T-LAB PATHWAYS TO THEMATIC ANALYSIS

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

ABSTRACT

Focusing on some issues concerning methods for thematic analysis, a conceptual distinction between 'pattern' recognition and 'theme' interpretation is proposed, which takes into account word co-occurrences within textual units. Subsequently, three 'postulates' concerning the automated processes for thematic analysis implemented in the T-LAB system are presented. The ways that the software tools allow the user to manage mixed strategies which combine bottom-up and top-down approaches are pointed out. In order to assess both the external and the internal reliability of some software procedures, a couple of experiments are performed by using the Reuter-21758 database. The description of the two experiments, which follows a step by step logic, allows the reader to fully understand how the presented T-LAB tools for thematic analysis can be used.

Citation:
Lancia, F. (2012). *T-LAB Pathways to Thematic Analysis*. Retrieved [date] from http://www.tlab.it/en/tpathways.php.

Word count: 5154.

---

1 - Introduction

There are many ways for doing thematic analysis of textual data, and the related methodological issues are widely discussed by scholars (see for examples: Boyatzis, 1998; Braun & Clarke, 2006; Fereday & Muir-Cochrane, 2006; Guest, MacQueen & Namey, 2012; Patton, 2002). Among these issues, there is one concerning two primary ways[1] by which 'themes' should be identified, that is:

[1] R. E. Boyatzis (1998, p. vii) argues that 'The themes may be initially generated inductively from the raw information or generated deductively from theory and prior research'; whereas, according to N. Hayes (1997, p. 6), 'Whether a researcher uses an inductive or deductive methodology is again less of a dichotomy that it may seem, and the distinction is also less tightly linked with the use of qualitative methods than some have implied.'

(a) by an 'inductive'[2] (or bottom-up) way, which allows themes to 'emerge' from textual data;

(b) by a 'deductive' (or top-down) way, which applies pre-defined categories (or 'themes') to textual units.

Even if the above dichotomy deals with questionable issues, at the moment we are not interested in discussing its epistemological foundations; rather we would like to point out that – in scientific literature - there isn't agreement about what counts as a 'theme' (or as a 'topic')[3]. This is probably because, in 'qualitative' thematic analysis, usually the 'how' (i.e. the methods and their steps) and the 'what' (i.e. themes as objects of investigation) are in circular relationship between them. Consequently, the 'analysis' and 'interpretation' processes are usually overlapping and iterative.

To be more specific, let's suppose that our thematic analysis is 'inductive' and that it requires to identify 'chunks' of text to be classified into mutually exclusive categories (i.e. themes). The fact is that manually identifying such chunks and determining their borders (for example, highlighting text passages manually), as well as comparing such chunks and exploring their similarities/differences, are not just 'analytical' steps. In fact they require that the researcher have some ideas about the 'themes' h/she is looking for, and that h/she is inferring meaning from 'data'.

For example, while describing the 'inductive' process of thematic analysis, Boyatzis (1998, p. 3) argues:

'They [i.e. the researchers] perceived a pattern, or theme, in seemingly random information. *They saw a pattern*! The perception of this pattern begins the process of the thematic analysis. It allows these people to continue to the next major step, classifying or encoding the pattern. They give it a label or definition or description. This allows them to proceed to the third major step in thematic analysis, interpreting the pattern' (ib., p. 3. Author emphasis).

According to this author, 'A theme is a pattern found in the information that at the minimum describes and organizes possible observations and at maximum interprets aspects of the phenomenon' (ib. p. 4).

Actually, in scientific literature, 'pattern' is often used as synonymous of 'theme'. However, when using software systems like T-LAB, the researcher should be aware that (a) 'pattern' recognition on the one hand and (b) 'theme' interpretation on the other are very different tasks. In fact the former

---

[2] According to J. M. Morse and C. Mitcham (2002, p. 30), 'The issue is not *if* the inductive process can be used in qualitative research, but *how* induction should be used'; however, we argue that the type of inference which allows themes to 'emerge' from data would be properly described as 'abductive' rather than as 'inductive' (see Lancia, 2007).

[3] So, paraphrasing Binet *('*Intelligence is whatever intelligence tests measure'), we could say that - in software – a 'theme' or a 'topic' is whatever 't' (i.e. 'theme' or a 'topic') a specific algorithm allows us to detect and measure.

(i.e. pattern recognition) can be performed algorithmically[4], whereas the latter (i.e. theme interpretation) requires human intervention. So we could also argue that, in computer aided text analysis, pattern recognition deals mostly with *co-textual* relations (i.e. relations 'internal' to text)[5], whereas theme interpretation deals both with *co-textual* and *contextual* relations, the latter referring to text as communication event.

As a matter of fact, in T-LAB logic 'pattern' always refers to a specific phenomenon, that is word co-occurrences within textual units (i.e. sentences, paragraphs, text segments, documents). So, being a software system mostly oriented to automated text analysis, it requires that the user agrees with the following postulates:

1. the borders of textual units are defined in advance[6] (see the corpus pre-processing options of the software[7]) and any 'unitizing' process is applied automatically;
2. any comparison between textual units and their patterns (e.g. similarities between co-occurrence vectors) is performed algorithmically;
3. grouping textual units into 'categories' (i.e. themes), each of which – according to Patton (2002, p. 465) – should exhibit 'internal homogeneity' and 'external heterogeneity'[8], is the result of a clustering algorithm.

However, by using various tools, the user is enabled to choose his list of relevant words, as well as to group and label them as he wishes. More to the point: the user is also enabled to build/apply customised dictionaries, to decide the number of themes to consider (where 'theme' usually is a cluster of textual units the co-occurrence profiles of which are the most 'similar' to each other and the most 'different' from those belonging to the other themes/clusters), as well as to measure, label and 'illustrate' the characteristics of such themes in a variety of ways.

Obviously, the use of the above 'procedures' doesn't allow the user to highlight themes which are sensitive to 'nuances' and metaphors; however, as would be expected by many scholars, each T-LAB analysis is completely replicable[9] and the researcher is enabled to 'identify and examine themes from textual data in a way that is transparent and credible' (Greg, MacQueen & Namey, 2012, p. 15).

[4] In some ways, all T-LAB tools deal with 'pattern recognition'; in fact each T-LAB output of statistical analysis (i.e. charts and tables) show items combined in an ordered way. However this paper deals with the T-LAB tools which are grouped in the 'thematic analysis' sub-menu of the software (see Lancia, 2012a).

[5] Actually T-LAB allows the user also to map the relationships between descriptive characteristics of the data sources (i.e. categorical variables used to encode texts).

