THE LOGIC OF THE T-LAB TOOLS EXPLAINED

by Franco Lancia
(© October 2012)
web: www.tlab.it; mail: franco.lancia@tlab.it

It is the theory that decides what can be observed. (A. Einstein)

ABSTRACT

While commenting on the architecture of the T-LAB system for text analysis, the author addresses several methodological issues, also maintaining that 'true' textual (or content) analysis deals just with what precedes and follows any use of statistical tools. By using an intuitive and geometric approach, he clarifies the processes through which the reciprocal relationships between textual units (e.g. words, sentences, documents, etc.) can be explored and automatically analysed by means of the various software tools. He comments on several works carried out by qualified researchers and on the questions/problems which animate their analysis strategies. The logic of the T-LAB tools they are using, as well as the conceptual framework within which both the 'building blocks' and the design of the software system make sense, are fully explained.

Citation:
Lancia, F. (2012). *The Logic of the T-LAB Tools Explained*. Retrieved [date] from http://www.tlab.it/en/toolsexplained.php.

Word count: 11970.

1 – Rethinking 'textscope' and its uses

This paper was initially conceived as a revised version of *The Logic of a Text-scope*, which was published over the Internet on October 2002 (that is, ten years ago). At that time 'textscope' was the neologism I invented to present T-LAB as an 'observation instrument'[1] and to help users better understand its logic[2]. From that time forward – and to my

[1] In the paper quoted above the meaning of such a neologism was explained as follows: "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". Actually it would seem that the neologism I invented has had some interesting fortune and use over the Web.

[2] As a matter of fact, both the 'logic' of the software and that of its possible uses are explained in various documents available for free download (see http://www.tlab.it/en/download.php), as well as in a book I published some years ago (Lancia, 2004). Moreover, by making reference to the notions of *isotopy* and *abductive inference*, in another work (Lancia, 2007) I tried to explain the logic of two interrelated processes which – actually – do not relate to the use of the T-LAB tools only, but rather to the use of a big 'family' of software tools. Such processes refer to: (a) what happens when – within the vector space model approach - words become numbers and (b) what happens when attempting to interpret the multi-semiotic texts produced by the software (i.e. outputs like tables and graphs).

wonderment - thousands of people have downloaded and, perhaps, also read this piece of writing. Meanwhile, both the architecture of the T-LAB system and that of several tools integrated in it have undergone relevant changes. As a consequence, now (i.e. 2012) much of the information contained in the paper quoted above would require critical examination.

So, after more than twelve year of ongoing research and development, I thought that it would be useful to address some points concerning the *conceptual framework* within which both the design and the architecture of such a product make sense. In doing so, I have also made reference to the literature dealing with its uses and, as a sort of peer-review process, I have submitted a draft to a dozen colleagues and qualified researchers for comment[3]. Some of them suggested that I should have made this paper suitable for a scientific journal and so not just publish it on the web. I will consider such a suggestion in the future; however, at the moment, I prefer dealing with scientific issues whilst using a figurative approach in my writing.

In order to work in this way, before entering into details, and taking a moment to focus the 'external' architecture of the software, I would like to recall that the way the T-LAB sub-menus[4] group the various analysis tools is just indicative and that it works like the 'signs' which help people when visiting a building with several floors and many offices, each one of them – in this case - can be used as a specific 'laboratory' (in fact 'text laboratory' is the word phrase which 'T-LAB' refers to). However the true way of connecting the uses of various tools is – at all times - the path followed by any researcher, that is *his* method[5], which usually is a 'process' involving *his* theories, *his* study subject, and *his* reference context (e.g. colleagues, clients etc.).

Actually, unlike that of most similar software, the T-LAB architecture is neither *one-way moving* nor *algorithm centered.* More specifically, it is not 'one-way moving' because the researcher is allowed to choose between several paths of analysis; furthermore it is not 'algorithm centered' because it relies on the assumption that statistical algorithms are just 'tools' for extracting patterns which are already present in the data structures. Moreover, I must to point out that, when text analysis software is algorithm-centered, its supposed competitive advantage (e.g. a specific clustering method or a specific topic model approach) is usually also a sort of *Trojan horse*, the implicit theories of which tend to replace the 'constructive' work of the researcher.

Obviously, there is a number of 'pre-set programmes' in the T-LAB system which can be used for convenience and which do their job properly; however the system has been designed to be *question-answer oriented*, to allow the use of interim outputs as inputs to further analysis, and to make any customisation easy[6]. So, in principle, any result of any analysis process should be perceived as being strictly connected with the user strategy rather than with software magic.

In my view, the added 'scientific' value of the T-LAB system resides precisely in the uniqueness of its architecture, that is to say in its *flexibility* and in its *transparence*. In

[3] I wish to acknowledge the valuable suggestions and feedback received from Sergio Salvatore, Cheryl Schonhardt-Bailey, Guendalina Graffigna, Heike Klüver, Lorenzo Montali and Alberto Trobia.

[4] See, for example, the distinction between tools for 'co-occurrence analysis', tools for 'thematic analysis' and tools for 'comparative analysis'.

[5] Etymologically, 'method' comes from the Greek word 'methòdos' ('meta' + 'hòdos'), meaning 'to follow a path'.

[6] See, for example, how T-LAB allows the user to build, import, export and use various dictionaries and word lists.

other words, its software architecture should be in fact 'observed'[7] as an implementation of the principle that 'science is about method'. So, in the following pages, I will try to clarify the 'logic' and the rationale of such architecture also by commenting on the literature dealing with its uses. However, I am aware that several other scientific issues would deserve to be carefully discussed: some more 'general' ones, which question any method for automatic textual analysis (see, for example, the so-called 'qualitative/quantitative divide'), and some more 'specific' which lead to the assessment of the reliability of the various T-LAB tools.

In relation to the 'general' issues, here I limit myself to arguing that, as lots of researchers well know, *it is not true* that the 'information' provided by the T-LAB tools (or by any similar software) is not relevant within the so-called *grounded theory* approach. Equally I argue that *it is not true* that the main reason for using software like T-LAB is that it allows us to automatically process a huge 'quantity' of documents that otherwise (i.e. by reading and manually coding) would not be analysable. Simply put, this type of software provides *new* information and *new* ways for knowledge discovery, either by spotting *patterns* or *linkages* within (and between) texts.

In relation to the more 'specific' issues, I would like to point out that – in research - the need to assess the reliability of any software tool originates from 'substantive' matters concerning the study subject and the data under examination. In fact, qualified researchers are usually interested in finding reliable answers to their *questions/problems* and not just in commenting on the software outputs. To this purpose, I would like to refer to the findings of two recent works which actually deal with 'substantive' issues which are very different from each other:

1. Cheryl Schonhardt-Bailey (2012) – who is a Reader in Political Science (London School of Economics) and is working on a project that 'seeks better to understand deliberations on US monetary policy' over many decades - while trying 'to assess the extent to which different automated content analysis software yield broadly similar results', has compared the results of *Alceste* (author: Max Reinert*)*, *Dtm-Vic* (author: Ludovic Lebart) and that provided by a T-LAB tool for thematic analysis. So, through such a sort of triangulation, she has concluded that 'We are more certain that *our results and interpretations* of the oversight hearing in the House and Senate banking committees are sound' (ib., p.18. My emphasis).
2. A team coordinated by Professor Sergio Salvatore (University of Salento, Italy) – which uses automated content analysis as 'a device for psychotherapy process research' - seems to have assessed that, when classifying text segments of a fixed length, blind human coders and T-LAB obtain very similar results; so, by using the Cohen's Kappa as inter-coder agreement measure, the researchers argue that the logic of a T-LAB tool for thematic analysis satisfy the 'Turing-like' criterion of validity (Salvatore et al., 2012). It is noteworthy that the main aim of such researchers is to validate *their* method and that they consider the above findings just a 'first step', and the results are considered 'encouraging but far from definitive' (ib., p. 17).

Not by chance, both the above researches refer to a T-LAB tool (i.e. *Thematic Analysis of Elementary Contexts*) which uses an 'unsupervised' method for clustering textual units,

[7] See Einstein's quotation in the epigraph of this paper.

and such a method (not just the specific algorithm used) is, for complementary reasons, at the same time ‘powerful’ and ‘weak’. In fact, it is powerful because it looks for similarities in a ‘human-like’ way, and – for this very reason – it is also weak: in fact, the way data are ‘partitioned’ into groups (i.e. clusters) needs the human being as a sort of referee.

Furthermore, by considering the relevance of the thematic approach in text analysis, in section ‘7’ of this paper (see below) I will try to explain how various T-LAB tools deal with such a difficult matter[8]. Now, without discussing the above and in order to introduce some architectural issues, I would like to recall another idea of Alan Turing: that of the so-called ‘universal machine’, i.e. a machine which – through software ‘logic’ – is not ‘single-purpose’ and allows the user to manage very different tasks. In my opinion, the best way of ‘imagining’ such a machine is not to think that it is a ‘one’ all-powerful device, rather that it can be a virtual architecture which allows connecting the *building blocks* of any ‘procedure’.

Just to give an example, *littleBits* – as stated in its website (see http://littlebits.cc/) which the Figure 1 below comes from – ‘is an open source library of electronic modules that snap together with tiny magnets for prototyping and play’. In fact, each one of the below little and colored electronic modules has a specific function (e.g. make lights, make sounds, be a motor, a sensor, etc.) and the way they can be assembled (i.e. by tiny magnets) is very simple, so that people – including kids – can easily build their preferred device.

Figure 1: The *littleBits* modules

Actually, the T-LAB tools can be considered like *littleBits’* modules waiting to be connected to each other by some skilled user while trying to follow ‘his’ research path. And, and at the same time, the system architecture is ready for new modules to be added, which could perform specific tasks like lemmatisation in more languages, new kinds of statistical analyses, and so on.

2 - Dealing with some architectural issues

In relation to the T-LAB system I can simply say that presently it is the result of a sort of *triangulation* between three points: (a) the logic of specific *procedures*, (b) the model of an ‘imaginary’ *universal machine*, and – above all - (c) the *needs of users*[9] working in

[8] For further information on how T-LAB deals with thematic analysis, see Lancia (2012b).

[9] Actually, beyond its ‘logic’, the fortune of the T-LAB system resides in the *virtuous circle* between our

different fields (e.g. social psychology, marketing research, political science, linguistics, etc.).

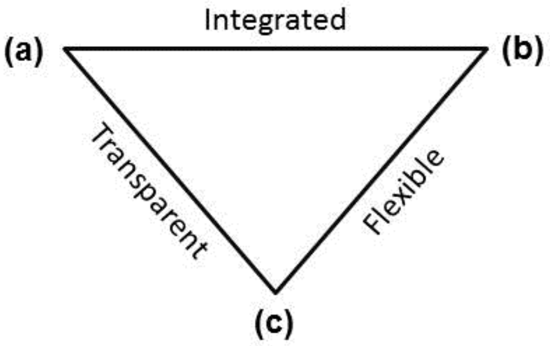


Figure 2: The triangulation logic

The first result of the above triangulation is the logic through which T-LAB *stores* any textual information. In fact, being a question-answer oriented system, all relationships between the user questions (i.e. queries) and the T-LAB answers depend – in the first instance - on how any text collection has been segmented and on how any textual unit (i.e. word, sentence, paragraph, document) has been indexed.

Now, in order to make clear the reasons why the storage logic is so important in text analysis, I would like to introduce a few simple and 'abstract' concepts that should be helpful. To start with, let's think:

a) *how*, for 'The Rocks Aroma Festival' in Sidney (Australia), in 2009 the Mona Lisa was *recreated* with 3,604 cups of coffee, each filled with varying amount of milk;

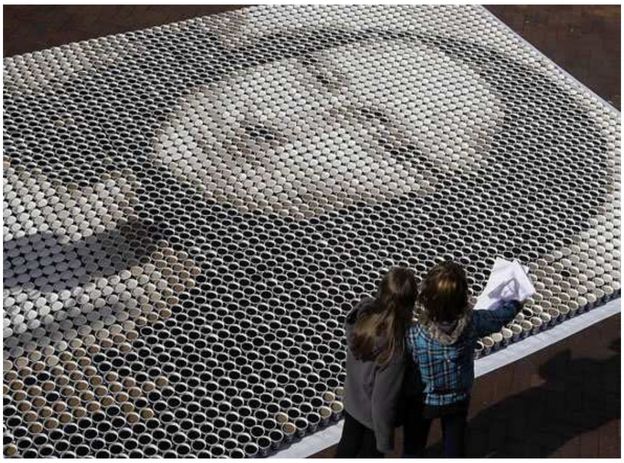


Figure 3: Mona Lisa made with cups of coffee

team and the *users* who, largely, are very skilled and competent researchers. For this very reason it is not easy to decide when a product like T-LAB has reached its 'maturity'.

b) *how* a simple ‘feature vector’[10] (‘x’) used in pattern recognition can be *conceived* as a piece of *Lego*.

$$x = \left[x_{1,}x_2 \dots x_n\right] \;\; ; \;\; x = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_m \end{bmatrix}$$

Figure 4: The building blocks

Actually, in order to perform any analysis, T-LAB represents and stores any text collection by means of various ‘pixel-grids’, i.e. by means of data matrices the rows and columns of which are ‘feature vectors’ which shape the *corpus* as it were ‘made with cups of coffee’. So, by using *Mona Lisa* (see Figure 3 above) as a metaphor of any text to be analysed, while performing its tasks T-LAB always allows the user to *retrieve* relevant textual units from the ‘original’ copy of the corpus and, at the same time, in order to *extract* relevant information, it allows the user to modify the resolution of any ‘pixel grid’ and to manage any feature vector like a piece of *Lego*. Out of the metaphor, the ‘pixel grid’ (i.e. any data matrix) refers to the reciprocal relationships between context units and lexical units (see section ‘3’ below) and the way they are managed has relevant effects on any statistical computations. However T-LAB also provides the user with a sort of ‘automatic pilot’ and – to reassure the reader - the performances of such a pilot are usually quite good.

