

CENTRO NAZIONALE UNIVERSITARIO DI CALCOLO ELETTRONICO

PISA
Via S.Maria, 36

ASSOCIAZIONE ITALIANA PER IL CALCOLO AUTOMATICO
(A.I.C.A.)

Giornate di studio sul tema:

"La preparazione del Personale per l'elaborazione automa
tica dei dati, in Italia"

ROMA, 2-3 ottobre 1969

Appunti per l'intervento del Dr. Antonio Zampolli

coordinatore della sezione linguistica del Centro Nazionale Uni
versitario di Calcolo Elettronico dell'Università di Pisa

A. Mi occupo della elaborazione automatica dei dati in quanto può interessare la linguistica e le discipline umanistiche in genere. Attualmente dirigo la sezione linguistica del Centro Nazionale di Calcolo Elettronico di Pisa (C.N.U.C.E.); costituita tre anni fa, essa conta 12 persone, utilizza gli elaboratori del C.N.U.C.E., in particolare le è riservato a tempo pieno un sistema IBM 360/30, dotato di dispositivi specialmente concepiti per le nostre esigenze, soprattutto per la stampa di alfabeti e lingue diverse. L'attività della sezione si articola in tre settori.

1° Numerosi Istituti Universitari e di ricerca, italiani e stranieri, (circa 50) si avvalgono delle procedure e dei programmi da noi approntati e dei nostri elaboratori. A questi progetti partecipiamo per tutto quanto concerne l'analisi, la programmazione e l'esecuzione della ricerca per mezzo del calcolatore. I lavori in atto possono essere classificati nei 10 tipi esposti alla nota 1.

2° La nostra équipe svolge, accanto ai lavori in collaborazione, ricerche proprie. Alcune importanti fasi elaborative sono comuni alla maggior parte delle applicazioni. Per esempio riconoscere, analizzare, classificare, le unità linguistiche che compongono il testo (fonemi, monemi, parole, sintagmi, ecc.), e identificare le strutture nelle quali sono inserite. Gli algoritmi suppongono modelli linguistici a livelli relativamente semplici (per es. fonematici e morfologici) e a livelli più complessi (per es. di strutture sintattiche e semantiche). Le nostre ricerche sono fin qui condotte per il latino e l'italiano moderno.

I materiali analizzati per conto dei diversi istituti costituiscono un corpus operabile elettronicamente che è probabilmente il più vasto mai realizzato (50 milioni di parole in 15 lingue), ed è registrato con uniformità di criteri scientifici e tecnici. Ci si offre, pertanto, la possibilità di condurre, finalmente su una base adeguata e sufficiente, lo studio dei modelli e delle strutture quantitative del linguaggio, per superare l'"impasse" nella quale la statistica linguistica si è venuta a trovare dopo i primi facili entusiasmi.

3° L'attività didattica della nostra sezione si svolge a tre livelli.

- a) I ricercatori che iniziano un loro progetto presso di noi, vengono innanzi tutto introdotti ai metodi e ai problemi dell'automazione applicata alla linguistica, per mezzo di riunioni e del lavoro in comune; spesso gli Istituti utenti specializzano una o più persone che seguono la ricerca presso di noi.
- b) Svolgo, da due anni nell'ambito del C.N.U.C.E., un corso di ap

plicazioni 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.).

Il programma del corso è allegato.

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

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

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

B. L'esperienza acquisita suggerisce alcune constatazioni.

Le applicazioni fino ad oggi tentate sono raggruppabili in prima approssimazione nei 5 filoni principali esposti nella nota n° 2. Essi richiedono una preparazione specifica del ricercatore, che deve essere fornita dai corsi di studio esistenti. E' però indispensabile isolare alcuni aspetti comuni, che possono essere trat

tati solo in corsi e seminari specializzati.

