

# LA LOGICA DI UN TESTOSCOPIO

**Franco Lancia**

(© 28 ottobre 2002)

web: [www.tlab.it](http://www.tlab.it); mail: [franco.lancia@tlab.it](mailto:franco.lancia@tlab.it)

## Avvertenza:

Questo scritto è stato pubblicato quando, nel 2002, fu rilasciata la versione 2.0 di T-LAB. Il suo scopo era quello di presentare la “logica” e l’architettura del software. Poiché questo ha subito molti cambiamenti (settembre 2009: versione 7.0), si consiglia di integrare questa lettura con la breve introduzione disponibile al seguente indirizzo: <http://www.tlab.it/it/download.php>

## INTRODUZIONE

Nella storia delle scienze le relazioni tra teorie e strumenti sono in qualche modo bi-direzionali: lo sviluppo di nuove teorie porta a costruire nuovi strumenti, reciprocamente l’uso di nuovi strumenti porta a sviluppare nuove teorie. Ad esempio, è stato possibile costruire antenne e satelliti per le telecomunicazioni solo dopo aver sviluppato ricerche e teorie nell’ambito della fisica e dell’ingegneria (teorie -> strumenti); reciprocamente, molte scoperte e teorie della biologia sono impensabili senza l’uso del microscopio, così come la moderna astronomia è impensabile senza l’uso del telescopio (teorie <- strumenti).

In qualche modo, tutti gli strumenti tecnologici – soprattutto quelli orientati a potenziare le nostre capacità di osservazione – sono la oggettivazione di qualche teoria.

Anche T-LAB è uno strumento di osservazione: appartiene alla famiglia dei software orientati a produrre mappe che rappresentano i contenuti dei testi, sia presi singolarmente che confrontati tra loro (somiglianze e differenze). Le teorie che regolano il suo funzionamento, quelle che si traducono in regole di trasformazione e che, in esso, organizzano le relazioni tra dati e loro rappresentazioni, appartengono a due gruppi di discipline: la linguistica e la statistica.

Obiettivo di questo documento è appunto quello di chiarire il ruolo che queste discipline hanno nella logica degli strumenti T-LAB.

Il documento è suddiviso in cinque sezioni:

- 1- **Unità di analisi e teoria del significato**, in cui vengono proposti i modelli della linguistica che consentono di definire gli “oggetti” di osservazione;
- 2- **Itinerari di analisi**, in cui vengono individuate due diverse strategie che presuppongono la scelta di “cosa” si vuole osservare e analizzare;
- 3- **La logica delle co-occorrenze**, in cui vengono descritti e commentati gli strumenti che consentono di conseguire i risultati tipici del primo tipo di strategia: associazioni, mappe dei nuclei tematici, tipologie dei contesti elementari;
- 4- **La logica delle comparazioni**, in cui vengono descritti e commentati gli strumenti che consentono di conseguire i risultati tipici del secondo tipo di strategia: specificità, analisi delle corrispondenze e cluster analysis;

5- **Dai testi alla cultura**, in cui vengono proposte alcune riflessioni sulla funzione culturale dell'analisi dei testi.

## 1. UNITA' DI ANALISI E TEORIA DEL SIGNIFICATO

In T-LAB, il riferimento alla linguistica è utilizzato soprattutto per definire e organizzare i “dati” che sono oggetto di osservazione; in altri termini, per definire le cosiddette “unità di analisi” (AU = Analysis Units). Queste sono di due tipi:

- a) le **unità di contesto** (CU = Context Units), ovvero i sottoinsiemi derivati dalla scomposizione del corpus inteso come insieme di uno o più testi. Ad esempio, se il corpus in analisi è costituito da un insieme di articoli di quotidiani, le unità di contesto possono essere: i singoli articoli, i sottoinsiemi di articoli classificati in base a un criterio (testata, giorno di pubblicazione, tema dominante, etc), le singole frasi in cui ogni articolo può essere segmentato (i contesti elementari).
- b) le **unità lessicali** (LU = Lexical Units), ovvero le singole parole, sia utilizzate come “forme grafiche”, sia ricondotte al loro lemma (ad esempio “lavorano” -> “lavorare”), sia ricondotte a classi semantiche (ad esempio “bronchite” -> “malattia”) o a categorie di un qualche dizionario (vedi i coding schemes in uso nella Content Analysis), in ogni caso tutte associate ad indici che le riconducono ai loro contesti di appartenenza (CU).

Questa distinzione, che ha un fondamento teorico nell'ambito della linguistica e della semiologia, ha notevole rilevanza pragmatica: essa, infatti, consente le traduzioni operazionali che sono alla base della statistica testuale.

Quanto al fondamento teorico, esso va ricondotto all'ipotesi inizialmente proposta da F. de Saussure (1916), e successivamente ripresa da vari studiosi (R. Jakobson, 1963; R. Barthes, 1964), secondo la quale le relazioni tra gli elementi linguistici possono essere analizzate come **rapporti sintagmatici** e/o come **rapporti paradigmatici**. I primi regolano la combinazione degli elementi linguistici entro i contesti (l'uno “accanto” all'altro: CU), le seconde presiedono alla possibilità di sostituire un elemento linguistico con altri che con esso hanno in qualcosa in comune (l'uno “al posto” dell'altro: LU).

La rilevanza pragmatica della distinzione proposta è data dal fatto che le relazioni tra i due tipi di unità di analisi (CU e LU) possono essere rappresentate come matrici (o tabelle) i cui valori numerici indicano i fenomeni di **occorrenza e co-occorrenza**.

Ad esempio, dato un insieme di unità lessicali ( $lu_1, lu_2, lu_3, \dots, lu_m$ ) e un insieme di unità di contesto ( $cu_1, cu_2, cu_3, \dots, cu_n$ ), rispettivamente con “m” e “n” elementi, è possibile costruire una matrice rettangolare “m X n”, con “m” righe e “n” colonne, ai cui singoli incroci ( $lu_i cu_j$ ) corrispondono valori di occorrenza che indicano quante volte ciascuna unità lessicale (ad esempio una parola) occorre in ciascuna unità di contesto (ad esempio un articolo di quotidiano). Ugualmente, se le unità di contesto sono costituite da frasi, possiamo costruire una matrice quadrata m X m, ai cui singoli incroci ( $lu_i lu_j$ ) corrispondono valori di co-occorrenza, ovvero il numero di frasi in cui ciascuna unità lessicale (ad esempio una parola) è presente insieme a ciascuna delle altre.