[6] Here 'in advance' means before performing any statistical computation. However some algorithms for topic analysis have been proposed which don't use a-priori segmentation of the text (Canny, 2004; Canny & Rattenbury, 2006).

[7] Lancia (2012a)

[8] As a matter of fact, these two criteria are used for defining the main task of any cluster analysis.

[9] For the study to be replicable, the researcher needs to monitor and report his analytical procedures and processes as completely and truthfully as possible; where procedures and processes refer to 'identifying, coding, categorising, classifying, and labeling the primary patterns in the data' (Patton, 2002, p. 463).

2- T-LAB pathways to thematic analysis

As recalled above, T-LAB tools enable the user to follow various paths to thematic analysis. To illustrate such a variety, a simple two-way table (see Figure 1 below) can be useful.

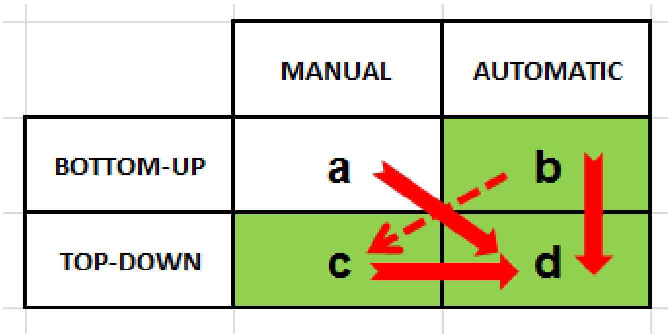

Figure 1

Key:

- 'manual' = manually defining 'themes' and their characteristics;
- 'automatic' = use of statistical algorithms to detect 'themes' and their characteristics;
- 'bottom-up' = data-driven (i.e. 'inductive') approach which doesn't use pre-existing coding frames (i.e. prefixed categories);
- 'top-down' = analytical approach which applies pre-existing coding frames to textual units;
- red arrows = paths allowed by the T-LAB tools for thematic analysis;
- red solid arrows = T-LAB paths illustrated in this document[10];
- green squares = analyses enabled by T-LAB.

So, for example, the classical use of software for 'qualitative' analysis deals mostly with methods falling into the 'a' and 'c' cells (see Figure 1 above), whereas the use of software which only uses unsupervised clustering or topic-model approaches[11] falls into the 'b' cell.

In detail, by making reference to cells in Figure 1, the T-LAB tools which enable the performance of the various thematic analyses (and so grouping and tagging textual units) are the following:

[10] The dashed arrow in Figure 1 refers to the use of the topic model approach (see Blei, Ng & Jordan, 2003) implemented in T-LAB, which however is not discussed in this paper.

[11] The main difference between the unsupervised clustering and topic-model approaches resides in the fact that the former assigns each textual unit to a 'theme' (or to a 'topic'), whereas the latter considers each textual unit as a 'mixture' of various topics (or themes). Unlike other software for text analysis, T-LAB enables the researcher to use both the above mentioned approaches. For more information about the T-LAB architecture see the paper 'The Logic of the T-LAB tools explained' (Lancia, 2012b).

BOTTOM-UP/AUTOMATIC (b)	**TOP-DOWN/MANUAL** (c)	**TOP-DOWN/AUTOMATIC** (d)
Thematic Analysis of the Elementary Contexts (unsupervised clustering); **Thematic Document Classification** (unsupervised clustering); **Modeling of Emerging Themes** (probabilistic approach).	**Dictionary Based Classification**; **Modeling of Emerging Themes**.	**Thematic Analysis of the Elementary Contexts** (supervised classification); **Thematic Document Classification** (supervised classification); **Dictionary Based Classification** (supervised classification).

Table 1

The reason why in the above table some tools are listed in more than one column is that their menus contain various options. Moreover it is worth noting that in the (d) cases (i.e. top-down/automatic classification) all listed tools use the same three step procedure, that is:

1. normalization of the seed vectors (i.e. co-occurrence profiles) corresponding to the 'k' categories (or 'themes') of the dictionary[12] used;
2. computation of Cosine similarity and of Euclidean distance between each 'i' textual unit and each 'k' seed vector (both normalized using Euclidean norm);
3. assignment of each 'i' textual unit to the 'k' class or category for which the corresponding seed is the closest (N.B.: In this case, maximum Cosine similarity and minimum Euclidean distance must coincide, otherwise T-LAB considers the 'i' textual unit as unclassified).

In the pages below – and in the first instance – we will concentrate on the use of an automatic/top-down (i.e. 'd') approach after having performed any thematic analysis (i.e. 'a'. 'b', 'c') the results of which imply that 'n' textual units have been tagged with 'k' themes or topics. So the reader is invited to think about situations in which any researcher – either by manually coding, or by using T-LAB (or other software) tools – has a dataset consisting of 'n' textual units (i.e. elementary contexts[13] or documents) subdivided into 'k' groups (where 'k' can vary from 2[14] to 50) in such a way that each textual unit is tagged with only one of 'k' themes under examination.

[12] See below for more explanations.

[13] Depending on the user's choice, in T-LAB 8.0 the elementary contexts 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, twitties etc.).

[14] For algorithmic reasons (e.g. performing a Correspondence Analysis of contingency tables), two tools listed in column 'c' of Table 1 (i.e. 'Thematic Analysis of the Elementary Contexts' and 'Thematic Document Classification') require a minimum of 3 categories (or themes).

The situations in which the use the 'top-down' method implemented in T-LAB could be useful are many. For example, there are the following:

- assessing a previous 'manual' content analysis;
- perform a sentiment analysis;
- apply the same criteria (i.e. the same coding frames) when analysing various corpuses, the vocabularies (and the 'themes') of which are quite homogeneous;
- etc.

Obviously, having being obtained by a software procedure, the results of the top-down method implemented in T-LAB are 'stable' and 'reproducible'; however the 'accuracy' of the method (i.e. the way it conforms to its specifications and yields what it is designed to yield[15]) must be assessed through measures concerning both the 'external' and 'internal' reliability. In detail:

a) the 'external reliability' requires that the above automatic method, after having 'learned' any coding frame used by human coders when classifying textual units belonging to a 'C1' corpus, is able to classify textual units belonging to a 'C2' corpus by obtaining results which are in agreement with those obtained by the same human coders when using the same criteria for classifying textual units belonging to the 'C1' corpus (see inter-rater reliability);
b) the 'internal' reliability requires that, after having performed an unsupervised classification of 'n' textual units belonging to a 'C1' corpus, and after having stored 'its' criteria for classifying such textual units into 'k' groups, the T-LAB automatic method is able to classify the same 'n' textual units into the same 'k' groups even when such textual units are included in a 'C2' corpus which is different from the 'C1' one.