Now try to think about any ‘textual unit’, i.e. any ‘analysis unit’ which is relevant to text analysis (e.g. word, concept, sentence, segment, paragraph, document, etc.), as if it were an abstract ‘x’ entity; then consider that the T-LAB architecture has been designed for enabling you to manage the following basic tasks:

a) Detect any ‘x’ in a standardised way;
b) Make – automatically - any ‘x’ a member of *equivalence classes*[11] labeled as you wish (e.g. a group of words belonging to the same ‘lemma’ or to the same ‘semantic class’, a group of sentences referring to the same ‘thematic cluster’, a group of documents referring to the same category of customers, etc.);
c) Represent any ‘x’ (or the equivalence class to which it belongs) as a *feature vector* the numerical values of which refer to phenomena like ‘presence/absence’ (e.g. presence/absence of the ‘x’ word within any sentence), ‘occurrence’ (e.g. how many times the ‘x’ word is repeated within the same document), ‘sequential order’ of words within a sentence and of sentences within a text;
d) Consider each ‘x’ feature vector as a row or a column of any *data table* (i.e. matrix);
e) Interpret any *query* of yours (i.e. your research question) as a task which either requires to explore the *relationships* between any pair of rows or columns, or to explore the multidimensional relationships within rectangular (i.e. ‘*n* x *m*’) or square (i.e. ‘*n* x *n*’) data tables;
f) Do any analysis (see ‘e’ above) by assuming your context of reference as the

[10] A feature vector ‘x’ can be represented as x = (x_1, x_2, …x_n), where each ‘x_i’ is a numerical feature of ‘x’.

[11] If *‘A’* is the set of all cars, and ‘R’ is the equivalence relation ‘has the same color as’, then one particular equivalence class consists of all green cars (see http://en.wikipedia.org/wiki/Equivalence_class).

whole *corpus* you have imported or as a *sub-set* of it. Moreover you are allowed to create and ‘extract’ any sub-corpus including all elementary contexts (i.e. sentences or paragraphs) which fit your query (i.e. a selection of relevant words);

g) Do any analysis (see ‘e/f’ above) either by making an automatic or a customised selection of *key words* to be used[12].

In other words– as stated in the introductory part of the user’s manual (Lancia, 2012a, p. 3) - by using T-LAB you are enabled to manage tasks like the following along with many others:

- measure, explore and map the *co-occurrence relationships* between key-terms;
- perform either unsupervised or supervised clustering of textual units, i.e. perform a *bottom-up clustering* which highlights *emerging themes* or perform *top-down classification* which uses a set of *predefined categories*;
- check the *lexical units* (i.e. words or lemmas), *context units* (i.e. sentences or paragraphs) and *themes* which are typical of specific text subsets (e.g. newspaper articles from specific time periods, interviews with people belonging to the same category);
- apply categories for *sentiment analysis*;
- perform various types of *correspondence analysis* and *cluster analysis*;
- create *semantic maps* that represent *dynamic* aspects of the discourse (i.e. sequential relationships between words or themes);
- customise and apply various types of *dictionaries* for both lexical and content analysis;
- perform *concordance* searches;
- create, explore and export numerous *contingency tables* and *co-occurrences matrices*.

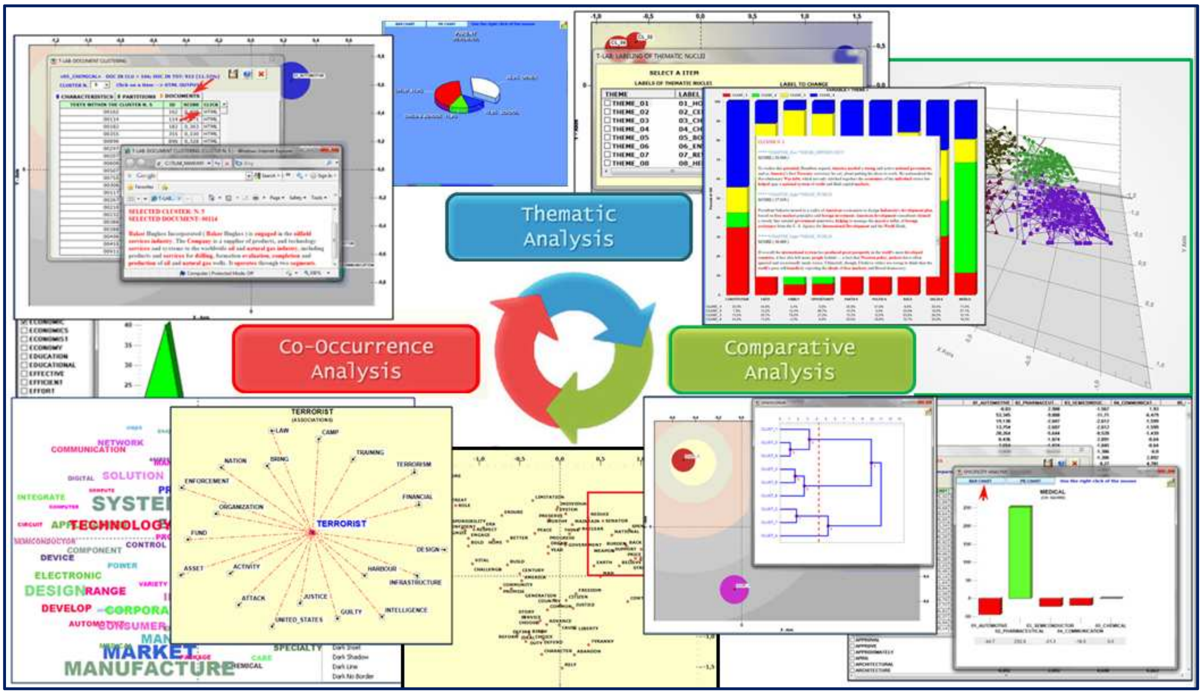


Figure 5: The T-LAB system

[12] The fact that T-LAB allows you to select the reference context (see ‘f’ above) and the key-word list (see ‘g’ above) implies that through this type of flexibility you can manage the dimensions of any table to be analysed (see ‘e’ above).

3 – Metaphors and pixel grids

As researchers well know, in text analysis, choosing not to choose is still a choice. So, given that in T-LAB everything depends on how the *textual units* are detected and stored, as well as on which tables are built and on how they are analysed, we should ask: is it that of 'text-scope' a good metaphor?
According to a colleague of mine (i.e. a psychologist) this is not the case, and I full agree with him. In fact the T-LAB tools do not 'mirror' phenomena, but rather 'construct' the phenomena in conjunction (or in spite of) the researcher 'point of view'. This is true for various reasons that I will explain below.

During the *preprocessing phase*, in order to allow *co-occurrence* analysis – at the moment (i.e. version 8.0) - T-LAB segments any text collection in *elementary contexts* which, depending to the user's choice, can be of four types: (a) sentences; (b) textual 'chunks' (i.e. textual segments) of comparable length made up of one or more sentences; (c) paragraphs; (d) short texts the length of which can be up to 2,000 characters (e.g. responses to open-ended questions, tweets etc.)[13]. Once again during the preprocessing phase T-LAB allows the user to work with or without *automatic lemmatization*, with or without *stop-word*[14] detection, with or without *multi-word*[15] detection. Moreover, even if the software provides automatic methods for building, importing and exporting lists of relevant words, any word list (including the stop-word and the multi-word lists) can be customised in various ways and the selection of *relevant words*[16], as well as the *equivalence classes* which each of them belongs to, can be reviewed at any stage of the analysis process. That means that, beyond the 'how to' logic, the user should be aware that any change in his word lists affects the way both *occurrence* and *co-occurrence* values are computed, whereas the way texts are segmented into elementary contexts affects the co-occurrence values only[17].

In any case, when the user starts any analysis, T-LAB builds some data tables first, and such tables contain *patterns* which simply need to be 'extracted' by statistical algorithms[18]. In other words, from the point of view of 'constructivist logic', *the way textual units became number is more relevant than the statistical algorithms to be applied*. And this is the very reason why, given the same table and the same standard measures, different algorithms produce very similar results. This means that, when using software like T-LAB, *'true' textual (or content) analysis deals with what precedes and follows any use of statistical algorithms*. So, as already recalled (Lancia, 2007), researchers must be aware of (a) what happens when words become numbers and (b) what happens when attempting to interpret the multi-semiotic texts (i.e. outputs like tables and graphs) produced by the software.

Actually, *no statistical algorithm is 'per se' an algorithm for text analysis*. In fact measures like association indexes (e.g. Cosine, Jaccard, Dice, etc.) and Chi-square test, as well as Markov chains, clustering methods and any multidimensional analysis (e.g. MDS, SVD, Correspondence Analysis, etc.) are not 'specific' to textual analysis. Biologists, ethologists, geologists and physicists use the

[13] When the corpus consists of short texts, the default option is 'd' (see above), otherwise it is 'b'. All the details are explained in the user's manual.

[14] Stop words are words considered irrelevant to the analysis. Typically a stop-word list includes prepositions, articles and other 'empty' words.

[15] A multi-word expression is made up of a sequence of two or more lexemes that stand for only one mining (e.g. 'Unites States', 'public transportation', etc.).

[16] When the corpus consists of two or more subsets, T-LAB allows the user to choose between two methods for selecting 'relevant' words: the one uses a chi-square measure, the other the TF-IDF.

[17] Text segmentation is also relevant for the 'Thematic Analysis' tools; in fact, in the first instance, they refer to co-occurrence tables.

[18] See sections 5-6-7 of this paper.

same algorithms too. More to the point: the very reason why *textual statistics really doesn't exist* is the same reason why 'school psychology' really doesn't exist. In other words, just as 'school' is not a psychological construct, equally 'text' is not a statistical construct. So, when doing text analysis, which phenomena are we 'really' studying?

At least in its standard use, the *rationale* which shapes the building blocks of the T-LAB tools refers to *linguistics*, and the linguistic nature of such tools doesn't concern processes like the automatic lemmatization and the dictionary customization only, rather it involves the definition of the analysis units as well as their reciprocal relationships. In detail, the pieces of *Lego* (see Figure 4 above) managed by the T-LAB tools and their users are of two types:

a- *Context Units* (CU), which are analysis units resulting from the corpus segmentation. So, 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, year of publication, topic, etc.), the single sentences into which every article can be split up (the elementary contexts), etc.;

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 (e.g. the coding schemes used in Content Analysis), or to 'labels' (or tags), each of which is indexed by its context of origin (i.e. CU).

This distinction, which has a theoretical foundation in linguistics and semiology, is of great practical importance; in fact, it allows all 'transformations' that are the basis of any statistical analysis.

With regards to the theoretical foundation, it goes back to the hypothesis initially proposed by F. de Saussure (1916), and subsequently by several authors (Jakobson, 1963; Barthes, 1964), according to whom the relationships between the linguistic elements can be analysed 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 deals with the LU 'selection' and determines the possibility of replacing any LU with one that has something in common with it (one 'in place of' the other: LU).

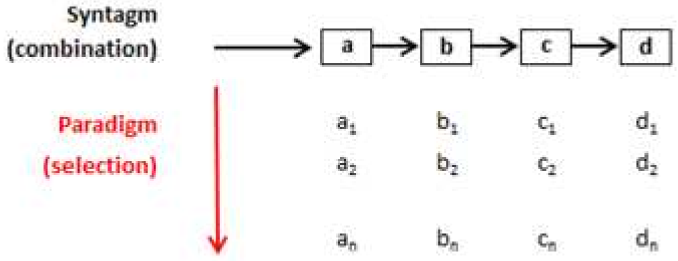


Figure 6: Syntagmatic and paradigmatic relationships

The practical relevance of the above distinction arises from the fact that the reciprocal relationships between CUs and LUs can be represented as vectors and matrices (i.e. tables) whose numerical values can indicate the instances of *occurrence* and *co-occurrence*, as well as the *sequential order* of the analysis units within texts.

