Giuseppe Gigliozzi

Introduzione all'uso del computer negli studi letterari

A cura di Fabio Ciotti

Bruno Mondadori

BIBLIOGRAFIA CONSIGLIATA

Abercrombie J.R., *Computer Programs for Literary Analysis*, University of Pennsylvania Press, Philadelphia 1984.

Avalle D'Arco S., *Il lessico italiano delle origini e l'informatica linguistica*, in Fattori M., Bianchi M.L. (a c. di), *Ordo. Atti del II colloquio internazionale del lessico intellettuale europeo*, Edizioni dell'Ateneo & Bizzarri, Roma 1979, pp. 749-760.

Busa R., *Fondamenti di informatica linguistica*, Vita e Pensiero, Milano 1987.

Savoca G. (a c. di), *Lessicografia, filologia e critica*, Olschki, Firenze 1986.

Zampolli A., *Problemi di linguistica applicata*, s.e., Pisa 1974.

Zampolli A., Cappelli A. (a c. di), *The Possibilities and Limits of the Computer in Producing and Publishing Dictionaries*, Giardini, Pisa 1984.

3.3.1 Le concordanze, gli indici e le frequenze

Varrà la pena di aprire una piccola parentesi sul problema degli spogli lessicali e del trattamento del testo con metodi statistici, anche se così facendo ci spingeremo in territori che non sono di esclusiva pertinenza dell'informatica (infatti, praticamente tutte le applicazioni che verranno trattate possono essere, con un po' di pazienza, fatte "a mano"). Questo settore, però, è quello che storicamente ha aperto la strada all'uso dell'informatica nello studio della lingua, tanto che l'estrazione delle varie occorrenze dalla loro distribuzione naturale nella pagina, assieme alla registrazione degli indici numerici che è possibile ricavare dal testo, si pongono come primo passo in quasi tutte le applicazioni informatiche nelle discipline linguistiche e letterarie.

Le concordanze

Le concordanze sono, quindi, l'elenco delle parole contenute in un testo ordinate alfabeticamente, accompagnate da alcuni riferimenti che ne rendono possibile il ritrovamento nell'originale e seguite da un contesto che aiuta a interpretare il significato dell'occorrenza esaminata.

Ovviamente il primo serio problema che si incontra nella realizzazione delle concordanze, una volta effettuata o meno la selezione delle parole da espungere o da includere, è proprio quello della dimensione del contesto.

L'uso corrente è di utilizzare il verso in caso di testi di poesia e la riga tipografica nel caso della prosa, ma questo tipo di contesto, anche se comodo per una facile consultazione ed economico dal punto di vista delle dimensioni delle concordanze, spesso può risultare scarsamente significativo. Immaginiamo quanto poco potrebbe dirci sull'ultima parola di una riga.

Meglio sarebbe usare l'intera frase o una porzione di testo co-

munque valida da un punto di vista semantico, magari definendo il contesto, anziché con una lunghezza fissa, con i segni di interpunzione (secondo una precisa gerarchia che tenga conto del diverso peso, per esempio, del punto rispetto alla virgola) o con un segno di codifica appositamente individuato. In questo modo sarebbe possibile presentare anche contesti fortemente significativi: basti pensare a quell'unità genetica costituita dalla battuta di un personaggio di un lavoro teatrale.

Così facendo, però, è facile ottenere un contesto lungo anche decine di righe per parola, provocando un'esplosione delle dimensioni dell'output che renderebbe la concordanza scarsamente maneggevole. In ogni caso la selezione del contesto sarà condizionata anche dal supporto che è destinato a contenere le concordanze (se pensiamo al libro è un conto, se pensiamo al cd-rom è un altro) e bisognerà, inoltre, decidere in base allo scopo finale della concordanza (uso privato a scopo di ricerca, tesi di laurea o pubblicazione).

I formati *Kwic* e *Kwoc*

I moderni programmi dedicati alla realizzazione delle concordanze offrono di solito la possibilità di scegliere tra più forme di output che possono essere divisi in due grandi famiglie: le concordanze di tipo *Kwic* e le concordanze di tipo *Kwoc*.

