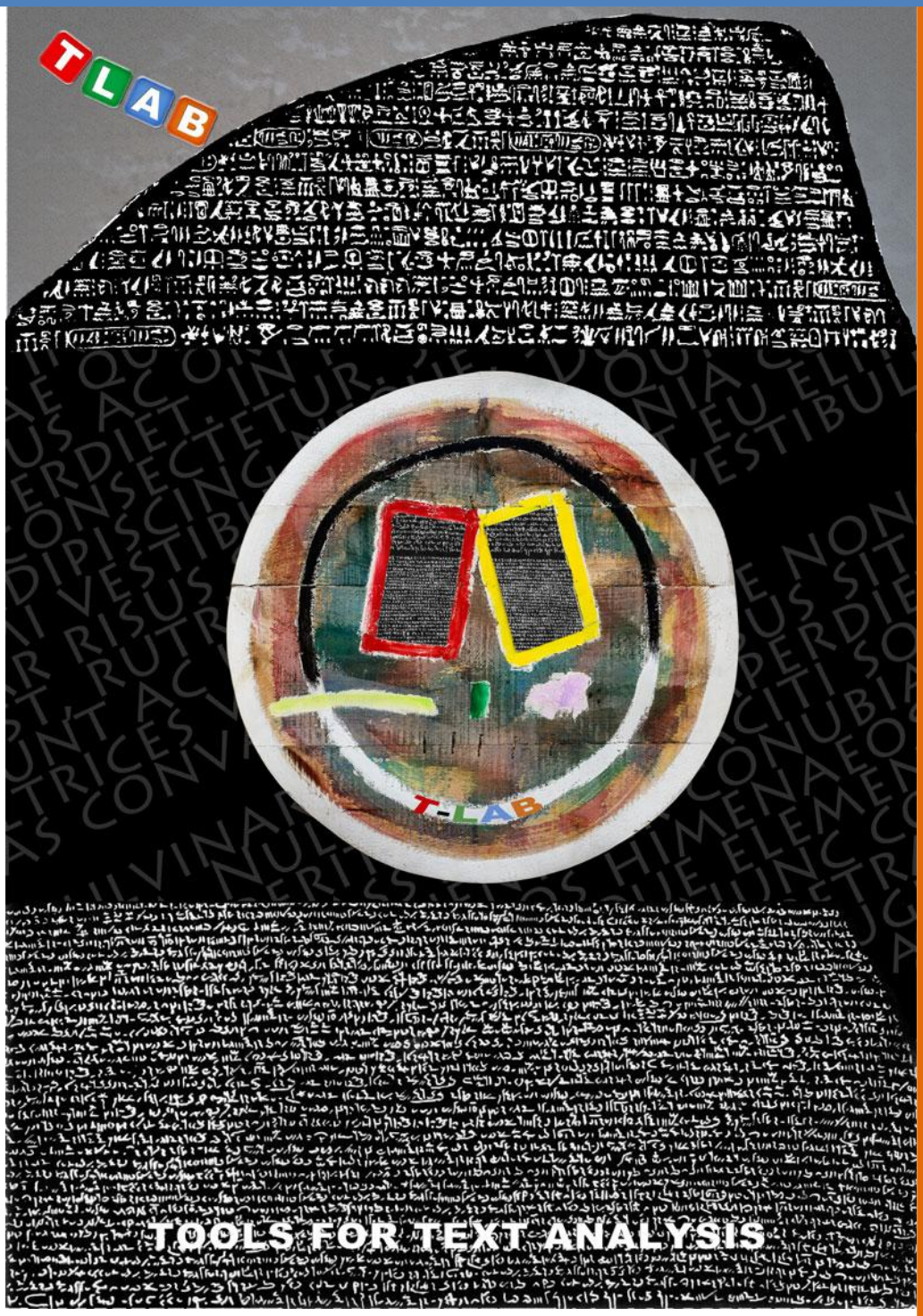
T-LAB 10

User's Manual



The above artwork has been realized for T-LAB by Claudio Marini (http://www.claudiomarini.it/) in collaboration with Andrea D'Andrea.

Tools for Text Analysis

Copyright © 2001-2023
T-LAB by Franco Lancia
All rights reserved.

Website: https://www.tlab.it/
E-mail: info@tlab.it

T-LAB is a registered trademark

CONTENTS

Installation and system requirements 2
What T-LAB does and what it enables us to do 3
SETTINGS 34
Automatic and Customized Settings 35
Dictionary Building 41
CO-OCCURRENCE ANALYSIS 44
Word Associations 45
Co-Word Analysis and Concept Mapping 54
Comparison between Pairs of Key Words 65
Sequence and Network Analysis 71
Concordances 84
THEMATIC ANALYSIS 87
Thematic Analysis of Elementary Contexts 88
Modeling of Emerging Themes 110
Thematic Document Classification 123
Dictionary-Based Classification 127
Texts and Discourses as Dynamic Systems 142
COMPARATIVE ANALYSIS 160
Specificity Analysis 161
Correspondence Analysis 170
Multiple Correspondence Analysis 178
Cluster Analysis 180
Singular Value Decomposition (SVD) 187
CORPUS PREPARATION 191
Corpus Preparation 192
Structural Criteria 193
Formal Criteria 194
FILE 196
Import a single file… 197
Prepare a Corpus (Corpus Builder) 202
Open an existing project 212
LEXICAL TOOLS 213
Text Screening / Disambiguations 214
Corpus Vocabulary 217
Stop-Word list 219
Multi-Word list 221
Word Segmentation 223
OTHER TOOLS 225
Variable Manager 226
Advanced Corpus Search 229
Classification of New Documents 231
Key Contexts of Thematic Words 233
Export Custom Tables 237
Editor 241
Import-Export Identifiers List 243

GLOSSARY 245
Analysis Unit 246
Association Indexes 247
Chi-Square 249
Cluster Analysis 250
Coding 251
Context Unit 251
Corpus and Subsets 252
Correspondence Analysis 254
Data Tables 255
Disambiguation 256
Dictionary 257
Elementary Context 258
Frequency Threshold 260
Graph Maker 261
Homographs 263
IDnumber 264
Isotopy 265
Key-Terms 265
Lemmatization 266
Lexical Unit 267
Lexie and Lexicalization 267
Markov Chain 268
MDS 269
Multiwords 270
N-Grams 271
Naïve Bayes Classifier 272
Normalization 273
Occurrences and Co-occurrences 274
Poles of Factors 276
Primary Document 277
Profile 277
Specificity 278
Stop Word List 279
Test Value 280
TF-IDF 281
Thematic Nucleus 282
Variables and Categories 282
Words and Lemmas 283
BIBLIOGRAPHY 284

Installation and system requirements

Minimum configuration required:

- Windows 7 or later
- 4 Gb RAM
- Full HD screen resolution (1920 x 1080 recommended)

Installation:

- Double click on Setup.exe
- Follow the instructions on the screen
- Exit from the program
- Wait for a reply from info@tlab.it to get your activation key
- For more info, see https://www.mytlab.com/T-LAB_Installation.pdf

What T-LAB does and what it enables us to do

T-LAB software is an all-in-one set of **linguistic**, **statistical and graphical tools for text analysis** which can be used in research fields like Content Analysis, Sentiment Analysis, Semantic Analysis, Thematic Analysis, Text Mining, Perceptual Mapping, Discourse Analysis, Network Text Analysis, Document Clustering, Text Summarization.

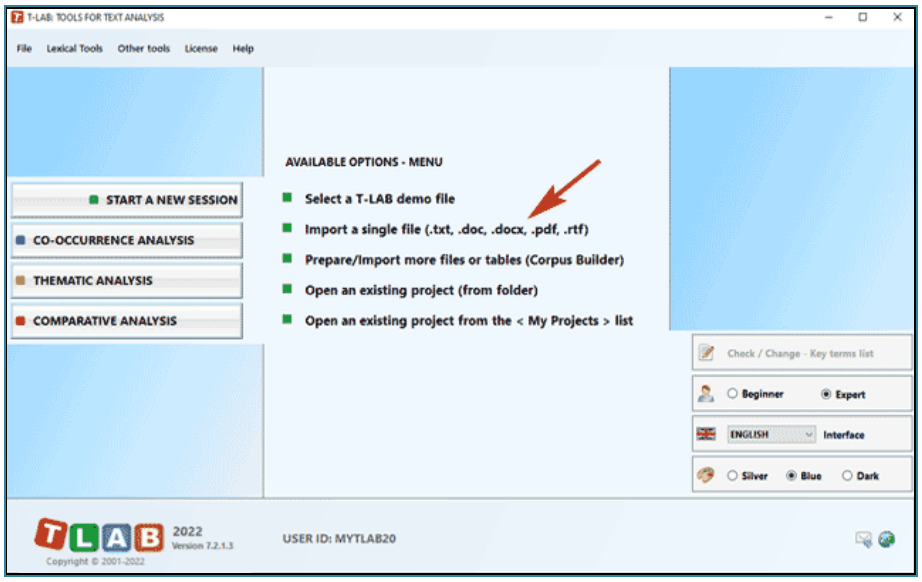
In fact **T-LAB** tools allow the user to easily manage tasks like the following:

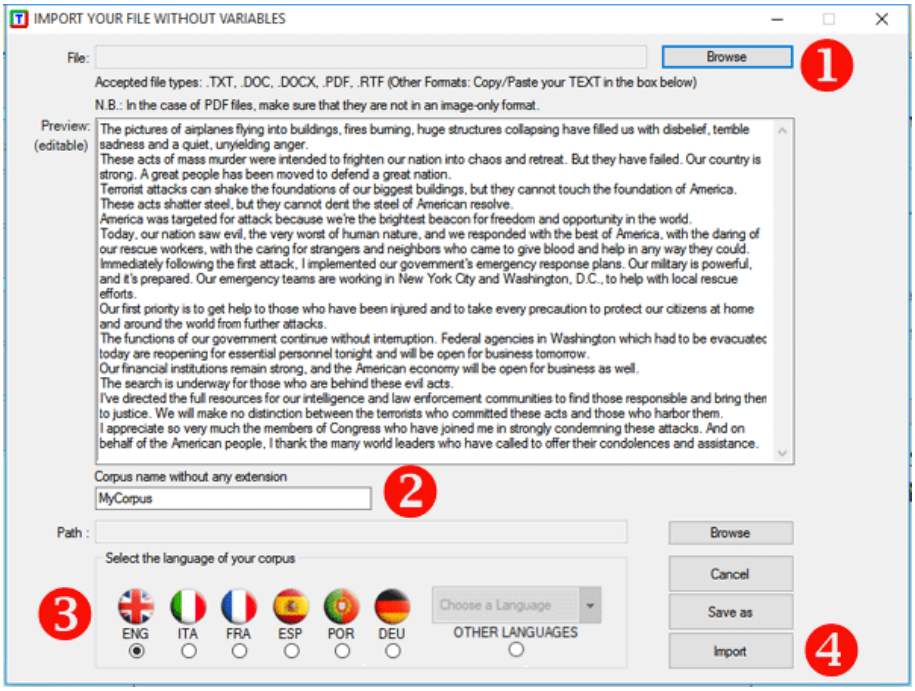
- measure, explore and map the **co-occurrence relationships** between key-terms;
- perform either unsupervised or supervised clustering of textual units and documents, i.e. perform a **bottom-up clustering** which highlights **emerging themes** or a perform **top-down classification** which uses a set 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);
- represent and explore any text as a **network**;
- obtain measures and graphical representations concerning **texts and discourses** treated as **dynamic systems**;
- customize and apply various types of **dictionaries** for both lexical and content analysis;
- perform **concordance** searches;
- analyse all the **corpus** or its **subsets** (e.g. groups of documents) by using various key-term lists;
- create, explore and export numerous **contingency tables** and **co-occurrences matrices**.

The **T-LAB** user interface is very **user-friendly** and various types of texts can be analysed:
- a single text (e.g. an interview, a book, etc.);
- a set of texts (e.g. a set of interviews, web pages, newspaper articles, responses to open-ended questions, Twitter messages, etc.).

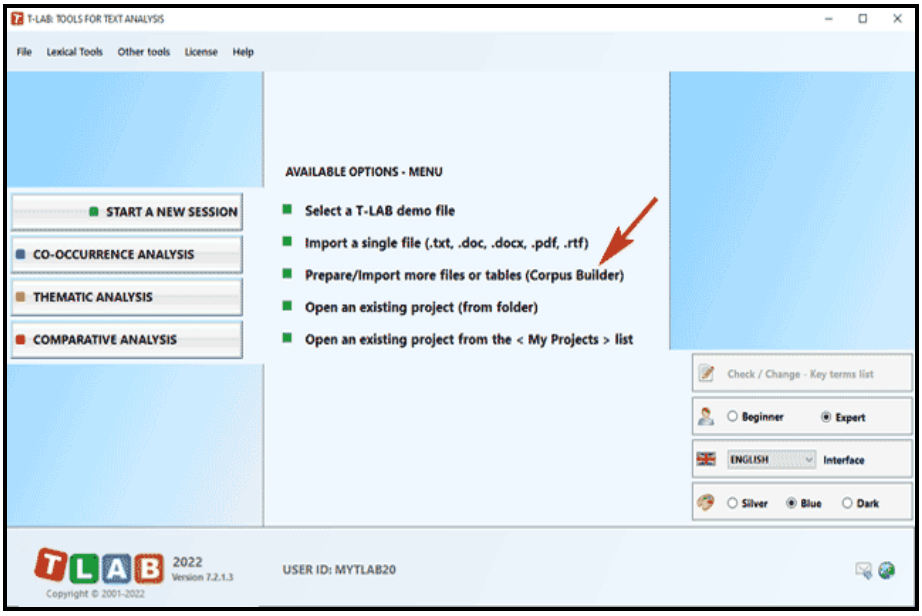
All texts can be encoded with categorical **variables** and/or with **Unique Identifiers** that correspond to context units or cases (e.g. responses to open-ended questions).

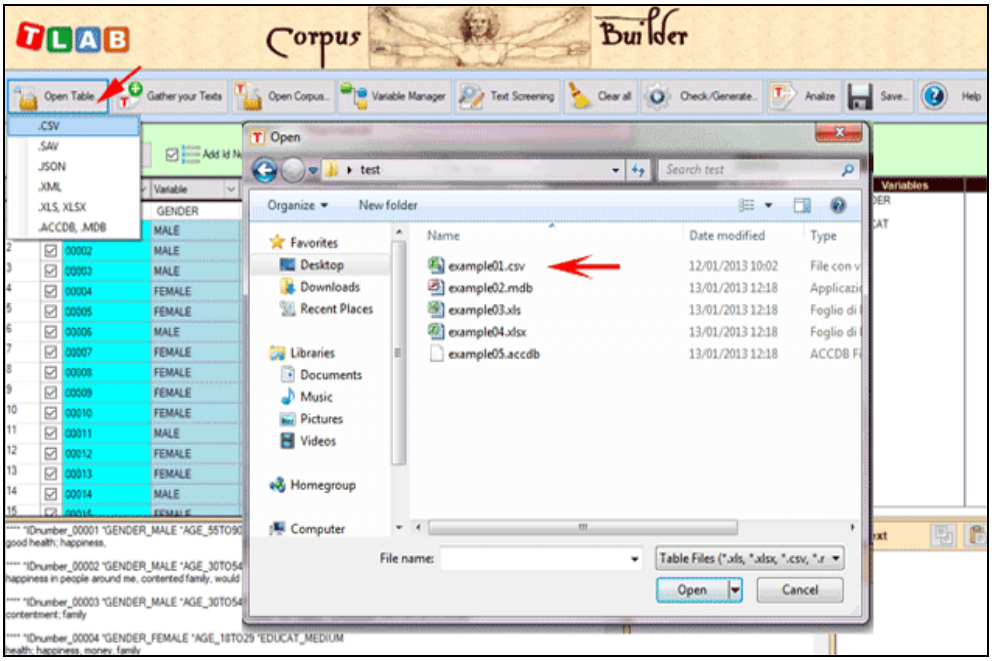
In the case of a **single document** (or a corpus considered as a single text) **T-LAB** needs no further work: just select the 'Import a single file…' option (see below) and proceed as follows.

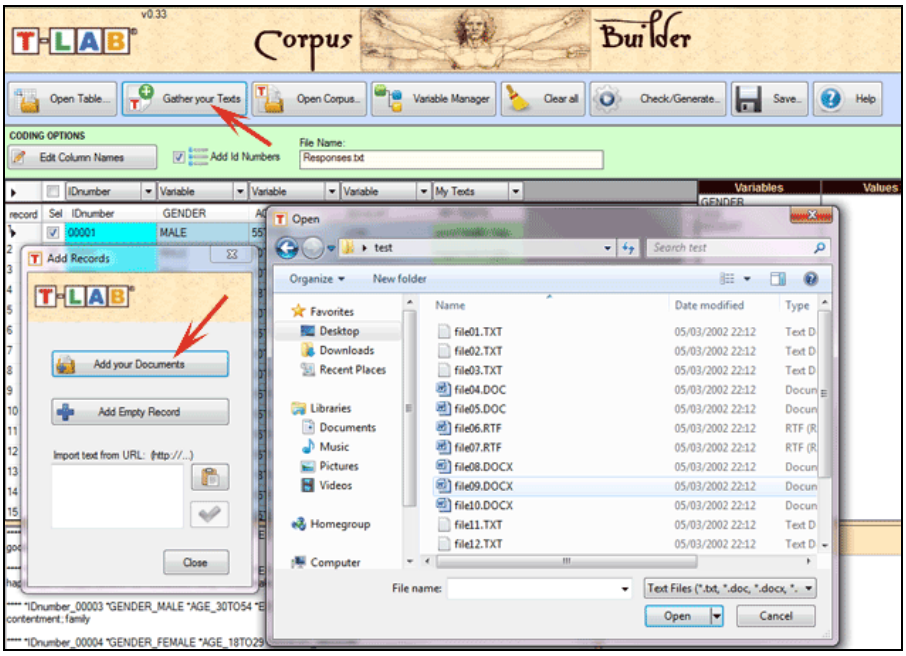




When, on the other hand, the corpus is made up of various texts and/or categorical variables are used the **Corpus Builder** tool (see below) must be used. In fact, such a tool automatically transforms any textual material and various types of files (i.e. up to ten different formats) into a corpus file ready to be imported by **T-LAB.**



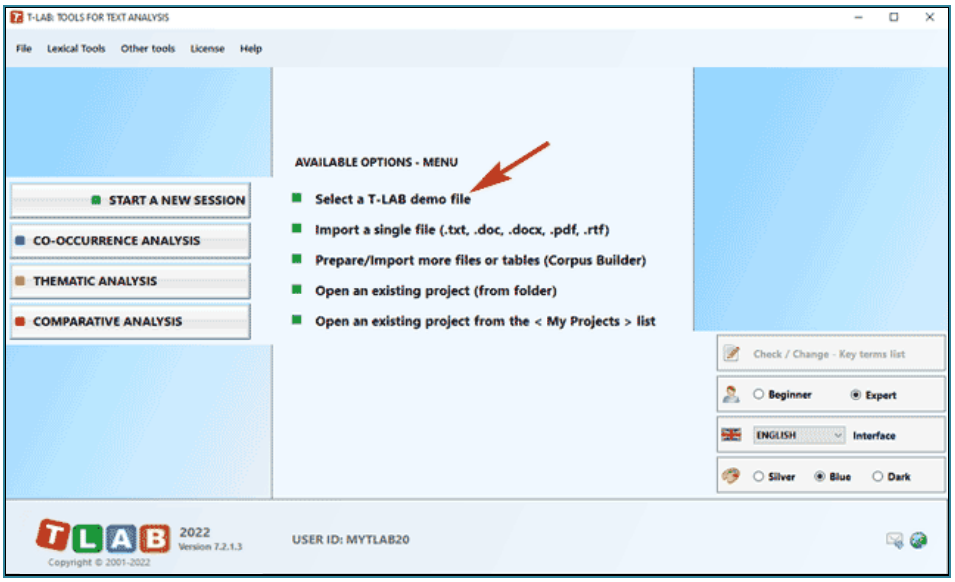




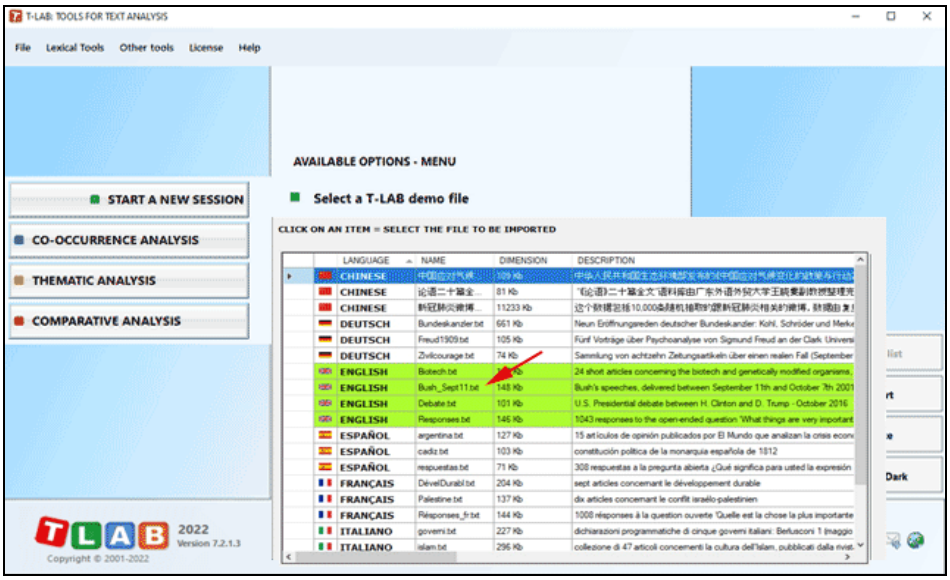
N.B.: At the moment, in order to ensure the integrated use of various tools, each corpus file shouldn't exceed 90 Mb (i.e. about 55,000 pages in .txt format). For more information, see the 'Requirements and Performances' section of the Help/Manual.

Six steps are that is required to perform a quick check of the software functionalities:

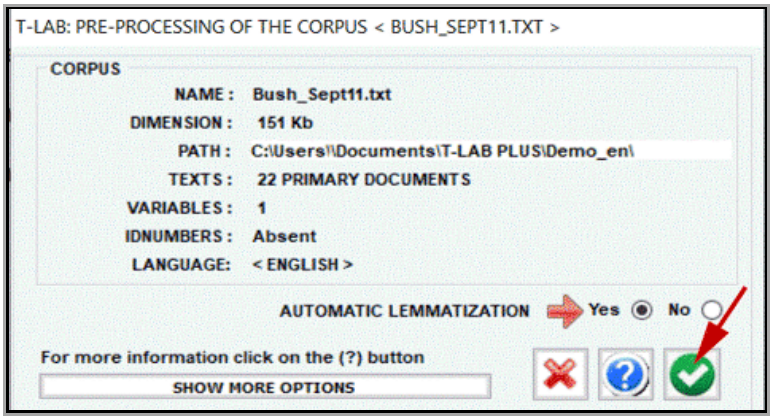
1 – Click on the 'Select a T-LAB demo File' option



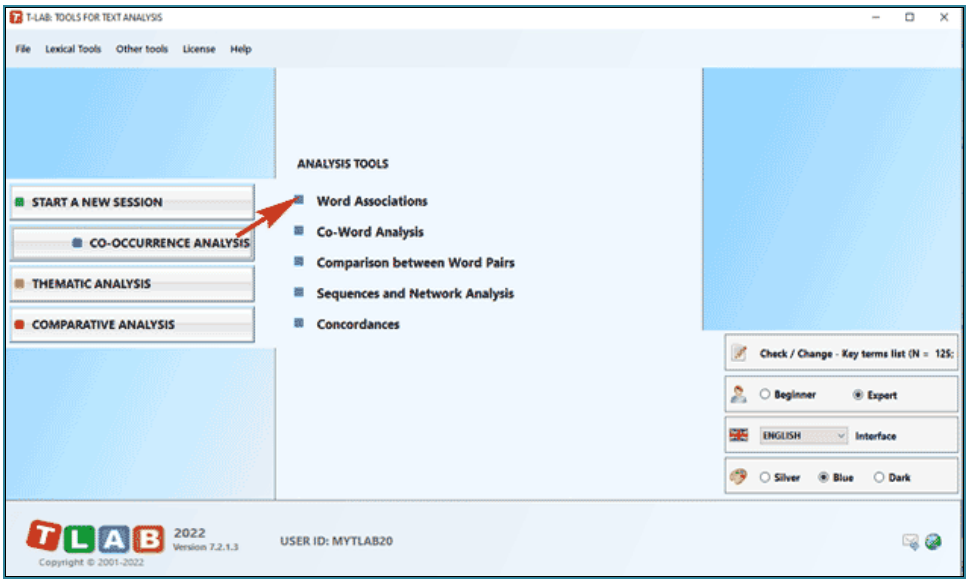
2 - Select any corpus to analyse



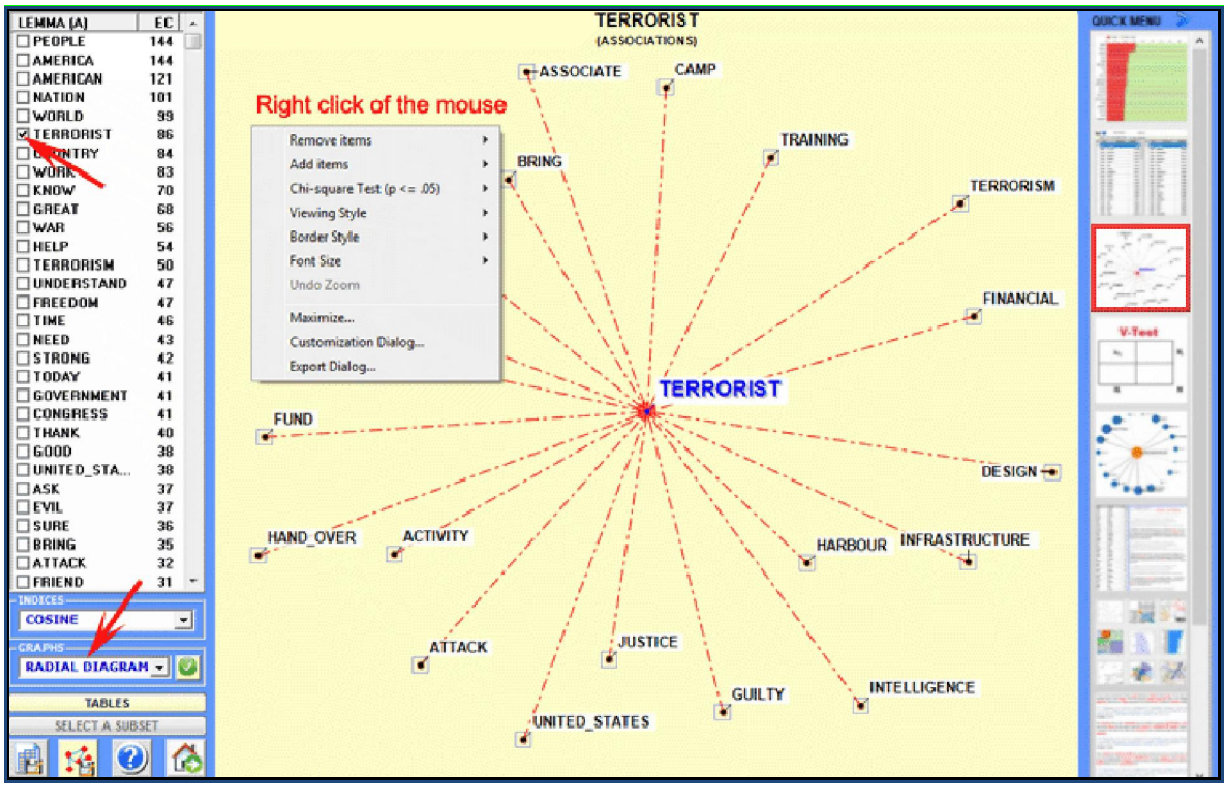
3 - Click "ok" in the first Setup window

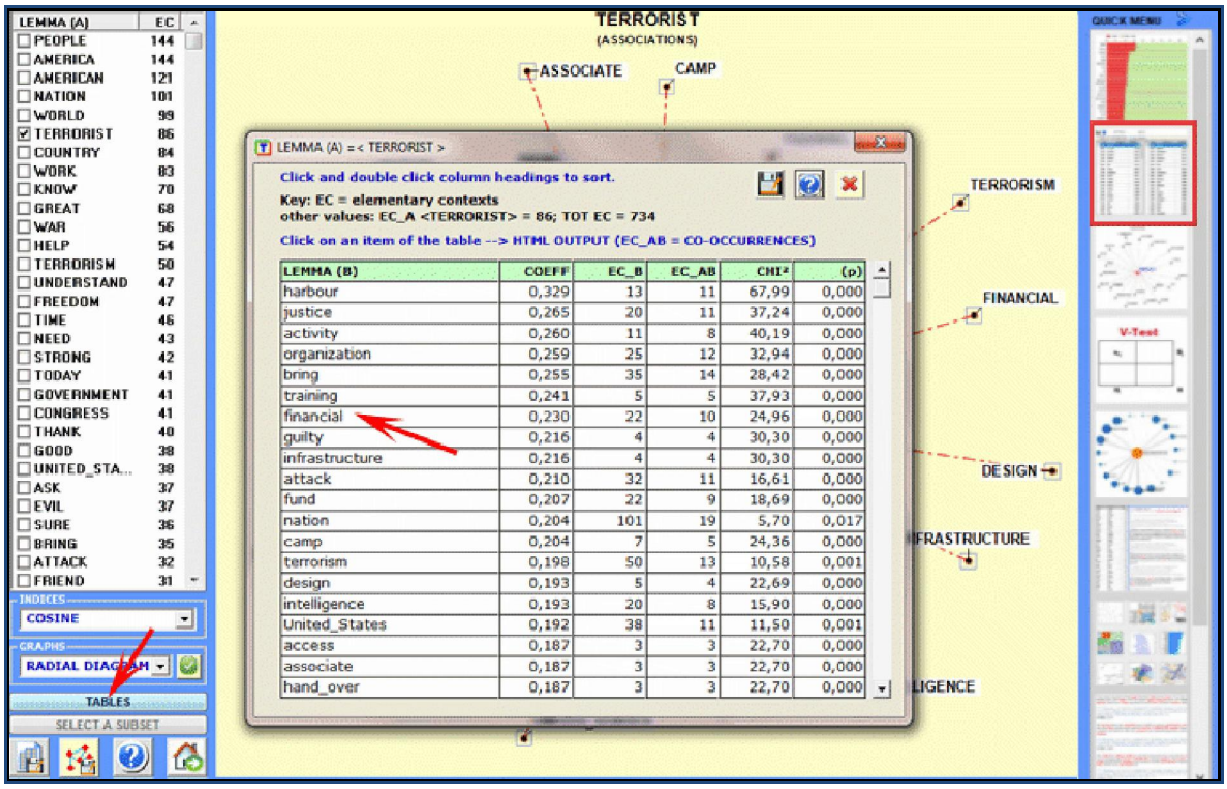


4 - Select a tool from one of the "Analysis" sub-menus

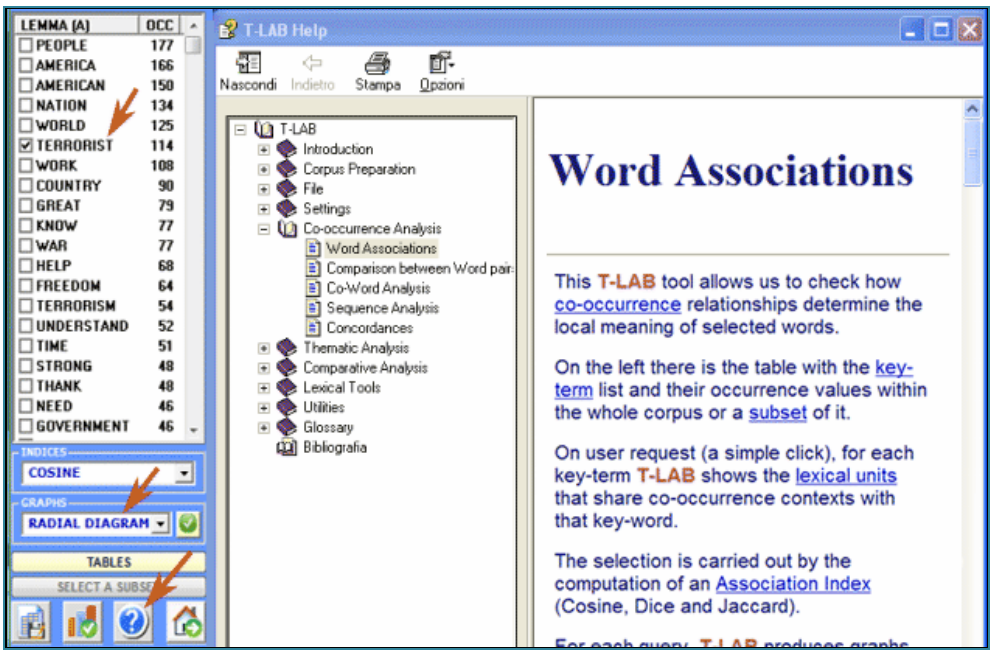


5 - Check the results





6 - Use the contextual help function to interpret the various graphs and tables



The following information is provided to help the user to better understand what **T-LAB** does and how to make full use of it.

Apart from the user interface, the **T-LAB** system is organized into two main components:

- the **database**, the "place" where the input **corpus** (the text or the set of texts to be analysed) is represented as a set of **tables** in which the **analysis units**, their characteristics and their mutual relationships are recorded.
- the **algorithms**, which are subsets of **instructions** that allow us to use the interface, to consult and modify the database, to produce further tables with the available data, to perform **statistical computations** and to produce **outputs** that represent the relationships between the analysed data.

To understand how **T-LAB** works and how it can be used, it is essential to have a clear idea as to which **analysis units** are filed in its database and what statistical **algorithms** are used in the various analyses. In fact, the analysed data tables always consist of rows and columns the headings of which correspond to the analysis units filed in the database, while the algorithms regulate the processes that make it possible to detect significant relationships between the data and to extract useful information.

The **analysis units** used in **T-LAB** are of two types: **lexical units** and **context units**.

A - the **lexical units** are words and multi-words, filed and classified on the basis of a criterion. More precisely, in the **T-LAB** database each lexical unit consists of a classified record with two fields: **word** and **lemma**. In the first field ("**word**"), the words are listed as they appear in the corpus, while in the second ("**lemma**") the **labels** attributed to groups of lexical units are listed and classified according to linguistic criteria (e.g. **lemmatization**) or by dictionaries and semantic grids defined by the user.

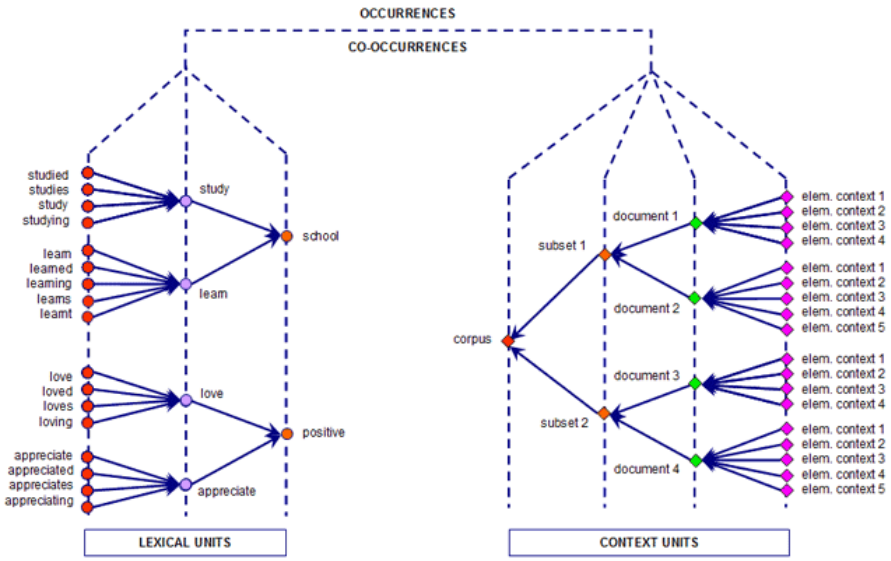
B - the **context units** are portions of text that the corpus can be divided into. More precisely, according to **T-LAB** logic, there can be three types of context units:

B.1 primary documents, which correspond to the "natural" subdivision of the corpus (e.g. interviews, articles, answers to open-ended questions, etc.), that is the **initial context** defined by the user;

B.2 elementary contexts, which correspond to syntagmatic units (i.e. fragments, sentences, paragraphs) in which each primary document can be subdivided;

B.3 corpus subsets, which correspond to groups of primary documents which lead to the same category (e.g. interviews with "men" or "women", articles in a specific year or a particular magazine and so on), including **thematic clusters** of documents or elementary contexts obtained by using the corresponding **T-LAB** tools.

The picture below illustrates the possible relationships between lexical and context units which **T-LAB**, through statistical and graphical tools (see section 5 below), allows us to analyse.

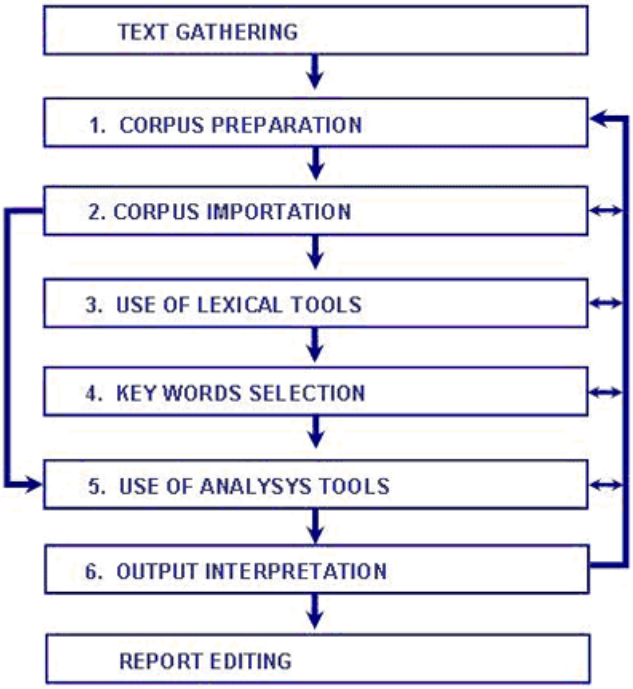


Starting from this database organization, **T-LAB** makes it possible - in automatic mode - to explore and to analyse the relationships between the analysis units of the whole **corpus** or its **subsets.**

In **T-LAB**, the selection of any analysis tool (click of the mouse) always activates a semi-automatic process that, with a few simple operations, generates an input table, applies some statistical algorithms and produces some outputs.

Let's consider how a typical work **project** which uses **T-LAB** can be managed. Hypothetically, each project consists of a set of analytical activities (operations) which have the same **corpus** as their subject and are organized according to the user's **strategy** and **plan**. It then begins **gathering the texts** to be analysed, and concludes with a **report**.

The succession of the various phases is illustrated in the following diagram:



N.B.:
- The six numbered phases, from the corpus preparation to the interpretation of the outputs, are supported by **T-LAB** tools and are always reversible;

- By using **T-LAB automatic settings** it is possible to avoid two phases (3 and 4); however, in order to achieve high quality results, their use is, nevertheless, advisable.

Now let's try to comment on the various steps.

1 - CORPUS PREPARATION: transformation of the texts to be analysed in a file (**corpus**) that can be processed by the software.

In the case of a single text (or a corpus considered as a single text) **T-LAB** needs no further work. When, on the other hand, the corpus is made up of various texts and/or categorical variables are used the Corpus Builder tool must be used, which automatically transforms any textual material and various types of files (i.e. up to eleven different formats) into a corpus file ready to be imported by **T-LAB**.

N.B.:

- At the end of the corpus preparation phase it is recommended that a new folder be created which contains only the corpus to be imported;

- When analysing any corpus, it is recommended that the working files (i.e. the working folder of the corpus) reside on a hard disk of the computer where **T-LAB** is installed. Otherwise, the various procedures could slow down and the software may report errors.

2 - CORPUS IMPORTATION: a series of **automatic processes** that transform the corpus into a set of tables integrated in the **T-LAB database**.

During the pre-processing phase, **T-LAB** carries out the following treatments: Corpus **Normalization**; **Multi-Word** and **Stop-Word** detection; **Elementary Context** segmentation; Automatic **Lemmatization** or **Stemming**; **Vocabulary** building; **Key-Terms s**election.

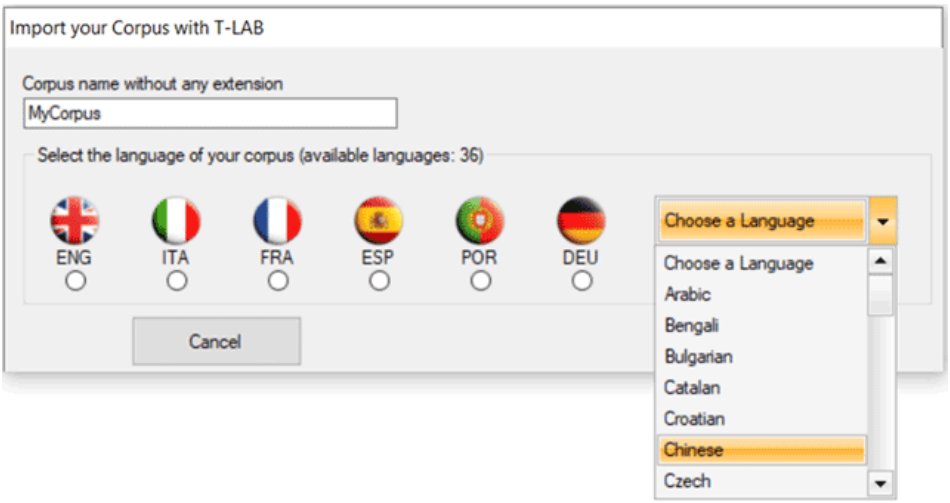
Here is the complete list of the languages for which specific pre-processing options are available.

LEMMATIZATION: Catalan, Croatian, English, French, German, Italian, Latin, Polish, Portuguese, Romanian, Russian, Serbian, Slovak , Spanish, Swedish, Ukrainian.

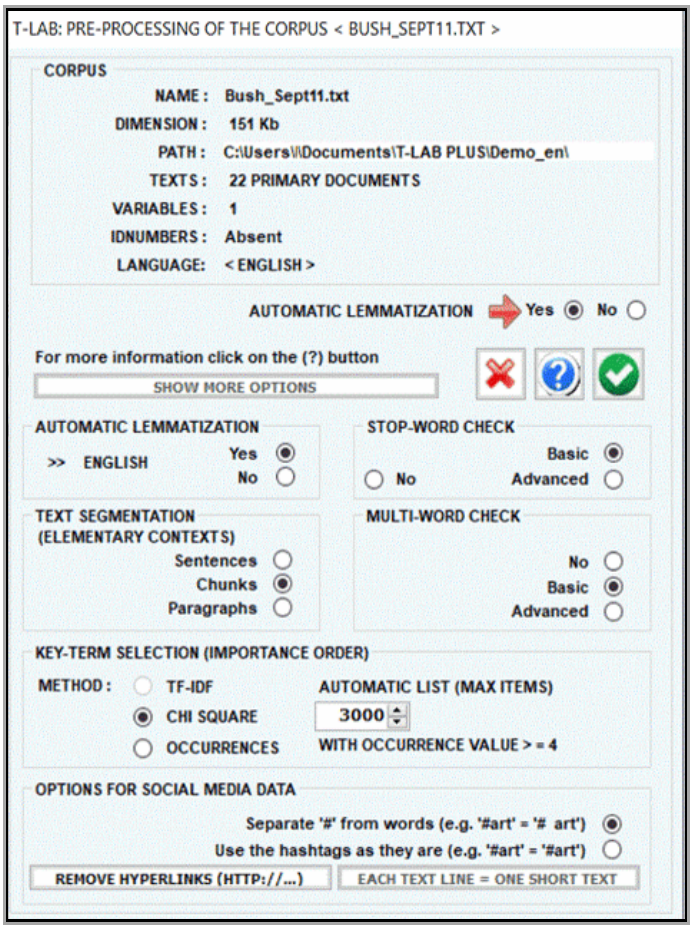
STEMMING: Arabic, Bengali, Bulgarian, Czech, Danish, Dutch, Finnish, Greek, Hindi, Hungarian, Indonesian, Marathi, Norwegian, Persian, Turkish.

WORD SEGMENTATION: Chinese and Japanese.

In any case, without automatic lemmatization and / or by using customized dictionaries the user can analyse texts in **all languages**, provided that words are separated by spaces and / or punctuation.



The setup form in which the user can select the pre-processing options which fit his needs is the following:

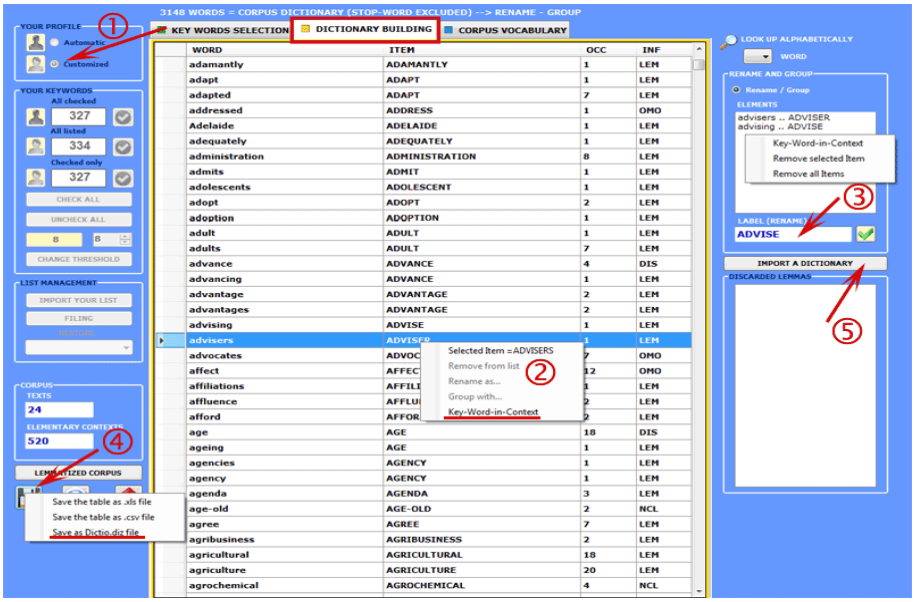


N.B.: As the pre-processing options determine both the kind and the number of analysis units (i.e. context units and lexical units), different choices determine different analysis results. For this reason, all **T-LAB** outputs (i.e. charts and tables) shown in the user's manual and in the on-line help are just indicative.

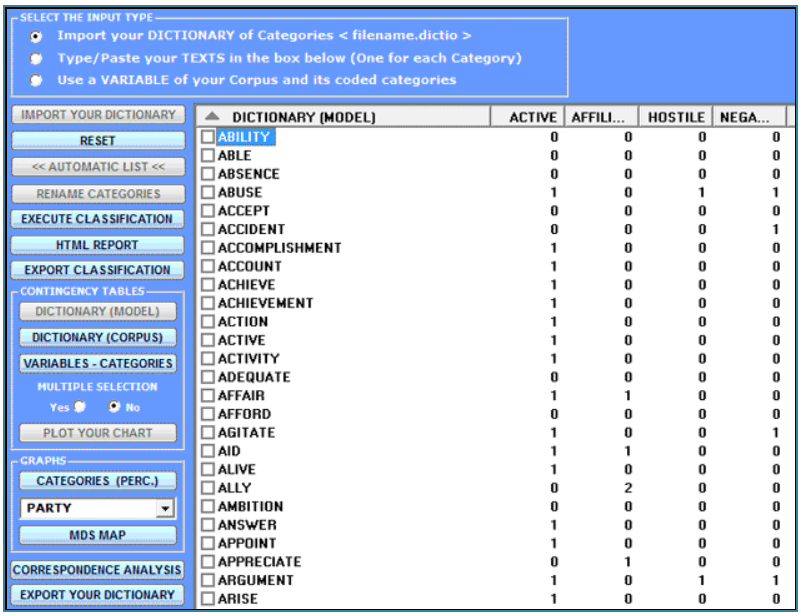
3 - THE USE OF LEXICAL TOOLS allows us to verify the correct **recognition** of the lexical units and to customize their **classification**, that is to verify and to modify the automatic choices made by **T-LAB.**

The procedures of the various interventions are illustrated in the corresponding help sections (and in the manual).

In particular the user is requested to refer to the corresponding help section for a detailed description of the **Dictionary Building** process (see below). In fact any change concerning the dictionary entries affects both the occurrence and the co-occurrence computation.

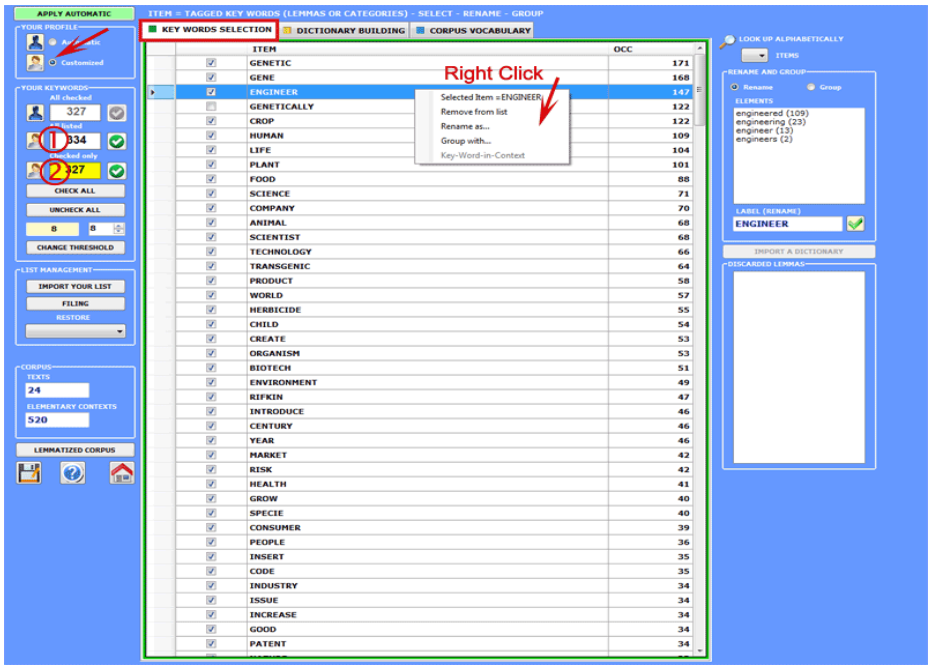


N.B.: When the user, without losing any lexical information, intends to apply coding schemes which group words or lemmas in a few categories (i.e. from 2 to 50) it is advisable to work with the **Dictionary-Based Classification** tool included in the **Thematic Analysis** sub-menu (see below).



4 - THE KEY-WORD SELECTION consists of the arrangement of one or more lists of lexical units (words, lemmas or categories) to be used for producing the data tables to be analysed.

The **automatic settings** option provides the lists of the **key-words** selected by **T-LAB**; nevertheless, since the choice of the analysis units is extremely relevant in relation to subsequent elaborations, the use of **customized settings** (see below) is highly recommended. In this way the user can choose to modify the list suggested by **T-LAB** and/or to arrange lists that better correspond to the objectives of his research.



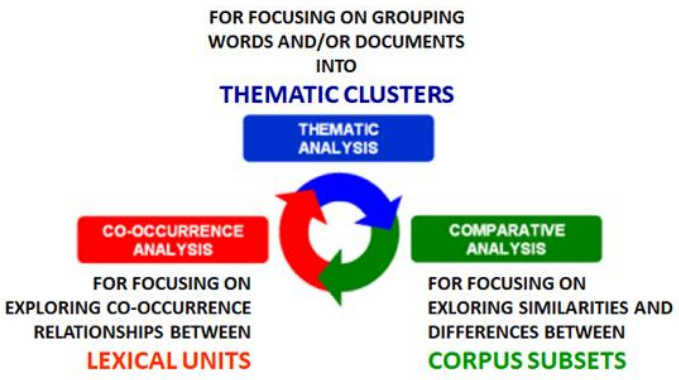
In any case, while creating these lists, the user can refer to the following criteria:

- check the quantitative (total of the occurrences) and qualitative **importance** of the various items;
- check the **limitations** of the analytical tools that you intend to use (see at the end of this chapter);
- check whether the set of items is compatible with your own research **strategies** (see item : 5 to follow).

5 - THE USE OF ANALYSIS TOOLS allows the user to obtain outputs (tables and graphs) that represent **significant relationships** between the analysis units and enables the user to make **inferences**.

At the moment, **T-LAB** includes twenty different analysis tools each of them having its own specific logic; that is, each one generates specific tables, uses specific algorithms and produces specific outputs.

Consequently, depending on the structure of texts to be analysed and on the goals to be achieved, the user has to decide which tools are more appropriate for their analysis strategy every time.

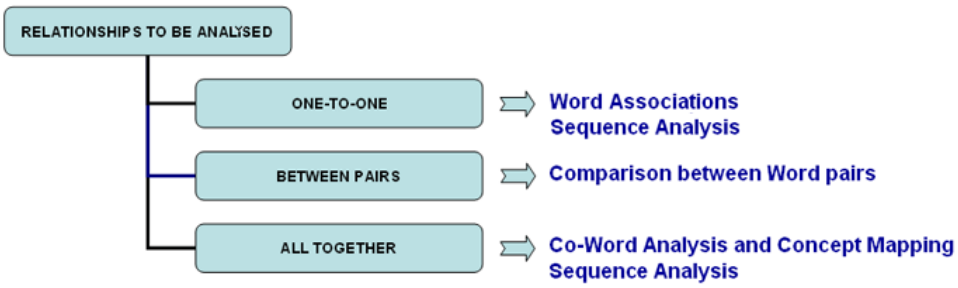


N.B.: Besides the distinction between tools for **co-occurrence**, **comparative** and **thematic** analysis, it can be useful to consider that some of the latter allow us to obtain new corpus subsets which can be included in further analysis steps.

Even though the various **T-LAB** tools can be used in any order, there are nevertheless three ideal starting points in the system which correspond to the three ANALYSIS sub-menus:

A : TOOLS FOR CO-OCCURRENCE ANALYSIS

These tools enable us to analyse different kinds of relationships between lexical units (i.e. words or lemmas)

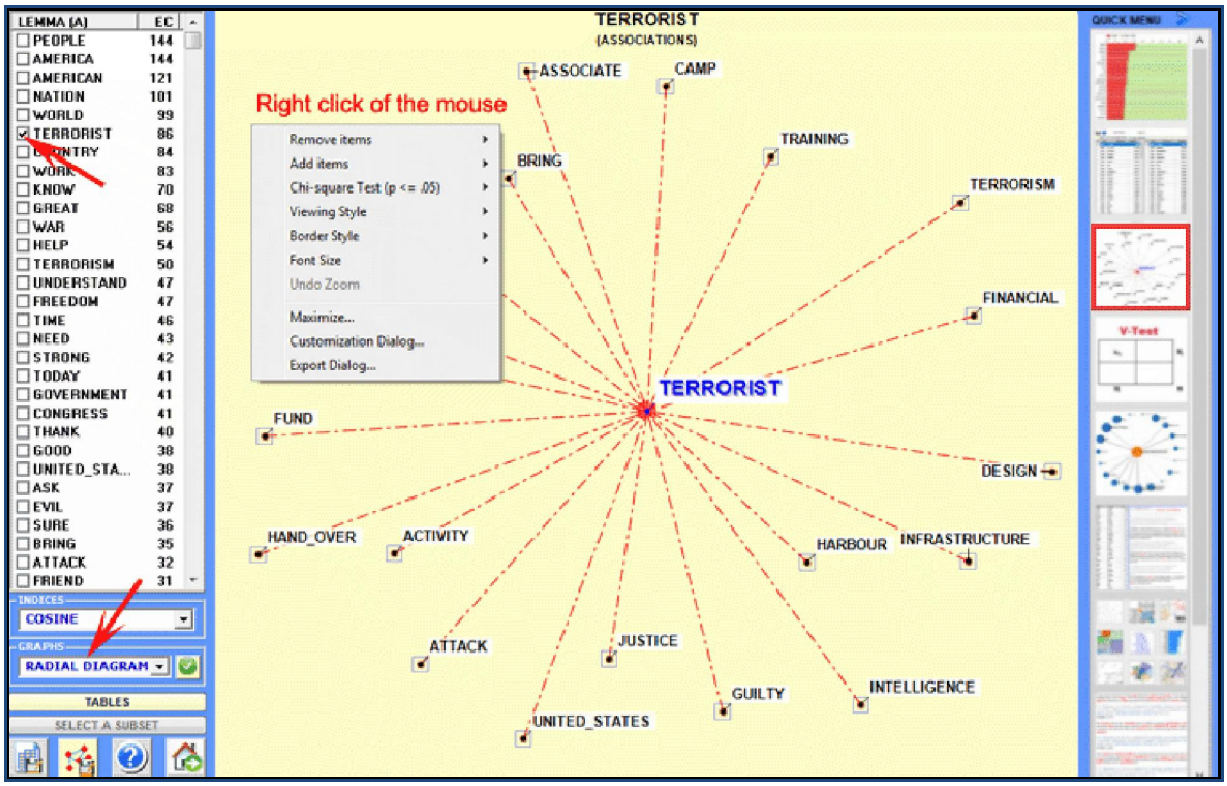


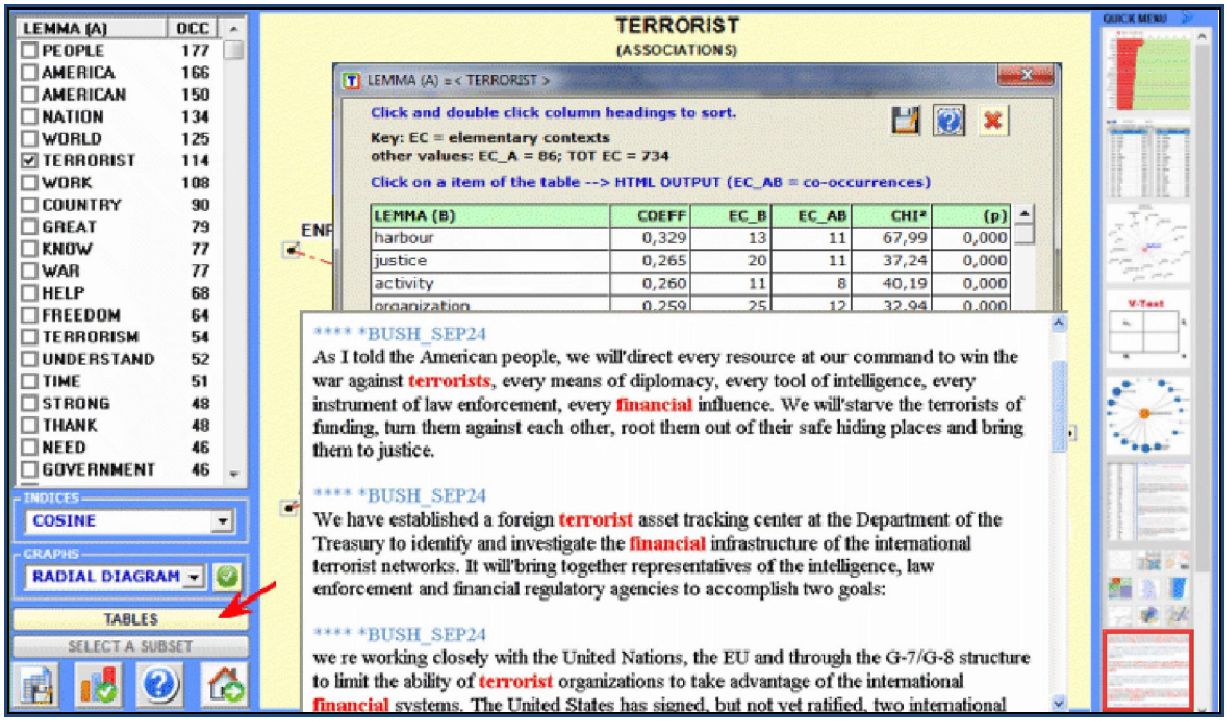
According to the types of relationships to be analysed, the **T-LAB** options indicated in this diagram use one or more of the following statistical tools: **Association Indexes**, **Chi Square Tests**, **Cluster Analysis**, **Multidimensional Scaling**, **Principal Component Analysis, t-SNE** and **Markov chains**.

Here are some examples (N.B.: for more information on how to interpret the outputs please refer to the corresponding sections of the help/manual).

- Word Associations

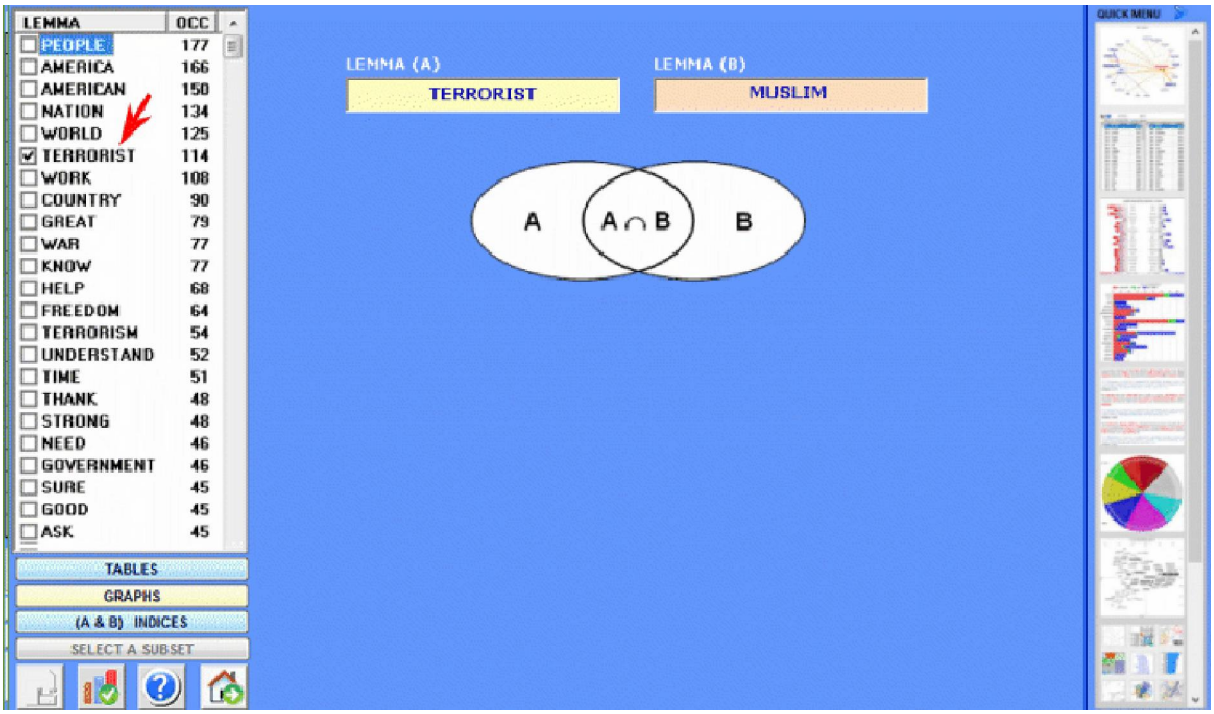
This **T-LAB** tool allows us to check how **co-occurrence** relationships determine the **local meaning** of selected words.

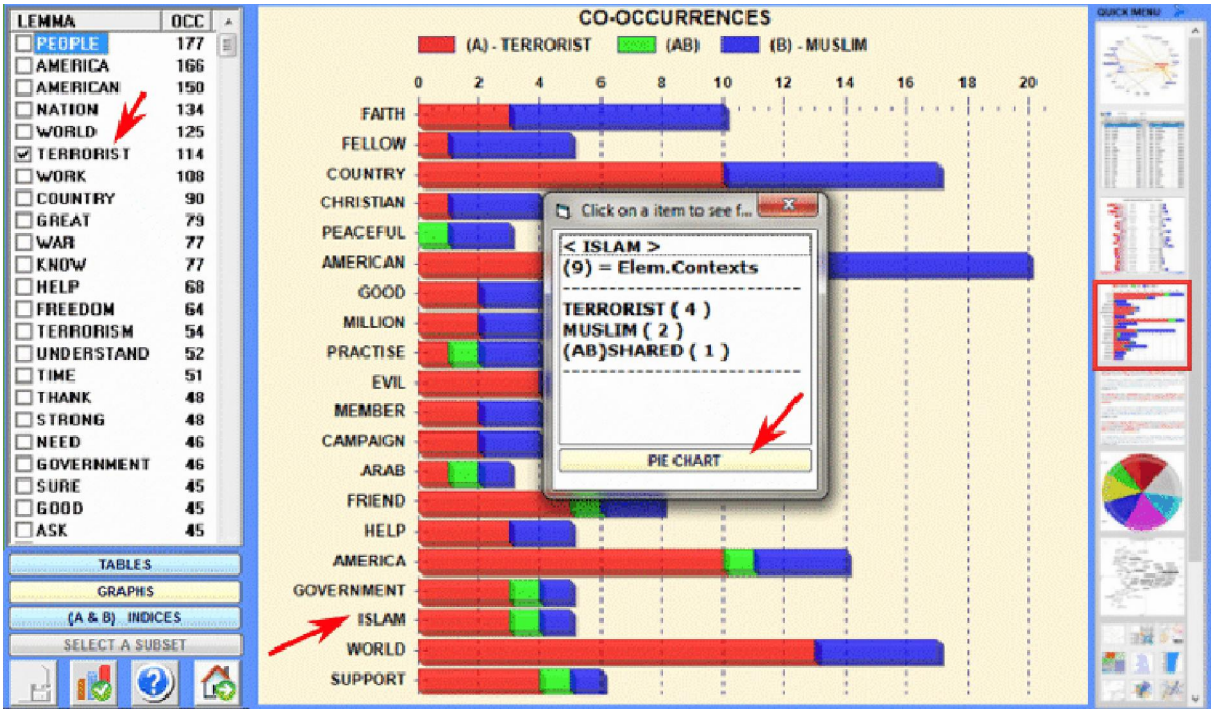




- Comparison between Word Pairs

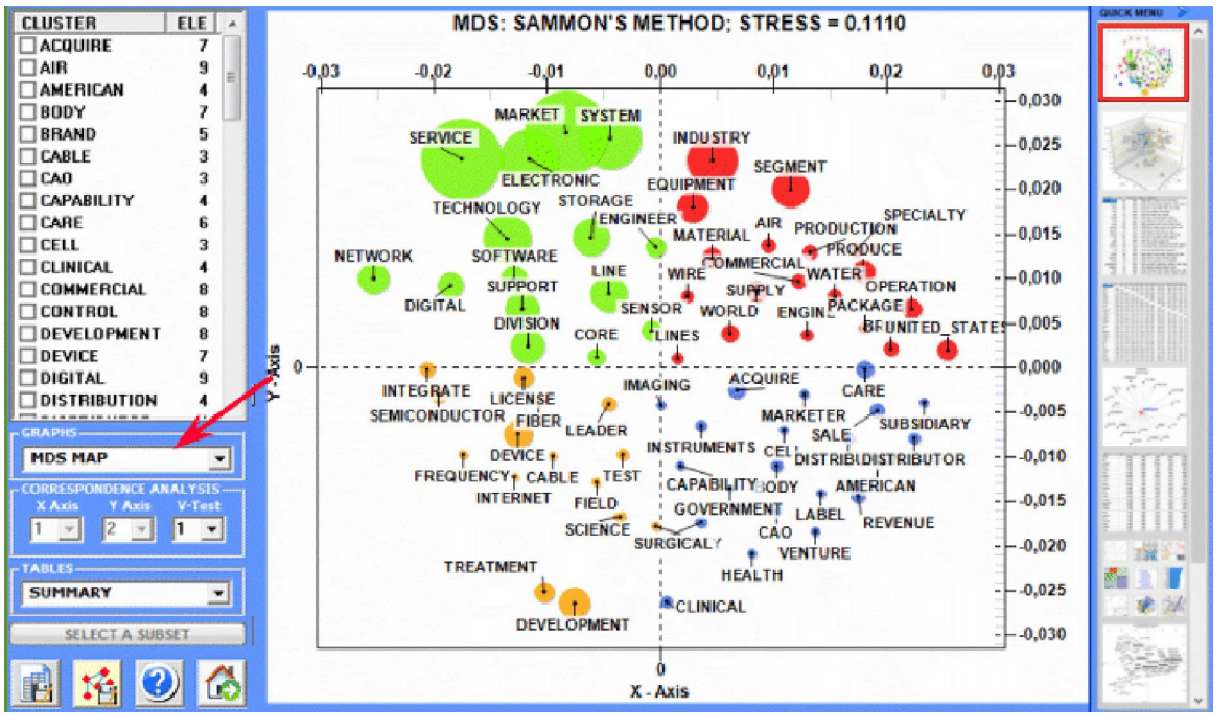
This **T-LAB** tool allows us to compare sets of elementary contexts (i.e. co-occurrence contexts) in which the elements of a pair of key-words are present.

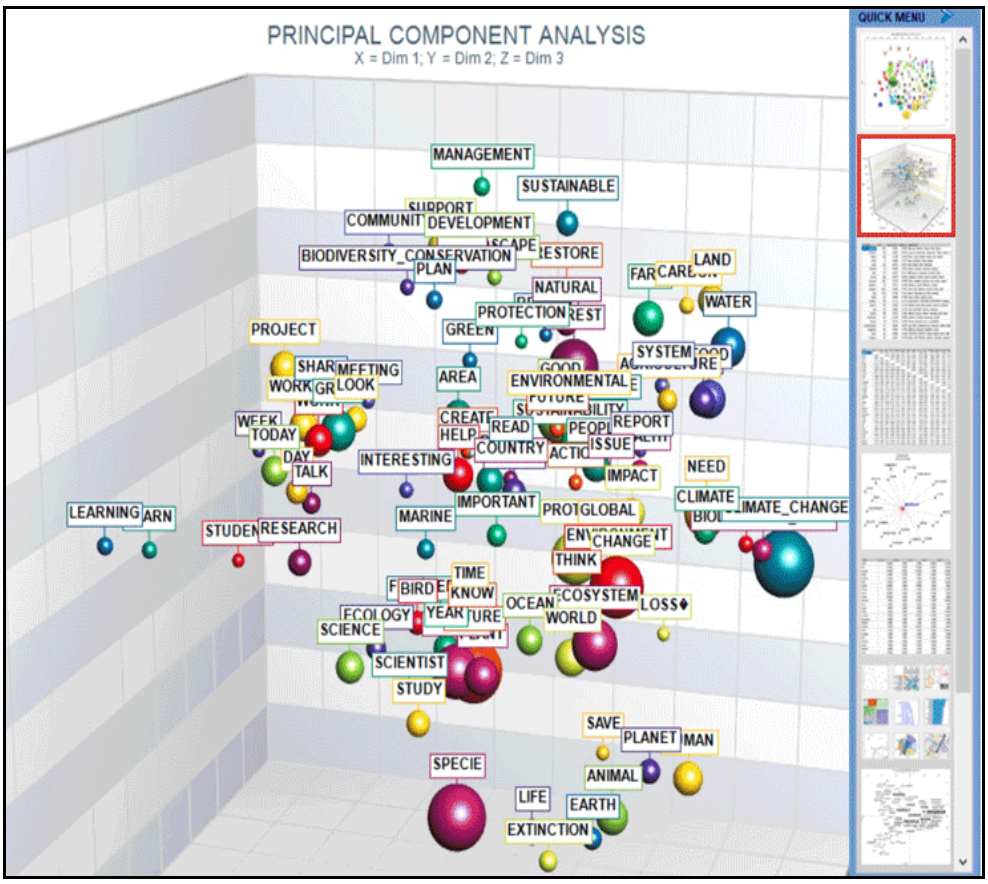




- Co-Word Analysis

This **T-LAB** tool allows us to find and map co-occurrence relationships within (and between) **sets** of key-words.



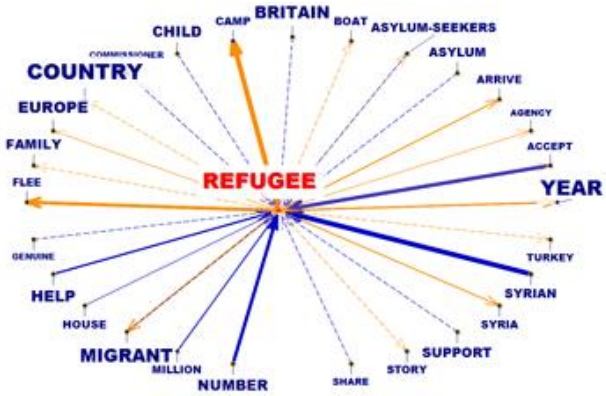


- Sequence and Network Analysis

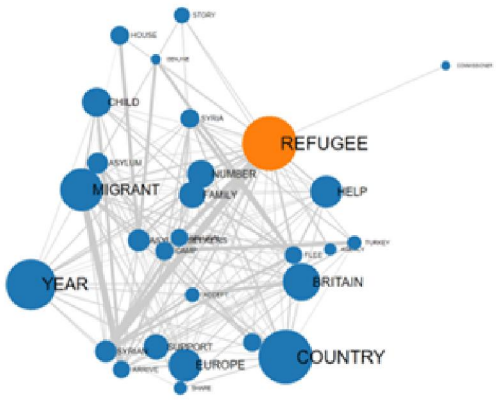
This **T-LAB** tool, which takes into account the positions of the various lexical units relative to each other, allows us to represent and explore any text as a network.

That means that the user is allowed to check the relationships between the 'nodes' (i.e. the key-terms) of the network at different levels: a) in one-to-one connections; b) in the 'ego' networks; c) within the 'community' to which they belong; d) within the entire text network.

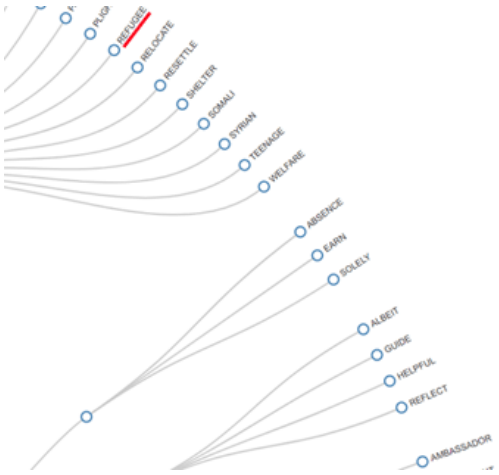
ONE-TO-ONE



EGO-NETWORK

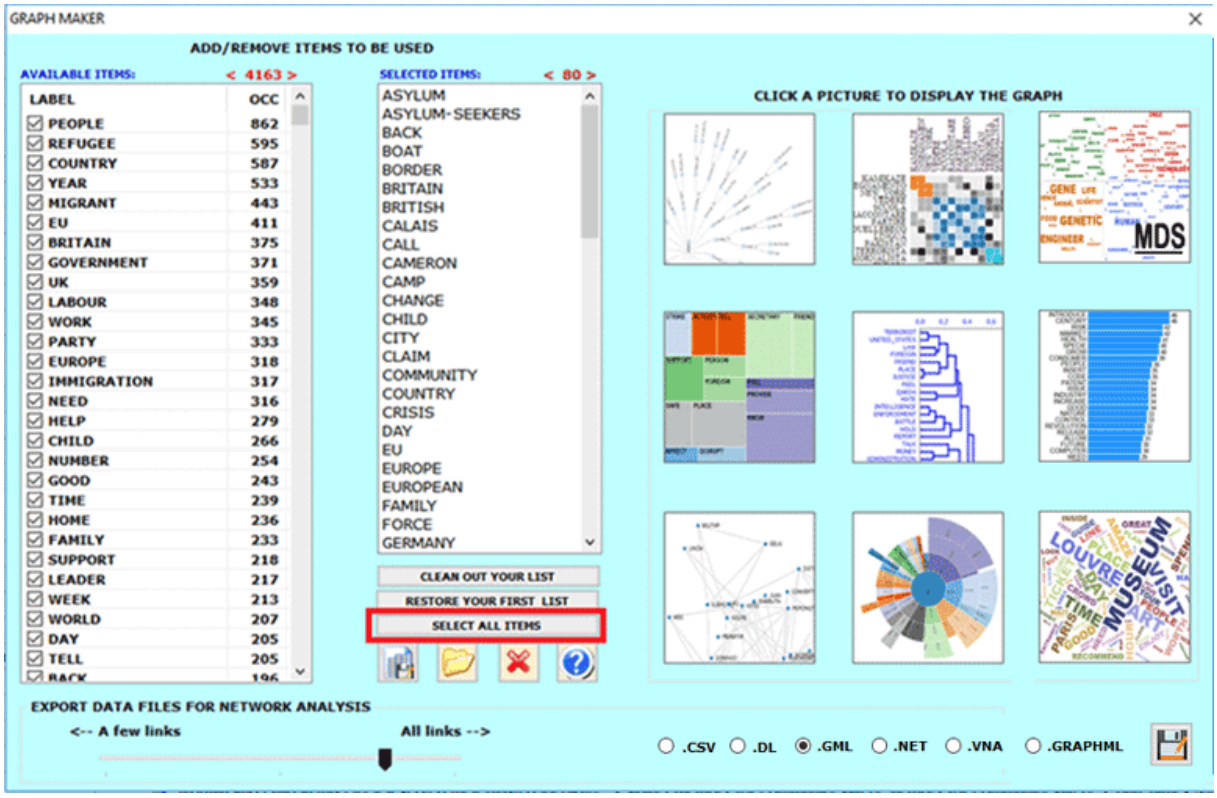


COMMUNITY



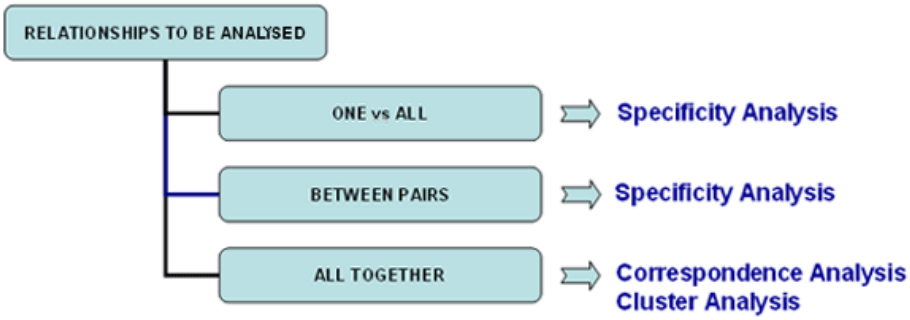
ENTIRE NETWORK

Moreover, by clicking the **GRAPH MAKER** option, the user is allowed to obtain various types of graphs by using customized lists of key words (see below).

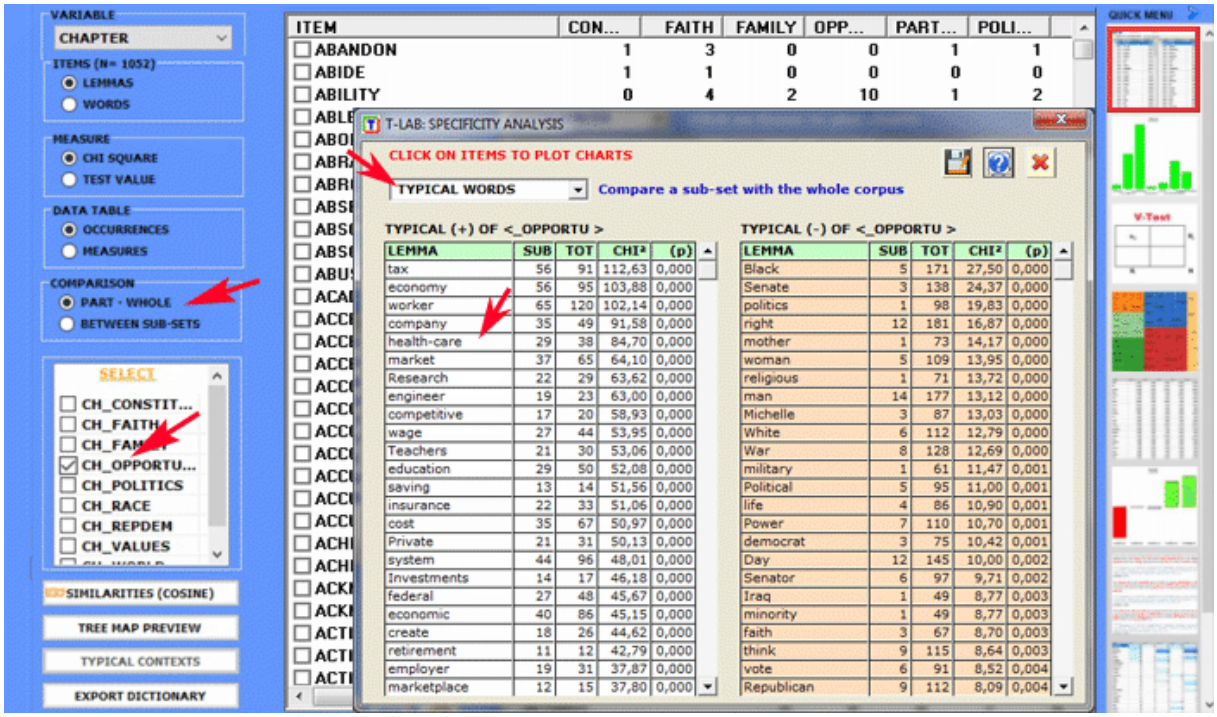


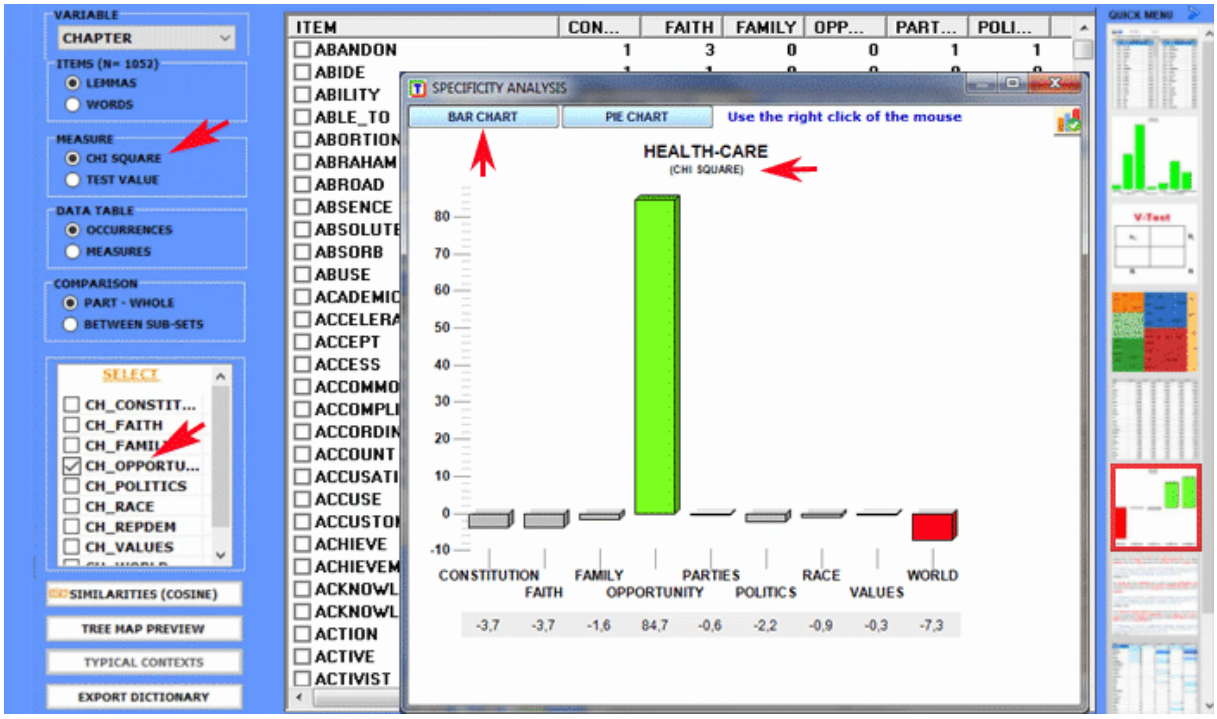
B : TOOLS FOR COMPARATIVE ANALYSIS

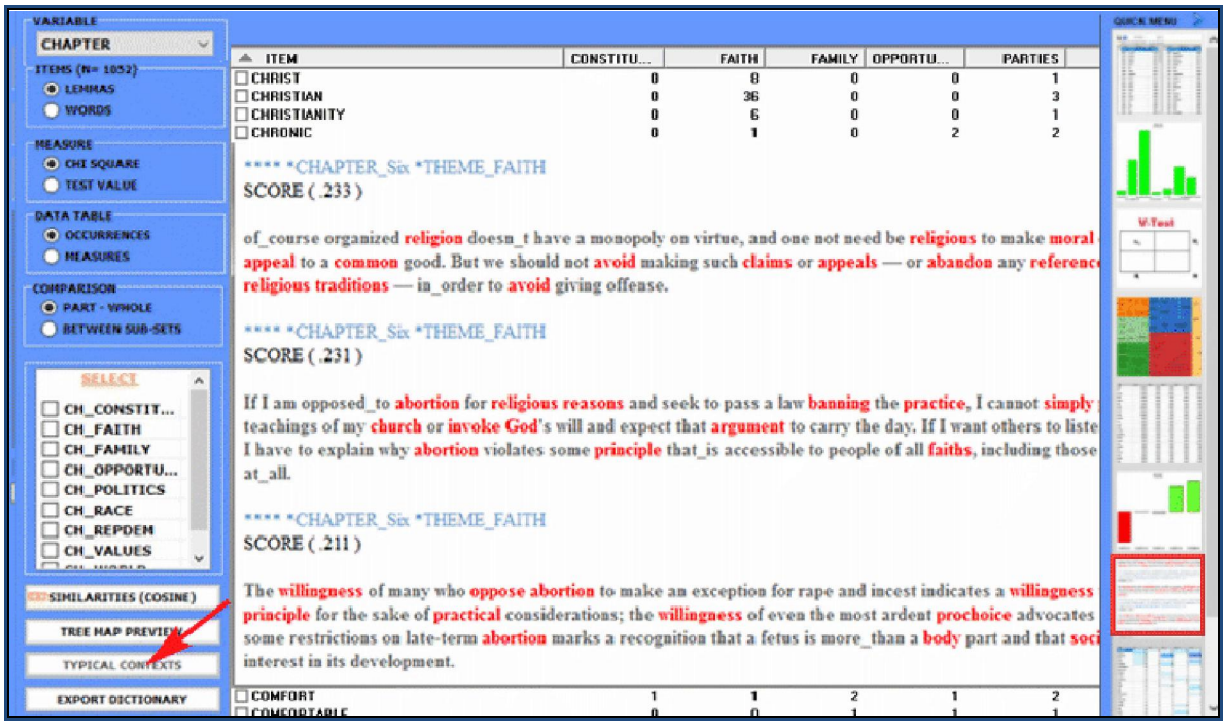
These tools enable us to analyse different kinds of relationships between context units (e.g. documents or corpus subsets)



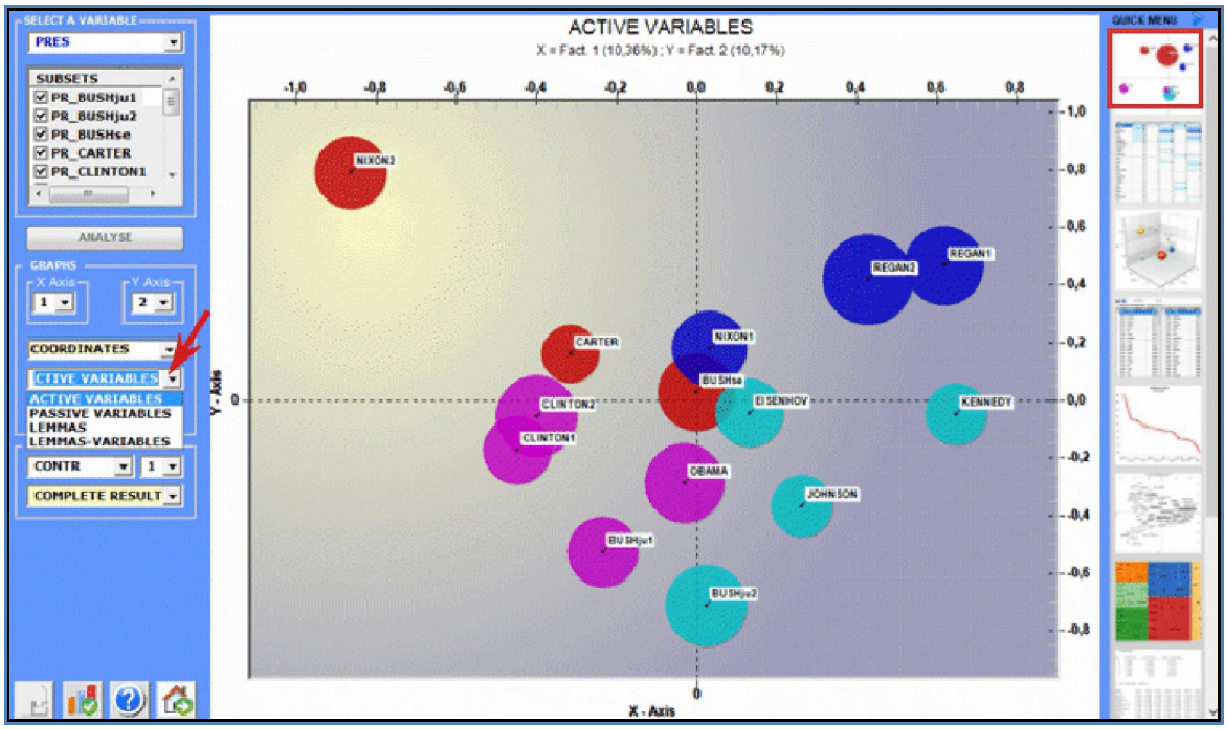
Specificity Analysis enables us to check which words are **typical** or **exclusive** of a specific corpus subset, either comparing it with the rest of the corpus or with another subset. Moreover it allows us to extract the **typical contexts** (i.e. the characteristic elementary contexts) of each analysed subset (e.g. the 'typical' sentences used by any specific political leader).

