CRONACA

Notizie, spunti e appunti

a cura di

A. ZAMPOLLI

Estratto dall'*Archivio Glottologico Italiano*

Volume LV - Fasc. 1-2 (1970)

FIRENZE

CASA EDITRICE FELICE LE MONNIER

1970

* Alla terza *International Conference on Computational Linguistics*, organizzata a Stoccolma dal KVAL (*Forskningsgruppen för Qvantitativ Lingvistik*) nei giorni 31 agosto - 5 settembre 1969, hanno preso parte più di 200 persone di 22 Paesi, appartenenti a Facoltà, Dipartimenti o Centri di Linguistica (57%), di Scienza dell'informazione (25%), di Medicina, Psicologia, Pedagogia, Traduzione Automatica, Documentazione Automatica, ecc. (18%). Le percentuali, se raffrontate a quelle dei convegni precedenti (per es. a Grenoble, agosto 1967), confermano indirettamente la tendenza della Linguistica Computazionale (LC), espressa in questo Convegno, verso una propria autonomia nei confronti della *linguistica applicata* e della *linguistica matematica*. In questo contesto assumono particolare significato comunicazioni come quella di P. A. Verburg (*Hobbes' Calculus of Words*), che ha ricercato i fondamenti della LC nella filosofia di Hobbes e di Leibniz. La caratteristica essenziale della linguistica applicata sembra essere quella di *applicare* ritrovati e teorie

della linguistica al perseguimento di *scopi pratici*, generalmente raggruppati in due grandi categorie (vedi J. P. VINAY, *Enseignement et apprentissage d'une langue seconde*, in *Le langage*, [a cura di A. Martinet], Parigi, 1968, p. 699): la glottodidattica e il trattamento dell'informazione contenuta in testi e messaggi formulati in una lingua naturale. Sono stati invece descritti molti studi e progetti che non presentano questa caratteristica: concordanze a livello fonematico e lessicale, archivi di dati per lo studio di una lingua, ricerche definite di « linguistica storica e comparativa », elaborazioni di grandi *corpus* dialettali, statistiche fonematiche e lessicali, modelli di grammatica, ecc. Soprattutto si è notato un nuovo atteggiamento nello studio e nella verifica di modelli linguistici a livello sintattico e semantico: in precedenza si mirava direttamente allo scopo ambizioso e utopistico di simulare globalmente con il programma la « competenza linguistica » del parlante, così da poter operare sul linguaggio nel suo carattere di supporto e di veicolo di informazione; ora invece il calcolatore viene usato soprattutto come uno strumento per verificare descrizioni di settori limitati di una lingua, e per analizzare o generare automaticamente delle frasi al fine di valutare il grado di adeguatezza di un modello grammaticale. Il problema se e fino a che punto sia possibile formalizzare la descrizione di una lingua è stato molto discusso (ci si riferisce per es. esplicitamente al ben noto teorema di Tarski, secondo il quale ogni teoria semantica consistente e completa di un linguaggio deve essere formulata in un metalinguaggio di ordine più elevato del linguaggìo descritto: A. TARSKI, *The concept of truth in formalized languages*, in *Logic, Semantics, and Metamathematics*, Oxford, 1956, pp. 152-278): finalmente è stata formulata in termini espliciti la distinzione tra linguaggi artificiali e linguaggi naturali, al di là delle utili analogie metodologiche che in precedenza avevano ingenerato confusioni teoriche.

La LC viene distinta anche dalla linguistica matematica: nella linguistica matematica la formalizzazione e l'adozione di tecniche matematiche è richiesta per conferire certe caratteristiche al modello (struttura assiomatica, esplicitazione nelle definizioni, verificabilità delle regole, ecc.); nella LC la formalizzazione è richiesta per la eseguibilità meccanica delle elaborazioni, e non sempre come caratteristica intrinseca della ricerca. A ciò si aggiunga che «si può meccanizzare più di quanto si possa formalizzare » come ha osservato S. Gorn (*The Identification of the Computer and Information Sciences*, in « Foundations of Language », IV, 4, 1968, pp. 339-372). Si sono costituiti un insieme di nozioni, di metodi, di tecniche, di procedure, e una terminologia, che stanno diventando patrimonio comune a quanti lavorano nel campo della LC.