1. L'elaborazione elettronica dei dati è oggi di fatto applicata in quasi tutte le discipline umanistiche, e non può più essere ignorata dai loro cultori. Si stanno costituendo, presso i centri di calcolo europei e americani, immense biblioteche elettroniche, e diventa sempre più probabile per un ricercatore trovare registrato su schede o nastro il testo che lo interessa, con relativi programmi e procedura di analisi, con notevole risparmio di spesa e di tempo.

Il primo e più importante compito, io credo, è di informare studiosi e studenti sulle applicazioni già in atto nella loro specifica disciplina e nelle discipline vicine. I progetti si sono moltiplicati enormemente, e il compito non è facile, tanto che è l'argomento principale di tutti gli ultimi congressi del settore.

Sono già funzionanti organismi e riviste per lo scambio e la raccolta di informazioni e di materiali, presso i quali rappresentiamo l'Italia.

- 2° E' necessario diffondere i linguaggi di programmazione che la esperienza mostra più adatti sia ai problemi da elaborare sia alla formazione dei ricercatori.

Questo problema è stato collegialmente discusso all'ultimo congresso del KVAL, ove è stata messa in evidenza la opportunità di diffondere tra i linguisti linguaggi del tipo PLI, e di sfatare alcuni luoghi comuni, per es. sulla superiorità, per i linguisti, dei linguaggi manipolatori di liste. D'altra parte ai colleghi di molte nazioni è noto che numerosi linguisti hanno rinunciato ai loro progetti di impiego dei calcolatori per difficoltà incontrate in corsi di programmazione di orientamento matematico.

- 3° L'approccio e gli esercizi dovrebbero avere carattere prevalentemente linguistico. Soprattutto la terminologia deve essere curata in tal senso.

- 4° Come hanno mostrato le esperienze statunitensi, gli allievi linguisti devono avere accesso al calcolatore secondo modalità di tempo diverse da quelle in uso nei centri di calcolo per gli studenti di materie scientifiche, e ciò a motivo della estensione dei dati e della natura di certe operazioni (per es. di SORT). E' inoltre indispensabile che gli allievi si esercitino nella elaborazione di testi reali, e pertanto deve essere disponibile un corpus di testi sufficientemente ampio e variato.
- 5° Le applicazioni linguistiche sono caratterizzate, generalmente, dalle notevoli dimensioni dell'input e dell'output, e dal continuo intercalarsi di operazioni umane alle operazioni di macchina. Questi fattori pongono problemi di organizzazione, di procedura, di esecuzione, di controlli, di conservazione, particolarissimi, senza riscontro forse in altri settori. Questo aspetto è di fondamentale importanza nella preparazione del ricercatore. Inoltre è necessaria la conoscenza di macchine e dispositivi speciali: esempio chiarissimo quelli per la stampa degli elaborati, soprattutto per la pubblicazione.
- 6° I diversi settori hanno in comune, come si è detto, alcune fasi fondamentali, per le quali sono già state proposte e sperimentate diverse procedure e strategie, le quali costituiscono un argomento fondamentale di insegnamento. Ma soprattutto sono richieste alcune conoscenze scientifiche specifiche, che non sempre sono comprese negli altri corsi universitari. Sono le più svariate: vanno dai problemi filologici di edizione di un testo, per la preparazione dell'input, alle grammatiche formali per l'analisi sintattica, alle nozioni statistiche elementari. Alcuni problemi posti dall'automazione in linguistica sono risolvibili solo per mezzo di convenzioni e concetti accettati nel nostro settore: i problemi connessi alla definizione dei lemmi e alla individuazione delle loro occorrenze forniscono un buon esempio.

Questi problemi e le loro soluzioni convenzionali costituiscono un altro capitolo fondamentale dell'insegnamento.

- C. Nella facoltà di lettere, l'insegnamento potrebbe essere reso autonomo data la varietà delle discipline interessate. Per questo stesso motivo, esso può essere inserito anche in una facoltà di informatica, ove sarebbe possibile creare degli specialisti in questo peculiare settore di studi.