In this paper's sections below several examples will make such 'logic' a bit more intuitive; however, at this moment, let's think in an abstract way and possibly realize that any CU can be represented by means of vectors (i.e. profiles) the features of which can be either CUs or LUs and that the same is true for any LU. For example, the features of any document (CU) can be either elementary contexts (CUs) or words (LUs), just as the features of any word can be either the CUs where it is present and the other words (LUs) which co-occur with it. In other words, as linguists say, all reciprocal relationships between such analysis units (i.e. CUs and LUs) can be regarded as *contiguity* and *similarity* phenomena.

Obviously, the software doesn't know the *meanings* (or contents), but only the signifiers[19], that is the 'strings' and the 'labels' that individualise the LUs and the CUs respectively; however, the relationships between signifiers (that is the syntagmatic relationships) assume meaningful shapes that, through co-occurrence patterns, propose a *contextual representation of the meaning*.

Consequently we could say that, in T-LAB's logic, the meaning of each single word is known only through its relationships with the contexts, viz. through the *distribution* of its occurrences (or co-occurrences) within the Context Units (CU). Equally we could say that such a semantic rationale refers to Greimas' notions of *contextual semes* and *isotopy* (iso=same; topos=place), even if the recognition of any isotopy is not simply the observation of the 'given' but the result of an interpretative process which requires abductive inference (Rastier, 1987, pp. 11-12; Lancia, 2007, p. 25)

Now, let's consider how various events can affect the above 'logic'.

Firstly, as a sort of mental experiment, let's imagine that the above CU and LU refer to any ecosystem whatsoever and to an animal or vegetal species respectively (i.e. CU = ecosystem and LU = species). So consider that a naturalist had decided to 'code' each vegetal species with a 'string' (i.e. LU) corresponding to a combination of alphabetic characters (e.g. 'oak tree' = 'ABRGG'; 'olive tree' = 'BCFQT', and so on) and had 'segmented' an ecosystem in CUs. Let's also imagine that this naturalist's 'description' had been transformed into a digital 'text' which can be imported through T-LAB. Does it make sense, for *his* science, to use the statistical tools for co-occurrence or occurrence analysis? The answer is obviously 'yes'. So, *what is a text and what is text analysis about*?

Now consider a more 'realistic' case. Due to the *flexibility* of the T-LAB system, any researcher can easily arrange any 'pixel grid' which maps the reciprocal relationships between LUs and CUs as h/she wishes, so that both the *dimensions* of any pixel (i.e. the values in any 'ij' cell) and the *area* under examination (i.e. the corpus subset) can vary. For example, let's imagine that both the grids below (see Figure 7 and Figure 8) represent the same corpus as a word-by-word matrix, and that the red squares delimit the areas under examination. Let's also imagine that the difference in the pixel dimension (and so in the pixel-grid 'resolution') is due to the fact that in the first case (i.e. Matrix 'A') the researcher has resorted to the uses of automatic lemmatisation only, whereas in the second case (i.e. Matrix 'B') h/she has applied a coding scheme which - according to his theories - grouped words into a few 'categories'. So, for example, in the first case ('A') a row of the matrix could be 'work' and it would include the occurrences of such a verb only (i.e. 'work', 'working', 'worked', 'works'), whereas in the second case ('B') all occurrences of 'work' (including its inflexions) could be recorded in a row of the matrix labeled as 'economy', which would group several lemmas and their corresponding inflexions (e.g. 'business', 'economy', 'money', 'worker', 'wage', and so on). The question is: can we reasonably affirm that, in the two different cases, the researcher is studying the same phenomenon and/or the same 'text'?

[19] 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).

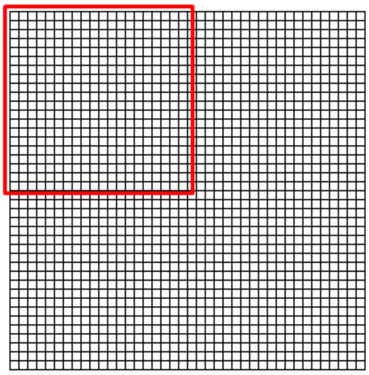


Figure 7: Matrix 'A'

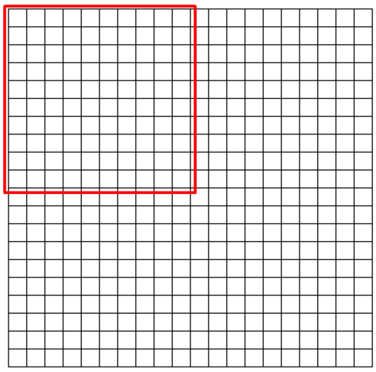


Figure 8: Matrix 'B'

Further question: when working within our discipline, *why* are we interested in analysing texts? For example, when a social psychologist who refers to the 'social representation' approach decides to analyse some interviews or some 'free associations' by mean of any T-LAB tool is h/she just interested in text analysis? If so, probably h/she is doing the wrong profession. The same would be true if a biologist were just interested in looking through a microscope.

4 - Starting from simple questions

Recently, while reading some published papers[20] which report the use of T-LAB, I 'discovered' that – from more than *fifteen* tools – the most used are just *four*[21]. The following table summarizes the results of my exploration. Actually in the same table the correspondence between authors and types of tools used is not accurate; in fact, following their research paths, several authors resorted to the use of tools listed in more than one column. So in such cases I added one asterisk to their reference in Table 1.

CO-OCCURRENCE ANALYSIS	COMPARATIVE ANALYSIS	THEMATIC ANALYSIS
1 - Word Associations	1 - Specificity Analysis 2 - Correspondence Analysis	1 -Thematic Analysis of Elementary Contexts
Capone & Petrillo (2011)* McNeely & Hopewell (2010) Perriton (2009)* Sengers et al. (2010)	De Rosa & Holman (2011) Grion & Varisco (2007) Greener (2009)* Margola et al. (2010)	Gambetti & Graffigna (2010)* Montali et al. (2011) Salvatore et al. (2010, 2012)* Veltri (2012)*

Table 1: The most used T-LAB tools

At this stage, I am not interested in commenting on the reasons why several T-LAB tools result under-used; rather I would like to outline the 'logic' of the four tools listed in the above Table 1. For this purpose, while making reference to the *questions/problems* the users are interested in, as a sort of 'ascetic' choice I will renounce including in the text any direct reference to the T-LAB outputs[22].

To start with, professor Ian Greener, who has been using T-LAB for many years, while examining the 'policy documents concerned in broad terms with the organization of health services in the UK'

[20] In this case I make reference to a few of English paper that I read personally. For more information about the T-LAB bibliography, see http://www.tlab.it/en/bibliography.php.
[21] In T-LAB sub-menus, each column of Table 1 includes five tools.
[22] The reader interested in screenshot gallery can explore the on-line help and other links available at http://www.tlab.it/.

(Greener, 2009), decided to use just two 'simple' tools: the one (a) which detects 'over- and under-used words' within corpus subsets, the other (b) which measures and maps the co-occurrence relationships between key words[23]. In doing so, he focused on the 'uses' of two key-words (i.e. 'choice' and 'responsiveness'[24]), then he has explored their different 'associations' with other words within subsets of documents belonging to four different time periods. So his findings are like 'Responsiveness in 1989 was about increasing patient choice through staff treating patients not as consumers of health care' (ib., p. 316), 'If patient choice was largely absent in 1977, the need for increased responsiveness to patients was very apparent' (ib., p. 317). 'Responsiveness was also very important in 2000… The idea of responding to the 'individual' patient is ubiquitous' (ib., p. 318).

Being rightly more interested in understanding the 'problems' of the National Health Service (i.e. NHS) over time than in describing the software tools he was using, the author didn't include in his paper any specific software output (i.e. tables and charts) obtained by the *Specificity Analysis* and *Word Associations* T-LAB tools. However in 'this' context it would be useful to recall the logic of two simple tools just recalled, which – like the *littleBits* modules - 'snap together' in professor Greener's method.

Let's start with the *Specificity Analysis*[25] tool, which allows us to check which lexical units (words, lemmas or categories) are *typical* or *exclusive* in a text or a corpus subset defined by any categorical *variable*, as well as to check the 'typical contexts' of each analysed subset (e.g. the 'typical' sentences used by any specific political leader). In detail:

- the 'typical' *lexical units*, defined for over-using or under-using, are detected by means of the *chi-square* or the *test value* computation;
- the 'typical' *elementary contexts* are detected by computing and summing the normalised *TF-IDF* values assigned to the words which each sentence or paragraph consists of. (Lancia, 2012a, p. 112)

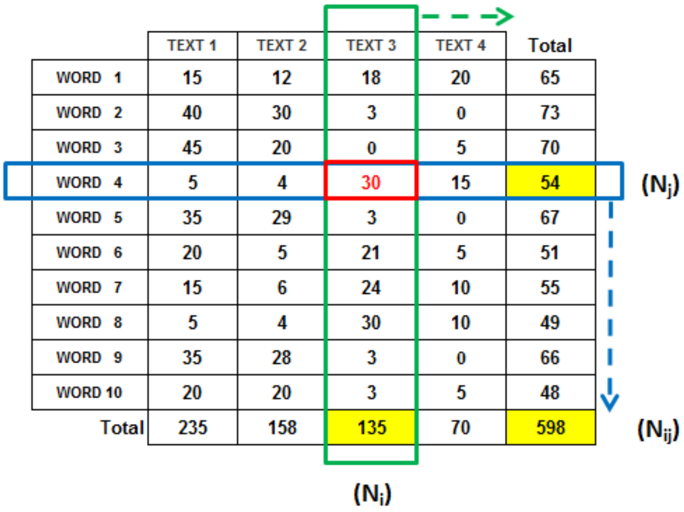
As, in such a case, the real tables analysed can include thousands of rows and hundreds of columns[26], in order to be didactic, I will refer to the example below (see Table 2), which includes just 10 'words' (i.e. LUs) and their **occurrences** within 4 'texts' (i.e. CUs). Now, let's imagine that when you ask for over-used and under-used words in any text, T-LAB builds a table like the following and moves two different rulers, the one from left to right (see the 'green' ruler below), the other from top to bottom (see the 'blue' ruler below); so that – for each step – just one cell comes into the focus (see the 'red' box below) and its 'statistical significance' can be measured. In order to achieve this, at the moment, T-LAB allows the user to choose between two different measures; however a brief explanation of how the chi-square test is applied should be sufficient.

[23] A similar two-step method has been used by L. Perriton (2009) while studying corporate discourses of gender.

[24] He has also included all inflexions and some synonyms of such words.

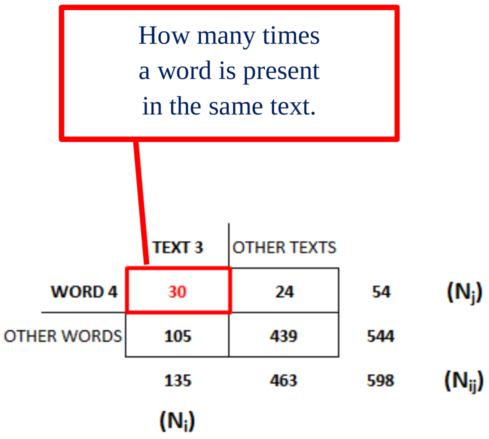
[25] Probably the name of such a tool is not intuitive for English people. In fact it derives from a literal translation of the French phrase 'analyse des spécificités'.

[26] Where columns are the categories of any variable.



	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 2: A contingency table



How many times a word is present in the same text.

	TEXT 3	OTHER TEXTS		
WORD 4	30	24	54	(N$_j$)
OTHER WORDS	105	439	544	
	135	463	598	(N$_{ij}$)
	(N$_i$)			

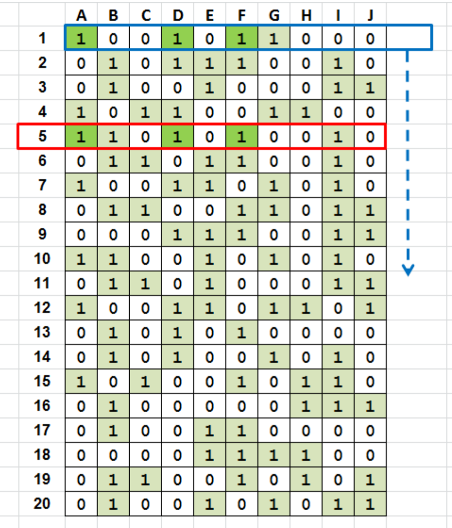
Table 3 – Relevant values for the Chi-square test

The above Table N. 3 shows which 'values' are manipulated at any stage by applying a simple formula that you can find in the Glossary section of the User's Manual (Lancia, 2012a).

