

THE LOGIC OF A TEXT-SCOPE

by **Franco Lancia**

(© October 28, 2002)

web: www.tlab.it; mail: franco.lancia@tlab.it

Notice: This paper was written when the 2.0 version of T-LAB was first released, in order to help users understand the software logic better. As the current (September 2009) version is 7.0 and its logic is quite different, it is advisable to read the quick introduction available at <http://www.tlab.it/en/download.php>

INTRODUCTION

In the history of science the relationships between theories and instruments are in some ways reciprocal: the development of new theories leads us to construct new instruments, and in turn the use of new instruments leads us to develop new theories. For example, it was possible to construct parabolic antennas and communication satellites only after developing research and theories in physics and engineering (theories → instruments); likewise, many discoveries and theories of biology are unthinkable without the use of the microscope just as modern astronomy is unthinkable without the use of the telescope (theories ← instruments).

In fact, all technological instruments, above all those which aim to increase our powers of observation, are the concrete expression of some theory.

T-LAB is just one observation instrument: it belongs to the family of the software designed to produce maps that represent text content, whether single texts or several texts being compared (similarity and dissimilarity). The theories which regulate its operations, which are translated into transformation rules and organize the relationships between the “data” and their representation, belong to two disciplines: linguistics and statistics.

The aim of this paper is to clarify exactly the role that these scientific subjects play in the logic of T-LAB’s operations.

It is subdivided into five sections:

- 1- **Analysis Units and Theory of Meaning**, where the linguistics models are proposed which allow us to define the objects of observation;
- 2- **Two Analysis Routes**, where two kinds of strategies are featured that presuppose the choice of "what " we like to observe and to analyse;
- 3- **The Logic of Co-occurrences**, in which the instruments that produce the results typical of the first kind of strategy are described and commented on: associations, co-word mapping, elementary context analysis;
- 4- **The Logic of Comparisons**, in which the instruments that allow us to achieve the typical results of the second kind of strategy are described and commented: specificities, correspondence analysis, cluster analysis;
- 5- **From Texts to Cultures**, where some thoughts on the cultural purpose of text analysis are proposed.

1. ANALYSIS UNITS AND THEORY OF MEANING

In T-LAB, reference to linguistics is used principally in order to define and organize the “data” that are the objects of observation; in other words, in order to define the so-called Analysis Units, which are of two types:

- a- **Context Units** (CU), which are subsets resulting from the breakdown of the corpus into one or more texts. For example, if the corpus analysed consists of a set of newspaper articles, the context units can be: the single articles, the subsets of articles classified by a criterion (such as mast-head, day of publication, topic, etc), the single phrases into which every article can be split up (the elementary contexts).
- b- **Lexical Units** (LU), which are the single words, either used as "row forms", or taken back to lemmas (e.g. "working" -> "work"), or taken back to semantic classes (e.g. "bronchitis" -> "disease") or to dictionary categories (refer to the coding schemes used in the Content Analysis), in all cases they are indexed by their context of origin (CU).

This distinction, which has a theoretical foundation in linguistics and semiology, is of great practical importance: in fact, it allows the operational translations that are the basis of textual statistics.

As regards the theoretical foundation, it goes back to the hypothesis initially proposed by F. de Saussure (1916), and subsequently by several authors (R. Jakobson, 1963; R. Barthes, 1964), according to which the relationships between the linguistic elements can be analyzed as **syntagmatic relationships** and/or as **paradigmatic relationships**. The former regulate the combination of linguistic elements within contexts (one “near to” the other: CU), the latter determine the possibility of replacing a linguistic element with one that has something in common with it (one “in place of” the other: LU).

The practical importance of the proposed distinction arises from the fact that the relationships between the two types of Analysis Units (CU and LU) can be represented as matrices (or tables) whose numerical values indicate the instances of **occurrence** and **co-occurrence**.

For example, given a set of lexical units ($lu_1, lu_2, lu_3, \dots, lu_m$) and a set of context units ($cu_1, cu_2, cu_3, \dots, cu_n$), with "m" and "n" elements respectively, it is possible to construct a rectangular matrix "m X n", with "m" rows and "n" columns, whose single intersections (lu_i, cu_j) correspond to the occurrence values that indicate how many times each lexical unit (e.g. a word) occurs in each context unit (e.g. a newspaper article). Equally, if the context units are made up of phrases, we can construct one “m X m” square matrix, whose single intersections (lu_i, lu_j) correspond to co-occurrence values, that is the number of phrases in which the lexical unit (e.g. a word) co-occurs with each other.

In both the cases, in rectangular and in square matrices, each row and each column reassumes the "profile" of one Analysis Unit (CU or LU). As an example, in rectangular matrices "m X n" every row can be used to represent the profile of a word as a "distribution"¹ of its occurrences within the analyzed contexts.

¹ For this purpose reference to distributional linguistics (L. Bloomfield, 1933) is pertinent.

So the notions of **distribution**, **profile**, **occurrence** and of **co-occurrence**, are the keys to understanding the theories of meaning that are implemented in T-LAB's instruments.