In entrambe i casi, cioè sia nelle matrici rettangolari che in quelle quadrate, ciascuna riga e ciascuna colonna riassume il “profilo” di una unità di analisi (CU o LU). Ad esempio, nelle matrici rettangolari “m X n” ogni riga può essere usata per rappresentare il profilo di una parola inteso come “distribuzione”<sup>1</sup> delle sue occorrenze entro i contesti considerati.

Ebbene le nozioni di **occorrenza** e di **co-occorrenza**, di **profilo** e di **distribuzione**, sono le chiavi per capire le teorie del significato che sono implementate negli strumenti T-LAB.

In effetti, nella logica del software ogni unità di analisi (CU o LU) è “conosciuta” solo in base al suo profilo. Più esattamente:

- ogni CU è conosciuta in base alla distribuzione delle occorrenze di ogni LU in essa presente;
- ogni LU è conosciuta in base alla distribuzione delle sue occorrenze all’interno di ogni CU, come anche in base alla distribuzione delle sue co-occorrenze con ciascuna delle altre LU.

Evidentemente, **il software non conosce propriamente i significati** (o contenuti), bensì solo i significanti<sup>2</sup>, cioè le “stringhe” che individuano le varie LU e CU; tuttavia, le relazioni tra significanti (ovvero i rapporti sintagmatici) assumono delle forme significative che, in qualche modo, propongono una **rappresentazione contestuale dei significati**.

Prendiamo, ad esempio, l’unità lessicale per eccellenza: la singola “parola”. Se prescindiamo dalle categorie morfologiche e dalle funzioni sintattiche<sup>3</sup>, ogni singola parola che utilizziamo si differenzia da tutte le altre in base alle “unità minimali” o “tratti distintivi” in cui può essere scomposta, sia sul versante del significante (o espressione) che su quello del significato (o contenuto): nel primo caso (significante/espressione) il risultato della scomposizione è costituito dai fonemi, nel secondo caso (significato/contenuto) dai “semi” o tratti semantici. Questi ultimi, nel modello proposto da A. J. Greimas (1966) risultano dalla combinazione di due tipi di elementi:

- a) il **nucleo semico**, supposto invariante e considerato come una sorta di minimo comun denominatore di tutte le accezioni della parola considerata<sup>4</sup>;
- b) i **semi contestuali**, costituiti dagli “effetti di senso” determinati da insiemi (o classi) di contesti.

Ebbene, nella logica di T-LAB, il “significato” di ogni singola parola è conosciuto solo attraverso le sue relazioni con i contesti, cioè attraverso la distribuzione delle sue occorrenze (o co-occorrenze) all’interno delle Unità di Contesto (CU). Il riferimento ai “semi contestuali” di A. J. Greimas è dunque più che pertinente.

Tuttavia, gli algoritmi statistici implementati nel software non possono essere considerati come mere traduzioni delle teorie proposte nell’ambito della semantica strutturale.

---

<sup>1</sup> A questo proposito è pertinente il riferimento alla linguistica distribuzionale (L. Bloomfield, 1933).

<sup>2</sup> Secondo De Saussure (1916) ogni segno (ad es. una parola) ha due facce: il significante, costituito dal suo aspetto fisico, e il significato, costituito dalla rappresentazione mentale a cui è associato.

<sup>3</sup> Le categorie morfologiche vengono utilizzate per classificare le parole in sostantivi, aggettivi, verbi etc. Le funzioni sintattiche sono quelle del tipo “soggetto” e “predicato”.

<sup>4</sup> Nella terminologia di Greimas le parole, considerate sotto l’aspetto del significato, sono “sememi”.

## 2. ITINERARI DI ANALISI

Nell'architettura di T-LAB, i vari strumenti (le "funzioni")<sup>5</sup> e gli algoritmi in essi implementati acquistano senso solo all'interno di una qualche **strategia di analisi**. In effetti, per ogni architettura vale il principio che può essere esplorata per verificare "come è fatta" o "come funziona", ma che – tuttavia – acquista il suo vero significato solo quando è utilizzata per qualche scopo. Detto con una metafora: una cosa è "visitare" una casa, altra cosa è "abitarci".

Immaginiamo quindi un ipotetico ricercatore alle prese con l'analisi di un qualche corpus. Se ha deciso di utilizzare T-LAB, le sue strategie sono determinate da due tipi di scelte che, in ogni caso, sono reversibili:

- 1- allo scopo di osservare l'organizzazione semantica dei contenuti entro il corpus o entro suoi sottoinsiemi, (a) voglio **dare la priorità all'analisi delle co-occorrenze tra LU**; oppure (b) voglio **dare priorità all'analisi delle somiglianze/differenze tra sottoinsiemi del corpus (CU)**, così come risultano dai profili delle occorrenze e dall'uso di qualche variabile?
- 2- in entrambi i casi di cui sopra (a, b), le LU che voglio utilizzare sono le parole e/o i lemmi, così come T-LAB li classifica (sia automaticamente che attraverso interventi dell'utilizzatore), oppure preferisco utilizzare un dizionario o un "coding scheme" da importare?

Le modalità per rendere operative queste scelte sono ben illustrate nell'help del software. Le note che seguono sono destinate a chiarire le implicazioni delle due strategie del punto (1), le quali – rispettivamente – fanno leva sulla **logica delle co-occorrenze** (vedi sotto 3.1) e sulla **logica delle comparazioni** (vedi sotto 3.2). Quanto al punto (2), mi limito a dare solo qualche chiarimento.

Nei databases T-LAB, tutte le singole parole del corpus vengono indicizzate in base alla loro posizione nelle varie CU. Successivamente, in modo automatico, a ciascuna di esse è attribuita una "label" che generalmente corrisponde a un lemma dei dizionari standard. Infine è previsto che l'utilizzatore, vuoi tramite interventi mirati, vuoi tramite dizionari importati, vuoi utilizzando risultati di precedenti analisi, possa cambiare le denominazioni delle "labels".

Lo schema adottato è il seguente:

---

<sup>5</sup> Alcuni strumenti T-LAB hanno la funzione di consentire analisi statistiche, altri (ad esempio: Disambiguazione, Liste di Poliformi, Concordanze, etc) sono solo a supporto del lavoro di analisi.