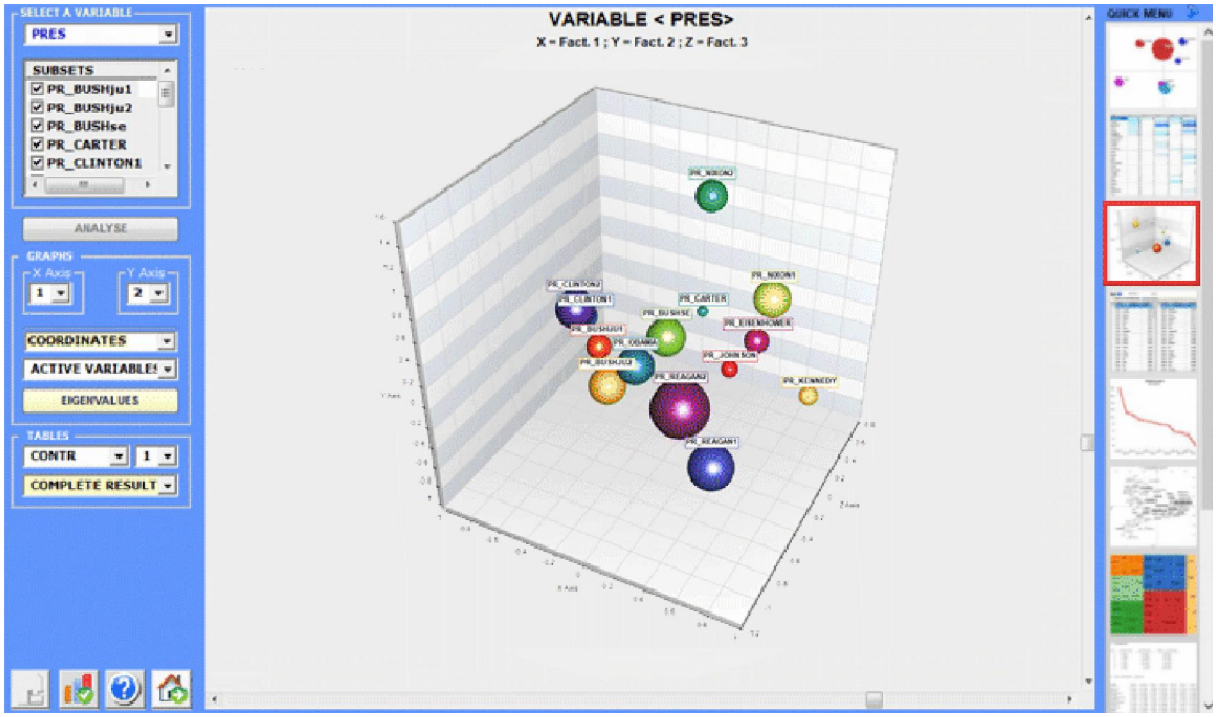

VARIABLE
CHAPTER
ITEMS (N= 1052)
LEMMAS
WORDS
MEASURE
CHI SQUARE
TEST VALUE
DATA TABLE
OCCURRENCES
MEASURES
COMPARISON
PART - WHOLE
BETWEEN SUB-SETS
SELECT
CH_CONSTIT...
CH_FAITH
CH_FAMILY
CH_OPPORTU...
CH_POLITICS
CH_RACE
CH_REPDEM
CH_VALUES
SIMILARITIES (COSINE)
TREE MAP PREVIEW
TYPICAL CONTEXTS
EXPORT DICTIONARY
ITEM CON... FAITH FAMILY OPP... PART... POLI...
ABANDON 1 3 0 0 1 1
ABIDE 1 1 0 0 0 0
ABILITY 0 4 2 10 1 2
T-LAB: SPECIFICITY ANALYSIS
CLICK ON ITEMS TO PLOT CHARTS
TYPICAL WORDS
Compare a sub-set with the whole corpus
TYPICAL (+) OF <_OPPORTU >
LEMMA SUB TOT CHI² (p)
tax 56 91 112,63 0,000
economy 56 95 103,88 0,000
worker 65 120 102,14 0,000
company 35 49 91,58 0,000
health-care 29 38 84,70 0,000
market 37 65 64,10 0,000
Research 22 29 63,62 0,000
engineer 19 23 63,00 0,000
competitive 17 20 58,93 0,000
wage 27 44 53,95 0,000
Teachers 21 30 53,06 0,000
education 29 50 52,08 0,000
saving 13 14 51,56 0,000
insurance 22 33 51,06 0,000
cost 35 67 50,97 0,000
Private 21 31 50,13 0,000
system 44 96 48,01 0,000
Investments 14 17 46,18 0,000
federal 27 48 45,67 0,000
economic 40 86 45,15 0,000
create 18 26 44,62 0,000
retirement 11 12 42,79 0,000
employer 19 31 37,87 0,000
marketplace 12 15 37,80 0,000
TYPICAL (-) OF <_OPPORTU >
LEMMA SUB TOT CHI² (p)
Black 5 171 27,50 0,000
Senate 3 138 24,37 0,000
politics 1 98 19,83 0,000
right 12 181 16,87 0,000
mother 1 73 14,17 0,000
woman 5 109 13,95 0,000
religious 1 71 13,72 0,000
man 14 177 13,12 0,000
Michelle 3 87 13,03 0,000
White 6 112 12,79 0,000
War 8 128 12,69 0,000
military 1 61 11,47 0,001
Political 5 95 11,00 0,001
life 4 86 10,90 0,001
Power 7 110 10,70 0,001
democrat 3 75 10,42 0,001
Day 12 145 10,00 0,002
Senator 6 97 9,71 0,002
Iraq 1 49 8,77 0,003
minority 1 49 8,77 0,003
faith 3 67 8,70 0,003
think 9 115 8,64 0,003
vote 6 91 8,52 0,004
Republican 9 112 8,09 0,004
QUICK MENU
V-Test


VARIABLE
CHAPTER
ITEMS (N= 1052)
LEMMAS
WORDS
MEASURE
CHI SQUARE
TEST VALUE
DATA TABLE
OCCURRENCES
MEASURES
COMPARISON
PART - WHOLE
BETWEEN SUB-SETS
SELECT
CH_CONSTIT...
CH_FAITH
CH_FAMILY
CH_OPPORTU...
CH_POLITICS
CH_RACE
CH_REPDEM
CH_VALUES
SIMILARITIES (COSINE)
TREE MAP PREVIEW
TYPICAL CONTEXTS
EXPORT DICTIONARY
ITEM CON... FAITH FAMILY OPP... PART... POLI...
ABANDON 1 3 0 0 1 1
ABIDE
ABILITY
ABLE_TO
ABORTION
ABRAHAM
ABROAD
ABSENCE
ABSOLUTE
ABSORB
ABUSE
ACADEMIC
ACCELERA
ACCEPT
ACCESS
ACCOMMO
ACCOMPLI
ACCORDIN
ACCOUNT
ACCUSATI
ACCUSE
ACCUSTOM
ACHIEVE
ACHIEVEM
ACKNOWL
ACKNOWL
ACTION
ACTIVE
ACTIVIST
SPECIFICITY ANALYSIS
BAR CHART
PIE CHART
Use the right click of the mouse
HEALTH-CARE
(CHI SQUARE)
80
70
60
50
40
30
20
10
0
-10
CONSTITUTION
FAITH
FAMILY
OPPORTUNITY
PARTIES
POLITICS
RACE
VALUES
WORLD
-3,7 -3,7 -1,6 84,7 -0,6 -2,2 -0,9 -0,3 -7,3
QUICK MENU
V-Test

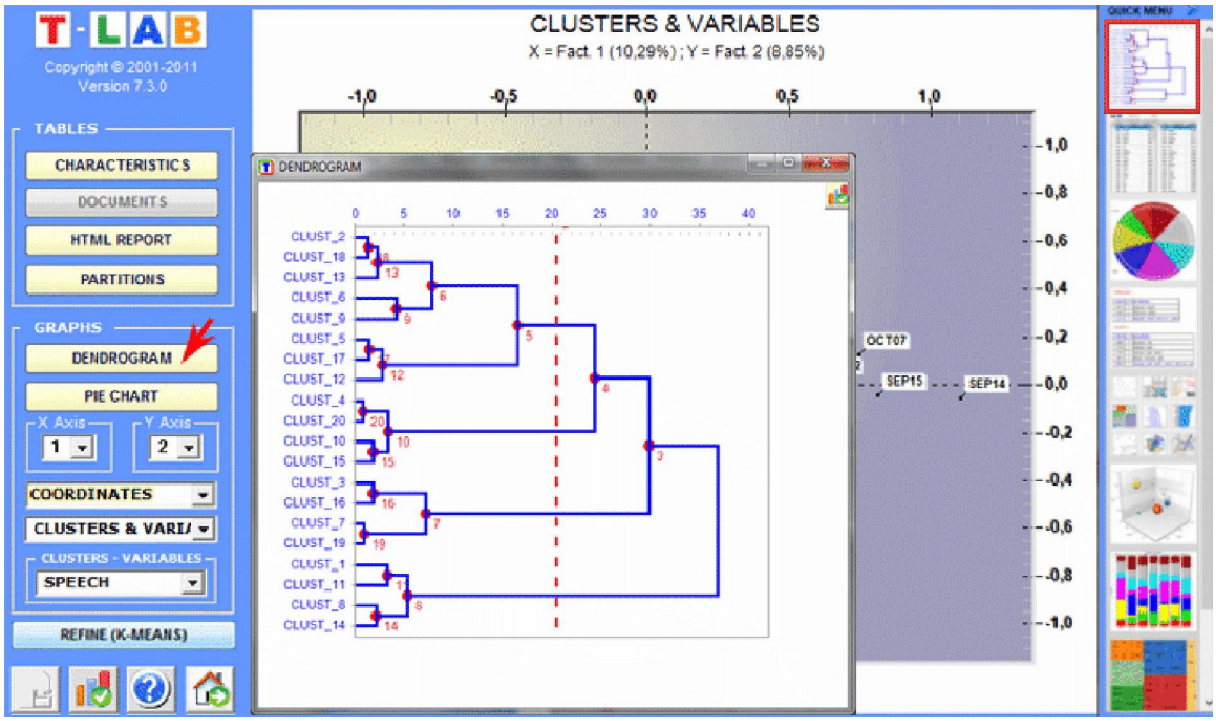


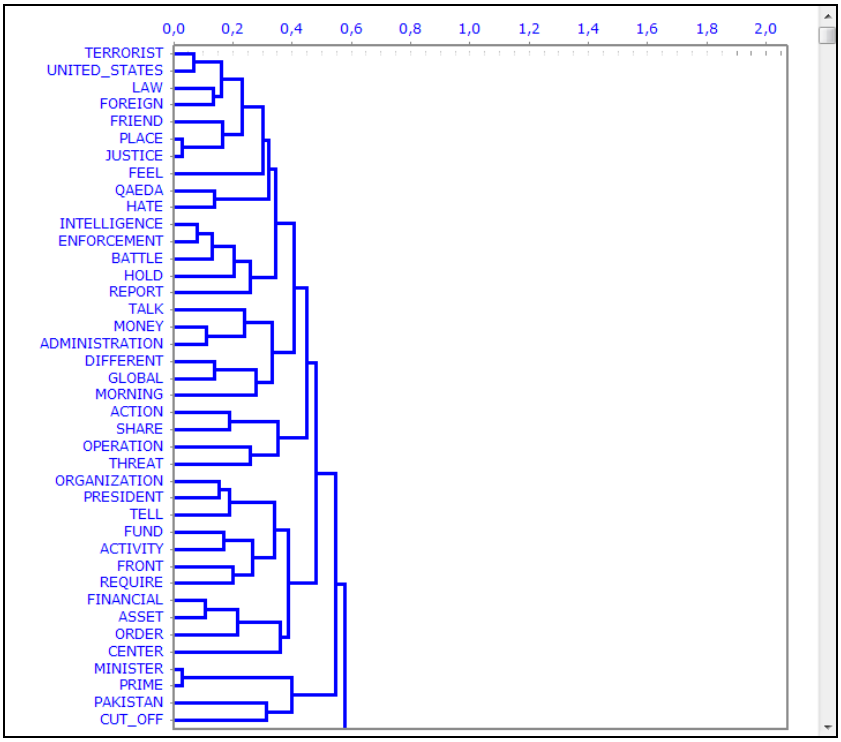
Correspondence Analysis allows us to explore similarities and differences between (and within) groups of context units (e.g. documents belonging to the same category).





Cluster Analysis, which can be carried out using various techniques, allows us to detect and explore groups of analysis units which have two complementary features: high internal (within cluster) homogeneity and high external (between cluster) heterogeneity.





C : TOOLS FOR THEMATIC ANALYSIS

These tools enable us to discover, examine and map "themes" emerging from texts.

As **theme** is a polysemous word, when using software tools for thematic analysis we have to refer to operational definitions. More precisely, in these **T-LAB** tools, "theme" is a label used to indicate four different entities:

1- a **thematic cluster of contexts units** characterized by the same patterns of key-words (see the Thematic Analysis of Elementary Contexts, Thematic Document Classification and Dictionary-Based Classification tools);

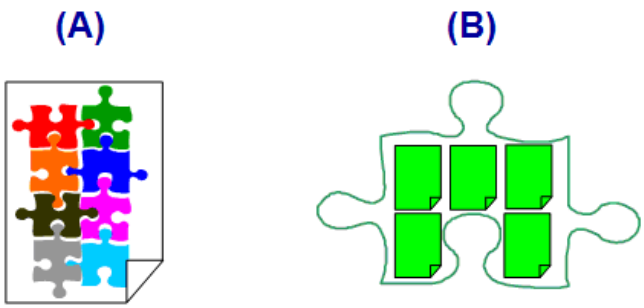
2- a **thematic group of key terms** classified as belonging to the same category (see the Dictionary-Based Classification tool);

3 – a **mixture component** of a probabilistic model which represents each context unit (i.e. elementary context or document) as generated from a fixed number of topics or "themes" (see the Modeling of Emerging Themes and the Texts and Discourses treated ad Dynamic Systems tools).

4- a **specific key term** used for extracting a set of elementary contexts in which it is associated with a specific group of words pre-selected by the user (see the Key Contexts of Thematic Words tool).

For example, depending on the tool we are using, a single document can be analysed as composed of various 'themes' (see 'A' below) or as belonging to a set of documents

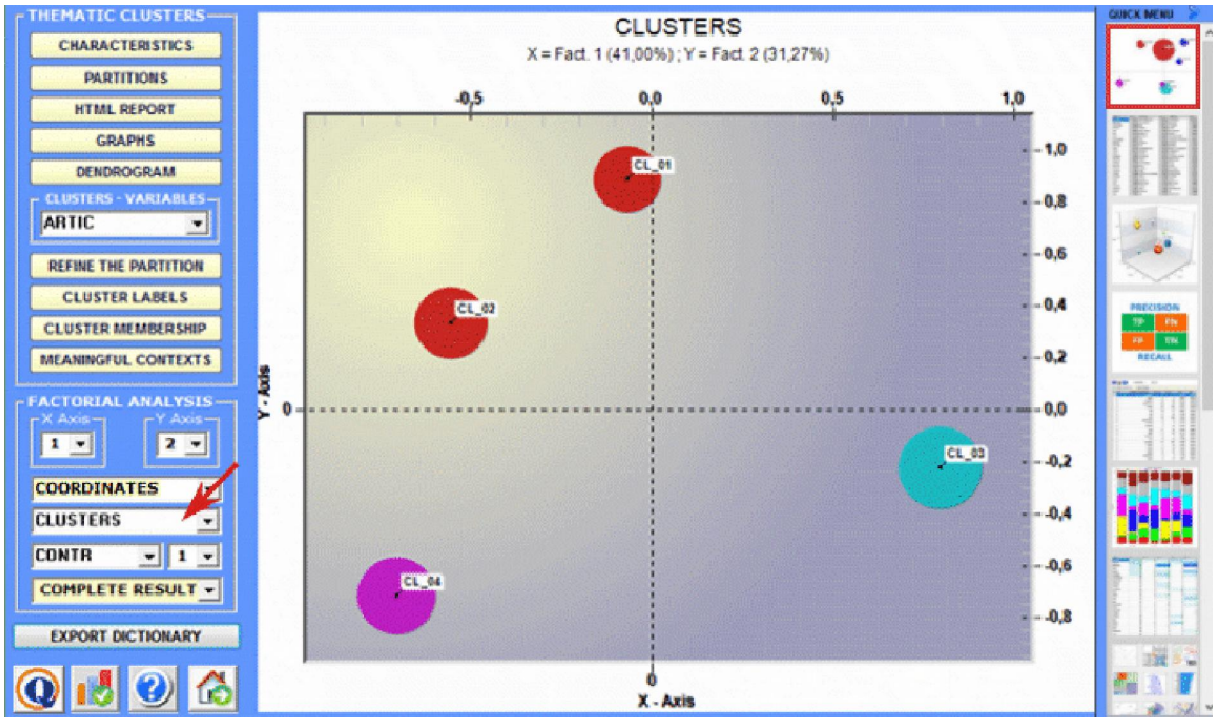
concerning the same 'theme' (see 'B' below). In fact, in the case of 'A' each theme can correspond to a word or to a sentence, whereas in the case of 'B' a theme can be a label assigned to a cluster of documents characterized by the same patterns of key-words.

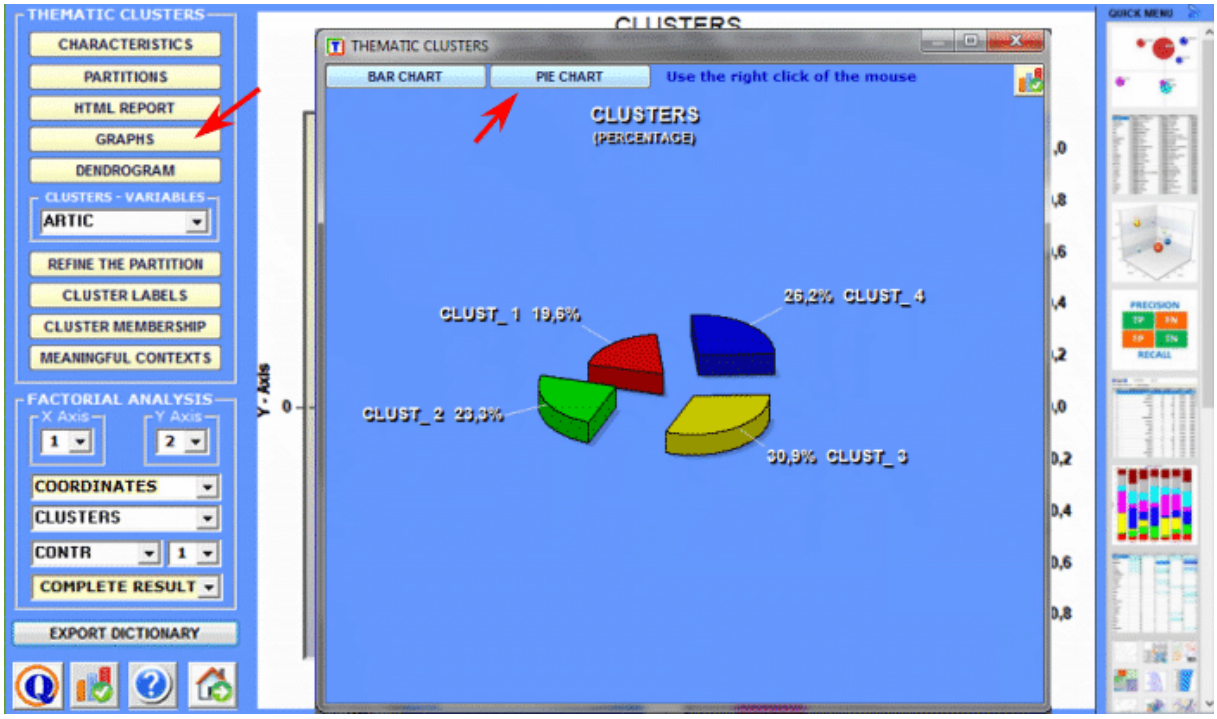


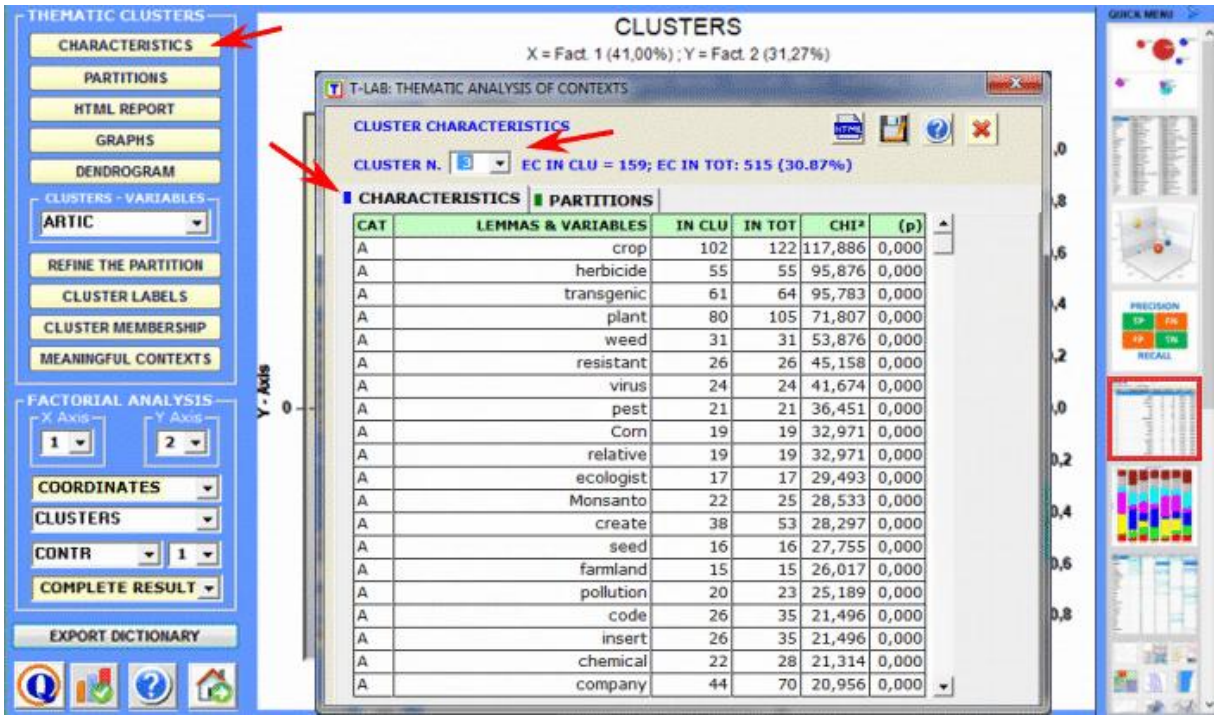
In detail, the ways how **T-LAB** 'extracts' themes are the following:

1 - both the **Thematic Analysis of Elementary Contexts** and the **Thematic Document Classification** tools, when performing an unsupervised clustering, work in the following way:

a - perform **co-occurrence analysis** to identify thematic clusters of context units;
b - perform **comparative analysis** of the profiles of the various clusters;
c - generate various types of graphs and tables (see below);
d - allow you to save the **new variables** (thematic clusters) for further analysis.

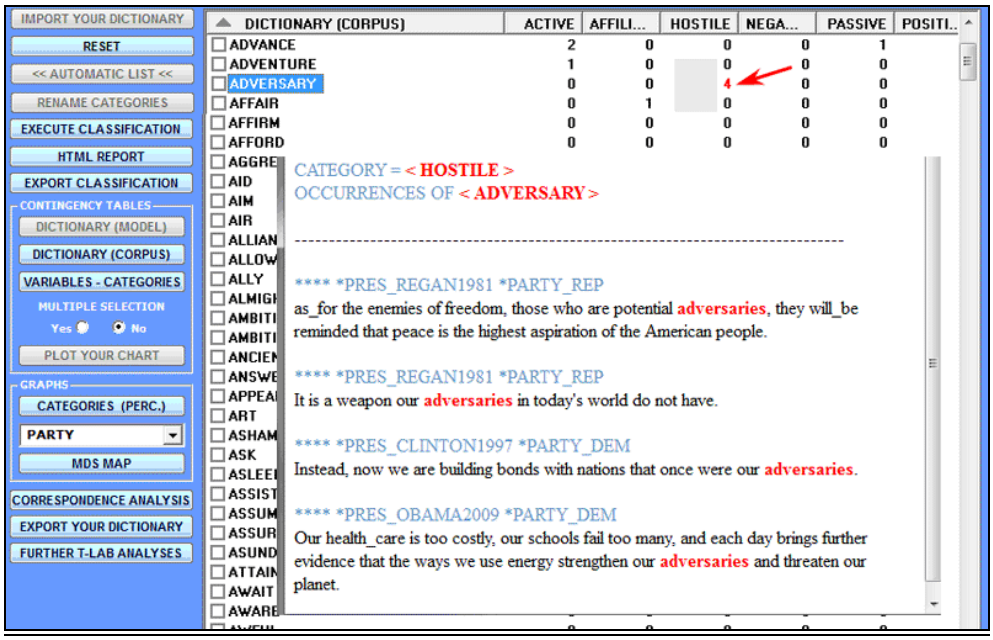


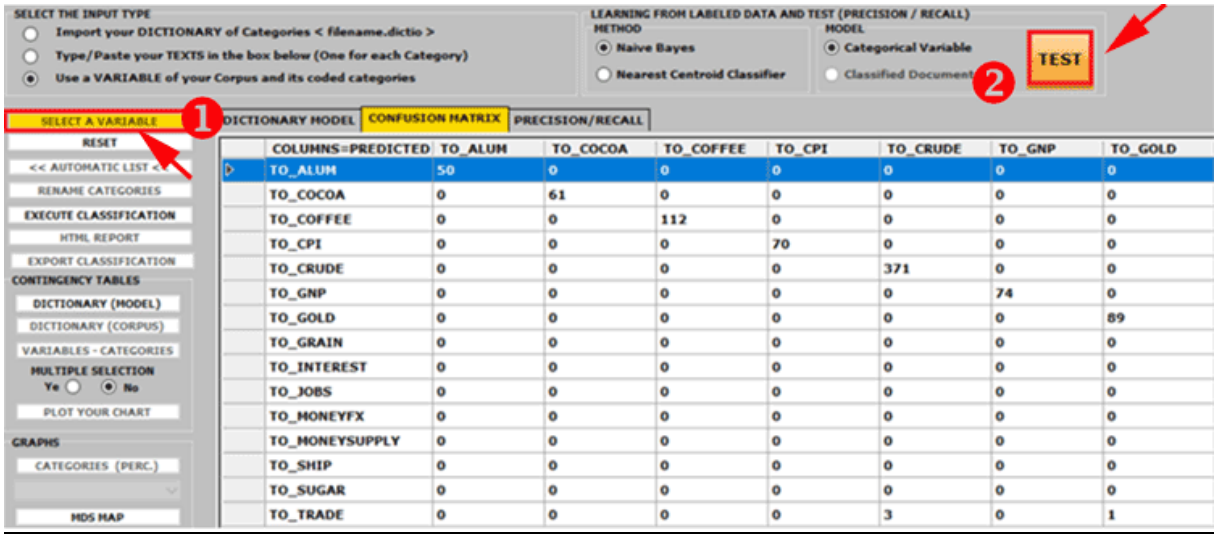




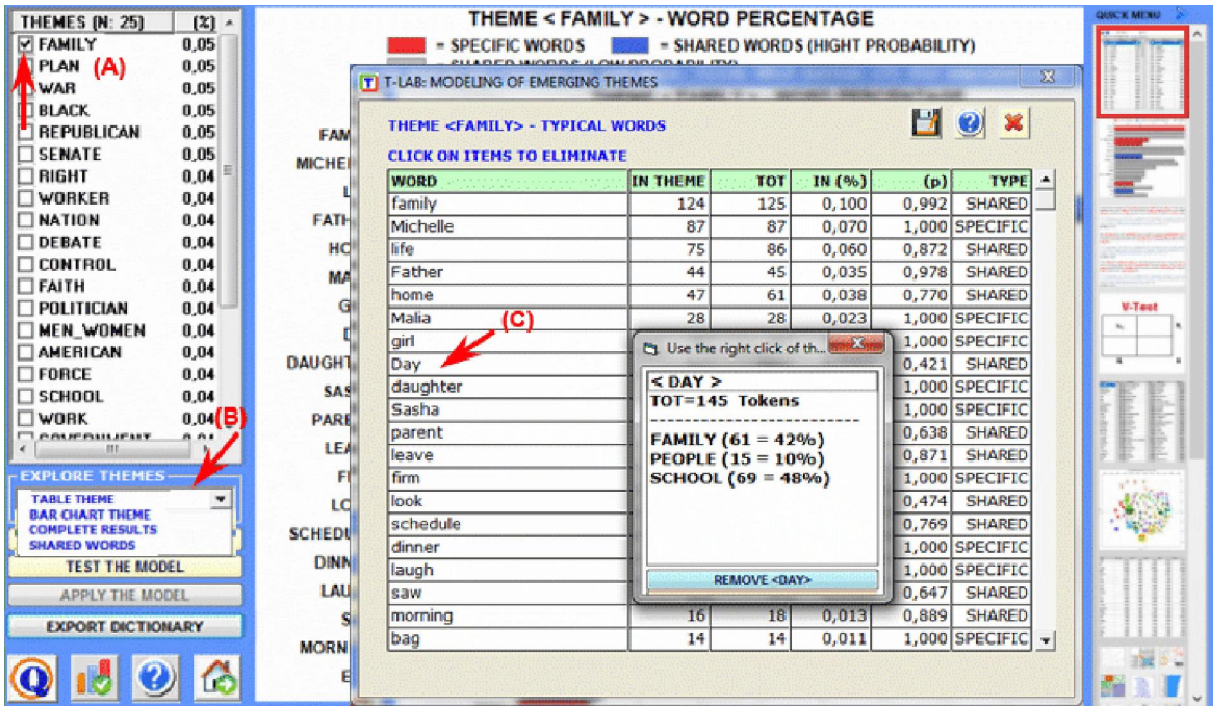
CAT	LEMMAS & VARIABLES	IN CLU	IN TOT	CHI²	(p)
A	crop	102	122	117,886	0,000
A	herbicide	55	55	95,876	0,000
A	transgenic	61	64	95,783	0,000
A	plant	80	105	71,807	0,000
A	weed	31	31	53,876	0,000
A	resistant	26	26	45,158	0,000
A	virus	24	24	41,674	0,000
A	pest	21	21	36,451	0,000
A	Corn	19	19	32,971	0,000
A	relative	19	19	32,971	0,000
A	ecologist	17	17	29,493	0,000
A	Monsanto	22	25	28,533	0,000
A	create	38	53	28,297	0,000
A	seed	16	16	27,755	0,000
A	farmland	15	15	26,017	0,000
A	pollution	20	23	25,189	0,000
A	code	26	35	21,496	0,000
A	insert	26	35	21,496	0,000
A	chemical	22	28	21,314	0,000
A	company	44	70	20,956	0,000

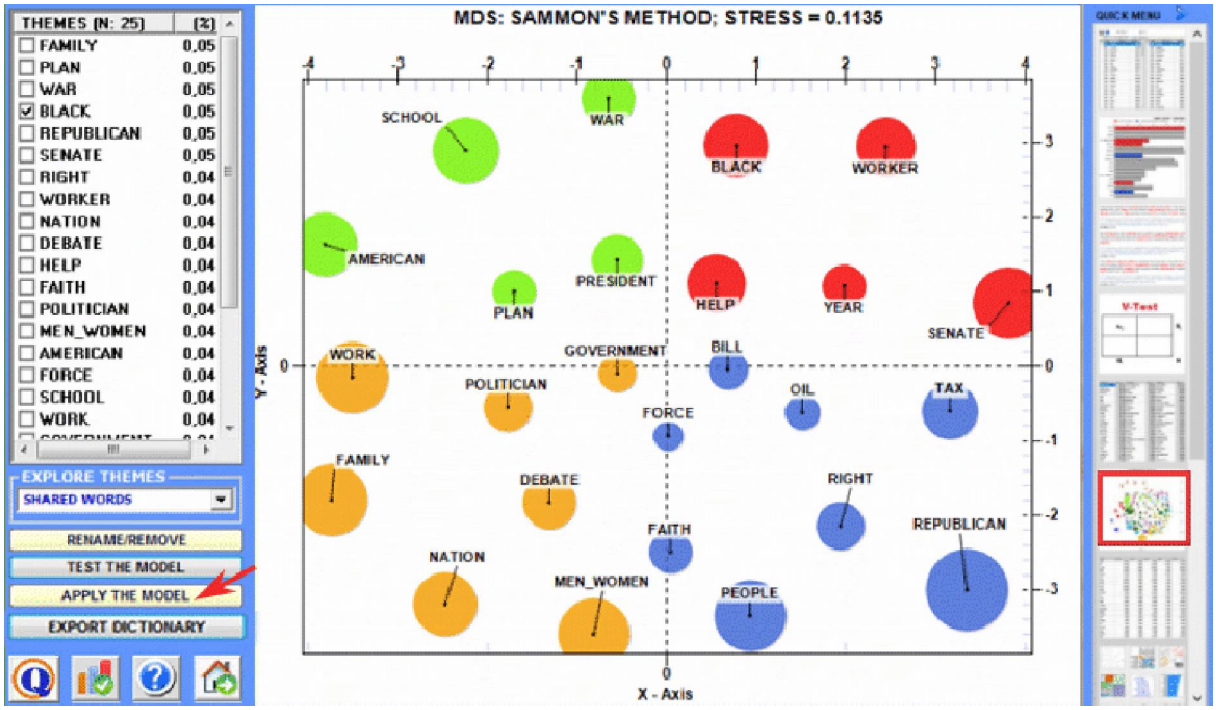
2 - through the **Dictionary-Based Classification** tool we can easily build/test/apply models (e.g. dictionaries of categories or pre-existing manual categorizations) both for the classical qualitative content analysis and for the sentiment analysis. In fact such a tool allows us to perform an automated top-down classification of lexical units (i.e. words and lemmas) or context units (i.e. sentences, paragraphs and short documents) present in a text collection.


IMPORT YOUR DICTIONARY
RESET
<< AUTOMATIC LIST <<
RENAME CATEGORIES
EXECUTE CLASSIFICATION
HTML REPORT
EXPORT CLASSIFICATION
CONTINGENCY TABLES
DICTIONARY (MODEL)
DICTIONARY (CORPUS)
VARIABLES - CATEGORIES
MULTIPLE SELECTION
Yes No
PLOT YOUR CHART
GRAPHS
CATEGORIES (PERC.)
PARTY
MDS MAP
CORRESPONDENCE ANALYSIS
EXPORT YOUR DICTIONARY
FURTHER T-LAB ANALYSES
DICTIONARY (CORPUS) ACTIVE AFFILI... HOSTILE NEGA... PASSIVE POSITI..
ADVANCE 2 0 0 0 1
ADVENTURE 1 0 0 0 0
ADVERSARY 0 0 4 0 0
AFFAIR 0 1 0 0 0
AFFIRM 0 0 0 0 0
AFFORD 0 0 0 0 0
CATEGORY = < HOSTILE >
OCCURRENCES OF < ADVERSARY >
**** *PRES_REGAN1981 *PARTY_REP
as_for the enemies of freedom, those who are potential adversaries, they will_be reminded that peace is the highest aspiration of the American people.
**** *PRES_REGAN1981 *PARTY_REP
It is a weapon our adversaries in today's world do not have.
**** *PRES_CLINTON1997 *PARTY_DEM
Instead, now we are building bonds with nations that once were our adversaries.
**** *PRES_OBAMA2009 *PARTY_DEM
Our health_care is too costly, our schools fail too many, and each day brings further evidence that the ways we use energy strengthen our adversaries and threaten our planet.

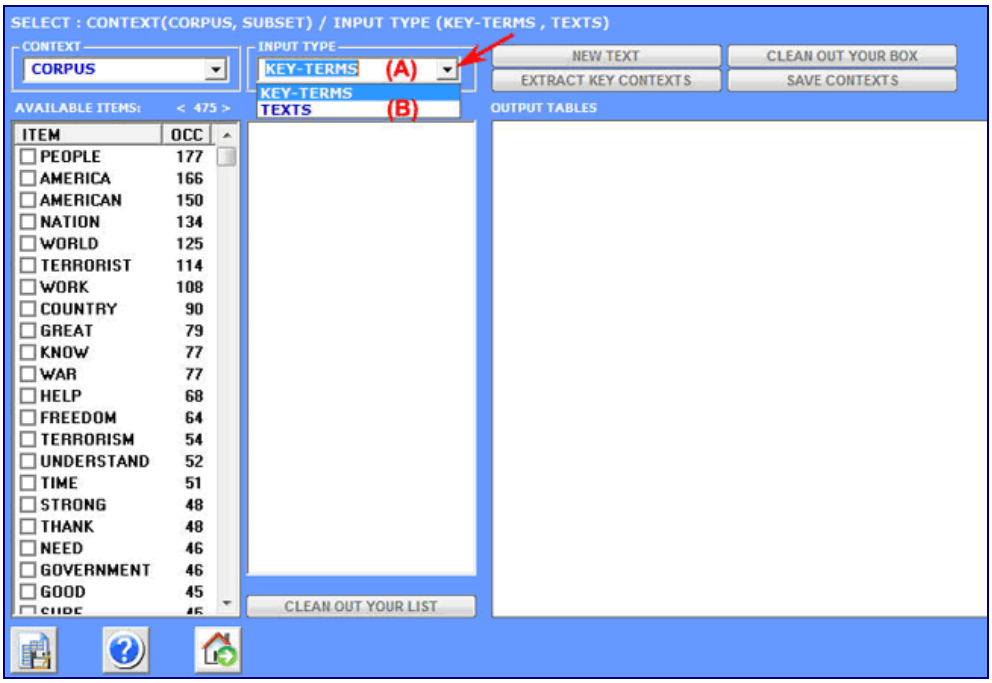

SELECT THE INPUT TYPE
Import your DICTIONARY of Categories < filename.dictio >
Type/Paste your TEXTS in the box below (One for each Category)
Use a VARIABLE of your Corpus and its coded categories
LEARNING FROM LABELED DATA AND TEST (PRECISION / RECALL)
METHOD
Naive Bayes
Nearest Centroid Classifier
MODEL
Categorical Variable
Classified Document
TEST
SELECT A VARIABLE
DICTIONARY MODEL
CONFUSION MATRIX
PRECISION/RECALL
RESET
<< AUTOMATIC LIST <<
RENAME CATEGORIES
EXECUTE CLASSIFICATION
HTML REPORT
EXPORT CLASSIFICATION
CONTINGENCY TABLES
DICTIONARY (MODEL)
DICTIONARY (CORPUS)
VARIABLES - CATEGORIES
MULTIPLE SELECTION
PLOT YOUR CHART
GRAPHS
CATEGORIES (PERC.)
MDS MAP
COLUMNS=PREDICTED TO_ALUM TO_COCOA TO_COFFEE TO_CPI TO_CRUDE TO_GNP TO_GOLD
TO_ALUM 50 0 0 0 0 0 0
TO_COCOA 0 61 0 0 0 0 0
TO_COFFEE 0 0 112 0 0 0 0
TO_CPI 0 0 0 70 0 0 0
TO_CRUDE 0 0 0 0 371 0 0
TO_GNP 0 0 0 0 0 74 0
TO_GOLD 0 0 0 0 0 0 89
TO_GRAIN 0 0 0 0 0 0 0
TO_INTEREST 0 0 0 0 0 0 0
TO_JOBS 0 0 0 0 0 0 0
TO_MONEYFX 0 0 0 0 0 0 0
TO_MONEYSUPPLY 0 0 0 0 0 0 0
TO_SHIP 0 0 0 0 0 0 0
TO_SUGAR 0 0 0 0 0 0 0
TO_TRADE 0 0 0 0 3 0 1

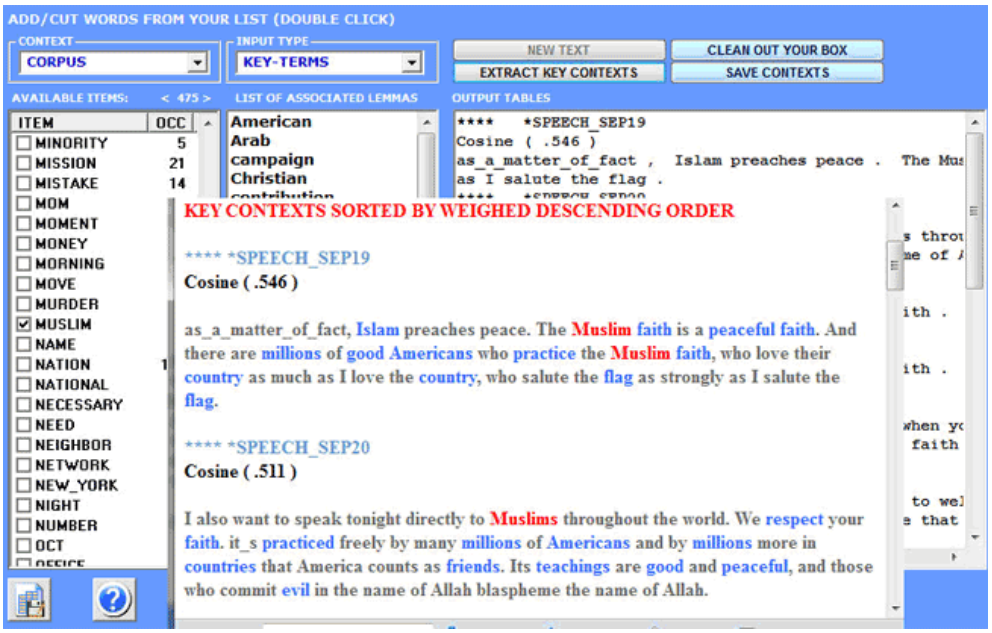
3 - through the **Modelling of Emerging Themes** tool (see below) the **mixture components** described through their characteristic vocabulary can be used for building a coding scheme for qualitative analysis and/or for the automatic classification of the context units (i.e. documents or elementary contexts).





4 - the **Key Contexts of Thematic Words** tool (see below) can be used for two different purposes: (a) to extract lists of meaningful context units (i.e. elementary contexts) which allow us to deepen the thematic value of specific **key words**; (b) to extract context units which are the most similar to sample **texts** chosen by the user.

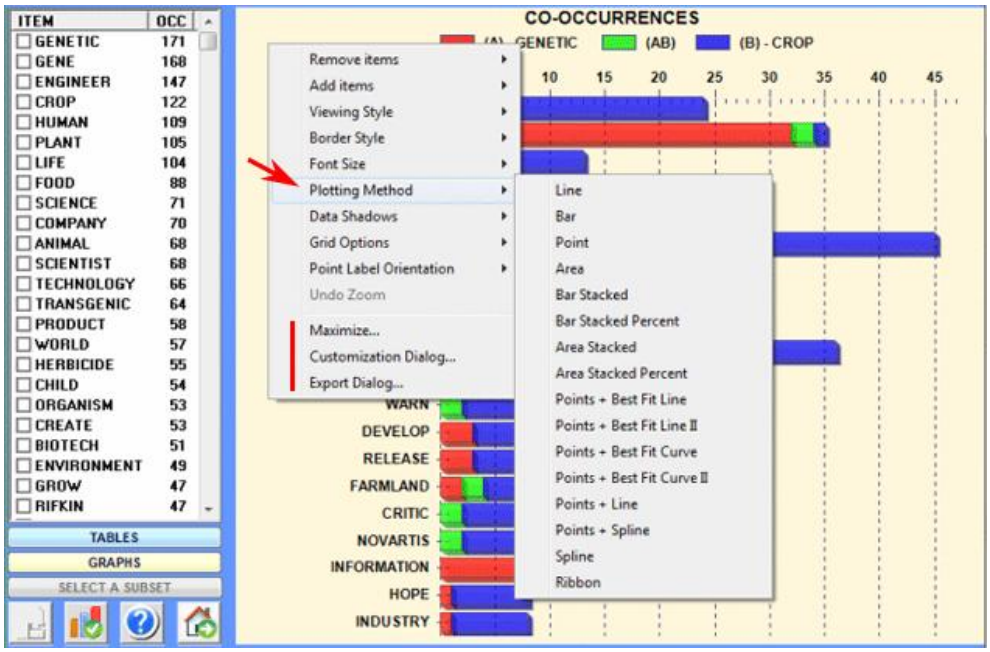


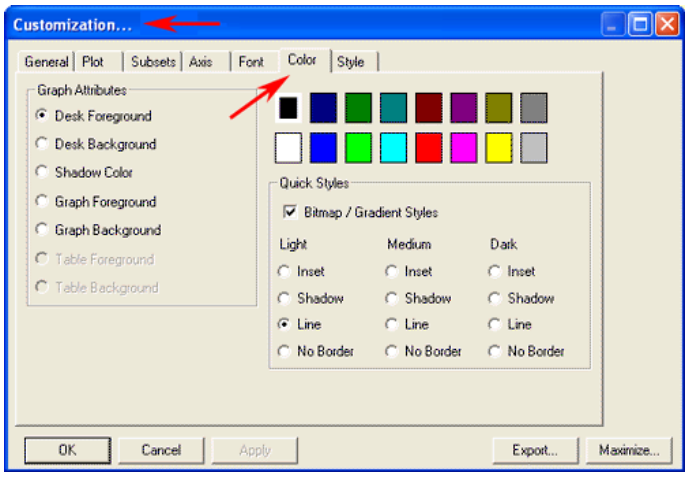


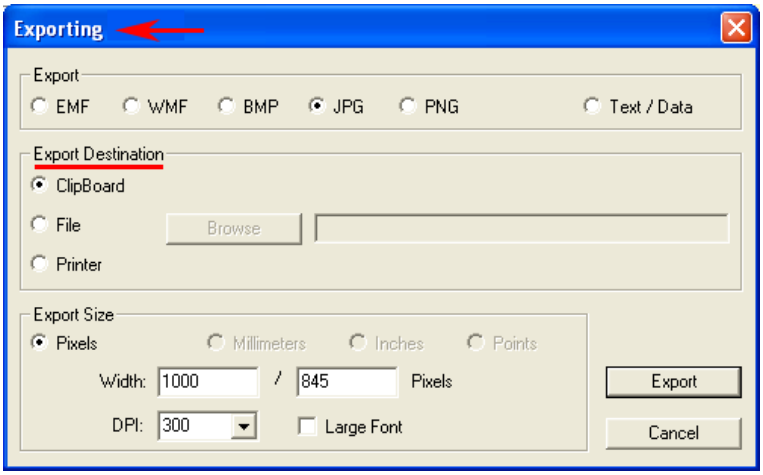
6 - INTERPRETATION OF THE OUTPUTS consists in the consultation of the tables and the graphs produced by **T-LAB**, in the eventual customization of their format and in making inferences on the meaning of the relationships represented by the same.

In the case of **tables**, according to each case, **T-LAB** allows the user to export them in files with the following extensions: **.DAT**, **.TXT**, **.CSV**, **.XLXS**, **.HTML**. This means that, by using any text editor program and /or any Microsoft Office application, the user can easily import and re-elaborate them.

All **graphs and charts** can be zoomed (left-click and drag), maximized, customized and exported in different formats (right click to show popup menu)







Some general criteria for the interpretation of the **T-LAB** outputs are illustrated in a paper quoted in the **Bibliography** and are available from the https://www.tlab.it website (Lancia F.: 2007). This document presents the hypothesis that the statistical elaboration outputs (tables and graphs) are particular types of texts, that is they are multi-semiotic objects characterized by the fact that the relationships between the signs and the symbols are ordered by measures that refer to specific **codes**.

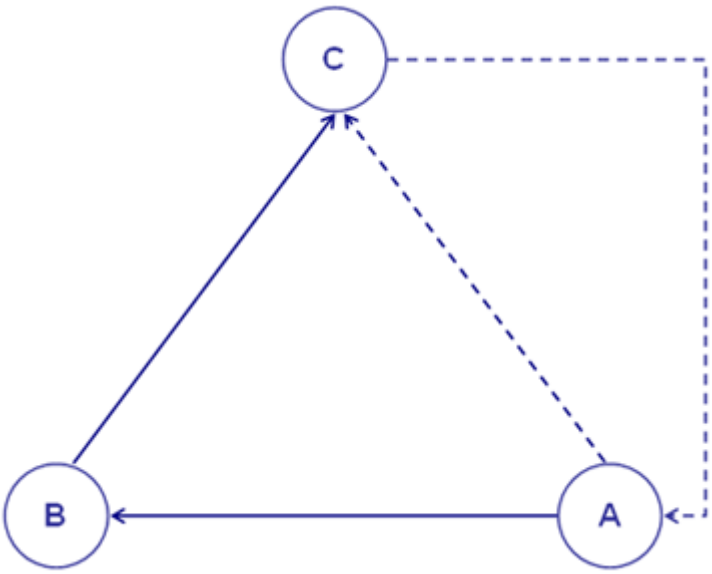
In other words, both in the case of texts written in "natural language" and those written in the "statistical language", the possibility of making inferences on the relationships that organize the **content forms** is guaranteed by the fact that the relationships between the **expression forms** are not random; in fact, in the first case (natural language) the significant units follow on and are ordered in a linear manner (one after the other in the chain of the discourse), while in the second case (tables and graphs) the organization of the multidimensional **semantic spaces** comes from statistical measures.

Even if the semantic spaces represented in the **T-LAB** maps are extremely varied, and each of them require specific interpretative procedures, we can theorize that - in general - the logic of the inferential process is the following:

A – to detect some significant relationships between the units "present" on the expression plan (e.g. between table and/or graph labels);

B – to explore and compare the semantic traits of the same units and the contexts to which they are mentally and culturally associated (content plan);

C – to generate some hypothesis or some analysis categories that, in the context defined by the corpus, give reason for the relationships between expression and content forms.



At present, **T-LAB** options have the following **restrictions**:

- corpus dimension: max 90Mb, equal to about 55,000 pages in .txt format;
- primary documents: max 30,000 (max 99,999 for short texts which do not exceed 2,000 characters each, e.g. responses to open-ended questions, Twitter messages, etc.);
- categorical variables: max 50, each allowing max 150 subsets (categories) which can be compared;
- modelling of emerging themes: max 5,000 lexical units (*) by 5,000,000 occurrences;
- thematic analysis of elementary contexts: max 300,000 rows (context units) by 5,000 columns (lexical units);
- thematic document classification: max 99,999 rows (i.e. documents) by 5,000 columns (lexical units);
- specificity analysis (lexical units x categories): max 10,000 rows by 150 columns;
- correspondence analysis (lexical units x categories): max 10,000 rows by 150 columns;
- correspondence analysis (context units x lexical units): max 10,000 rows by 5,000 columns;
- multiple correspondence analysis (elementary contexts x categories): max 150,000 rows by 250 columns;
- singular value decomposition: max 300,000 rows by 5,000 columns;
- cluster analysis that uses the results of a previous correspondence analysis (or SVD): max 10,000 rows (lexical units or elementary contexts);
- word associations, comparison between word pairs: max 5,000 lexical units;
- co-word analysis and concept mapping: max 5,000 lexical units;
- sequence analysis: max 5,000 lexical units (or categories) by 3,000,000 occurrences.

(*) In **T-LAB**, 'lexical units' are words, multi-words, lemmas and semantic categories. So, when the automatic lemmatization is applied, 5,000 lexical units correspond to about 12,000 words (i.e. raw forms).

SETTINGS

Automatic and Customized Settings

The choice of **automatic** (A) or **customized** (B) settings relates to the list of Key-words used in all analyses performed with **T-LAB**. This choice is reversible until the user performs operations that modify the dictionary of the corpus.

A) AUTOMATIC SETTINGS

When choosing **automatic settings** the list of keywords includes **up to a maximum of 5,000 lexical units** automatically selected by **T-LAB**, which belong to the category of content words: nouns, verbs, adjectives and adverbs.

The selection criterion varies according to the kind of file analysed.

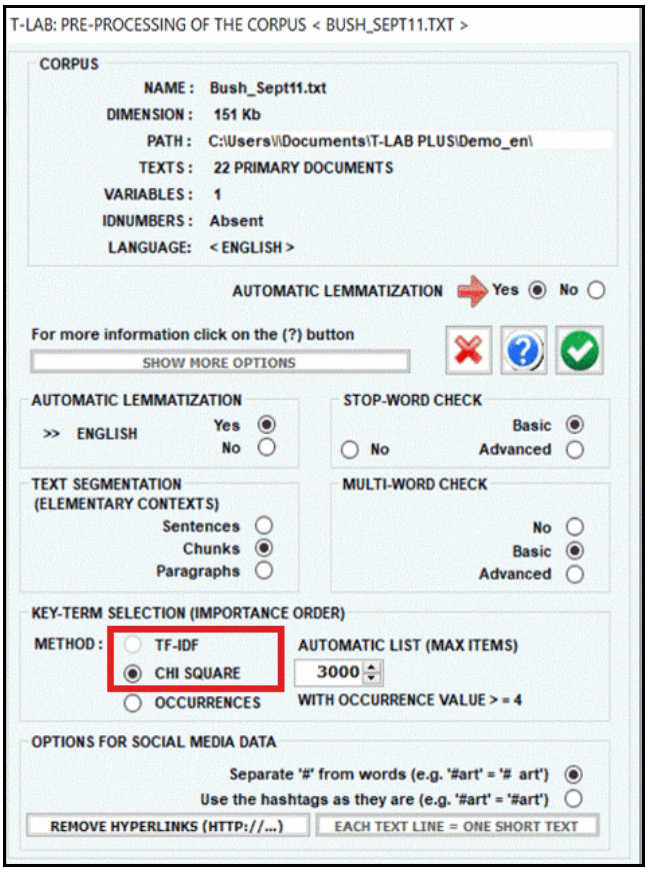
If the corpus is a single text **T-LAB** simply selects the lexical units with the highest **occurrence** values.

If the corpus is made up of two or more texts **T-LAB** uses the following algorithm:

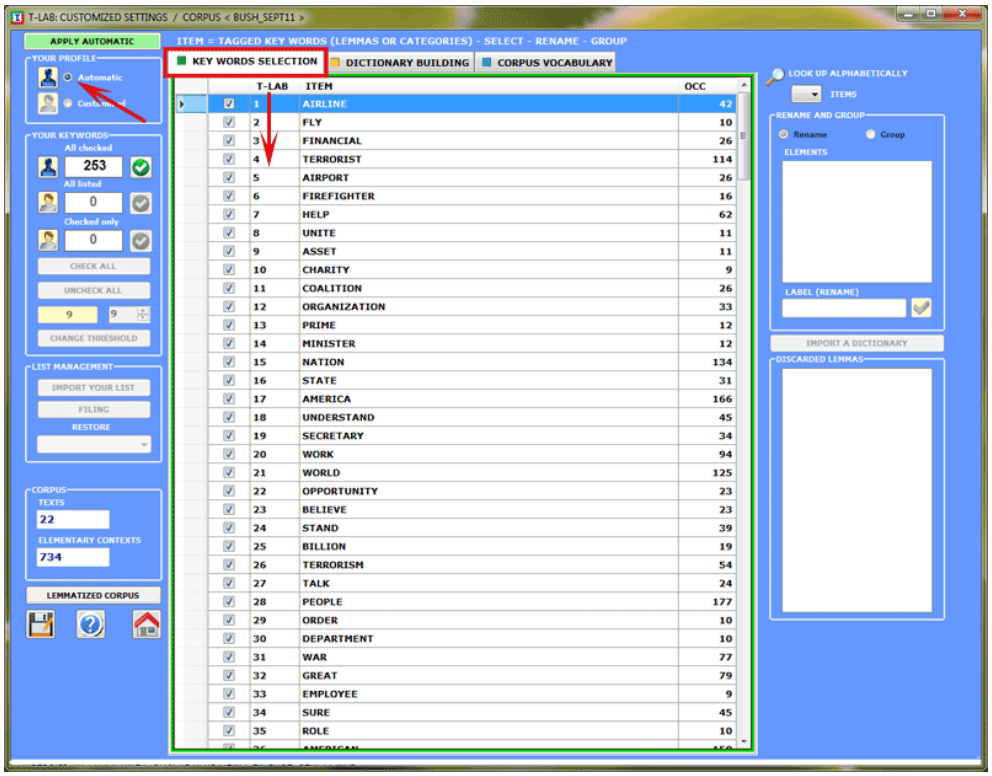
it selects the words with occurrence values higher than the minimum threshold;

- *it selects the words with occurrence values higher than the minimum threshold;*
- *it computes the TF-IDF or applies the chi-square test to all the crosses of each selected word for all the texts being analysed (N.B.: In the case of chi square test, the maximum number of text allowed is 500);*
- *it selects the words with the TF-IDF or the chi-square highest values, that is those words that, in the corpus, make the difference.*

N.B.:
- In the case that the corpus consists of two or more texts, the user can choose the selection criterion (CHI square or TF-IDF) in the import stage (see below);



- when the automatic settings option is enabled, the table with the list of Keywords includes a 'T-LAB' column which indicates the importance of each item according to the selected criterion (see below).



B) CUSTOMIZED SETTINGS

When choosing **customized settings** the user is allowed to **select**, **rename** and **group** the lexical units (i.e. words, lemmas or categories) to be included in subsequent **T-LAB** analyses.

All the lexical units with an occurrence value which is equal or superior to the preset **threshold** are listed (list 1). Some of these, that is those indicated with "☑", belong to a sub-list (list 2) create by **T-LAB** (see automatic settings).

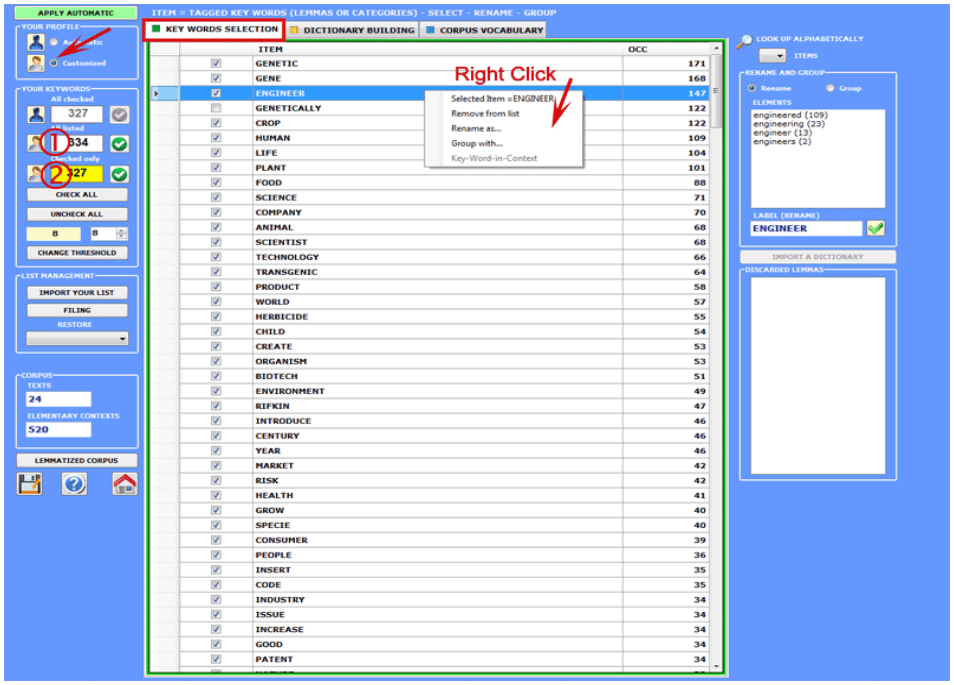
Depending on the analyses to be performed, the user can decide whether to use/modify the more extended list (1) or the sub-list (2) of the **T-LAB** key terms.

In both cases the operations available are the following:

- **change** the threshold value;
- **select** which lemmas are to be excluded from analysis
- **restore** one or more lemmas for use;
- **select/deselect** items.

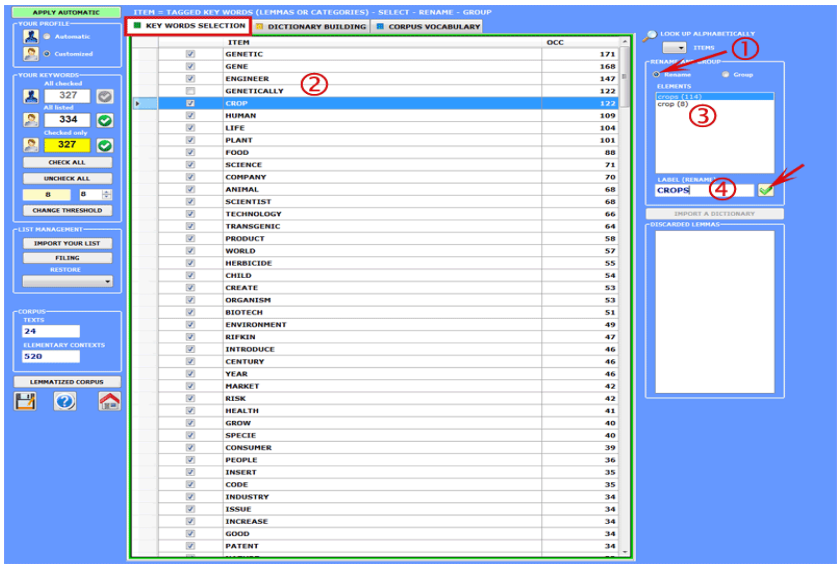
By clicking either the **list (1)** or the **list (2)** button the customized option of Analysis Settings is enabled (see below).

All options regarding interventions on individual entries are accessed by right clicking the mouse on any item of the table.



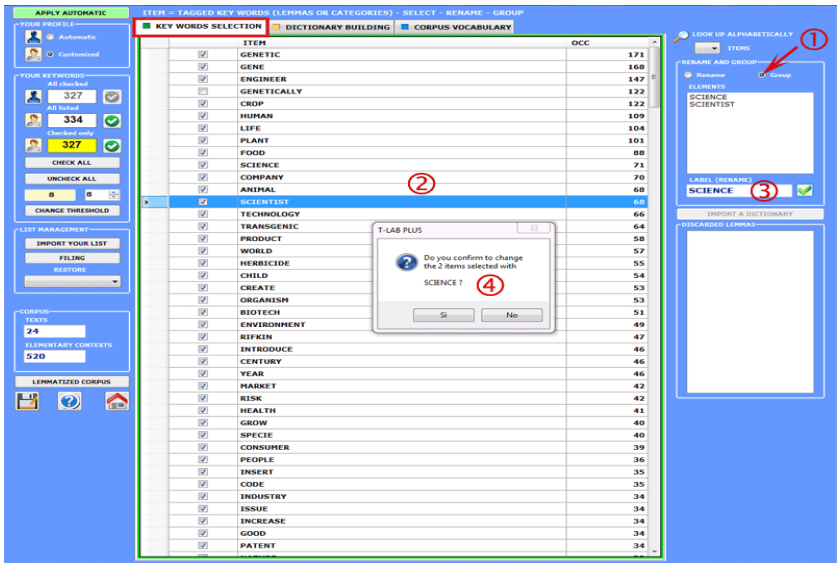
In order to **change the label (i.e. rename) of a single lemma**, proceed as follows:

1. be sure that the "RENAME" option is selected;
2. click on an item of the list;
3. select one word or type a new label in the appropriate box:
4. click on "RENAME".

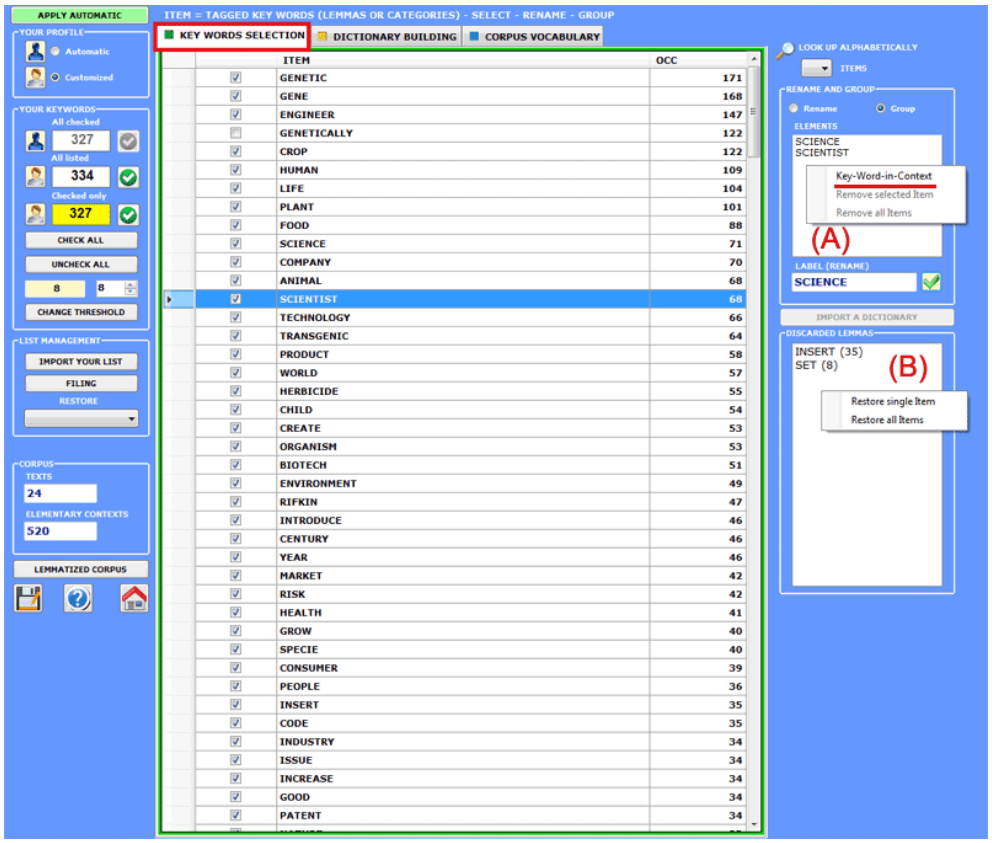


In order to **group two or more lemmas**, proceed as follows:

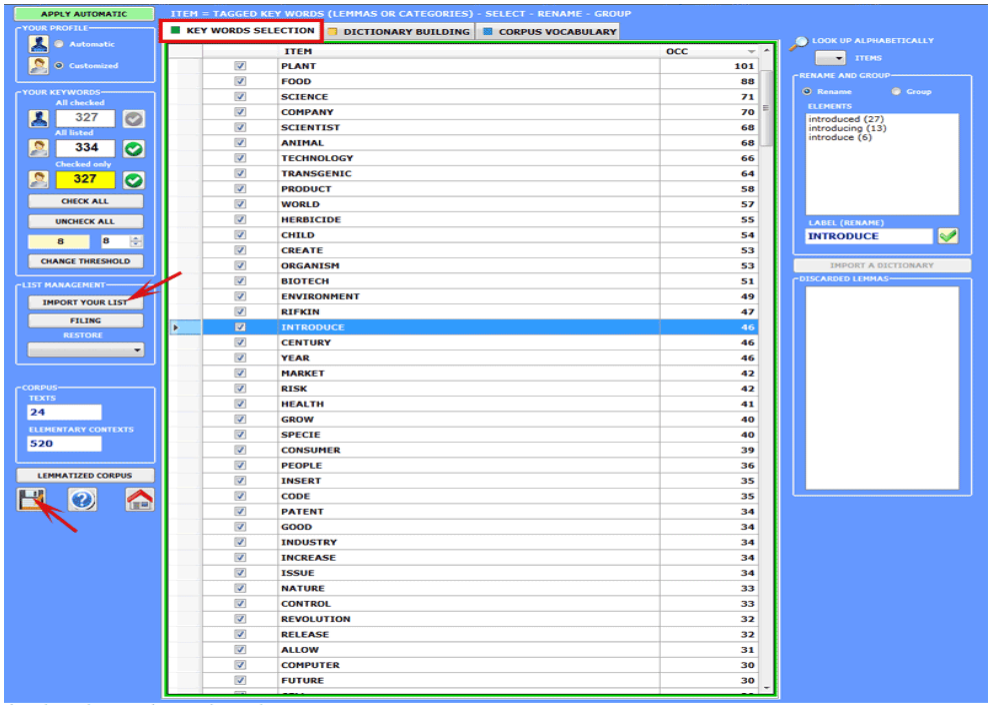
1. be sure that the "GROUP" option is selected;
2. click on two or more items of the list;
3. select one lemma or type a new label in the appropriate box;
4. click on "REPLACE".



Additional options can be enabled using the right click in the box with the items to be renamed / grouped (A) or in the box with the 'discarded lemmas' (B).
In particular, when - in the (A) case – the 'Key-Word-in-Context' option is selected, the user can automatically access the Concordances tool and check the occurrence contexts of the various items (see below).



A specific button (see below) allows you to **import customized lists of key-terms**.
Each list to be imported, named MyList.diz, can include up a maximum of 10,000 records (min = 20).
Each record of your list must be a single word with no spaces and no punctuation marks.
A model of MyList.diz file is automatically created by **T-LAB** when saving any your word list (see the appropriate button in the lower left).



The **settings** of each analysis (up to a maximum of 10) can be saved and restored. That means that the same corpus - without need of further importation - can be analysed with several dictionaries and various word selections (see the "Filing" and the "Restore" options).

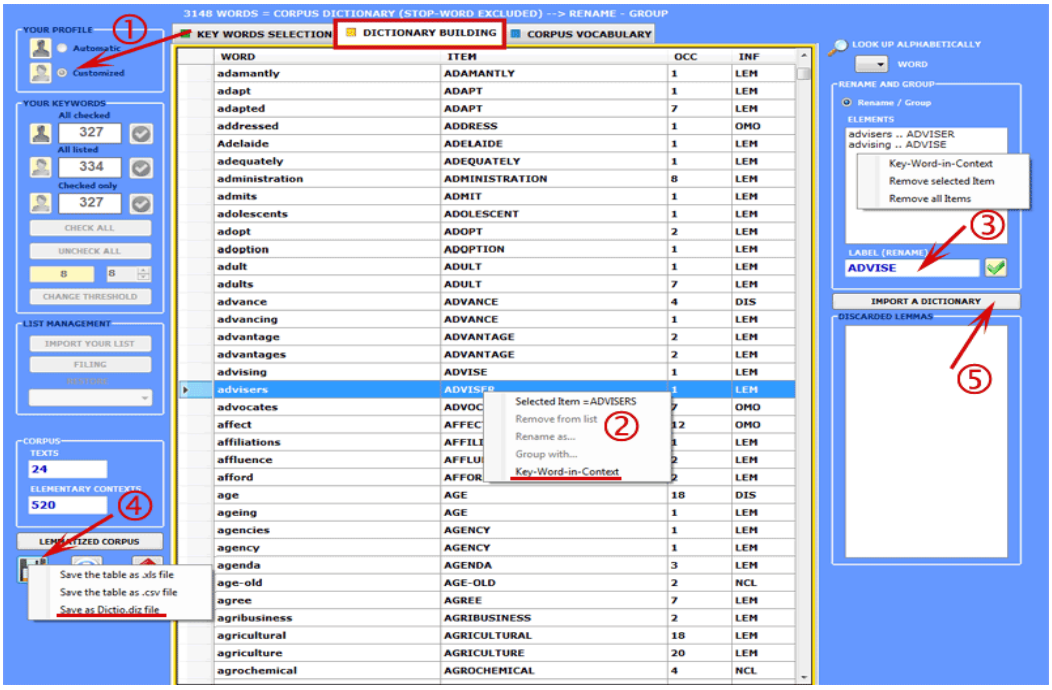
T-LAB allows customized settings to be carried out and modified over several sessions, even after operations like **Dictionary building.**

Dictionary Building

The option **Dictionary building** opens a window in which the user can carry out some operations on the **corpus** dictionary.

The user can rename or group the avalaible **lemmas** (see option '3' below); furthermore he can export the dictionary (see option '4' below); or import a **customized dictionary** (see option '5' below).

The starting point is a table (the **Corpus Dictionary**) that reports the following information:
- word/lemma correspondences;
- word occurrences;
- some labels which refer to the **automatic lemmatization** (see the **INF** column).



Before any intervention, by selecting a specific word and by using the right click of the mouse, the user can check the concordances (Key-Word-in-Context) which interests him (see the above option '2'). In any case, after clicking the "keyword selection" tab, the customized settings must be selected (see the above option '1').

The **possible operations**, even though different in their goals (revision of the lemmatizations and/or applications of grids for content analysis), all give a reorganization of the **T-LAB** database, thus creating different tables used to analyse data. Therefore all operations must be done for the words (lemmas or categories) considered to be interesting for the subsequent analyses. **T-LAB**, in fact, makes a further option available, **Key Words Selection**, with which users can decide which lemmas to "keep" and which to "discard".

The two functions (Dictionary Building and Key Words Selection) are strongly interconnected and the user can easily move from one to the other, also in order to change one's choices.

In **Dictionary building** there are **two operating modes**:

- one which allows you to move the selected words (just click) to the box on the right and, afterwards, re-denominating them by using the option "replace" (N.B.: In this case, the new label can be chosen from the selected lemmas. See the above option '3') or by typing a new label in the appropriate box;
- the other by using the "import a dictionary" option when the user intends to apply his list for classifying the words (see the above option '5').

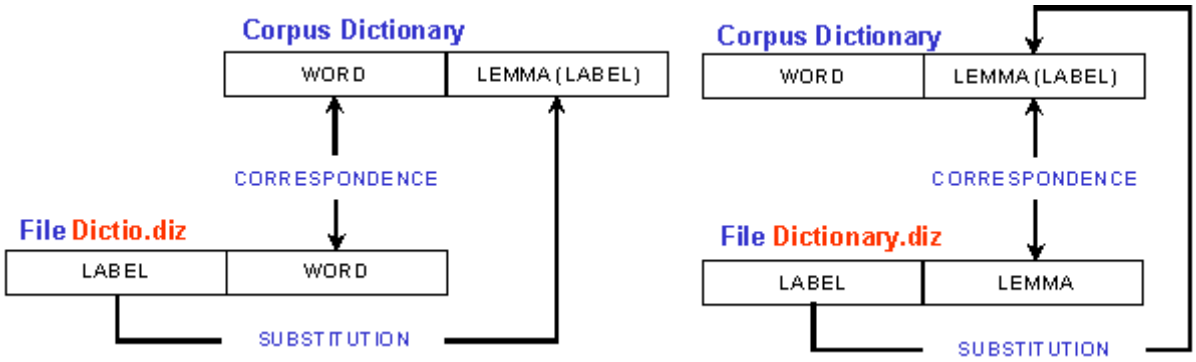
N.B.: The right-click in the Rename / Group box enables a context menu which allows three operations: a) verify the concordances (Key-Word-in-Context) of the selected item; b) remove the selected item from the box; c) remove all selected items from the box.

In order to import a **customized dictionary**, it is required that the user has set up a **Dictio.diz** or a **Dizionary.diz** file.
These files can be made up of "n" lines, each with a couple of strings separated by the character ";".
The maximum length of a string (word, lemma or category) is 50 characters: no blank spaces must be included.
For each couple, the first string - on the left - indicates the label (lemma or the category) defined by the user, the second indicates the corresponding word (**Dictio.diz** case) or lemma (**Dictionary.diz** case) already present in **T-LAB** dictionary**.**

These are some examples:

(File **Dictio.diz**)	(File **Dictionary.diz**)
ACCEPT;accept ACCEPT;accepted ACCEPT;accepting ACCEPT;accepts ------ CHILD;child CHILD;children WOMAN;woman WOMAN;women	BIOTECH;biotech BIOTECH;biotechnology --- ABSTRACT_TOUGHT;distinctness ABSTRACT_TOUGHT;distinguish ABSTRACT_TOUGHT;diversification ABSTRACT_TOUGHT;diversif

According to the type of file you import, the changes will be as follows:

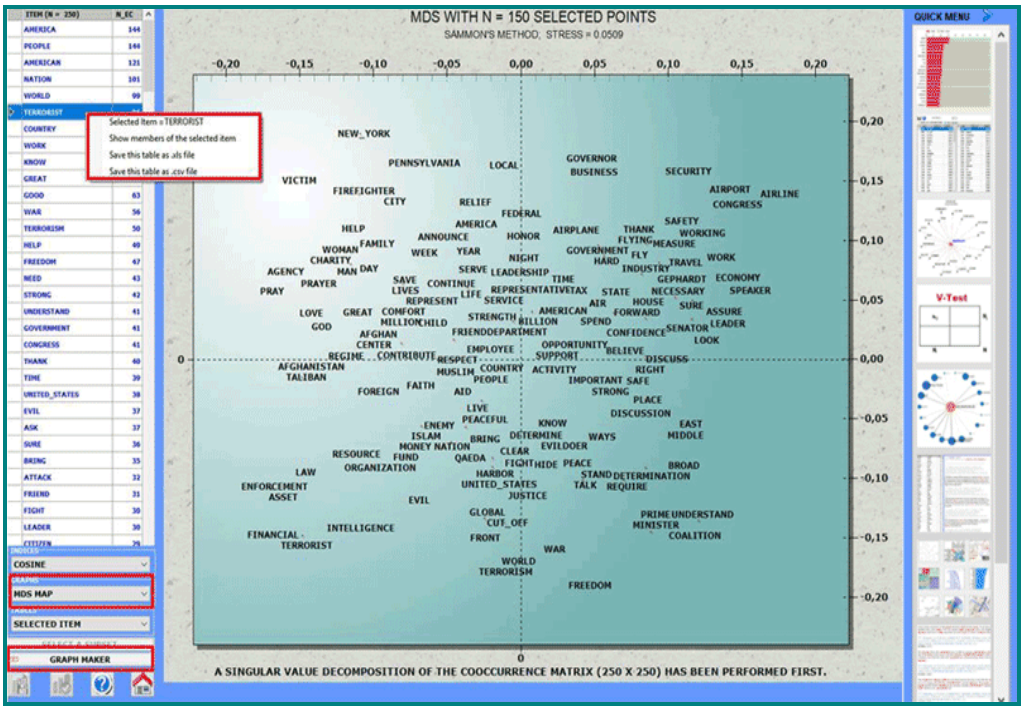


N.B. :
- Using the option **Lemmatized Corpus** it is possible to export a copy of the corpus (.txt file) in which every word will be replaced by the corresponding lemma or category;
- When the dictionary has been modified, the following analyses on the same corpus are available only as "customized settings".

CO-OCCURRENCE ANALYSIS

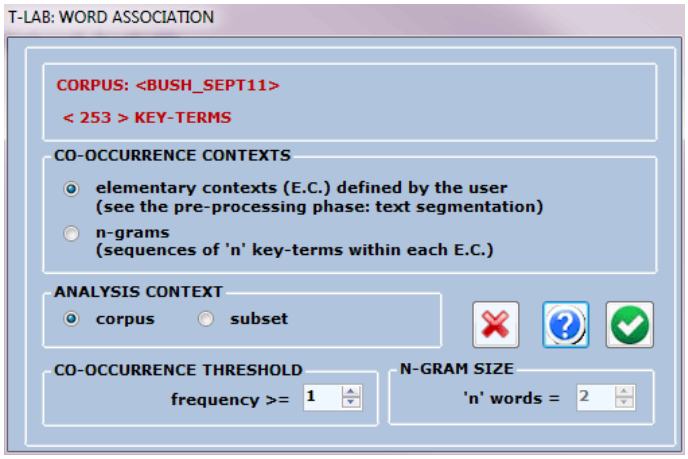
Word Associations

N.B.: The pictures shown in this section have been obtained by using a previous version of **T-LAB**. These pictures look slightly different in **T-LAB 10**. Also: a) there is a new option which allows the user to plot a **MDS Map Overview** with the most relevant words; b) there is a new button (**GRAPH MAKER**) which allows the user to create several dynamic charts in HTML format; c) by **right clicking** on the keyword tables , additional options become available; d) a quick access gallery of pictures which works as an **additional menu** allows one to switch between various outputs with a single click. Some of these new features are highlighted in the below image.



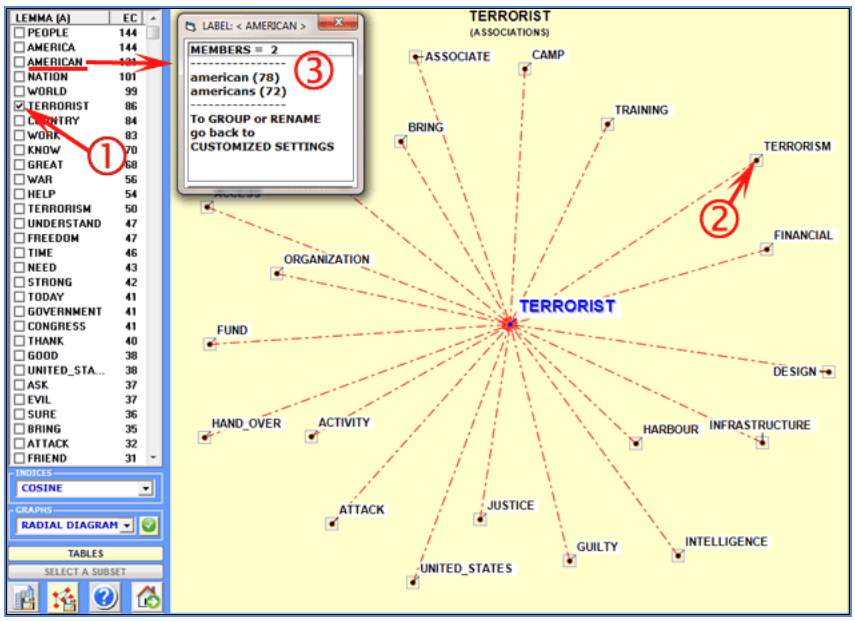
This **T-LAB** tool allows us to pick out **co-occurrence** and **similarity** relationships which, within any corpus or its subset, determine the local meaning of selected **key-terms**.

Such a tool can be used with the default options (A) or through options selected by the user (B).



In the first case (A: default) word co-occurrences are computed within the **elementary contexts** (e.g. sentences, fragments, paragraphs). In the second one (B: options selected by the user) word co-occurrences can also be computed within **n-grams** (i.e. sequences of two or more words) and the user is also enabled to choose the minimum threshold of co-occurrences to be considered.

The working window (see below) is made available immediately after the computation of co-occurrences between all the words included in the list selected by the user has been done.



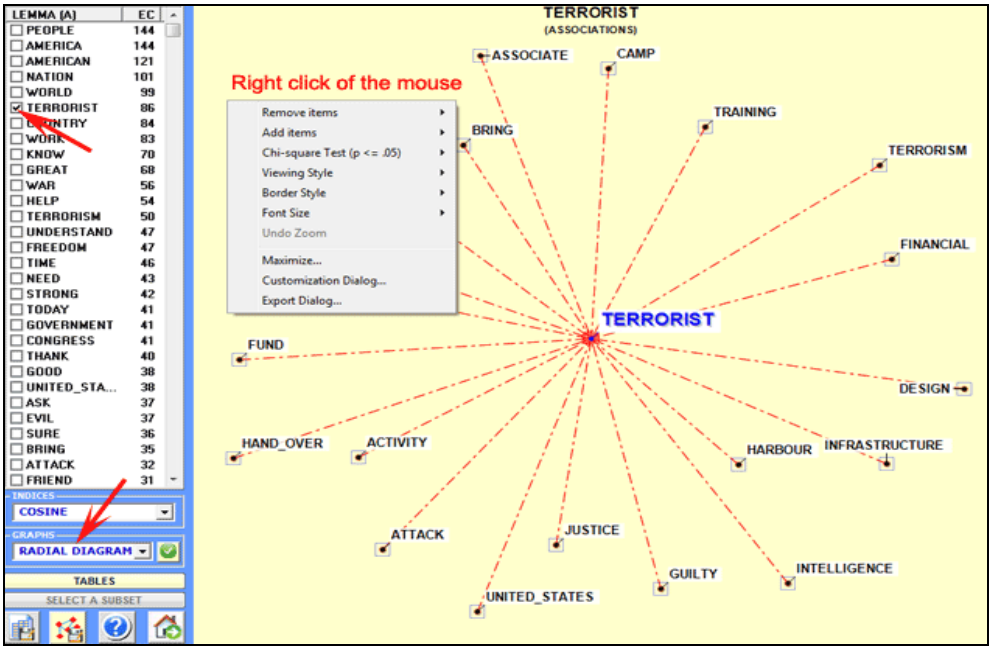
On the left of the above window there is a table with the key-term list and numerical values indicating the number of elementary contexts (EC) or n-grams where each key-term is present.

Either by clicking a item in the table (see '1' above) or by clicking on a point of the charts (see '2' above) it is possible to check the associations of each target word. Moreover, by clicking any item of the table (see '3' above') it is possible to check which words are included in the corresponding lemma or semantic class.

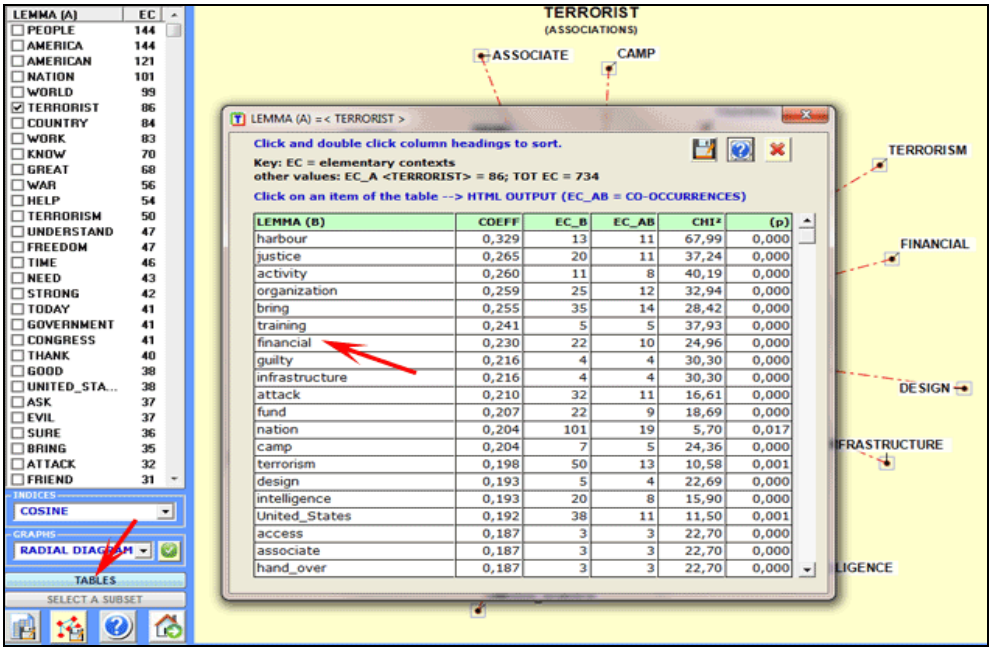
Each time the selection of associated words is carried out by the computation of an **Association Index** (see the corresponding item of the glossary) or by the computation of second order similarities (see the note at the end of this page). In the first case the available indexes are six (Cosine, Dice, Jaccard, Equivalence, Inclusion and Mutual Information) and their computation is quite fast. In the second case (i.e. second order similarities), as the computation requires lots of comparisons, it can take a number of minutes. Moreover the user has to take into account that the greater the number of words included in his list, the more reliable the similarity values become.

For each query, **T-LAB** produces graphs and tables. Both graphs and tables can be saved using the appropriate buttons.

In the **radial diagrams** the lemma selected is placed in the center. The others are distributed around it, each at distance proportional to its degree of association. The significant relationships are therefore one-to-one, to the central lemma and to each of the others.
Each click on a item produces a new chart and, by using the right click of the mouse, it is possible to to open a dialog box which allows several customizations (see below).



Tables reporting various measures allow us to check the relationships between occurrences and co-occurrences concerning the words (up to 50) that are most associated to the target ones.



The reading keys are as follows:

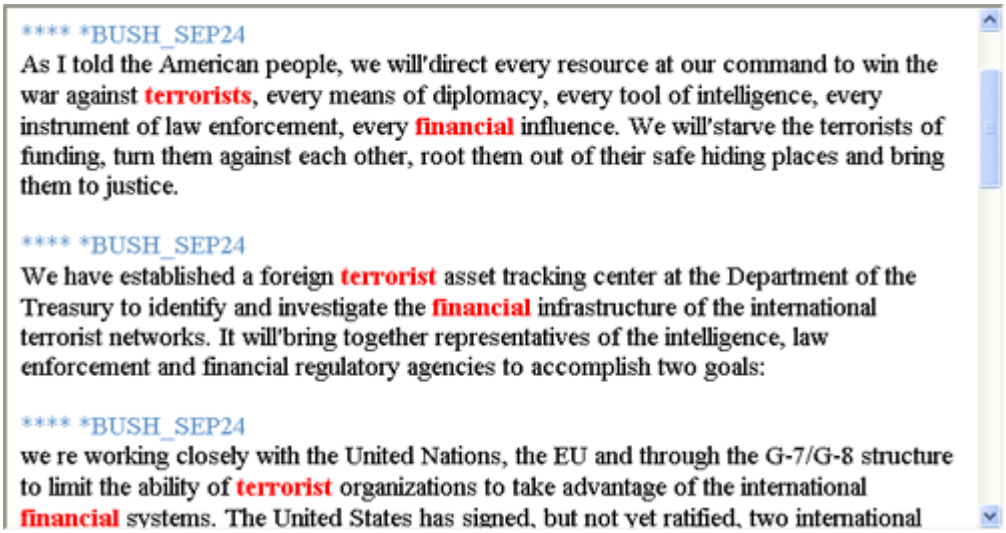
- LEMMA (A) = selected lemma;
- LEMMA (B) = lemmas associated with LEMMA (A);
- COEFF = value of the selected index;
- TOT EC = total amount of elementary contexts (EC) or n-grams in the corpus or in the analysed subset;
- EC_A = total amount of EC that contains the selected lemma (A);
- EC_B = total amount of EC that contains every associated lemma (B);
- EC_AB = total amount of EC where lemmas "A" and "B" are associated (co-occurrences);
- CHI2 = chi square value concerning the co-occurrence signifiance;
- (p) = probability associated with the chi square value (def=1).

In the case of **chi square** test, for each couple of lemmas ("A" and B") the structure of the analysed table is the following:

		LEMMA "B"		
		+	-	
LEMMA "A"	+	n_{ij}		N_j
	-			
		N_i		N

Where : n_{ij} = EC_AB; N_j = EC_A; N_i = EC_B; N = TOT EC.

A click on each table item (e.g. 'financial') allows us to save a HTML file with all the elementary contexts (i.e. sentences or paragraphs) where the selected lemma co-occurs with the central word (e.g. 'financial' and 'terrorist').

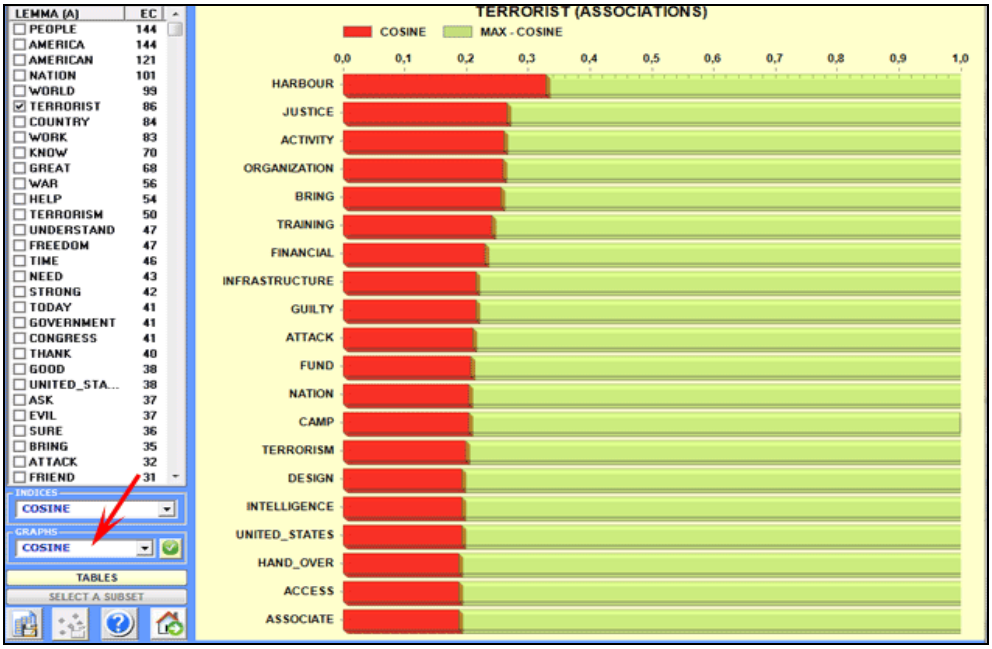


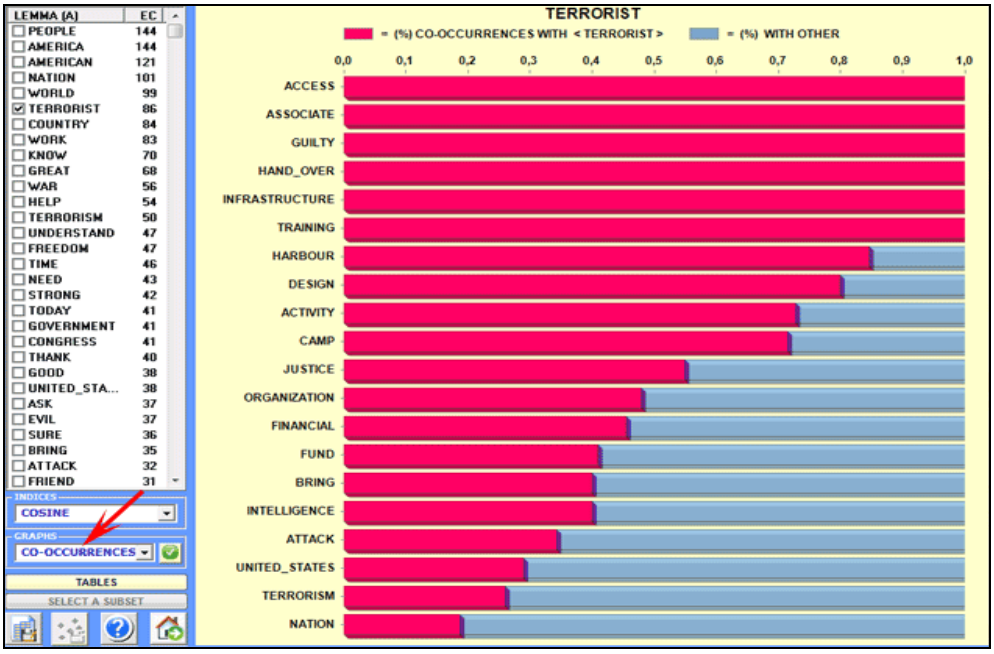
**** *BUSH_SEP24
As I told the American people, we will'direct every resource at our command to win the war against **terrorists**, every means of diplomacy, every tool of intelligence, every instrument of law enforcement, every **financial** influence. We will'starve the terrorists of funding, turn them against each other, root them out of their safe hiding places and bring them to justice.

**** *BUSH_SEP24
We have established a foreign **terrorist** asset tracking center at the Department of the Treasury to identify and investigate the **financial** infrastructure of the international terrorist networks. It will'bring together representatives of the intelligence, law enforcement and financial regulatory agencies to accomplish two goals:

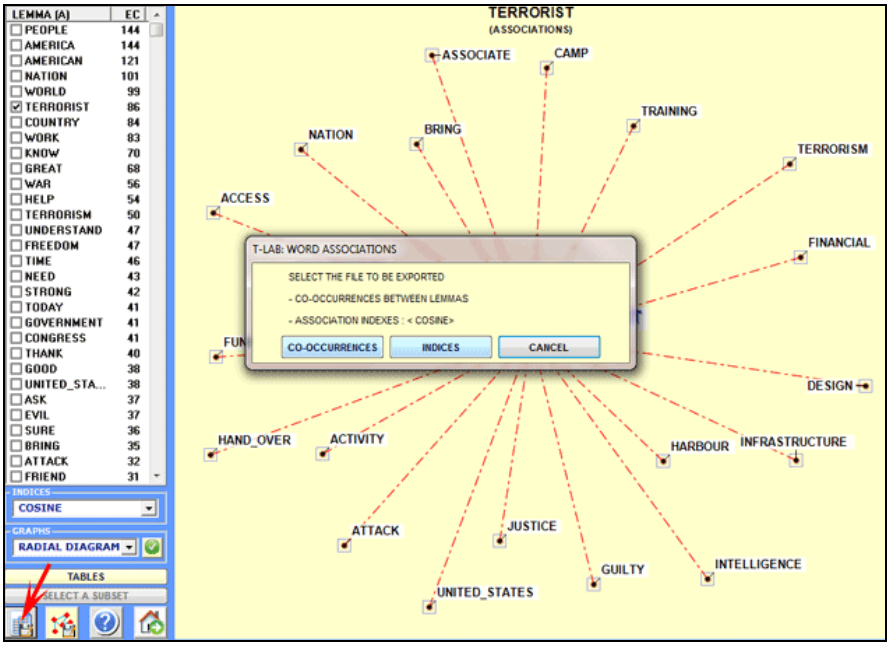
**** *BUSH_SEP24
we re working closely with the United Nations, the EU and through the G-7/G-8 structure to limit the ability of **terrorist** organizations to take advantage of the international **financial** systems. The United States has signed, but not yet ratified, two international

Further graphs (bar charts) allow us to appreciate the values of the **coefficient** used and the **percentage** of co-occurrence contexts (see below).





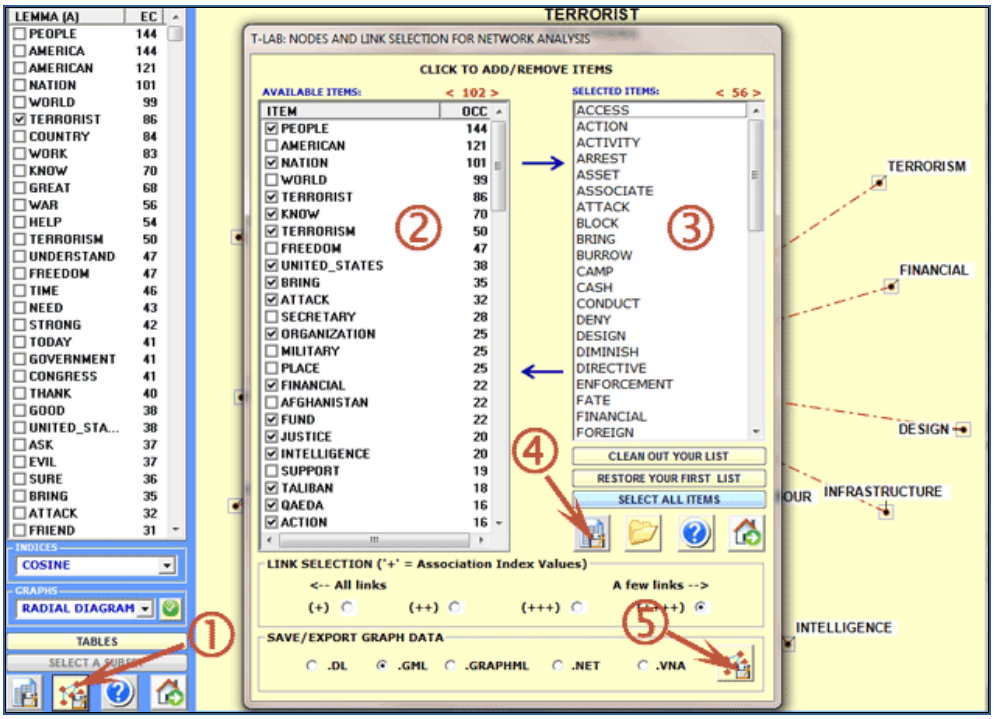
By clicking the button at the bottom left, the user can export various types of tables (see the picture below).



