BOZZE non corrette dall'autore

PROGETTI E METODI DELLA SEZIONE LINGUISTICA DEL C.N.U.C.E.

1. La Sezione Linguistica del CNUCE

Nel 1966 l'Index Thomisticus e l'opera del vocabolario dell'Accademia della Crusca cominciarono ad utilizzare gli elaboratori IBM del Centro Nazionale Universitario di Calcolo Elettronico di Pisa (1) per gli spogli lessicali automatici previsti dalle loro ricerche (2). Ben presto altri Istituti seguirono il loro esempio, e nel 1968 i progetti di impiego del calcolatore in diversi settori delle discipline umanistiche erano così numerosi, che la Direzione del CNUCE decise di costituire una *Sezione Linguistica* (3), la quale da un lato promuove ricerche e studi originali, dall'altro fornisce agli Utenti umanisti non solo l'uso degli elaboratori elettronici e il personale tecnico necessario per l'esecuzione delle operazioni di macchina, ma anche la consulenza scientifica, l'analisi e la stesura di nuovi programmi, quando non siano sufficienti i programmi generalizzati (*utility programs*) già disponibili.

Gli Istituti Universitari e di ricerca, italiani e stranieri, che svolgono presso il CNUCE ricerche nel settore umanistico sono poco meno di 50 (4) ed

hanno registrato e ci hanno elaborato poco meno di 50 milioni di parole in più di 20 lingue. I vantaggi scientifici ed economici di questa riunione di progetti sono notevoli : si pensi per esempio al fatto che l'adozione di uno stesso sistema generale di registrazione e di elaborazione permette di adoperare programmi già sperimentati, fa sí che ogni nuovo programma scritto per un progetto particolare sia utilizabile anche per gli altri e rende possibile lo scambio di testi tra Istituti diversi. In pratica, uno studioso può svolgere le proprie ricerche in una grande biblioteca elettronica operabile automaticamente, e non solo nel testo da lui stesso registrato. D'altro lato, la presenza di un corpus così esteso sollecita i ricercatori che intraprendono nuovi spogli ad adottare gli *standards* da noi proposti, i quali sono frutto di esperienze collettive più che decennali e comportano uno schema di registrazione e di elaborazione ben formalizzato. Esso richiede che non siano trascurate informazioni le quali, pur non essendo immediatamente necessarie per la specifica ricerca proposta, sono tuttavia indispensabili per garantire l'utilizzazione dei materiali da parte di altri ricercatori. (5)

2. Progetti degli Istituti utenti

In questa sede non mi sembra possibile elencare i progetti in corso ad una; è forse più utile indicare i principali settori di attività degli Istituti utenti. (6)

2.1 Documentazione lessicografica per la redazione di grandi dizionari storici di una lingua

Esempi : *Dizionario storico* della lingua italiana dell'Accademia della Crusca (7); Dizionario dei testi storici della eteo cuneiforme, dell'-Istituto di Glottologia di Pavia, per il Centro di Studi Egeo-anatolici del CNR. È 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 le attestazioni di una medesima voce, riunite assieme dalle fonti più disparate, l'esatta cosa è, quanto più possibile e immediata comprensione, delle informazioni riportare.

Si pensi alla varietà dei tipi di riferimento, alla necessità di unificare i segni dell'apparato critico (integrazioni, espunzioni, lacuna, ecc.) e in genere tutti i diversi modi tra i quali oscillano gli editori nel rappresentare a livello grafemico uno stesso fenomeno. Si pensi sopratutto all'esigenza di garantire coerenza di criteri nella scelta e nella formulazione del lemma, che vengono eseguite in tempi successivi da lemmatizzatori diversi su testi che appartengono a strati di lingua differenziati sia sincronicamente (stili, dialetti) sia diacronicamente.

2.2 Documentazione lessicografica per la redazione di dizionari storici di discipline particolari, o di particolari strati di lingua

A titolo d'esempio ricordo il progetto per la compilazione del "Lessico Intellettuale Europeo" (8) e quello del "Vocabolario Giuridico" della lingua italiana (9) : entrambi questi progetti seguono le tecniche e le metodologie messe a punto per l'Accademia della Crusca, e 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 del Crusca. Questi progetti presentano pertanto le difficoltà esaminate al punto precedente, con in più la complicazione di dover operare una maggiore selezione nello spoglio : infatti interessano qui soprattutto termini pertinenti lo stato o lo stile di lingua studiata che si trovano immersi nel lessico comune.

2.3 Documentazione lessicografica per la redazione lessici, indici e concordanze dell'opera omnia di un autore

Sempre a titolo di esempio, nomino gli spogli dell'*opera omnia* di Dante Alighieri, di Antonio Rosmini e di L.A. Seneca. L'Index Thomisticus svolge le sue elaborazioni al CNUCE ancora in questi giorni, ma P. Busa sta per trasferirsi per alcuni anni negli U.S.A. Egli lascerà però una copia dei materiali (nastri e schede) presso di noi. Noi ci sembra casuale il fatto che in questa categoria rientrino soprattutto filosofi e pensatori. L'indice del lessico viene considerato, mi disse più volte P. Busa, come una chiave d'ingresso al pensiero dell'autore, come un mezzo di ricerca dei luoghi

ove vengono definite o esplicitate le idee, e lo studio della evoluzione dell'uso della terminologia, è rivelatore della evoluzione del pensiero del filosofo; a tutto ciò, conviene naturalmente lo spoglio dell'opera omnia.