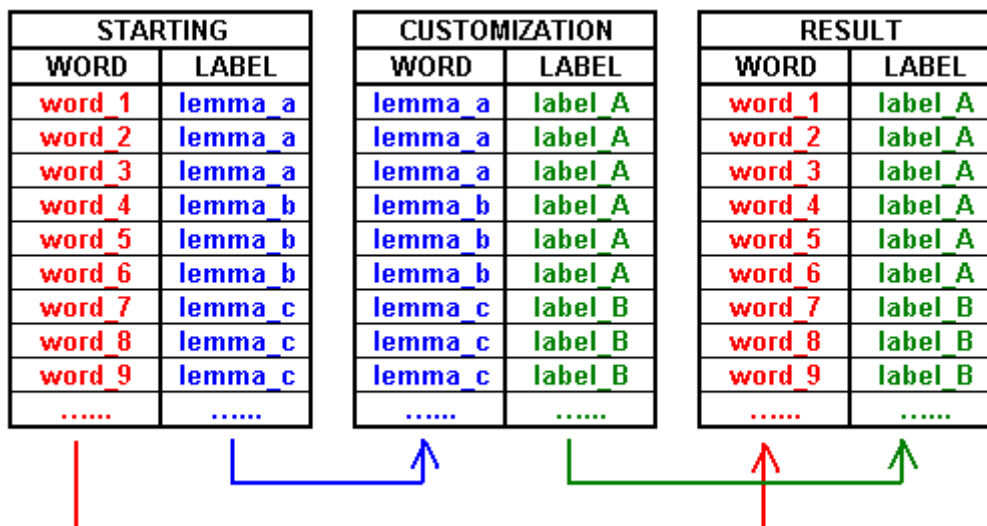


Tabella 2.1

In altri termini, per cambiare “label” alle parole, l’utente deve mettere in corrispondenza bi-univoca le entrate del suo dizionario con i lemmi proposti da T-LAB. In questo modo può prendere “n” decisioni reversibili del tipo “Questa parola, invece di chiamarla X, la voglio chiamare Y”<sup>6</sup>. Ovviamente, **tutte le “sostituzioni” hanno senso solo se rispettano un qualche criterio paradigmatico**, ovvero se le “associazioni” del tipo parola-label rispettano un qualche criterio semantico (vedi la definizione di unità lessicale al punto 1).

### 3. LA LOGICA DELLE CO-OCCORRENZE

#### 3.1 ASSOCIAZIONI

Per iniziare ad esplorare la logica delle co-occorrenze, partiamo dalla funzione “Associazioni”. Nella libreria dei programmi T-LAB gli algoritmi che in essa sono implementati sono tra i più semplici: per ogni LU selezionata (parola, lemma o categoria), essendo note le sue occorrenze e le sue co-occorrenze all’interno delle varie CU, è possibile calcolare un indice di associazione con ciascuna delle altre LU (parole, lemmi o categorie) del corpus o di un sottoinsieme di esso.

In questo caso, l’indice utilizzato è il **coefficiente del coseno** (G. Salton, 1989).

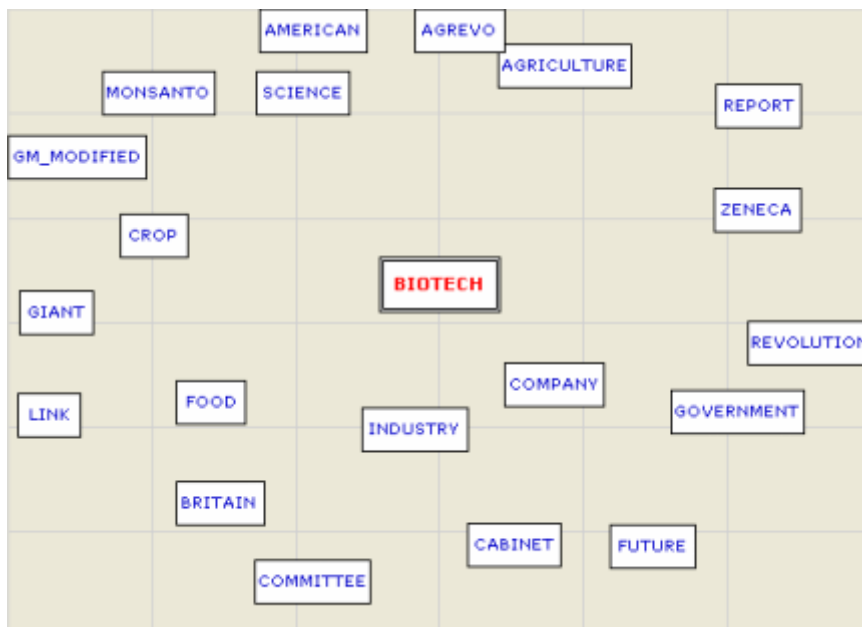
La sua formula è la seguente:

$$C(X,Y) = \frac{X \cap Y}{\sqrt{X} \times \sqrt{Y}}$$

<sup>6</sup> La funzione T-LAB che consente questo tipo di operazioni è “Categorizzare”.

dove il coefficiente del coseno (C) tra ogni coppia di LU (X,Y) è definito come rapporto tra la quantità delle loro co-occorrenze e quella ottenuta moltiplicando le radici quadrate delle rispettive occorrenze entro le CU considerate <sup>7</sup>.

A partire dai calcoli effettuati, T-LAB produce grafici come quello seguente:



**Figura 3.1**

La parola (LU) selezionata, quella di cui si vuol verificare il **significato contestuale**, è posta al centro; tutte le altre sono distribuite intorno ad essa, ciascuna a una distanza proporzionale al suo grado di associazione (le parole più prossime al centro sono quelle con il più elevato coefficiente del coseno).

Le relazioni significative sono quindi del tipo uno-ad-uno, tra la parola centrale e ciascuno delle altre.

Evidentemente, il significato non è “dato” dall’analisi realizzata con il software. In analogia con altri strumenti di osservazione, potremmo dire che T-LAB funziona come un testoscopio: mostra cose che possono essere “viste” solo da chi è in grado di interpretarle.

### **3.2 MAPPE DEI NUCLEI TEMATICI**

Gli algoritmi statistici implementati in questa funzione vengono applicati sempre e soltanto a matrici di co-occorrenze. Più specificamente, la matrice iniziale è costituita da tante righe e tante colonne (matrice quadrata) quante sono le LU selezionate (parole, lemmi, categorie).

Gli steps previsti dall’analisi sono i seguenti:

---

<sup>7</sup> In questo caso, la quantità delle co-occorrenze è data dal numero di contesti elementari (frasi o segmenti ad essi equivalenti) in cui X e Y sono contemporaneamente presenti.



In particolare, le mappe prodotte (vedi sopra) sono organizzate da assi fattoriali; quindi tutte le relazioni in essa presenti possono essere lette e interpretate attraverso tabelle che riportano varie misure. Tra queste, quelle di più immediato utilizzo sono costituite dai Valori Test <sup>10</sup>; misura, quest'ultima, che ha due proprietà rilevanti: un valore di soglia per il rifiuto dell'ipotesi nulla ( $p = 0.05$ ) e un segno (-/+). Ciò significa che ordinando i valori in modo crescente o decrescente, a seconda che si considerino i valori sul polo "negativo" (-) o su quello positivo (+), è possibile apprezzare il peso di ciascun oggetto su ciascun asse fattoriale e – quindi – definire le **opposizioni semantiche** che organizzano i vari spazi bi-dimensionali.