Below two different experiments will be illustrated, dealing with the 'external' and 'internal' validation respectively. So, more information will be provided about the methods used by the T-LAB tools and their respective analysis steps.

As a matter of fact, the T-LAB tools for thematic analysis allow the user to classify various textual units (e.g. words, elementary contexts and short documents); however the experiments below refer to the document classification only, the logic and the performances of which are similar to those of the elementary context classification.

[15] See K. Krippendorff (2004, p. 215)

3- Dataset and corpuses

For the purpose of our experiments, a dataset has been extracted from the Reuters-21578 corpus[16] and has been subdivided into three corpuses (see below for more details). As stated in a web page of the Stanford University[17], such a corpus – which is the main benchmark for text classification evaluation – 'is a collection of 21,578 newswire articles, originally collected and labeled by Carnegie Group, Inc. and Reuters, Ltd. in the course of developing the CONSTRUE text classification system […] The articles are assigned classes from a set of 118 topic categories. A document may be assigned several classes or none, but the commonest case is single'.

Without going into lots of technical details, given that – at the moment – the T-LAB tool for Document Classification (as well as the tool for 'Thematic Analysis of Elementary Contexts') allows the user to obtain up to 50 classes/clusters, by using a reduced version of the Reuters-21578 corpus which includes documents assigned to 90 classes, we proceed as follows:

- firstly we selected only the articles assigned to one class;
- secondly we selected 15 classes which weren't too unbalanced in the number of documents (N.B.: in fact, in the Reuters collection mentioned above, four classes include more than 2000 documents, whereas about forty classes include less than 20 documents);
- thirdly we randomly split the documents belonging to selected classes into two datasets, respectively 'A' and 'B', the sum of which is a 'C' dataset including 2255 documents.

The structure of three datasets is illustrated in Table 2 below and their logic relations are the following:

- $(A \cap B) = \varnothing$ (i.e. no document in 'A' is also in 'B', and no document in 'B' is also in 'A');
- $C = A + B$.

[16] See http://www.daviddlewis.com/resources/testcollections/reuters21578/
[17] See http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-text-classification-1.html

TOPIC	**(A)** 25%	**(B)** 75%	**(C)** 100%
alum	13	37	50
cocoa	15	46	61
coffee	28	84	112
cpi	18	53	71
crude	94	280	374
gnp	19	55	74
gold	23	67	90
grain	13	38	51
interest	68	204	272
jobs	12	37	49
money-fx	77	232	309
money-supply	37	112	149
ship	36	108	144
sugar	31	91	122
trade	82	245	327
Total (Documents)	**566**	**1689**	**2255**

Table 2

Following the aim of the experiment, first of all the three datasets have been transformed into corpuses ready to be imported by T-LAB. So, for example, at the end of corpus preparation phase, the 566 documents belonging to the 'A' dataset have been assembled in a text file where each document was preceded by a coding line (see below for one example).

```
**** *IDNUMBER_00274 *TOPIC_ship
STORMY WEATHER TO DISRUPT NORTH SEA SHIPPING
STATE COLLEGE, PA., March 27 - Very stormy weather is likely in the North Sea through Saturday, disrupting shipping in the region, private forecaster Accu-Weather Inc said. Rain will accompany the strong winds that are expected over the North Sea today into tonight. Saturday will also be very windy and cooler with frequent showers. Winds today will be southwest at 30 to 60 mph, but will become west to northwest tonight and Saturday at 25 to 50 mph. Waves will build to 20 to 30 feet today and tonight and continue Saturday. Wind and waves will not diminish until late in the weekend.
```

4 - Experiment N. 1: Assessing External Reliability

After having imported the 'A' corpus by T-LAB (default options[18]), to ensure that the dictionary[19] to be created included most of relevant words, a word list was selected which included 2,427 'lemmas' (distinct words: 3,979; occurrence values: min. 3, max. 753; word tokens: 44,963). Subsequently, in order to create/save the dictionary of the fifteen categories (i.e. topics), the T-LAB tool named 'Dictionary Based Classification' has been selected and the first results have been obtained with just three mouse clicks (see below).

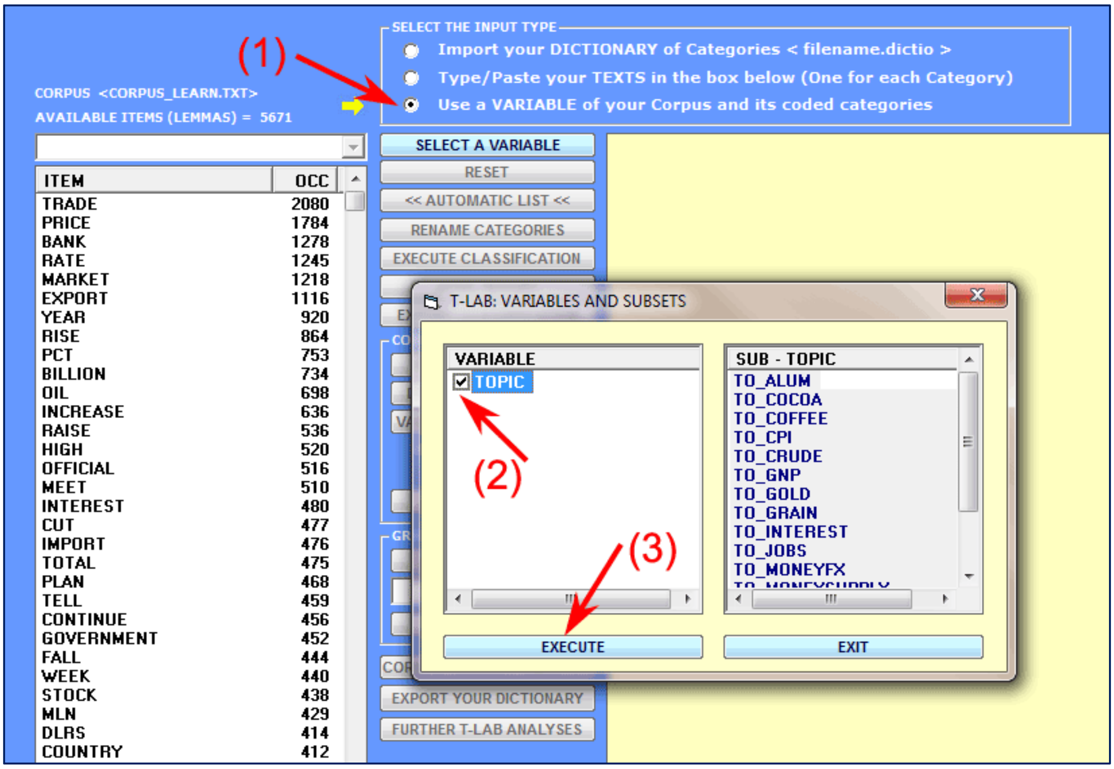

Figure 2

In such cases the first results are tables the columns of which contain the word occurrence vectors for each category (see Figure 3 below). Even if T-LAB allows the user to customise such dictionaries in a variety of ways, in this case no customisation has been performed and the only option selected has been 'export your dictionary'. So, a *.dictio* file has been automatically created, the format of which is illustrated in Table 3 below.