Nella forma *Kwic* (*Key Word in Context*) le forme vengono allineate tutte a partire da una stessa colonna di stampa. Le forme, cioè, possono essere sia centrate, sia allineate al margine destro o sinistro. L'allineamento delle forme aumenta significativamente la leggibilità della concordanza, soprattutto quando si cercano delle co-occorrenze, che risaltano immediatamente, ma ne restringe le possibilità di impiego ai casi in cui il contesto desiderato sia breve, al massimo una riga. Come è facile immaginare osservando l'esempio seguente, al di sopra di una sola riga di contesto la leggibilità delle concordanze in formato *Kwic* diventerebbe veramente scarsa.

Riferimenti			Contesto a sinistra	Parola	Contesto a destra
Proemio	1	40	fra' quali se	alcuno	mai n'ebbe bisogno o
Proemio	1	17	piacevoli ragionamenti d'	alcuno	amico e le sue
Proemio	1	40	a' quali fa luogo,	alcuno	alleggiamento
Proemio	1	69	Rimuoverlo almeno per	alcuno	spazio di tempo,

Nella forma *Kwoc* (*Key Word Out of Context*) le forme normalmente non vengono allineate (anche se è sempre possibile prendere il meglio di ogni formato e ottenere un output misto). La *Key Word* viene collocata esternamente, come esponente, e poi

riportata all'interno del contesto nella posizione indicata dal tipo di formato scelto. La concordanza *Kwoc* è sostanzialmente un *indice* a cui viene associata una riga di contesto. Questa è la forma migliore nel caso di contesti ampi, sviluppati anche su più righe.

Esponente	Riferimento al testo			
Abbian	*Decameron* (proemio)			
	le quali quanto più di forza	> abbian <	che le palesi	1.49
	Contesto a sinistra	Parola	Contesto a destra	Rif.

Gli indici

Un indice, al contrario, può essere considerato un caso particolare di concordanza priva di contesto. L'indice delle parole (*index verborum*) o l'indice dei luoghi (*index locorum*) è una lista delle parole contenute in un testo o in un *corpus* testuale, accompagnata dai riferimenti ai luoghi in cui è possibile rintracciare le occorrenze individuali della parola.

In genere l'indice viene organizzato alfabeticamente in forma ascendente (dalla A alla Z), meno spesso in forma discendente (cioè al contrario). In alcuni casi possono poi essere convenienti organizzazioni particolari dell'indice. *Sortare* (ordinare alfabeticamente) un indice prendendo in considerazione le parole rovesciate (partendo dall'ultima lettera per arrivare alla prima) può risultare molto utile, come vedremo poi, per ottenere il cosiddetto "rimario". Quello che segue è un esempio di indice con organizzazione alfabetica:

Occorrenza	Riferimento brano	Riferimento luogo
ABBIAM	0001	49
ABISOGNA	0001	40
ACCESO	0001	7
ACCIO	0001	72
ADUNQUE	0001	72, 90
AFFANNO	0001	28, 50, 100
AFFLIGGE	0001	13
AFFLITTI	0001	1
AGLI	0001	47
AGO	0002	16
AL	0001	25, 81
ALCUNA	0001	57, 63
ALCUNE	0002	20
ALCUNI	0001	4
ALCUNO	0001	4, 17, 40, 69
ALL'	0001	3, 75, 47, 49

La lista di frequenze

Ultimo caso della nostra breve carrellata: la *lista di frequenze*. La lista di frequenze di un testo, generalmente, mostra le parole che lo compongono accompagnate dal numero di volte che ricorrono, e dalla percentuale rispetto al numero totale di parole. Il posto in "classifica" che ogni parola riporta (prima, seconda, terza e, ignominiosamente, ultima) viene detto *rango* (e chissà se sia più meritevole essere un luogo comune o un *hapax*). Comunque, può essere interessante stabilire una soglia al di sotto o al di sopra della quale le parole vengono considerate rare o frequenti, per poi attribuire alla rarità o alla frequenza un certo valore e/o significato.

La lista di frequenze si distingue dagli indici poiché (come è ovvio e inevitabile) mancano le indicazioni sulla collocazione della parola nel testo, ma anche la lista di frequenze deve trovare una sua organizzazione e quindi le frequenze vengono solitamente organizzate o in ordine alfabetico oppure in ordine crescente o decrescente di frequenza.