In altri termini, le rappresentazioni proposte da questa funzione consentono di osservare e di interpretare l'organizzazione semantica del corpus in analisi o dei suoi sottoinsiemi.

### 3.3 **TIPOLOGIE DEI CONTESTI ELEMENTARI**

In T-LAB, la logica delle “Associazioni” è in qualche modo inclusa in “Mappe dei Nuclei Tematici”; analogamente la logica di quest'ultima è inclusa in “Tipologie dei Contesti Elementari” (E.C.A. = Elementary Context Analysis). In effetti, in tutti e tre i casi, le matrici dei dati sono costituite dalle co-occorrenze e, in tutti e tre i casi, l'obiettivo è quello di fornire rappresentazioni che consentano di esplorare l'organizzazione contestuale dei significati senza la necessità di prefigurare l'uso delle variabili eventualmente scelte dal ricercatore.

Più precisamente, nella logica E.C.A. le variabili sono presenti solo come “labels” che consentono di reperire le appartenenze delle unità di analisi (CU e LU) a sottoinsiemi del corpus. In altri termini, vengono usate “a posteriori”, e solo per caratterizzare i clusters individuati.

Gli obiettivi e gli outputs di questa funzione (E.C.A.) sono molto simili a quelli di ALCESTE <sup>11</sup>; tuttavia le tabelle dati e gli algoritmi utilizzati sono diversi.

Il punto di partenza è costituito da una matrice delle co-occorrenze, esattamente quella ottenuta tramite Mappe dei Nuclei Tematici (vedi 3.2 step “d”). Successivamente viene costruita una matrice rettangolare ( $m \times n$ ) con tante righe quanti sono i “contesti elementari” (E.C. = Elementary Contexts, cioè frasi o segmenti di testo ad esse assimilabili) e tante colonne quanti sono i “nuclei tematici”; i contesti elementari vengono assunti come “oggetti” da classificare (clustering) e i nuclei tematici vengono assunti come loro “caratteristiche”. Più specificamente, il profilo di ogni riga è costituito da valori “1” e “0” che marcano la presenza/assenza di ogni nucleo al suo interno.

Gli steps successivi sono i seguenti:

- a- analisi delle corrispondenze applicata alla matrice ( $m \times n$ );
- b- una prima cluster analysis (del tipo K-means) che usa le coordinate di ogni “oggetto” sui primi tre fattori ottenuti tramite analisi delle corrispondenze;
- c- una seconda cluster analysis (metodo di Ward) che definisce la classificazione delle E.C.;

---

<sup>10</sup> Per una descrizione dettagliata di questa misura si può consultare il volume di Lebart L. Morineau A. Piron M. (1995: 123-125).

<sup>11</sup> ALCESTE (Analyse des Lexèmes Co-occurents dans les Énoncés Simples d'un TExt) è stato sviluppato da Max Reinert.

- d- rappresentazioni grafiche dei risultati mediante analisi delle corrispondenze (righe: LU e CU; colonne: i clusters ottenuti);
- e- caratterizzazione di ogni cluster mediante test del chi-quadro.

Le mappe prodotte consentono di verificare le relazioni tra clusters e variabili entro spazi bi-dimensionali. Varie tabelle consentono di apprezzare le caratteristiche dei clusters e delle polarità fattoriali, quelle risultate significative tramite l'uso del chi-quadro (per i clusters) o del Valore Test (per i fattori). In entrambe i casi, le "caratteristiche" dei cluster sono costituite da LU o da CU.

Per molti versi, il risultato dell'analisi può essere considerato come una mappatura delle **isotopie** (iso = uguale; topoi = luoghi) <sup>12</sup>, le quali rinviano a una concezione del significato come "effetto del contesto", cioè come qualcosa che non appartiene alle parole prese singolarmente, bensì che risulta dai loro rapporti all'interno delle unità di contesto (CU). In effetti, ciascun cluster individua un contesto di riferimento "condiviso" da più parole, ma che non deriva dai loro specifici significati. Ciò nella logica che l'insieme è qualcosa di diverso dalla sommatoria dei suoi elementi. Di conseguenza, ciascun cluster consente di ricostruire "un filo" del discorso all'interno della trama complessiva costituita dal corpus in analisi o da un suo sottoinsieme.

Tuttavia va ricordato che il riconoscimento di un'isotopia non è la mera constatazione di un "dato", bensì il risultato di un processo interpretativo (F. Rastier 1987).

### 3.4 ANALISI DELLE SEQUENZE

La funzione "Analisi delle Sequenze" costruisce e analizza due matrici delle co-occorrenze (m X m) i cui valori sono rispettivamente costituiti dal conteggio di quante volte, all'interno del corpus considerato, ciascuna unità lessicale risulta **predecessore** o **successore** di ciascuna delle altre nella struttura lineare (sequenziale) del discorso. Subito dopo, gli stessi valori vengono trasformati in valori di probabilità, più esattamente in probabilità di transizione, in quanto le relazioni tra gli elementi considerati (predecessori e successori) costituiscono una catena markoviana.

Come è noto, dato un qualunque insieme finito di eventi (i cosiddetti "stati") che si alternano in successione (le cosiddette "transizioni"), le catene markoviane consentono di ricostruire la "rete" delle loro reciproche relazioni attraverso legami di tipo probabilistico. Per l'applicazione di questo modello all'analisi del discorso un riferimento d'obbligo è costituito dal software DiscAn, ideato dall'antropologo canadese P. Maranda (1990) e proposto come uno strumento che – a partire da un'analisi di contenuto – consente di costruire mappe semantiche di tipo dinamico.

Quanto a T-LAB, le sue ambizioni sono più circoscritte; tuttavia esso rende molto più agevoli le operazioni del tipo analisi di contenuto (funzioni di tagging e uso di dizionari) e – soprattutto – consente di esplorare le relazioni tra un numero considerevole di "nodi" costituiti dalle LU (max 250).

---

<sup>12</sup> La nozione di isotopia è stata inizialmente proposta da A.J. Greimas (1966) per definire la ricorrenza, all'interno delle unità sintagmatiche (frasi e/o testi), di più parole con tratti semantici in comune tra loro.

## 4. LA LOGICA DELLE COMPARAZIONI

### 4.1 SPECIFICITA'

Un corpus può essere suddiviso in “n” sottoinsiemi e attraverso vari criteri di partizione (le variabili). Ad esempio, un libro può essere suddiviso in capitoli oppure in brani che trattano temi diversi; un insieme di interviste può essere suddiviso in base al sesso o alla professione dell'intervistato, in base alla data o ai temi affrontati; ugualmente si può procedere per analizzare le risposte a questionari con domande aperte, e così via.

