A. Zampolli

DUE CONVERSAZIONI SUL PANORAMA ATTUALE DELLA LINGUISTICA COMPUTAZIONALE

Tenute a Firenze presso l'Istituto di Matematica Ulisse Dini, nel mese di Giugno 1969.

1. In questa conversazione non intendo formulare un giudizio critico sulle ricerche condotte nei settori in questione; mi propongo invece di fornirvi alcune informazioni che permettano una prima rappresentazione intuitiva del loro contenuto. Esporrò dati e giudizi raccolti da studiosi direttamente impegnati in queste ricerche; spesso infatti si tratta più di progetti e di studi preliminari che di realizzazioni avanzate.

I problemi linguistici, teorici e particolari, sollevati direttamente o indirettamente sono numerosissimi, e sarebbe impossibile esaminarli in questa sede ogniqualvolta vengono coinvolti, spesso con termini gravemente approssimativi e inesatti dagli studi che passeremo in rassegna. In un certo senso, le informazioni che esporrò daranno un quadro sommario di questi settori più come li vedono i loro cultori, i quali sono di provenienza varia e di attendibilità diversissima, che come essi si presentano oggettivamente. Però dato il poco tempo disponibile, un simile approccio può forse più efficacemente servire per una rapida informazione. I settori in questione sono in rapida e disordinata espansione, mancano strutture informative sufficienti e riconosciute; per questo la nostra scelta degli esempi, degli autori, dei progetti citati è episodica, non risponde a un giudizio di valore, ha valore, per così dire, di aneddoto.

Si parla spesso indifferentemente di linguistica applicata (LA), linguistica matematica (LM), linguistica computazionale (LC), linguistica algebrica, linguistica algoritmica, ecc.: non voglio qui definire l'area pertinente a ciascun termine, ma solo mostrare le sovrapposizioni parziali, con le quali si spiega la molteplicità delle designazioni che vengono applicate di volta in volta alla stessa ricerca.

2. Linguistica Applicata.

LA è stato per lungo tempo, soprattutto negli Stati Uniti d'America, "un termine elegante per l'impiego di idee e di scoperte della linguistica nell'insegnamento delle lingue"/138 p.1/ Anche in Italia è spesso usato come sinonimo di glottodidattica. Per la natura stessa dell'insegnamento, "una glottodidattica scientifica si suppone integrata con gli apporti di tutte le scienze di confine, dalla linguistica - che ne è la matrice formante - all'antropologia, alla psicolinguistica". /14 p.3/.

Gli sviluppi delle dottrine linguistiche, "determinati dal profondo rinnovamento e ripensamento di una teoria generale del linguaggio", /142 p.V/ hanno profondamente interessato studiosi di altre discipline, i quali, dovendo operare sul linguaggio,sia pure partendo da prospettive diverse e perseguendo scopi diversi, hanno sentito l'esigenza sempre più profonda di rinnovare i propri metodi, e ricercano nelle grandi costruzioni teoriche i fondamenti più sicuri, e una più adeguata conoscenza del complesso fenomeno lingua. "Se, fino a qualche tempo fa, le indagini e le speculazioni degli specialisti rimanevano nel chiuso delle scuole linguistiche, oggi esse vanno sempre più diventando patrimonio culturale comune". (ibidem). Jean-Paul Vinay ha paragonato la linguistica applicata al diritto anglosassone : LA (...)c'est une réalité flottante qui s'appuie sur des précédents: c'est en somme l'art d'accomoder les postulats des linguistes". /320 p.698/.

Il linguista osserva il comportamento dei soggetti parlanti. Egli è portato a mettere a punto delle tecniche descrittive che permettono di rendere conto della struttura delle lingue descritte. Spesso i linguisti hanno deluso gli specialisti di altre discipline, o meglio questi hanno sopravvalutato il grado delle no-

stre conoscenze del fenomeno lingua, probabilmente perchè ne sottovalutano la complessità. In alcuni settori, hanno reagito tentando in proprio di procurare le conoscenze mancanti, come è successo ad esempio per la scienza delle comunicazioni, per la foniatria e per la traduzione meccanica. Del resto non pochi lessicografi hanno dichiarato esplicitamente l'inadeguatezza per i loro scopi delle teorie della linguistica, con particolare riguardo alla semantica. /266/. Nel lavoro quotidiano al C.N.U.C.E., questo tipo di esperienze mi occorre non di rado.

Il caso limite è forse quello di alcuni psichiatri i quali hanno proposto di consegnarci le conversazioni registrate di alcuni loro pazienti, chiedendoci di analizzarle statisticamente e di comunicare loro, quasi con un referto di laboratorio, "se e in cosa il loro campione differiva dalla norma". Altri si attendono che la linguistica definisca e identifichi in modo univoco e operativamente applicabile nei testi le unità che intendono spogliare (fonemi, parole, frasi, ecc.).

Le esigenze e le richieste non sono restate sempre senza conseguenza. M.Joos in un articolo che egli stesso definisce speculativo, riconosce ai compiti della linguistica applicata la funzione "di stimolo dello sviluppo della teoria linguistica negli USA", al punto che conoscere la natura di questi stimoli è indispensabile per comprendere la direzione degli sviluppi teorici /163 p.11/ degli ultimi trenta anni.

Numerosi settori della linguistica applicata ricorrono all'uso dei calcolatori, e pertanto ricorrono anche a quell'insieme di metodologie e di tecniche per la elaborazione di materiali linguistici che sono il nucleo della linguistica computazionale.

"Quando le preoccupazioni pedagogiche dei linguisti si sono cristallizzate verso il 1950 sotto il vocabolo assai generale di Linguistica Applicata, ci si è accorti che si delineavano due grandi tendenze - osserva J.P. Vinay nel saggio citato - che si sono sviluppate parallelamente ma che si devono distinguere accuratamente. Per alcuni la LA dovrebbe occuparsi esclusivamente dei problemi dell'insegnamento, per altri si tratterebbe di applicare le teorie attuali a delle macchine elettroniche che permettono il trattamento automatico dell'informazione, la traduzione automatica, la protesi uditiva, la sintesi della parola " /320 p.699-700/.

L'Associazione Internazionale di Linguistica Applicata, concepita nel 1963 a Stoccolma, si è costituita definitivamente all'occasione del Congresso di Nancy del 26-31 Ottobre 1964,("I^{er} Colloque International de linguistique appliquée") /4 pp.7-9/ organizzato attorno a due nuclei d'interesse : 1. l'automazione in linguistica, 2. la pedagogia delle lingue viventi. Il primo è un tema specifico della linguistica computazionale, e fu articolato nei settori della traduzione meccanica, della documentazione automatica e della linguistica quantitativa. Il secondo riguardava le teorie linguistiche e le loro possibilità di applicazione all'insegnamento, e fu articolato in tre punti : a) strutture lessicali e insegnamento del vocabolario, b) strutture sintattiche fondamentali e loro insegnamento, c) problemi pedagogici della traduzione. Il calcolatore era utilizzato anche da progetti esposti nel secondo gruppo, sia per simulare modelli di competenza linguistica e verificare le regole relative, sia per la sua "abilità" statistica, come sussidio nelle grandi inchieste sulle lingue parlate e sulle lingue di specializzazione /269/.

L'impostazione data al II° Congresso Internazionale di LA che

si tiene nel Settembre di quest'anno a Cambridge (Gran Bretagna), ha accentuato questa plurivalente caratterizzazione della LA, ed è descritto come un Congresso interdisciplinare che interessa : linguistica, psicologia, sociologia, insegnamento delle lingue, gestione dell'istruzione, neurofisiologia, logopedia e foniatria, scienza dell'informazione, ingegneria delle comunicazioni, lessicografia. /289 p.1/.

Il programma prevede 15 sezioni : /290/ 1) Linguistica applicata a testi letterari, 2) analisi computazionale di testi, 3) ricerche nella psicologia dell'apprendimento della prima lingua, 4) ricerche nella psicologia dell'apprendimento della seconda lingua, 5) ingegneria delle comunicazioni, 6) tecnologia dell'insegnamento delle lingue, 7) materiali per l'insegnamento delle lingue, 8) aspetti linguistici dei disordini e della terapia del linguaggio, 9) lessicografia, 10) materiali per il 'test' linguistico, 11) misura e classificazione degli errori nella seconda lingua, 12) teoria della traduzione, 13)linguistica contrastiva, 14) sociolinguistica, 15) generalità di LA.

Le sezioni 2 e 5 hanno come oggetto specifico l'uso del calcolatore per la elaborazione di dati linguistici, e le comunicazioni della sezione 9 (lessicografia) sono tutte centrate su problemi e i metodi delle 'nuove vie' /79/ che il calcolatore apre alla lessicografia. Ma in quasi tutte le altre sezioni vengono descritti progetti che si avvalgono della elaborazione elettronica dei dati. Molte comunicazioni riguardano l'utilizzazione,nell'insegnamento, delle tecniche dell'istruzione programmata basata sul calcolatore /290p.15/; si va affermando una nuova concezione del "Laboratorio linguistico" (sezione 6 : /290 p.24 e p.67/ nel quale il calcolatore ha un ruolo di coordinatore. Nelle comunicazioni conclusive (sez. 15) si prospetta la funzione del calcolatore in termini molto generali; per es. H. Richter afferma : "La linguistica dipende dai documenti linguistici. Spogliare i documenti è tanto costoso quanto lungo. Progetti di razionalizzazione su larga scala avrebbero successo limitato, senza il contributo delle tecniche dell'elaborazione (elettronica) dei dati". /290 p.73/. E' interessante confrontare queste proposizioni con quelle di Hays a proposito degli scopi generali e delle prospettive future della linguistica applicata computazionale, che qui sotto riferiremo.

3. Linguistica matematica.

" 'Linguistica matematica' - ha scritto G.Lepsky - è una etichetta che ha almeno due valori nettamente distinti. Uno di 'linguistica quantitativa', in cui entrano considerazioni numeriche, e che potremo chiamare in senso lato 'statistica linguistica' ; l'altro di linguistica per così dire 'algebrica' o 'algoritmica', tale cioè che nel corso dell'analisi si usano simboli e si compiono operazioni su simboli che appaiono a volte raccolti sulla pagina in 'formule matematiche' : in questo secondo senso il termine 'matematica' è usato come nell'espressione 'logica matematica'; si tratta di un senso completamente diverso da quello quantitativo o statistico." /194 p.190/.

Ciononostante, oltre che nella denominazione, le due materie sono riunite in numerose trattazioni. E' il caso per esempio della sezione che la Bibliografia Linguistica pubblicata dall'Unesco /25/ dedica alla linguistica matematica, e di alcune sintesi apparse nell'ultimo ventennio, come quelle di P. Guiraud /126/, J. Whatmough /326/, A.G. Oettinger /233/, J. Plath Warren /25/, F.Papp /245/, F.Kiefer /179/, e soprattutto nell'ottimo libro di O. S.

Akhmanova, I.A. Mel'chuk, et alii /7/, "Exact Methods in Linguistic Research". Proprio questo titolo ci indica un motivo, un denominatore comune che origina l'accostamento dei due settori. La caratteristica comune a questi studi è l'introduzione, nelle ricerche linguistiche, di "metodi esatti", che, nelle scienze, "sono inseparabilmente uniti alle matematiche". "L'essenza di questa direzione e del suo contenuto reale consiste non nel creare un nuovo speciale genere di 'linguistica', ma piuttosto nel perfezionare, rendere accurati, attendibili, e moderni, i metodi della ricerca linguistica nel senso usuale del termine",/7-VII°/ Nella maggior parte delle sintesi citate, un altro elemento ritenuto comune e fondamentale è l'impiego del concetto di 'modello'.

3.1. Una analisi accurata della funzione del concetto di modello e di simulazione in linguistica è dovuta a I.I. Revzin, nel suo "Modeli Jazyka", dalla cui introduzione riassumiamo le considerazioni seguenti. /264 cap. I°/. Non esistono delle scienze puramente induttive o puramente deduttive, anche se può essere ragionevole distinguere le scienze deduttive dalle scienze induttive, secondo il metodo che predomina nella scienza considerata. Per la sua stessa natura, la linguistica deve utilizzare innanzi tutto i metodi induttivi; essa descrive dei fatti concreti di parola, riferendosi a delle lingue naturali. D'altra parte, ci si può domandare se l'esistenza di un insieme infinito di atti di parola, studiati dai linguisti, permetta di formulare le nozioni fondamentali della lingua grazie a una generalizzazione per induzione. Per ottenere un sistema di nozioni generali, che permettano di interpretare i dati ricavati con l'analisi delle lingue naturali, i linguisti hanno bisogno non solo di metodi induttivi, ma anche di metodi euristici deduttivi. Inoltre, le applicazioni tecniche nuove della lingua esigono la messa a punto di sistemi deduttivi rigorosi. Nella sua parte deduttiva, la linguistica può essere - sembra - costruita al modo della matematica e della logica: cioè enucleando un numero minimo di nozioni prime, non definibili, a partire dalle quali si definiscono tutte le altre. Le numerose controversie degli ultimi anni sulla linguistica matematica, si riportano nella loro essenza al problema seguente : la scienza del linguaggio può essere costruita in questo modo? Revzin non vuole dare una risposta definitiva; ma osserva che attualmente non si vede bene come far rientrare nel quadro di una teoria deduttiva rigorosa tutta la molteplicità dei fatti anche contraddittori della realtà della parola, come sono descritti in numerose grammatiche di lingue naturali. Questa conclusione non lo induce al pessimismo. La scienza possiede oggi uno strumento potente per studiare i fenomeni complessi del mondo reale: il metodo di simulazione. "Esso consiste nel costruire una serie determinata di schemi astratti capaci di costituire una approssimazione più o meno buona dei fatti concreti". La simulazione, nel senso in cui la intende Revzin, "è un metodo secondo il quale lo sperimentatore, a partire dai tratti più generali delle lingue naturali, formula certe ipotesi sulla struttura della lingua, in tanto che sistema semiotico astratto, e in seguito stabilisce i rapporti tra le conseguenze di queste ipotesi e i fatti delle lingue reali come le descrivono le discipline linguistiche concrete". E' necessario, cioè, esaminare prima dei modelli generali, non legati alle particolarità delle lingue naturali; questo in seguito permetterà di simulare dei gruppi particolari di lingue e fors'anche delle lingue naturali. (Sulla scorta delle precedenti affermazioni dubito che questi modelli generali previsti da Revzin rientrino nella

controversia sugli universali del linguaggio: non si afferma infatti che essi appartengono alle lingue naturali; hanno piuttosto un valore e una funzione metodologica ed epistemologica inerente al processo descrittivo. E infatti, più oltre, Revzin, aggiunge che il modello non è una parte del linguaggio concepito come sistema, ma appare come una costruzione scientifica ipotetica, un edificio fabbricato. In effetti, questo punto merita di essere ulteriormente chiarito, ma non è qui il luogo di farlo. Basta osservare che il metodo dei modelli è innanzitutto uno strumento scientifico per avvicinarsi alla conoscenza dei fenomeni, e non, o almeno non essenzialmente, una spiegazione totale dei fenomeni. /192 p.248-250/. Anche quando tutte le conseguenze del modello combaciano con la realtà, il modello non spiega per questo solo fatto le connessioni reali dei fatti. Ciò è tanto vero che uno stesso insieme di sequenze corrette può essere ottenuto da due modelli sintattici diversi: una stessa lingua può essere descritta da grammatiche diverse. /264 p.7/.

3.2. Un'altra chiara discussione del concetto di modello è nelle pagine di S.K. Saumian comprese nel libro Problèmes du Langage /284/, ora tradotto in italiano, e,soprattutto nel polemico articolo di F. Kiefer Some aspects of mathematical models in linguistics, /180/. Innanzitutto, egli dice, si può pensare a una applicazione diretta di metodi matematici, già esistenti, a problemi linguistici, cioè l'applicazione di tecniche e risultati di una o più discipline matematiche. I metodi matematici sarebbero direttamente applicati alla elaborazione di dati linguistici in alcuni campi relativi alla statistica linguistica e alla teoria dell'informazione.

Più importante - a suo avviso - l'applicazione indiretta di metodi matematici, e cioè l'adozione del modo di pensare usato finora nelle matematiche e del modo di costruire una teoria matematica (con la sua struttura assiomatica, definizioni esatte, argomentazione deduttiva, ecc.). Nel caso di applicazione diretta dei metodi matematici egli parlerebbe piuttosto di matematica applicata che di LM; ma non appena è riconosciuta "la similarità tra struttura del linguaggio e alcuni tipi di strutture che sono state previamente studiate dalla matematica (la LM) appare una disciplina linguistica a proprio diritto: la LM applica metodi analoghi a quelli che sono abituati per lo studio delle proprietà e delle caratteristiche di ogni struttura matematica. La LM è allora generata non introducendo qualche genere di 'conteggio' nelle ricerche linguistiche, ma usando un concetto interamente nuovo del linguaggio". /180 p.8/.

(Anche se, come ho prima dichiarato, mi limito in questo discorso a una semplice rassegna informativa, debbo far notare la differenza di impostazione teorica tra queste affermazioni, e quelle per es. di A. Kulagina, e di N. Chomsky, qui sotto riportate).

In conseguenza a questo nuovo concetto, la LM costruisce dei modelli che sono concepiti come uno strumento efficace per l'esatto trattamento dei dati linguistici, da un lato, e la base per una teoria linguistica generale, dall'altro.

"By a mathematical linguistic model we shall understand a more or less exactly formalized system, trowing light upon one or more aspects of language" /180 p.9/.

La formalizzazione sembra essere una delle nozioni più importanti nella costruzione del modello. Ogni segno, secondo Kiefer, ha una funzione eidetica e una operativa. Parliamo della prima,

"se consideriamo ciò che è significato, denotato dal segno"; la seconda "si riferisce invece alle operazioni alle quali il segno può essere sottoposto." "Formalization in linguistics means that we consider only the operative functions of the language signs and neglect their eidetic functions" /180 p.13/. Ma il processo di formalizzazione è più difficoltoso in linguistica che in matematica: le due funzioni del segno linguistico, eidetica e operativa, non sono univocamente separabili. Molti modelli formali per la linguistica sono costruiti con la procedura logica seguente : "we ascribe to the linguistic units meaningful signs, thus,signs having both operative and eidetic functions." In seguito, si astrae dalla loro funzione eidetica, si trascurano alcune irregolarità; siamo allora in grado di operare con i segni, che dispongono solo delle funzioni operative (ivi).

La formalizzazione è indispensabile nella costruzione dei modelli, e presenta numerosi vantaggi: assicura la univocità terminologica; rende possibili rappresentazioni esatte e più economiche, e controllabili e riproducibili più semplicemente; inoltre, quando si lavora con sistemi formali possono diventare evidenti analogie tra diverse aree di ricerca.

La formalizzazione non deve essere fine a se stessa. I risultati ottenuti in un sistema formale devono essere 'interpretabili eideticamente' /180 p.14), anche se in un secondo tempo. In questo modo la LM può essere considerata come una 'teoria sulla interpretabilità linguistica di modelli formali' : il grado di adeguatezza con cui le lingue naturali possono essere descritte da modelli formali è allora il problema cruciale della LM in quanto disciplina. (ivi.).

Con una seconda eccezione alla condotta acritica che mi sono proposto, ricordo che questo punto diviene perciò stesso cruciale anche nella discussione tra le correnti linguistiche contemporanee : in effetti sono proprio le stesse inevitabili circoscrizioni metodologiche dei sostenitori dei modelli - il modello non spiega la realtà ma è solo un mezzo di conoscenza; esso tratta un sistema astratto in se stesso, non vuole considerare più di un preciso unico delimitato aspetto del fenomeno, staccato dagli altri, e anche nel trattare questo aspetto esso trascura deliberatamente irregolarità minori che possono non essere in accordo, (cfr. lo stesso Kiefer, /180 p.12/ e Revzin, /264 p.4/ - a far sorgere la domanda : ma allora, in che modo il modello ha qualche presa sulla realtà? Non è forse inutile, se non porta addirittura a risultati inesatti e fuorvianti?

I modelli linguistici possono essere classificati da diversi punti di vista. La classificazione tradizionale si basa sulle discipline matematiche che prevalgono nella loro costruzione: si può parlare di modelli linguistici di teorie degli insiemi, di logica matematica, algebrici, statistici, e di teoria dell'informazione. Kiefer non condivide questo criterio di classificazione.Essi possono in definitiva essere ridotti a due tipi (cfr. la suddivisione, sopra citata, di G. Lepsky) : 1) modelli algebrici che posseggono le proprietà stabilite per il concetto di modello, 2) modelli statistici che le hanno solo in parte, e quindi sono modelli in senso diverso. Essi non sono rappresentabili come un sistema formale; è meglio parlare di "teorie" anzichè di modelli statistici.Molto è stato scritto sulla elaborazione statistica di dati linguistici, ed è di grande interesse distinguere ciò che contribuisce ad una teoria e ciò che può essere considerato solo una applicazione

di tecniche statistiche a sussidio di ricerche linguistiche. Senza dubbio, la statistica linguistica rappresenta un altro approccio alla struttura del linguaggio, complementare a quello algebrico, nel quale ultimo secondo Kiefer /180 p.25/ confluiscono sia i modelli della teoria dell'insieme, sia quelli logico-matematici, sia quelli algebrici propriamente detti.

3.3. Questo schema di classificazione è sostanzialmente accolto in gran parte dei lavori che, con maggiore o minore estensione, propongono un panomara sintetico della LM, fin dai primi saggi; lo troviamo infatti già nell'articolo di A.G.Oettinger "Linguistics and Mathematics" del 1957 /233/. Egli dà innanzitutto una rappresentazione intuitiva del concetto di modello, che egli giudica centrale ai fini di comprendere il possibile contributo della ML.

La parola modello -egli dice - ha molti sensi. Il senso specializzato nel quale è qui usato è vicino, secondo Oettinger, a quello definito dal Webster come "a miniature representation of a thing; sometimes, a facsimile". L'idea chiave è quella di 'rappresentazione', di astrazione, ma un buon modello gioca anche un ruolo più attivo. Oettinger fornisce un esempio semplice di questo tipo di utilità.

Per stabilire la migliore disposizione di due tavoli o di due sedie in un ufficio, li possiamo spostare qua e là nella stanza fino a che si trova una disposizione soddisfacente. Un'altra alternativa è disegnare una 'piantina' del pavimento della stanza, e collocarvi i modelli dei tavoli e delle sedie fino a trovare una soluzione soddisfacente. In questo modello le distanze e le dimensioni in scala sulla carta stanno per le distanze e le dimensioni nella stanza, e il cancellare e il disegnare tavole e sedie nel diagramma sostituiscono gli spostamenti concreti. Per quanto concerne la configurazione, lo spazio, le proporzioni, ecc., il modello riflette adeguatamente la stanza reale. D'altronde è importante osservare che il modello non vuole rappresentare certi fattori come la luce, la disposizione della finestra, ecc. Si è scelto che il modello sia una mappa del pavimento, e non si può attendere che dia informazioni su questi altri aspetti della stanza ai quali non corrispondono elementi del modello.

La scoperta che la sostituzione degli oggetti concreti con un diagramma, e delle operazioni su questi oggetti con operazioni sul diagramma, può dare risultati utili e risparmiare tempo e fatica, fornisce una buona giustificazione per l'uso del modello, diffusa per es. in ingegneria. Ma modelli di questo tipo, che servono per la sintesi, appaiono di valore limitato per la linguistica, che è una disciplina in primo luogo storica e descrittiva. (Oggi una utilizzazione di modelli sintetici, produttivi, è più diffusa; si pensi ad es., alla verifica di regole della grammatica generativa, alle applicazioni della traduzione meccanica (MT) per la produzione del testo nella lingua di uscita e dell'information retrieval (IR) per la produzione di 'risposte' alle richieste di documentazioni).

Oettinger sottolinea l'aspetto rappresentativo del modello, e la sua funzione esplicativa, che pure con alcune riserve, lo rendono utilizzabile in linguistica.

Il modello in scala di una casa non è certo la casa stessa. Per es., è possibile metterlo su un tavolo, è possibile operarvi, ma certamente le persone non possono abitarci.

Esso è utile, però, perchè rappresenta certe caratteristiche scelte della casa. Anche se le dimensioni non sono quelle della

casa, le proporzioni relative di tutte le dimensioni sono conservate. La composizione dei vari materiali edili e i dettagli figurativi possono essere imitati così da rendere l'idea di come la casa apparità. Così i modelli giuocano un ruolo descrittivo, e permettono la valutazione di certe qualità di una casa, o il confronto tra quelle di molte, prima della costruzione reale. Spesso le equazioni matematiche sono usate in circostanze analoghe. A questo livello può essere considerato il modello del rapporto rango-frequenza delle forma di un testo proposto da Zipf ($f = \frac{K}{r}$ dove f = frequenza; r = rango; K = costante); questa formula può essere una previsione, più o meno esatta, del rapporto che osserveremo in testi reali via via sottoposti a spoglio. E' naturale, a questo punto, se la formula è in maggiore o minor grado verificata, ricercare le ragioni per le quali il modello 'funziona' in circostanze diverse.

E' questa la utilità che Oettinger attribuisce al modello quantitativo in linguistica: di richiedere una spiegazione. Il modello è usato per 'spiegare', sovrapponendo una struttura a un sistema concreto, una struttura definita dagli elementi del modello e dalle relazioni tra essi, anche quando non è possibile una verifica diretta della esistenza indipendente di elementi e relazioni concreti corrispondenti. L'esame delle interpretazioni che della legge di Zipf vengono proposte, con riferimenti alla teoria della teoria dell'informazione di Shannon /290/,da Mandelbrot /200/ e /201/, Belevitch /24/,e soprattutto da Herdan /144/, lo porta a raccomandare la prudenza nella interpretazione dei modelli matematici in genere. La questione cruciale è "quale grado di realtà può utilmente riflettere il modello?" /233 p.184/. L'uso di un modello matematico non ben interpretato, è particolarmente pericoloso (capita, egli dice, che un matematico sia oscuro o vago, o che usi parole comuni in senso tecnico senza una spiegazione adeguata), in vista "della facilità con la quale i linguisti spesso abdicano al loro senso comune di fronte alla autorità di argomenti matematici supposti esatti". /233 p.185/ (E queste considerazioni non sono certo, ancora oggi, fuori luogo!) Solo se le limitazioni reali, come pure il potere dei modelli matematici sono chiaramente compresi dal linguista,la matematica può essere usata effettivamente a servizio della linguistica.

Dopo i modelli quantitativi e probabilistici, egli passa nel campo della linguistica algebrica. Egli propone una connessione tra il problema (evidentemente posto in una prospettiva distribuzionalista) della segmentazione nella trascrizione 'fonemica' e la descrizione (di Schutzenberger /287/), delle condizioni alle quali una sequenza discreta di simboli elementari ('letters'), ha una unica scomposizione in gruppi di questi simboli ('words'). Egli cita appena i primi tentativi di Chomsky /50/ 261) di applicare le tecniche dell'algebra e della logica matematica nella ricerca di modelli capaci di "generare tutte le frasi grammaticali di una lingua (....) e capaci anche di caratterizzare la struttura interna di tali frasi" /233 p.185/.

Egli nota altresì le feconde prospettive di applicare questi modelli a problemi della MT e della IR.

3.4. Anche Warven Plath, nella sintesi documentatissima (la bibliografia occupa sei pagine e ad essa ci si può riferire per il periodo precedente il 1960) 'Mathematical linguistics'del 1962, attribuisce parte dello sviluppo della linguistica al successo e alla potenza di modelli discreti nella pratica rappresentazione

e nella descrizione teorica delle proprietà formali del linguaggio. "Se i modelli linguistici sono considerati come sistemi astratti di elementi discreti, può essere loro applicata una varietà di concetti e di tecniche matematiche, dall'idea elementare di numero a complesse operazioni logiche, statistiche, e di teoria degli insiemi". /253 p.22/. L'idea che la semplice aggiunta di numeri o di operazioni matematiche agli elementi del sistema rendano le affermazioni espresse nei termini del sistema più 'esatte' o 'scientifiche' è errata. Ciò che si deve verificare, è se il nuovo sistema così ottenuto sia in qualche modo un modello più soddisfacente del sistema originale, o perchè permette affermazioni teoriche più semplici o generali (cfr. le considerazioni, qui sotto riportate, di Chomsky) sulla materia "modellata", o perchè le operazioni sul modello danno nuova penetrazione e comprensione dei risultati di corrispondenti operazioni nel settore "modellato". Warren Plath ribadisce le considerazioni di Oettinger sui pericoli insiti nella costruzione di modelli linguistici, soprattutto quantitativi; oltre alla conoscenza delle pertinenti branche della matematica, sono richieste una profonda competenza linguistica e la comprensione della natura del problema linguistico cui il modello è applicato.

3.4.1. Modelli statistici.

Anche nella sua rassegna vengono primi i modelli statistici /pp.24-38/ : egli distingue i modelli macrolinguistici (come esempio dei quali discute la legge di Zipf e le diverse interpretazioni proposte) dalle statistiche per lo studio dello stile, e per le attribuzioni d'autore. In contrasto con l'enfasi posta dai modelli macrolinguistici sulle analogie generali nelle distribuzioni di frequenza, le statistiche su singoli testi avrebbero lo scopo radicalmente diverso di trovare misure statistiche che possano servire a definire differenze stilistiche tra opere di diversi autori. Questi criteri quantitativi vengono applicati per risolvere problemi quale la incerta attribuzione, la cronologia relativa di opere di uno stesso autore, e la descrizione degli stili letterari in generale, nella speranza dichiarata che detti criteri portino ad un giudizio più oggettivo. L'uso del termine 'stile letterario' è flessibile, e va dalla caratterizzazione di una singola opera (o parte di una opera) di un singolo autore,a quella di più opere di un autore, di un gruppo di autori, di un intero periodo. Warren Plath rileva che la descrizione dello stile coinvolge considerazioni "di forma e di contenuto" / p.28 /, e può andare dalla discussione delle configurazioni sonore in poesia all'esame della disposizione degli argomenti in una intera opera di prosa, e, conclude che un concetto così indefinito e generale non può essere preso come 'un tutto', e ridotto direttamente in termini matematici. Ciò che è possibile, è trattare quantitativamente un particolare attributo stilistico, di solito sulla base della frequenza relativa di forma linguistiche di un tipo specifico. Egli cita gli studi di Guiraud /128/ di Yule /332/ di Herdan /144/ e /145/ e di Fucks /102/.

3.4.2. Per quanto concerne i modelli di teoria della informazione, egli ricorda il problema della non applicabilità della teoria dell'informazione alle ricerche semantiche; come già avevano osservato Joos e Oettinger /233/, la misura dell'informazione non si riferisce al contenuto semantico dei messaggi, ma opera esclusivamente con la struttura statistica delle loro rappresentazioni formali. Egli espone alcune applicazioni del concetto di entropia e di ridondanza alla fonetica: la prospettiva della natura binaria

dei tratti distintivi proposta da Jakobson e dai suoi collaboratori /48/, rende particolarmente conveniente l'impiego del 'bit', l'unità binaria di informazione selettiva. Greenberg, Osgood e Saporta /120/ hanno proposto, come misura della efficenza di un sistema fonematico dato, il rapporto fra il numero dei tratti distintivi di fatto usati e il numero minimo di distinzioni binarie necessarie per codificare univocamente ciascun fonema.

3.4.3. Particolarmente accurata e documentata è la sua esposizione dei metodi quantitativi nella linguistica storica e comparativa (pp.33-38). Egli rifà la storia di questa applicazione, riferendosi alle comunicazioni di J.Whatmough, che riporta come primo esempio in ordine cronologico il lavoro sulla metrica Vedica di Arnold (1905), il quale ha incorporato semplici criteri di frequenza come aiuto nel determinare la cronologia relativa di varie porzioni del Rigveda /15/. Warren Plath non prende posizione nella polemica che proprio in quegli anni divampava tra sostenitori e critici dei metodi della glottocronologia e della lessicostatistica. Osserva però che l'introduzione dell'uso del calcolatore, avrebbe avuto benefici effetti: da un lato avrebbe lasciato liberi i linguisti di concentrarsi sui problemi puramente linguistici insiti nella questione,dall'altro avrebbe permesso di portare fino in fondo i calcoli per un controllo estensivo, prerequisito essenziale per una accurata valutazione dei risultati.

3.4.5. Modelli strutturali del linguaggio
Teoria degli insiemi

Plath (pp.38-40) menziona le applicazioni nel settore fonematico di Harary e Paper /131/ per lo studio delle relazioni sintagmatiche in una prospettiva distribuzionalista, e di Ungeheuer /311/ per l'applicazione della logica booleana nella prospettiva binarista Jacobsoniana. Segue una breve esposizione del modello sviluppato da O.S. Kulagina. In questa teoria ci sono tre nozioni fondamentali date assiomaticamente, non definite:parola, frase, famiglia di parole e di frasi. Il linguaggio è definito come un sistema dell'insieme di parole, di frasi grammaticali,di famiglie.

La teoria di Kulagina può essere divisa in due parti: la prima, attraverso la relazione di equivalenza, arriva a suddivisioni dell'insieme di parole, chiamato tipo, che coincide più o meno con la nozione di parte del discorso della grammatica tradizionale.

Nella seconda parte, definisce le cosiddette configurazioni, che sono uno strumento per analizzare e descrivere la struttura delle frasi. Secondo Kiefer /179 p.20/ il modello di Kulagina è un modello di tipo analitico, e può essere visto, almeno in parte, come il corrispondente 'inverso' del modello 'sintetico' proposto da Chomsky con la grammatica trasformazionale. Questa fondamentale equivalenza, secondo Kiefer (ibid) sarebbe dimostrata dal carattere algebrico dei più recenti lavori, ad es. quello di S.Marcus /203/. Plath discute poi le divergenze e le equivalenze tra la terminologia dell'algebra glossematica presentata da Uldall, e quella della teoria degli insiemi. /149/.

3.4.6. La sezione riservata da Plath ai modelli sintattici,è divisa in due parti. La prima (modelli per la sintesi e la descrizione sintattica delle frasi) esamina soprattutto tre diversi modelli proposti da Chomsky in Sintactic Structures /51/; la seconda (modelli per l'analisi sintattica), esamina quei tipi di modelli che Kiefer /179 p.21-22/ chiama modelli logico-matematici,non

perchè trattino dei linguaggi artificiali, che secondo lo stesso Kiefer hanno poco più in comune con quelli naturali del nome 'languages', ma perchè cercano di interpretare alcuni fenomeni linguistici in un modo che è simile a quello seguito dalla logica matematica nell'esaminare linguaggi formali (artificiali). Plath cita i modelli proposti da Bar-Hillel /20/ Lambek /191/ Rhodes /267/ Oettinger /235/.

Osservo solo che, secondo la dimostrazione data da Chomsky /56/ e da alcuni suoi collaboratori, /123/ queste teorie, malgrado le divergenze terminologiche siano considerevoli, sarebbero delle varianti delle grammatiche sintagmatiche classificate da Chomsky stesso. Ciò varrebbe sia per le grammatiche categoriali proposte dai logici (per es. Bar-Hillel e Lambek), sia per l'analisi predittiva proposta soprattutto dagli ingegneri in vista della traduzione automatica. (cfr. Ruwet /281 pp.164-165/).

3.4.7. Lo schema di Plath è concluso da un breve capitolo (pp.51-52) dedicato alla LM applicata.

La presenza di questa categoria è motivata dal fatto che alcuni modelli e tecniche della LM hanno avuto importanti applicazioni fuori della linguistica in se stessa: (si vedano le osservazioni di G.Lepsky a questa separazione della LA dalla linguistica 'pura' /194 pp.189-190/) per es. i metodi statistici di analisi linguistica applicata alla crittografia e alla decifrazione di lingue ignote.

Nel campo della TM e dell'IR sono stati applicati anche modelli quantitativi, come per es. quello di Luhn che ha trattato i problemi dell'abstract automatico con criteri paragonabili a quelli per lo studio statistico del vocabolario letterario /199/, ma indubbiamente i modelli più usati sono quelli strutturali.

Plath conclude con l'osservazione che la collaborazione tra matematica e linguistica, è intimamente legata all'uso dei calcolatori.

3.5. Sintesi successive seguono pressapoco lo stesso schema di classificazione. E' il caso per es. del capitolo sulla LM di M.Ivic /159/, diviso in tre sezioni : statistica linguistica, teoria delle informazioni, traduzione meccanica (nella quale è appena accennata la parte strutturale, che è trattata come capitolo a sè). La bibliografia è anche qui ricca, anche se il capitolo sulla LM non è certo al livello del libro nel suo assieme.

Anche nell'Europa Orientale, il termine LM è usato ambiguamente; sembra però accettato generalmente che la LM copra tre campi di ricerca (F.Kiefer /180 p.1/) :
1 - linguistica statistica, 2) - modelli non - statistici del linguaggio in generale, grammatica trasformazionale in particolare; e, infine 3) - matematica linguistica applicata (traduzione automatica, linguistica computazionale, ecc.). (Lo stesso F. Kiefer discute e rifiuta altrove dettagliatamente come abbiamo visto la classificazione dei diversi aspetti della LM.)

Anche il libro di F.Papp /245/ "Mathematical Linguistics in the Soviet Union" ha tra capitoli corrispondenti :
1) - Statistical modelling (pp.62-79)
2) - Structural modelling (pp.80-99)
3) - Machine translation (pp.99-126)

L'autore aggiunge una quarta sezione : Other applications of mathematical linguistics (pp.127-142) nei quali parla di applicazioni dei metodi e dei risultati della LM all'insegnamento della lingua, alla traduzione non automatica, a questioni di ortografia e traslitterazione, ecc.

3.6. A conclusione di questa breve elencazione, solo due brevi osservazioni.

La distinzione, spesso una opposizione, tra linguistica e LM; come si delineava prima del decennio che sta per concludersi, sembra risolversi diversamente per i tre grandi settori tradizionalmente identificati nella LM.

3.6.1. La linguistica algebrica, (o dei modelli strutturali), sopratutto nei suoi aspetti di grammatica formale, viene oggi accolta, o almeno discussa, come una parte intrinseca della linguistica teorica (cf. per es. i cap. 4 e 6 della Introduction to Theoretical Linguistics di J. Lyons /198/, e il capitolo sulla grammatica trasformazionale nella Linguistica Strutturale di Lepsky /194 pp.175-188/), anche se, almeno in una prospettiva genetica, è ricollegata alla LM.

Nella introduzione del libro di Maurice Gross e André Lentin "Notions sur les grammaires formelles" /59/, Noam Chomsky analizza le ragioni della coincidenza dello sviluppo della linguistica algebrica e dell'interesse redivivo per la grammatica universale. "Nello studio tradizionale del linguaggio - egli dice - si faceva una importante distinzione tra 'grammatica particolare' e'grammatica universale'; la prima si occupava delle proprietà di lingue specifiche, la seconda degli aspetti generali di ogni linguaggio umano." L'interesse di questa distinzione è risorto negli ultimi anni in coincidenza con l'inizio di serie ricerche in "quel ramo della linguistica matematica che è talora chiamata 'linguistica algebrica', per distinguerla dalla investigazione statistica del linguaggio, o dallo studio di modelli probabilistici della performance".

La coincidenza di questi sviluppi sarebbe del tutto naturale sotto molti aspetti. La linguistica algebrica è lo studio delle proprietà formali del linguaggio naturale, astratte dalla realizzazione specifica in lingue particolari. Così definita,essa è pressoché indistinguibile dalla grammatica universale,anche se in pratica c'è stata una divisione di compiti tra gli studi di grammatica universale orientati più empiricamente, e gli studi strettamente matematici delle strutture formali suggerite dalle ricerche linguistiche. Ma può emergere una vera teoria linguistica matematica, la quale studi in modo puramente astratto la classe dei sistemi definiti dai principi della grammatica universale, la classe 'dei possibili linguaggi umani'.

Un altro interesse - dice Chomsky - era presente nella tradizione classica della linguistica, e poi si era perduto; la proprietà del linguaggio che può essere chiamata il suo 'aspetto creativo', cioè il fatto che una lingua provvede meccanismi ricorsivi che permettono l'espressione di diversi infiniti pensieri convinzioni intenzioni, ecc."Certamente almeno una ragione, forse la ragione maggiore, di questo slittamento di interessi (....) sta nel fatto che mancavano tecniche valide e comprensive di processi formali, che rendessero possibile studiare i meccanismi ricorsivi della sintassi in modo chiaro e produttivo." Infatti è solo durante la scorsa generazione che concetti appropriati sono stati sviluppati e formalizzati nel corso delle ricerche fondamentali della matematica (vedi anche /124/p.11).

Con la nuova comprensione dei meccanismi ricorsivi e della natura degli algoritmi che è stata sviluppata negli ultimi 30 anni - dice Chomsky - diviene possibile tornare a studiare l'aspetto creativo dell'uso di una lingua, e mirare a formulare, in modo preciso, i meccanismi che ogni linguaggio mette a disposizione per

l'uso libero e illimitato del linguaggio.

Lo studio di questo meccanismo è ora di solito riferito come grammatica generativa. La grammatica generativa è intesa come un sistema di regole e di procedimenti, che caratterizzano la classe potenzialmente infinita delle frasi di una lingua naturale, e che assegna a ciascuna di queste una descrizione strutturale che rappresenta le sue proprietà significative, fonetiche, sintattiche e semantiche. Così lo studio della grammatica generativa diventa fattibile come risultato degli sviluppi in matematica, e non è pertanto sorprendente che l'interesse per le proprietà formali delle grammatiche, nella linguistica algebrica, sia un naturale sviluppo di questo nuovo approccio allo studio del linguaggio.(Sto sempre riassumendo il brano citato di Chonisky).