[18] When using the default options, T-LAB performs an automatic lemmatisation of the corpus, uses a stop-word list and detects a limited number of multi-word phrases.

[19] In such a case, the 'dictionary' refers to the coding frame for thematic analysis (see below for more information).

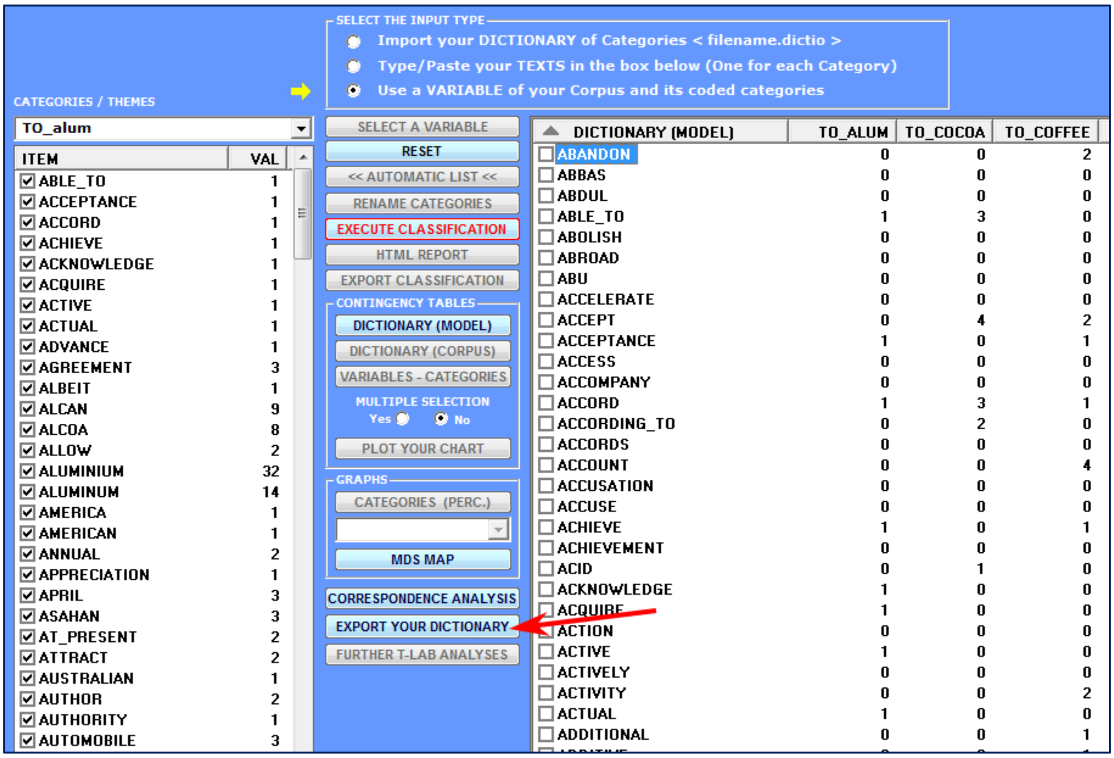

Figure 3

TO_ALUM;ALCAN;9
TO_ALUM;ALCOA;8
TO_ALUM;ALLOW;2
TO_ALUM;ALUMINIUM;32
TO_ALUM;ALUMINUM;14
…
TO_COCOA;COAST;6
TO_COCOA;COCOA;92
TO_COCOA;COMMENT;1
TO_COCOA;COMMISSION;2
TO_COCOA;COMMISSIONERS;1
…
TO_COFFEE;COAST;1
TO_COFFEE;COFFEE;120
TO_COFFEE;COLLAPSE;3
TO_COFFEE;COLOMBIA;19
TO_COFFEE;COLOMBIAN;2
…
TO_CRUDE;CROWE;3
TO_CRUDE;CROWN;5
TO_CRUDE;CRUDE;122
TO_CRUDE;CRUDES;19
TO_CRUDE;CTS;55
…

Table 3

At this point, after having closed the 'A' project, the 'B' corpus including 1,689 documents of the Reuters-21578 (see Table 2 above) has been imported by using the same criteria applied in the case of 'A' corpus. It is worth noting that, in this case, the only difference between the two projects (i.e. 'A' and 'B') relies on the word list; in fact, since the T-LAB tool for document classification presently allows users to select up to 3,000 lemmas, a word list of the 'B' corpus has been automatically generated including 2,840[20] lemmas (distinct words: 5,162; occurrence values: min. 6, max. 2,370; word tokens: 138,569).

Subsequently the 'Thematic Document Classification' tool has been selected by using the 'supervised classification (dictionary of categories)' option. Then, by using the default option (i.e. min. 2 word co-occurrences within the context units) the dictionary created through the 'A' corpus has been imported and applied to the classification of documents belonging to the 'B' corpus (see Figure 4 below).