In fact, in the software logic, each Analysis Unit (CU or LU) is known "only" according to its profile. More exactly:

- each CU is known according to the distribution of occurrences of each LU present within it;
- each LU is known according to the distribution of its occurrences within every CU, and also according to the distribution of its co-occurrences with each of the other LU.

Obviously, the software doesn't know the meanings (or contents), but only the signifiers ², that is the "strings" that individualize the LU and the CU; however, the relationships between signifiers (that is the syntagmatic relationships) assume meaningful shapes that, in some form, propose a **contextual representation of the meanings**.

Let us take, as an example, the lexical unit *par excellence*: the single "word". Leaving aside from the morphological categories and syntactic functions ³, each single word that we use is unlike others according to the "minimal units" or "distinctive traits" into which it can be broken down, either by signifier (or expression) or by meaning (or content): in the first case (signifier/expression) the result of the decomposition consists of the phonemes, in the second case (meaning/content) of the "semes" or the semantic traits. These last ones, in the model proposed by J. Greimas (1966) result from the combination of two types of elements:

- a) the **semic nucleus**, assumed invariable and considered as a lowest common denominator of all the meanings of the word being examined ⁴;
- b) the **contextual semes**, consisting of the "sense effects" determined from sets (or classes) of contexts.

Now, in T-LAB's logic, the meaning of each single word is known only through its relationships with the contexts, that is through the distribution of its occurrences (or co-occurrences) within the Context Units (CU). The reference to Greimas's "contextual semes" is therefore more than pertinent.

However, the statistical algorithms implemented in the software cannot be considered as mere translations of the theories proposed by structural semantics.

² According to De Saussure (1916) the "sign" (e.g. a word) is double-faced: it combines the signifier (or acoustic image) and the signified (or concept).

³ The morphological categories are "noun", "adjective", "verb" etc. The syntactic functions are "subject" and "predicate".

⁴ In the Greimas's terminology words, considered as units of meaning, are "sememes".

2. TWO ANALYSIS ROUTES

In T-LAB's architecture, the various tools (the "functions")⁵ and their algorithms are meaningful only if they are used within one **analysis strategy**. In effect, for every architecture the principle applies that it can be explored in order to see "how it is made" or "how it works", but it only acquires its proper meaning when used for some purpose. To put it metaphorically: it's one thing to visit a house, another matter to live in it .

Let's imagine a hypothetical investigator grappling with the analysis of some corpus. If he has decided to use T-LAB, its strategies are determined from two kinds of choice that, in any case, are reversible:

- 1- in order to observe the semantic organization of the contents within the corpus or within its subsets, (a) I want **to give priority to analysis of the co-occurrences between LU**; or (b) I want **to give priority to analysis of the similarities/dissimilarities between the corpus subsets (CU)**, resulting from the occurrence profiles and/or from the use of some variable?
- 2- in either case (see the previous "a" and "b"), the LU that I want to use are the words and/or the lemmas, as classified by T-LAB (either automatically or by the user), or do I prefer to use a "coding scheme" or a dictionary to be imported?

The procedures for activating these choices are fully explained in the help of the software. The notes that follow are destined to clarify the implications of two strategies (point 1), which - respectively - appeal to the logic of co-occurrences (see below 3.1) or to the logic of comparisons (see below 3.2). As regards point (2), I shall limit myself to a few clarifications.

In T-LAB's databases, all single words of the corpus are indexed according to their position in the different CU. Subsequently, each word is automatically assigned a **label** that generally corresponds to a standard dictionary lemma. Finally, either by targeted interventions, by using imported dictionaries or by using the results carried out by previous analyses, the user can change the label denominations.

The outline adopted is as follows:

⁵ Some T-LAB's tools allow us to do statistical analyses, others (e.g. Disambiguation, Multi-Word list, Concordance, etc) offer support to the analyses.

STARTING		CUSTOMIZATION		RESULT	
WORD	LABEL	WORD	LABEL	WORD	LABEL
word 1	lemma a	lemma a	label A	word 1	label A
word 2	lemma a	lemma a	label A	word 2	label A
word 3	lemma a	lemma a	label A	word 3	label A
word 4	lemma b	lemma b	label A	word 4	label A
word 5	lemma b	lemma b	label A	word 5	label A
word 6	lemma b	lemma b	label A	word 6	label A
word 7	lemma c	lemma c	label B	word 7	label B
word 8	lemma c	lemma c	label B	word 8	label B
word 9	lemma c	lemma c	label B	word 9	label B
.....

Table 2.1

Briefly, in order to change the word label, the user must put in dual matching entries from his dictionary and the lemmas proposed by T-LAB. In this way it can take "n" reversible decisions such as "Instead naming this word X, I want to name it Y"⁶. Obviously, **all the substitutions make sense only if they respect some paradigmatic criterion**, that is if the word-label associations respect a semantic criterion (see the definition of LU: 1).