Lo scopo che si propone la linguistica moderna è di costruire delle grammatiche per lingue naturali che siano interamente esplicite in forma di automa, o di grammatica formale, o di costruzione algebrica. Esigere che le descrizioni abbiano tali forme è interamente giustificato dal punto di vista teorico.Nel corso dei 15 ultimi anni un certo numero di studi importanti, effettuati con l'idea di raggiungere delle descrizioni formalizzate su basi empiriche solide, ha mostrato che i diversi sistemi combinatori di tipo sintagmatico presentano dei numerosi inconvenienti quando si cerca di descrivere dei fenomeni sintattici in maniera significativa. Questi inconvenienti spariscono in parte quando si utilizza la tecnica di descrizione conosciuta sotto il nome di analisi trasformazionale /124 P.185/.

3.6.2. Diversa è la situazione per la statistica linguistica (modelli quantitativi), che, nonostante alcune affermazioni dei suoi cultori (per es. /271/) non è riuscita ad acquisire - ci sembra - una propria autonomia. Le sue origini sono spesso ricercate molto lontano nel tempo: si risale ai conteggi degli Alessandrini, e attraverso gli indici delle edizioni francesi dei classici 'ad usum delphini', si giunge, con l'esplosione dei grandi spogli per scopi crittografici e pedagogici /304/, alle soglie della prima presa di coscienza teorica della possibile presenza di strutture quantitative nella lingua. Accennerò ai motivi di questo ripensamento nei riguardi della ling. statistica nella 2° conferenza : strettamente legato alla ling. statistica è lo sviluppo delle applicazioni linguistiche della teoria dell'informazione.

3.6.3. Nella linguistica matematica applicata, si sono venuti delineando diversamente i tre principali settori.

La applicazione all'insegnamento ha avuto un notevole sviluppo; mi riferisco a quanto detto nel paragrafo precedente sulla LA in generale, sopratutto per l'istruzione programmata. La documentazione automatica (IR), sotto la spinta prepotente delle esigenze pratiche dell'industria e del commercio, come pure delle discipline bibliografiche e di fronte al crescere delle biblioteche, si è rivolta, per soluzioni immediate, a tecniche e metodi che hanno solo marginali contatti con la linguistica matematica, e si rifanno, almeno su un piano logico, ai tradizionali sistemi della tecnica documentaristica (parole schiave, classificazione decimale universale, ecc.).

L'approccio di tipo linguistico, sia per la fase di classificazione automatica dei documenti, sia per quella di estrazione del contenuto, sia per la fase di "risposta" a conclusione della ricerca, ha subito una battuta di arresto, lucidamente puntualizzata da Bar-Hillel, per es. nel suo intervento alla Scuola e-

stiva della NATO del 1963 a Venezia /22/. Il problema documentaristico e di analisi del contenuto è ora da poco raffiorato con una maggior coscienza della vastità e della difficoltà dei problemi implicati, e in una prospettiva volontaristica molto più ambiziosa di sviluppo, come vedremo riassumendo la terza parte della già citata comunicazione di Hays /138/.

In traduzione automatica (MT), la situazione presenta alcune analogie.

Anche qui, un periodo di esplosione di entusiasmo, di progetti, di mezzi, tra il '50 e il '60, inquadrato tra l'altro, sopratutto all'inizio, nel clima della guerra fredda, "nelle rigide strutture politico-amministrative del mondo scientifico del dopoguerra". /8/.

Le nuove leve reclutate al di fuori delle scuole linguistiche tradizionali, tra logici, statistici, esperti di programmazione e di linguaggi artificiali, hanno spesso duramente scontato con l'insuccesso la presunzione e la insufficiente valutazione della complessità dei fenomeni linguistici. Anche questa crisi è stata evidenziata, (clamorosamente, come ricordo, essendomi trovato presente), alla Scuola Estiva Nato del 1962 in 4 celebri lezioni di Bar-Hillel /21/; e Chomsky, fatto che spesso non è stato visto dai cultori della MT, che sono spesso tra i suoi più ardenti fautori, ha, come del resto Bar-Hillel, provato su un piano teorico la inadeguatezza dei modelli usati. La crisi ha portato addirittura alla brusca sospensione di numerosi progetti di traduzione automatica, e oggi sopravvivono solo alcune équipes, che, come quelle di Vauquoïs a Grenoble, hanno affrontato con scadenze non immediate e con interessi eminentemente teorici il problema della traduzione.

Oggi occorre sempre più spesso di trovare nella introduzione a studi di LM e di MT frasi di questo tipo : "Nelle manifestazioni culturali dell'uomo esistono fattori che non possono essere ridotti a valori logici di 'vero' e di'falso', e prendono in considerazione forze che distinguono l'uomo come animale razionale dalla macchina" /237 p.58/. E il problema delle ambiguità sintattiche e delle ambiguità semantiche (ambiguità dal punto di vista di una analisi automatica), fa dire per es."There is no way, at present, for sistematically representing this knowledge of ours about the universe. These problems of syntactic and semantic anbiguities in natural languages are the basis for our pessimism concerning high-quality fully-automatic mechanical translation" /299/p.31/.

Si ripiega per ora, mentre si ridimensiona approfondendolo il campo di ricerca, su sistemi di traduzione parzialmente automatica, nei quali il calcolatore aiuta o chiede aiuto a un traduttore umano /207/.

4. In tutti i settori della LM, il calcolatore si è imposto come strumento di indiscussa utilità. Il rapporto tra LM e LC deve essere visto chiaramente e correttamente. Mentre l'esistenza e l'uso di un calcolatore o in genere di un automa, è essenziale per la concezione stessa di alcuni settori della LA, proprio per la loro caratteristica fattuale che richiede un automatismo che esegua alcune operazioni (TM, IR, istruzione programmata, ecc.), il calcolatore è uno strumento prezioso, a volte praticamente insostituibile, ma non essenziale al concetto di LM. Questa affermazione è immediatamente evidente per l'aspetto quantitativo sia nella sua parte negativa - la legge di Zipf, sul piano delle teorie, e i grandi dizionari di frequenza, sul

piano dei conteggi e della raccolta dei dati, sono antecedenti all'introduzione dei calcolatori -, sia in quella positiva:gli spogli necessari non possono essere esaustivi e qualificati, e i calcoli sui dati non possono essere così sofisticati, come si vuole, senza uno strumento che libera lo studioso della fatica delle operazioni manuali e di routine. Naturalmente, l'uso del calcolatore contribuisce anche alla evoluzione teorica e, per così dire, qualitativa, della disciplina, che dipende, per sua natura, pressochè del tutto dal suo aspetto sperimentale,aspetto che il calcolatore permette di espandere meravigliosamente.

Per il settore algebrico invece, ci sono state delle confusioni. Pensiamo per esempio alla descrizione della grammatica in una prospettiva Chomskiana, come è quella di N.Ruwet: "une grammaire doit être capable d'énumérer explicitement toutes les phrases qui sont incontestablement grammaticales, ou bien formées, dans la langue étudiée, et d'exclure explicitement toutes les séquences qui sont incontestablement agrammaticales dans cette langue. Nous pouvons nous représenter cette grammaire sous la forme d'un mécanisme d'une certaine sorte, analogue à une machine à calculer, et qui énumère (ou engendre) les phrases grammaticales au moyen d'un ensemble d'instructions qui sont l'équivalent de règles grammaticales"./281 p.32/ Lo stesso Chomsky ricorda che si è talvolta espressa l'idea che le ricerche di grammatica generativa siano il prodotto di tentativi per utilizzare i calcolatori elettronici in un modo o nell'altro./57 p.61 /. Questa idea, controbatte anche Ruwet, /281 p.371/ è completamente falsa. Al contrario, gli sviluppi della grammatica generativa sono spesso serviti a mostrare le insufficienze e i limiti dei progetti di questo tipo. E lo prova il fatto che i maggiori,originali contributi alla linguistica algebrica vengono da studiosi che non si sono mai cimentati in pratica con il calcolatore.

Del resto, il modello trasformazionale ha tre componenti : semantico, sintattico e fonologico. /195 p.13/.

Perchè una applicazione funzioni, così come sarebbe richiesto dalle applicazioni computazionali di MT, IR, esso dovrebbe funzionare globalmente, e quindi essere completo anche nel componente semantico, il che è chiaramente, almeno oggi, irrealizzato.

Ciò non toglie che il calcolatore possa rendere degli utili servizi al linguista teorico, come è universalmente riconosciuto. Cito, uno per tutti, le affermazioni di I.I. Revzin a proposito del significato e valore del metodo di simulazione e delle applicazioni pratiche computazionali, come criterio di verifica della validità dell'interpretazione di un modello linguistico,e cioè delle corrispondenze tra il modello e l'originale /264 p.4/.

Recentemente, come vedremo più avanti, sono stati compilati dei programmi che permettono, proprio per mezzo delle simulazioni, di testare praticamente le regole del modello linguistico, all'atto stesso della loro concezione.

Ma il vero punto di contatto tra grammatica generativa e calcolatore è un altro: la grammatica generativa ha stretti rapporti con lo studio astratto degli automi, studio che si è sviluppato a partire da certe ricerche matematiche (specialmente la teoria delle funzioni ricorsive). Secondo Ruwet /281 p.371/, un modello della competenza linguistica presenta certe analogie con un automa, pur supponendo dei meccanismi più complessi di quelli che caratterizzano la maggior parte degli automi conosciuti attual

mente.

Riassumendo, abbiamo visto che la LM si definisce, a seconda dei casi, per diversi modi di applicazione dei metodi matematici, anche essi diversi, ai problemi linguistici.

Essa è, per sé, indipendente sia da applicazioni, in quanto, sopratutto nella costruzione di modelli strutturali, fa parte della linguistica teorica, sia dall'uso dei calcolatori.

La LA, è stata caratterizzata come l'impiego di ritrovati e di teorie della linguistica, sia generale sia di lingue particolari, a scopi pratici alquanto diversi, ma aventi in comune delle operazioni sulla facoltà del linguaggio, o su una simulazione di questa facoltà, o sul linguaggio stesso come dato di input. Essa applica spesso teorie e metodi della LM, perchè il carattere fattuale pragmatico e operativo dei suoi compiti richiede completezza, esplicitazione, rigorosità, riproducibilità, ed economia che sono quasi esclusivamente garantite dal carattere formale delle descrizioni e dei modelli della LM. Si parla allora di LM applicata; spesso le applicazioni consistono nell'automatizzare per mezzo di un meccanismo, nella maggior parte dei casi un calcolatore, alcune operazioni sul linguaggio, sopratutto per quanto attiene al suo carattere di supporto e veicolo di informazione.

5. Giungiamo così al nostro ultimo, e centrale, argomento, la linguistica computazionale (LC).

La congerie di attività che oggi ha il nome di LC (ingl. computational linguistics; franc. linguistique computationelle) in America si è sviluppata attorno alla MT come primo nucleo, e spesso, nel passato, è stata con essa identificata /138 p.1/. Se la MT fosse tutta la LC, si potrebbe parlare di una crisi di quest'ultima. La MT è stata giudicata troppo costosa per essere utilizzabile, almeno nella situazione attuale della linguistica e della elaborazione dei dati, dall'Automatic Language Processing, Advisory Committee. (1966) /17/.

Spesso, la LC viene considerata come una parte della LM e della LA : dopo quanto si è detto, è chiaro che l'area della LC interseca quella della LA, nella misura in cui quest'ultima utilizza il calcolatore come strumento operativo, e interseca l'area della LM in quanto quest'ultima usa il calcolatore sia per raccogliere i propri dati di partenza, sia per raccogliere e verificare le proprie regole. Esiste anche un'area di intersezione tra LA, LM e LC : quando la LA si serve, per operare in base ai portati della LM, delle tecniche e degli strumenti della LC; si potrebbe, per amore di precisione, parlare di linguistica matematica applicata computazionale.

Si può parlare di una linguistica computazionale al di fuori di queste intersezioni? Certamente: è vero che LC applica sempre delle conoscenze linguistiche; ma ciò che è necessario per parlare di LA è, abbiamo visto, l'applicazione per uno scopo pratico, per così dire esterno alla linguistica: esistono invece, e lo vedremo, numerosi casi nei quali il calcolatore è usato come sussidio, di grande valore pratico e metodologico, per ricerche assolutamente interne alla linguistica, e quindi non si può parlare di LA.

D'altra parte, è vero che la natura algoritmica dei programmi esige un grado di formalizzazione che naturalmente è prerogativa della LM.

Ma spesso la formalizzazione dell'algoritmo non riguarda teorie linguistiche nel senso tradizionale della LM e il carattere

linguistico è dato più che dalla natura delle operazioni, dai materiali elaborati e dai risultati conseguiti. A mio avviso, si deve distinguere tra LM e algoritmi di elaborazioni linguistiche: nella LM la formalizzazione e l'adozione di tecniche matematiche sono richieste per conferire certe caratteristiche al modello (struttura assiomatica, esattezza di definizioni, possibilità di astrazioni verso una generalità della formulazione, esplicitazione e verificabilità delle regole); nella linguistica algoritmica, la formalizzazione è richiesta per la eseguibilità meccanica delle elaborazioni, e non sempre come caratteristica intrinseca della ricerca. A ciò si aggiunge che si può meccanizzare più di quanto si possa formalizzare, come osserva S. Gorn /115 p.341/.

Ciò che sembra essenziale alla LC è dunque il fatto di operare con un calcolatore per elaborarè dei dati che o rappresentano elementi di lingue naturali (grafemi, parole, frasi, testi, ecc.) o comunque sono coinvolti in ricerche che rientrano nel dominio della linguistica o di altre scienze che hanno interesse a ricerche sul linguaggio.

Sopratutto per quanto concerne la manipolazione di simboli o codici rappresentanti segni linguistici (o unità inferiori di seconda o terza articolazione per usare la terminologia impiegata per es. da Z.Muliačić /224 pp.15-18/), si è costituito un insieme di nozioni, di metodi, di tecniche, di procedure, e una terminologia, che stanno diventando pian piano patrimonio comune a quanti lavorano nel campo della LC.

6. Già tre anni fa D.G. Hays /135 p.1/ riteneva impossibile riunire in una sintesi tutte le attività della LC.

Nel tentativo di presentare un quadro generale delle diverse applicazioni dei calcolatori, alcuni hanno affrontato il problema di classificare o almeno ordinare le applicazioni in modo significativo.

6.1. S.Lamb se ne è occupato brevemente una prima volta, a conclusione di un articolo apparso in Language nel 1961./187/, nel quale egli descrive i concetti fondamentali della struttura del calcolatore e della programmazione, ritagliando nella grande massa delle nozioni quella parte di conoscenza che possono servire ai linguisti per liberarsi dalla "misconception.... that the computer is primarly, if not entirely, a glorified mathematical calculator" /187 p.382/ e per rendersi conto "of what the computer can do in the area of linguistics" /187 p.409/. Negli scritti di quegli anni la preoccupazione di smitizzare il computer è pressochè costante; non dimentichiamo che proprio in quel periodo di tempo si ebbero, tra i lessicografi europei, non poche discussioni sulla convenienza e addirittura sulla possibilità di sostituire il calcolatore alle macchine elettrocontabili tradizionali (UR) fino allora adoperate per la lessicografia automatizzata. (cfr. il Congresso di Besançon,1961 /3/).

Lamb avverte che ogni genere di elaborazione che involve la manipolazione di materiale esprimibile in forma di scrittura è virtualmente trasformabile in operazioni di calcolatore, purchè si specifichi esattamente e in dettaglio ciò che deve essere eseguito.

Il problema di cosa può fare il calcolatore nell'area della linguistica non è tanto questione di cosa il calcolatore può fare, quanto di quali procedimenti il linguista può specificare nei dettagli. Per ogni progetto di questa specie ci sono due operazioni : (1) formulare la procedura; (2) tradurre questa procedura in

istruzioni di macchina. Di questi due compiti, nella maggior parte dei casi, il secondo è più facile del primo. Perciò a chi vuole usare il calcolatore in una ricerca linguistica sono richieste molte conoscenze linguistiche, ma egli non deve possedere, del calcolatore, molto di più che poche nozioni sulla sua struttura e sulla logica della programmazione. Questa ultima affermazione alla luce dello sviluppo raggiunto oggi dalla LC, resta in linea di principio ancor valida, ma in pratica la situazione attuale richiede la presenza di linguisti con una specializzazione nel settore della LC. Non si tratta di una sub=specializzazione professionale, ma di un insieme di teniche utili per tutti i linguisti /137/ che richiede però dei cultori e sopratutto dei docenti /329/.

Lamb riprende e precisa la sua classificazione in una comunicazione dal titolo 'Linguistic Data Processing' presentata nel 1962 al Congresso di Burt Wartenstein (Austria), e pubblicata nel 1965 /189/.

L'espressione linguistic data processing (LDP), può avere tre valori, esprimibili così : 1. elaborazione dei dati linguistici, 2. elaborazione di dati per scopi linguistici, 3. elaborazione linguistica di dati. In altre parole, l'attributo linguistico o non linguistico va riferito o ai dati, o a gli scopi, o alla procedura. Lamb riduce a 6 le 8 combinazioni possibili, supponendo che procedure di tipo linguistico si applichino solo a dati linguistici, e ottiene così un primo schema qui riportato (Fig.1).

Egli fornisce, per ogni 'casella', una applicazione a titolo di esempio :

1. Ricerche sul parsing automatico come approccio allo studio di strutture linguistiche.
2. Studio statistico dei testi come sussidio per delle strutture linguistiche.
3. Simulazione della dinamica delle aree linguistiche.
4. MT di letteratura etnografica.
5. Costruzione e uso di indici automatici per l'etnografia.
6. Calcolare le orbite dei satelliti terrestri.

Uno schema più completo si ha considerando che le applicazioni dei calcolatori possono presentare differenza o somiglianza se considerate secondo: 1. il tipo di uso per le quale sono destinate; 2. i dati elaborati ; 3. il tipo di elaborazioni richieste; 4. l'obiettivo della elaborazione. Non si tratta solo di sostituire gli scopi del 1° schema, con i parametri obiettivo e destinazione (per esempio le concordanze sono un obiettivo che può essere usato da lessicografi, da etnografi, da documentalisti; la MT può essere usata da linguisti, da chimici, da fisici, ecc.). Si tratta sopratutto di sostituire all'alternativa 'linguistico non linguistico', alternative più specifiche : per es. se i dati siano costituiti da testi, o da regole grammaticali, o da dati bibliografici: oppure distinguere, tra le procedure, la semplice riorganizzazione dei dati, dalla generazione di frasi, dalla analisi automatica, ecc. . Una classificazione basata su questi 4 criteri avrebbe il pregio di coprire sicuramente l'intera gamma delle applicazioni - dice Lamb - ma le applicazioni realizzate verrebbero concentrate in poche classi, mentre altre classi rimarrebbero vuote o quasi. A un approccio di questo tipo, esclusivamente deduttivo, egli preferisce un approccio induttivo, nel quale la classificazione è basata sugli oggetti da classificare

piuttosto che sull'intelaiatura fornita dai parametri. Egli divide il campo del L.D.P. in 8 aree; i 4 parametri su citati servono come riferimento, ma il riferimento non è sistematico: le applicazioni vengono riunite in una classe perchè presentano alcune somiglianze particolarmente significative che le caratterizzano con grande evidenza. Nella fig. N.2, nella col. 1 sono elencate le aree. La col. 2 stabilisce i parametri secondo i quali le applicazioni sono raggruppate in un'unica area. La col.3 riguarda ciò che rende l'area interessante per lo studioso. Le altre tre colonne sono riferite a tre diversi significati dell'espressione LDP. Una x significa sì, uno spazio significa no. In particolare le colonne 6 e 7 dicono se le applicazioni dell'area hanno o no, rispettivamente, un uso linguistico e un uso extralinguistico. La col.3 richiede qualche spiegazione. Alcune applicazioni di LDP sono valide a motivo dell'output, dei risultati, mentre altre sono interessanti in primo luogo a causa della elaborazione che il calcolatore esegue. X

Nel caso degli studi statistici di materiali verbali, l'utente normalmente non si interessa a come il calcolatore consegua i risultati.

Egli usa il calcolatore solo per ottenerli agevolmente; in alcuni casi perchè i risultati non potrebbero essere ottenuti senza un calcolatore, l'insieme dei calcoli essendo troppo voluminoso per altra via. Egli non ha alcun interesse a quale procedura la macchina segua; ma applica una delle numerose routines statistiche standard disponibili presso i centri di calcolo. Dall'altro lato, il linguista che lavora alla simulazione dell'evoluzione delle lingue è interessato quasi esclusivamente al procedimento che il calcolatore esegue, poichè è questo procedimento che egli studia. La col.3 mostra per ogni area se l'interesse del linguista, in quanto tale, è legato in primo luogo al procedimento, o ai risultati, o a entrambi. Ciò premesso, Lamb commenta le aree ottenute, e noi coglieremo degli spunti dal suo commento, per nominare alcune delle principali applicazioni di fatto tentate.

6.1.1. Simple text processing.

Il termine 'semplice' va riferito ai procedimenti eseguiti dai programmi di calcolatore impiegati in quest'area.

Esso implica: 1. l'utente li usa non per l'interesse della elaborazione ma per i risultati; 2. i procedimenti sono di tipo meccanico, sostituiscono il lavoro d'ordine ('clerical type') e non implicano decisioni da parte della macchina; 3. sono quel tipo di elaborazioni che, eseguendole a mano, l'uomo giudicherebbe tediose e non richiedenti riflessioni o giudizio: hanno in comune l'operare su uno stesso tipo di dati, cioè su testi naturali. Si può aggiungere che l'elaborazione ha per scopo di fornire informazioni convenientemente organizzate estratte dal testo, che sono d'aiuto nello studio del testo stesso.

Il tipo più semplice di operazione in quest'area, che costituisce la base di altre operazioni - è - secondo Lamb - la 'semplice ricerca' ('the simple search'). Nella sua forma più semplice, consiste nel ricercare in un testo tutte le occorrenze di una unità chiaramente definita, per es. di una parola : quando la trova, il programma può fare cose diverse: annotare le parole immediatamente precedenti e seguenti, o un contesto più ampio, o solamente l'ubicazione.

Se queste annotazioni sono fatte per tutte le parole, si hanno rispettivamente delle concordanze o degli indici; le concordan-

ze possono essere elaborate ulteriormente, per es. rese selettiva con l'esclusione di alcune categorie di parole (di solito le parole così dette grammaticali); nella stessa area rientrano gli indici 'a contrario', i rimari, gli incipitari, ecc.: le utilizzazioni sono numerosissime, anche in settori extra-linguistici (lo vedremo più avanti parlando del CNUCE).

Lamb giudica, s'è visto, 'non linguistici' i procedimenti implicati. Egli ammette però che 'It is perhaps largely a matter of opinion wich of the operations are to be regarded as linguistic and wich are not.' Per questo una x minuscola e tra parentesi indica nella casella 4 che questi procedimenti possono essere considerati parzialmente linguistici.

6.1.2. Automatic linguistic analysis.

Ciò che differenzia quest'area dalla precedente, è che nel caso dell'analisi linguistica automatica spetta al programma prendere decisioni e esercitare un giudizio; invece nel caso della semplice ricerca è il linguista che prende le decisioni e trae le conclusioni, mentre il calcolatore lo aiuta nel lavoro di routine, e semplicemente riorganizza i dati, così che sia più facile, per il linguista, esaminarli.

Lamb distingue due grandi aree, la sincronica e la diacronica. Come esempio della prima area si riferisce a due esperienze, una propria /188/ l'altra di Garvin /109/. Accenno brevemente, poichè Lamb non li descrive, agli intendimenti del programma di Garvin: esso raduna dei dati, non analizzati, in maniera ordinata; per mezzo di un suo dato componente di deduzione('elicitation') applica assunzioni teoriche ai dati e produce ipotesi attorno alle strutture morfologiche della lingua rappresentata dai dati. Più che 'testare' la grammatica di una lingua, obiettivo immediato del programma è 'testare' una teoria dell'analisi linguistica come è rappresentata da un procedimento di scoperta e l'obiettivo finale è spiegare le assunzioni universali o quasi universali (universali del linguaggio) che sono impliciti nell'analisi linguistica.

Il programma così include tutte le operazioni che un linguista descrittivo esegue, meno la verifica delle ipotesi: esso è costruito su principi euristici, piuttosto che algoritmici. /220 p.10/.

Una concezione di questo genere richiama immediatamente alla memoria la presa di posizione di Chomsky relativamente ai cosidetti 'procedimenti di scoperta' della linguistica tassonomica, per es. in "Syntactic Structures", al 6° capitolo, e in 'A Trasformational Approach to Syntax' (cfr. a pag.95 dell'edizione italiana). Rinvio appunto a queste pagine per la discussione, nonchè all'articolo di G.Nickel, che discutendo i tipi dello strutturalismo tassonomico americano nei confronti della grammatica generativa trasformazionale, afferma che nella descrizione di lingue sconosciute i metodi automatici di analisi sviluppati dallo strutturalismo tassonomico sono ancora i soli utilizzabili / 230 p.18/. Mi limito ad osservare che, al di fuori dei due studi citati, uno dei quali prosegue ancora /110/ non conosco altri tentativi computazionali.

Come esempio dell'area diacronica, Lamb pone il problema di automatizzare i metodi comparativi : egli può citare pochi nomi (Glaeson, Swadesh, Oswald, Naim /229/.)

Neppure oggi la situazione è cambiata, anche se si nota recentissimamente un principio di ripresa dopo la crisi dei primi anni del '60, come è dimostrato da alcune comunicazioni all'ICCL.

6.2.3. Linguistic Automation

Cioè l'automazione dell'uso del linguaggio: " avere un calcolatore che adopera il linguaggio come la gente che conosce una lingua. Usare una lingua implica due processi, che possono essere chiamati produzione e decodifica. In altre parole, chi conosce una lingua è capace di produrre un discorso in questa lingua e di decodificare i discorsi altrui per capire ciò che è stato detto". /351 pp.174-5/.

Mentre l'analisi linguistica automatica può essere pensata come una simulazione del linguista, l'automazione del linguaggio può essere pensata come la simulazione di chi usa il linguaggio. In entrambi i casi il concetto simulazione deve essere inteso in un senso ristretto, poichè non si richiede di simulare il processo umano, ma solo i risultati.(Rinvio alle considerazioni sul valore dei modelli discusso a proposito della LM). Al limite di quest'area troviamo i progetti di MT, e ancora più utopisticamente avanzati, i progetti di un interlocutore automatico. Al più alto livello di immaginazione si può porre un sistema che abiliti la macchina ad apprendere l'uso del linguaggio.

6.1.4. Processing of Linguistic Files.

Con 'linguistic files' Lamb intende una lista o una tavola o altro aggregato di informazione linguistica, in forma operabile dalla macchina, per esempio su nastro magnetico: tabelle di regole grammaticali, dizionari, dati linguistici parzialmente analizzati e organizzati in qualche modo. L'elaborazione di tali 'files' include la compilazione e il loro aggiornamento, con aggiunte, soppressioni, correzioni, trasformazioni, ecc.. Quest'area è di importanza essenzialmente pratica, come supporto operativo e informativo per le aree precedenti.

6.1.5. Simulation of Language Dynamics.

Lamb vi distingue due sotto-aree.

"La simulazione dei mutamenti linguistici presuppone di introdurre una parte di una certa lingua nel calcolatore e di applicare ad essa certi principi generali di mutamento, con in più, forse, certe influenze esterne; si osservano poi i risultati /151 p.178/. In questo tipo di sistema il programma incorpora i principi di mutamento, e lo scopo è di incrementare la conoscenza del processo di evoluzione.

La dinamica delle aree linguistiche si riferisce al processo per cui alcune lingue, col passare del tempo, espandono le proprie aree geografiche, di solito a spese di altre lingue. Lamb discute le possibilità di un simile studio con il calcolatore a livello puramente teorico, poichè non conosce alcun tentativo reale.

6.1.6. Knowledge Retrieval

La richiesta di base a un sistema di 'retrieval' su una biblioteca, è che esso, rispondendo alla domanda posta dal ricercatore, elenchi i documenti, presenti nella biblioteca, pertinenti all'argomento proposto. Il sistema ha due versanti principali: la classificazione dei documenti, e la gestione del linguaggio di domanda-risposta. Lamb esamina le diverse strategie proposte. (A questo proposito, come pure per la MT, è da notare che la conferenza di Lamb è precedente alle grosse crisi e al ridimensionamento delle ambizioni in questi due settori.)

Un altro tipo di 'retrieval' è quello che Lamb chiama 'fact retrieval'. Un sistema cioè che non fornisce indicazioni bibliografiche, ma che risponde alle domande circa i fatti che sono riportati nei testi della biblioteca.

6.1.7. Statistical Studies of Verbal Data

Quest'area è una parte del LDP perchè implica dati linguistici. Costituisce un'area in vista del tipo di procedimento richiesto. L'interesse del linguista è per i risultati, più che per i metodi.

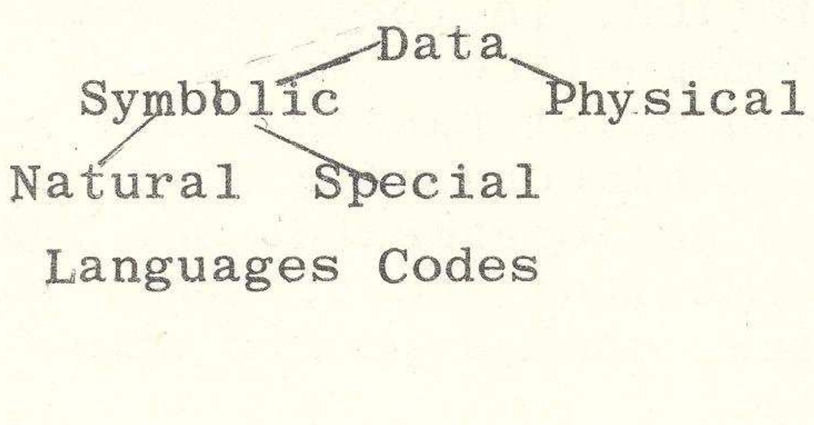
A conclusione, Lamb stesso avverte che molte delle aree sono pressochè prive di ricerche di fatto esistenti. E in effetti le sue descrizioni, in molti casi, possono essere accettate,nonostante l'enfasi utopistica tipica del tempo, non come un resoconto, ma come la segnalazione di direzioni possibili di studio, estrapolate alle estreme conseguenze.

6.2. Nello stesso volume J.C. Gardin si occupa di come ordinare le applicazioni dei calcolatori all'antropologia, intesa, secondo l'uso americano, nelle sue diverse branchie: antropologia fisica, antropologia culturale, sociologia, folklore, linguistica,ecc.

Egli sceglie come parametri i dati da elaborare e le operazioni di elaborazione.

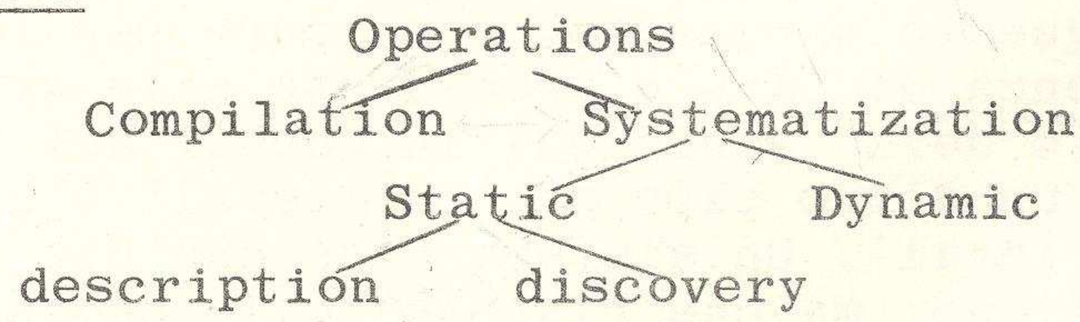
6.2.1. D a t i

Egli li classifica per mezzo di un sistema /106 p.105/ con due opposizioni binarie. "La prima opposizione coincide con la distinzione classica 'segni' e 'segnali' ". Nel primo caso si tratta di manipolare dei 'segni', con funzione simbolica; nel secondo di elaborare 'segnali' di natura fisica; egli non considera questi ultimi, (come esempio dei quali porta il 'pattern recognition' e l'analisi automatica del suono, come vedremo più avanti, quest'ultimo argomento è invece accolto nell'ultimo congresso di LC,l'ICCL di Stoccolma,1969). Il primo caso invece copre gran parte della elaborazione di dati non numerici, nel quale l'input del calcolatore consiste in 'informazioni verbali', che sono sottoposte a conteggi (p.e. statistica linguistica), compilazioni (p.e. indici automatici), trasformazioni, ecc. /106 p.107/.



La seconda opposizione distingue, fra i dati simbolici,quelli che rappresentano un linguaggio naturale e quelli che corrispondono a 'codici speciali'. La opposizione, i cui esempi si trovano nella Fig.3, riquadri A2 e B2, va intesa più che tra(1.)linguaggi naturali e(2.) codici, tra linguaggi e codici considerati 1. come oggetto di studio, e(2.) come mezzo per studiare altri dati non linguistici. Per es., dice Gardin, supponiamo di estrarre con il calcolatore tutte le occorrenze della parola 'innatural', in Shakespeare. Possiamo esaminarle in quanto vogliamo studiare la lingua inglese del 16° secolo, oppure in quanto vogliamo studiare la personalità. La differenza risiede nel propositi della ricerca, linguistico del primo caso, non linguistico nel secondo.E' questo il senso della distinzione che, con una denominazione semplificata, costituisce la seconda opposizione.

6.2.3. O p e r a z i o n i



A un polo della prima opposizione stanno le applicazioni 'triviali': il calcolatore funziona come una specie di contabile, 'pedina' ogni unità delle sue 'files', ne indica la frequenza, non altera i dati, ma li riordina semplicemente secondo ordini convenzionali, per es. alfabetico, ecc. . Osserva cioè le occorrenze di date sequenze di simboli, per es. parole o combinazioni di parole, e le tabula, con o senza conteggi, per es. per fare indici, concordanze, o liste bibliografiche.

All'altro polo dell'opposizione, le operazioni di compilazione e conteggio sono ancora presenti, ma servono a uno scopo più sofisticato, che Gardin chiama di 'sistematizzazione': per es. la MT, tassonomie culturali, simulazione di comportamento, ecc., dove alcune teorie (linguistiche, statistiche, psicologiche), sono usate in aggiunta alle operazioni proprie del calcolatore.

La seconda opposizione è tra sistematizzazioni statiche e dinamiche: può essere resa anche con i termini di 'classificazione' e 'simulazione': da un lato la organizzazione sincronica di ampi insiemi di dati, in gruppi, classi tipi, che tendono ad una rappresentazione ordinata dell'intero insieme; dall'altro il 'comportamento' di un sistema in azione, per es. ordinamenti diacronici, di solito basati su relazioni postulate gradino per gradino tra eventi singoli successivi (evoluzione linguistica, tendenze comportamentalistiche, ecc.).

La terza dicotomia si riferisce, a un polo, alla conoscenza empirica di un sistema culturale, a dati la cui distribuzione è nota : si cerca allora di costruirne un modello 'economico'in linguaggio matematico. All'altro polo si tenta di scoprire attraverso calcoli matematici, - nel senso largo del termine, sia numerico, sia non numerico - relazioni sistematiche o raggruppamenti, partendo da aggregati estesi di osservazioni indipendenti / 106 p.109/. L'esempio di Gardin (del resto collegato alla sua personale attività: cfr. /106/ e /105/)è particolarmente efficace. Supponiamo di avere centinaia di migliaia di reperti fittili, radunati da diversi siti archeologici. Se possediamo evidenti informazioni, stratificate o di altro genere, possiamo organizzare la collezione in gruppi di varia natura. Le tecniche di calcolo servono a costruire un modello matematico delle differenziazioni culturali osservate, in termini di quantificazioni numeriche o verbali, di tratti descrittivi. Otteniamo, allora, una 'mappa' economica delle analogie e delle differenze conosciute, che potrà servire ad esempio per collegare oggetti o tratti di origine ignota ai gruppi culturali organizzati. Se invece non possediamo nessuna informazione distribuzionale,e non disponiamo di raggruppamenti empirici, il calcolo è diverso nei metodi e nello scopo: l'obiettivo è, allora, suggerire distribuzioni probabili dei tratti osservati, in 'tipi' o 'culture' inferite dall'evidenza statistica. (A proposito di questa classificazione automatica in linguistica, rinvio alle osservazioni formulate più sopra all'area 'analisi automatica' di Lamb).

6.2.4. Le applicazioni del calcolatore

Combinando le principali categorie di dati e di operazioni in una tabella a due entrate, si ottiene una distribuzione delle applicazioni in 9 tipi, riassunte nella tabella di Fig.3,che mi sembra chiara senza altre spiegazioni.

6.3. P.L. Garvin in un articolo apparso in Language nel 1962 /108/ si propone di considerare il diverso grado di partecipazione con la quale un programma di calcolatore può essere integrato

nel procedimento stesso della ricerca linguistica. Egli si limita alla linguistica descrittiva, e non mira a un panorama dell'intero campo del LDP. Egli individua tre gradi di partecipazione: 1. La semplice collezione di dati: questi vengono ordinati sistematicamente e resi operabili per permettere una conveniente ispezione o consultazione da parte del ricercatore. Come esempio cita le concordanze nelle loro diverse strutturazioni, purchè il contesto non sia costruito con un algoritmo che sceglie le parole in base alla struttura sintattica del testo. (E' interessante osservare che Garvin anticipa una applicazione che prende piede oggi, e cioè la compilazione di concordanze 'bilingui', in cui sono fusi un testo e la sua traduzione in altra lingua, per la compilazione di dizionari /148/ e per l'analisi contrastiva./38/). 2. Programmi di calcolatore che usano i risultati di ricerche linguistiche; sia per scopi pratici,(MT e IR), sia per verificare teorie linguistiche, in particolare modelli e insiemi di regole. 3. L'automazione delle procedure di ricerca linguistica: egli afferma esplicitamente che essa è basata sull'utilità che il linguista attribuisce alla descrizione distribuzionale dei fenomeni linguistici.

In 1. il programma non contiene alcuna informazione linguistica, ed è capace solo di differenziare l'aspetto formale delle unità di input, e di operare delle selezioni su questa differenziazione. In 2. il programma contiene informazioni su una lingua particolare, che gli permettono di differenziare talune proprietà di coocorrenza di una lingua 'conosciuta' e di usare questa diversificazione per scopi vari. In 3. il programma contiene informazioni provenienti da assunzioni e metodi della linguistica generale. Queste informazioni gli permettono, per es. di 'scoprire' proprietà di coocorrenza di un linguaggio 'sconosciuto'. Un altro esempio di 3. è dato dalla semantica distribuzionale, basata sulla assunzione che certe similarità semantiche implicano certamente certe similarità in contesti rilevanti. Se dunque questi possono essere specificati, è possibile raggruppare automaticamente tutte quelle unità linguistiche che occorrono in contesti definibili come similari, e si assume che questi gruppi prodotti automaticamente siano di interesse semantico. (Questo tipo di approccio semantico, le cui premesse teoriche sono chiaramente esposte, per es., in un articolo di Apresjan /133/, è usato da alcuni ricercatori per progetti di information retrieval, /290/: si vedano, per es. le osservazioni da un punto di vista applicativo, di Rubenstein nell'articolo Direction in Semantic Research /279 pp.97-98/.)

Vediamo ora due tipi di classificazione proposte quasi 10 anni più tardi.

6.4. C.A. Montgomery, in una comunicazione all'ICCL del 1969, si preoccupa dei rapporti tra linguisti da un lato e persone imgnate nel settore del LDP dall'altro. C'è una certa intersezione dei due gruppi, ma si lamenta che l'intersezione sia troppo limitata rispetto alle dimensioni dei due gruppi: ciò dipende -secondo Montgomery - da una insufficiente comprensione delle comuni sorgenti di interesse per le lingue naturali. I linguisti valuterebbero insufficientemente il fatto che ogni elaborazione di informazioni in linguaggio naturale - scientifico, tecnico, letterario -, è essenzialmente un problema linguistico.

Il trattamento dell'informazione in lingua naturale, per fare degli indici, degli 'abstracts', della traduzione, richiede

una analisi del contenuto del testo e la sua rappresentazione in una qualche forma standard. Idealmente, l'analisi del contenuto consiste nel determinare i concetti presenti e le interrelazioni tra essi. La prima operazione sarebbe basata su una qualche forma di analisi semantica, la seconda implicherebbe l'analisi sintattica, anche se le due forme di analisi sono considerevolmente interdipendenti. Una volta identificati, i concetti e le relazioni devono essere trasposti in una serie di 'frasi' canoniche che rappresentano il contenuto del documento. Nel caso dell'I.R. o di un sistema che risponde a delle domande, le frasi canoniche rappresentano la 'conoscenza' del sistema e servono come base per generare risposte concrete a questioni specifiche. Anche se, di fatto, le procedure per l'analisi del contenuto sono lontane da questo ideale, è chiaro - secondo l'autore - che l'analisi deve essere basata su principi semantici e sintattici, oggetti naturali dell'interesse linguistico. A ciò si aggiunge l'alto grado di utilità del calcolatore come strumento delle ricerche linguistiche, che l'autore divide in due categorie.

6.4.1. Collezione e organizzazione dei dati, operazioni sulla base dei dati.

Ci sono due sensi nei quali i dati sono collezionati nell'analisi linguistica. Il primo senso si riferisce alla raccolta iniziale dei dati per includerli nel corpus di base: l'autore si riferisce alla strategia euristica per scoprire gli elementi base e le relazioni in una data lingua, e cita l'unico tentativo,compiuto da Garvin, di automatizzare la deduzione dei paradigmi per lingue sconosciute /111/, del quale abbiamo già parlato.

Nel secondo senso, la raccolta dei dati è la selezione di particolari unità da un corpus compilato in precedenza: poichè in un approccio tassonomico le classificazioni contenute nelle raccolte di dati fornirebbero in realtà la base della grammatica,e poichè l'analisi sintattica e la semantica richiede una organizzazione altamente raffinata e estensiva dei dati, questi aspetti della ricerca linguistica vengono facilitati dal calcolatore,che eccelle nelle operazioni 'clerical' di selezione e di stampa veloce e flessibile dei dati. Si ottengono così una varietà di ordinamenti di grandi volumi di dati, che possono evidenziare configurazioni sottostanti, non identificabili, o identificabili con difficoltà, coi mezzi tradizionali della schedatura manuale.