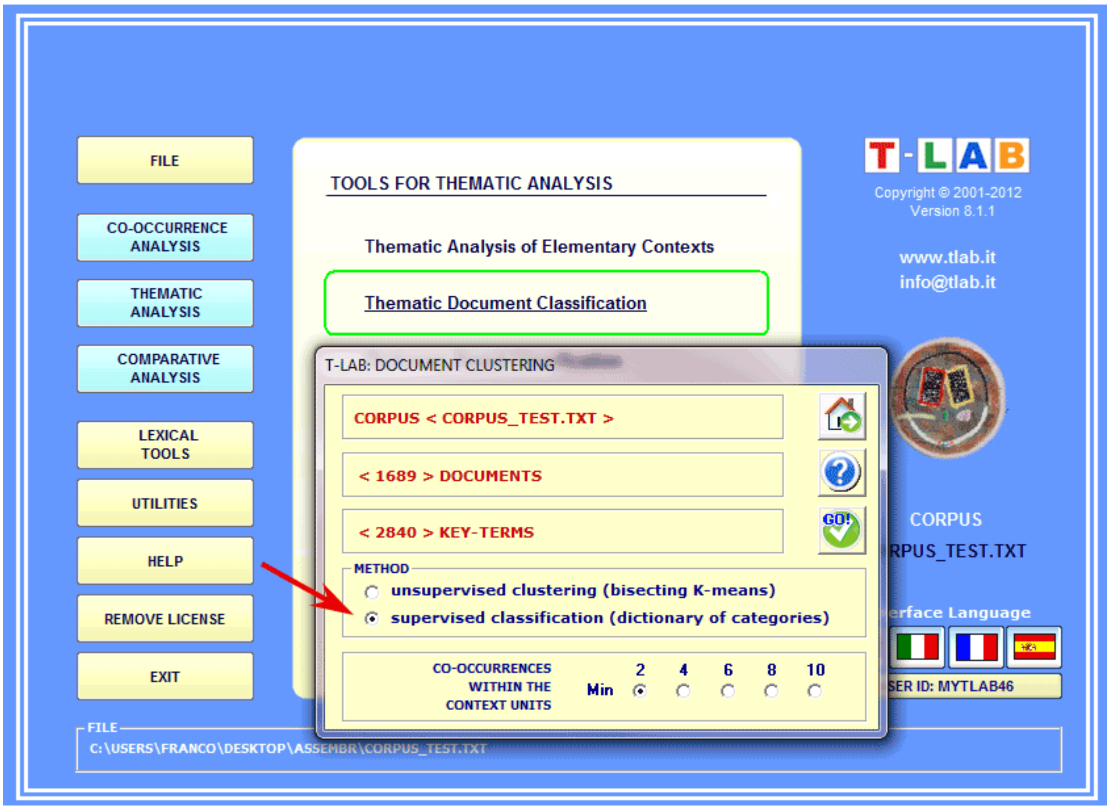

Figure 4

At the end of the classification process, T-LAB produced lots of outputs (i.e. tables and charts); however, for the purpose of our experiment, only a table containing documents and their tags (i.e. their 'thematic clusters') has been exported (see Figure 5 below).

---

[20] In such a case (i.e. automatic classification), the larger is the word list, the more accurate is the analysis.

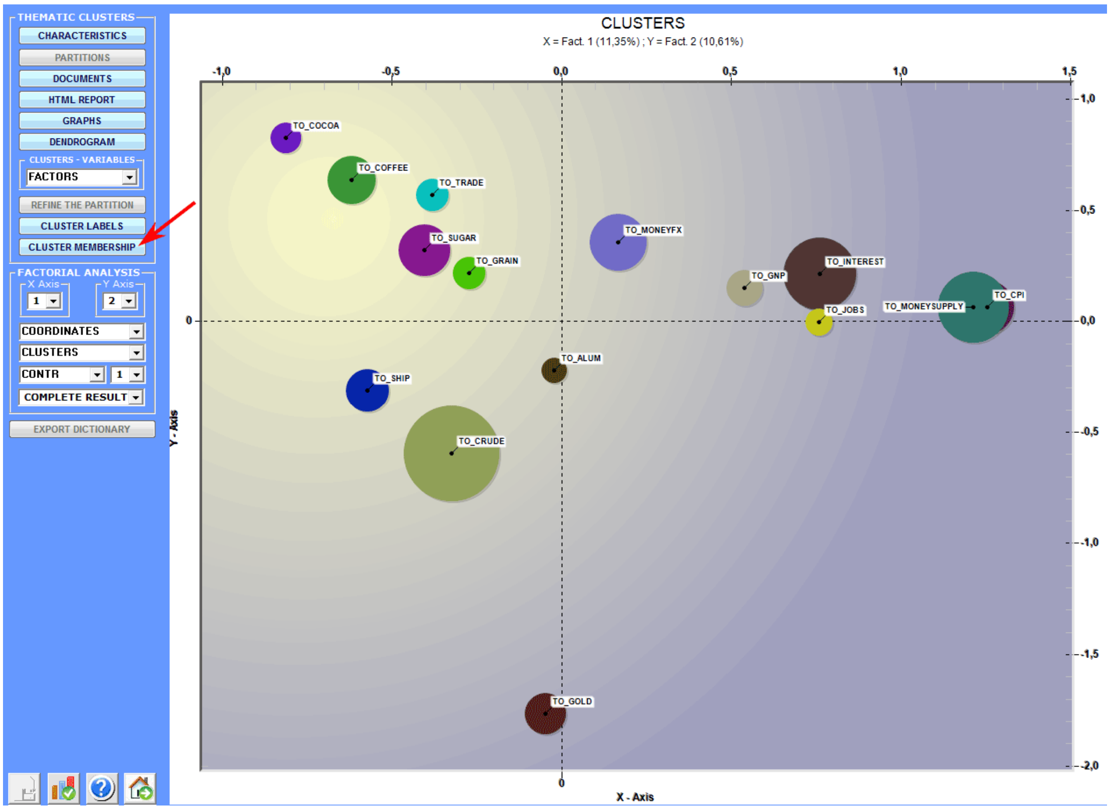

Figure 5

In order to make a correct comparison between the 'manual' coding (i.e. Reuters-21578 corpus sample) and the automatic/top-down classification performed by T-LAB, various measures have been obtained, some typical of the information retrieval field[21] (i.e. precision and recall; see Table 5 below), some mostly used in the content analysis field (e.g. Cohen's Kappa coefficient for estimating inter-rater agreement). All measures have been obtained by analysing the data in Table 4 below, the rows ('manual') and columns ('automatic') of which refer to the 'B' corpus documents classified by T-LAB (i.e. 1,293, out of a total of 1,689).

[21] See C. J. van Rijsbergen (1979)
van Rijsbergen, Cornelis Joost "Keith" (1979); *Information Retrieval*, London, GB; Boston, MA: Butterworth, 2nd Edition

	ALUM	COCO/	COFFE	CPI	CRUDE	GNP	GOLD	GRAIN	INTERE	JOBS	MONE	MONE	SHIP	SUGAI	TRADE	Total
ALUM	**20**	0	0	2	0	1	0	0	1	1	0	1	0	2	0	**28**
COCOA	0	**38**	1	0	0	1	0	0	0	0	1	0	0	0	0	**41**
COFFEE	0	0	**80**	0	0	0	0	0	1	0	0	0	0	0	0	**81**
CPI	0	0	0	**46**	0	0	0	0	1	0	0	0	0	0	0	**47**
CRUDE	0	0	0	3	**219**	1	4	0	1	1	0	0	3	0	0	**232**
GNP	0	0	0	0	0	**34**	0	0	0	0	0	0	0	0	0	**34**
GOLD	0	0	0	0	0	0	**65**	0	0	0	1	0	0	0	0	**66**
GRAIN	0	0	0	1	0	0	0	**30**	1	0	0	0	0	3	0	**35**
INTEREST	0	0	0	15	0	1	0	0	**123**	0	0	22	0	0	0	**161**
JOBS	0	0	0	6	0	1	0	0	0	**21**	0	0	0	0	0	**28**
MONEY-FX	0	0	0	2	0	1	0	0	22	0	**97**	32	0	0	0	**154**
MONEY-SUPPLY	0	0	0	26	0	1	0	0	1	0	0	**72**	0	0	0	**100**
SHIP	2	1	0	0	0	0	0	3	0	3	2	1	**70**	0	0	**82**
SUGAR	0	0	1	0	0	0	0	1	0	0	0	0	0	**87**	0	**89**
TRADE	2	0	7	0	0	10	0	6	8	1	11	24	0	4	**42**	**115**
Total	**24**	**39**	**89**	**101**	**219**	**51**	**69**	**40**	**159**	**27**	**112**	**152**	**73**	**96**	**42**	**1293**