Il tipo di domande alle quali la funzione “Specificità” consente di rispondere hanno la forma seguente: data una variabile (il criterio di partizione) e un sottoinsieme del corpus definito da una sua modalità <sup>13</sup>, quali caratteristiche sono “tipiche” di questo sottoinsieme e lo rendono diverso dai rimanenti sottoinsiemi del corpus?

Ovviamente, le caratteristiche analizzabili da T-LAB sono le parole utilizzate, cioè le Unità Lessicali (LU) e i loro profili. Quanto ai confronti, sono del tipo uno-ad-uno tra le occorrenze di ogni LU nel sottoinsieme considerato e quelle presenti nel resto del corpus. L'algoritmo di calcolo è quello del chi-quadro.

Proviamo a spiegarci con un esempio. Supponiamo che il corpus sia costituito da quattro articoli di altrettante testate giornalistiche (A, B, C, D) e che ci interessi calcolare le “specificità” della testata “A”. In questo caso, per ogni parola in analisi (LU) la tabella di riferimento è analoga alla seguente, in cui si ipotizza che: a) la LU in esame (“x”) ha “6” occorrenze nell'articolo della testata “A” e “8” occorrenze all'interno dei rimanenti articoli (testate: B, C, D); b) il totale delle occorrenze in “A” è “456”, mentre il totale delle occorrenze negli altri articoli è “3708”.

	in "A"	not in "A"	
LU "x"	6	8	14 (N <sub>i</sub> )
other LU	450	3700	4150
	456	3708	4164 (N <sub>ij</sub> )
	(N <sub>i</sub> )		

Tabella 4.1

<sup>13</sup> Ad esempio, per la variabile “professione”, le modalità possono essere le seguenti: medico, avvocato, psicologo, etc.

Per ogni cella, le frequenze “attese” (E) si calcolano nel modo seguente:  $(N_i \times N_j) / N_{ij}$ ; mentre il chi-quadro si calcola nel modo seguente:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

dove “O” sta per frequenze osservate (6, 8, 450, 3700). Ne risulta che il valore del chi-quadro è 14,664. Stando che in questo caso (df = 1; p. 0.05) il valore critico del chi-quadro è 3,84, l’ipotesi nulla viene rifiutata e la LU in esame può essere considerata “specificata” dell’articolo “A”.

## 4.2 ANALISI DELLE CORRISPONDENZE

In T-LAB, Analisi delle Corrispondenze (CA) è - allo stesso tempo - il nome di una specifica funzione e una procedura di calcolo implementata in altre funzioni. La descrizione dei suoi algoritmi richiederebbe più di qualche pagina e sarebbe necessariamente corredata da molte formule; quindi preferiamo non soffermarci su questi aspetti e, al lettore interessato, suggeriamo i seguenti riferimenti bibliografici: J.P. Benzecri (1984) e M. J. Greenacre (1984) Lebart L. Morineau A. Piron M. (1995).

Per dirla nel modo più semplice, la CA è una tecnica per rappresentare graficamente le relazioni tra i profili riga e i profili colonna delle matrici con valori di frequenza. Per descrivere la logica del suo funzionamento, almeno in prima istanza, si può prescindere da nozioni matematiche e statistiche. Basti pensare che le operazioni implementate nei suoi algoritmi consentono di ottenere due tipi di risultati:

- a) **rintracciare regolarità** nelle tabelle dati, ciò attraverso un confronto incrociato di tutti i profili (righe e colonne) nelle reciproche relazioni di somiglianza-differenza, con un risultato che – attraverso una serie di permutazioni – la tabelle possono essere ri-ordinate e le informazioni in esse contenute diventano “leggibili”.
- b) **ridurre le dimensioni** entro le quali i dati possono essere rappresentati, ciò attraverso la creazione di nuove variabili (i fattori) i cui valori corrispondono alle coordinate spaziali dei profili (righe e colonne). In questo modo i dati, inizialmente dispersi in modo random in uno spazio multi-dimensionale, risultano “agglomerati” entro uno spazio a dimensioni ridotte, quelle definite dai pochi fattori che, in modo statisticamente significativo, spiegano la loro variabilità.

Proviamo a piegarci con un esempio. Si consideri la tabella seguente, che simula una matrice delle occorrenze con 10 righe e 4 colonne <sup>14</sup>:

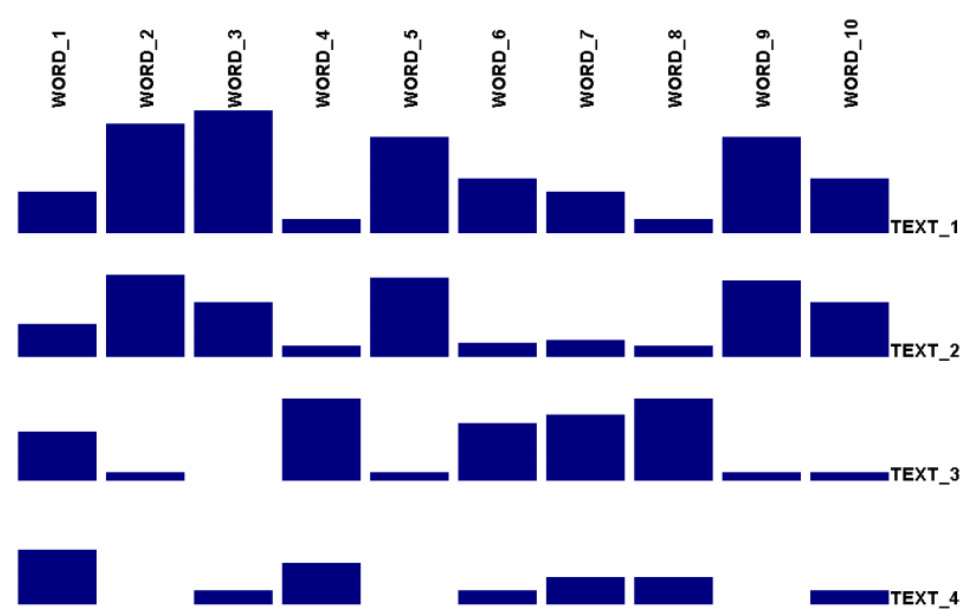
---

<sup>14</sup> Le matrici utilizzate da T-LAB possono avere migliaia di righe e qualche centinaio di colonne.

