistema elettronico è suddiviso in parti uguali tra l'Università di Pisa, la IBM e le altre Università ed Istituti partecipanti (...).
Il Sistema viene utilizzato per l'insegnamento e la ricerca; ricerca non solo orientata nei campi matematici, fisici, chimici ed ingegneristici, ma soprattutto alla applicazione delle tecniche direzionali ed alla diffusione in genere delle possibilità dell'elaborazione dei dati, con riferimento alla ricerca operativa, alla econometria e alle scienze sociali.
I risultati delle ricerche e delle elaborazioni svolte con lo impiego del calcolatore saranno oggetto di divulgazione sulle pubblicazioni scientifiche.
Infine presso il Centro Nazionale Universitario di Calcolo Elettronico vengono tenute conferenze e seminari su problemi inerenti l'elaborazione dei dati".
/Il Centro Nazionale Universitario di Calcolo Elettronico, 1964/.

3) - La produzione delle schede contesto avviene in due tempi:
a. il calcolatore le perfora
b. il calcolatore le stampa.
La stazione di lettura funzionerà durante la fase b, per controllare la corrispondenza tra dati perforati e dati da stampare.

4) - A Gallarate abbiamo adoperato per molti anni la carda-type, dotata di una IBM 866, come quella del LASLA, per la stampa di alfabetici fonetici, e dei testi ebraici del Mar Morto (Dead Sea Scroll), sfruttando la ricchezza e la possibilità di rapida sostituzione dei caratteri.
Questa soluzione non può essere adottata a Pisa, dove la velocità di stampa deve essere tenuta in grande considerazione, data la mole dei lavori in corso.

5) - Ricordo a questo proposito le concordanze fotocomposte di Tito Livio, curate da Pakard /1968/.

6) - Per un confronto tra le due forme più diffuse di consultazione (ricerca di lemmi e ricerca di forme) si veda /Zampolli 1968d/.

7) - Rinvio per questo problema al mio intervento apparso nella Revue di Liegi, in occasione del congresso di Pisa sui dizionari latini di macchina /Zampolli,1968/.

8) - Si veda per es. /Greenberg, 1957, p.27/, /Martinet 1968/ e /Gammon 1969/.

9) - Uso qui il termine matrice nel senso di /H.H.Josselson,1969/.

10) - E' in corso un esperimento di lessicografia sanscrita meccanizzata. Tale ricerca, di cui si occupa il mio assistente dott. G.Ferrari con la collaborazione dell'Istituto di Glottologia dell'Università di Pisa, viene condotta, per il momento, su alcuni testi non eccessivamente lunghi (Īśā-, Kena-, Kaṭha- e Kauṣītaki-upniṣad): ci si propone di ottenere, da un solo input, e cioè con una unica operazione di trascrizione e perforazione, due forme di trascrizione una in padapatha e una in samdhi, su cui operare separatamente, sia nell'ambito di una normale indagine lessicografica, sia in ogni altro campo dell'indagine statistica e algoritmica.

11) - Traggo queste notizie da /Tagliavini,1966/ e da /Pop,1950/.

12) - Si pensi per es. alla possibilità di eseguire conteggi statistici, per i diversi dialetti, sulle frequenze di fonemi e varianti e soprattutto sulle loro distribuzioni contestuali.

13) - Si tratta, come è noto, della più importante raccolta manoscritta di canti popolari italiani dovuta alla appassionata fatica di Michele Barbi. Iniziata intorno al 1887 come raccolta di soli canti pistoiesi, andò col passare degli anni estendendosi prima ai canti toscani e poi a quelli di tutte le altre regioni italiane; nè il Barbi si limitò a raccogliere i canti dalla viva voce del popolo, ma integrò la raccolta in loco con spogli bibliografici (numerose le stampe popolari). Tutti i dialetti e tutte le forme metriche popolari sono rappresentate nella RB; del resto un paio di esempi penso che serviranno meglio a chiarirne la

enorme importanza nel campo degli studi di poesia popolare (La povera Cecilia: 37 versioni edite, 82 nella RB; Buondì bella pastora: 26 lezioni edite, 68 RB). Questo ricco materiale è rimasto fino ad ora inedito e quindi scarsamente utilizzabile.

