FONDAZIONE ANGELO RIZZOLI
6

IL GIORNALE E IL NON-LETTORE

ATTI DEL CONVEGNO DEL 15/17 GIUGNO 1979 ORGANIZZATO DA WALTER TOBAGI

A cura di Walter Tobagi e Carlo Remeny

Sansoni Editore

ANTONIO ZAMPOLLI *

LE PAROLE DEI TITOLI (I)

Quando, verso la fine di aprile, la Fondazione Rizzoli chiese la nostra collaborazione per lo spoglio del *corpus* dei titoli, accettammo senza esitazioni.

Se da un lato, infatti, l'impiego delle metodologie automatiche da noi approntate per le analisi lessicali e statistiche ha consentito alla Fondazione di ottenere una prima — sia pur provvisoria — serie di risultati in tempo utile per la presentazione a questo convegno, dall'altro ci è stata offerta la opportunità di arricchire la nostra « banca » di dati linguistici (cioè l'insieme di testi — circa 5000 in più di 20 lingue — registrati a Pisa su memorie elettroniche) con materiale di un settore (i titoli dei giornali) non ancora adeguatamente rappresentato e, quel che più conta, con l'aiuto degli esperti del settore in questione.

In particolare, il *corpus* dei titoli continua il processo di completamento e di stratificazione, con nuovi e diversi « generi », del *corpus* di italiano contemporaneo che è alla base di numerosi progetti di ricerca, soprattutto nel settore della statistica linguistica.

Prima di entrare in qualche dettaglio sul progetto dei « titoli », del quale si presentano qui i primi risultati, conviene forse delineare il quadro generale delle ricerche entro le quali viene a collocarsi.

Definire in poche parole la linguistica computazionale — che i francesi, con un termine forse più chiaro chiamano « trattamento

* La relazione di Antonio Zampolli, momentaneamente assente, è stata letta da Alberto Angelucci dell'IBM.

automatico della lingua » — non è certo agevole. Sembra preferibile indicare — a titolo di esempio — alcune delle attività più comuni in questo settore.

L'elaborazione di *corpora* — a volte molto estesi — di testi si propone molto spesso di rendere più agevole ed economico il reperimento e lo studio dei materiali linguistici in essi contenuti, soprattutto quando è importante l'analisi esaustiva di grandi quantità di dati.

Indici alfabetici, diretti e inversi, indici di frequenza, rimari, incipitari, concordanze, elenchi di strutture e tipi sintattici ecc., costituiscono la documentazione più tradizionale fornita dal calcolatore. Oggi si va anche verso la costituzione di vere e proprie « banche di dati linguistici », biblioteche elettroniche opportunamente organizzate nella memoria del calcolatore, corredate da strumenti di analisi automatica (dizionari-macchina, analizzatori sintattici, « thesauri » di vario tipo, ecc.), con le quali l'utente possa « dialogare », a esempio, per mezzo di un terminale, per individuare ed estrarre i materiali di volta in volta rilevanti per le proprie ricerche.

Gli esempi di utilizzatori sono i più diversi. Ne cito alcuni tra i progetti degli Istituti italiani e stranieri con i quali collaboriamo. Lessicografi per la redazione di dizionari storici, sincronici, di lessici di autore, di lessici terminologici, ecc. Filologi e storici della lingua per la ricerca di attestazioni particolari, per la preparazione di edizioni critiche, per il confronto di redazioni o versioni diverse di un testo. Grammatici per l'inventario di forme, strutture sintattiche. Storici, per il recupero e la classificazione delle informazioni contenute nelle fonti testuali. Dialettologi e sociolinguisti per l'ordinamento e lo spoglio di materiali raccolti con le inchieste sul campo. Studiosi di tradizioni popolari, per repertoriare elementi, per identificare strutture compositive. Studiosi di stilistca, psicolinguisti, glottodidatti, e quanti altri sono interessati a rilevare le frequenze d'uso delle diverse unità linguistiche.