Generally, T-LAB applies this test to '2 x 2' tables; then the threshold value is 3.84 (df = 1; p. 0.05) or 6.64 (df = 1; p. 0.01). So, following the above example (Table 3), the CHI value is equal to 36.94. And, since its value is greater than the critical value (i.e. 6.64; df = 1; p. 0.01), the null hypothesis (i.e. absence of meaningful difference) can be rejected. In other words 'word 4', when 'comparing' the texts under examination, results to be 'over-used' within 'text 3'.

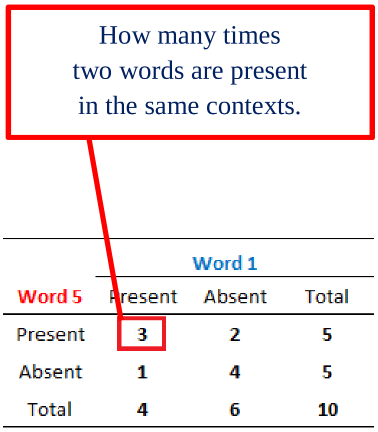
Actually in a similar way professor Greener (2009) picked up – within the corpus he was examining - 'over-used' words like 'choice' and 'responsiveness'.

Now let's explain something concerning the logic of the *Word Association* tool. Also in this case I will refer to an 'imaginary' table (see Table 4 below), the rows of which correspond to 'words' (i.e. 1, 2, 3 etc.) and the columns of which correspond to 'elementary contexts' (i.e. A, B, C, etc.) , while the cell values mark either the *presence* (i.e. '1') or the *absence* (i.e. '0') of any 'i' word within any 'j' elementary context. So, in this case, we are talking about **co-occurrences**.



	A	B	C	D	E	F	G	H	I	J
1	1	0	0	1	0	1	1	0	0	0
2	0	1	0	1	1	1	0	0	1	0
3	0	1	0	0	1	0	0	0	1	1
4	1	0	1	1	0	0	1	1	0	0
5	1	1	0	1	0	1	0	0	1	0
6	0	1	1	0	1	1	0	0	1	0
7	1	0	0	1	1	0	1	0	1	0
8	0	1	1	0	0	1	1	0	1	1
9	0	0	0	1	1	1	0	0	1	1
10	1	1	0	0	1	0	1	0	1	0
11	0	1	1	0	1	0	0	0	1	1
12	1	0	0	1	1	0	1	1	0	1
13	0	1	0	1	0	1	0	0	0	0
14	0	1	0	1	0	0	1	0	1	0
15	1	0	1	0	0	1	0	1	1	0
16	0	1	0	0	0	0	0	1	1	1
17	0	1	0	0	1	1	0	0	0	0
18	0	0	0	0	1	1	1	1	0	0
19	0	1	1	0	0	1	0	1	0	1
20	0	1	0	0	1	0	1	0	1	1

Table 4: A co-occurrence table[27]



How many times two words are present in the same contexts.

	Word 1		
Word 5	Present	Absent	Total
Present	3	2	5
Absent	1	4	5
Total	4	6	10

Table 5 – Relevant values for association measure

[27] In order to apply MDS (i.e. Multidimensional Scaling) and clustering algorithms the T-LAB tools also 'build' square co-

Now let’s imagine that any time the user selects a single word (e.g. the word corresponding to the feature vector ‘5’ in the above Table 4), T-LAB moves from the first to the last row of the table looking for ‘similarities’ (see the bleu ruler in Table 4). As explained in the corresponding section of the User’s Manual (see Glossary/Association Indexes), at the moment T-LAB allows the user to choose between three of the most popular similarity measures: Cosine, Jaccard and Dice. In any case, before applying any of the above measures – and for each word pair – T-LAB builds cross-tables like the above Table 5 . Subsequently, by using the corresponding formulas[28], the similarity between ‘word 5’ and ‘word 1’ can be expressed as follows: Jaccard = 0.50 ; Dice = 0,67; Cosine = 0.67. So, by using similar measures, professor Greener (2009, p. 312) made statements like the following: ‘The strongest co-association with *choice* (that is, the word most likely to appear with it) was *independence*’.

To sum up, the logic of the two simple tools just recalled deals with ***measures*** concerning similarities or differences between single ‘vectors’ (i.e. rows or columns) of matrices (i.e. data tables) the values of which correspond to word ***occurrences*** and word ***co-occurrences*** respectively. To be less ‘formal’ - and by considering each word as a ‘human’ individual - in the first case (i.e. occurrences) the *information* provided relates to events like how many times ‘John’ entered the same restaurant (i.e. same ‘place’); whereas in the second case (i.e. co-occurrences) the *information* relates to events like how many times ‘John’ met ‘George’, ‘Maria’ and ‘Rudolph’ (i.e. his ‘friends’).

Actually both the T-LAB tools referred to the above allow the users to obtain several outputs (i.e. customisable tables and charts). Among these I will just mention the possibility of extracting and visualising in HTML format all elementary contexts where two key-words co-occur (see *Word Associations*) and the ‘typical’ elementary contexts which characterise any corpus subset (see *Specificity Analysis*).

5 – Interlude concerning the geometric logic of matrices

Section above made reference to two ‘typical’ matrices which represent ‘events’ like word occurrences (see Table 2 above) and word co-occurrences (see Table 4 above). However, in order to better understand the logic of the T-LAB tools, I invite the reader to think in a more general way. More specifically, I would like to point out that such tools allow the user to build, to explore and to analyse, matrices which represent the entire *corpus* or any *subset* of it. So, if the corpus is ‘partitioned’ by means of categorical variables, any category can be used for building both occurrence and co-occurrence matrices. For example, if the documents under examination include tags referring to three categorical variables, the *corpus* can be represented by means of three different contingency tables which cross the ‘*n*’ words by the ‘*m*’ categories of each variable. Moreover, by ‘extracting’ a subset of documents belonging to a variable category (e.g. sex→female) it is possible to build contingency tables which cross words and variables categories within the chosen *corpus subset* (e.g. female). Equally, given that any subset includes a number of ‘elementary contexts’, it is possible to build and to analyse the corresponding co-occurrence tables.

Actually the reasons why the T-LAB tools are grouped into three sections are related to the ‘type’ of tables which – in first instance – they refer to. To be more specific:

- the *co-occurrence* analysis tools deal with matrices like the one in Table 4 above and its various transformations (e.g. matrices word-by-word);

occurrence tables word-by-word.

[28] See the T-LAB user manual (Lancia, 2012a).

- the *comparative* analysis tools deal with contingency tables (see Table 2 above) the columns of which result from an 'a priori' or 'a posteriori' partitioning, where the former refer to categorical variables used to tag the corpus before its importation and the latter refer to categorical variables obtained by thematic analysis (e.g. thematic clusters);
- the *thematic* analysis tools deal with both the above types; in fact a sort of co-occurrence analysis is performed for highlighting patterns and, when required, the same patterns can be transformed into categorical variables.

At this stage I would like to give you an idea of the geometric transformations allowed by the T-LAB tools and here I accept that I run the *risk* of being criticised for lack of 'precision'. So, the questionable picture that I have arranged (see Figure 9 below) aims to communicate the following ideas:

- when doing 'occurrence' analysis each word (i.e. each row of the corresponding matrix) is a feature of each corpus subset;
- the relation between 'occurrence' and 'co-occurrence' analysis can be dynamic;
- 'patterns' (e.g. thematic clusters) can be transformed into variable categories (i.e. corpus subsets).

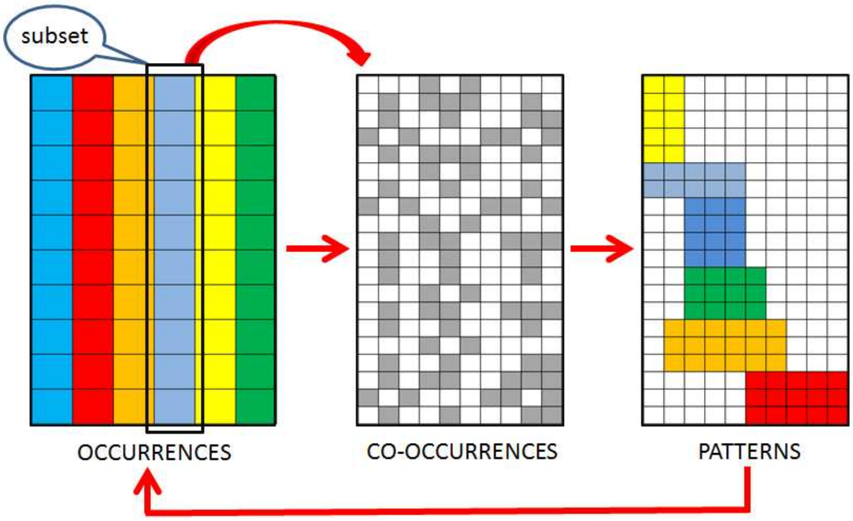


Figure 9: Geometric transformations of matrices

6 – From measures to patterns

To introduce the shift form *measures* (see the above section 4) to *patterns*, I make reference in the first instance to a paper (Gambetti & Graffigna, 2010) which reports on an 'exploratory and systematic content analysis' the aims of which were two:

a- to explore the main dimensions associated with '*engagement*' in marketing and communication literature (i.e. 'How is engagement conceived? Which variables and issues are related to it? Are there different types of engagement or is it a unified concept?');

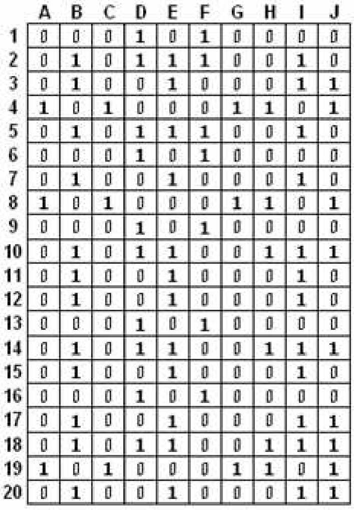
b- to identify similarities and differences in the vocabulary most often associated with 'engagement' in academic peer-reviewed journals and professional journals.

The path followed by the authors is quite interesting and - as luck would have it – by analysing journal titles and journal abstracts, they used only the four T-LAB tools listed in the above Table 1. To be more specific, they carried out:

- a *Thematic Analysis of Elementary Contexts* of the entire corpus (i.e. all titles and abstracts included in the research) as a result of which four 'conceptual clusters' were identified and interpreted;
- a *Specificity Analysis* of two sub-corpora (i.e. academic sub-corpus vs. professional sub-corpus) showing the corresponding over-used and under-used words;
- a *Word Association Analysis* of the above sub-corpora by focusing on key-words linked to 'engagement';
- a *Correspondence Analysis* of all five corpus sub-sets codified by the authors to map their 'thematic' similarities and differences.

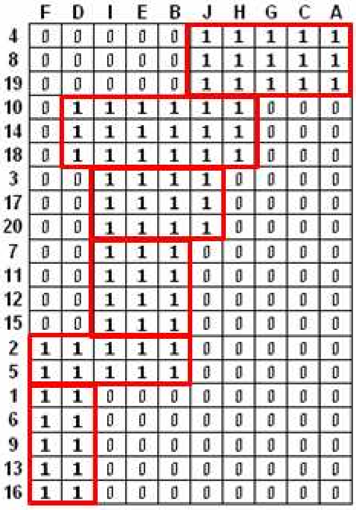
Having explained in the above section the logic of two tools dealing with 'simple' measures, here and in the section below I will concentrate on the logic of two complementary algorithms which look for *patterns*, viz. the 'Simple' (or binary) *Correspondence Analysis* (CA) and the specific kind of *Cluster Analysis* (i.e. the *bisecting K-means* algorithm, hereinafter BKM) which constitutes the core process implemented in the T-LAB tool named *Thematic Analysis of Elementary Contexts*.

To start with I would like to point out that the 'transformation' of Table 6 below into Table 7 can be obtained either by CA[29] or by Cluster Analysis (including BKM[30]). However, as we will see, in the first case the algorithm (i.e. CA) has the task of extracting new variables (i.e. the factors) which work like 'classification principles' (Burt, 1940), whereas in the second case (e.g. BKM) the algorithm has the task of assigning 'objects' to groups (i.e. clusters). So, by considering that – in T-LAB - Table 6 can be either a document-by-word matrix or context-by-word matrix, in both cases (i.e. CA and BKM) the result (i.e. the partitioning into ***patterns*** and/or ***clusters***) is quite important.