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

Ciascun tipo di studio meriterebbe un discorso a sé, che potete facilmente immaginare : infatti sono qui presenti esperti di ciascun settore. Mi limito perciò a citare alcune delle lingue (greco classico, latino medioevale rinascimentale e moderno, spagnolo, francese, tedesco, sanscritto (10), ebraico, aramaico, paleoslavo, italiano dalle origini ai giorni nostri, ecc.) e alcuni degli autori (Pindaro, Bacchilide, Aristotile, Petronio, Baumgarten, Kant, Goethe, Canovacci della Commedia dell'Arte, Fabbri, Machado, Gide, K. Marx, Saba, Gozzano, ecc.).

2.5 Dialettologia

Alcune ricerche recenti svolte indipendentemente in Europa e in America hanno dimostrato che la dialettologia è un campo della linguistica particolarmente aperto all'uso dei calcolatori. Interessa soprattutto la possibilità di combinare assieme tre funzioni importanti per la dialettologia : la automazione degli archivi di dati dialettali, i rilievi statistici, e la compilazione di carte e di mappe (quest'ultima tramite la "capacità" di recente acquisita dai calcolatori grazie al collegamento con diversi tipi di *plotter* et di *fotocompositrici*). (11)

In collaborazione con il *laboratorio dialettologico* dell'Università di Torino, abbiamo compilato per ricerche di dialettologia diacronica concordanze – per così dire contrastive o sinottiche – di versioni successive di testi dialettali appartenenti a periodi diversi; affrontiamo ora il problema di meccanizzare l'archivio delle risposte dialettali al questionario dell'*Atlante Linguistico Italiano* (12) – circa 9 milioni di parole in trascrizione fonetica – per trarne poi automaticamente le carte dialettali e gli indici relativi. (13)

Già si pensa a un nuovo progetto di raccolta di materiali in tutte le regioni italiane, è, i promotori ipotizzano un nuovo archivio le cui voci potranno divenire lo *heading* di una enciclopedia elettronica delle conoscenze dialettali.

Va tenuto presente a questo proposito il rinnovato interesse per la dialettologia in Italia, motivato tra l'altro, nel contesto italiano attuale di mobilità sociale e geografica di gruppi e di individui, dalla posizione privilegiata degli studi dialettologici e regionali nel quadro delle ricerche sociolinguistiche.

2.6 Ricerche demologiche

Un ottimo esempio è costituito dai progetti del *Gruppo Italiano di Studi Demologici* e del *Comitato per la "Raccolta Barbi"*, nel campo della poesia di tradizione orale (14) : una larga varietà di generi e tipi, 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 opere 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. (15)

E' evidente come il riscontro delle analogie, che era in pratica affidato alla memoria dello studioso, o a una lunga schedatura manuale, sia enormemente facilitata dalle concordanze, le quali 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. (16)

Il progetto della raccolta Barbi la quale esiste al presente come collezione di manoscritti che attende di essere edita, fornisce, a mio avviso, anche un notevole esempio di applicazione nel settore delle edizioni con l'uso di calcolatori.

2.7 Psichiatria, Psicologia, Pedagogia

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.

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. Lo studio psicologico del comportamento linguistico ha ricevuto un nuovo impulso dell'influenza chomshiana (17).

In Italia, una equipe del Centro di Psicologia del CNR a Roma situa i propri studi in questo contesto. In particolare, in accordo con i più recenti sviluppi cosidetti "postchomskiani", dirige le ricerche verso la descrizione formalizzata di alcuni settori di lessico. Queste ricerche utilizzano il materiale documentario e gli esempi prodotti dai nostri spogli (18).

La premessa teorica di altre ricerche che psichiatri e psicologi dell' Università di Pisa conducono con la nostra collaborazione su registrazioni di conversazioni (o comunque di testi orali) di malati mentali o di bambini ritardati, 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" (19).

I primi risultati, già pubblicati, hanno mostrato che, tra i normali e le diverse categorie nostalgiche, soprattutto nell'uso delle categorie grammaticali – in particolare degli aggettivi e dei tempi futuri dei verbi –, esistono differenze statistiche che potrebbero essere utilizzate per la diagnosi e per una maggiore comprensione della malattia mentale. Ad-
analoghi risultati ha portato l'analisi morfologico-lessicale del linguaggio di bambini intellettualmente ipodotati operata in correlazione con bambini di normale dotazione mentale, provenienti dallo stesso ambiente socio-economico e culturale (20).

2.8 "Information retrieval" e ricerche documentatistiche

Un esempio notevole di attività che tradizionalmente vengono raggruppate sotto l'etichetta di *information retrieval*, è constituito dagli esperimenti che l'*Istituto per la documentazione giuridica del CNR* ha iniziato presso di noi, e che dovranno in primo luogo rendere operante a livello nazionale una *banca di dati* giuridici.

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 informazione, 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. (21)

In assenza di un modello realmente funzionante di analisi del contenuto, non mancano però tecniche e procedure che sono già di grande aiuto nell'analisi di documenti. Mi riferisco per es. alla comunicazione di L. Fossier, che abbiamo sentito oggi, sull'analisi del contenuto di fonti diplomatiche medioevali.

Anche al CNUCE abbiamo in corso progetti analoghi, 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.).

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.

Stiamo anche mettendo a punto un sistema di "descrittori", inseriti nel testo, per individuare le unità referenziali, e un sistema per correlarle tra di loro.

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.

