LINGUISTICA COMPUTAZIONALE E DIDATTICA DELLE LINGUE: I DIZIONARI ELETTRONICI

Antonio Zampolli

L'istituto di linguistica computazionale del CNR (Università di Pisa) che io dirigo non si occupa direttamente di creare sistemi e/o strumenti per l'insegnamento. Tuttavia, lavorando nel settore della linguistica computazionale, siamo interessati ad avere la collaborazione di insegnanti e ricercatori impegnati nel settore della didattica per cercare di valutare se i dati, gli strumenti e i metodi che stiamo sviluppando con la ricerca che facciamo, possano essere di qualche utilità al mondo della scuola. Il mio sarà, pertanto, un intervento di tipo informativo su quei settori della nostra attività che ritengo possano avere uno sbocco applicativo nella didattica delle lingue.

Il CNR ha lanciato ultimamente una serie di progetti, chiamati *progetti strategici*, ed io sono responsabile del coordinamento del *progetto strategico nazionale* per il trattamento automatico del linguaggio naturale.

Fra gli obiettivi che ritengo più vicini agli interessi di questo convegno:

- la creazione di un *corpus* di riferimento italiano;
- l'attuazione di una stazione di lavoro linguistica;
- la creazione di una base lessicale-terminologica di dati mono e multilingui;
- la creazione di strumenti per l'analisi automatica, semantica e sintattica dell'italiano, ecc.

Tra i vari progetti in atto, c'è quello di un analizzatore sintattico e semantico italiano che sarà collegato, nell'ambito di un progetto della CEE, a quelli per tutte le altre lingue della Comunità.

Costruire un analizzatore vuol dire scrivere esplicitamente una grammatica, scrivere un componente lessicale, morfologico, con i quali il calcolatore automaticamente analizza e genera frasi.

Questo, per esempio, credo che potrebbe essere utilizzato per l'inse-

gnamento della grammatica di una lingua a un livello abbastanza alto. Non lo facciamo solo per le lingue contemporanee, lo facciamo, per esempio, anche per il latino, ed è soprattutto nell'insegnamento delle lingue classiche che i nostri sistemi hanno trovato già delle applicazioni. Un altro tipo di progetto, di attività, che noi abbiamo è quello di mettere a disposizione dei ricercatori delle università italiane e straniere, un insieme di procedure generalizzate per lo spoglio elettronico di testi nella loro forma più semplice, oramai di *routine*. Non c'è più nulla di nuovo nelle procedure con le quali è possibile introdurre nel calcolatore delle opere di qualsiasi lingua ed epoca per fare concordanze, indici di frequenze, studi statistici, ecc.

Dato che queste procedure sono utilizzate da una cinquantina di istituti in Italia ed anche da molti istituti italiani all'estero, abbiamo costituito una specie di biblioteca elettronica in 32 lingue che contiene più di 5000 opere. La strategia che seguiamo ora è quella di concentrarci sull'italiano e di creare una *rete di centri* europei con i quali siamo collegati tramite un *network*: qualsiasi ricercatore italiano ha, tramite noi, direttamente accesso ai *corpora* registrati in altri centri specializzati europei. Con la collaborazione del Consiglio d'Europa abbiamo anche in corso un progetto di costruzione di *corpora* di riferimento per le varie lingue.

Cominceremo con le lingue romanze, ma si sono già associati anche gli inglesi.

Il concetto di *corpora* di riferimento è chiaro: si tratta di raccogliere in forma leggibile per il calcolatore, e di analizzare a molti livelli linguistici, testi che siano rappresentativi della lingua contemporanea nei suoi vari aspetti e stratificazioni (parlato, scritto, vari generi letterari, ecc.).