	A	B	C	D	E	F	G	H	I	J
1	0	0	0	1	0	1	0	0	0	0
2	0	1	0	1	1	1	0	0	1	0
3	0	1	0	0	1	0	0	0	1	1
4	1	0	1	0	0	0	1	1	0	1
5	0	1	0	1	1	1	0	0	1	0
6	0	0	0	1	0	1	0	0	0	0
7	0	1	0	0	1	0	0	0	1	0
8	1	0	1	0	0	0	1	1	0	1
9	0	0	0	1	0	1	0	0	0	0
10	0	1	0	1	1	0	0	1	1	1
11	0	1	0	0	1	0	0	0	1	0
12	0	1	0	0	1	0	0	0	1	0
13	0	0	0	1	0	1	0	0	0	0
14	0	1	0	1	1	0	0	1	1	1
15	0	1	0	0	1	0	0	0	1	0
16	0	0	0	1	0	1	0	0	0	0
17	0	1	0	0	1	0	0	0	1	1
18	0	1	0	1	1	0	0	1	1	1
19	1	0	1	0	0	0	1	1	0	1
20	0	1	0	0	1	0	0	0	1	1

Table 6: A co-occurrrence table



	F	D	I	E	B	J	H	G	C	A
4	0	0	0	0	0	1	1	1	1	1
8	0	0	0	0	0	1	1	1	1	1
19	0	0	0	0	0	1	1	1	1	1
10	0	1	1	1	1	1	1	0	0	0
14	0	1	1	1	1	1	1	0	0	0
18	0	1	1	1	1	1	1	0	0	0
3	0	0	1	1	1	1	0	0	0	0
17	0	0	1	1	1	1	0	0	0	0
20	0	0	1	1	1	1	0	0	0	0
7	0	0	1	1	1	0	0	0	0	0
11	0	0	1	1	1	0	0	0	0	0
12	0	0	1	1	1	0	0	0	0	0
15	0	0	1	1	1	0	0	0	0	0
2	1	1	1	1	1	0	0	0	0	0
5	1	1	1	1	1	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0
6	1	1	0	0	0	0	0	0	0	0
9	1	1	0	0	0	0	0	0	0	0
13	1	1	0	0	0	0	0	0	0	0
16	1	1	0	0	0	0	0	0	0	0

Table 7: Table 6 reorderd

Actually, in T-LAB, *Correspondence Analysis* (CA) is both the name of a specific tool[31] and a visualization technique implemented in several procedures, including that of the thematic tool mentioned above. In text analysis, CA has become popular thanks to the researches carried out by J.P.

[29] More specifically, by using the object coordinates on the first factorial axis (see also Table 9 below).

[30] The core process of CA is the Singular Value Decomposition (i.e. SVD) and, as S.M. Savaresi and D.L. Booley (2001) have shown, when 'partitioning' a large sparse matrix, methods which use SVD (e.g. PDDP) and the BKM obtain very similar results.

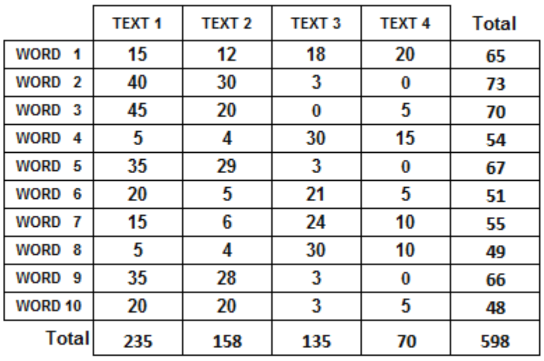
[31] T-LAB also includes a 'Multiple Correspondence Analysis' tool; but this one analyses different tables and its algorithm is quite different from the 'simple' CA.

Benzecri (1984) and his school, which have proven that such a technique is very useful for mapping the interrelationships between words and, mutually, between words and categorical variables present in any text collection. Its elective use is related to ***occurrence*** analysis[32] and ‘partitioned’ data (see variable categories); so – when using T-LAB - it is often ‘combined’ with the *Specificity Analysis* tool (see, for example: Grion & Varisco, 2007; Margola et al., 2010). However, as the ‘thematic clusters’ produced by other T-LAB tools are for all practical purposes ‘categories’ which group textual units, their reciprocal relationships (as well their relationships to relevant words), being represented in contingency tables, can be explored by using the same technique (i.e. CA) too. In fact this is what was done by the researchers mentioned above (Gambetti & Graffigna, 2010).

To put it simply, CA is a multidimensional technique that allows us to represent the relationships between *all* the ‘feature vectors’ – i.e. all rows and columns - of any matrix with frequency or presence/absence values. In order to describe how it works, at least in the first instance, we can put aside mathematical and statistical notions[33]. Moreover as the data matrices analysed by this T-LAB tool can include several thousand rows by several hundred columns, let’s focus on the same table used for illustrating the ‘logic’ of *Specificity Analysis* tool (see Table 2 above and Table 8 below). In fact, when comparing *occurrences*, both tools refer to the same tables and the user is allowed to display and export them in various ways.

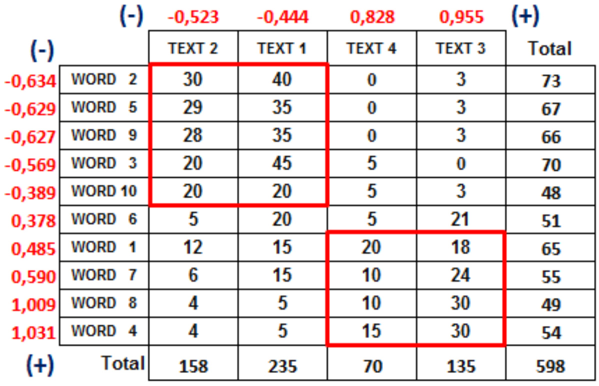
Here, it is sufficient to recall that the operations implemented in the CA algorithm allow us to obtain two kinds of results:

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 (i.e. *patterns*) they contain is then made ‘readable’. See, for example table 8 below, which ‘transformation’ has been obtained by using just the ‘coordinates’ of rows and columns on the CA first factor ‘extract’, which have been properly ordered (see decimals in red in Table 9).



	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 8: A contingency table



	(-)	-0,523	-0,444	0,828	0,955	(+)
(-)		TEXT 2	TEXT 1	TEXT 4	TEXT 3	Total
-0,634	WORD 2	30	40	0	3	73
-0,629	WORD 5	29	35	0	3	67
-0,627	WORD 9	28	35	0	3	66
-0,569	WORD 3	20	45	5	0	70
-0,389	WORD 10	20	20	5	3	48
0,378	WORD 6	5	20	5	21	51
0,485	WORD 1	12	15	20	18	65
0,590	WORD 7	6	15	10	24	55
1,009	WORD 8	4	5	10	30	49
1,031	WORD 4	4	5	15	30	54
(+)	Total	158	235	70	135	598

Table 9: Table 8 reorderd

[32] Some T-LAB tools allow the user to map – by Correspondence Analysis – also tables including ‘co-occurrence’ values (e.g. tables the rows and columns which are ‘elementary contexts’ and words respectively).

[33] Basically CA requires that the chi-square distance be used for measuring similarities between all rows and all columns of any ‘A’ contingence table and that to a square matrix, obtained through an appropriate transformation of ‘A’, be applied a *Singular Value Decomposition* (SVD). The description of its algorithms would require several pages and would necessarily include many formulas; therefore I would prefer not to spend time on these concerns and, to the interested reader, I suggest the following bibliographical references: J.P. Benzecri (1984), M. J. Greenacre (1984) and L. Lebart, A. Morineau M. Piron (1995).

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 plotted within a reduced space defined by the few factors that, in a statistically significant way, explain their variability[34]. So the same data in Table 9 (see above) can be represented by means of a classic two-dimensional chart (Figure 10) that, as can be seen, is coherent with the representation of the profiles (Table 9). In both cases the first factor turns out characterised by the prevailing 'weight' of two CUs (Text_2, Text_1) and of two LUs (word_2, word_5) on the negative (-) pole, while on the positive pole (+) two other couples of CUs (Text_4, Text_3) and of LUs (word_4, word_8) prevail.

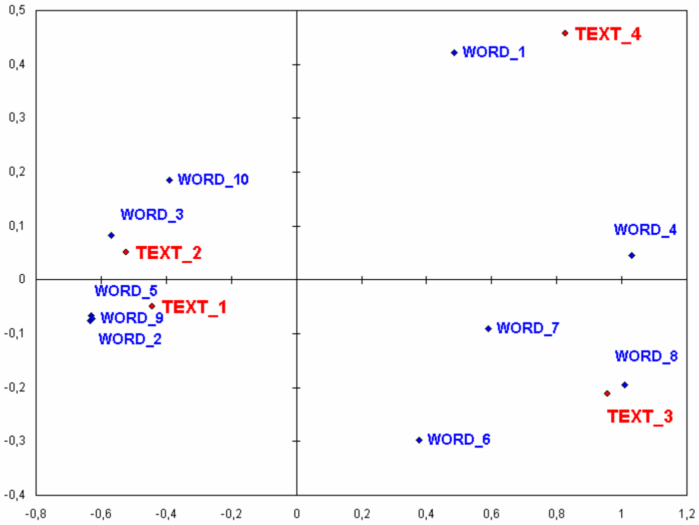


Figure 10: Table 8 displayed by CA (N.B.: This is not a T-LAB output)

When interpreting the results of a factorial analysis, we often risk getting lost in a maze of tables and charts[35]. In order to avoid this risk, or at least to reduce it, reference to some definitions can be useful. In fact, the factors can be considered as *classification principles* (Burt, 1940) - i.e. as organizers of the relationships between the data - that put similar things together, distinguish them from different things and 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 to the right of the origin (barycentre), and secondly all that is to the left of it, and then expressing concisely and exactly the opposition between the two extremes*" (1984, p. 302. My translation).

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

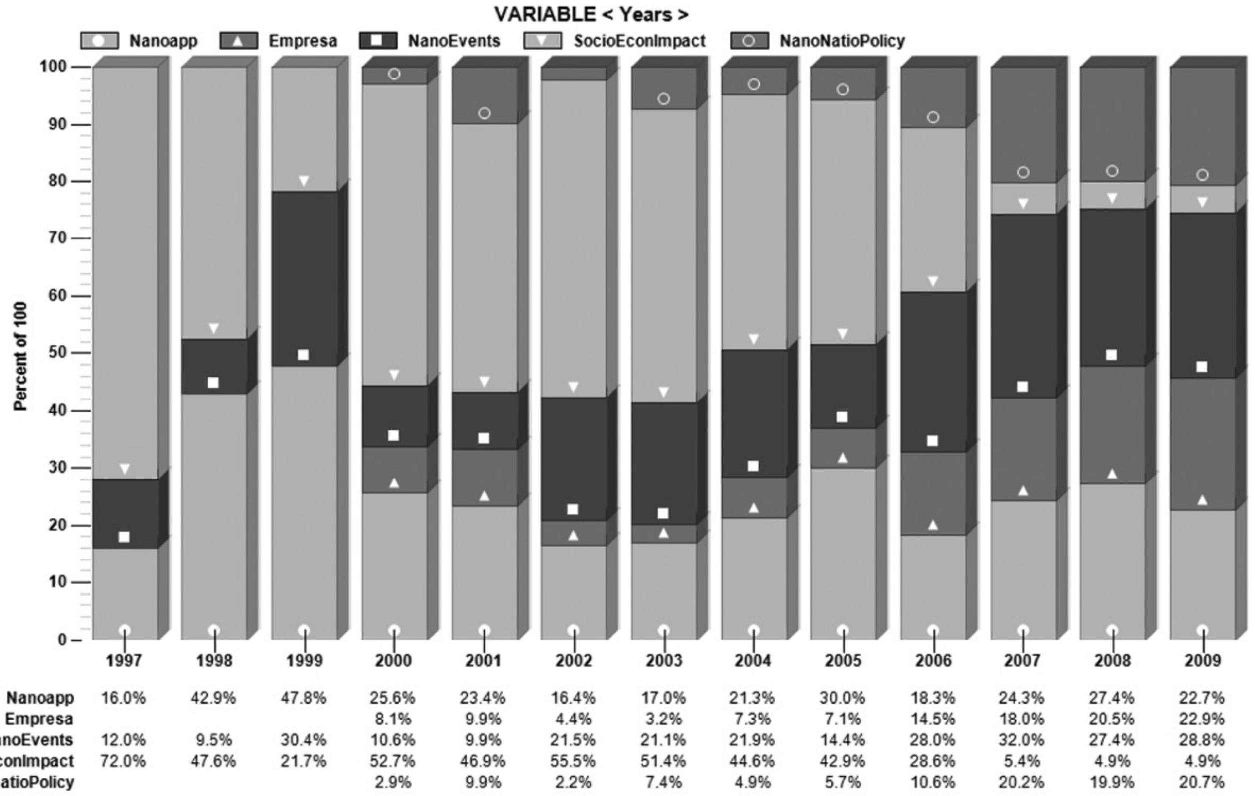
[34] By definition, the factors 'extracted' are n-1, where 'n' is the number of columns in the table.