ABBIAN	1	2.6	0.12	ALLA	2	5.1	0.24
ABISOGNA	1	2.6	0.12	ALLE	1	2.6	0.12
ACCESO	1	2.6	0.12	ALLEGGERIMENTO	1	2.6	0.12
ACCIO'	1	2.6	0.12	ALLEGGIARE	1	2.6	0.12
ADUNQUE	1	2.6	0.12	ALLEGRI	1	2.6	0.12
AFFANNO	1	2.6	0.12	ALLI	1	2.6	0.12
AFFLIGGE	1	2.6	0.12	ALMENO	2	5.1	0.24
AFFLITTI	1	2.6	0.12	ALTISSIMO	1	2.6	0.12
AGLI	1	2.6	0.12	ALTRE	2	5.1	0.24
AGO	1	2.6	0.12	ALTRI	1	2.6	0.12
AL	2	5.1	0.24	ALTRO	2	5.1	0.24
ALCUNA	2	5.1	0.24	AMANO	1	5.1	0.12
ALCUNE	1	2.6	0.12	AMATA	1	2.6	0.12
ALCUNI	1	2.6	0.12	AMENDI	1	2.6	0.12
ALCUNO	4	10.3	0.47	AMICO	1	2.6	0.12
ALL	1	2.6	0.12	AMORE	4	10.3	0.47

Liste di frequenze in ordine alfabetico.

CHE	39	100.0	4.61	COME	6	15.4	0.71
DI	23	59.0	2.72	LA	6	15.4	0.71
PER	22	56.4	2.60	QUALE	6	15.4	0.71
DA	17	43.6	2.01	COLORO	5	12.8	0.59
NON	17	43.6	2.01	DÈ	5	12.8	0.59
IL	16	41.0	1.89	DONNE	5	12.8	0.59

IN	14	35.9	1.65	GIA'	5	12.8	0.59
SI'	13	33.3	1.54	MI	5	12.8	0.59
PIU'	11	28.2	1.30	NOIA	5	12.8	0.59
QUALI	11	28.2	1.30	TEMPO	5	10.3	0.59
SE	10	25.6	1.18	ALCUNO	4	10.3	0.47
NE	9	20.5	1.06	AMORE	4	10.3	0.47
CIO'	8	23.1	0.95	COSE	4	10.3	0.47
LE	8	20.5	0.95	DELLE	4	10.3	0.47
IO	7	17.9	0.83	EGLI	4	10.3	0.47
SIA	7	17.9	0.83	FORZA	4	10.3	0.47

Lista di frequenze in ordine di frequenza.

CHE	37	100.0	4.61	TEMPO	5	1.28	0.59
NON	17	4.36	2.01	ALCUNO	4	1.03	0.47
SI	13	3.33	1.54	AMORE	4	1.03	0.47
PIU'	11	2.82	1.30	COSE	4	1.03	0.47
QUALI	11	2.82	1.30	EGLI	4	1.03	0.47
SE	10	2.56	1.18	FORZA	4	1.03	0.47
NE	9	2.31	1.06	LOR	4	1.03	0.47
CIO'	8	2.05	0.95	LORO	4	1.03	0.47
IO	7	1.79	0.83	ME	4	1.03	0.47
SIA	7	1.79	0.83	QUANTUNQUE	4	1.03	0.47
COME	6	1.54	0.71	QUELLE	4	1.03	0.47
QUALE	6	1.54	0.71	ASSAI	3	7.7	0.35
COLORO	5	1.28	0.59	BISOGNO	3	7.7	0.35
DONNE	5	1.28	0.59	CREDO	3	7.7	0.35
GIA'	5	1.28	0.59	DOVE	3	7.7	0.35
MI	5	1.28	0.59	FU	3	7.7	0.35
NOIA	5	1.28	0.59	HA	3	7.7	0.35

Lista di frequenze in ordine di frequenza (articoli e preposizioni espunte).

La lemmatizzazione

Ottenute le nostre liste, se abbiamo deciso di imboccare la via delle concordanze, ci troviamo di fronte a una scelta veramente delicata. Quando avremo effettuato il primo spoglio linguistico, infatti, ci troveremo di fronte a un risultato che a seconda dei punti di vista potrà essere considerato definitivo, oppure un semplice stadio di passaggio dell'elaborazione.