6.4.2. Formulazione, organizzazione di ipotesi, verifica, operazioni di analisi e di sintesi.

In ordine a testare delle ipotesi, il calcolatore obbliga ad essere il più espliciti possibile: è difficile che il linguista possa seguire - nel verificare una grammatica trasformazionale - le tortuose ramificazioni di regole coordinate, senza l'aiuto di una verifica automatica. Sono stati scritti programmi in questo senso, sia per le regole fonologiche, sia per quelle sintattiche. Un algoritmo di analisi serve invece come verifica di una grammatica tradizionale di tipo descrittivo, e l'autore cita gli algoritmi più diffusi.

6.5. D.G.Hays /138/ in una comunicazione al citato congresso di Cambridge, parla di tre livelli di LC.

6.5.1. I sistemi di superfice.

Il linguaggio può essere considerato a più livelli, secondo gli scopi di chi lo considera. E' naturale che sia più comune lavorare ai livelli più superficiali. Un sistema lavora alla superfice del linguaggio se lo tratta come una catena di caratteri o

di fonemi. Sarebbero sistemi di superfice le applicazioni per indici, concordanze, analisi fonematiche, operazioni statistiche, i sistemi di documentazione automatica basati sulle parole chiavi, sulla statistica, e i primi tentativi di traduzione automatica.

6.5.2. I sistemi classici.

Piuttosto presto è stato proposto uno schema che Hays considera classico; con variazioni, è uno schema essenzialmente comune per la MT, l'IR, la correzione dei tests scolastici, ecc. . Questo schema classico consiste in una elaborazione superficiale (per es. la consultazione di un dizionario di macchina), una elaborazione sintattica, una elaborazione semantica. L'obiettivo di queste tre fasi, nell'ordine, è ottenere una espressione per il contenuto del testo di input. Certamente, il testo di input è esso stesso una espressione del proprio contenuto, ma l'espressione che lo schema classico intende ottenere ha alcuni vantaggi,come l'assenza di ambiguità, la perspicuità in relazione alla comprensione,o anche l'universalità (nel senso di indipendenza dal linguaggio naturale di origine). La traduzione meccanica consiste nello schema classico che opera prima l'analisi del testo di input, e poi,in direzione opposta,la sintesi per ottenere il testo di output. Per il controllo automatico, lo schema classico è applicato alle risposte degli studenti, e la logica è poi applicata per determinare se la risposta è conforme alla risposta richiesta, predeterminata. Come esempi di attività di questo tipo Hays ricorda il sistema di Joyce Friedman /100/ per la generazione di frasi con una grammatica trasformazionale; l'analisi sintattica per la traduzione dal russo al francese, dell'équipe di B. Vauqois a Grenoble, che è impegnata a compilare un linguaggio pivot.

Smith /293/ ha scritto un programma per applicare le regole storiche dei cambiamenti fonetici. Con 21 regole e 500 forme proto-indoeuropee ricostruite, egli ottiene alcune - ma non la maggioranza - delle moderne forme russe che egli ricerca. Pratt e Pacak lavorano sull'analisi automatica di referti medici. Simmons si occupa dell'analisi linguistica delle risposte costruite da studenti nella computer assisted instructions.Un rapido giro di orizzonte, convince facilmente che la tradizione 'classica'(nel senso di sistema computazionale classico) mantiene il proprio vigore, e che la utilità di questi approcci è ancora rilevante.

6.5.3. Una nuova concezione.

Hays, movendo, da un lato, da alcuni lavori recentissimi nel campo della semantica (attestati per esempio nella raccolta di saggi 'Semantic Information Processing' edita nel 1968 al MIT /219/. E in particolare i saggi /29/ e /260/), dall'altro, da una concezione della linguistica che egli definisce esplicitamente come non classica e non chomschiana, conclude anticipando i compiti che - a suo avviso - impegneranno nei prossimi decenni la LC. Si dovrà programmare nel calcolatore un modello dei due elementi che -nella sua concezione- costituiscono la 'facultè de langage' della persona umana: la capacità di memorizzare dei simboli, e la capacità di apprendere degli schemi di azione (action pattern).

La caratteristica essenziale del linguaggio è la sua connessione con l'informazione; il linguaggio ha molte altre funzioni (emotiva, evocativa, ecc.), ma la sua funzione primaria è supportare e trasmettere informazioni: "il linguaggio umano è la manifestazione esterna della capacità, unicamente umana, di elaborare i sim-

boli in modo che l'informazione è ritenuta, rivista, aggiornata." /138 p.8/.

La capacità più notevole dell'uomo è quella di convertire una formula memorizzata simbolicamente in uno schema di azione. L'esempio più usato per uno schema di azione è il saper andare in bicicletta. Tra gli schemi d'azione, ci sono quelli per la elaborazione di simboli. I modi per apprenderli sono diversi. Uno è quello per analogia. Quando costruisco una frase, posso ritenere il suo schema, - alcuni dei suoi tratti principali - come una formula simbolica. Più tardi posso costruire un'altra frase su questo modello, e, se lo ripeto altre volte, posso acquisirlo come schema di frase. Questo concetto sarà disapprovato, egli dice, dai trasformazionalisti; a me ha ricordato da vicino un modo di vedere dei neogrammatici, menzionato anche dal prof. P. Meriggi in un recente intervento al II° Congresso della Società Linguistica Italiana. /210/.

Sembra quindi che Hays riproponga, alla LC, il fine ambizioso di riprodurre l'attività di un "human information processor" /43 p.9/, uno scopo globale al quale, dopo i primi anni '60, sembrava evidente dover rinunciare, come ad una utopia. E ripropone di conseguenza il problema di un modello semantico globale, di un "reticolo di rapporti che colleghi tra loro tutte le parole del dizionario".

6.6. Come ho detto all'inizio, propongo in questa conferenza solo un rapido excursus informativo sulla LC; ho riportato alcuni tentativi di classificazione come strumento per riassumere, da più punti di vista, le ricerche della LC : del resto, queste riflessioni dei cultori di LC sulla attività proprie e dei colleghi, sono forse più efficaci a delineare le direttrici della disciplina che non un esame esterno obiettivo. Da questo punto di vista, anche eventuali proposizioni utopistiche sono illuminanti sulle caratteristiche dell'ambiente della LC, sia pure per un giudizio negativo di alcuni aspetti.

Mi astengo pertanto qui da ogni considerazione di tipo critico, che è esclusa a priori dalla natura informativa di questa conferenza. Mi limito solo a una osservazione di natura appunto informativa: si è sviluppata ed è esplosa una congerie di progetti che utilizzano i calcolatori come strumento ormai normalmente accettato nelle ricerche linguistiche e umanistiche in genere. Le classificazioni proposte sono visibilmente inadeguate a dominarle. Se un criterio, non di classificazione ma di ordine può essere proposto, è a mio avviso un raggruppamento a seconda del 'capitolo' della linguistica o delle discipline letterarie nel cui ambito il calcolatore viene adoperato. Sono le discipline linguistiche che impongono a ciascuna applicazione il proprio metodo. Il principio unitario della LC diventa allora l'esigenza di formalizzare i compiti da svolgere, e l'insieme di nozioni e di tecniche attinenti sia all'esperienza nella utilizzazione ottimale dei calcolatori, sia agli aspetti formalizzabili della elaborazione di dati linguistici.

7. Le relazioni che vengono presentate quest'anno al II° Congresso Internazionale di Linguistica Computazionale, che riunirà a Stoccolma nel 1969 molti dei cultori di LC /157/, costituiscono il panorama attuale, anche se certamente non completo, dello 'state-of-arts' della LC.

Seguo, nell'esposizione, la classificazione data dagli organizzatori alle comunicazioni anche se non sono d'accordo con i

raggruppamenti fatti, ovviamente influenzati da motivi pratici organizzativi.

Anche in questa esposizione come nella cernita degli argomenti, non intendo affatto valutare o criticare le affermazioni degli autori, ma riportarle il più obiettivamente possibile,avendo come unico scopo quello di informare su alcuni progetti di fatto in corso, oggi, nel settore della LC. Allo stesso modo, la scelta delle comunicazioni cui farò riferimento è basata solo su criteri di economia esemplificativa e non su un giudizio critico o di valore.

Completerò il quadro con una sintesi dei progetti di LC in corso presso la Sezione Linguistica del CNUCE.

8. L'ICCL del 1969 /157/.

8.1. Linguistica algebrica - Teoria e procedura di parsing - teoria della grammatica.

L'analisi sintattica automatica mira a scoprire per mezzo del calcolatore le strutture della frase nel quadro di una determinata grammatica.

Ottenere una struttura per una sequenza data, relativa a una grammatica data, è detto dagli anglosassoni parsing - tradotto spesso con analisi logica -, e 'parser' è il programma o sistema che analizza automaticamente la frase. Di solito, l'input del parser, comunemente ottenuto consultando un Dizionario, è una sequenza di insiemi di descrizioni grammaticali. Ogni insieme si riferisce a una unità di occorrenza (normalmente una parola o un sintagma lessicalizzato). Le descrizioni di ciascun insieme sono caratterizzazioni sintattiche alternative. Se un insieme ha due o più elementi, l'unità corrispondente è sintatticamente ambigua, e uno degli obiettivi del parsing è di eliminare le ambiguità, nel caso che le regole grammaticali rendano alcune delle descrizioni inaccettabili nel contesto dell'intera frase. Un altro obiettivo è la scoperta dei legami tra le unità, dei loro ruoli grammaticali.

Differenti teorie di grammatica richiedono differenti specie di parser . (Per una chiara esposizione delle relazioni tra tipo di grammatica e specie di parsing, si veda G.Hays, /134 pp.72-82/ e per le strategie del parsing si veda pure Hays /135 pp.106-132/ e Kuno /298/./Nota 1/.

Nella sua comunicazione Solomon Marcus dell''Institutul de Matematica' di Bucarest introduce un tipo di grammatica contestuale generativa che utilizza la teoria degli insiemi, e viene confrontato con un linguaggio 'context-free'.

Stanley Peters, dell'Università di Texas (Austin),discute un mezzo per l'uso di regole 'context-sensitive' nella descrizione di un linguaggio 'context-free', al fine di provvedere una maggiore economia di descrizione e di analisi. Oltre a un risparmio di memoria, prevede una semplificazione e una valorizzazione delle procedure.

Nel campo dei confronti tra diversi tipi di grammatiche formali si situa anche la comunicazione di H.Karlgren, del KVAL (Research Group for quantitative Linguistics - Stoccolma), che lavora sopratutto nell'analisi di frasi nominali svedesi come appaiono in titoli di documenti, particolarmente in articoli di periodici.

La comunicazione dell'ungherese F.Kiefer, ora al KVAL, concerne gli aspetti teorici della grammatica utilizzata nel sistema di parsing discusso da Karlgren. Il loro parsing utilizza in

formazioni lessicali e morfologiche. Petöfi, del Computing Center dell'Accademia Ungherese delle Scienze, ha sottolineato l'opportunità di utilizzare i progressi metodologici della teoria della documentazione, e in particolare dei metodi che suppongono un 'thesaurus', in un algoritmo che si proponga la analisi completa del testo, fino all'analisi tematica, che egli fonderebbe su una serie di 'abstracts' delle unità che rappresentano le unità "compositive" che costituiscono il testo. Si preoccupa soprattutto dei problemi semantici, che vede però strettamente connessi e integrati a quelli di una grammatica di tipo trasformazionale. D.Varga, dello stesso istituto, si pone il problema di provare l'efficenza di diversi tipi di parsing nella applicazione pratica.

R.C.Schank e L.Tesler del Computer Science Department della Università di Stanford (California); descrivono un "conceptual parser". Dato in input un testo in linguaggio naturale, il parser identifica e disambigua i concetti derivabili da questo input e li colloca in un reticolo che esplicita le interrelazioni. Esso utilizza una "conceptually-oriented dependency grammar",che lavora cioè con un sistema di regole il quale opera su categorie concettuali. (Esse, divise in un "gruppo reggente" (governing)e in un gruppo retto (assisting), sono del tipo attore, azione,oggetto, locazione e tempo di concettualizzazione, ecc.; attributi di un attore, di un'azione, ecc., che non corrispondono alle categorie sintattiche). Queste regole generano tutte le dipendenze possibili in una concettualizzazione. La loro semantica concettuale consiste essenzialmente nelle liste dei dipendenti potenziali di ogni concetto dato.

Prendiamo ad esempio la frase //^{1}the tall boy went to the park/^{2}with the girl// Supponiamo di avere già risolto la struttura di /^{1}, e di voler attaccare ora il costrutto 'with the girl', a un elemento di questa struttura. Esiste un problema, perchè esso può essere riferito a) all'azione 'go' (come sembra essere di fatto), oppure b) al 'parco'.Il problema è risolto dalla semantica concettuale. Essa contiene per 'go' una lista di preposizioni concettuali : sotto 'with' è elencato 'ogni oggetto fisico mobile', e poichè una ragazza è un oggetto fisico mobile, la dipendenza a) è ammessa. E' controllata anche la semantica per 'park'. Sotto 'with' per 'park' sono elencati i vari 'items' che un parco -notoriamente- può contenere: per es. statue, cespugli,ecc. 'Girl'non c'è, così il nesso b) è escluso. Ma potrebbe darsi il caso che si debba intendere 'park with girl' come opposto a un altro 'park'.E' come se il sistema non avesse mai 'incontrato' prima 'park with girl'; esso potrà 'apprendere' questa possibilità, se l'utilizzatore opera una aggiunta alla semantica.

Attualmente un vocabolario di poche centinaia di parole è stato usato per interviste psichiatriche.

A.Birbanescu, dell'IPA (Institutul de proiectari de automatizari - Bucarest) propone possibili analogie tra una grammatica di relazioni, rappresentata da un grafo, cioè da una rete di relazioni binarie tra parole e categorie di parole, e un diagramma elettrico, nel quale ogni parola è sostituita da un contatto, e ogni freccia da un conduttore elettrico. Egli pensa che sia così possibile giungere ad altri modelli di analisi e comprensione del linguaggio oltre a quelli forniti da programmi e calcolatori sequenziali. Egli crede che il processo di comprensione non può essere attualmente diviso in gradini elementari come in un algoritmo;

il calcolatore sequenziale non potrebbe trattare adeguatamente un problema combinatorio poichè non ha la possibilità di una 'percezione globale'.

A.V.Gladky e I.A.Mel'cuk del Sov'et po kompleksnoy problemy Kibernetika pri Prezidiuma AN SSR' di Mosca, suggeriscono un nuovo genere di grammatica formale (grammatica Δ ,da δένδρον ,albero), la quale per molti aspetti è strettamente collegata con le grammatiche di Chomsky, ma ne differisce perchè intende elaborare degli 'alberi' (nel senso della teoria dei grafi) e non delle catene (strings), come fanno le grammatiche chomskiane.Più precisamente pensano a un tipo di grammatica con regole di trascrizione del tipo " X $\longrightarrow$ Y ", dove X e Y sono degli alberi.Linguisticamente, gli alberi considerati sono alberi di dipendenza (non struttura di frase) che rappresentano frasi naturali a diversi livelli di profondità : 'surface sintax','deep sintax' , semantica. Pertanto la grammatica Δ non è concepita per generare delle frasi, ma per trasformare alberi dati in altri alberi. Essa non conosce nè la fase di estrazione da una descrizione semantica data della struttura più profonda ammissibile in forma di albero, né la linearizzazione dell'albero sintattico più superficiale per produrre una frase attuale. Essa procede solamente, a gradini successivi, da alberi profondi a alberi più superficiali.

A.K.Joshi, del Dipartimento di Linguistica dell'Università di Pennsylvania, Filadelfia, studia una classe di grammatiche con tipo misto di regole. La ragione per considerare una tale grammatica è che un solo stile (cioè il carattere formale delle regole) non è capace di rappresentare i vari aspetti della struttura delle lingue naturali.

Molte comunicazioni, pure raggruppate nella categoria 'teoria della grammatica', sono caratterizzate dal fatto che non tracciano una linea di distinzione tra semantica e sintassi. Particolare attenzione è rivolta a stabilire un equilibrio tra due alternative estreme: complicare le regole di sintassi o complicare le categorie del lessico. A questo proposito si veda anche Kiefer /181/. J.Mey, dell'Università del Texas (Austin), in polemica con i cecoslovacchi P.Sgall e E. Hajičovà, confronta con un modello trasformazionale, un modello recente di sintesi e analisi automatica, orientato verso la nozione di 'livello linguistico', e concepito nel quadro dello strutturalismo funzionale di tipo europeo.

L.Kalmár, del Research Group for Math. Logic and Theory of Automata dell'Accademia Ungherese delle Scienze, formula una rappresentazione intuitiva dei linguaggi context-free.

8.2. Grammatica trasformazionale.

8.2.1. Grafemica.

O.Fujimura, dell'Istituto di Logopedia e Foniatria dell'Università di Tokyo, e R.Kagaya, dell'Istituto per lo studio delle Lingue e Culture di Asia e Africa di Tokyo, espongono un modo di descrivere le costanti dei caratteri cinesi come modelli grafici, senza alcun esplicito riferimento al suono e al senso. La descrizione può essere considerata come data in forma di grammatica generativa di forme. S.C.Yang e C.W.Yang del 'Data and Computation Center' dell'Università del Wisconsin, descrivono come un plotter possa essere programmato a funzionare da produttore universale di caratteri grafici, per tutte le lingue alfabetiche, al posto di una macchina da scrivere. E' evidente che il metodo è particolarmente vantaggioso per lingue non alfabetiche, per le quali gli

autori hanno creato un sistema che può essere controllato dagli utilizzatori, attraverso degli 'statements'estratti dal vocabolario naturale. Il carattere è programmato come insieme di sottoparti separabili. Ogni sottoparte è definita e quindi rappresentata da una sequenza di'filevti" (stroke) concatenati da 'operatori'.

8.2.2. Fonologia.

Stephan Braun dell'Istituto Matematico della Technisher Hochschule di Monaco (BRD), investiga su base matematica le proprietà delle regole di rindondanza fonologica, nella prospettiva della teoria dei tratti distintivi binari di Jakobson e Halle, come è esposta per es. in 'Fundamentals of Language './160/ La rindondanza in questa teoria significa che la specificazione (+ o -) di certi tratti di un fonema è predicibile date le specificazioni di certi altri tratti dello stesso fonema e/o di fonemi vicini o di una sequenza di fonemi.

Queste restrizioni nelle specificazioni sono abitualmente espresse da regole di 'ridondanza'. Per es., in italiano tutti i fonemi nasali sono sonori, e ciò è espresso dalla regola (+ nasale) $\longrightarrow$ (+ sonoro), che può essere letta come "ogni fonema che è specificato come (+ nasale) deve essere specificato anche come (+ sonoro)". Si distinguono due tipi principali di regole di ridondanza. Quelle appena esemplificate, che esprimono una restrizione valida per un fonema, indipendentemente dal contesto fonetico, sono chiamate (regole - P. 'phoneme-structure rules') quelle esprimenti restrizioni in una sequenza accettata da una lingua. Secondo gli autori, dopo che Halle ha trattato le regole di rindondanza in modo essenzialmente intuitivo /130/, e dopo la ridiscussione recente di Chomsky e Halle /60/, è utile, sull'esempio di Stanley /294/ formalizzarle e trattarle matematicamente. Gli autori dimostrano che l'insieme completo delle "regole P.", per un insieme di fonemi interamente specificati può essere derivato da una certa funzione booleana, e, così computate senza ricorrere alla intuizione linguistica.

V.A.Fromkin e D.Lloyd Rice, del Dipartimento di Linguistica dell'Università di California, affermano che una delle più utili applicazioni di un calcolatore alle ricerche linguistiche è la valutazione di grammatiche generative. Molti sistemi di programmazione sono stati proposti per questo scopo. Gli autori descrivono un sistema per "testare" le regole del componente fonologico di una grammatica generativa. La complessità delle lingue naturali si riflette anche nelle regole fonologiche, particolarmente quando si vogliono concretamente applicare per convertire la struttura astratta di superficie di una frase nella sua rappresentazione fonetica.

Il programma di calcolatore è stato scritto per aiutare il fonetista nello scrivere le regole, testarle, e nell'insegnare la fonologia.

Esso consiste in un blocco di 'input', di una sequenza di regole fonologiche, e in un blocco per il 'printout'. Il blocco di input accetta dalla console una sequenza di caratteri, che rappresenta la forma sottostante di una parola, o di una frase, assunta come forma terminale di una derivazione nel componente fonologico. Questa forma è allora confrontata con le condizioni contestuali specificate dalle regole memorizzate, ed è modificata secondo queste regole ogni qualvolta si trova un riscontro. La sequenza di unità fonologiche, cioè segmenti e simboli di confine, e/o la matrice binaria che risulta nell'applicazione di ogni re-

gola può essere evidenziata sulla console dell'operatore dopo l'applicazione della regola. Le regole codificate inizialmente dagli autori sono prese dalla descrizione dell'inglese di Chomsky e Halle /60/. Però il programma non è limitato ad esse, ma può essere usato per testare ogni insieme di regole del componente fonologico di una grammatica generativa debitamente memorizzate.

(Anche noi, a Pisa, abbiamo agli inizi un progetto analogo per l'italiano).

8.2.3. Grammatica.

Anche J.Friedman, del 'Department of Computer and Communication Sciences' dell'Università del Michigan, constata che scrivere una grammatica trasformazionale, sia pure per un solo frammento di lingua naturale, è un compito di un elevato ordine di complessità. Non solo le regole individuali della grammatica devono funzionare, come è richiesto, prese isolatamente, ma esse devono lavorare correttamente anche insieme, per ottenere il risultato voluto. I dettagli della scrittura della grammatica sono considerati come secondari dal linguista, che è più interessato a ciò che è nella lingua e a come può essere generato, e generalmente preferisce prestare poca attenzione ai dettagli formali e notazionali. Così è naturale chiedersi se un calcolatore possa essere usato per assistere il linguista nello sviluppare una grammatica. Nel modello formale, c'è un gran numero di dettagli che devono essere elaborati, e le regole interagiscono le une con le altre in modo non sempre prevedibile.Molti degli errori che capitano nello 'scrivere' le grammatiche possono essere corretti, se solo essi sono portati all'attenzione del linguista.Il gruppo della Friedman ha scritto, in Fortran IV e per un sistema IBM 360/67, un sistema che accetta le grammatiche in input, come dati. "Il programma rappresenta la metateoria linguistica". Il modello linguistico incorporato è la teoria della grammatica trasformazionale, descritto da Chomsky nei suoi 'Aspects of the Theory of Syntax' /58/. " Il programma elabora tutti e tre i componenti di una grammatica trasformazionale: la struttura della frase, il lessico, e le trasformazioni". Esso compie l'intero processo di generazione della frase, compresa la generazione della struttura della frase, l'inserzione del lessico, e la trasformazione.

A Pisa abbiamo già ricevuto la documentazione relativa a questo programma, e ci proponiamo di adoperarlo nell'insieme degli studi da noi iniziati per l'italiano. Del resto, tale programma è stato usato per testare tre grammatiche dell'inglese moderno, una del swahili, una dell'inglese alfrediano, e una del francese: quest'ultima applicazione è illustrata, nella stessa sezione del Congresso, da A.M.Querido(Montreal).R.I. Binnick del 'Department of Linguistics' dell'Università di Chicago, discute la possibilità di applicare ai problemi della LC un modello di uso della lingua basato su una modificazione e estensione della teoria 'generativa semantica trasformazionale' della competenza linguistica recentemente sviluppata da P.Postal, G.Lakoff, J.R. Ross, H.D. McCawley, e altri. Egli presenta la teoria della semantica generativa come una conseguenza e una reazione alla teoria Chomschiana del 1965 sulla linguistica trasformazionale.

L. Karttunen del 'Department of Linguistics' dell'Università del Texas, considera un meccanismo concepito per leggere un testo in una lingua naturale, per interpretarlo, per poi memorizzare il contenuto in qualche modo, ad esempio con lo scopo di poter rispondere a delle domande su di esso contenuto. Per questo compito, la macchina dovrebbe avere almeno alcuni requisiti fondamen-

tali. Deve essere capace di costruire un 'file' che consiste nel le registrazioni di tutte le unità individue memorizzate nel testo (eventi, oggetti, ecc.). Per ogni 'individuo' dovrebbe essere registrato quello che di esso è detto. Naturalmente, almeno per ora, sembra che un simile interprete di testi non sia una idea realizzabile, ma questo non deve scoraggiarci dallo studiare in astratto quale specie di capacità la macchina deve possedere, visto che tale studio implica alcune ricerche sul linguaggio naturale in genere. L'autore discute in particolare una capacità che la macchina interprete deve avere: riconoscere quando è menzionata una nuova unità nel testo di input e memorizzarla con la caratterizzazione per futuri riferimenti.

Egli esamina sopratutto in che condizioni l'apparizione di un sintagma nominale indefinito implica l'esistenza di una qualche entità specifica che l'ipotetico interprete del testo deve registrare come una nuova individualità. L'esame è condotto alla luce dei problemi della 'referenza' nella grammatica trasformazionale.

8.3. Analisi computazionale.

8.3.1. Analisi di parole.

K.D.Bunting del 'Lehrstuhl fur Linguistik der Techinschen Universitä' di Berlino, parla di una ricerca empirica sui 'sistemi di derivazione' mediante affissi delle parole tedesche, condotta con l'aiuto del calcolatore all'Institut für Kommunikationsforschung und Phonetik' dell'Università di Bonn.

Egli definisce innanzitutto alcuni concetti del resto abituali negli spogli lessicografici: codice naturale (alfabetico grafemico), codice artificiale (notazione esplicita delle proprietà grammaticali di una data sequenza grafemica), unità linguistica o forma (determinata da certe proprietà grammaticali e dalle loro rappresentazioni per mezzo di una sequenza grafemica), alloforme (se un insieme di proprietà è rappresentato da più di una forma, le forme sono dette alloforme), omografi (se una forma rappresenta più di un insieme di proprietà), ecc. . Tenta di formalizzare questi concetti, per poter definire esplicitamente le unità di base del suo studio statistico sulla produttività dei diversi affissi nel lessico tedesco. Il materiale di base è un insieme di lemmi estratti da un dizionario tedesco. L'autore espone poi i risultati delle operazioni di conteggio. La comunicazione è interessante sopratutto per lo sforzo di definire rigorosamente gli elementi che occorrono in un lessico, sforzo che deve essere necessariamente compiuto da chiunque voglia intraprendere una statistica sul vocabolario, nel senso proposto, per es., da N.S.Troubetzkoy. /312 p.286/.

S.Zoltàn, romeno, affronta invece il problema della formazione per derivazione delle parole in una prospettiva trasformazionale, con esempi presi dall'ungherese.

E.Gammon del 'Department of Linguistics' del 'Fresno State College' di California, dichiara il suo interesse per l'eziologia delle idee sulle unità linguistiche. Egli vuole giungere a un modello che sia una formulazione particolarmente semplice di una 'procedura di scoperta' della parola come unità linguistica.

"La parola come unità occupa una posizione paradossale nella linguistica contemporanea. In una eccezione approssimativamente coincidente con i suoi impieghi nel linguaggio quotidiano e nei discorsi di scienze diverse dalla linguistica, essa occorre quasi universalmente nella pratica dei linguisti descrittivi come una linea divisoria fra i due livelli di costruzione, morfologico (in-

tra-word) e sintattico (extra-word). Non esiste ancora una definizione soddisfacente e generalmente accettata, e alcuni linguisti negano ogni validità della parola, relegandola alla linguistica popolare."(Si veda a questo proposito A.Martinet /206/ che Gammon non cita).

Gammon esamina i criteri proposti dalla scuola distribuzionalista per identificare le diverse unità linguistiche, e cerca dei corrispettivi statistici. Ecco per esempio uno dei criteri. Sia XABCDE una sequenza di morfemi appartenenti alla lingua da analizzare. Ci si chiede se tra B e C ci sia un confine di parola. Si conteggiano allora in un corpus (considerato come la totalità della lingua o come un campione di essa), i gradi di libertà e di combinazione di B a destra e a sinistra. In altre parole, si conteggia L (numero dei morfemi diversi, precedenti), e R (numero dei diversi morfemi seguenti). Se R ≠ L, allora si avrà un primo criterio per collocare il confine di parola tra B e C, e viceversa. Questo criterio si combina poi con altri, relativi ai gradi di combinabilità di C, di B con C, ecc.

8.3.2. Analisi di frasi.

Z.Bujas, del dipartimento di Inglese dell'Università di Zagabria,descrive un progetto di analisi contrastiva condotta con il calcolatore su 680mila parole serbocroate e sulla loro traduzione inglese. L'insistenza su un corpo così esteso, è dovuta alla convinzione, espressa dall'autore, che solo una ricerca estensiva delle corrispondenze (tra elementi della lingua originale e la loro traduzione), può adeguatamente rilevare le forme meno prevedibili, che avrebbero un considerevole potenziale informativo ai fini dell'analisi contrastiva. Lo strumento scelto per facilitare l'analisi è una serie di concordanze "contrastive": l'autore chiama così delle concordanze nelle quali ogni parola sia accompagnata da un doppio contesto: cioè dal contesto nella lingua cui appartiene e dal corrispondente contesto nell'altra lingua. Il calcolatore genera separatamente le concordanze del testo inglese e le concordanze del corrispondente testo serbocroato, e poi accoppia i contesti meccanicamente. La traduzione è avvenuta in modo che a una frase inglese corrisponde una e una sola frase serbocroata (e viceversa), e le frasi sono state numerate progressivamente. Perciò una frase originale e la frase che la traduce hanno lo stesso numero progressivo, e su questo numero il calcolatore può operare l'accoppiamento.

8.3.3. Lessico.

Harry H.Josselson dello 'Slavic Department Research' dell'Università del Wayne, propone il concetto di 'matrice', come strumento per organizzare un lessico corredato di una descrizione (codificazione) consistente delle parole. Egli vuole cioè assicurare che i codici posti ad alcune parole vengano posti a tutte le parole della stessa classe. Egli osserva che l'introduzione del calcolatore ha generato un cambiamento nel pensiero, nelle tecniche e nei risultati della linguistica. L'essenza del cambiamento sta nel fatto che generalizzazioni vaghe, espresse da frasi del tipo 'parole che hanno questo significato generale sono spesso incontrate in queste e in simili strutture', sono sostituite con la precisa definizione di regole e con la enumerazione di insiemi completi di parole definite da una proprietà data.

Egli si preoccupa di applicare questi principi a un lessico operabile dal calcolatore, ed enumera le diverse fasi necessarie alla sua compilazione.

a) Decidere quali parole introdurre nel lessico, (lo 'stock lessicale'). Idealmente un lessico di una lingua deve contenere tutte le sue parole; ma ci sono parole che escono gradualmente dall'uso, mentre altre vengono continuamente formate e aggiunte. Le parole da immettere nel lessico possono essere desunte da sorgenti esistenti (lessici e dizionari tecnici) e completate da neologismi presi da nuovi testi. Il lessicografo deve anche essere in allarme per registrare eventuali nuovi significati o contesti in cui la parola 'vecchia' può apparire.

b) Dividere in sottoinsiemi lo 'stock' lessicale.
Lo 'stock' lessicale può essere suddiviso in classi di parole, parole che hanno alcune proprietà in comune: le proprietà possono essere morfologiche, funzionali, ecc. I sottoinsiemi non sempre coincidono con quelli delle grammatiche tradizionali.

c) Chiedersi quali informazioni devono essere codificate per ogni sottoinsieme. La scelta è evidentemente una funzione degli scopi proposti al lessico; in altre parole il compilatore deve codificare le informazioni che saranno necessarie quando il lessico sarà utilizzato per gli scopi previsti. (Osservo che buona parte della letteratura sull'argomento (cfr. Zampolli /333/ e /338/ consiglia di assegnare a ciascuna parola il maggior numero di informazioni possibili: per es. di natura fonologica, stilistica, etimologica, ecc.). Josselson accenna all'opinione di ricercatori che lavorano in quest'area: è possibile identificare e codificare le proprietà sintattiche, ma trovano impossibile fare lo stesso nella sfera della semantica. Eppure molto del nuovo lavoro in lessicologia include l'analisi dei predicati e dei loro argomenti (soggetti, e complementi, sintagmi nome/aggettivo, sintagmi preposizionali, ecc.). Josselson ritiene che Ch.J.Fillmore con i suoi casi grammaticali (agente, strumento, oggetto, ecc.) abbia iniziato la transizione da una codificazione puramente sintattica a una codificazione semantica.

d) Compilare la informazione.
Il fattore chiave - secondo Josselson - è che quando uno "statement" è fatto per un membro di una classe di parole, tutti gli altri membri di questa classe devono essere codificati secondo il modo nel quale questo "statement" si applica loro. Egli suggerisce, per garantirsi da lacune in questo senso, di costruire una matrice per ogni classe di parole. Ogni colonna della matrice rappresenta un 'bit' di informazione pertinente a questa classe, e ciascuna parola di questa classe occupa una riga. Ogni intersezione deve essere riempita, per dire se la parola ha o no tale proprietà, e per specificare la modalità di tale proprietà, quando appaia. La matrice sarebbe utile sia in fase di compilazione sia in fase di controllo semiautomatico della compilazione.

e) Strutturare la registrazione della informazione.
Josselson discute la forma nella quale deve essere rappresentata l'unità lessicale: lemma -tema- tutte le forme flessive, e i vantaggi e gli svantaggi di ciascun organizzazione. (Poichè anche a Pisa stiamo compilando un dizionario di macchina per l'italiano, tornerò sù questi argomenti più oltre).

8.3.4. Riconoscimento automatico del suono.

A.Iivonen, dell'Istituto di fonetica dell'Università di Oulu, (Finlandia), propone diversi tipi di strategie per il riconoscimen-

to automatico dei "Suoni del discorso" per mezzo di un calcolatore digitale. Gli esperimenti sono condotti per le vocali e le consonanti nasali del finlandese e del tedesco. La prima strategia, applicata a fenomeni vocalici lunghi del tedesco, ha dato una probabilità di correttb riconoscimento del 94%, ed è basata sull'analisi discriminate di un input spettrografico. Le varianti considerate sono: le frequenze delle prime quattro formanti, la loro ampiezza, il minimo di energia tra di esse, e la durata. Nel secondo esperimento, per otto vocali brevi e tre nasali finniche, i dati di input sono costituiti da descrittori numerici dei suoni. Essi sono formati usando punti costanti nella misurazione delle spettro dei suoni. I descrittori di un suono consistono in una serie di numeri, che indicano l'ampiezza a frequenza costante selezionata. I dati così ottenuti sono stati memorizzati e sottoposti ad analisi. I punti di misurazione sono stati trattati come variabili. La probabilità di una identificazione corretta era del 60-70%. Nella terza strategia, alcuni descrittori numerici scelti opportunamente tra i diversi ottenuti, sono stati ordinati in modo da costituire un numero, che è adoperato come modello numerico per il riconoscimento di un gruppo di suoni nasali finlandesi.

8.5. Linguistica diacronica e comparativa, dialettologia.

8.5.1. Fonologia - Nello studio di E.A.Afendras dell'International Center for Research on Bilingualism' dell'Università Laval (Canada) " i sistemi vocalici di numerose lingue balcaniche (passate e presenti) sono analizzati in termini dei tratti distintivi di Jakobson". L'autore discute e valuta diversi metodi, per comparare e graduare il grado di similarità dei sistemi. In particolare introduce alcuni indici tipologigi; per es.quello proposto nel 1966 dalla linguistá sovietica V.I. Postovalova, relativi alla distribuzione dei tratti distintivi in un sistema fonologico.

a) Probabilità semplice: cioè la frequenza di utilizzazione di un tratto da parte del sistema per la composizione dei fonemi. Le tre possibilità +,-,O, sono considerate separatamente. Per il tratto a abbiamo $Pa = \frac{m}{n}$ dove m è il numero di righe con a, n il numero di righe complessivo.

b) Probabilità condizionale

Indicano quanto frequentemente coppie diverse di tratti caratterizzano i fonemi nel sistema; in altre parole, dato il tratto a per un fonema, quante probabilità ci sono che il tratto b si combini con a nello stesso fonema. Questa possibilità è :

$p_a(b) = \frac{m'}{n'}$ dove m' = numero di fonemi con a e b

n' = " " " " $\frac{a}{a}$

c) Infine, la valenza è definita per informare sulla combinabilità di un tratto con gli altri e anche sul sistema come intero.

$V_a(b) = \frac{Pa(b)}{n-1}$ dove n = numero complessivo di tratti del sistema.

L'autore, pur affermando esplicitamente le difficoltà poste alla sua impresa, per esempio della mancanza di una solida descrizione in termini di tratti distintivi di parte dei 51 sistemi confrontati, ha stabilito delle correlazioni statistiche tra le matrici vocaliche di questi sistemi, assumendo come parametri gli indici suddetti. Egli ha usato due diverse analisi fattoriali con vari programmi di calcolatore. Dai risultati conseguiti, l'autore conclude che gli sembrano dimostrati i vantaggi di questo approccio per una classificazione tipologica delle lingue, e propone ul-

teriori ricerche per modelli matematici nello stesso settore, così da approfittare del potere esplicativo di una teoria matematica ben sviluppata.
S.P.Durham e D.E.Rogers del 'Department of Romance Languages' dell'Università di Michigan, illustrano un programma per produrre delle concordanze di parole aventi come criterio di ordinamento la coincidenza dei segmenti (fonemi) e dei contesti fonematici. I fonemi sono definiti a priori, per mezzo di una lista di consonanti e di vocali. Un contesto è definito come il luogo di occorrenza di ciascun fonema. La concordanza dovrebbe facilitare il riconoscimento delle corrispondenze foniche per la ricostruzione di un protolinguaggio.
Peter Ladefoged, del 'Department of Linguistics' dell'Università di California, presenta un metodo per misurare la similarità fonetica tra lingue o dialetti appartenenti allo stesso gruppo. Nel Sud dell'Uganda vivono 6 milioni di persone, ci sono numerose lingue e dialetti bantu: ufficialmente ne sono recensiti 15. La similarità fonetica è ricercata come dato per decidere quali linguaggi usare per la radio (attualmente ne sono usati 9) e quali usare nelle scuole (attualmente ne sono usati 3 ufficialmente e 5 non ufficialmente).

Secondo R.N.Smith, del 'Department of Linguistics' della 'Northwestern University',(Illin.) i due maggiori problemi degli studi di linguistica storica sono la struttura dei cambiamenti fonetici, e la relazione tra questo cambiamento e gli altri,cioè la sua cronologia relativa. L'autore ritiene che la mancanza di rigore per quest'ultimo studio sia conseguenza, tra l'altro,della grande quantità di dati che devono essere elaborati dalle regole e dal gran numero di regole che devono essere verificate per lo studio della evoluzione. Un aiuto nella verifica può essere fornito da un calcolatore; egli ha scritto un programma per applicare le regole storiche del mutamento fonetico.

Con 21 regole e 500 forme protoindeuropee ricostruite tratte dal Vergleichendes Worterbuch der indogermanischen Sprachen di A.Walde e J.Pokorny /321/, egli ha ottenuto alcune - ma non tutte - le forme del russo moderno che egli si attendeva.

Willian S-Y.Yang, autore tra l'altro, in collaborazione con Anatole Lyovin, di una Chinese Linguistics Bibliography on Computer, in stampa presso la 'Cambridge University Press', espone una ricerca analoga per la ricostruzione della evoluzione fonologica dei maggiori dialetti cinesi.

8.5.2. Dialettologia.

S.Klein, M.A.Kuppin, K.A.Meives, del 'Computer Sciences Department' dell'Università di Wisconsin, presentano la simulazione col 'metodo Monte Carlo' dell'evoluzione linguistica in Tikopia e Maori. Le persone rappresentate nella simulazione dalle loro grammatiche, "eseguono" delle conversazioni e alterano le loro regole per analizzare quello che "odono".

W.Skalmowski e M.Van Overbeke, dell'Istituto per la Linguistica Applicata di Lovanio, presentano i risultati del confronto tra testi olandesi scritti da autori bilingui (parlanti francese e olandese : bilingui nel senso definito da E.Haugen /133 p.9/), con testi considerati come rappresentativi dell'olandese scritto 'standard' (testi della letteratura contemporanea olandese e fiamminga), allo scopo di studiare i fenomeni di interferenza lessicale, particolarmente importanti per il neerlandese, che contiene in gran numero prestiti francesi ed è ancora una

'porta aperta' all'intrusione di gallicismi. L'autore nota una rassomiglianza con la situazione dell'inglese con le sue parole anglosassoni e romanze, e studia un modello di interferenze, basato da un lato su certe ipotesi sulla diversità di comportamento delle due classi suddette di parlanti, e dall'altro sulle considerazioni statistiche generali di Herdan, sull'interferenza lessicale e sulla legge di Zipf-Mandelbrot. /146/.