Gli organizzatori hanno suddiviso le circa 70 comunicazioni in 7 sezioni concomitanti. Seguo la loro ripartizione, rilevando però che essa non ha valore classificatorio: come è noto, i tentativi di classificare la congerie di attività riunite sotto il termine LC non hanno dato finora risultati soddisfacenti (basti pensare al disaccordo tra le classificazioni

più note, per es. quelle proposte da S. M. Lamb, *The digital Computer as an Aid in Linguistics*, « Language », XXXVII, 3, 1961, pp. 382-412 e *Linguistic Data Processing*, in *The Use of Computers in Anthropology* [a cura di D. Hymes], L'Aia, 1965, pp. 159-188; da P. L. Garvin, *Computer Participation in Linguistic Research*, « Language », XXXVIII, 4, 1962, pp. 385-389; da J. C. Gardin, *A Typology of Computer Uses in Anthropology*, in *The Use of Computers in Anthropology* (*op. cit.*), pp. 103-118; da C. A. Montgomery, *Linguistics and Automated Language Processing*, Stoccolma, 1969, e da D. G. Hays, *Applied Computational Linguistics*, Cambridge U. K., 1969). Dati i limiti rigorosi di spazio disponibile (per una esposizione più particolareggiata si veda la mia relazione, *Due conversazioni sullo stato attuale della linguistica computazionale*, Pisa, 1969), riporto il più oggettivamente possibile le affermazioni degli autori, senza una valutazione critica: osservo una volta per tutte che i problemi linguistici, teorici e particolari, coinvolti, sono talora espressi dagli autori con termini gravemente approssimativi e inesatti, e che spesso si tratta più di progetti e studi preliminari che di realizzazioni avanzate. La scelta delle comunicazioni che cito è episodica, ha valore di esempio e non di giudizio.

Linguistica teorica e algebrica, suddivisa in: *linguistica algebrica, procedure e teorie del parsing* (1), *teoria della grammatica*. Nella maggior parte delle comunicazioni sono stati discussi e confrontati tra loro diversi tipi di grammatiche generative denominate e classificate secondo le distinzioni ormai classiche di N. Chomsky (*Formal Properties of Grammars*, in *Handbook of Mathematical Psychology*, a cura di R. D. Luce, R. R. Bush e E. Glanter, vol. II, New York, 1963, pp. 323-418): per usare la terminologia di N. Ruwet (rispettivamente al cap. 2° e 3° di N. Ruwet, *Introduction à la grammaire générative*, Parigi, 1967), sono stati esaminati sia modelli di « sintassi elementari » sia « modelli sintagmatici ». Solomon Marcus ha introdotto un tipo di grammatica generativa contestuale, che utilizza la teoria degli insiemi; sembra confermare così, indirettamente, le affermazioni di Chomsky, riprese da F. Kiefer (*Some Aspects of Mathematical Models in Linguistics*, « Statistical Methods in Linguistics », III, 1964, pp. 8-26), secondo le quali i modelli teorici formulati in termini di teoria degli insiemi sarebbero riconducibili a qualche tipo di modello chomskiano. A V. Gladky e I. A. Mel'cuk hanno suggerito un nuovo genere di grammatica formale (grammatica **Δ** = **δένδρον**, albero), la quale per molti rispetti è collegata con le grammatiche di

---

(1) Trovare 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 che analizza automaticamente una frase, comunemente partendo da una sequenza di descrizioni ottenuta consultando un dizionario registrato su disco o nastro magnetico.