3. THE LOGIC OF CO-OCCURRENCES

3.1 ASSOCIATIONS

Setting out to explore the logic of the co-occurrences, we start from the function "Associations". In the T-LAB's library, the algorithms implemented in it are the simplest: for each selected LU (word, lemma or category), its occurrences and its co-occurrences within the different CU are known so it is possible to compute an association index with each of the other LU (words, lemmas or categories) of the corpus or of a corpus subset.

In this case, the index used is the **cosine coefficient** (G. Salton, 1989). Its formula is as follows:

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

where the cosine coefficient (C) between every LU pair (X,Y) is defined as the ratio of the sum of their co-occurrences and the sum obtained by multiplying the square roots of the respective occurrences within the CU⁷.

⁶ In T-LAB the function that allows us to do this kind of operation is "Tagging".

⁷ In this case, the sum of the co-occurrences is the result of the number of elementary contexts (phrases or segments) where X and Y are present at the same time.

From the computations carried out, T-LAB produces charts like the following:

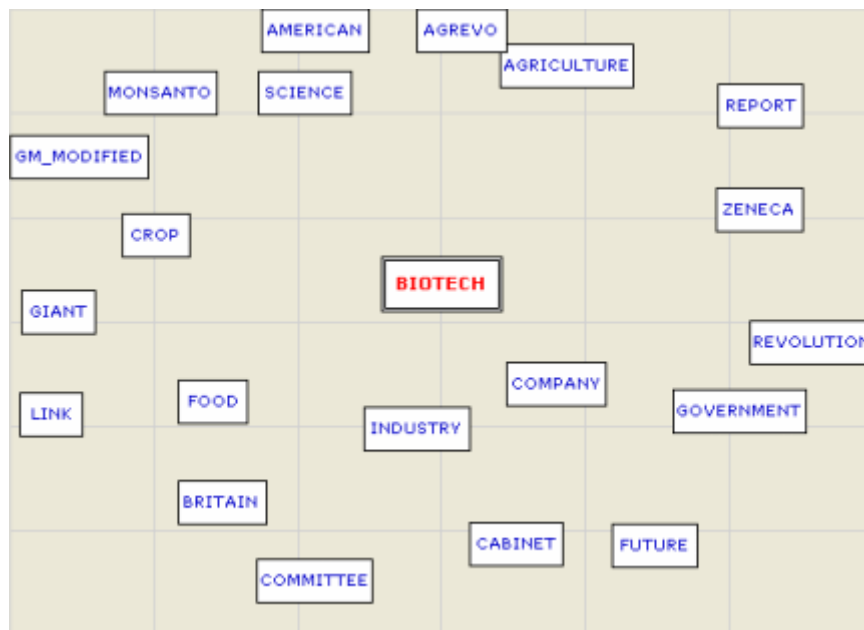


Fig. 3.1

The selected word (LU), whose **contextual meaning** is to be verified, is placed in the centre; all the others are distributed around it, each at a distance proportional to its degree of association (the words nearer to the centre have the highest cosine coefficient).

The meaningful relationships are therefore one-to-one, between the word in the centre and each of the others.

Obviously, the meaning isn't "given" by the software. In analogy with other observation instruments, we could say that T-LAB works like a text-scope: it shows things that can be "seen" only by someone who is able to interpret them.

3.2 CO-WORD MAPPING

The statistical algorithms implemented in this function are applied, always and only, to co-occurrence matrices. More specifically, the starting matrix consists of as many rows and as many columns (square matrix) as there are selected LU (words, lemmas, categories).

The fixed analysis steps are as follows:

- a) construction of one co-occurrence matrix ($l u_i \times l u_j$);
- b) computation of cosine coefficient for all the matrix values (A);
- c) hierarchical clustering with stop to "N" nucleuses, each with a number of LU (min 3, max 13);
- d) construction of a second co-occurrence matrix (B) with dimensions $N \times N$;
- e) graphical representation of (B) by means of correspondence analysis.

At the end of the computations a map is shown as follows:

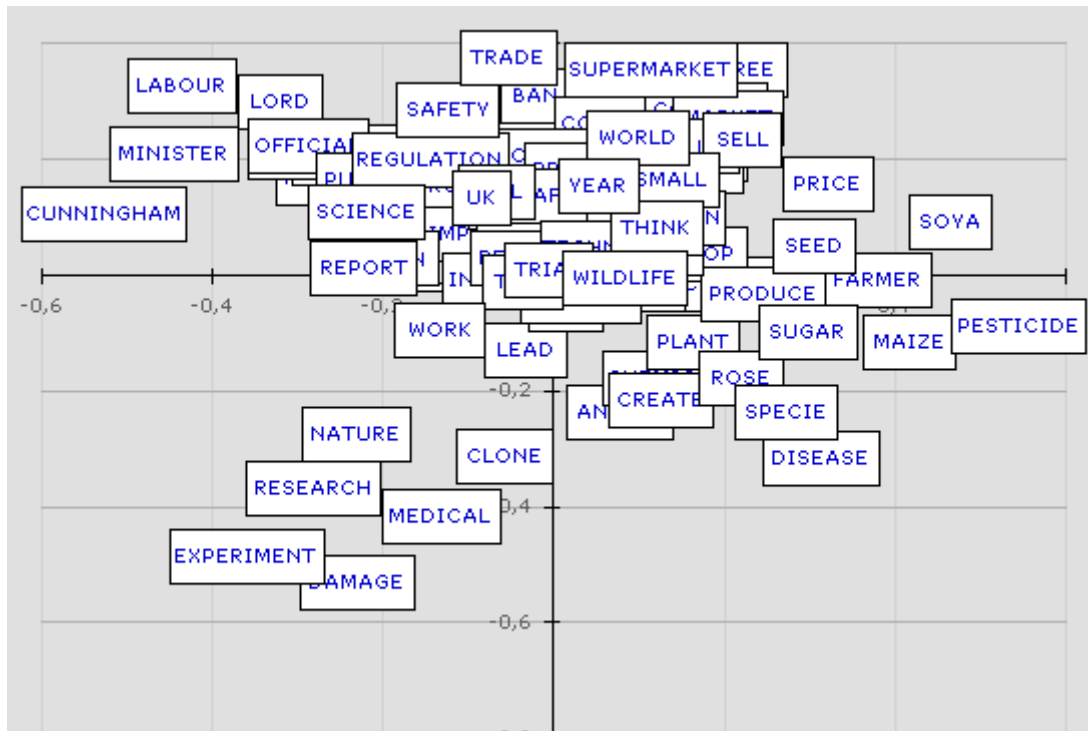


Fig. 3.2

Each label of the map represents a cluster (or nucleus) of "associated" words (LU), co-occurring in the considered CU, and the composition of every cluster can be verified with a simple click of the mouse.

In some ways, the logic of this function corresponds to the "Hume machine" described by G. Teil and B. Latour (1995); in fact, T-LAB's ancestors include both software applications cited by these authors: Leximappe and Candide⁸. However, in T-LAB the way in which "the simple network of unstructured co-occurrences enables us to produce structuring differentiations"⁹ is different because it consists of two methods: cluster analysis and correspondence analysis.

In particular, the maps produced (see above) are organized by factorial axes; therefore all the relationships shown can be read and interpreted through tables with several measures. Of these, the most useful are Test Values¹⁰; this measure has two important properties: a threshold value to reject the null hypothesis ($p = 0.05$) and a sign (-/+). This means that by sorting the values in increasing or decreasing order, according to whether the values are considered to be on the "negative" pole (-) or "positive" pole (+), it is possible to appreciate the weight of each object on each factorial axis and - therefore - to define the **semantic oppositions** that organize various two-dimensional spaces.

In other words, the representations proposed by this function enable us to observe and interpret the semantic organization of the corpus or of the corpus subsets.

⁸ Leximappe and Candide were developed by the « Centre de Sociologie de l'Innovation (CSI) de l'Ecole des Mines de Paris (ENSM) ».

⁹ Quotation from G. Teil and B. Latour (1995)

¹⁰ A detailed description of this measure is found in Lebart L. Morineau A. Piron M. (1995: 123-125).

3.3 ELEMENTARY CONTEXT ANALYSIS

In T-LAB, the logic of "Associations" is in some form included in "Co-Word Mapping"; the logic of which is in turn included in "Elementary Context Analysis" (E.C.A.). In fact, in all three cases, the data matrices consist of the co-occurrences and, in all three cases, the aim is to provide representations that help to explore the contextual organization of meanings without the need to pre-figure the use of the variables chosen by the researcher.

To be exact, in E.C.A. logic the variables are only "labels" that enable us to find out how the Analysis Units (CU and LU) belong to the corpus; and they are used only "a posteriori" in order to characterize the detected clusters.

The aims and the outputs of this function (E.C.A.) resemble ALCESTE¹¹; however the data tables and the algorithms used are very different.

The starting point is a co-occurrence matrix, to be exact the one obtained by Co-Word Mapping (see 3.2 step "d"). Subsequently a rectangular matrix is constructed (m X n) with as many rows as there are Elementary Contexts (E.C. = phrases or text segments) and as many columns as many are the "thematic cores"; the elementary contexts are assumed as "objects" to be classified (clustering) and the "thematic cores" are assumed as their characteristics. More specifically, the profile of each row consists of one/zero (1/0) values marking the presence/absence of every nucleus within it.

The consecutive steps are the following:

- a- a correspondence analysis applied to (m X n) matrix;
- b- a first cluster analysis (K-means clustering) that uses the object coordinates on the first three factors obtained through the previous correspondence analysis;
- c- a second cluster analysis (Ward's method) that defines the E.C. classification;
- d- graphical representations by means of correspondence analysis (rows: LU and CU; columns: the obtained clusters);
- e- characterization, using chi-square tests, of each cluster obtained.

The maps produced enable us to see the relationships between clusters and variables within two-dimensional spaces. Various tables concur to evaluate the characteristics of the clusters and factorial poles, those that turn out to be meaningful through the use of chi-square test (clusters) or of Test Values (factors). In both cases the characteristics are made up of LU and CU.