Table 4

ANALYSED DOCUMENTS	**1689**	i.e. all documents of the 'B' corpus
CLASSIFIED DOCUMENTS	**1293**	i.e. documents automatically classified by T-LAB
HITS (TRUE POSITIVES)	**1044**	
FALSE HITS (FALSE POSITIVES)	**249**	
MISSES (FALSE NEGATIVES)	**396**	
PRECISION	**0.8074**	HITS/(HITS + FALSE HITS)
RECALL	**0.7250**	HITS/(HITS + MISSES)
F	**0.7640**	2*((PRECISION*RECALL)/(PRECISION + RECALL))

Table 5

Cohen's Kappa = 0.7883[22];
Kappa error = 0.0121; Kappa C.I. (alpha = 0.0500) = 0.7647-0.8119;
Variance = 0.0001; z = 91.2476; p = 0.0000.

So, according to the above results, we can conclude that - after having 'learned' the coding frame used by human coders when classifying textual units (and provided that, like in the Reuters-21578 corpus, such a coding frame deals with word co-occurrence patterns) – the top-down method implemented in T-LAB performs quite well (see 'external' reliability as defined in section 2 above).

[22] Common criteria for evaluating Cohen's Kappa (see Landis & Koch, 1977) are: POOR agreement = less than 0.20; FAIR agreement = 0.21 to 0.40; MODERATE agreement = 0.41 to 0.60; GOOD agreement = 0.61 to 0.80; VERY GOOD agreement = 0.81 to 1.00.

5 - Experiment N. 2: Assessing Internal Reliability

The aim of the experiment we report below deals with 'internal' reliability, and so it intends to measure the T-LAB agreement with 'itself', that is the agreement between different T-LAB procedures used for classifying textual units.

For algorithmic reasons, when the top-down automatic classification (see steps from 1 to 3 illustrated in section '2' above) is applied, provided that:

(a) a dictionary automatically produced by a T-LAB tool which performs an unsupervised clustering[23] has been used,

(b) the same corpus through which the dictionary has been created is analysed,

(c) the chosen cluster partition for generating the dictionary has been 'refined' by using the appropriate re-classification method,

both the precision and recall parameters are equal to 100%[24] and the Cohen's Kappa is equal to 1. In fact, in the above cases, the 'export dictionary' option (see below for further explanations) has been implemented also to allow the researcher to quickly repeat (i.e. same results) any analysis of the same data with the same criteria.

However this second experiment deals with situations where, when performing a top-down classification ('TDC') which uses a dictionary obtained through an unsupervised clustering ('UC'), the corpus is not the same. In detail, by using the three corpuses extracted from the Reuter dataset, the structure of which is reported in Table 2 above, the following four analyses have been performed (see Table 7 below).

ANALYSIS	UNSUPERVISED CLUSTERING ('UC')	TOP-DOWN CLASSIFICATION ('TDC')
A1	CORPUS 'A'	CORPUS 'C'
A2	CORPUS 'B'	CORPUS 'C'
A3	CORPUS 'C'	CORPUS 'A'
A4	CORPUS 'C'	CORPUS 'B'

23 Both in the case of 'Thematic Document Classification' and 'Thematic Analysis of Elementary Contexts', the unsupervised clustering implemented in T-LAB follows the same procedure, the main steps of which are four: a - construction of a data table context units x lexical units; b - TF-IDF normalization and scaling of row vectors to unit length (Euclidean norm); c - clustering of the context units (measure: cosine coefficient; method: bisecting K-means); d – storage of the obtained partitions in 'k' clusters. For more information about the bisecting K-means, see Steinbach, Karypis & Kumar (2000), Savaresi & Boley (2001), and Lancia (2012b).

24 To obtain this result, four steps are required: 1 – select a cluster partition; 2 – save the corresponding dictionary; 3 - 'refine' the partition through the reclassification based on typical words; 4 – repeat any analysis by using the 'top-down' approach (i.e. by importing the dictionary saved in step 2). As steps 3 and 4 use the same procedure above described, the results will be identical.

Table 7

In all the four analyses above the problem was the same, that is to assess how many documents correctly classified through the unsupervised clustering (see the 'classified' column in Table 8 below) were retrieved and assigned to the same clusters by using a top-down approach. In all the four above analyses the same word lists of the experiment N. 1 have been used; moreover, in all the above four analyses a partition in 14 classes (i.e. thematic clusters) has been used. It is worth noting that in such cases the fact that the number of classes (14) obtained by the T-LAB unsupervised clustering doesn't match the number (15) of topics of the Reuter dataset is irrelevant; in fact, here the issue is not the 'external' validation, but rather the 'internal' one, that is the relationships between the 'UC' and 'TDC' analyses performed by the same software tool.

In all the four cases above the analysis procedure has been the same. In greater detail, each time the 'Thematic Document Classification' tool has been used as follows:

Step 1: an 'unsupervised clustering' has been performed (see Figure 6 below);

Step 2: a partition into 14 'thematic clusters' has been selected (see '1' in Figure 7 below);

Step 3: the corresponding dictionary has been exported (see '2' in Figure 7 below);

Step 4: the above partition (i.e. 14 clusters) has been refined (see '3' in Figure 7 below) by using a method which uses the same dictionary saved in step 3 above (see Figure 8 below). As a consequence, for each analysed corpus (i.e. 'A', 'B', 'C') a different number of documents has been classified (see 'UC' column in Table 7 above).

Step 5: when exiting, the results (i.e. the classification of 'n' documents into 14 clusters) have been saved for comparison.