Questi *corpora* saranno costruiti in questo modo: ci saranno opere italiane registrate solo in italiano, opere francesi in francese, opere inglesi in inglese, spagnole in spagnolo, ecc. Una parte di questi *corpora* sarà, però, costituita da traduzioni, in modo tale che ci sia la possibilità di fare degli studi contrastivi sulle varie lingue. La Comunità Europea, per esempio, sta per metterci a disposizione un insieme enorme di opere e di testi; essa deve, infatti, per legge pubblicare tutto nelle varie lingue della comunità. Questi testi vengono prodotti con i sistemi di *word processing* e sono, quindi, già in forma leggibile dal calcolatore. Sarà così possibile accedere a un testo in italiano ed avere immediatamente tutti i corrispondenti nelle altre lingue.

Noi crediamo che ciò potrebbe essere utilizzato nell'ambito dell'insegnamento. In collaborazione con l'Istituto Italiano di Cultura di Barcellona, stiamo cercando di realizzare un sistema di questo tipo, vale a dire di avere la possibilità di disporre di un corpo di testi italiani per l'insegna-

mento dell'italiano all'estero. Lo stesso si potrebbe realizzare per lo studio delle lingue straniere in Italia. Tali procedure sono utilizzate oggi da linguisti, da letterati, ecc.; ma noi ci chiediamo se l'insegnante da un lato e lo studente dall'altro non potrebbero utilizzare questi sistemi. Uno studente, per esempio, potrebbe sedersi davanti a un terminale e chiedere al calcolatore (che gli risponde subito) una parola ed avere tutti i contesti in cui la parola si trova; oppure chiedere un lemma e tutti i contesti in cui il lemma si trova; oppure una parola seguita da un'altra parola. Potrebbe anche chiedere al calcolatore di fornire quali sono le combinazioni più frequenti di una parola con altre parole, oppure, a livelli forse più interessanti, proporre al calcolatore una o più famiglie di parole e chiedere tutti i luoghi nel testo in cui qualsiasi parola di una famiglia occorre con un'altra famiglia di parole.

Se, per esempio, uno vuole studiare che rapporti ci sono in *Gide* tra il concetto di luce e il concetto di tristezza o il concetto di malattia, lo studente può fornire al calcolatore le parole che in quella lingua, secondo lui, rappresentano quei concetti e può chiedere in quali appare una parola che spiega questo concetto.

Queste procedure possono essere oggi potenziate tramite un *vocabolario di macchina*. Concetto che spiegheremo subito e che costituisce il secondo settore in cui crediamo che i nostri sistemi siano più direttamente utilizzabili per scopi didattici.

Cosa intendiamo per vocabolario di macchina? Oggigiorno i vocabolari non sono più stampati in piombo, ma sono registrati su calcolatore e stampati.

Con i fondi stanziati dal governo inglese e dal governo canadese stanno, per esempio, traducendo tutto l'*Oxford English Dictionary* su calcolatore e vi è anche l'intenzione di metterlo a disposizione del mondo della ricerca su *compact disk*.

Stanno, quindi, creando dizionari di vario tipo — dai dizionari storici ai dizionari per l'insegnamento della lingua, ai dizionari bilingui — leggibili dal calcolatore.

A Pisa, per esempio, abbiamo stipulato accordi con varie case editrici e ci troviamo ad avere disponibili nel calcolatore dei vocabolari monolingui e bilingui. Questi vocabolari, per ora, sono semplicemente la trascrizione del testo leggibile dal calcolatore. Noi, invece, abbiamo delle procedure per trasformare questi vocabolari in quelli che noi chiamiamo *basi di dati lessicali*. Naturalmente noi siamo interessati a fare ciò.

Un vocabolario stampato in fotocomposizione contiene solamente le informazioni che appaiono scritte; invece noi le strutturiamo in una gerarchia di campi (dell'etimologia, ecc.). Ovviamente facciamo questo perché

le applicazioni sono di vario tipo: traduzione automatica, linguaggio uomo-macchina, documentazione automatica, ecc.; ma una delle possibili applicazioni è anche quella di mettere a disposizione di qualsiasi utilizzatore — ricercatore, insegnante, studente, ecc. — un vocabolario sul calcolatore. Molte case editrici, del resto, sono orientate a mettere sul mercato opere di tipo enciclopedico su *compact disk*. Noi crediamo che l'avvenire sia in queste cose e crediamo non solo che esse siano di grande utilità per la ricerca, ma anche che dovrebbero rendere migliore, più potente, più intelligente e creativo l'accesso alle informazioni da parte di qualsiasi utente.