14) - Dice il prof. Cirese:
"Queste analisi metriche, partendo da una corretta nozione della rima, non come una qualità, ma come una relazione di identità totale o parziale delle desinenze di due o più parole, la prima delle quali costituisce la proposta, e la seconda (o le altre) la risposta di rima, si propone il riconoscimento dei gradi di omofonia che nelle possibilità combinatorie teoriche e nella pratica effettiva della tradizione orale sono molto più numerosi delle due o tre abitualmente riconosciute dalla trattatistica metrica. Si è intanto impegnati nel riconoscimento preliminare delle possibilità combinatorie teoriche, per il quale si è utilizzata anche la teoria dei grafi, ma al calcolatore sarà affidato il compito di riconoscere all'interno del quadro teorico delle possibilità combinatorie generali, i gradi e le linee di omofonia presenti concretamente nei diversi componimenti. Un tale riconoscimento richiede un numero di operazioni che sarebbe impossibile eseguire manualmente e, qui appunto, la velocità dell'elaboratore diviene condizione indispensabile per la identificazione, questa volta non più intuitiva e soggettiva, ma analitica e oggettiva dei gradi e delle rime".

15) - Per esempio, nella /Raccolta Barbi,1967/, abbiamo compilato un Rimario, nel quale per ogni parola 'in posizione di rima', sono riportate tutte e sole le parole che, nei canti esaminati sono collocate di fatto in reciproca relazione di 'proposta e risposta' di rima, a differenza di quanto si fa nei rimari tradizionali, ove sotto la rima posta in esponente si elencano tutte le parole (e relativi versi) la cui terminazione coincide con la rima in questione, e non si tiene conto del fatto se in realtà rimino o no, in qualche componimento. Evidentemente il nostro tipo di rimario è molto utile quando si studino componimenti, come quelli popolari, in cui oltre la rima perfetta, funzionano altre corrispondenze, quali per es. la assonanza, ecc.
Si genera così una serie di coppie di parole, il cui numero n è dato dalla formula $n(n-1)$, dove n è il numero delle parole in relazione, nello stesso componimento, secondo una determinata rima.

Per es., se lo schema del componimento è $a\ b\ a^1\ b^1\ a^2\ b^2\ c\ c$:

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

16) - Anche G.Lepsky /1969/ dice: "Nel campo della psicologia, e in contatto con la linguistica trasformazionale, si è avuto un profondo rinnovamento" (p.14).

17) - Del resto, anche gli studiosi generativo-trasformazionali, affermano che gli "atti di parola", la performance dei parlanti non dipendono unicamente dalla loro competenza linguistica; essi variano in funzione di un gran numero di altri fattori, come la memoria, l'attenzione, l'emotività, ecc.

18) - Lo studio dell'apprendimento del linguaggio nei bambini riveste grande importanza per la psicologia /H.Sinclair-De Zwart, 1967, p.11/, per comprendere la natura e il comportamento dell'organismo che acquisisce 'un sistema di comunicazione unico, tra quelli degli altri animali', /Menyuk; 1969, p.IX/. Esso ha sempre interessato i linguisti /Tagliavini,1966,I,n.66/ cfr. la bibliografia di /W.F.Leopold,1952/, e in particolare evidenza è messo da Chomsky e dai suoi seguaci (/Chomsky, 1969, p.30/ e /Menyuk,1969,cap.1/).
Per le relazioni tra linguistica e problema dei ritardati mentali, si veda per es. J.B.Carrol, in /Schiefelbusch et alii, 1967, pp. 39-53 e 178-181/.