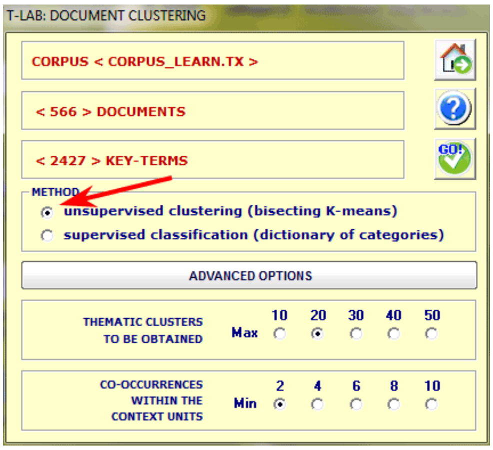

Figure 6

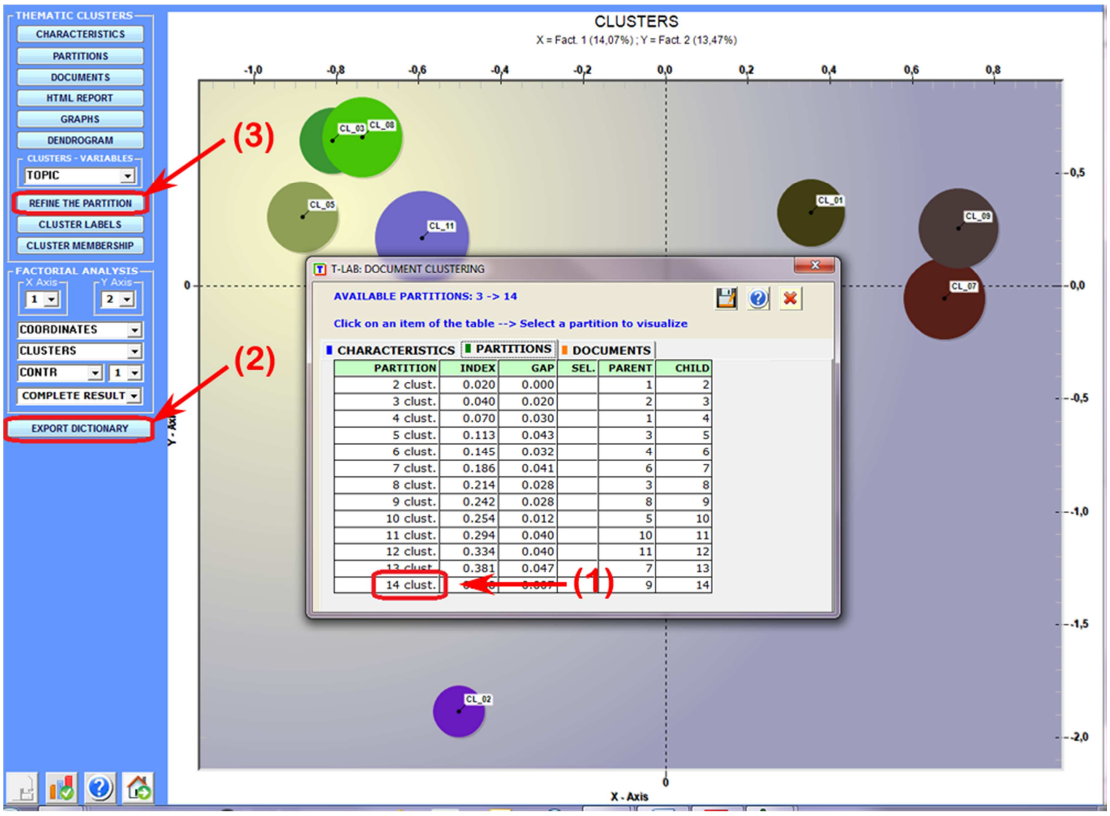

Figure 7

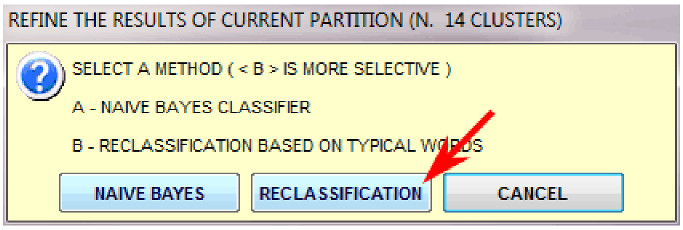

Figure 8

Table 8 below summarises the results of the unsupervised clustering (‘UC’) applied to the three corpuses under examination:

	DOCUMENTS			
CORPUS	ANALYSED	CLASSIFIED	UNCLASSIFIED	PERCENTAGE
A	566	446	120	78,80%
B	1689	1429	260	84,60%
C	2255	1936	319	85,85%

Table 8

So, in order to assess the 'internal' reliability of the top-down procedure implemented in T-LAB, by using the dictionary resulting from the corresponding unsupervised clustering (see 'UC' column in Table 7 above), the four analyses described in Table 7 above ('TDC' column) have been performed. In all the four cases listed in Table 7 above the structure of the dictionary, which is similar to that illustrated in Table 3 above, was the following:

```
THEME_01;ABOLISH;7
THEME_01;ACCESS;27
THEME_01;ACCOUNT;7
THEME_01;ACCUSE;25
...
THEME_02;ACCUSATION;41
THEME_02;ACQUIRE;39
THEME_02;ACTION;14
THEME_02;AGENCY;18
...
THEME_03;AGGREGATE;27
THEME_03;ASSET;10
THEME_03;AUSTRALIA;61
THEME_03;AUSTRALIAN;53
...
THEME_04;AFFAIR;5
THEME_04;AFRICA;40
THEME_04;AFRICAN;4
THEME_04;ALCAN;138
...
```

Table 9

Each time (that is 4 times) a supervised classification has been performed by using the same steps of the N. 1 Experiment (see Figure 4 above). Each time the problem was the same, that is – as already recalled - to assess how many documents correctly classified through the unsupervised clustering (see the 'classified' column in Table 8 above) were retrieved and assigned to the same clusters by using a top-down approach. The results are illustrated by the four tables below, where columns refer to clusters obtained by the unsupervised method and rows refer to the top-down classification. In all the four tables, the 'UNCLASS' row refers to documents resulting as unclassified through the top-down method.