Nelle università, dove esiste un istituto o un centro di calcolo elettronico, esso è forse la sede più adatta, per il carattere di interdisciplinarietà e per i mezzi e i testi a disposizione. Negli altri paesi, soprattutto negli USA, conosco esempi di tutte e tre le soluzioni.

D'altra parte, alcune tra le considerazioni esposte al n. B, consigliano l'opportunità di seminari sia per rendere concreta la possibilità di ricerche ed esperienze reali, sia per permettere l'incontro con esperti e ricercatori rappresentativi dei diversi settori di applicazione e per agevolare la diffusione e lo scambio delle informazioni tra paesi diversi.

Almeno a livello informativo, è forse necessario creare anche nelle scuole medie superiori un interesse e un certo grado di conoscenza sulle possibilità e sulle realizzazioni dei calcolatori nell'ambito delle scienze umane.

Il problema, qui, sta nella informazione degli insegnanti, e potrebbe essere risolto, fino a che non vi siano corsi nelle facoltà di lettere, intervenendo nei corsi di aggiornamento per insegnanti delle scuole medie, oggi sempre più diffusi.

NOTA N° 1

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

Ricordo in particolare le imprese per l'ittita, per il romeno, e soprattutto per l'archivio e il dizionario della lingua italiana dell'Accademia della Crusca, al cui esempio è dovuto molto dello sviluppo nel settore.

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

Per esempio, il dizionario storico italiano della lingua giuridica e il lessico intellettuale europeo.

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

Per esempio gli spogli dell'opera omnia di Antonio Rosmini, di L.A.Seneca, di S.Tommaso di Aquino, ecc.

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

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

5. Inchieste dialettali, atlanti linguistici, testi dialettali.

E' possibile ora tracciare diagrammi e carte sotto controllo del calcolatore. E' così possibile combinare la compilazione delle carte con operazioni statistiche e con la gestione automatica dello schedario delle risposte dialettali.

6. Elaborazioni filologiche e documentalistiche di grandi corpus di origine diversa.

Come esempio cito il progetto per la edizione della raccolta Barbi di canti popolari italiani, quello per la elaborazione metrica e la classificazione di altre raccolte già edite di canzoni popolari, lo studio dei canovacci della Commedia dell'Arte, ecc.

7. Statistica linguistica (dizionari di base, frequenze fonematiche, grammatica dell'uso, ecc.).
(di questo punto parlo nella relazione).
8. Psichiatria, psicologia, pedagogia.
Per es. analisi statistica di conversazioni di diverse categorie di malati mentali e lo studio della evoluzione del linguaggio in fantile.
9. Simulazioni di modelli linguistici morfologici grammaticali e semantici.
(di questo punto parlo nella relazione).
10. 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 diplomatica, (atti notarili in latino fino al 1100), ecc.

NOTA N° 2

Le applicazioni fino ad oggi tentate sono raggruppabili, in prima approssimazione, in questi filoni principali:

- analisi e sintesi del suono: questo settore, solo di rado è considerato dai linguisti.
- spoglio di testi a diversi livelli (fonematico, lessicale, morfologico, ecc.) per organizzare i materiali linguistici secondo modelli di documentazione, nuovi o tradizionali: per es. concordanze, indici lessicali, indici inversi, indici fonematici, ricerca di stilemi, reperimento di termini o cose notevoli, ecc.
- spogli di testi per conteggi statistici sulle diverse unità linguistiche (fonemi, monemi, parole, sintagmi, ecc.), nell'ambito degli studi di statistica linguistica o linguistica quantitativa.
- costruzione e verifica di modelli di operazioni e strutture lin

guistiche (per es. analisi sintattica e semantica) sia in sincronia ("linguistica algoritmica") sia in diacronia (lessicostatistica e glottocronologia).

- traduzione automatica tra lingue diverse.
- analisi documentalistica, per operazioni di information retrieval. Si applicano qui, come per la traduzione automatica, risultati e metodi dei tre precedenti settori, anche se, per quanto concerne la documentalistica, l'approccio linguistico è oggi generalmente criticato.