19) - Altre fonti testimoniano, se pur in settori diversi, la variabilità della frequenza percentuale degli aggettivi, e indicano possibili interpretazioni.
Sembra per es. che la prosa di carattere tecnico e scientifico contenga una percentuale di aggettivi sensibilmente maggiore rispetto alla prosa letteraria e al dialogo.

20) - La collaborazione tra il CNUCE e l'Istituto del Prof. Alinei è iniziata da alcuni anni, da quando cioè l'Accademia della Crusca riceve da Utrecht la registrazione su nastro-parola dei testi italiani del '200. /Alinei 1968/.

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

Accademia della Crusca, Inni Sacri di A.Manzoni. Indici e concordanze, 1967.

Accademia della Crusca, Novella del Grasso Legnaiuolo. Testo - Frequenze - Concordanze, Firenze, 1968.

Actes du Colloque International sur La Mécanisation des Recherches Lexicologiques (Besançon, 1961), Parigi, 1962.

Alinei, M.L., Spogli elettronici dell'italiano delle origini e del duecento, L'Aia, 1968.

Alinei M., Due frammenti di grammatica italiana: 1. Il tipo sintagmatico 'quel matto di Giorgio' 2. Il nesso temporale, Comunicazione al Convegno sulla grammatica trasformazionale italiana, Roma, 1969.

Antinucci F., Crisari M., Parisi D., Analisi semantica di alcuni verbi italiani, comunicazione al Convegno sulla grammatica trasformazionale italiana, Roma, 1969.

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

Bach E., Harms R.T. (edit.), Universals in linguistic Theory, New York, 1968.

Beazley D., Attic read-figure vase painters-, Oxford.

Binnick, R.I. The Application of an Extended Generative Semantic Model of Language to Man-Machine Interactions, ICCL.

Boldrini M., Statistica, teoria e metodi, Milano, 1960.

Borsellino A., Neuroni, Percezioni, Intelligenza, in /Atti,1968/, pp. 219-226.

Bortolini U., Tagliavini C., Zampolli A., Dizionario di frequenze lessicali dell'italiano scritto contemporaneo, (in stampa).

Busa R. S.J., Les Travaux du Centro per l'Automazione dell'analisi letteraria di Gallarate, in /Actes 1962/ pp. 64-68.

Busa R. S.J., Zampolli A., Centre pour l'automation de l'analyse linguistique (C.A.A.L.), Gallarate, Les machines dans la linguistique, Praga, 1968, pp. 25-34.

Calboli G., Costrittori nelle proposizioni complemento: i modi del verbo e l'infinito, Comunicazione al Convegno sulla grammatica trasformazionale italiana, Roma, 1969.

Ceccato S., A Model of the Mind, in Computers in Psychological Research, Parigi, 1966, pp.105-175.

Chenhall R.G., "The impact of Computers on Archeological Theory", Computers and the Humanities, 3,1,1968.

Chomsky N. e Miller G.A., "Finite State Languages", Information and Control, II, 1958, pp.137-167.

Chomsky N., Prefazione all'edizione italiana dei Saggi Linguistici, vol.1, L'analisi formale del linguaggio, Torino, 1969.

Cohen M., "Statistique linguistique," Actes du VIème Congrès des Linguistes, (Parigi 1948), Parigi 1949, pp. 83-87.

Colombo A. Appunti per una grammatica delle proposizioni completive, Comunicazione al Convegno sulla grammatica trasformazionale italiana, Roma, 1969.

Convegno Maschinelle Methoden der Literaryschen Analyse und der Lexicographie, Tubinga, 1960 - Preprints.

De Mauro, T., Proposta per una teoria formalizzata del noema lessicale e della storicità e socialità dei fonemi linguistici, relazione letta nella 1^ giornata del Convegno "Linguaggi nella società e nella tecnica." Milano, 14-17 Ottobre, 1968.

De Tollenaere F., Nieuwe Wegen in de Lexicologie, Amsterdam,1963.