2.9 Storia dell'arte, Ricerche musicali, ecc.

L'Istituto di Archeologia dell'Università di Pisa ha progettato una elaborazione sul catalogo della ceramica greca di D. Beazley, come primo 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. (22)

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 CNUCE per lo studio dello stile di partiture, sono paragonabili allo studio statistica dello stile in poesia e nella letteratura in generale. Per la seconda il laboratorio Fonologico del Conservatorio di Firenze, ha in corso l'elaborazione di programmi per usare il calcolatore come esecutore di musica, che arricchisca la già cospicua serie degli strumenti elettronici oggi in uso nei laboratori musicali. (23)

3. Ricerche in corso

Come s'è detto, la Sezione Linguistica del CNUCE, oltre a collaborare con gli Istituti Utenti per i quali funge anche da vero e proprio Centro-Servizi (*Service-Bureau*), svolge ricerche proprie, alle quali vengono assegnati borsisti dell'Università.

3.1 Procedure e programmi di utilità

Non credo occorra aggiungere altri argomenti a proposito della importanza e della urgenza — nel contesto attuale della *computational linguistics* — di una standardizzazione o almeno della interscambiabilità dei materiali, delle procedure e dei programmi di base, temi, oltre tutto, di questo nostro convegno. Come ho detto in precedenza, ci proponiamo di confrerire ai nostri programmi e ai nostri sistemi di codificazione carattere di generalità, e la loro flessibilità è dimostrata se non altro dalla varietà delle lingue e dei testi elaborati. Possiamo affermare, con un ragionevole grado di certezza, che se un ricercatore ci propone un testo da spogliare in una lingua qualsiasi, purchè trascrivibile con un alfabeto, l'esecuzione dello spoglio e delle relative statistiche è semplice routine, dalla codificazione del testo alla pubblicazione dei risultati.

I nostri programmi si dividono in 5 gruppi principali :

a. programmi per lo spoglio lessicale di testi

A partire da *schede-testo* generano i risultati consueti degli spogli lessicali : elenchi di lemmi e forme variamente ordinate alfabeticamente, per frequenza decrescente, all'inverso : *revers-index*, ecc.; index locorum, concordanze, schede-contesto (24), incipari, rimari, ecc. Esistono anche programmi che provvedono a convertire i risultati in modo che possano essere stampati per mezzo delle tecniche di fotocomposizione.

b. programmi per l'analisi di testi a diversi livelli (fonematico, lessicale, morfologico, sintattico, ecc.) mediante consultazione di un Dizionario di macchina (= DM)

A partire da un *lemmario* un algoritmo genera le flessioni possibili nel sistema linguistico considerato, e, viceversa, a partire da un *testo*, associa a ciascuna unità del livello linguistico in esame le informazioni relative contenute nel D.M. (25).

c. programmi per elaborazioni di testi in trascrizione fonetica

Esistono algoritmi per eseguire automaticamente, con o senza un dizionario di macchina (DM), la trascrizione fonetica di un testo, a partire dalla quale altri algoritmi operano le analisi ormai tradizionali in questo settore (26).

d. programmi per conteggi ed elaborazioni statistiche

A partire dai risultati dei programmi precedenti, eseguono i conteggi e applicano le formule statistiche più diffuse. (27)

3.2 Lessico automatico della lingua italiana

Ho già elencato in occasione del Seminario "*Lexicon Electronicum Latinum*" (Pisa, 1968) i motivi che ci hanno indotti ad intraprendere la compilazione di un DM dell'italiano, e rinvio al n. di questa stessa rivista che del Seminario ha pubblicato gli Atti. Un DM non solo rende parzialmente automatica e molto più veloce la fase di *lemmatizzazione*, ma contribuisce alla *coerenza della lemmatizzazione*, che è indispensabile negli spogli di grandi corpùs (per es. grandi dizionari storici, che impegnano numerose persone per lungo tempo ed è necessaria per poter confrontare tra loro statistiche lessicali su autori diversi eseguito da ricercatori diversi.

Il DM permette anche una trascrizione fonematica automatica (tranne per i rari omofoni) richiesta da statistiche fonematiche di testi, da indici inversi, da rimari, da studi metrici, ecc., e, fornisce i dati per tutte una serie di statistiche sul vocabolario che linguisti insigni (28) hanno ritenuto di essere il necessario complemento alle statistiche sui testi : rendimento funzionale delle opposizioni fonematiche, rendimento delle diverse strutture (sillabiche, accentuali, ecc.) delle parole, proporzioni relative delle diverse provenienze etimologiche, rendimento dei sistemi suffissali e dei sistemi di derivazione, ecc.

Il DM è anche parte integrante di qualsiasi modello di analisi (e cioè di qualsiasi sistema che proponga quell'analisi sintattica automatica - parsing - o comunque quel riconoscimento della struttura delle frasi che rientra nei compiti e nei propositi di tutte le attività della cosiddetta *linguistica*

applicata computazionale. - (applied computational linguistics) - (nota) e di qualsiasi *modello di sintesi.* E' nota per esempio, la complessità delle regole di una grammatica trasformazionale quando si vogliano far interagire tra loro per descrivere compiutamente anche solo un frammento di una lingua naturale. Nel modello formale c'è un gran numero di dettagli notazionali che devono essere elaborati e generalmente il linguista oreferisce prestarvi poca attenzione. Il programma può portare all'attenzione del linguista gli errori di "scrittura".