	TEXT_1	TEXT_2	TEXT_3	TEXT_4	Total
WORD_1	15	12	18	20	65
WORD_2	40	30	3	0	73
WORD_3	45	20	0	5	70
WORD_4	5	4	30	15	54
WORD_5	35	29	3	0	67
WORD_6	20	5	21	5	51
WORD_7	15	6	24	10	55
WORD_8	5	4	30	10	49
WORD_9	35	28	3	0	66
WORD_10	20	20	3	5	48
Total	235	158	135	70	598

**Tabella 4.2**

Una prima rappresentazione dei dati mostra i profili di righe e colonne in forma di istogrammi.



**Figura 4.1**

Utilizzando il software grafico AMADO, ideato da J. Bertin (1967), possiamo ri-ordinare righe e colonne tramite l'applicazione del modello CA. Più precisamente, in questo caso il ri-ordine è effettuato tramite l'uso del primo fattore estratto. Come si può rilevare, a partire da una tabella con dati "dispersi" si ottiene la rappresentazione di una struttura ordinata che, in qualche modo, è leggibile.

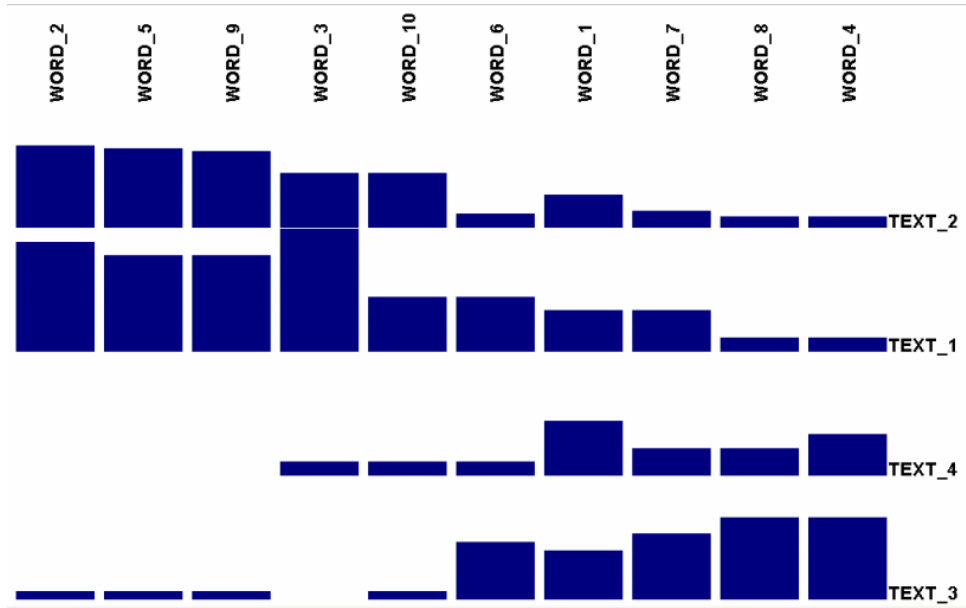


Figura 4.2

Gli stessi dati della tabella 4.2 possono essere rappresentati tramite un classico grafico bi-dimensionale (Figura 4.3) che, come si può rilevare, è coerente con la rappresentazione dei profili (Figura 4.2). In particolare, in entrambe i casi il primo fattore risulta caratterizzato dal “peso” prevalente di due CU (Text\_2, Text\_1) e due LU (word\_2, word\_5) sul polo negativo (-), mentre sul polo positivo (+) prevalgono altre due coppie di CU (Text\_4, Text\_3) e di LU (word\_4, word\_8).

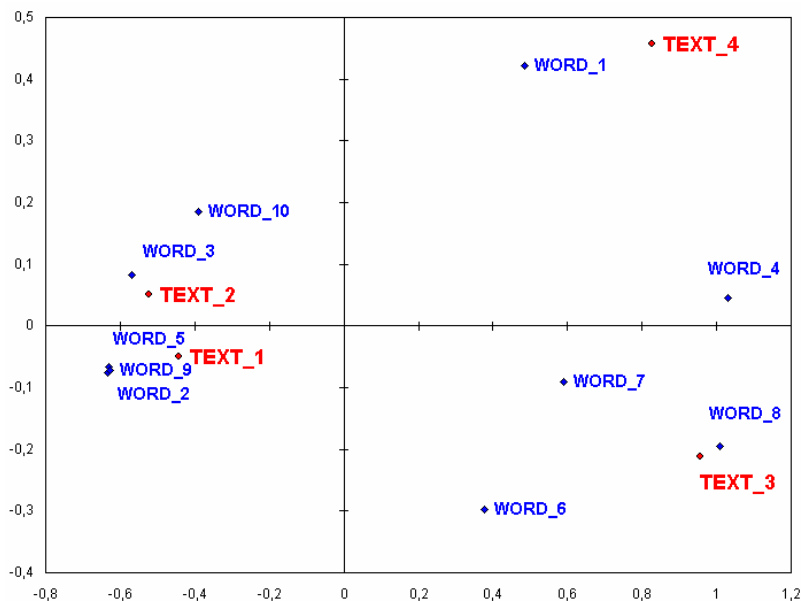


Figura 4.3

Per interpretare i risultati delle CA, in genere vengono utilizzate alcune misure, sia concernenti il “peso” di ogni fattore (Eigenvalue o Inertia), sia riferite alle coordinate e ai contributi (assoluti e

relativi) di ogni oggetto (riga o colonna) su ogni asse fattoriale. Tutte queste misure sono fornite anche da T-LAB; tuttavia, in via prioritaria, il software mostra alcune tabelle con i Valori Test, misura sulla quale ci siamo già soffermati (vedi 3.2) e la cui formula è la seguente <sup>15</sup>:

$$t\alpha(j) = \sqrt{n_j \frac{n-1}{n-n_j}} \varphi_{\alpha j}$$

Per esempio (vedi sopra Tabella 4.2), dato un elemento  $j$  (TEXT\_4) con un totale di occorrenze pari a  $n_j = 70$  e una coordinata sul primo asse fattoriale  $\varphi_{\alpha j} = 0.8276$ , poiché - nell'esempio - il totale delle occorrenze nel corpus è pari a  $n = 598$ , il Valore Test si calcola nel modo seguente:

$$t\alpha(j) = \sqrt{70 \frac{598-1}{598-70}} 0.8276 = 7.3627$$

Quando si interpretano i risultati di un'analisi fattoriale spesso si rischia di “perdersi” negli anfratti delle varie tabelle. Per evitare questo rischio, o comunque per ridurlo, può essere utile far riferimento a una qualche definizione. Per molti versi, i fattori possono essere considerati dei **principi di classificazione** (C.L. Burt, 1940), ovvero degli organizzatori delle relazioni tra dati che, secondo la logica della categorizzazione, mettono insieme cose simili, le distinguono dalle diverse, costruiscono “parentele” tra categorie di cose.