Devoto G., Sémantique et Syntaxe, Conférences de l'Institut de linguistique de l'Université de Paris, XI, 1952-1953, pp. 51-62.

Durand M., Spang Hanssen etc., Communications au VIème Congrès des Linguistes, Actes du VI C.L. (Parigi 1948), Parigi 1949, 87.

Duro A., La technique de dépouillement xérographique employée par l'Académie de la Crusca, Les machines dans la linguistique, Praga, 1968a, pp. 285-291.

Duro A., Spogli elettronici. Compiti dell'Accademia della Crusca per il vocabolario storico della Lingua Italiana, in L'Uomo e la Macchina, Pisa, 1968b.

Duro A., Zampolli A., Analisi lessicali mediante elaboratori elettronici, "Atti del Convegno sul tema "L'automazione elettronica e le sue implicazioni scientifiche tecniche sociali". (Accademia Nazionale dei Lincei, Roma, 1967), Roma 1968, pp. 121-139.

Erickson R.F. "Musical Analasysis and the Computer: A Report on Some Current Approaches and the Outlook for the Future", Computer and the Humanities, 3,2,1968.

Fillmore Ch.J., Toward a Modern Theory of Case, RF Project 1685-6, Columbus, Ohio, 1966.

Fillmore Ch.I., Lexical Entries for Verbs, Working Paper in Linguistics, n° 2, Columbus, Ohio, 1968a.

Fillmore Ch.J., The Case for Case, in Bach e Harmon, 1968b.

Fossier L., Recherches sur une méthode d'exploitation des sources diplomatiques médiévales, Comunicazione al"Colloque International sur la Recherche computationelle en Philologie", Liegi, 1969.

Francis A.N., Rubin G.M., Svartvik, J., A method of Computer-produced graphical representation of dialectal Variation in initial Fricatives in Southern British English. Preprint, ICCL, 1969.

Gammon E., Computational Approximation to the Word, ICCL, 1969.

Gardin, J.C., A Typology of Computer Uses in Anthropology, The Use of Computer in Anthropology, edit., Hymes, D., 1965, pp. 104-117.

Gemelli A., Zunini G., Introduzione alla psicologia, Milano, 1947.

Greenberg, J. Essays in Linguistics, Chicago, 1957.

Grassi C., Terracini S.A., Comunicazione al Convegno Internazionale sul tema "Atlanti linguistici; problemi e risultati" (Roma,1967), Accademia Nazionale dei Lincei, (in corso di stampa).

Gross M., Lexique des constructions completives, CNRS, 1969.

Grossi P., Studio di Fonologia musicale di Firenze; Attività didattica e di ricerca con la collaborazione del CNUCE, comunicazione interna, Pisa, 1969.

Gruber J.S., "Topicalization in Child Language", Fundations of Language, 3,1,1967, pp. 37-65.

Guiraud P., Les Caractères statistiques du Vocabulaire, Parigi, 1954.

Guiraud P., Problèmes et méthodes de la statistique linguistique, Dordrecht, 1959.

Hays D.G., Processing natural Language Text, "Seminar on Computational Linguistics", 1966, pp. 69-73. Bethseda.

Hays D.S. Applied Computational Linguistics, Preprints, for the Second International Congress of Applied Linguistics, Cambridge, (G.B.),1969.

Heilmann L., "Considerazioni statistico-matematiche e contenuto semantico", Quaderni dell'Istituto di Glottologia di Bologna, VII (1962-63),pp. 35-45.

Herdan G., Type-Token Mathematics, L'Aia, 1960.

ICCL, Preprints, International Conference on Computational Linguistics, KVAL, Stoccolma, 1969.

Il Centro Nazionale Universitario di Calcolo Elettronico, IBM Italia, 1964.

Josselson, H.H., "Automatization of lexicography", Cahiers de Lexicologie, 9,1966, pp.73-87.