Negli ultimi tempi si è affermata un'altra importante funzione del DM : il suo impiego nel compilare, organizzare, aggiornare, in una parola nel *gestire* un archivio delle conoscenze, un repertorio dei dati di fatto noti per una lingua. In Italia è in atto uno sforzo considerevole per la raccolta e la organizzazione di materiali in questa direzione, per colmare alcuni gravi lacune scientifiche – mancanza di una grammatica storica – di una grammatica sincronica, di una descrizione dei vari "italiani renali" ecc. e per rispondere alle urgenti richieste di sussidi didattici per l'insegnamento dell'italiano sia come lingua prima sia come lingua seconda, i quali possono essere ricavati solo da una documentazione completa e sistematica. (29)

Siamo confortati nelle nostre convinzioni 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 della 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 (30), 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 alcuni 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, che verrà integrato mediante il confronto con altri dizionari (Dizionario Enciclopedico Italiano, Devoto-Oli, Garzanti, ecc.).

3.3 Repertorio degli spogli manuali ed automatici di testi italiani

Ci si sta avviando ad una situazione di fatto nella quale è sempre più probabile che un ricercatore trovi già spogliato da altri il testo che gli interessa : spesso però l'interessato non ne riceve notizia, perchè lo spoglio resta confinato nell'Istituto che lo ha richiesto come tesi o perchè l'autore lo usa per le proprie ricerche senza avere la voglia o i mezzi di pubblicarlo.

In collaborazione con la Società di Linguistica Italiana abbiamo compilato un *questionario* che invieremo a tutti gli Istituti stranieri che a nostro avviso, potrebbero compiere o aver compiuto spogli di testi italiani, e in seguito, con procedimento meccanografici pubblicheremo i risultati dell'inchiesta.

3.4 Statistica linguistica

L'insieme dei testi sottoposti a spoglio presso il CNUCE offre alle ricerche di statistica linguistica un corpus straordinariamente vasto e una grande quantità di dati sulle occorrenze delle diverse unità del sistema linguistico. Come è noto, i cultori della linguistica quantitativa, riflettendo sui risultati degli spogli elettronici moltiplicatisi in questi ultimi anni, hanno sottoposto a esame critico le teorie formulate, forse con entusiasmo di pionieri (31), dai loro predecessori, hanno più volte affermato la necessità di verificarle induttivamente, partendo da spogli di testi e di corpus rigorosamente definiti, soprattutto per quanto riguarda il problema dei rapporti *parti di testo/testo, testi/corpus, corpus/lingua* equiparati al rapporto *campione/universo statistico* (12).

Abbiamo già iniziato ad usare i nostri materiali per ricerche in questa direzione, e per l'italiano contemporaneo stiamo conducendo in modo sistematico analisi statistiche a diversi livelli linguistici. Uscirà entro l'anno un mio lavoro sulla fonematica, è, agli inizi, un progetto sulla sintassi, è molto avanzato uno studio a livello lessicale, che si concreterà nella pubblicazione di un *lessico di frequenza dell'italiano,* che tra l'altro sopporirà, almeno in parte, alla mancanza di sussidi didattici di questo tipo, molto lamentata da quanti insegnano la nostra lingua.

3.5 Studi sociolinguistici e di italiano regionale

Le grammatiche tradizionali e moderne prendono generalmente a modello la lingua scritta e letteraria; soprattutto per l'italiano mancano sia repertori di dati sia descrizioni tradizionali o formalizzate della lingua parlata.

La lacuna è particolarmente grave data la complessità dei rapporti tra dialetti, varietà regionali e "lingua comune" (33) che caratterizzano la recente storia linguistica italiana, e dato il contesto sociolinguistico, caratterizzato da notevoli immigrazioni interne. Esse pongono a contatto modelli linguistici culturali e regionali diversi, con consequenza rilevanti sul piano sociale, per es. per l'integrazione degli immigrati nei centri industriali, e per l'adattamento e il profitto in una stessa comunità scolastica di alunni provenienti da regioni o classi sociali diverse. La moderna glottodidattica consiglia di dare carattere *contrastivo* all'insegnamento, e cioè di insistere sulla differenza tra lingua prima e lingua seconda. In molti casi l'italiano insegnato nelle scuole della penisola si presenta con caratteristiche di lingua seconda rispetto al dialetto o all'italiano regionale degli allievi.

5. Attività didattica

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) Oltre a un corso accademico, in Facoltà di Lettere, sulla linguistica matematica e computazionale, svolgo da due anni nell'ambito del CNUCE, 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 (professori, 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 hai loro concreti diversi interessi.

c) Stiamo organizzando una scuola estiva internazionale sull'esempio recente di alcune università USA, che hanno messo in 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.

Voglio esprimere la mia riconoscenza al Prof. Delatte e ai suoi collaboratori che ci hanno permesso di incontrarci e di scambiare le nostre esperienze, 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 in questo congresso confermata dalla constatazione di quanti lavori vengano svolti e siano 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 sarebbero potuti rispamiare con una migliore informazione reciproca : l'impegno reciproco in questo senso mi sembra

rivestire le caratteristiche di un preciso dovere morale. Il CNUCE, esempio, riteniamo almeno per ora, fortunato di collaborazione tra Enti diversi, 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.

Centro Studi IBM-Pisa

A. ZAMPOLLI

NOTES

1) Nel maggio del 1964 l'IBM Italia metteva a disposizione dell'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 il 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.
All'iniziativa del Centro possono partecipare tutte le Università italiane e gli Istituti di ricerca.
La convenzione è stata rinnovata quest'anno e prevede un IBM 360/67.