Ricerche di questo tipo richiedono procedure, programmi, strumenti linguistici, generali o specifici per ogni lingua, che si fondano su teorie o modelli più o meno formalizzati. Alcune ricerche, anzi, si propongono innanzitutto la formalizzazione e la verifica dei modelli, per esempio si cerca di incrementare nel calcolatore modelli e regole a diversi livelli (morfologico, lessicale,

sintattico, semantico, ecc.), e si chiede al calcolatore di aiutare il linguista nello studio delle caratteristiche del modello, della coerenza della notazione, della esaustività delle regole. Si memorizzano lessici e grammatiche, si costruiscono dei *parser*, che dovrebbero analizzare automaticamente parole e frasi di una lingua storico-naturale, fornendo per le prime l'analisi morfologica e il lemma, per le seconde una rappresentazione formalizzata della loro struttura sintattica e, al limite, del loro « significato ». Alcune scuole si spingono fino a considerare questi sistemi come modelli che potrebbero aiutare — mediante simulazione — lo studio dei processi mentali del parlante.

Nel nostro Paese, nell'ultimo decennio, si sono moltiplicati i progetti in questo settore. In particolare presso il Laboratorio di linguistica computazionale, sono stati messi a punto programmi e metodologie per i principali settori della linguistica computazionale. I progetti in corso riguardano dizionari di macchina per l'italiano, lo spagnolo, il latino, analizzatori sintagmatici e sintattici per le prime due lingue, modelli semantici, morfologici, fonetici.

Per il lavoro che viene qui presentato abbiamo utilizzato le procedure approntate dal Laboratorio per il *test processing* e per l'analisi statistica dei risultati degli spogli.

Il *corpus* campione, scelto dalla Fondazione Rizzoli, è costituito da circa 50.000 parole, proveniente dai titoli di 20 testate, suddivise in 4 gruppi:

1. quotidiani
2. settimanali politico-culturali
3. settimanali familiari
4. settimanali popolari

Per ognuno dei settimanali si è preso in considerazione il primo numero di ogni mese per gennaio, febbraio, marzo e aprile 1979; per i quotidiani, 40 prime pagine (una ogni tre giorni) dal 2 gennaio al 30 aprile 1979.

Come parole componenti i titoli, si sono considerate sia quelle dei titoli propriamente detti, sia quelle degli occhielli.

Naturalmente, titoli e occhielli sono distinti da opportuni contrassegni, e quindi possono, eventualmente, essere elaborati anche separatamente.

Dopo una rapida operazione di preedizione, con la quale sono stati identificati i titoli da elaborare e sono stati contraddistinti

i nomi propri, i titoli sono stati ricopiati su schede perforate e memorizzati nel calcolatore. Il breve tempo a disposizione ci ha costretti a omettere l'operazione di controllo della ricopiatura, e ciò spiega la presenza di alcuni errori negli elenchi forniti dall'opuscolo.

La serie di programmi generalizzati di spoglio disponibili presso il Laboratorio ed eseguiti sui calcolatori IBM 370/158 e 370/168 dal CNUCE di Pisa, ha prodotto i seguenti risultati:

a) indici alfabetici, e per frequenza decrescente, delle forme presenti nel *corpus*, nelle singole testate, nei 4 sottogruppi;
b) indici delle forme ordinati secondo alcune misure statistiche comunemente adoperate dalla statistica linguistica. In particolare, sono qui disponibili gli indici di *dispersione* e di *uso*;
c) « concordanze » delle forme, e cioè l'elenco delle forme che compaiono nei titoli, ciascuna seguita da tutti i contesti (in pratica i titoli) nei quali essa appare, e dalla indicazione della testata del numero e della pagina dal quale il titolo è tratto.