Josselson H.H., The Lexicon: A System of Matrices of Lexical Units and their Properties, ICCL, 1969.

Juilland A., Chan-Rodriguez, E., Frequency Dictionary of Spanish Words, L'Aia, 1964.

Juilland A., Dictionnaire inverse de la Langue Française, L'Aia, 1965a.

Juilland A., Edwards, P.M.H., Juilland I., Frequency Dictionary of Rumanian Words, L'Aia, 1965b.

Kay M. "Standards for Encoding Data in a Natural Language", Computers and the Humanities, 1,5,1967, pp. 170-177.

Kay M., Ziehe T., Natural Language in Computer Form, The Rand Corporation, RM-4390, 1965.

Karttunen L., Discourse Referents, ICCL, 1969.

Kuno S., Automatic Syntactic Analysis, in Seminar on Computational Linguistics, Bethseda, 1966, pp. 19-41.

Ledley R., Use of Computers in Biology and Medicine, McGraw Hill, 1965.

Ledley R., Outline of Statistical Techniques, Applications and Programs, IBM C20-1645, 1967.

Leopold W.F., Bibliography of Child Language, Evanston, Ill., 1952.

Lepsky G., "Prefazione alla edizione italiana dei 'Saggi Linguistici' di Noam Chomsky" in Chomsky, N. L'analisi formale del Linguaggio, Saggi Linguistici, Vol. 1, 1969, pp. 9-17.

Maccacaro G.A., Colombi A., Applicazione dell'elaborazione elettronica alla Medicina, in /Atti,1968/, pp. 177-210.

Martinet A., "Mot et Synthème", Lingua, 21, 1968, pp. 294-302.

Menyuk P., Sentences Children Use, The M.I.T. Press, 1969.

Minsky M. (edit.) Semantic Information Processing, The M.I.T. Press, 1968.

Moreau, R.M., "Intervento", Statistique et analyse linguistique (Colloque de Strasburgo, 1964), Parigi, 1966.

Moreau R., Préface, in Computers in psycological research, Parigi,1966b.

Muller Ch., Essai de statistique lexicale, Paris, 1964.

Muller Ch., Fréquence, dispersion et usage: à propos des dictionnaires de fréquence, Cahiers de Lexicologie, 2, 1965, pp. 32-42.

Muller, Ch., Initiation à la Statistique linguistique, Parigi, 1968.

Orton, Harold e Eugen Dieth, Survey of English Dialects. Vol. 4, The Southern Counties, Leeds, 1967.

Pakard, D.W., A Concordance to Livy, Cambridge (U.K.), 1968.

Papp F., "Le vocabulaire du hongrois contemporain sur cartes perforées", Cahiers de Lexicologie, 7,1965, pp.103-117.

Parisi D., Intervento, in L'automazione elettronica, ecc., 1968, p.161.

Parisi D., Analisi componenziale del Lessico in Psicolinguistica in La Grammatica e la Lessicologia, Atti del I e del II Convegno di Studi, SLI, Roma, 1969, pp. 129-159.

Pop S., La dialectologie. Aperçu historique et méthodes d'enquêtes linguistiques, Lovanio, 1950.

Raccolta Barbi di Canti popolari italiani, Esperimento di elaborazione elettronica, E 1/RB, Pisa, 1967.

Reed D.W., "A Statistical Approach to Quantitative Linguistics Analysis", Word, 5,9,1949, pp. 235-247.

Rezvin I.I., Les modèles linguistiques, Paris 1968. trad. de Y.Gentilhomme dal russo Modeli Jazyka, Mosca, 1961.

Roceric-Alexandrescu, A., Fono-statistica limbii Române, Bucarest, 1968.

P.Sarteschi, Castrogiovanni P., Del Carlo G. Giannini, Maffei G., Pasquinucci P.P., Faedo A., Lytmaer N., Torrigiani G., Zampolli A., Il linguaggio nel test di Rorschach: 1.Metodologia e primi risultati di una analisi mediante elaboratore elettronico, Pisa, 1968.