Potremmo cioè decidere di accontentarci del risultato ottenuto fino a questo punto (e preferire un output che si presenti in modo *non lemmatizzato*) oppure potremmo ritenere insufficiente il livello di elaborazione raggiunto (e decidere per il modo lemmatizzato).

Nel secondo caso sarà necessario effettuare una serie complessa di operazioni, mentre nel primo ci si potrà quasi accontentare del risultato restituito dal calcolatore.

In entrambe le situazioni, infatti, sembra decisivo mantenere uno stretto collegamento con il testo originale, e forse l'informatica, svincolandoci dai tradizionali supporti e dall'idea stessa di "libro", è in grado di indicarci nuove strade e di fornirci strumenti più potenti.

Il problema è che nelle cosiddette "lingue naturali", e figuriamoci nell'italiano, non troviamo solamente le forme così come sono attestate in un vocabolario, ma le troviamo praticamente sempre "flesse". Troviamo, cioè, i vari tempi dei verbi (tempi semplici e composti), troviamo maschile e femminile, singolare e plurale, pronomi legati a verbi e così via. Troviamo anche omografi, parole uguali con diverso significato, che devono essere distinti. Ci imbattiamo in tutti quei casi che hanno reso la vita dura alla linguistica computazionale, alla comprensione del linguaggio naturale, per non parlare, poi, della traduzione automatica.

L'operazione di lemmatizzazione si propone di riportare le diverse *forme* sotto un unico *lemma*, di distinguere invece gli omografi, di ricostruire, insomma, una sorta di dizionario del testo. Fatto questo sarà opportuno inserire anche alcune specificazioni di carattere grammaticale, che in alcuni casi potranno essere individuate molto semplicemente ma che in altri potrebbe non essere facile e immediato definire.

La lemmatizzazione di un testo, come appare evidente, tutto è, meno che un'operazione indolore. Il lemmatizzatore potrà essere aiutato da alcuni strumenti informatici (che vanno da un dizionario macchina completamente automatico, fino a un *lemmatizzatore elettronico interattivo* che, utilizzando la competenza dell'operatore, superi la mancata formalizzazione di conoscenza), ma sarà sempre l'uomo a dover intervenire in prima persona per dirimere le questioni più delicate.

Al contrario del modo lemmatizzato, il modo non lemmatizzato non prevede alcun procedimento per normalizzare le varie occorrenze. Le diverse forme andranno al loro posto senza essere radunate sotto un unico lemma e anche in questo caso sarà possibile inserire alcune specificazioni grammaticali.

Il vantaggio del modo lemmatizzato sta nella facilità di consultazione: vi è una logica precisa che regola la posizione delle entrate nelle concordanze. Lo svantaggio naturalmente consiste nella distanza quasi asettica che si crea tra il lemma e la parola che l'autore aveva voluto inserire nel proprio testo con il rischio di una perdita o di una mimetizzazione dell'informazione.

Il vantaggio del modo non lemmatizzato risiede nell'aderenza perfetta dello spoglio alle parole del testo. Verrà messa in eviden-

za proprio quella forma che l'autore aveva voluto mettere nel testo e non un lemma di dizionario in qualche modo astratto. Lo svantaggio consiste essenzialmente in una difficoltà di consultazione, che tende ad aumentare man mano che ci si occupa di testi più antichi. Potrebbe, al limite, diventare impossibile ritrovare una certa forma e collegarla a forme dal significato o dalla funzione simile, ma di diversa grafia.

In conclusione: come è facile costatare, le tabelle più sopra riportate potrebbero essere realizzate anche da un paziente contabile, mentre la lemmatizzazione è da sempre un esercizio che si è portato a termine manualmente. L'informatica, in questo caso come in tutti i casi in cui vengono messe in gioco le sue abilità più note e ovvie, non produce risultati diversi da quelli ottenibili con metodi tradizionali (forse non abita qui lo specifico dell'applicazione dell'informatica alle discipline umanistiche), ma permette di ottenere con maggior rapidità e su una massa di dati più ampia dei risultati più attendibili.

ALCUNI PROGRAMMI PER L'ANALISI TESTUALE

Per effettuare questo genere di elaborazione di materiali testuali e linguistici sono stati sviluppati numerosi applicativi. Qui di seguito esaminiamo quelli attualmente più diffusi.