W.N.Francis, J.Svartvik, G.M.Rubin del 'Department of Linguistics' della Brown University, affermano che la dialettologia, sopratutto nella sua fase interpretativa, è un ramo della linguistica particolarmente adatto per l'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 ad ordinare e comparare elementi individuali su diverse basi: fonetiche, morfologiche, sintattiche, geografiche, ecc. Il maggior ostacolo che fino a qui ha allontanato dall'uso dei calcolatori per lo studio dei dialetti, è il fatto che i dati per i dialetti più importanti sono stati raccolti ed editi prima dell'epoca dei calcolatori. (Apro una breve parentesi,per addurre a conferma l'esperienza da me acquisita a questo proposito nello studio per la trasposizione, su supporti accettabili dal calcolatore, dei quasi 5 milioni di risposte dialettali,raccolte in più di 40 anni di inchieste, al questionario dell'ALI (Atlante Linguistico Italiano), che saranno elaborate elettronicamente al C.N.U.C.E., al fine di ottenere la pubblicazione del copiosissimo materiale, in forma di indici a più entrate, e, in un secondo tempo, la messa in carta automatica dei materiali stessi

A differenza degli autori, contiamo di usare, al posto di un plotter, una macchina fotocompositrice, che assicura, accanto a una maggiore velocità di stampa, una presentazione grafica nettamente superiore della pubblicazione. L'idea di combinare la recente capacità di 'map making' del calcolatore, con le possibilità di operazioni statistiche, e con la creazione di un archivio dialettale gestibile automaticamente, ci è venuta indipendentemente dall'esempio americano, e questo fatto è una ulteriore conferma della esistenza del problema e della possibilità dell'uso del calcolatore in dialettologia. Recentemente ho saputo che progetti analoghi sono in corso in Germania, e all'Università del Wisconsin, ove sotto la direzione di G.Cassidy, si prepara il DARE (Dictionary of American Regional English); si veda anche il lavoro di R.W.Shuy, a proposito di un programma di 'automatic retrieval' per l'Atlante Linguistico degli Stati Uniti e del Canada /301/.

Il problema di preparare una grande quantità di dati, per lo più in trascrizione fonetica, per il calcolatore, è un compito formidabile - osservano gli autori - ; lo scopo della loro comunicazione è mostrare che i risultati, ottenuti in modo relativamente agevole, con un enorme risparmio di ore-uomo per ordinare e mettere in carta, rispetto ai metodi tradizionali, sono tali da incoraggiare senza dubbio altri ad investire tempo e danaro nel preparare i dati per il calcolatore piuttosto che nell'ordinare e fare la carta a mano. In particolare, essi illustrano un esperimento di utilizzazione dei materiali dialettali, registrati nel 4° volume del 'Survey of English Dialects' /238/, per studiare le variazioni nella situazione delle fricative iniziali in 10 contee del sud dell'Inghilterra. Dopo aver trascritto su schede, per ciascuna risposta, la parola inglese standard, l'equiva-

lente dialettale in trascizione fonetica e il numero codice della località, hanno chiesto al calcolatore elenchi di vario tipo, nei quali le classi di informazione registrate sono ordinate variamente le une rispetto alle altre. Ulteriori liste e conteggi statistici sono stati eseguiti per rilevare la distribuzione dei fenomeni fricativi all'iniziale in relazione con i contesti fonici seguenti. Contemporaneamente, il calcolatore stampava una carta per ciascuna delle distribuzioni studiate.
Gordon R.Wood,dell'English Faculty, Humanities Division' della 'Southern Illinois University', illustra una applicazione analoga, per la 'scoperta' e lo studio dei diversi tipi di Americano regionale, lungo la costa dell'Atlantico. Dopo considerazioni sulla difficoltà della raccolta e della preparazione dei dati,e sul risparmio di tempo permesso dal calcolatore per il loro ordinamento e spoglio esaustivo,l'autore illustra i metodi seguiti, e in particolare l'utilità della lista di frequenza per identificare le parole che, scelte a preferenza di altre di significato equivalente, caratterizzano una preferenza tipica di un sistema linguistico regionale; e l'utilità delle concordanze, nel senso tradizionale del termine, per studiare le varietà grammaticali e sintattiche.

8.6. Semantica. Come abbiamo già detto, i problemi che vengono genericamente riferiti come appartenenti alla semantica,si sono affacciati ripetutamente in tutte le sezioni del Congresso. I temi sono quelli che occorrono spesso nelle discussioni tra strutturalisti, trasformazionalisti chomskiani e postchomskiani e che sono emersi anche in due recenti convegni sulla grammatica in Italia organizzati rispettivamente dalla SLI a Roma e dal Centro per l'Insegnamento dell'Italiano a Trieste. Si discute il ruolo dalla descrizione del lessico, e la distinzione tra sintassi e semantica da un lato, e tra conoscenza linguistica e 'conoscenza del mondo' dall'altro. I lavori più spesso citati sono quelli di A.J.Fodor, J.J.Katz, P.M.Postal, M.R.Quillian, G. e R. Lackoff, J.G.MacCawley, J.R.Ross, P. e C.Kiparsky, e sopratutto CH.J.Fillmore; frequenti sono anche i riferimenti alla logica formale. Alcune comunicazioni trattano solo problemi teorici,senza riferimento esplicito all'uso dei calcolatori. Altre, lo propongono per la compilazione e gestione di thesaurus, o addirittura come tipo autonomo di approccio alla formalizzazione della semantica delle lingue naturali. E' il caso di R.M. Schwarcz (California) che esamina questo problema con grande cura : egli discute le diverse teorie linguistiche della semantica e i problemi logico-teorici connessi, passa in rassegna i progetti computazionali significativi e propone un tipo di approccio basato sulla teoria 'operazionale' del significato. Di carattere teorico e generale sono anche la comunicazione di I.Bellert (Varsavia) e quella di J.Rouault (Grenoble) che applica categorie della logica alla semantica di lingue naturali. B.Vauquois et alii (Grenoble), propongono, per un progetto di traduzione meccanica, un linguaggio pivot (metalinguaggio intermediario tra lingua d'entrata e lingua d'uscita) come sistema di notazione per registrare 'il significato del testo indipendentemente da particolarità grammaticali e lessicali di lingue naturali'. Alla traduzione automatica lavorano anche Dugas et alii (Montreal), che praticano un'analisi dei lessemi in tratti semantici.E.Vasiliu (Bucarest) studia la 'categoria-tempo' e la sua interpretazione semantica riferendosi ai sistemi logici proposti da Carnap. D.Wunderlich (Berlino) si riferisce invece all'opera dei logici R.Montague e

J.A.W.Kamp nella sua comunicazione 'On time reference and tense'.

8.7. Information Retrieval Secondo C.A.Montgomery,(California) gli scienziati dell'informazione sono interessati al linguaggio per la sua funzione di veicolo primario di comunicazione di informazioni nella società umana, e ricercano un sistema automatico per analizzare il contenuto di testi naturali. Il sistema ideale dovrebbe identificare i concetti presenti e le loro interrelazioni, sulla base di una qualche forma di analisis sintattica e semantica (questo è il punto di convergenza con gli interessi dei linguisti), e tradurli poi in un insieme di 'frasi canoniche'. Queste rappresenterebbero il contenuto del documento, avendo i vantaggi della non ambiguità e della indipendenza da una lingua particolare; costituirebbero, per così dire, le 'conoscenze' del sistema, e servirebbero di base per generare risposte fattuali a richieste specifiche. Quanto si sia lontani da questo ideale è dimostrato se non altro dal fatto che i maggiori successi nel settore dell'information retrieval li hanno ottenuti, come osserva D.G.Hays /138/ coloro che hanno rinunciato ad applicare teorie linguistiche, e si sono riferiti a modelli puramente documentaristici. Tuttavia non sono mancate comunicazioni molto interessanti, tra le quali cito quelle di G.Salton (Cornell Univ.), che descrive un sistema 'conversazionale' on line utente-calcolatore, di A.V.Pratt e M.G.Pacak (Bethseda) che lavorano all'analisi automatica di documenti medici, e di Szanser A.J., che descrive un metodo per la correzione della registrazione di testi di lingua naturale. Il problema della correzione dell'input è fondamentale per la CL, sopratutto per progetti che prevedono l'elaborazione di grandi corpus, come per es. quello per il Grande Vocabolario Storico dell'Accademia della Crusca.

9. LA SEZIONE LINGUISTICA DEL C.N.U.C.E.

"Non ho una memoria elettronica (....) Non c'è che da cominciare a preparare un po' di schede perforate per il nostro 'robot' filologico: allestire spogli o anzi glossari completi di più testi o autori che si possa, anche di breve respiro. Come già mi accadde di suggerire altre volte, io vi esorto alle Concordanze." (G. Contini, 68a p.272).

Numerosi linguisti si son rivolti all'uso del calcolatore per "semplificare e sveltire certe ricerche linguistiche e letterarie che per loro natura comportano una grande quantità di lavoro materiale....., da eseguire su un gran volume di dati". Su questo terreno, essi si sono trovati a dover formulare gli algoritmi risolutivi delle proprie elaborazioni, (Nota 3) e quindi a riconoscere la possibilità o la necessità di una formalizzazione. Questa esperienza può diventare uno stimolo intellettuale fortissimo a cercare algoritmi anche per tipi di problemi che "per consuetudine o timidezza intellettuale si sono sempre affrontati con metodi euristici" /241,p.11/. Linguisti e filologi, di formazione e interessi dichiaratamente tradizionali, si sono trovati ad affrontare problemi comuni, in fin dei conti, a quella parte della linguistica matematica che Lepsky chiama algoritmica : per esempio filologi classici compilano regole generative per la flessione automatica di un lemmario che faciliti le operazioni di spoglio, e lessicografi impegnati nella redazione di grandi vocabolari storici nazionali esplorano le possibilità di rendere (almeno parzialmente) automatica la distinzione tra le forme omografe per mezzo dell'analisi automatica del contesto. D'altra parte, la registrazione di corpus sempre più estesi di

testi (in linguaggio naturale) per le più svariate ricerche linguistiche e letterarie, pone il problema della definizione delle unità linguistiche a tutti i livelli; il problema è comune alla statistica, alla quale nel contempo si forniscono i dati quantitativi fondamentali, realizzando così i voti espressi dal grande semitista e comparatista /304/ Marcel Cohen al VI° Congresso Internazionale dei Linguisti tenutosi a Parigi nel 1948 /66/.

Parafrasando uno slogan ormai scontato, potremo dire che la LC sta entrando nella seconda generazione mentre i calcolatori entrano nella terza. Si scrivono manuali di LC /134/ e /135/ si tengono corsi di questa materia nella principali università americane (e ormai da 3 anni ne svolgo uno al CNUCE dell'Università di Pisa), si compilano linguaggi programmativi ad uso specifico dei linguisti, come il COMIT e lo SNOBOL, ecc. Indubbiamente i progressi dello hardware e la specializzazione del software offrono alla linguistica computazionale nuove possibilità.(Nota 4).

L'input è facilitato da testi già perforati su banda dagli editori; i linguaggi convenzionali di programmazione per il dialogo diretto uomo-macchina permettono l'intervento immediato nel processo d'analisi dei testi e nella simulazione dei modelli linguistici; la macchina fotocompositrice unisce alla varietà pressochè illimitata dei caratteri e alle qualità estetiche della stampa, l'esattezza assoluta della composizione. (Nota 6).

In questa prospettiva, si situa come fattore di principale importanza il costituirsi di immense biblioteche elettroniche, e diventa sempre più probabile per uno studioso trovare già registrati elettronicamente i testi che gli interessano.

In particolare nel settore della documentazione lessicale, l'utilità dell'impiego dei calcolatori è ormai un dato di fatto che oggi nessuno più discute perchè risponde ad esigenze da sempre drammaticamente presenti nella storia della disciplina, che R. Wisbey ha pittorescamente rappresentate. /328/.

I centri per la automazione delle ricerche lessicografiche e linguistiche si sono moltiplicati in tutto il mondo, e sono così numerosi che una loro rassegna richiede ormai delle pubblicazioni a parte: posso qui solo rinviare ad alcune tra esse, negli atti dei convegni che periodicamente riuniscono gli esperti del settore,/3/ /5/ /16/ 197/ e ai bollettini e alle riviste e ai panorami editi dai centri stessi. /16/ /45/ /64/ /65/ /39/ /67/ /91/ /263/ /308//305/ /209/ /79/.

9.1. La sezione linguistica del C.N.U.C.E.

In Italia, il primo progetto di lessicografia automatizzata è quello dell'Index Thomisticus, iniziato da Padre Busa S.J. a Gallarate nel 1949 : al suo progetto, anzi, è riconosciuta,quasi universalmente, la priorità internazionale. Nel 1964, per iniziativa e sotto la direzione del Prof. A.Duro, furono eseguiti a Gallarate i primi esperimenti di spoglio per il grande Dizionario della Crusca; nel 1966 entrambi questi progetti si trasferirono sugli impianti del CNUCE. L'esempio, l'autorità e il successo delle applicazioni della Crusca, suscitarono presto altri propositi, e così oggi in Italia tutti o quasi tutti i progetti in atto utilizzano le attrezzature della Sezione Linguistica del Centro Nazionale Universitario di Calcolo Elettronico dell'Università di Pisa, costituita due anni or sono sotto la mia direzione.

Questa sezione, (Nota 6) ha approntato un insieme di programmi di utilità (utility programs), esperimentati in diversi setto-

ri della analisi di testi in linguaggio naturale, che vanno dalla statistica fonematica e dalla statistica lessicale agli spogli lessicografici.

Gli Istituti Universitari e del C.N.R. che si avvalgono dei programmi, degli impianti, della collaborazione scientifica e tecnica della sezione linguistica del C.N.U.C.E. sono oggi poco meno di 50, distribuiti in tutte le regioni italiane; (Nota 7) per la loro elencazione e descrizione rinvio a /336/ /340/ /342/ e /88/. Qui elenco solamente i principali settori nei quali operano.

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

Ricordo in particolare le imprese per la lingua italiana/87/, per l'ittita, per il rumeno.

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

Ricordo a titolo di esempio, il dizionario storico italiano della lingua giuridica e il lessico intellettuale europeo.

9.2.3. Documentazione lessicografica per la redazione di dizionari glossari o lessici dell'opera omnia di un autore.

Sempre a titolo di esempio nomino gli spogli dell'opera omnia di Antonio Rosmini, di L. A Seneca, di S.Tommaso di Aquino, ecc.

9.2.4. Spoglio di opere di singoli autori, per studi diversi (metrici, grammaticali,stilistici,tematici,ecc.) su base lessicale e statistica:

Opere di Pindaro, Bacchilide, A.Gide, U.Saba, A.Pavese, A. Macado, ecc.

9.2.5. Inchieste dialettali, atlanti linguistici, testi dialettali

Le inchieste si concretano normalmente in uno schedario delle risposte dialettali al questionario. La gestione di uno schedario di informazioni è una delle applicazioni tipiche, fondamentali della meccanografia. Ricordo in particolare l'impresa della edizione e degli indici dell'Atlante Linguistico Italiano e le inchieste minori associate. (cfr. 8.5.2.).

9.2.6. Elaborazioni di grandi corpus di origine diversa

Riunisco in questa categoria alcuni progetti che, pur non avendo come oggetto principale delle loro ricerche lo studio della lingua dei corpus, operano su di essa per i loro scopi: edizione, sistemazione, catalogazione degli elementi del corpus.Come esempio cito il progetto per la edizione della raccolta Barbi di canti popolari italiani /335/,quello che la elaborazione metrica e la classificazione di altre raccolte già edite di canzoni popolari, lo studio dei canovacci della Commedia dell'Arte, ecc. Questi studi, naturalmente, producono nello stesso tempo una documentazione linguistica di estremo interesse.

9.2.7. Statistica linguistica

I testi dei diversi settori registrati su nastro magnetico 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, (ittica, sanscritto, greco antico, latino della classicità, ebraico, aramaico, nabateo, latino mediovale rinascimentale e scientifico, italiano dalle origini ai giorni nostri

scritto e parlato, francese, tedesco, spagnolo, poleoslavo, 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: come vedremo più avanti,la applicazione e la verifica delle teorie e delle formule della statistica linguistica a un gran numero di testi nei quali le unità linguistiche siano uniformemente e rigorosamente definite, è condizione essenziale per la sopravivenza stessa della statistica linguistica come scienza autonoma.

Ciò premesso, rinvio alla lezione di domani per la esemplificazione di alcuni progetti di statistica linguistica in corso al C.N.U.C.E.

9.2.8. Psichiatria, psicologia, pedagogia.

Il calcolatore ha trovato numerose applicazioni in questa disciplina /63/, in particolare per l'analisi del comportamento linguistico, attività privilegiata dell'uomo.

Ricordo il progetto per l'analisi statistica di conversazioni di diverse categorie di malati mentali e per lo studio della evoluzione del linguaggio infantile. /283/.

9.2.9. Altre discipline (che gli americani designerebbero con il collettivo "Humanities").

Fanno capo alla sezione linguistica del C.N.U.C.E. anche progetti nei settori della musicologia (testi e musica del rinascimento), della archeologia (catalogo della ceramica greca), della storia (analisi statistiche, concordanze, indici di nomi e cose notevoli di documenti notarili e atti legali del IX, X, XI secolo in Italia.

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

a) I ricercatori che iniziano un loro progetto presso di noi,vengono innanzitutto 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 (professori, assistenti, studenti, documentalisti, ecc.).
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.

9.4. Terminologia delle operazioni di spoglio.

Tutti questi settori si propongono, come prima tappa, la registrazione dei testi (parlati o scritti) in forma operabile dalle macchine, l'individuazione delle unità che le compongono e la classificazione e il censimento di queste unità in forme diverse, alcune delle quali sono ormai consacrate dalla tradizione. Prima di descrivere le procedure elaborative e i problemi tecnici o scientifici connessi, preciso alcuni termini propri, per così dire,del

nostro gergo.

Insisto sul fatto che, data la natura di questo scritto, non do delle definizioni scientifiche, ma indico semplicemente il contenuto dei termini in questione, o al più ne do una rappresentazione intuitiva o una esemplificazione. Le parole sono in corsivo nel testo là dove vengono così illustrate.

Chiamiamo occorrenze (o parole) le unità successive di cui è costituito un testo. Per la macchina, la parola è una unità grafica; dal punto di vista della sua delimitazione, i caratteri del testo si dividono funzionalmente in due categorie : caratteri costituenti la parola (lettere; segni modificatori di lettere: accenti; contrassegni di corpo: maiuscolo, corsivo, ecc.; sostitutivi di lettera: apostrofo, lacuna, abbreviazioni) e caratteri separatori (punteggiatura, spazio, segni di ripartizione del testo : semiverso, verso, fine riga, paragrafo, ecc.).

Una occorrenza è definita allora come una serie ininterrotta di caratteri costituenti tra due separatori. (Nota 8).

L'insieme delle indicazioni che individuano la posizione della parola nel testo costituiscono il suo riferimento. Esso può consistere semplicemente in un numero progressivo, ma di solito è maggiormente articolato per permettere il ritrovamento della parola nel testo stampato, e rientra in una di queste due categorie principali:

- riferimento tipografico : per es. volume, tomo, pagina, riga del testo dove appare la parola.
- riferimento organico : quando l'autore o l'editore hanno suddiviso il testo in parti, la parola è identificabile per mezzo del nome e/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. Il riferimento organico è preferito a quello tipografico quando si presume che gli utilizzatori degli spogli si serviranno di edizioni diverse. Al C.N.U.C.E., d'abitudine, registriamo entrambi i tipi.

Non tutte le parole del testo sono diverse, alcune sono ripetute due o più volte: l'elenco delle forme grafiche è l'elenco delle parole diverse, nel quale appaiono una sola volta sia le parole ripetute sia le parole non ripetute. Accanto a ciascuna forma apparirà il numero delle volte che ricorre nel testo, cioè la sua frequenza assoluta.

Il calcolatore considera eguali le parole composte da una serie assolutamente identica di caratteri : è noto che a una stessa forma grafica possono corrispondere significati (uso questa parola non nel senso tecnico di una qualche scuola linguistica, ma nella sua accezione corrente) diversi : per es. la forma danno può essere sia la 1ª persona del presente indicativo di dannare , sia la 3a persona plurale del presente indicativo di dare , sia il sostantivo sinonimo di guasto , lesione .Diciamo che la forma grafica danno è omografa, e corrisponde a tre forme lessicali distinte.

Il calcolatore, operando sulla base della identità grafiche, accomuna tutte le occorrenze di danno in un unica forma grafica: chiamiamo smistamento degli omografi l'operazione di ripartire le occorrenze omografe tra forme lessicali distinte.(Nota 9).

Se abbiamo verificato che le diverse forme lessicali corrispondenti ad una forma omografa, sono presenti di fatto nello stesso testo, diremo che tale forma è omografa attuale nel testo.

Prima di tale verifica,oppure se solo una delle forme lessicali omografe appare nel testo, non possiamo dire che la forma sia attualmente omografa, ma solo che può essere omografa nel sistema linguistico considerato, e la chiameremo forma omografa possibile.

Chiamiamo lemma l'insieme di quelle forme lessicali che sono usualmente raggruppate dai dizionari in una unica voce o unità lessicale(pensiamo di dover cercare le forme dell'esempio dantesco in un vocabolario: non cercheremo certo "va", ma "andare", e non troveremo "perduta" ma solo "perdere" o "perduto") e che spesso sono flessioni di un unico paradigma: naturalmente negli indeclinabili la flessione è unica, ma le forme possono essere più di una; per es. "e" "ed" "ancora" "ancor".

Chiamiamo esponente del lemma la forma di base, quella cioè con cui ciascuna unità lessicale viene registrata usualmente nel vocabolario. (Nel linguaggio corrente si abbrevia "esponente del lemma" in "lemma".

Chiamiamo "lemmatizzazione"o, come preferirebbe dire G.Devoto, lemmazione, l'operazione di ricondurre ciascuna forma e le relative occorrenze al rispettivo lemma. L'operazione di smistamento degli omografi costituisce normalmente una parte della lemmatizzazione.

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

Forme e lemmi reperiti nello spoglio possono essere ordinati secondo diversi criteri. Per esempio in ordine alfabetico diretto : il calcolatore, normalmente, dispone di 26 lettere dell'alfabeto inglese con cui opera nell'ordine lessicografico tradizionale; il problema si complica quando tra i componenti della parola appaiono modificatori di lettera o caratteri sostitutivi,che il calcolatore dispone dopo la Z se sono codificati con una cifra, e prima della A se sono codificati con uno dei cosidetti segni speciali (special marks : + - =,ecc.).
Ancora più complesso il caso di un testo traslitterato in alfabeto latino da un alfabeto diverso; per es. nella traslitterazione da noi usata per i testi ittiti elaborati per il dizionario di P.Meriggi, la seguenza delle lettere è :

In questo caso si suole accoppiare ad ogni carattere traslitterato un numero di 2 cifre, che corrisponde al rango del carattere nella sequenza richiesta, e su di esso il calcolatore opera la selezione.

Mentre nell'ordine alfabetico diretto le parole sono ordinate in modo da raggruppare assieme prima le parole con la stessa iniziale, poi, nell'ambito della stessa iniziale, quelle con lo stesso digramma iniziale, quindi nell'ambito dello stesso digramma quelle con lo stesso trigramma iniziale, e così via, nell'ordine alfabetico inverso, le parole vengono invece ordinate prima secondo la lettera finale, poi, nell'ambito di una stessa finale, secondo il digramma finale, e così via. (Nota 9).

Il rimario (nella nostra terminologia) è l'indice alfabetico inverso delle sole parole in posizione di rima nei versi /71/, spesso accompagnate dal verso.
Il rimario attuale /335/ è un rimario nel quale, per ciascuna forma in posizione di rima, sono elencate le forme che con essa si trovano di fatto a corrispondere (rima, assonanza, consonanza , ecc.) in posizione di rima, secondo la struttura del componimento poetico.

Nell'elenco dei lemmi (o forme) per frequenza decrescente, i

lemmi o (le forme) sono elencati dal più frequente al meno frequente: lemmi (o forme) di frequenza uguale sono ordinati alfabeticamente entro la stessa classe di frequenza.

Se numeriamo progressivamente, in tali liste, i lemmi (o le forme), il rango del lemma (o della forma) è il suo numero progressivo. Se numeriamo progressivamente le diverse frequenze, il numero assegnato a lemmi (o forme) di eguale frequenza, è la loro classe di frequenza.

Se alla frequenza del 1° lemma (o forma) sommiamo la frequenza del II° e del III°, ... dell'n°, il totale così ottenuto per l'n° lemma (o forma) è la sommatoria delle frequenze corrispondenti al lemma (o forma) di rango n. (Nota 11).

Chiamiamo rationarium l'elenco alfabetico dei lemmi, a ciascuno dei quali sono aggiunte, intercalate in ordine alfabetico, grammaticale, o altro, le forme che gli appartengono.

Se in una lista alfabetica di lemmi, o di forme, e in un rationarium, aggiungiamo i riferimento delle occorrenze di ciascun lemma o forma, abbiamo un index verborum.

Chiamiamo contesto di una occorrenza, un insieme variamente delimitato di occorrenze che le sono contigue, o comunque associate, nel testo, scelto per lo più fra quelle che la precedono o la seguono immediatamente.

Se ad un index verborum aggiungiamo, per ciascuna occorrenza del lemma o della forma, oltre al riferimento, anche il suo contesto, otteniamo quella che chiamiamo lista di concordanze o concordanza.

Chiamiamo rispettivamente lemma esponente e forma esponente del contesto il lemma e la forma per cui il contesto è stato composto e a cui è assegnato.

I contesti (per lo più in ordine di riferimento) possono essere raggruppati secondo criteri diversi. I più comuni sono:

a) - All'interno delle rispettive forme, ordinate alfabeticamente.
b) - All'interno dei rispettivi lemmi, ordinati alfabeticamente.
c) - All'interno delle forme, raggruppate a loro volta alfabeticamente (o in altro modo) sotto i rispettivi lemmi, lessicograficamente ordinati.

Nel caso a) parliamo di concordanze per forme, nel caso b) di concordanze per lemmi, nel caso c) di concordanze per lemmi e forme.

A seconda del grado che occupano nella gerarchia, chiamiamo gli esponenti maggiori, minori ed eventualmente intermedi e chiamiamo codici maggiori, minori ed intermedi di selezione i parametri dell'ordinamento delle concordanze. Per es. nel caso c) il lemma è l'esponente maggiore e la forma il minore; l'ordine alfabetico del lemma è il codice di selezione maggiore, l'ordine alfabetico della forma il codice intermedio, l'ordine progressivo dei riferimenti il codice minore.

Chiamiamo concordare l'operazione con cui si associa ad ogni occorrenza un contesto.

II IL DIZIONARIO DI MACCHINA C.N.U.C.E.

1. Prima di passare ad esporvi brevemente un esempio della cosiddetta linguistica statistica o quantitativa, vorrei aggiungere alcune considerazioni a proposito del dizionario di macchina (DM), strumento utilissimo a molti fini.

Per esempio, proprio per facilitare lo sviluppo di quelle ricerche statistiche delle quali dovrò dare un esempio.

Per limitarmi alla statistica lessicale, anch'essa, come la lessicografia, esige uno spoglio esaustivo di testi o parti di testo, e il DM è utile in questa fase allo stesso modo che per gli spogli lessicografici. In un secondo tempo il calcolatore servirà ad operare i calcoli sui dati numerici risultanti,tramite lo spoglio, dal conteggio. Esaminiamo pertanto,brevemente, una procedura automatica di spoglio, ai fini statistici, e le caratteristiche richieste a un eventuale DM inserito nel ciclo delle operazioni.

2.1. Lo spoglio lessicografico : procedimento e problemi(Nota 12).

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

Solitamente vengono riportati, oltre le parole e la punteggiatura, gli eventuali segni propri dell'edizione critica, informazioni sulla struttura grafica (divisioni in pagine e in righe; maiuscolo, minuscolo, corsivo, ecc.), la suddivisione del testo nelle parti (capitoli, strofe, paragrafi,ecc.) utilizzabili come riferimento; spesso vengono aggiunte informazioni non presenti a livello grafemico nel testo stampato ma inserite con una fase di preedizione: ad esempio frasi che sono citazioni,o brani interpolati da omettere /41/.

Di norma il numero di caratteri diversi da rappresentare oltrepassa il centinaio, mentre le macchine perforatrici in uso oggi offrono solo 48 o 64 segni diversi; si deve pertanto mettere in atto una particolare tecnica di codificazione. Questo problema è meno drammatico nel caso si perfori non una intera riga, ma una sola parola per scheda(come fa ad esempio LASLA di Liegi/73/). Si parlerà in questo caso di schede parola e non di schede testo. In questa prospettiva, i diversi tipi di informazione (parola, punteggiatura, diacritici) possono essere perforati ciascuno in zone fisse della scheda, in ognuna delle quali sono dunque disponibili tutti i caratteri della tastiera.

Le soluzioni più comuni consistono nel rappresentare un carattere del testo con una sequenza di due o più perforazioni,oppure nell'adottare l'equivalente funzionale del tasto delle maiuscole nella macchina da scrivere.Tra i 64 codici disponibili se ne scelgono alcuni con funzione di "chiave". Ciascuno dei codici restanti assume, quando viene letto dal calcolatore, un significato diverso a seconda dell'ultima chiave precedente. E' chiaro che se le chiavi sono n, i caratteri rappresentabili complessivamente sono n.(64 - n).

L'operazione di ricopiatura del testo su schede, con tutti i controlli inevitabili, è costosa e lunga a confronto con le operazioni successive del calcolatore. Di qui l'esigenza (Nota 13) che i diversi centri operanti in questo settore adottino un metodo comune nel perforare i testi, in modo che l'insieme di perforazioni praticate secondo un sistema standard garantisca l'utilizzazione del testo perforato per altre ricerche, anche se non previste nel programma iniziale /342/.

In questo senso già si opera al Centro Studi IBM di Pisa, dove il sistema di perforazione messo a punto per l'Accademia della Crusca viene normalmente impiegato per gli altri progetti linguistici in corso.

Il problema della sperequazione tra segni disponibili e segni da rappresentare si pone anche per la stampa dei risultati delle elaborazioni.

Al Centro Nazionale Universitario di Calcolo Elettronico di Pisa lo si sta risolvendo con il dotare la macchina stampatrice veloce IBM di catene di stampa, intercambiabili facilmente, fornite di un massimo di 120 caratteri componibili tra loro per mezzo di sovra e sotto impressioni successive: e alcune catene sono costruite per lingue scritte con alfabeto non latino. Nel frattempo, si codifica anche in sede di stampa, con delle "chiavi" sottoscritte ai caratteri usati non univocamente.

Il calcolatore "legge" le schede o il nastro perforato e registra il testo su nastro magnetico, scomponendolo nelle diverse unità elementari di elaborazione (nel nostro caso le parole; in altri lavori i grafemi, i fonemi, le sillabe, i periodi, i sintagmi, ecc.) e ripete per ciascuna unità informazioni che la riguardano, anche se nel testo stampato e perforato occorrono una sola volta, quali numero di pagina e di riga, riferimento, caratteri dell'apparato critico, ecc. A questo punto può avere inizio il ciclo delle elaborazioni vere e proprie la cui velocità è molto superiore alle operazioni precedenti: per perforare un testo di mezzo milione di parole (circa 60.000 righe) occorrono 350 ore-uomo circa; per conteggiare le frequenze con il calcolatore bastano due ore.

I problemi posti dalle elaborazioni sono in primo luogo problemi di tecnica di programmazione (indici alfabetici diretti e inversi, rimari, frequenze varie, ecc.), che tuttavia non possono prescindere da considerazioni di carattere più strettamente linguistico. Questo si verifica già nella costruzione delle concordanze /57/. I criteri di delimitazione del contesto vanno ormai uniformandosi presso i vari Centri, e praticamente possono essere raggruppati nei tipi seguenti :

a. - La parola esponente è sempre al centro del suo contesto; ossia, è sempre preceduta e seguita da un ugual numero di battute o di parole; le concordanze di Tito Livio /242/ recentemente pubblicate presso il M.I.T. sono un esempio notevole.

b. - Il contesto è scelto in base alla natura della parola: per le parole grammaticali è spesso un 'trinomio', per le preposizioni sono prese per lo più le due parole seguenti, ecc. /44/. In questo caso sarebbero interessanti le possibilità di automatizzare la analisi sintattica; si potrebbe scegliere per ogni costituente la parte della struttura da riportare come contesto.

c. - Il contesto è costituito da una intera unità di riferimento: il verso, il paragrafo, il versetto, il comma, ecc.; il metodo è seguito frequentemente per opere in versi; esempio notevole le concordanze della Divina Commedia pubblicate dal Prof. Tagliavini per l'inaugurazione del C.N.U.C.E. nel 1965 /305/.

d. - I limiti di contesto sono segnati in fase di preedizione: in altre parole il testo viene suddiviso in pericopi, ciascuna delle quali funge da contesto per tutte le parole che la compongono; P.Ficher sta compilando così le concordanze della Vulgata /246/.

e. - Il contesto è regolato tenendo conto di determinati segni quali l'interpunzione, il cambio di riferimento, ecc. Ottimo esempio le concordanze della Novella del Grasso Legnaiuolo, pubblicate dall'Accademia della Crusca /2/.

L'insieme di operazioni comunemente raggruppate sotto il termine "Lemmatizzazione" richiede una serie di interventi umani che spezzano il ritmo delle elaborazioni interamente automatiche con le quali dal testo perforato si producono i prospetti statistici, le concordanze, le schede contesto e in generale la stampa dei risultati finali.

In una procedura di lemmatizzazione interamente artigianale il lemma va scritto dapprima accanto a ciascuna delle parola del testo disposte alfabeticamente, e successivamente va perforato su schede, unitamente a un numero di codice per mezzo del quale il calcolatore associa il lemma a ciascuna delle occorrenze nel nastro-testo.

La sproporzione esistente tra velocità dell'elaboratore e tempi di intervento umano rende particolarmente delicato il compito di coordinare le fasi del lavoro, per sfruttare al massimo le disponibilità della macchina.

Tuttociò spiega come gli sforzi dei ricercatori tendano a ridurre sempre più, per mezzo del cosiddetto "dizionario di macchina" la necessità di interventi umani. A raggiungere questo obiettivo sono interessati non solo i lessicografici ma anche quanti, per studi di stilistica e di linguistica quantitativa, devono operare su lemmi e non sulle forme intese a livello grafemico, e quanti operano nei diversi settori della linguistica applicata computazionale (traduzione meccanica, linguistica automatica).

In termini generali, la struttura di un dizionario di macchina e il meccanismo della consultazione non sono dissimili da quelli di un normale dizionario stampato. Il dizionario di macchina è costituito da una serie di 'articoli' (Nota 14) registrati nella memoriale centrale o nelle memorie ausiliarie (nastro e dischi) del calcolatore, ciascuno dei quali si compone di due parti: un termine di ingresso (indice), che serve per la ricerca, e una serie di informazioni (funzione) su questo termine. Dall'altro lato si ha una parola, estratta da un contesto, che deve essere 'ricercata' nel vocabolario; le informazioni in esso reperite devono essere associate alla parola quando viene reinserita nel suo contesto.

Che cosa siano le funzioni risulterà chiaro dai seguenti esempi. Nella traduzione automatica, le informazioni consistono principalmente in :

- un codice sintattico-semantico che specifica le proprietà distribuzionali e semantiche da utilizzare per collegare secondo schemi strutturali definiti la parola in questione con quelle che la attorniamo; (parsing)
- un insieme di istruzioni (o indirizzi di istruzioni) per la soluzione (tentativa) dell'eventuali omografie;
- la parola (o il gruppo di parole) che traduce nella lingua di uscita la parola in questione.

Nelle applicazioni lessicografiche, la funzione è costituita principalmente dal lemma (inteso come forma di base cui la parola intestigata va assegnata nei conteggi e sopratutto negli ordinamenti alfabetici delle operazioni finali) e da una serie di codici che classificano morfologicamente la parola, oppure la qualificano per un particolare trattamento nelle elaborazioni succes

sive.

2.2. Le tecniche di consultazione del vocabolario di macchina sono riconducibili a due tipi fondamentali :

- il primo, usato sopratutto in sede di traduzione e documentazione automatica, ma anche, recentemente, in lessicografia (ad esempio a Liegi) /73/ prevede che le parole da lemmatizzare vengano ricercate nel vocabolario di macchina rimanendo, sia in input sia in output, in ordine di testo;
- il secondo esige invece che le parole del testo siano previamente ordinate alfabeticamente, così come sono ordinate alfabeticamente le forme elencate nel dizionario di macchina.

E' evidente che, mentre nel secondo caso la ricerca si riduce a un confronto tra due serie di parole disposte nello stesso ordine, nel primo caso si deve ricorrere a una tecnica più complessa non esistendo alcun rapporto tra l'ordine delle parole nel testo e la successione degli articoli nel dizionario. Per facilitare e velocizzare la ricerca, si mira a comprimere le dimensioni del dizionario, in modo che possa essere contenuto per intero nella memoria centrale, o registrato su memorie periferiche ad eccesso rapido casuale (ramac, dispac).

Il metodo più generalmente seguito è quello di registrare, non tutte le forme possibili a livello grafemico, ma solo le parole 'invariabili' e i 'temi' (Nota 15) di quelle 'flessibili'. Ogni 'tema' è accompagnato da un codice di 'paradigma', e per ogni tipo di 'paradigma' possibile sono registrate tutte le desinenze in forma di tabella.

In programmi molto avanzati, la distinzione non è solo tra 'tema' e 'desinenza' ma addirittura tra 'forma base' e 'affissi' (prefissi e suffissi) cosicchè la riduzione del numero delle voci di vocabolario è ancora maggiore. Alcuni dizionari impiegati nella traduzione automatica coprono con circa 20.000 temi alcune centinaia di migliaia di forme grafiche.

Ogni parola del testo da lemmatizzare viene innanzitutto "segmentata" dal calcolatore il quale procede tentativamente per scomposizioni successive, fino a che perviene a isolare un tema e una desinenza che oltre a essere previsti dal dizionario, siano compatibili fra loro. Se il calcolatore esaurisce tutte le possibili scomposizioni senza trovarne una corretta, la parola viene segnalata e l'operatore umano interviene, o immediatamente a console, o in un secondo tempo per mezzo di schede perforate.

Nella prospettiva invece di una ricerca operata con le parole in ordine alfabetico non è necessario che il dizionario sia accessibile contemporaneamente in tutte le sue parti; può essere registrato su nastro magnetico e letto in memoria centrale un poco per volta, a mano a mano che avanza il nastro delle parole. Il risparmio di spazio non costituisce più un problema, e le diverse forme flessive possono essere registrate integralmente, tutte.

Il programma prevede, in linea di massima, 2 nastri in input e 3 nastri in output. Attualmente per alcuni progetti usiamo al C.N.U.C.E. questo schema : (v. fig. 5).

1. il nastro-vocabolario, contenente
   a) le forme certamente univoche, con relativo lemma,
   b) le forme omografe possibili senza lemma;
2. il nastro delle parole in ordine alfabetico, ciascuna accompagnata da riferimento e da relativo contesto.

Ogni parola viene confrontata con il dizionario e, a seconda del risultato del confronto, viene smistata su uno dei 3 nastri di output che sono :

1. - Nastro delle "parole nuove", cioè non ancora previste nel vocabolario. La loro lemmatizzazione viene sospesa. Il calcolatore stampa tutte le occorrenze delle forme nuove, con il relativo contesto, e predispone contemporaneamente le schede cui verrà apposto artigianalmente il lemma. Queste schede in una fase successiva, permettono di associare il lemma alle parole momentaneamente accantonate, e nello stesso tempo aggiornano il dizionario, in modo che se nella prossima elaborazione si riscontreranno ancora queste forme, esse potranno essere lemmatizzate automaticamente.

2. - Nastro delle parole trovate univoche.
La parola è prevista nel vocabolario come univoca; le viene associato immediatamente l'unico lemma possibile, e la fase di lemmatizzazione per essa è chiusa.

3. - Nastro delle parole trovate omografe.
La parola è prevista, ma il calcolatore non è istruito a scegliere tra i due o più lemmi diversi cui può essere assegnata. La sua lemmatizzazione viene rinviata a una fase successiva, in cui si stampano i contesti in base ai quali il lemmatizzatore decide a quale lemma attribuire ogni singola occorrenza.

E' possibile inserire nel vocabolario per ogni forma omografa tutti i lemmi prevedibili. Il calcolatore li stampa, allora, ciascuno contraddistinto da un numero progressivo, in capo ai contesti da esaminare; il lemmatizzatore si limita a scrivere a lato di ciascun testo il numero codice di ciascun lemma.

2.3. Il Dizionario di macchina della sezione linguistica del C.N.U.C.E.

Si può facilmente dedurre da quanto è stato detto che il vero rischio del metodo sta nella possibilità di inserire nel vocabolario come univoca una forma che invece può risultare di fatto omografa. Il calcolatore in tal caso lemmatizzerebbe senz'altro le parole corrispondenti come univoche, senza proporle all'esame del linguista.

Questo rischio, come già è stato sopra accennato, è tanto maggiore quanto più si estende diacronicamente lo strato di lingua esaminato e quanto più si torna indietro nel tempo verso sistemi non perfettamente presenti alla coscienza linguistica di chi compono il dizionario di macchina.

Si pensa però di superare questo ostacolo, almeno fino a che l'esperienza d'uso del dizionario non ne avrà comprovata la completezza, col considerare provvisoria la lemmatizzazione operata per suo tramite. Basterà stampare anche il nastro delle forme univoche dotandole dei relativi contesti per consentire il controllo dei lemmi assegnati automaticamente. Si può facilmente calcolare che, pur con questa fase di controllo, il risparmio di tempo rispetto a una procedura interamente artigianale sarebbe notevolissimo (del 75% circa).

A favore dell'impiego del dizionario di macchina stanno anche le considerazioni che esso, evitando gran numero delle scrizioni di lemmi necessarie nella lemmatizzazione artigianale, diminuisce in maniera notevole il rischio di errori casuali di trascrizione, e nel contempo garantisce una maggiore sistematicità e coerenza nei casi, più numerosi di quanto si possa pensare, in cui la formulazione del lemma richiederebbe una discussione o il ricorso a un corpo di regole prefissate; il dizionario funziona come registrazione delle decisioni prese e, potendo essere stampato facilmente, fornisce in ogni momento il quadro completo delle forme esaminate e del trattamento scelto.