In some ways, the result of this analysis can be considered as **isotopy mapping** (iso = same; topoi = places)¹², which refers to a conception of meaning as "contextual effect", that is to say something that does not belong to words considered one by one, but results from their relationships within context units (CU). In fact, each cluster characterizes a context "shared" by several words, but not resulting from their specific meanings. This is the logic whereby the whole is something different from the sum of its parts. Consequently, each cluster allows the retracing of a "thread of speech" within the whole web made up of the corpus under analysis or its subsets.

¹¹ ALCESTE (Analyse des Lexèmes Co-occurents dans les Énoncés Simples d'un TExt) was developed by Max Reinert.

¹² The notion of isotopy was initially proposed by A.J. Greimas (1966) in order to define the recurrence, in syntagmatic units (phrases and/or texts), of several words sharing semantic features.

However we have to remember that isotopy recognition is not the mere ascertainment of a "datum", but the result of an interpretation process (F. Rastier 1987).

3.4 SEQUENCE ANALYSIS

The function "Sequence Analysis" constructs and analyzes two co-occurrence matrices (m X m) whose respective values are the count of how many times, within the analyzed corpus, each lexical unit precedes or follows the other in the linear (sequential) structure of the texts. Soon after, the same values are converted to probability values, or rather to transition probabilities, because the relationships between the elements considered (**predecessors** and **successors**) constitute a markov chain.

As we know, given any finite set of events (the so-called "states") that occur in succession (the so-called "transitions"), markov chains enable us to find the network of their mutual relationships by means of probability links. An application of this model to discourse analysis is DiscAn, a software programme designed by a Canadian anthropologist (P. Maranda, 1990) and proposed as an instrument that - beginning from a content analysis - enables the user to construct semantic maps and dynamic networks.

As far as T-LAB is concerned, its ambitions are more circumscribed; however it facilitates the process of content analysis (tagging function and use of dictionaries) and - above all - it allows you to explore the relationships between a considerable number of "nodes" arising from the LU (max 250).

4. THE LOGIC OF COMPARISONS

4.1 SPECIFICITIES

A corpus can be subdivided into "n" subsets and also by several partition criteria (such as variables). For example, a book can be subdivided in chapters or in extracts dealing with different topics; a set of interviews can be subdivided according to the sex or profession of the interviewee, according to the date or according to the topics considered; equally it can be processed in order to analyze the answers to questionnaires with open-ended questions, and so on.

The questions that the "Specificities" function enables us to answer have the following form: given a variable (the partition criterion) and a corpus subset defined by a category ¹³, what are the "typical" characteristics of this subset that differentiate it from the other corpus subsets?

Obviously, the characteristics analyzable by T-LAB are the words used, that is the Lexical Units (LU) and their profiles. As in comparisons, they are one-to-one between the occurrences of each LU in the subset considered and in the rest of the corpus. The computation algorithm uses the chi-square test.

Here is an example. Let's assume that the corpus consists of four articles from as many newspapers (A, B, C, D) and let's assume that we are interested in calculating the "specificities" of the "A" mast-head. In this case, for each word under analysis (LU) the reference table is similar to the one below, in which it is assumed that: a) the LU under examination ("x") has 6 occurrences within the article "A" and 8 occurrences in the other articles (B, C, D); b) the total of the occurrences in "A" is 456, while the total of the occurrences in the other articles is 3708.

	in "A"	not in "A"	
UL "x"	6	8	14 (N _i)
other UL	450	3700	4150
	456	3708	4164 (N _{ij})
	(N _i)		

Table 4.1

For each cell the expected frequencies (E) are estimated in the following way: $(N_i \times N_j) / N_{ij}$; while the chi-square is estimated by the following formula:

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

¹³ For example, using "occupation" as a variable, the categories can be the following: physician, lawyer, psychologist, etc.

where "O" stands for the observed frequencies (6, 8, 450, 3700). It follows that the chi-square value is 14.664. Since in this case ($df = 1$; $p = 0.05$) the chi-square critical value is 3.84, the null hypothesis is refused and the LU under examination can be considered "specific" to article "A".

4.2 CORRESPONDENCE ANALYSIS

In T-LAB, Correspondence Analysis (CA) is both the name of a specific function and a computation procedure implemented in various functions. The description of its algorithms would require several pages and would necessarily include many formulas; therefore we would prefer not to spend time on these concerns and, to the interested reader, we suggest the following bibliographical references: J.P. Benzecri (1984) and M. J. Greenacre (1984) Lebart L. Morineau A. Piron M. (1995).

To put it simply, CA is a technique that allows us to represent the relationships between the row profiles and the column profiles of the matrices with frequency values. In order to describe how it works, at least in the first instance, we can leave aside mathematical and statistical notions. Suffice it to say that the operations implemented in its algorithms allow us to obtain two kinds of results:

- a) **to trace the regularities in the data tables** through a crosscheck of all the profiles (rows and columns) in the mutual similarity-difference relationships, with the result that – through a series of permutations – the tables can be re-sorted and the information they contain made “readable”.
- b) **to reduce the dimensions within which data can be represented**, by means of new variables (the factors) which correspond to the spatial coordinates of profiles (rows and columns). In this way, the data initially scattered at random in a n-dimensional space, are assembled within a reduced space, defined by the few factors that, in a statistically significant way, explain their variability.