Tutti questi risultati verranno ottenuti a breve scadenza anche per i *lemmi*, oltre che per le forme. Il tempo disponibile non era sufficiente, infatti, per eseguire la operazione di « lemmatizzazione ». Questa operazione consiste nel distinguere, mediante l'esame dei contesti, i diversi significati di una forma omografa nel sistema lessicale italiano (per esempio *conquista* verbo e sostantivo) e nel ricondurre le diverse flessioni e varianti di un paradigma alla loro forma di base (il lemma): per esempio: *amo*, *amava*, *amerai*, *amò*, ecc. al lemma *amare*.

La lemmatizzazione è forse meno importante nei titoli che nei testi ma appare in ogni caso indispensabile. Si considerino per esempio, nel fascicolo distribuito, *affrontare*, *ammazzare*, *illusione*. Essi non compaiono nell'elenco delle forme più frequenti perché nessuna delle loro forme, presa separatamente, raggiunge il limite inferiore fissato.

Il tempo richiesto per la lemmatizzazione deriva dal fatto che questa operazione non è ancora interamente automatizzabile. Il Laboratorio ha già costruito per l'italiano un dizionario macchina. Esso consiste, essenzialmente, in 120.000 lemmi italiani registrati nella memoria del calcolatore, corredati da informazioni linguistiche di diverso genere (fonetico, morfologico, sintattico, semantico, ecc.). Un insieme di regole morfologiche consente al calcolatore di produrre automaticamente tutte le diverse

possibili flessioni (circa un milione) di questi lemmi, e, viceversa, di riconoscere nelle parole dei testi, mediante un opportuno algoritmo di analisi, i rispettivi lemmi. Il dizionario macchina consente quindi la esecuzione automatica della lematizzazione, fatta eccezione per la distinzione degli omografi. Anche con questa limitazione, il risparmio dei tempi e dei costi è di circa il 75%. Per la verità stiamo cercando (e non solo per l'italiano ma anche per lo spagnolo) di « insegnare » al calcolatore a distinguere automaticamente gli omografi (o almeno la maggior parte di essi) sulla base di una analisi automatica del contesto: ma non è questa la sede per discutere il problema.

Prospettive di ricerca

Gli ideatori del progetto si sono proposti, come è naturale, obbiettivi specifici relativi all'ambito giornalistico. Il nostro gruppo vede in questo progetto alcune interessanti prospettive di ricerca.

Il *corpus* dei titoli acquisito grazie alla Fondazione Rizzoli e alla IBM va a integrare, come si è detto, l'archivio di materiali linguistici registrati in *machine readable forme* che la collaborazione di numerosi Istituti italiani e stranieri (prima fra tutti l'Accademia della Crusca) ha contribuito a costituire a Pisa presso il nostro Laboratorio.

Il *corpus* dei titoli è stato registrato per il calcolatore ed elaborato secondo gli *standard* e i criteri di analisi proposti dal Laboratorio, che vengono oggi adottati dalla quasi totalità dei progetti, in corso in Italia, che impiegano i calcolatori nel settore linguistico-filologico.

Questa uniformità di metodi, che anche i Paesi tecnologicamente più avanzati ci invidiano, presenta alcuni notevoli vantaggi sul piano scientifico ed economico.

Essa consente di considerare ogni nuovo *corpus* come un sottoinsieme che completa i *corpora* preesistenti nell'archivio e che può essere immediatamente confrontato con essi.

Inoltre i programmi già disponibili possono essere immediatamente applicati con evidente risparmio di tempi e di costi.

L'idea di contare le frequenze del discorso è molto antica: già gli Alessandrini compilavano elenchi degli *apax*. Verso la fine del secolo scorso si moltiplicarono spogli di frequenze grafemiche e fonetiche, intesi a facilitare la decifrazione di codici, la

razionalizzazione delle tastiere delle macchine da scrivere e i sistemi di dattilografia.

Tra le due guerre mondiali si moltiplicarono gli spogli di testi, soprattutto nei Paesi anglosassoni per la creazione di dizionari di frequenza orientati all'insegnamento delle lingue straniere. Negli anni '50, per merito soprattutto di Pierre Guiraud e Gustav Herdan, la statistica linguistica cercò i propri fondamenti teorici nei principi della linguistica strutturale. Negli anni '60 l'utilità dei metodi statistici fu messa in discussione da più parti: in particolare, in linguistica, dalle scuole generativo-trasformazionali.