Naturalmente l'economia di tempo realizzata con l'uso di un dizionario di macchina è in relazione al numero delle forme univoche incluse. La percentuale di forme omografe possibili dipende dal grado di "profondità" al quale si vuole spingere l'analisi e la distinzione delle omografie e della polisemia. Del resto, non sembra che le categorie tradizionali, allo stato attuale delle conoscenze e sopratutto quando si applichino a grandi quantità di esempi, permettano una risposta coerente e integrale al problema della individuazione della unità lessicale. Pertanto entrano in gioco considerazione di ordine economico e organizzativo: si tratta di trovare il giusto punto di equilibrio tra le analisi che si devono eseguire nella prima fase di raccolta e di organizzazione automatica della documentazione e le analisi demandate agli utenti successivi.

Le tecniche per la generazione di un vocabolario di forme attualmente in uso nei Centri lessocografici di cui abbiamo conoscenza si possono ricondurre a due tipi principali.

Alcuni procedono, per così dire, induttivamente: iniziano senza un vocabolario di macchina l'elaborazione della prima opera, e ne lemmatizzano artigianalmente tutte le forme, utilizzate per costituire un primo vocabolario che viene via via arricchito con le 'forme nuove' delle opere successive.

Altri iniziano invece con il vocabolario di macchina già approntato, ottenuto operando tutte le possibili flessioni di un lemmario assunto come base di partenza.

Il lemmario può essere desunto da un solo dizionario, o dalla fusione di più dizionari o con procedimenti ancora più complessi, che garantiscano maggiore completezza (un esempio è il progetto di H.H.Josselson che prepara un DM russo /166/). In questa fase preparatoria il calcolatore può essere di aiuto non solo con il confronto e la fusione di dizionari diversi, ma anche per flettere i lemmi automaticamente anzichè a mano. Questa operazione che è, per così dire, l'inverso di quella di segmentazione esaminata precedentemente, richiede che al calcolatore vengano comunicate le tabelle con le desinenze dei vari paradigmi, e che i lemmi da flettere siano prima ridotti ciascuno a un 'tema' (nel senso già definito o, nel caso di lemmi irregolari, a due o più temi) accompagnato da un codice di rinvio a una determinata tabella. Il calcolatore genera tutte le questioni possibili in alcuni casi, semplicemente giustapponendo al tema, una alla volta, le desinenze previste della tabella; in altri casi modificando tema e desinenza prima di unirli.

Con tale procedimento si otterrebbe anche un notevole aiuto per la ricerca della omografia possibile. Sarebbe sufficiente infatti disporre col calcolatore le forme in stretto ordine alfabetico e far segnalare le forme ripetute almeno una volta.

Il DM al quale lavora la sezione linguistica del C.N.U.C.E. conterrà in due sezioni distinte sia i lemmi sia le loro flessioni.

D'altronde l'esplicitazione di tutte le forme potenziali è di per sè utile. Essa permette per es. l'elencazione e la statistica completa delle omografie possibile nel sistema.

Il nostro algoritmo di flessione funziona anche a livello fonematico, per consentire la trascrizione fonematica di testi italiani dall'alfabeto tradizionale. Inoltre lo spoglio e la statistica delle coppie minime è indispensabile allo studio del rendimento funzionale delle opposizioni fonematiche /198/ /268/. Sarà

così possibile eseguire a livello fonematico, quelle statistiche sulle frequenze del sistema che, da N.Troubetzkoy /312 p.286/, ad A.Martinet /206 p.187 e seg/ e C.Tagliavini /304 p.117/ sono considerate necessario completamento alle statistiche sulle frequenze nel discorso, e delle quali abbiamo eccellenti esempi da parte della scuola rumena di linguistica quantitativa /282/.Tra gli altri settori di ricerca che richiedono un algoritmo di trascrizione fonematica, innanzitutto cito i dizionari inversi e i rimari: si veda per ēs. il dizionario inverso dello Juilland /170/ per il francese. Non mancano gli studi per la scansione automatica, per es. per quella dell'esametro latino. /240/.

Per quanto riguarda l'italiano, ricordo il progetto di analisi metrica dei canti popolari delle raccolte citate più sopra. E' nostra intenzione enucleare il tessuto sottile e vario di corrispondenze foniche che intuiamo accettate in posizione di rima dalla coscienza popolare, accanto alle forme canoniche della rima, della assonanza, della consonanza.

Servirà perciò la trascrizione fonematica almeno delle parole e dei sintagmi in posizione di rima.

La statistica lessicografica, anche al di fuori del livello fonematico, ha e ha avuto numerosi cultori.

Come esempio, pensiamo agli studi sulle proporzioni relative delle " diverse classi sintagmatiche, sulle strutture statistiche dei derivati e dei composti, sulla estensione relativa delle basi, sul rendimento delle formanti lessicali", illustrate da J.Dubois /84/.

Si contano ricerche sulla proporzione dei termine suffissati e sulla disponibilità relativa di ogni formante suffissale sia nell'intero lessico, sia nell'interno di un micro-sistema definito, come è il caso per il lavoro di C. Tagliavini e U.Bortolini in relazione al "Dizionario dei toponimi e degli etnici"/304 p.147/.

Altri settori per lo studio statistico del lessico, di cui terremo conto nella compilazione del nostro D.M. sono quelli delle annotazioni etimologiche e stilistiche /305/ /297/ /154/ /165/ /244 pp.107-109/. La codificazione di categorie di stile contribuirà, tra l'altro, a selezionare nel lessico totale del D.M.,dei D.M. più o meno ridotti in funzione della natura del testo da lemmatizzare.

2.3.1. Nella introduzione alla pubblicazione di qualsiasi spoglio lessicografico, troviamo dibattuti numerosi problemi concernenti la definizione di lemma e di parola o occorrenza. I punti controversi per una stessa lingua sono quasi sempre gli stessi, e spesso sono analoghi in lingue affini: ad es. si confronti per l'italiano /333 pp.135-155/, per il francese /225/ per lo spagnolo /169 XXVII-XXXV/, per il rumeno /170/ e per il latino /43/. Nella notazione più comune si simbolizza con N il numero delle occorrenze, o lunghezza di un testo, e con V il numero dei suoi lemmi, con V_1 V_2 ..... V_n il numero dei lemmi che appaiono V volte nel testo con frequenza 1, 2,..... n /228 p.142/ /129 p.88/.

Abbiamo definito fin qui l'occorrenza come unità grafica : è evidente che qualsiasi modifica a questa definizione si riflette sul valore di N, così come una diversa definizione del lemma influisce sia su V sia su Vn. Di qui le preoccupazioni /222 p.126/ degli studiosi di statistica lessicale: "L'un des prècepts élémentaires de toute statistique est que l'on n'opère valablement que sur des populations, des individus et des caractères bien definis.

- ha detto Ch.Muller - Aussi peut- il sembler hasardeux de faire des statistiques sur les'mots', puisque les linguistes sont d'accord pour constater qu'aucune definition satisfaisante de cet élement du language ne peut être proposée" /225/.

La crisi del concetto di parola /121 p.27/ non ha però ancora aperto la strada alla applicazione in statistica del concetto di monema anche se ne è tentata una definizione algoritmica./1/.

2.3.2. Il calcolo di N presuppone la segmentazione del testo in unità di testo (mot unitè de texte, di Muller), che in buona parte dei casi corrispondono alle occorrenze grafiche, ma spesso se ne differenziano o perchè una unità grafica contiene più unità di testo, o, viceversa, perchè una unità di testo è composta di più unità grafiche.Non è difficile trovare degli esempi in italiano. Nota 16.

2.3.2.1 Per il primo caso pensiamo alle preposizioni articolate, alcune delle quali hanno una frequenza elevata. Di fatto le soluzioni alternative proposte sono di assegnare la preposizione articolata : (a) solo al lemma della preposizione semplice; (b) al lemma dell'articolo e alla preposizione semplice,(c) solo all'articolo. Quando la si assegna anche all'articolo, si deve scegliere tra : (a) trattare la preposizione articolata come una flessione autonoma dell'articolo,(b) cumulare le frequenze delle diverse preposizioni articolate in un unica forma dell'articolo, definita come "forma dell'articolo unita a preposizione";(c) aggiungere la frequenza della preposizione articolata alle corrispondenti flessioni dell'articolo (per es. aggiungere a la la frequenza di alla, a le quella di alle, ecc. Per inciso osservo che questa soluzione obbliga a smistare le omografie di forme del tipo all', che può corrispondere ad allo, alla, alle, e che sono di solito di elevata frequenza). Del resto analoghe alternative sono possibili sotto i lemmi della preposizione.

Le particelle encitiche (-mi,-ti, -ci, ecc.) presentano una situazione analoga. Spesso si trascura l'enclitica e una forma come parlarmi si assegna solo a parlare.Ma altre volte si assegna un lemma -mi enclitico, separato dal corrispondente lemma proclitico mi (come in mi parla). Molti si chiedono poi se, per esempio le occorrenze di farmi, farti, fargli, farvi, farlo, ecc.,debbono essere, sotto l'esponente fare, tenute distinte tra loro, oppure riunite in una unica forma far- , oppure semplicemente aggiunte alla frequenza di far o di fare.

I casi di 'trait d'union' in italiano non sono frequenti come in francese, e di solito non sollevano dubbi (es. il diretto Pisa-Roma, la partita Genova-Livorno). Semmai può dare adito a dubbi la categoria dei così detti prefissoidi,/217/ che possiedono la possibilità pressochè illimitata di formare dei nuovi composti /218 p.337/ (auto, foto, elettro, galvano, anti, para, mini, ecc.) che talvolta si trovano scritti con trattino, talora (e il caso è imbarazzante per il lemmatizzatore) addirittura separati. Nè suscita dubbi l'apostrofo in italiano contemporaneo, a differenza per es. del francese aujourd'hui (una parola):l'agneau (2 parole). /225/. Questa situazione si presenta invece nell'italiano dei primi secoli (es. in tra'mi per traimi).

2.3.2.2. Diamo ora alcuni esempi del caso inverso: più unità grafiche = 1 unità di testo. Questi casi sono complessi, e sono fonte di dubbi nel lemmatizzatore prima per la loro individuazione che per la decisione del loro trattamento. Mi riferisco in particolare alle forme composte dei verbi frequentissime; sopratutto in quelle composte con l'ausiliare essere + participio passato,è

non di rado difficile decidere se si ha a che fare con forme verbali composte oppure con participi autonomi.

Se lo scopo dello spoglio è limitato alle concordanze, le risorse tipografiche, sopratutto se si usa la fotocomposizione (nel qual caso la parola esponente è di solito rilevata in corsivo all'interno dell'esempio), si possono adottare delle soluzioni intermedie di disimpegno; per esempio mettendo in corsivo, nell'esempio citato sotto il participio, oltre il participio stesso, anche l'ausiliare o sospettato tale. Ma quando si propongono conteggi statistici, una decisione si impone tra alternative di questo tipo : (a) o non si computa l'ausiliare, che si unisce al participio, in una forma verbale distinta dalle altre del lemma (così si diminuisce di una unità N ma si aumenta di 1 il numero delle forme del lemma); (b) o si attribuisce l'ausiliare a un suo proprio lemma.

Per le locuzioni, la situazione è, direi, disperata, siano esse composte da parole lessicali (segnare il passo, bagnato fradicio, avere fame)o da parole grammaticali (al di là, al di sopra, se non altro). La loro definizione è difficile e, per rendersene conto, basta confrontare i silenzi e le discordanze dei cosidetti buoni dizionari. Alcuni progetti di lessicografia meccanizzata ricercano delle conoscenze più precise su questi gruppi di parole che, in un certo senso,"préexistent au discours, sont assemblés en langue" /225/, studiano con metodi vari le frequenze di accoppiamento delle parole nei testi (per es. Sintagmi dell'Index Thomisticus /165/ e groups binaires del Trésor de langue française di Nancy). Da questo punto di vista è utile non trascurare il problema della locuzione negli spogli, anche se, data l'estrema incertezza del problema, quasi tutti consigliamo di non raggruppare in alcun modo le locuzioni, se non forse mettendole in rilievo all'interno dell'esempio. (Nota 17).

2.3.3. Anche la riconduzione al lemma delle unità del testo è cosparsa di punti controversi: basteranno anche qui alcuni esempi.

In italiano contemporaneo sono piuttosto rari i casi di grafie diverse di uno stesso lemma. (ricuperare, recuperare; dinanzi, dinnanzi; interpretare, interpetrare, ecc.). V'è incertezza tra grafia unita e grafia divisa in alcune locuzioni (per lo più e perlopiù, in fine e infine). Sono da ricordare anche i dittonghi mobili (sonare e suonare), le forme sincopate cfr.2 a pag.20 (torre e togliere) aferetiche ('sto per questo) e con i aferetico (scherzo, ischerzo). Non è sempre facile dire quale delle due grafie è la più comune, e solo delle assicurano la coerenza nelle scelte: poichè, quasi sempre, si adotta un solo lemma, l'incertezza sta nella scelta dell'esponente più che nel trattamento per la statistica.

Molto più rilevanti le varianti grafiche quando lo spoglio si estende diacronicamente, come è il caso per l'Archivio della Lingua negli spogli dell'Accademia della Crusca per il grande Dizionario Storico Italiano. I ricercatori della Crusca adottano una procedura di normalizzazione della varianti. Essi registrano la forma sotto l'esponente oggi più comune, ma creano un sottolemma che evidenzia, all'interno del lemma, le diverse variazioni della grafia. Spesso il lemmatizzatore si impegna in riflessioni e consultazioni di natura etimologica per distinguere le semplici varianti da casi di lemmi distinti e per formulare coerentemente gli esponenti. E' probabile che una forma si ripresenti in altri testi. Il D.M. fornisce allora prontamente la soluzione de-

cisa la prima volta.

I sostantivi, che di solito si registrano con la forma del singolare, non presentano molti problemi, se si eccettua il caso dei cosidetti 'nomi mobili' del tipo 'animato', che hanno entrambi i generi, maschili e femminili, ai quali corrispondono, ma non sempre, forme diverse, e degli 'alterati', che talora i dizionari registrano come lemmi indipendenti: panca e panchina, tavolo e tavolino, ma non libro e libretto.

L'esponente degli aggettivi è di solito la forma del singolare maschile. Certi comparativi o superlativi di formazione 'irregolare' sono eccezionalmente costituiti in lemmi autonomi (ottimo, massimo) e così pure dei superlativi regolari di accezione caratteristica (chiarissimo, altissimo). Come dice Muller, in mancanza di una regola generale, il legame semantico dovrebbe essere valutato caso per caso nella coscienza dei parlanti, che del resto può essere sondata obiettivamente solo con inchieste sul tipo del 'differenziale semantico' degli psicologi. Questa considerazione va a maggior ragione riferita ai lemmi che, pur in una unica funzione sintattica, esprimono due o più sensi diversi; quando cioè si deve decidere se si è di fronte a un caso di polisemia (un solo lemma) o a un caso di omografia (più lemmi). Tutti gli studiosi di cui ho notizia concordano nell'affermare che la soluzione praticamente realizzabile consiste nel seguire coerentemente un dizionario, scelto come arbitro, in quanto rappresentativo della competenza linguistica della comunità dei parlanti.

I verbi hanno generalmente per esponente l'infinito; ma di alcuni verbi difettivi l'infinito è raro oppure non documentato. Di più alcuni verbi si presentano nel contesto sia con coniugazione pronominale (riflessiva o intransitiva : bagnarsi, dolersi, alzarsi) si in forma normale (bagnare, alzare, dolere).

Le alternative sono ancora più frequenti per le parole grammaticali.

Alcuni pronomi hanno una flessione regolare (questo, codesto, quello, tutto), ma i pronomi personali (io, tu, lui, ci, ecc.) possono essere messi ciascuno in esponente come lemmi autonomi, oppure essere raggruppati diversamente in serie diverse del tipo, per es. io, tu, egli, noi, ecc.; mi, ti, si, ci, ecc.; io, me, mi, ecc. ; oppure essere raggruppati semplicemente sotto un unico esponente: il pronome personale.

Le diverse specie tradizionalmente riconosciute dei pronomi possono generare altrettante omografie (es. che relativo, interrogativo, ecc.)

Sopratutto è ricca di casi la categoria delle parole che possono assumere valori sintattici diversi, cambiando di parte del discorso. Addirittura alcuni, per es. P.Busa /42/ hanno rinunciato a elaborare questo tipo di omografia, che avrebbe aumentato enormemente la percentuale di omografi da smistare. In italiano, si pensi per esempio a infiniti sostantivati, participi aggettivati e sostantivati, aggettivi sostantivati e sostantivi aggettivati, nomi propri formati da nomi comuni (Palazzo Reale, Casa Bianca), pronomi e aggettivi pronominali, aggettivi e avverbi (tanto, poco, molto), preposizioni e avverbi (su, sotto) e a molti altri scambi tra categorie grammaticali, per es. (ho visto dei fiori, uno dei tanti). Alcuni lemmatizzatori considerano omografi solo i casi in cui differenze di funzione si sono consolidate in differenze semantiche rilevanti, tanto da essere codificate nei dizionari come lemmi distinti (per es. amante, ma non bollente; fatto

e detto ma non stimato, illustrato).

Spesso il contesto non chiarisce il dubbio, ma chi mira a una statistica di parti del discorso deve pure distinguere.

2.3.4. Questo stato di cose mi offre l'occasione di rilevare che, sopratutto per queste categorie, una opportuna codificazione dei lemmi può rendere il D.M. operativamente valido per le diverse possibili prospettive di lemmatizzazione.

Supponiamo di registrare queste omografie con un codice che le distingua in categorie. Questo codice può essere usato per annullare, all'atto pratico della consultazione del D.M., la possibilità di omografia per una o più categorie.

In italiano, per es., l'omografia tra flessioni diverse di uno stesso lemma è assai frequente: 2a persona plurale dell'indicativo presente e participio passato femminile plurale (pensate, stordite); 3a persona singolare e 2a plurale dell'indicativo presente e forme rispettive dell'imperativo, ecc. . C. Tagliavini /302/ ha distinto queste omografie nelle concordanze della Divina Commedia e così anch'io in /333/; invece, per es. l'Accademia della Crusca le trascura. Queste considerazioni confermano l'importanza operativa di una classificazione dei tipi di omografia in una struttura organizzata più o meno gerarchicamente, cosicchè sia possibile predeterminare un livello di non accettazione, come abbiamo proposto io /333/ e P. Busa /42/.

La presenza di un D.M. assicura la normalizzazione nella lemmatizzazione a due livelli:(a)assicura la coerenza dei lemmatizzatori con un dizionario scelto come arbitro in casi particolari, per es. di polisemia, dove non esistono regole comunemente accettate;(b)d'altra parte garantisce un comportamento uniforme in tutti i casi che costituiscono una classe che presenta le stesse alternative di lemmatizzazione.

Gli esempi esaminati mostrano, sia pure sommariamente, che gran parte delle incertezza nella lemmatizzazione sono riconducibili a pochi punti controversi. La decisione presuppone delle convenzioni, poichè non abbiamo una definizione operativamente efficente di "lemma".

Non è probabile che i ricercatori accettino di impiegare le stesse convenzioni per degli spogli che hanno prospettive diverse,sopratutto se non sono direttamente interessati ai compiti della statistica lessicale.Ciò non impedisce, se grazie al D.M. le condizioni sono rese esplicite e se le scelte sono seguite coerentemente per tutti i lemmi di una stessa classe, che i risultati degli spogli lessicografici siano trasformabili o riducibili a standars comparabili statisticamente con un numero limitato di trasformazioni precise, come si augura anche Ch. Muller.

2.3.5. Il D.M. può aiutare il lessicografo anche nella formulazione dell'esponente, che spesso è costituito, oltre che dalla forma base, da un codice di parte del discorso, e da indicazioni complementari (sinonimi, specificazioni grammaticali) per distinguere esponenti omografi attuali nel testo, o possibili nel sistema linguistico.

Non è facile ricordare tutti i lemmi che richiedono queste specificazioni, sopratutto relativamente al sistema, e ancora una volta il D.M. assicura la coerenza del comportamento.

Il fatto che esponenti omografi possibili in una certa 'tranche' sincronica, non lo siano in un'altra (per es. caro, oggi non più usato come sostantivo nel senso di carestia, cosicchè le annotazioni necessarie variano per testi appartenenti a strati di lingua diacronicamente distinti, non è una difficoltà decisiva.Basta

codificare queste differenze diacroniche nel D.M. e renderle operative.

2.3.6. Il problema se e come sia possibile operare automaticamente lo scioglimento dell'omografia si è posto, sopratutto sotto l'influsso della LA, con soluzioni diverse, in genere fondate sul riconoscimento della struttura sintattica della frase, operato secondo le tecniche di parsing, fondate a loro volta sui diversi tipi di grammatica (context-free; context-sensitiv, ecc.)

Non voglio comparare qui le diverse tecniche impiegate, e le relative grammatiche. Le pubblicazioni in questo settore sono ormai numerosissime, anche se spesso difficilmente accessibili, trattandosi per lo più di "reports" interni dei singoli istituti. A titolo di esempio cito la descrizione delle impostazioni degli algoritmo risolutivi, esposta nelle opere di G. Hays /136/; un breve riepilogo è costituito dall'articolo di S.Kuno sulla Art of Syntactic Analysis /299/ al congresso di Bethesda del 1966. Gli atti della 2a Conferenza Internazionale sul trattamento automatico delle lingue di Grenoble, 1967, fanno il punto dei progetti in corso, anche nell'est europeo.

Tra i programmi di parsing, alcuni, sperimentati su un vasto corpus di testi, per ciò che concerne l'inglese, sono stati messi a disposizione del pubblico. Ricordiamo come esempio l'English Parser di Robinson /270/ e il Predictive Analyzer di Kuno e Oettinger /298/. Si servono entrambi di una grammatica inglese compilata nella prospettiva di context-free phrase structure gramlars /234/ /58/. La versione corrente del Predictive Analyzer adoperata a Harvad su un IBM 7094 , si compone di circa 3500 regole e di un dizionario di 25000 forme, ed elabora 70 frasi al minuto. Per la lingua russa, ricordo l'algoritmo inserito nel programma di traduzione automatica iniziato dall'équipe di Georgetown /36/, attualmente a disposizione presso il CETIS dell'EURATON a Ispra /248/.

2.3.7. In questi programmi ci sono molte strutture che ancora non possono essere analizzate, oppure vengono assegnate delle descrizioni strutturali corrette rispetto alla grammatica usata ma non rispondenti alla lingua inglese. Non c'è speranza, nell'immediato futuro, per la realizzazione di un analizzatore automatico completo che produca solamente analisi corretta di ogni frase non ambigua /299 p.19/. Ciò non toglie che gli studi in questa direzione siano utili sia nel piano pratico, per i programmi di MT non completamente automatizzata (traduzione semiautomatica), sia sul piano teorico per lo sviluppo degli studi sintattici, come chiaramente affermano alcuni studiosi di linguistica generale. /195 p.205/ Nota 18.

2.3.8. Per risolvere il caso di parole omografe che appartengono alla stessa parte del discorso, il caso di frasi che ammettono più descrizioni sintattiche, si ricorre anche a una classificazione semantica-tentativa del lessico, nella quale sono specificate le possibilità di relazione e di compatibilità tra le diverse categorie.

Durante un lungo periodo la semantica "a été la parente pauvre des sciences linguistiques". /309 p.5/. A partire dal 1960, ricerche interessanti sono state condotte un pò ovunque, specialmente in Francia, negli U.S.A., in Gran Bretagna e nell'Unione Sovietica. Anche se, come ha detto N.Ruwet, "L'étude du contenu est à l'ordre du jour" /280/, i grandi problemi della semantica sono ben lontani dall'essere risolti. Da molti, il valore delle attuali teorie è riconosciuto nel fatto che si è scelto un buon

punto di partenza /309 p.43/. H.Rubestein, /279/ nel tracciare un breve panorama delle teorie semantiche attuali, dal punto di vista dalla LC, distingue tre filoni principali : il primo (scaling method) ottiene dei giudizi da parte dei parlanti secondo più parametri relativi al significato di una espressione data. Il "Semantic Differential"sviluppato dallo psicologo Charles Osgood /239/ è probabilmente l'esempio meglio conosciuto di questo approccio. Alcuni psicologi hanno usato i calcolatori del C.N.U.C.E. per la valutazione statistica delle risposte. Ma io sono d'accordo con Rubestein sulla estraneità di questo metodo ai nostri fini.

Il secondo metodo (statistical association) o analisi distribuzionale è stato impiegato sopratutto nell'IR /199/ e /185/, ed è stato molto discusso, per es. al seminario estivo della NATO del 1963 a Venezia,già citato. Fondamentalmente il metodo richiede l'analisi statistica delle distribuzioni di due parole nei contesti per un giudizio sul loro grado di similarità semantica. E' evidente la derivazione dalla concezione di Z.S. Harris /132/. "Due unità che hanno dei significati diversi, differiscono anche in qualche parte della loro distribuzione". In un certo senso,è il programma del calcolatore che costituisce un thesaurus a partire da un certo numero di frasi di input. Io credo che questo metodo possa rappresentare nel nostro caso sopratutto un mezzo per il completamento e la verifica di un D.M. già costituito, e tutt'al più potrà costituire una notevole base di documentazione.

Il terzo metodo individuato da Rubenstein è quello dell'analisi componenziale, (componential analysis). Probabilmente più corretta è la classificazione di Todorov nell'articolo citato/309/ di sintesi, che distingue tra analisi semica e analisi combinatoria.Tra i rappresentanti della prima sarebbero rilevanti i francesi B.Pottier e A.J. Greimas e M.Joos negli Stati Uniti, per la seconda J.J. Katz e i suoi collaboratori J.A. Fodor e P.Postal /174/ /175/ e U. Weinrich /325/. Come lo stesso Todorov riconosce /309 p.33/ i due indirizzi hanno in comune la convinzione che :

a) - Il senso di una parola non è una unità indivisibile, ma composta.
b) - Gli stessi atomi di senso (gli stessi semi) si trovano in tutto il vocabolario.

Condizione della applicabilità alla descrizione del lessico, che il numero dei semi sia finito e non troppo elevato, in ogni caso inferiore a quello dei componenti del lessico. /309 p.18/. Inoltre i semi sarebbero organizzati in strutture determinate. La innovazione di Katz e dei suoi collaboratori, che si pongono nel quadro di una grammatica generativa, sta nell'oggetto, che non è più la descrizione statica del vocabolario, ma il funzionamento del linguaggio nel suo aspetto semantico /309 p.33/. Le regole di proiezione rendono conto delle relazioni semantiche tra i morfemi e dell'interazione tra il significato e la struttura sintattica,determinando l'interpretazione semantica corretta di tutte le frasi in numero infinito che la grammatica genera.

Una tale prospettiva sopratutto nella formulazione di Weinrich che insiste particolarmente sulla connessione tra semantica e sintassi /324/ non poteva non attirare l'attenzione di quanti vogliono per mezzo della formalizzazione della semantica risolvere le ambiguità dell'analisi sintattica automatizzata, e una applicazione è stata tentata dallo stesso Rubestein /279 pp.100-101/. Anche Greimas afferma esplicitamente che "l'analisi semica del livello semiologico del linguaggio appare non solo come il primo compito della

semantica strutturale, ma renderà dei grandi servizi ai diversi compiti della linguistica applicata", e in particolare alla MT per la analisi delle frasi. /122 p.65/.

Una della prime conseguenze di queste teorie è la formalizzazione della descrizione semantica delle unità lessicali nel vocabolario: non v'è dubbio che questi concetti costituiscono la premesse insostituibile per la eventuale parte semantica di un dizionario di macchina. Finora, i tentativi di applicazione si sono ristretti a una campionatura su un lessico ridottissimo. In Italia, l'équipe di Parisi e Crisari ha presentato un simile progetto di analisi componenziale al Congresso annuale della Società di Linguistica Italiana a Roma nel 1968. /246/. Eppure queste applicazioni presentano un certo interesse per la teoria stessa. L'analisi semica - dice Todorov - non ha che una quindicina di anni d'esistenza e numerosi aspetti della teoria saranno senza dubbio rivisti e perfezionati. E' auspicabile, nello stesso tempo, intraprendere uno studio simile del lessico intero di una lingua sviluppata (diciamo le 5000 parole più frequenti); una tale impresa sarà superiore alla campionatura attuale. "Ci si potrà rendere conto delle possibilità effettive di economia del linguaggio descrittivo, facendo apparire numerosi problemi che oggi non avvertiamo" /309 pp.17-18/.

Del resto all'VIII° Congresso Internazionale dei Linguisti a Oslo nel 1957, S.Ullmann ha assegnato come scopo alla semantica strutturale la descrizione integrale della struttura del vocabolario.(Rinvio alla 1^ lez.per l'atteggiamento dei'postchmskiani').

Siamo consci non solo delle difficoltà di concreta realizzazione della analisi semica applicata a un vocabolario esteso,ma anche delle difficoltà teoriche ad essa connesse, e che sono state segnalate da molti /70 pp.57-60/. E', come ha detto F.De Tollenaere, "le problème débattu, si le vocabulaire a une structure, et si jamais une description structurale du vocabulaire sera possible". /80 p.29/. E rimane sempre il problema del linguaggio descrittivo: si tratterà di un metalinguaggio puramente artificiale, oppure di un lessico descrittivo scelto nel vocabolario della lingua ? Penso che il nostro atteggiamento, almeno al presente, nel programmare il vocabolario di macchina, sia un pò quello che traspare dalle conclusioni di F. De Tollenaere di fronte alle prospettive della lessicografia ideologica e del Begriffsystem di von Wartburg /330/. "Pourquoi ne pas essayer l'outil mis à notre disposition.....? Bien sûr les difficultés surgiront nombreuses, et beaucoup de problèmes resteront à résoudre, mais on peut procéder a tâtons et sur une échelle réduite"./80 p.29/.

2.3.9. Per lo scioglimento almeno parzialmente automatico delle omografie, in pratica, si fa intervenire il macrocontesto o, in altre parole, il tema generale del testo, facendo ricorso ai microglossari: quando una parola può avere, oltre ed altre accezioni, un senso tecnico specifico, è molto probabile che sia quest'ultimo a ricorrere in un testo specializzato. (Per esempio, fattore in un libro di matematica molto probabilmente non significherà l'amministratore di una fattoria).

Da più parti si avverte però che i metodi più economici e più sicuri sono quelli che cercano di rilevare il valore semantico e morfosintattico di una parola dal suo contesto immediato, con regole stabilite individualmente per ogni forma omografa, come è stato appunto esemplificato nella parte precedente. /316/.

Il problema della soluzione automatica della omografia e della polisemia è solo parzialmente comune alla lessicografia e alla

MT. Per esempio, è inutile distinguere nella lingua di partenza i segmenti la cui polisemia è omogenea nella lingua di arrivo, mentre nella lessicografia, in cui il punto di arrivo è il lemma come unità del sistema, tutte le omografie devono essere risolte. Vi è, per così dire, una precedenza, una dipendenza tra i due settori : la LA attende dalla lessicografia e dalla linguistica non solo l'inventario completo delle omografie, ma anche la descrizione delle differenze che, formalizzate, dovrebbero permettere l'automazione delle operazioni di discriminazione.

In effetti bisogna riconoscere che gli inventari e le descrizioni esistenti sono incompleti e parziali, e che appunto la raccolta e l'analisi di concordanze stabilite su una grande quantità di testi è la prima tappa indispensabile per il completamento delle conoscenze già acquisite. Ciò non toglie che anche nelle elaborazioni lessicografiche si studi la possibilità di automatizzare la distinsione degli omografi. /89/.

3. ESEMPI DI RICERCHE DI STATISTICA LINGUISTICA IN ATTO AL C.N.U.C.E.

3.0. Premessa

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. /9, 179, 310, 315/.

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 /125 pp.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. /375/.

I linguisti cominciarono a prendere coscienza delle possibilità di analisi fondate su un metodo rigoroso /126 p.1/ e particolare attenzione fu rivolta alla statistica linguistica dalla scuola di Praga. Questo interesse si accentuò nel secondo dopo guerra, come risulta tra l'altro dalle comunicazioni al VI° Congresso dei Linguisti del 1948, di linguisti di formazione tradizionale, come il già citato Marcel Cohen /66/ I. Whatmough /326/ e M.Durand /85/. Press'a poco in quegli anni, D.W.Reed /262 pagine 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. Nella metodologia scientifica - egli ricorda - si riconoscono normalmente due ben definiti tipi di procedura: l'analisi qualitativa, un insieme di metodi per mezzo dei quali si osserva la natura e la varietà dei fenomeni; l'analisi quantitativa, un insieme di metodi per mezzo dei quali si osserva la loro grandezza e frequenza.

Evidentemente l'analisi quantitativa presuppone un minimo di analisi qualitative che generano uno schema di classificazione e individuano le unità strutturali di base, delle quali deve essere determinata la grandezza e la frequenza. Come non si può ricercare il peso atomico dell'idrogeno, senza aver prima inferito l'esistenza di unità dette atomi, e di certi elementi uno dei quali idrogeno; così non si può determinare la frequenza del fonema p in una data lingua senza aver inferito l'esistenza di unità chiamate fonemi, una varietà dei quali è detta p. L'analisi linguistica è stata largamente qualitativa - egli dice -; ora le

tecniche descrittive della fonetica, della fonologia, della sintassi, "has been pursued for so long and with such thoroughness that it will now be fruitful for linguists to turn some of their attention to quantitative analysis" (Nota 19) /262. p.235/.

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. /129 p. 16 e p.25/.

Egli ha anche formulato le basi teoriche della possibilità di analisi della "langue" e della "parola" 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 généreaux et prélinguistiques communs à toute une catègorie de signes informationnels". /127 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 significato e su quello del significato), caratterizzati da forma contenuto e frequenza, attuantesi in un processo individuale fisiopsichico sottoposto al determinismo statistico". /141 p.137/.

Un ulteriore sviluppo di questi principi, è costituito dalla equiparazione, proposta da Herdan, tra l'antinomia saussuriana "langue-parola"da una parte e il rapporto "popolazione statistica-campione" nell'altra.

La langue equivale alla "totality of engrams in the brains members of the speech community of occurence"toglether with their probabilities of occurence"e la parole"for random samples for it"/145 p. 34/. Queste premesse, secondo alcuni, costituirebbero le basi teoriche sulle quali la statistica linguistica si propone come disciplina indipendente /141 p.39/ /27 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 sempre più numerosi e attendibili. Già nel 1964 al Convegno di Strasburgo Statistique et analyse linguistique, R. Moreau delineava esplicitamente questa situazione 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. M.me Hirschberg et Ch. Muller ont montré que, sauf cas particulier, cette règle de jeu n'était pas vérifiée". /222 p.130/.

L'atteggiamento odierno è frutto di un processo dialettico.

Benoît Mandelbrot, concludeva un articolo programmatico sulle possibilità di applicare allo studio del linguaggio i concetti della teoria della comunicazione, esprimendo la speranza di veder avvicinarsi il giorno nel quale "lo studio della lingua oltrepasserà definitivamente la barriera fittizia che l'ha sin qui separata dalle scienze esatte" /201/.

E Anthony G.Oettinger all'inizio dell'articolo citato /233/ afferma che la linguistica, almeno in molti aspetti "is evolving from a humanity into a science. There is a growing recognition of languages ad natural phenomena whose secrets may yeld to methods of analysis and description akin to those that have proved fruitful in the physical sciences".

Affermazioni di questo tenore, hanno provocato naturalmente vivaci reazioni da parte dei linguisti tradizionali /196 pp.101 e 102/. F.Papp /245 pp.7-8) ha delineato chiaramente il processo dialettico dell'incontro tra linguisti tradizionalisti e linguistici matematici nell'Unione Sovietica e Istvàn Fodor nell'Ungheria /97 p.6-7/. Si veda sopratutto la limpida e profonda sintesi di G.Devoto 'La méthode comparative et les conceptions actuales de la linguistique' presentata al X° Congresso Internazionale dei Linguisti nel 1967 a Bucarest. /81/.

Gli anni '50 hanno indubbiamente costituito un periodo di acuta crisi nello sviluppo della linguistica. L'attitudine dei tradizionalisti è stata critica sopratutto per il rifiuto al dialogo, a prendere conoscenza e familiarità con i metodi della nuova scuola; e non è stato certamente estraneo il disagio di fronte al complesso apparato simbolico e matematico, alla cui "lettura" la maggior parte dei linguisti non era adeguatamente preparata./151 p.5/./140 p.130/. Si stigmatizzavano le nuove correnti come manifestazioni di una errata scuola epistemologica /97 p.5/, come un modo di formulare fatti ben noti in un linguaggio complesso,senza produrre nessuna conoscenza di fatti nuovi e di nuovi aspetti dei fenomeni linguistici. /145 p.32/ /139 p.173/.

Nel 1962, nella relazione al X° Congresso Internazionale di Linguistica e di Filologia Romanza di Stasburgo, J.Iordan così descriveva le esitazioni e le resistenze dei romanisti: "Ils négligent et, aussi souvent, ils méprisent peut-être les courants modernes, tels le descriptivisme, le structuralisme, la linguistique mathématique (.....) Ce qui surprend, d'une manière assez pénible, ce n'est pas la réserve à l'ègard de ces nouvelles méthodes, mais le fait que leurs adversaires, qui sont très nombreux parmi les romanistes, n'interviennent pas dans les discussions théoriques qu'elles suscitent". /161/.

Per conto, il pregiudizio di molti aderenti ai nuovi metodi è ben riassunto nel frequente uso del termine "prescientifico" applicato all'intera eredità della scienza linguistica. A partire dai primi del '60, lo sforzo di riconciliazione delle due scuole diviene sempre maggiore.

In Italia nel 1959-60 Carlo Tagliavini assegna le prime tesi di laurea nel settore della statistica linguistica e computazionale /333/.

T.De Mauro, autore della voce 'statistica linguistica' dell'Enciclopedia Italiana /76/, presenta una memoria attinente la linguistica quantitativa dell'Accademia dei Lincei /75/ e pubblica successivamente altri studi sull'argomento /77/. Nell'anno 1962 l'almanacco letterario Bompiani è dedicato alla linguistica statistica e computazionale e svolge una inchiesta tra eminenti linguisti e pensatori italiani.

I fatti hanno reso ovvio che le nuove scuole forniscono importanti e utili prospettive teoriche alla disciplina linguistica, e ottengono risultati che non eranp acquisibili con i mezzi e le attitudini tradizionali /97 p.6/. L.Heilmann, in un articolo del 1964, afferma:"Chi è interessato ai problemi generali della disciplina non può disconoscere che il modo come questi vengono proposti ed affrontati da qualche tempo non consente più di sottrarsi alla discussione". Di fronte al moltiplicarsi delle indagini e al progressivo affinamento del metodo, egli propone al linguista tre quesiti:

a) - "la statistica linguistica è un semplice sussidio descrittivo o rappresenta una integrazione metodologica della nostra disciplina?
b) - l'indagine statistica è necessaria al pieno intendimento del fatto linguistico?
c) - quali sono per i linguisti, i limiti della statistica linguistica? " /141 p.36/.

D'altra parte, i sostenitori delle nuove teorie, sopratutto le nuove leve, reclutate dalla LM al di fuori delle scuole linguistiche tra logici, matematici, statistici, documentaristi, studiosi di linguaggi artificiali, hanno spesso duramente scontato con l'insuccesso la presunzione e la insufficiente valutazione della complessità dei fenomeni linguistici, determinata, sul piano dell'indagine diacronica, dal continuo riflettersi dei fatti sociali e storici nell'evoluzione del linguaggio /82/, sul piano dell'indagine sincronica, dal suo carattere di veicolo,sociale e individuale del pensiero.

Oggi, occorre sempre più spesso di trovare espressa, nella introduzione a scritti di linguistica matematica, la preoccupazione dell'autore di qualificare il proprio atteggiamento a questo riguardo, con frasi di questo tipo: " Nelle manifestazioni culturali dell'uomo esistono fattori che non possono essere ridotti a valori logici di 'vero' e 'falso', e che prendono in considerazione forze che distinguono l'uomo come animale razionale da una macchina" /237 p.58/.

La contestazione dell'eccessivo entusiasmo delle prime leve dei "linguisti matematici", non si limita certamente ad affermazioni generali e filosofiche, ma si concreta nella constatazione documentata delle contraddizioni, e della sterilità nei risultati, che verificano le teorie formulate in alcuni settori della linguistica matematica.

Nel campo della statistica linguistica, viene affermata la inadeguatezza della legge di Zipf, anche nella versione perfezionata da B.Mandelbrot /213/ /147p.438/ /52/ /146/ /193/ /214/(ma si vedano anche le risposte di Mandelbrot /202/).

I metodi e le procedure della glottocronologia e della lessicostatistica, intese a misurare con metodi esatti la cronologia storica delle lingue, (cfr./155 e 156/ ricchi di dettagli bibliografici) vengono ampiamente criticati /147 pp.444-5/.

Il risultato di questi insuccessi, fu da un lato l'allontanamento verso altre discipline di molti studiosi non formati a una scuola linguistica, dall'altro la rivalutazione dei metodi delle scuole tradizionali /97 p.5/ /245 pp.8-9/,dei loro problemi e una maggiore cautela e sobrietà nell'impiego del linguaggio e delle formule della matematica. /237 p.5 e pp.58-59/.

Nel settore della statistica linguistica un'atmosfera di meditata prudenza caratterizzò gli interventi al Congresso di Strasburgo del 20-22 Aprile 1964 "Statistique et analyse linguistique".

Nel volume degli atti apparso due anni dopo, l'introduzione a firma di Charles Muller e Bernard Pottier ne dà testimonianza: "..... Comme dans toute évolution des méthodes de recherche, après une période d'expansion, voir d'excès, il vient un moment d'èquilibre. Les linguistes savent à présent qu'il faut tenir compte des critères quantitatifs; ils ignorent dans quelle mesure il faut les considérer comme pertinentes, en fonction des champs spécifiques d'études". /295 p.1/.