Chomsky, ma ne differisce perché intende elaborare degli « alberi » (nel senso della teoria dei grafi), e non « sequenze » (*strings*): essa non è concepita per generare delle frasi, ma per trasformare alberi di struttura data in altri alberi, a diversi livelli di profondità. A. Joshi (Filadelfia), studia una classe di grammatiche con un tipo misto di regole: una grammatica di un solo stile (cioè il carattere formale delle regole) non sarebbe capace di rappresentare i vari aspetti della struttura delle lingue naturali. J. Mey, in polemica con i cecoslovacchi P. Sgall e E. Hajicova, ha confrontato, con un modello trasformazionale, un modello recente orientato verso la nozione di « livello linguistico », concepito nel quadro dello strutturalismo funzionale di tipo europeo. Molte comunicazioni hanno messo in discussione la distinzione tra semantica e sintassi, riflettendo così le tendenze affermatesi tra i cosiddetti linguisti postchomskiani; il problema della formalizzazione della semantica, o almeno della classificazione del lessico, è stato un *leit-motiv* che ha percorso tutte le sezioni del convegno. F. Kiefer ha discusso le due alternative: complicare le regole di sintassi o complicare le categorie del lessico. J. S. Petöfi ha discusso la compilazione e formalizzazione di un *thesaurus*. R. C. Schank e L. G. Tesler hanno descritto una *conceptually-oriented dependency grammar*, che lavora cioè con un sistema di regole il quale opera su categorie concettuali (del tipo: attore, azione, locazione, ecc.) le quali non corrispondono alle categorie sintattiche. La loro semantica concettuale consiste essenzialmente nelle liste dei dipendenti potenziali di ogni concetto dato.

Grammatica trasformazionale, suddivisa in *grafemica*, *fonologia* e *grammatica trasformazionale*. O. Fujimura e R. Kagaya hanno esposto un metodo per descrivere le costanti dei caratteri cinesi come modelli grafici: la descrizione può essere considerata come data in forma di grammatica generativa di forme. S. C. e C. W. Yang hanno descritto come un *plotter* possa essere programmato a funzionare da produttore universale di caratteri grafici, per tutte le lingue, anche quelle con scrittura non alfabetica. S. Braun investiga su base matematica le proprietà delle regole di ridondanza fonologica, nella prospettiva della teoria dei tratti distintivi binari di Jakobson e Halle, come è esposta, per es., in *Fundamentals of Language*. È stato seguito con vivo interesse un gruppo di comunicazioni che hanno descritto programmi generalizzati, concepiti per *valutare* le grammatiche generative e le regole di trasformazione. Già I. I. Revzin (*Les modèles linquistiques*, traduz. dal russo *Modeli Jazyka*, Mosca, 1961) aveva chiaramente espresso le possibilità offerte dal calcolatore per la verifica di modelli linguistici. J. Friedman ha constatato che scrivere una grammatica trasformazionale, sia pure per un solo frammento di lingua naturale, è compito di un elevato ordine di complessità: 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. Il programma scritto dalla *équipe* della Friedman

« rappresenta la metateoria linguistica »: genera frasi applicandole a un lessico, e verifica la coerenza formale e notazionale delle regole; il modello linguistico incorporato è la teoria della grammatica trasformazionale descritto da Chomsky nei suoi *Aspects of the Theory of Syntax*. Il programma è già stato usato per testare tre grammatiche dell'inglese moderno, una del swahili, una del francese (comunicazioni di A. Querido) ecc.: a Pisa, presso il CNUCE, verrà applicato ad alcuni frammenti di grammatica dell'italiano. V. A. Fromkin e D. L. Rice hanno descritto un programma analogo per verificare le regole del componente fonologico di una grammatica generativa, in particolare per convertire, tenendo conto delle condizioni contestuali, la struttura astratta di superficie di una frase nella sua rappresentazione fonetica. R. I. Binnik ha discusso la possibilità di applicare alla LC un modello della lingua basato su una modificazione e estensione della teoria « semantica generativa (trasformazionale) » della competenza linguistica recentemente sviluppata da P. M. Postal, G. Lakoff, J. R. Ross, J. D. Mc Cawley e altri. La comunicazione è stata particolarmente ricca di informazioni bibliografiche sulle più recenti teorie della semantica.