Oggi la statistica linguistica conosce un periodo di rivalutazione. I metodi statistici vengono usati per costruire modelli teorici, per ricerche stilistiche (soprattutto nello studio del vocabolario), per identificare autori anonimi, per definire la cronologia relativa di opere diverse di uno stesso autore, ecc.

In Italia, devono essere ricordati gli importanti contributi di alcuni insigni linguisti, quali Marcello Durante, Luigi Heillman, Luigi Rosiello, Tullio De Mauro.

Nei primi anni '70 sono usciti due dizionari di frequenza dell'italiano: il *Frequency Dictionary of Italian Words* di Juilland e Traversa, relativo al periodo tra le due guerre, e il LIF (*Lessico di Frequenza dell'Italiano contemporaneo*) di Bortolini, Tagliavini, Zampolli, relativo agli anni 1948-68.

Entrambi questi dizionari contengono 5000 lemmi identificati come i più usati in un *corpus* di 500.000 occorrenze, considerato come rappresentativo, entro certi limiti, dell'italiano scritto.

I dati forniti da questi dizionari hanno prodotto conoscenze interessanti sulle strutture quantitative dell'italiano. Si è riscontrato per esempio che anche la frequenza delle parole grammaticali (congiunzioni, preposizioni, avverbi, pronomi, ecc.) che venivano di solito ritenute stabili, e cioè indipendenti dal tipo di lingua (generi letterari) variano invece sensibilmente tra i cinque sottoinsiemi in cui è suddiviso il LIF: romanzi, testi scientifici, giornali, cinema, teatro.

Per esempio, l'articolo *il* è usato nei giornali cinque volte più che nel cinema, tre volte più che nei romanzi. Diverse misure statistiche delle strutture quantitative hanno mostrato che i sottoinsiemi del LIF possono esser ordinati significativamente secondo la sequenza: cinema, teatro, romanzi, testi scientifici, giornali. In altre parole le strutture linguistiche del teatro sarebbero

dal punto di vista quantitativo più simili a quelle del cinema, i giornali sarebbero invece i più diversi, e così via. Ciò che ci interessa è individuare strutture quantitative « caratteristiche » dei vari « generi » letterari, caratteristiche che possono aiutare a descrivere quelli che percepiamo come fatti di stile.

In questo quadro, ci ripromettiamo di applicare al *corpus* dei titoli tutta una serie di misure e di analisi statistiche, considerandolo da due punti di vista:

a) come un *corpus* suddiviso in sottoinsiemi (le diverse testate, i quattro gruppi, ecc.) allo scopo di mettere in evidenza ciò che è comune e ciò che è diverso tra questi sottoinsiemi, e perciò « tipico » di ciascuno;
b) come un *corpus* da confrontare globalmente con altri *corpora* disponibili nella nostra « banca ». In particolare, con il *corpus* del LIF, considerato nella sua globalità e nei suoi sottoinsiemi: con il *corpus* costituito da circa 350.000 parole tratte dai 10 quotidiani di maggiore diffusione nel 1974 (tesi di M. Crocetti); il *corpus* di italiano parlato raccolto da B. Marioni (50.000 parole registrate in negozi, case private, trasmissioni radiofoniche e televisive); il *corpus* di giornali lombardi della prima metà dell'Ottocento (raccolto in collaborazione con M. Ciccone della British Columbia University di Vancouver).

Ci proponiamo di condurre le ricerche in questione, non solo a livello lessicale, ma anche ad altri livelli: grafemico, fonologico, grammaticale, sintattico.

Riteniamo che sarebbe interessante ripetere, se possibile, la campionatura dei titoli a intervalli regolari di tempo (uno o due anni) per un certo periodo.