Nelle conclusioni che ricordo approvate dall'assemblea dei congressisti (e riportate a chiusura degli atti /295 p.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 e di utilità nella scelta dei problemi da sottoporre al trattamento statistico. A questo scopo, si auspica che le nozioni matematiche e statistiche di base applicabili in linguistica (algebra, logica, teoria degli insiemi, teoria dell'informazione, calcolo delle probabilità, metodi statistici) prendano posto nella formazione data dall'Università ai linguisti.

Queste raccomandazioni non sono rimaste inascoltate, come dimostrano la pubblicazione di alcuni manuali introduttivi dedicati ai linguisti, per es. /48/, l'introduzione di insegnamenti relativi ai diversi settori di linguistica matematica a livello accademico in diverse nazioni europee, americane, e asiatiche, la revisione critica e programmatica di alcune convinzioni di base della statistica linguistica. /147 pp.438-445/, l'ideazione e la verifica di nuove tecniche per lo spoglio, le molteplicazioni degli spogli di testi per fornire materiali di base per le operazioni di induzione e di verifica delle ipotesi della statistica linguistica.

I materiali linguistici analizzati e registrati su nastro magnetico presso il C.N.U.C.E. e presso altri centri analoghi, costituiscono una fonte preziosa d'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 /641/, è agli inizi un progetto sulla sintassi, è molto avanzato uno studio a livello lessicale, che scelgo come esempio per illustrare alcuni problemi in questo settore.

3.1. I dizionari di frequenza.

Si è già rilevato qui sopra che gli spogli dei pedagoghi hanno contribuito ad attirare l'attenzione dei linguisti e degli statistici sui problemi della frequenza e della distribuzione delle parole. La prima della grandi inchieste statistiche moderne, e la più estesa fino alla utilizzazione dei calcolatori, è lo spoglio organizzato da Kaeding per la lingua tedesca; pubblicata a Berlino nel 1898, e destinato in un primo tempo al miglioramento dei sistemi di stenografia, richiese il concorso di circa 6000 collaboratori per lo spoglio di quasi 11 milioni di parole /172/. I dati quantitativi pubblicati da Eldridge /90/ nel 1911 per l'"inglese comune", furono, unitamente a quelli di Kaeding, utilizzati da G.K. Zipf per le formulazioni statistiche esposte nel secondo numero (The form and behavior of words) del suo "The Psycho-Biology of Language" /345/, e in particolare per la celebre quanto contestata legge di Zipf.

La grande compilazione di Keading mostra che:

- le 15 parole più frequenti rappresentano 25,22% (1/4 del totale);

- le 66 parole più frequenti rappresentano 50,06%(1/2 del totale);
- le 320 parole più frequenti rappresentano 72,25% (3/4 del totale).

Queste osservazioni, che hanno trovato conferma nelle compilazioni operate per altre lingue, hanno provocato ricerche miranti a stabilire dei vocabolari di base destinati all'insegnamento /128/. La loro compilazione sottintende un ragionamento di questo tipo: le osservazioni dimostrano che un numero relativamente piccolo di parole molto frequenti costituisce la parte maggiore di qualsiasi testo; queste parole sarebbero per lo più le medesime in testi diversi, cosicchè il loro insieme potrebbe essere identificato spogliando un numero sufficiente di testi; è da queste parole che è più economico e produttivo iniziare l'apprendimento di una lingua.

Già nel 1953, nella "Bibliographie de la statistique linguistique" di Pierre Guiraud, sono radunati più di 108 titoli di dizionari di base fondamentali, e di liste di frequenza, per 15 lingue /126, n.58, pp.44-50/.

Negli anni successivi i progetti di dizionari di frequenza sono continuati numerosi. Quasi contemporaneamente alla "Bibliographie" (e perciò non sono in essa citati) furono pubblicati i lavori di L.Rodriguez Bou /273/ basato sullo spoglio di 7.067.000 parole, per lo spagnolo, e di H.H.Josselson per il russo con spoglio di 1000000 di parole /165/.

Un nuovo impulso venne dato dal "Centre de Recherches et d'Etudes pour la diffusion internationale du Français (C.R.E.D.I.F.)", fondato nel 1959, presso l'Ecole Normale Supérieure de Saint-Cloud che ereditò l'iniziativa per il francese elementare o fondamentale, dovuta a Marcel Abraham nel 1951 /117 p.55/(Nota 21).

Gli autori hanno rinnovato la metodologia degli spogli, e il loro esempio è stato seguito da numerose iniziative, che ne accettarono sostanzialmente il metodo. Le ha passate in rivista P. Rivenc, al primo Congresso Internazionale di linguistica applicata (Nancy, 1964 /269/). Ricordo tra esse :

- Inchiesta sul tedesco parlato, diretta da J. Alan Pfeffer,della Università di Pittsburgh, /251/.
- Inchiesta per un vocabolario e una grammatica dello spagnolo parlato, affidata a P.Rivenc e a A. Rojo - Sastre. Nota 22.

In Irlanda Padre O'Huallachaine dirige inchieste sul piano fonematico e lessicale. /154/ . Le università di Montréal e di Québec conducono inchieste per il francese del Canada /274/.

Il C.R.E.D.I.F., in collaborazione con diverse università africane, conduce inchieste sul francese parlato e scritto nell'Africa francofona /114/.

Un'altra iniziativa di grande rilievo è la progettata serie di Dizionari di frequenza dedicata alle lingue romanze scritte, in corso di realizzazione sotto la direzione di Alphonse Juilland all'Università di Stanford. Sono già apparsi nel 1964 il volume per lo spagnolo /169/ e nel 1965 quello per il rumeno /170/, che viene così ad aggiungersi, per queste lingue, ai già esistenti lavori di A.Graur /118/ (che però non è basato su criteri statistici) /119 e 272/ V. Suteu /300/ e G . Bolocan /33/.

Gli spogli per questa bella collezione dell'editore Mouton sono condotti esclusivamente su materiali scritti, e gli autori giustificano l'aggiungersi di un nuovo dizionario di frequenza

delle parole in spagnolo a quelli esistenti citati più sopra, "by progress made in the new discipline of mechanolinguistics, wich gives acces to mechanical and electronic devices allowing for a more thorough scanning of larger samples of data; and by progress made in mathematical linguistics, wich permits a more reliable weighting of word usage" /169 p.XIII/. Ma sopratutto sottolineano il diverso intento della loro ricerca rispetto alle precedenti: si vogliono stabilire inventari di unità linguistiche corredati da accurati indici numerici che servano di base a investigazioni strutturali, e non più, o non solo, stabilire liste graduate di parole per l'insegnamento elementare delle lingue (ivi).

In una direzione analoga si colloca il progetto di un dizionario di frequenza dell'italiano scritto contemporaneo che vuol colmare la lacuna in questo settore per la lingua italiana, grave per molti aspetti, /31/ e che sta per essere pubblicato, con il patrocinio della IBM Italia, a cura di U.Bortolini, G.Tagliavini e A.Zampolli(Nota 23).

3.2. I problemi di composizione del campione lessicale.

Gli autori dei dizionari di frequenza hanno quasi sempre composto il corpus da spogliare con brani scelti da testi diversi, definiti da una differenza di origine (autore o soggetto parlante) o da una differenza di natura (genere letterario, ambiente sociale, situazione concreta, ecc.). Non solo le dimensioni complessive del campione scelto a rappresentare una determinata lingua sono, come abbiamo visto, notevolmente diverse, ma la scelta stessa dei testi risponde a criteri molto differenti. (Nota 24.)

Molti sforzi sono volti a riprodurre nella composizione del corpus campione la complessità di strati e di stili della lingua nel suo insieme. Si hanno così classificazioni e suddivisioni particolarmente complesse : (Nota 25) quali sono i motivi di queste a volte incredibili e variopinte diversità di composizione e di dimensioni del corpus assunto come campione? (Nota 26.)

3.3. 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"./227 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 /30 p.181 e segg.).

Da un lato occorre delimitare l'universo (Nota 27) dall'altro assicurarsi della rappresentatività del campione /228 p.14-15/. 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 stratifica-

zione sottile (ivi).(Nota 28.)

Che non sia questo il caso della lingua, è facilmente intuibile.

Le frequenze si osservano in un testo, cioè in una realizzazione del sistema da parte di un individuo, che è determinata da un lato dalle sue caratteristiche personali, dall'altro da quelle dalla comunità cui appartiene, caratterizzata sul piano sociale, geografico, storico, ecc. Se il testo è letterario, nel la sua produzione influiscono anche le regole e le peculiarità del genere letterario e in ogni caso la situazione in cui viene prodotto. I lessicografi hanno da sempre parlato di un linguaggio tecnico e di uno poetico, di uno popolare e di uno dotto, ecc. E' legittimo chiederci se attualmente possediamo, della struttura dell'universo linguistico, una conoscenza sufficiente a decidere le dimensioni e la composizione di un campione rappresentativo, e, addirittura, a stabilire se e in che limiti il rapporto tra testo e lingua sia legittimamente equiparabile, con rigore metodologico, al rapporto campione-universo.

Vi sono almeno due maniere diverse di concepire l'universo statistico.

3.4.1(a) La prima concezione, quella a cui si riferisce Muller, ha ricevuto formulazione teorica esplicita, come abbiamo visto, da P.Guiraud; alcuni cultori della materia vi riconoscono il fondamento teorico sul quale la statistica linguistica pone la propria autonomia come scienza.

Le unità del sistema linguistico sarebbero caratterizzate, oltre che dai tratti qualitativi emergenti dalle opposizioni e dalle relazioni che formano la struttura del sistema stesso, anche dalle loro rispettive probabilità di uso. Queste probabilità non sono direttamente osservabili, come non lo è del resto il sistema; esse però si tradurrebbero nel fatto (dato per certo) che le unità linguistiche ricorrerebbero nei testi, parlati e scritti, con frequenze relativamente stabili. In questa prospettiva, le frequenze osservabili nei testi, vengono assunte come approssimazioni delle probabilità non osservabili del sistema.

Sono interessanti a questo proposito le considerazioni di Ch.Muller sui rapporti tra lessico (lexique) e vocabolario (vocabulaire): egli riserva il primo di questi termini alla "langue", il secondo al "discours": le unità che compongono il lessico sono "lessemi" (lexème); quando queste unità virtuali sono attualizzate nel discorso, ciascuna di esse è un vocabolo (vocable), cui corrisponde un certo numero di occorrenze nel testo.

Si definisce dunque un vocabolario come l'insieme di vocaboli presenti, ciascuno con una propria frequenza, nel testo in esame. La definizione del lessico è più complessa. Il vocabolario di un testo presuppone l'esistenza di un lessico, che trascende il testo, del quale il vocabolario non è che un campione. Si può concepire una serie di insieme lessicali, ciascuno dei quali contiene il successivo. Ch. Muller li elenca così: il lessico di un idioma, nel senso estensivo del termine ma entro certi limiti cronologici; il lessico di una lingua, considerata sincronicamente; il lessico di un gruppo umano limitato appartenente alla stessa comunità linguistica; il lessico di un individuo di questo gruppo;(Nota 29) il lessico di questo individuo in una situazione definita stilisticamente o tematicamente.(Nota 30.)

Tra le unità lessicali o lessemi e gli insiemi citati si possono stabilire delle relazioni di inclusione o di esclusione.

Se queste relazioni potessero essere stabilite con rigore,

sarebbe possibile anche la quantificazione del lessico, mentre è possibile quantificare con rigore solo il vocabolario.

La relazione di inclusione può essere stabilita di solito attraverso la recensione del vocabolario; ma la relazione di esclusione può essere affermata solo per ragioni diacroniche, e dunque in un piccolo numero di casi. Per un individuo vivente, si potrebbe, almeno in linea di principio, pensare di apprezzare quantitativamente il suo lessico sottoponendogli la nomenclatura di un buon dizionario, per osservare l'accettazione o il rifiuto di ciascun lessema. Ma quanto si volesse delimitare il lessico di un autore scomparso, avremmo solo la certezza del suo vocabolario; l'inventario del suo lessico è impossibile, anche se ciò non impedisce di stimarne l'estensione mediante un processo di estrapolazione (su basi statistiche) reso di necessità approssimativo dalla nostra ignoranza della parte passiva o virtuale del lessico, la cui proporzione varia verosimilmente secondo gli individui.

Un lessico e un vocabolario sono caratterizzati, oltre che dalla estensione, da un contenuto e da una struttura.

Il contenuto è definito dalla natura delle unità che compongono il lessico o il vocabolario; la struttura è composta dagli elementi quantitativi semplici, che in un vocabolario sono facilmente conoscibili per mezzo dello spoglio: numero dei vocaboli del testo, frequenza di ciascuno, classi di frequenza e rapporto tra essi: l'insieme insomma dei dati che chiamiamo "distribuzione delle frequenze".

Si può trasportare questa nozione di struttura nel dominio del lessico? (come è possibile trasportare la nozione del contenuto, per lo meno per quanto riguarda la nozione di inclusione). E' il nostro problema di partenza : "Peut-on considérer la fréquence des vocables dans un texte ou un corpus comme une estimation satisfaisante de leur probabilité en langue?" /228 p.141/.

La probabilità di un lessema nel lessico di un individuo può essere conosciuta o stimata - risponde Muller - solo se la frequenza è stabile (tenuto conto delle variazioni aleatorie di cui conosciamo la natura) in tutti i discorsi di questo individuo, indipendentemente da ciò che noi abbiamo definito come "situazione". Il caso è -secondo Muller- molto più raro "qu'on ne l'a admis de façon précipitée". (Nota 31).

Quanto alla probabilità di un lessema nel lessico di un gruppo umano, si dovrebbe supporre che la sua frequenza sia costante presso tutti gli individui del gruppo, e Muller giudica questa possibilità ancora meno verosimile. Si è potuto sperare - egli dice - che la frequenza delle unità lessicali fosse stabile presso l'individuo, variabile nel gruppo: per ciò si attribuivano alla statistica lessicale straordinarie possibilità per risolvere problemi filologici di attribuzione. Egli teme che questa stabilità sia eccezionale, perfino "pour les 'mots de relation',qui sont les éléments le moins thématiques du lexique, car beaucoup d'entre eux, la plupart même, subissent visiblement l'effet des situations stylistiques".Detto per inciso, i dati da noi ottenuti negli spogli per il dizionario di frequenza dell'italiano,confermano l'esistenza di variazioni, nella frequenza delle "parole di relazione", tra i 5 sottoinsiemi esaminati; variazioni che appaiono probabilmente connesse alla natura del sottoinsieme, dal momento che hanno lo stesso andamento per la maggior parte delle unità appartenenti alla medesima categoria grammaticale .

Ciononostante Muller prende in considerazione egualmente una struttura quantitativa in materia di lessico, definita da probabilità associate ai lessemi, ma non le riconosce "quelque réalité que dans un lexique de situation". Per il lessico di un individuo, o di un gruppo, bisognerebbe invece fare una media di tutte le situazioni di discorso nel quale l'individuo o gli individui del gruppo potrebbero trovarsi, ponderando ciascuna con la propria probabilità. Questo compito gli sembra meno utopistico nel caso di un gruppo, nel quale il grande numero dei membri diminuisce la difficoltà di ponderare la situazione.

Questa conclusione presenta, a mio avviso, parziali analogie con la conclusione di Moreau, qualora si riconoscano delle analogie tra il concetto "lessico di situazione" di Muller e il concetto "centro di interesse" di Moreau.

3.4.2.b) Nelle considerazioni esposte da Moreau nel suo intervento "Au sujet de l'utilisation de la notion de fréquence en linguistique" al "Colloque International sur la Mécanisation des Recherches Lexicologiques"/221/, l'universo statistico è definito come l'insieme di tutti i testi (parlati e scritti) prodotti in un certo periodo di tempo. Anche ammettendo che l'insieme sia definibile esattamente, un gran parte di questo insieme sfugge ad ogni misura: per es. le lettere, le conversazioni private, ecc.... Moreau descrive la frequenza di una parola di questo insieme come una superficie in uno spazio (v. Fig.3) le cui 3 dimensioni sono la frequenza, il tempo (variazioni nel tempo), lo spazio dei messaggi (libri, riviste, lettere, rapporti, conversazioni, discorsi, ecc...., formano altrettanti sotto-spazi più o meno omogenei, più o meno strutturati, più o meno indipendenti).

La superficie presenta degli accidenti più o meno rilevanti a seconda che, per esempio, una parola molto adoperata si usuri, o che cambi il centro d'interesse.

Come definire allora la frequenza di una parola nella lingua? Con il suo valore medio? Innanzi tutto non è possibile contare il numero complessivo delle occorrenze della parola, e si dovrà fare una stima a partire da quel che si può conoscere. Inoltre, che significato avrebbe questa misura media? Se attingiamo a una presa di corrente, esemplifica Moreau, con una tensione di 120 Volts, un'ora su due, non parleremo di tensione media di 60 Volts. Se la superficie che rappresenta la frequenza di una parola nella lingua (nello spazio a tre dimensioni precedentemente definito) fosse un piano orizzontale, il calcolo della frequenza della parola (che è allora la stessa in tutto il piano) non necessiterebbe di alcuna precauzione.

Ma tale rappresentazione esige che la parola sia impiegata in modo identico in tutti i centri di interesse e che il suo impiego non vari nel corso del tempo.

Secondo Moreau, queste condizioni sarebbero realizzate, approssimativamente, dalle parole di "classe fermée". La stima delle loro frequenze a partire da testi campione non pone problemi. Conoscendo la loro frequenza nella lingua, d'altra parte, si può egualmente stimare la probabilità di apparizione di un testo dato: egli chiama tali parole "mots ergodiques": hanno la caratteristica essenziale di essere "pas trop inégalement répartis dans le corpus" (Nota 32)

Esistono di tali parole? Moreau lo afferma; sarebbero in particolare tutte le parola "athématiques", secondo l'espressione di

Michéa; quelle che servono "à nous exprimer au sujet des choses plutôt qu' à exprimer les choses elles-mêms, des termes plus ou moins communs à tous les sujets et à toutes les situations". Si collocherebbero in questa categorie un certo numero di aggettivi e di verbi correnti, le parole grammaticali, alcuni nomi assai generali.

Ma che fare per il caso di parole di "classe ouverte", che, al contrario, presentano oscillazioni di frequenza da un testo all'altro, e spesso tra brani di uno stesso testo?

Per valutare la loro frequenza il solo modo operativo conveniente consisterà nello stratificare a priori la lingua, nel delimitarvi dei centri di interesse (Centres d'intérêt, strates) /221 p.154/, all'interno dei quali sono tematiche e quindi ergodiche parole che pure sono inegualmente ripartite nell'insieme della lingua.

L'idea di Moreau è, appunto, che parole non ergodiche nella lingua, lo siano invece in testi con un tema, un centro di interesse comune.

Per esempio nel centro d'interesse "algebra elementare" parole tematiche come eguaglianza, diseguaglianza, equazione, incognita, variabile, funzione saranno frequenti ed ergodiche. I risultati ottenuti sui testi accessibili alla campionatura, potranno essere estesi agli altri testi, alla parte inesplorata, del centro d'interesse. Come ha senso classificare per frequenza le parole generalmente ergodiche nella lingua, così potranno essere messe in ordine di frequenza, all'interno del loro centro d'interesse, le parole tematiche. Ma vi sono almeno due problemi relativi alla determinazione dei centri d'interesse. In primo luogo si dovranno cercare per ogni parola, anche di frequenza debole, i centri d'interesse che la concernano, generando così, alla fine, una straordinaria moltiplicazione degli strati? O invece si definirà a priori, nella lingua, un numero ragionevole di centri d'interesse, e si cercherà all'interno di ciascuno solo la frequenza di alcune parole? Per Moreau, la risposta sta nella constatazione che, per alcune parole, è impossibile delimitare un qualsiasi centro di interesse nel quale siano ergodiche; sarebbero, approssimativamente, quelle che, come abbiamo visto, il CREDIF, nella propria terminologia, chiama mots disponibles. Definire una frequenza per le parole disponibili non ha senso - dice Moreau - infatti il CREDIF ha preferito non indicarne la frequenza nel francese fondamentale.

Moreau propone di stratificare una volta per tutte la lingua di una data epoca in un certo numero di gruppi omogenei, in centri di interesse, che potranno essere anche molto numerosi, del tipo, per es. matematica, fisica, chimica, ecc.. All'interno di ciascuno strato, fissato dunque non più in funzione di una parola, ma della lingua, sarà determinata la frequenza delle sole parole frequenti ed ergodiche in questo insieme. Sembra più interessante sapere in quali centri di interesse si incontra una parola, e quale è la sua frequenza in questi centri di interesse, che conoscere la frequenza media delle parole nella lingua. (Nota 33.) Egli suggerisce una presentazione dei dati di frequenza del tipo riprodotto nella Fig. 4 (Nota 34).

Una lista delle parole frequenti stabilita con un procedimento di questo tipo avrà un certo valore, almeno fino a che le fluttuazioni riscontrate nella ergodicità non superano un certo limie.

3.5. La determinazione dei centri di interesse avrà evidentemente qualcosa di arbitrario e la varietà di strutture e partizio-

ni che abbiamo riscontrato nella campionatura dei Dizionari di Frequenza lo conferma.

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

D'altra parte la linguistica quantitativa ha già ottenuto alcuni importanti risultati, sopratutto per quanto riguarda la applicazione e l'adattamento dei metodi classici della statistica ai problemi della campionatura lessicale. /226 p.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 ed ugualmente 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 ergotiche.

La determinazione degli strati di lingua potrà allora derivare da un lavoro accurato e documentato di induzione 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 /222 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 /226 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, oggi, sembra estremamente ragionevole.

4. Il dizionario di frequenza dell'italiano contemporaneo.

In questa stessa direzione, gli spogli e le elaborazioni per il dizionario di frequenza dell'italiano contemporaneo (Nota 35) si pongono come primo contributo (Nota 36) sistematico allo studio statistico dell'universo lessicale, attorno al quale potrà organizzarsi la elaborazione, per lo stesso fine, dei dati provenienti dagli ormai numerosissimi spogli che arricchiscono progressivamente la biblioteca elettronica del C.N.U.C.E.

I testi spogliati si distribuiscono entro limiti cronologici ben precisi, tra il 1948 e il 1968.

Essi sono stati scelti in 5 categorie diverse :

1) - sceneggiature cinematografiche
2) - copioni teatrali
3) - romanzi
4) - giornali, quotidiani e periodici
5) - testi di divulgazione scientifica.

La scelta dei testi di ciascuna categoria è stata fatta con

la consulenza di esperti del settore, per mancanza di informazioni obiettive sulla composizione dei sottoinsiemi di lingua, già lamentata. Per ciascuna categoria sono state sottoposto a spoglio esattamente 100000 occorrenze, dunque un totale complessivo di 500000 occorrenze. Nel numero non sono compresi i nomi propri, e le parole non appartenenti al sistema linguistico italiano (dialettali o straniere, secondo la competenza linguistica degli autori degli spogli).

I testi sono stati perforati ed elaborati secondo la metodologia attualmente in uso alla sezione linguistica del C.N.U.C.E. (descritta al punto 1.1.).

Particolare attenzione è stata dedicata alla coerenza dei criteri di lemmatizzazione, che sarebbe troppo lungo esporre in questa sede. Mi limito ad osservare che le omografie possibili e attuali sono state separate e opportunamente contraddistinte a livello di lemma, mentre non si è distinto tra flessioni omografe di uno stesso lemma.

I dizionari presi come base (Cappuccini-Migliorini,Dizionario Enciclopedico Italiano, Garzanti, ecc.) sono stati di utile sussidio, ma spesso si sono rilevati insufficienti, per la presenza di numerosi neologismi e di nuove accezioni nei nostri testi.

4.1. Il procedimento di stratificare il corpus, suggerisce immediatamente l'idea di completare la nozione di frequenza con quella della stabilità della frequenza, o, se si preferisce, di correggere la frequenza con il modo nel quale si distribuisce nel corpus: è inevitabile che la frequenza globale di un elemento sia soggetta a degli accidenti che porterebbero a sopravvalutarla, per esempio nel caso che una situazione imprevedibile accumuli in una parte del corpus un numero elevato di occorrenze di questo elemento /227 p.34/. Ci sono due tipi di tecniche fondamentali per tener conto di questo fatto: (1) /169 p.XLV/: una tecnica semplice che consiste nell'assumere come indice dell'uso di una unità lessicale il numero di sottoinsiemi nel quale occorre (che chiamerò d'ora in poi R : se il numero dei sottoinsiemi è N, R potrà variare da 1 a N); (2) una varietà di tecniche più complesse, che producono diverse formule per misurare la distribuzione della frequenza nei sottoinsiemi.

Un esempio della prima tecnica è il lavoro pionieristico del Kenniston /177 p.86/, che divide il suo vocabolario dello spagnolo in 6 classi : le parole della lista 1° sono trovate in almeno l'80% dei testi studiati, quelle della lista 2° in almeno il 67,7%, ecc.

Il principio sottointeso è che la distribuzione è meno soggetta della frequenza alle variazioni della selezione casuale,e che essa provvede un indice migliore dell'uso della unità lessicale che non il numero globale delle occorrenze nell'universo considerato come blocco unico.

Un altro esempio è quello di G.E.Vander Beke /322/. Egli chiamò range (R) il numero dei sottoinsiemi nei quali la unità lessicale è rappresentata: poichè egli operava su 88 testi distinti, R era dunque un numero intero che variava da 1 a 88. Il suo vocabolario di base elimina le unità lessicali con R inferiore a 5, e ordina le restanti secondo il decrescere di R, indipendentemente dalla frequenza. Anche il vocabolario russo della Josselson /165/, ad eccezione delle prime 204 parole scelte tenendo conto anche delle frequenze, ordina tutte le altre, estratte da un dni-

verso di 1000000 di occorrenze e diviso in 100 sottoinsiemi,esclusivamente in base ad R, senza che la frequenza abbia alcuna influenza sull'ordinamento.

Vediamo ora alcuni esempi di tecniche miste.

Nel Francaise fondamental, du premier degré del C.R.E.D.I.F., (Nota 37) gli autori chiamano répartition /116 p.69/ l'indice R sopra definito.

La priorità è data alla frequenza, che fornisce la soglia di eliminazione (fissata a 20 occorrenze) e determina l'ordinamento dei lemmi; la ripartizione interviene solo in un secondo tempo: per eliminare i lemmi che, con una frequenza uguale o superiore a 20, appaiono in meno di 5 dei 163 testi spogliati, e poi per ordinare tra loro i lemmi che hanno uguale frequenza. (Per una valutazione del metodo si veda Ch. Muller /227 p.35/). Victor Garcia Hoz distingue il vocabolario risultante dai suoi spogli in tre zone, per così dire concentriche:

a) - vocabulario usual. Opposto a vocabulario total, comprende sia il léxico activo costituito dall'insieme delle parole che l'uomo comune impiega correntemente nella conversazione e scrittura spontanea, sia il léxico latente, formato dalle parole che senza essere usate in modo spontaneo, sono comprese senza difficoltà, quando si leggono e si odono. (Nota 38). Egli si chiede: come stabilire il limite preciso tra le parole usuali e le non usuali? Si potrebbe cercare di stabilire il numero di parole conosciute dall'uomo medio: ma ogni uomo concreto ha nel suo vocabolario parole speciali della sua professione e del suo ambiente; inoltre le investigazioni in questo senso hanno condotto a risultati contrastanti /153 pp.57-58/. Per esempio C.K.Ogden segnalò il vocabolario di una cameriera tra le 7 e le 8mila parole; quello di un dirigente sindacalista tra 10 e 15mila, quello di W.Churchill tra 20 e 30mila, quello di uno scienziato tra 60 e 80mila.

A.F. Watts /323 pp.57-58/ suggerisce una media di 10000 parole all'età di 14 anni, e, ammettendo un accrescimento di 400 parole all'anno, colloca a 15000 il vocabolario della maggioranza degli adulti. Garcia Hoz accetta questa cifra, e decide di accumulare spogli successivi fino al raggiungimento di un tal numero di lemmi diversi, che accoglie nel vocabulario usual. (In realtà egli si ferma prima, dopo aver spogliato 400000 parole e ottenuto 13000 lemmi, supponendo che i rimanenti 2000 possano appartenere alla porzione specialistica del lessico individuale).E' da notare che il nostro spoglio ci ha fornito circa 15700 lemmi,cifra curiosamente vicina a quella proposta da Watts. Questa coincidenza è probabilmente attribuibile solo alle dimensioni del nostro campione.

b) - Vocabulario comun.

Una volta registrati tutti i lemmi con le frequenze rispettive in ciascun sottoinsieme, si separano dagli altri quelli che hanno almeno una occorrenza in ciascun sottoinsieme (complessivamente 1971). Si tratta, evidentemente, di un procedimento nel quale interviene solo R. Ma Garcia Hoz si rende conto dello scarso potere discriminativo di questa formula. Egli confronta, come esempio, probabilidad che ha 4 occorrenze complessive, ripartite una in ciascuno dei 4 sottoinsiemi, e elemento che, con 104 occorrenze complessive, figura solo in 3 sottoinsiemi: manca solo nella corrispondenza privata. Egli sente l'esigenza di combinare assieme frequenza e ripartizione; "esta última clase de palabras no puede figurar en el vocabulario común; mas puede decirse que sean de menos interés que algunas de los que figuran en él, teniendo

menos frecuencias en total?" Egli aggiunge perciò al vocabolario comune una relacion adicional /153 p.386/: altri 212 lemmi, che pur non apparendo in tutti e 4 i sottoinsiemi, superano la frequenza di 40, da soli, o unendo insieme (sic!) lemmi diversi,ma strettamente affini tra loro. Il numero di 40 è stato fissato in relazione al calcolo di correlazione menzionato al punto c). Nell'insieme, al vocabulario comun corrispondono 339110 occorrenze.

c) - Vocabulario fundamental.

E' costituito da lemmi che hanno la frequenza distribuita praticamente in parti uguali nei sottoinsiemi. Per misurare la uniformità della distribuzione, G.Hoz ha applicato il calcolo delle correlazioni, in base al quale ha scelto 208 parole (che rappresentano 245384 occorrenze) (Nota 40).

Buchanan /37/ propone una formula che ha un certo potere discriminativo, ma solo in relazione alle parole di frequenza uguale distribuita in un numero diverso di sottoinsiemi.

La formula è

$$U = \frac{F}{10} + n_i$$

dove U sta per valore di uso, coefficiente a un tempo, per l'ordinamento e per la soglia di accettazione delle parole; F è la frequenza complessiva nel corpus; n_i è il numero dei sottoinsiemi nei quali appare. Evidentemente questa formula non ha potere discriminativo tra parole con la stessa frequenza ripartita non allo stesso modo, in uno stesso numero di sottoinsiemi e,dà un peso troppo grande alla ripartizione nei confronti della frequenza sopratutto nel caso di un grande numero di sottoinsiemi.

4.2. A.Juilland e E.Chang Rodriguez /169 pp. LII-LIV/ hanno introdotto una diversa misura della distribuzione, che chiamano dispersione (inglese = dispersion), nel loro Frequency Dictionary of Spanish Words, basato sullo spoglio di 500.000 occorrenze circa, ripartite in 5 sottoinsiemi di eguali dimensioni, cioè di 100.000 occorrenze ciascuno:

1° - Teatro
2° - Romanzo, novelle
3° - Saggistica
4° - Periodici
5° - Letteratura scientifica e tecnica.

Dopo aver esaminato criticamente le soluzioni dei predecessori, essi cercano un metodo migliore.

Una prima formula per la misura della dispersione di una data parola potrebbe essere :

$$D = \frac{f^M - f^m}{F}$$

dove la D è la dispersione, f^M e f^m sono rispettivamente la frequenza maggiore e la frequenza minore tra le diverse frequenze dei sottoinsiemi, e F è la frequenza complessiva nel corpus.

La formula è giustificata col seguente ragionamento.

Se consideriamo parole con frequenze (F) eguali, una possibile misura della dispersione è data dalla differenza tra la frequenza più alta (f^M) e la frequenza più bassa (f^m) degli n sottoinsiemi.

Prendiamo per esempio 4 parole con 10 occorrenze diversamente distribuite in 5 sottoinsiemi. Siano P^a= 2,2,2,2,2, p^b=3,2,2,2,1 P^c= 4,3,2,1,0 e P^d = 0,0,0,0,10. Le differenze fra la frequenza parziale più alta e la frequenza parziale più bassa sarà allora:

$P^a = 2 - 2 = 0$; $P^b = 3 - 1 = 2$; $P^c = 4 - 1 = 3$; $P^d = 10 - 0 = 10$. La differenza 0 per Pa, 2 per P^b, 3 per P^c, 10 P^d indicano in modo inverso la uniformità, la "bontà" della dispersione: più piccola è la differenza, migliore è la dispersione, e viceversa.

Ma questa differenza è ingannevole quando operiamo con parole di diversa frequenza; la stessa differenza, per es. 100, può indicare una dispersione uniforme se il numero delle occorrenze (F) è grande, e una dispersione poco uniforme se F è piccolo.

Prendiamo due esempi estremi; siano

$P^x = 1,1,1,1, 100$, in cui F è 104, $f^{Ma} = 100$, $f^m = 1$, $\frac{f^m}{f^{Ma}} = \frac{1}{100}$

$P^y = 900,900,900,900,1000$, in cui F è 4600, $f^{Ma} = 1000$,

$f^{mi} = 900$ $\frac{f^{mi}}{f^{Ma}} = \frac{9}{10}$

Un mezzo semplice di correzione è dividere la differenza per il totale delle occorrenze nell'universo (F), donde la formula predetta.

Nell'ultimo esempio in questione

$$D^x = \frac{100 - 1}{104} = 0,95 \qquad D^y = \frac{1000 - 900}{4600} = 0,02$$

in cui D^x (cattiva dispersione) è più vicina al 1, mentre D^y (dispersione migliore) è più vicina a 0.

E' da notare che la dispersione sarà rappresentata da un coefficiente variante da 0.00 (dispersione ideale) fino a 1.00 (disperzione peggiore).

per es.

per $^a = 2,2,2,2,2$, $\qquad D^a = \frac{2 - 2}{10} = 0$

per $^d = 0,0,0,0,10$ $\qquad D^d = \frac{10 - 0}{10} = 1$

Diciamo ora x_i, dove $i = 1,.....n$, la frequenza con cui l'unità lessicale compare nel sottoinsieme S_i. La frequenza nel corpus sarà :

$$F = \sum_{i=1}^{n} x_i$$

la frequenza media sarà

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{F}{n}$$

Consideriamo le seguenti distribuzioni di 7 parole con un totale di 5 occorrenze in un universo diviso in 5 sottointesi

	x1	x2	x3	x4	x5
p^a	1	1	1	1	1
p^b	2	1	1	1	0
p^c	2	2	1	0	0
p^d	3	1	1	0	0
p^e	3	2	0	0	0
p^f	4	1	0	0	0
p^g	5	0	0	0	0

si avrà $\overline{x} = \frac{5}{5} = 1$
Chiediamo ora D, la dispersione, come la somma delle deviazioni da $\overline{x}$ (frequenza ideale), trascurando il segno positivo o negativo della deviazione. Avremo :

		(x1 - $\overline{x}$)	+ (x2 - $\overline{x}$)	+ (x3 - $\overline{x}$)	+ (x4 - $\overline{x}$)	+ (x5 - $\overline{x}$)	= D
D^a	=	0	0	0	0	0	0
D^b	=	1	0	0	0	1	2
D^c	=	1	1	0	1	1	4
D^d	=	2	0	0	1	1	4
D^e	=	2	1	1	1	1	6
D^f	=	3	0	1	1	1	6
D^g	=	4	1	1	1	1	8

Per parole con la stessa frequenza, la somma delle deviazioni fornisce una misura abbastanza soddisfacente: più piccola la somma, migliore la dispersione e viceversa. Ma per parole di frequenza diversa essa è ingannevole allo stesso modo della formula precedente.

La stessa deviazione totale può avere significati diversi, a seconda che la frequenza totale sia grande o piccola; più grande è la frequenza, meno significativa è la deviazione totale, e viceversa.

Ma anche per le parole con la stessa frequenza, la deviazione totale non ha un potere discriminativo sufficiente. Negli esempi precedenti, la stessa deviazione totale di 4 risulta sia per D^c= 2,2,1,0,0, sia per D^d = 3,1,1,0,0, e la stessa deviazione totale di 6 per D^e = 3,2,0,0,0, e per D^f = 4,1,0,0,0, mentre la dispersione di P^c è migliore di quella di P^d, e quella di P^e è migliore di quella di P^f, come mostra la differenza tra il coefficiente più alto e il coefficiente più basso.
Infatti :

$$f^{Ma} - f^{mi} = (2 - 0) = 2 \text{ per } P^c$$
$$= (3 - 0) = 3 \text{ per } P^d$$
$$= (3 - 0) = 3 \text{ per } P^e$$
$$= (4 - 0) = 4 \text{ per } P^f$$

Per questo gli Autori propongono una formula diversa, (Nota 41) che qui espongo nella notazione che ne ha dato Muller nell'articolo citato Fréquence dispersion et usage /227/(Cfr. anche la versione di Torrigiani /310 pp.14-18/.
Si stabilisce dapprima la varianza, rappresentata con σ^2, che è la media dei quadrati degli scarti dalla frequenza teorica $\overline{x}$:

$$\sigma^2 = \frac{\Sigma i \, (x_i - \overline{x})^2}{n}$$

e lo scarto tipo $\sigma = \sqrt{\sigma^2}$ che, a dispersione uguale, cresce con la frequenza F, dunque con la frequenza media $\overline{x}$.
Chiameremo coefficiente di variazione il rapporto

$$V = \frac{\sigma}{\overline{x}}$$

che è indipendentemente dalla frequenza (F). Risulterà V = 0 quan-

do $\sigma = 0$, cioè quando $x_i = \overline{x}$ (i = 1,....., n), cioè quando l'unita lessicale si presenta un egual numero di volte negli n settori.

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

quando l'unità lessicale appare in un solo dei sottoinsiemi.
Nota 44.

Per ottenere un indice che vari tra 0 (dispersione ottimale, uniforme nei vari sottoinsiemi) e 1 (dispersione pessima, in un solo sottoinsieme), basterà dividere V per il suo valore massimo. Il rapporto

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

cresce col diminuire della uniformità della distribuzione. Se preferiamo un indice che, invece, cresca concordemente al crescere del grado di uniformità, basterà, per invertire il senso della variazione, introdurre il complemento all'unità. Questa formula

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

qualifica la dispersione di ogni parola del corpus per mezzo di un coefficiente di dispersione (D) che varia tra 0 (dispersione pessima) e 1 (dispersione ottima).
Verifichiamo per le parole P^a e P^g del precedente esempio di 7 parole, con F = 5

per P^a avremo :

$x1 = x2 = x3 = x4 = x5 = \overline{x} = 1$

e dunque

$(x1 - \overline{x}) = (x2 - \overline{x}) = (x3 - \overline{x}) = (x4 - \overline{x}) = (x5 - \overline{x}) = 0$

e dunque

$$\Sigma(x_i - \overline{x})^2 = 0$$

per cui la varianza è

$$\sigma^2 = \frac{\Sigma(xi - \overline{x})^2}{n} = \frac{0}{5} = 0$$

e anche

$$\sigma = \sqrt{\sigma^2} = 0$$

e

$$V = \frac{\sigma}{\overline{x}} = \frac{0}{1} = 0$$

poichè

$$\sqrt{n - 1} = \sqrt{5 - 1} = 2$$

si avrà

$$D = 1 - \frac{V}{\sqrt{n - 1}} = 1 - \frac{0}{2} = 1$$

Invece, per D^g = 5,0,0,0,0 avremo

xi	$(xi - \overline{x})$	$(xi - \overline{x})^2$
x1 5	+ 4	16
x2 0	- 1	1
x3 0	- 1	1
x4 0	- 1	1
x5 0	- 1	1
		20 = $\Sigma(xi - \overline{x})^2$

da cui abbiamo $\sigma^2 = \frac{\Sigma(xi - \bar{x})^2}{n} = \frac{20}{5} = 4$

lo scarto tipo

$$\sigma = \sqrt{\sigma 2} = 2$$

mentre il coefficiente di variazione sarà

$$V = \frac{\sigma}{x} = \frac{2}{1} = 2$$

e il coefficiente di dispersione sarà

$$D = 1 - \frac{V}{\sqrt{n - 1}} = 1 - \frac{2}{\sqrt{5 - 1}} = 0$$

D è un coefficiente indipendente dalla frequenza F; pertanto sarà tra parole di frequenza eguale che andrà dimostrato il suo potere di discriminazione. Possiamo agevolmente verificare questa proprietà di D, calcolandolo per tutte e 7 le parole del nostro esempio, e ricordando le esigenze avanzate a proposito delle coppie di parole P^c, P^d e P^e : P^f.
Avremo infatti

$D^a = 1.000$
$D^b = 0.685$
$D^c = 0.577$
$D^d = 0.450$
$D^e = 0.368$
$D^f = 0.225$
$D^g = 0.000$

Uso (usage). Gli autori ottengono un terzo indice, dopo la frequenza e la dispersione, che chiamiamo usage, con la formula seguente:

$$U = \frac{F \cdot D}{100}$$

Gli autori discutono a lungo le diverse maniere di collegare la frequenza e la dispersione, e propongono una complessa casistica per la quale rimandiamo alle pagg. LXI - LXII della loro opera. In accordo con il concetto assunto che il coefficiente di uso di una parola mira a predire o ad approssimare l'occorrenza della parola in un campione "ideale", cioè perfettamente rappresentativo e senza difetti, la formula ha per effetto di correggere la frequenza con la dispersione, piuttosto che viceversa, anche se si potrebbe, forse, dire che i due fattori sono coordinati. Abbiamo visto che la dispersione varia da O ad 1; la formula di U tratta il coefficiente di dispersione come una percentuale della frequenza. U è uguale a F quando la parola è ripartita uniformemente nel corpus, poichè in tal caso D = 1. U = O, quando le occorrenze sono concentrate tutte in un solo sottoinsieme, perchè D = O. A frequenza uguale, U è tanto più vicino a F quanto più la ripartizione della parola è regolare. A partire da U=F per una ripartizione perfetta, U decresce proporzionalmente al decrescere di D.