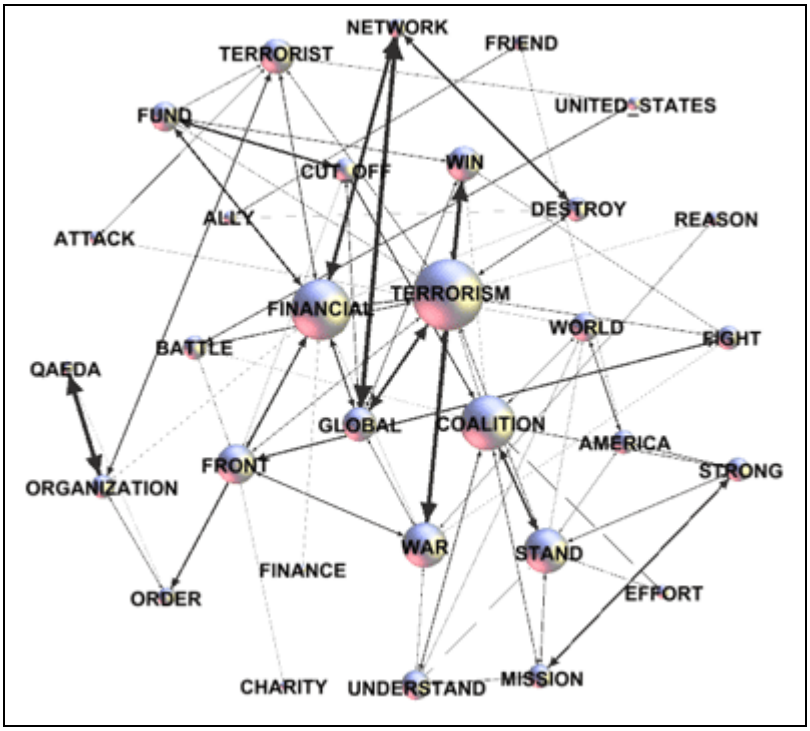
A specific **T-LAB** window (see the picture below) allows us to create various files which can be edited by software for **network analysis** (e.g. Gephi, Pajek, Ucinet, yEd and others). In this case the **nodes** are words associated with the target key-term; so each time it is possible to map the local network of such a term. The available options are the following: select the words (i.e. the 'nodes') to be inserted into the graph (see steps 2 and 3 below), export the

corresponding adjacency matrix (see step 4 below), export the selected graphical file (see step 5 below).

N.B.: In **T-LAB 10** the following window has been replaced by the **GRAPH MAKER** tool.



For example, .gml files exported by **T-LAB** can allow us to create graphs like the following.



N.B.: The above graph has been created by means of Gephi (https://gephi.org/), which is an open source software.

The way **T-LAB** computes the association (or proximity) indexes is illustrated in the corresponding section of the Manual/Help (see the glossary). All these 'first order' indexes are obtained through a normalization of the co-occurrence values concerning word pairs; so, in such computation, two words which never co-occur have an association index equal to '0' (zero). Differently, the 'second order' indexes highlight similarity phenomena which are not directly related to co-occurrences between word pairs; in fact, in this case, two words which never co-occur can nevertheless have a high similarity index.
By making reference to structural linguistics, we could say that 'first order' indexes point out phenomena concerning the sintagmatic axis ('in praesentia' combination and proximity, i.e. each word 'near to' the other), whereas 'second order' indexes point out phenomena concerning the paradigmatic axis ('in absentia' association and similarity, i.e. quasi-synonymity between key-terms used within the same corpus).
In order to understand how **T-LAB** computes 'second order' similarities it is useful to recall that all 'first order' indexes can be gathered in proximity matrices like the following (Matrix 'A').

	w_01	w_02	w_03	w_04	w_05	w_06	w_07	w_08	w_09	w_10
w_01	0,000	0,006	0,052	0,000	0,002	0,050	0,031	0,015	0,041	0,063
w_02	0,006	0,000	0,014	0,000	0,001	0,006	0,001	0,022	0,002	0,022
w_03	0,052	0,014	0,000	0,024	0,092	0,139	0,018	0,117	0,064	0,373
w_04	0,000	0,000	0,024	0,000	0,004	0,004	0,000	0,003	0,002	0,013
w_05	0,002	0,001	0,092	0,004	0,000	0,026	0,000	0,017	0,007	0,055
w_06	0,050	0,006	0,139	0,004	0,026	0,000	0,020	0,063	0,044	0,270
w_07	0,031	0,001	0,018	0,000	0,000	0,020	0,000	0,001	0,007	0,016
w_08	0,015	0,022	0,117	0,003	0,017	0,063	0,001	0,000	0,007	0,208
w_09	0,041	0,002	0,064	0,002	0,007	0,044	0,007	0,007	0,000	0,046
w_10	0,063	0,022	0,373	0,013	0,055	0,270	0,016	0,208	0,046	0,000

Matrix 'A' - First Order Similarities

In the above 'A' symmetric matrix the values in yellow (0.373) correspond to the highest 'first order' similarity between the selected words and indicate the association between words 'w_03' and 'w_10'. More specifically, 0.373 is an equivalence index obtained by dividing their squared co-occurrences by the product of their occurrences (360^2/627*553).

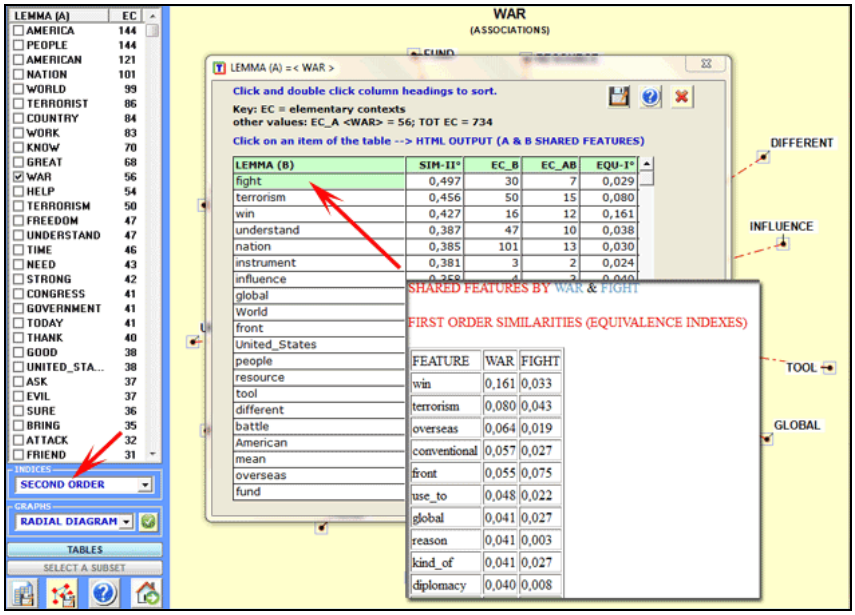
Starting from the above 'A' matrix, T-LAB builds a second matrix (see 'B' below) obtained by computing all cosine coefficients between all 'A' columns. For example, in matrix 'B' below the highest similarity index (the one in green: 0.905) has been obtained by computing the cosine coefficient between the corresponding columns of the 'A' matrix (i.e. w_06 and w_10), the 'first order' similarity of which is quite low (0.063).

	w_01	w_02	w_03	w_04	w_05	w_06	w_07	w_08	w_09	w_10
w_01	0.000	0.581	0.674	0.564	0.694	0.679	0.724	0.647	0.675	0.616
w_02	0.581	0.000	0.784	0.663	0.727	0.820	0.536	0.755	0.665	0.660
w_03	0.674	0.784	0.000	0.548	0.602	0.844	0.553	0.804	0.652	0.407
w_04	0.564	0.663	0.548	0.000	0.863	0.751	0.438	0.779	0.690	0.711
w_05	0.694	0.727	0.602	0.863	0.000	0.807	0.573	0.824	0.770	0.782
w_06	0.679	0.820	0.844	0.751	0.807	0.000	0.593	0.905	0.740	0.496
w_07	0.724	0.536	0.553	0.438	0.573	0.593	0.000	0.580	0.752	0.620
w_08	0.647	0.755	0.804	0.779	0.824	0.905	0.580	0.000	0.717	0.539
w_09	0.675	0.665	0.652	0.690	0.770	0.740	0.752	0.717	0.000	0.707
w_10	0.616	0.660	0.407	0.711	0.782	0.496	0.620	0.539	0.707	0.000

Matrix 'A' - Second Order Similarities

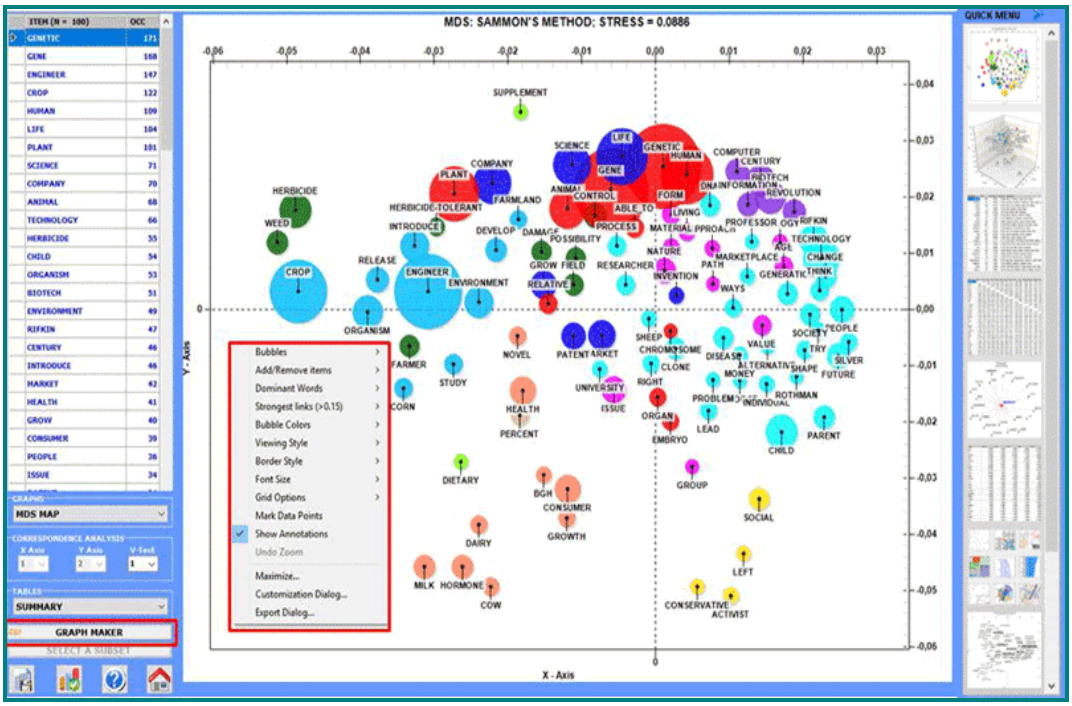
In other words, a 'first order' index is obtained by a formula which includes co-occurrence and occurrence values, whereas a 'second order' index is obtained by multiplying two normalized feature vectors.
Beyond any computational issue, we have to recall that in the above two cases ('A' and 'B' matrices) we are dealing with two very different phenomena. In fact, in the case of 'A' we are focusing on the co-occurrences between word pairs, whereas in the case of 'B' - and without any reference to their direct co-occurrences - we are focusing on the 'similarity' between feature vectors (see the matrix 'A' columns) which refers to the use (and so to the meaning) of the corresponding words.
For example, by analysing 'The Audacity of Hope' (i.e. a book written by B. Obama) it is possible to point out that - when using 'first order' measures - the word 'nuclear' is strongly associated with co-occurrent words like 'weapon', 'option', 'arms' etc.; whereas, when using 'second order' measures, 'nuclear' results strongly associated (i.e. similar) to 'destruction', even so the co-occurrence value of this word pair (i.e., 'nuclear' and 'destruction') is just '1' (one).

The tables shown by **T-LAB** allow the user to check both the second order similarities (see column SIM-II below) and the first order indexes (see column EQU-I, i.e. Equivalence Index). Moreover, by clicking any item of such a table, it is possible to generate HTML files which allow the user to check which features determine the similarity between each word pair. For example, the following table shows that the second order similarity between 'war' and 'figh' is - above all - determined by shared words like 'win', 'terrorism', etc..



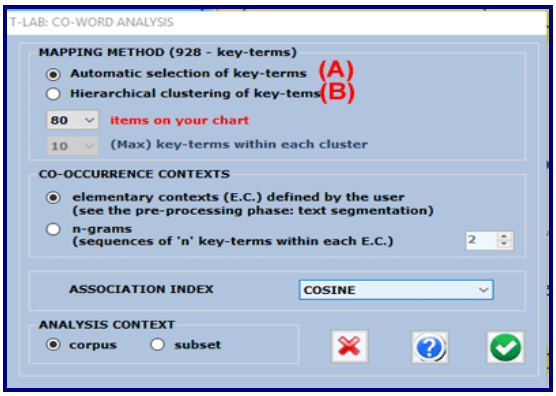
Co-Word Analysis and Concept Mapping

N.B.: The pictures shown in this section have been obtained by using a previous version of **T-LAB**. These pictures look slightly different in **T-LAB 10**. Also: a) when the 'automatic selection of key terms' is selected, different colours are used for different groups of items in the MDS map; b) the visualization technique called t-SNE (t-Distributed Stochastic Neighbor Embedding) has been added; c) there is a new button (**GRAPH MAKER**) which allows the user to create several dynamic charts in HTML format; d) by **right clicking** either the charts or the keyword tables, additional options become available; e) a quick access gallery of pictures which works as an **additional menu** allows one to switch between various outputs with a single click. Some of these new features are highlighted in the below image.



This **T-LAB** tool allows us to find and map two kinds of relationships concerning **word co-occurrences**:

A - between **single key-words** (lemmas or categories), if their number does not exceed 500 elements (min 10);
B - between/within **clusters** (i.e. **Thematic Nuclei**), if the number of **key-words** selected exceeds 100 elements (max 3,000).

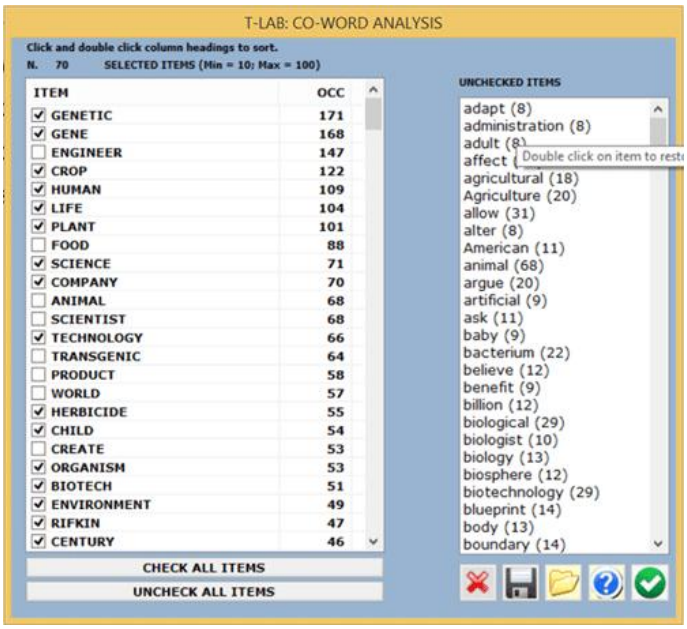


The user may choose which **association index** to be used and, for option 'B' only, he may also choose both the maximum number of clusters to be obtained (from 50 to 100) and the maximum number of key-terms within each cluster.

The computation process includes the following steps:

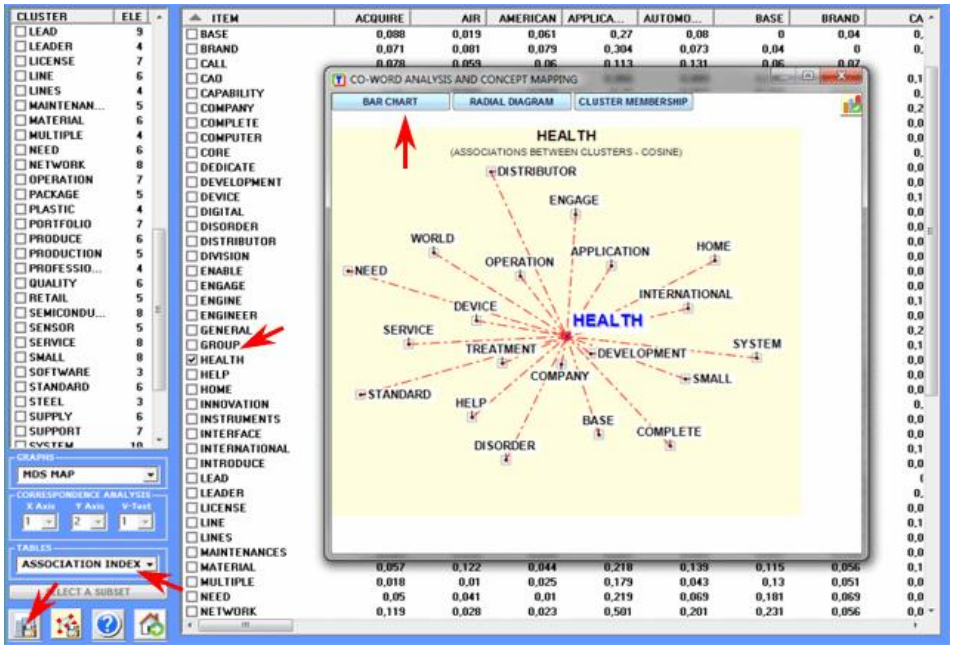
1- building a co-occurrence matrix (word x word);
2- computing the selected association indexes (Cosine, Dice, Jaccard, Equivalence, Inclusion, Mutual Information);
3- hierarchical clustering of the dissimilarity matrix;
4- building a second dissimilarity matrix (cluster x cluster);
5- graphic representation by Multidimensional Scaling and Correspondence Analysis.

N.B.:
- in 'A' cases (see the below image), the user can review the key-term selection and **T-LAB** doesn't carry out steps 3 an 4;



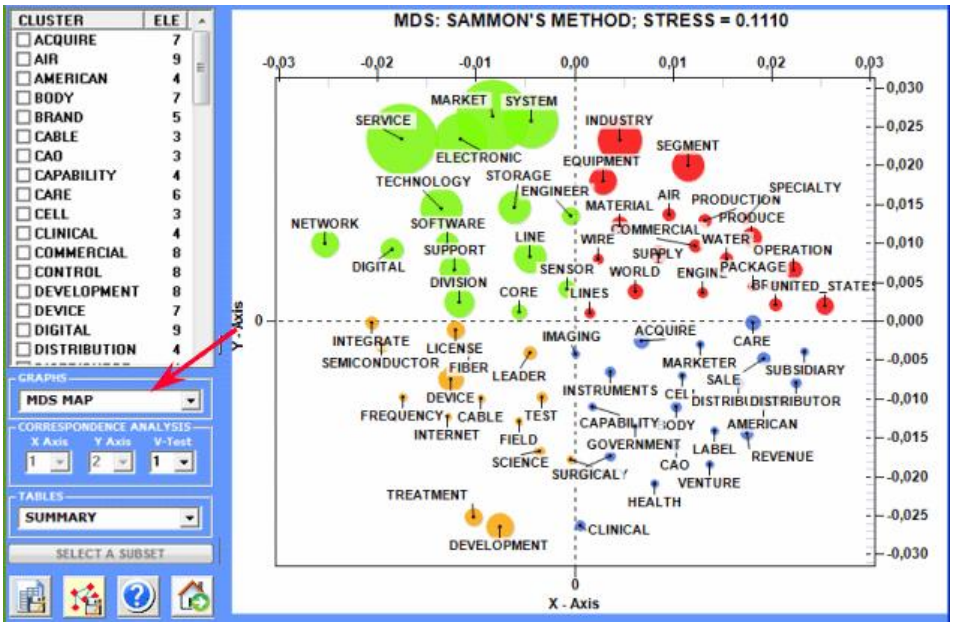
-the quality of results depends on a thorough selection of key-words;
-as the **multiwords** unclassified by **T-LAB** are specific cases of co-occurrence and the 'B' option treats them like little clusters (e.g. "Twin" + "Towers"), the user is advised to resolve these cases during the **pre-processing phase**. Anyway, without repeating the corpus importation, it is possible to make changes by means of the **Dictionary Building** function (e.g. by assigning the label "Twin_Towers" to the two different items "Twin" and "Towers");

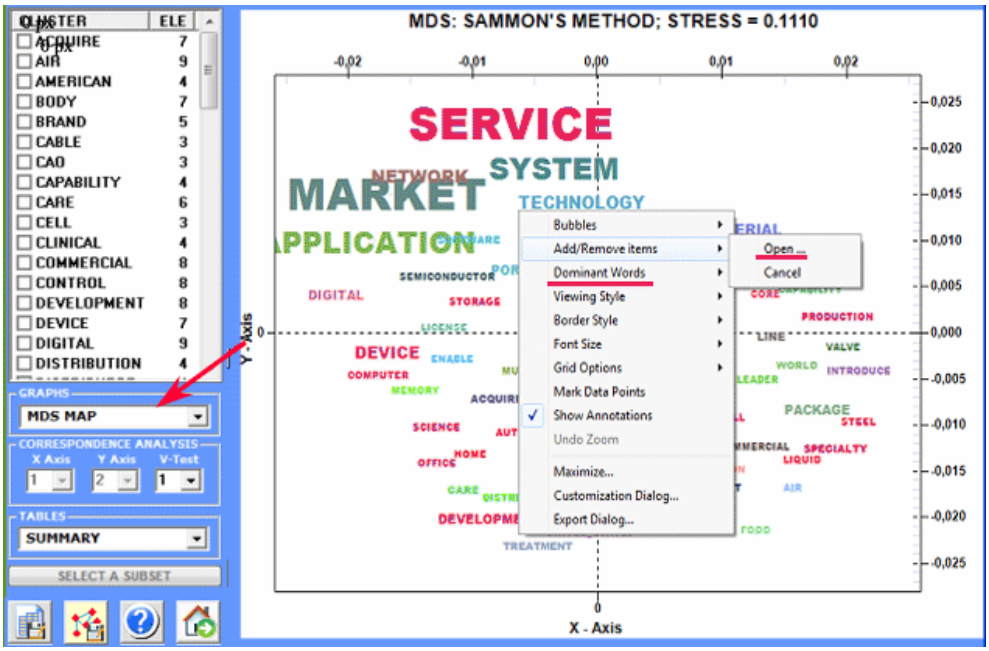
-by clicking on the appropriate buttons all data tables can be checked (see the picture below).



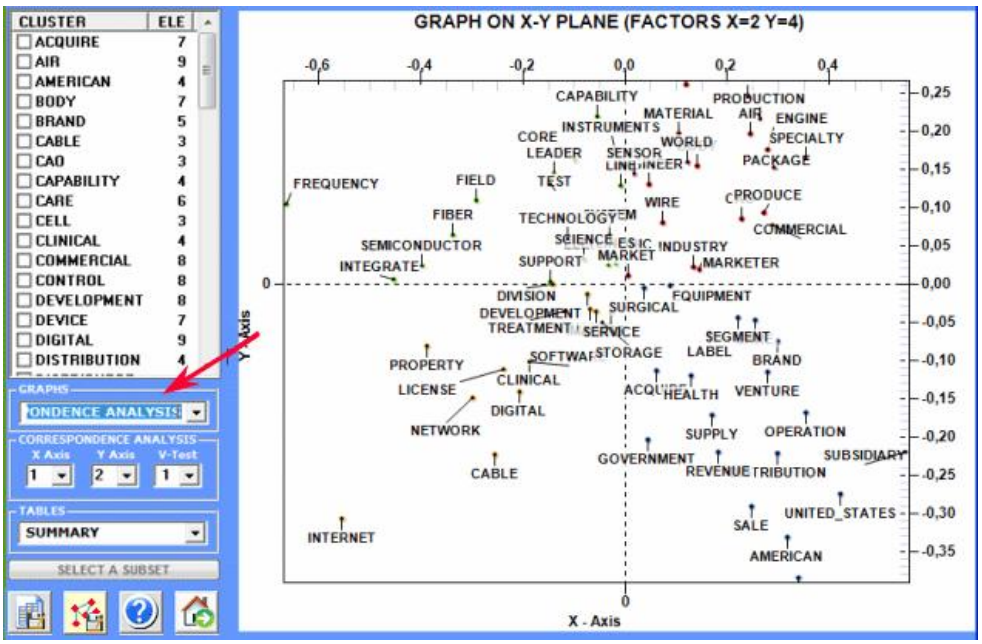
When the automatic analysis is over, four kinds of charts are available (see below) and each of them can be customized by using the appropriate dialog box (just right click on the chart).

1 - **MDS Map**

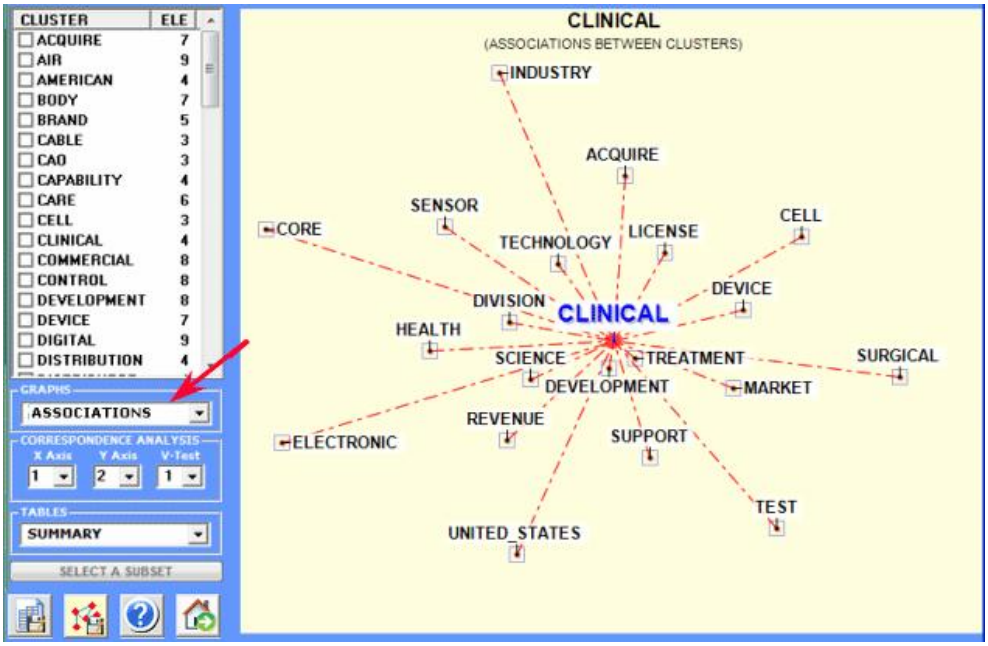




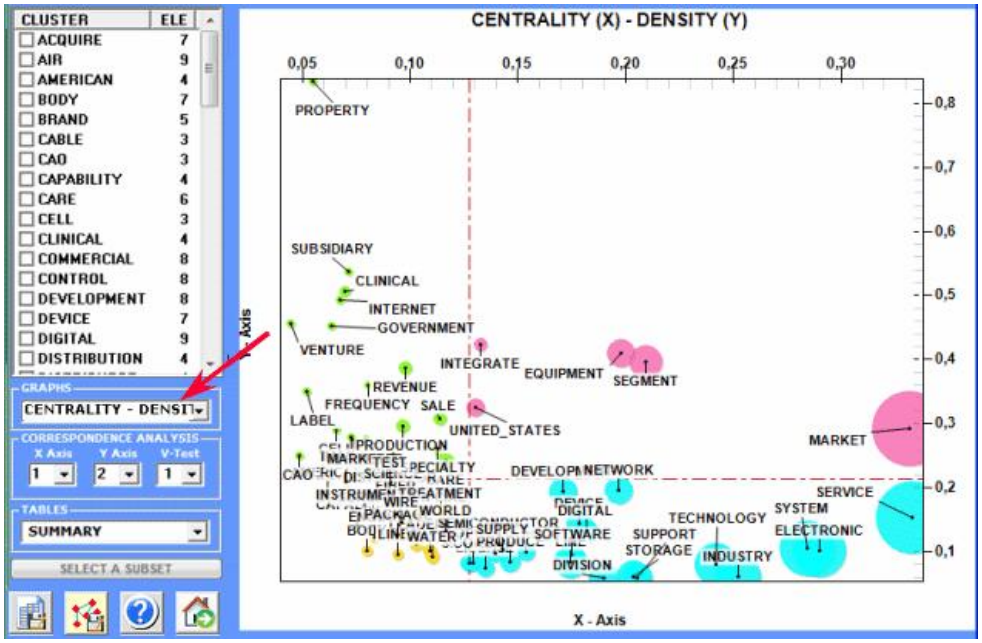
2 – Factorial Analysis of Correspondences



3 - **Association Diagram**

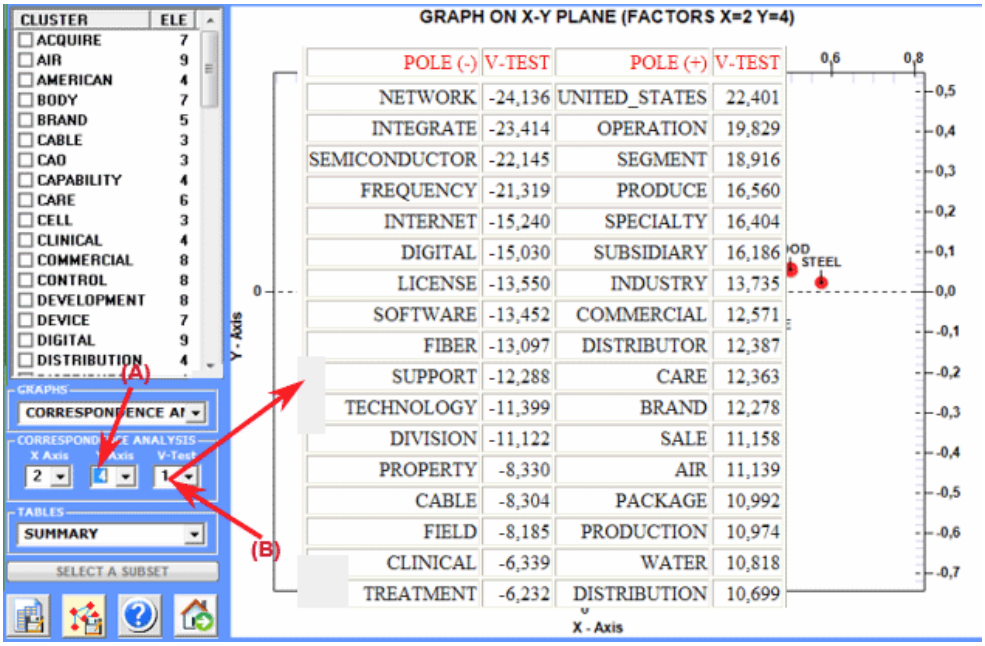


4 - Diagram of **Centrality-Density** measures (after a cluster analysis only)



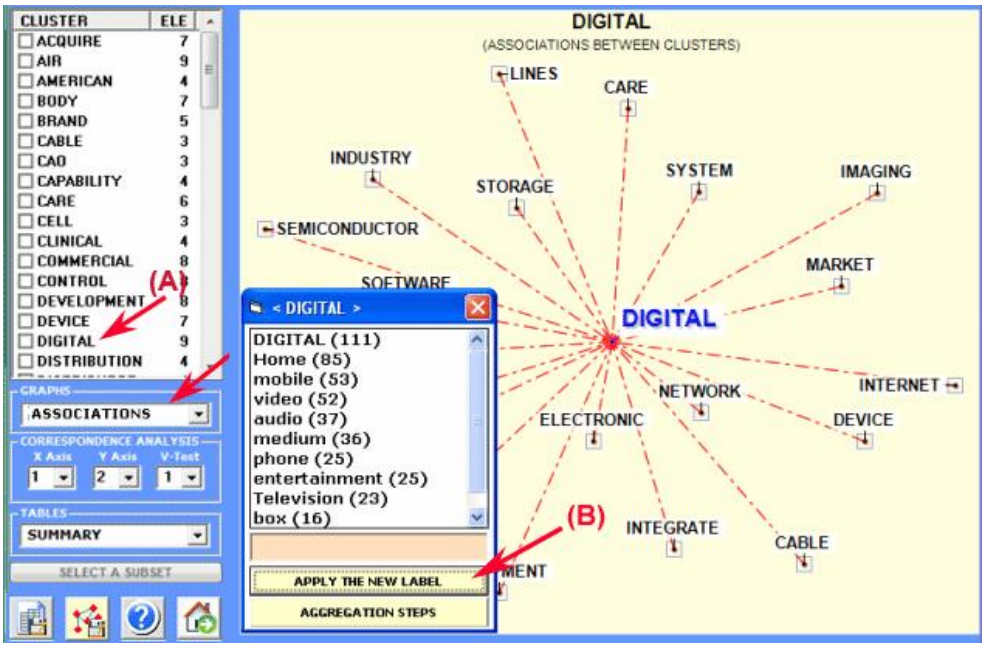
In particular, the results obtained by **Correspondence Analysis** can be mapped using the coordinates of the first ten axes (see "A" below).

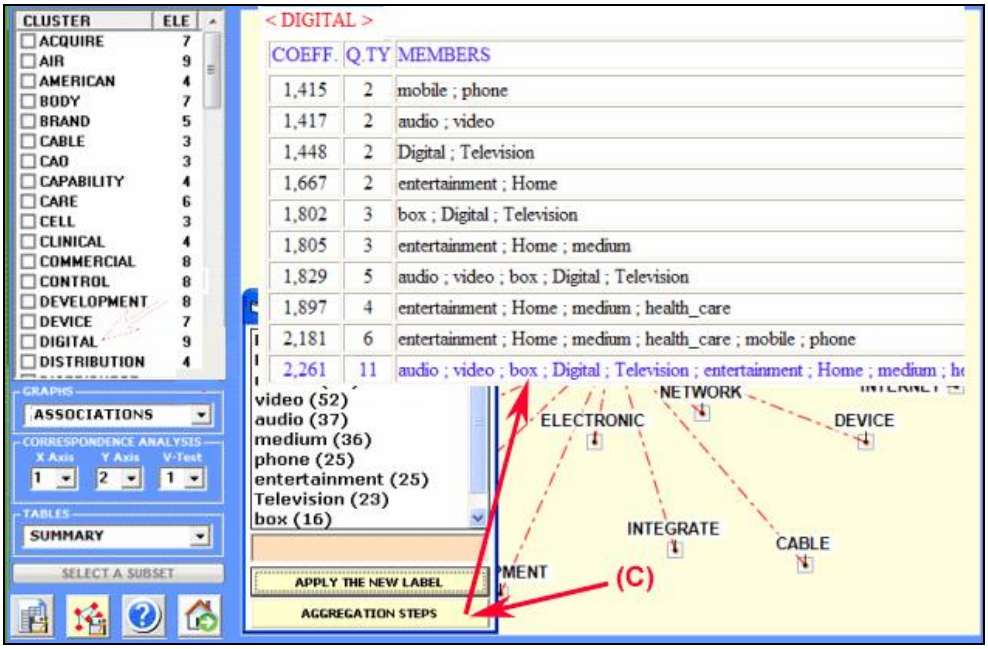
As **T-LAB** allows us to verify the Test Values of each factor (see "B" below), this kind of output can be useful for an accurate interpretation of the relationships between cluster and/or key-words.

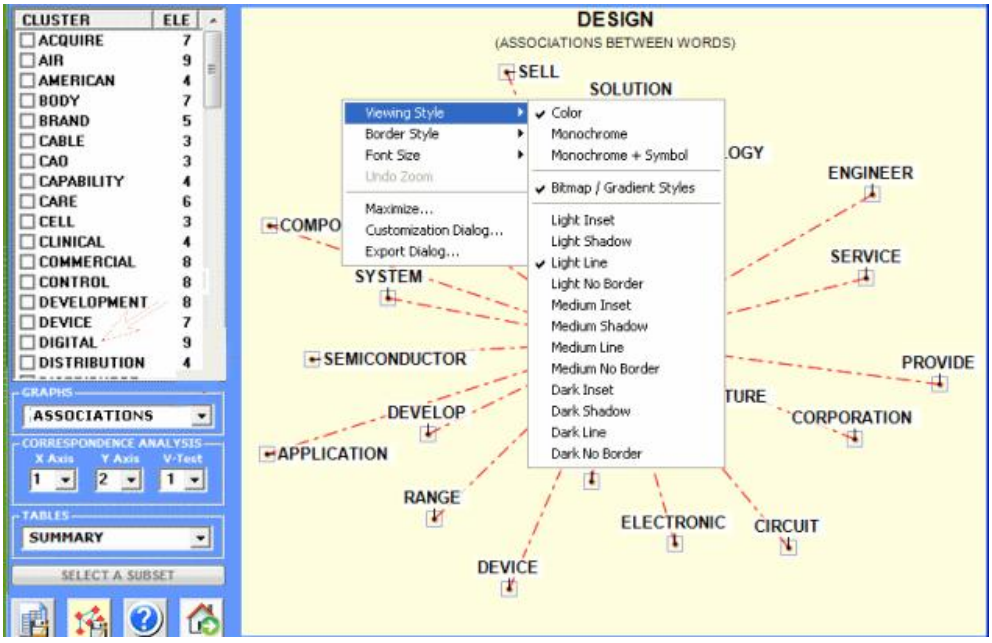


The charts can be explored and customized in the following ways:

ACTION	RESULT
click on a table item or on a chart point	diagram of corresponding associations
click on a label of "CLUSTER" column (see "A" below)	list of cluster elements
click on "apply the new label" (see "B" below)	new label assigned to the cluster
click on "aggregation steps" (see "C" below)	word aggregation within the cluster
right click on the chart	open the dialog menu

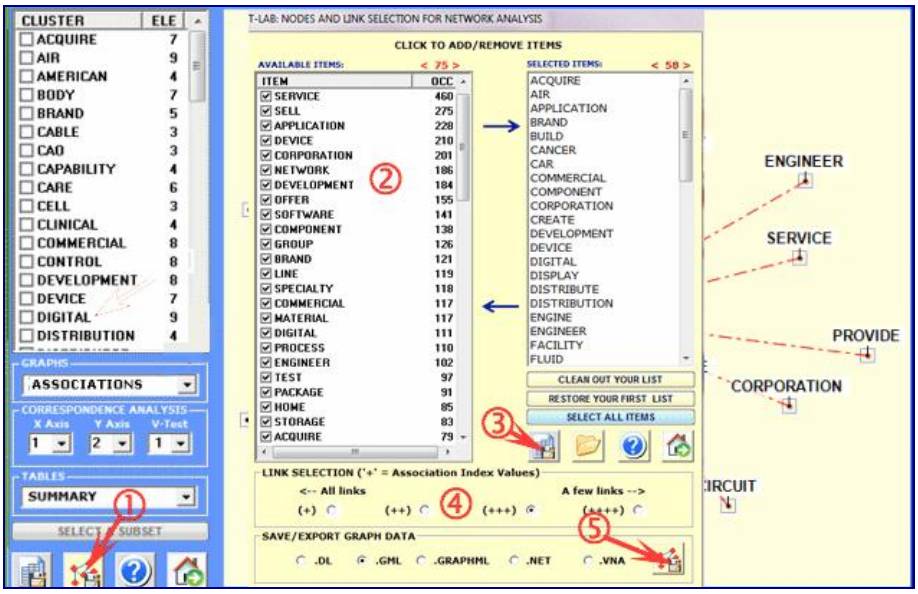

CLUSTER
ELE
ACQUIRE 7
AIR 9
AMERICAN 4
BODY 7
BRAND 5
CABLE 3
CAD 3
CAPABILITY 4
CARE 6
CELL 3
CLINICAL 4
COMMERCIAL 8
CONTROL 8
DEVELOPMENT 8
DEVICE 7
DIGITAL 9
DISTRIBUTION 4
(A)
GRAPHS
ASSOCIATIONS
CORRESPONDENCE ANALYSIS
X Axis
Y Axis
V-Test
TABLES
SUMMARY
SELECT A SUBSET
DIGITAL
(ASSOCIATIONS BETWEEN CLUSTERS)
LINES
CARE
INDUSTRY
STORAGE
SYSTEM
IMAGING
SEMICONDUCTOR
MARKET
SOFTWARE
< DIGITAL >
DIGITAL (111)
Home (85)
mobile (53)
video (52)
audio (37)
medium (36)
phone (25)
entertainment (25)
Television (23)
box (16)
DIGITAL
NETWORK
INTERNET
ELECTRONIC
DEVICE
(B)
INTEGRATE
CABLE
APPLY THE NEW LABEL
AGGREGATION STEPS


CLUSTER
ELE
ACQUIRE 7
AIR 9
AMERICAN 4
BODY 7
BRAND 5
CABLE 3
CAD 3
CAPABILITY 4
CARE 6
CELL 3
CLINICAL 4
COMMERCIAL 8
CONTROL 8
DEVELOPMENT 8
DEVICE 7
DIGITAL 9
DISTRIBUTION 4
GRAPHS
ASSOCIATIONS
CORRESPONDENCE ANALYSIS
X Axis
Y Axis
V-Test
TABLES
SUMMARY
SELECT A SUBSET
< DIGITAL >
COEFF. Q.TY MEMBERS
1,415 2 mobile ; phone
1,417 2 audio ; video
1,448 2 Digital ; Television
1,667 2 entertainment ; Home
1,802 3 box ; Digital ; Television
1,805 3 entertainment ; Home ; medium
1,829 5 audio ; video ; box ; Digital ; Television
1,897 4 entertainment ; Home ; medium ; health_care
2,181 6 entertainment ; Home ; medium ; health_care ; mobile ; phone
2,261 11 audio ; video ; box ; Digital ; Television ; entertainment ; Home ; medium ; he
video (52)
audio (37)
medium (36)
phone (25)
entertainment (25)
Television (23)
box (16)
NETWORK
ELECTRONIC
DEVICE
INTEGRATE
CABLE
APPLY THE NEW LABEL
AGGREGATION STEPS
(C)

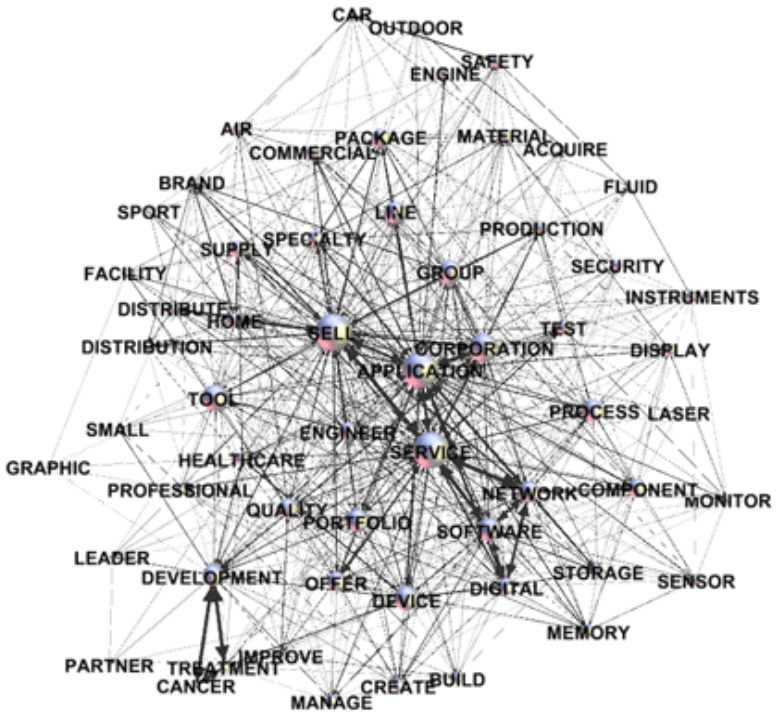


A further option allows us to select the items (i.e. the 'nodes') for **Network Analysis** (see the image below, step 1 and 2), to export the corresponding adjacency matrix (step 3), select the links on the basis of their range of probability value (step 4) and export different types of files (step 5) which can be edited by software such as Gephi, Pajek, Ucinet, yEd and others.

N.B.: In **T-LAB 10** the following window has been replaced by the **GRAPH MAKER** tool.

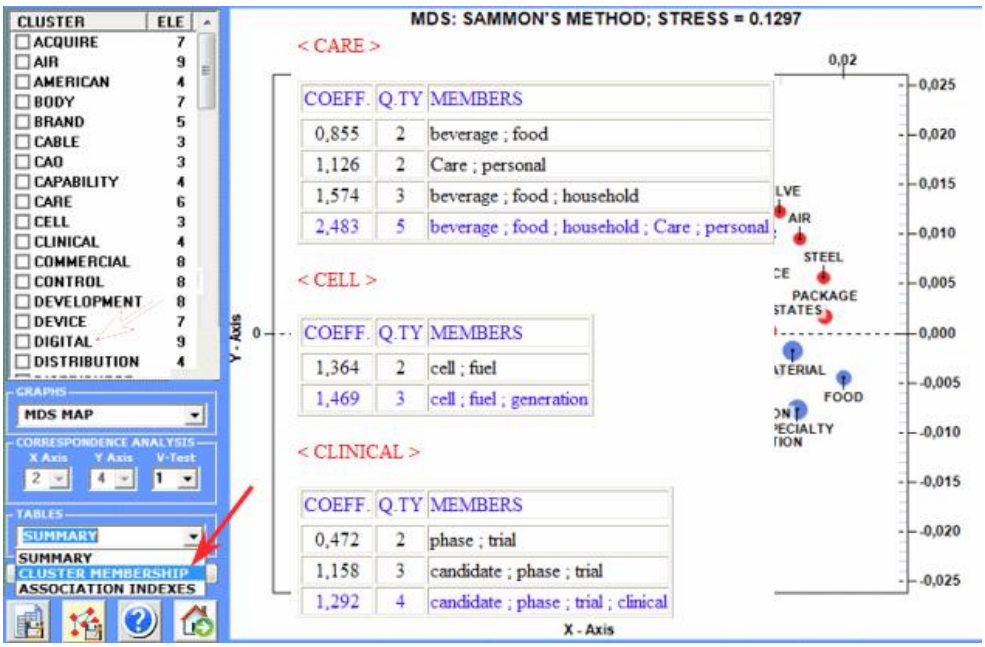


For example, files exported by **T-LAB** can allow the user to create graphs like the following.



There are available three tables which can be exported by this **T-LAB** tool:

1 - "**Cluster Membership**" table (see below) deals with the hierarchical aggregation of words within each cluster;



2 - "**Summary**" table (see below) includes the following measures:

- **ECQ** = Quantity of Elementary Contexts in which two or more word clusters are co-occurring;
- **Centrality** = average of association indexes concerning cluster relationships;
- **Density** = average of word association indexes within each cluster.

CLUSTER	ECQ	CENTRALITY	DENSITY	MEMBERS
CARE	95	0,117	0,242	BEVERAGE; CARE; FOOD; HOUSEHOLD; PERSONAL
CELL	17	0,066	0,291	CELL; FUEL; GENERATION
CLINICAL	92	0,070	0,506	CANDIDATE; CLINICAL; PHASE; TRIAL
COMMERCIAL	35	0,105	0,155	APPAREL; COMMERCIAL; FABRIC; RESIDENTIAL
CORE	77	0,129	0,083	BROAD; CORE; DEMAND; ENVIRONMENT; EXPERTISE; HIGHLY; MEET; NEED; SCIENTIFIC; WORK
DEVELOPMENT	322	0,171	0,194	BIOPHARMACEUTICAL; COMMERCIALIZATION; DEVELOPMENT; DISCOVERY; FOCUS; PHARMACEUTICAL; PROGRAM; RESEARCH
DEVICE	160	0,179	0,144	DEVICE; DIAGNOSTIC; DISPOSABLE; HEALTHCARE; HOSPITAL; LABORATORY; MEDICAL
DIGITAL	236	0,182	0,132	AUDIO; BOX; DIGITAL; ENTERTAINMENT; HEALTH_CARE; HOME; MEDIUM; MOBILE; PHONE; TELEVISION; VIDEO

3 - "**Association Indexes**" table (see below) includes measures of the between and the within cluster relationships.

Between

< ELECTRONIC >

CLUSTER	INDEX
market	0,801
system	0,786
service	0,707
Digital	0,693
industry	0,681
technology	0,633
network	0,568
segment	0,548
equipment	0,506
STORAGE	0,495
engineer	0,476
Support	0,456
semiconductor	0,449
device	0,447
software	0,412
integrate	0,384
Division	0,381
Care	0,359
supply	0,356
license	0,344
material	0,324
CORE	0,317
line	0,316
specialty	0,310
commercial	0,303

Within

< ELECTRONIC >

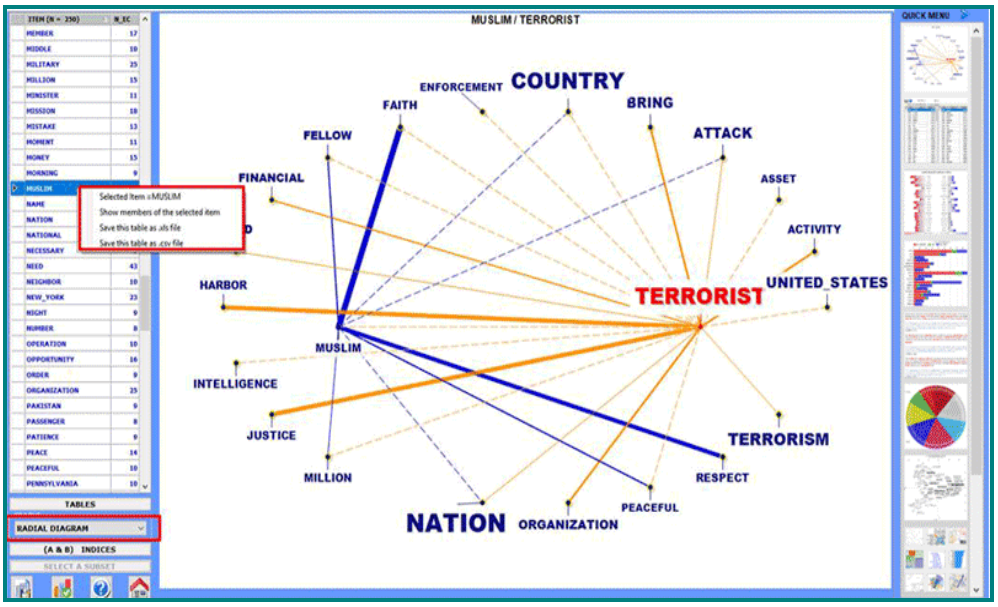
LEMMA_A	LEMMA_B	INDEX
computer	peripheral	0,279
consumer	electronic	0,279
application	industrial	0,260
computer	electronic	0,206
application	consumer	0,201
consumer	industrial	0,200
Instrumentation	interconnect	0,196
computer	consumer	0,193
application	electronic	0,175
application	variety	0,165
industrial	process	0,162
application	computer	0,152
electronic	variety	0,152
application	military	0,151
interconnect	peripheral	0,135
electronic	interconnect	0,134
industrial	Instrumentation	0,133
consumer	variety	0,133
application	process	0,127
appliance	variety	0,126
electronic	security	0,126
military	security	0,123
computer	variety	0,121
appliance	electronic	0,120
consumer	military	0,118
Instrumentation	peripheral	0,114

N.B.:
- when a Cluster Analysis has not been carried out, the "Cluster Membership" table is not available, consequently the "Summary" is simplified and the "Association Indexes" table refers to word co-occurrences only;

- when exiting from this analysis, the dictionary of Thematic Nuclei (i.e. the list of labels assigned to each word cluster) can be exported and, after a thorough revision, can be imported by means of the **Dictionary Building** function. In this way the user will be able to perform certain second order analyses (i.e. analysis concerning **themes** or **concepts**).

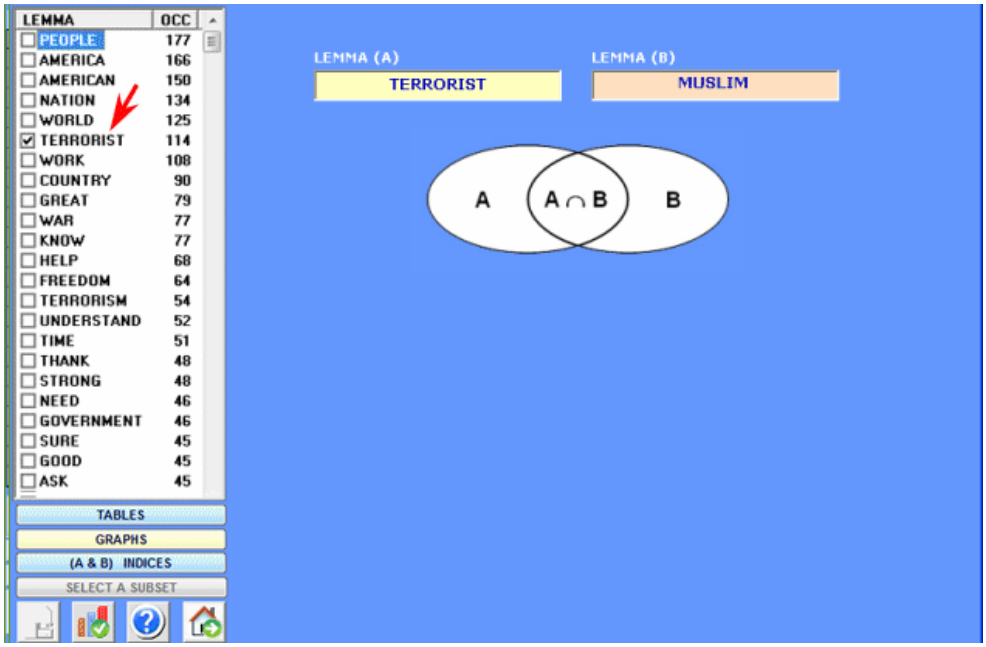
Comparison between Pairs of Key Words

N.B.: The pictures shown in this section have been obtained by using a previous version of **T-LAB**. These pictures look slightly different in **T-LAB 10**. Also: a) a **new radial diagram** is now available which allows to quickly appreciate differences in word associations; b) **right clicking** on the keyword tables makes additional options available; c) a quick access gallery of pictures which works as an **additional menu** allows one to switch between various outputs with a single click. Some of these new features are highlighted in the below image.



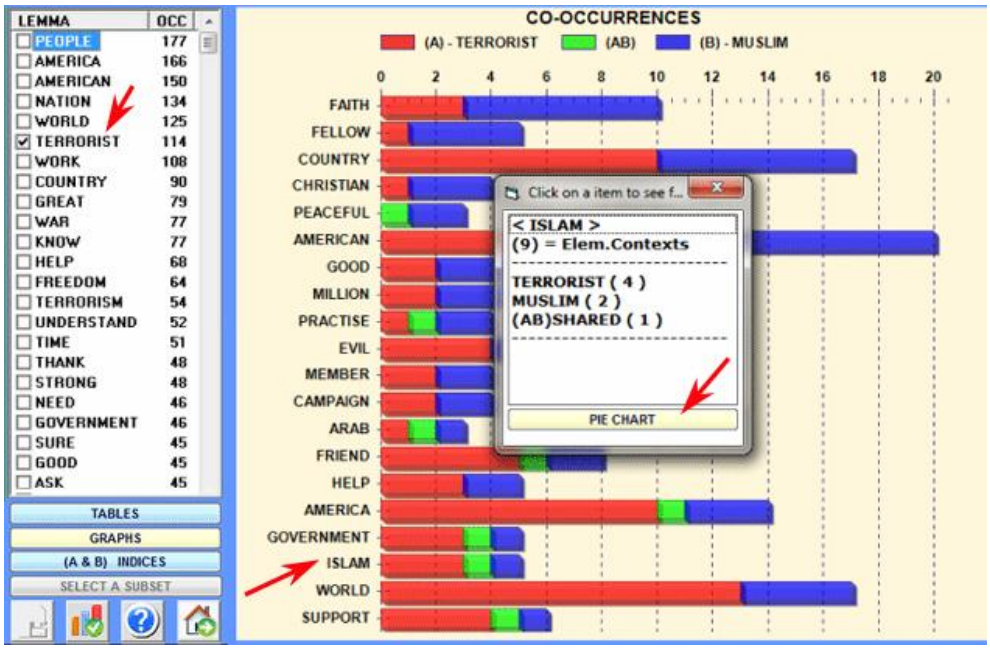
This **T-LAB** tool allows us to compare sets of **elementary contexts** (i.e. co-occurrence contexts) in which the elements of a pair of key-words are present.

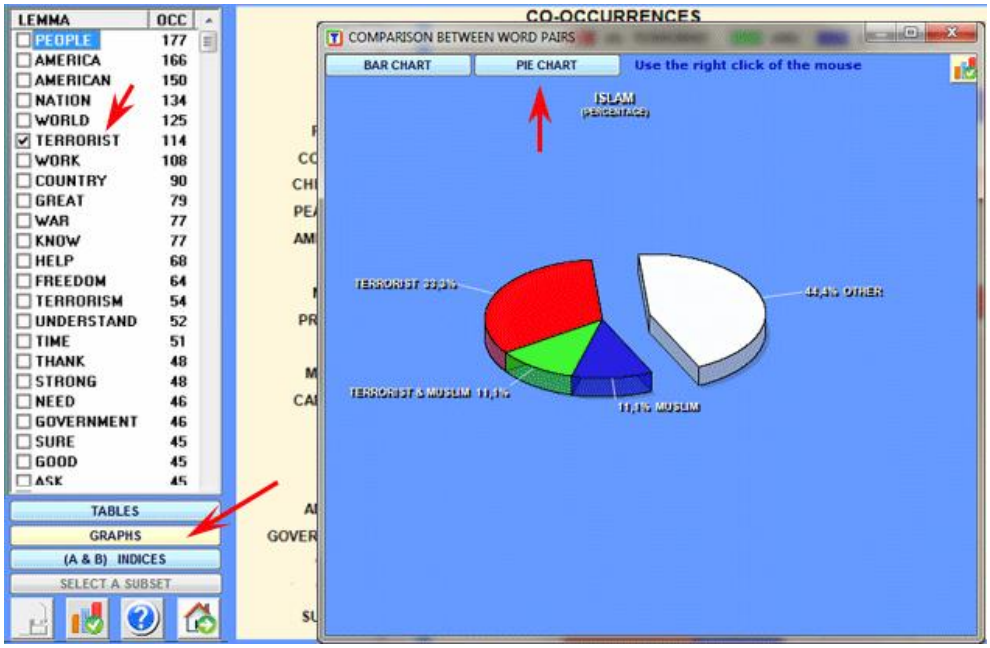
The table on the left shows the list of selected lemmas and their corresponding **occurrence** values within the whole **corpus** or a **subset** of it.
The user is invited to select - one after the other - two of these (a "pair") with a single click (see below).

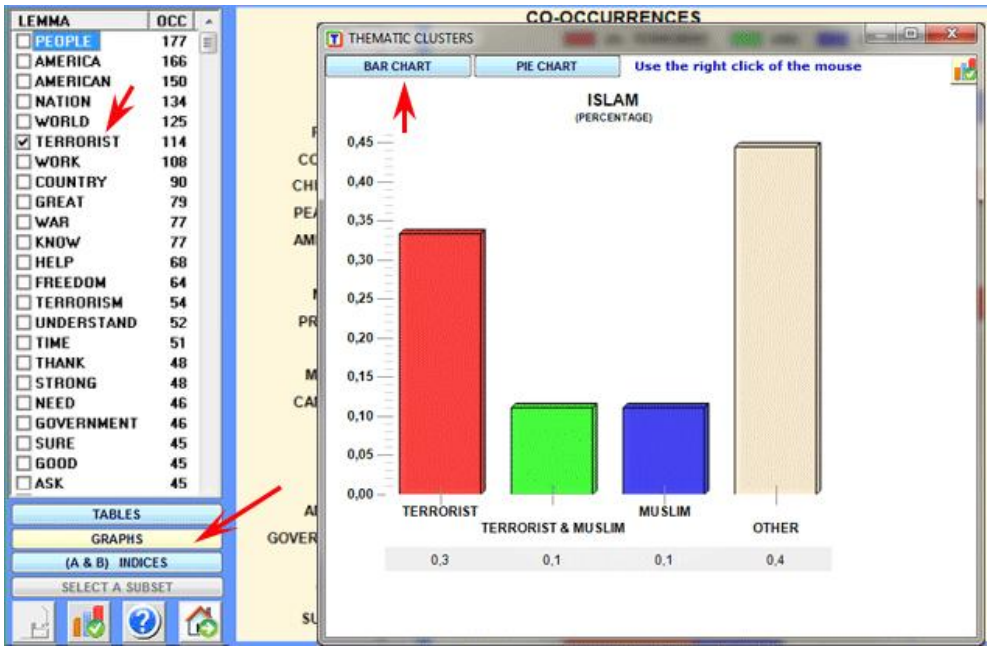


The bar chart (see below) allows us to appreciate the number of elementary contexts in which each lemma co-occurs with the "A" term (red colour), with the "B" term (blue colour) and with both "A" and "B" (green colour).

By double-clicking each label of the chart it is also possible to check its corresponding co-occurrence values and obtain a pie chart with the co-occurrences values concerning each selected word (see below).





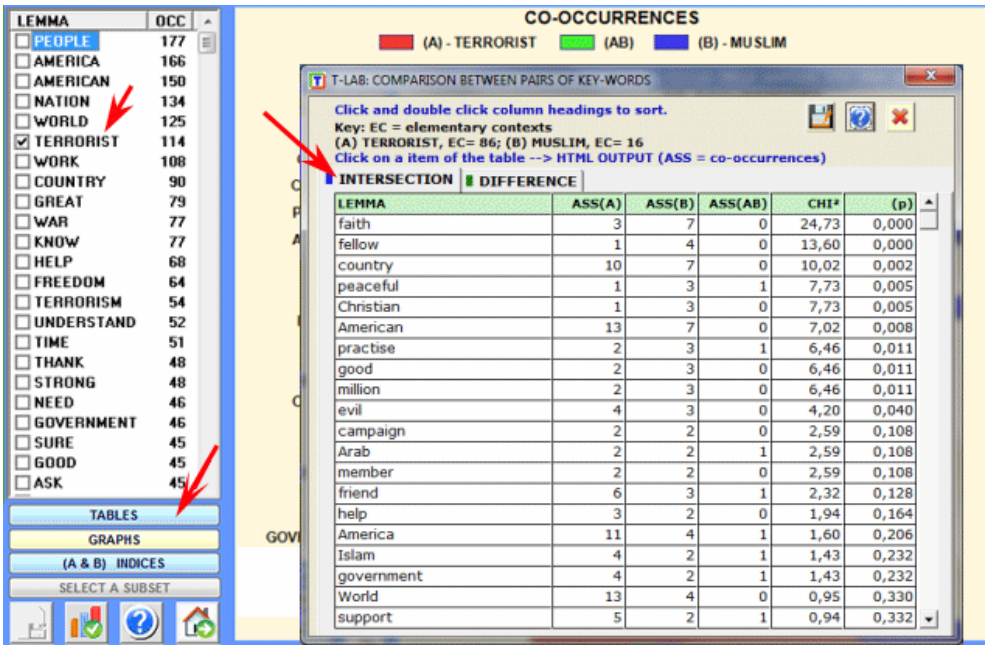


The comparisons provided by **T-LAB** involve the co-occurrences of the elements in the "pair" and each of the words contained in the table (see below).

Let:

A = set of **elementary contexts** (TOT. E.C. = 86) in which the first word of the pair (e.g. "Terrorist") is present;
B = set of elementary contexts (TOT. E.C. = 16) in which the second word of the pair (e.g. "Muslim") is present.

The **first type of comparison** concerns the **shared associations** (see the **intersection** button) and takes into account all words which are present both in "A" and in "B".
In the output table every row shows the values corresponding to the comparisons of each lemma.



LEMMA	ASS(A)	ASS(B)	ASS(AB)	CHI²	(p)
faith	3	7	0	24,73	0,000
fellow	1	4	0	13,60	0,000
country	10	7	0	10,02	0,002
peaceful	1	3	1	7,73	0,005
Christian	1	3	0	7,73	0,005
American	13	7	0	7,02	0,008
practise	2	3	1	6,46	0,011
good	2	3	0	6,46	0,011
million	2	3	0	6,46	0,011
evil	4	3	0	4,20	0,040
campaign	2	2	0	2,59	0,108
Arab	2	2	1	2,59	0,108
member	2	2	0	2,59	0,108
friend	6	3	1	2,32	0,128
help	3	2	0	1,94	0,164
America	11	4	1	1,60	0,206
Islam	4	2	1	1,43	0,232
government	4	2	1	1,43	0,232
World	13	4	0	0,95	0,330
support	5	2	1	0,94	0,332

The reading keys are as follows:

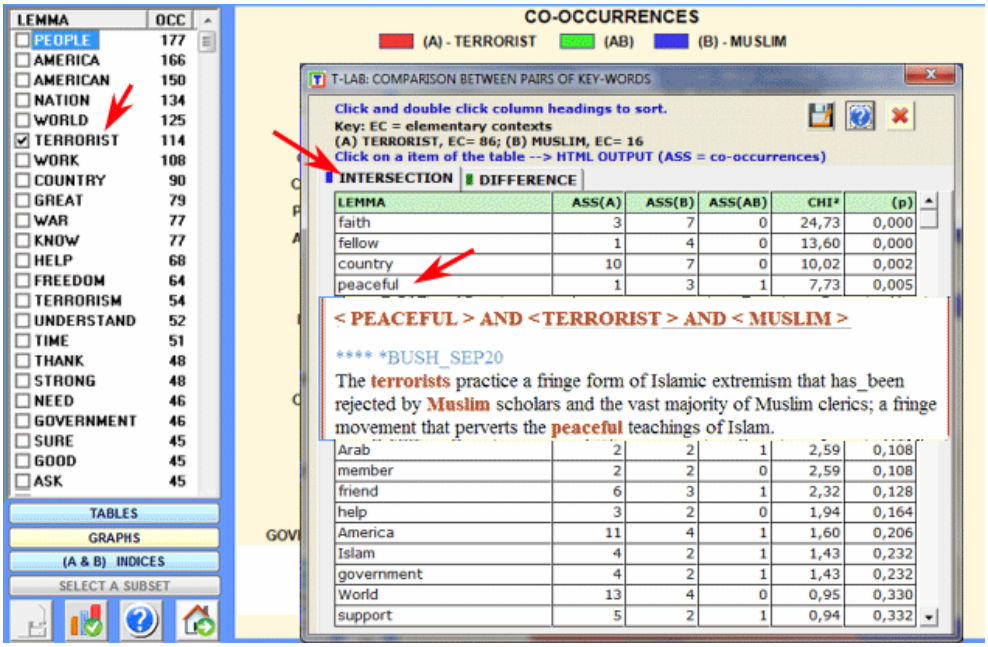
- **ASS (A)** = number of elementary contexts in which each lemma is associated (i.e. co-occurrence) with (A);
- **ASS (B)** = number of elementary contexts in which each lemma is associated with (B);
- **ASS (AB)** = number of elementary contexts in which each lemma is associated with (A) and (B);
- **CHI2** = chi-square values;
- **(p)** = probability associated with the chi square value (def=1).

In this case, for each key word (e.g. "country") **T-LAB** builds a table like the one below and applies the **CHI-square** test to it:

	ASSOC.	NO ASSOC.	TOT.
A	10	76	86
B	7	9	16
	17	85	102

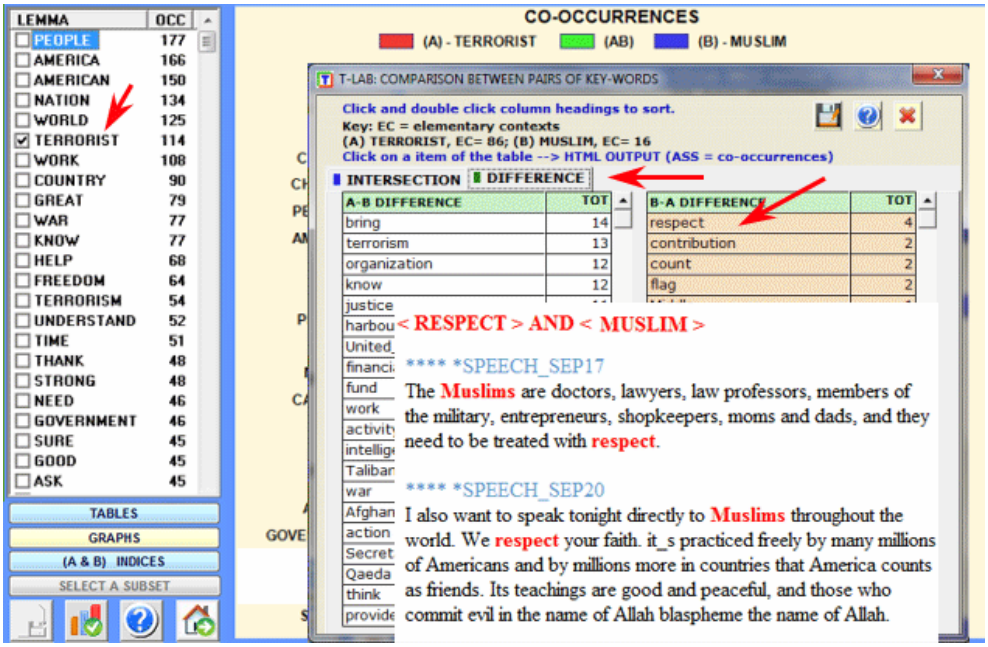
In this table:
- the (**A**) row shows the number of elementary contexts in which "country" is present (10) or absent (76) in the set of contexts (86) containing the first word of the pair ("Terrorist");
- the (**B**) row shows the number of elementary contexts in which "country" is present (7) or absent (9) in the set of contexts (16) containing the second word of the pair ("Muslim").

N.B.: In this case the Chi-square value is 10.02.
Moreover a double click on each item of the output table allows us to save a HTML file with the number of elementary contexts in the corresponding column.

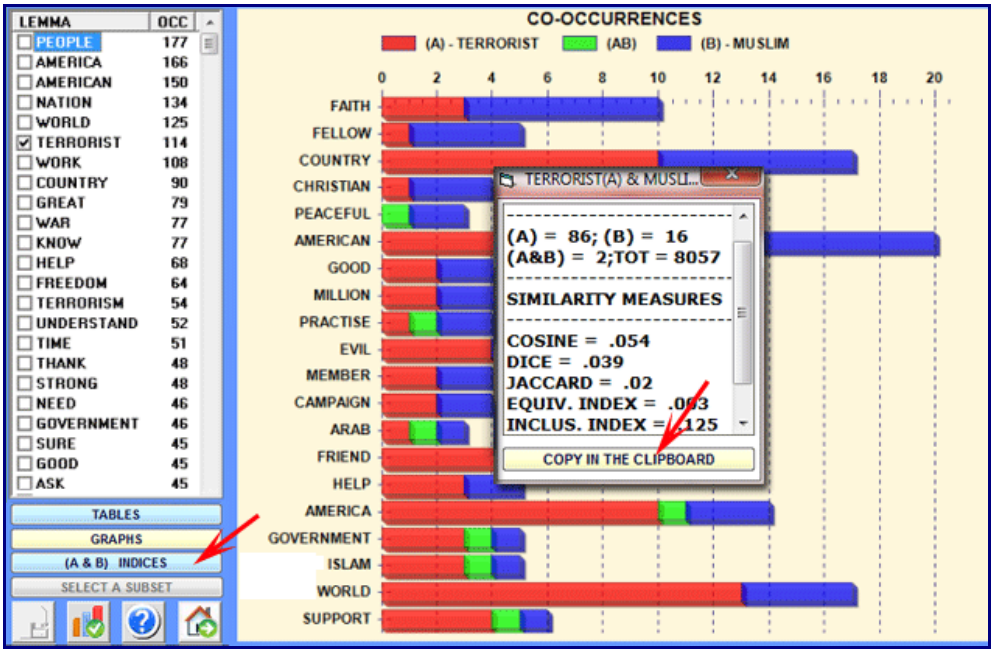


The **second type of comparison** involves the **differences** between A and B (A-B and B-A).

In this case **T-LAB** offers two tables showing the key words which are exclusively associated with the first (A) "or" second (B) word in the pair.
In both tables the "TOT" column shows the number of elementary contexts in which each lemma is associated with only one of the two terms of the pair.



Eventually, by clicking the appropriate button (see the picture below) it is possible to check and export all similarity indexes concerning any word pair.



Sequence and Network Analysis

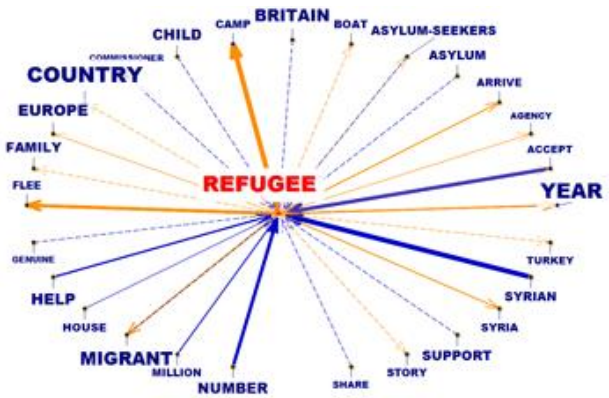
This **T-LAB** tool, which takes into account the positions of the various lexical units relative to each other, allows us to represent and explore any text as a network.

Various options are available which can be used both for performing a **Co-Word Analysis** and a **Thematic Analysis**, as well as **Disambiguation** tasks.

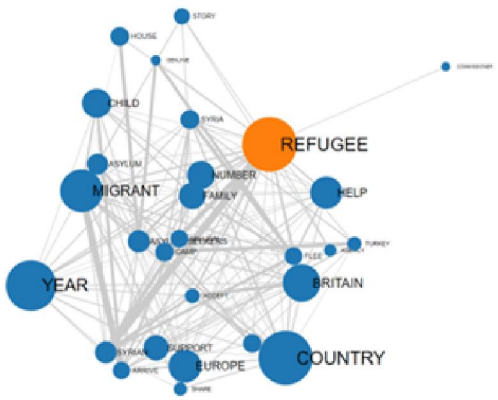
In fact, after building two matrices in which all pairs of predecessors and successors are recorded, **T-LAB** calculates the **transition probabilities** (markov chains) and provides various outputs concerning the target words.
Moreover, it is possible to perform a **cluster analysis** of the network data and explore the semantic relationships between words either within or between the various 'thematic clusters'. To this purpose, the Louvain method for community detection is used (see Blondel V.D., Guillame J.-L , Lambiotte R., Lefebre E., 2008; N.B.: In T-LAB, the analised network consists of directed and weighted links).

That means that the user is allowed to check the relationships between the 'nodes' (i.e. the key-terms) of the network at different levels: a) in one-to-one connections; b) in the 'ego' network; c) within the 'communities' to which they belong; d) within the entire text network.

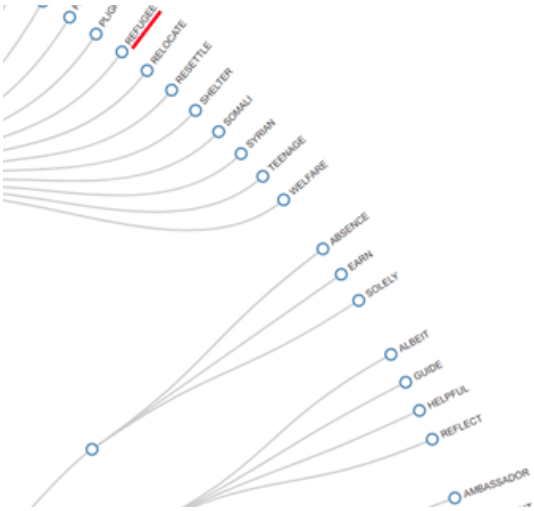
ONE-TO-ONE



EGO-NETWORK



COMMUNITIES



ENTIRE NETWORK

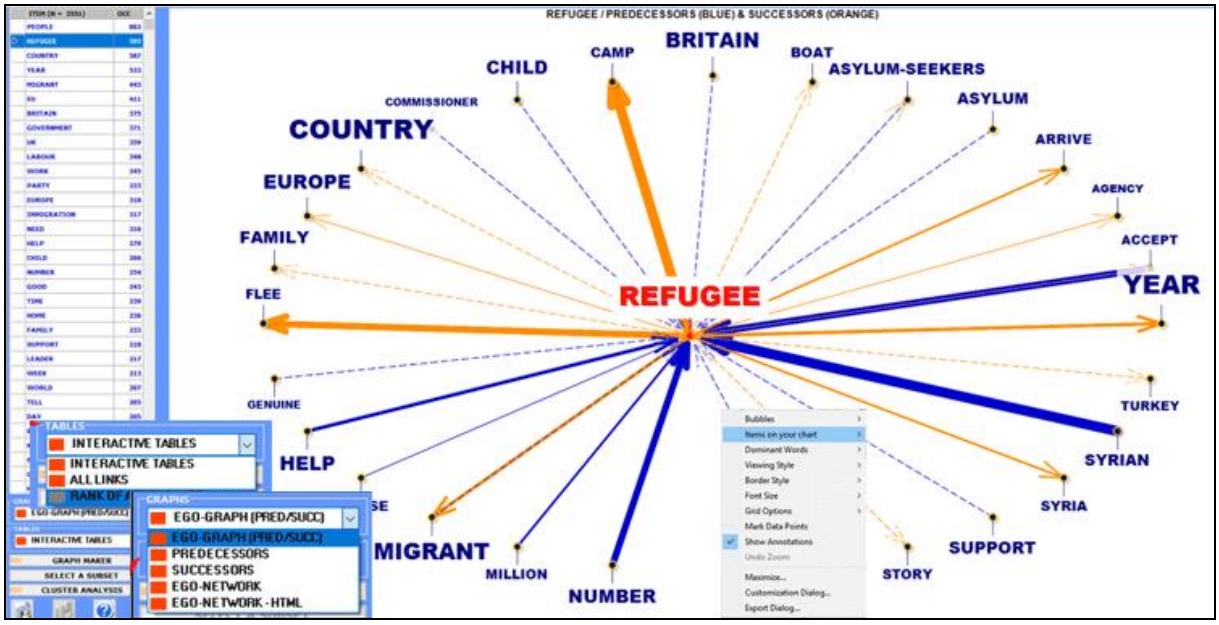
The information concerning how to use the above options is organized in three sections:

A. Exploring one-to-one connections and 'ego' networks;
B. Exploring 'communities' (i.e. thematic clusters) and the entire network;
C. Some technical details.

A - EXPLORING ONE-TO-ONE CONNECTIONS AND 'EGO' NETWORKS

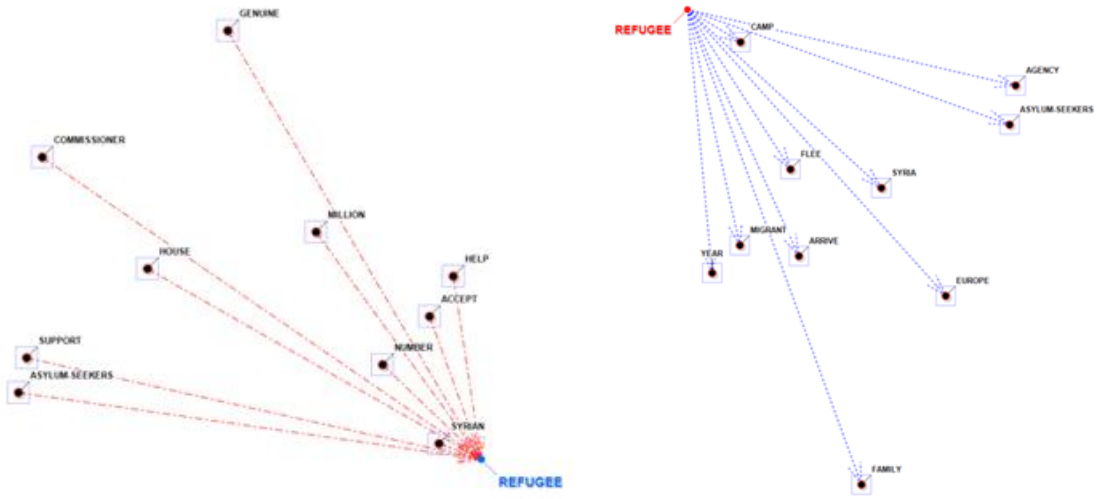
When the automatic analysis is over, several **graphs** and **tables** are available which allow us to ckeck the relationships and the data concerning target words (just click any item in the tables or any point on the graphs).

All graphs can be customized and exported in different formats (right click to show pop-up menu).

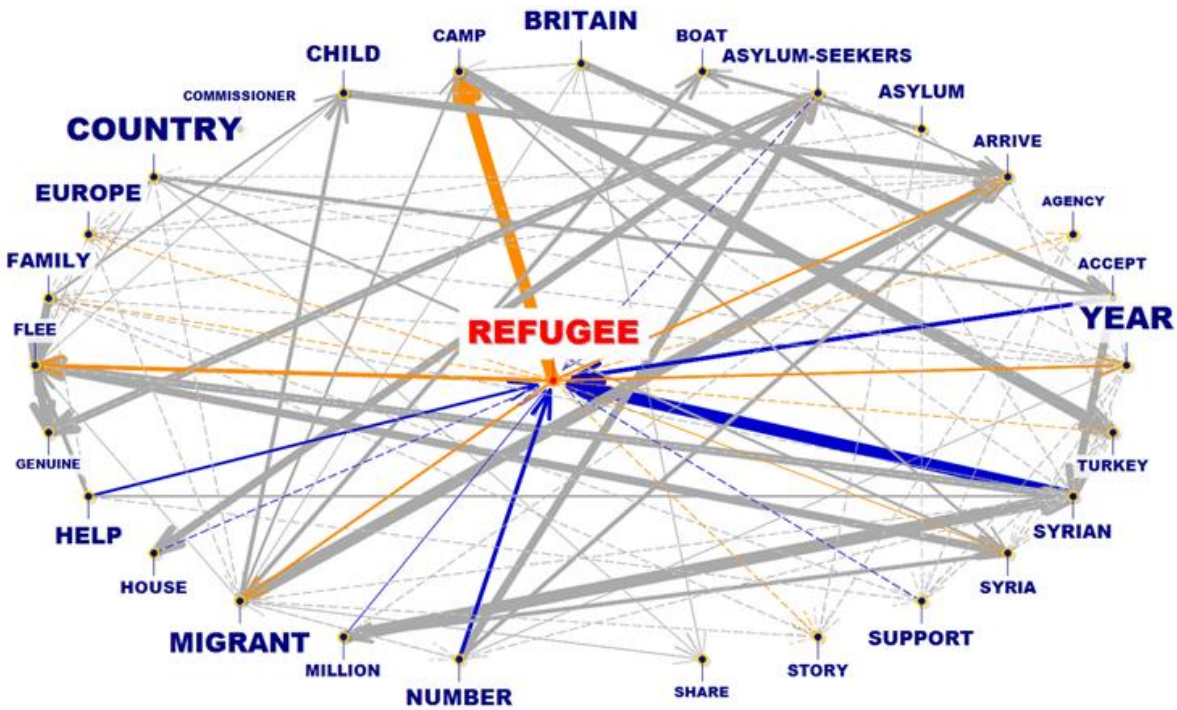


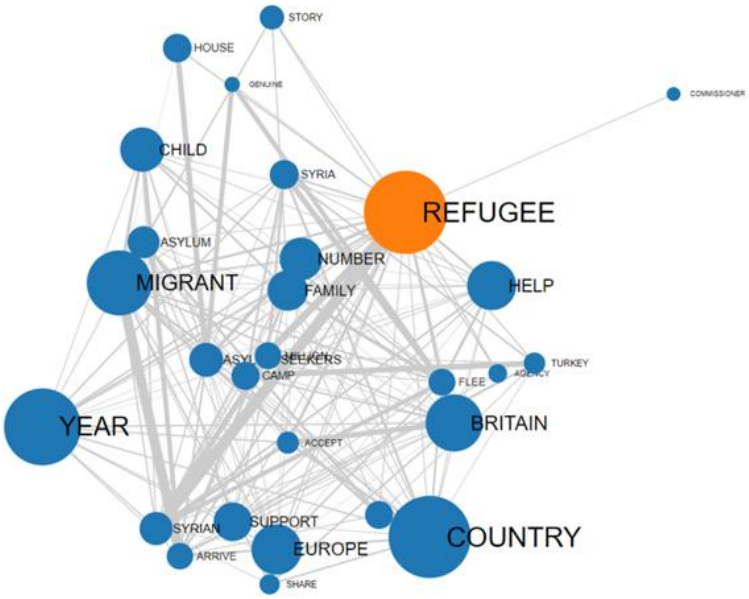
In two of graphs the items that are closer to the selected one are those that have the higher probability of coming before (predecessors) and after (successors).

PREDECESSORS **SUCCESSORS**



In the other cases, the closeness between key-terms is represented by means of the arrow tickness (see below).



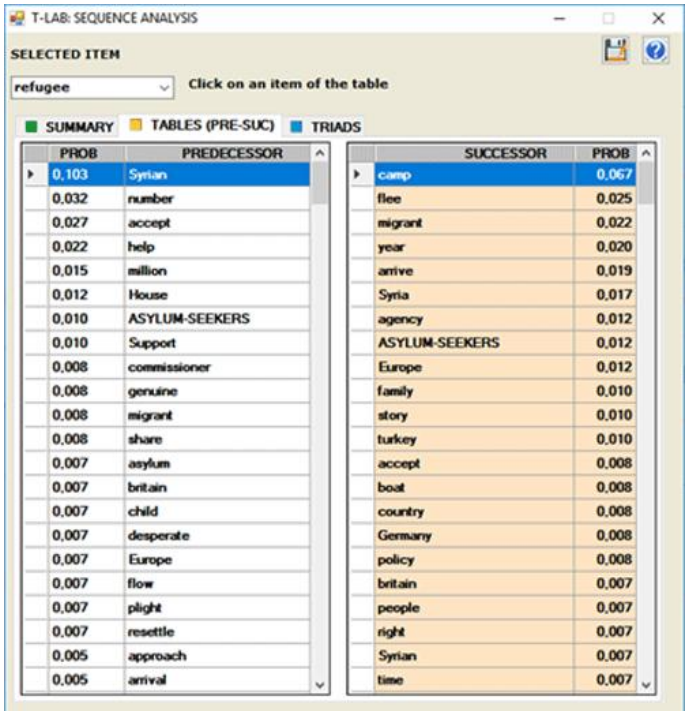


All data can be checked by means of various **tables**.

In detail:

The **INTERACTIVE TABLES** show the sorted list of predecessors and successors of each selected item.

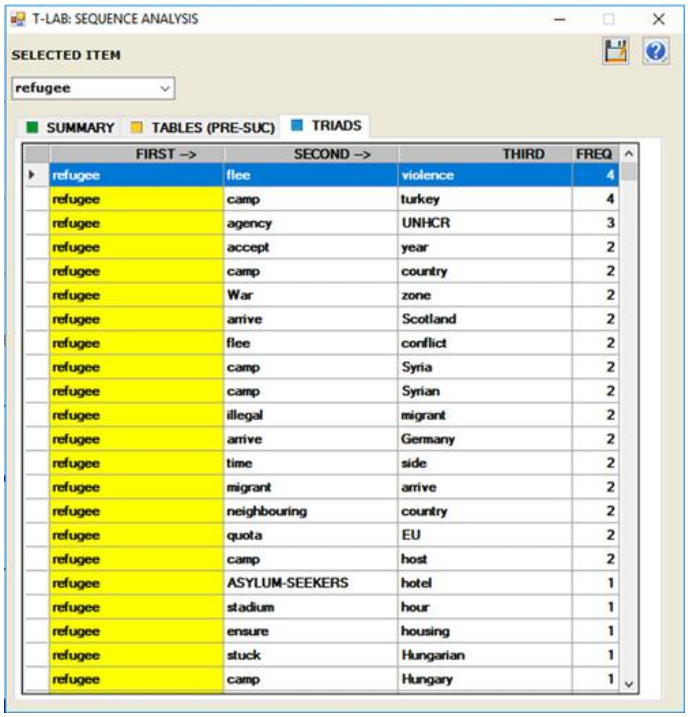
The list is in descending order according to the probability values ("PROB"). For example, in the following table, the probability that "camp" will follow "refugee" is equal to 0.067, that is 6.7%.



PROB	PREDECESSOR
0,103	Syrian
0,032	number
0,027	accept
0,022	help
0,015	million
0,012	House
0,010	ASYLUM-SEEKERS
0,010	Support
0,008	commissioner
0,008	genuine
0,008	migrant
0,008	share
0,007	asylum
0,007	britain
0,007	child
0,007	desperate
0,007	Europe
0,007	flow
0,007	plight
0,007	resettle
0,005	approach
0,005	arrival

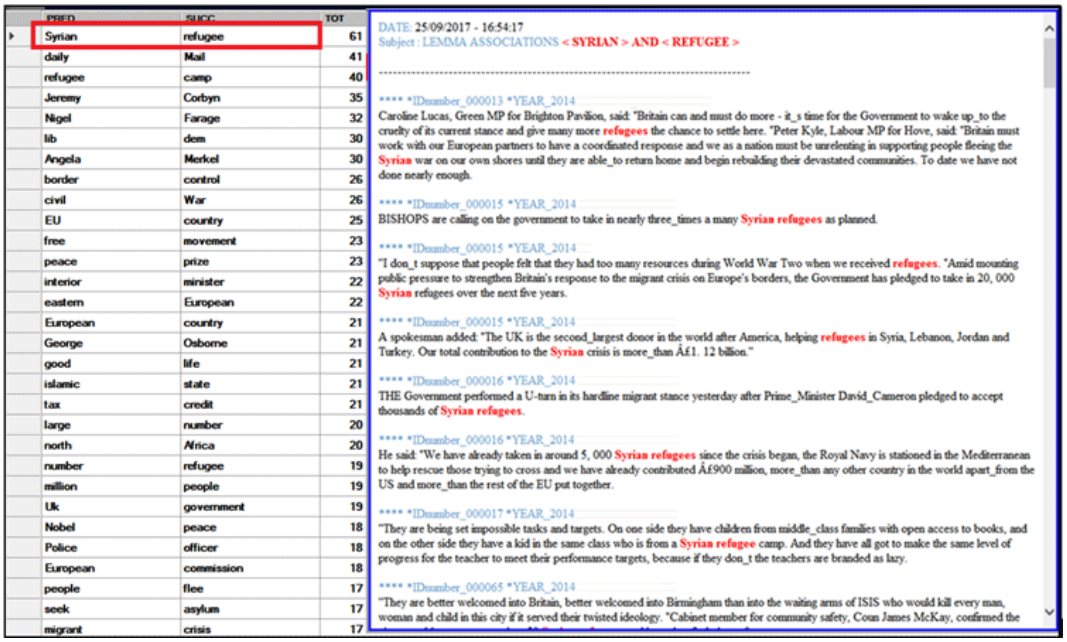
SUCCESSOR	PROB
camp	0,067
flee	0,025
migrant	0,022
year	0,020
arrive	0,019
Syria	0,017
agency	0,012
ASYLUM-SEEKERS	0,012
Europe	0,012
family	0,010
story	0,010
turkey	0,010
accept	0,008
boat	0,008
country	0,008
Germany	0,008
policy	0,008
britain	0,007
people	0,007
right	0,007
Syrian	0,007
time	0,007

The option **TRIADS** (see below) allows us to visualize some tables with sequences of three elements in which the selected item is in the first, in the second or in the third position. For each triad **T-LAB** shows the corresponding occurrence values. (N.B.: Within the triads the empty words are not included.)



FIRST ->	SECOND ->	THIRD	FREQ
refugee	flee	violence	4
refugee	camp	turkey	4
refugee	agency	UNHCR	3
refugee	accept	year	2
refugee	camp	country	2
refugee	War	zone	2
refugee	arrive	Scotland	2
refugee	flee	conflict	2
refugee	camp	Syria	2
refugee	camp	Syrian	2
refugee	illegal	migrant	2
refugee	arrive	Germany	2
refugee	time	side	2
refugee	migrant	arrive	2
refugee	neighbouring	country	2
refugee	quota	EU	2
refugee	camp	host	2
refugee	ASYLUM-SEEKERS	hotel	1
refugee	stadium	hour	1
refugee	ensure	housing	1
refugee	stuck	Hungarian	1
refugee	camp	Hungary	1

The **ALL LINKS** table (see below), which is particularly useful for word-sense disambiguation, contains all word pairs (i.e. predecessor and successor), as well as their occurrence values. Moreover, by clicking any row of this table, all text segments (i.e. elementary contexts) where the two members of each pair are present at same time (i.e. co-occurrences) will be displayed in HTML format on the right side of the form.



PRED	SUCC	TOT
Syrian	refugee	61
daily	Mail	41
refugee	camp	40
Jeremy	Corbyn	35
Nigel	Farage	32
lib	dem	30
Angela	Merkel	30
border	control	26
civil	War	26
EU	country	25
free	movement	23
peace	prize	23
interior	minister	22
eastern	European	22
European	country	21
George	Osborne	21
good	life	21
islamic	state	21
tax	credit	21
large	number	20
north	Africa	20
number	refugee	19
million	people	19
Uk	government	19
Nobel	peace	18
Police	officer	18
European	commission	18
people	flee	17
seek	asylum	17
migrant	crisis	17

DATE: 25/09/2017 - 16:54:17
Subject : LEMMA ASSOCIATIONS < SYRIAN > AND < REFUGEE >

**** *IDnumber_000013 *YEAR_2014
Caroline Lucas, Green MP for Brighton Pavilion, said: "Britain can and must do more - it_s time for the Government to wake up_to the cruelty of its current stance and give many more refugees the chance to settle here. "Peter Kyle, Labour MP for Hove, said: "Britain must work with our European partners to have a coordinated response and we as a nation must be unrelenting in supporting people fleeing the Syrian war on our own shores until they are able_to return home and begin rebuilding their devastated communities. To date we have not done nearly enough.

**** *IDnumber_000015 *YEAR_2014
BISHOPS are calling on the government to take in nearly three_times a many Syrian refugees as planned.

**** *IDnumber_000015 *YEAR_2014
"I don_t suppose that people felt that they had too many resources during World War Two when we received refugees. "Amid mounting public pressure to strengthen Britain's response to the migrant crisis on Europe's borders, the Government has pledged to take in 20, 000 Syrian refugees over the next five years.

**** *IDnumber_000015 *YEAR_2014
A spokesman added: "The UK is the second_largest donor in the world after America, helping refugees in Syria, Lebanon, Jordan and Turkey. Our total contribution to the Syrian crisis is more_than Â£1. 12 billion."

**** *IDnumber_000016 *YEAR_2014
THE Government performed a U-turn in its hardline migrant stance yesterday after Prime_Minister David_Cameron pledged to accept thousands of Syrian refugees.