Analisi computazionale, a livello di lessico, di frasi, di testi. K. D. Büntig ha parlato di una ricerca empirica sui sistemi di derivazione, mediante affissi, delle parole tedesche contenute in un lessico ritenuto rappresentativo. E. Gammon vuole formulare una « procedura di scoperta » della parola come unità linguistica, di cui ha puntualizzato « lo stato paradossale » nella linguistica contemporanea. Il modello quantitativo proposto cerca dei corrispettivi statistici ai criteri della scuola distribuzionalista americana. Z. Bujas ha descritto un progetto di analisi contrastiva condotta sulle concordanze di 680.000 parole serbocroate e sulla loro traduzione inglese. In queste concordanze « contrastive » a ogni parola serbocroata è associato automaticamente sia il contesto serbocroato sia la frase inglese che lo traduce, e viceversa per le parole inglesi. Ovviamente la traduzione è avvenuta in modo che a una frase serbocroata corrispondesse una e una sola frase inglese, e le frasi sono state numerate progressivamente. Harry H. Josselson ha proposto alcuni metodi per organizzare un lessico su un nastro magnetico e per codificare le diverse caratteristiche morfologiche, sintattiche, stilistiche, semantiche, ecc. di ciascun lemma (2); il lemmario su cui lavora risulta dalla fusione dei

---

(2) Un « dizionario di macchina », oltre che come aiuto nella lemmazione di testi, può fungere come archivio, gestibile automaticamente, di conoscenze linguistiche. Progetti in questo senso sono in corso per l'italiano (A. Zampolli, *Le dictionnaire italien de machine*, « Calcolo », V, 2, 1968, pp. 109-126), per il francese (M. Gross, *Lexique des constructions complétives*, CNRS, 1969), per il ceco (J. Stindlova, *Le Dictionnaire de la langue tchèque littéraire : enregistrement des données sur cartes méchanographiques*, « Cahiers de Lexicologie », X, 1967, pp. 103-113), ecc.

principali dizionari della lingua russa. A. Livonen ha confrontato diversi tipi di strategie per il riconoscimento automatico dei « suoni del discorso » per mezzo di un calcolatore digitale. E. Glasersfeld e P. P. Pisani hanno esaminato gli aspetti teorici e computazionali di una classificazione semantica e sintattica delle parole, alla luce della grammatica di correlazione della scuola di S. Ceccato.

Linguistica diacronica e comparativa ([3]), *e Dialettologia.* E. A. Afendras ha analizzato i sistemi vocalici di numerose lingue balcaniche secondo i tratti distintivi di Jakobson. Egli ha discusso e valutato diversi metodi per misurare e comparare il grado di similarità dei sistemi, e ha proposto ulteriori ricerche per modelli matematici nello stesso settore, così da approfittare del potere esplicativo di una teoria matematica ben sviluppata. R. N. Smith ritiene che il calcolatore possa aiutare a verificare la grande quantità di dati e di regole che entrano in gioco negli studi di linguistica storica relativi alla struttura dei cambiamenti fonetici; con 21 regole applicate a 500 forme protoindoeuropee ricostruite tratte dal *Vergl. Worterbuch der Idg. Sprachen* di A. Walde e J. Pokorny egli ha ottenuto alcune forme russe corrispondenti, ma non tutte quelle che si attendeva. W. S. J. Wang ha esposto una ricerca analoga per la ricostruzione della evoluzione fonologica di alcuni dialetti cinesi. W. Skalmowski e M. van Overbeke hanno studiato statisticamente i fenomeni di interferenza lessicale in testi neerlandesi di autori bilingui (parlanti francese e neerlandese), confrontandoli con testi ritenuti rappresentativi del neerlandese scritto contemporaneo *standard*. W. N. Francis, J. Svartvik, G. M. Rubin, per studiare la situazione delle fricative iniziali di 10 contee del sud dell'Inghilterra, hanno elaborato con il calcolatore i materiali dialettali registrati nel 4° volume del « Survey of English Dialects ». Gordon R. Wood ha illustrato una ricerca statistica per la identificazione e lo studio dei diversi tipi di americano regionale lungo la costa dell'Atlantico ([4]).

([3]) Rinvio alla recente comunicazione che è stata tenuta da P. RAMAT (*L'indoeuropeo nella linguistica contemporanea*) in occasione del convegno annuale del Circolo Linguistico Fiorentino (1969), e al dibattito che ne è seguito, in special modo agli interventi di G. DEVOTO e di C. A. MASTRELLI, per il valore dei « modelli tipologici » in questo settore.