TACT <http://www.chass.utoronto.ca/cch/tact.html>

TACT (acronimo di *Textual Analysis Computing Tools*) è un pacchetto di programmi per l'analisi dei testi letterari di pubblico dominio messo a punto dall'università di Toronto nel 1984 e da allora periodicamente aggiornato. TACT è particolarmente efficace per l'analisi di testi, sia perché capace di elaborare ricerche estremamente complesse, mediante operatori e raggruppamenti dei termini di ricerca, sia perché permette di codificare il testo di input mediante il linguaggio di mark-up COCOA. Tuttavia l'utilizzo di questo prodotto è piuttosto ostico, non solo per la complessità delle operazioni da svolgere, ma anche perché i singoli programmi girano solo in ambiente MS-DOS e devono essere utilizzati singolarmente, in assenza di un ambiente integrato. Esiste comunque un modulo (TACTWeb) che consente di interrogare gli indici *full-text* in ambiente Web.

Concordance <http://www.rjcw.freeserve.co.uk>

Concordance 3.0 è un programma funzionale e *user-friendly*, basato su ambiente Windows. In pochi secondi genera da uno o più testi codificati in ASCII un file con l'estensione *.concordance, che contiene le concordanze del singolo testo o di un *corpus*, ed è convertibile automaticamente in una pagina Web. All'interno delle concordanze così ottenute si possono effettuare ricerche di singole parole (senza però la possibilità di utilizzare operatori) e ottenere elementari statistiche. Le occorrenze indicano come riferimento per la citazione il numero di linea del file di input utilizzato, anche nel caso di corpora, e non l'indicazione della suddivisione testuale alla quale quel contesto appartiene. È possibile utilizzare Concordance con i file XML, grazie al comando *ignore* che esclude i caratteri compresi tra due marcatori, ma non è possibile operare ricerche che tengano conto dei *tag* come in Wordsmith e Monoconc.

Wordsmith <http://www.lexically.net/wordsmith/index.html>

Insieme di *tool* per l'analisi linguistica particolarmente versatile e complesso, presenta alcune interessanti funzioni, come la possibilità di dividere il testo analizzato in sezioni; statistiche complesse; una modalità di ricerca che utilizza operatori avanzati.

Particolarmente interessante è la possibilità di lavorare con testi codificati in XML. I *tag* possono essere ignorati, ma anche utilizzati per ricavare statistiche dettagliate (per esempio le concordanze del primo capitolo di un romanzo, oppure delle battute di un singolo personaggio, o ancora l'elenco dei nomi propri di un saggio).

Estremamente interessante per chi studia lessicologia, ma anche ai fini dell'analisi testuale, è la possibilità di effettuare una lemmatizzazione automatica. Quest'ultima operazione richiede, in ogni caso, la stesura di un elenco delle desinenze della lingua italiana, in quanto il programma è impostato per la lingua inglese; particolarmente difficile, ovviamente, è il caso dei lemmi verbali.

Completano il prodotto un *viewer*, utilizzabile anche come *browser* SGML, e una documentazione dettagliata.

Monoconc Pro <http://www.athel.com/mono.html>

Prodotto professionale, in grado di gestire *corpora* complessi, composti da milioni di parole. Simile al suo concorrente Wordsmith, è particolarmente versatile nella visualizzazione delle occorrenze: utilizza il formato *Kwic* oltre a quello per frase, elimina le occorrenze che non interessano, visualizza un contesto molto ampio, segnala tag particolari all'interno del contesto.

(scheda a cura di Lorenzo Gerl)

BIBLIOGRAFIA CONSIGLIATA

Howard-Hill T.H., *Literary Concordance. A Guide to the Preparation of Manual and Computer Concordance*, Pergamon Press, Oxford-New York-Toronto 1979.

Lamb S.M., Gould L., *Concordances from Computers*, University of California, Berkeley 1964.

Lana M., *L'uso del computer nell'analisi dei testi*, Franco Angeli, Milano 1994.

Lancashire I., *Using Tact with Electronic Texts*, MLA, New York 1996.

McCarty W., *Finding Implicit Patterns in Ovid's "Metamorphoses" with TACT*,