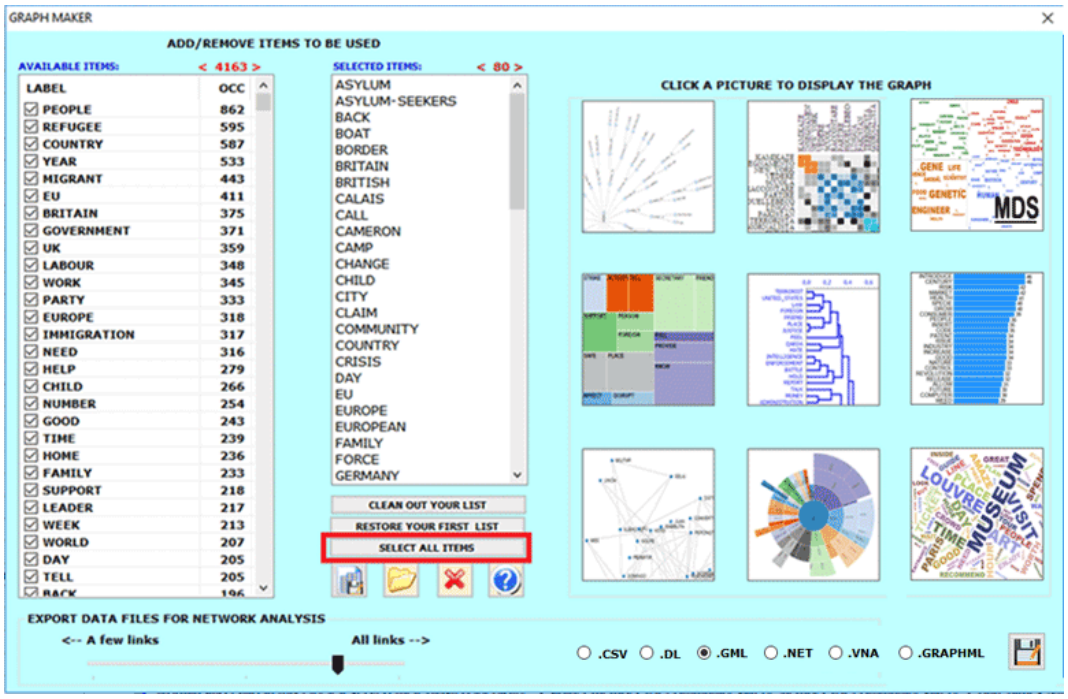
**** *IDnumber_000016 *YEAR_2014
He said: "We have already taken in around 5, 000 Syrian refugees since the crisis began, the Royal Navy is stationed in the Mediterranean to help rescue those trying to cross and we have already contributed Â£900 million, more_than any other country in the world apart_from the US and more_than the rest of the EU put together.

**** *IDnumber_000017 *YEAR_2014
"They are being set impossible tasks and targets. On one side they have children from middle_class families with open access to books, and on the other side they have a kid in the same class who is from a Syrian refugee camp. And they have all got to make the same level of progress for the teacher to meet their performance targets, because if they don_t the teachers are branded as lazy.

**** *IDnumber_000065 *YEAR_2014
"They are better welcomed into Britain, better welcomed into Birmingham than into the waiting arms of ISIS who would kill every man, woman and child in this city if it served their twisted ideology. "Cabinet member for community safety, Coun James McKay, confirmed the

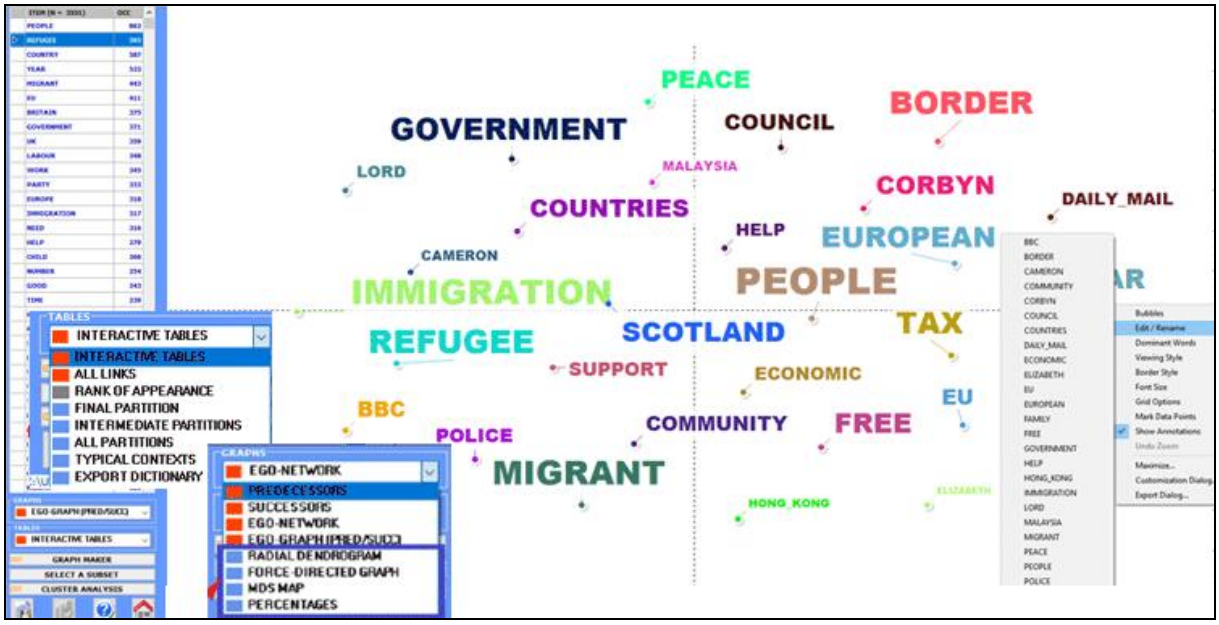
The **RANK OF APPEARANCE** table, with the frequency and the average order of appearance (or evocation) of each term within the text segments, is only provided when the corpus consists of short texts, such as responses to open-ended questions.

Anytime, by clicking the **GRAPH MAKER** option, the user is allowed to obtain various types of graphs by using customized lists of key words (see below)
N.B.: Experienced users who are interested in exporting files in different formats (e.g. .dl, .gml, .net etc.) with data concerning ALL links, may click the 'SELECT ALL ITEMS' button.



B - EXPLORING THE THEMATIC CLUSTERS AND THE ENTIRE NETWORK

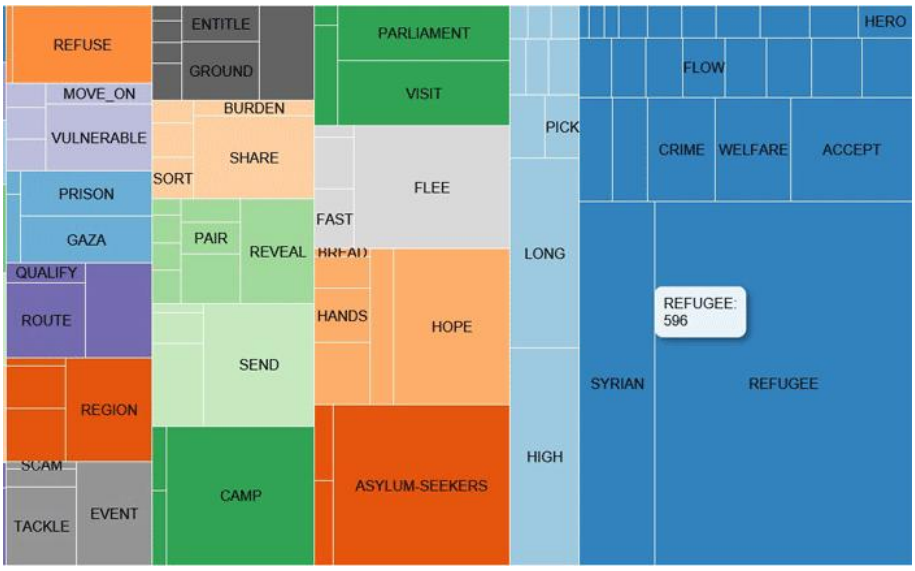
When performing a cluster analysis, further **graphs** and **tables** become available, which allow the easy exploration of all levels of the network hierarchy (see the items marked with the blue rectangles in the below picture).



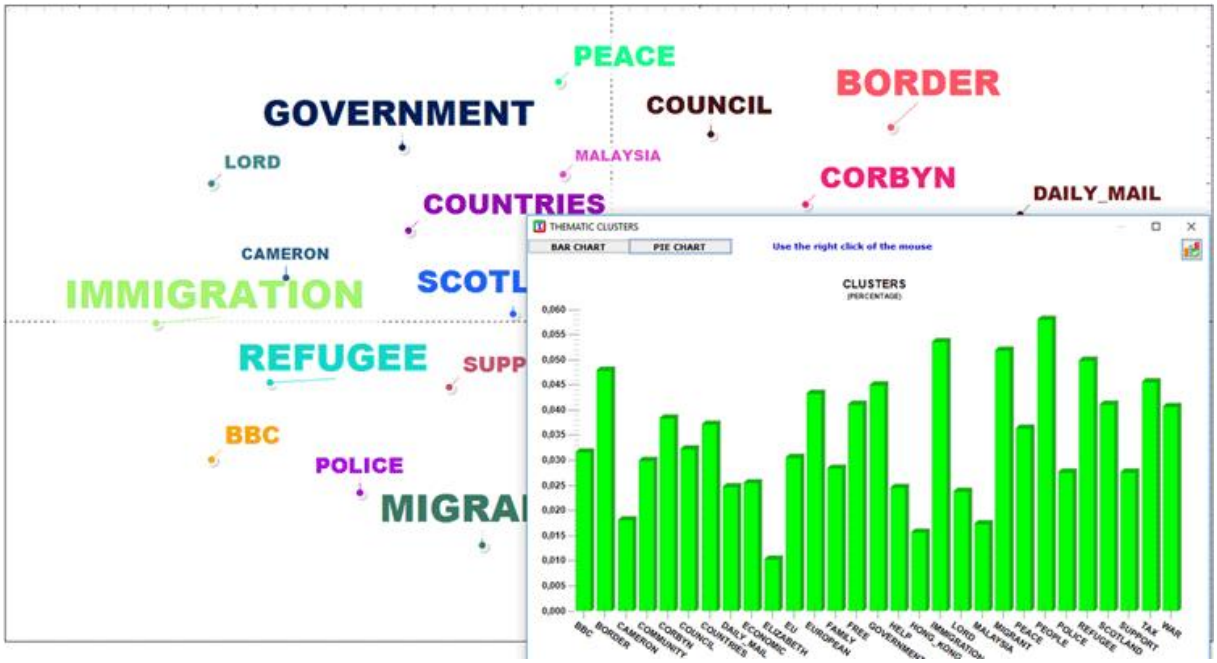
A first table summarizes the characteristics (i.e. key-terms) of the **FINAL PARTITION** obtained by the clustering algorithm.
In such a table, the characteristics of each thematic cluster are sorted by the TF-IDF value (see below).
N.B.: When a cluster of the final partition consists of only two words, usually that means a multiword case has not been resolved during the pre-processing phase.

10_REFUGEE	TF-IDF_10	11_NICK	TF-IDF_11	12_KONG	TF-IDF_12	14_MIGRANT	TF-IDF_14
REFUGEE	692,605	NICK	50,809	KONG	45,461	MIGRANT	203,235
SYRIAN	288,808	CLEGG	45,461	HONG	42,786	MINISTER	112,314
CAMP	187,190	CAMERON	34,764	CHARGE	37,438	BOAT	101,618
ASYLUM-SEEKERS	101,618	FOOTBALL	26,741	NETWORK	34,764	CHANGE	90,921
FLEE	96,269	CAR	24,067	TRAFFIC	29,416	CLAIM	90,921
ACCEPT	90,921	CASE	21,393	VIOLENCE	29,416	RESCUE	74,876
SCHEME	61,505	LEGACY	21,393	FARM	29,416	INTERIOR	66,854
HIGH	53,483	PRIME_MINISTER	21,393	FOOD	29,416	SMALL	64,180
SHARE	45,461	THOUGHT	18,719	INDUSTRY	26,741	BUSINESS	64,180
REFUSE	45,461	UNHAPPY	16,045	VICTIM	26,741	BENEFIT	58,831
RESETTLEMENT	42,786	RECALL	16,045	DOMESTIC	24,067	ROMANIAN	58,831
VULNERABLE	42,786	SERIOUS	16,045	INFRASTRUCTURE	21,393	ITALIAN	56,157
RESETTLE	40,112	HIT	13,371	ABUSE	21,393	MILLION	50,809
COMMISSIONER	37,438	MATCH	13,371	SMUGGLE	21,393	WORKER	50,809
HOPE	34,764	BELIEVE	13,371	SPENCER	21,393	NAVY	48,135
PERIOD	34,764	FAN	13,371	TERMS	18,719	BULGARIAN	48,135
RELOCATION	32,090	DELIGHT	13,371	MARKS	18,719	FISH	48,135
SEND	32,090	DEVOTE	10,697	PRODUCTION	18,719	SHIP	45,461
HOST	32,090	FEDERATION	10,697	SEXUAL	18,719	VESSEL	42,786
CURRENTLY	32,090	BLAIR	10,697	BRISTOL	18,719	CLIMATE	40,112
EVENT	32,090	BOMBER	10,697	BOOST	16,045	LIFE	37,438
UNHCR	32,090	ABSOLUTELY	10,697	CAMPAIGN	16,045	LAUNCH	37,438
CRIME	29,416	ACQUIRE	10,697	HOSPITALITY	16,045	ROYAL	34,764
LONG	26,741	MOUTH	10,697	WAIT	16,045	EXAMPLE	34,764
VISIT	26,741	HONOUR	10,697	PREPARATION	13,371	CONTRIBUTE	32,090
MAIN	24,067	ROBINSON	10,697	BREATH	13,371	PORT	32,090
REGION	24,067	THREAT	10,697	COAT	13,371	TAXPAYER	32,090
REFLECT	21,393	SUICIDE	10,697	AGRICULTURAL	13,371	SKILLED	29,416
PRISON	21,393	SURE	10,697	BRAVE	10,697	AFRICAN	29,416
PROGRAM	21,393	POOR	10,697	COMPETITION	10,697	APPLY	26,741
PAIR	21,393	WIFE	10,697	CHARACTER	10,697	AUSTRALIA	26,741
TRAFFICKER	21,393	TERRIFY	8,022	GLASS	10,697	CABINET	26,741
SHELTER	21,393	TRANSPORT	8,022	EXPORT	10,697	COASTGUARD	26,741

By clicking any word in the above table (as well as in the **ALL PARTITIONS** table), a TreeMap allows us to check the communities to which it results to belongs (see below).

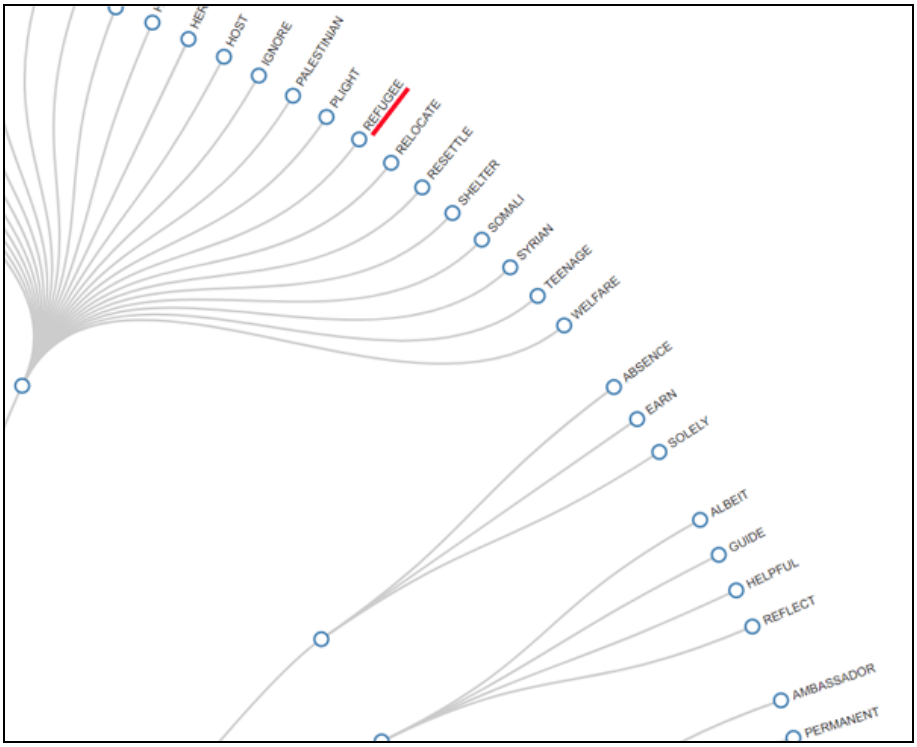


The **MDS MAP** and the **PERCENTAGES** charts (see below) allow us to check the weight of each cluster as well as their relationships within the final partition (see below).

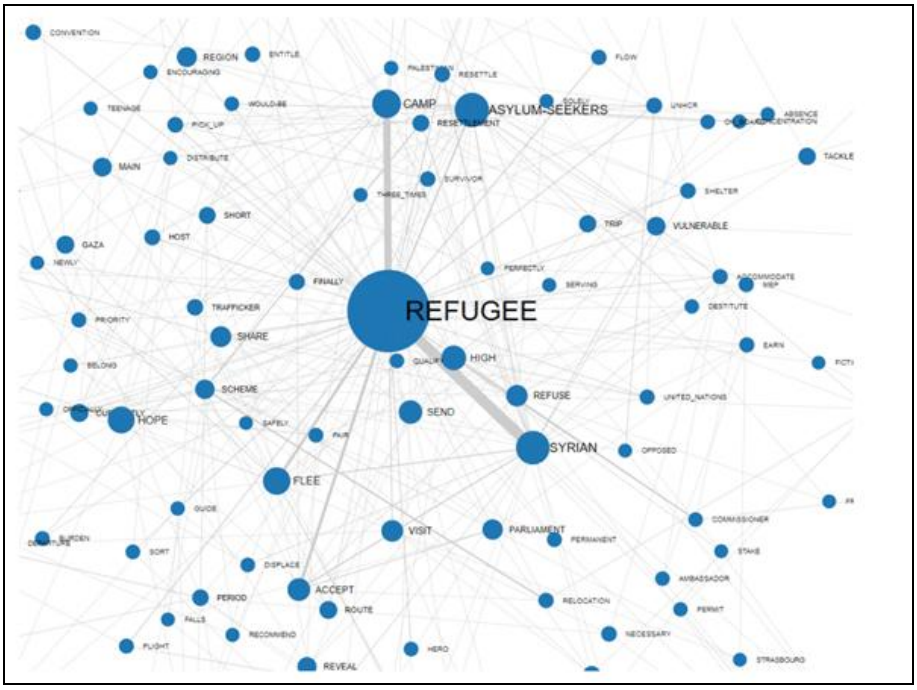


Depending on the number of key-words, two graphs in HTML format allow us to check the relationships between them, either within the entire network or within the cluster they belong to (see below).

RADIAL DENDROGRAM



NETWORK GRAPH (FORCE-DIRECTED GRAPH)



Three other tables provide us with further outputs of the cluster analysis.

In detail:

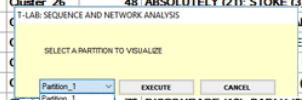
The **ALL PARTITIONS** table allows us to check how the key-words have been grouped at each cluster partition (see the below table, in which the numbers in the partition columns refer to the various clusters).
N.B.: In such a table, which – by default – is ordered on the first partition, each shift from one small cluster to the other is marked by highlighting in green the first word which belongs to it.

Save this table as .xls file
Save this table as .csv file

Final_Partition	Partition_3	Partition_2	Partition_1	Lemma	OCC	PERC
24	26	36	60	IRAQ	37	
24	26	36	60	AFGHANISTAN	19	
24	26	36	60	ERITREA	19	
24	26	36	60	SUDAN	17	
[illegible]	[illegible]	36	60	POLAND	10	
[illegible]	[illegible]	36	60	SOMALIA	8	
[illegible]	[illegible]	46	61	DOCUMENT	28	
4	4	46	61	SO-CALLED	19	
4	4	46	61	PASSPORT	18	
4	4	46	61	AFRAID	10	
4	4	46	61	KNIFE	5	
4	4	46	61	STAMP	5	
4	4	46	61	EXPIRE	2	
24	26	36	62	NORTH	74	
24	26	36	62	AFRICA	63	
24	26	36	62	MIDDLE_EAST	35	
14	14	39	63	BOAT	130	
14	14	39	63	AFRICAN	30	
14	14	39	63	SINK	23	
14	14	39	63	FISH	20	
14	14	39	63	CAFE	11	
14	14	39	63	EGYPTIAN	10	
14	14	39	63	SAIL	6	
14	14	39	63	LAKE	4	
14	14	39	63	OVERCROWDED	4	
11	11	47	64	CHAOS	24	
11	11	47	64	WEDNESDAY	22	
11	11	47	64	AFTERMATH	4	
16	17	20	65	YESTERDAY	167	
16	17	20	65	LOCAL	129	
16	17	20	65	AFTERNOON	8	
16	17	20	65	PROVINCE	2	
18	19	23	66	TALK	134	
18	19	23	66	AGE	59	
18	19	23	66	TEACHER	26	

The **INTERMEDIATE PARTITIONS** table allows us to check how the key-words have been grouped at any selected cluster partition.
In such a table, the characteristics of each thematic cluster are sorted by their occurrence value (see below).

Partition_3	Higher_Level	Members	Features
Cluster_01	Cluster_01	119	BUS (24); ABAAOUD (17); AFGHAN (12); ACTUAL (9); EMIGRATE (6); OMAR (6); ALBANIAN (3); ADAM (12); HOLMES (2); LE
Cluster_02	Cluster_13	148	MANCHESTER (33); HOTEL (31); ABANDON (23); FLIGHT (15); GATWICK (4); HOSTEL (4); HEATHROW (2); FRIDAY (27); EN
Cluster_03	Cluster_20	123	PERSECUTION (43); TRUCK (12); REACH_OUT (8); REPORTEDLY (8); ABANDONED (5); TURN_DOWN (4); PURCHASE (3);
Cluster_04	Cluster_22	132	TONY (30); MISSING (27); MONDAY (26); ABBOTT (25); SERIE (19); CHEF (17); FAMILIAR (13); RACHEL (8); OXFORD (7); CA
Cluster_05	Cluster_12	135	DIFFICULT (48); IMPACT (31); ABILITY (22); SPELL (8); ADAPT (4); ACTION (65); IMMEDIATE (12); APOLOGIZE (6); INDIVID
Cluster_06	Cluster_21	83	CHILD (268); ABOARD (4); ARRIVE (117); FEATURE (18); SONG (17); SUDDENLY (10); ALBUM (10); HANDFUL (6); FOLK (2);
Cluster_07	Cluster_14	111	TEMPORARY (23); PARK (19); ADOPT (12); CAMBRIDGE (12); STYLE (10); ABOLISH (5); OLYMPIC (4); ACCOMMODATION (2
Cluster_08	Cluster_11	157	OPPORTUNITY (39); CANADA (19); CONDEMN (16); ABORTION (4); INTOLERANCE (4); AIM (39); COUNT (13); STRENGTHE
Cluster_09	Cluster_29	147	ABROAD (28); TOMORROW (19); TALE (8); DIVERSE (4); WORK (404); HARD (70); EMPLOYMENT (24); BATTLE (24); CARR
Cluster_10	Cluster_25	149	REFUGEE (596); SYRIAN (174); ACCEPT (80); WELFARE (50); CRIME (45); HATE (23); HOST (22); SHELTER (19); RESETTL
Cluster_11	Cluster_26	48	ABSOLUTELY (21); STOKE (3); CHAOS (24); WEDNESDAY (22); AFTERMATH (4); BELIEVE (106); POOR (58); HUNGRY (6);
Cluster_12			BUSE (25); DOMESTIC (17); SLAVERY (7); HORRENDOUS (3); PAIN (21); ADDITIONAL (13); NE
Cluster_13			E (16); SCORE (16); BRUTAL (11); COMMUNICATION (10); LOCK (10); GCSE (8); EXCELLENT (8)
Cluster_14			ROYAL (33); NAVY (21); WATERS (12); NO_DOUBT (11); NON-EU (9); ACADEMY (8); ORIGINA
Cluster_15			(18); WINTER (16); FRIGHTEN (11); IMPRESSIVE (5); ACCENT (5); BOAST (5); FORMAL (3); IM
Cluster_16		70	DISCOURAGE (10); BADLY (8); ACCEPTABLE (2); ACTIVIST (39); MINUTE (28); SWEAR (2); WESTERN (24); MANIFESTO (20
Cluster_17		152	EUROPEAN (189); PRESIDENT (85); EASTERN (45); COMMISSION (41); JUNCKER (20); RUSSIA (20); UKRAINE (9); ACCESS
Cluster_18	Cluster_18	102	AGREE (64); AT_ALL (23); ACCOMPANY (8); STANLEY (5); DAVIS (2); KEY (40); ADMIT (38); TOOL (12); CONCEDE (10); JAC
Cluster_19	Cluster_07	102	ACCORDING_TO (72); LEAVING (29); MOTIVATE (2); TALK (134); AGE (59); TEACHER (26); WORK_OUT (11); COMPUTER (2
Cluster_20	Cluster_15	185	ACCOUNT (29); FRANK (6); ENGLISH (75); ADMISSION (7); ANNIVERSARY (6); VETO (6); PREVENT (30); ALLEGE (10); HOT
Cluster_21	Cluster_23	115	TORY (178); ACCUSATION (6); MIGRATION (153); COMMITTEE (46); ADVISORY (5); AFFAIR (27); SELECT (10); SOLICITOR
Cluster_22	Cluster_06	169	HOUSING (56); ACCUSE (50); SOUTH_EAST (2); ACUPUNCTURE (18); SESSION (8); JOB (141); ENGLAND (74); WAVE (23);
Cluster_23	Cluster_16	113	LEADER (217); ACHIEVE (20); SPIRITUAL (8); TIBETAN (4); ACTIVITY (12); WELCOMED (11); INTENSE (8); SCRUTINY (6);
Cluster_24	Cluster_05	32	ACQUIRE (7); SMART (4); CAR (48); CRASH (7); ADVENTURE (6); GOLF (3); EASE (9); AUTUMN (7); TERRIFY (7); JUDGMEN
Cluster_25	Cluster_24	104	BIG (103); ACTOR (11); STABLE (10); VETERAN (9); IDEAL (6); HOLLYWOOD (3); SCHENGEN (36); AGREEMENT (25); TREA



The **TYPICAL CONTEXTS** table allows us to check the text segments which have the highest score of association with the clusters of the final partition. In such tables the 'score' refers to the similarity (cosine index) between the feature vector of each cluster and the vector in which each text segment is represented.
N.B. In this table, the most significant text segment of each cluster is highlighted in yellow.

CLUSTER	SEG_ID	SCORE	TEXT
EU	22386	0,0794	' ' THE PRINCIPLES MUST SUPPORT THE INTEGRITY OF THE EUROPEAN SINGLE MARKET , THAT INCLUDES THE RECOGNITION THAT T
EU	22385	0,0725	' ' What we seek are principles embedded in EU law and binding on EU institutions that safeguard the operation of the union for all 28 member_st
EU	6105	0,0625	The only good news is that when the impact sinks in , it will_be another nail in the coffin of our disastrous EU membership .
EU	6558	0,0590	There is no better symbol of the EU ambition to banish the old world of competing nation_states , each with their own laws , borders and currency
EU	7633	0,0538	All are governed by our relationship with the EU - a relationship that we now know will_be renegotiated before a referendum is put to the British_
EU	19880	0,0538	Brussels has demanded pounds 600m extra from Britain next year to meet the pounds 5 . 5bn increase in the EU budget . While the countries of th
EU	1685	0,0529	The new workers , many of whom were from Poland , coincided with a nationwide influx of new economic_migrants , which began when 10 new s
EU	5381	0,0526	Without a formal renegotiation of our relationship with the EU , all these transfers of power from Westminster to Brussels are irreversible .
EU	3237	0,0518	Even the EU pretext that cod stocks must be protected is a sham .
EU	10665	0,0513	The EU has a track record of guaranteeing democracy , often only recently achieved , in its member_states and ending cross-border conflicts . C
EU	15916	0,0505	The EU foreign policy chief , Mr Javier Solana , declared that , as they approach the end of their six-month EU presidency , the Belgians have n
EU	17173	0,0496	Any EU citizen will_do .
EU	6596	0,0496	They are coming and the EU has no answer .
EU	6566	0,0471	For all their rhetoric about open internal borders and a brotherhood of nations under one flag , the reality across the EU is rather less edifying .
EU	8192	0,0466	Unless the EU elite recognises that nations must control their borders , no deal they can offer will convince the electorate the cost of EU members
EU	8717	0,0442	He added that as France would hold the rotating EU presidency when the Games take place it would_be up_to him to sound out member_states on
EU	12969	0,0427	Now Leave . EU , which Farage supports , has criticised Lawson strongly . As Sebastian Payne reports at Coffee House , Leave . EU has issued
EU	13163	0,0411	Sipping on mint tea , Hajj said : ' ' After the revolution we wanted to return the favour to the EU because they stood with us against the tyrant
EU	16564	0,0408	As the death toll in the Mediterranean continues to rise week by week , those seeking asylum in Europe will_be hoping EU leaders take their pledg
EU	19199	0,0406	British acceptance of genuine asylum_seekers is the lowest of the EU member_states .
EUROPEAN	15573	0,0686	THE FRENCH PRIME_MINISTER , MANUEL VALLS , AND THE EUROPEAN COMMISSION PRESIDENT , JEAN-CLAUDE JUNCKER , YESTERD
EUROPEAN	6469	0,0678	The suspension of free travel by the Germans was backed by the European Commission as being within the rules . However , Commission Preside
EUROPEAN	6589	0,0617	So much , then , for European brotherhood and the principle of an ever-closer union ' .
EUROPEAN	14463	0,0442	Antonio Guterres , the head of UNCHR , warned European countries yesterday to keep out the welcome mat for genuine Iraqi asylum-seekers or r
EUROPEAN	21229	0,0437	Mr Brown angered the European Commission and his European counterparts on Monday by announcing that he was going to a finance ministers ' m
EUROPEAN	21793	0,0417	5 Which two European nations failed to win a game ?
EUROPEAN	19694	0,0417	And they are being joined by failed asylum_seekers and Eastern European economic_migrants .
EUROPEAN	6635	0,0417	They make_up around half the 1 . 3million eastern Europeans in the UK .
EUROPEAN	15582	0,0413	Europe should embrace more refugees fleeing war and dictatorship while also tightening border controls and more strictly enforcing its returns polic
EUROPEAN	24152	0,0410	' ' The government needs to stop apportioning blame by pushing the responsibility back onto the Muslim community . Instead , those professiona
EUROPEAN	13117	0,0386	His brief covers European integration , international patterns of economic growth , investment , productivity , wages and employment .
EUROPEAN	16024	0,0385	The ' ' vice-president foreign minister ' ' who is to be appointed under the new European constitution will_be assisted by a European external

Like other cases of thematic analysis, **T-LAB** allows us to **export the dictionary** of the final partition which can be used for further analyses.

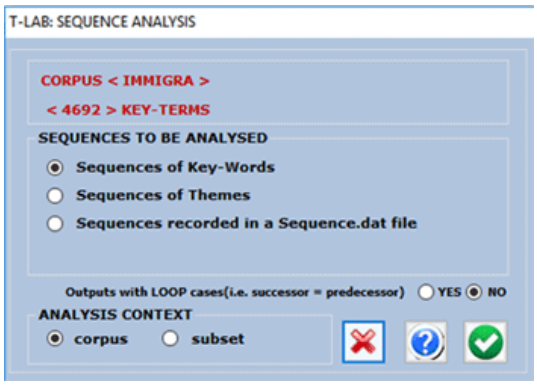
C – SOME TECHNICAL DETAILS

The types of sequences that this tool allows us to analyse are the following:

A) **Sequences of Key-Words**, the items of which are lexical units (i.e. words or lemmas) present in the the corpus or in a subset of it. In this case the maximum number of nodes (i.e. 'types' of lexical units) is 5,000;
N.B.: When the automatic lemmatization is applied, this limit corresponds to about 12,000 words (i.e. raw forms).

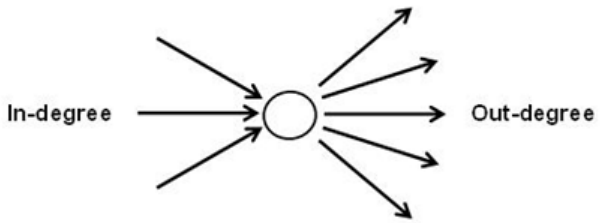
B) **Sequences of Themes**, the items of which are context units (i.e. elementary contexts) tagged by a **T-LAB** tool for thematic analysis;
N.B.: Since the sequence of elementary contexts (sentences or paragraphs) characterises the entire 'chain' (predecessors and successors) of the corpus, in this case **T-LAB** performs a specific form of **Discourse Analysis** the nodes of which (i.e. 'themes') can vary from 5 to 50.

C) **Sequences recorded in a Sequence.dat file** made by the user (see the the explanation at the end of this section). In this case the maximum number of records is 50,000 and the number of 'types' (i.e. nodes) must not exceed 5,000.



The following information is provided to help the user to better understand the data reported in the **SUMMARY** table.

According to the graph theory, the predecessors and the successors of each **node** (in this case, lexical unit or theme) can be represented by means of arrows (arcs) coming to (in-degree = types of predecessors) or going out (out-degree = types of successors).



As an example, in the following table table "people" has 412 types of successors and 449 types of predecessors.And its centrality degree is 0.243.

T-LAB: SEQUENCE ANALYSIS

SUMMARY

SUMMARY | TABLES (PRE-SUC) | TRIADS

NODE	FREQ	PRED	SUCC	RATIO	COVER	CENTR
people	862	449	412	0,918	0,849	0,243
country	587	291	336	1,155	0,813	0,177
year	533	271	275	1,015	0,751	0,154
refugee	595	269	268	0,996	0,856	0,151
migrant	443	255	231	0,906	0,861	0,137
britain	375	215	220	1,023	0,812	0,123
EU	411	230	204	0,887	0,848	0,122
government	371	188	243	1,293	0,840	0,121
work	345	216	186	0,861	0,833	0,113
Uk	359	185	193	1,043	0,783	0,106
Europe	318	164	213	1,299	0,832	0,106
party	333	176	196	1,114	0,799	0,105
labour	348	189	182	0,963	0,762	0,105
need	316	184	182	0,989	0,840	0,103
immigration	317	176	159	0,903	0,833	0,094
help	279	169	155	0,917	0,821	0,091
child	266	139	167	1,201	0,806	0,086
time	239	146	155	1,062	0,776	0,085
family	233	147	150	1,020	0,822	0,084
good	243	149	135	0,906	0,823	0,080
number	254	126	152	1,206	0,860	0,078
Support	218	148	130	0,878	0,826	0,078

According to their ratio (successors/predecessors), it is possible to verify the semantic variety engendered by each node:

- if the ratio is greater than 1, the node is defined "source";
- if the ratio is equal to 1, the node is defined "relay";
- if the ratio is lower than 1, the node is defined "well".

In the same table, for each lexical unit, the column "cover" (coverage) indicates the percentage of its occurrences preceded or followed by lexical units included in the user list.

When the analysed units "cover" the totality of those present within the corpus, the "cover" value is equal to 1; otherwise, it is a lower value.

Moreover: when the "cover" value is equal to 1, the summations of the probability values (both of predecessors and of successors) are also equal to 1; otherwise, they have lower values.
In both cases, the "residual" percentage is determined by the fact that there are predecessors and successors not included in the analysis.

For example, the sequence represented in the following image is constituted by 39 events: of these, only 16 (the hypothetical units in analysis) are "covered" (gray boxes). That is because some of them, e.g. those corresponding to the occurrences of the lexical unit "A", have predecessors and successors not included in the analysis (white boxes).

Differently, when the user analyses **sequences of themes** or **sequences recorded in external files** all the events are "covered".

N.B.: In order to analyse an external file, the user must prepare a 'Sequence.dat' file; then, after opening an existing project, he must select the 'Sequences recorded in a Sequence.dat file' option.

The calculation method, the graphs and the tables are analogous to those already described (see above).

The Sequence.dat file, which can contain numerous kinds of tags (e.g. names of speakers in a conversation, categories obtained by content analysis, kinds of events, etc.), must be made up by "N" lines (min 50 max 50,000), each with a tag of a max of 50 characters, without punctuation marks or blank spaces.

Tag types must be max 5,000.

Here are some lines of Sequence.dat files in the correct format:

EXAMPLE_01	**EXAMPLE_02**	**EXAMPLE_03**
Hamlet	activist	event_01
King	food	event_03
Hamlet	genetic	event_02
Queen	conservative	event_03
Hamlet	activist	event_03
Queen	genetic	event_01
Hamlet	conservative	event_05
King	activist	event_02
Queen	commerce	event_05
Hamlet	conservative	event_01
King	activist	event_02
... ...	... ...	... ...

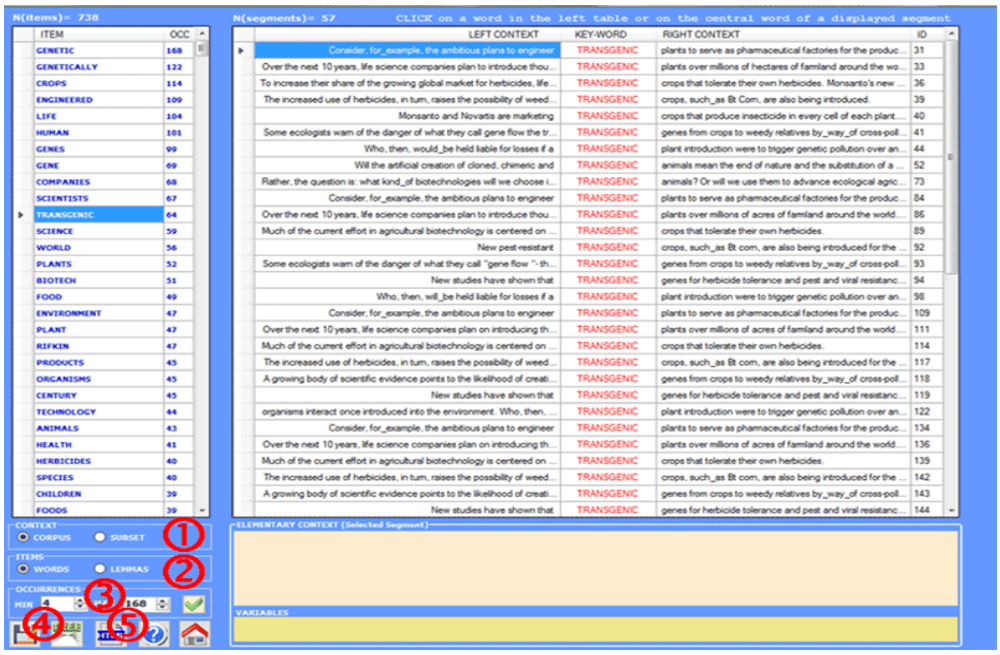
Both in the case of sequences concerning the corpus lexical units (or themes) and of those included in an external file (Sequence.dat), **T-LAB** several working tables which can be found in the MY-OUTPUT folder.

Concordances

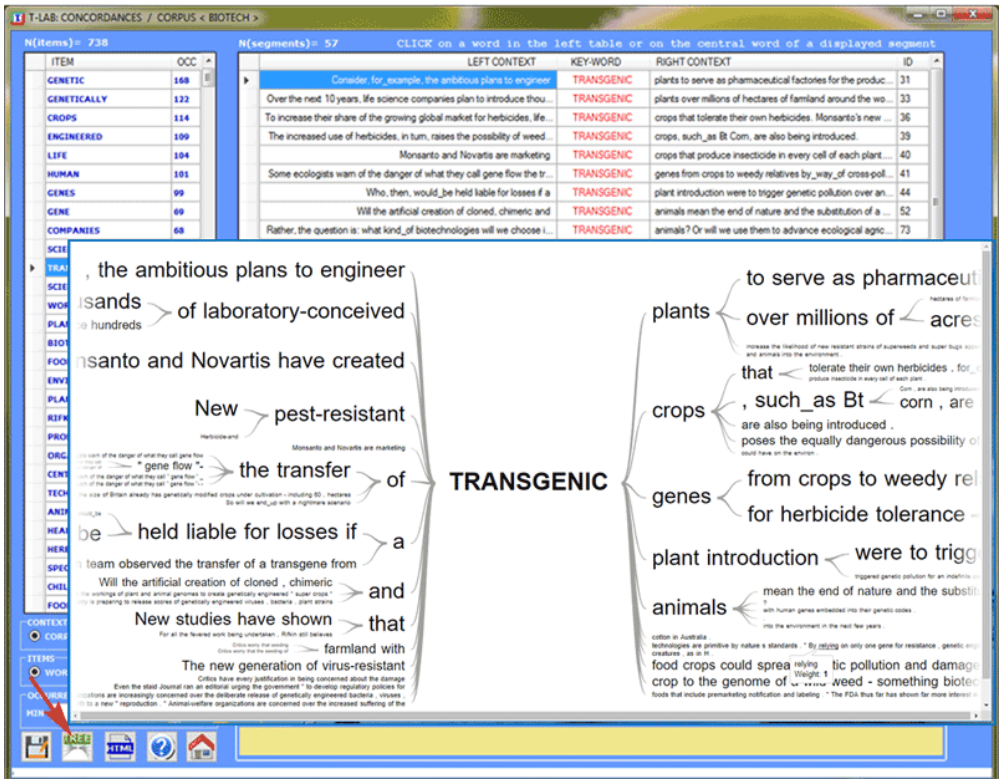
This **T-LAB** tool allows us to check the occurrence contexts of each lexical unit.

The KWIC (key-word in context) search can be carried out using two criteria: by **word** and by **lemma** (see option '2' below), both within the entire **corpus** or within a **subset** of it (see option '1' below).

It is also possible to define the occurrence (min. and max.) range (see option '3' below).



With a simple click on the corresponding column, you can, for each corpus lexical unit, verify what its occurrence contexts (the **elementary contexts**) are; furthermore, it is possible to create a dynamic **Word Tree** (see above option '4') or to save a HTML file with all the selected contexts (see above option '5').



DATE: 20/06/2015 - 09:29:24
CONCORDANCES **TRANSGENIC**

**** *ARTIC_0100
Consider, for_example, the ambitious plans to engineer **TRANSGENIC** plants to serve as pharmaceutical factories for the production of chemicals and drugs.

**** *ARTIC_0100
Over the next 10 years, life science companies plan to introduce thousands of laboratory-conceived **TRANSGENIC** plants over millions of hectares of farmland around the world. Ecologists tell us that the risks in releasing these crops into the biosphere are similar to those we_ve encountered in introducing exotic organisms into North America.

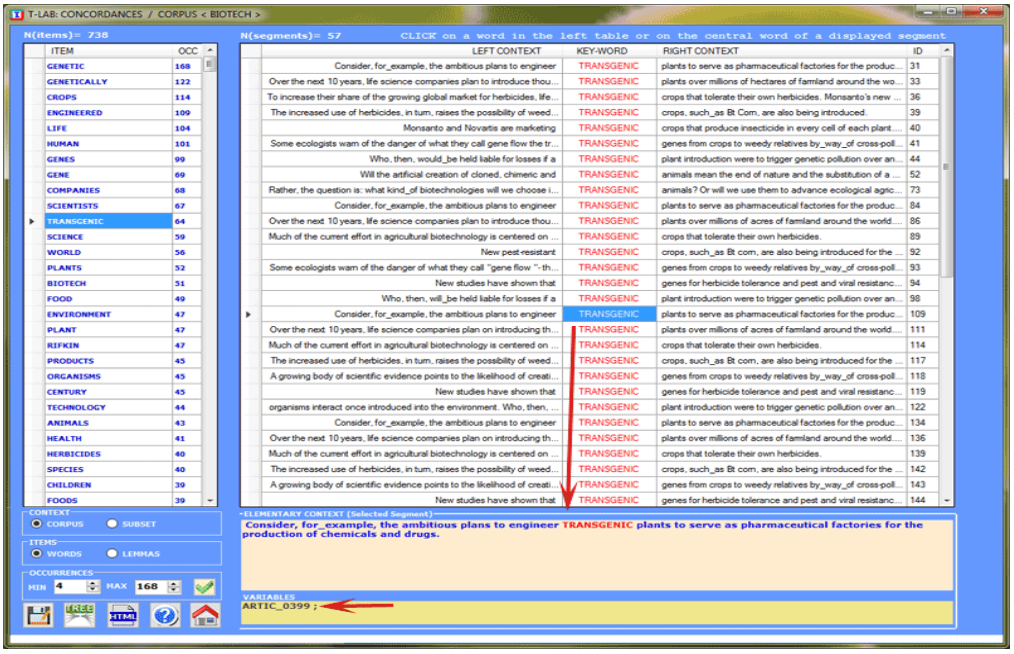
**** *ARTIC_0100
To increase their share of the growing global market for herbicides, life science companies such_as Monsanto and Novartis have created **TRANSGENIC** crops that tolerate their own herbicides. Monsanto's new herbicide-resistant patented seeds, for_example, are resistant to its best-selling chemical herbicide, Roundup.

**** *ARTIC_0100
The increased use of herbicides, in turn, raises the possibility of weeds developing resistance, forcing an even greater use of herbicides to control the more resistant strains. New pest-resistant **TRANSGENIC** crops, such_as Bt Corn, are also being introduced.

**** *ARTIC_0100
Monsanto and Novartis are marketing **TRANSGENIC** crops that produce insecticide in every cell of each plant. A growing body of scientific evidence points to the likelihood of creating "super bugs", resistant to the effects of the new pesticide-producing genetic crops.

**** *ARTIC_0100
Some ecologists warn of the danger of what they call gene flow the transfer of **TRANSGENIC** genes from crops to weedy relatives by_way_of cross-pollination. New studies have shown that transgenic genes for herbicide tolerance and pest and viral resistance can spread pollen and insert themselves into the genomes of relatives, creating weeds that are resistant to herbicides, pests and viruses.

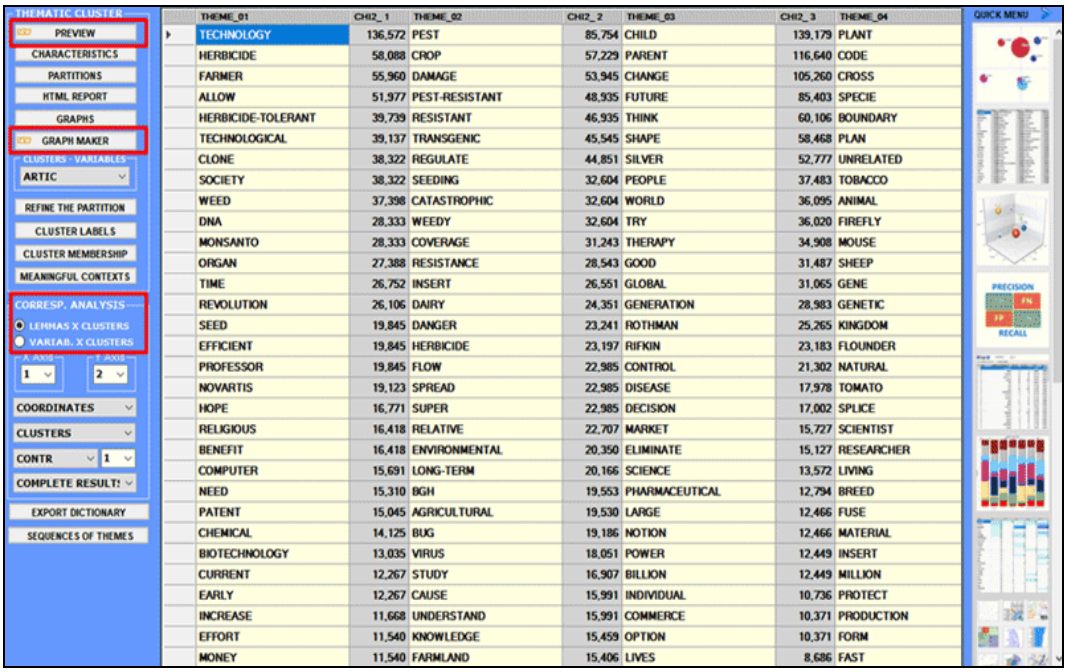
Moreover, by clicking the centre of displayed segment it is possible to visualise all its content and to check the variable categories used in its coding lines (see below).



THEMATIC ANALYSIS

Thematic Analysis of Elementary Contexts

N.B.: The pictures shown in this section have been obtained by using a previous version of **T-LAB**. These pictures look slightly different in **T-LAB 10**. Also: a) there is a new button (**TREE MAP PREVIEW**) which allows the user to create dynamic charts in HTML format; b) the DENDROGRAM button has been replaced by the **GRAPH MAKER** tool; c) a new table that shows in different columns the typical words of each cluster is available; d) when analysing a corpus which includes variable attributes, it is now possible to build and analyse tables which cross the themes and the attributes of each variable; e) a quick access gallery of pictures which works as an **additional menu** allows one to switch between various outputs with a single click. Some of these new features are highlighted in the below image.



THEME_01	CHI2_1	THEME_02	CHI2_2	THEME_03	CHI2_3	THEME_04
TECHNOLOGY	136,572	PEST	85,754	CHILD	139,179	PLANT
HERBICIDE	58,088	CROP	57,229	PARENT	116,640	CODE
FARMER	55,960	DAMAGE	53,945	CHANGE	105,260	CROSS
ALLOW	51,977	PEST-RESISTANT	48,935	FUTURE	85,403	SPECIE
HERBICIDE-TOLERANT	39,739	RESISTANT	46,935	THINK	60,106	BOUNDARY
TECHNOLOGICAL	39,137	TRANSGENIC	45,545	SHAPE	58,468	PLAN
CLONE	38,322	REGULATE	44,851	SILVER	52,777	UNRELATED
SOCIETY	38,322	SEEDING	32,604	PEOPLE	37,483	TOBACCO
WEED	37,398	CATASTROPHIC	32,604	WORLD	36,095	ANIMAL
DNA	28,333	WEEDY	32,604	TRY	36,020	FIREFLY
MONSANTO	28,333	COVERAGE	31,243	THERAPY	34,908	MOUSE
ORGAN	27,388	RESISTANCE	28,543	GOOD	31,487	SHEEP
TIME	26,752	INSERT	26,551	GLOBAL	31,065	GENE
REVOLUTION	26,106	DAIRY	24,351	GENERATION	28,983	GENETIC
SEED	19,845	DANGER	23,241	ROTHMAN	25,265	KINGDOM
EFFICIENT	19,845	HERBICIDE	23,197	RIFKIN	23,183	FLOUNDER
PROFESSOR	19,845	FLOW	22,985	CONTROL	21,302	NATURAL
NOVARTIS	19,123	SPREAD	22,985	DISEASE	17,978	TOMATO
HOPE	16,771	SUPER	22,985	DECISION	17,002	SPLICE
RELIGIOUS	16,418	RELATIVE	22,707	MARKET	15,727	SCIENTIST
BENEFIT	16,418	ENVIRONMENTAL	20,350	ELIMINATE	15,127	RESEARCHER
COMPUTER	15,691	LONG-TERM	20,166	SCIENCE	13,572	LIVING
NEED	15,310	BGH	19,553	PHARMACEUTICAL	12,794	BREED
PATENT	15,045	AGRICULTURAL	19,530	LARGE	12,466	FUSE
CHEMICAL	14,125	BUG	19,186	NOTION	12,466	MATERIAL
BIOTECHNOLOGY	13,035	VIRUS	18,051	POWER	12,449	INSERT
CURRENT	12,267	STUDY	16,907	BILLION	12,449	MILLION
EARLY	12,267	CAUSE	15,991	INDIVIDUAL	10,736	PROTECT
INCREASE	11,668	UNDERSTAND	15,991	COMMERCE	10,371	PRODUCTION
EFFORT	11,540	KNOWLEDGE	15,459	OPTION	10,371	FORM
MONEY	11,540	FARMLAND	15,406	LIVES	8,686	FAST

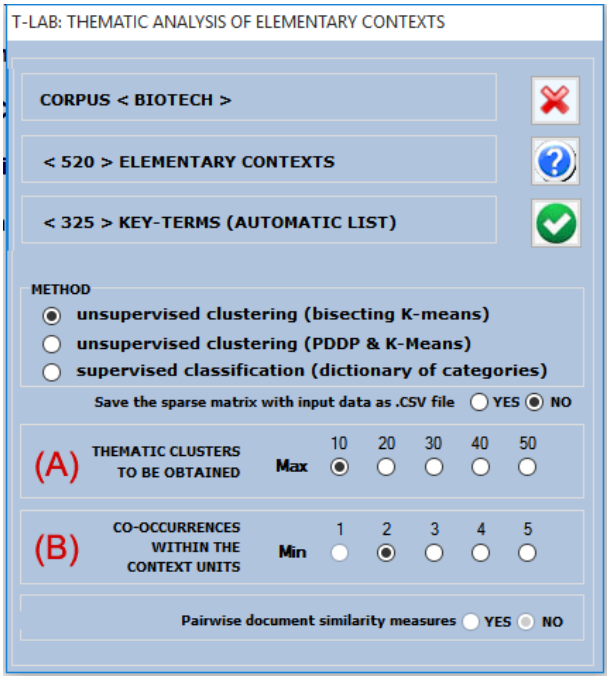
This **T-LAB** tool allows you to obtain and explore a **representation of corpus contents** through few and significant **thematic clusters** (from 3 to 50), each of which:

a) consists of a set of **elementary contexts** (i.e. sentences, paragraphs or short texts like responses to open-ended questions) characterized by the same patterns of key-words;

b) is described through the **lexical units** (i.e. words, lemmas or categories) and the **variables** (if present) most characteristic of the context units from which it is composed.

In many ways, analysis results can be considered as an **isotopy** (iso = same; topoi = places) map where each of them, as "generic" or "specific" theme (Rastier, 2002: 204), is characterized by the co-occurrences of semantic traits.

The analysis process can be performed through an unsupervised clustering (i.e. **bottom-up approach**), which is the default option, or a supervised classification (i.e. **top-down approach**). When choosing the latter (i.e. supervised classification), a dictionary of categories must be imported, either created by means of a previous **T-LAB** analysis or made up by the user.



A **T-LAB** dialog box (see above) allows the user to set some analysis parameters.

In particular:

- the (A) parameter allows the user to fix the maximum number of cluster partitions to be included in **T-LAB** outputs. Nonetheless, the clustering algorithm used stops when any further partition doesn't match statistical criteria;

- the (B) parameter allows the user to exclude from the analysis any context unit that doesn't contain a minimum number of key-words included in the list which is being used.

N.B.:
- When selecting the 'supervised classification' option, as the number of clusters to be obtained coincides with the number of categories present in the dictionary, the (A) parameter is not available;

Both the above parameters produce significant changes in the analysis results only when the number of context units is very large and/or when they are short texts.

In the case of **unsupervised clustering** (default option), the analysis procedure consists of the following steps:

a - construction of a data table context units x lexical units (up to 300,000 rows x 3,000 columns), with presence/absence values;
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; references: Steinbach, Karypis, & Kumar, 2000; Savaresi, Booley, 2001);
d - filing of the obtained partitions and, for each of them:
e- construction of a contingency table lexical units x clusters (n x k);
f- chi square test applied to all the intersections of the contingency table;
g- correspondence analysis of the contingency table lexical units x clusters ((references: Benzécri, 1984; Greenacre, 1984; Lebart, Salem, 1994).

N.B.: Starting from T-LAB Plus 2016, the unsupervised clustering of the context units (see step 'c' above) can be performed either by using the bisecting K-means algorithm (1) or by using a not centered version of PDDP(i.e. *Principal* Direction Divisive Partitioning) proposed by D. Boley (1998) for selecting the seeds of each K-means bisection.
The main differences between the above methods relies on how the two seeds of each bisection are computed; in fact, in the (1) case they result from an iterative procedure whereas in the (2) case they are computed through SVD (i.e. Singular Value Decomposition), i.e. through a 'one-shot' algorithm (see Savaresi, S.M., & Boley, D.L. , 2004).

This procedure therefore performs a type of **co-occurrence analysis** (steps a-b-c) and, subsequently, a type of **comparative analysis** (steps e-f-g). In particular, comparative analysis uses the categories of the "new variable" derived from the co-occurrence analysis (categories of the new variable = thematic clusters) to form the contingency table columns.

In the case of **supervised classification** the steps of comparative analysis are the same (see e-f-g above), whereas co-occurrence analysis is performed as follows:

a) normalization of the seed vectors (i.e. co-occurrence profiles) corresponding to the 'k' categories of the dictionary used;

b) computation of Cosine similarity and of Euclidean distance between each 'i' context unit and each 'k' seed vector;

c) assignment of each 'i' context unit to the 'k' class or category for which the corresponding seed is the closest (In this case, maximum Cosine similarity and minimum Euclidean distance must coincide, otherwise T-LAB consider the 'i' context unit as unclassified).

N.B.: When the user decides to repeat/apply the results of a previous analysis (i.e. a **Thematic Analysis of Elementary Contexts** or a **Modeling of Emerging Themes**), **T-LAB** performs a comparative analysis only (steps e-f-g).

On completion of the analysis you can easily perform the following operations:

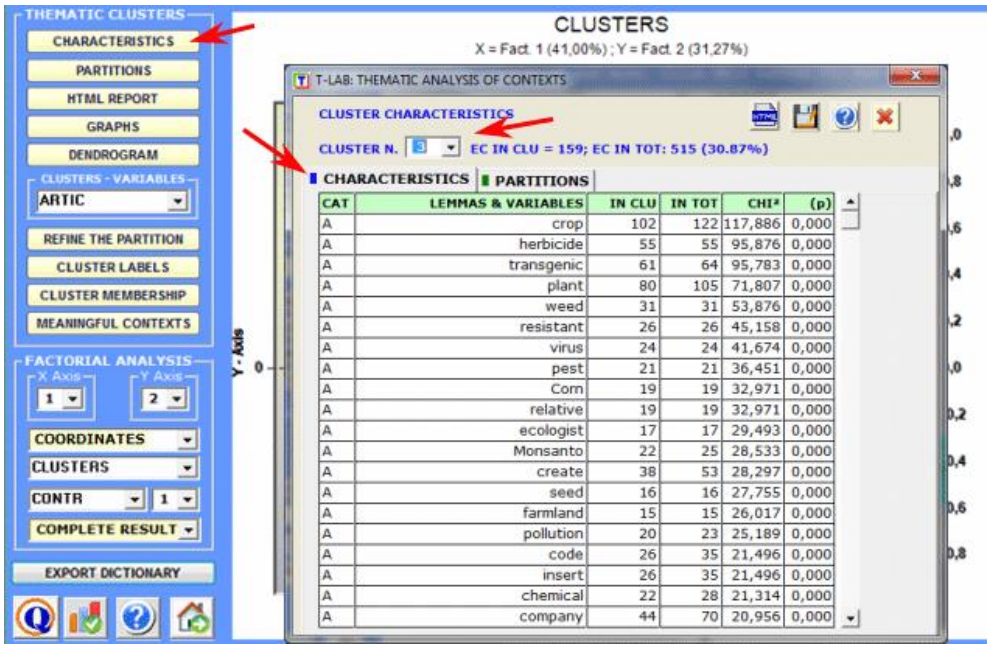
1 - explore the characteristics of the clusters;
2 - explore the relationships between the clusters;
3 - explore the relationships between clusters and variables;
4 - explore the various cluster partitions (from 3 to 50);
5 – refine the results of the chosen partition and, if necessary, repeat the above steps (1,2,3);
6 - assign labels to the clusters;

7 - verify which elementary contexts belong to each cluster;
8 - verify the score of each elementary context within the cluster to which it belongs;
9 – export a thematic document classification (only provided when the corpus is made up of at least 2 primary documents and when they are not short texts like responses to open ended questions);
10 - save the selected partition for exploration with other **T-LAB** tools;
11 – export a dictionary of categories;
12 – validate the chosen partition and assess the semantic coherence of each theme;

13 - moreover, when the corpus is structured like a discourse or like a conversation (i.e. the context units follow each other in a temporal order), it is possible to explore theme sequences by means of animated charts (see below, the final part of this section).

In detail:

1 - Explore the characteristics of the clusters



CAT	LEMMAS & VARIABLES	IN CLU	IN TOT	CHI²	(p)
A	crop	102	122	117,886	0,000
A	herbicide	55	55	95,876	0,000
A	transgenic	61	64	95,783	0,000
A	plant	80	105	71,807	0,000
A	weed	31	31	53,876	0,000
A	resistant	26	26	45,158	0,000
A	virus	24	24	41,674	0,000
A	pest	21	21	36,451	0,000
A	Corn	19	19	32,971	0,000
A	relative	19	19	32,971	0,000
A	ecologist	17	17	29,493	0,000
A	Monsanto	22	25	28,533	0,000
A	create	38	53	28,297	0,000
A	seed	16	16	27,755	0,000
A	farmland	15	15	26,017	0,000
A	pollution	20	23	25,189	0,000
A	code	26	35	21,496	0,000
A	insert	26	35	21,496	0,000
A	chemical	22	28	21,314	0,000
A	company	44	70	20,956	0,000

Clicking on the **CHARACTERISTICS** button shows the lexical units and the variable values which characterize each cluster: Chi-square values and the sums of the elementary contexts in which it is found, both in the selected cluster ("IN CLUST") and in the analysed total ("IN TOT"). The "CAT" column also indicates whether the characteristic has been selected by the user ("A") with the **Customized Settings** function or has been suggested by **T-LAB** as a "supplementary" description ("S").

In the case of the chi square test the structure of the analysed table is the following:

	Cluster "A"	Other Clusters	
Word "a"	n_{ij}		N_j
Other Words			
	N_i		N

Where:

n_{ij} refers to occurrences of word (a) within the selected cluster (A)
N_j refers to all occurrences of word (a) within the corpus (or the corpus sub-set) analysed
N_i refers to all word occurrences within the selected cluster (A)
N refers to all word occurrences of the contingency table word by cluster.

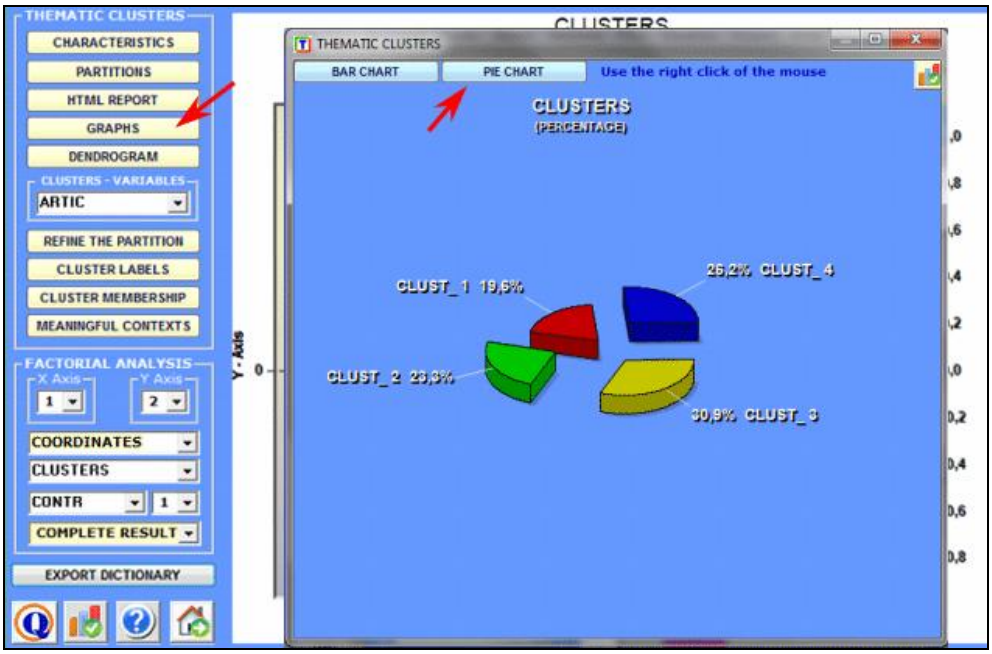
An **HTML report** (see below) is generated to permit detailed analysis of the cluster characteristics. In the report, in addition to the list of typical words, the most characteristic elementary contexts of the selected cluster are shown in descending order according to their respective score.

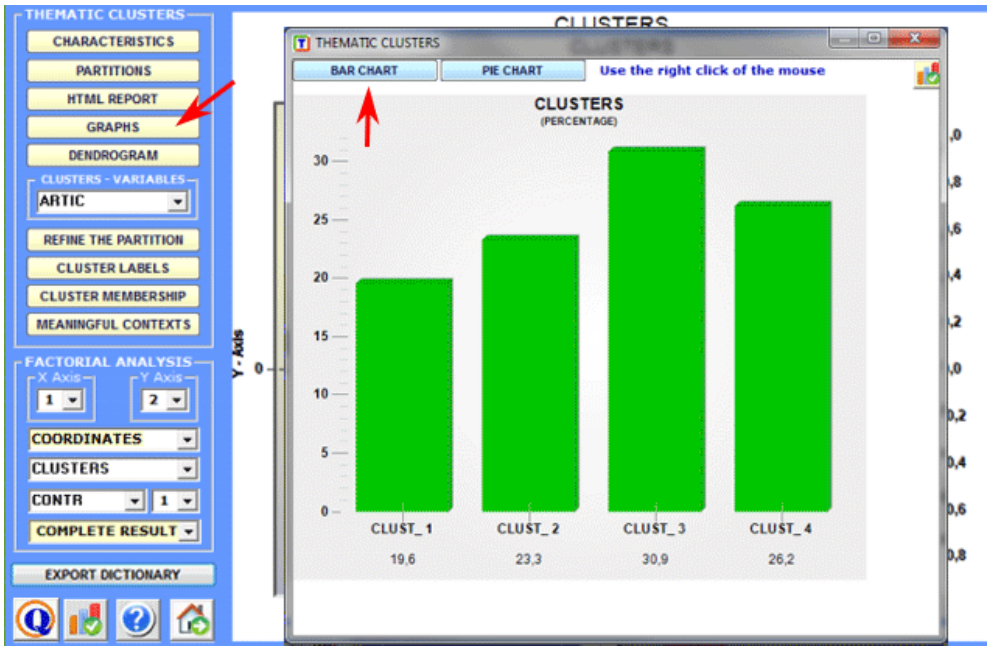
LEMMA	CHI SQUARE	WORD	OCC
crop	117.886	crop	8
crop	117.886	crops	94
herbicide	95.876	herbicide	15
herbicide	95.876	herbicides	40
transgenic	95.783	transgenic	61
plant	71.807	plant	43
plant	71.807	planted	2
plant	71.807	plants	35
weed	53.876	weed	2

**** *ARTIC_0100
SCORE (238.757)

Some ecologists warn of the danger of what they call gene flow the transfer of transgenic genes from crops to weedy relatives by_way_of cross-pollination. New studies have shown that transgenic genes for herbicide tolerance and pest and viral resistance can spread pollen and insert themselves into the genomes of relatives, creating weeds that are resistant to herbicides, pests and viruses.

Pie charts and **bar charts** are used to verify the percentage of context units (i.e. elementary contexts) that belong to each cluster



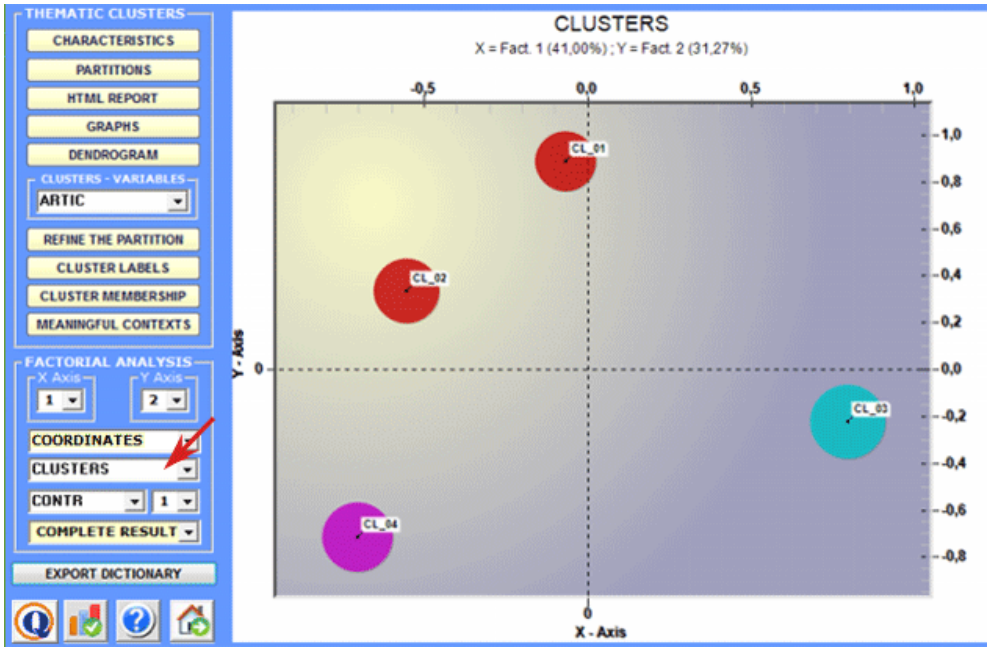


2 - **Explore the relationships between the clusters**

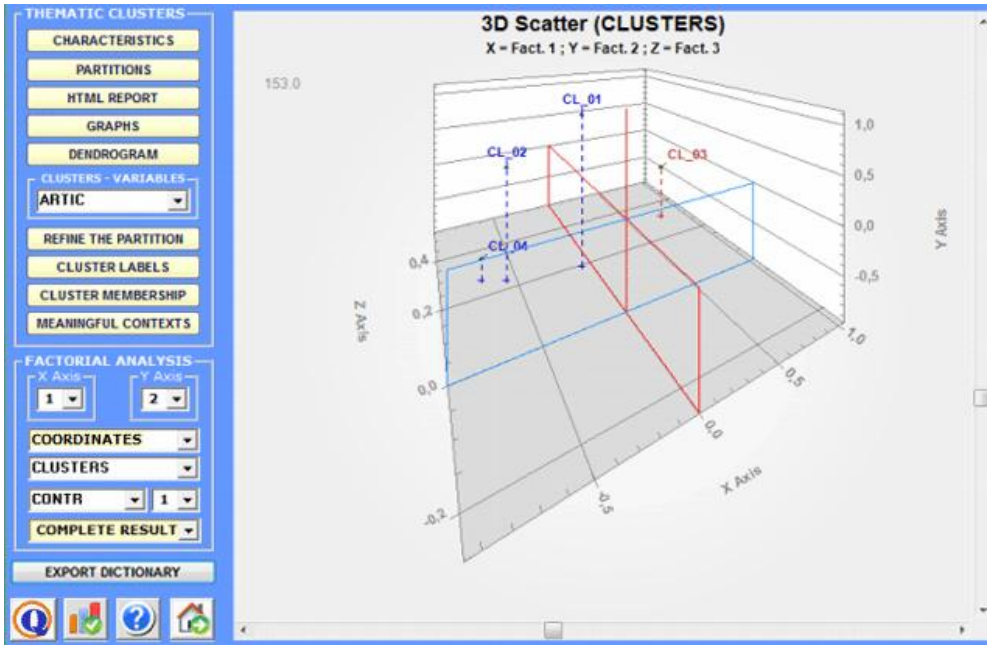
Some of the graphs obtained by **Correspondence Analysis** enable you to explore the relationships between clusters in bi-dimensional spaces.

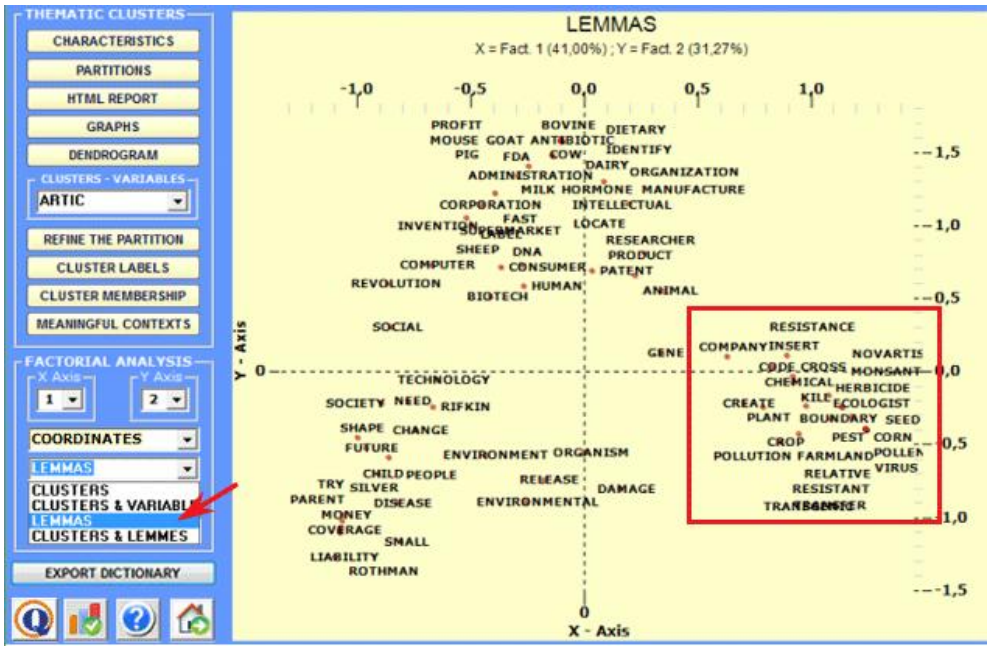
More specifically:

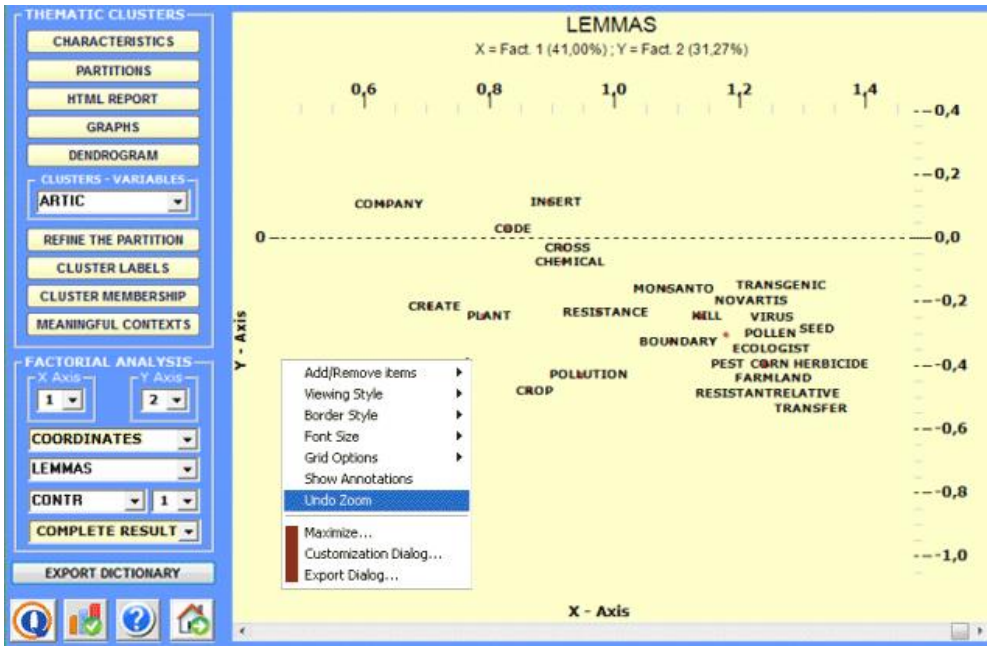
- You can explore the various combinations of factorial axes, simply by selecting them in the appropriate boxes ("X axis", " Y axis");
- For each of the combinations (X-Y), you can display various types of elements (clusters, lemmas and variables).



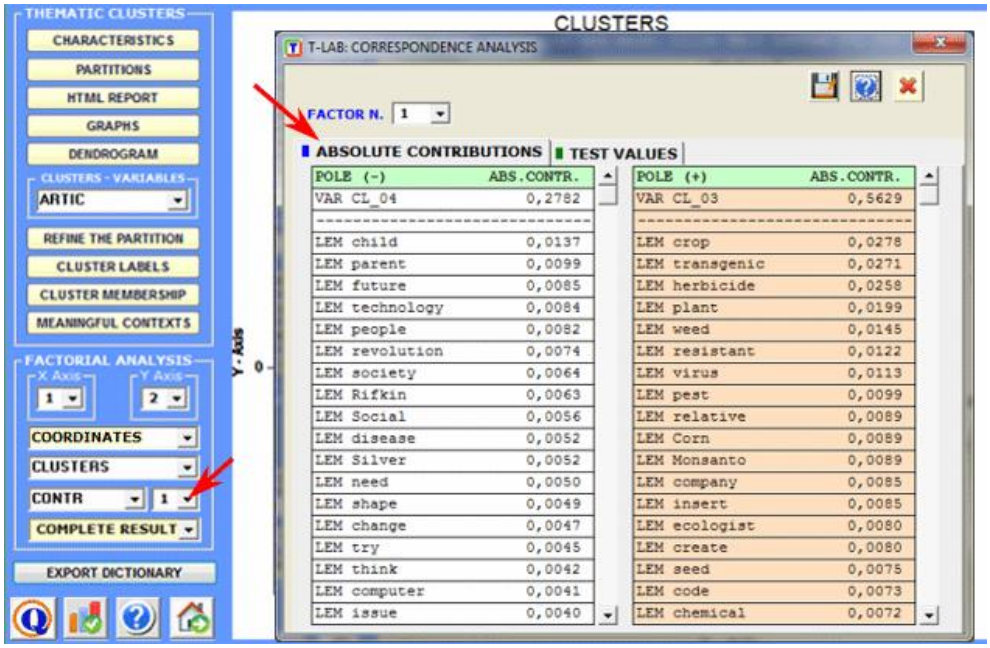
All the graphs can be maximized and customized by using the appropriate dialog box (just right click on the chart). Moreover, when thematic clusters are 4 or more, their relationships can be explored through **3d** moving (see below).



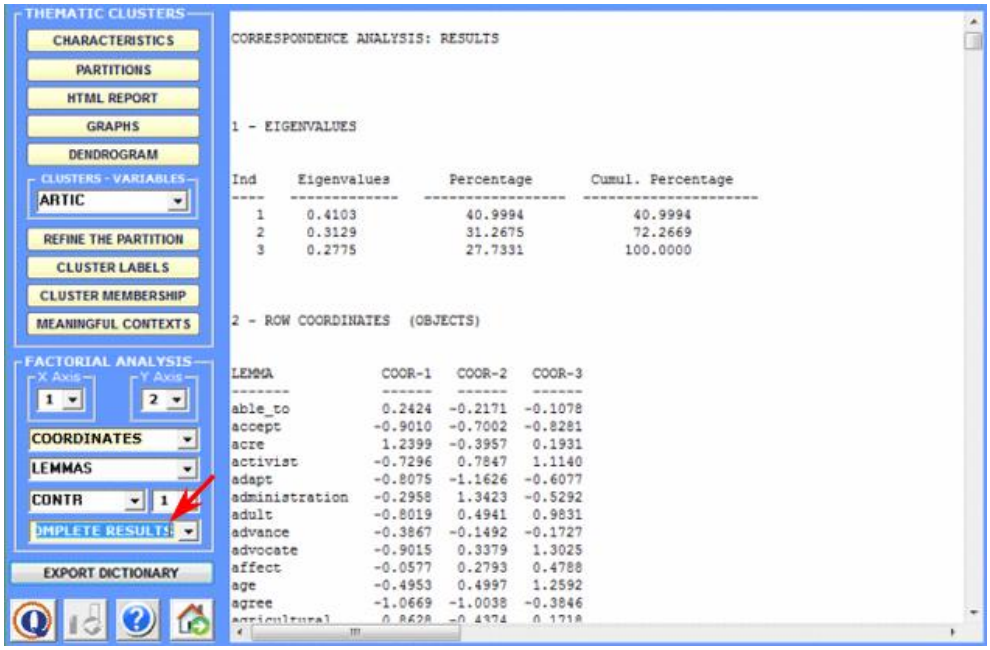




Moreover, for every factorial axis, **T-LAB** supplies two tables that facilitate the interpretation. These are shown after every selection in the appropriate boxes (see below).

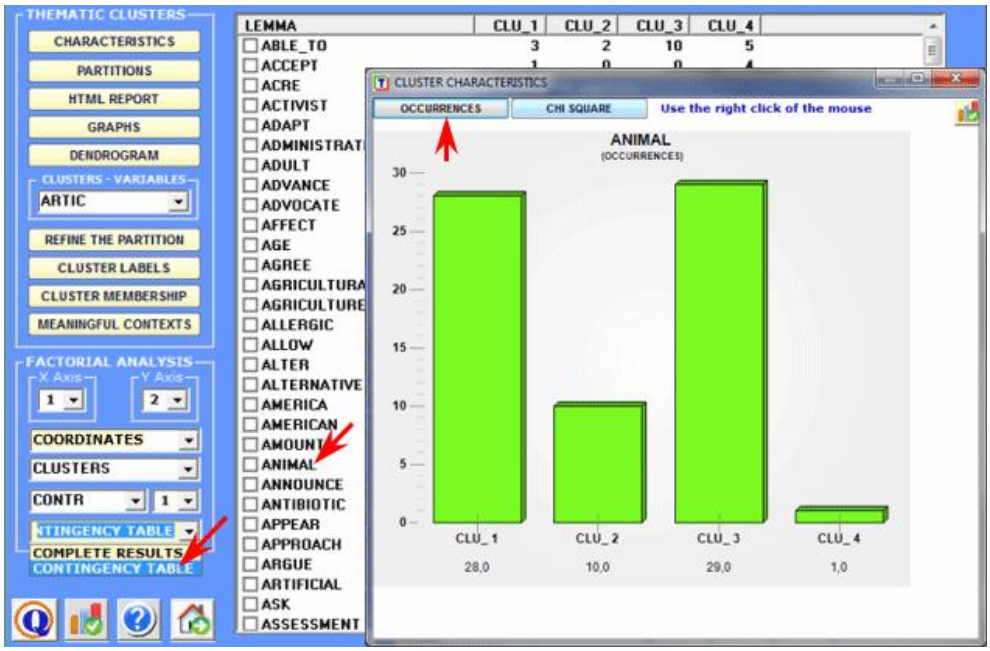


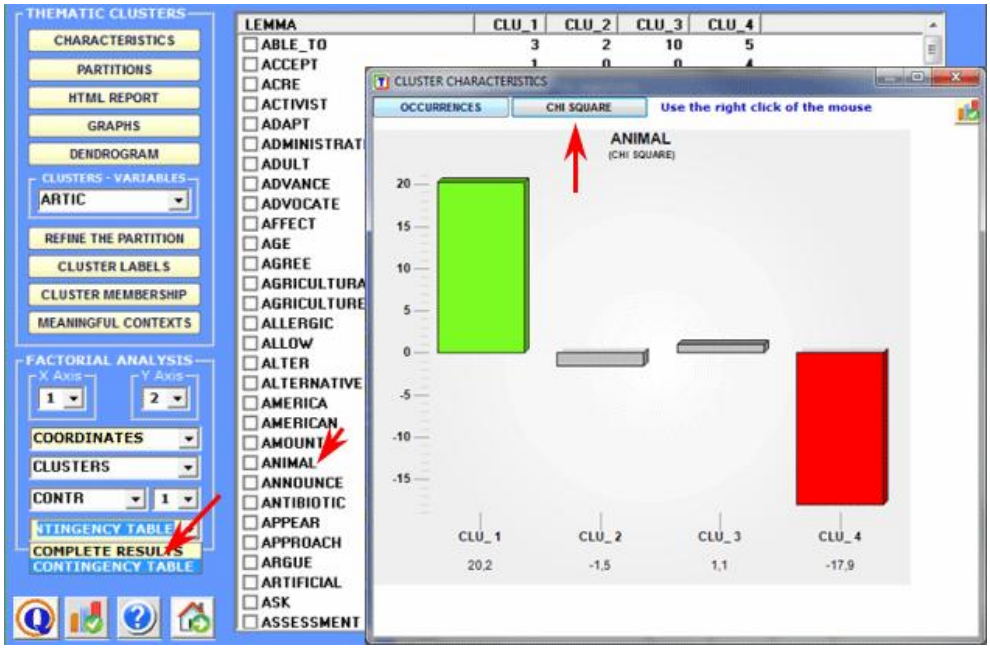
By selecting the **Complete Results** option it is possible to check all the results of the Correspondence Analysis lexical units x clusters.



A specific option (see below) allows us to visualise/export the contingency table and to create charts showing the distribution of each word within the clusters and their corresponding chi-square value.
Moreover, by clicking on specific cells of the table, it is possible to create a HTML file including all elementary contexts where the word in row is present in the corresponding cluster.

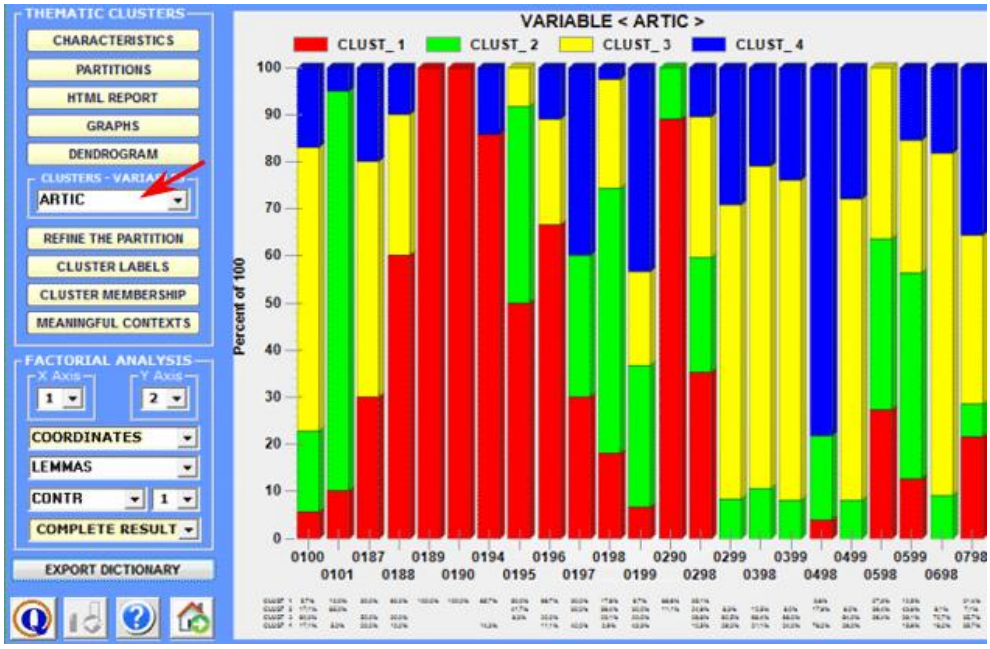
N.B.: Such a table includes both active ('A') and supplementary ('S') key-words.





3 - **Explore the relationships between clusters and variables**

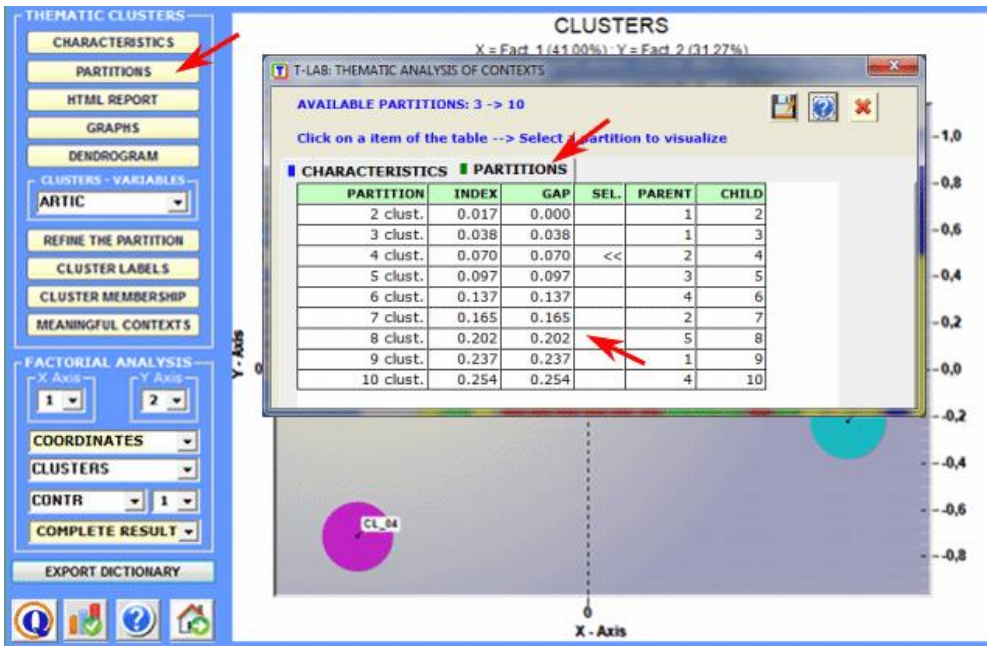
Bar charts allow you to verify the relationships between clusters and variables.



You can explore additional relationships between clusters and variables using the functions provided in the **Factor Analysis** section (see above).

4 - **Explore the various cluster partitions**

Because the algorithm used (bisecting K-means) produces a hierarchical clustering, the user can explore various analysis solutions: partitions from 3 to 50 clusters.

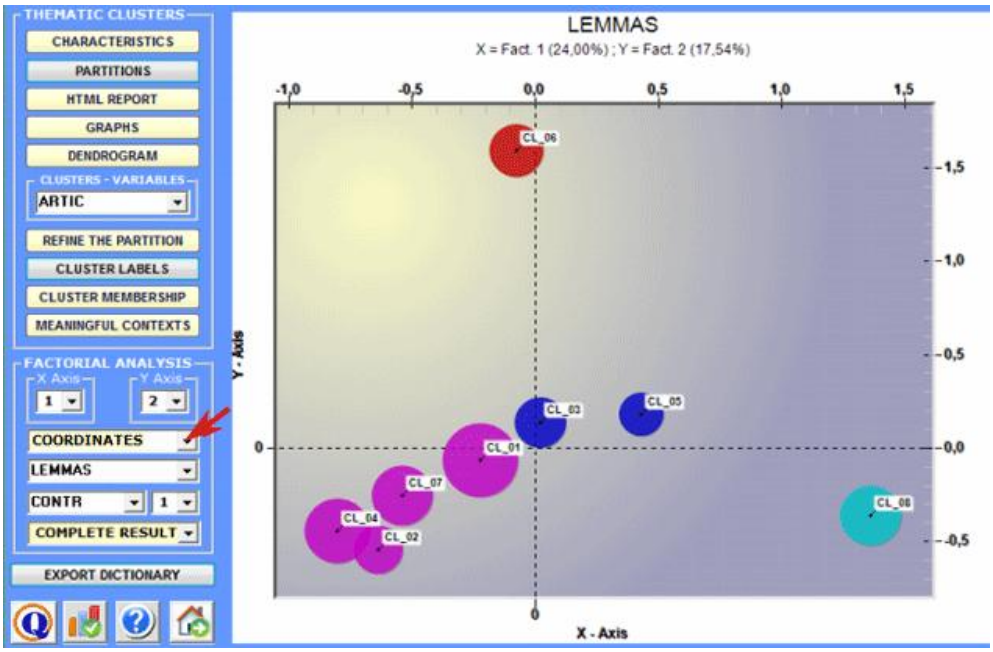


For each partition obtained, a specific table (see above) lists the following values:

- "Index", obtained by dividing the between cluster variance by the total variance;

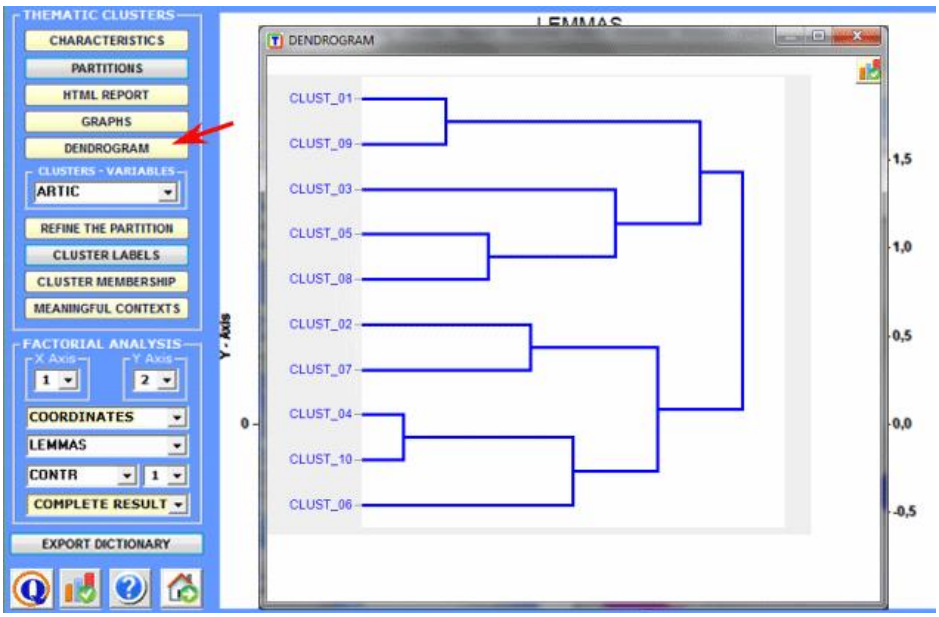
- "Gap", corresponding to the difference between the index value and the value of the immediately previous partition;
- the number of the "child" cluster obtained from the bisection of the corresponding "parent".