2) In Italia il primo progetto di lessicografia automatizzata è stato quello dell'Index Thomisticus, iniziato da R. Busa S.J., a Gallarate nel 1949 (cfr. Busa, R., Zampolli, A., *Le Centre*, 1968). Nell'anno accademico 1959-60 fu discussa all'Università di Padova la tesi di laurea di A. Zampolli, *Studi di Statistica Linguistica eseguita con impianti IBM*, (cfr. Zampolli, A., *Recherche* ..., 1968), sotto la direzione C. Tagliavini, che costituisce il primo saggio di applicazione dei calcolatori allo spoglio e all'analisi di un testo italiano a livello fonematico, lessicale e morfologico (cfr. Tagliavini, C., *Applicazione*, 1968). Nel 1964 furono eseguiti sotto la direzione di A. Duro i primi esperimenti di spoglio per il grande dizionario storico della Lingua Italiana affidato dal CNR all'Accademia della Crusca

(cfr. Duro A., Zampolli A., *Analisi*, 1968). Nel 1965, in occasione dell'inaugurazione del CNUCE a Pisa, venne offerta al Capo dello Stato una copia delle concordanze e degli indici lessicali della Divina Commedia, elaborati elettronicamente per iniziativa della IBM Italia a cura di C. Tagliavini. Nel 1966 l'Index Thomisticus e l'Accademia della Crusca affidarono l'elaborazione elettronica dei rispettivi progetti di spoglio del CNUCE; il loro esempio e il successo delle applicazioni suscitò presto altri progetti, cosicchè oggi oltre 50 Istituti Universitari del CNR di tutta Italia e di alcuni paesi europei si avvalgono dègli impianti, della collaborazione tecnica e scientifica e dei programmi di utilità della Sezione Linguistica del CNUCE, costituita nel 1968 sotto la direzione di A. Zampolli.

3) La *Sezione Linguistica* conta oggi 5 programmatori-operatori, 3 operatrici-perforatrici, una segretaria, 4 ricercatori con laurea in discipline umanistiche. Attualmente vi è riservato a pieno tempo un IBM 360/30 64K dotato di 4 unità nastro, 2 unità disco, 1 lettore-perforatore, una stampatrice veloce 1403/n.1 (1100 righe al minuto) e una stampatrice/lettrice di schede 1404. Quest'ultima è destinata in particolare alla stampa di *schede contesto*. Entrambe le stampatrici sono corredate di catene speciali a 120 segni, da noi progettate e richieste all'IBM per diversi alfabeti : ci sono pervenute finora catene con alfabeto greco e latino.
I caratteri sono distribuiti negli *slugs* della catena a diverse altezze relative, in modo che è possibile combinarli tra loro usando il dispositivo di *space-suppression* : eliminando a programma la spaziatura verticale automatica tra riga e riga, si può battere più volte sulla stessa linea.

Così per es. i caratteri a . _ fisicamente distinti nella catena, possono essere combinati assieme per formare il carattere a̲
Queste apparecchiature dovrebbero risolvere il problema della stampa di elaborati di controllo e di tabulati definitivi a tiratura limitata che sarà possibile riprodurre in *offset* (Tagliavini, C., *Divina Commedia,* 1955; Zampolli, A., *Nota tecnica,* 1967; Accademia della Crusca, *Inni Sacri*, 1967; sono in corso le concordanze non lemmatizzate dell'opera omnia di Seneca). La moderna tecnica della fotocomposizione è però preferibile (Hays, *Processing,* 1966) poichè 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. Nel dicembre del 1968 abbiamo stampato, in collaborazione con l'Accademia della Crusca, un primo volumetto di indici e di concordanze di una novella anonima del 400, che è il primo esperimento nel suo genere in Italia e, a quanto sappiamo, in Europa. Alla fine dell'anno pubblicheremo indici, concordanze, rimario, ecc. del Canzoniere del Petrarca.

4) Il *Rapporto* periodico del CNUCE contiene anno per anno un breve resoconto dei lavori in corso e del loro progresso.

5) Si dice spesso 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 i quali il sistema ottimale *non* è quello standard, a motivo, per restare nell'esempio scelto, di una diversa distribuzione di frequenza dei grafemi; si conclude che la standardizzazione va operata 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. (Kays, M., *Natural Language*, 1965 e *A system*, 1967).

Noi siamo fondamentalmente d'accordo con tutte queste osservazioni, 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.*

Il nostro *programma di carico* è munito di una routine generalizzata di conversione, che assegna le diverse informazioni contenute nelle schede a categorie funzionali distinte. La standardizzazione è resa possibile proprio da una classificazione dei grafemi e delle informazioni generalmente 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 successive elaborazioni dell'intera procedura.

6) I raggruppamenti sequenti (da 2.1 e 2.10) non vogliono in alcun modo essere una classificazione, ma corrispondono solo a raggruppamenti di comodo — nell'esposizione — delle attività di fatto in corso

al CNUCE. E' noto come le classificazioni fino a qui proposte di quella congeria di attività che gli anglosassoni raggruppano sotto il termine *computational Linguistics* non abbiano dato risultati completamente soddisfacenti. Ricordo quelle di S. Lamb, una in *Language* del 1961 e una nel volume *The Use of computer in Anthropology* del 1965; quella di J. Gardin nello stesso volume; quella di Garvin in *Language* del 1962; quella di C.A. Montgomery presentato all'ICCI del 1969 e quella di D. G. Hays presentato alla *II/TH International Conference on Applied Linguistics* del 1969.