Key:

A1-A2-A3-A4 analyses refer to Table 7 above;

Table 10: ‘A1’ analysis;

Table 11: A2’ analysis;

Table 12: ‘A3’ analysis;

Table 13: ‘A4’ analysis;

	CLUST_01	CLUST_02	CLUST_03	CLUST_04	CLUST_05	CLUST_06	CLUST_07	CLUST_08	CLUST_09	CLUST_10	CLUST_11	CLUST_12	CLUST_13	CLUST_14	TOTAL
CLUST_01	23														23
CLUST_02		26													26
CLUST_03			51												51
CLUST_04				28											28
CLUST_05					48										48
CLUST_06						25									25
CLUST_07							27								27
CLUST_08								10							10
CLUST_09									3						3
CLUST_10										16		1			17
CLUST_11											26				26
CLUST_12												43			43
CLUST_13													33		33
CLUST_14	1													63	64
UNCLASS	1	1	4		2	1	1	2	1	2	1	1	2	3	22
TOTAL	25	27	55	28	50	26	28	12	4	18	27	45	35	66	446

Table 10

	CLUST_01	CLUST_02	CLUST_03	CLUST_04	CLUST_05	CLUST_06	CLUST_07	CLUST_08	CLUST_09	CLUST_10	CLUST_11	CLUST_12	CLUST_13	CLUST_14	TOTAL
CLUST_01	106														106
CLUST_02		53													53
CLUST_03			168												168
CLUST_04				84											84
CLUST_05					72										72
CLUST_06						188									188
CLUST_07							107								107
CLUST_08								32							32
CLUST_09									44						44
CLUST_10					1					90				1	92
CLUST_11											156				156
CLUST_12												130			130
CLUST_13													72		72
CLUST_14														110	110
UNCLASS	1			2	2	2		2				3	3		15
TOTAL	107	53	168	86	75	190	107	34	44	90	156	133	75	111	1429

Table 11

	CLUST_01	CLUST_02	CLUST_03	CLUST_04	CLUST_05	CLUST_06	CLUST_07	CLUST_08	CLUST_09	CLUST_10	CLUST_11	CLUST_12	CLUST_13	CLUST_14	TOTAL
CLUST_01	15	1													16
CLUST_02		54													54
CLUST_03			27												27
CLUST_04				42											42
CLUST_05					34										34
CLUST_06						36									36
CLUST_07							29								29
CLUST_08								9							9
CLUST_09									23						23
CLUST_10										67					67
CLUST_11											35				35
CLUST_12												37			37
CLUST_13													7		7
CLUST_14														65	65
UNCLASS	1	2		1	2			1	1		4	1	2	1	16
TOTAL	16	57	27	43	36	36	29	10	24	67	39	38	9	66	497

Table 12

	CLUST_01	CLUST_02	CLUST_03	CLUST_04	CLUST_05	CLUST_06	CLUST_07	CLUST_08	CLUST_09	CLUST_10	CLUST_11	CLUST_12	CLUST_13	CLUST_14	TOTAL
CLUST_01	63														63
CLUST_02		120													120
CLUST_03			89												89
CLUST_04				88											88
CLUST_05					99										99
CLUST_06			1			111									112
CLUST_07							79								79
CLUST_08								61							61
CLUST_09									91						91
CLUST_10										172					172
CLUST_11											149				149
CLUST_12												106			106
CLUST_13													75		75
CLUST_14													1	119	120
UNCLASS	1	3	1		2		1		1	3	1		2		15
TOTAL	64	123	91	88	101	111	80	61	92	175	150	106	78	119	1439

Table 13

Table 14 below summarises the measures concerning the internal reliability of the four analysis reported in this section.

ANALYSIS	PRECISION	RECALL	F	CHOEN'S KAPPA
A1	0.9953	0.9505	0.9724	0.9953
A2	0.9986	0.9895	0.9940	0.9986
A3	0.9979	0.9677	0.9826	0.9979
A4	0.9986	0.9896	0.9941	0.9986

Table 14

So, according to the above results, we can conclude that - after having performed an unsupervised classification of 'n' textual units belonging to a 'C1' corpus, and after having stored 'its' criteria for classifying such textual units into 'k' groups - the top-down method implemented in T-LAB is able to classify the same 'n' textual units (as well as other textual units characterized by similar co-occurrence patterns) into the same 'k' groups even when the textual units are included in in a 'C2' corpus which is different from the 'C1' one (see 'internal' reliability as defined in section 2 above).

6 - Concluding remarks

While the 'qualitative' vs. 'quantitative' dichotomy risks to obscure the ways software like T-LAB can allow researchers to experiment new paths in thematic analysis of textual data, a better understanding of how the software works can lead to the clarification of several epistemological and methodological issues. In particular, such an understanding can lead the researcher to question the same notion of 'theme' and to frame any thematic analysis as a problem concerning 'pattern recognition', viz. as a problem concerning the analysis of word co-occurrence patterns. Consequently the uses of automated methods for textual analysis, as well as the relations between 'bottom-up' and 'top-down' approaches, should be carefully reconsidered.

Even if the external and internal reliability of many automated methods for textual analysis is unquestionable, by means of a couple of experiments, this paper offers a way to evaluate the specific performances of some T-LAB tools for thematic analysis.

REFERENCES

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

BOYATZIS, R.E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Thousand Oaks, CA: Sage.

BRAUN, V. & CLARK, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3 (2), pp. 77-101.

CANNY, J.F. (2006). *GaP: a Factor Model for Discrete Data.* ACM Conference on Information Retrieval (SIGIR 2004), PP. 122-129.

CANNY, J.F. & RATTEMBURY, T.L. (2006). *A Dynamic Topic Model for Document Segmentation.* Technical Report, Dept. of Electrical Eng. And Computer Sciences, Univ. of California at Berkeley.

FEREDAY, J. & MUIR- COCHRANE, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, *5*(1), Article 7. Retrieved on Sept. 15 2012 from http://www.ualberta.ca/~iiqm/backissues/5_1/html/fereday.htm

GUEST, G.S., MACQUEEN, K. M. & NAMEY, E.E. (2012). *Applied Thematic Analysis*. Thousand Oaks, CA: Sage

HAYES, N. (1997). Qualitative research and research in psychology. In N. Hayes (Ed.), *Doing Qualitative Analysis in Psychology*. Hove, UK: Psychology Press.

KRIPPENDORF, K. (2004). *Content Analysis. An Introduction to Its Methodology* (2nd ed.).Thousand Oaks, CA: Sage:

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). *The Logic of T-LAB Tools Explained*. Retrieved on October 26 2012, from the T-LAB website: http://www.tlab.it/en/toolsexplained.php.

LANDIS, J.R. & KOCH, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, pp. 159-174.

MORSE, J., & MITCHAM, C. (2002). Exploring qualitatively derived concepts: Inductive- deductive methods. *International Journal of Qualitative Methods*, *I*(4), Article 3. Retrieved from http://ejournals.library.ualberta.ca/index.php/IJQM/article/view/4589

PATTON, M. Q. (2002). *Qualitative Research & Evaluation Methods* (3 ed.).Thousand Oaks, CA: Sage:

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.

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

VAN RIJSBERGEN, C. J. K. (1979). *Information Retrieval* (2 ed.). London, GB; Boston, MA: Butterworth.