Let's consider a table that reproduces an occurrence matrix with ten rows and four columns ¹⁴:

¹⁴ In the data matrices used by T-LAB there can be several thousand rows by several hundred columns.

	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

Table 4.2

At first we can represent the profiles (rows and columns) by means of histograms.

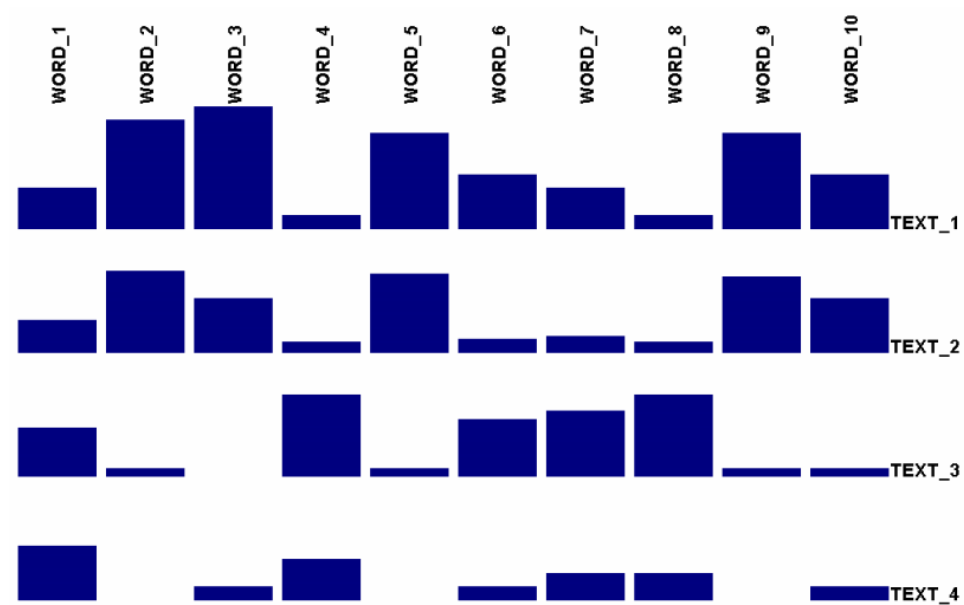


Fig. 4.1

Subsequently, using AMADO, a graphical software application developed by J. Bertin (1967), we can rearrange the relationships between rows and columns by means of correspondence analysis. More exactly, the re-sort is carried out using the first extracted factor. As can be seen, from a table with scattered data we obtain an assembled and, in some measure, readable structure.

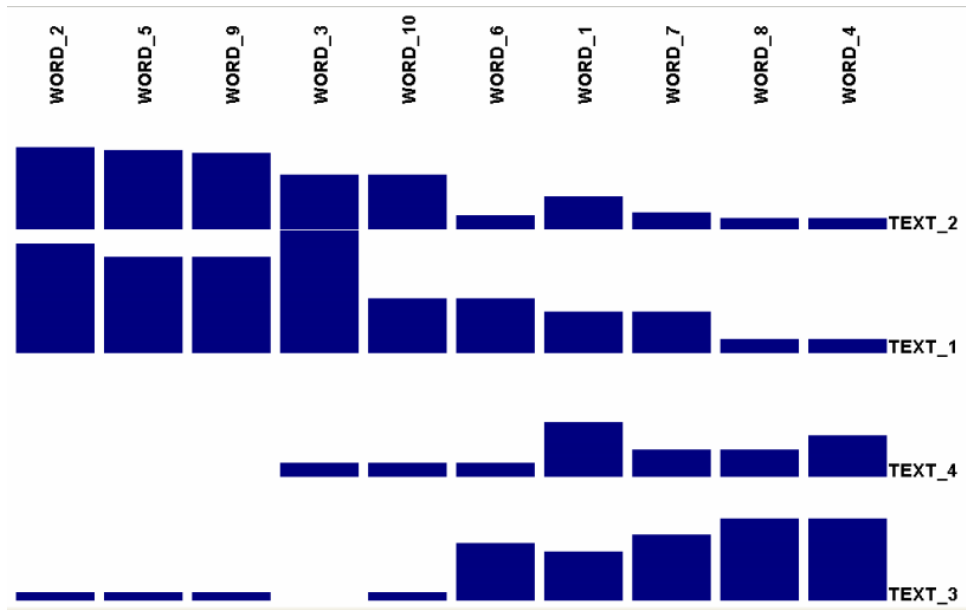


Fig. 4.2

The same data in table 4.2 can be represented by means of a classic two-dimensional chart (Fig. 4.3) that, as can be seen, is coherent with the representation of the profiles (Fig. 4.2). Particularly, in both cases the first factor turns out characterized by the prevailing "weight" of two CU (Text_2, Text_1) and of two LU (word_2, word_5) on the negative (-) pole, while on the positive pole (+) two other couples of CU (Text_4, Text_3) and of LU (word_4, word_8) prevail.