7) 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 sospesa 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). 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, innanzi tutto, *un tesoro delle origini della lingua italiana,* nel quale confluirà gran parte del materiale documentario relativo al periodo delle origini, fino alla data convenzionale del 1375. Seguirà poi il grande *Vocabolario Storico* che, assorbendo in sé, convenientemente diradato, il materiele del Tesoro, proseguirà a fare la storia della lingua italiana fino ai giorni nostri. Si tratta di mettere assieme decine di milioni di schede, estratte da non meno di 20.000 volumi.

8) "The notion of an intellectual lexicon has become more precise in the most recent linguistic studies; this notion designates the section of a lexicon of a language concerning aspects of the intellectual life of individuals or groups; this section also includes the so-called culture words, that is, words which have been remarkable in the history of thought, in that they express notion which are essential in one current of thought or in another". (*Lessico Intellettuale Europeo,* p. 1).

9) Si vedano gli articoli di P. Fiorelli citati nella bibliografia.

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 (ISA - Kena -, Katha - e Kausitaki - upnisad) : ci si propone di ottenere, da un solo input, e cioè con un'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) Ricordo i lavori di W. N. Francis, J. Svartik, G. M. Rubin della Brown University sui materiali registrati nel 4° volume del *Survey of English Dialects;* di G. Cassidy all'Università di Wisconsin per il *Dictionary of American Regional English*; di R. W. Shuy per l'*automatic retrieval* dell'Atlante Linguistico degli USA e del Canada, e di G. R. Wood per la "scoperta" e lo studio dei diversi tipi di americano regionale lungo la costa atlantica.

12) 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), 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.
(Si veda il *Bollettino dell'Atlante Linguistico* pubblicato a Torino e, inoltre Tagliavini, C., *Introduzione,* 1966 e Pop, *Dialectologie,* Bolelli, *Per una storia,* 1965).

13) 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 "tipicizzare", 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
  d) publicare, prima ancora delle carte, 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 parola 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 fotocomposizione 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).

14) Si tratta, come è noto, della più importante raccolta manoscritta di canti popolari italiani dovuta all'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.

15) Cfr. Cirese, AM, *Note per una nuova indagine,* 1965.

16) Accanto alle forme "canoniche" della rima della consonanza e della assonanza, si intuisce un tessuto più sottile e una varietà di rispondenza foniche accettate dalla coscienza popolare, che sono in parte da enunciare e da inventariare. Lo strumento classico è il rimario, che noi abbiamo sviluppato in forme diverse.
Per esempio 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 componenti, 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 $a^{1}\ b^{1}\ a^{2}\ b^{2}\ a\ b\ c\ c$:

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

17) "Nel campo della psicologia, e in contatto con la linguistica transformazionale, si è avuto un profondo rinnovamento". (Lepschi, G., *Prefazione*, 1969, p. 14).

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 che prendono i loro esempi presso gli scrittori.

Le scuole generative osservono 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 ammesse, sarebbero considerate dai parlanti come facenti parte della lingua". (Ruwet, *Introduction*, 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 letterari, ecc.), che permettono di affrontare lo studio delle 'competenza' " (Ruwet, ivi, p. 18).

19) Abbiamo scelto lo psicodiagnostico di Rorschach in quanto le dieci tavole costituiscono uno stimolo altamemente 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. (Sarteschi e alii, *Il linguaggio nel test*, 1968).

20) L'analisi morphologico-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 testimoniara la diminuita attitudine a qualificare o differenziare le caratteristiche della realtà.
I primi risultati dello studio per i bambini intellettualmente ipodotati 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.
Nei normali la percentuale degli aggettivi sul totale delle parole è risultata del 6,34%, nei subnormali del 4,61%.
Il rapporto tra frequenze dei verbi e frequenze dei sostantivi matura allo stesso modo nello sviluppo dell'insufficiente mentale a del bambino normale. Nel subnormale la difficoltà starebbe dunque nell'entrare in possesso del patrimonio linguistico, ma quando ciò è possibile il linguaggio verrebbe acquisito rispettando certe strutture che sono costanti e peculari del linguaggio umano.

21) Analizzare in questa sede le difficoltà connesse a questo problema, equivarrebbe in buona parte 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, 1966. 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 e di Schwartz, all'ICCL, 1969, 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 "*Syntaxe et sémantique*" di G. Devoto, 1953).
Oltre ai gruppi già citati di Parisi, si hanno altri studi in questo settore, per es. a Bologna (cfr. le comunicazioni Calboli e Colombo nel convegno della SLI 1970. Mi preme rilevare come studi di questo tipo conducano a un recupero 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.).
I problemi che generalmente vengono riferiti alla semantica, avevano 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, *automatic Syntactic*, 1966 e Bar-Hillel, *Four Lectures*, 1964). Oggi, non indipendentemente forse dal risveglio degli studi della cosiddetta "semantica generativa" negli USA dopo la stasi beaviorista e distribuzionalista, *anche la linguistica applicata* ritorna ad affrontare il problema della semantica in una prospettiva nuova (certo molto più ambiziosa e forse utopistica, per le difficoltà teoriche di una formalizzazione della descrizione semantica - De Mauro, *Nota*, 1970) descritta criticamente da D. G. Hays, *Applied Computational*, 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.

22) Le applicazioni della elaborazione automatica all'Archeologia hanno avuto un notevole sviluppo. Sono ben note le classificazioni di queste attività proposte da Gardin, 1955), e un quadro aggiornato della situazione è fornito dai *Preprints* per il "Colloque international sur l'emploi des calculateurs en archéologie : problèmes sémiologiques et mathématiques" organizzato dal CNRS francese a Marsiglia lo scorso aprile.

23) 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 la 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, *Rapporto interno*, 1969).

24) Intendiamo con il termine scheda-contesto 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).