Scriva J.P. Benzecri, uno dei matematici che più ha contribuito a definire il modello dell'Analisi delle Corrispondenze:

"Interpretare un asse fattoriale significa trovare ciò che vi è di analogo, da una parte tra tutto ciò che è situato a destra dell'origine (o baricentro), dall'altra tra tutto ciò che è alla sinistra di questo, ed esprimere poi con concisione ed esattezza l'opposizione tra i due estremi" (1984: 302).

Con questa affermazione, mentre descrive un metodo di interpretazione, di fatto l'autore ci comunica una specifica concezione dei fattori intesi come organizzatori di relazioni oppostive tra insiemi o classi ("tutto ciò" che sta a destra e “tutto ciò” che sta a sinistra dell'origine). Quanto a dire che egli condivide una concezione dei fattori come **principi di classificazione**.

In effetti, nonostante la parola fattore evochi un significato di causa, le analisi di tipo fattoriale consentono “soltanto” di individuare un ordine nella complessità dei dati oggetto di trattamento, ovvero consentono di ridurre le dimensioni spaziali in cui questi possono essere rappresentati. Ma, evidentemente, una cosa è il senso statistico (o geometrico) dei fattori, altra cosa sono i modelli che, all'interno di ogni disciplina, fondano la possibilità di interpretarli. D'altro canto, se le scienze non

---

<sup>15</sup> Per una descrizione dettagliata della formula si può consultare il volume di [Lebart L. Morineau A. Piron M. \(1995: 123-125\)](#).

cercassero di spiegare i fattori che generano un qualche tipo di ordine nei fenomeni che studiano, esse stesse non avrebbero alcuna ragione di esistere.

In conclusione, può essere utile distinguere due tipi di fattori:

- i fattori ( $\alpha$ ) evidenziati dalle elaborazioni statistiche, quelli corrispondenti alle strutture di tipo matematico e geometrico (le dimensioni) che organizzano le relazioni tra dati, e che sono stati denominati come **principi di classificazione**.
- i fattori ( $\beta$ ) che, attraverso l'uso di processi inferenziali e facendo riferimento a modelli teorici, vengono evocati per fondare interpretazioni e/o spiegazioni delle forme assunte dai fattori ( $\alpha$ ). Questi fattori - di secondo tipo ( $\beta$ ) - possono essere denominati **principi di spiegazione** e sono specifici di ogni disciplina scientifica.

### 4.3 CLUSTER ANALYSIS

Gli algoritmi di Cluster Analysis implementati negli strumenti T-LAB sono di tre tipi diversi; uno di questi è specifico della funzione che può essere usata solo dopo una precedente Analisi delle Corrispondenze.

In questa sede ci soffermiamo a considerare solo la "logica" di quest'ultimo. Quanto alla letteratura sulla Cluster Analysis, ogni lettore può trovare facilmente più di un riferimento. Tuttavia qualche semplice definizione può risultare utile. Le seguenti sono tratte dall'Help on-line di T-LAB :

In generale, le tecniche statistiche della Cluster Analysis hanno l'obiettivo di individuare raggruppamenti di oggetti che abbiano due caratteristiche complementari:

- al loro interno, la massima somiglianza tra gli elementi che li costituiscono (gli oggetti appartenenti a ciascun cluster);
- tra di loro, la massima differenza.

In generale, i metodi della Cluster Analysis vengono distinti in due tipi:

- **hierarchical methods**, i cui algoritmi ricostruiscono l'intera gerarchia degli oggetti in analisi (il cosiddetto "albero"), vuoi in senso ascendente, vuoi in senso discendente;
- **partitioning methods**, i cui algoritmi prevedono che l'utilizzatore abbia preventivamente definito il numero di clusters in cui l'insieme degli oggetti in analisi va diviso.

Il metodo di analisi implementato nella funzione T-LAB che stiamo considerando è appunto di tipo gerarchico, più esattamente di tipo gerarchico-ascendente: cioè, parte dai singoli oggetti e - via via - utilizzando delle misure di prossimità, li aggrega fino a ricomporre l'intero insieme.

In questo caso gli “oggetti” sono costituiti dalle LU (parole, lemmi o categorie), ciascuno dei quali è caratterizzato da un profilo ottenuto tramite Analisi delle Corrispondenze: quello delle sue coordinate sui tre primi assi fattoriali <sup>16</sup>. Il criterio di aggregazione degli oggetti, in base alle loro distanze, è costituito dal metodo di Ward, uno dei più adatti nei metodi di clustering gerarchico che utilizzano coordinate fattoriali (Bolasco S., 1999).

Quanto al criterio di partizione, quello che determina il numero di clusters ottenuti, in T-LAB è implementato un algoritmo che utilizza il rapporto tra varianza inter-cluster e varianza totale <sup>17</sup> e che – in modo automatico – assume come “partizione ottimale” quella in cui questo rapporto supera la soglia del 50%.

Gli output sono di due tipi:

- alcuni grafici consentono di verificare la posizione dei clusters entro spazi bi-dimensionali che corrispondono a quelli ottenuti tramite la precedente Analisi delle Corrispondenze;
- alcune tabelle, per ciascun cluster, riportano le LU (parole, lemmi o categorie) che lo costituiscono (i suoi “elementi”) e le CU (modalità della variabile utilizzata) che li caratterizzano <sup>18</sup>.

I cluster ottenuti, analogamente a quelli della funzione E.C.A, propongono delle **isotopie** (vedi sopra 3.3). Ovviamente, nei due casi, la prospettiva di osservazione è diversa; mentre i dati analizzati dalla E.C.A. sono costituiti dai profili delle co-occorrenze entro i contesti elementari, i dati della cluster analysis, all’origine, sono costituiti dai profili delle occorrenze. Ciò porta a riflettere sulla particolare natura del rapporto tra dati e strumenti di osservazione nell’ambito dell’analisi dei testi. In effetti, a partire dal momento in cui diventano oggetti di osservazione, **i dati testuali sono costruiti dagli strumenti che utilizziamo per analizzarli**: diventano rappresentazioni del rapporto tra due tipi di modelli culturali, quelli “contenuti” nei testi e quelli utilizzati da chi li analizza.

## 5. DAI TESTI ALLA CULTURA

L’analisi dei testi, soprattutto quando è orientata all’analisi del contenuto, porta necessariamente a riflettere sul rapporto tra linguaggio e cultura, oppure – più in generale – sulle relazioni tra pensiero, linguaggio e cultura.