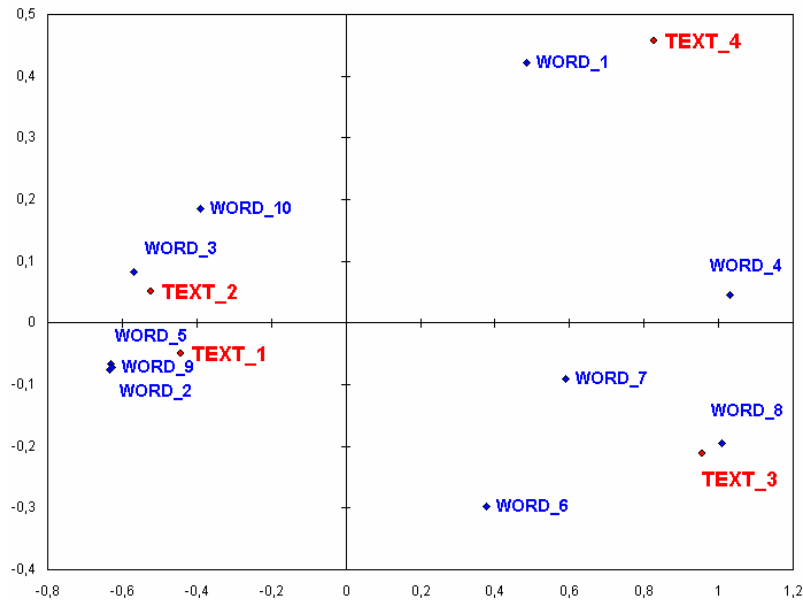


Fig. 4.3

In order to interpret the AC results, various measures are used, either concerning the "weight" of each factor (Eigenvalue or Inertia), or reporting the coordinates and the contributions (absolute and relative) of each object (row or column) on each factorial axis. All these measures are also

provided by T-LAB; however, as a priority, the software shows some tables with the Test Values, which we have already discussed (see above 3.2) and whose formula is the as follows¹⁵:

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

For example (see over Table 4.2), let's an element j (TEXT_4) with a total occurrences $n_j = 70$ and a coordinate on the first factorial axe $\varphi\alpha_j = 0.8276$. Because the corpus has a total occurrences $n = 598$, the Test Value is computed as following:

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

When we interpret the results of a factorial analysis, we often risk getting lost in a maze of tables. In order to avoid this risk, or at least to reduce it, reference to some definitions can be useful. In some ways, the factors can be considered as **classification principles** (C.L. Burt, 1940), as organizers of the relationships between the data that put together similar things, distinguish them from different things, construct kinship between categories of things. J.P. Benzecri, one of the mathematicians that has contributed most to defining the CA model, wrote: "Understanding a factorial axis means finding what is similar, firstly all that is on the right of the origin (barycentre), and secondly all that is on the left of it, and then expressing concisely and exactly the opposition between the two extremes" (1984:302).

With this assertion, while describing an interpretation method, the author in effect communicates a specific conception of factors as organizers of contrasting relationships between sets or classes ("all that" is on the right and "all that" is on the left of the origin), going as far as to say that he shares a conception of factors as classification principles.

In effect, even if the word factor suggests a causal relationship, the factorial analyses "only" serve to find an order in the complexity of the data analyzed, helping to reduce the space dimensions in which the data can be represented. But, obviously, the statistical (or geometric) meaning of the factors is one thing and the models for interpreting them within each discipline are another. On the other hand, if science did not try to explain the factors that generate some order in the phenomena studied, it would have no reason to exist.

In conclusion, it can be useful to distinguish two types of factors:

- (α) factors indicated by statistical computations, corresponding to mathematical and geometric structures (the dimensions), which organize the relationships between data and which have been called **classification principles**;
- (β) factors that, using inferential processes and making reference to theoretical models, are evoked in order to find interpretations and/or explanations for the forms taken by the (α) factors. These factors – of the second type (β) - can be called **explanation principles** and they are specific to each scientific subject

¹⁵ A detailed description of this measure is found in Lebart L. Morineau A. Piron M. (1995: 123-125).

4.3 CLUSTER ANALYSIS

In T-LAB's tools three kinds of clustering algorithms are implemented; one of these is specific to the function that can be used only after a previous correspondence analysis.

Now we must pause to consider its "logic". As regards literature on Cluster Analysis, every reader can easily find various references. However, some simple definitions may be useful. The following are taken from T-LAB's help:

Generally speaking, technical statistics of Cluster Analysis aim to detect object clusters with two complementary features:

- High internal (within cluster) homogeneity;
- High external (between cluster) heterogeneity.

In general, there are two kinds of Cluster Analysis techniques:

- **Hierarchical methods**, whose algorithms rebuild the whole hierarchy of the objects under analysis (the so called "tree"), in ascending or descending order;
- **Partitioning methods**, where the user previously defines the cluster numbers into which the set of objects under analysis is divided.