The **Partition** option allows you to easily explore the characteristics of the available clustering solutions (just click on a table item)

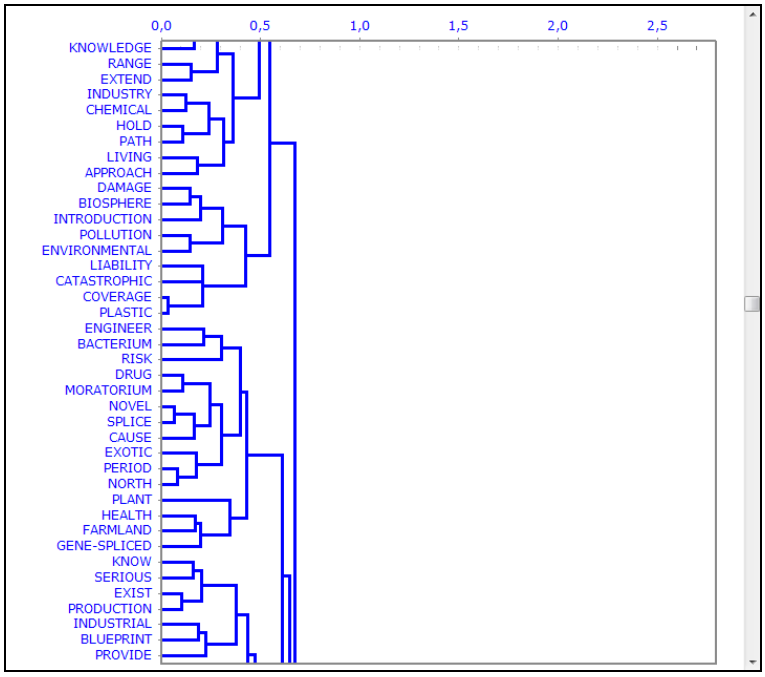


Moreover the **dendrogram** option (see below) allows two possibilities:

a) to check the tree structure of the various bisections;



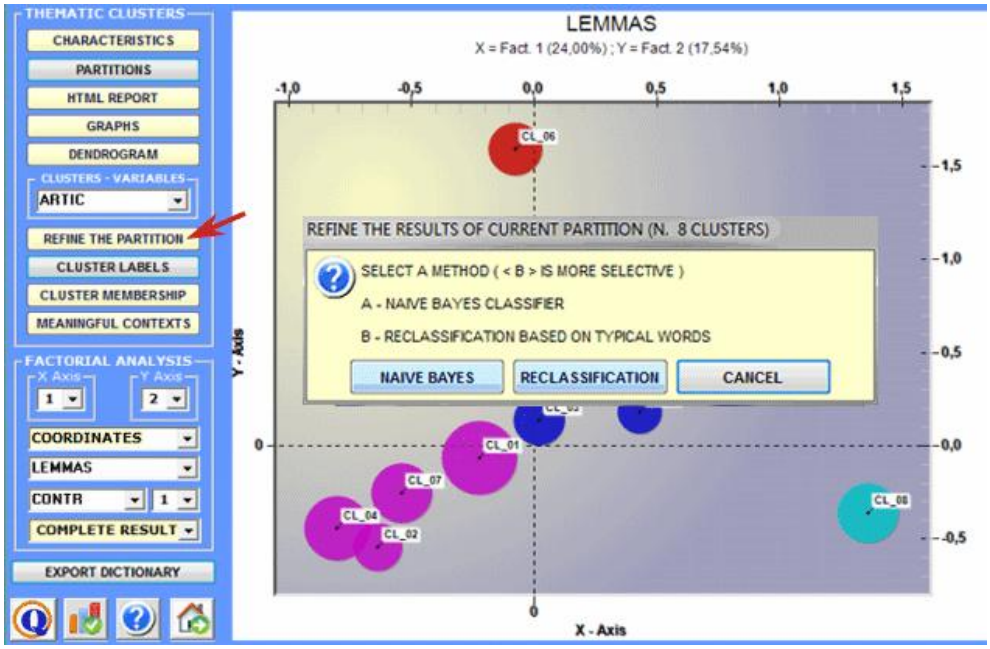
b) to check the tree with the characteristic words of each cluster.



5 – **Refine the results of the chosen partition**

After having explored different solutions, the user can refine the results of the chosen partition and, if necessary, repeat some of the three operations above illustrated.

For this purpose two methods are available (see the picture below).



When the '**A**' method (i.e. Naïve Bayes Classifier) is chosen, this step allows the user to delete from the analysis all context units of which cluster membership doesn't fit either of the following criteria:

a) the cluster memberships of the i-context unit, determined by the bisecting K-means first (unsupervised clustering) and by a Naïve Bayes Classifier later (supervised clustering), must be the same;
b) the maximum posterior value (see below) corresponding to the i-context unit cluster membership must be, in percentage terms, at least 50% higher than its remaining values (i.e. posterior value in other clusters).

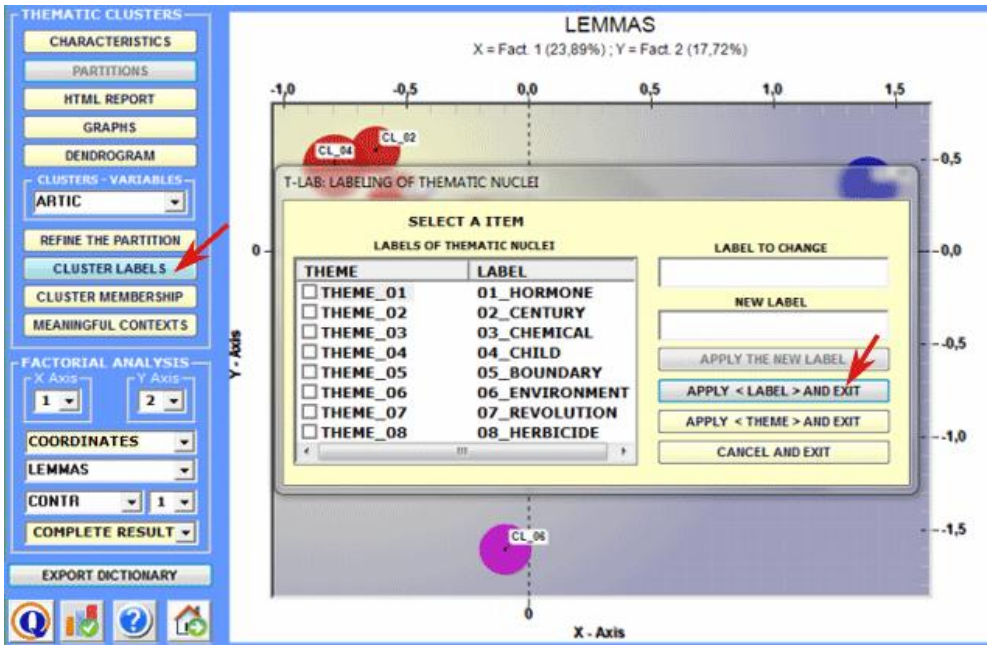
Whereas, in the case of '**B**' method (i.e. Reclassification Based on Typical Words) **T-LAB** considers the cluster characteristics - i.e. the words with a significant Chi-Square value - like items of a category dictionary and performs the three steps of 'supervised classification' described at the beginning of this section. So, when the user is interested in re-using dictionaries and in comparing the analysis results, this method is highly recommended.

All the results of this computation are in the following table exported by **T-LAB** (see below), where the posteriori values for each cluster are in percentage format.

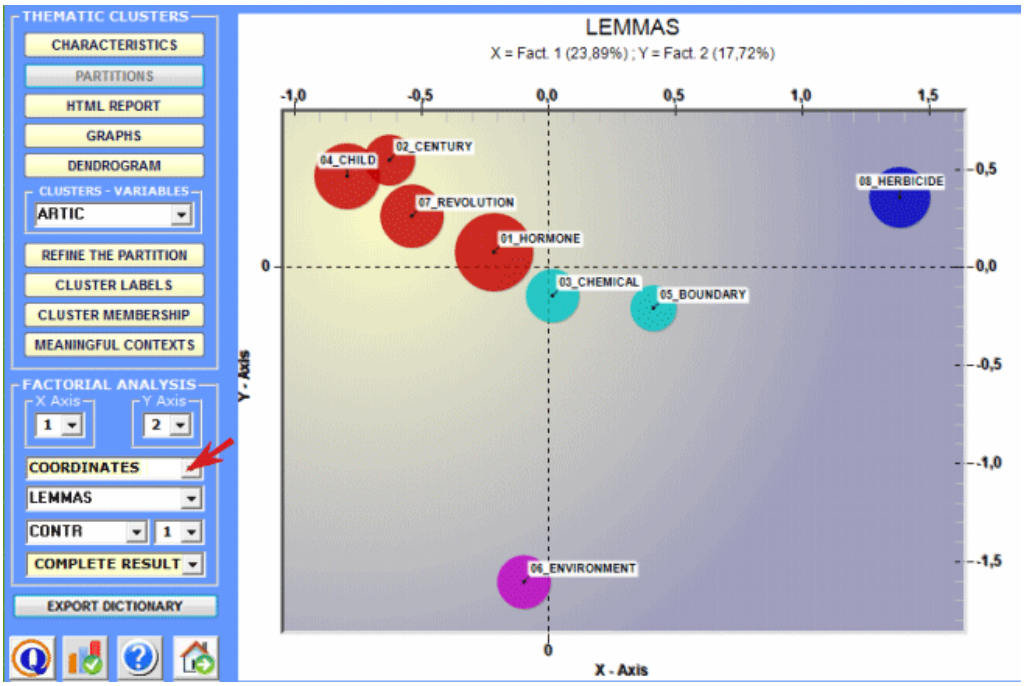
Context_ID	OLD	NEW	MATCH	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CONTEXT
'00001000001	2	2	YES	0	1	0	0	0	0	0	0	Left and right are finding common ground in op
'00001000002	2	2	YES	0	1	0	0	0	0	0	0	The current debate over embryo stem cell rese
'00001000003	2	2	YES	0	1	0	0	0	0	0	0	Although reluctant to acknowledge it , both sc
'00001000004	2	2	YES	0	1	0	0	0	0	0	0	The threads that unite these two groups are th
'00001000005	4	4	YES	0,006	0	0	0,994	0	0	0	0	The former crusade for what they regard as the
'00001000006	2	2	YES	0	1	0	0	0	0	0	0	But , as we make the great transformation fro
'00001000007	2	2	YES	0	1	0	0	0	0	0	0	that_is because , while the industrial age divic
'00001000008	7	7	YES	0	0	0	0	0	0	1	0	with those who champion the intrinsic value of
'00001000009	1	1	YES	1	0	0	0	0	0	0	0	The latter say that any ' vitalistic ' notion of li
'00001000010	7	7	YES	0	0	0	0	0	0	1	0	Increasingly , in the years ahead , individuals
'00001000011	2	2	YES	0	1	0	0	0	0	0	0	The emerging connection between social cons
'00001000012	2	2	YES	0	1	0	0	0	0	0	0	Left activists , on the other hand , are more a
'00001000013	1	1	YES	1	0	0	0	0	0	0	0	Both sides come together in their opposition to
'00001000014	7	7	YES	0	0	0	0	0	0	1	0	The former argue that life is God ' s creation ,
'00001000015	2	2	YES	0	1	0	0	0	0	0	0	The genetic foods issue has also brought toge
'00001000016	2	2	YES	0	1	0	0	0	0	0	0	Finally , on the subject of designer babies , s
'00001000017	2	2	YES	0	1	0	0	0	0	0	0	The point is that , although social conservative
'00001000018	2	2	YES	0	1	0	0	0	0	0	0	What is equally true , however , is that on so
'00001000019	2	2	YES	0	1	0	0	0	0	0	0	and the left activists with their alliances with s
'00001000020	2	2	YES	0	1	0	0	0	0	0	0	Of this much we can be sure : the biotech era
'00002000001	7	7	YES	0	0	0	0	0	0	1	0	With genetic modification crossing plant , ani
'00002000002	7	7	YES	0	0	0	0	0	0	1	0	WHILE the biotech revolution will reshape the
'00002000003	7	7	YES	0	0	0	0	0	0	1	0	the genetic components that help orchestrate
'00002000004	3	3	YES	0	0	1	0	0	0	0	0	Imagine the wholesale transfer of genes betwe
'00002000005	2	2	YES	0	1	0	0	0	0	0	0	Then , with clonal propagation , mass-produc
'00002000006	3	3	YES	0	0	1	0	0	0	0	0	Genetic pollution is already appearing and is li
'00002000007	5	5	YES	0	0	0	0	1	0	0	0	Troubling questions are already being raised a
'00002000008	5	5	YES	0	0	0	0	1	0	0	0	for_example , scientists have introduced an a
'00002000009	5	5	YES	0	0	0	0	1	0	0	0	other animals and plants into the genetic code
'00002000010	3	3	YES	0	0	1	0	0	0	0	0	Consider , for_example , the ambitious plans

6 - Assign labels to the clusters

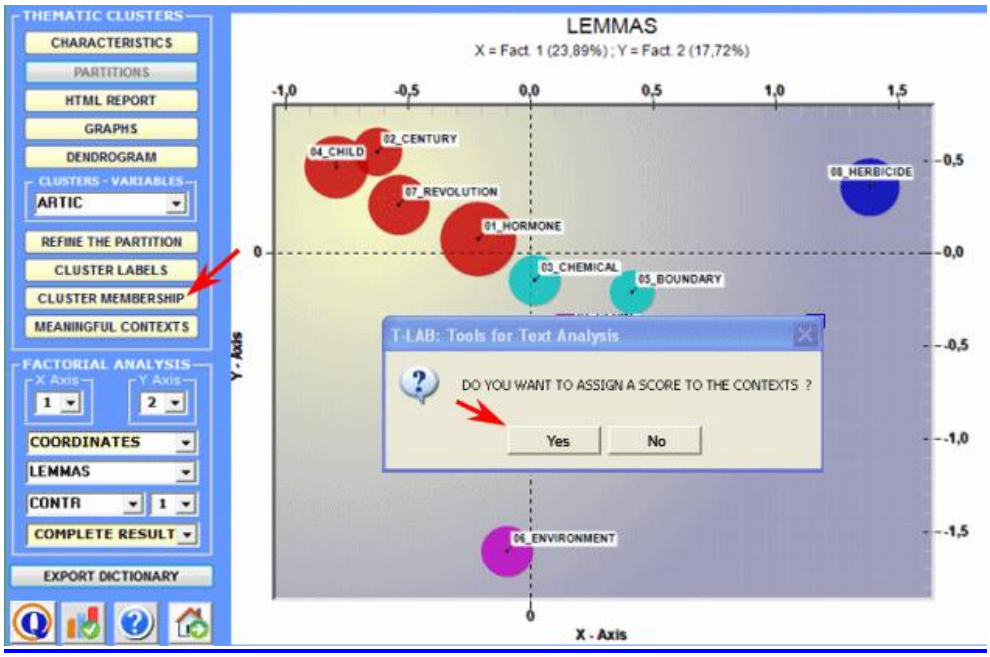
A specific **T-LAB** function allows you to assign labels to clusters. (N.B: The software proposes a number of labels automatically the first time you use this function.)



Labels assigned to clusters can be displayed in the various graphs available (see below).



7 - Verify which elementary contexts belong to each cluster; (8) Verify the score of each elementary context within the cluster to which it belongs; (9) Obtain a thematic document classification



In fact the **Cluster Membership** button lets you export three types of tables (see below) in MS Excel format:

a – "Cluster_Partitions.xls" listing all the context unit correspondence for each cluster within the various partitions;

IDNUMBE	PART-2	PART-3	PART-4	PART-5	PART-6	PART-7	PART-8	PART-9
1	2	2	4	4	4	4	4	
2	1	3	3	3	3	3	3	
3	1	3	3	3	6	6	6	
4	1	1	1	5	5	5	5	
5	2	2	4	4	4	4	4	
6	1	3	3	3	6	6	6	
7	1	3	3	3	3	3	3	
8	1	3	3	3	6	6	6	
9	2	2	4	4	4	4	4	
10	1	3	3	3	6	6	6	
11	1	1	1	1	1	1	1	
12	1	1	1	1	1	1	1	
13	1	3	3	3	3	3	3	
14	1	3	3	3	6	6	6	
15	1	1	1	1	1	1	1	
16	1	3	3	3	6	6	6	
17	1	3	3	3	3	3	3	
18	1	3	3	3	6	6	6	
19	1	1	1	5	5	5	5	
20	1	1	1	5	5	5	5	
21	1	3	3	3	3	3	3	
22	2	2	2	2	2	7	7	
23	1	3	3	3	3	3	3	
24	1	1	1	5	5	5	5	
25	2	2	2	2	2	2	2	
26	2	2	2	2	2	2	2	
27	2	2	2	2	2	2	2	
28	1	3	3	3	6	6	6	

b – Themes-Contexts.xls (see below) listing the context unit correspondences for each cluster within the selected partition.

(IDNUMBER)	THEME	SCORE	CONTEXT
'00001000001	02_CENTURY	53,61	Left and right are finding common ground in opposition to a utilitarian view of
'00001000002	02_CENTURY	44,2	The current debate over embryo stem cell research , as_well_as the debates
'00001000003	02_CENTURY	55,75	Although reluctant to acknowledge it , both social conservatives and left acti
'00001000004	02_CENTURY	94,74	The threads that unite these two groups are their belief in and commitment to
'00001000005	04_CHILD	4,1	The former crusade for what they regard as the rights of the unborn , and rail
'00001000006	02_CENTURY	36,18	But , as we make the great transformation from the age of physics and cher
'00001000007	02_CENTURY	38,49	that_is because , while the industrial age divided people from right to left bas
'00001000008	07_REVOLUTION	24,62	with those who champion the intrinsic value of life on one pole and those who
'00001000009	01_HORMONE	2,63	The latter say that any ' vitalistic ' notion of life is a throwback to religious r
'00001000010	07_REVOLUTION	1,11	Increasingly , in the years ahead , individuals , families , communities and
'00001000011	02_CENTURY	113,75	The emerging connection between social conservatives and left activists is al
'00001000012	02_CENTURY	48,23	Left activists , on the other hand , are more ambiguous about the status of t
'00001000013	01_HORMONE	34,97	Both sides come together in their opposition to the cloning of human embryo
'00001000014	07_REVOLUTION	30,11	The former argue that life is God ' s creation , not a human invention , and tl
'00001000015	02_CENTURY	120,28	The genetic foods issue has also brought together social conservatives wary
'00001000016	02_CENTURY	25,39	Finally , on the subject of designer babies , social conservatives believe tha
'00001000017	02_CENTURY	60,59	The point is that , although social conservatives and left activists view the wo
'00001000018	02_CENTURY	41,51	What is equally true , however , is that on some of the most important ques
'00001000019	02_CENTURY	25,52	and the left activists with their alliances with social democratic parties . of_co
'00001000020	02_CENTURY	84,11	Of this much we can be sure : the biotech era will bring with it a very differen
'00002000001	07_REVOLUTION	2,51	With genetic modification crossing plant , animal and human boundaries , a
'00002000002	07_REVOLUTION	8,42	WHILE the biotech revolution will reshape the global economy and remake ou
'00002000003	07_REVOLUTION	9,18	the genetic components that help orchestrate the developmental processes i
'00002000004	03_CHEMICAL	19,54	Imagine the wholesale transfer of genes between totally unrelated species an
'00002000005	02_CENTURY	10,08	Then , with clonal propagation , mass-producing countless replicas of these
'00002000006	03_CHEMICAL	16,55	Genetic pollution is already appearing and is likely to spread in the biotech c
'00002000007	05_BOUNDARY	22,34	Troubling questions are already being raised about the widespread introductio
'00002000008	05_BOUNDARY	358,54	for_example , scientists have introduced an anti-freeze gene from flounder fis
'00002000009	05_BOUNDARY	45,7	other animals and plants into the genetic code of traditional food crops . Eco
'00002000010	03_CHEMICAL	38,17	Consider , for_example , the ambitious plans to engineer transgenic plants

In particular, the relevance value (Score) assigned to each elementary context (j) belonging to the cluster (k) comes from the following formula

$$score_j = \sum_{j \in k} X_{i,j} \times \frac{n_j}{N}$$

Where:

Score$_j$ = relevance value assigned to the elementary context (j);

ΣX_{ij} = sum of the Chi-square values corresponding to the key-words (i) found in the elementary context in question (j) which are typical of the cluster (k);

n_j = number of key-words (distinct words), typical of the cluster (k), found in the elementary context (j);

N = number of key-words (distinct words) typical of the cluster (k).

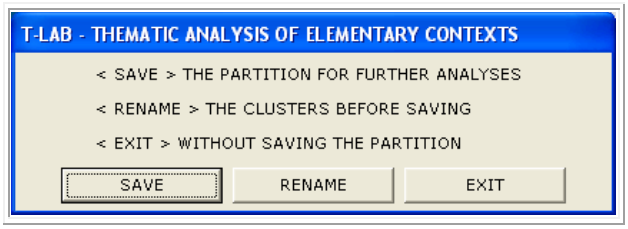
c – "Ec_Document_Classification.xls" (only provided when the corpus is made up of at least 2 primary documents and when they are not short texts like responses to open ended questions) listing the mixed cluster membership of each document (see below).

In this case the values come from the above formula (see "b") by summing the scores of elementary contexts belonging to each document and by applying a percentage calculation.

DOC_ID	VAR_01	BEST_CLU	CLUST_ 1	CLUST_ 2	CLUST_ 3	CLUST_ 4	CLUST_ 5	CLUST_ 6	CLUST_ 7	CLUST_ 8
1	AR_0101	2	0,04	0,896	0	0,004	0	0	0,059	0
2	AR_0100	8	0	0,004	0,029	0	0,2	0,167	0,029	0,572
3	AR_0199	2	0,023	0,358	0,077	0,206	0,127	0,08	0,095	0,033
4	AR_0299	8	0	0	0,015	0	0,209	0,169	0,026	0,581
5	AR_0399	8	0	0	0,014	0	0,199	0,167	0,023	0,597
6	AR_0499	8	0	0	0,013	0	0,2	0,149	0,022	0,616
7	AR_0599	2	0,046	0,225	0,223	0,123	0,068	0,038	0,168	0,109
8	AR_0198	7	0,081	0,309	0,069	0	0,032	0,001	0,329	0,18
9	AR_0298	1	0,403	0,09	0,318	0,057	0,019	0,001	0,066	0,044
10	AR_0398	8	0	0,038	0,084	0	0,105	0,13	0,019	0,623
11	AR_0498	4	0,012	0,066	0	0,894	0	0,015	0,014	0
12	AR_0598	2	0,141	0,374	0,139	0	0	0	0,09	0,257
13	AR_0698	8	0	0,025	0,097	0	0,108	0,125	0,015	0,63
14	AR_0798	3	0,239	0,067	0,394	0,02	0,261	0,019	0	0
15	AR_0197	1	0,427	0	0	0,133	0	0,274	0,165	0
16	AR_0196	3	0,267	0	0,585	0	0	0,148	0	0
17	AR_0195	1	0,683	0,032	0,02	0	0	0	0,264	0
18	AR_0194	1	0,993	0	0	0	0	0,007	0	0
19	AR_0190	1	1	0	0	0	0	0	0	0
20	AR_0290	1	0,969	0	0	0	0	0	0,031	0
21	AR_0189	1	1	0	0	0	0	0	0	0
22	AR_0188	1	0,674	0	0,326	0	0	0	0	0
23	AR_0187	3	0,276	0	0,37	0,059	0	0,295	0	0
24	AR_0100	5	0,056	0	0,131	0	0,404	0,13	0	0,279

10 - Save the selected partition for exploration with other T-LAB tools

When you exit the **Thematic Analysis of Elementary Contexts** function, the software displays messages to remind you that you can use other **T-LAB** tools to explore the clusters obtained.



If you select **Save**, the < **CONT_CLUST** > variable (clusters of elementary contexts) remains available only for certain types of analysis (e.g. Sequences of Themes, Word Associations, Comparison between Pairs of Key-Words, Co-Word Analysis and Concept Mapping) and until the user modifies his word list.

11 – Export a dictionary of categories

When this option is selected, T-LAB allows the user to create two files:
- a dictionary file with the **.dictio** extension which is ready to be imported by any **T-LAB** tool for thematic analysis. In such a dictionary each cluster is a category described by its characteristic words, i.e. by all words with a significant Chi-Square value within it;
- a **MyList.diz** file ready to be imported via the 'Customized Settings' tool. Since this file contains a list of all words with a significant chi-square value (i.e. all words that determine the differences between thematic clusters), its use may allow the user to repeat some analyses in a 'more selective' way.

12 – Validate the chosen partition and assess the semantic coherence of each theme

When clicking the 'Quality Indices' button (see the picture above), **T-LAB** creates a HTML file listing various measures.

The first ones, i.e. the measures of 'cluster quality', refer to the quality of the chosen partition.

The second ones, i.e. the measure of 'semantic quality', refer to the similarities between the top 10 words of each theme. More specifically:

- the top 10 words are those with the highest chi-square values over themes;
- the average similarity is computed using the cosine index;
- the cosine index of each word pairs, like the **Word Association** tool, is computed at the text segment (i.e. elementary context) level .

13 – Sequences of Themes

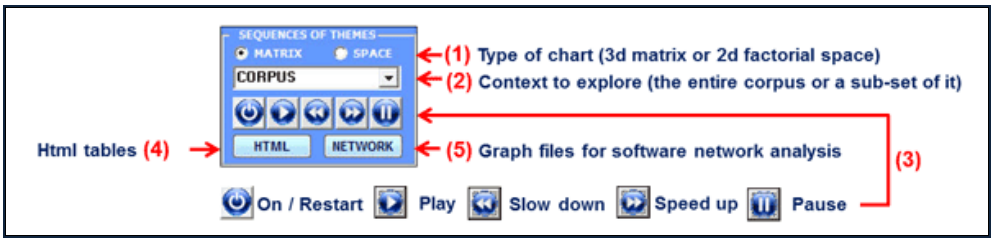
Unlike the 'Sequences of Themes' tool included in the co-occurrence analysis sub-menu, this one has been specifically designed to complement the thematic analysis of elementary contexts. More specifically its use makes sense only when the entire corpus can be considered like a **discourse** and/or its various sections (e.g. chapters of a book, parts of an interview, turns in a conversation or a debate, etc.) follow each other in a temporal order.

In fact this tool deals with the relationships between elementary contexts (up to 100,000) along the linear chain of the corpus, by considering each of them - either predecessor or successor - as an analysis unit belonging to a thematic cluster (or as unclassified). Accordingly, all available outputs allow the user to explore sequential relationships between 'themes', either by means of static charts and tables or by means of animated charts showing changes over time. This way the user can check either when people are engaged in specific themes (e.g. by looking at a diagonal of the matrix below) or when they shift from a dominant theme to another.

Step by step, here is a short description of how to proceed.

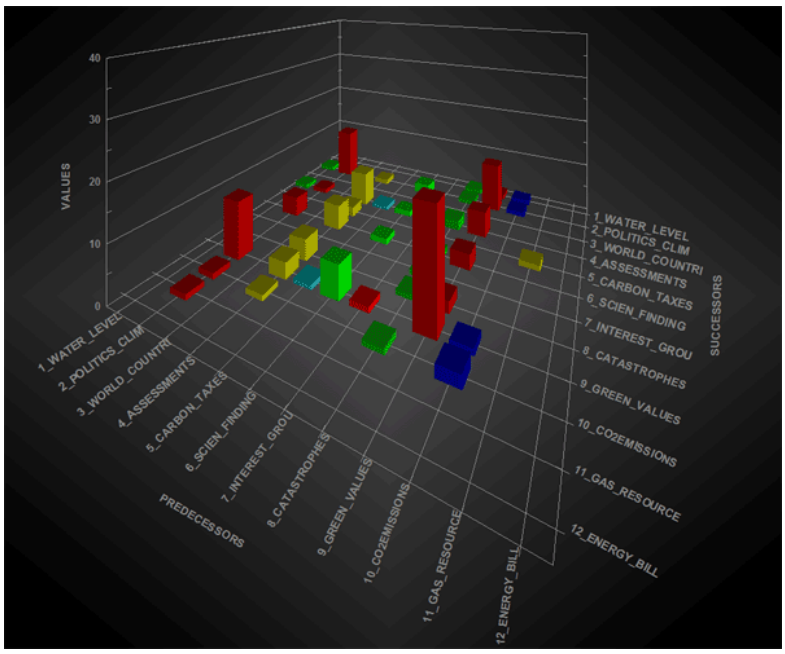
(N.B.: All the following outputs refer to a thematic analysis of the book 'The Politics of Climate
Change' by Antony Giddens published in the T-LAB website).

When the 'Sequence of themes' button is enabled, by clicking it the following 'player' becomes visible and active in the T-LAB working window.

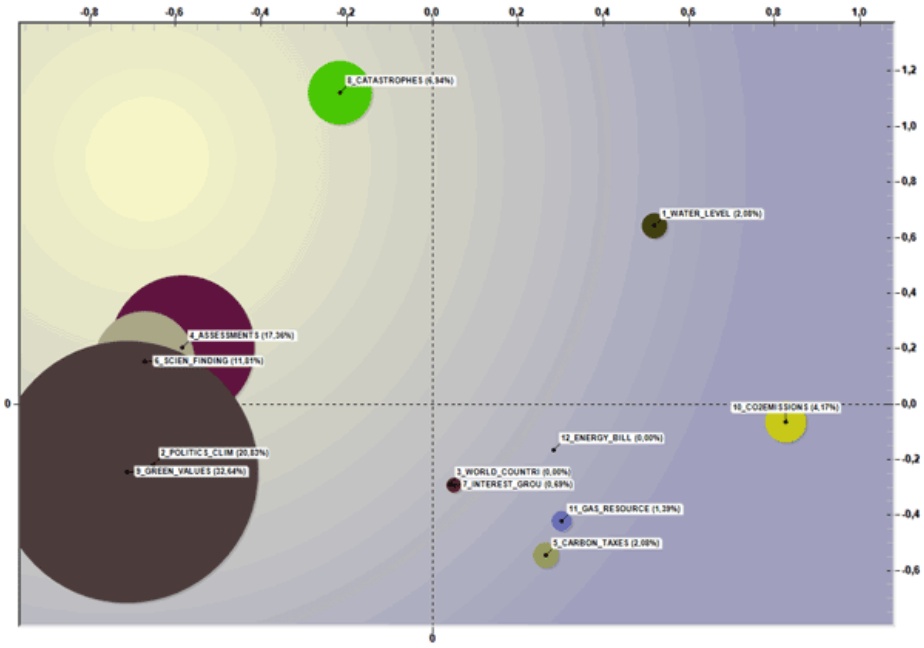


Option '1' (see matrix / space above) refers to the type of chart for visualizing theme sequences, either within the entire corpus or within a part of it (see option 2 above).

When checking 'matrix', a 3d chart is available which summarizes the relationships between predecessors and successors. In this case, while exploring 3d animated charts the bar dimensions are continuously readjusted to indicate how the occurrences of each sequence (i.e. two way relationship between a 'predecessor' and a 'successor') increases (see below).



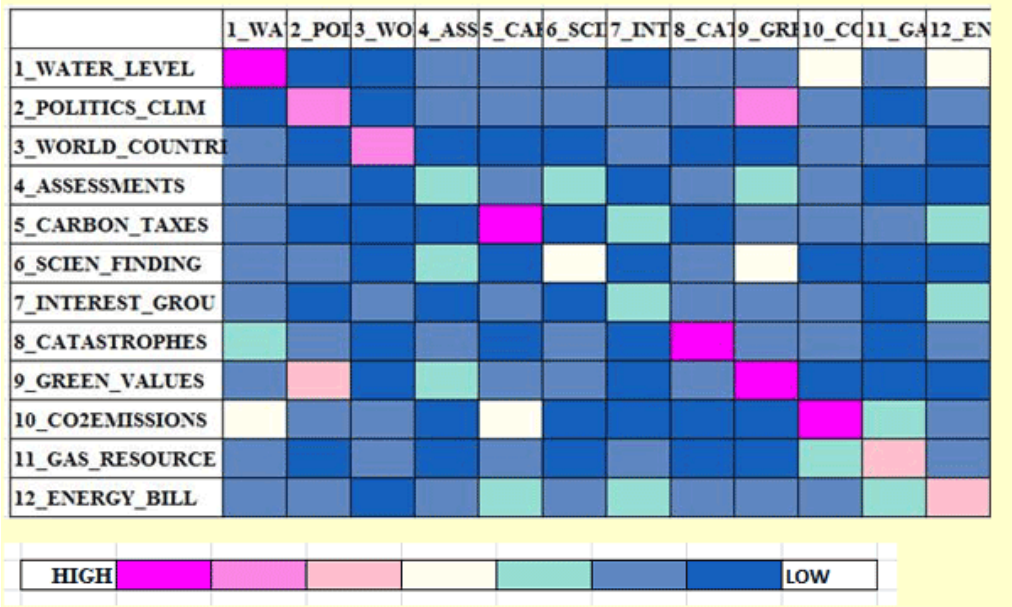
When checking the 'space', a 2d scatter chart is available which summarizes both the dimensions (i.e. percentages) and the relationships between thematic clusters. In this case, while exploring 2d animated charts the bubble dimensions – which are continuously readjusted to a total equal to 100% - indicate how the percentage of each thematic cluster changes over time; meanwhile the moving arrows indicate how themes follow each other (see below).



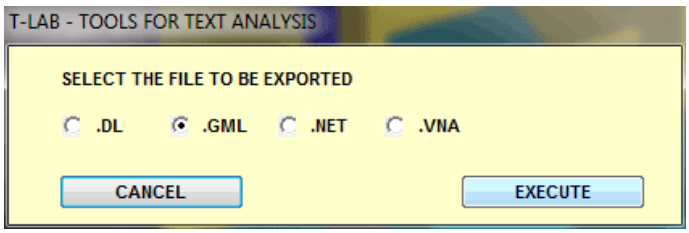
Moreover, at each step – after stopping the video (see the 'pause' button) - it is possible to obtain two further outputs:

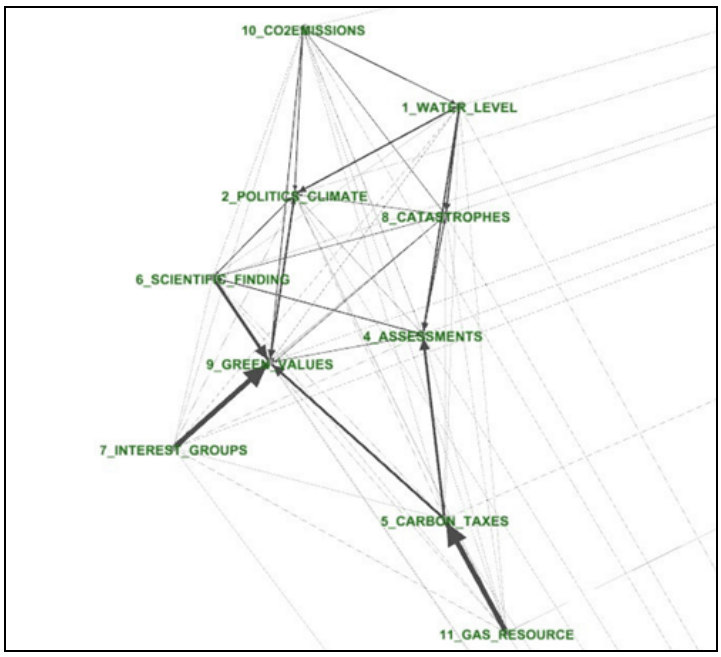
A – html tables which summarize the relationships between predecessors and successors (see below);

	1_WA	2_POI	3_WO	4_ASS	5_CAI	6_SCI	7_INT	8_CA	9_GRI	10_CC	11_GA	12_EN	TOT
1_WATER_LEVEL	41	4	4	8	5	6	3	9	6	15	8	18	127
2_POLITICS_CLIM	4	24	4	9	5	8	5	6	26	5	1	5	102
3_WORLD_COUNTRI	5	3	24	2	3	2	6	2	1	6	6	4	64
4_ASSESSMENTS	7	8	3	12	5	13	3	9	10	5	3	3	81
5_CARBON_TAXES	9	3	4	4	31	1	11	3	9	8	8	11	102
6_SCIEN_FINDING	5	9	2	11	1	17	1	9	16	2	0	2	75
7_INTEREST_GROU	8	2	6	1	6	0	10	5	6	5	3	10	62
8_CATASTROPHES	12	9	4	5	3	7	4	30	5	8	2	5	94
9_GREEN_VALUES	6	22	2	12	8	9	3	8	41	3	4	3	121
10_CO2EMISSIONS	18	7	6	2	15	1	2	3	3	48	13	9	127
11_GAS_RESOURCE	7	4	9	4	9	2	5	2	2	12	22	5	83
12_ENERGY_BILL	8	6	2	6	10	6	10	5	5	8	10	21	97



B – graph files which can be imported by software for network analysis.



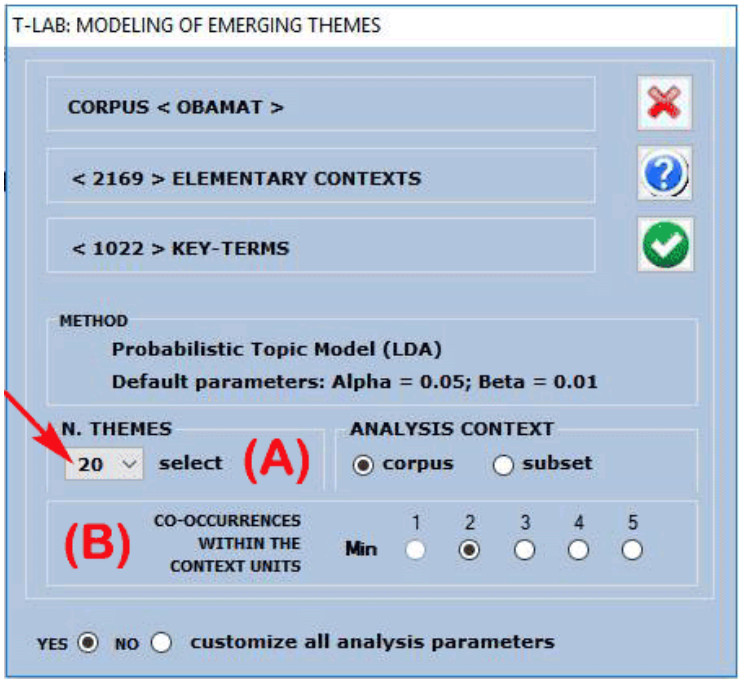


N.B.: The above graph, which refers to the third chapter of Giddens' book (i.e. 'The Greens and After') has been created by means of Gephi (see https://gephi.org/).

Modeling of Emerging Themes

This **T-LAB** tool provides a simple way of discovering, examining and modeling, the main **themes** or **topics** (henceforward 'theme' and 'topic' will be used synonymously) emerging from texts. Subsequently they can be explored further with several tools, either by keeping separate or by combining **qualitative** and **quantitative** approaches.

In fact, themes - which are described through their characteristic vocabulary and consist of **co-occurrence patterns** of key-terms - can be used as categories in further analyses or for automatically **classifying the context units** (i.e. documents or elementary contexts).



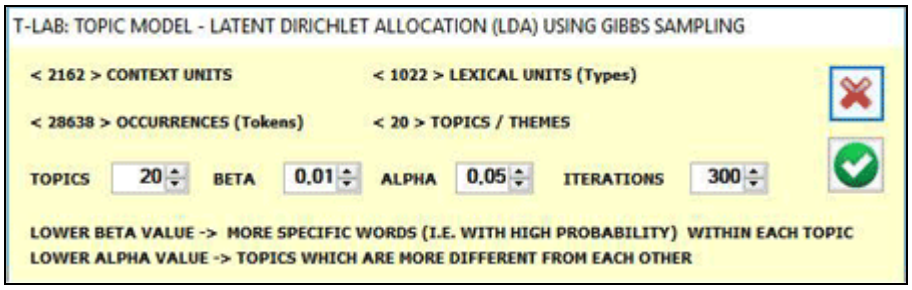
A **T-LAB** dialog box (see above) allows the user to set two analysis parameters.

In particular:

- the (A) parameter allows the user to set the amount (i.e. a fixed number) of themes to be obtained. (Note that the higher this number is the more consistent the co-occurrence patterns are; moreover, if necessary, some themes - e.g. those that are redundant or difficult to interpret - can be discarded later);

- the (B) parameter allows the user to exclude from the analysis any context unit that doesn't contain a minimum number of key-words included in the list which is being used.

Only when choosing to customize all analysis parameters (see the above 'Yes' option'), the following window will appears and more options will be available. (Note that in the below picture the number of context units is determined by the above parameter 'B').



The **analysis procedure** consists of the following steps:

a - construction of a document per word matrix, where documents are always elementary contexts corresponding to the context units (i.e. fragments, sentences, paragraphs) in which the corpus has been subdivided into;

b - data analysis by a probabilistic model which uses the Latent Dirichlet Allocation and the Gibbs Sampling (see the related information on Wikipedia:
http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation;
http://en.wikipedia.org/wiki/Gibbs_sampling;

c - description of themes by means of the probability of their characteristic words, either "specific" or "shared" by two or more themes.

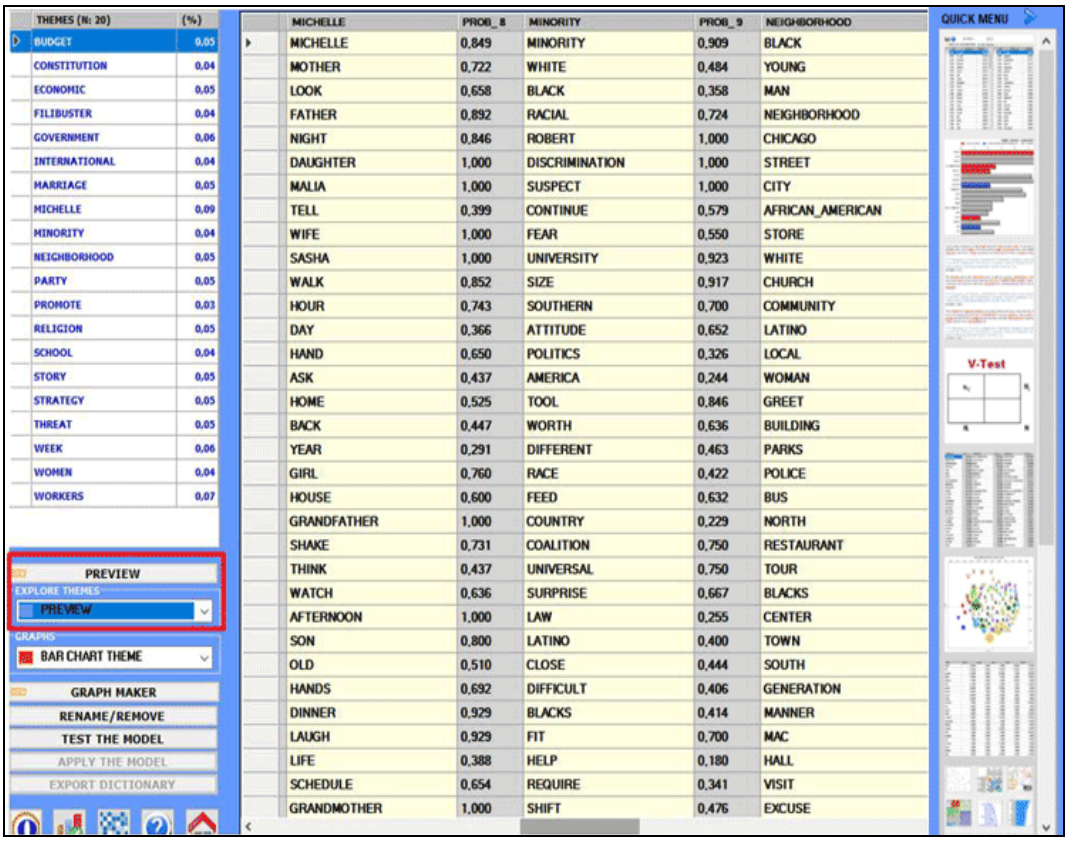
On completion of the analysis you can easily perform the following operations:

1 – explore the characteristics of each theme;
2 – explore the relationships between the various themes;
3 – rename or discard specific themes;
4 - assess the semantic coherence of each theme;
5 – test the model and assign context units (i.e. documents and/or elementary contexts) to themes;
6– apply the model by creating a new thematic variable, the values of which are the chosen topics;
7 – export a dictionary of categories, which can be used in further analyses.

In detail:

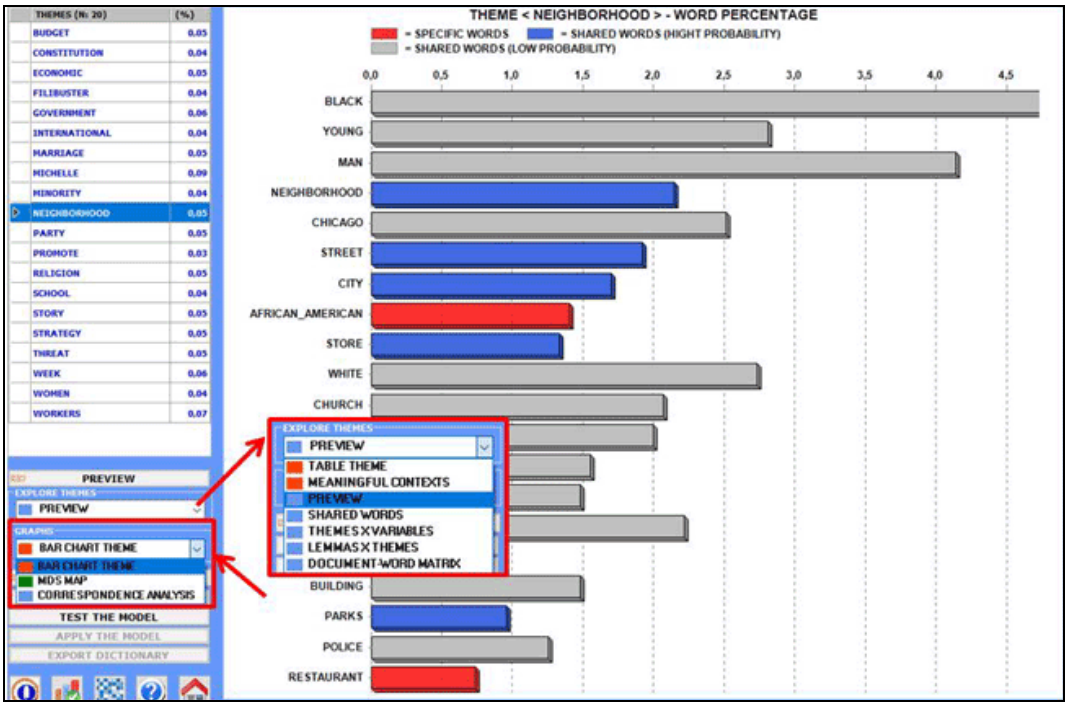
1 – Explore the characteristics of each theme

An overview of all themes is the first output which can be checked and saved, and it can be easily re-accessed by using the 'Preview' button (see below).



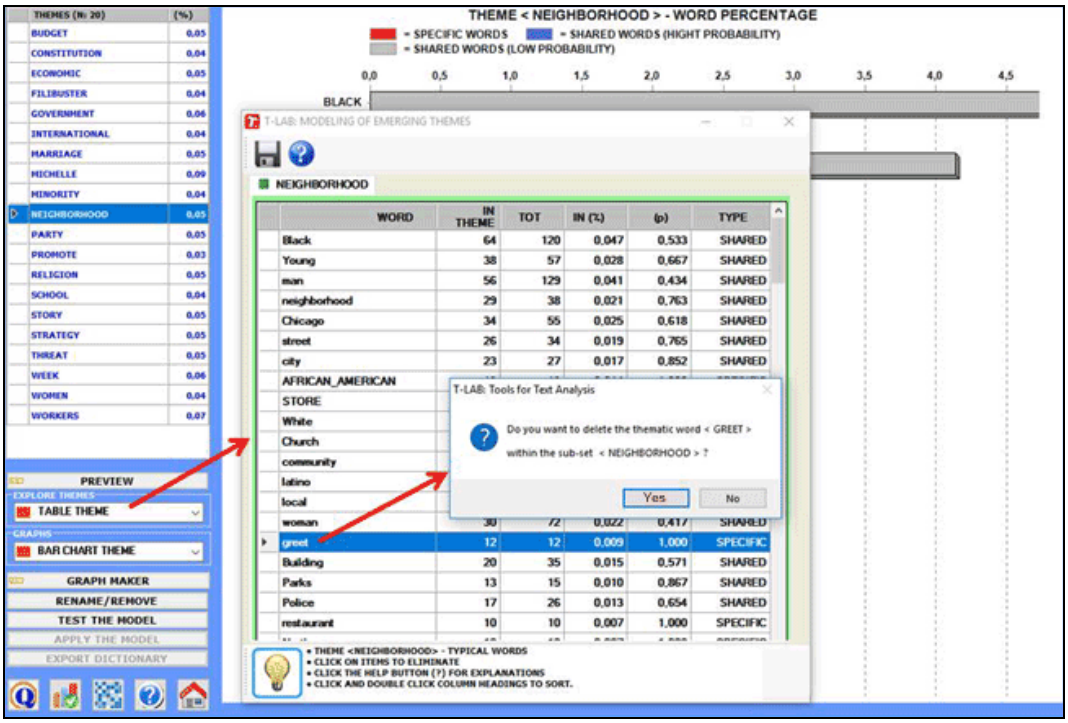
MICHELLE	PROB_8	MINORITY	PROB_9	NEIGHBORHOOD
MICHELLE	0,849	MINORITY	0,909	BLACK
MOTHER	0,722	WHITE	0,484	YOUNG
LOOK	0,658	BLACK	0,358	MAN
FATHER	0,892	RACIAL	0,724	NEIGHBORHOOD
NIGHT	0,846	ROBERT	1,000	CHICAGO
DAUGHTER	1,000	DISCRIMINATION	1,000	STREET
MALIA	1,000	SUSPECT	1,000	CITY
TELL	0,399	CONTINUE	0,579	AFRICAN_AMERICAN
WIFE	1,000	FEAR	0,550	STORE
SASHA	1,000	UNIVERSITY	0,923	WHITE
WALK	0,852	SIZE	0,917	CHURCH
HOUR	0,743	SOUTHERN	0,700	COMMUNITY
DAY	0,366	ATTITUDE	0,652	LATINO
HAND	0,650	POLITICS	0,326	LOCAL
ASK	0,437	AMERICA	0,244	WOMAN
HOME	0,525	TOOL	0,846	GREET
BACK	0,447	WORTH	0,636	BUILDING
YEAR	0,291	DIFFERENT	0,463	PARKS
GIRL	0,760	RACE	0,422	POLICE
HOUSE	0,600	FEED	0,632	BUS
GRANDFATHER	1,000	COUNTRY	0,229	NORTH
SHAKE	0,731	COALITION	0,750	RESTAURANT
THINK	0,437	UNIVERSAL	0,750	TOUR
WATCH	0,636	SURPRISE	0,667	BLACKS
AFTERNOON	1,000	LAW	0,255	CENTER
SON	0,800	LATINO	0,400	TOWN
OLD	0,510	CLOSE	0,444	SOUTH
HANDS	0,692	DIFFICULT	0,406	GENERATION
DINNER	0,929	BLACKS	0,414	MANNER
LAUGH	0,929	FIT	0,700	MAC
LIFE	0,388	HELP	0,180	HALL
SCHEDULE	0,654	REQUIRE	0,341	VISIT
GRANDMOTHER	1,000	SHIFT	0,476	EXCUSE

Other kinds of outputs are accessible by choosing the options highlighted in the below picture.



N.B.: In the above chart "high probability" indicates a probability >=.75.

When a topic is selected, by clicking the 'table theme' option, you can check its characteristics and - by clicking on any word in the table - a further option becomes available which allows you to "remove" the selected item (see the below picture).



The reading keys of the above table are as follows:

IN THEME = tokens of each word within the selected theme;

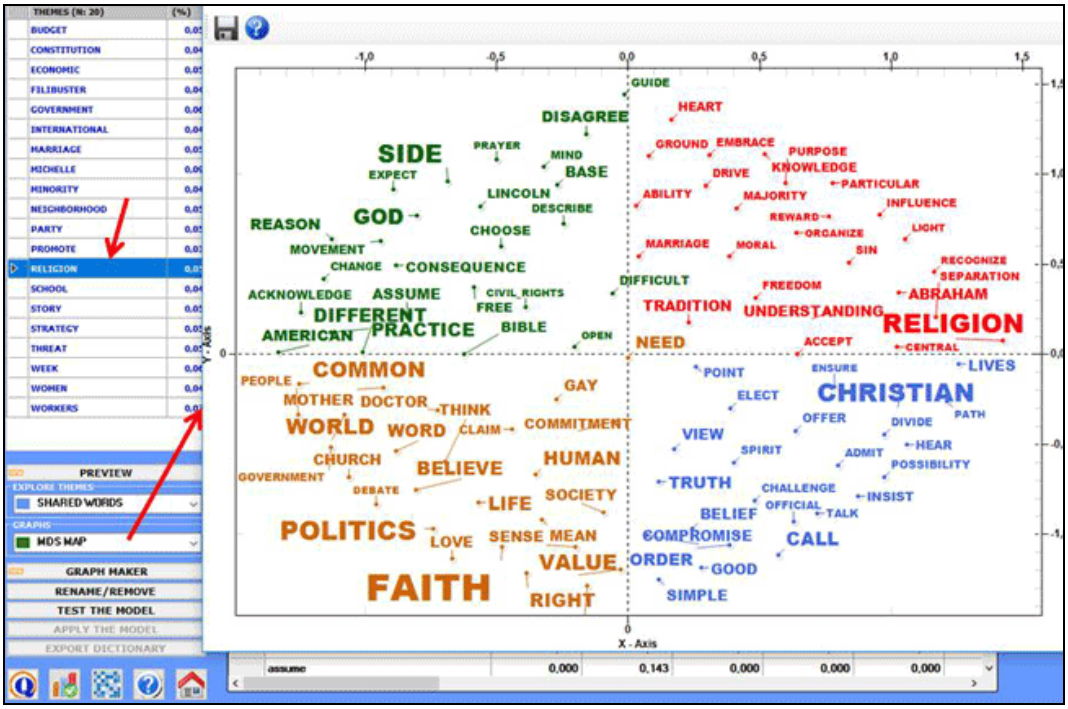
TOT = total tokens of each word within the corpus (or the subset) analysed;

IN (%) = percentage values of each word within the selected theme;

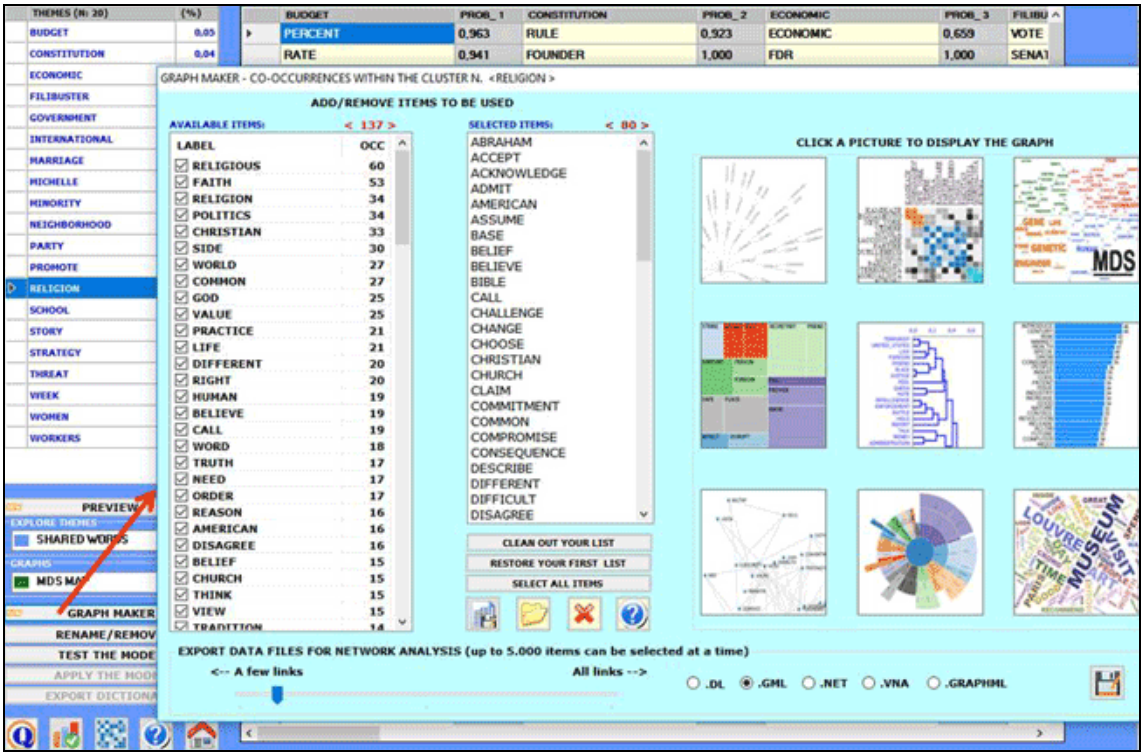
(p) = probability value of each word over themes;

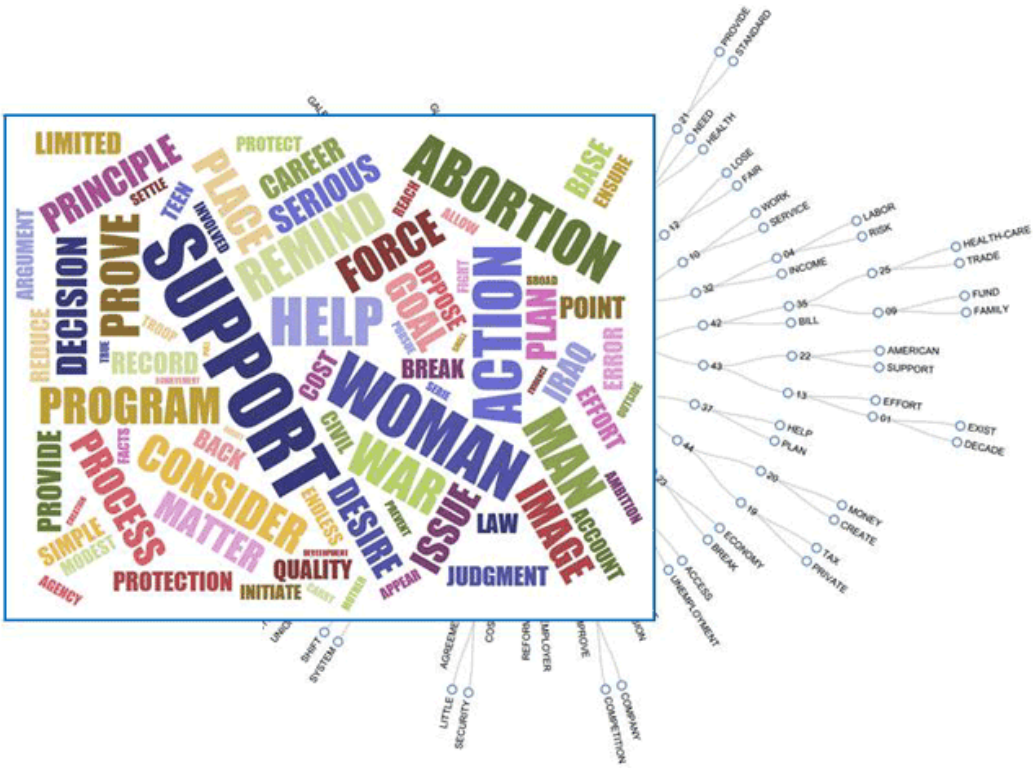
TYPE = "specific" when the word belongs to the selected theme only (i.e. p=1); "shared" in the other cases.

When a topic is selected, by clicking the 'MDS Map' option, the semantic relationships between its most characteristic words can be easily explored (see the below picture).

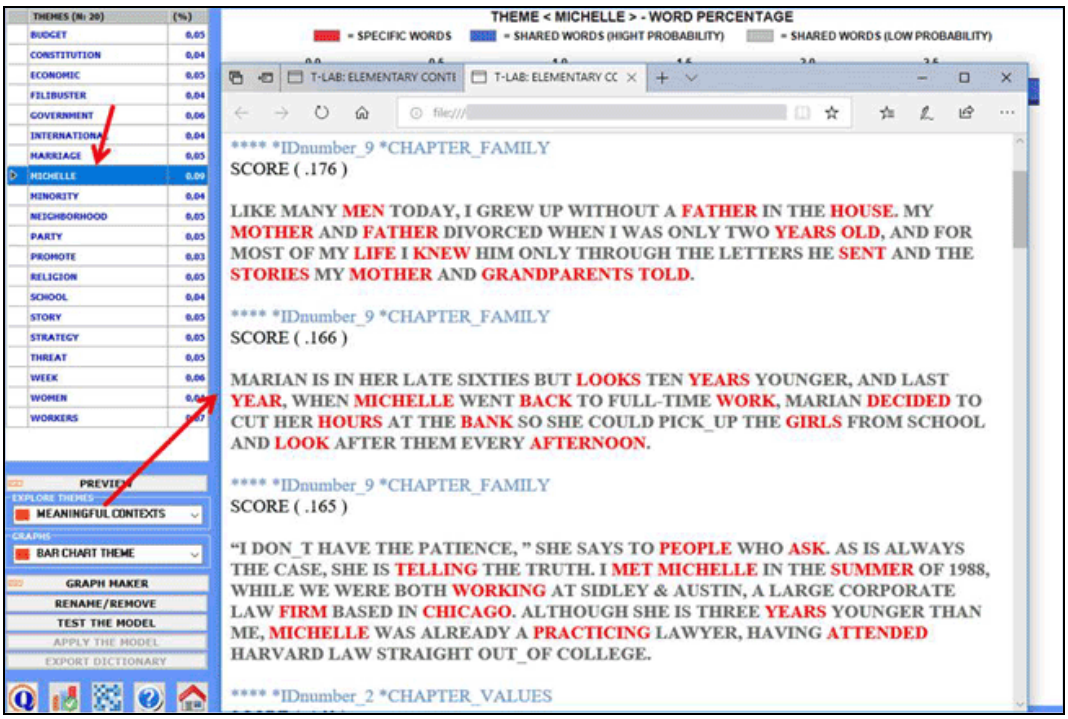


Moreover, by using the '**Graph Maker**' tool, more graphic options become available (see the below pictures).





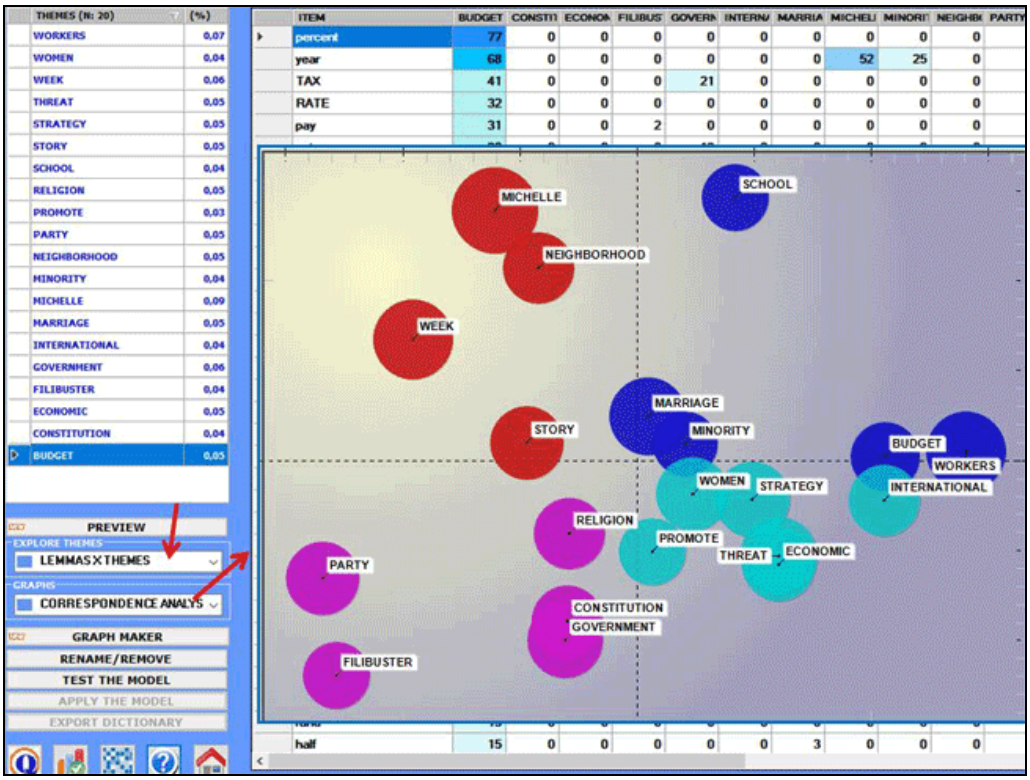
When a topic is selected, by clicking the 'meaningful contexts' option, a HTML file is created where the top 20 text segments – which most closely match the topic characteristics - are displayed (see the below picture).



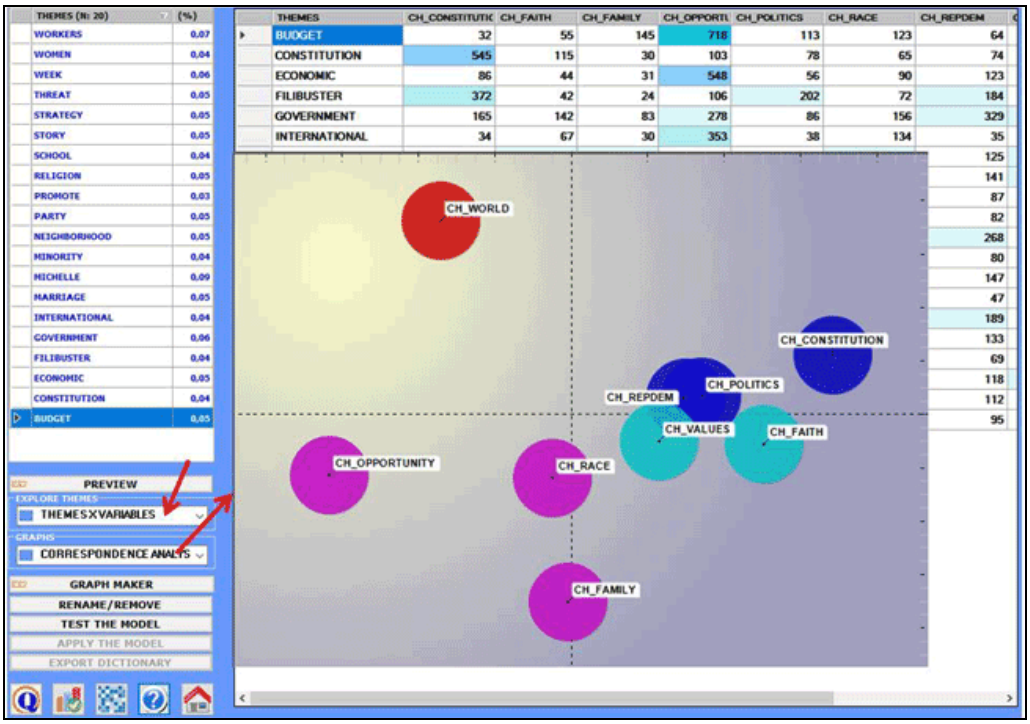
2 – Explore the relationships between the various themes

Two kinds of contingency tables can be created and explored through the Correspondence Analysis tool:

2.1) a word per topic table (see below)

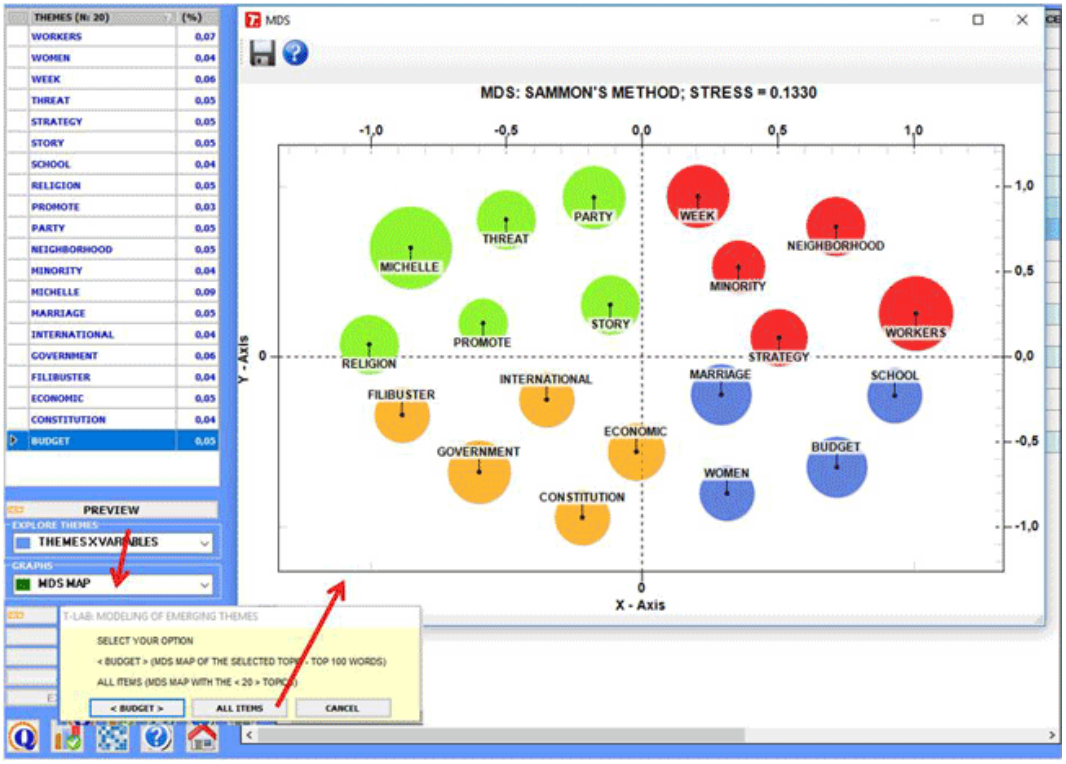


2.2) a topic per variable table (N.B.: In the below chart the nine bubbles correspond to the chapters of one of Obama's book)

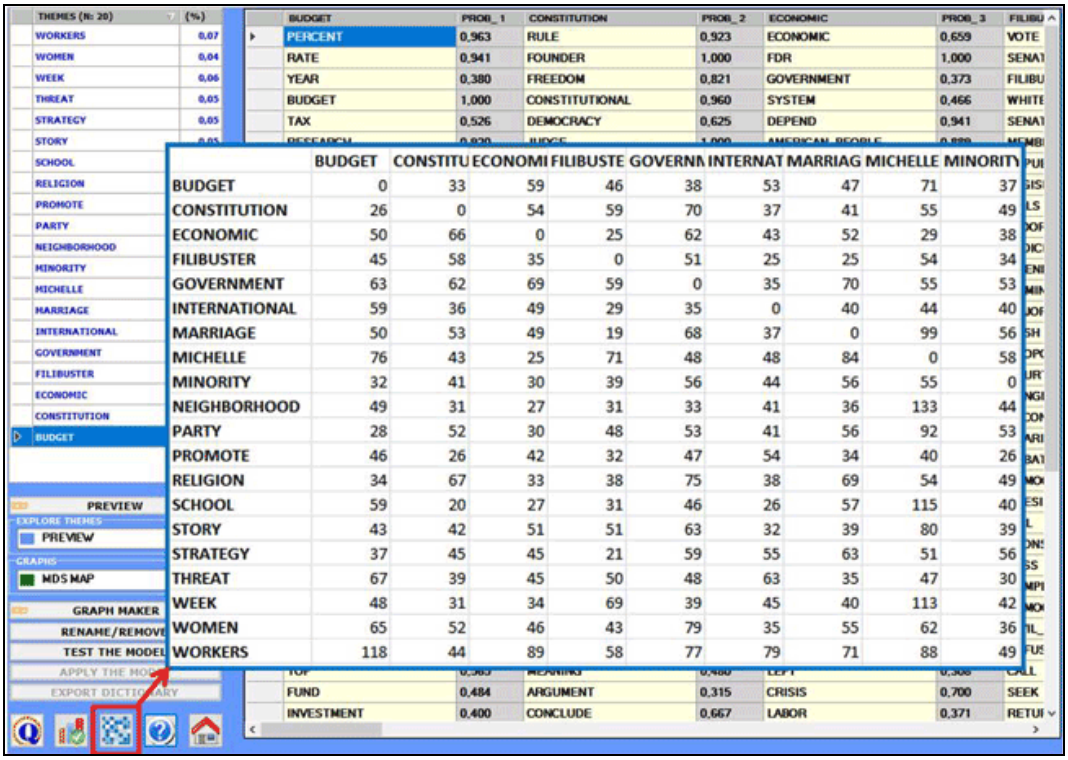


Two more graphic options are available which allow us to map the relationships between the various topics/themes:

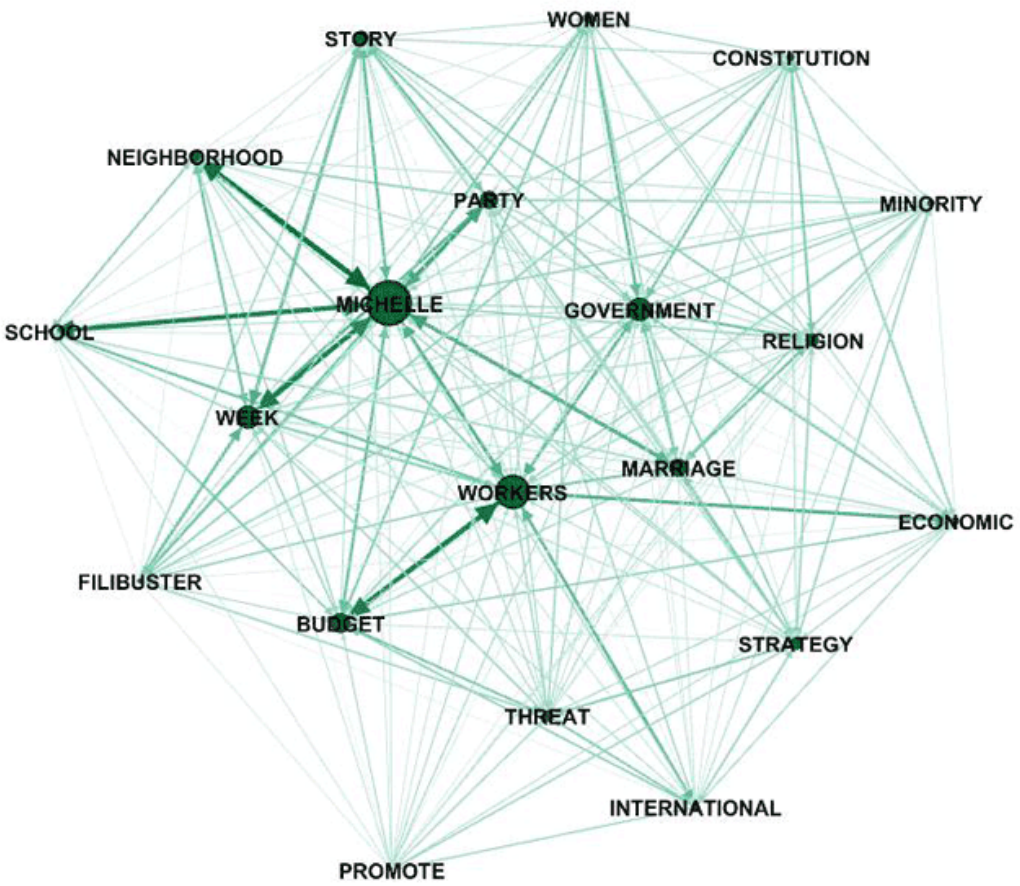
2.3) a **MDS Map**



2.4) a **Network Graph** obtained by exporting/importing the adjacency table created by T-LAB (see below)



	BUDGET	CONSTITU	ECONOMI	FILIBUSTE	GOVERNM	INTERNAT	MARRIAG	MICHELLE	MINORITY
BUDGET	0	33	59	46	38	53	47	71	37
CONSTITUTION	26	0	54	59	70	37	41	55	49
ECONOMIC	50	66	0	25	62	43	52	29	38
FILIBUSTER	45	58	35	0	51	25	25	54	34
GOVERNMENT	63	62	69	59	0	35	70	55	53
INTERNATIONAL	59	36	49	29	35	0	40	44	40
MARRIAGE	50	53	49	19	68	37	0	99	56
MICHELLE	76	43	25	71	48	48	84	0	58
MINORITY	32	41	30	39	56	44	56	55	0
NEIGHBORHOOD	49	31	27	31	33	41	36	133	44
PARTY	28	52	30	48	53	41	56	92	53
PROMOTE	46	26	42	32	47	54	34	40	26
RELIGION	34	67	33	38	75	38	69	54	49
SCHOOL	59	20	27	31	46	26	57	115	40
STORY	43	42	51	51	63	32	39	80	39
STRATEGY	37	45	45	21	59	55	63	51	56
THREAT	67	39	45	50	48	63	35	47	30
WEEK	48	31	34	69	39	45	40	113	42
WOMEN	65	52	46	43	79	35	55	62	36
WORKERS	118	44	89	58	77	79	71	88	49

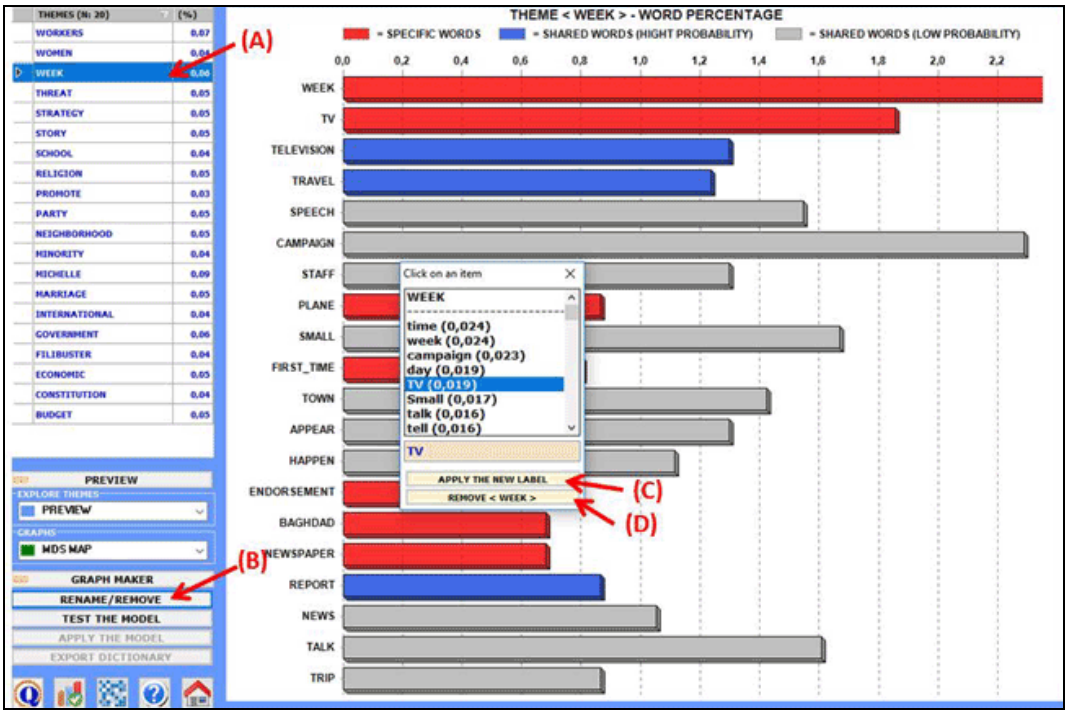


N.B.: The above graph has been created by means of Gephi (https://gephi.org/), which is an open source software, after importing a table created by **T-LAB**.

3 - **Rename or discard specific themes**

In order to rename or discard specific themes, just select one of them (see "A" below) and click on the "rename/remove" button (see "B" below).

When the appropriate box appears, depending on your goals, you can change the label by choosing among the available words or by typing a new label in the appropriate field (see "C" below); otherwise you can discard the selected theme just by clicking on the corresponding button (see "D" below)



4 – Assess the semantic coherence of each theme

When clicking the 'Quality Indices' button, **T-LAB** computes the average similarity between the top 10 words of each theme. More specifically:

- the top 10 words are those with the highest probability values over themes;
- the average similarity is computed using the cosine index;
- the cosine index of each word pairs, like the **Word Association** tool, is computed at the text segment (i.e. elementary context) level .

As a result, T-LAB creates a HTML table where the 'k' themes are listed according to their 'semantic coherence' (i.e. the first theme in the list is the one with the highest average similarity index).

N.B.: Because the above measures vary according to the selected words, the user is advised to repeat the procedure each time any of the top 10 words of each theme is removed.

5 – **Test the Model**

At the end of the analysis procedure (see above the "a" and "b" points of the analysis procedure) each context unit (i.e. primary documents or elementary contexts) is represented as mixture of different topics; differently the classification process used in this step assigns each context unit to the topic which is the most characteristic of it.

For this reason, when the "Test the Model" option is selected, **T-LAB** creates a HTML file and two XLS two files including the classification of contexts units (see below).

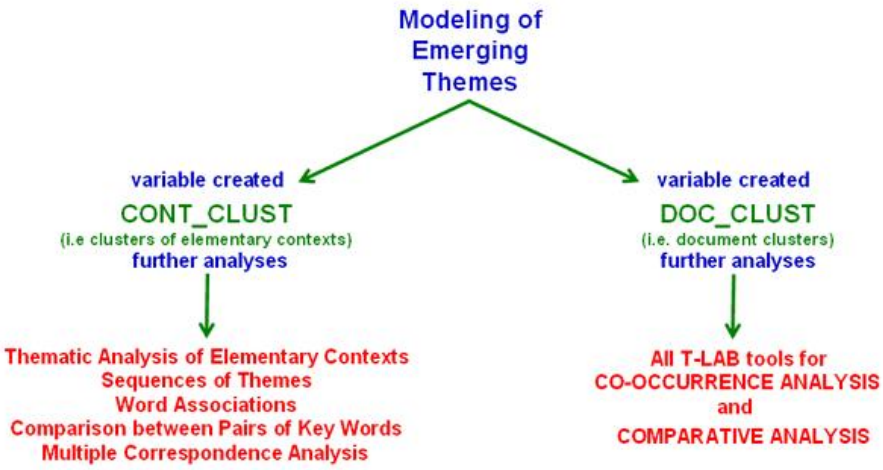
ID_DOC	BEST	BUDGET	CONSTITUTIO	ECONOMIC	FILIBUSTER	GOVERNME	INTERNATIO	MARRIAGE	MICHELLE	MINORITY	NEIGHBORHO	PARTY
1	5	0.842	0.834	1.573	2.715	4.229	0.391	1.144	1.136	1.279	1.083	3.243
2	7	0.727	0.945	1.162	1.459	3.118	0.694	3.664	1.630	0.914	1.129	1.384
3	2	0.455	7.560	0.887	6.329	2.543	0.451	1.124	1.489	0.960	0.715	1.892
4	11	1.783	1.031	0.566	3.642	0.851	0.492	0.843	1.603	1.371	0.855	4.996
5	20	10.599	1.248	7.583	1.569	3.347	4.337	1.946	1.930	1.560	1.275	0.942
6	13	0.684	1.586	0.560	0.509	1.637	1.067	2.062	2.007	0.933	1.829	2.076
7	10	1.763	0.836	1.226	1.010	1.749	2.041	2.081	2.556	8.744	8.868	2.261
8	17	2.441	2.331	3.105	2.148	1.543	8.905	1.354	1.718	1.542	1.809	1.331
9	8	2.369	0.340	0.361	0.840	1.110	0.350	2.536	8.624	0.454	1.283	0.467

In the above table, each document has a probability value associated with each topic.

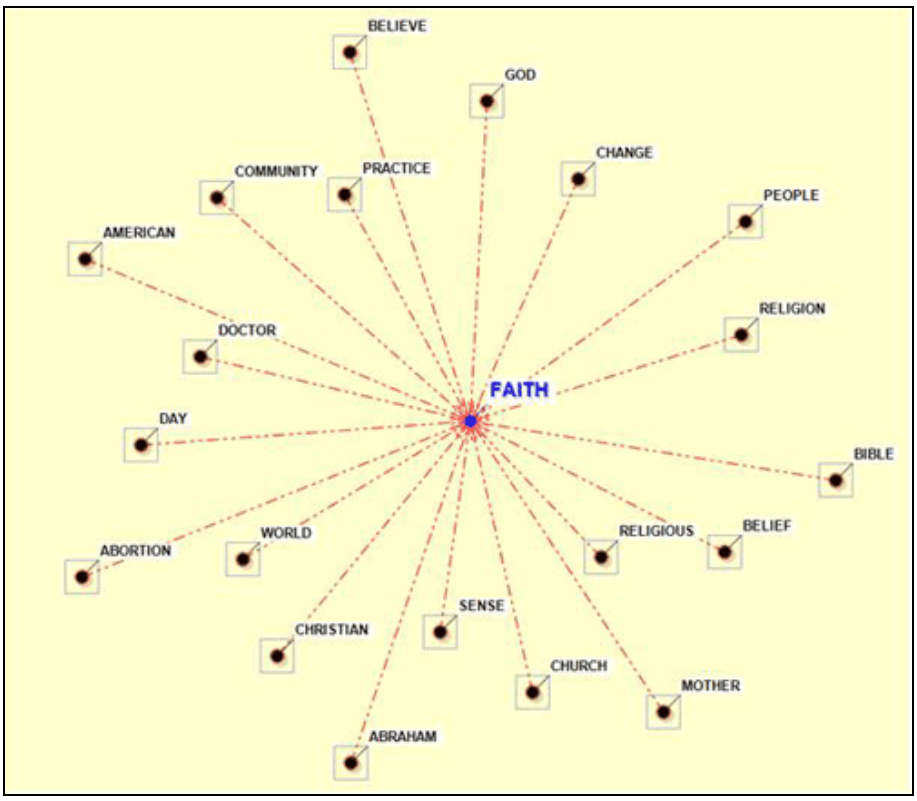
IdDoc	IdSeg	Topic	Score	Segm
5	1122	BUDGET	0.502	INSTEAD , THE BULK OF THE DEBT IS A DIRECT RESULT OF THE PRESIDENT 'S TAX CUTS , 47 . 4 PERCENT OF WHICH WEN
5	934	SCHOOL	0.459	WHERE A CHILD IN LOS_ANGELES HAS TO COMPETE NOT JUST WITH A CHILD IN BOSTON BUT ALSO WITH MILLIONS OF
7	1464	SCHOOL	0.439	PARENTS WHO COACHED LITTLE LEAGUE GAMES AND BAKED BIRTHDAY CAKES AND BADGERED TEACHERS TO MAKE SU
5	948	SCHOOL	0.342	RECENT STUDIES SHOW THAT THE SINGLE MOST IMPORTANT FACTOR IN DETERMINING A STUDENT 'S ACHIEVEMENT IS
9	2100	SCHOOL	0.324	IT_S ALSO TIME TO REDESIGN OUR SCHOOLS—NOT JUST FOR THE SAKE OF WORKING PARENTS , BUT ALSO TO HELP PR
5	1132	BUDGET	0.319	"THOUGH I_VE NEVER USED TAX SHELTERS OR HAD A TAX PLANNER , AFTER INCLUDING THE PAYROLL TAXES WE EACH
9	2034	SCHOOL	0.317	WE_RE AL EVEN THE MOST FAIR-MINDED OF WHITES , THOSE WHO WOULD GENUINELY LIKE TO SEE RACIAL INEQUALI
9	2040	SCHOOL	0.316	CHILDREN IN SINGLE-PARENT HOMES ARE ALSO MORE LIKELY TO DROP_OUT OF SCHOOL AND BECOME TEEN PARENTS
3	481	SCHOOL	0.314	SHOULD WE LET TEACHERS LEAD OUR CHILDREN IN PRAYER AND LEAVE OPEN THE POSSIBILITY THAT THE MINORITY FA
5	1104	WORKERS	0.313	IF NECESS AT SOME LEVEL I_M JUST GOING_THROUGH THE SAME CONFLICTING EMOTIONS THAT OTHER FATHERS EXP
8	1640	INTERNATIONAL	0.312	WITH MORE_THAN 240 MILLION PEOPLE , INDONESIA 'S POPULATION RANKS FOURTH IN THE WORLD , BEHIND CHINA
3	433	FILIBUSTER	0.302	DECADE AFTER DECADE , COURTLY , ERUDITE MEN LIKE SENATOR RICHARD B . RUSSELL OF GEORGIA (AFTER WHOM T
9	2039	SCHOOL	0.300	WHATEVER THE EFFECT ON ADULTS , THOUGH , THESE TRENDS HAVEN_T BEEN SO GOOD FOR OUR CHILDREN . MANY
7	1614	WORKERS	0.299	TO ADDRESS THIS PROBLEM , I SUCCEEDED IN INCLUDING LANGUAGE REQUIRING THAT ANY JOB FIRST BE OFFERED TO
9	2057	SCHOOL	0.298	SEVENTY PERCENT OF FAMILIES WITH CHILDREN ARE HEADED BY TWO WORKING PARENTS OR A SINGLE WORKING PAR
8	1784	FILIBUSTER	0.287	WITH AN EYE ON THE MIDTERM ELECTIONS , REPUBLICANS STEPPED_UP THE ATTACKS AND PUSHED FOR_A VOTE AUT
4	756	FILIBUSTER	0.286	"MAKE THESE CHANGES , " THE SENATOR TOLD ROVE , "AND NOT ONLY WILL I VOTE FOR THE BILL , BUT I GUARANTEE
6	1330	SCHOOL	0.285	IT IS DOUBTFUL THAT CHILDREN RECITING THE PLEDGE OF ALLEGIANCE FEEL OPPRESSED AS A CONSEQUENCE OF MUTT
7	1390	MINORITY	0.280	"THERE IS NOT A BLACK AMERICA AND WHITE AMERICA AND LATINO AMERICA AND ASIAN AMERICA—THERE 'S THE U
8	1933	INTERNATIONAL	0.277	IF OVERALL THE INTERNATIONAL SYSTEM HAS PRODUCED GREAT PROSPERITY IN THE WORLD 'S MOST DEVELOPED CO
9	2036	SCHOOL	0.275	BETWEEN 1960 AND 1995 , THE NUMBER OF AFRICAN_AMERICAN CHILDREN LIVING WITH TWO MARRIED PARENTS DR
1	52	FILIBUSTER	0.275	SENATOR BARBARA BOXER OF CALIFORNIA AGREED TO SIGN THE CHALLENGE , AND WHEN WE RETURNED TO THE SEN
8	1952	INTERNATIONAL	0.266	INDEED , COUNTRIES THAT HAVE SUCCESSFULLY DEVELOPED UNDER THE CURRENT INTERNATIONAL SYSTEM HAVE AT
3	465	GOVERNMENT	0.261	THE RIGHT TO WORSHIP HOW AND IF WE WISH
5	946	SCHOOL	0.261	AND IN FACT WE ALREADY HAVE HARD EVIDENCE OF REFORMS THAT WORK : A MORE CHALLENGING AND RIGOROUS
2	276	MARRIAGE	0.260	WE VALUE THE IMPERATIVES OF FAMILY AND THE CROSS-GENERATIONAL OBLIGATIONS THAT FAMILY IMPLIES . WE VA
5	956	SCHOOL	0.260	WORKING WE HEARD ABOUT SONS AND DAUGHTERS ON THEIR WAY TO IRAQ AND THE NEED TO TEAR_DOWN AN OLD
8	1657	SCHOOL	0.258	WITHOUT THE MONEY TO GO TO THE INTERNATIONAL SCHOOL THAT MOST EXPATRIATE CHILDREN ATTENDED , I WENT
5	957	SCHOOL	0.255	AND WE CAN MAKE SURE THAT NONPERFORMING TEACHERS NO_LONGER HANDICAP CHILDREN WHO WANT TO LEAR
5	1042	FILIBUSTER	0.254	THE PRESI SHE WAS STUDYING GOVERNMENT IN SCHOOL , SHE SAID , AND WOULD SHOW IT TO HER CLASS . I ASKED
5	1072	WORKERS	0.249	THE EARNED INCOME TAX CREDIT , A PROGRAM CHAMPIONED BY RONALD REAGAN THAT PROVIDES LOW-WAGE WOR

6 - Apply the model

After having applied and saved the, given that after exiting from the analysis themes are recorded as clusters of context units (i.e. like the **Thematic Analysis of Elementary Contexts** and **Thematic Document Classification** results), the new thematic variables just created (i.e. **CONT_CLUST** and/or **DOC_CLUST**) can be explored by using various **T-LAB** tools (see below).



For example, by using the **Word Associations** tool and by selecting the sub-set (i.e. topic) 'Religion' the following graph can be created.



7 – Export a dictionary of categories

When this option is selected a dictionary file with the .dictio extension is created which is ready to be imported by any **T-LAB** tool for thematic analysis. In such a dictionary each theme (or category) is described by its characteristic words.

Thematic Document Classification

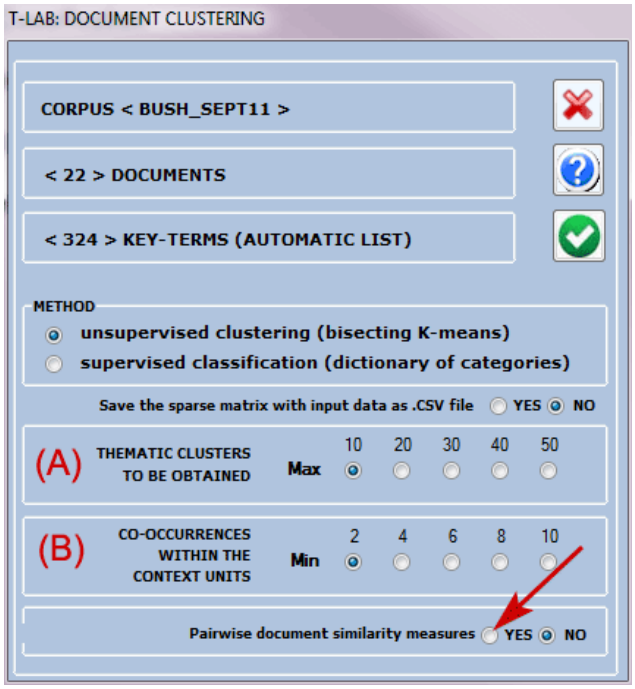
This function is only enabled when the corpus under analysis includes from 20 (min) to 99,999 (max) primary documents.

The analysis process can be performed through an unsupervised clustering (i.e bottom-up approach), which is the default option, or a supervised classification (i.e. top-down approach). When choosing the latter (i.e. supervised classification), a dictionary of categories must be imported, either created by means of a previous T-LAB analysis or made up by the user.

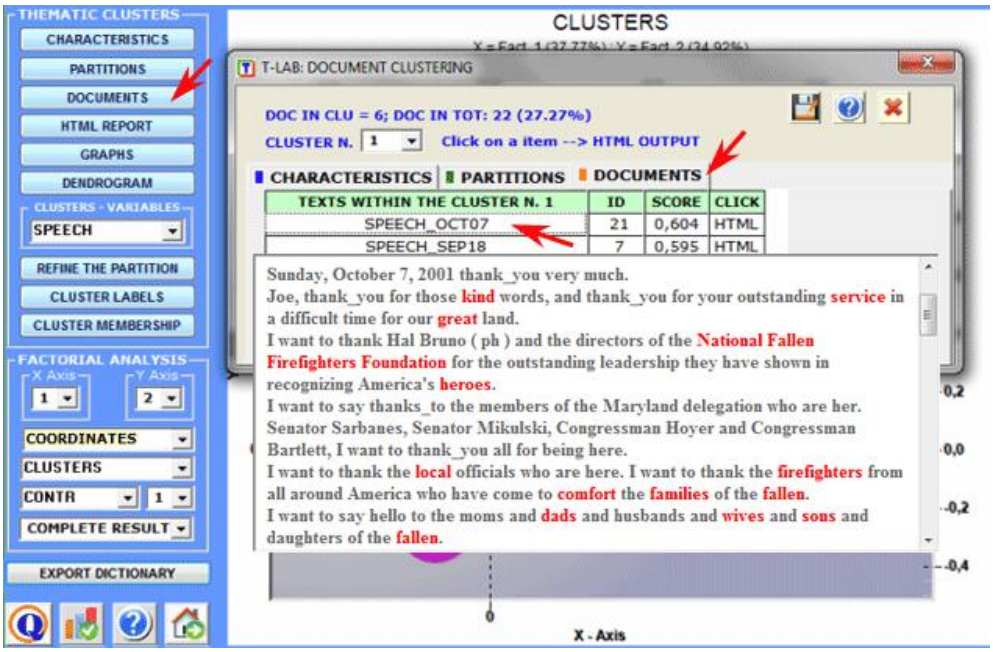
You can use this function to construct document clusters and explore their characteristics by means of operations (including algorithms) similar to those described in the section of the manual dedicated to **Thematic Analysis of Elementary Contexts.**

The specificity of this function is that the table analysed consists of one line for each document in the corpus, each of which is represented as a vector of values indicating the occurrences of the words found in it.