A questo proposito, il primo riferimento che viene alla mente è la cosiddetta “Sapir-Whorf Hypothesis”, nome con il quale vengono evocati alcuni contributi teorici di E. Sapir (1949) e del

---

<sup>16</sup> Generalmente, in questo tipo di analisi, la percentuale cumulativa della varianza (o inerzia) spiegata dai primi tre fattori risulta più che sufficiente (Lebart L. e Salem A., 1994).

<sup>17</sup> La varianza inter-cluster risulta dalla dispersione dei cluster nello spazio n-dimensionale; mentre la varianza totale risulta dalla sommatoria tra varianza inter-cluster e quella intra-cluster, la quale ultima risulta dalla dispersione degli elementi entro ciascun cluster.

<sup>18</sup> In questo caso, la misura della caratterizzazione è costituita dal chi-quadro, in modo analogo a quanto descritto nel punto 4.1 (funzione Specificità).

suo allievo B. Whorf (1956), che da decenni sono al centro di molte discussioni scientifiche. Una parte di questa ipotesi concerne il cosiddetto “relativismo linguistico”, ovvero l’affermazione che le differenze tra culture sono largamente determinate dalle differenze tra i rispettivi linguaggi, i quali determinano i differenti modi in cui “vediamo” il mondo e ce lo rappresentiamo: “We see and hear and otherwise experience very largely as we do because the language habits of our community predispose certain choices of interpretation. No two languages are ever sufficiently similar to be considered as representing the same social reality. The worlds in which different societies live are distinct worlds, not merely the same world with different labels attached” (Sapir, 1949: 162).

Ovviamente vale anche l’ipotesi inversa: le differenze tra culture determinano i differenti modi in cui usiamo le parole e il linguaggio. Tutto sta ad intendersi su cosa intendiamo per **culture**. Se le culture corrispondono ai diversi modi in cui i gruppi sociali costruiscono le loro rappresentazioni del mondo, costruiscono le loro esperienze, organizzano le loro relazioni e attribuiscono valore alle cose, le differenze tra culture non riguardano solo le “tribù” (vedi gli Indiani Hopi studiati da B. Whorf) o le nazioni. Differenti culture infatti sono quelle dei gruppi di adolescenti, quelle di alcune aziende, quelle di alcune professioni, quelle proposte dalle diverse religioni, e così via.

In ogni caso, tutte le culture costruiscono diverse rappresentazioni dei contesti; e queste, generalmente, si traducono in testi che possiamo analizzare.

A rigor di termini, nell’analisi dei testi, ogni volta abbiamo a che fare con un contesto diverso: quello definito dal corpus che abbiamo deciso di analizzare, più esattamente quello definito dalle relazioni tra CU e LU. Tuttavia, quello che osserviamo ci porta a costruire altri tipi di relazioni. Ad esempio se capita (come è capitato) di rilevare che sugli articoli di una rivista italiana, pubblicati dopo l’11 settembre 2001, la parola “Islam” è fortemente associata a “terrorismo”, l’interpretazione del “dato” non può non tener conto di quanto è accaduto - e sta accadendo - nel contesto mondiale.

Per molti versi, chi analizza testi analizza rappresentazioni di contesti e, a sua volta, propone altri contesti di rappresentazione; ma che tipo di cultura è quella prodotta da chi analizza i testi?

In ipotesi, l’analisi dei testi porta a costruire una sorta di meta-linguaggio, cioè apre la possibilità di non “essere parlati” dalla cultura, ma di poter parlare e di poter riflettere “su” di essa.

Anni fa, J.M. Lotman e B.A. Uspenskij (1973) hanno proposto di distinguere due tipi di culture: quelle che assumono i testi come **codici** di comportamento e quelle che sono più orientate a costruire **regole** per la creazione di meta-testi. Le prime, denominate **culture testualizzate**, sono mitiche e mono-linguistiche, cioè parlano soltanto il linguaggio della loro cultura; le seconde, denominate **culture grammaticalizzate**, introducono la possibilità di usare un metalinguaggio e, per definizione, sono poli-linguistiche e de-mitologizzanti.

Per molti versi, le culture de-mitologizzanti sono le migliori risorse per la costruzione di una società aperta.

La storia ci dice che la pratica dell’analisi dei testi è nata all’interno di culture mitiche, fortemente orientate a considerare la “sacralità” dei loro oggetti di studio. La scienza e le nuove tecnologie hanno aperto la strada a nuove possibilità di analisi; tuttavia molte risorse appaiono sottoutilizzate. Paradossalmente, gli analisti dei testi possono costruire una loro cultura, restare “chiusi” in essa e parlare solo il loro linguaggio.

---

## BIBLIOGRAFIA

- BARTHES R. (1964), *Eléments de sémiologie*, Paris, Seuil
- BENZECRI J.P & F. (1984), *Pratique de l'analyse des données. Analyse des correspondances & Classification*, Paris, Dunod
- BERTIN J. (1967), *Sémiologie graphique*, Paris, Gauthier-Villars Mouton
- BLOOMFIELD L. (1933), *Language*, New York, Holt, Reinehart & Winston
- BOLASCO S. (1999), *Analisi Multidimensionale dei dati. Metodi, strategie e criteri di interpretazione*, Roma, Carocci
- BURT C.L. (1940), *Factors of the Mind*, University of London Press, London
- GREENACRE M.J. (1984), *Theory and Applications of Correspondence Analysis*, New York, Academic Press
- GREIMAS A.J. (1966), *Sémantique structurale*, Paris, Larousse
- JAKOBSON R. (1963), *Essais de linguistique générale*, Paris, Editions de Minuit
- LEBART L. MORINEAU A. PIRON M. (1995), *Statistique exploratoire multidimensionnelle*, Paris, Dunod
- LEBART L. SALEM A. (1994), *Statistique textuelle*, Paris, Dunod
- LOTMAN J.M. USPENSKIJ B.A. (1973), *Tipologia delle cultura*, Milano, Bompiani
- MARANDA P. (1990), *DisCan: User's Manual*, Québec, Nadeau Caron Informatique
- RASTIER F. (1987), *Sémantique interprétative*, Paris, PUF
- SALTON G. (1989), *Automatic text processing: the transformation, analysis, and retrieval of Information by Computer*, Addison-Wesley, Reading, Massachussets
- SAPIR E. (1949), *Culture, Language and Personality*, The Regents of the University of California
- SAUSSURE (de) F.(1916), *Cours de Linguistique générale*, Lusanne-Paris, Payot
- TEIL G. E LATOUR B. (1995), *The hume machine. Can association networks do more than formal rules?*, in SEHR, volume 4, issue 2: Constructions of the Mind
- WHORF B. (1956), *Language, Thought and Reality*, Cambridge, Mass. MIT Press