Schiefelbusch R.L., Copeland R.H., Smith J.O., Language and Mental Retardation, New York, 1967.

Schlesinger, J.M., "A Note on Relationship between Psychological and Linguistic Theories", Foundations of Language, 3,4,1967, pp.397-402.

Schwarcz R.M., Towards a Computational Formalization of Natural-Language Semantics, ICCL, 1969.

Shuy R.W., "An Automatic Retrieval Program for the Linguistics Atlas of the United States and Canada", Computation in Linguistics, edit. P.L. Garvin e B.Spolsky, Bloomington, 1966.

Silva G., "An Automatic Ortographic-to-Phonetic Conversion System for French", Computer and the Humanities, 3,5,1969, pp. 257-266.

Sinclair ...-De Zwart, H. Acquisition du langage et développement de la pensée, Parigi, 1967.

Soporta S., edit., Psycholinguistics, New York, 1961.

Statistique et analyse linguistique, (Colloque de Strasburg , 1964), Parigi, 1966.

Stindlova J., "Le dictionnaire de la langue tchèque lettéraire: enregistrement des données sur cartes mécanographiques", Cahiers de Lexicologie, 10, 1967, pp. 103-113.

Szanser A.J., Automatic Error-correction in Natural Language, ICCL, 1969.

C.Tagliavini, Indici e Concordanze della Divina Commedia, Pisa, 1965.

Tagliavini C., Introduzione alla Glottologia, Bologna, 1966.

Tagliavini C., Applicazione dei calcolatori elettronici all'analisi e alla statistica linguistica,"L'automazione elettronica e le sue implicazioni scientifiche tecniche e sociali". (Accademia Nazionale dei Lincei, Roma, 1967), Roma 1968, pp. 111-118.

Titone R., La Psicolinguistica oggi, Zurigo, 1964.

Todorov T., "Recherches Sémantiques", Langages, I, 1966, pp. 5-43.

Torrigiani G., Problemi metodologici dell'analisi linguistica mediante elaboratori elettronici presso il CNUCE, Pisa, 1968.

Whatmough, J. "Selectives Variations", Actes du VIème Congrès des Linguistes (Parigi, 1948), Parigi, 1949, pp. 347-348.

Wisbey R. "Computers in Lexicography", The Use of Computer in Anthropology, edit. Dell Hymes, l'Aia, 1965, pp. 215-234.

Wood G.R., Dialectology by Computer, ICCL, 1969.

Zampolli A., Studi di statistica linguistica dell'italiano, eseguiti con impianti IBM, Tesi di Laurea, Università di Padova, 1960.

Zampolli, "Nota", Esperimento elettronico di elaborazione di canti popolari, Pisa, 1967.

Zampolli, La Sezione Linguistica del CNUCE, rapporto interno, C.N.U.C.E., Pisa, 1968a.

Zampolli, "Recherche statistique sur la composition phonologique de la langue italienne", Les Machines dans la Linguistique, L'Aia, 1968b. pp.159-176

Zampolli, Il calcolatore elettronico negli studi linguistici, Rivista IBM, Milano, 1968b.

Zampolli, "Intervento", Calcolo, V.suppl. N.2; Pisa, 1968d, pp. 109-126.

Zampolli, Appunti per l'intervento alle Giornate di Studio sul tema "La preparazione del Personale per la elaborazione automatica dei dati in Italia", AICA, Roma, 1969.

Zampolli, Studi di fono-statistica della lingua italiana, Bologna 1969, (in pubblicazione).

Zampolli, "Computer analysis of texts in Italy" negli Atti del Second International Congress of Applied Linguistics, Cambridge, 1969 (in pubblicazione).

Zipf, G.K., "Statistical Methods and dynamic philology", Language, 13, 1937, pp. 60-70.