N.B.: When doing an unsupervised clustering and the number of analysed documents doesn't exceed 3,000, it is possible to obtain similarity measures (i.e. cosine) between each pair of them (see below). However only the similarities with a cosine coefficient greater or equal to 0.05 are recorded.



Accordingly, the following outputs are different:



The documents belonging to each cluster are ordered by their decreasing relevance value (see above) and can be browsed in HTML format.

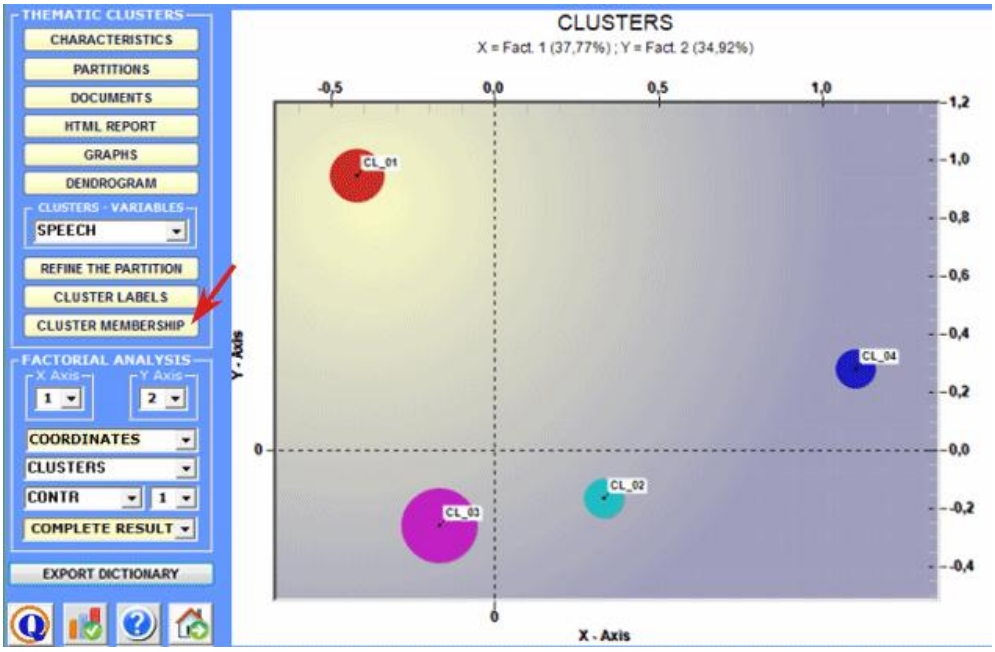
In this case the relevance value (score) assigned to each document (i) in the cluster (k) is obtained by applying the following formula:

$$score_{i,k} = \cos(d_i, c_k)$$

Where:

i – refers to document ***i***;
k – refers to cluster ***k***;
cos – is the cosine symbol;
$\mathbf{d}_i$ – is the normalized vector of $TF_{j,i}IDF_j$, where **j** refers to word in document **i**;
$\mathbf{c}_k$ – is the normalized vector of $TF_{j,k}IDF_j$, where **j** refers to word in cluster **k**;

By using the scores obtained by the above formula, transformed into percentage values, the file "Document_Membership_Degree.xls" (see below) - containing the clusters to which the documents are assigned, either by the bisecting K-Means (mutual exclusive memberships) or the TF-IDF (mixed or fuzzy memberships) – is made available by **T-LAB**.

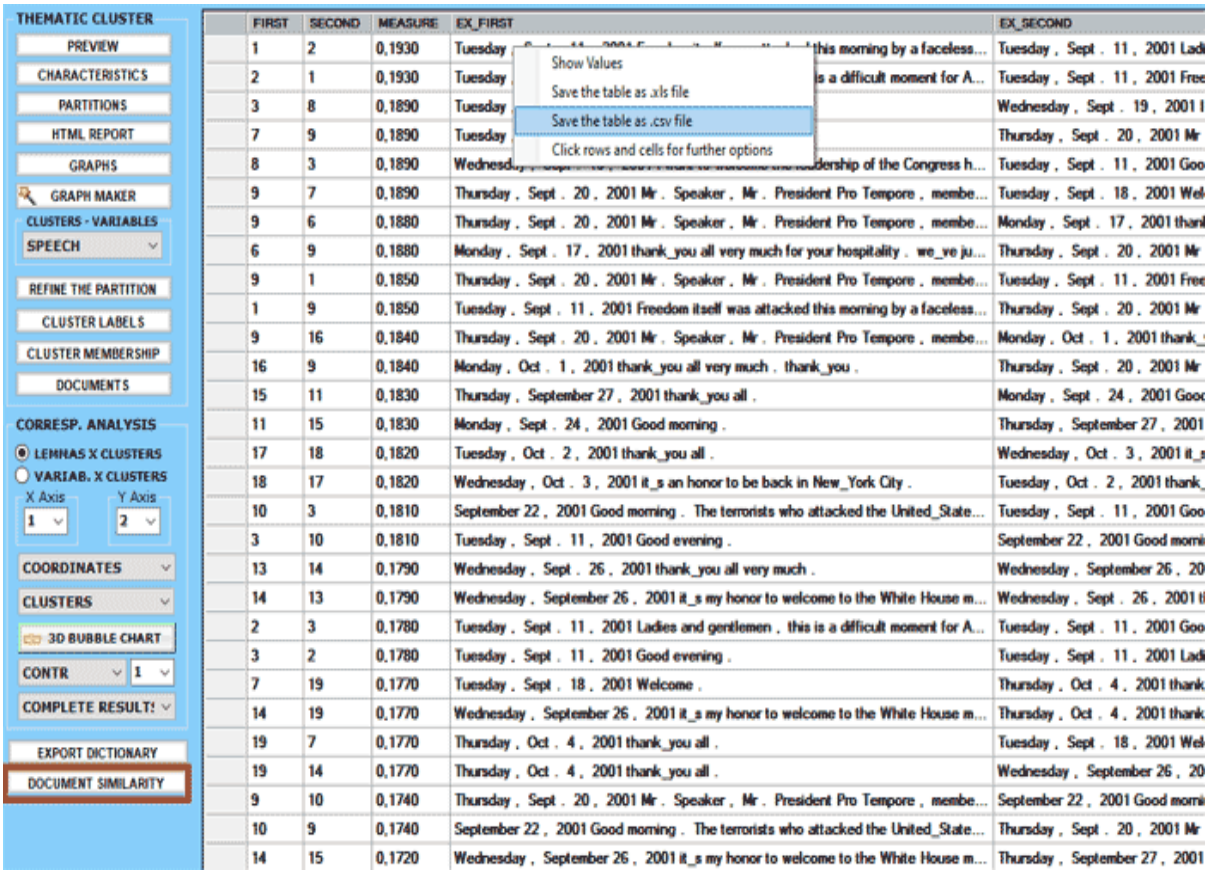


DOC_ID	VAR_01	CLUST_K-	BEST_TF-	MATCHIN(	CLUST-1	CLUST-2	CLUST-3	CLUST-4
1	SP_SEP1	1	1	1	0,456	0,13	0,215	0,199
2	SP_SEP1	1	1	1	0,451	0,098	0,216	0,235
3	SP_SEP1	1	1	1	0,431	0,174	0,23	0,165
4	SP_SEP1	1	1	1	0,601	0,116	0,217	0,066
5	SP_SEP1	3	1	0	0,387	0,155	0,333	0,125
6	SP_SEP1	2	2	1	0,142	0,6	0,141	0,118
7	SP_SEP1	1	1	1	0,544	0,122	0,182	0,152
8	SP_SEP1	2	2	1	0,115	0,49	0,194	0,201
9	SP_SEP2	3	3	1	0,259	0,227	0,372	0,143
10	SP_SEP2	4	4	1	0,129	0,141	0,162	0,568
11	SP_SEP2	3	3	1	0,145	0,198	0,454	0,204
12	SP_SEP2	3	3	1	0,089	0,204	0,522	0,184
13	SP_SEP2	3	3	1	0,159	0,196	0,403	0,242
14	SP_SEP2	2	2	1	0,091	0,59	0,182	0,136
15	SP_SEP2	3	3	1	0,165	0,177	0,388	0,27
16	SP_OCT0	3	3	1	0,21	0,203	0,479	0,108
17	SP_OCT0	4	4	1	0,135	0,121	0,172	0,571
18	SP_OCT0	4	4	1	0,089	0,158	0,146	0,608
19	SP_OCT0	3	3	1	0,203	0,211	0,42	0,166
20	SP_OCT0	3	3	1	0,15	0,151	0,598	0,101
21	SP_OCT0	1	1	1	0,691	0,086	0,132	0,091
22	SP_OCT0	3	3	1	0,223	0,168	0,488	0,121

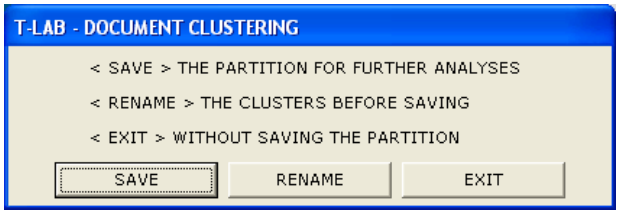
When the 'Document Similarity' button is enabled, by clicking it is possible to check how each document is similar to the others.

As in other cases, the similarity measure is the cosine coefficient and this can vary according to how many features (i.e. words) have been used for the thematic classification.

Below is a short description of how this tool works.



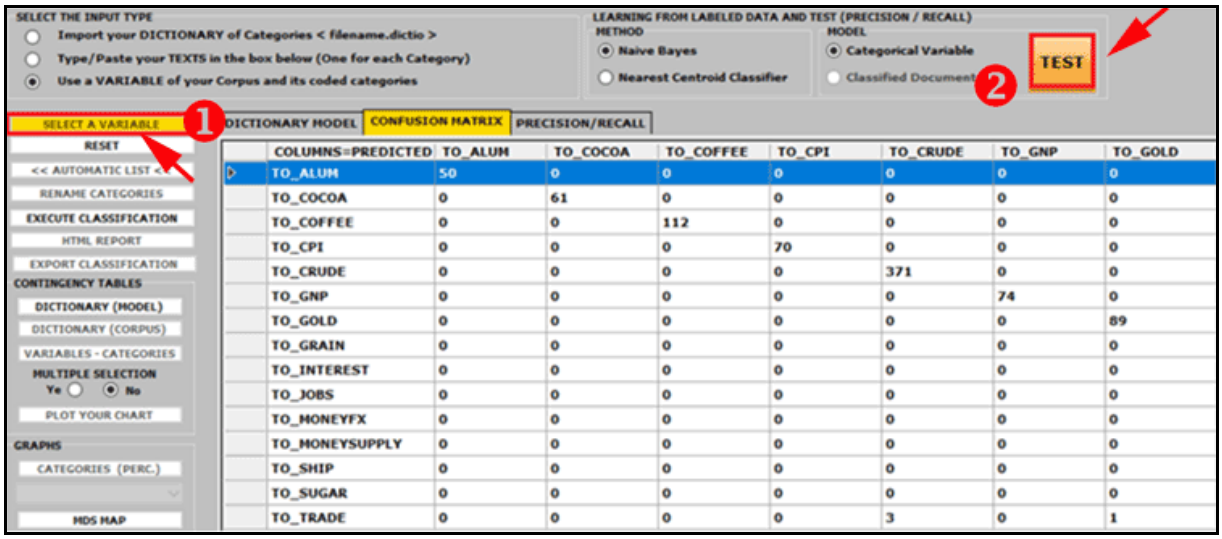
When you exit this function, the software displays messages to remind you that you can use other **T-LAB** tools to explore the clusters obtained.



If you select "Save", the < **DOC_CLUST**> variable (document cluster) remains available for all subsequent analyses of the corpus performed with other **T-LAB** tools.

Dictionary-Based Classification

N.B.: The pictures shown in this section have been obtained by using a previous version of **T-LAB**. These pictures look slightly different in **T-LAB 10**. In particular, starting from the 2021 version, a new feature allows one to easily test any model on labeled data (e.g. data which includes themes obtained from a previous qualitative analysis) and obtain outputs like confusion matrices and precision/recall metrics (see picture below).



COLUMNS=PREDICTED	TO_ALUM	TO_COCOA	TO_COFFEE	TO_CPI	TO_CRUDE	TO_GNP	TO_GOLD
TO_ALUM	50	0	0	0	0	0	0
TO_COCOA	0	61	0	0	0	0	0
TO_COFFEE	0	0	112	0	0	0	0
TO_CPI	0	0	0	70	0	0	0
TO_CRUDE	0	0	0	0	371	0	0
TO_GNP	0	0	0	0	0	74	0
TO_GOLD	0	0	0	0	0	0	89
TO_GRAIN	0	0	0	0	0	0	0
TO_INTEREST	0	0	0	0	0	0	0
TO_JOBS	0	0	0	0	0	0	0
TO_MONEYFX	0	0	0	0	0	0	0
TO_MONEYSUPPLY	0	0	0	0	0	0	0
TO_SHIP	0	0	0	0	0	0	0
TO_SUGAR	0	0	0	0	0	0	0
TO_TRADE	0	0	0	0	3	0	1

This **T-LAB** tool allows the user to perform an **automated classification** of **lexical units** (i.e. words and lemmas, including multi-word phrases) or **context units** (i.e. sentences, paragraphs and short documents) present in a text collection (i.e. a corpus) according to a set of categories constructed by the researcher.

Depending on the type of categories used, such a classification may be considered a classical **content analysis** or a lexicon-based **sentiment analysis**. Moreover, as the analysis process can create new variables and category dictionaries which can be exported and re-imported in further analysis projects, such a tool allows the user to explore the same corpus from varied perspectives as well as analyse two or more text collections by using the same models.

Here are some examples of **possible uses** of this tool:

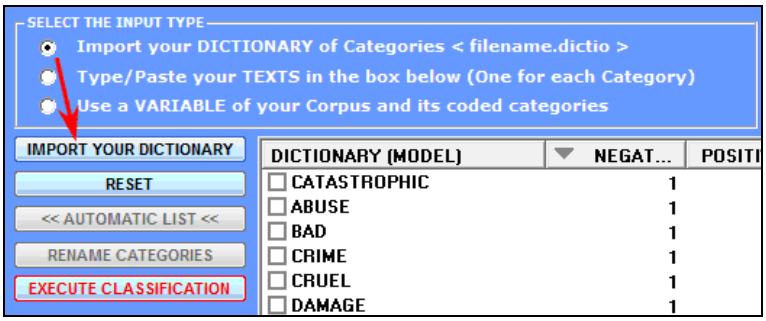
- automated coding of open-ended surveys;
- top-down analysis of political speeches;
- sentiment analysis of customer comments;
- verification of psychotherapy processes;
- validation of methods for qualitative content analysis.

Below is a short description of the four main phases of the analysis process, which are, however, independent one from the other. In fact, the researcher can also use this tool just for customizing his dictionaries or for exploring his data set.

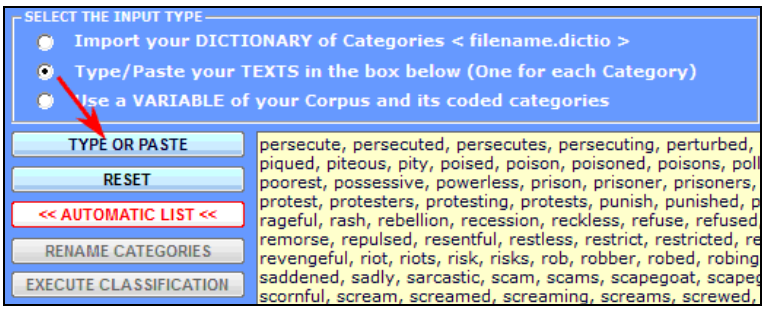
A) - PREPROCESSING PHASE

The starting points and the corresponding **input types** of the pre-processing phase can be three:

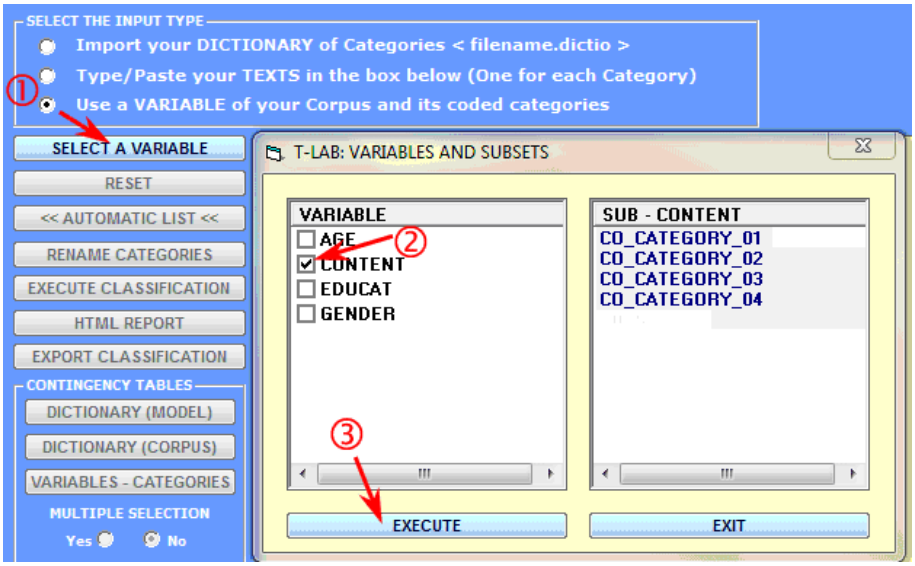
1 - a ready-made **dictionary** in the appropriate format is already available (see all related information in the 'E' section of this document). In this case just click the '**Import your Dictionary**' button (see below);



2 - a dictionary has to be distilled from **sample texts** or **word lists** made up by the user. In this case just type or copy/paste the sample texts into the appropriate box (one sample for each category, one after the other, max 100,000 characters each);

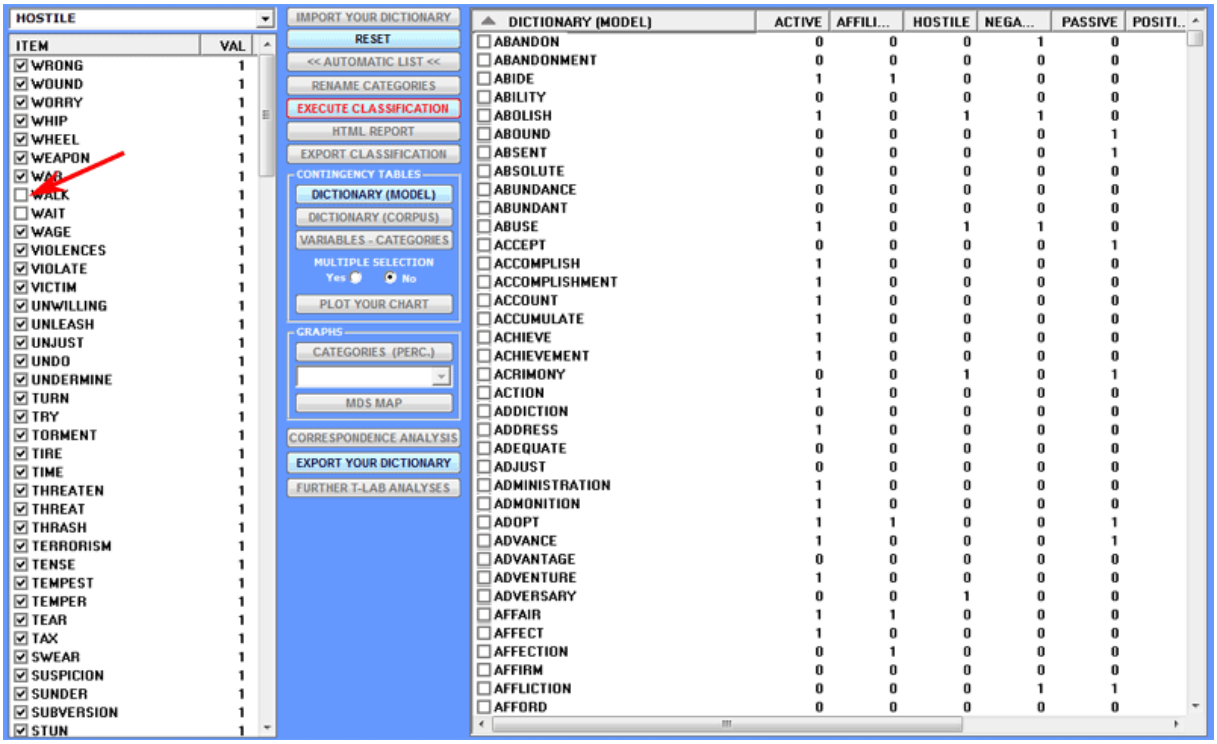


3 - a dictionary has to be distilled from the **categories** of a variable resulting from a previous content analysis. In this case just click the '**Select a Variable**' button and make the appropriate choices (see below).

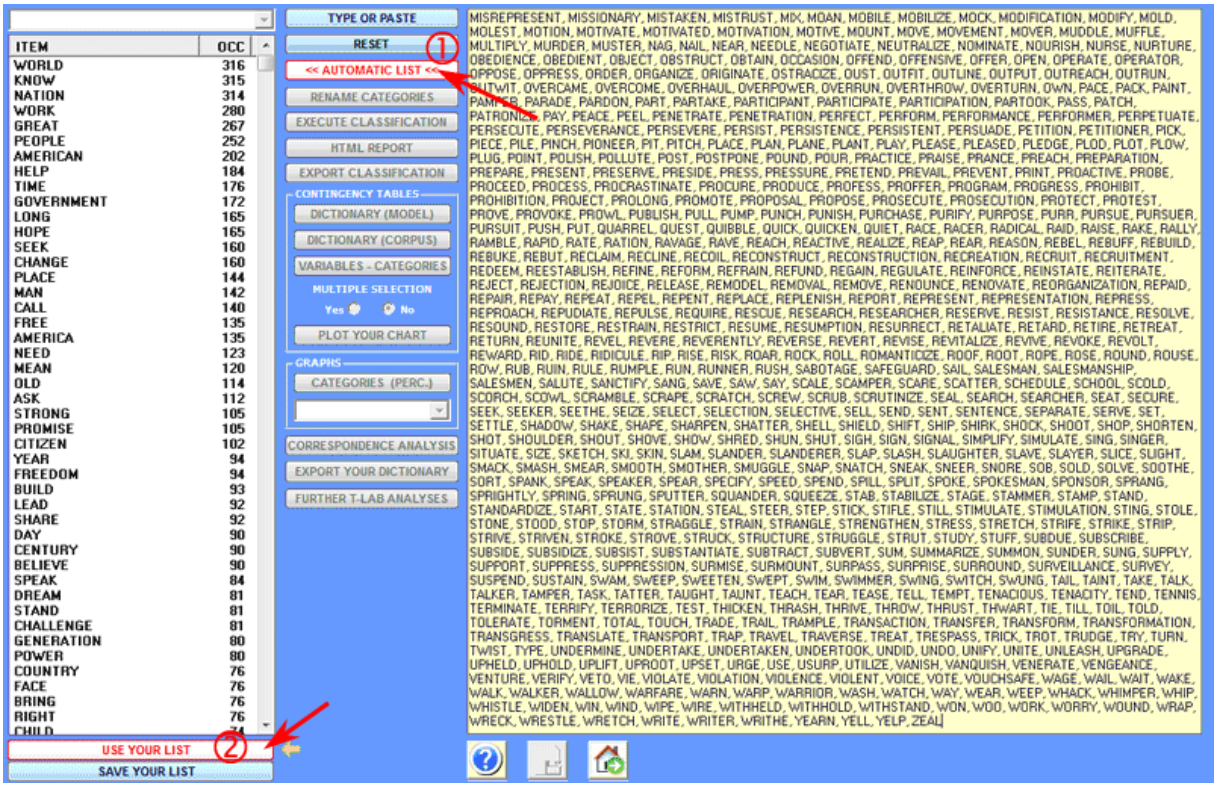


According to the above three cases, before performing the classification of selected textual units, **T-LAB** works in the following ways:

1 - the ready-made dictionary is transformed into a contingency table. Subsequently the user can explore such a table in various ways (see the 'C' section of this document); moreover, by selecting each category, he can remove one or more of the corresponding items (see the below picture).

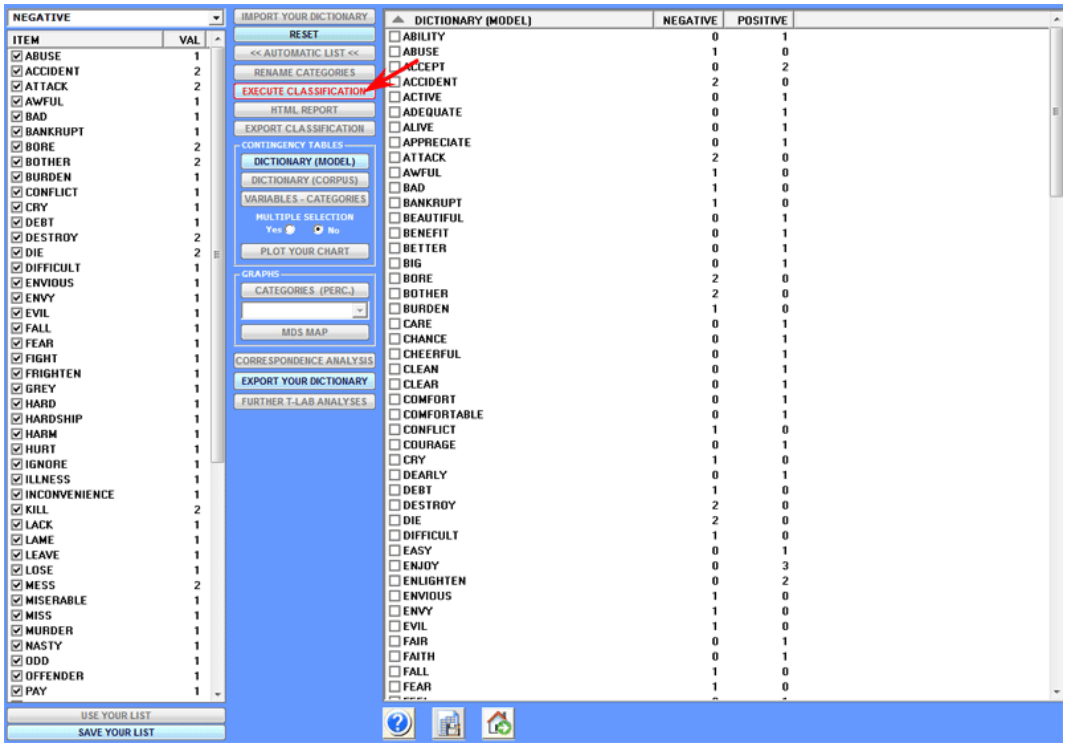


2 - when sample texts are inserted in the appropriate box, after clicking the '**Automatic List**' button (see '1' below), **T-LAB** performs a specific kind of lemmatization which uses only the vocabulary of the selected corpus (see the list of available items on the left of image below), then it transforms each text into a word list the items of which can be selected and deselected. Subsequently, in order to validate each word list (i.e. each category of your dictionary), just click the '**Use Your List**' button (see '2' below). All the above operations must be repeated for each category of the dictionary, then the user can perform all operations described in the 'C' section of this document.

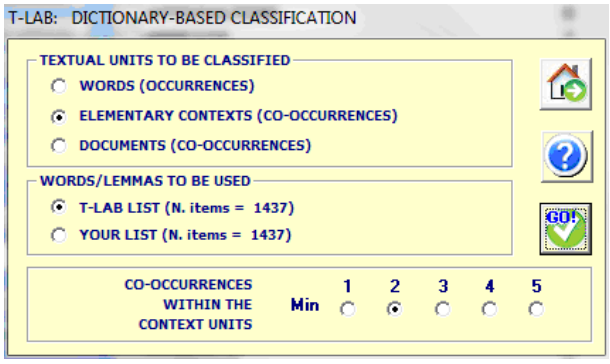


3 - when a variable resulting from a previous content analysis is selected, **T-LAB** makes available the corresponding term-category contingency table and the user can perform all operations of '**Data Exploration**' (see the 'C' section of this document).

B) - CLASSIFICATION PROCESS



After clicking the '**Execute Classification**' button (see the above picture), depending on the type of corpus under analysis, the user can make the following choices:



At this stage, if the user decides to **classify words**, no further choices are available; in fact, in such a case, the occurrences of each word (i.e. the word tokens) are simply counted as occurrences of the corresponding category. For example, if a category of our dictionary is 'religion' and this includes words like 'faith' and 'prayer', when analysing a document which contains the above two words, **T-LAB** simply groups their occurrences (e.g. 2 occurrences of 'faith' and 3 occurrences of 'prayer' = 5 occurrences of 'religion').

Differently, when the user decides to **classify context units** (i.e. 'elementary contexts' like sentences and paragraphs or 'documents'), **T-LAB** considers both the dictionary categories and the context units to be classified as co-occurrence profiles (i.e. term vectors) and then calculates their similarity measures. To this purpose, the co-occurrence profiles can be filtered by a 'T-LAB list' (i.e. all key-words which occurrence values are greater than or equal to the minimum threshold of 4) or a by 'user's list' (e.g. all key-words resulting from a customized settings) which, however, can sometimes be equal. Moreover in such cases **T-LAB** allows the user to exclude from the analysis any context unit that doesn't contain a minimum number of key-words included in the list which is being used (see above the 'co-occurrences within context units' parameter).

In detail, when classifying context units, **T-LAB** performs the following steps:

a) normalization of the 'seed vectors' corresponding to the 'k' categories (i.e. column profiles) of dictionary used;
b) normalizations of term vectors corresponding to the context units analysed;
c) computation of the Cosine similarity and Euclidean distance between each 'i' context unit and each 'k' seed vector of the dictionary used;
d) assignment of each 'i' context unit to the 'k' class or category for which the corresponding seed is the closest. (N.B.: In all cases there must be a correspondence between the maximum Cosine similarity and the minimum Euclidean distance of each 'context unit'/'category' pair, otherwise T-LAB considers the 'i' context unit as 'unclassified').

In other words, in the above case **T-LAB** uses a sort of K-means clustering where the 'k' centroids have a priori patterns and they are not updated during the analysis process.
Being a top-down classification, in such a case the quality of the analysis results will depend on two main factors:
1 - the 'relevance' of the dictionary used for the analysed corpus;
2 - the 'discriminant' capacity of the categories used.
In fact, if the optimum of the above characteristics is reached, both the 'precision' and the 'recall' parameters (see http://en.wikipedia.org/wiki/Precision_and_recall) have values between 80% and 95%.
Please note that, at the moment, **T-LAB** doesn't take into account negations. So, when doing sentiment analysis, a sentence like 'Do not hate your enemy' can be classified as 'negative'. Advanced users can manage this issue during the corpus importation (see the use of stop-word and multiword lists). For example, the word phrase 'do not hate' can be transformed into 'do_not_hate' and consequently included in the 'positive' category'.

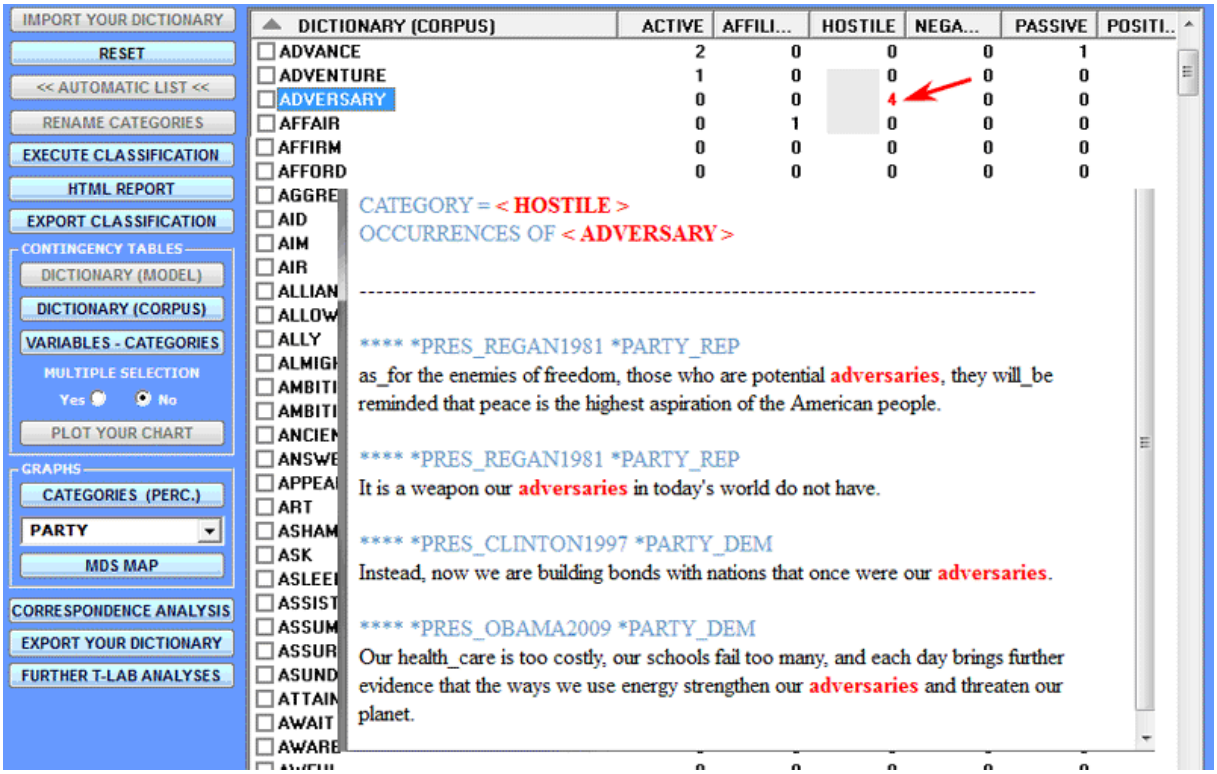
C) - DATA EXPLORATION

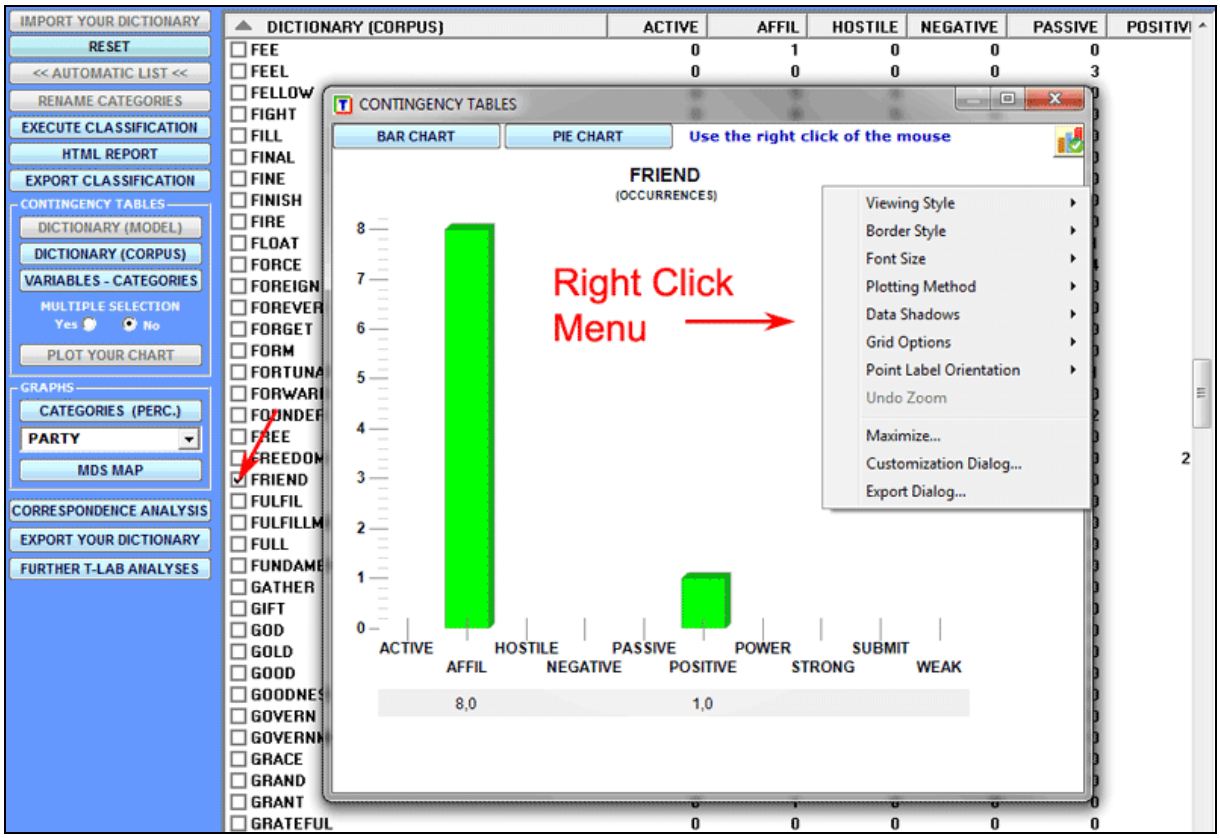
Any data exploration activity uses **contingency tables** where both the dictionaries and the results of the classification process can be represented.

Depending on the textual units classified - respectively (a) 'words', (b) 'elementary contexts' or (c) 'documents' - the cells of such tables contain the following values:
a) number of occurrences of each word (i.e. the 'i' row) which, within the analysed corpus or a subset of it, has been assigned to a predefined category (i.e. the 'j' column). So, in such a case, words belonging to two or more predefined categories have the same values repeated in the corresponding columns;

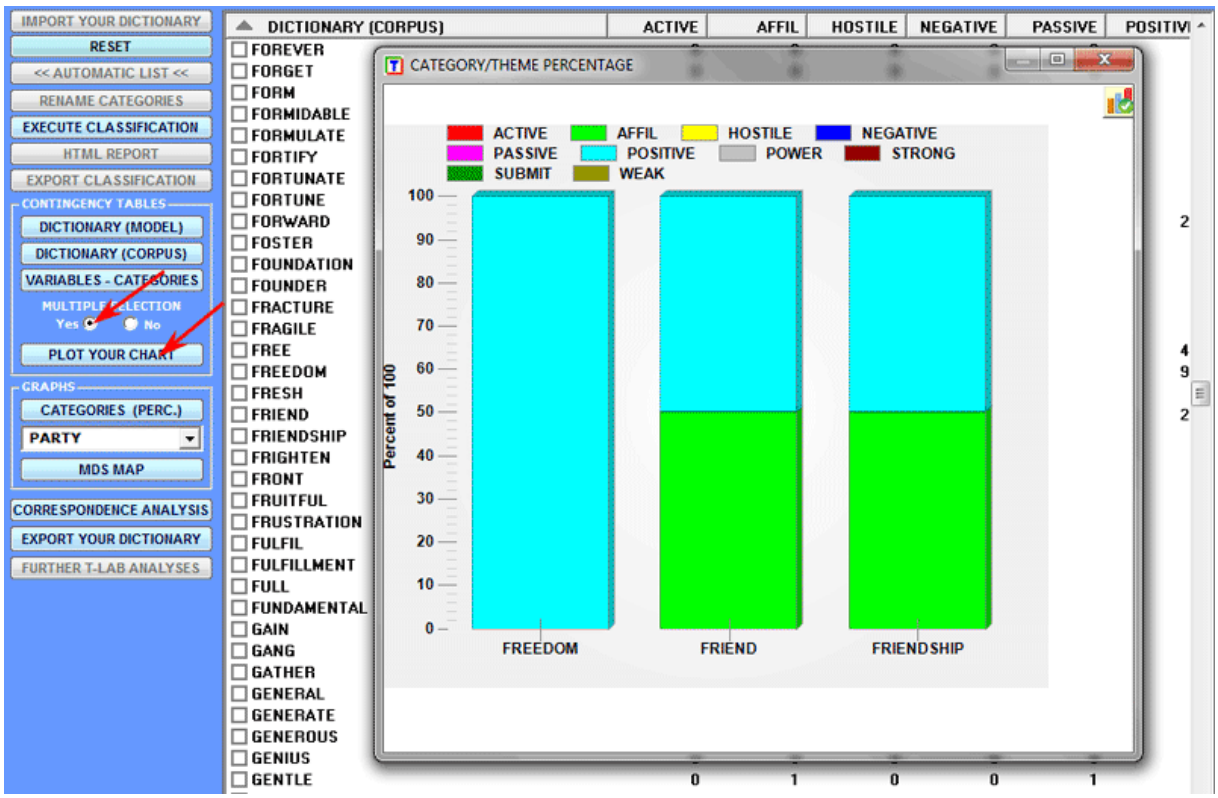
b) number of elementary contexts, which contain the word in the 'i' row, assigned to a given category (i.e. the 'j' column);
c) number of occurrences of each word (i.e. the 'i' row) within the documents assigned to each category (i.e. the 'j' column).

By clicking their respective check-boxes, it is possible to check the occurrence contexts of each listed word (N.B.: this option is available only for the 'b' case above, for which you click the appropriate cell) as well as to create customized charts concerning each item of the tables (N.B.: In the examples below some categories of the Harvard IV-4 dictionary have been applied to the analysis of inaugural addresses of US presidents).

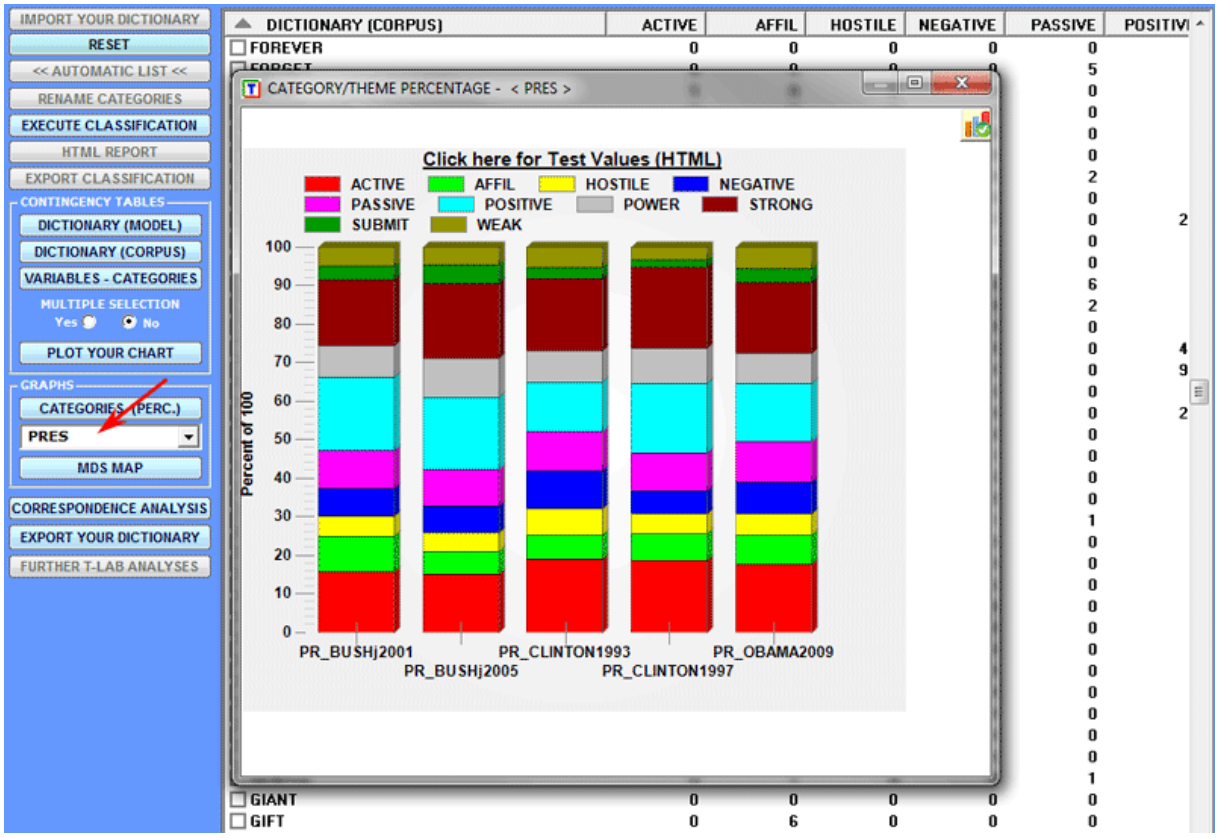




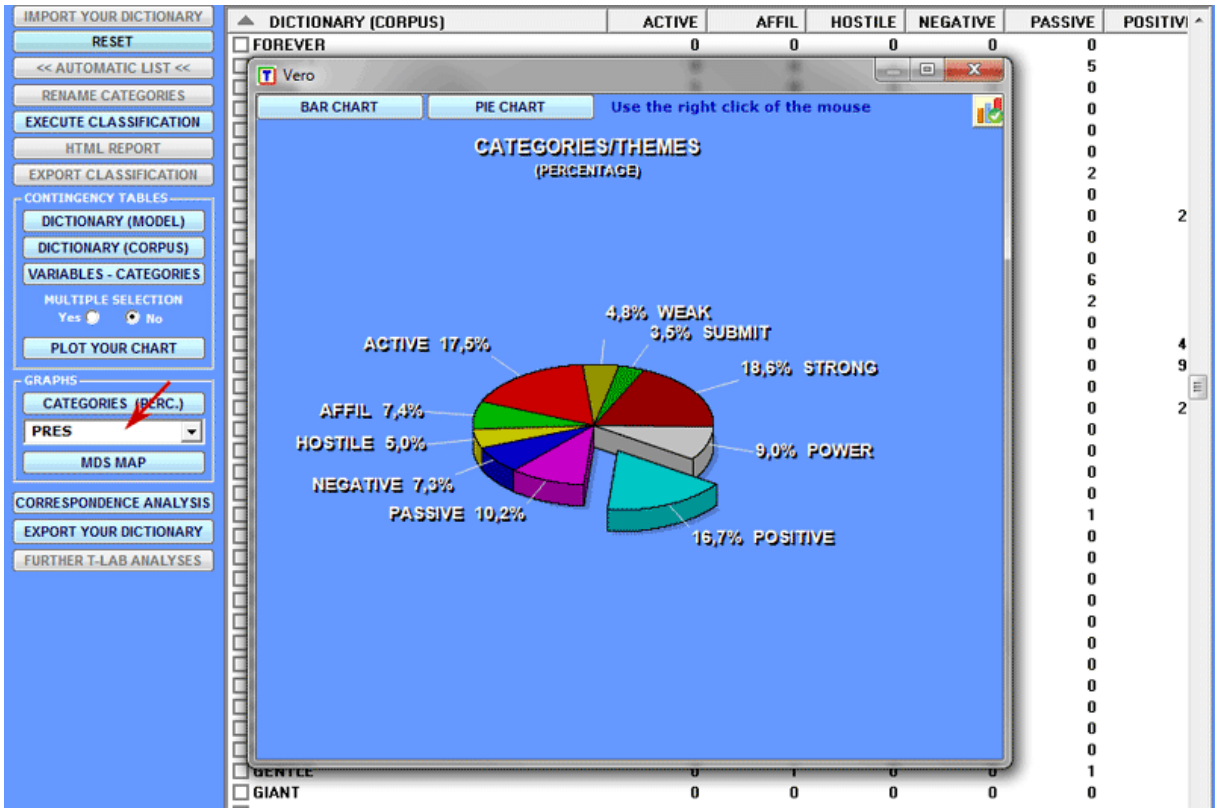
In order to plot charts with multiple data series, just choose '**Multiple Selection**' ('Yes' option), select up to 20 items and click the '**Plot your chart**' button (see below).



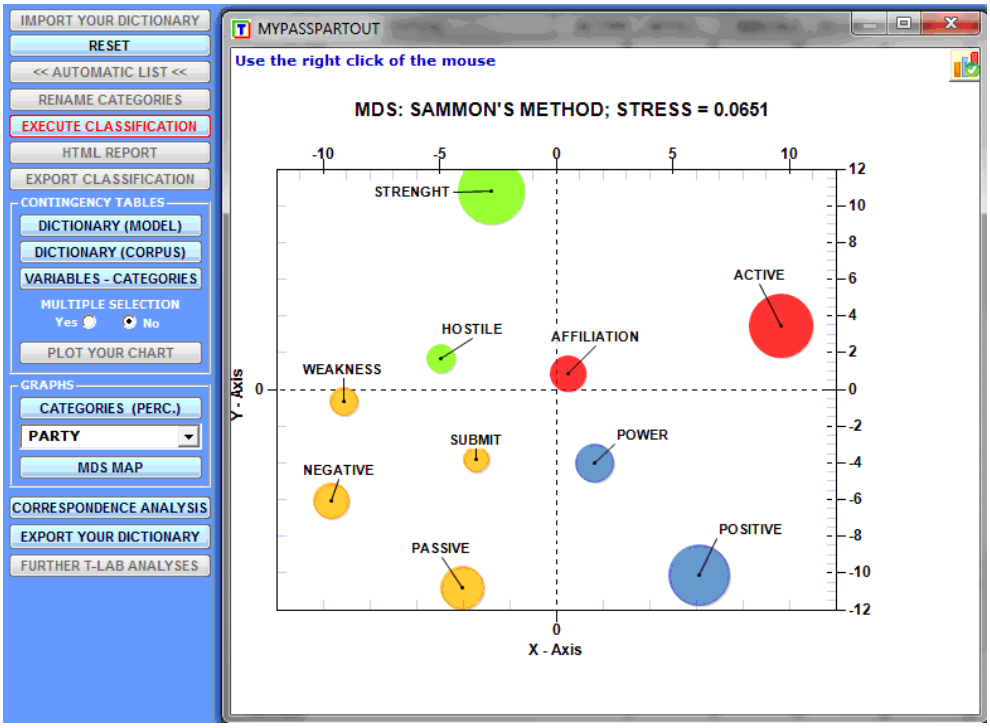
The above two options are also available for tables with variable values.

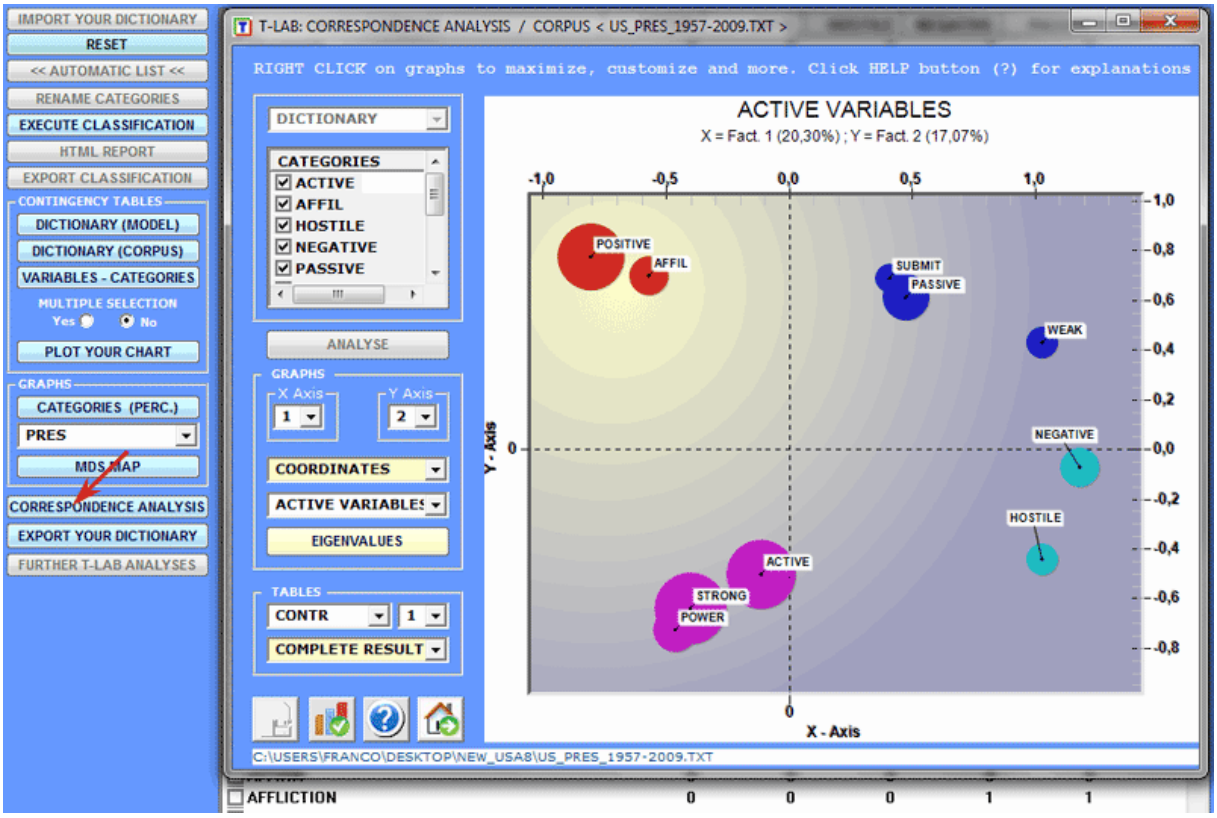


The percentage of categories can be appreciated in various ways (see below)



It is also possible to explore the data structure by using the '**MDS**' or the '**Correspondence Analysis**' tool (see below).





Only in the case where context units have been classified it is possible to obtain and export two more outputs; moreover, in such a case, it is possible to save the analysis results in a new variable and pursue the data exploration with other **T-LAB** tools.

For example, by clicking the '**HTML Report**' button it is possible to visualize some results of the classification process where the 'elementary contexts' or the 'documents' are assigned to their respective category with a Cosine similarity score (see the images below concerning a corpus of documents containing short descriptions of companies).

THEME < MEDICAL >

SCORE (.143)

Cytokinetics, Incorporated (Cytokinetics) is a biopharmaceutical company focused on developing small molecule therapeutics for the treatment of cardiovascular diseases and cancer. The Company's development efforts are directed to advancing multiple drug candidates through clinical trials to demonstrate proof-of-concept in humans in two markets: heart failure and cancer.

SCORE (.119)

Pharmacopeia, Inc. is a clinical development stage biopharmaceutical company dedicated to discovering and developing small molecule therapeutics to address medical needs. It has a portfolio of clinical and preclinical candidates under development internally or by partners, including eight clinical compounds in Phase II or Phase I development addressing multiple indications,

SCORE (.115)

Dyax Corp. (Dyax) is a clinical stage biotechnology company focused on the discovery, development and commercialization of biotherapeutics for unmet medical needs, with an emphasis on oncology and inflammatory indications. Dyax uses the drug discovery technology, known as phage display, to identify antibody, small protein and peptide compounds for clinical development.

SCORE (.111)

Rigel Pharmaceuticals, Inc. (Rigel) is a clinical-stage drug development company that discovers and develops small molecule drugs for the treatment of inflammatory/autoimmune diseases, cancer and viral diseases. The Company's research focuses on intracellular signalling pathways and related targets that are critical to disease mechanisms.

SCORE (.111)

It is also awaiting a decision from the United States Food and Drug Administration (FDA) regarding its application to market VELCADE for patients with diagnosed multiple myeloma. Millennium Pharmaceuticals, Inc. has a development pipeline of clinical and preclinical product candidates in its therapeutic focus areas of cancer and inflammatory diseases.

DOCUMENT	THEME	SCORE	BEGINNING
00001	SEMICONDUCTOR	0,051	2Wire , or not 2Wire , that is the question ...
00002	SEMICONDUCTOR	0,125	3Com Corporation (3Com) provides secure ...
00003	SEMICONDUCTOR	0,059	3D Systems Corporation is a holding company ...
00004	CHEMICAL	0,065	3M Company (3M) is a diversified technology ...
00005	SEMICONDUCTOR	0,095	What We Build 3PAR® (NYSE Arca : PAR ...
00006	MEDICAL	0,102	Abbott Laboratories is engaged in the discovery ...
00007	MEDICAL	0,071	ABIOMED , Inc . (ABIOMED) , provides ...
00008	CHEMICAL	0,046	Manufactures turbines & turbine generator ...
00009	CHEMICAL	0,085	ACCO Brands Corporation is a supplier of ...
00010	MEDICAL	0,013	focused on the casino industry . Developing ...
00011	CHEMICAL	0,078	Slides rule at Accuride International . ...
00012	MEDICAL	0,102	Established : Acorn Cardiovascular™ is ...
00013	SEMICONDUCTOR	0,094	Actel Corporation is a supplier of low-power ...
00014	MEDICAL	0,120	ActivBiotics , Inc . (ActivBiotics) ...
00015	SEMICONDUCTOR	0,129	ActivIdentity Corp . is a provider of digital ...
00016	CHEMICAL	0,126	Actuant Corporation (Actuant) is a manufacturer ...
00017	CHEMICAL	0,094	Acuity Brands , Inc . (Acuity Brands ...
00018	CHEMICAL	0,041	The Adams Manufacturing Company cares for ...
00019	SEMICONDUCTOR	0,145	Adaptec , Inc (Adaptec) , designs ...
00020	SEMICONDUCTOR	0,183	ADC Telecommunications , Inc . (ADC ...
00021	SEMICONDUCTOR	0,118	Adobe Systems Incorporated is a diversified ...
00022	MEDICAL	0,089	Adolor Corporation is a development-stage ...
00023	SEMICONDUCTOR	0,159	ADTRAN , Inc . (ADTRAN) designs , ...
00024	SEMICONDUCTOR	0,124	Advanced Analogic Technologies Incorporated ...
00025	MEDICAL	0,033	Advanced Ceramic Research was founded in ...

Similar data can be exported in .XLS files (see below) with all the information regarding the elementary contexts ('Context_Classification.xls') or the documents ('Document_Classification.xls') which have been correctly classified;

(1) - Context_Classification.xls

'IDNUMBER	THEME	SCORE	CONTEXT
'0000100001	SEMICONDUCTOR	0,017	2Wire , or not 2Wire , that_is the question : Whether ' tis nobler in networks to suffer the slings a
'0000100002	SEMICONDUCTOR	0,044	2Wire ' s HomePortal and OfficePortal networking devices combine router and firewall functions , ar
'0000100003	SEMICONDUCTOR	0,01	2Wire also makes DSL filters and adapters . Alcatel-Lucent owns one-quarter of 2Wire . For in bro
'0000200001	SEMICONDUCTOR	0,065	3Com Corporation (3Com) provides secure , converged networking solutions on a global scale to
'0000200002	SEMICONDUCTOR	0,081	3Com ' s long-term , technology-based strategy centers on enterprises and public_sector organizat
'0000300001	CHEMICAL	0,033	3D Systems Corporation is a holding company that operates through subsidiaries in the United_Sta
'0000300002	SEMICONDUCTOR	0,043	The Company ' s systems are used by its customers to produce physical objects from digital data u
'0000400001	CHEMICAL	0,035	3M Company (3M) is a diversified technology company with a presence in various businesses , i
'0000400002	CHEMICAL	0,024	3M manages its operations in six business segments : Industrial and Transportation ; health_care
'0000400003	CHEMICAL	0,032	The Company ' s products are sold through numerous distribution channels , including directly to u
'0000500001	SEMICONDUCTOR	0,018	What We Build 3PAR® (NYSE Arca : PAR) is the leading global provider of utility storage , a c
'0000500002	SEMICONDUCTOR	0,008	Next-generation storage is a category of arrays developed to address the limitations of traditional st
'0000500003	SEMICONDUCTOR	0,03	The Problem We Solve 3PAR Utility Storage is designed to address the problem of costly , comple
'0000500004	SEMICONDUCTOR	0,066	Our Customers 3PAR customers are organizations for whom delivering IT as a service is mission-cr
'0000500005	SEMICONDUCTOR	0,038	The Value We Bring 3PAR Utility Storage enables customers to cut Total Cost of Data by up_to 50
'0000600001	MEDICAL	0,033	Abbott Laboratories is engaged in the discovery , development , manufacture and sale of a diversifi
'0000600002	MEDICAL	0,042	The Diagnostic Products segment ' s products include diagnostic systems and tests for blood bank
'0000600003	MEDICAL	0,034	The Vascular Products segment ' s products include a line of coronary , endovascular and vessel c
'0000700001	MEDICAL	0,022	ABIOMED , Inc . (ABIOMED) , provides medical products and services in the area of circulator
'0000700002	MEDICAL	0,044	The Company ' s products can be used in a range of clinical settings , including by heart surgeons
'0000700004	MEDICAL	0,008	intra-aortic balloons (IABs) , and ventricular assist devices (VADs) .
'0000800001	CHEMICAL	0,046	Manufactures turbines & turbine generator sets & parts ; manufactures motor vehicle parts & acces
'0000900001	CHEMICAL	0,052	ACCO Brands Corporation is a supplier of select categories of branded office products (excluding f
'0000900002	CHEMICAL	0,03	personal computer accessory products , paper-based time management products , presentation a
'0000900003	CHEMICAL	0,013	During the year ended December 31 , 2007 , these markets represented 61% , 28% and 8% of its
'0001000001	MEDICAL	0,013	focused on the casino industry . Developing innovative new games , dazzling visual environments ,
'0001100001	CHEMICAL	0,017	Slides rule at Accuride International . Accuride International designs and makes ball bearing slides
'0001100002	CHEMICAL	0,072	The company ' s slides are also found in automotive accessories , including storage units and arm
'0001200001	MEDICAL	0,009	Establishe Acorn Cardiovascular™ is a privately held medical device company that was incorporate
'0001200002	MEDICAL	0,047	Mission : Acorn Cardiovascular develops innovative solutions to successfully treat patients with he
'0001200003	MEDICAL	0,031	Backgroun Heart failure (HF) is a condition that_is caused by damage to the heart muscle , whic
'0001200004	MEDICAL	0,027	An estimated 550 , 000 new HF cases are diagnosed each year in the United States alone . Heart
'0001200006	MEDICAL	0,033	It is intended to prevent and reverse the progression of heart failure by improving the heart ' s structu
'0001300001	SEMICONDUCTOR	0,065	Actel Corporation is a supplier of low-power field-programmable gate arrays (FPGAs) and progran
'0001300002	SEMICONDUCTOR	0,039	programming hardware and starter kits ; and a variety of design services . Its Flash-based solutions
'0001400001	MEDICAL	0,063	ActivBiotics , Inc . (ActivBiotics) is a biopharmaceutical company focused on the discovery , d

(2) - Document_Classification.xls

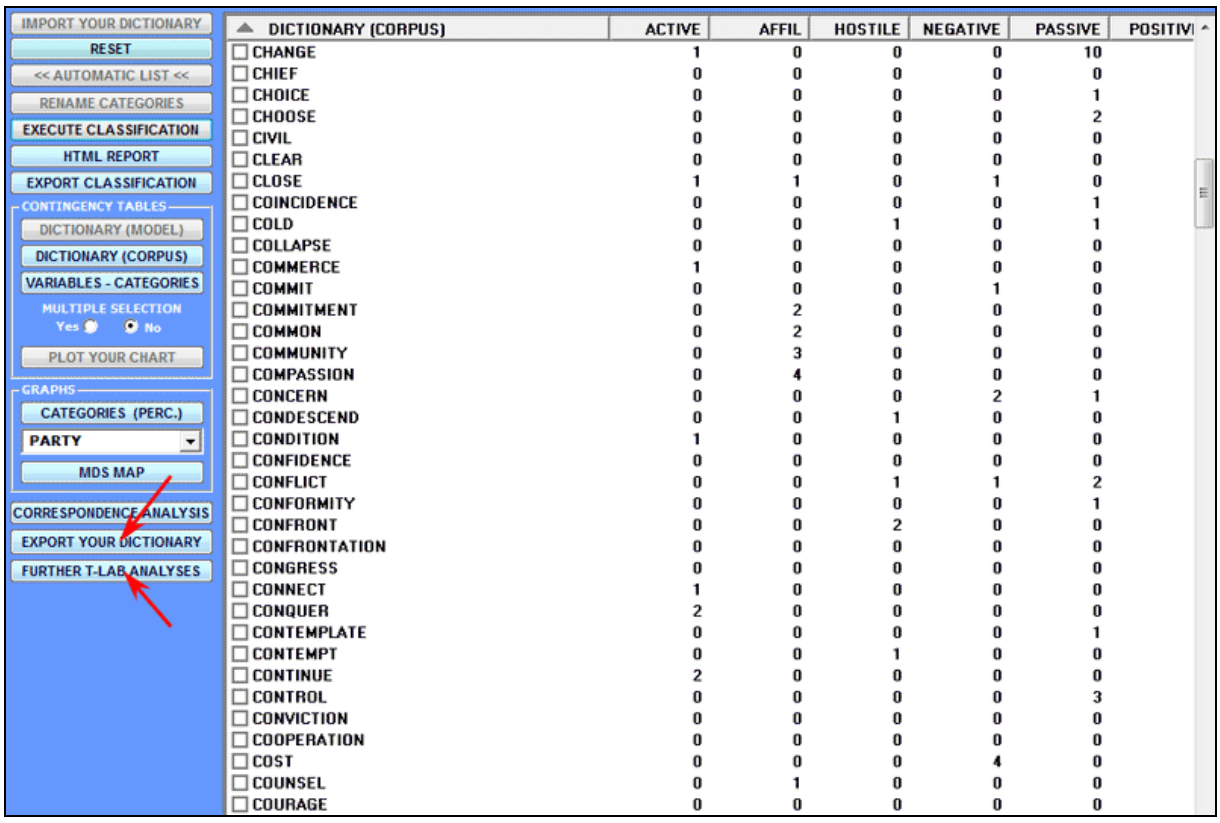
1	'IDNUMBER	THEME	SCORE
2	'00001	SEMICONDUCTOR	0,051
3	'00002	SEMICONDUCTOR	0,125
4	'00003	SEMICONDUCTOR	0,059
5	'00004	CHEMICAL	0,065
6	'00005	SEMICONDUCTOR	0,095
7	'00006	MEDICAL	0,102
8	'00007	MEDICAL	0,071
9	'00008	CHEMICAL	0,046
10	'00009	CHEMICAL	0,085
11	'00010	MEDICAL	0,013
12	'00011	CHEMICAL	0,078
13	'00012	MEDICAL	0,102
14	'00013	SEMICONDUCTOR	0,094
15	'00014	MEDICAL	0,12
16	'00015	SEMICONDUCTOR	0,129
17	'00016	CHEMICAL	0,126
18	'00017	CHEMICAL	0,094
19	'00018	CHEMICAL	0,041
20	'00019	SEMICONDUCTOR	0,145
21	'00020	SEMICONDUCTOR	0,183
22	'00021	SEMICONDUCTOR	0,118
23	'00022	MEDICAL	0,089
24	'00023	SEMICONDUCTOR	0,159
25	'00024	SEMICONDUCTOR	0,124
26	'00025	MEDICAL	0,033
27	'00026	SEMICONDUCTOR	0,045
28	'00027	SEMICONDUCTOR	0,046
29	'00028	CHEMICAL	0,057
30	'00029	MEDICAL	0,082
31	'00030	SEMICONDUCTOR	0,058
32	'00031	CHEMICAL	0,051
33	'00033	MEDICAL	0,138
34	'00034	CHEMICAL	0,129
35	'00035	CHEMICAL	0,035
36	'00036	SEMICONDUCTOR	0,064

D) - FURTHER ANALYSIS STEPS

When the classification process is over, two further options are available:

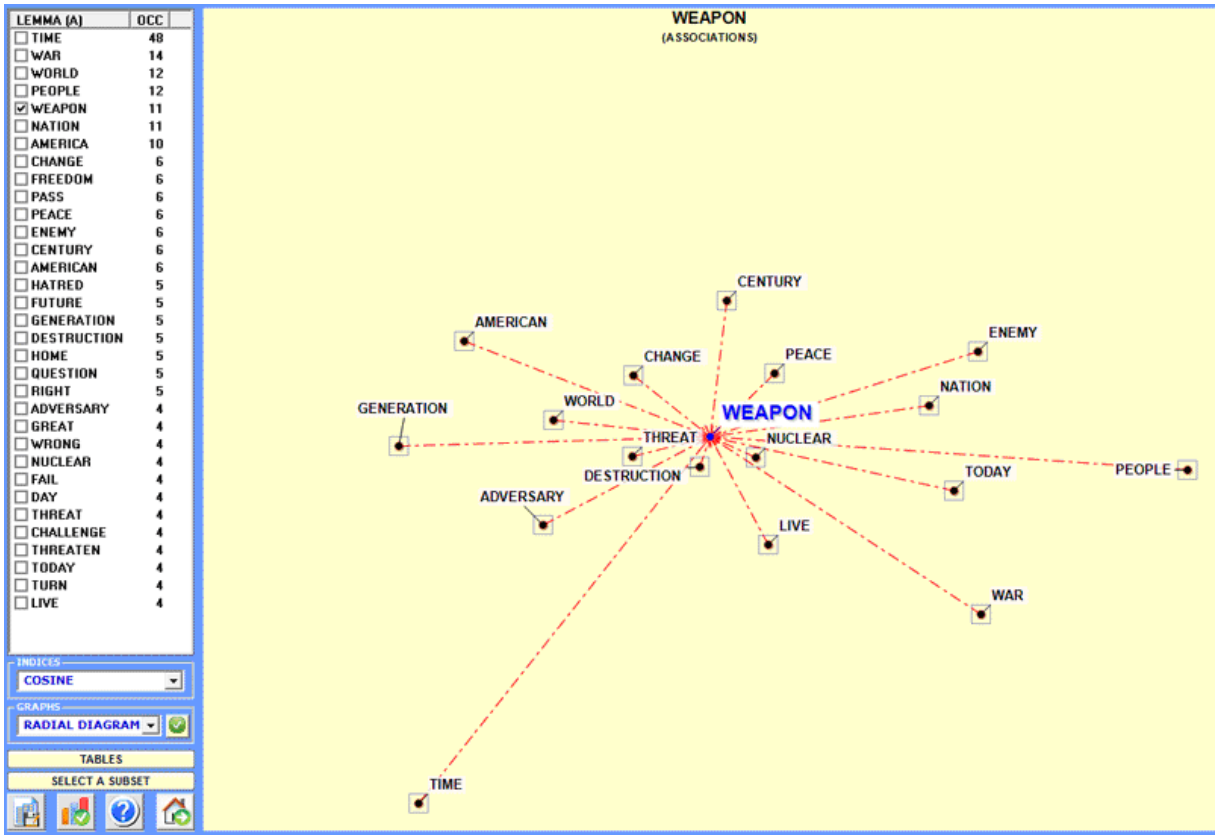
- '**Export Your Dictionary**', which creates a ready-made dictionary to be imported by other **T-LAB** tools for Thematic Analysis;

- '**Further T-LAB Analyses**', which, depending on the structure of the corpus analysed, on the kind of classification performed and on the number of categories used, makes a new variable for further **T-LAB** tools available (see below).



DICTIONARY (CORPUS)	ACTIVE	AFFIL	HOSTILE	NEGATIVE	PASSIVE	POSITIVI
CHANGE	1	0	0	0	10	
CHIEF	0	0	0	0	0	
CHOICE	0	0	0	0	1	
CHOOSE	0	0	0	0	2	
CIVIL	0	0	0	0	0	
CLEAR	0	0	0	0	0	
CLOSE	1	1	0	1	0	
COINCIDENCE	0	0	0	0	1	
COLD	0	0	1	0	1	
COLLAPSE	0	0	0	0	0	
COMMERCE	1	0	0	0	0	
COMMIT	0	0	0	1	0	
COMMITMENT	0	2	0	0	0	
COMMON	0	2	0	0	0	
COMMUNITY	0	3	0	0	0	
COMPASSION	0	4	0	0	0	
CONCERN	0	0	0	2	1	
CONDESCEND	0	0	1	0	0	
CONDITION	1	0	0	0	0	
CONFIDENCE	0	0	0	0	0	
CONFLICT	0	0	1	1	2	
CONFORMITY	0	0	0	0	1	
CONFRONT	0	0	2	0	0	
CONFRONTATION	0	0	0	0	0	
CONGRESS	0	0	0	0	0	
CONNECT	1	0	0	0	0	
CONQUER	2	0	0	0	0	
CONTEMPLATE	0	0	0	0	1	
CONTEMPT	0	0	1	0	0	
CONTINUE	2	0	0	0	0	
CONTROL	0	0	0	0	3	
CONVICTION	0	0	0	0	0	
COOPERATION	0	0	0	0	0	
COST	0	0	0	4	0	
COUNSEL	0	1	0	0	0	
COURAGE	0	0	0	0	0	

Below is an example obtained by analysing a 'subset' of classified contexts by means of the **Word Associations** tool (see the T-LAB main menu).



E) - INPUT AND OUTUT FORMAT OF T-LAB DICTIONARIES

Here is all the information about the format of dictionaries which can be imported by this **T-LAB** tool:
- they are plain text files with a '.dictio' extension (e.g. Mycategories.dictio);
- all dictionaries created by **T-LAB** thematic tools, including those created by the **Dictionary-Based Classification**, are ready to be imported and no further user intervention is required;
- other dictionaries, either 'standard' or customized must be saved by following the guidelines below:

1- they can include up to 100,000 records (i.e. lines), each consisting of strings separated by semicolons (e.g. economic;loan);
2- for each line, the first string must be a 'category', the second a 'word' (or lemma), the third - if present - must be a positive real number (i.e. a integer) from '1' to '999' which represents the 'weight' of each word within the corresponding category;
3- the maximum length of a string (word, lemma or category) is 50 characters: neither blank spaces no apostrophes can be included;
4- when multi-word phrases are included, blank spaces must be replaced by the underscore ('_') character (e.g. Federal_Government);
5- the number of categories used can vary from 2 (minimum) to 50 (maximum). When the number of categories is higher than 50 it is advisable to use a different format and import the dictionary by the **Dictionary Building** tool (see the '**Lexical Tools**' sub-menu of **T-LAB**). In such a case there must be univocal correspondence between each single word and the corresponding category.

The following are two excerpts from **T-LAB** .dictio files, with two or three strings per line respectively:

a) case with two strings (i.e. categories and words only)

…
negative;catastrophic
negative;bad
…
positive;outstanding
positive;supportive
…

b) case with three strings (i.e. categories, words and numbers)

…
negative;catastrophic;10
negative;bad;7
…
positive;outstanding;9
positive;supportive;8

Texts and Discourses as Dynamic Systems

This **T-LAB** tool provides several **integrated analysis options** (see picture below) which can be used in various combinations for obtaining measures and graphical representations concerning **texts treated as dynamic systems**.

In particular this tool allows us to verify how texts are organized in time, how the **recurring themes** and the **sequential order** of utterances relate to each other and how **similarities** and **differences** between them evolve in time. For these reasons this tool – more than other **T-LAB** tools - challenges the divide between qualitative and quantitative approaches in text analysis.

In principle the objects of this type of integrated analysis should be texts in which – like discourses and conversations – the **sequence** and the temporal flow of utterances is important (i.e. transcripts of focus group sessions, interviews, speeches, debates, doctor/patient iterations, novels etc.).

However, as this tool provides us with **similarity measures** concerning all pairs of text segments (both within the whole corpus and within its subsets), it may be also useful in other cases. Just remember that - when text segments are not in sequential order – the use of RQA Analysis and/or Sequence Analysis options does not produce proper results.

To begin with, two things must be taken into consideration:

-as the granularity is important, the key-word list chosen before using this tools should contain as many items as possible;
-at the moment, this tool allows us to analyse a corpus which includes up to 30,000 text segments (i.e. about 5,000 pages), which can even be organized in two or more sub-sections (i.e. corpus subsets). However, due to some limitations concerning the visualization of recurrence plots, both the RQA Analysis and the Similarities Measures are available only for corpuses consisting of up to 3,000 text segments (i.e. about 500 pages, and a bit more when the corpus has been segmented into paragraphs).

The **analysis procedure** consists of the several steps, some of which are automatic and others which – when desired - can be manually performed by the user.

The **initial steps** performed automatically by **T-LAB** are the following:

a - construction of a **document-term matrix**, where documents are always text segments (i.e. text fragments, sentences, paragraphs) into which the corpus has been subdivided (see the **T-LAB** initial settings options);
b - **topic analysis** based on a probabilistic model which uses the Latent Dirichlet Allocation and the Gibbs Sampling (see the related information on Wikipedia);

c – use of a **Naïve Bayes classifier** for estimating the probability values of each topic within each text segment, and for assigning each text segment to the topic (or theme **) it most closely resembles.

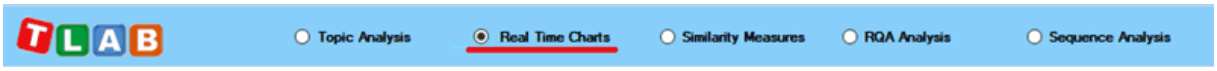
(**) 'Topic' and 'Theme' will be hereafter treated as synonymous terms.

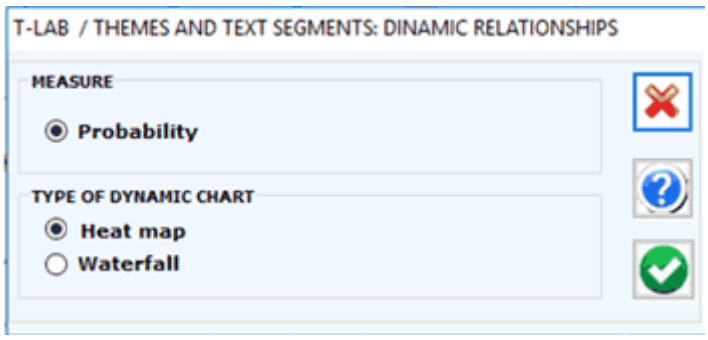
Please note that the main goal of the above automatic steps is to extract 'k' latent dimensions (where 'k' varies from 20 to 30) which determine the content structure of the analysed text and which – like a mixture model - can be used for exploring both text dynamics and similarities between text segments. For this reason the segments used for building the model are only those in which at least two key-terms included in the user list are present. Differently, after building the model, every text segment – even by maintaining the mixed nature of its content - is assigned to the topic to which it most closely resembles.

At the end of automatic steps, **five options** are made available, two of which correspond to two analysis tools already present in the **T-LAB** menu – namely the Topic Analysis (i.e. Modelling of Emerging Themes) and the Sequence Analysis of themes – and which, for this very reason, do not need further explanations. Just consult the parts of this help/manual where the main options depicted in the below section 'F' are commented.

Regarding the **new tools**, here is – for each of them - the required information.

A) Real Time Charts





When plotting real time charts, which allow us to **dynamically visualize** the time sequence of the text segments from the beginning to the end, the measures used are always the probability values that the Bayes classifier has assigned – for each of the 'k' topics - to each text segment.

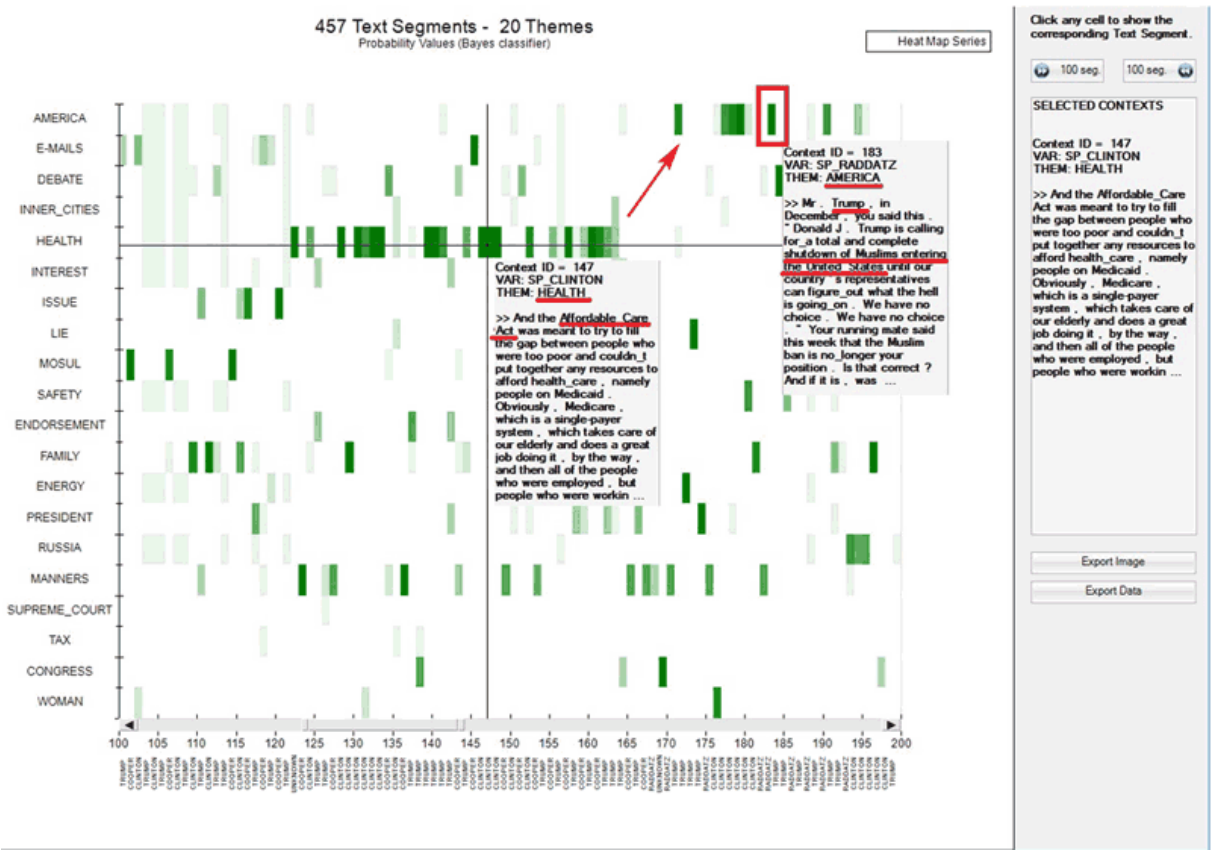
Two complementary charts allows us to easily appreciate various types of events, including the **strong recurrences** of some themes or the **shifts** from a theme to another (see the below pictures, obtained by analysing a presidential debate between Hillary Clinton and Donald Trump which took place on October 2016. N.B.: In this case the corpus was automatically segmented into paragraphs and a multi-word list was applied).

From a semiotic point of view, we may argue that both these types of charts deal with the relationships between **paradigm** and **syntagm** or – in other words – between the synchronic and diachronic axes, where paradigm/synchronic refers to the various themes and syntagm/diachronic refers to the temporal sequence of the 'N' text segments.

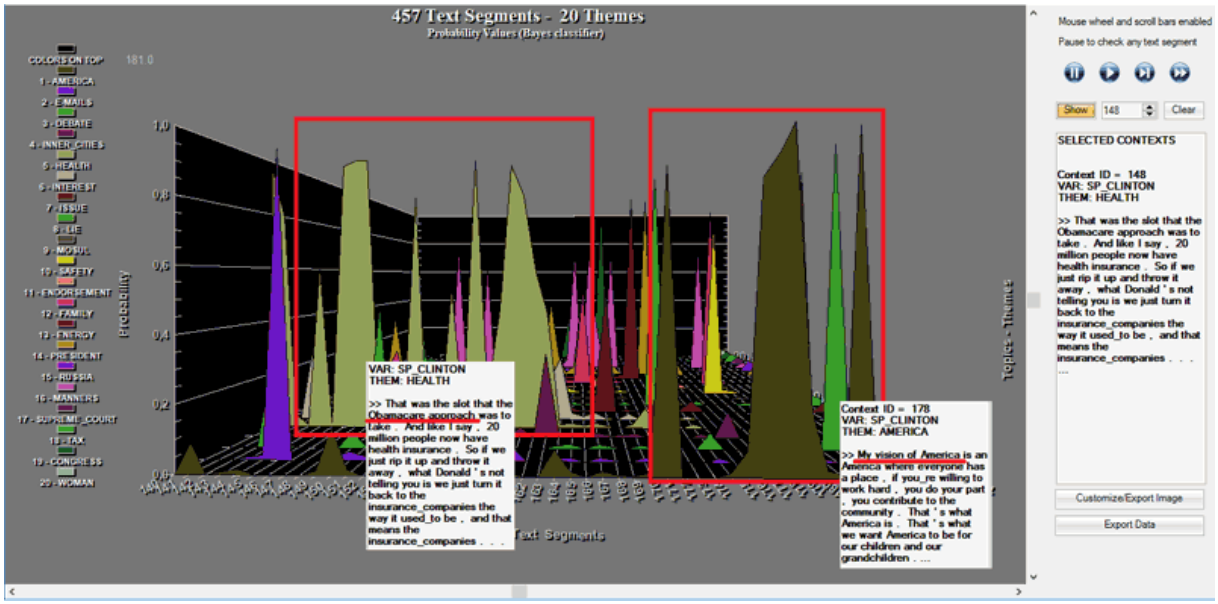
As the information summarized by these types of charts mainly refers to formal aspects of text contents, the same charts may be regarded as some sort of musical scores where the sequence of themes and their 'intensity' (i.e. probability) vary in time.

Anytime, in order to check 'who' is speaking and about 'what', just click the corresponding point.

A.1 - Heat map



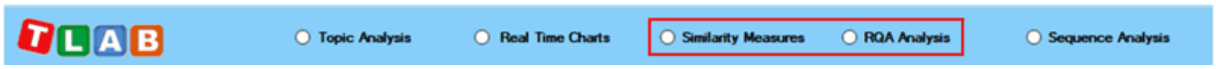
A.2 - Waterfall



Please note that in the real time charts all text segments are present, and each of them is represented as a mixture of probability values associated with the various topics which the model consists of. In fact, when clicking the 'Export Data' option, all this information is made available in a data table in CSV format like the following.

SPEAKER	THEME	ID_Segm	Selected	AMERICA	E-MAILS	DEBATE	INNER_CITIES	HEALTH	INTEREST	ISSUE	LIE	
SP_RADDATZ	MANNERS	1	16	0.0159	0.0003	0.0003	0.0027	0.0029	0.0006	0.0003	0.0003	...
SP_COOPER	MANNERS	2	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...
SP_UNKNOWN	DEBATE	3	3	0.0062	0.0000	0.9929	0.0000	0.0000	0.0000	0.0000	0.0000	...
SP_CLINTON	AMERICA	4	1	0.5593	0.1448	0.0002	0.0002	0.0006	0.0055	0.0148	0.0002	...
SP_CLINTON	AMERICA	5	1	0.9999	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	...
SP_CLINTON	AMERICA	6	1	0.9997	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...
SP_CLINTON	E-MAILS	7	2	0.1328	0.4183	0.3872	0.0130	0.0005	0.0003	0.0005	0.0001	...
SP_CLINTON	AMERICA	8	1	0.9969	0.0000	0.0000	0.0026	0.0000	0.0000	0.0000	0.0001	...
SP_COOPER	MANNERS	9	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...
SP_TRUMP	FAMILY	10	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...
SP_TRUMP	LIE	11	8	0.0000	0.0000	0.0000	0.0001	0.0244	0.0000	0.0000	0.9740	...
SP_TRUMP	LIE	12	8	0.0000	0.0000	0.0000	0.2745	0.0000	0.0001	0.0000	0.7248	...
SP_TRUMP	FAMILY	13	12	0.0003	0.0000	0.0252	0.0000	0.0000	0.0000	0.0000	0.0028	...
SP_TRUMP	INNER_CITIES	14	4	0.0016	0.0001	0.0001	0.7819	0.0002	0.0001	0.0007	0.1364	...
SP_COOPER	ISSUE	15	7	0.0000	0.0000	0.0071	0.0000	0.0000	0.0000	0.8903	0.0000	...
SP_TRUMP	E-MAILS	16	2	0.0002	0.7197	0.0000	0.0038	0.0000	0.0028	0.0000	0.0000	...
SP_TRUMP	FAMILY	17	12	0.0000	0.0000	0.0003	0.0046	0.0014	0.0769	0.0003	0.0001	...
SP_TRUMP	INNER_CITIES	18	4	0.0319	0.0004	0.0001	0.7348	0.0015	0.0152	0.0001	0.0835	...
SP_TRUMP	ENDORSEMENT	19	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...
SP_COOPER	MANNERS	20	16	0.0143	0.0139	0.0139	0.0161	0.0161	0.0245	0.0113	0.0117	...
SP_TRUMP	SUPREME_COURT	21	17	0.0230	0.0062	0.0017	0.0019	0.0154	0.0017	0.0014	0.0014	...
SP_COOPER	WOMAN	22	20	0.0004	0.0003	0.0003	0.0030	0.0027	0.0043	0.0003	0.0003	...
SP_TRUMP	WOMAN	23	20	0.0087	0.0011	0.0011	0.0013	0.0013	0.0011	0.0009	0.0352	...
SP_COOPER	ENDORSEMENT	24	11	0.0410	0.0398	0.0398	0.0460	0.0460	0.0398	0.0323	0.0336	...
SP_TRUMP	WOMAN	25	20	0.0002	0.0000	0.0000	0.0004	0.0000	0.0002	0.0000	0.0000	...
...	...	...	...	...	...	...	...	...	...	...	...	...

B) Preliminary information about the Recurrence plots



Both the 'Recurrence Quantification Analysis (RQA)' and the 'Similarity Measures' tools use the **recurrence plot** technique. That is to say they build a N × N matrix, the rows and columns of which – in our case - are text segments ordered according to their temporal sequence. However in the two cases the recorded information is different. In fact, in the first case (i.e. RQA) any **recurrence** – marked with an unshaded dot - refers to the presence (absence in the case of white spaces) of the same theme in the 'i' and 'j' items (i.e. where the 'X' and 'Y' values are the same) and uses a categorical time series as input; differently, in the second case (i.e. Similarity Measures) any recurrence – marked with a shaded dot - refers to the similarity (i.e. Cosine) concerning the 'i' and 'j' items, the values of which are continuous (i.e. they vary from 0 to 1).

N.B.: In the case of recurrence plots with similarity measures the cut-off limit used by **T-LAB** is 0.0001 (Cosine measure). This because many scholars tend to count all nonzero entries of the similarity matrix.

Though the two types of recurrence plots may highlight similar patterns (see the below Fig. 1 and Fig. 2, which have been obtained by analysing a legislative text), by default **T-LAB** uses the first (i.e. Fig. 1) for computing the RQA measures and it uses the second (i.e. Fig. 2) for exploring similarities and differences concerning text segments.

However, by clicking the appropriate button, the user is also allowed to obtain the RQA measures for the recurrence plots with the similarity measures. Just remember that, as in this case the percentage of recurrent points is higher, all RQA measures are somehow inflated.
The fact remains that, like the 2D barcodes used for marketing purposes, both the below recurrence plots can be seen as unique fingerprints of the analysed text.