Per altri commenti rinvio alla citata recensione di Muller. Egli avanza qualche critica: per esempio, nel caso particolare del campione a 5 sottoinsiemi, sono avvantaggiate le parole con frequenza multipla di 5. Così una parola con frequenza 10 può avere teoricamente un indice U = 10. Ma una parola di frequenza 9 non può raggiungere che U = 8, e per F = 11, non si può sorpassare U = 10. Questo difetto si riflette nell'ordinamento dei lemmi, e potrebbe essere ridotto aumentando il numero dei sottoinsie

mi.

La difficoltà di tale aumento risiede nel dover specializzare maggiormente il campione, cioè nell'immaginare un numero maggiore di strati o sottoinsiemi, e nel reperire i testi relativi. La elaborazione e il calcolo non ne sarebbero però minimamente appesantiti o complicati ; usando il calcolatore, infatti, nè la fase di lemmatizzazione nè la fase di applicazione della formula comporterebbero maggiori difficoltà di programmazione o di esecuzione.

Un'altra modifica è suggerita da G.Torrigiani /144/, il quale propone di rendere più significativo il coefficiente di dispersione assumendo

$$D = \sqrt{1 - \frac{v^2}{n - 1}}$$

4.3.4. Nella elaborazione del dizionario italiano, abbiamo per ora preferito seguire per U la formula di Juilland e Rodriguez, per garantire la confrontabilità tra i nostri dati e i loro.

Ci siamo però discostati per quanto riguarda la soglia di esclusione della parola. Essi hanno scelto il materiale da pubblicare su un elenco di circa 9.000 lemmi, rimasti dopo aver eliminato, in un primo tempo, tutti i lemmi con $F < 4$ (circa 6.000) e poi quelli con $R < 3$ (cioè presenti solo in 1 o 2 sottoinsiemi: circa 5.000). Essi spiegano queste decisioni con motivi di convenienza pratica, per semplificare le operazioni di calcolo.

Dal punto di vista della elaborazione con il calcolatore,il loro procedimento ci è sembrato invece più complesso: (Nota 45) pertanto abbiamo calcolato D e U per tutti i lemmi. Al momento attuale, stiamo elaborando liste diversamente ordinate dei 15.700 lemmi e delle 36.000 forme trovate nel nostro corpus. Eccone alcune :

- in ordine di frequenza complessiva (F)
- in ordine di frequenza per ciascun sottoinsieme (xi)
- in ordine di dispersione (D)
- in ordine di uso (U)
- in ordine di (F,xi,D,U), all'interno di ogni singola categoria grammaticale,ecc.

In ciascuna di queste liste, il lemma e la forma sono accompagnati, oltre che dal codice di categoria grammaticale, da F, da xi, da D, da U, dal numero del rango occupato nell'ordine scelto per la lista, e dalla sommatoria di F per tutte le unità che in tale ordine precedono (questo per poter conoscere, ad esempio, quando F sia l'elemento ordinatore, quanto occorrenze corrispondono alle prime n unità, a una tranche compresa tra il rango n ed $\underline{n + x}$, ecc.

Da queste liste estrarremo l'insieme dei materiali lessicali per pubblicare il nostro contributo a un dizionario di frequenza dell'italiano di oggi. Probabilmente la scelta dei lemmi verrà operata disponendoli in ordine di U, e fissando poi un certo valore di U come soglia di eliminazione.

L'analisi descrittiva e la interpretazione dei dati raccolti, appena iniziata, si preannuncia estremamente interessante, per esempio dal punto di vista lessicografico (Neologismi, nuove accezioni, ecc.). (Nota 42).

Mi preme qui sottolineare l'orientamento della ricerca verso il problema della stratificazione del corpus e delle relative strutture quantitative, nel contesto teorico presentato, alla fine del

paragrafo precedente.

Anche prescindendo dal confronto tra i diversi testi spogliati, i 5 sottoinsiemi (nei quali li abbiamo raggruppati per il calcolo di D) sembrano possedere caratteristiche quantitative ben differenziate. Mi limito ad un esempio. 4 dei 5 sottogruppi sembrano disporsi in questo ordine : C T R G

C - Cinema
T - Teatro
R - Romanzi
G - Giornali.

Per esempio :

a) cresce secondo il predetto ordine :
   - il numero dei lemmi diversi
   - il numero delle forme diverse
   - la frequenza del lemma di rango 1
b) decresce invece nello stesso ordine
   - la frequenza del lemma di rango 50
   - la frequenza del lemma di rango 100
c) Il numero dei lemmi, il numero delle forme, il totale delle occorrenze, crescono secondo il predetto ordine per:
   - articoli
   - sostantivi
   - aggettivi
   - preposizioni
d) decrescono invece per :
   - esclmazioni
   - pronomi
   - verbi
   - avverbi
   - congiunzioni
   - avverbi

(non seguono quest'ordine i numerali).

Una posizione particolare occupano i testi scientifici, i quali, per le categorie grammaticali, (c e d), si allineano all'estremità opposta al cinema, a destra dei giornali (l'allineamento dunque è C,T,R,G,S); mentre si situano, per i valori a.b., tra i romanzi e i giornali (la sequenza è dunque C T R S G).

Potremo dunque descrivere la situazione anche con una sequenza di questo tipo

I	II	III
C+T	R	S/G

L'allineamento cinema-teatro è sempre rispettato per i parametri suddetti: ciò potrebbe indicare che il fattore o i fattori che determinano la caratterizzazione quantitativa reciproca dei due sottoinsiemi operano concordemente. Questa osservazione varrebbe anche per la caratterizzazione reciproca nei gruppi I, II, III; il gruppo I e il gruppo III si dispongono simmetricamente attorno al gruppo II: i valori del sottoinsieme R (II), nei parametri considerati, coincide sempre, più o meno approssimativamente, col il valore medio del parametro del corpus.

A quali motivi attribuire queste regolarità? Significano forse esse che i denominatori comuni, la classificazione in base ai quali i testi sono stati raggruppati in sottoinsiemi, sono rilevanti rispetto alla struttura della stratificazione? (Nota 43).

N O T E

N° 1 - Per le comunicazioni all'ICCL qui passate in rassegna,non riporto accanto all'autore citato il numero di rinvio alla bibliografia: infatti la comunicazione è univocamente indicata dal nome dell'autore e dalla sigla ICCL che la segue.

N° 2 - Mentre nello stesso volume 3 articoli sono dedicati interamente o in parte allo sviluppo della linguistica matematica negli USA nello stesso periodo.

N° 3 - Vale a dire un elenco di istruzioni che specificano la serie di operazioni con le quali è possibile risolvere ogni problema di un dato tipo /311 p.5/.

N° 4 - Non so se si realizzerà il futuro ipotizzato da G.Hays nel recente "Seminar on Computational Linguistics" di Bethesda (USA, 1956) /134 p.69/ : la Typewriter dello studioso collegata, 'time-sharing', al computer; un D.M. corregge i primi errori, risolve le abbreviazioni, completa i riferimenti bibliografici memorizzati in file speciale. La fotocompositrice pone sotto il suo diretto controllo anche l'architettura della pagina.

N° 5 - Mentre il Type-setting, a causa degli errori meccanici assai frequenti nella caduta dei caratteri, trascinati alla fusione per forza di gravità, richiedeva la lettura delle bozze (Questa annulla uno dei vantaggi base della meccanografia : se la prima registrazione dei dati è esatta, essa è mantenuta in tutte le trasmissioni di questi dati nelle diverse fasi dell'elaborazione meccanografica). La collaborazione tra la sezione linguistica del C.N.U.C.E. e l'Accademia della Crusca ha già prodotto un primo volumetto,che, a quanto sappiamo, costituisce la prima applicazione della fotocomposizione nel settore in Europa. /2/.

N° 6 - Essa conta attualmente 14 collaboratori, tra ricercatori, programmatori e operatori di macchina, e si avvale anche, per le proprie ricerche, di ricercatori con borse di studio.

N° 7 - (Torino,Milano,Padova,Genova,Pavia,Bologna,Firenze,Pisa, Cagliari,Roma,Napoli,ecc.) e in altre nazioni (Romania, Cecoslovacchia, USA, Canada).

N° 8 - Rinvio alle pagine seguenti per i limiti della corrispondenza parola unità grafica : parola unità di testo.

N° 9 - Per es. nella terzina dantesca "Per me si va ne la città dolente, Per me si va ne l'etterno dolore, Per me si va tra la perduta gente" le parole o occorrenze sono 24, ma le forme grafiche sono 14 (per, me, si, va, ne, la, città,dolente, l'etterno, dolore, tra, perduta, gente)/89 p.122/.

N°10 - Di come sia interessante far precedere alla creazione del rimario la trascrizione fonetica del testo, parleremo più avanti.

N°11 - Nell'uso dei termini rango e classe di frequenza non c'è accordo generale. La definizione riportata nel testo si riferisce alla consuetudine invalsa al C.N.U.C.E. Altri usi sono reperibili.Per es. vedi De Mauro.

N°12 - Seguo in più punti letteralmente in questo paragrafo il testo di due mie precedenti comunicazioni, all'Accademia dei Licei /89/ e al Convegno L.E.L. di Pisa /338/.

N°13 - Sottolineata anche di recente al Congresso di Grenoble,23-25 Agosto 1967 "International Conference on Computational Linguistics" e vedi anche 'Computer and the Humanity'.

N° 14 - o unità lessicale, i francesi dicono unite lexicale e gli inglesi dictionary entrie.

N° 15 - Il concetto di tema è naturalmente adattato alle possibilità operative di riconoscimento da parte della macchina e non coincide perciò con il valore scientifico del termine.

N° 16 - Gli esempi che riporto in questo paragrafo, si riferiscono tutti a soluzioni alternative che di fatto ho visto proporre per la lemmatizzazione di testi italiani nelle mie collaborazioni con Enti utenti del C.N.U.C.E.

N° 17 - Nello spoglio di Veglia d'Armi di Diego Fabbri, /333 pp. 135-145/ ho compilato una lista di possibili locuzioni e, a scopo sperimentale, ho calcolato 3 diverse statistiche: a) senza alcuna considerazione per le locuzioni; b) contando solo le locuzioni considerate come lemmi autonomi nel vocabolario italiano prescelto come guida alla lemmatizzazione; c) considerando tutte le locuzioni che mi sembrano tali.

N° 18 - G. Lepsky ha chiarito storicamente i rapporti di reciprocità tra linguistica matematica e linguistica strutturale /194 p.189/ T. De Mauro afferma /78/ che il rapporto tra identificazione del carattere strutturale dei fenomeni linguistici e il processo di formalizzazione fu evidente già in Saussure. I. I Revzin /264 p.2/ mostra la necessità di impiegare il metodo della simulazione nel quadro dei metodi deduttivi in linguistica e l'importanza delle applicazioni computazionali per la teoria dei modelli. Si potrà impiegare il calcolatore per la verifica dei modelli proposti. /264 p.23/.
I modelli linguistici sono classificati in modelli "analitici"(o analizzanti), e in modelli inversi "sintetici" (o sintetizzanti), a seconda che, si parta da sequenze o frasi date (modello analitico)o che si ottengano queste sequenze in seguito a certe operazioni (modello sintetico) /264 p.7/. Nella traduzione automatica sono adoperati entrambi, il primo per la lingua di partenza (langue source), il secondo per la lingua di arrivo (langue cible)/194 p.205/ mentre in lemmatizzazione nel senso qui discusso, si utilizza essenzialmente l'analisi.
In ogni caso i modelli operano con degli elementi o dei costituenti ai quali si applicano le regole della grammatica. Al livello dell'applicazione, solo il D.M. può procurare queste unità fondamentali, con le informazioni relative e le funzioni grammaticali. Sul piano dell'analisi, per realizzare la segmentazione, la scomposizione della catena parlata o scritta con criteri formali chiari e per assegnare a ogni unità le sue "valenze"/194 p.210/ di combinabilità. Sul piano della sintesi per sostanziare con elementi appropriati le diverse terminazioni inferiori della struttura.
Nel compilare il D.M. del C.N.U.C.E., ci chiediamo se nei diversi tipi di grammatiche e di tecniche di approach al parsing, le informazioni da associare ad ogni unità del D.M. siano le stesse, o almeno convertibili con regole formali. Sul piano pratico, ciò vuol dire chiedersi se è possibile classificare le unità lessicali del D.M. senza scegliere dei modelli e delle tecniche già così definitivi da

segue Nota N° 18

escludere degli altri. La nostra esperienza non ci permette ancora di dare una risposta in base a dati di fatto; ma l'argomento è già stato trattato, con discussioni positive, al citato Congresso di Bethesda /134, Discussion, p.72/. Il problema - sembra - va affrontato in termini di una codificazione, complessa e strutturata gerarchicamente, delle informazioni grammaticali. E' possibile pensare un D.M. equivalente a dizionari diversi,come è stato fatto da Josselson /166/ per i dizionari russi, alla Wayne State University, ed è questa la soluzione verso cui ci stiamo orientando.

N° 19 - Potremmo rilevare che, a 20 anni di distanza, questa affermazione appare ottimistica: il problema della identificazione delle unità linguistiche ai diversi livelli fonematico, lessicale, sintattico, e addirittura della esistenza di alcune è ancora oggi una delle difficoltà contro cui si dibatte la stat. ling. Ci importa invece qui osservare che l'affermazione del Reed, quasi banale nella sua semplicità, concorda col fatto già rilevato che l'interesse per la stat.ling. è stato naturalmente vivo fin dal principio presso le correnti strutturalistiche dei praghensi e dei loro continuatori, come per es. A. Martinet, i quali si propongono la rigorosa identificazione e descrizione delle unità linguistiche e delle opposizioni distinte all'interno di un sistema strutturato. Sono esemplari a questo proposito le pagine di N. Troubetzkoy nei Principes e gli studi di V. Mathesius.

N° 20 - Per l'inglese sono riportati 54 titoli; tra essi sono molto conosciuti i lavori di Thorndike /307/ di Palmer /243/ di Ogden /236/. Il tedesco ha 13 titoli, fra i quali, oltre il noto Kaeding, ricordo Purin /257/ e Muegge /223/; per il francese 12 titoli, sono noti il "French Word Book" di Hemmon /143/, e sopratutto il "French Word Book" di George E.Vander Beke, che è basato sullo spoglio di 88 testi di 1300 parole ciascuno, per un totale complessivo di 1.200.000 parole /322/.
Per lo spagnolo la bibliografia del Guiraud cita 4 titoli, tra i quali sono rivelanti i lavori di M.A. Buchanan, basato sullo spoglio di 1.200.000 parole /37/ e di H.A. Kennston /177/.

N° 21- Il CREDIF ha proposto due tappe nell'acquisizione del vocabolario e della grammatica francese. Le premier degré du francais fondamental, è dedicato sopratutto " à l'adulte débutant ", cioè a chi desidera acquistare rapidamente i primi "moyens d'expression, simples sans doute, mais réels et solides" /177 p.56/.
"Le deuxième degré", risponde invece ai bisogni "des personnes qui, déjà initiées à la pratique orale du français, sont désireuses d'acquérir le moyen d'exprimer des notions plus étendues et plus variées, et d'aborder la lecture des journaux et des livres" /117 p.65/.
Dopo la tappa del 2° degré, resterebbero le difficoltà degli studi e delle professioni che reclamano dei vocabolari specializzati. Sono perciò in preparazione dei vocabolari tecnici, dei lessici specializzati per ciascuna disciplina, e in più un vocabolario generale di orientamento scientifi-

segue Nota N° 21

co, che riunirà le parole essenziali comuni alle diverse scienze. /117 p.72/.
I dizionari di frequenza citati venivano per lo più stabiliti sullo spoglio di opere letterarie e sopratutto di testi studiati nell'insegnamento secondario, in coerenza con la concezione dell'insegnamento delle lingue viventi che predominava verso il 1930 negli Stati Uniti: si trattava di mettere l'allievo in condizioni di leggere dei libri in una lingua straniera e non di parlare questa lingua. La conoscenza delle parole più frequenti risparmiava molte consultazioni del dizionario /117 p.57/.
Gli autori del 1° degré, desiderando prendere per base "la lingua parlata autentica (ibidem), hanno registrato 163 conversazioni dal vivo, i cui interlocutori appartenevano ai "milieux" più diversi, e hanno spogliate le 312.000 parole raccolte, ottenendo 8.000 lemmi che hanno disposto in ordine di frequenza decrescente. Essi hanno constatato, anche per la lingua parlata, che le frequenze delle prime parole nella lista sono considerevoli. Le prime 38 parole coprono già il 50% del testo. Ma, secondo un recente articolo di R.Michéa /214/, la struttura quantitativa della lingua parlata sembra essere diversa da quella approssimativamente verificata, nei testi letterari, dalla legge di Zipf: le parole di alta frequenza sarebbero più frequenti nella lingua parlata: le prime 50 parole coprono nel"français fondamental" il 56,02%, mentre le prime 50 della lista del Dewey /83/, scelta come termine di confronto per i testi scritti, coprono meno del 47%. Questa differenza andrebbe progressivamente riducendosi, e verso la 220/ma parola, la percentuale di testo coperta da una data "tranche" della lista comincia a divenire più alta presso Dewey che nel francese fondamentale. La spiegazione di queste variazioni, secondo l'autore, è da ricercare nel fatto che, nella lingua parlata, i termini più frequenti si ripetono più spesso che nella lingua scritta, dove ci si sforza di evitare ripetizioni troppo vicine.
Dall'analisi della lista di frequenza risulta che nelle prime 100 parole, al di fuori delle parole grammaticali, appaiono solo 4 verbi di carattere generale, il cui uso "risulta più o meno da vincoli linguistici (avoir, être, voir, dire, faire, ect.) /214 p.12/, qualche aggettivo e solamente 3 nomi, i quali appaiono con un rango elevato: heure (r.82), jour (r.84) e chose (r.88).
Insomma alla evidenza della frequenza, sfugge la "plupart de noms concrèts" /117 p.59/. Essi non hanno, in generale, che una debole frequenza, e per di più, come ha dimostrato R. Michéa, instabile, fortemente legata al contenuto dei testi : "il ne faut en effet jamais perdre de vue que l'on ne peut tirer de dépouillements statistiques que ce que l'on y a mis préalablement" /117 p.60/. E' certo che una parola frequentemente pronunciata o frequentemente scritta è utile a conoscersi. "On peut donc poser : 'Tout mot fréquent est utile'. Le parologisme consiste à renverser les termes et à dire 'Tout mot utile est fréquent'. Et même : 'Seul les mots fréquents sont utiles' /.../ L'opinion commune /..../ confond sous la dénomination d' 'usuels' tous les mots qui lui semblent utiles à connaître

segue Nota N° 21

et /..../ elle identifie 'usuel' et 'fréquent' /..../. Nous distinguerons donc, parmi les mots 'usuels', les mots fréquents et les mots utiles non fréquents"./117 p.61/
Ma come può essere giustifcata questa ultima categoria,dato il carattere negativo dell'assenza di frequenza?
Gli autori hanno deciso di porre l'uno di faccia all'altro due categorie di vocabolari : le vocabulaire de la fréquence, che comprende essenzialmente le parole grammaticali e i verbi, e accessoriamente degli aggettivi e qualche nome di carattere generale, e le vocabulaire de la disponibilité,costituito essenzialmente da nomi concreti, sussidiariamente da aggettivi e da verbi che si riferiscono a nozioni concrete particolari. Ai nomi frequenti si oppongono i nomi disponibili.
"Ceux-ci sont à notre disposition dans notre esprit. Mais nous ne les prononçons ou les écrivons effectivement que lorsque nous y sommens amenés par les circostances"./117 p.61/.
Per esempio noi pronunceremo nomi di strumenti o di vestiti solo nelle circostanze particolari di un accidente o un acquisto. Tale distinzione sarebbe confermata dalla medicina la quale constata che, nell'afasia progressiva, i nomi propri e i nomi concreti sono i primi a scomparire, e anche dall'osservazione stessa della vita quotidiana la quale mostra che le parola "qui nous échappent" appartengono a queste due categorie. Esistono dei mezzi scientifici per valutare il grado di disponibilità? "Les mots disponibles étant liés aux circostances, on ne peut mesurer de disponibilité absolue /...../ Les degrés de disponibilité ne peuvent être déterminés que par rapport à un intérêt".
Gli autori hanno prospettato 16 centri di interesse,parti del corpo, vestiti, abitazioni, ecc., chiedendo agli allievi delle scuole primarie di 4 dipartimenti, lontani geograficamente per eliminare le influenze locali, di scrivere i 20 nomi che sembravano loro più utili per ciascun centro di interesse. I nomi concreti forniti da questa inchiesta sono stati aggiunti alle 800 parole ritenute dall'esame delle frequenze dei testi registrati al magnetofono, e il "francais élemtaire" conta in tutto 1400 lemmi.
Il francese fondamentale di 'deuxième degré' dovrebbe, come s'è detto, rispondere invece ai bisogni di persone che, già iniziate alla pratica orale del francese della vita quotidiana, desiderano acquisire il mezzo per esprimere nozioni più estese e più variate e per affrontare la lettura di libri e giornali.
Furono spogliati 425 brani di 500 parole ciascuno, estratti da articoli di giornali e periodici scelti in 14 rubriche (Arts et lettres, La politique, La Vie sociale, L'économie, Les sciences et techniques, ecc.). Per ciascuna rubrica fu preso un diverso numero di brani, proporzionale alla importanza attribuita dagli autori alla rubrica nella composizione dello strato di lingua che si voleva rappresentare. Furono tralasciati nello spoglio le parole grammaticali in genere e tutte quelle registrate per il 1° degré. Si vide che" La langue des journaux d'aujourd'hui est (.....) plus abstraite que littéraire". /177 p.70/.

segue Nota N° 21

Essa fornisce gli strumenti necessari all'espressione di un giudizio, di un pensiero, ma sarebbe povera sul piano affettivo, estetico.
Per questo settore del vocabolario (nomi designanti i sentimenti, verbi esprimenti i sentimenti e le loro manifestazioni, ecc.) gli autori posero 11 domande dettagliate a 160 allievi delle Ecoles Normales d'Instituteurs.
Ulteriori arricchimenti al vocabolario sul piano culturale vennero da altre fonti, con spoglio di opere specializzate, in particolare per la terminologia della attività amministrativa e civica.

N° 22- Entrambe queste ricerche si rivolgono, per le operazioni meccanografiche dello spoglio, al Centre d'étude du vocabulaire Francais di Besançon. /258/ e /269.

N° 23- Ricordo le considerazioni di S.E.M. Boldrini a questo proposito, nel corso di una visita al CAAL di Gallarate nel 1963: egli affermò chiaramente l'importanza di potenziare gli strumenti per la diffusione dell'italiano, e tra essi poneva un dizionario di frequenza.

N° 24- Per es. Thordike ha usato le fonti più svariate, dalla Bibbia e dalle opere drammatiche di Shakespeare, fino alla corrispondenza privata e ai libri di testo, senza dichiarare le ragioni della scelta /307 p.224/. Più spesso invece ci si preoccupa di definire più precisamente almeno i limiti cronologici: Vander Beke /322/ ha spogliato 33 romanzi, 13 drammi teatrali, 14 tra giornali e riviste,13 opere scientifiche e filosofiche, 16 opere storiche e critiche, che vanno da Balzac a Musset fino a Proust e Bergson; P.Guiraud /128 p.63/ li ritiene rappresentativi della prosa letteraria dalla fine del XIX° al principio del XX° secolo, e pensa che possano costituire un termine di paragone conveniente per lo studio stilistico del vocabolario poetico dell'epoca.

N° 25- come quelle dei più che 7 milioni di parole esaminate da Rodriguez Bou /273/; esse si articolano in 3 aree principali, ciascuna suddivisa in gruppi minori:

1) - Vocabulario de expresión (circa 3.400.000 parole di testo), consistente in : a) vocabolario parlato (circa 1.073.000 parole); b) vocabolario di associazione libera e controllata (circa 926.000 parole); c) vocabolario di composizioni di scuole medie e superiori (circa 586.000 parole).
2) - Vocabulario de reconocimiento (circa 3.011.000 parole) a) giornali (1.050.000 parole); b) programmi radio (465.000 parole) ; c) letteratura religiosa (295.000 parole); d) spoglio di Buchanan (1.200.000 parole).
3) - Vocabulario de autores (circa 666.000 parole) : a) libri di testo (circa 255.000 parole); b) letture supplementari (circa 410.000 parole);

Si è già detto della composizione del campione scelta dal CREDIF: conversazioni registrate in diversi ambienti sociali per il vocabolario delle frequenze; inchieste sui sostantivi e sui verbi disponibili per diversi centri di interesse; spoglio di giornali per il vocabolario letterario e di opere specializzate per il vocabolario scientifico.
Buchanan /37/ divide il suo campione di circa 1.200.000 pa-

segue Nota N° 25

role in 7 zone, come segue :

1) - Drammi (7 unità)
2) - Novelle (8 unità)
3) - Poesie (3 unità)
4) - Folklore (3 unità)
5) - Prose varie (8 unità)
6) - Tecnica (7 unità)
7) - Periodici (4 unità)

L'estesa spiegazione di Garcia Hoz per la ripartizione delle sue fonti in 4 categorie (corrispondenza privata, periodici, documenti ufficiali-politici,religiosi,sindacali-, libri vari), ci sembra un buon esempio per illustrare la soggettività che informa generalmente (e come vedremo, non potrebbe essere altrimenti) le scelte dei testi.

Egli dice di voler tener conto delle manifestazioni fondamentali della vita umana, o, più precisamente, delle relazioni umane, dal momento che il linguaggio "es principalmente un medio de relación" /153 p.18/. Quattro aspetti della vita gli paiono chiaramente differenziabili, perchè sono diversi in se stessi e perchè possiedono un proprio mezzo di espressione: /153 p.18-19/.

1) Vida familiar, rappresentata nel suo campione dalla corrispondenza privata: 620 lettere, raccolte in 15 diverse province spagnole tra gente di non elevata condizione.
2) Vida social indiferenciada, dell'uomo della strada, che vive le quotidiane preoccupazioni e conquiste sociali. La rappresentano articoli di 25 periodici editi in diverse regioni. Gli articoli scelti in modo da coprire l'intero arco dei mesi dell'anno, sono ripartiti in modo da rispecchiare quantitativamente la complessa composizione dei periodici. Delle 4.000 parole prese per ogni giornale, 500 provengono da articoli dottrinali, 2.000 dalla cronaca, 1.000 dagli spettacoli, 500 dagli annunci.
3) Vida social regulada, dell'uomo in quanto membro della società organizzata. La prima società che sembra regolare la vita dell'uomo è lo stato. "Pero no es sólo el Estado la sociedad que protege o apresa al individuo; la vida de este resulta incompleta y sin sentido cuando no tiene trascendencia sobrenatural. En España, tal vez como en ningun otro país, nos econtramos al hombre des envolviendo su vida religiosa en el marco de una nueva sociedad : La Iglesia. Por último, ha de subrayarse el hecho de que, en la actualidad, la vida profesional se desenvuelve regulada por los sindicatos; y no hay or qué considerar que lo sindicatos, en su forma actual, tengan o no garantías de continuidad, sean de una u otra forma las organizaciones, el hecho sindical tiene alcance universal. De aquí el considerar que la vida social regulada tiene su expresión en le Boletin Oficial del Estado, en los Boletines Eclesiásticos y en las publicaciones sindicales". Queste tre categorie di documenti sono ugualmente rappresentate : 33.400 parole per ciascuno.

segue Nota N° 25

4) Vida cultural, in quanto patrimonio culturale cui partecipa l'uomo comune. Il criterio per la scelta dei 22 libri non è il valore intrinseco, ma la diffusione, appurata dalle statistiche editoriali. Inoltre le parole di un libro non sono prese da un unico blocco di pagine consecutive, ma da pagine singole distribuite a intervalli regolari nel testo.

N° 26- Come è possibile accettare, senza mettere in discussione la validità stessa del concetto di dizionario di frequenza, che per una stessa lingua si senta il bisogno di nuovi dizionari quando già esistono dizionari pressochè contemporanei basati su spogli estremamente ampii, di parecchi milioni di parole?

N° 27- Il primo problema è la definizione dell'universo o popolazione statistica. Esso è comune, tra l'altro all'applicazione di criteri quantitativi nelle ricerche stilistiche. "Le style - ha detto P.Guiraud - est un écart qui se définit quantitativament par rapport à une norme" /129p.19/. Ne risulta che lo studio dello stile, lo studio degli scarti che costituiscono lo stile "postule l'établissement d'une norme". /128 p.5/. Se dunque si vogliono mettere alla prova i metodi quantitativi nello studio stilistico di un'opera - commenta Ch.Muller /226 p.5/ - il primo compito, e il più decisivo, sarà la scelta della norma.
"L'oeuvre littéraire est en effet à l'intersection d'ensembles divers dont les limites ne peuvent être déterminées que par convention". (ivi) Supponiamo di voler definire lo stile del Cid : lo faremo con riferimento alle tragedie precedenti di Corneille, o all'insieme del suo teatro, o alle sole sue tragedie, o alla totalità della sua opera pubblicata? Oppure, ancora, in rapporto a tragedie di autori contemporanei, o, ancora, a un corpus di testi diversi degli anni 1630-1640 ? Tutte queste soluzioni e altre ancora possono essere sostenute; "toutes sont arbitraires et discutables" (ivi). Charles Muller preferisce, nella fase attuale degli studi di stilistica quantitativa, ricerche che, prendendo come punto di partenza un insieme ben definito di una unità incontestabile, si dirigano verso le parti di questo insieme, egualmente coerenti e ben definite. La norma è allora la struttura dell'insieme e gli scarti appaiono nei sotto-insiemi. Alle tecniche statistiche spetta dire se sono significativi, e alla filologia di decidere in che misura corrispondano a fatti di stile.

N° 28- E' questo il caso, per esempio, dei sondaggi d'opinione prima delle elezioni. Per gli iscritti alle liste elettorali, si conoscono con precisione la ripartizione geografica, socio-professionale, per età, per sesso, ecc.

N° 29- E' questo il caso più semplice, e si definisce come l'insieme dei vocaboli che l'individuo in questione utilizza o potrebbe utilizzare nel discorso.
Il vocabolario di uno dei testi prodotti dall'individuo, è una parte del suo lessico, un campione. Per avere identità tra di essi, occorrerebbe che l'individuo non conoscesse altri vocaboli oltre quelli che ha utilizzati nel testo, in altri termini che egli avesse utilizzato nel testo tutto il suo lessico virtuale. E' certo che questa condizione non si rea-

segue Nota 29

lizza nell'esercizio normale del linguaggio. A un dato momento della esistenza, il lessico di un individuo ingloba, oltre ai vocaboli che si possono osservare nei suoi discorsi, dei lessemi virtuali (Muller intende con questo termine quelle nuove unità che grazie alle "possibilités de formations léxicales libres" dell'idioma, l'individuo può creare senza averle mai prima percepite, ritenendo che l'interlocutore le interpreterà immediatamente anche se sono nuove per lui) e dei lessemi passivi. Questi sono connessi alla nozione di disponibilità, che abbiamo visto utilizzata dal CREDIF. Alcuni lessemi, per quanto perfettamente conosciuti e compresi dal soggetto, hanno una probabilità di impiego minima, che si traduce in frequenze molto piccole o addirittura nell'assenza dei vocaboli corrispondenti. Ci sono inoltre dei lessemi che il soggetto utilizza intanto che uditore o lettore, ma che non impiega mai in quanto locutore; possono essere almeno parzialmente provocati con tecniche sperimentali (come quelle delle inchieste dialettali o quelle delle inchieste per i vocabolari fondamentali connesse alla nozione di disponibilità), se il soggetto è vivente, ma, presso uno scrittore del passato, sono per definizioni inconoscibili.

N° 30- 31 Si aggiunge inoltre la nozione di lessico di situazione (lexique de situation). Essa ingloba due tipi di elementi; i primi sono d'ordine stilistico legati all'interlocutore e all'effetto che il locutore vuole produrre; i secondi sono di ordine tematico, legati a ciò che il locutore vuole comunicare, al contenuto del messaggio. La situazione provoca, all'interno del lessico individuale "un déplacèment des probabilités d'emploi", certe unità sono escluse dal discorso, e cadono a probabilità 0 . Si può ammettere cioè che al momento della redazione di un testo e della concezione di un enunciato l'autore abbia un certo numero di unità del suo lessico 'in gioco', mentre le altre sono escluse per motivi stilistici e tematici.

N° 32- Sia un corpus di 10.000 parole, un romanzo poliziesco per esempio. Ammettiamo che la parola gorilla vi appaia 100 volte, dunque con una frequenza relativa di 100/10.000 = = 1/100. Possiamo dedurre che in un brano qualsiasi di 1.000 parole consecutive, (la cui prima parola cioè è estratta a caso) la parola gorilla apparirà circa 10 volte? Ciò sarà vero solo se ammettiamo che le 100 occorrenze di gorilla siano uniformemente ripartite nel romanzo, con un intervallo costante a partire da una certa parola (la parola gorilla sarebbe per es. la 50/ma la 150/ma, la 250/ma ecc.). Non sarebbe però vero se ammettiamo, invece, che, per esempio, nelle prime 9.000 parole del romanzo l'autore non abbia mai utilizzato la parola gorilla, e solamente all'ultimo brano (tranche) di 1.000 parole la si incontri 100 volte. (Se questo fatto rispondesse alle esigenze dell'argomento trattato, la parola gorilla sarebbe chiamata, nella concezione di Moreau "thématique".

N° 33- Ritorniamo al nostro romanzo poliziesco e alla presenza in fine di questo romanzo della parola gorilla. Prendiamo

segue nota N° 33

un altro romanzo che non contiene la parola gorilla. Il primo romanzo ha N = 10.000 occorrenze e il secondo N'=0 occorrenze.
Il corpus formato dalla addizione dei due romanzi ha N + N' parole. Il numero relativo di occorrenze della parola gorilla è allora 100/N + N' e non è più 100/N. E' dunque cambiato. Conoscere la frequenza del nome gorilla nel primo romanzo non ci insegna nulla sulla frequenza della stessa parola nell'insieme dei due romanzi. Essi non sono omogenei rispetto alla parola gorilla. Aggiungiamo allora un terzo romanzo poliziesco che contiene una volta la parola gorilla. La frequenza relativa precedente non è più quella dell'insieme formato dai tre romanzi. Aggiungiamo via via altri testi di vario genere: la frequenza relativa della parola gorilla è ogni volta diversa. Questa frequenza relativa non ci insegnerà nulla su quella della parola nella lingua scritta, se noi non consideriamo la quasi totalità delle pubblicazioni di questa lingua, ciò che è impossibile, come abbiamo visto. Il procedimento è dunque cattivo.

N° 34- Questa presentazione permette d'altronde di stimare, per mezzo di una semplice proporzione, la frequenza della parola nella lingua.

N° 35- (il quale come si è detto, sta per essere pubblicato a cura di U. Bortolini, C.Tagliavini, A.Zampolli).

N° 36- Per i motivi esposti, altri progetti di dizionari di frequenza annunciati per l'italiano, sono da riguardare come iniziative complementari (e non concorrenti) alla nostra, ed anzi sembra più che mai doveroso assicurare la comparabilità dei dati, attraverso la comparabilità delle metodologie e delle modalità di registrazione.

N° 37- abbiamo già visto come lo spoglio dei testi abbia fornito solo una parte del lessico accettato nel francese fondamentale. Qui ci interessa solo riassumere il procedimento con il quale sono stati utilizzati i dati provenienti dai 163 testi disponibili, lunghi, in media, 2.000 occorrenze ciascuno .

N° 38- E' rilevante il parallelismo con la distinzione francese del lessico passivo (Muller) o disponibile (Moreau,CREDIF). Del resto questa dicotomia si trova anche in J.Casares,in un discorso del 1920 /47/, al quale G.Hoz si rifà esplicitamente.

N° 39-40 La frequenza media dei lemmi del vocabolario usuale è di 31, nel comune di 172, nel fondamentale di 1324. Ma se si considerano i lemmi che appaiono rispettivamente (a) solo nel vocabolario usuale, (b) solo nel vocabolario comune e il vocabolario fondamentale, (c) solo nel vocabolario fondamentale, le frequenze medie rispettive sono (a)4,2 (b)52 e (c)1324. La ricerca, egli conclude, ha reso evidente che il vocabolario usuale, per lo meno quello scritto, non è tutto omogeneo nei diversi sottoinsiemi (en las distintas manifestaciones de la vida humana), ma che vi è un piccolo numero (200 lemmi) indifferenziato, che è presente in tutti gli aspetti. Altre 2.000 parole sono presenti in tutte le manifestazioni, ma non in modo indifferenziato, sibbene cariche di un valore specifico che le rende più rappresentative di un

segue nota N° 39-40

aspetto sociale che di un altro. Se ammettiamo che in una conversazione si pronunciano 160-170 parole al minuto, 10.000 all'ora, le parole del vocabolario fondamentale avrebbero,in media,33 ripetizioni ciascuna; quelle del vocabolario comune tra una o due volte, quelle dell'usuale non hanno praticamente probabilità di apparire (0,10$\overline{5}$): in altri termini le parole del vocabolario fondamentale appaiono una volta ogni due minuti, quelle del comune una volta all'ora, quelle dell'usuale una ogni 10 ore. (Il metodo di Garcia Hoz è discusso ampiamente da Juilland. /169 pp.XLVII e segg./.

N° 41- Gli autori hanno preferito un'altra formula, che equipara la dispersione alla deviazione standard da una media stabilita. Nelle loro intenzioni, questa formula rappresenta essenzialmente la deviazione totale delle occorrenze di una parola dalla dispersione ideale ($\overline{x}$) rappresentata dal totale delle occorrenze (F) diviso per il numero dei sottoinsiemi.

N° 42- Le concordanze del corpus saranno indubbiamente utili come documentazione per lo studio di altri livelli di lingua, per esempio per la sintassi dell'italiano scritto.

N° 43- Alcuni, cui abbiamo mostrato questi risultati, hanno tentato di qualificare l'allineamento dei 3 gruppi in relazione a una crescente distanza dalla lingua parlata,dalla lingua "comune",dal dialogo. Per ora, questa è solo appunto una 'tentazione' : nuovi dati apporterà lo spoglio di un insieme di conversazione registrate dal vivo nella zona di Pisa.

N° 44- Per la dimostrazione si veda Muller /227 p.37 nota 8/ e Torrigiani /310 p.16-17/.

N° 45- In effetti, nel dizionario dello spagnolo di Juilland, compaiono anche lemmi con R = 2, e non so quale sia la spiegazione di questo fatto, contrastante con le dichiarazioni dell'autore: per es. arollador a pag. 35,col.a.

TABLE I

The Scope of Linguistic Data Processing and of Mechanolinguistics

PROCESSING	Of Linguistic Data: With Linguistic Processes	Of Linguistic Data: With Non-linguistic Processes	Of Non-linguistic Data
For Linguistic Purposes	1 $LDP_{1,2,3}$ ML	2 $LDP_{1,3}$ ML	3 LDP_3 ML
For Non-linguistic Purposes	4 $LDP_{2,3}$ ML	5 LDP_1	6

Examples:
1. Research on automatic parsing as an approach to the study of linguistic structure.
2. Statistical studies of texts to throw light on linguistic structure.
3. Simulation of language-area dynamics.
4. Machine translation of Russian ethnographic literature (for anthropologists).
5. Construction and use of an automatic citation index for ethnography.
6. Calculating orbits of earth satellites.