[35] In order to interpret the CA results, various measures are used, either concerning the 'weight' of each factor (e.g. Eigenvalues and Inertia), or reporting the coordinates and the contributions (absolute and relative) of each object (row or column) by each factorial axis, just as their corresponding Test Values. All these measures are provided by T-LAB in an interactive way.

In effect, even if the word ‘factor’ suggests a sort of causal relationship between data, the factorial analyses only serve to find an order (i.e. patterns) in the complexity of the data analysed, 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 scientific discipline is 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 for existing.

7 – Patterns and ‘themes’

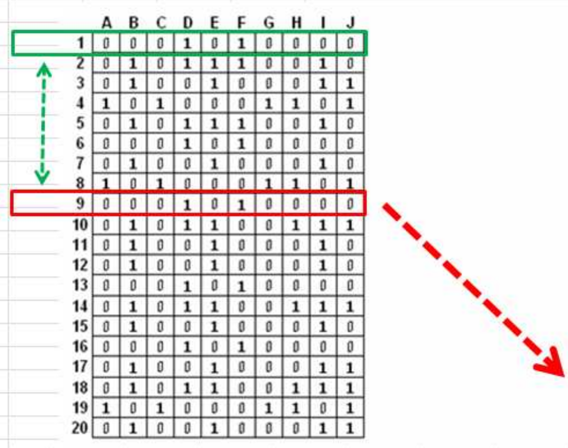
In keeping with my intention to connect the logic of any T-LAB tool with its uses, here I refer to a paper of G. A. Veltri (2012), the title which is ‘*Viva la Nano-Revolution! A semantic Analysis of the Spanish National Press*’. In such a paper the author, by making reference to ‘social representation’ studies, addresses three ‘research questions’, the third of which is the following: ‘Which themes are present in the representation of nanotechnology in the Spanish national press? And complementary to the previous question: Does the Spanish press display an initial emphasis on economic potential followed by an increasing salience of risks?’ (ib., p.5). The author clearly explains ‘why’ and ‘how’ he used the T-LAB tool named *Thematic Analysis of Elementary Contexts*. The analysed corpus was made up of 646 articles published – between January 1997 and August 2010 – from three main Spanish newspapers of different political orientations: *El Pais*, *El Mundo* and *ABC*. The thematic clusters that – in the first instance – professor Veltri focuses on are five and their changes over the period of time are illustrated by the following customised diagram:



	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Nanoapp	16.0%	42.9%	47.8%	25.6%	23.4%	16.4%	17.0%	21.3%	30.0%	18.3%	24.3%	27.4%	22.7%
Empresa				8.1%	9.9%	4.4%	3.2%	7.3%	7.1%	14.5%	18.0%	20.5%	22.9%
NanoEvents	12.0%	9.5%	30.4%	10.6%	9.9%	21.5%	21.1%	21.9%	14.4%	28.0%	32.0%	27.4%	28.8%
SocioEconImpact	72.0%	47.6%	21.7%	52.7%	46.9%	55.5%	51.4%	44.6%	42.9%	28.6%	5.4%	4.9%	4.9%
NanoNatioPolicy				2.9%	9.9%	2.2%	7.4%	4.9%	5.7%	10.6%	20.2%	19.9%	20.7%

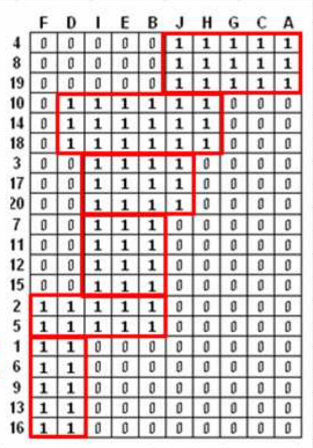
Figure 11: A T-LAB customised output (Veltri, 2012, p. 15)
(Relative weight of thematic clusters in the corpus across years)

The reader interested in better understanding how this tool for thematic analysis can be used can read both the Veltri article and the corresponding section of the T-LAB manual (Lancia, 2012a) . Now, without entering into technical details, I would like to explain the ‘simple’ logic of the Bi-secting K-means algorithm (i.e. BKM), which - among other things – ends up being a sort of mix of the classic ‘partitioning’ and ‘hierarchical’ methods. To start with, let’s think that it ‘looks for similarities’ like the *Word Association* tool (see Figure 12 below, where the table crosses elementary contexts by words and the ‘feature vector’ in red stands for a cluster ‘centroid’ whatsoever) and, like the *Correspondence Analysis* tool, it has the task of ‘finding patterns’ (see Figure 13 below).



	A	B	C	D	E	F	G	H	I	J
1	0	0	0	1	0	1	0	0	0	0
2	0	1	0	1	1	1	0	0	1	0
3	0	1	0	0	1	0	0	0	1	1
4	1	0	1	0	0	0	1	1	0	1
5	0	1	0	1	1	1	0	0	1	0
6	0	0	0	1	0	1	0	0	0	0
7	0	1	0	0	1	0	0	0	1	0
8	1	0	1	0	0	0	1	1	0	1
9	0	0	0	1	0	1	0	0	0	0
10	0	1	0	1	1	0	0	1	1	1
11	0	1	0	0	1	0	0	0	1	0
12	0	1	0	0	1	0	0	0	1	0
13	0	0	0	1	0	1	0	0	0	0
14	0	1	0	1	1	0	0	1	1	1
15	0	1	0	0	1	0	0	0	1	0
16	0	0	0	1	0	1	0	0	0	0
17	0	1	0	0	1	0	0	0	1	1
18	0	1	0	1	1	0	0	1	1	1
19	1	0	1	0	0	0	1	1	0	1
20	0	1	0	0	1	0	0	0	1	1

Figure 12: It looks for similarities



	F	D	I	E	B	J	H	G	C	A
4	0	0	0	0	0	1	1	1	1	1
8	0	0	0	0	0	1	1	1	1	1
19	0	0	0	0	0	1	1	1	1	1
10	0	1	1	1	1	1	1	0	0	0
14	0	1	1	1	1	1	1	0	0	0
18	0	1	1	1	1	1	1	0	0	0
3	0	0	1	1	1	1	0	0	0	0
17	0	0	1	1	1	1	0	0	0	0
20	0	0	1	1	1	1	0	0	0	0
7	0	0	1	1	1	0	0	0	0	0
11	0	0	1	1	1	0	0	0	0	0
12	0	0	1	1	1	0	0	0	0	0
15	0	0	1	1	1	0	0	0	0	0
2	1	1	1	1	1	0	0	0	0	0
5	1	1	1	1	1	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0
6	1	1	0	0	0	0	0	0	0	0
9	1	1	0	0	0	0	0	0	0	0
13	1	1	0	0	0	0	0	0	0	0
16	1	1	0	0	0	0	0	0	0	0

Figure 13: It finds patterns

In detail the BKM algorithm starts with a single cluster of *all* ‘objects’ (e.g. ‘feature vectors’ encoding textual units) and it works in the following manner (Steinbach, Karypis, & Kumar, 2000):

1 – Pick a cluster to split;

2 – Find 2 sub-clusters using the basic K-means algorithm;

3 – *Repeat* step 2, the bisecting step, for a fixed number of times and take the split that produces the cluster with the highest overall similarity;

4 – *Repeat* steps 1, 2 and 3 until the desired number of clusters is reached[36].

The following picture illustrates the ‘logic’ of the above steps.

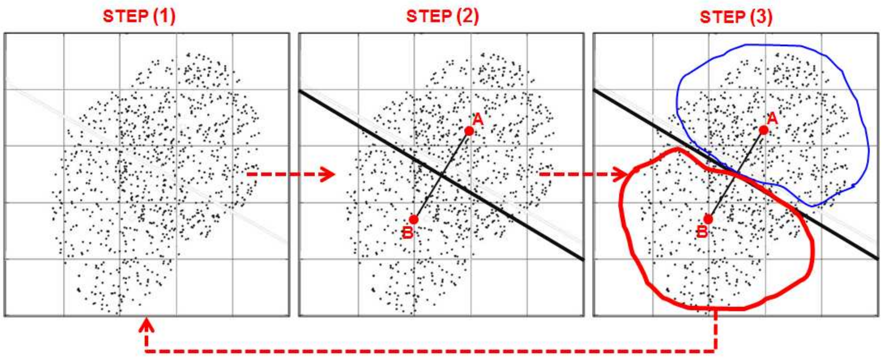


Figure 14: How the bisecting K-means works

[36] Actually the classical implementation of the BKM requires that the ‘desired number of clusters’ is fixed in advance; whereas T-LAB stores a number of cluster partitions, selects the possible ‘best’ solution by using the 'intracluster correlation coefficient' (i.e. between cluster variance / total variance, where ‘total variance’ = between cluster + within cluster variance) and enables the user to quickly explore other solutions.

In technical language, all depends on how the 'A' and 'B' points (i.e. the centroids) have been picked up[37], on how they are moved until 'convergence', and on how the next cluster to split is selected[38]. As a matter of fact, by proceeding with consecutive bisections and by allowing the storage of the various cluster partitions, such an algorithm can also be regarded as the implementation of a *hierarchical divisive* (or *descending*) clustering (see Figure 15 below).

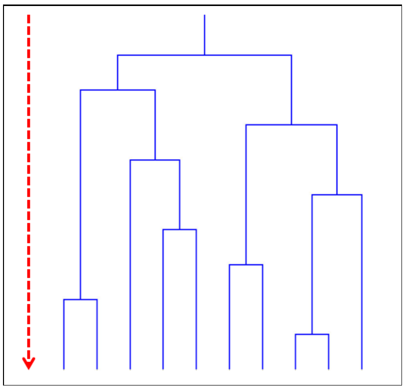


Figure 15: Dendrogram

In T-LAB the BKM algorithm is integrated within two *twin* tools (i.e. *Thematic Analysis of Elementary Contexts* and *Thematic Document Classification*[39]) which allow the user to explore the relationships between 'thematic clusters', words, context units and categorical variables in a variety of ways. In fact, each selected partition is a 'new' categorical variable which can be used for building word-by-category tables; and the relationships between rows (i.e. words) and columns (i.e. 'thematic clusters') can be measured and explored by means of methods which use the same 'logic' of two tools described above. In fact, the 'characteristics' of each cluster are obtained by the Chi-square test (see the 'Specificity Analysis' tool) and the scatter charts are obtained by the 'Simple Correspondence Analysis' tool. Moreover, when saving a partition, the corresponding variable categories can be used in further T-LAB analyses (e.g. in *Multiple Correspondence Analysis, Word Associations,* etc.).

Now, let's come back to what I have argued in section '1' about the limits of any 'unsupervised' (i.e. bottom-up) approach to thematic analysis. To this purpose, firstly and above all, I would like to remind the reader that when thinking that 'themes' are just the result of a software procedure we – certainly – are not on the right path. In fact – in text analysis – 'themes' are always the result of an interpretation process and, for this very reason, any thematic analysis relies on the researcher competencies. However, sometimes the 'architecture' of the software system can help, that is it can make the most of the user's 'thinking'.

Actually - starting from the 8.0 release (may 2012) - the T-LAB tool we are talking about, is surprisingly *flexible*; in fact, in addition to allowing the user to browse and check any cluster partition (see the previous version of the software), now it allows the user to 'refine' the chosen partition in two different ways[40], and it allows the user to import/export any thematic 'dictionary', which - being an array of 'semes' – enables a combination of the 'bottom-up' and 'top-down' approaches (see the

[37] The T-LAB algorithm follows the method outlined by S.M. Savaresi and D.L. Booley (2001).

[38] Actually several 'criterion functions' can be used that measure various aspects of intra-cluster similarity and inter-cluster dissimilarity (Zhao & Karypis, 2004).

[39] In the current literature several studies can be found which analyse the performances of the BKM either in clustering 'documents' (e.g. Krishna, Satheesh, Suneel Kumar, 2012) or in clustering 'text segments' (e.g. Tagarelli, Karypis, 2008).

[40] For more explanation, see the corresponding section of the T-LAB User Manual (Lancia, 2012a).

below Figure 16)[41].