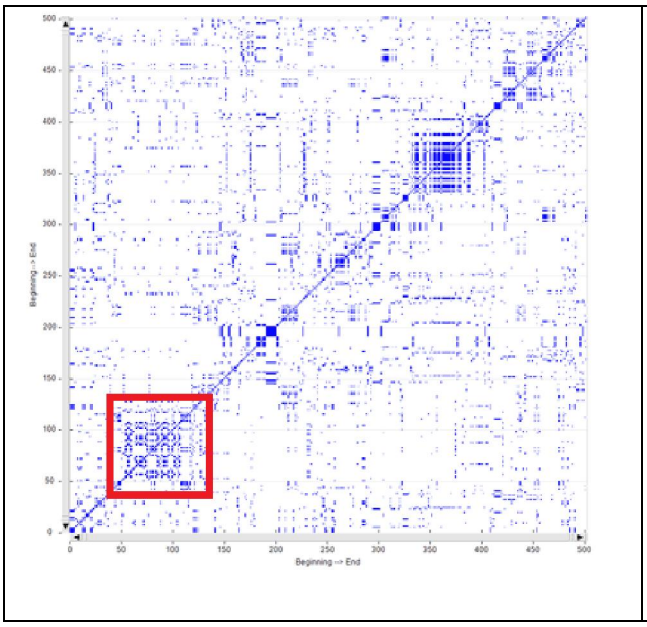


Fig. 1 - Time series

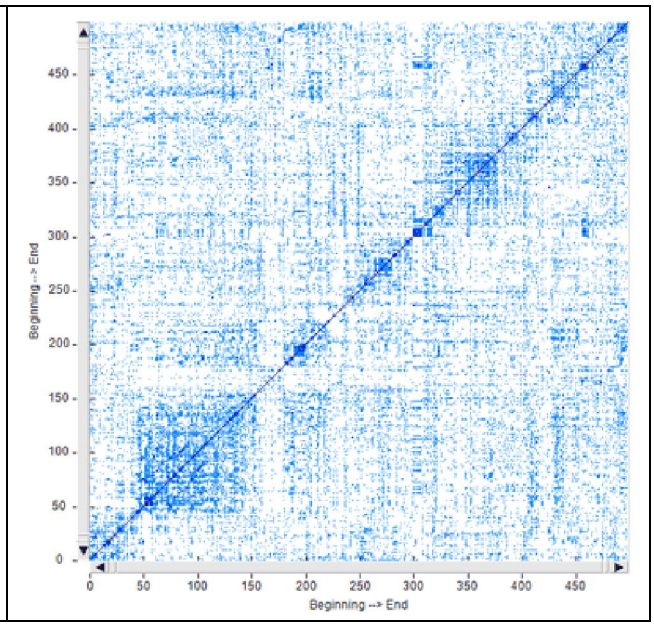
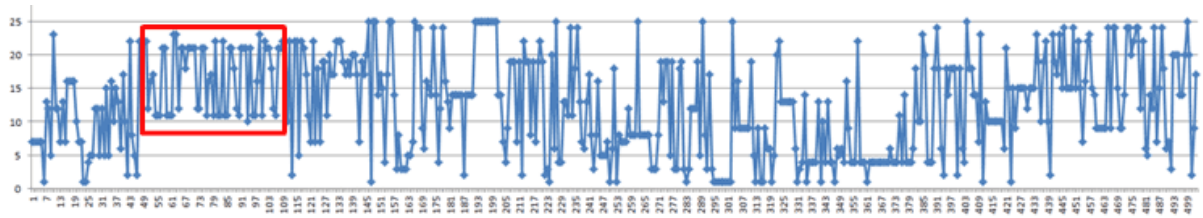
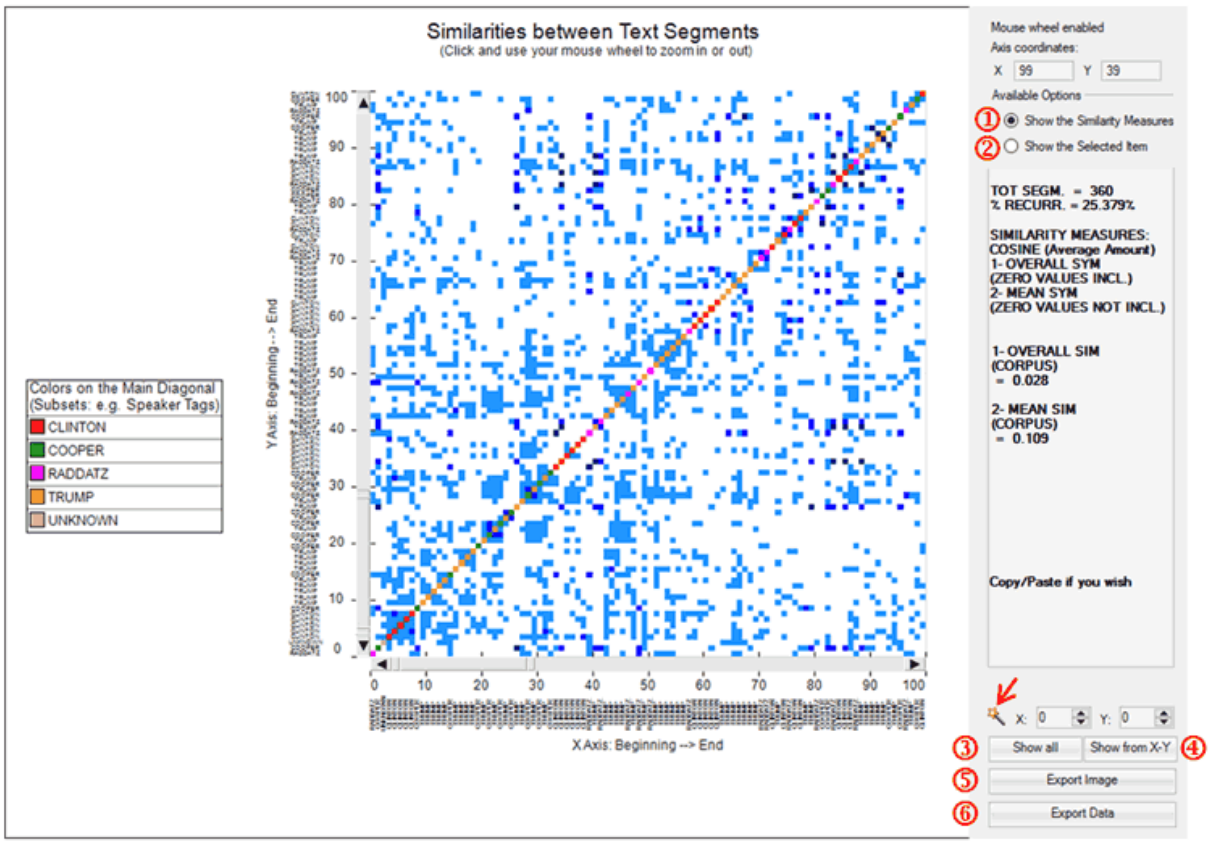


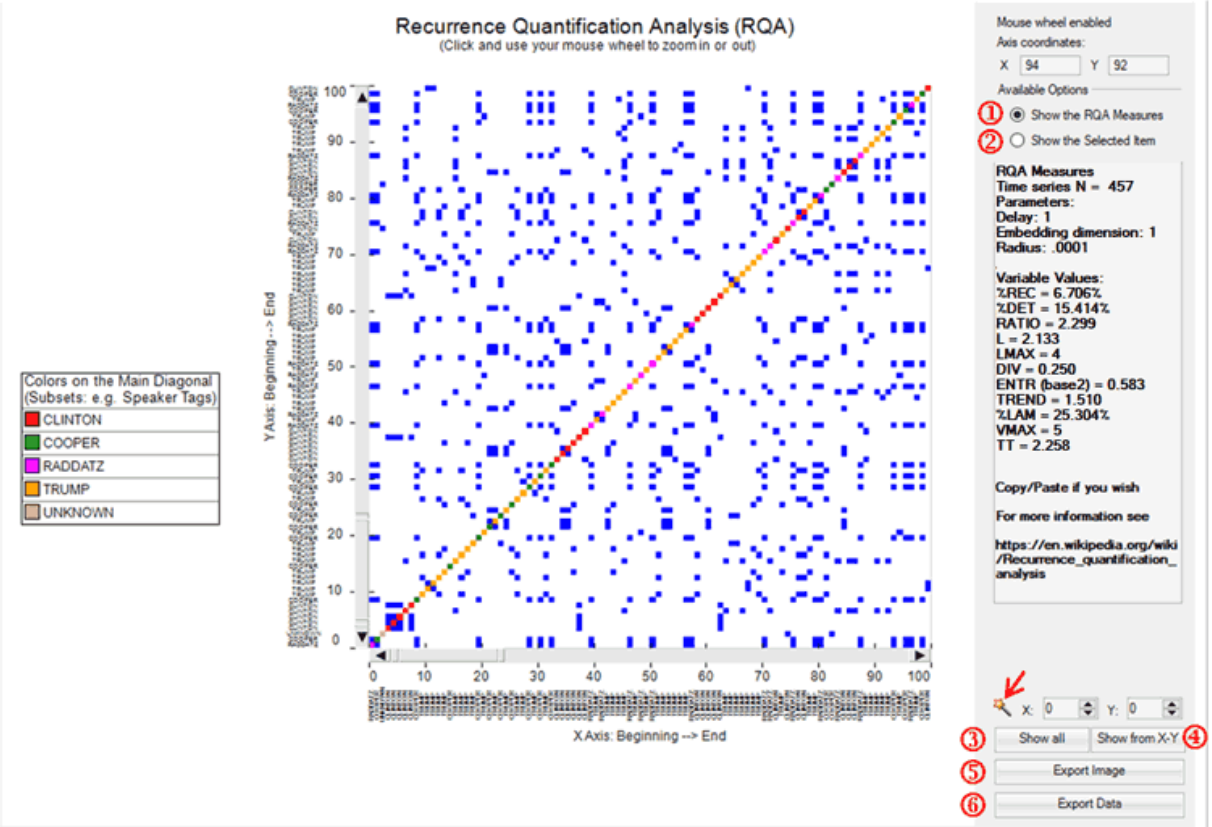
Fig2 - Similarities

N.B. The time series used for the recurrence plot in Fig. 1 is the following:



Both when clicking 'Similarity Measures' and 'Recurrence Quantification Analysis (RQA)' the default **T-LAB** chart shows a 100x100 recurrence plot which however **can be zoomed in and out** by using the mouse wheel. Moreover in both cases **six different options** allow us to perform different operations (see pictures below).



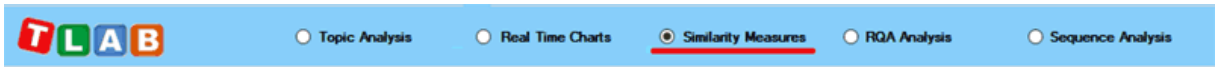


In particular:
-options ‘1’ and ‘2’ allow us to visualize the general measures (‘1’) or the transcript of the selected segment (‘2’);
-options ‘3’ and ‘4’ allow us to visualize the complete recurrence plot (‘3’) or a subsection of it (‘4’);
-options ‘5’ and ‘6’ allow us to export the image in different formats (‘5’) or to export a data table with all the analysed values (‘6’).

Please note:
-in the RQA case the magic wand button (🪄) allows us to check some characteristics which will be explained in the below section ‘D’. Differently, in the case of similarities, the same button may be used for obtaining the RQA measures for the shown recurrence plot;
-when exporting the similarity data, all measures concerning ‘Self-Similarity’ and ‘Other-Similarity’ are included (see table below).

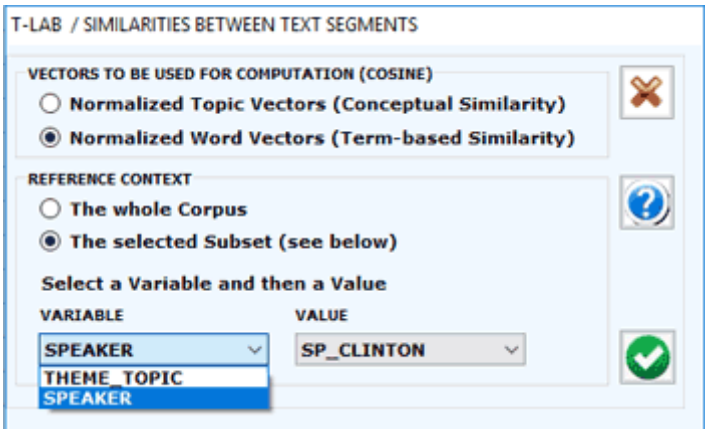
FIRST	SECOND	Cosine
SP_CLINTON	SP_CLINTON	0.0961
SP_CLINTON	SP_COOPER	0.1099
SP_CLINTON	SP_RADDATZ	0.1025
SP_CLINTON	SP_TRUMP	0.0847
SP_CLINTON	SP_UNKNOWN	0.1087
SP_COOPER	SP_CLINTON	0.1099
SP_COOPER	SP_COOPER	0.3106
SP_COOPER	SP_RADDATZ	0.2359
SP_COOPER	SP_TRUMP	0.1432
SP_COOPER	SP_UNKNOWN	0.1446
SP_RADDATZ	SP_CLINTON	0.1025
SP_RADDATZ	SP_COOPER	0.2359
SP_RADDATZ	SP_RADDATZ	0.2121
SP_RADDATZ	SP_TRUMP	0.1103
SP_RADDATZ	SP_UNKNOWN	0.1161
SP_TRUMP	SP_CLINTON	0.0847
SP_TRUMP	SP_COOPER	0.1432
SP_TRUMP	SP_RADDATZ	0.1103
SP_TRUMP	SP_TRUMP	0.1003
SP_TRUMP	SP_UNKNOWN	0.0958
...	...	...

C) Similarity Measures

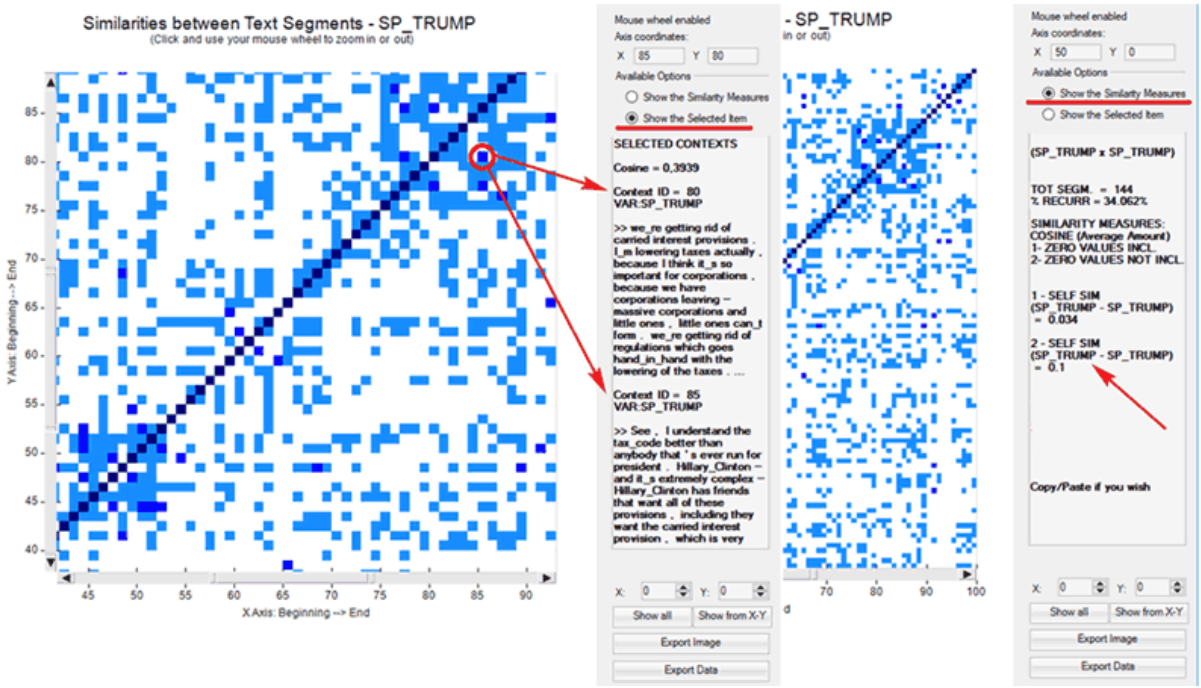


When choosing 'Similarity Measures', several options are made available (see picture below) which allow the user to select both the vectors to be used for the similarity computation and the reference context to be analysed (i.e. either the entire corpus or a subset of it).

N.B.: The difference between 'conceptual' (1) and 'term-based'(2) similarities is that in the first case (1) each text segment is represented by a feature vector concerning topics, whereas in the second case (2) each text segment is represented by a feature vector concerning words. In both cases the similarity measure used is the Cosine coefficient.

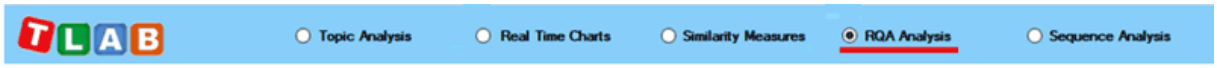


According to the design of the user interface, in this case - like in the RQA analysis (see section 'D' below) - the user can choose between visualizing the global measures or the transcripts of recurrent segments (see picture below). Moreover, when a corpus subset is selected, two further measures are provided concerning the 'self-similarity' (i.e. averaged cosine similarity) between all pairs of text segments within the chosen corpus subset, one (1) with and the other (2) without zero values included. Other measures concerning similarities between all pairs of corpus subsets can be exported by clicking the 'Export Data' button.



Please remember that, unlike the RQA, the 'Similarity Measures' option considers only those text segments in which at least two key-terms included in the user list are present. This is in order to reduce biases in the Cosine computation.

D) Recurrence Quantification Analysis (RQA)



RQA is a method of nonlinear data analysis for the investigation of dynamical systems which quantifies the information contained in a recurrence plot and detects the transitions in the systems by analysing time series (see https://en.wikipedia.org/wiki/Recurrence_quantification_analysis).

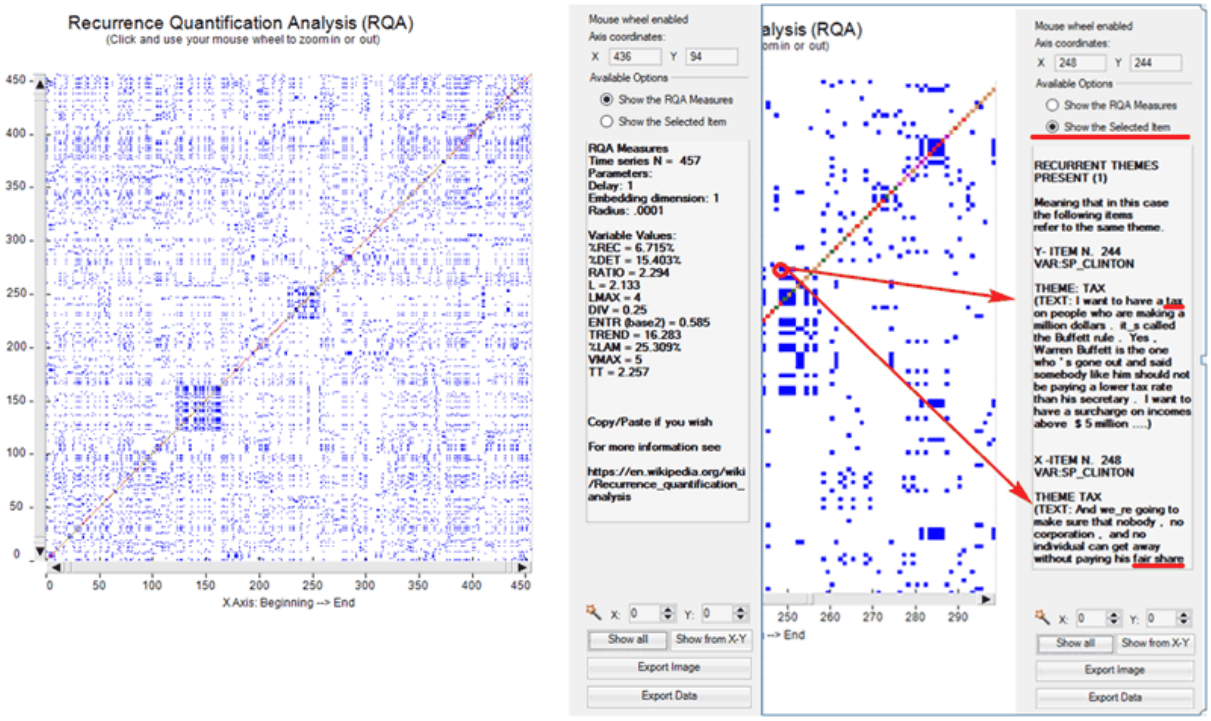
In this **T-LAB** tool, both in the case of the RQA Analysis and in the case of the Sequence Analysis (i.e. Markovian Analysis), a time series is represented by a categorical vector where each element is an integer which corresponds to the topic assigned to the 'i' text segment. However only in the case of the RQA a square matrix is built where the time series is both in rows and in columns.

When using the RQA tool, two main options are made always available (see pictures below):

1-Show the RQA Measures;
2-Show the Selected Item.

In the first case, the **standard measures** of RQA are provided (e.g. %REC, %DET, ENTR etc.**). In the second case the excerpts of recurring text segments are displayed. In both cases, the mouse wheel allows zooming in and out. Moreover two buttons allow the user to export both the picture and the analysed data.

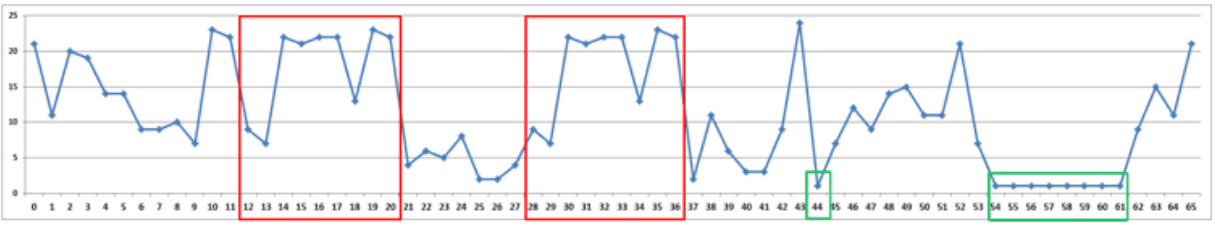
(**) For more information about the RQA measures see section 'E' below.



Please note that in the recurrence plot analysed with RQA the representation is symmetric across the main diagonal and two types of lines are particularly important: the **diagonals** parallel to the main diagonal and the **vertical lines** (**). In fact these lines mark the **transitions** present in the system and they are the base for obtaining the various RQA measures.

(**) In any recurrence plot vertical lines and horizontal lines mirror each other. In fact vertical lines in the upper part of the plot correspond to horizontal lines in the lower part, and vice versa.

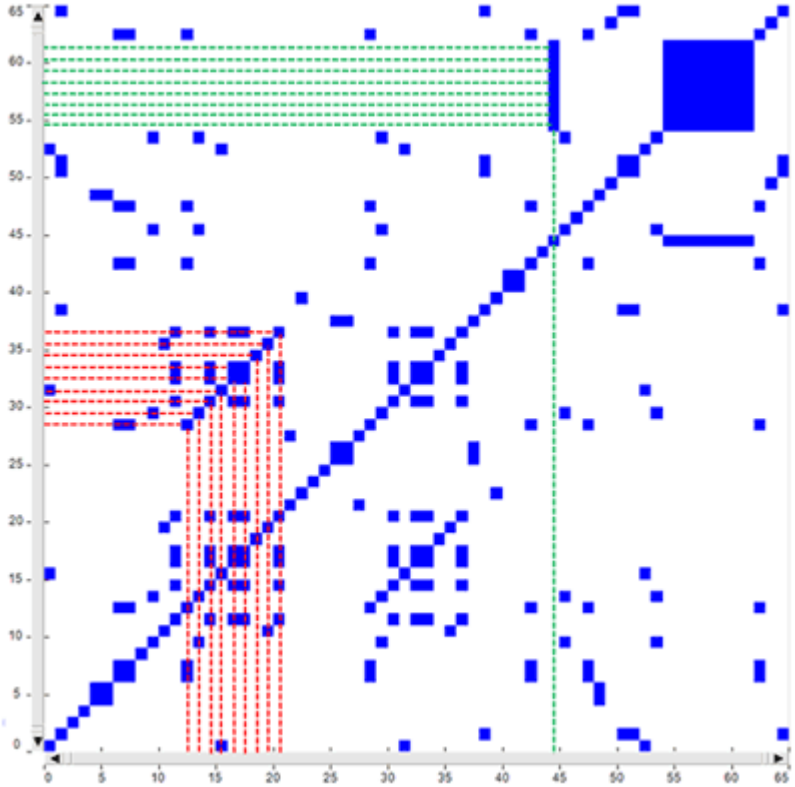
In particular, the distribution of diagonal lines allows for the investigation of **determinism** (i.e. the predictability of the system) and the distribution of vertical lines allows for the investigation of **intermittency** (i.e. the sequences which are interspersed by erratic breaks).



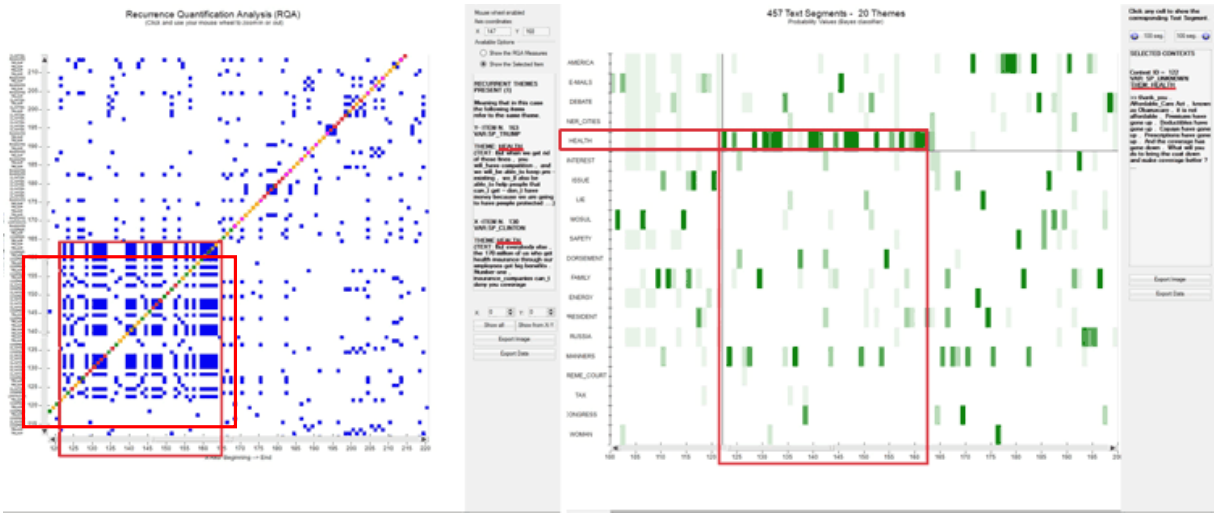
As an example, just consider the above fictitious time series. In it the same sequence of nine points/themes is repeated two times in different time spans (see the above red rectangles), respectively from t-12 to t-20 and from t-28 to t-36, where each 't' stands for a different text segment. In the same series there is also a sequence – from t-54 to t-61 - in which the same theme which appears at t-44 is repeated eight times (see the above green rectangle).

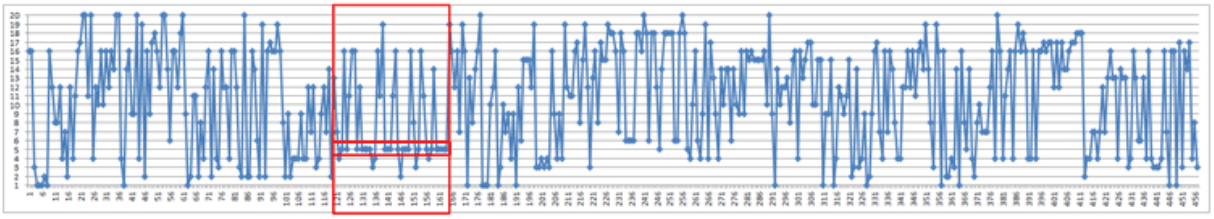
The corresponding recurrence plot (RP) - which has the same time series on the 'X' and the 'Y' axes - is that depicted in the image below.
Please note that in the case of diagonal line each point on the 'X' axis (i.e. from t-12 to t-20) recurs with the corresponding point on the 'Y' axis (i.e. from t-28 to t-36); differently the eight points which form the vertical line recur with just one point (i.e. t-44). Accordingly, in musical terms we may say that diagonal lines refer to a restatement of a motif (i.e. a pattern is repeated), whereas vertical lines refer to a repetition of a single note which somehow breaks the thematic variation.

Please note that when a monothematic sequence like that form t-54 to t-61 is repeated two or more times, usually in the recurrence plot it is represented by a square or by a rectangle.



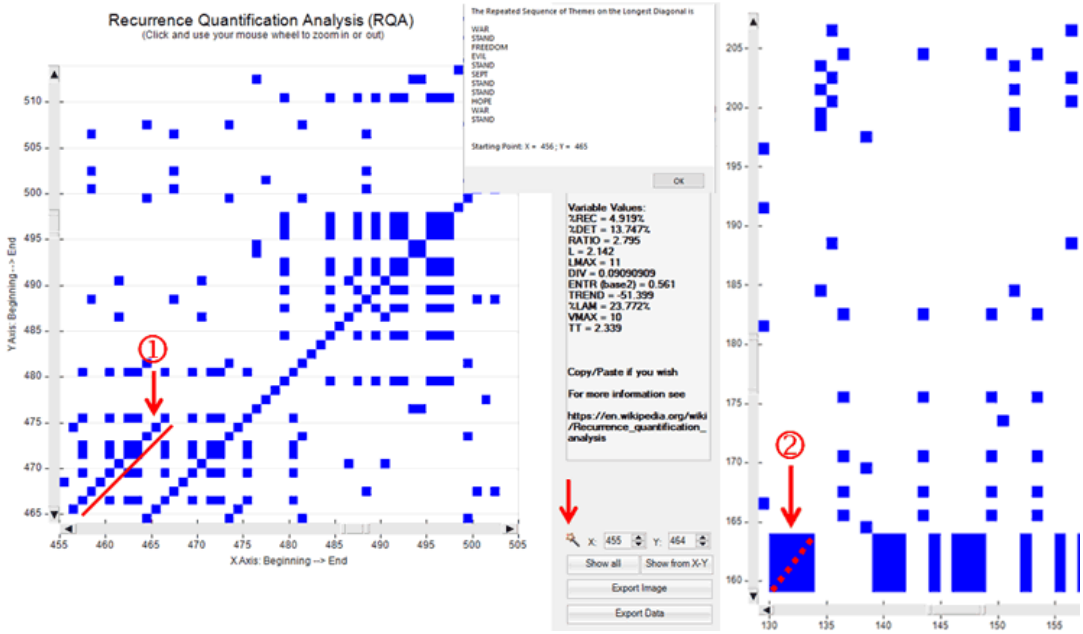
Regarding the **rectangular block** structures – which actually include both vertical and diagonal lines - they can be seen as referring to recurrences of the same topics in sub sections of the time series, i.e. to groups of overall similar feature vectors. In fact each dot in the graph represents a revisit of the same state and there is a correspondence between the rectangular blocks of the recurrence plot, the rectangles highlighted in the real time heat map and the chart of the time series (see pictures below). In other words we may say that in this cases speakers are repeatedly engaged on the same topic/theme, which appears to be 'hot'.



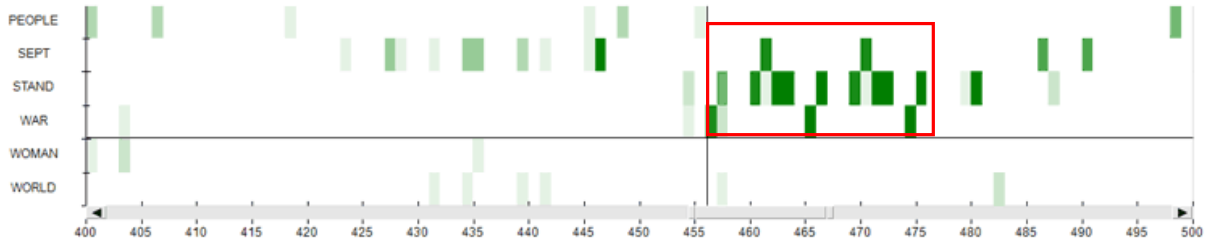


As stated above, in the RQA outputs the **longest diagonals** parallel to the main diagonal allow us to detect interesting repetitions of the same thematic sequence. However their shapes are not so evident as the rectangular block structures, also because sometimes they can be hidden inside one of them (see the below case marked with '2'). For this reason T-LAB includes a specific option (see the magic wand below) which automatically detects the longest diagonal, informs the user about the sequence of repeated themes included in it and automatically positions the cursor in the corresponding X-Y coordinates.

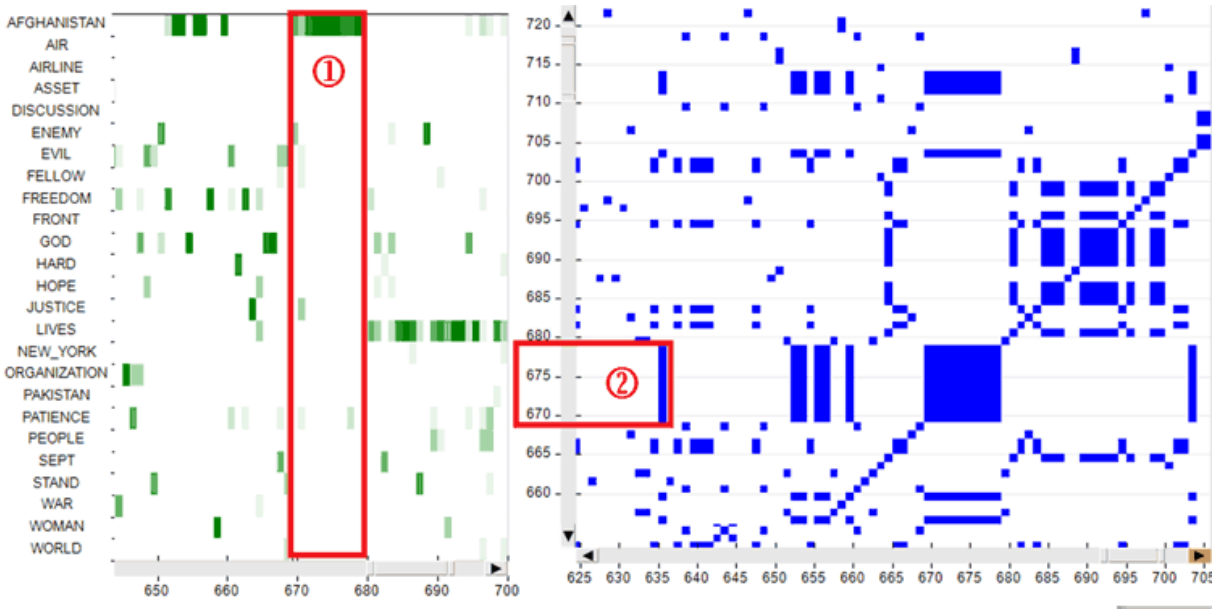
N.B.: Soon after the longest diagonal is detected **T-LAB** allows the user to export a file with the most frequent **repeated sequences**, each one of them including at least three concatenated themes. Such a file can be considered a sort of summary of the main themes - and of the corresponding variations - present in the corpus.



N.B.: In the case of the above diagonal '1', one of the corresponding patterns on the heat map is the following.



Regarding the vertical/horizontal lines they can be easily checked by exploring the heat map first (see case '1' in the image below) and then the recurrence plot (see case '2' in the image below).



E) Some notes about the RQA measures

When talking about the RQA measures, we have to make a clear distinction between their technical definitions (1) and their relevance in a thematic text analysis (2).
In fact the technical definitions correspond to formulas and are the same in all sciences using RQA for the study of dynamic systems and their time series (e.g. physics, physiology, meteorology, finance, etc.). Differently, the relevance – and also the meaning – of the RQA measures in text analysis is a matter of debate.

Starting with the technical definitions (1), here is a table which summarizes the relevant information for the most used RQA measures.

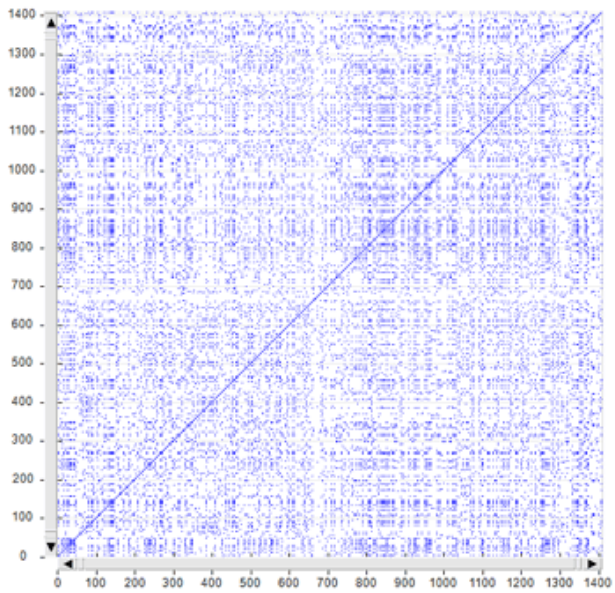
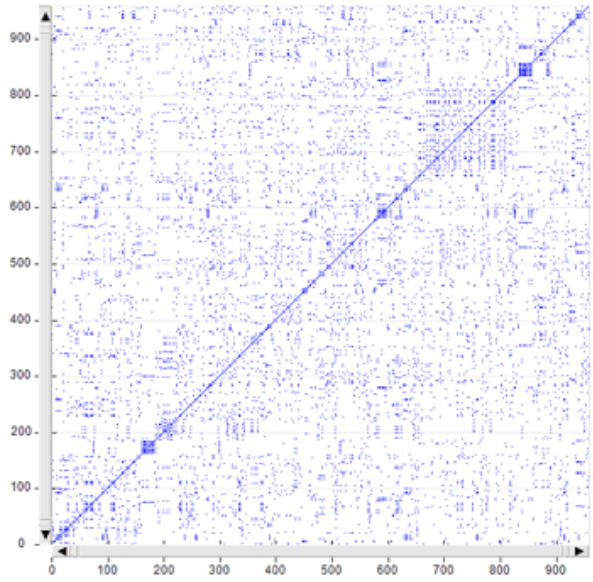
Measure	**Definition**
%REC - Recurrence Rate	The percentage of recurrence points in a Recurrence Plot which fall within a specified radius.
%DET - Determinism	The percentage of recurrence points which form diagonal line structures, main diagonal not included (N.B.: In RQA the main diagonal is also called LOI, i.e. Line of Identity, because in it each point recurs with itself).
RATIO	The ratio between %DET and %REC.
L	The average length of the diagonal lines.
LMAX	The length of the longest diagonal line.
DIV - Divergence	The inverse of LMAX.
ENTR - Entropy	The Shannon entropy of all diagonal line lengths distributed over integer bins in a histogram (Webber, C. L., & Zbilut, J. P. , 2005, p. 48). Accordingly, if there are lots of diagonal lines with varying lengths, the entropy will be high. Please note that, as in the RQA case entropy reflects the complexity of the RP in respect of the diagonal lines, here the definition of entropy does not correspond to the entropy of physical systems, where the higher the entropy the greater the disorder.
TREND	The degree of system stationarity . Accordingly, when recurrent points are homogeneously distributed across the recurrence plot, TREND value will be close to zero. Differently, when points 'fade away' from the central diagonal, the trend will have a negative value.
%LAM - Laminarity	The percentage of recurrence points which form vertical lines.
VMAX	The length of the longest vertical line.
TT – Trapping time	The average length of the vertical lines.

Regarding the relevance of RQA measures in text analysis (2) both **%DET** and **TREND** deserve special attention. In fact higher determinism (%DET) values indicates that the same thematic patterns are repeated more often and that – accordingly – the dynamic of analysed system is somehow more predictable. On the other hand TREND can be interpreted as a measure referring to how quick the transitions are from some themes to others, where lower TREND values indicate quicker transitions.

For example, when comparing RQA measures obtained by analysing a scientific essay ('a') and a novel ('b'), we can find out that in the first case ('a') the %DET value is higher than 'b' and that in the second case ('b') the TREND value is very low (often below zero).
Below is a comparison of the RQA measures obtained by analysing the essay 'On the Origin of Species' (C. Darwin) and the novel 'The adventures of Pinocchio' (C. Collodi).

On the Origin of Species (a)	The adventures of Pinocchio (b)
%REC = 8.201%	**%REC = 3.525%**
%DET = 16.474%	**%DET = 9.676%**
RATIO = 2.009	RATIO = 2.745
L = 2.093	L = 2.089
LMAX = 6	LMAX = 5
DIV = 0.167	DIV = 0.2
ENTR (base2) = 0.460	ENTR (base2) = 0.435
TREND = 4.705	**TREND = -5.599**
%LAM = 30.717%	%LAM = 23.194%
VMAX = 7	VMAX = 6
TT = 2.263	TT = 2.267

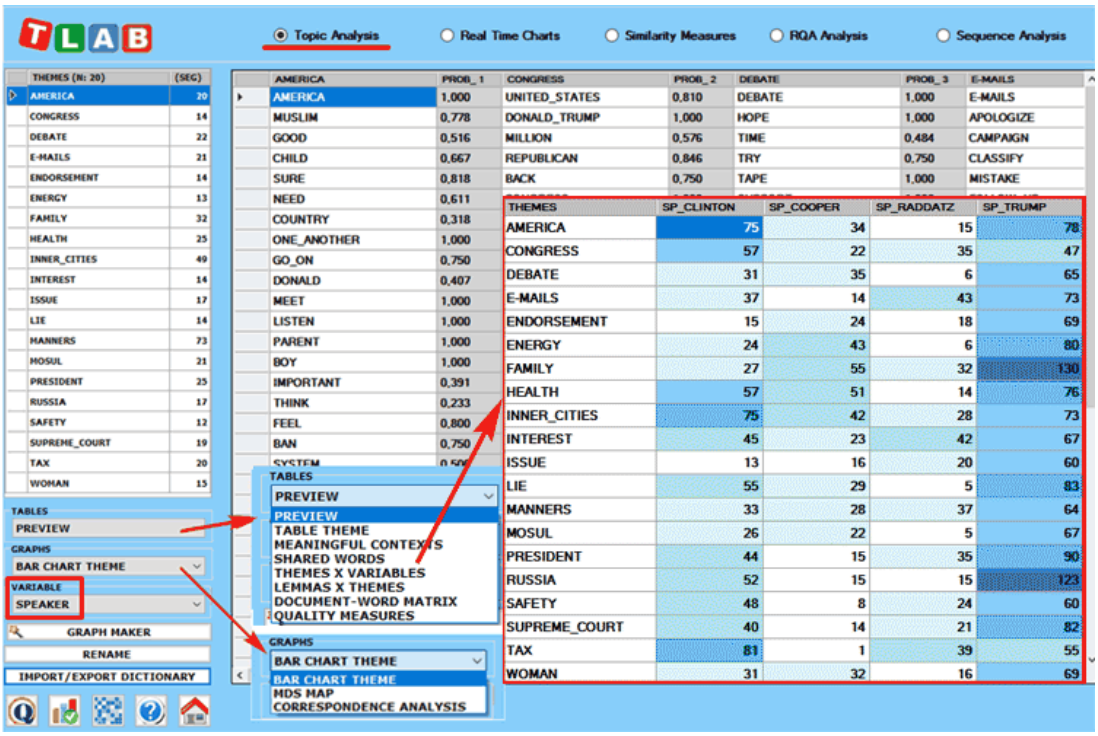
Here are the two corresponding recurrence plots.

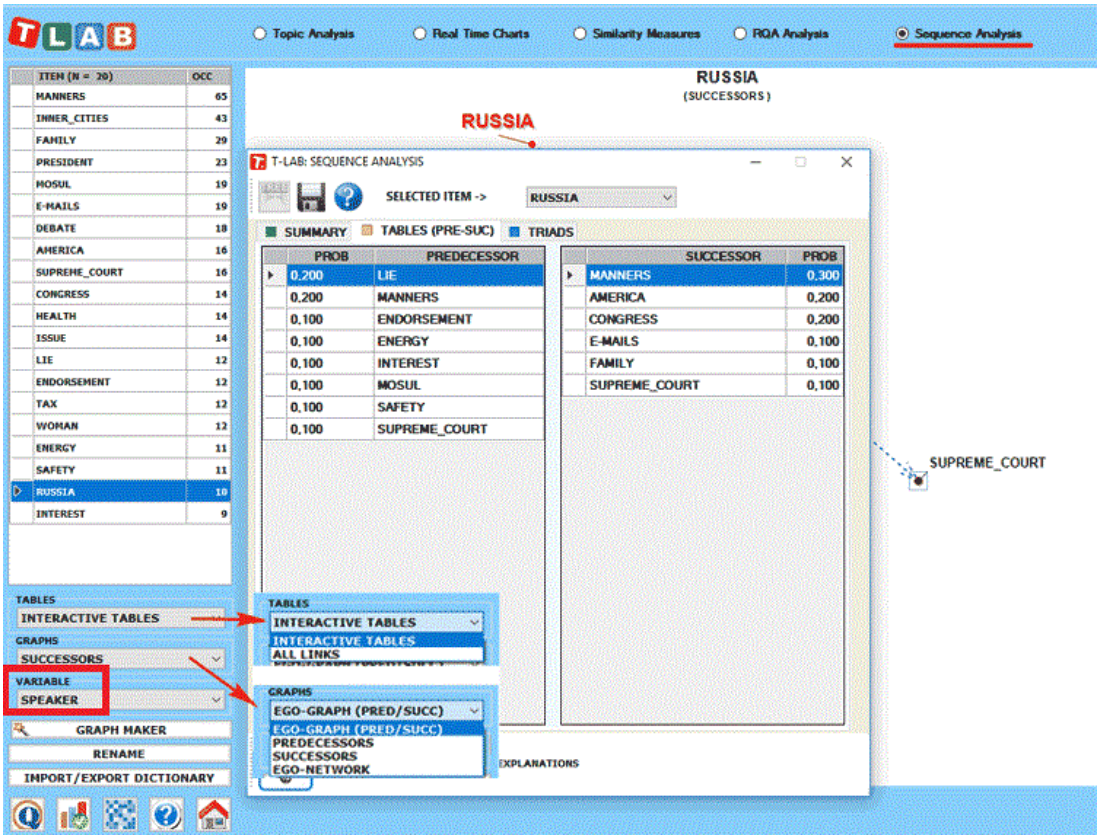
On the Origin of Species (a)	The adventures of Pinocchio (b)
	

N.B.: A table which summarizes the meanings of typical patterns in recurrence plots can be found at page 251 of the following article:
N. Marwan, M. Romano, M. Thiel and J. Kurths, "Recurrence Plots for the Analysis of Complex Systems", Phys. Rep. 438, 240-329 (2007).

F) Topic Analysis and Sequence Analysis

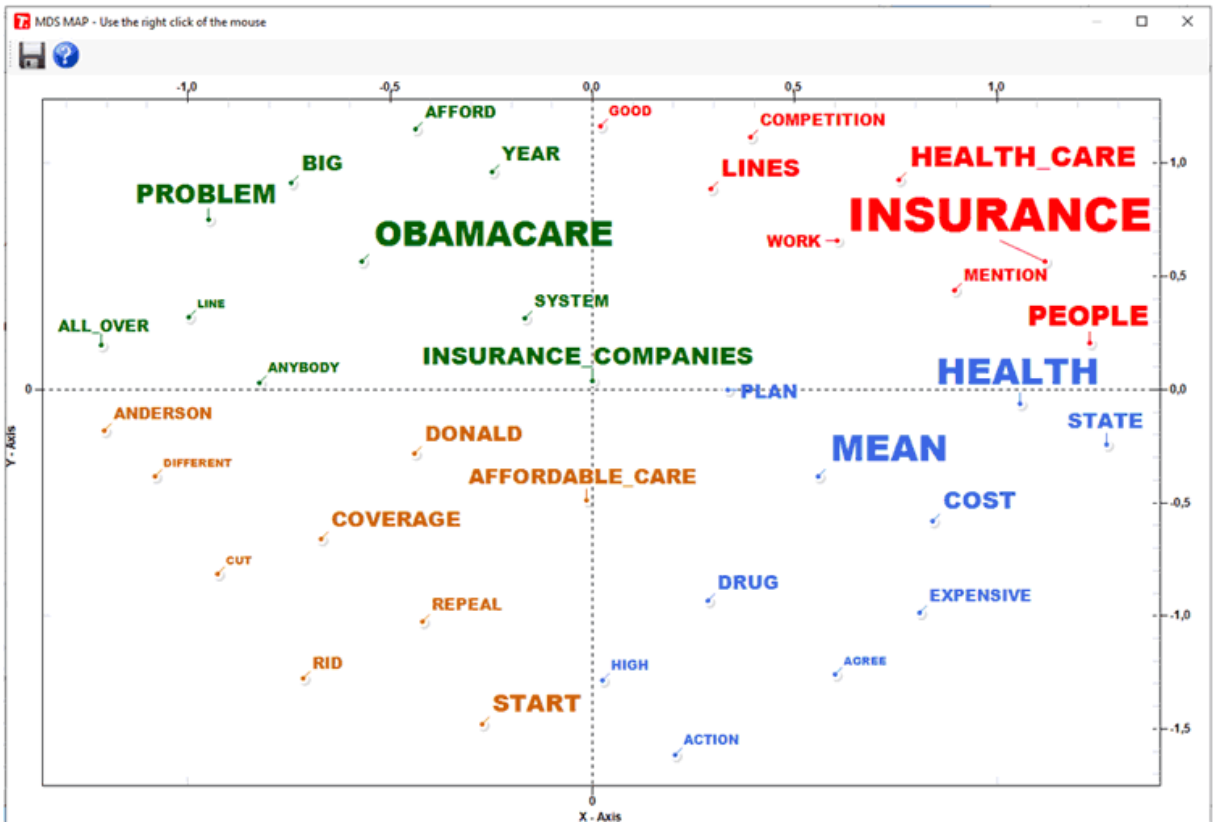
The below pictures summarize the main options of two tools already present in the **T-LAB** menu, which are integrated with the new ones and which are explained in the corresponding sections of this manual/help, i.e. 'Modeling of Emerging Themes' and 'Sequence and Network Analysis'.





N.B.:
-Any variable selected in the above forms (see the label highlighted by a red rectangle) will be used in the outputs provided by the various tools (Please note that only categorical variables with up to 20 values are made available) ;
-The 'Export/Import Dictionary' option, which is no longer available after performing a Sequence Analysis, is intended to allow the user to save time when repeating the same analysis by using topic labels manually assigned previously. In other words: just export the topic dictionary after completing - if desired - all renaming operations and import the same dictionary when repeating the same analysis with the same corpus, the same key-word list and the same parameters;
-While the Correspondence Analysis option allows us to explore the relationships between the various topics and the various speakers, the 'Graph Maker' tool allows us to explore the relationships between key-terms within each selected topic (see pictures below).

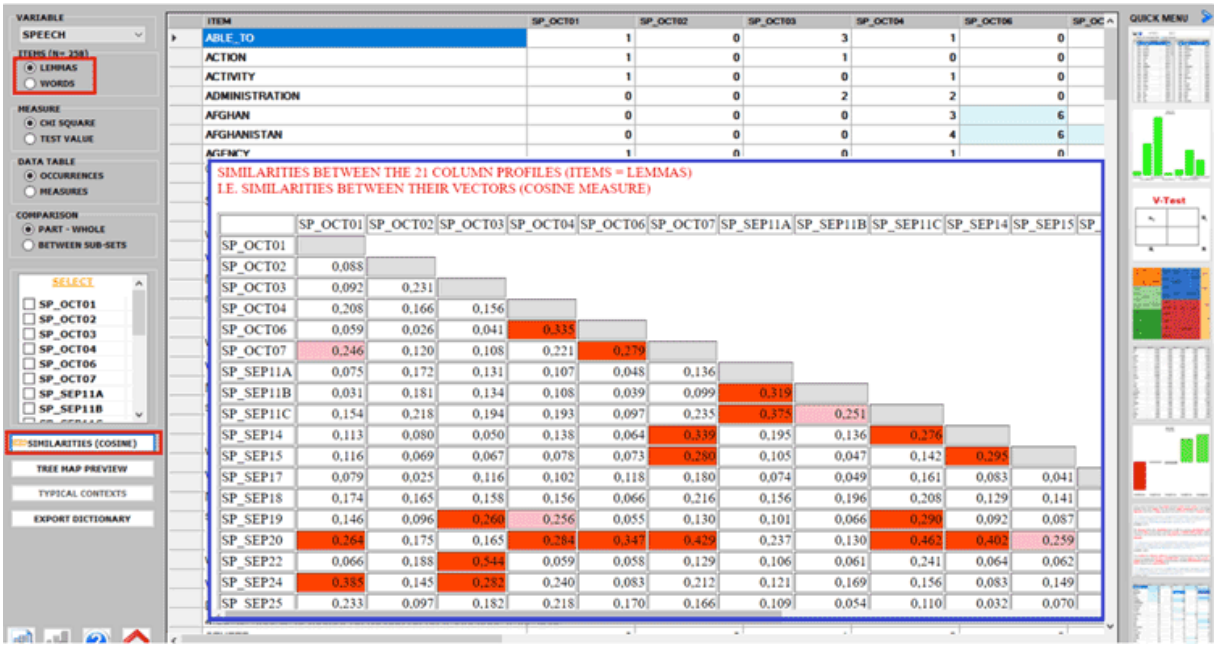
Topic Analysis
Real Time Charts
Similarity Measures
RQA Analysis
Sequence Analysis
GRAPH MAKER - CO-OCCURRENCES WITHIN THE CLUSTER N. <HEALTH>
ADD/REMOVE ITEMS TO BE USED
CLICK A PICTURE TO DISPLAY THE GRAPH
EXPORT DATA FILES FOR NETWORK ANALYSIS (up to 5.000 items can be selected at a time)
GRAPH MAKER
RENAME
IMPORT/EXPORT DICTIONARY

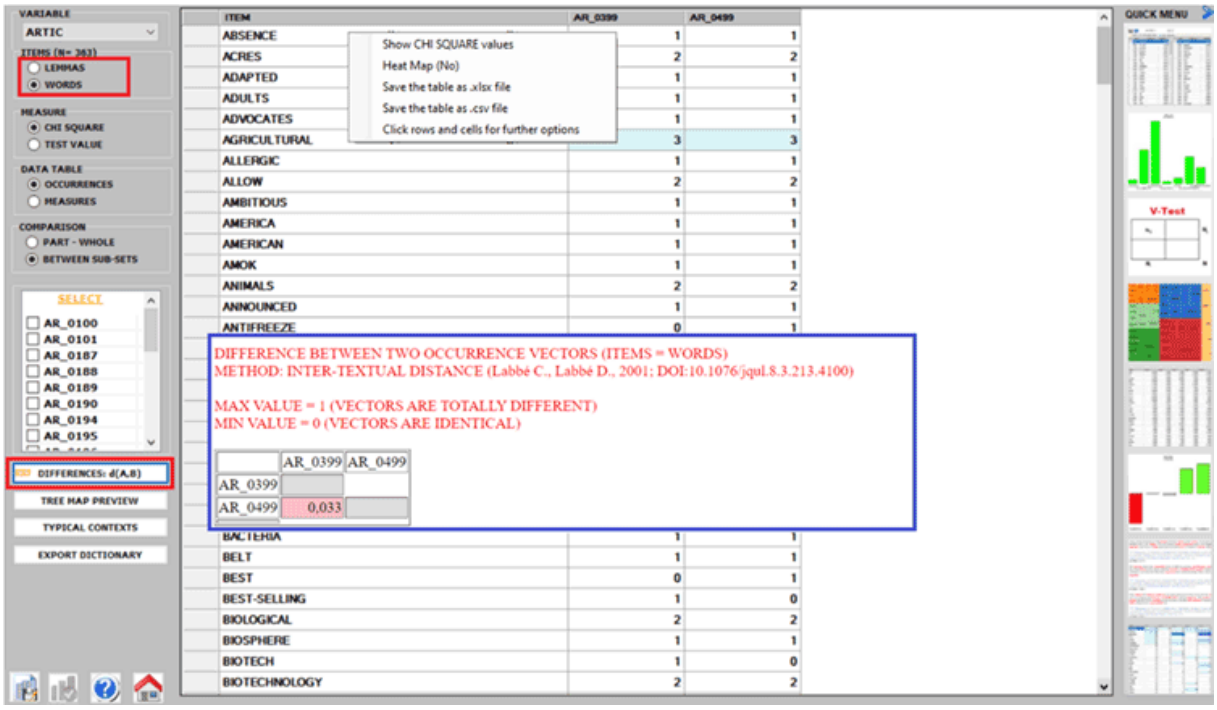

MDS MAP - Use the right click of the mouse
AFFORD
GOOD
COMPETITION
BIG
YEAR
LINES
HEALTH_CARE
PROBLEM
OBAMACARE
INSURANCE
WORK
MENTION
LINE
SYSTEM
PEOPLE
ALL_OVER
INSURANCE_COMPANIES
ANYBODY
HEALTH
PLAN
ANDERSON
STATE
DONALD
MEAN
DIFFERENT
AFFORDABLE_CARE
COST
COVERAGE
CUT
DRUG
EXPENSIVE
REPEAL
RID
HIGH
AGREE
START
ACTION
X - Axis
Y - Axis

COMPARATIVE ANALYSIS

Specificity Analysis

N.B.: The pictures shown in this section have been obtained by using a previous version of **T-LAB**. These pictures look slightly different in **T-LAB 10**. In particular, starting from the 2021 version, a quick access gallery of pictures which works as an **additional menu** allows one to switch between various outputs with a single click. Moreover the user is enabled to easily evaluate **similarities** (i.e. Cosine) and **differences** (i.e. Inter-Textual Distance) between corpus subsets (from 2 to 150), and so also to detect duplicate and near-duplicate documents (see pictures below).





This **T-LAB** tool enables us to check which lexical units (words, lemmas or categories) are **typical** or **exclusive** in a text or a corpus subset defined by a 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:

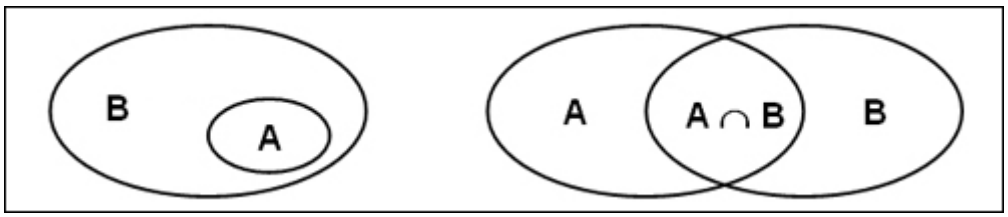
The 'typical' **lexical units**, defined for over-using or under-using, are detected by means of the **chi-square** or the **test value** computation.

The 'typical' **elementary contexts** are detected by computing and summing the **normalized TF-IDF values** assigned to the words which each sentence or paragraph consists of.

Specificity Analysis allows us to carry out two types of **comparisons**:

1- between a **part** (e.g. the subset "A") and the **whole** (e.g. the corpus under analysis, "B");

2- between couples of corpus **subsets** ("A" and "B").



In either instance Specificities involving both the **intersection** (tipical words) and the **differences** (exclusive words) can be analysed.

The computation modalities are shown in the corresponding **glossary** entry.

The considered lexical units can be all (see automatic settings) or only those selected by the user (see customized settings).

The four types of possible comparisons are as follows:

1.1 - **part/whole**: "typical" lexical units

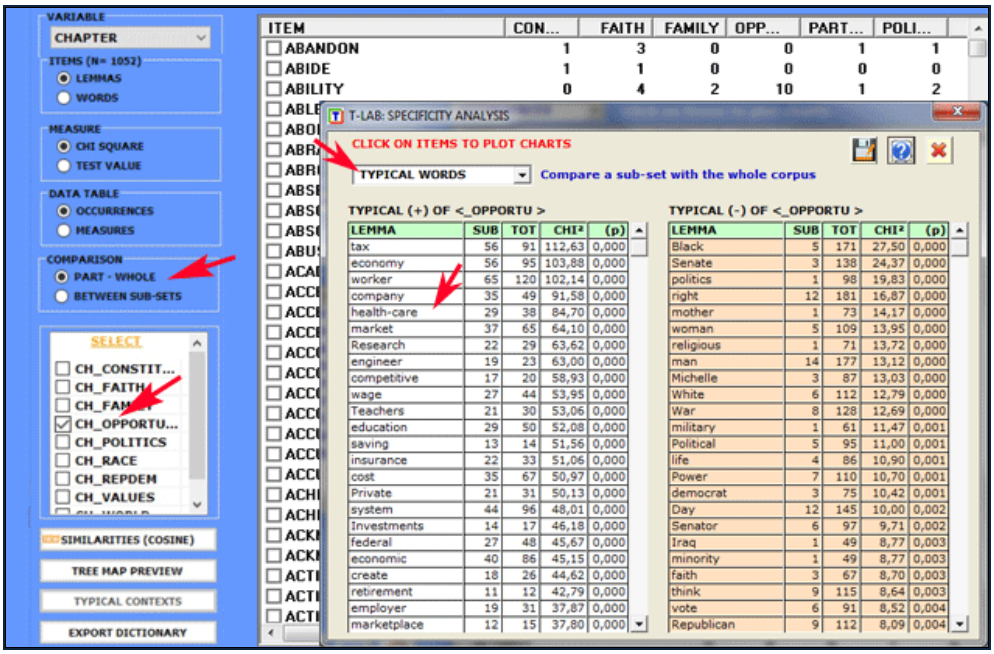
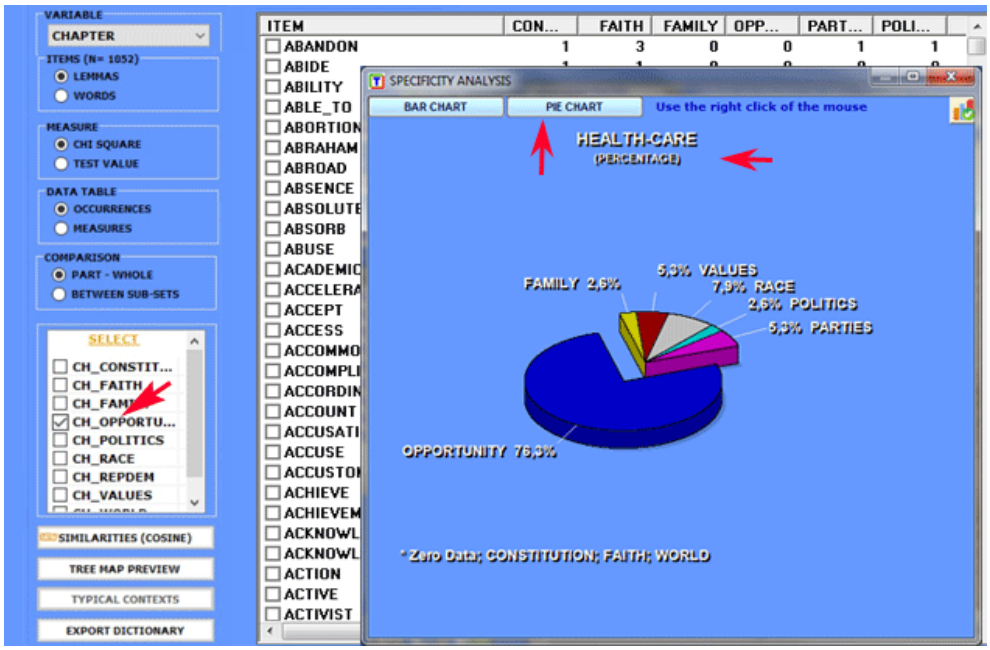
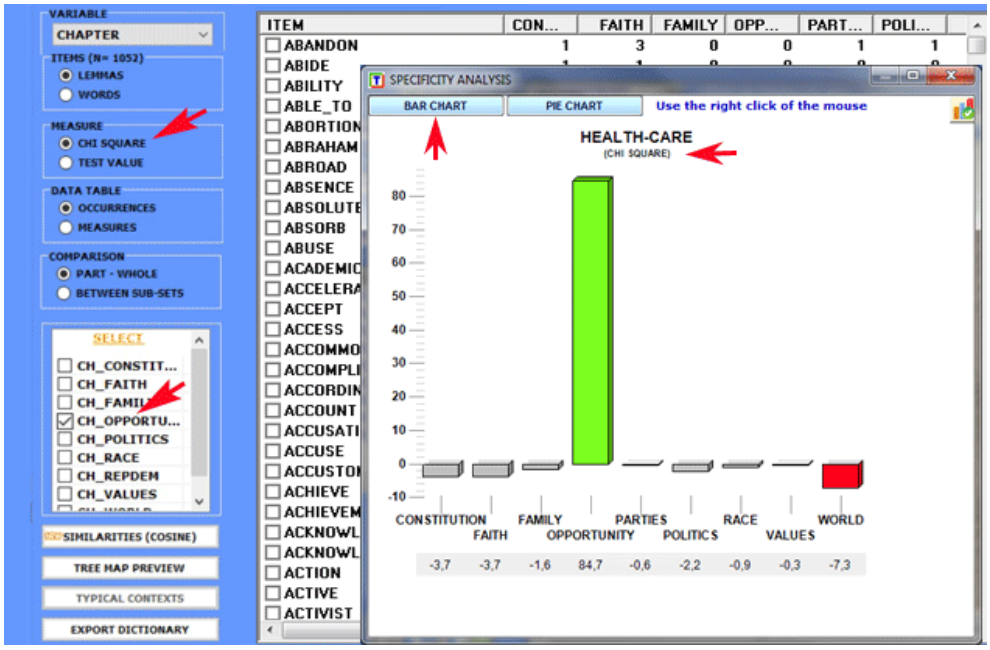


Table reading keys are as follows:

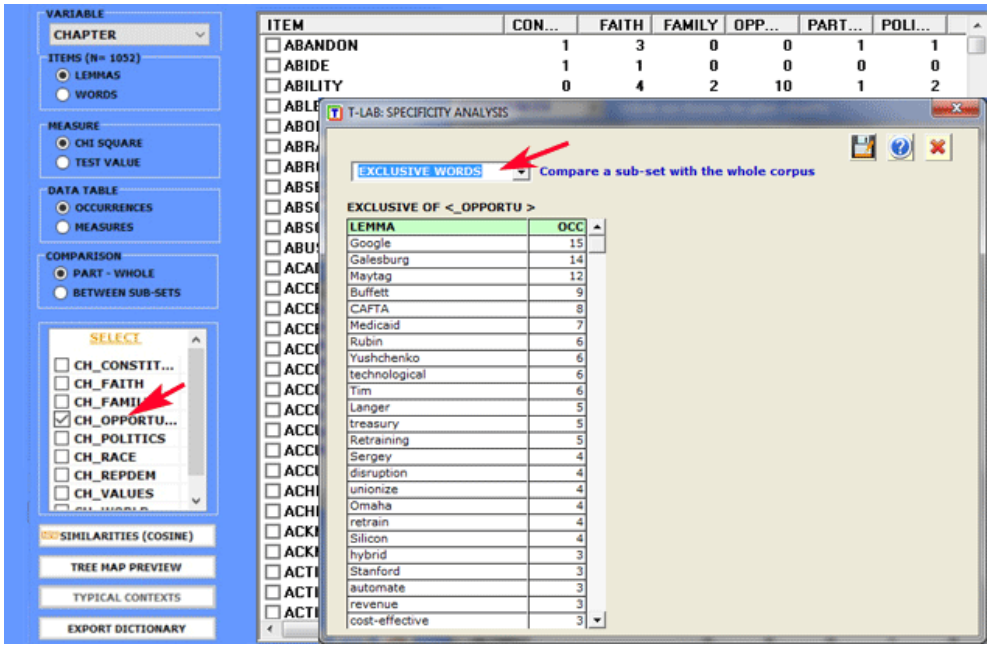
- LEMMA = specific lexical units (over/under used);
- SUB = occurrences of each LEMMA in the subset;
- TOT = occurrences of each LEMMA in the corpus or in the two compared subsets (see 2.1 below);
- CHI2 = CHI square value (or VTEST = Test Value);
- (p) = probability associated with the chi square value (def=1).

By clicking on the listed items it is possible to create various charts (see below).

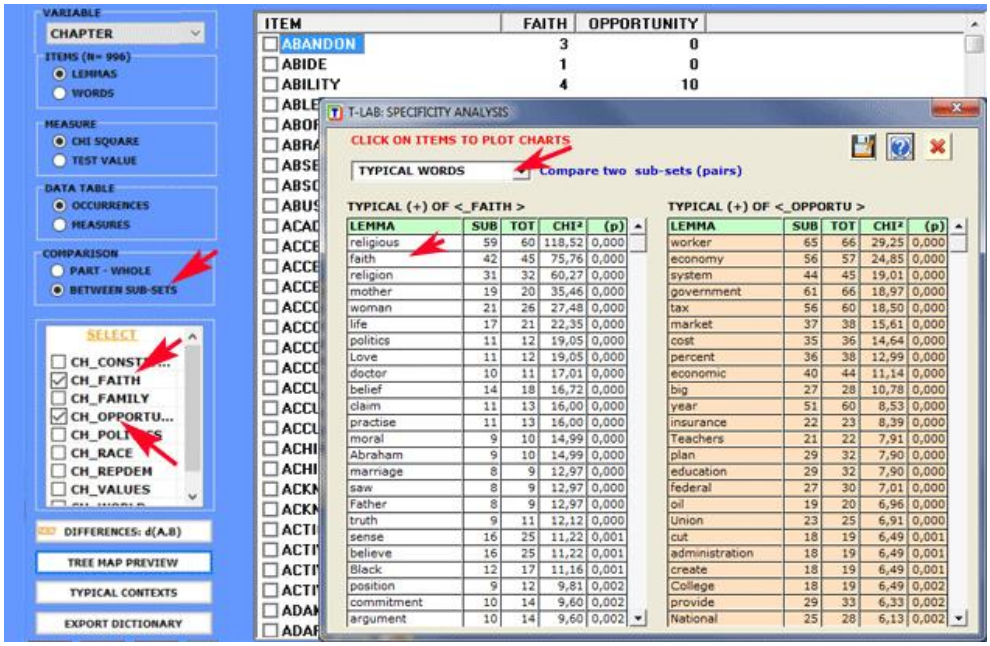

VARIABLE
CHAPTER
ITEMS (N= 1052)
LEMMAS
WORDS
MEASURE
CHI SQUARE
TEST VALUE
DATA TABLE
OCCURRENCES
MEASURES
COMPARISON
PART - WHOLE
BETWEEN SUB-SETS
SELECT
CH_CONSTIT...
CH_FAITH
CH_FAMI
CH_OPPORTU...
CH_POLITICS
CH_RACE
CH_REPDEM
CH_VALUES
SIMILARITIES (COSINE)
TREE MAP PREVIEW
TYPICAL CONTEXTS
EXPORT DICTIONARY
ITEM
CON...
FAITH
FAMILY
OPP...
PART...
POLI...
ABANDON
ABIDE
ABILITY
ABLE_TO
ABORTION
ABRAHAM
ABROAD
ABSENCE
ABSOLUTE
ABSORB
ABUSE
ACADEMIC
ACCELERA
ACCEPT
ACCESS
ACCOMMO
ACCOMPLI
ACCORDIN
ACCOUNT
ACCUSATI
ACCUSE
ACCUSTOI
ACHIEVE
ACHIEVEM
ACKNOWL
ACKNOWL
ACTION
ACTIVE
ACTIVIST
SPECIFICITY ANALYSIS
BAR CHART
PIE CHART
Use the right click of the mouse
HEALTH-CARE
(PERCENTAGE)
FAMILY 2,6%
5,3% VALUES
7,9% RACE
2,6% POLITICS
5,3% PARTIES
OPPORTUNITY 76,3%
* Zero Data: CONSTITUTION; FAITH; WORLD

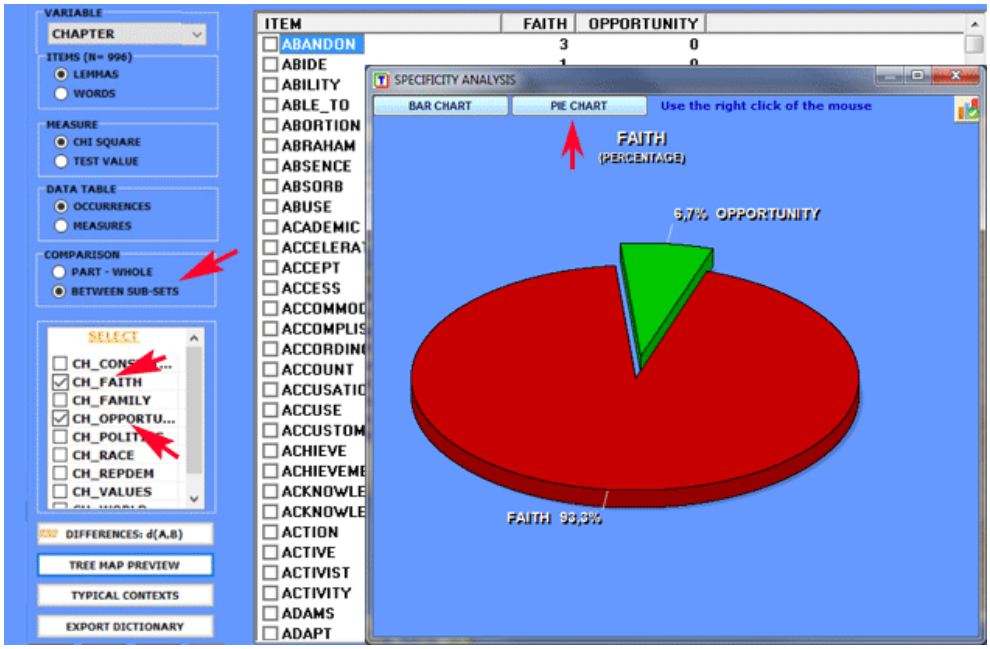

VARIABLE
CHAPTER
ITEMS (N= 1052)
LEMMAS
WORDS
MEASURE
CHI SQUARE
TEST VALUE
DATA TABLE
OCCURRENCES
MEASURES
COMPARISON
PART - WHOLE
BETWEEN SUB-SETS
SELECT
CH_CONSTIT...
CH_FAITH
CH_FAMI
CH_OPPORTU...
CH_POLITICS
CH_RACE
CH_REPDEM
CH_VALUES
SIMILARITIES (COSINE)
TREE MAP PREVIEW
TYPICAL CONTEXTS
EXPORT DICTIONARY
ITEM
CON...
FAITH
FAMILY
OPP...
PART...
POLI...
SPECIFICITY ANALYSIS
BAR CHART
PIE CHART
Use the right click of the mouse
HEALTH-CARE
(CHI SQUARE)
80
70
60
50
40
30
20
10
0
-10
CONSTITUTION
FAITH
FAMILY
OPPORTUNITY
PARTIES
POLITICS
RACE
VALUES
WORLD
-3,7
-3,7
-1,6
84,7
-0,6
-2,2
-0,9
-0,3
-7,3

1.2 - **part/whole**: "exclusive" lexical units

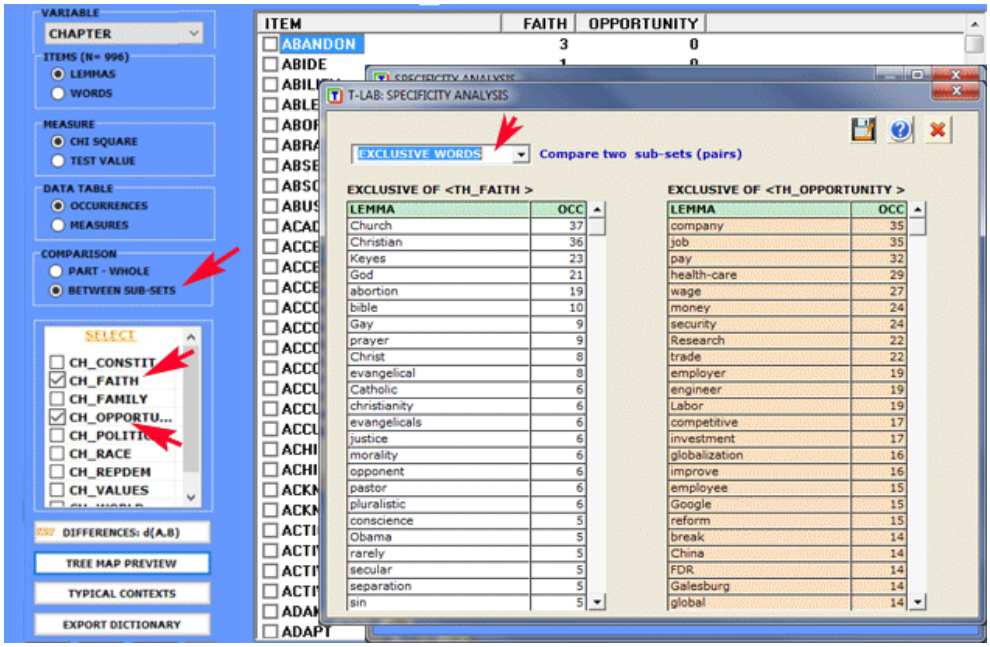


2.1 - **subset/subset**: "typical" lexical units

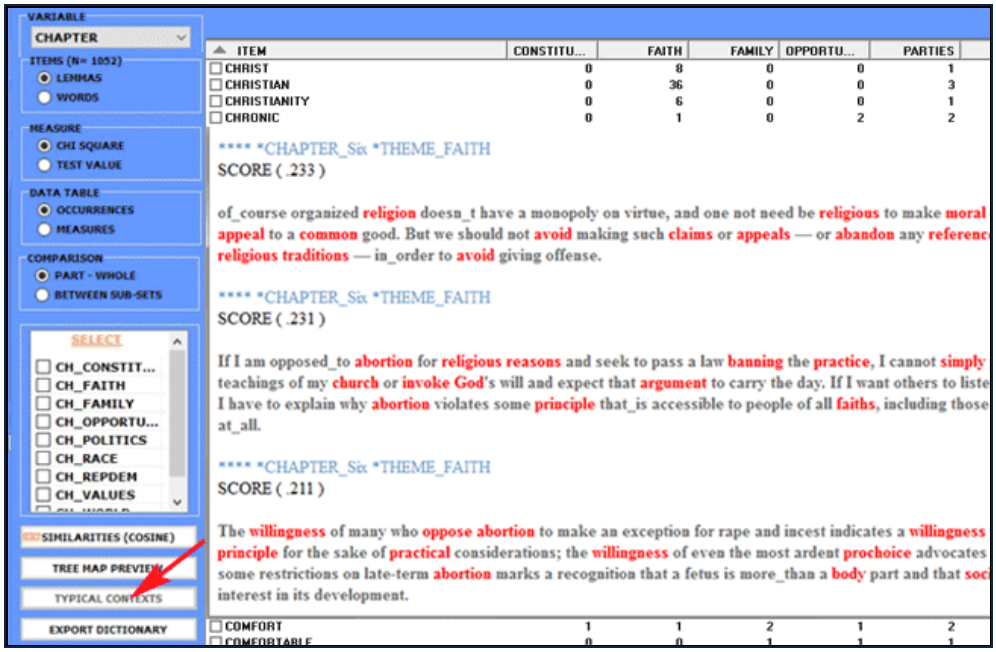




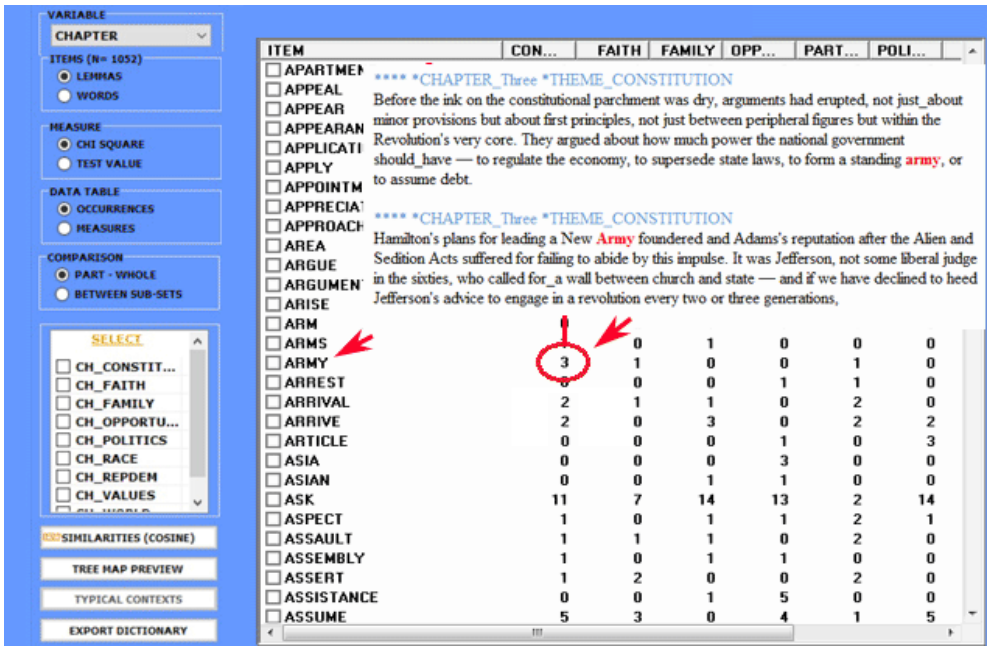
2.2- **subset/subset**: "exclusive" lexical units

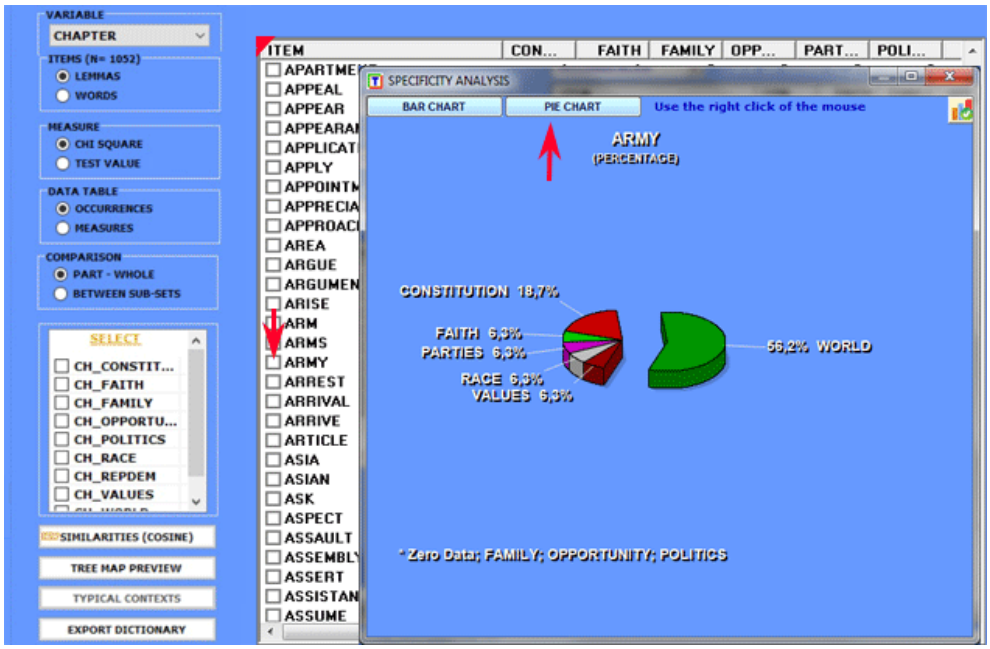


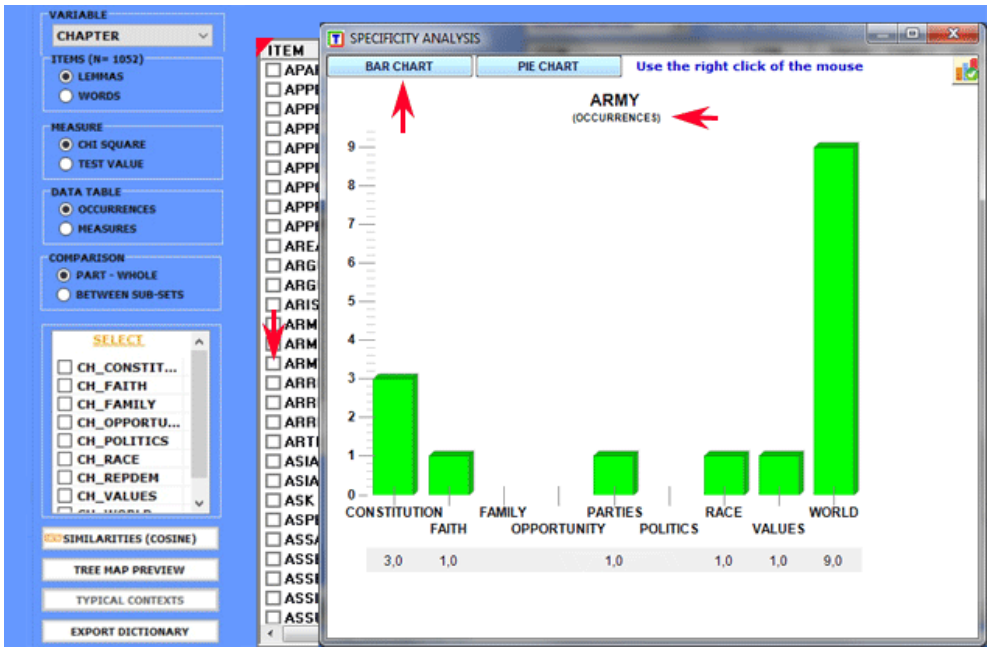
For each targeted subset it is also possible to check its 'typical' elementary contexts, the 'specificity' of which is a result of the computation of normalized **TF-IDF** values. More specifically, the 'score' assigned to each elementary context (see the picture below) results from the sum of **TF-IDF** values assigned to the words which it consists of.

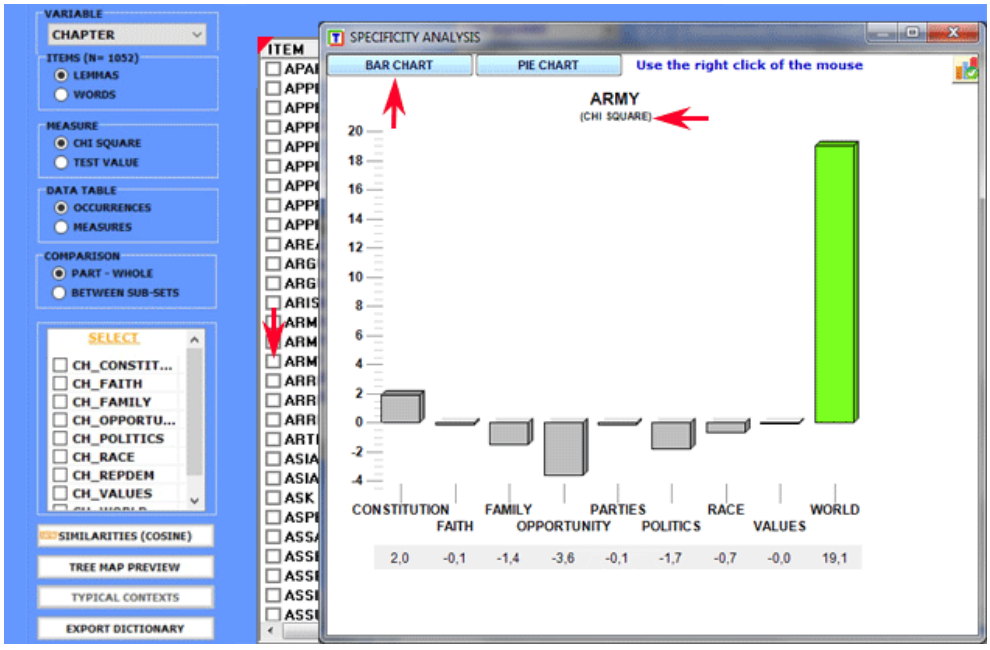


All contingency tables can be easily exported and allow us to create various charts. Moreover, by clicking on specific cells of the table (see below), it is possible to create a HTML file including all elementary contexts where the word in row is present in the corresponding subset.

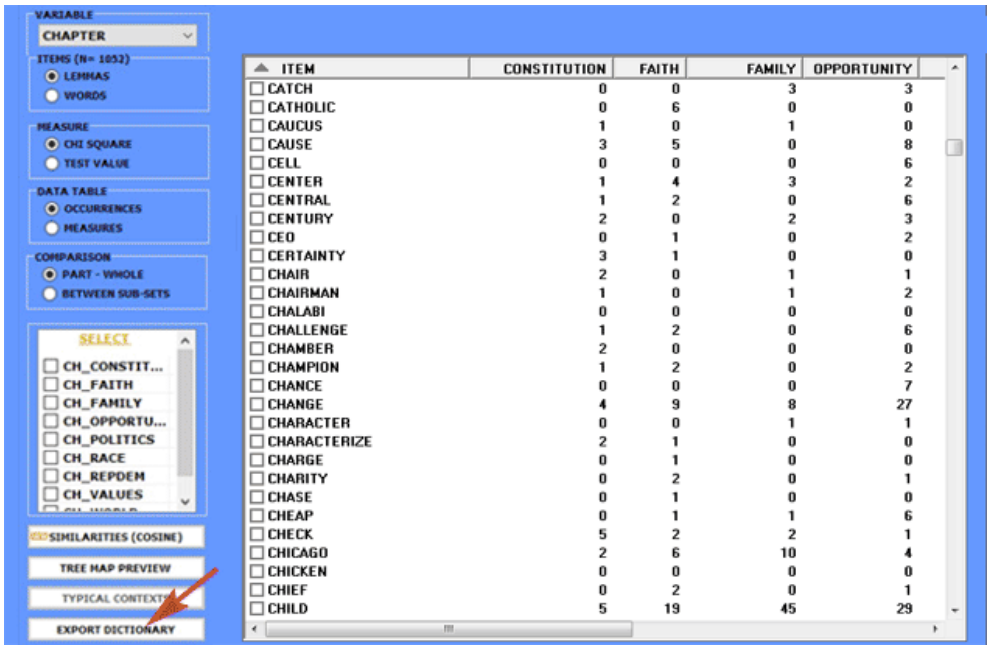



VARIABLE
CHAPTER
ITEMS (N= 1052)
LEMMAS
WORDS
MEASURE
CHI SQUARE
TEST VALUE
DATA TABLE
OCCURRENCES
MEASURES
COMPARISON
PART - WHOLE
BETWEEN SUB-SETS
SELECT
CH_CONSTIT...
CH_FAITH
CH_FAMILY
CH_OPPORTU...
CH_POLITICS
CH_RACE
CH_REPDEM
CH_VALUES
SIMILARITIES (COSINE)
TREE MAP PREVIEW
TYPICAL CONTEXTS
EXPORT DICTIONARY
ITEM
CON...
FAITH
FAMILY
OPP...
PART...
POLI...
SPECIFICITY ANALYSIS
BAR CHART
PIE CHART
Use the right click of the mouse
ARMY
(PERCENTAGE)
CONSTITUTION 18,7%
FAITH 6,3%
PARTIES 6,3%
RACE 6,3%
VALUES 6,3%
56,2% WORLD
*Zero Data; FAMILY; OPPORTUNITY; POLITICS


VARIABLE
CHAPTER
ITEMS (N= 1052)
LEMMAS
WORDS
MEASURE
CHI SQUARE
TEST VALUE
DATA TABLE
OCCURRENCES
MEASURES
COMPARISON
PART - WHOLE
BETWEEN SUB-SETS
SELECT
CH_CONSTIT...
CH_FAITH
CH_FAMILY
CH_OPPORTU...
CH_POLITICS
CH_RACE
CH_REPDEM
CH_VALUES
SIMILARITIES (COSINE)
TREE MAP PREVIEW
TYPICAL CONTEXTS
EXPORT DICTIONARY
ITEM
SPECIFICITY ANALYSIS
BAR CHART
PIE CHART
Use the right click of the mouse
ARMY
(OCCURRENCES)
9
8
7
6
5
4
3
2
1
0
CONSTITUTION
FAITH
FAMILY
OPPORTUNITY
PARTIES
POLITICS
RACE
VALUES
WORLD
3,0
1,0
1,0
1,0
1,0
9,0



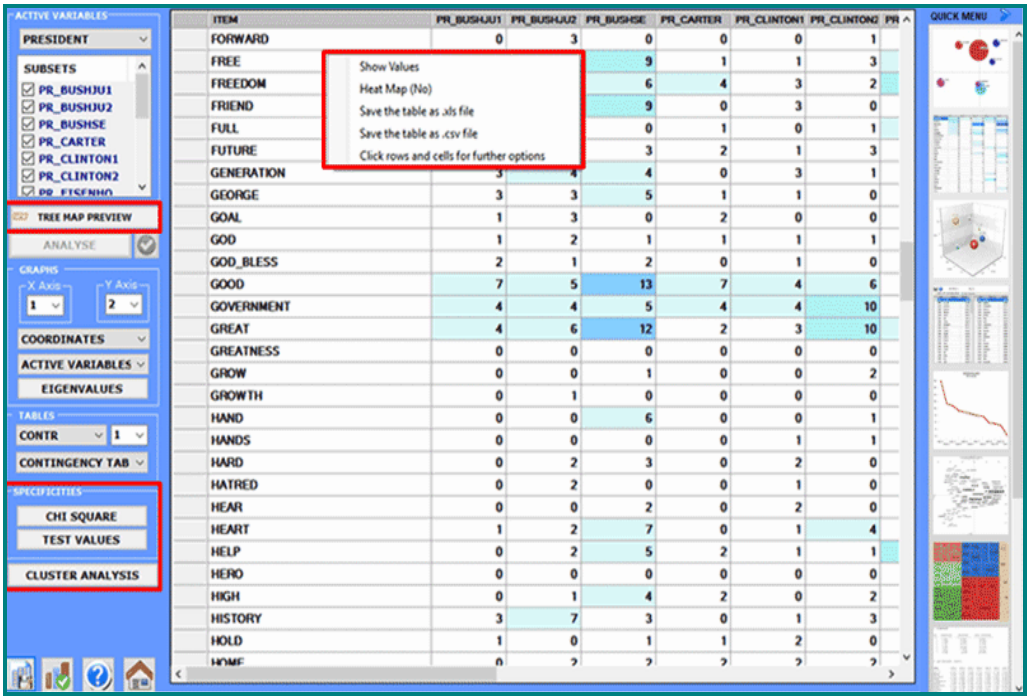
Eventually, by clicking the appropriate button (see below), a **dictionary** file with the .dictio extension is created which is ready to be imported by any **T-LAB** tool for **thematic analysis**. Such a dictionary includes all typical words of the selected categorical variable.



ITEM	CONSTITUTION	FAITH	FAMILY	OPPORTUNITY
CATCH	0	0	3	3
CATHOLIC	0	6	0	0
CAUCUS	1	0	1	0
CAUSE	3	5	0	8
CELL	0	0	0	6
CENTER	1	4	3	2
CENTRAL	1	2	0	6
CENTURY	2	0	2	3
CEO	0	1	0	2
CERTAINTY	3	1	0	0
CHAIR	2	0	1	1
CHAIRMAN	1	0	1	2
CHALABI	0	0	0	0
CHALLENGE	1	2	0	6
CHAMBER	2	0	0	0
CHAMPION	1	2	0	2
CHANCE	0	0	0	7
CHANGE	4	9	8	27
CHARACTER	0	0	1	1
CHARACTERIZE	2	1	0	0
CHARGE	0	1	0	0
CHARITY	0	2	0	1
CHASE	0	1	0	0
CHEAP	0	1	1	6
CHECK	5	2	2	1
CHICAGO	2	6	10	4
CHICKEN	0	0	0	0
CHIEF	0	2	0	1
CHILD	5	19	45	29

Correspondence Analysis

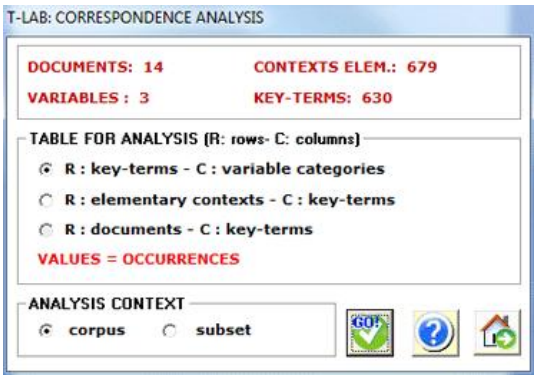
N.B.: The pictures shown in this section have been obtained by using a previous version of **T-LAB**. These pictures look slightly different in **T-LAB 10**. Moreover: a) by **right clicking** on the keyword tables, additional options become available; b) a new button (**TREE MAP PREVIEW**) which allows the user to create dynamic charts in HTML format; c) two new buttons allows us to check the **specificities** of each variable values either by using the **chi-square** test or the **test value**; d) a **new button** allows the user to carry out a **cluster analysis** that uses the coordinates of the objects (i.e. either lexical units or context units) on the first factorial axes (up to a maximum of 10); e) a quick access gallery of pictures which works as an **additional menu** allows one to switch between various outputs with a single click. Some of these new features are highlighted in the below image.



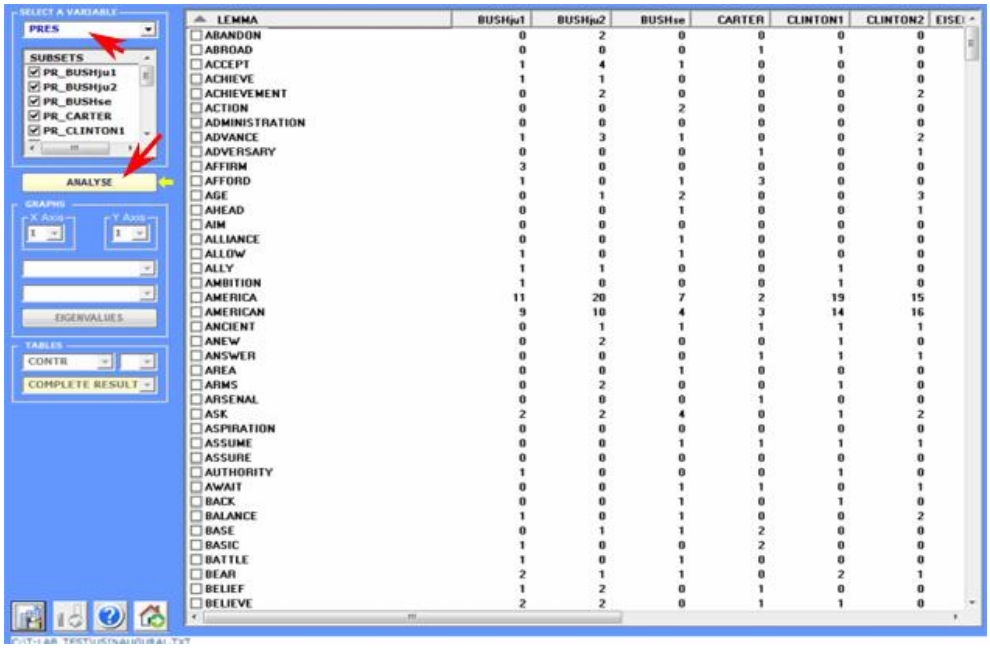
This **T-LAB** tool highlights the **similarities** and the **differences** between **context units**.