The method of analysis implemented in the T-LAB function that we are considering is actually of hierarchical type or, to be more precise, of hierarchical-ascending type: meaning it starts from as many elements as there are single objects to be clustered and – using proximity measures – aggregates them until it regroups the whole set. In this case the "objects" are the LU (words, lemmas or categories), each characterized by the profile of its coordinates on three first factorial axes¹⁶ obtained through a correspondence analysis. The aggregation criterion, according to their distances, is Ward's method, one of the most suited to hierarchical clustering which uses factorial coordinates (Bolasco S., 1999).

With regard to the partition criterion, which determines the number of clusters obtained, in T-LAB an algorithm is implemented that computes the ratio of inter-cluster variance to total variance¹⁷ and which - automatically - assumes as "optimal partition" the one where the ratio exceeds the .50 (50%) threshold.

There are two kinds of outputs:

- charts show the position of the clusters within two-dimensional spaces that correspond with those obtained through the previous correspondence analysis;

¹⁶ Generally, in this type of analysis, the cumulative percentage of the variance (or inertia) explained by the first three factors turns out more than sufficient (Lebart L. e Salem A., 1994).

¹⁷ The inter-cluster variance is the result of the dispersion of the cluster in the n-dimensional space; while the total variance turns out from adding the variance inter-cluster to the intra-cluster variance, which in turn is the result of the dispersion of the elements within each cluster.

- tables, one for each cluster, list the LU (words, lemmas or categories) that comprise it (its "elements") and the CU (variable categories) that characterize them ¹⁸.

The clusters obtained, as in E.C.A, show some instances of isotopy (see above 3.3). Obviously, in both cases, the observation perspective is different; while the data analyzed by E.C.A. come from the co-occurrence profiles within the elementary contexts, the data of cluster analysis, at source, come from the occurrence profiles. This leads us to think about the particular kind of relationships between data and observation instruments in the sphere of text analysis. In fact, from the moment they become observation objects, **textual data are constructed by the instruments that we use to analyze them**: in other words they become representations of the relationship between two kinds of cultural models: those contained in the texts analysed and those used by whoever analyzes them.

5. FROM TEXTS TO CULTURES

Text analysis, especially when its aim is content analysis, leads us to think about the relationship between language and culture, or - more broadly - about the relationships between thought, language and culture.

On this topic, the first reference that comes to mind is the so-called "Sapir-Whorf Hypothesis", named after some theoretical contributions by E. Sapir (1949) and his disciple B. Whorf (1956), which for many decades has been at the heart of many scientific arguments. One part of this hypothesis concerns so-called "linguistic relativism", which states that differences between cultures are widely determined by differences between their respective languages, which determine the different ways we "see" the world and represent it: "We see and hear and otherwise experience things 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).

Obviously the inverse hypothesis is valid: the differences between cultures determine the different ways in which we use words and language. It all depends on what we mean by "cultures". If cultures correspond to the different ways in which social groups construct their world representations, construct their experiences, organize their relationships and assign value to things, differences between cultures do not only concern "tribes" (see the Hopi Indians studied by B. Whorf) or nations. Different cultures are in groups of adolescents, in some companies, in some professions, in proposals from the various religions, and so on.

In any case, all cultures construct different representations of contexts; and these, generally, are turned into texts which can be analyzed.

Strictly speaking, in text analysis, we are dealing every time with a different context: the one defined by the corpus that we have decided to analyze; more exactly the one defined by specific relationships between CU and LU. However, what we observe leads us to construct other kinds of relationships. For example if - as has happened - we find in some articles in an Italian magazine, published after September 11th 2001, that the word "Islam" is strongly associated with "terrorism", the interpretation of the "datum" cannot neglect what happened - and what is happening - in the world-wide context.

¹⁸ In this case, as in 4.1 (Specificities), the measure used is the chi-square test.

In some ways, anyone who analyzes texts also analyzes context representations and, in their turn, proposes other representation contexts; but what sort of culture is produced by whoever analyzes the texts?

In theory, text analysis leads us to construct a kind of meta-language, that is to say it opens up the possibility of not “being spoken” by the culture, but of being able to speak and think “about” that culture.

About thirty years ago, J.M. Lotman and B.A. Uspenskij (1973) posited two distinct types of cultures: those that assume texts as **codes** of behaviour and those that are more oriented towards constructing **rules** for the creation of meta-texts. The first, called **textualized cultures**, are mythical and mono-linguistic, that is they speak only the language of their culture; the second, called **grammarized cultures**, start using meta-languages and, by definition, they are poly-linguistic and de-mythologizing.

In some ways, de-mythologizing cultures are the best resources for constructing an open society.

History says that the practice of text analysis was born in mythical cultures, strongly oriented towards considering the sacredness of their study objects. Science and the new technologies have opened a door to new possibilities of analysis; and yet many resources appear underused. Paradoxically, text analysts can construct their own culture, remain closed in it and only speak their own language.

REFERENCES

- 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), *Factor 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