Ci domandiamo, in particolare, se l'utente non potrebbe essere l'insegnante o lo studente.

Che cos'è, dunque, una *base di dati lessicali?* È un insieme di informazioni lessicali strutturate e accessibili *on line*. Una *base dati* è incrementabile a piacere; e non esistono i problemi di spazio tipici della stampa.

Oggi è possibile accedere al vocabolario solamente con la parola, se non si sa la parola non si accede.

La ricchezza di informazioni che c'è in un vocabolario è incredibile, siano esse informazioni esplicite o informazioni implicite.

Con il calcolatore chiaramente si può accedere al vocabolario da qualsiasi unità, attraverso una qualsiasi delle informazioni che il vocabolario contiene. Potrei, per esempio, chiedere: dammi tutte le parole che finiscono con questo suffisso e nella cui definizione c'è la parola "atto" oppure c'è la parola "strumento". Non solo, ma, se, per esempio, si opera sulle definizioni, è possibile strutturare il vocabolario nelle tassonomie che contiene, cioè rendere espliciti tutti i rapporti di iperonimia e di iponimia. Noi abbiamo strutturato tutto il vocabolario di italiano e nostri colleghi lo stanno facendo per l'inglese. Lo studente potrebbe, pertanto, chiedere al calcolatore di fornire l'iperonimo "tessuto" e chiedere tutti i termini che hanno nella definizione "tessuto". Il vocabolario di italiano che abbiamo noi ne contiene fra i 900 e i 1000: lino, lana, seta, ecc. È di notevole utilità poter interrogare il vocabolario per campi lessicali: per esempio, tutti i termini di suono, tutti i termini di colore, ecc. Il calcolatore può, inoltre, proiettare tutte queste famiglie, questi campi semantici sui testi.

Quindi, mi posso sedere davanti ad un terminale e dire: dammi tutti i termini di colore. La richiesta va nel vocabolario e, con le procedure di accesso che già abbiamo, identifica tutti i termini che hanno a che fare con colore, va nei testi e fornisce tutti i contesti in cui appaiono, ecc. Il concetto è quello di avere un vocabolario che non è più qualcosa di fisso, ma di flessibile, ed al quale si può accedere da diversi punti.

Naturalmente — come fanno rilevare le case editrici orientate a produrre su *compact disk* — occorre mettere ben a punto procedure intelli-

genti per consultare tale vocabolario e sfruttare adeguatamente il fatto che non è più un oggetto statico, ma dinamico e flessibile.

Sarebbe, però, opportuno poter collaborare anche con gli insegnanti in modo che le procedure che noi già stiamo realizzando con alcune case editrici siano pensate e tagliate anche secondo le esigenze dell'insegnamento. Noi ci interessiamo anche molto al fatto di riuscire a collegare i vocabolari monolingui attraverso dei vocabolari bilingui. È già in atto il progetto di collegare una *base dati* "italiano" con una *base dati* "inglese" attraverso un vocabolario bilingue.

Con il vocabolario bilingue basta che si trovi un termine ed il calcolatore va direttamente al monolingue inglese e stabilisce tutta una serie di relazioni.Vi accedo con un termine e trovo una famiglia di parole, un campo semantico, l'insieme delle parole che derivano da un'altra parola, tutti i sinonimi, tutti gli iperonimi, gli iponimi ecc.

Non sappiamo valutare se strumenti di questo tipo siano utilizzabili anche per l'insegnamento. D'altra parte ci rendiamo conto che Comunità Europea, Consiglio d'Europa, *European Science Foundation*, grosse ditte di calcolatori e case editrici stanno facendo uno sforzo grandissimo e stanno investendo ingenti somme di danaro nell'industria delle lingue.

Ci chiediamo: Quale profitto può trarre la scuola da tutto ciò? Che tipo di collaborazione può offrire? Quale spazio è opportuno riservarle?