([4]) L'ALI ha da tempo iniziato esperimenti analoghi presso il CNUCE di Pisa: pensiamo però di usare, anziché il *plotter*, la *fotocomposizione* (già usata per stampare indici e concordanze dell'Accademia della Crusca) che assicura una maggiore velocità di stampa e una presentazione grafica nettamente superiore; questo metodo è stato scelto anche per l'ALALP. Contiamo di combinare le capacità di produrre grafici di recente acquisite dal calcolatore, con la possibilità di operare statistiche sui materiali registrati e di creare un archivio dialettale gestibile automaticamente.

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 (a Roma, a cura della « Società di Linguistica Italiana »; a Trieste, a cura del « Centro per l'insegnamento dell'italiano »). Si discute il ruolo della descrizione del lessico, e la distinzione tra sintassi e semantica (5) da un lato, e tra conoscenza linguistica e « conoscenza del mondo » dall'altro. I lavori più spesso citati sono quelli di J. A. Fodor, J. J. Katz, P. M. Postal, M. R. Quillian, G. e R. T. Lackoff, J. D. Mc Cawley, J. R. Ross, P. Kiparsky, e soprattutto Ch. J. Fillmore (per la bibliografia si veda, per es. *Universals in Linguistic Theory*, a cura di E. Bach e R. T. Harms, New York, Chicago, ecc., 1968; *Semantic Information Processing*, a cura di M. Minsky, The MIT Press, 1968); frequenti sono anche i riferimenti alla logica formale. Alcune comunicazioni hanno trattato solo problemi teorici, senza riferimento esplicito all'uso dei calcolatori. Altre lo hanno invece esplicitamente proposto per la compilazione e la gestione di *thesaurus*, o addirittura come tipo autonomo di approccio alla formalizzazione della semantica delle lingue naturali. È il caso di R. M. Schwarcz che ha esaminato questo problema con grande cura: egli ha discusso le diverse teorie linguistiche della semantica e i problemi logico-teorici connessi, ha passato in rassegna i progetti computazionali più significativi e ha proposto un tipo di approccio basato sulla teoria « operazionale » del significato. Di carattere teorico e generale sono state anche la comunicazione di I. Bellert e quella di J. Rouault, che applica categorie della logica alla semantica di lingue naturali. B. Vauquois *et alii* hanno proposto, 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 A. Dugas *et alii*, che praticano un'analisi dei lessemi in tratti semantici. E. Vasiliu studia la « categoria-tempo » e la sua interpretazione semantica riferendosi ai sistemi logici proposti da Carnap. D. Wunderlich si è riferito invece all'opera dei logici R. Montague e J. A. W. Kamp nella sua comunicazione « On Time Reference and Tense ».

Documentazione automatica. Secondo C. A. Montgomery, gli scienziati dell'informazione sono interessati al linguaggio per la sua funzione di veicolo primario di comunicazione di informazioni nella società umana,

---

(5) Il problema non è certo nuovo. Si veda, per es., G. Devoto, *Sémantique et Syntaxe*, *Conférences de l'Institut de linguistique de l'Université de Paris*, XI, 1952-53, pp. 51-62.

e ricercano un sistema automatico per analizzare il contenuto di testi naturali. Il sistema ideale dovrebbe identificare i concetti presenti nel testo e le loro interrelazioni, sulla base di una qualche forma di analisi sintattica e semantica (questo è il punto di convergenza con gli interessi dei linguisti), e tradurli poi in un insieme di « frasi canoniche » indipendenti da una lingua particolare; esse costituirebbero, per così dire, le « conoscenze » del sistema, e servirebbero di base per generare risposte fattuali a richieste specifiche di informazione. Quanto si sia lontani da questo ideale è dimostrato, se non altro, dal fatto che i maggiori successi nel settore della *information retrieval* li hanno ottenuti, come ha osservato D. G. Hays (*Applied Computational Linguistics*, Cambridge U. K., 1969), 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 che ha descritto un sistema « conversazionale » *on-line* utente-calcolatore, di A. W. Pratt e M. G. Pacak che lavorano all'analisi automatica di documenti medici, e di Szanser A. J., che ha approntato un metodo per la correzione semiautomatica di testi perforati. Il problema della correzione dell'*input* è fondamentale per la LC, soprattutto per progetti che prevedono l'elaborazione di grandi *corpus*, come per es. quello per il Vocabolario Storico della Lingua Italiana dell'Accademia della Crusca.

A. Zampolli