TABLE II

Areas Within the Field of Linguistic Data Processing

1 Area		2 Involves Type Of:	3 Primary Interest In:	4 Linguistic Process	5 Linguistic Data	6 Linguistic Use	7 Extra-Linguistic Use
Simple Text Processing		Process Data	Results	(x)	X	X	X
Automatic Linguistic Analysis	Synchronic	Objective Use Data	Process and Results	X	X	X	
	Diachronic	Objective Use	Process and Results	X	X	X	X
Linguistic Automation		Objective	Process and Results	X	X	X	X
Processing of Linguistic Files		Data Use	Results		X	X	
Simulation of Language Dynamics	Of Linguistic Change	Process Use Objective	Process	X	X	X	
	Of Language-Area Dynamics	Process Use Objective	Process			X	X
Knowledge Retrieval		Use	Results	(x)	X	X	X
Statistical Studies of Verbal Data		Process	Results		X	X	X

Fig.3 Da [138] pag. 151

Fig.4, da 138, pag. 156

Annèe del l'enquete	Désignation du mot	Centres d'interet ou' il a eté recontre	Collectif relatif de ces centres d'interet par rapport a vocabulaire total	Frequence relative du mot mesurèe dans le centre d'intèret
1961	Officier	Justice	x%	a %
		Police	y%	b %
		Armée	z%	c %

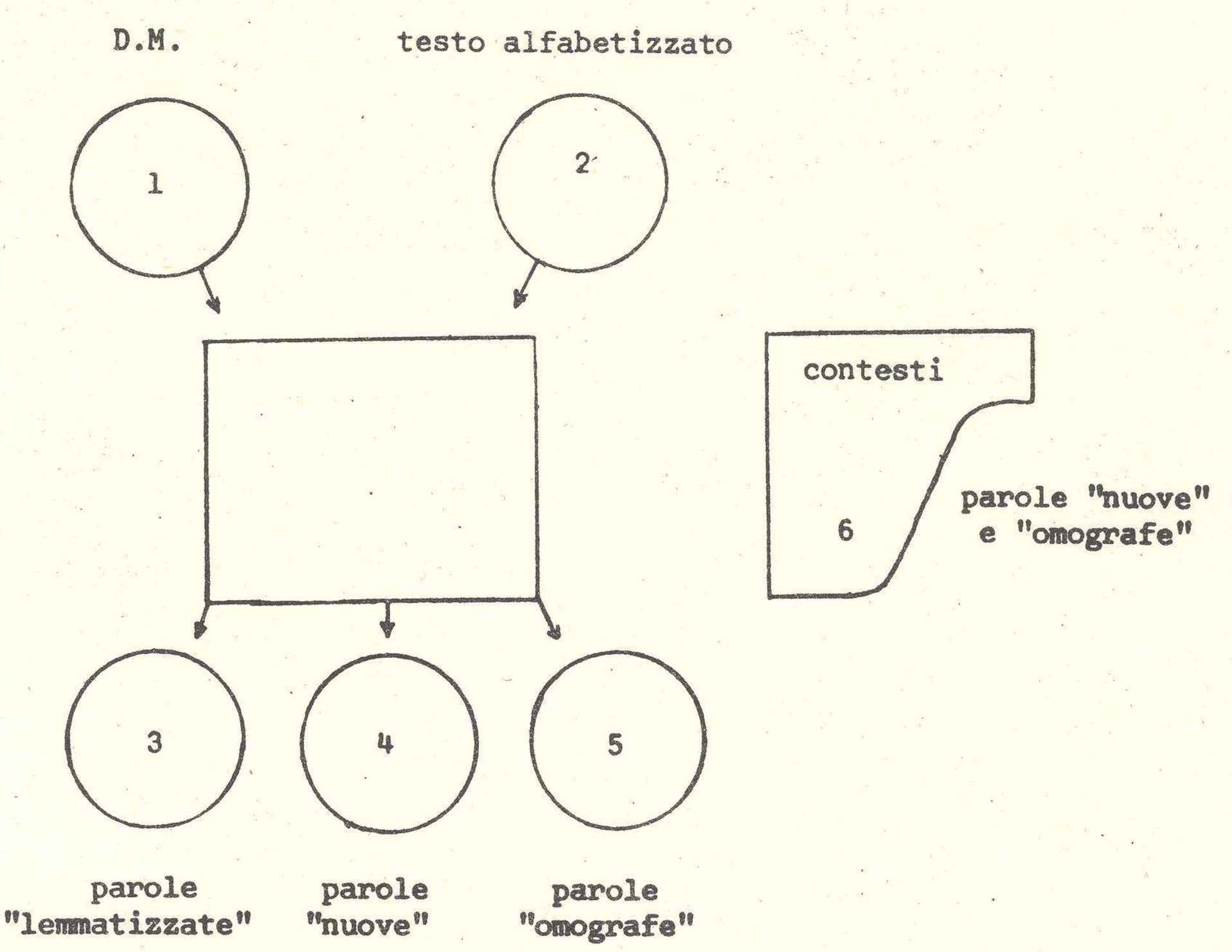


Fig. 5, da 310 / p. 9

BIBLIOGRAFIA.

1. Abraham,S., Kiefer F.,"An algorithmic definition of the morpheme", Statistical Methods in Linguistics, IV, 1965, pp.4-9.

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

2a Accademia della Crusca,Norme per la schedatura lessicografica, Firenze 1967.

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

4. Actes du premier Colloque international de Linguistique Appliquée.(1964), Nancy, 1966.

5. Actes du Séminaire International sur le Dictionnaire Latin de machine, (Pisa, CNUCE,1967) Calcolo,V suppl. al n.2 ,1968.

6. Afendres,E.A., "The Balkan Linguistic Area:A Study in Phonological Convergence", ICCL .

7. Akhmanova,O.S.,Mel'chuk,I.A., Frumkina,R.M.,Paducheve, E.V., Exact Methods in Linguistic Research, Berkeley and Los Angeles, 1963, (trad.dal russo di D.G.Hays e Dolores V.Mohr).

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

9. Aliprandi , G., "Frequenze dattilografiche", Bollettino dell'Accademia Italiana di studi stenografici di Padova XV, 1940, p.273.

10. Almanacco letterario Bompiani, Milano, 1962

11. Andreyev, N.D., "The Models as a Tool in the Development of Linguistic Theory", Word, XVIII (1962)pp.187-197.

12. An Introduction to COMIT Programming, the Research Laboratory of Electronics and Computation Center,M.I.T., 1961.

13. Apresjan,J., "Analyse distributionnelle des significations et champs sémantiques structurés",(trad.dal russo di Yvan Mignot) Langages I,1967,pp.44-74.

14. Arcaini,E., Principi di linguistica applicata, Bologna, 1967.

15. Arnold,E.V., Vedic Metre in its Historical Development, Cambridge,1905.

16. Atti del Congresso, Maschinelle Methoden der leteraryschen Analyse und der Lexicographie, Tubingen, (1960).

17. Automatic Language Processing Advisory Committee. Language and machines computers in translation and linguistics. National Academy of Sciences-National Research Council, Washington, 1966.

18. Automatic Typographic-Quality Typesetting Techniques: A State of-the-art Review, National Bureau of Standar,Washington, D.C., 1967.

19. Baldinger, K., "Sémantique et structure conceptuelle", Cahiers de Lexicologie,8,1966 pp.3-46.

19a Bakker,I.I.M., "Frequency in usage and in the Lexicon", Lingua, 21, 1968, pp.13-22.

20. Bar Hiller, Y., "A Quasi-arithmetical Notation for Syntactic Description" Language, 29, 1953 pp. 47-58.

21. Bar-Hillel,Y., Four Lectures on Algebraic Linguistics and Machine Translation (versione riveduta di una serie di 4 lezioni tenute nel Luglio 1962, al 'NATO advanced Summer Institute on Automatic Translation of Languages', Venezia, incluse nel volume Language and Information, Gerusalemme, 1964 pp.185-218).

22. Bar-Hillel,Y., "Theoretical aspects of the mechanization of Literature Searching, ibidem pp.330-364.

23. Battisti,C., "Orientamenti generali della Linguistica in Italia", 1930-1960 in Trends in European and American Linguistics, Utrecht, 1963, pp.240-282.(Mohrmann,C., Sommerfelt, A., Watmough, J.edit.).

24. Belevitch, V., Langage des machines et langage humain, Bruxelles, 1956.

25. Bibliographie Linguistique,publiée par le Comité International Permanent des Linguistes, Sprectrum,Utrecht - Anvers.

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

27. Birbanescu,A.: "Networks of Binary Relations in Natural Languages. Electrical Analogues". ICCL

28. Bloomfield, L., Language, New York, 1933.

29. Bobrow, D.G., Natural Language Input for a Computer Problem-Solving System, in Minsky,M.edit.(1968) pp.135-215.

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

31. Boldrini, M., Discorso al CAAL di Gallarate, 1964.

32. Bolelli, T., Per una storia della ricerca linguistica, Testi e note introduttive, Napoli, 1965.

33. Bolocan, GH., "Unele caracteristici ale stilului publicistic al limbii române literare", Studi si cercetàri linguistice, XII,1,1961.

34. Bortolini, U., Tagliavini, C., Zampolli, A., Dizionario di Frequenza delle parole italiane, in pubblicazione.

35. Braun, St., "Some Formal Properties of Phonological, Redundancy Rules", ICCL

36. Brown,A.F.R., The 'SLC' Programming Language and System for MT, Washington, D.C.,1962.

37. Buchanan, A., Graded Spanish Word Book, Toronto,1927.

38. Bujas, Z.; "Computers in Serbocroat-English Contrastive Analysis Project", ICCL

39. Bulletin d'information du laboratoire d'analyse lexicologique, Besançon.

40. Bünting, K.D.: "Empirical Investigation of German Word Derivation with the Aid of a Computer", ICCL

41. Busa, R.,"Les Travaux du Centro per l'Automazione del=l'analisi letteraria di Gallarate",Cahiers de Lexicologie III, Parigi, 1962, pp.64-68.

42. Busa,R.,"L'analyse des homographes et des formes flexionelles", Cahiers de Lexicologie, 3,1962,pp.134-139.

43. Busa, R., De lexico electronico latino, in "Acta omnia Gentium ac Nationum Conventus Latinis Litteris",Roma, 1968 pp.385-410.

44. Busa, R., Traitement de mots d'une fréquence extrême,in "Les machines dans la linguistique",Praga,1968,pp.137-142.

45. Cahiers de lexicologie, Parigi

46. Calculi, Hanover (N.Hampsire).

47. Casares, J., Nuevo concepto del Diccionario de la Lengua, Madrid, 1921.

48. Cherry,E.C., Halle,M. e Jakobson,R., "Toward the Logical Description of Languages in Their Phonemic Aspects", Language, Vol.29, 1953, pp.34-46.

49. Chirvai, L., Hales E., Marculetii, S., Soltyscki,S.,"Observations sur la grammaire mécanique de la langue roumaine", Cahiers de linguistique théorique et appliquée,IV, 1967,pp.47-54.

50. Chomsky,N.,"Three Models for the Description of Language", IRE Trans, of Information Theory, Vol.IT-2,1966,pp.113-124.

51. Chomsky, N., Syntactic Structures, L'Aia, 1957.

52. Chomsky, N., "Recensione di: Beleviych, V., Langage des machines et langage humain, Bruxelles, 1956", Language, XXXIV,1,1958, pp.99-105.

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

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

55. Chomsky, N., A Transformational Approach to Syntax, in "Proceedings of the Third Texas Conference on Problems of Linguistic Analysis in English", Austin, 1962 pp.124-158.

56. Chomsky, N., Formal Properties of Grammars in Luce,R.D., Bush,R.R. e Galanter,E. ed.Handbook of Mathematical Psy= cology, New York, Vol.2,pp.323-418 (1963)

57. Chomsky, N., Current Issues in Linguistic Theory, L'Aia, 1964.

58. Chomsky, N., Aspects of the Theory of Syntax,Cambridge (Mass.), 1965.

59. Chomsky, N., Introduction, in Gross,M. e Lentin, A., Notions sur les grammaires formelles, Paris,1967,pp.7-8.

60. Chomsky, N., e Halle, M., The Sound Pattern of English, New York, 1968.

61. Chomsky, N., A Transformational Approach to Syntax,pre= sentato nel 1958,traduzione italiana in: Chomsky,N., L'analisi formale del linguaggio, Vol.I,Torino,1969, pp.53-105.

62. Chrétien, C.D.,"The mathematical Models of Glottochro= nology", Language, Vol. 38,1,1962,pp.11-37.

63. Computer in psychological research, Proceedings of IBM Symposium, (1964), Parigi, 1966.

64. Computer and the Humanities, New York.

65. CNRS-CETA, Description Sommaire du Système de diction= naire automatique, Etude Préliminaire à la Traduction Automatique, note N° 8,Parigi,1961.

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

67. Colombo, C., I principali centri europei di lessico= grafia meccanizzata, Milano 1967.

68. 2/ème Conférence Internationale sur le traitement auto= matique des langues, Grenoble, 1967.

69. Coseriu, E. "Structure lexicale et enseignement du vo= cabulaire",Actes du 1/er Colloque International de Linguistique Appliquée (1964) Nancy,1966,pp.175-217.

70. Coseriu, E., "La struttura del Lessico", La Grammatica e la lessicologia, Roma 1969, pp.55-72

68a Contini; G., "Esperienze di un antologista del duecento poetico italiano",Studi e problemi di critica testuale, 1961, pp. 241-272.

71. Cirese, A.M., Note per una nuova indagine sugli stram= botti delle origine romanze, della società quattrocen= tesca e della tradizione moderna, in "Giornale Storico della Letteratura Italiana", CXVIV,fasc.445 pp.1-45; fasc.448 pp.491-566.

72. Davies, D.W., The organisation of a Russian - English Stem Dictionary on Magnetic Tape (mimeo).

73. Delatte, P., Evrard, E., Bodson, A.,"L.A.S.L.A.: Docu= ments",Calcolo, V, supp. N°2 (1968) pp.11-39.

74. Delavenay,E. e Delavenay,K., Bibliographie de la tradu= ction automatique, L'Aia, 1960.

75. De Mauro,T. Frequenza e funzione dell'accusativo in gre= co. Accademia Nazionale dei Licei- Rendiconti della Clas= se di Scienze Morali, Storiche e Filosofiche, Serie VIII, Vol. XV, fasc.5-6,132.

76. De Mauro, T.,"Statistica linguistica" in Enciclopedia Ita= liana, App.III, vol.II, Roma,1962.

77. De Mauro,T., Il linguaggio della critica d'arte,Firenze, 1965.

78. De Mauro, T., Proposta per una teoria formalizzata del noe= ma lessicale e della storicità e socialità dei fenomeni linguistici, relazione letta nella I^giornata del Conve= gno "Linguaggi nella società e nella tecnica". Milano,14-17 ottobre, 1968.

79. De Tollenaere, F., Nieuwe Wegen in de Lexicologie,Amster= dam, 1963.

80. De Tollenaere, F., "Lexicographie alphabétique o idéologi= que", Cahiers de Lexicologie, 2, 1960,pp.19-29.

81. Devoto,G., La mèthode comparative et les conceptions actuel= les de la linguistique. Preprints per il 10^Congresso Inter= zionale dei linguisti, Bucarest, 1967.

82. Devoto, G., Relazione letta nella 1^giornata del Convegno, Linguaggi nella Società e nella tecnica, Milano, 14-17 Ottobre 1968.

83. Dewey, G., Relative frequency of English Speech, Sounds, Cambridge, 1923.

84. Dubois, J., Utilisation des statistiques lexicographiques pour l' étude structurale du lexique, in "Statistique et analyse linguistique", (Colloque de Strasbourg, 1964), Parigi 1966, pp. 95-98.

85. Durand, M., Spang, Hanssen, etc., Communications au VIème Congrès de Linguistes, Actes du VI C.L.(Parigi 1948) Parigi 1949, pp. 87-91.

86. Durham, St.P., and D.E. Rogers,: "An application of computer programming to the reconstruction of a protolanguage.", ICCL

87. Duro, A., " Notizie sul vocabolario", Studi di Filologia Italiana, XXV, 1967, pp.315-321.

88. Duro, A., "Humanities Computing Activities in Italy,"Computer and the Humanities, III, n.1,1968,pp.49-52.

89. 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 1969, pp.121-139.

90. Eldridge, R.C.,Six Thousand common English Words: their comparative frequency and wath can be done with them, Niagara Falls N.Y. 1911.

91. Etudes de linguistique appliquée , Parigi.

92. Evans, G.W., Wallace, G.F., Sutherland, G.L., Simulation Using Digital Computers, Prentice Hall, 1967.

93. Evseev, I., "La caractérisation des mots du point de vue d' une sémantique structurale", Cahiers de linguistique théorique et appliquée, IV, 1967, pp.55-64.

94. Farber D.I., Griswold, R.E., Polonsky I.L., The Snobol 3 Programmy Language, in The Bell System Technical Journal", XLV, n.6, (1966), pp.895-944.

95. Fillmore, Ch.,J., Toward a Modern Theory of Cases , RF Project 1685-6, Columbs, Ohio, 1966.

96. Fiorelli, P., "Vocabolari Giuridici fatti e da fare". Rivista Italiana per le scienze giuridiche, Nuova Serie,I, 1947 pp.293-327.

97. Fodor, I., The Rate of Linguistic Change, L' Aia, 1965.

98. Francini, G.L., De Biasi, G.B., "Study of a System of Minimal Speech-Reproducing Units for Italian Speech" The Journal of the Acoustical Society of America, 43,6,1968, pp.1282-1286.

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

100. Friedman, J.: "Applications of a Computer System for Transformational Grammar", ICCL.

101. Fromkin, V.A., Rice, D.L.,: "An Interactive Phonological Rule Testing System", ICCL.

102. Fucks, W., "Mathematishe Analyse von Sprachelementen, Sprachstil und Sprachen", Arbeitsgemeinschaft fur Forschung des Landes Nordrhein-Westfalen, Colonia,1955.

103. Fujimura, O., Kagaya, R., 'Structural Patterns of Chinese Characters ', ICCL.

104. Gammon, E., "Computational Approximation to the Word", ICCL.

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

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

107. Garvin, O.L., "Heuristic Syntax in Russian-English Machine Traslation ", To Honor Roman Jakobson, 1967, L'Aia, Vol.I, pp. 718-726.

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

109. Garvin, P.L., "Automatic linguistic analysis - a heuristic problem", in 1961 International Conference on Machine Translation of Languages and Applied Language Analysis, London, 1962.

110. Garvin, P.L., "The place of heuristic in the fulcrum approach to machine traslation",Lingua,21, 1968,pp.162-182.

111. Garvin, P.L., Simulation and analysis of intelligent behavior. Preprint, University of Chicago, 1969.

112. Giuliano, V.E. e Jones, P.E., Studies for the Design of an English Command and Control Language System, Cambridge, (Mass.), 1963.

113. Gladky, A.V., Mel'cuk, I.A., Tree Grammars (-Grammars), ICCL.

114. Golliet, A., Interventions préparées sur le rapport de M. Rivenc, (vedi 269)pp.337.

115. Gorn, S., "The Identification of the Computer and Information Sciences: Their Fondamental Semiotic Concepts and Relationships", Foundations of Language, Vol.4,4,1968, pp. 339-372.

116. Gougenheim,G. Michéa,R., Rivenc,P.,Sauvageot,A., L'élaboration du français fondamental. Paris,1964.

117. Gougenheim, G., Rivenc,P., M/me Hassan, Le français fondamental, in "Tendances nouvelles en matierè de recherche linguistique",Strasburgo,1964,pp.53-72.

118. Graur, A., Incercare asupra fondului principal lexical limbii române, Bucarest,1954.

119. Graur, A., "Statistique et Précision".Révue Roumaine de Linguistique, IX n.2 1966, pp.123-132.

120. Greenberg, J.H.,Osgood, C.E. e Saporta,S.,"Language Change", nel Supplemento al International Journal of American Linguistics,Vol.20, 1954.

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

122. Greimas A.J., Sémantique Structurale, Parigi, 1966.

123. Gross, M.,"On the equivalence of models of language used in the fields of mechanical translation and information retrieval",in Information Storage and Retrieval, II,1964, pp.43-57.

124. Gross, M.,Lentin, A., Notions sur les Grammaires formelles, Paris, 1967.

125. Gross, M., "Les Modéles en linguistique", Langages, 1968.

126. Guiraud, P., Bibliographie critique de la statistique linguistique,Utrecht, 1954.

127. Guiraud P. - Langage et communication,le substrat informationnel de la sémantisation, Bulletin de la Societé de Linguistique de Paris (1954)

128. Guiraud,P., Les Caractères Statistiques du Vocabulaire, Parigi, 1954.

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

130. Halle, M.,The Sound Pattern of Russian, L'Aia, 1959.

131. Harary, F. e Paper, H.H., "Toward a general calculus of Phonematic Distribution", Language, 33, 1957, pp.143-169.

132. Harris Z.S., "Distributional Structure", Word, 1954,pp.142-162.

133. Haugen, E., Bilingualism in the Americas, Alabama,1956.

134. Hays D.G., Processing Natural Language Text, in "Seminar on Computational Linguistics", 1966, pp.69-73.

135. Hays, D.G., Readings in Automatic Language Processing, New York, 1966.

136. Hays, D.G., Introduction to Computational Linguistics, New York, 1967

137. Hays, D.G., "Computational Linguistics: Introduction" in Encyclopaedia of Linguistics, Information and Control, Pergamon Press, 1969, pp.49-51.

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

139. Heilmann,L., "Statistica linguistica e critica del testo", Studi e problemi di critica testuale, 1961,pp.173-182.

140. Heilmann,L., "Recensione di: Herdan, G., The Calculus of Linguistic Observations, L'Aia, 1962" in Quaderni dell'Istituto di Glottologia, VII (1962-1963), pp.129-130.

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

142. Heilmann, L., "Presentazione" in: Arcaini, E., Principi di linguistica applicata, Bologna,1967,pp.V-VIII.

143. Hemmon, V.A.C.,A French word book based on a count of 400.000 running words, Madison,1924.

144. Herdan,G., Language as Choice and Chance, Groningen, 1956

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

146. Herdan, G., "Mathematical Models of linguistics Distribution Functions", in Etudes de linguistique appliquée, II, 1963, pp.47-48.

147. Herdan, G., The Advanced Theory of Language as Choice and Chance, Berlino, 1966.

148. Herzog, R., Computational Studies in Terminilogy,ICCL

149. Hjelmslev, L., Uldall, H.J., Outline of Glossematics, Copenhagen, 1957.

150. Hiorth, F., Zur formalen Charakteriesirung des Satzes, L'Aia, 1962.

151. Hockett, Ch.F., Language, Mathematics, and Linguistics, L'Aia, 1967.

152. Householder, F., Saporta, S., edit, Problems in Lexicography, Bloomington 'Ind.) 1962.

153. Hoz, V.,G., Vocabulario Usual Comun y Fundamental, Madrid, 1953.

154. O'Huallachain, Enquêtes en Irlande. Interventions préparées sur le rapporte de M. Rivenc, (vedi 121) pp.339-340.

155. Hymes, D.H.,"Lexicostatistics So Far", Current Anthropology, I, 1960, pp.3-44

156. Hymes, D.H., "More on Lexicostatics", Current Anthropology, I, 1960, pp.338-345.

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

158. Iivonen, A., "Automatic Recognition of Speech Sounds by a Digital Computer", ICCL

159. Ivic, M., Trends in Linguistics, trad. dal russo di Heppell, M., L'Aia, 1965. (Il capitolo 'Mathematical Linguistics' è alle pp.212-242).

160. Jakobson, R. e Halle, M.,Fundamentals of Languages, L'Aia, 1956.

161. Jordan, J., "L'état actuel de la Linguistique Romane et ses Perspectives de Developpement", Actes du X/ème Congrès International de Linguistique et Philologie Romanes, Strasburgo, 1962, I-III, Parigi, 1965.

162. Joshi, A., "Properties of Formal Grammars with Mixed Types of Rules", ICCL.

163. Joos, M., Recensione a: Zipf,G.K., The Psycho-Biology of Language, in "Language", 1936, pp.196-210.

164. Joos, M., "Linguistics Prospects in the United States", in Mohrmann,C, Sommerfelt, A.,Whatmough, I, edit. in: Trends in European and American Linguistics 1930-1960,Utrecht, 1963, pp.11-20.

165. Josselson, H.H., The Russian Word Count and Frequency Analysis, Detroit, 1953.

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

167. Josselson, H.H., "Lexicography and the computer". To Honor Roman Jakobson, L'Aia, 1967, pp.1046-1059.

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

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

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

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

172. Kacding, F.W., Häufigkeitsworterbuch der deutschen Sprache, Berlino, 1897-1898.

173. Karttunen, F., "Formal Models for Rhonological Theory", ICCL.

174. Katz, Postal P., An Integraded Theory of Linguistic Descriptions, Cambridge,(Mass),1964.

175. Katz, J.J., Fodor J.A., "The structure of a Semantic Theory", in"Language",1963, pp.170-210. Trad. francese in Cahiers de Lexicologie, XI-X (1966-67).

176. Kay M., Ziehe T.W., Natural Language in Computer form, Santa Monica (California), 1965.

177. Keniston, H.A., Basic List of Spanish Words and Idioms, Chicago, 1933.

178. Kerckhoffs, A.,La cryptographie militaire, Parigi, 1883.

179. Kiefer, F., Some Aspects of Mathematical Models in Linguistics, in Statistical Methods in Linguistics, N.3, 1964, pp.8-26.

180. Kiefer, F., Mathematical Linguistics in Eastern Europe, New York, 1968.

181. Kiefer, F., "Slant Grammar Calculus Some Observations", ICCL

182. Klein, S., et al. "Monte Carlo Simulation of Language Change in Tikopia". ICCL

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

184. Kucera, H., e Monroe, G. K., A comparative Phonology of Russian, Czech and Germany, New York, 1968.

185. Kulagina, O.S., Ob odnom spobose opredeleniu grammaticheskih ponjatii na base teorii mnozhestv, in Problemy Kibernetiki I, Mosca, 1958, pp.203-215.

186. Ladefoged, P., "Quantitative Measures of Phonetic Similarity." ICCL

186a Lakoff G. On the Nature of Syntactic Irregularity, Report No. NSF-17, Camridge, (Mass.) 1965.

186b Lakoff, R.T., Abstract Syntax and Latin Complementation, The M.I.T Press, 1968.

187. Lamb, S.M., "The digital Computer as an Aid in Linguistics", Language, Vol.37,N.3 1961, pp.382-412.

188. Lamb, S.,M., "On the mechanization of syntactic analysis", in 1961 International Conference on Machine Translation of Languages and Applied Language Analysis, London, 1962.

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

190. Lamb S.M., Jacobsen W.H., Jr., A High-Speed Large -Capacity Dictionary System, in "Readings in Automatic Language Processing",(edit.Hays, D.G.), New York,1966,pp.51-72

191. Lambek, J., "The Mathematics of Sentence Structure", American mathematical Monthly, Vol.65,1958 pp.154-170.

192. Latil, de,P., Il pensiero artificiale, Milano, 1962. Trad. dal francese "La pensée artificielle", Paris, 1953.

193. Lees, R.B., "Recensione di: Apostel, L., Mandelbrot, B., Morf, A., Logique , Language et theorie de l'information",Parigi 1957", Language, XXXV, 2, 1959, pp.271-303.

194. Lepscky, G. ., La linguistica strutturale, Torino,1966.

195. 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.

196. Leroy,M., Les Grands Courants de la linguistique Moderne, Bruxelles,1963.

197. Les Machines dans la Linguistique. (Actes du Colloque International sur la Mécanisation et l'Automation des Recherches Linguistiques), Praga,1968.

198. Lyons, J., Introduction to Theoretical Linguistics, Cambridge, 1968.

199. Luhn,H.P., "The Automatic Creation of Literature Abstracts", IBM Journal of Research and Developement, 2, 1958, p.159.

200. Mandelbrot, B., "An informational Theory of the statistical Structure of language", Communication Theory, édit. W.Jackson, 1953,pp.486-502.

201. Mandelbrot, B., "Structure Formelle des Textes et Communication", Word, 10,1954, pp.1-27.

202. Mandelbrot, B., "Les constantes chiffrées du discours", Le Language (volume pubblicato sotto la direzione di André Martinet),1968, PP.46-56.

203. Marcus, S., Typologie des langues et modèles logiques,in Acta Mathematica Acad. Scientiarum Hung. Budapest, 1963.

204. Marcus, S., "Contextual Grammars", ICCL.

205. Martinet, A., "La double articulation linguistique" Travaux du Cercle Linguistique de Copenhague, V, 1949, pp. 30-37.

206. Martinet, A., Eléments de linguistique générale, Parigi, 1960, pp.2.

206a Martinet, A.,"Mot et Synthème", Lingua,21,1968,pp.294-302

207. Mastermann, M., Man-aided computer translation from english into french using an on-line system to manipulate a bi-lingual conceptual dictionary, or thesaurus, negli atti della "2/ème Conférence Internationale sur le "traitement automatique des langues", Grenoble, Agosto, 1967.

208. Mathesius, V., "La structure phonologique du lexique du Tchèque moderne, Travaux du Cercle linguistique de Prague, I, 1929, pp. 67-84.

209. Mechanical Translation and Computational Linguistics, Chicago

210. Mey,J., "Level-oriented Descriptions and Context-Free Grammars in Computational Linguistics",ICCL

211. Meriggi, P., I Casi in Italiano, Comunicazione al III Congresso della Società Linguistica Italiana, Roma,1969.

212. Mersel, J., Reitz, G., A Manual for Automatic Dictionary Revision, Los Angeles, Ramo-Wooldrige Corp., C97-OU7, 1960.

213. Miller, G.A., "Introduction " alla riedizione "Zipf,G.K., The Psycho-biology of Language, Cambridge, (Mass.)1965, pp.V-X.

214. Michéa, R., "La relation rang-fréquence et la structure statistique de la langue parlée", Bulletin de la Société de Linguistique de Paris, (1968), pp. 9-14.

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

216. Migliorini, B., Storia della Lingua Italiana,Firenze,1960

217. Migliorini, B., Saggi sulla lingua del novecento, Firenze, 1963.

218. Migliorini, B., Baldelli, I., Breve storia della lingua italiana, Firenze, 1964.

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

220. Mohrman, C., Sommerfelt, A., Whatmough,J., Trends in European and American Linguistics, 1930-1960, Utrecht, 1962.

220a Montgomery, C.A. "Linguistics and Automated Language Processing", ICCL, 1969.

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

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

223. Muegge, M.A., German frequency Vocabulary, London, 1936.

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

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

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

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

228. Muller, Ch., Initiation à la statistique linguistique, Parigi, 1968.

229. Najm, C.M.,"A program for partial automation of comparative reconstruction", Anthropopogical Linguistics,4, N. 9 (1962),pp.1-10.

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

231. Noomen, W., Recensione a Juilland, A.,ect., Frequency dictionary of Rumanian Words, Lingua,23,1969,pp.202-204.

232. Oettinger, A.G., A Study for the Design of an Automatic Dictionary, Ph.D. Thesis, Harvard University,1954

233. Oettinger, A.G.,"Linguistics and Mathematics", Studies Presented to Joshua Whatmough, L'Aia,1957,pp.179-186.

234. Oettinger, A.G., "Automatic Syntactic Analysis and Pushdown Store", Proceedings of the Symposia in Applied Mathematics,XII,1961,pp.104-129.

235. Oettinger, A.G., Automatic Processing of Natural and Formal Languages, in "Proceedings of the IFIP Congress", 65, Washington, 1965.

236. Ogden C.K., The ABC of Basic English, London,1932.

237. Ortiz, A., Zierer, E., Set Theory and Linguistics, L'Aia, 1968.

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

239. Osgood, C.E., Suci, G.J., Tannebaum, P.H., The measurement of Meaning, Urbana, 1957.

240. Ott, W., "Metrical Analysis of latin Hexameter by Computer", Revue, IV, 1966, pp.7-20.

241. Pacifico, M.,"I numeri e le lettere; automatizzazione nelle ricerche linguistiche e letterarie", in "Nuova Corrente", XIL, 1967,pp.2-29.

242. Pakard, D.W., A Concordance to Livy, Cambridge (Mass.) 1968

243. Palmer, H.E., Hornby, A.J.,Thousand Words English,London, 1937.

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

245. Papp, F., Mathematical Linguistics in the Soviet Union, L'Aia, 1966.

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

247. Percy, M., Tannembaum et al."A Stochastic Approach to the grammatical Coding of English", Computational Linguistics, VIII, n.6, 1965.

248. Perschke, S., The computer programs of the "S.L.C." System for Machine Traslation, Brussels, 1965.

249. Peters,P.S.,"The Use for Context-Sensitive Rules in Immediate Constituent Analysis", ICCL.

250. Petöfi, J.S., "On the Analysis of Narrative-Text-Structures",ICCL.

251. Pfeffer, J.A., Evolution of the Basic (Spoken) German Word List, Pittsburg, 1963

252. Phal, S., Interventions préparées sur le rapport de M.Rivenc, (vedi 269) pp. 340-345.

253. Plath-Warren, "Mathematical linguistics", Trends in European and American Linguistics, 1930-1960,Utrecht,1962, pp.21-57.

254. Postovalova,V.I., "O socetajemosti differencial'nyx priznakov soglasnyx fonem sovremennogo russkogo jazika". Problemy Linguisticeskogo. Mosca, 1966.

255. Pottier, B. "Vers une sémantique moderne" Travaux de linguistique et de littérature,p.p.le Centre de philologie et de litteratures romanes de l'Université de Strasbourg, II-1-1964, pp.107-137.

256. Progress Report, N°2, NSF- C 233, Computer-Aided Research in Machine Translation,Bunker-Ramo Corporation MT Group, 1963.

257. Purin C.M., A standard German vocabulary of 2000 words and idioms. Chicago,1931.

258. Quemada, B.,"L'inventaire statistique de l'Allemand parlé",Bulletin d'Information du Laboratoire d'Analyse Lexicologique, VII, 1963.

259. Querido,A.,"Computer Implementation of a Transformational Description of French", ICCL.

260. Quillian, M.R., Semantic Memory, in Minsky,M., edit (1968) pp.216-270.

261. Radzin, H., Statistical Method in Lexicological Research in the Baltic States. ICCL, 1969.

262. Reed D.W., "A Statistical Approach to Quantitative Linguistic Analysis", Word, 5,3,1949,pp.235-247.

263. Revue, Periodico dell'"Organisation internationale pour l'etude des langues anciennes par ordinateur", Liège

264. Revzin I.I., Les modèles linguistiques, Paris,1968, trad. di V.Gentilhomme dal russo"Modeli Jazyka", Mosca 1961.

265. Rey, A., "A propos de la définition lexicographique", Cahiers de Lexicologie, 6, 1965, pp.67-80.

266. Rey, A.-Debove, "Preprints " per il Second I. Congress of Applied Linguistics, Cambridge (G.B.),1969.

267. Rhodes, I., "A new approach to the mechanical syntactic analysis of Russian" Machine Translation,VI,1961,pp.33-50.

268. Rischel,J., "On functional lood in phonemics", Statistical Methods in Linguistics, Stoccolma, 1, 1961,pp.13-23.

269. Rivenc P., Etat actuel des enquêtes sur les langues parlées et les langues de spécialité, in "Actes du premier colloque international de linguistique appliquée",(1964 Nancy,1966, pp.313-336.

270. Robison, J., Parse,A. System for Automatic Syntactic Analysis of English text, Santa Monica, Calif.,1965.

271. Roceric Alexandrescu,A., Fono-statistica Limbii Române, Bucarest, 1968.

272. Roceric-Alexandrescu, A.,"Recensione di: Juilland A., ecc., Frequency Dictionary of Rumanian words"(vedi), Revue Roumaine de Linguistique, IX° n.2, 1966, pp.205-210.

273. Rodriguez, Bou , L., Recuento de Vocabulario Espanol, Puerto Rico 1952.

274. Rondeau,G.,Interventions préparées sur le rapport di Rivenc (vedi 269) pp.345-346.

275. Rosiello, L., Struttura uso e funzioni della lingua,Firenze, 1965.

276. Rosiello, L."Stilistica e strutturalismo linguistico", Sigma, 9, 1966, pp.5-21.

277. Ross, A.S.C., "Philologica mathematica", in Casopis pro Modernì Filologii, XXXI (1947-1948), Praga, 102-108, 185-191.

278. Rubenstein, H., Goodenough, J.B., Contextual Correlates of Synonymy, A.C.M. 8, 1965.

279. Rubenstein, H., "Direction in Semantic Research", Seminar on Computational Linguistics, Bethesda, 1966 pp.97-104.

280. Ruwet, N., "La linguistique générale aujourd'hui", in Archives européennes de Sociologie, 5, 1964, 277-310.

281. Ruwet, N., Introduction à la grammaire générative, Parigi,1967

282. Saramandu, N.,"Sur le rendement fonctionnel des types phonématiques en roumain", Cahiers de linguistique théorique et appliquée, III, 1966, pp.147-160.

283. Sarteschi, Castrogiovanni,P;, Del Carlo Giannini, G., Maffei, G., Pasquinelli, P.P., Faedo,A., Lijtmaer,N., Torrigiani,G., Zampolli, A., Il linguaggio nel test di Rorschach, Pisa, 1968

284. Saumjan, S.K., La cibernetica e la Lingua.Trad. italiana in "I problemi attuali della Linguistica", Milano 1968, pp.159-176.

285. Schank, R.C, L.G.,Tesler "A Conceptual Dependency Parser for Natural Language", ICCL

286. Schiaffini, A., Lessicologia italiana, Roma, 1961.

287. Schützenberger, "On a application of semi-groups methods to some problems in coding", IRE Transactions on Information Theory, Vol.IT., 2, N.3,1956,pp.47-60.

288. Second International Congress of Applied Linguistics, Abstracts, Cambridge (G.B.), 1969.

289. Second International Congress of Applied Linguistics: second circular Cambridge (G.B.), 1969.

290. Shannon, C.E.Weaver,W., The Mathematical Theory of Communication, Urbana, 1949.

291. Skalmowski,W., M. van Overbeke:"Computational Analysis of Language Interference on the lexical Level",ICCL.

292. Smith, P.H., Villani, R.D., A One-Tape-Pass Concordance Generator, IBM Research, Yorktown, N.Y.,1960.

293. Smith, R.N., "Automatic Simulation of Historical Change", ICCL.

294. Stanley, R., "Redundancy Rules in Phonology", Language, 43,1967,p.393.

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

296. Stevens, M.E., Giuliano, V.E., Heilprin Editori, Proceedings of the Symposium on Statistical Association Methods for Mechanized Documentation, Washington,1966.

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

298. Susumu Kuno, The predictive Analyzer and a Path Elimination Technique, Communications of the ACM, 8,1965.

299. Susumu, Kuno, "Automatic Syntactic Analysis", in Seminar on Computational Linguistics, Bethseda, 1966,pp.19-38.

300. Suteu, V., "Observatii asupra frecventei cuvintelor în operele unor scriitori români", Studi si cercetàri linguistice,4,1959.

301. 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.

302. Tagliavini, C., Concordanze della Divina Commedia, Pisa, 1965.

303. Tagliavini, C., Introduzione della Glottologia, Bologna, 1966.

304. Tagliavini, C., Applicazione dei calcolatori 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.

305. Tagliavini, C., Intervento in "Actes du Séminaire International ect." (vedi) pp.106-109.

306. The Finite String, Newsletter of the Association for Machine Traslation and Computational Linguistics. pp.215-234.

306a Tarski, A., "The concept of truth informalized languages", Logic, Semantic, and Mathematics, Oxford, 1957,pp.152-278.

307. Thorndike, E.L., e Lorge, I., The Teacher's Word Book of 30.000 words, New York, 1944.

308. Tijidschrift voor toegepaste Linguistiek, Lovanio.

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

310. Torrigiani, G., Problemi metodologici dell'analisi linguistica mediante eleboratori elettonici presso il CNUCE, Pisa, 1968.

311. Trakhtenbrot B.A., Algoritmi e macchine calcolatrici automatiche, trad. di Mangione, C., Milano, 1964.

312. Trubetzkoy, N.S., Principes de Phonologie (trad.franc. di J. Cantineau), Parigi, 1949.

313. Ullmann, S., Semantics An Introduction to the Science of Meaning, Oxford, 1962. Trad.Ital., Bologna, 1966.

314. Ungeheuer, G., "Das logistische Fundament binärer Phonemklassifikationen" in Studia Linguistica, 1959, pp.69-97.

314a Universal in Linguistic Theory a cura di Bach, E., e Harms, R.T., New York, Chicago, London, 1968.

315. Valero, C., De la Cryptographie, Parigi, 1893 e 1896.

316. Van Buren, P., "Preliminary aspects of mechanisantion in lexis", Cahiers de Lexicologie, 11, 1967, pp.89-111 e segue in 12, 1968, pp.71-84.

317. Varga Dénes, "Questions of Improving the Effectiveness of Context Sensitive Parsing Systems", ICCL.

318. Veillon G., Présentation de l'analyse morphologique et du programme du dictionnaire allemand, CETA, Document Scientiphique, n.5, 1963.

319. Viaris, Le Marquis de, L'art de chiffrer e déchiffrer dépêches secrètes, Parigi, 1893.

320. Vinay, J.P., Enseignement et apprentissage d'une langue seconde, in Le Langage (éd.A.Martinet), Gallimard, 1968, pp.685-725.

321. Walde, A., e Pokorny, J., Vergleichendes Worterbuch der indogermanichen Sprachen, Berlino, 1932.

322. Vander Beke, G.B., French word book, New York, 1929.

323. Watts, A.F., The Language and Mental Development of children, London, 1946, pp.57-58.

324. Weinrich, U., "Explorations in Semantic Theory", Current Trends in Linguistics, 3, pp.395-477.

325. Weinreich, U., "On the Semantic Strucure of Language", Universals of Language, Cambridge (Mass), 1963, pp.114-171.

326. Whatmough, J., "Mathematical Linguistics", Proceedings of the VIII International Congress of Linguists, Oslo, 1957, Oslo, 1958.

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

328. Wisbey, R. "Computers in Lexicography" in The Use of Computer in Anthropology, edito da Dell Hymes, L'Aia, 1965, pp.

329. Wood, C.R., "Dialectology by Computer", ICCL

330. W.Von Wartburg, R. Hallig, Begriffssystem als Grundlage für die lexicographie, 1952.

331. Yang, S.C., C.W. Yang, " A Universàl Graphic Character Writer", ICCL.

332. Yule; G.U., The Statistical Study of Literary Vocabulary, Cambridge, 1944

333. Zampolli, A., Studi di statistica linguistica eseguiti con impianti IBM, Tesi di laurea, Padova, 1960.

334. Zampolli, A., "Recherche statistique sur la composition phonologique de la langue italienne", Les Machines dans la Linguistique, L'Aia, 1968.

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

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

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

338. Zampolli, A., "Projet d'un dictionnaire italien de machine, Calcolo, V, suppl. N.2, Pisa, 1968, pp.109-126.

339. Zampolli, A., 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.

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

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

342. Zampolli, A.,"La section linguistique du C.N.U.C.E.: projets et méthodes", negli Atti del Colloque International sur la recherche computationelle en philologie, Liegi, 1969 (in pubblicazione).

343. Zipf G.K. (in risposta a nr.66) "Statistical Methods and dynamic philology", Language, 13, 1937, pp.60-70.

344. Zipf, G.K., Human Behavior and the Principle of Least Effort, Cambridge (Mass), 1949.

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

INDICE

I

- Linguistica Applicata, Linguistica Matematica, Linguistica Computazionale ..... Pag. 1
- Alcune classificazioni proposte per le attività della Linguistica Computazionale ..... " 17
- L'International Conference on computational linguistics del 1969 ..... " 27
- La sezione linguistica del CNUCE ..... " 40

II

- Lo spoglio lessicografico: procedimenti e problemi ... " 47
- Il Dizionario di macchina del CNUCE ..... " 51
- Un Esempio di ricerca nel settore della statistica lessicale:
  - Premessa ..... " 61
  - I dizionari di frequenza ..... " 65
  - Il dizionario di frequenza dell'italiano scritto " 72
- Note ..... " 82
- Figure ..... " 93
- Bibliografia ..... " 96