25) Si veda Duro A., Zampolli A., *Analisi lessicali*, 1968.

26) Si vedano per esempio :
Kramsky, J., *Quantitative Phonemics*, 1962
Juilland, A., *Dictionnaire inverse*, 1965
Kucera, H., Monroe, G.K., *A comparative Phonology*, 1968
Roceric Alexandrescu, A., *Fonostatistica*, 1968
Wood, G.R., *A computational analysis*, 1962
Zampolli, A., *Statistica fonematica dell'italiano*, 1970.

27) Cfr. le opere di Guiraud, Herdan, Muller, Juilland, ecc.

28) Cfr. : Tabliavini, C., *Applicazioni*, 1968
Troubetzkoj, N., *Principes*,
Martinet, A., *Eléments*, 1960
Dubois, J., *Utilisation*, 1964.

29) Cfr. l'articolo di Lepsky, G., nel 1° numero del *Bollettino* del Centro per l'insegnamento dell'italiano agli stranieri di Trieste.

30) 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 dell '200. (Alinei, M., *Spogli elettronici*, 1968).

31) 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 précisément 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 supposait donc que l'on pouvait assimiler le choix d'un mot au tirage d'une boule dans une urne dont la composition reste inchangée au cours du temps. Le (...) CREDIF (...) a fait une première brèche dans cette croyance en introduisant sa notion de mots disponibles. Mme Hirschberg et Ch. Muller ont montré que, sauf cas particulier, cette règle de jeu n'était pas vérifiée" (Moreau, R., *Intervento,* 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 une période d'expansion, voire d'excès, il vient un moment d'équilibre. Les linguistes savent à présent qu'il faut tenir compte des critères quantitatifs; ils ignorent dans quelle mesure il faut les considérer comme pertinents, en fonction des champs spécifiques d'études" (*Statistique et analyse linguistique,* 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.

32) Si vedano a questo proposito le pagine di Ch. Muller nella *Initiation à la statistique linguistique,* 1968; quella di Moreau, R., nel n° 3 dei *Cahiers de lexicologie,* 1962, pp. 140-158.

33) Si vedano a questo proposito i capitoli III° e IV° in De Mauro, *Storia Linguistica dell'Italia Unita,* 1970.

BIBLIOGRAFIA

Accademia della Crusca

- *Indici e concordanze degli Inni Sacri di A. Manzoni,* Firenze, 1967.

- *Novella del Grasso Legnaiuolo, Testo, Frequenze, Concordanze,* Firenze, 1968.

Alinei, M. L.,

- *Spogli elettronici dell'italiano delle origini e del duecento,* L'Aia, 1968.

Bar-Hillel, Y.,

- "Four Lectures on Algebraic Linguistics and Machine Translation", *Language and Information,* Gerusalemme, 1964, pp. 185-218.

Bach E., Harms R. T. (edit),

- *Universals in Linguistic Theory,* New York, 1968.

Binnick, R. I.,

"The application of an Extended Generative Semantic Model of Language to Man-Machine Interactions", *ICCL.*.

Bolelli, T.,

- *Per una storia della ricerca linguistica, Testi e note introduttive,* Napoli, 1965.

- *Bollettino dell'Atlante Linguistico Italiano,* Torino.

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 trasformazionele italiana, Roma, 1969.

Chomsky, N.,

- *Syntactic Structures,* L'Aia, 1957.

- "On certain formal Properties of Grammars". *Information and Control,* 1959, 2, pp. 137-167.

- *Aspects of the Theory of Syntax,* Cambridge (Mass.), 1965.

Chomsky, N., e Halle, M.,

- *The Sound Patterns of English,* New York, 1968.

Cirese, A. M.,

- *Note per una nuova indagine sugli strambotti delle origine romanze, della società quattrocentesca e della tradizione moderna,* in "Giornale Storico della Letteratura Italiana ", CXVIV, fasc. 445, pp. 1-45; fasc. 448, pp. 491-566.

C.N.U.C.E.,

Rapporto, 1965-1969.

Colombo, A.,

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

Francis, W. N., Rubi, G. M., Svartivik, J.,
- A Method of computer-produced graphical representation of dialectal variation in initial fricatives in Southern British English, *ICCL.*

Gardin, J. C.,
- "La mécanisation de certaines recherches historiques, à partir de textes préalablement analysés", *Cahiers de Lexicologie,* 3, 1962, pp. 177-184.

- "A Typology of Computer Uses in Anthropology", *The Uses of Computers in Anthropology,* edit. Hymes, D., 1965, pp. 104-117.

Garvin, P. L.,
- "Computer partecipation in linguistic research" *Language,* Vol. 38, 4, 1962, pp. 385-389.

Gross, M.,
- "Les modèles en linguistique", *Languages,* 1968.

Guiraud, P.,
- *Les Caractères Statistiques du Vocabulaire,* Parigi, 1954.

- *Problèmes et méthodes de la statistique linguistique,* Dordrecht, 1959.

Hays, D. G.,

- *Processing Natural Language Text,* in "Seminar on Computational Linguistics", 1966, pp. 69-73.

- *Applied Computational Linguistics.* Preprints for the Second International Congress of Applied Linguistics. Cambridge (G. B.) 1969.

Herdan, G.,

- *Type-Token Mathematics,* L'Aia, 1960.

- *The Advanced Theory of Language as Choice and Chance,* Berlino, 1966.

ICCL,

- *Preprints, International Conference on Computational Linguistics,* KVAL, Stoccolma, 1969.