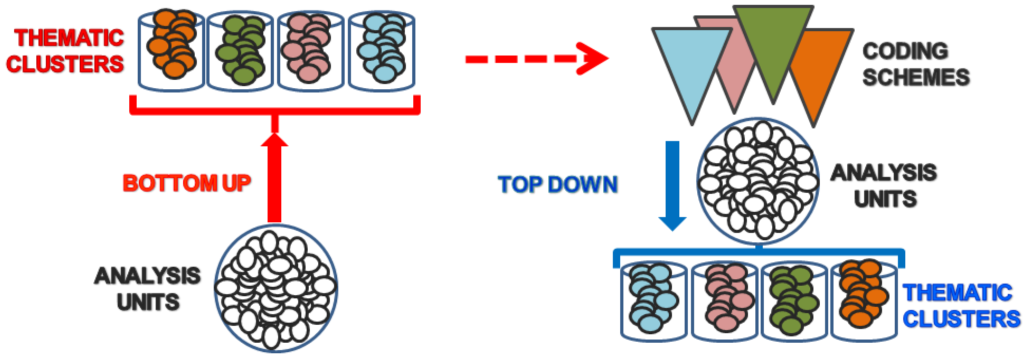


Figure 16: Bottom-up and Top-down approaches

Moreover advanced users who like to 'construct' themes in a variety of ways can experience how the *Dictionary-Based Classification* tool allows for the easy application of previously developed 'coding schemes' to obtain top-down classifications of any sort of textual units (i.e. words, elementary contexts, documents). Last and not least the *Modeling of Emerging Themes* tool, which works through a 'topic model approach' and a very sophisticated algorithm (i.e. a combination of *Latent Dirichlet Allocation*[42] and *Gibbs Sampling*), allows the user to decide both the number of themes/topics and the 'features' (i.e. words) which characterise each of them; moreover, after having been tested, any 'constructed' model can be applied to further analyses.

8 - Exploring new perspectives in text analysis

In my opinion, 'discovering' requires three attitudes: (true) *astonishment*, (true) *curiosity* and (true) *imagination*. As the latter (i.e. imagination) deals mostly with metaphors and with having 'an eye for resemblances'[43], here I will use some metaphors to defy the widespread opinion according to which – firstly and above all - 'text mining'[44] software like T-LAB should be used for analysing 'huge' quantities of text documents.

To start with, let's 'observe' T-LAB as a software system for 'text playing', where the 'resemblances' concern human activities like 'playing the piano' and 'playing with 'wooden blocks'(see Figure 17 below). Actually both T-LAB tools and the texts to be analysed are strange 'objects' to play with; in fact such 'objects' allow researchers to produce new knowledge, as well as to write their own texts (i.e. reports, journal articles, books, etc.) and – last but not least - to *make money* by helping individuals and organizations to manage their own problems better[45]. In doing so, there are people who like to go along the same path and people who try new ways for text analysis, or – rather - people who try to answer their new 'questions' by means of text analysis.

[41] The reader interested in such issues can download a working paper (Lancia, 2012b) which is available on the T-LAB website.

[42] See D. Blei, A.Y. Ng & M.I. Jordan (2003).

[43] 'For the right use of metaphor means an eye for resemblances' (Aristotle, *Poetics*, 1459a, 8).

[44] Actually the 'text mining' phrase is a metaphor too.

[45] For example, 'text mining' is a quite good business in marketing research.

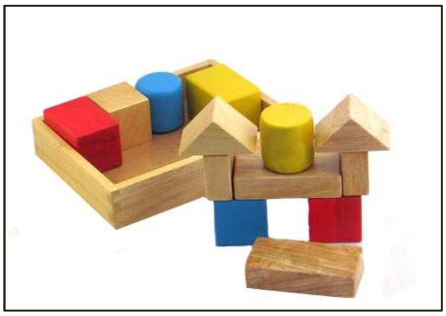


Figure 17: Children's wooden blocks[46]

In the above pages I have tried to explain the 'logic' of a few T-LAB tools, actually the most quoted in the current literature. However, there are several other tools in the software system and lots of possible combinations between them can be easily experienced[47]. In fact, when used by skilled researchers, their combinations can turn out to be 'mixed methods' and 'mixed strategies'.

Always looking for 'resemblances' - and hoping that the reader has the gift of self-irony – I would like to remind that a good meal isn't about having a lot of food. So, once you, the user, have selected high quality ingredients (i.e. the judiciously chosen texts for analysis) then T-LAB allows you to become the chef and actually put attractive food on the table. Just choose the more appropriate tools for *your* cooking.

Therefore, also to promote a sort of 'slow food' in text analysis, in this concluding section I would like to dedicate a few words to the illustration of a couple of new 'recipes' which – actually – have been created by Italian 'chefs', i.e. by Italian groups of research which – in my opinion – are making an interesting and *dynamic* use of the T-LAB tools. The first 'recipe' can be found in a book chapter (Trobia, Frazzica & Milia, 2012) which reports the application of a new methodological *mix* designed for analysing various kinds of data concerning *focus groups*, in particular: the 'dynamic' relationships between participants (e.g. sociomatrices and sociograms), their verbatim transcripts and the categorical variables which give information about their social status (e.g. sex, profession, etc.). Among other things, by studying three focus groups, the authors have assembled a corpus including all the 'textual' and 'contextual' data just quoted and – by using T-LAB – they have performed two kind of analysis:

- a *Multiple Correspondence Analysis* the 'active' variables of which include also *codes* concerning the network analysis (i.e. ego density, cliques, in-degree and out-degree, etc.);
- a *Thematic Analysis of Elementary Contexts* oriented to exploring the relationships between 'themes' and focus groups 'dynamics' (e.g. their 'phases', just as the in-degree and out-degree of participants).

The following pictures illustrate some findings of the above analyses:

[46] This picture has been downloaded from http://www.ecotoys.com.au .

[47] I have to say that - as far I know - T-LAB is not 'too demanding' as a software system; in fact lots of young researchers, who actually are not very skilled in statistics, 'enjoy' using it.

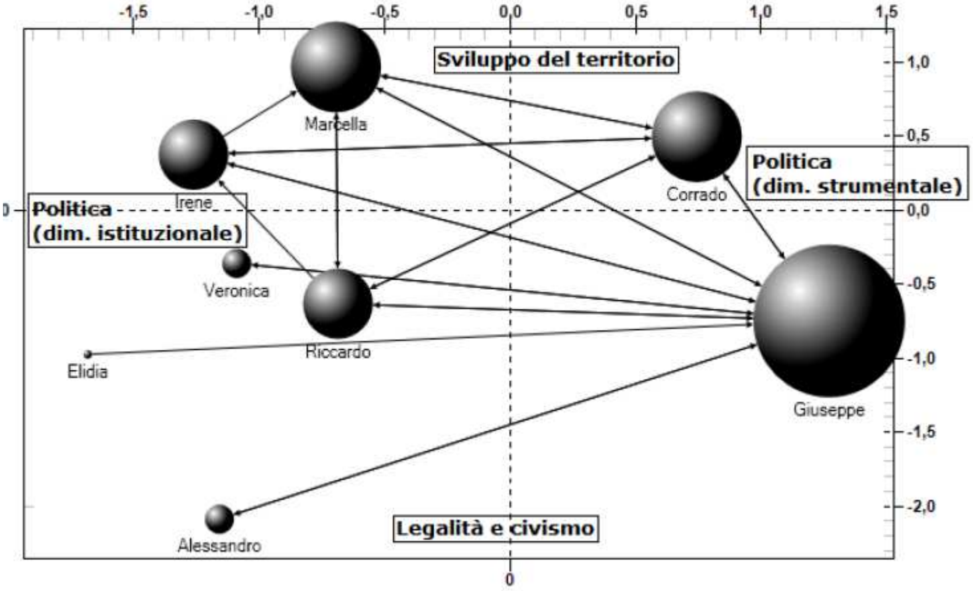


Figure 18 - Factorial space with the dynamic relationships between participants
(Trobia, Frazzica & Milia, 2012, p. 373)

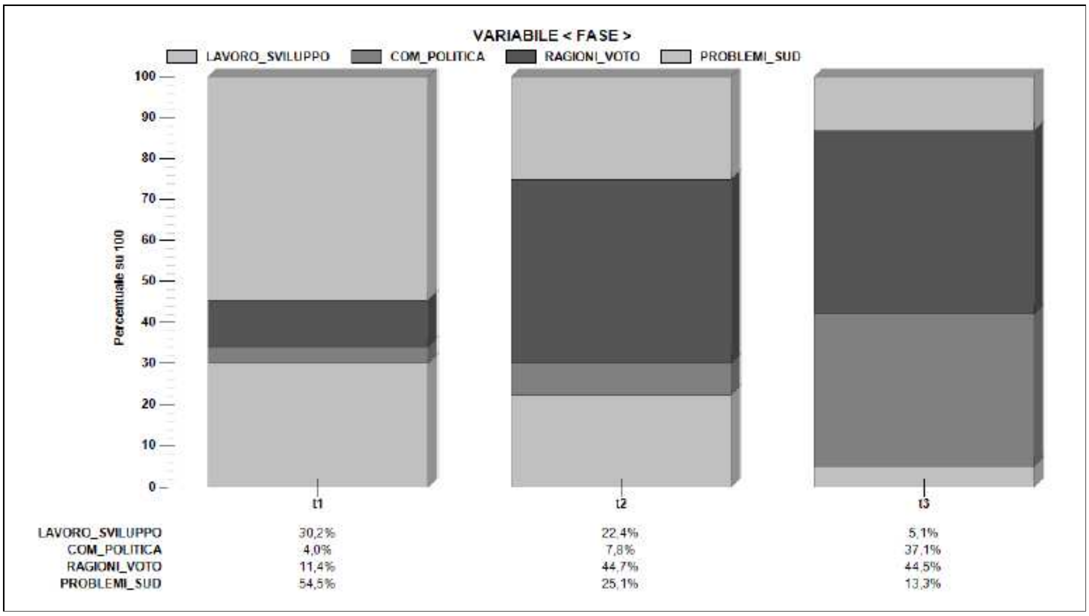


Figure 19 - Diachronic development of ‘themes’ within the three focus groups
(Trobia, Frazzica & Milia, 2012, p. 376)

Now, by taking one’s cue from the paper just quoted, which actually studies the *dynamic* relationship between *people* within specific *contexts* (i.e. focus groups), and before presenting the second ‘recipe’, I would like to point out that – among other under-used tools – T-LAB includes *Sequence Analysis*, which allows the researcher to study the *dynamic relationships within texts and discourses*, i.e. between words, themes, concepts and any ‘labels’.

In other words, such a tool brings into focus the ‘transitions’ (i.e. what comes before or after any ‘x’) and uses a Markovian approach. However its ‘logic’ is very simple.

	x_1	x_2	x_3	x_4	x_5	x_6	Tot
x_1	0	8	7	11	2	1	29
x_2	6	0	**24**	5	10	8	**53**
x_3	9	24	0	3	28	16	80
x_4	3	7	5	0	6	14	35
x_5	4	5	26	11	0	7	53
x_6	7	9	18	5	7	0	46

Table 10: Transitional values

	x_1	x_2	x_3	x_4	x_5	x_6	Tot
x_1	0.00	0.28	0.24	0.38	0.07	0.03	1.00
x_2	0.11	0.00	**0.45**	0.09	0.19	0.15	1.00
x_3	0.11	0.30	0.00	0.04	0.35	0.20	1.00
x_4	0.09	0.20	0.14	0.00	0.17	0.40	1.00
x_5	0.08	0.09	0.49	0.21	0.00	0.13	1.00
x_6	0.15	0.20	0.39	0.11	0.15	0.00	1.00

Table 11: Probability values

Basically it constructs and analyses two asymmetrical co-occurrence matrices (like the above Table 10) whose respective values are the count of how many times, within any text, each ‘x’ precedes or follows the other in the linear (sequential) order of the same text. Subsequently, the same values are converted into probability values (i.e. transition probabilities). For example, if ‘24’ is the count of how many times ‘x_2’ is followed by ‘x_3’, the corresponding transition value is ‘0.45’ (i.e. 24/53; see Table 11 above).

The relationships between the elements considered (i.e. *predecessors* and *successors*) constitute a Markov chain and the appropriate algorithm enables us to map their network links (see Figure 21 below).

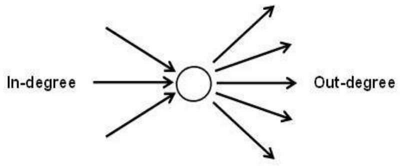


Figure 20 - In-degree and Out-degree

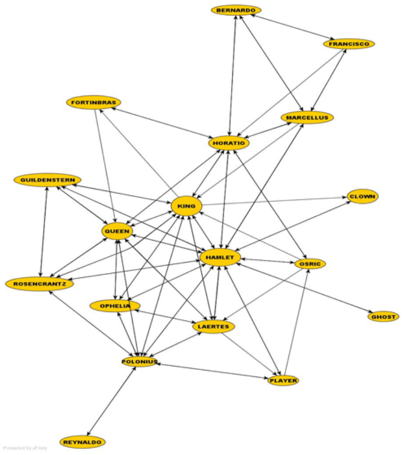


Figure 21 - A network

Such an approach for studying the ‘dynamic’ of discourses is popularly known as ‘network text analysis’ and the way the T-LAB tool is designed allows us to perform simple tasks like studying the sequential order of words in ‘free association’ and more complex tasks like mapping semantic links within a speech, an interview or a book. At the moment, T-LAB doesn’t provide network graphs ‘directly’ like the ones in Figure 21; however it allows the user to export both the *adjacency matrices* and *.graphML* files which can be easily imported by software like *Ucinet* and *yEd***,** the latter of which is available for free download.

More specifically, this T-LAB tool allows a Markovian analysis of three kinds of sequences:

a) Sequences of ‘Key-Words', the items of which are lexical units (i.e. words, lemmas, semantic class etc.) present in the corpus or in a subset of it;

b) Sequences of ‘Themes', the items of which are context units (i.e. elementary contexts) tagged by a T-LAB tool for thematic analysis (i.e. *Thematic Analysis of Elementary Contexts*, *Dictionary-Based Classification* or *Modeling of Emerging Themes*). Since the sequence of elementary contexts (sentences or paragraphs) characterises the entire 'chain' (predecessors and successors) of the corpus, in this case T-LAB performs a specific form of *Discourse*.

c) Sequences recorded in a ‘Sequence.dat’ file made up by the user.

Having said the above, let’s have a look at the second ‘recipe’, which – actually – deals with the *psychotherapy process* (Salvatore et al., 2010). Without entering into the theoretical and methodological issues widely discussed by the authors, I would like to focus on how – within their ‘*DFA’* model (i.e. ‘Discourse Flow Analysis’) - the use of some T-LAB tools in this instance makes sense.

Actually, by looking at the psychotherapy process as an ‘intersubjective dynamic of meaning-making’, the authors are not just interested in exploring the semantic (or thematic) ‘contents’ of the patient-therapist’s verbal exchanges, but rather they are interested in mapping the ways such ‘contents’ are combined ‘one after the other’ throughout the flow of discourse, i.e. the way they are associated for *adiacency* within a ‘time-dependent structure’. More specifically, by focusing on the alternation of ‘decontructive’ and ‘costructive’ phases, they study the transcripts of a 15-session ‘good outcome course of psychotherapy from the York Psychotherapy Depression Project’. The T-LAB analyses they report on are two:

- a *Thematic Analysis of Elementary Contexts* ‘used’ for obtaining 23 (twenty-three) clusters, each of which corresponding to a specific ‘content’ (e.g. ‘feeling/expression of impotence’; ‘desire to indulge own selfishness’, etc.);
- a *Sequence Analysis* of the above thematic contents considered as ‘nodes’ of a network.

Table 12 below reports some descriptive statistics of the first analysis.

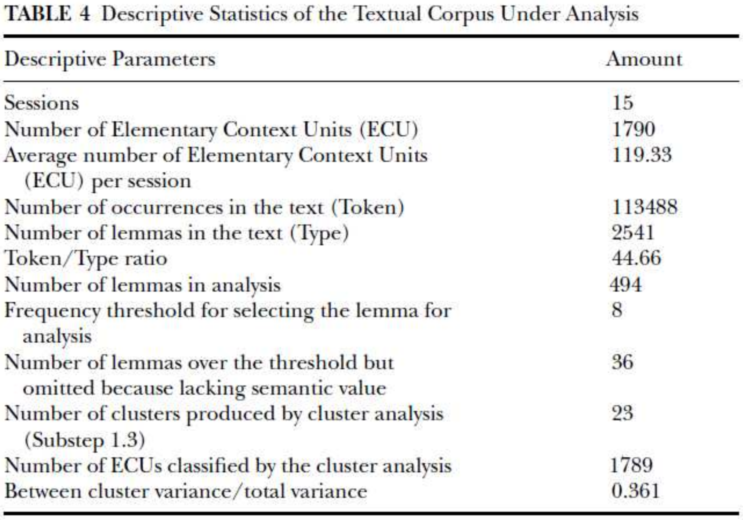


TABLE 4 Descriptive Statistics of the Textual Corpus Under Analysis

Descriptive Parameters	Amount
Sessions	15
Number of Elementary Context Units (ECU)	1790
Average number of Elementary Context Units (ECU) per session	119.33
Number of occurrences in the text (Token)	113488
Number of lemmas in the text (Type)	2541
Token/Type ratio	44.66
Number of lemmas in analysis	494
Frequency threshold for selecting the lemma for analysis	8
Number of lemmas over the threshold but omitted because lacking semantic value	36
Number of clusters produced by cluster analysis (Substep 1.3)	23
Number of ECUs classified by the cluster analysis	1789
Between cluster variance/total variance	0.361

Table 12: Descriptive statistics (Salvatore et al., 2010, p. 207).

Below is a graph which depicts the way seven of the ‘nodes’ (i.e., N_1, N_2 etc.) interact within the first session of the psychotherapy studied.

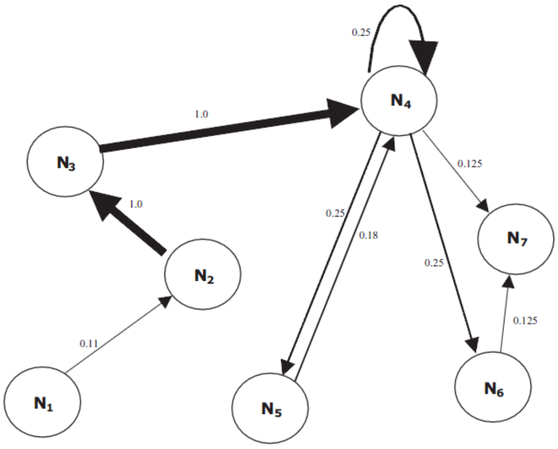


Figure 22: Discourse network of the Lisa psychotherapy's first session (Salvatore et al., 2010, p. 204).

In my opinion the fact that both sociologists (e.g. Trobia, Frazzica & Milia, 2012) and psychologists (e.g. Salvatore et al., 2010) are using T-LAB not just for exploring the text *structure*, but also the text *dynamics* is a good omen indeed.

To conclude I would like to recall that T-LAB is also used by lots of professionals working in private and public organizations around the world. As, at the moment, their reports are neither published in journal articles nor in book chapters, I haven't been able to make reference to them; however, being in contact with many such users, I know that they also use their imagination to play with the 'textscope', and are also able to transform any analysis into an effective product for their customers.

REFERENCES

BARTHES, R. (1964). *Eléments de sémiologie*. Communications 4, Paris, Seuil.

BENZECRI , J.P & F. (1984). *Pratique de l'analyse des données. Analyse des correspondances & Classification*. Paris, Dunod.

BLEI, D.M., NG, A.Y., JORDAN, M.I. (2003) , *Latent Dirichlet Allocation.* Journal of Machine Learning Reserach, 3, pp. 993-1022.

BURT, C.L. (1940). *Factor of the Mind*. London, University of London Press.

CAPONE, V. , & PETRILLO, G. (2011). *Health Promotion in International Documents: Strengths and Weaknesses from the Perspective of Community Empowerment a history of choice in UK health policy.* Journal of Community & Applied Social Psychology, 15 dec. 2012.

DE ROSA, A.S., & HOLMAN, A., (2011). *Social representations of female-male beauty and aesthetic surgery: a cross-cultural analysis.* Temas em Psicologia, Vol. 19, No 1, 79-98.

GAMBETTI, R.C., & GRAFFFIGNA, G., (2010). *The concept of engagement. A systematic analysis of the ongoing marketing debate*. International Journal of Market Research, Vol. 52, Issue 6, 801–826.

GREENACRE, M.J. (1984). *Theory and Applications of Correspondence Analysis*. New York, Academic Press.

GREIMAS, A.J. (1966). *Sémantique structural*. Paris, Larousse.

GREENER, I. (2009). *Towards a history of choice in UK health polic.* Sociology of Health & Illness, Research, Volume 31, Issue 3, 309–324.

GRION, V., & VARISCO, B.M. (2007). *On Line Collaboration for Building a Teacher Professional Identity.* PsychNology, Vol 5, No 3, 271-284.

JAKOBSON R. (1963). *Essais de linguistique générale*. Paris, Editions de Minuit.

KRISHNA, B.S., SATHEESH, P., SUNEEL KUMAR, R. (2012) , *Comparative Study of K-means and Bisecting K-means Techniques in Wordnet Based Document Clustering.* International Journal of Engineering and Advanced Technology (IJEAT), Volume-1, Issue-6, pp. 229-234.

LANCIA, F. (2004). *Strumenti per l'Analisi di Testi. Introduzione all'Uso di T-LAB[Tools for Text Analysis. An Introduction to the Use of T-LAB]*. Milano, Angeli.

LANCIA, F. (2007). *Word Co-occurrence and Similarity in Meaning*. Retrieved on July 15 2012 from http://www.mytlab.com/wcsmeaning.pdf.

LANCIA, F. (2012a). *T-LAB 8.0 - User's Manual.* . Retrieved on September 12 2012, from the T-LAB website: http://www.tlab.it/en/download.php.

LANCIA, F. (2012b). *T-LAB pathways to thematic analysis.* Retrieved on October 26 2012, from the T-LAB website: http://www.tlab.it/en/tpatways.php.

LEBART, L. MORINEAU, & A. PIRON, M. (1995). *Statistique exploratoire multidimensionnelle*. Paris, Dunod.

MARGOLA, D. ET AL. (2010). *Cognitive and Emotional Processing Trough Writing Among Adolescents Who Experienced the Death of a Classmate.* Psychological Trauma: Theory Research, Practice, and Policy, Vol 2, No 3, 250-260.

MCNEELY, C.L. , & HOPEWELL, L. (2010). *U.S. University Leader Pronouncements on Women and STEM Field.* International Journal of Gender, Science and Technology, Vol 2, No 2, 297-333.

MONTALI, L. ET AL., (2011). *Conflicting Representations of Pain: A Qualitative Analysis of Health care professional's Discourse.* Pain Medicine, 12, 1585-1593.

PERRITON, L., (2009). *'We Don't Want Compaining Women!'. A Critical Analysis of the Business Case for Diversity*. Management Communication Quarterly, 23 (2). 218-243.

RASTIER, F., (1987). *Sémantique interprétative*. Paris, P.U.F.

SALVATORE, S. ET AL., (2010). *Looking at the Psychotherapy Process as an Intersubjective Dynamic of Meaning-Making: A Case Study with Discourse Flow Analysis*. Journal of Constructivist Psychology, Volume 23, Isuue 3, 195-230.

SALVATORE, S. ET AL.., (2012). *Automated method of content Analysis. A device for psychotherapy process research*. Psychotherapy Research, 22 (3). 256-273.

SAUSSURE (de) F.(1916). *Cours de Linguistique générale*. Lusanne-Paris, Payot.

SAVARESI, S.M., & BOOLEY, D.L. (2001) , *On the performance of bisecting K-means and PDDP.* 1st OOSIAM Conference on DATA MINING, Chicago, IL, USA, April 5-7, paper n.5, 1-14.

SCHONHARDT-BALEY, C., (2012). *Looking at Congressional Committee Deliberations from Different Perspectives: Is the Added Effort Worth It?* In: NCRM Research Methods Festival 2012, 2nd - 5th July 2012, St. Catherine's College, Oxford.

SENGERS, F., RAVEN, R.P.J.M., & VAN VERNOOIJ, A., (2010). *From riches to rags: Biofuel, media discourse, and resistance to sustainable energy technologies.* Energy Policy Volume 38, Issue 9, September 2010, 5013–5027.

STEINBACH, M., KARYPIS, G., KUMAR, V. (2000) , *A comparison of Document Clustering Techniques.* Proceedings of World Text Mining Conference, KDD2000, Boston.

TAGARELLI, A., KARYPIS, G., (2008) , *A Segment-based Approach To Clustering Multi-Topic Documents.* Workshop on Text Mining, in conjunction with the 8th SIAM International Conference on Data Mining (SDM '08). Atlanta, Georgia, USA, April 24-26

TROBIA, A., FRAZZICA, G. &MILIA, V., (2012) , "L'Analisi del focus group: testi, contesti e reti d'interazione in una prospettiva dinamica" [Analysing Focus Group: texts, contexts and interaction networks in a dynamic perspective], in Cipolla, C., de Lillo, A., Ruspini, E. (ed.), *Il sociologo, le sirene e le pratiche di integrazione*. Milano, Franco Angeli, 361-380.

VELTRI, G.A. (2012). *Viva la Nano-Revolución! A Semantic Analysis of the Spanish National Press.* Science Communication, March 22, 2012, 1–25.

ZHAO, Y., KARYPIS, G., (2004). *Empirical and theoretical comparisons of selected criterion functions for document clustering.* Machine Learning, 55, 311–331.