Josselson, H. H.,

- "The Lexicon : A Sistem of Matrices of Lexical Units and their Properties", ICCL.

Juilland, A.,

- *Dictionnaire inverse de la Langue Française,* L'Aia, 1965.

Juilland, A., Chan-Rodriguez, E.,

- *Frequency Dictionary of Spanish Words,* L'Aia, 1964.

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

Kay, M.,
- "Standards for Enconding 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.

Kramsky, J.,
- "Quantitative Phonemics in Last Decade", *Phonetica,* 8, 1962, pp. 166-185.

Kucera, H., Monroe, G. K.,
- *A Comparative Phonology of Russian, Czech and German,* New York, 1968.

Kuno, S.,
- *The predictive analyzer and a Path Elimination Technique,* Communication of the ACM, 8, 1965.

Lakoff, R. T.,
- *Abstract Syntax and Latin Complementation,* The M.I.T. Press, 1968.

Lamb, S. M.,
- "The digital computer as an Aid in Linguistics", *Language*, vol. 37, N.3, 1961, pp. 382-412.

- "Linguistic Data Processing", *The Use of Computers in Anthropology*, edit. Hymes, D., 1965, pp. 159-188.

Lepsky, G.,
- "*La Linguistica strutturale*, Torino, 1966.

- "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.

- *Lessico Intellettuale Europeo*, Roma, Edizione dell'Ateneo (catalogo)

Martinet, A.,
- *Eléments de Linguistique générale*, Parigi, 1960.

De Mauro, T.,
- *Storia linguistica dell'Italia Unita*, Roma, 1970.

Convegno 1970.

Migliorini, B.,
- Che cos'è un vocabolario ? , Firenze, 1951.

Minsky, M., (edit.)
- *Semantic Information Processing*, Cambridge (Mass.) 1968.

Montgomery, C. A.,
- "Linguistics and Automated language Processing", *ICCL*, 1969.

Moreau, R.,
- "Au sujet de l'utilisation de la notion de fréquence en linguistique", *Cahiers de lexicologie,* 3, 1962, pp. 140-158.

- "Intervento", *Statistique et analyse linguistique* (Colloque de Strasbourg, 1964), Parigi, 1966.

Muljacic, Z.,
- *Fonologia generale e fonologia della lingua italiana,* Bologna, 1969.

Muller, Ch.,
- "Le mot, unité de texte et unité de lexique, en statistique lexicologique", *Travaux de Linguistique et Littérature,* I, Strasburgo, 1963, pp. 165-175.

- "Fréquence, dispersion et usage : à propos des dictionnaires de fréquences", *Cahiers de lexicologie,* 2, 1965, pp. 32-42.

Nickel, G.,
- Geschichte und Leistung des Taxonomischen Strukturalismus. *Zeitschrift für Dialektologie und Linguistik,* 1, 1969, pp. 2-18.

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

Parisi, D.,

- "Analisi componenziale del lessico in psicolinguistica", in *La grammatica e la lessicologia,* Roma, 1969, pp. 129-151.

Pop, S.,

- *La dialectologie. Aperçu historique et méthodes d'enquêtes linguistiques;* Lovanio, 1950.

Roceric- Alexandrescu, A.,

- "Recensione di : Juilland, A., ecc. Frequency Dictionary of Rumanian words", *Revue Roumaine de Linguistique,* IX° n. 2, 1966, pp. 205-210.

Ruwet, N.,

- *Introduction à la grammaire générative,* Parigi, 1967.

Schwarcz, R. M.,

- "Towards a Computational Formalization of Natural-Language Semantics", *ICCL,* 1969

- *Statistique et analyse linguistique* (Colloque de Strasbourg, 1964), Parigi, 1966

Stindlova, J.,

- "Le dictionnaire de la langue tchèque littéraire : enregistrement des données sur cartes mécanographiques", *Cahiers de lexicologie,* 10, 1967, pp. 103-113.

Shuy, R. W.,

- "An Automatic Retrieval Program for the linguistic Atlas of the United States and Canada", *Computation in Linguistics* edit. Garvin e B. Spolsky, Bloomington, 1966.

Tagliavini, C.,

- *Introduzione alla Glottologia,* Bologna, 1966.

- *Concordanze della Divina Commedia,* Pisa, 1965.

- *Applicazione dei calcoli elettronici all'analisi e alla statistica linguistica,* in "Atti del Convegno sul tema", l'automazione elettronica e le sue implicazioni scientifiche, tecniche e sociali (Accademia Nazionale dei Lincei, Roma 1967), Roma 1969, pp. 111-118.

Todorov, T.,

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

Trubetzkoy, N. S.,

- *Principes de Phonologie* (trad. franc. di J. Cantineau), Parigi, 1949.

Wood, G.,

- *Dialectology by Computers,* ICCL, 1969.

Zampolli, A.,

- *Studi di statistica linguistica eseguiti con impianti IBM,* Tesi di Laurea, Padova, 1960.

- Recherche statistique sur la composition phonologique de la langue italienne, *Les machines dans la linguistique,* L'Aia, 1968.

- "Nota" in *Esperimento elettronico di elaborazione di canti popolari,* Pisa, 1967.

- Projet d'un dictionnaire italien de machine, *Calcolo,* V, suppl. n° 2, Pisa, 1968, pp. 109-126.

- *Studi di fono-statistica della lingua italiana,* Bologna, 1970 (in pubblicazione)

Zipf, G. K.,

- *The Psycho-biology of Language,* 1935 (riedito da G.A. Miller, Cambridge, Mass., 1965 e 1968).