More precisely, in **T-LAB**, **correspondence analysis** can be applied to three kinds of tables:

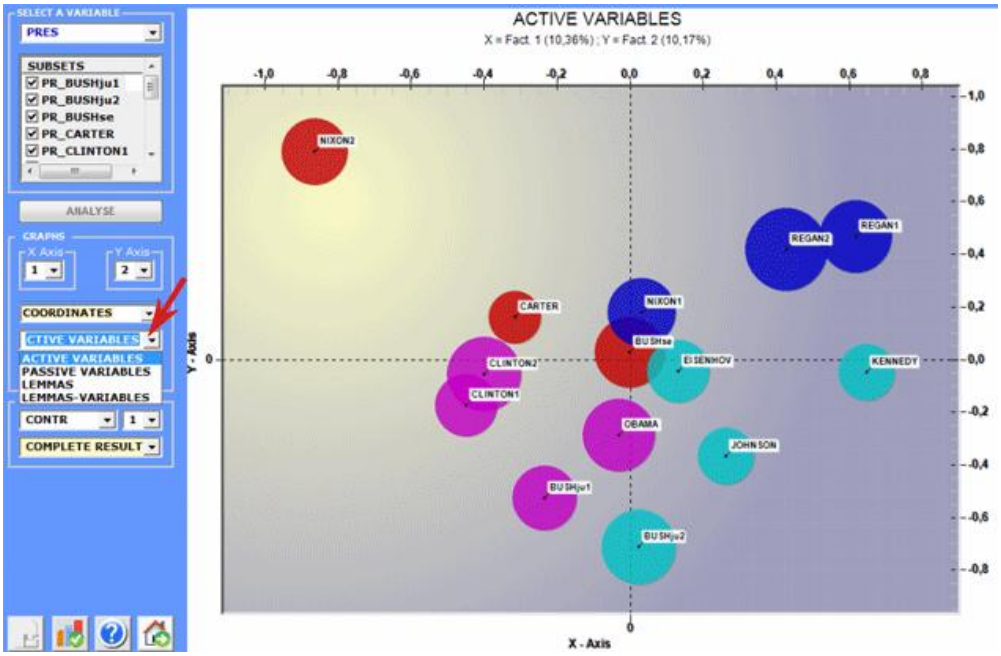
(**A**) tables of words by variables with **occurrence values**;
(**B**) tables of elementary contexts by words with **co-occurrence values**;
(**C**) tables of documents by words with **occurrence values**.



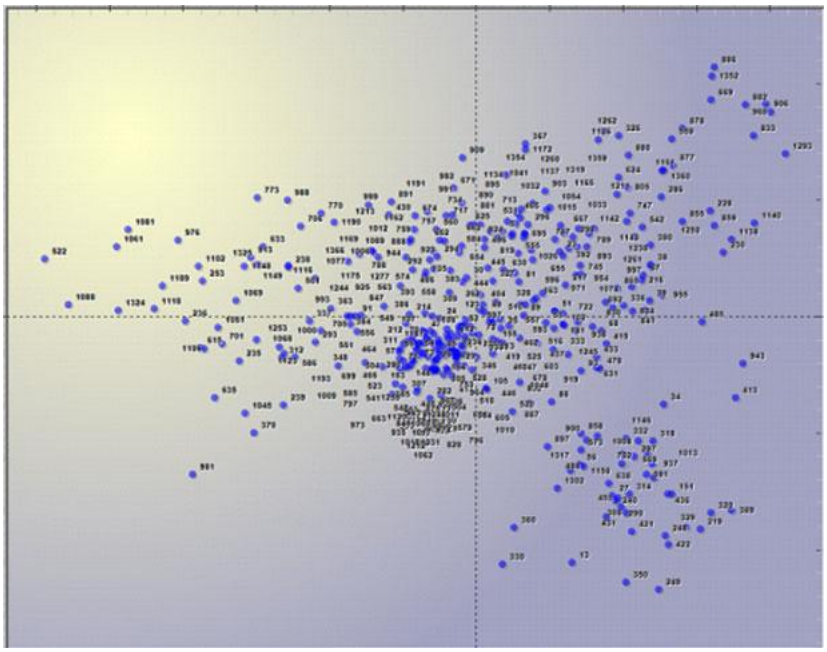
In order to analyse occurrence tables (**A**), the corpus should be made up of a minimum of three texts or should be codified with some **variables** (not less than three categories).
The variables are listed in an appropriate box and can be used one at a time.
After every selection - in sequence – the contingency table is dispalyed and **T-LAB** asks us to click on the **analyse** button (see below).



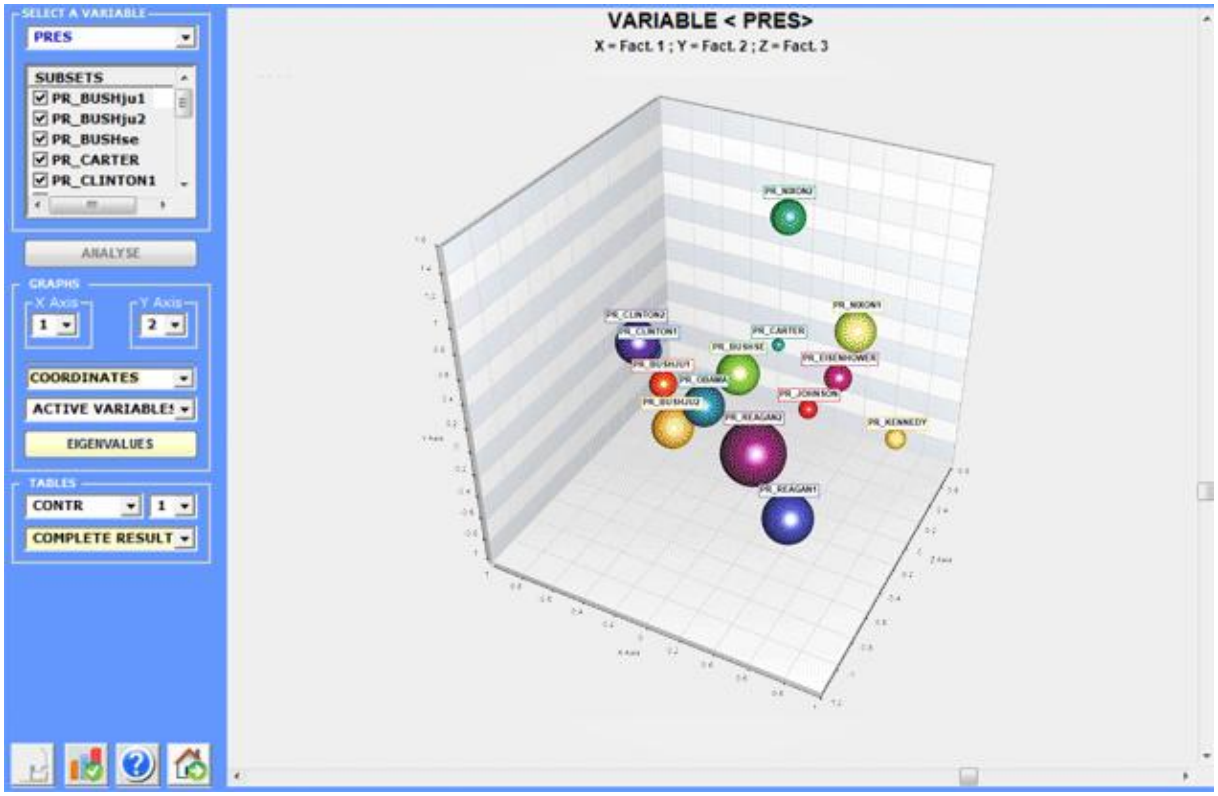
The analysis results allow the drawing of graphs in which the relationships between both the corpus subsets and the lexical units that make them up are represented.
More precisely, depending on the case, the types of graphs available show the relationships between **active variables**, between **illustrative variables**, between **lemmas** and between **lemmas and variables** (see below)**.**

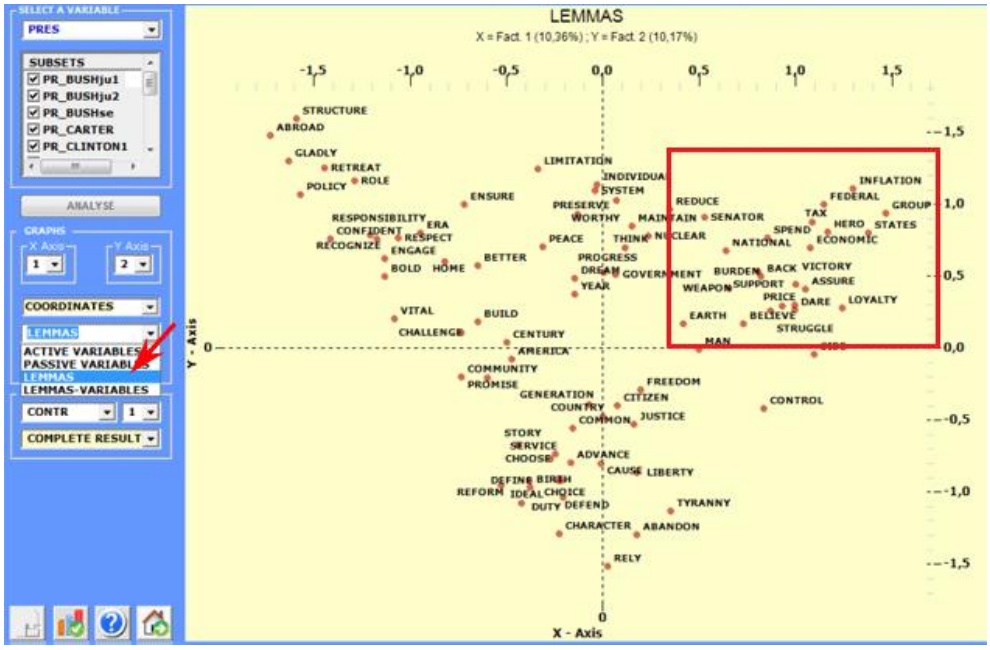


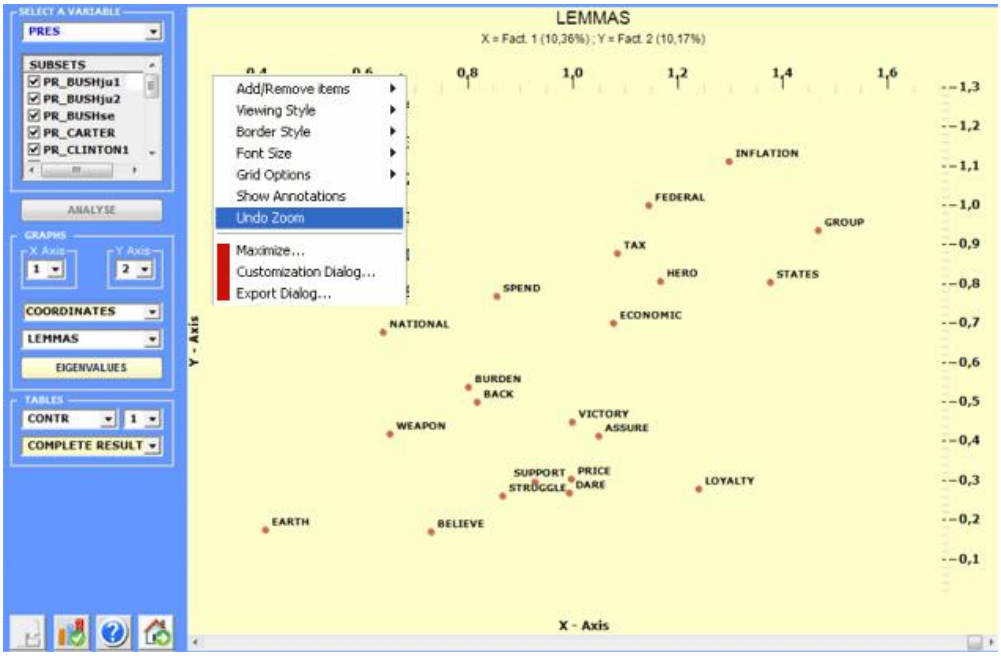
Moreover, when analysing a document by word table, it is possible to visualize the points (Max 3,000) corresponding to each document (see below).



All the graphs can be maximized and customized by using the appropriate dialog box (just right click on the chart). Moreover, when variable categories are 3 or more, their relationships can be explored through **3d** moving (see below).

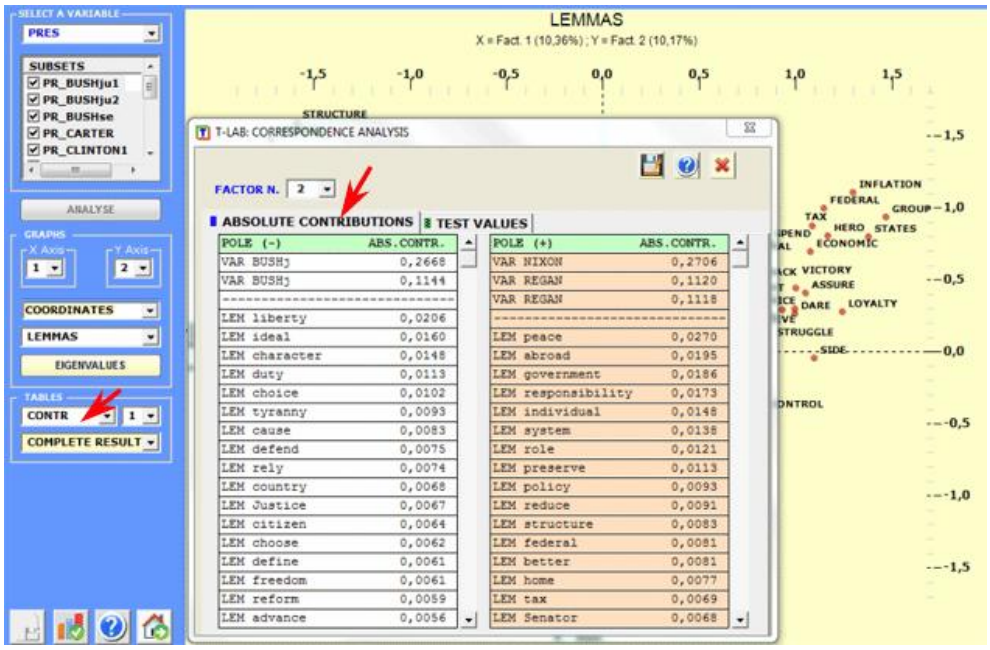

SELECT A VARIABLE
PRES
SUBSETS
PR_BUSHju1
PR_BUSHju2
PR_BUSHse
PR_CARTER
PR_CLINTON1
ANALYSE
GRAPHS
X Axis
Y Axis
COORDINATES
ACTIVE VARIABLES
EIGENVALUES
TABLES
CONTR
COMPLETE RESULT
VARIABLE < PRES>
X = Fact. 1 ; Y = Fact. 2 ; Z = Fact. 3


SELECT A VARIABLE
PRES
SUBSETS
PR_BUSHju1
PR_BUSHju2
PR_BUSHse
PR_CARTER
PR_CLINTON1
ANALYSE
GRAPHS
COORDINATES
LEMMAS
ACTIVE VARIABLES
PASSIVE VARIABLES
LEMMAS
LEMMAS-VARIABLES
CONTR
COMPLETE RESULT
LEMMAS
X = Fact. 1 (10,36%) ; Y = Fact. 2 (10,17%)
Y - Axis
X - Axis

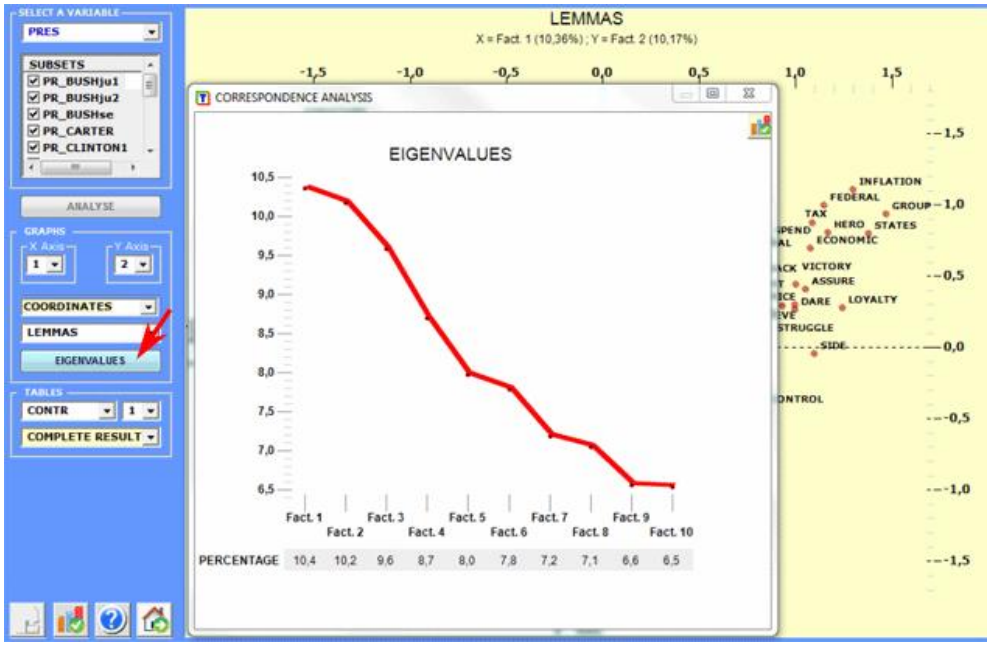


In order to explore the various combinations of the factorial axes it is sufficient to select them in the appropriate boxes ("X Axis", "Y Axis").

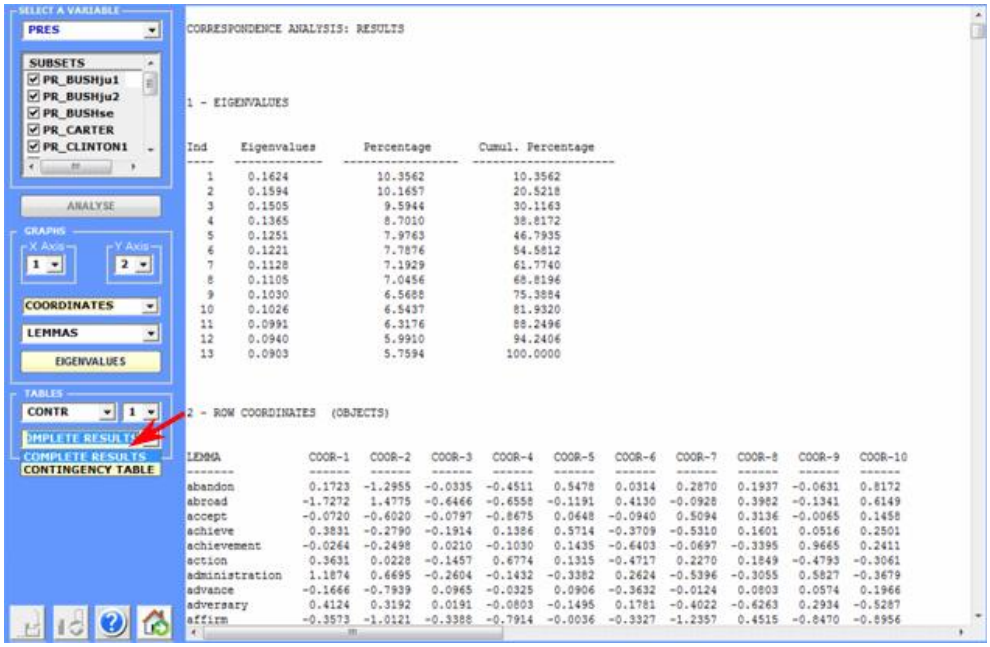
In **T-LAB** the characteristics of each **factorial pole** (i.e. the opposites on the horizontal and vertical axes) are shown using the **Absolute Contributions**, the threshold value of which is 1/N (in this case, N = rows of contingence tables), and the **Test Values**, the threshold value of which is +/- 1.96.



The eigenvalue chart enables the evaluation of the relative weight of each factor, that is the percentage of variance explained by each one of them.

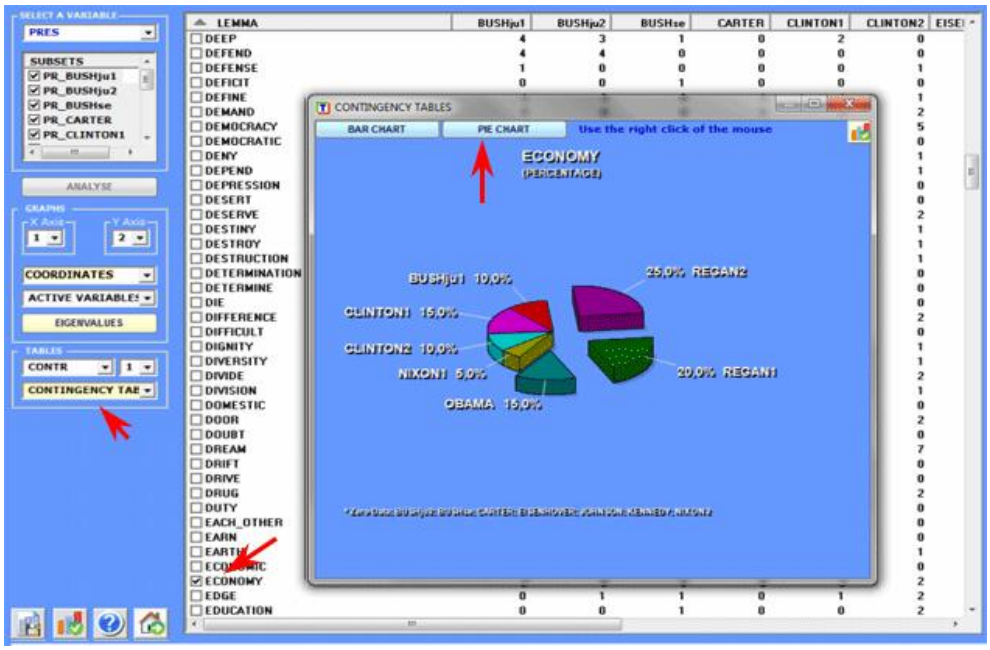


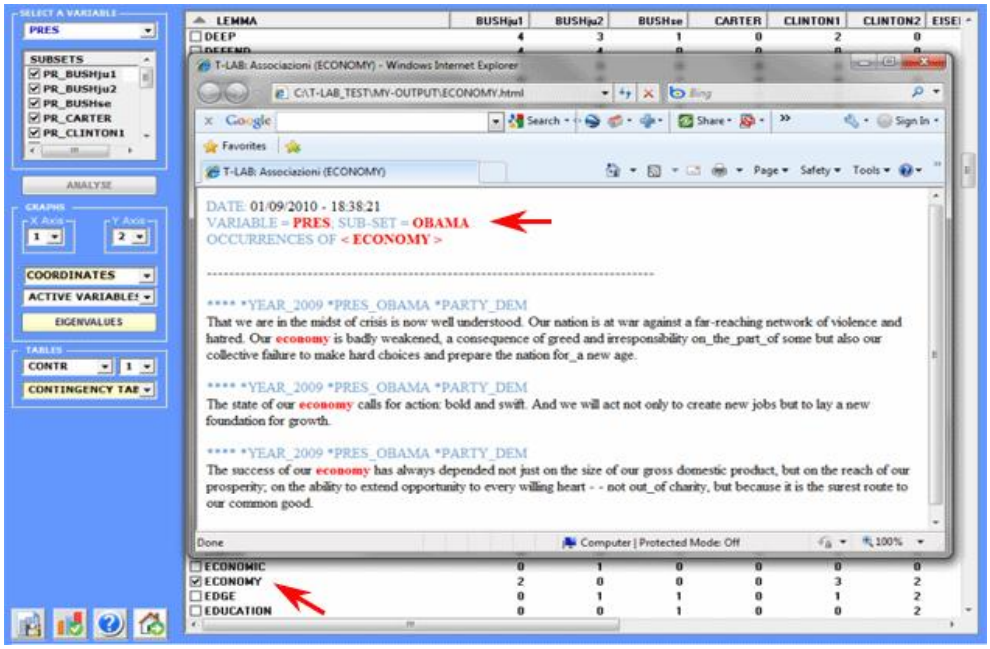
Finally, a click on the button "Complete Results" enables the user to visualize and export a file that contains the results of the analysis: eigenvalues, coordinates, absolute and relative contributions, and test values.



All contingency tables can be easily exported and allow us to create various charts.
Moreover, by clicking on specific cells of the table (see below), it is possible to create a HTML file including all elementary contexts where the word in row is present in the corresponding subset.

CONTINGENCY TABLES
BAR CHART
PIE CHART
Use the right click of the mouse
ECONOMY
(OCCURRENCES)
BUSHju1
BUSHju2
BUSHse
CARTER
CLINTON1
CLINTON2
EISENHOVER
JOHNSON
KENNEDY
NIXON1
NIXON2
OBAMA
REGAN1
REGAN2


CONTINGENCY TABLES
BAR CHART
PIE CHART
Use the right click of the mouse
ECONOMY
(PERCENTAGE)
BUSHju1 10,0%
25,0% REGAN2
CLINTON1 15,0%
CLINTON2 10,0%
NIXON1 5,0%
20,0% REGAN1
OBAMA 15,0%



In the case of the (**B**) or (**C**) tables (see above), they consist of as many rows as there are context units (max. 10,000) and as many columns as there are selected key words (max. 3,000).

The calculation algorithm and the outputs are similar to those of the analysis of lexical unit by variable tables, except that - in this case - in order to cut down processing time, **T-LAB** limits itself to the extraction of the first 10 factors, which is a more than sufficient number in order to summarize the variability of the data.

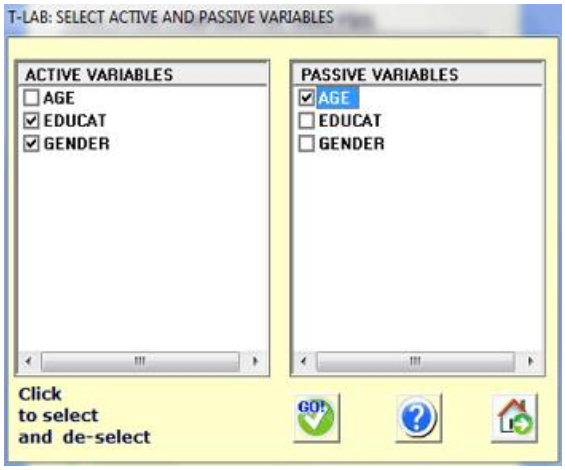
Moreover, subsequently it is possible to carry out a **Cluster Analysis**.

Multiple Correspondence Analysis

Multiple Correspondence Analysis, which may be considered an extension of the simple Correspondence Analysis (see above), allows us to analyse the relationships between two or more categorical variables.

In **T-LAB**, the limitations of this kind of analysis are the following:
- 150,000 elementary contexts as rows;
- 250 variable categories as columns;
- 3,000 key-words, as supplementary columns (Lebart L., Salem A., 1994)

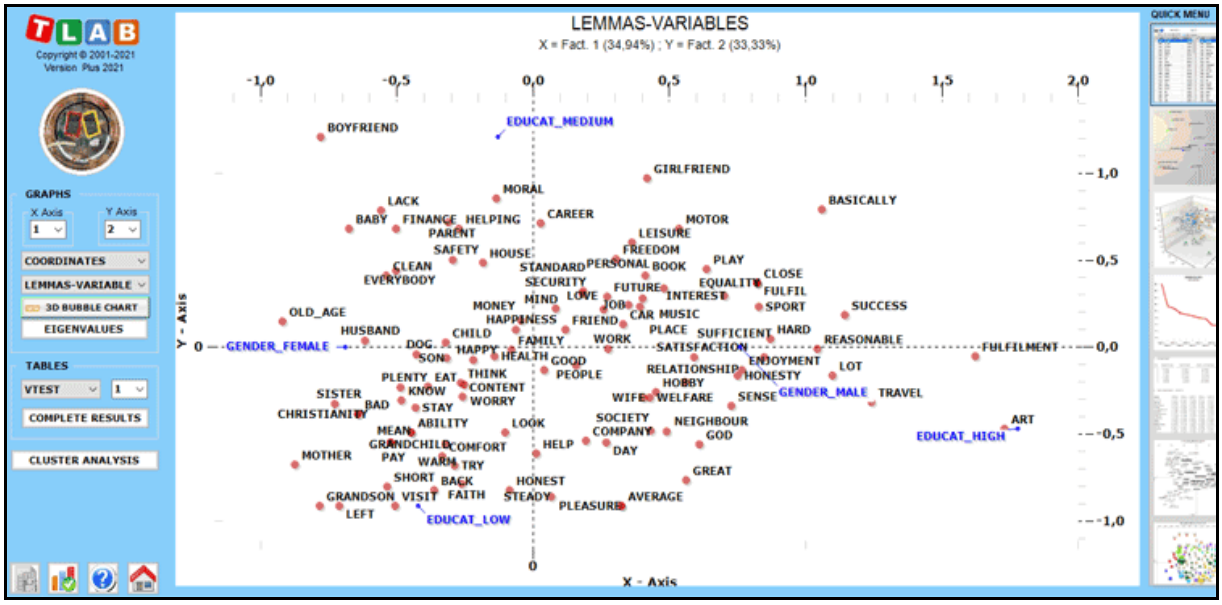
Multiple Correspondence Analysis, available only if the corpus includes at least two variables, requires that the user select his options within the following window:

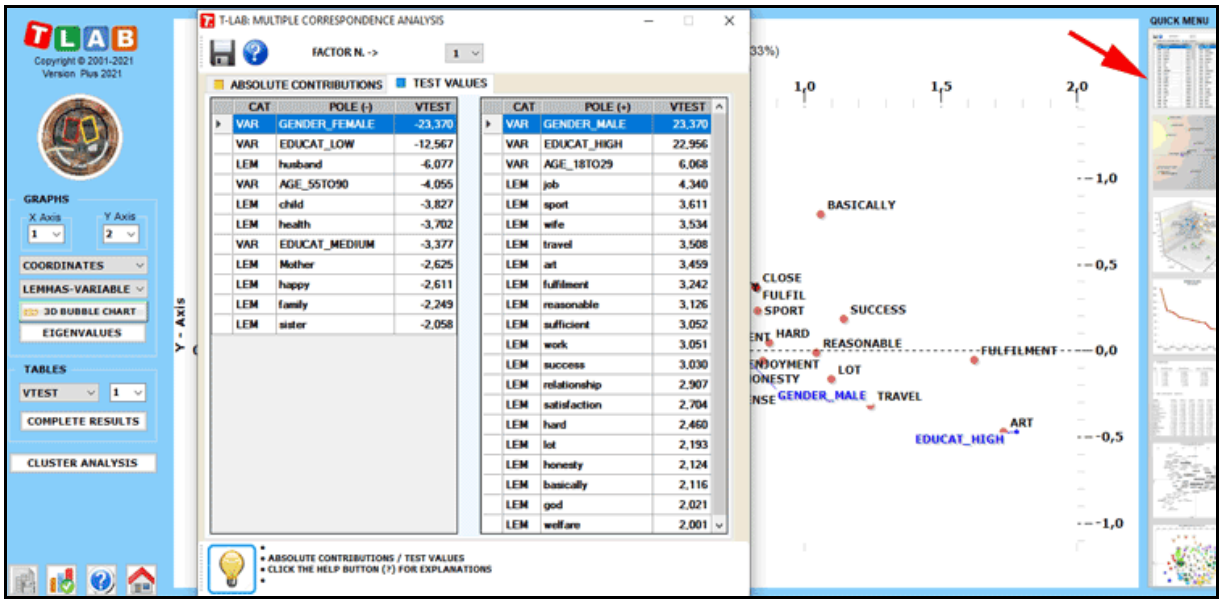


At the end of the analysis:

- **T-LAB** outputs are the same as correspondence analysis (see below) plus the Burt table (Burt_Table.xls) including all crossed variables;

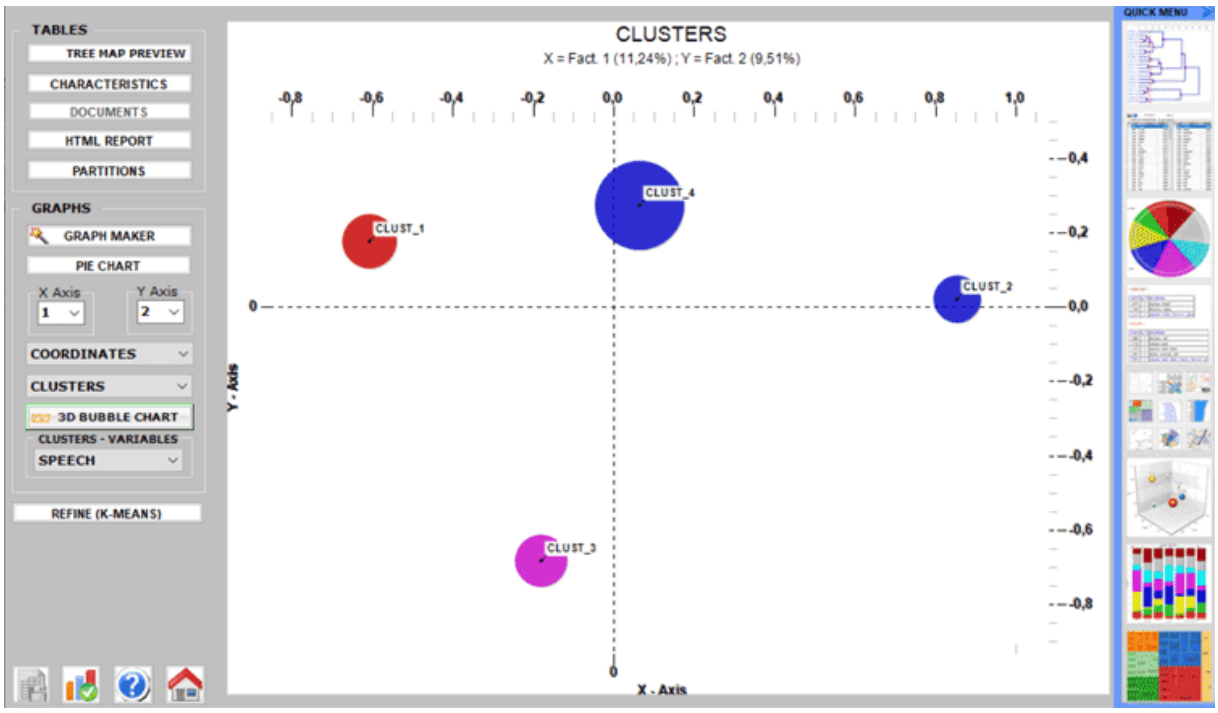
- only when the elementary contexts correspond to primary documents (e.g. responses to open-ended questions) it is possible to do a **cluster analysis**.


LEMMAS-VARIABLES
X = Fact. 1 (34,94%) ; Y = Fact. 2 (33,33%)
Copyright © 2001-2021
Version Plus 2021
GRAPHS
X Axis
Y Axis
COORDINATES
LEMMAS-VARIABLE
3D BUBBLE CHART
EIGENVALUES
TABLES
VTEST
COMPLETE RESULTS
CLUSTER ANALYSIS
QUICK MENU
Y - Axis
X - Axis
EDUCAT_MEDIUM
BOYFRIEND
GIRLFRIEND
BASICALLY
GENDER_FEMALE
GENDER_MALE
EDUCAT_HIGH
EDUCAT_LOW
FULFILMENT
SUCCESS
REASONABLE
TRAVEL
ART

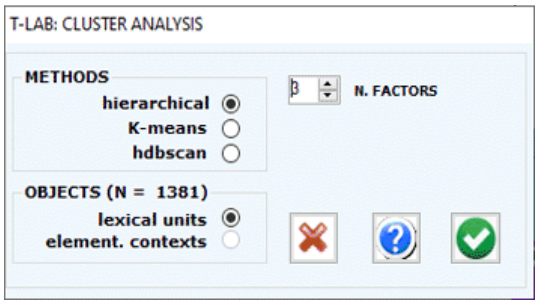

T-LAB: MULTIPLE CORRESPONDENCE ANALYSIS
FACTOR N. ->
ABSOLUTE CONTRIBUTIONS
TEST VALUES
CAT POLE (-) VTEST
VAR GENDER_FEMALE -23,370
VAR EDUCAT_LOW -12,567
LEM husband -6,077
VAR AGE_55TO90 -4,055
LEM child -3,827
LEM health -3,702
VAR EDUCAT_MEDIUM -3,377
LEM Mother -2,625
LEM happy -2,611
LEM family -2,249
LEM sister -2,058
CAT POLE (+) VTEST
VAR GENDER_MALE 23,370
VAR EDUCAT_HIGH 22,956
VAR AGE_18TO29 6,068
LEM job 4,340
LEM sport 3,611
LEM wife 3,534
LEM travel 3,508
LEM art 3,459
LEM fulfilment 3,242
LEM reasonable 3,126
LEM sufficient 3,052
LEM work 3,051
LEM success 3,030
LEM relationship 2,907
LEM satisfaction 2,704
LEM hard 2,460
LEM lot 2,193
LEM honesty 2,124
LEM basically 2,116
LEM god 2,021
LEM welfare 2,001
• ABSOLUTE CONTRIBUTIONS / TEST VALUES
• CLICK THE HELP BUTTON (?) FOR EXPLANATIONS

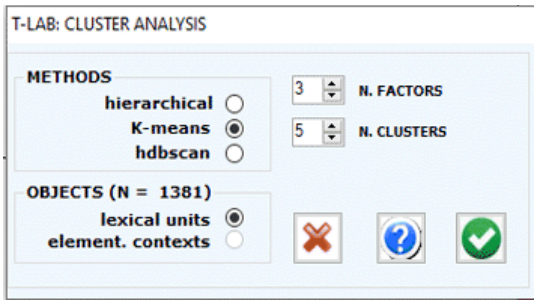
Cluster Analysis

N.B.: The pictures shown in this section have been obtained by using a previous version of **T-LAB**. These pictures look slightly different in **T-LAB 10**. Also: a) there is a new button (**TREE MAP PREVIEW**) which allows the user to create dynamic charts in HTML format; b) the DENDROGRAM button has been replaced by the **GRAPH MAKER** tool; c) a quick access gallery of pictures which works as an **additional menu** allows one to switch between various outputs with a single click (see the below image).



This **T-LAB** tool uses the results of a previous **Correspondence Analysis**; in particular, the computation uses the object coordinates (lexical units or context units) on the first factorial axes (until a maximum of 10).





Accordingly, the user can select from three clustering techniques:

a) **hierarchical** (Ward method);
b) **K-means** (MacQueen method);
c) **hdbscan** (hierarchical DBSCAN).

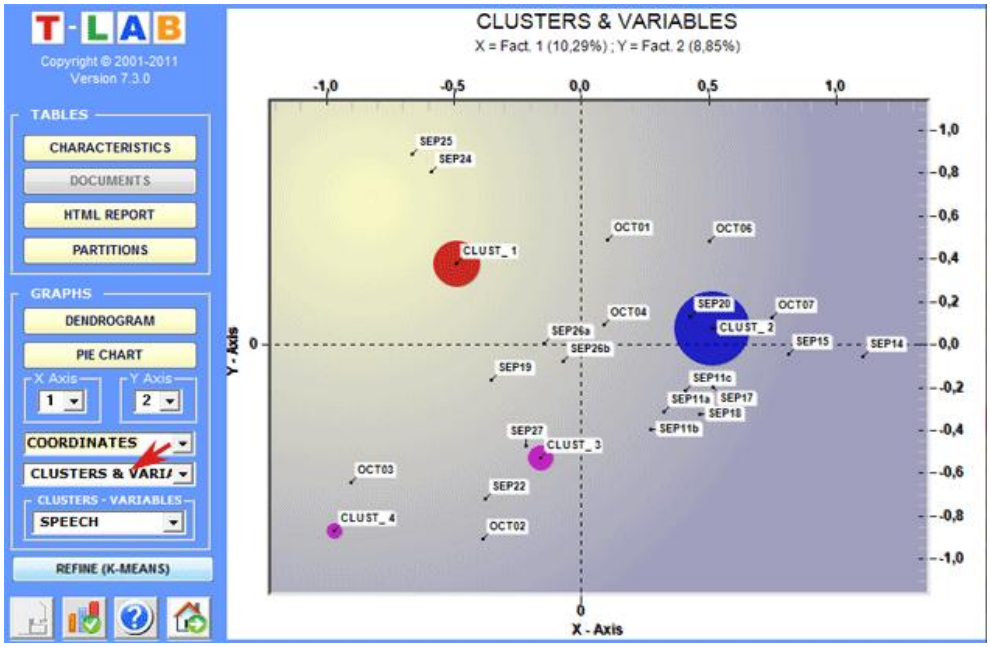
The first two (a, b) allow the user to explore (tables and graphs) solutions from **3** to **20** clusters; while the third (c), which requires an additional parameter (i.e. the minimum number of words within a cluster), allows the user to explore just one solution.

N.B.: When the hierarchical method is select **T-LAB** enables an option (see the 'Refine' button below) that allows the user to combine the Ward and K-Means methods.

A brief description of the three techniques is available in the **glossary** of this manual.

At the processing end, **T-LAB** shows graphs and tables.

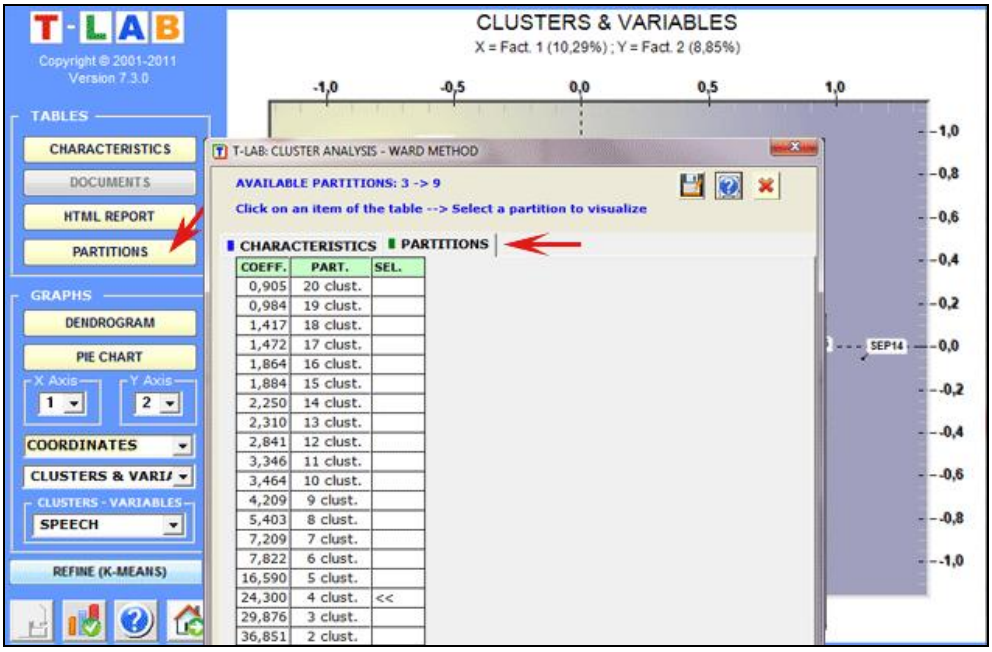
The graphs represent clusters in the space detected by the correspondence analysis (see below).



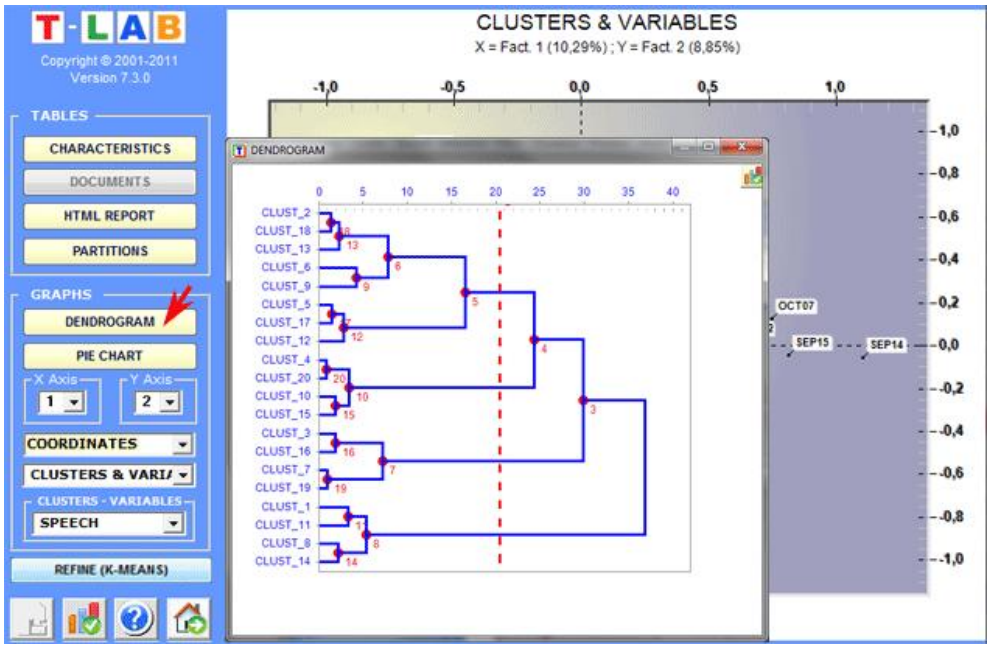
In order to explore the various combinations of the factorial axes it is sufficient to select them

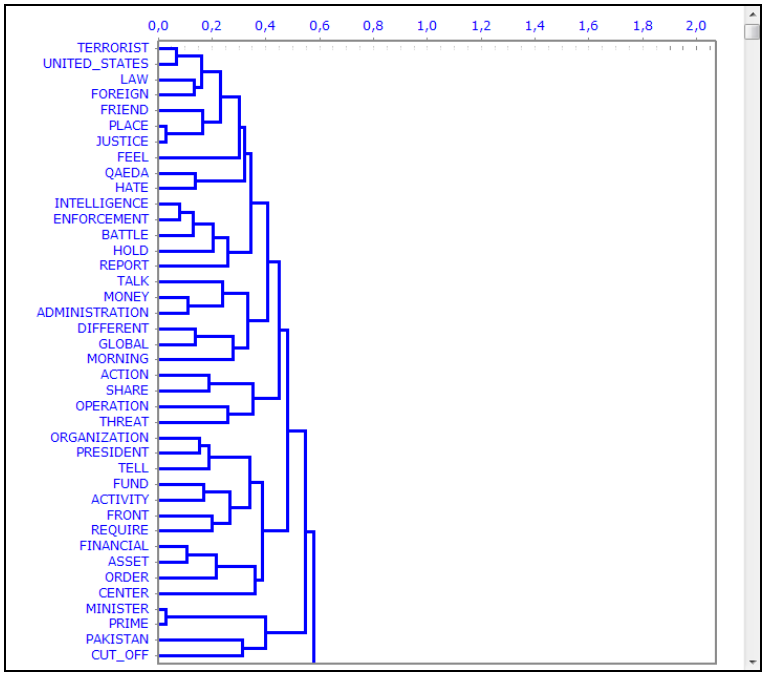
in the appropriate boxes ("X Axis", "Y Axis").

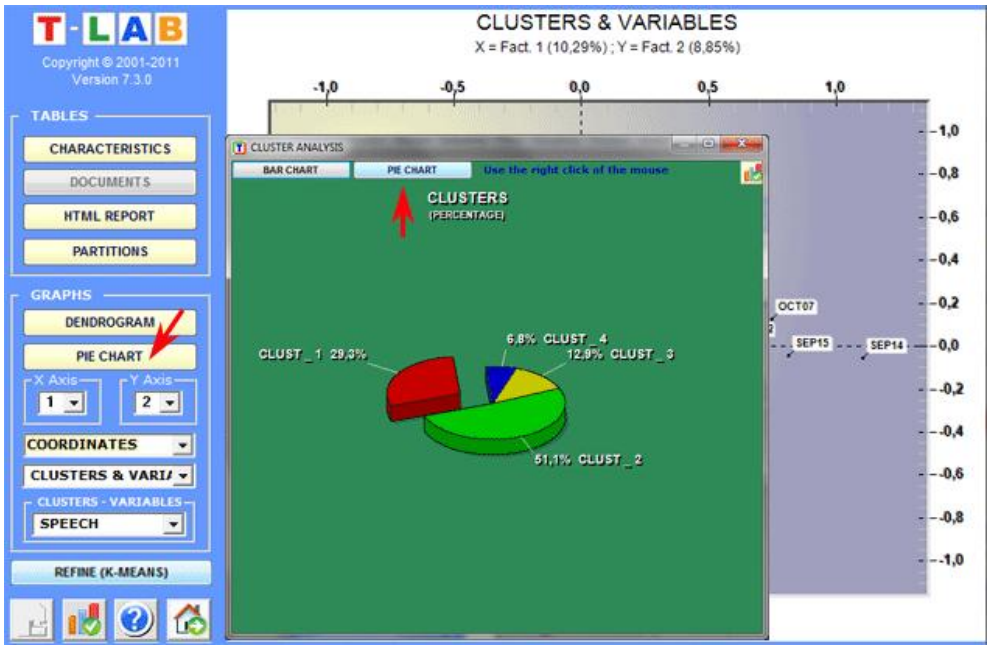
In the case of hierarchical clustering, the user can easily explore (graphs and tables) the different partitions.

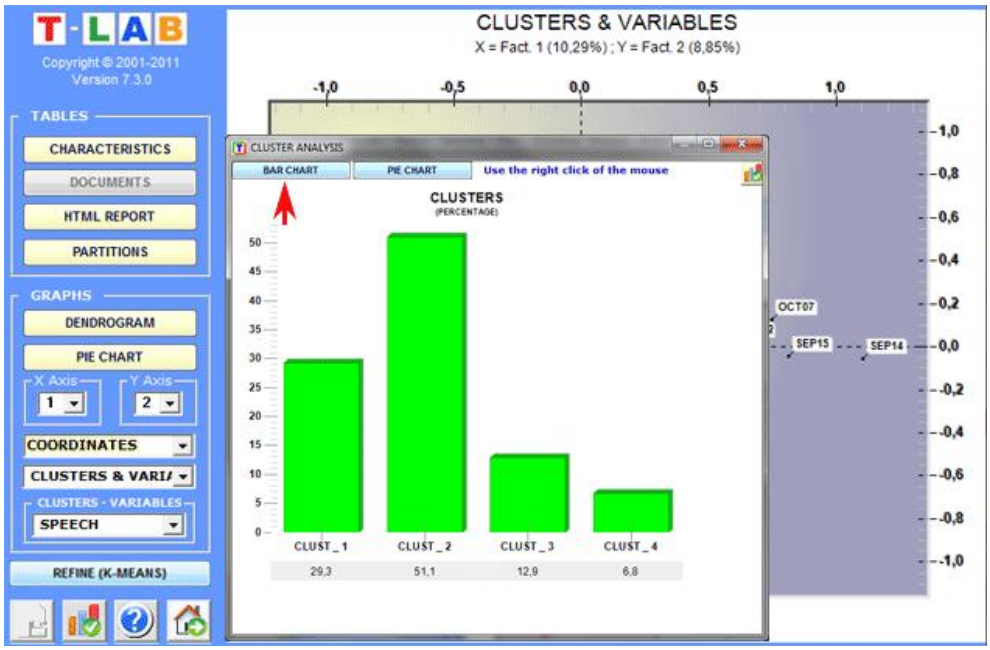


Dendrograms, pie charts and bar charts allow us to check the characteristics of each partition.

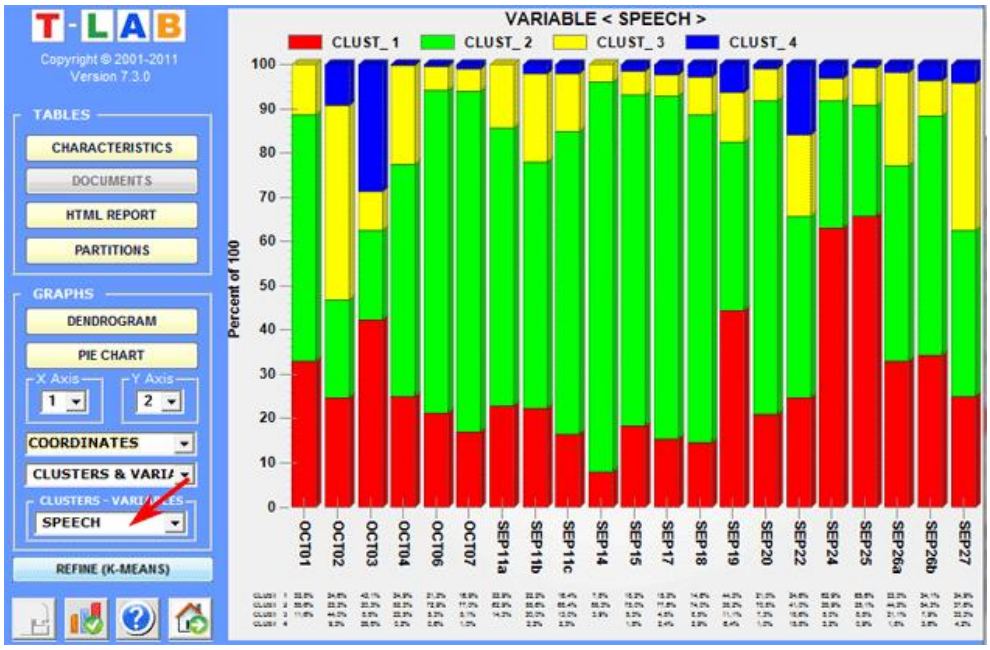



0,0
0,2
0,4
0,6
0,8
1,0
1,2
1,4
1,6
1,8
2,0
TERRORIST
UNITED_STATES
LAW
FOREIGN
FRIEND
PLACE
JUSTICE
FEEL
QAEDA
HATE
INTELLIGENCE
ENFORCEMENT
BATTLE
HOLD
REPORT
TALK
MONEY
ADMINISTRATION
DIFFERENT
GLOBAL
MORNING
ACTION
SHARE
OPERATION
THREAT
ORGANIZATION
PRESIDENT
TELL
FUND
ACTIVITY
FRONT
REQUIRE
FINANCIAL
ASSET
ORDER
CENTER
MINISTER
PRIME
PAKISTAN
CUT_OFF


T-LAB
Copyright © 2001-2011
Version 7.3.0
TABLES
CHARACTERISTICS
DOCUMENTS
HTML REPORT
PARTITIONS
GRAPHS
DENDROGRAM
PIE CHART
X Axis
Y Axis
1
2
COORDINATES
CLUSTERS & VARIA
CLUSTERS - VARIABLES
SPEECH
REFINE (K-MEANS)
CLUSTERS & VARIABLES
X = Fact. 1 (10,29%) ; Y = Fact. 2 (8,85%)
CLUSTER ANALYSIS
BAR CHART
PIE CHART
Use the right click of the mouse
CLUSTERS
(PERCENTAGE)
CLUST_1 29,3%
6,8% CLUST_4
12,9% CLUST_3
51,1% CLUST_2
OCT07
SEP15
SEP14

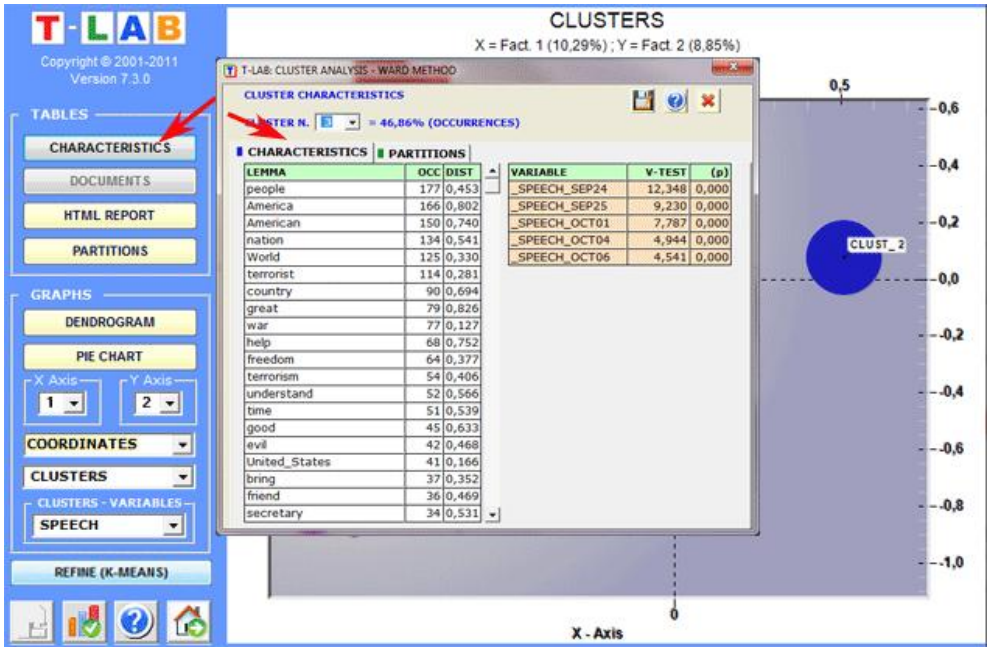


Bar charts allow us to check the relationships between clusters and variables.

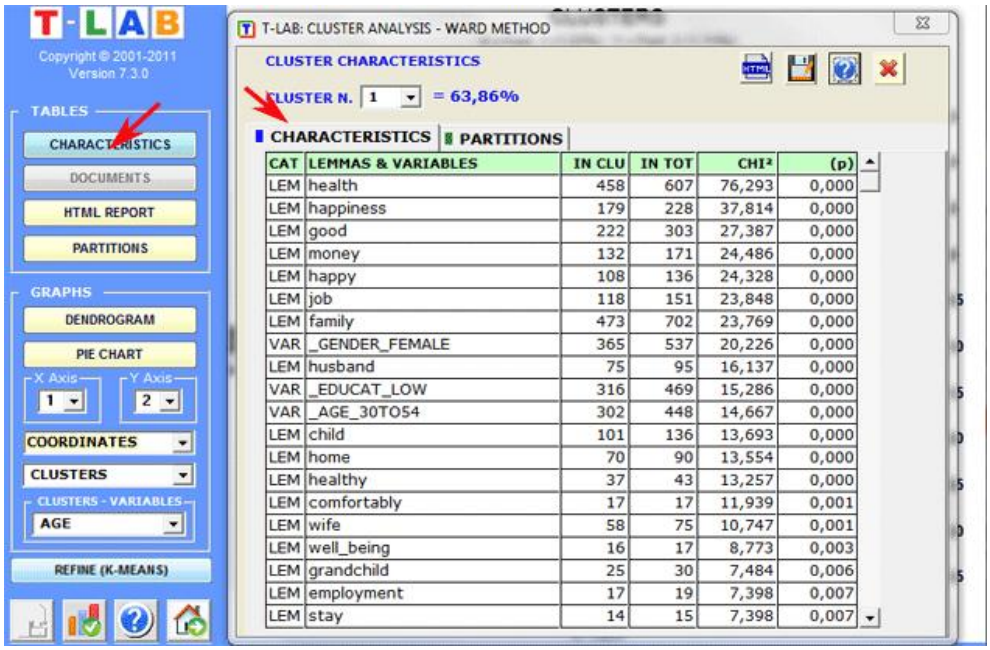


Two kinds of tables are available:

(**A**) if the clustered objects are lexical units, for each of them (and for each cluster) the respective occurrences ('OCC') and distances ('DIST') from the centroids are displayed are displayed; moreover, for each variable which is significantly associated with the cluster examined, the respective Test-Value is displayed.



(**B**) if the clustered objects are elementary contexts, the characteristics of each cluster (lexical units and variables) are described by means of the same method used in Thematic Analysis of Elementary Contexts. (see below).



In the case of analyses performed using the hierarchical or K-means methods, **T-LAB** allows the user to view and to export a file (see "HTML Output" key) in which the characteristics of the clusters and some measures relating to the quality of the partition are reported.

BETWEEN-CLUSTER VARIANCE (S2b) : 0,5498

WITHIN-CLUSTER VARIANCE (S2w) : 0,4578

CLUSTERS 1	0,0314
CLUSTERS 2	0,0734
CLUSTERS 3	0,1500
CLUSTERS 4	0,2030

S2b / (S2b + S2w) : 0,5457

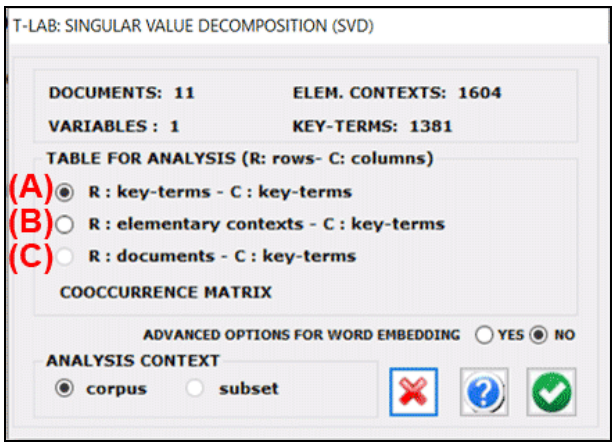
CENTROID COORDINATES

CLUSTERS 1	-0,2323	-0,7438	0,6919
CLUSTERS 2	0,7034	-0,1055	-0,2775
CLUSTERS 3	-0,5777	-0,1751	-0,5078

Singular Value Decomposition (SVD)

The **Singular Value Decomposition** (SVD) is a technique for dimensionality reduction, which - in Text Mining - can be used for discovering the **latent dimensions** (or components) which determine **semantic similarities** between words (i.e. lexical units) or between documents (i.e. context units).

T-LAB allows us to perform a Singular Value Decomposition of **three types of data tables**. In the first case (see 'A' below), the data table is a co-occurrence matrix whose rows and columns are key-terms. In the second case (see 'B' below), a data table elementary contexts x key-terms will be filled with presence/absence values (i.e. '1' and '0'). In the third case (see 'C' below), a data table documents x key-terms will be filled with occurrence values.
N.B.: Please note that, when analysing co-occurrence matrix whose rows and columns are key-terms (see 'A' below), T-LAB provides high-quality dense vectors (i.e. word embeddings).



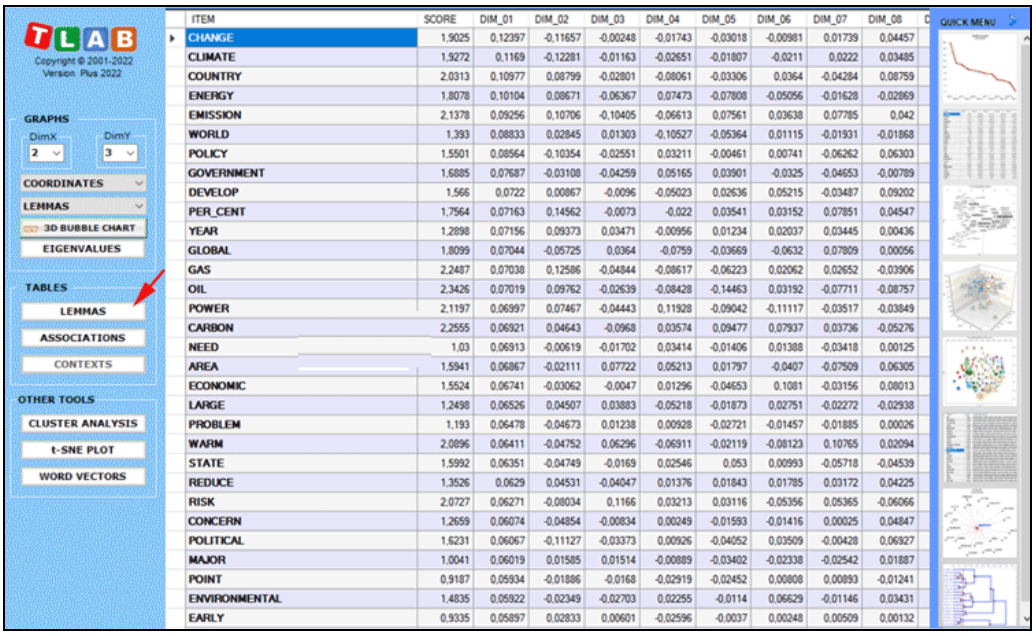
The analysis procedure consists of the following steps:
1 - construction of the data table to be analysed (up to 300,000 rows x 5,000 columns);
2 - TF-IDF normalization and scaling of row vectors to unit length (Euclidean norm);
3 - extraction of first 20 'latent dimensions' through the Lanczos algorithm.
N.B.:
-In the case of co-occurrence matrix (see ‘A’ above), data normalization is performed through the cosine measure;
-When the advanced options for word embedding are selected, T-LAB computes PPMI values (Positive Pointwise Mutual Information) and makes it possible to use the first 50 dimensions of the SVD.

The analysis results are displayed in **tables** and **charts**.

In detail:

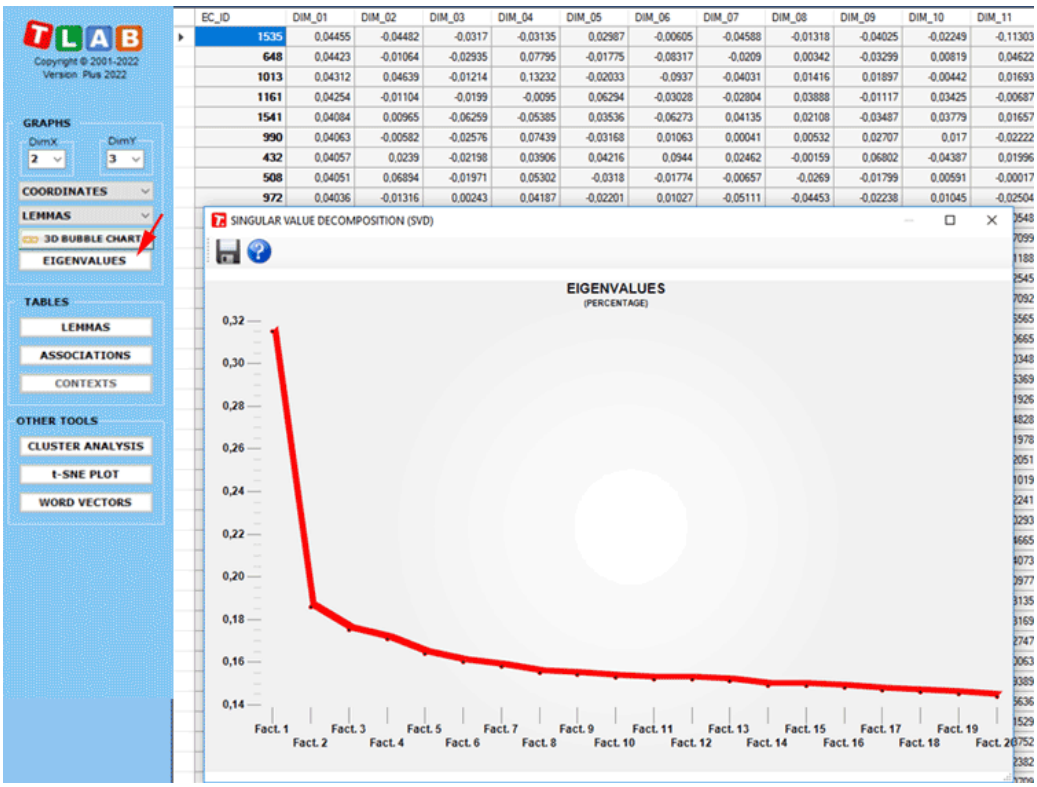
Two tables - the rows of which can be either lexical units or context units - have as many columns as the extracted dimensions.

In the case of the LEMMAS (i.e. lexical units) table, a further column is displayed, in which the importance scores are reported (see below).



ITEM	SCORE	DIM_01	DIM_02	DIM_03	DIM_04	DIM_05	DIM_06	DIM_07	DIM_08
CHANGE	1,9025	0,12397	-0,11657	-0,00248	-0,01743	-0,03018	-0,00981	0,01739	0,04457
CLIMATE	1,9272	0,1169	-0,12281	-0,01163	-0,02651	-0,01807	-0,0211	0,0222	0,03485
COUNTRY	2,0313	0,10977	0,08799	-0,02801	-0,08061	-0,03306	0,0364	-0,04284	0,08759
ENERGY	1,8078	0,10104	0,08671	-0,06367	0,07473	-0,07808	-0,05056	-0,01628	-0,02869
EMISSION	2,1378	0,09256	0,10706	-0,10405	-0,06613	0,07561	0,03638	0,07785	0,042
WORLD	1,393	0,08833	0,02845	0,01303	-0,10527	-0,05364	0,01115	-0,01931	-0,01868
POLICY	1,5501	0,08564	-0,10354	-0,02551	0,03211	-0,00461	0,00741	-0,06262	0,06303
GOVERNMENT	1,6885	0,07687	-0,03108	-0,04259	0,05165	0,03901	-0,0325	-0,04653	-0,00789
DEVELOP	1,566	0,0722	0,00867	-0,0096	-0,05023	0,02636	0,05215	-0,03487	0,09202
PER_CENT	1,7564	0,07163	0,14562	-0,0073	-0,022	0,03541	0,03152	0,07851	0,04547
YEAR	1,2898	0,07156	0,09373	0,03471	-0,00956	0,01234	0,02037	0,03445	0,00436
GLOBAL	1,8099	0,07044	-0,05725	0,0364	-0,0759	-0,03669	-0,0632	0,07809	0,00056
GAS	2,2487	0,07038	0,12586	-0,04844	-0,08617	-0,06223	0,02062	0,02652	-0,03906
OIL	2,3426	0,07019	0,09762	-0,02639	-0,08428	-0,14463	0,03192	-0,07711	-0,08757
POWER	2,1197	0,06997	0,07467	-0,04443	0,11928	-0,09042	-0,11117	-0,03517	-0,03849
CARBON	2,2555	0,06921	0,04643	-0,0968	0,03574	0,09477	0,07937	0,03736	-0,05276
NEED	1,03	0,06913	-0,00619	-0,01702	0,03414	-0,01406	0,01388	-0,03418	0,00125
AREA	1,5941	0,06867	-0,02111	0,07722	0,05213	0,01797	-0,0407	-0,07509	0,06305
ECONOMIC	1,5524	0,06741	-0,03062	-0,0047	0,01296	-0,04653	0,1081	-0,03156	0,08013
LARGE	1,2498	0,06526	0,04507	0,03883	-0,05218	-0,01873	0,02751	-0,02272	-0,02938
PROBLEM	1,193	0,06478	-0,04673	0,01238	0,00928	-0,02721	-0,01457	-0,01885	0,00026
WARM	2,0896	0,06411	-0,04752	0,06296	-0,06911	-0,02119	-0,08123	0,10765	0,02094
STATE	1,5992	0,06351	-0,04749	-0,0169	0,02546	0,053	0,00993	-0,05718	-0,04539
REDUCE	1,3526	0,0629	0,04531	-0,04047	0,01376	0,01843	0,01785	0,03172	0,04225
RISK	2,0727	0,06271	-0,08034	0,1166	0,03213	0,03116	-0,05356	0,05365	-0,06066
CONCERN	1,2659	0,06074	-0,04854	-0,00834	0,00249	-0,01593	-0,01416	0,00025	0,04847
POLITICAL	1,6231	0,06067	-0,11127	-0,03373	0,00926	-0,04052	0,03509	-0,00428	0,06927
MAJOR	1,0041	0,06019	0,01585	0,01514	-0,00889	-0,03402	-0,02338	-0,02542	0,01887
POINT	0,9187	0,05934	-0,01886	-0,0168	-0,02919	-0,02452	0,00808	0,00893	-0,01241
ENVIRONMENTAL	1,4835	0,05922	-0,02349	-0,02703	0,02255	-0,0114	0,06629	-0,01146	0,03431
EARLY	0,9335	0,05897	0,02833	0,00601	-0,02596	-0,0037	0,00248	0,00509	0,00132

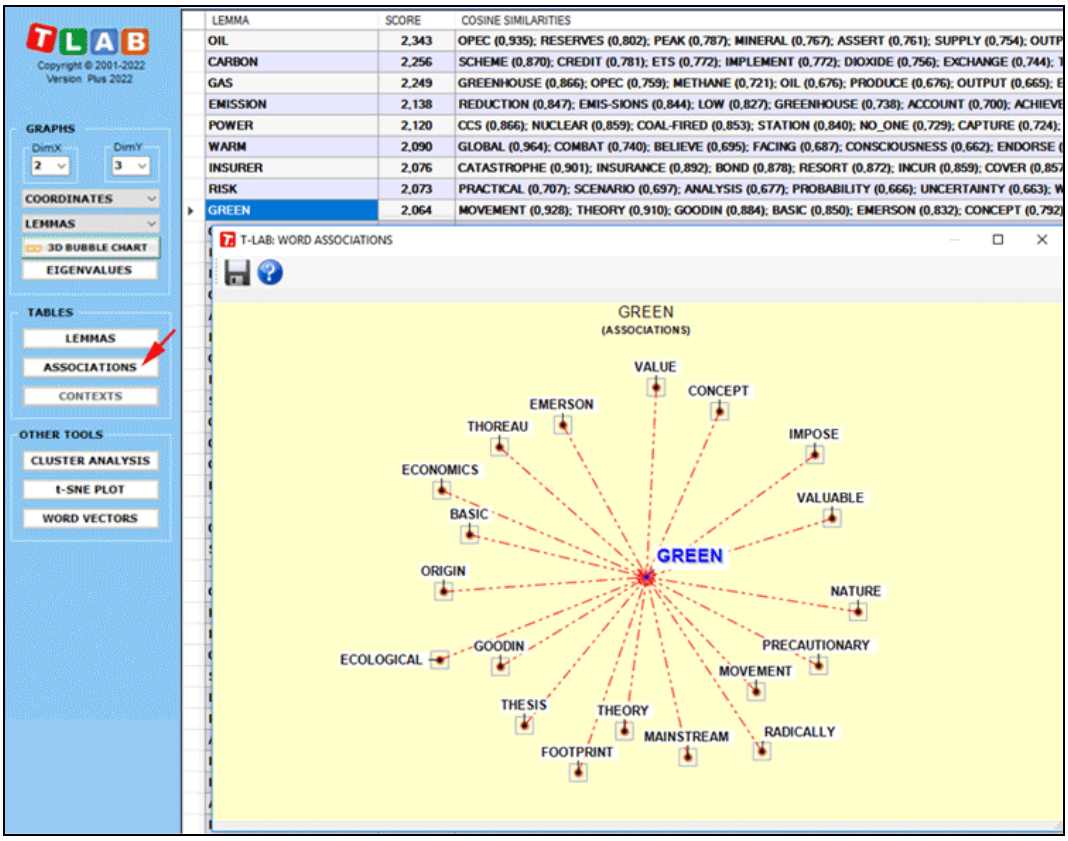
N.B.: The **importance** score of each lemma is computed by summing the absolute values of its first 20 coordinates (i.e. the eigenvectors), each one multiplied by its corresponding eigenvalue.



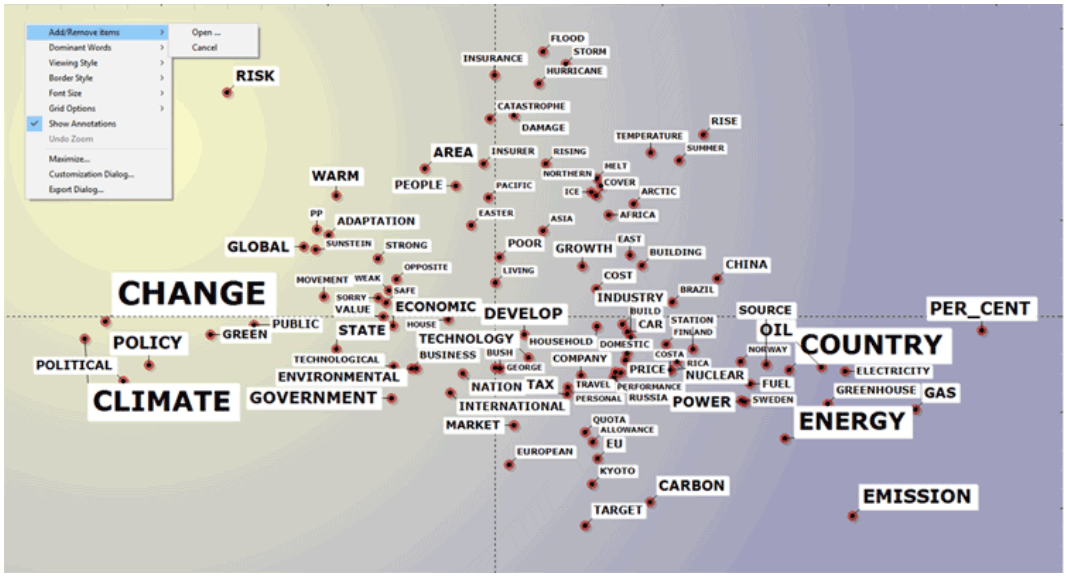
EC_ID	DIM_01	DIM_02	DIM_03	DIM_04	DIM_05	DIM_06	DIM_07	DIM_08	DIM_09	DIM_10	DIM_11
1535	0,04455	-0,04482	-0,0317	-0,03135	0,02987	-0,00605	-0,04588	-0,01318	-0,04025	-0,02249	-0,11303
648	0,04423	-0,01064	-0,02935	0,07795	-0,01775	-0,08317	-0,0209	0,00342	-0,03299	0,00819	0,04622
1013	0,04312	0,04639	-0,01214	0,13232	-0,02033	-0,0937	-0,04031	0,01416	0,01897	-0,00442	0,01693
1161	0,04254	-0,01104	-0,0199	-0,0095	0,06294	-0,03028	-0,02804	0,03888	-0,01117	0,03425	-0,00687
1541	0,04084	0,00965	-0,06259	-0,05385	0,03536	-0,06273	0,04135	0,02108	-0,03487	0,03779	0,01657
990	0,04063	-0,00582	-0,02576	0,07439	-0,03168	0,01063	0,00041	0,00532	0,02707	0,017	-0,02222
432	0,04057	0,0239	-0,02198	0,03906	0,04216	0,0944	0,02462	-0,00159	0,06802	-0,04387	0,01996
508	0,04051	0,06894	-0,01971	0,05302	-0,0318	-0,01774	-0,00657	-0,0269	-0,01799	0,00591	-0,00017
972	0,04036	-0,01316	0,00243	0,04187	-0,02201	0,01027	-0,05111	-0,04453	-0,02238	0,01045	-0,02504

Any table can be **sorted** in ascending or descending order by clicking on any column header. In order to **export** any table, just use the right click of the mouse when data are displayed. Please note that, the first time such a table is exported, the Eigenvalues are also exported. This way the user is allowed to evaluate the relative weight of each dimension, that is the percentage of variance explained by each one of them.

By clicking the **Associations** button (see below), a further table is displayed with the similarity measures (i.e. cosine coefficients) of each word. Moreover, when any row of such a table is clicked, a graph is displayed with the corresponding data.



The main **charts** shows the relationships between the key-terms (i.e. lemmas) on the selected dimensions (see below).



By default, the above chart includes the 100 most important lemmas. However the user is allowed to customize both the number of lemmas and the chart characteristics.

CORPUS PREPARATION

Corpus Preparation

In the case of a single document (or a corpus considered as a single text) **T-LAB** needs no further work: just select the 'Import a single file.." and proceed as explained in the corresponding section of this manual.

When, on the other hand, the corpus is made up of various texts and/or categorical variables are used, the Corpus Builder tool must be used, which automatically transforms any textual material and various types of files (i.e. up to eleven different formats) into a corpus file ready to be imported by **T-LAB**.

N.B.:

- we advise an orthographic review of the material to be analysed. Moreover, if some important acronyms are spaced out from punctuation (e.g. "U.N.") their transformation in single string (e.g. "U_N") is recommended; this is because, in the normalization phase, **T-LAB** interprets the punctuation marks like separators;

- at the end of the corpus preparation phase it is recommended that a new folder be created which should contain only the corpus to be imported.

Structural Criteria

There are two **structural criteria** which must be observed: the **corpus size** and its subdivision into **parts**.

As for the size, all **T-LAB** tools have been tested with a 90Mb corpus, approximately equivalent to 55,000 pages in text format.

Minimum size limits require different evaluation criteria, because, under a certain threshold, the corpus size can prejudice the reliability of many statistical analyses. Just follow these simple instructions: use corpora with at least 5,000 occurrences (approximately 30 Kb); otherwise, in the case of open-ended questions, a minimum of 50 answers.

In order to be processed, a corpus can be made up of: a single text without further partitions; a single text subdivided according to criteria established by the user (for example, a book divided into chapters); a number of texts (for example, different interviews or documents) classified through the use of labels linked to as many **variables** or **IDnumber**. In any case, the corpus is subdivided into parts that must be defined by precise **formal criteria**.

Formal Criteria

In the case of a **corpus** made up of a **single text**, and when the user doesn't resort to variables, there are no further operations required: it is possible to continue with the **importation** phase.

When, on the other hand, the corpus is made up of **various text documents** and/or categorical variables are used, the corpus preparation must be done by means of the **Corpus Builder** tool which **automatically** applies the following criteria:

Each text or subset of it (the "parts" defined by variables and/or IDnumber) is preceded by **a coding line**.

Each coding line has this format:

- It **begins** with a **four asterisks** string (****) followed by a blank space. **T-LAB** reads this string as: "here begins a user-defined text or a context unit".
- It **goes** on with the addition of strings made up by **single asterisks** and labels that define **cases** (**IDnumber**), **variables** and respective **categories**.
- It **ends** with the return key.

Here are some examples.

The following line introduces a text (or a corpus subset) codified with three variables - AGE, SEX and OCC (occupation) - and their categories (ADUL, FEM, PROF).

**** *AGE_ADUL *SEX_FEM *OCC_PROF

The following line introduces a text (or a corpus subset) codified with the same variables and the **IDnumber** label

**** *IDnumber_0001 *AGE_ADUL *SEX_FEM *OCC_PROF

The following line introduces a text (or a corpus subset) codified with two variables: YEAR, NEWSP.

**** *YEAR_98 *NEWSP_TIMES

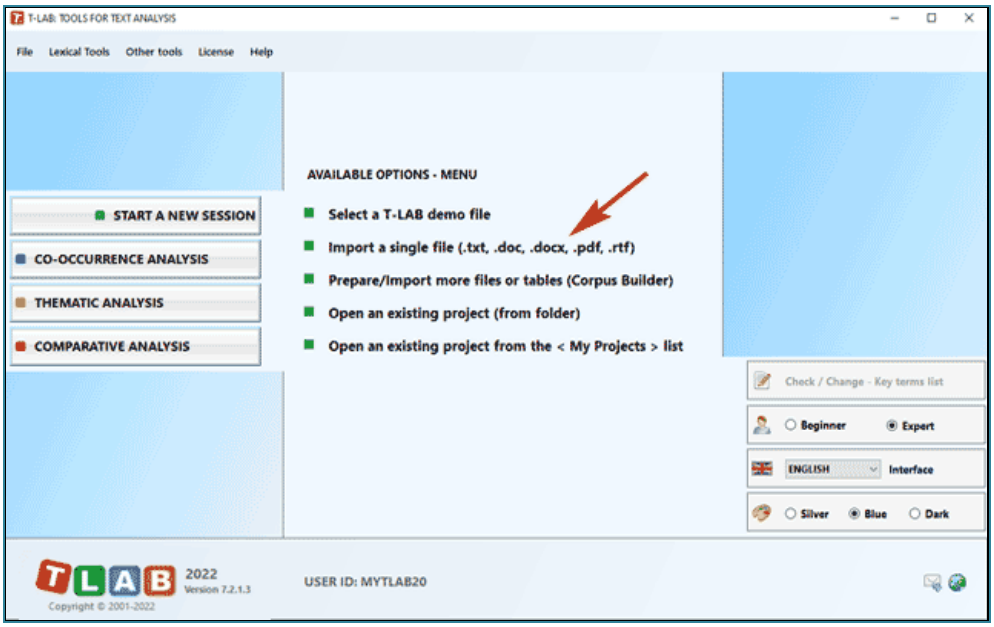
In **each coding line** these **T-LAB** rules are observed:

1. Each label (IDnumber, variables and variable categories) cannot be spaced out by blank spaces;
2. Each label - both for variables and variable categories - cannot be longer than **25** characters (min. 2);
3. Each variable label must be linked to the respective category using an underscore ("_");
4. Between two different variables, that is before the next asterisk, a blank space must be inserted;
5. Each variable and respective category must be assigned for each corpus subset;
6. We can use a maximum of 50 variables, each allowing a max of 150 categories which can be compared;
7. The maximum IDnumbers is fixed at 99.999 for short texts (Max. 2,000 characters each, e.g. responses to open-ended questions, twitter messages, etc.) at 30,000 for the other cases.

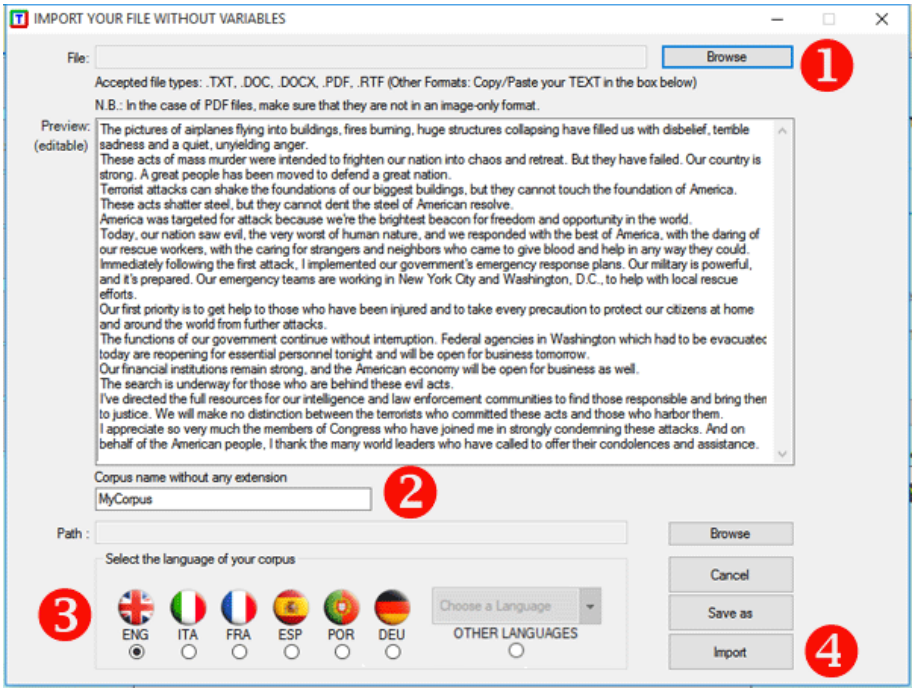
FILE

Import a single file…

In the case of a single document (or a corpus considered as a single text) **T-LAB** needs no further work: just select the 'Import a single file…' option.

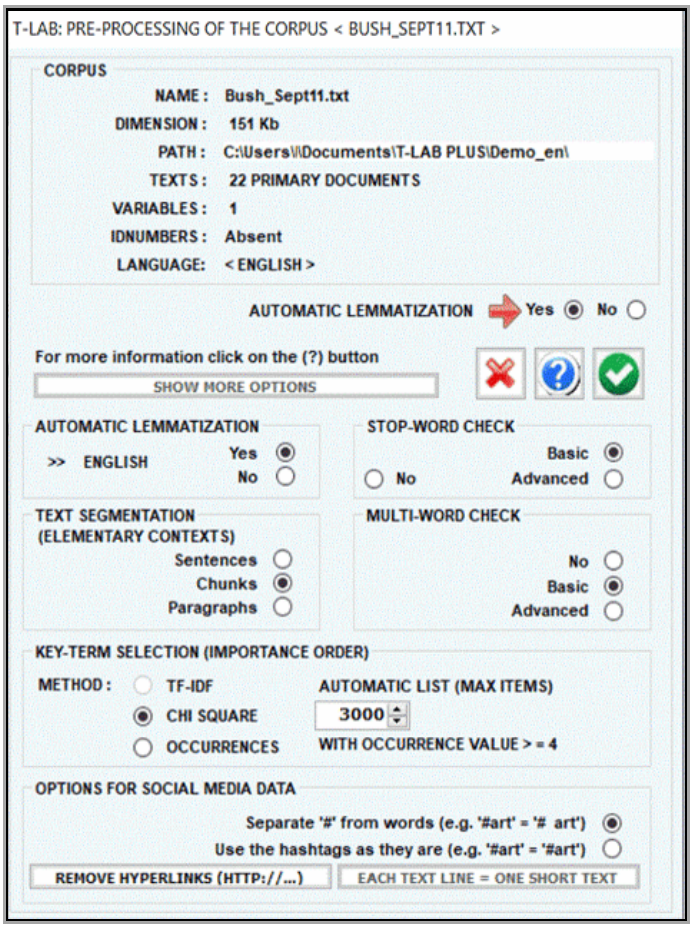


Then perform the following steps: (see the image below): (1) select any file; (2) choose the project name; (3) select the language of your text; (4) click on 'Import'.



Subsequently a setup form appears (see below) in which the user can make his choices.

N.B.:
- As the pre-processing options determine both the kind and the number of analysis units (i.e. context units and lexical units), different choices (see below the advanced options) determine different analyis results. For this reason, all **T-LAB** outputs (i.e. charts and tables) shown in the user's manual and in the on-line help are indicative only;
- All pre-processing steps are performed when importing any type of corpus.



1 - AUTOMATIC LEMMATIZATION OR STEMMING

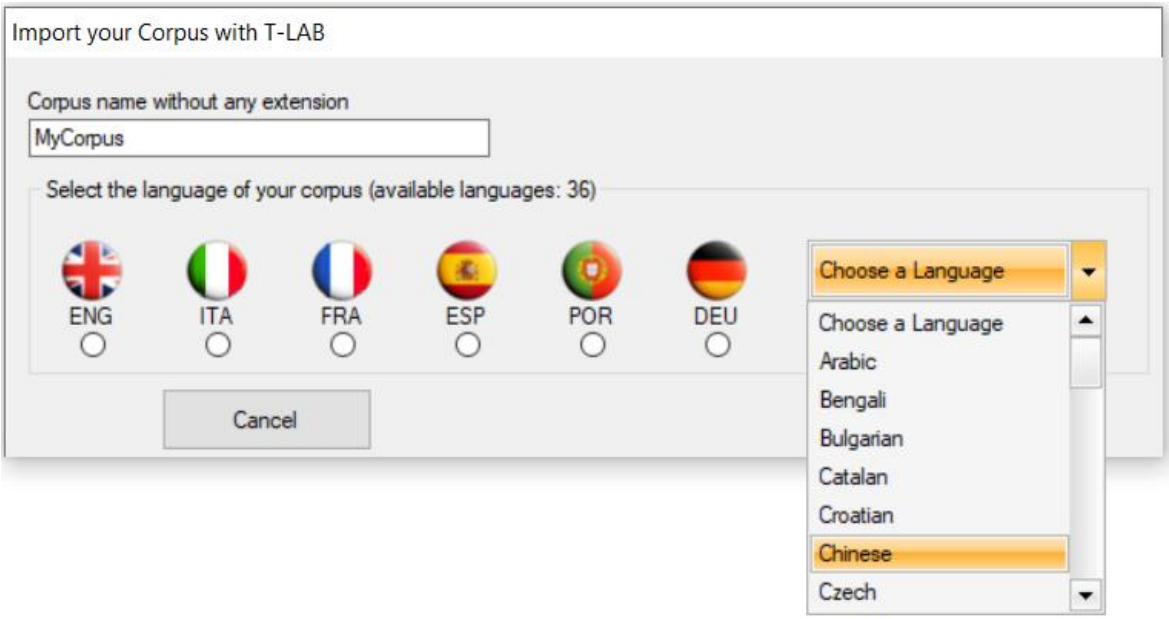
Here is the complete list of the thirty (30) languages for which automatic lemmatization or the stemming process is supported by **T-LAB.**

LEMMATIZATION: Catalan, Croatian, English, French, German, Italian, Latin, Polish, Portuguese, Romanian, Russian, Serbian, Slovak , Spanish, Swedish, Ukrainian.

STEMMING: Arabic, Bengali, Bulgarian, Czech, Danish, Dutch, Finnish, Greek, Hindi, Hungarian, Indonesian, Marathi, Norwegian, Persian, Turkish.

In any case, without automatic lemmatization and / or by using customized dictionaries the

user can analyse texts in **all languages**, provided that words are separated by spaces and / or punctuation.



The result of the lemmatization (or stemming) process can be verified by means of the **Vocabulary** function and can be modified by means of the **Dictionary Building** function.

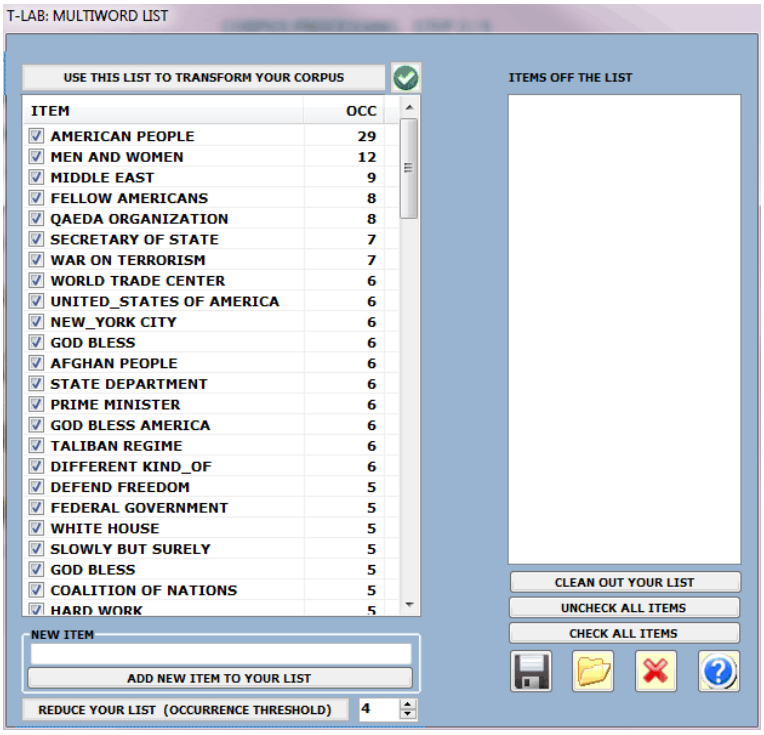
2 - TEXT SEGMENTATION (ELEMENTARY CONTEXTS)

According to the user's choices, the **elementary contexts** for the computation of **co-occurrences** can be four: sentences, chunks of comparable length, paragraphs or short texts (e.g. responses to open-ended questions).

The corpus_segments.dat file allows the user to verify the result of corpus segmentation.

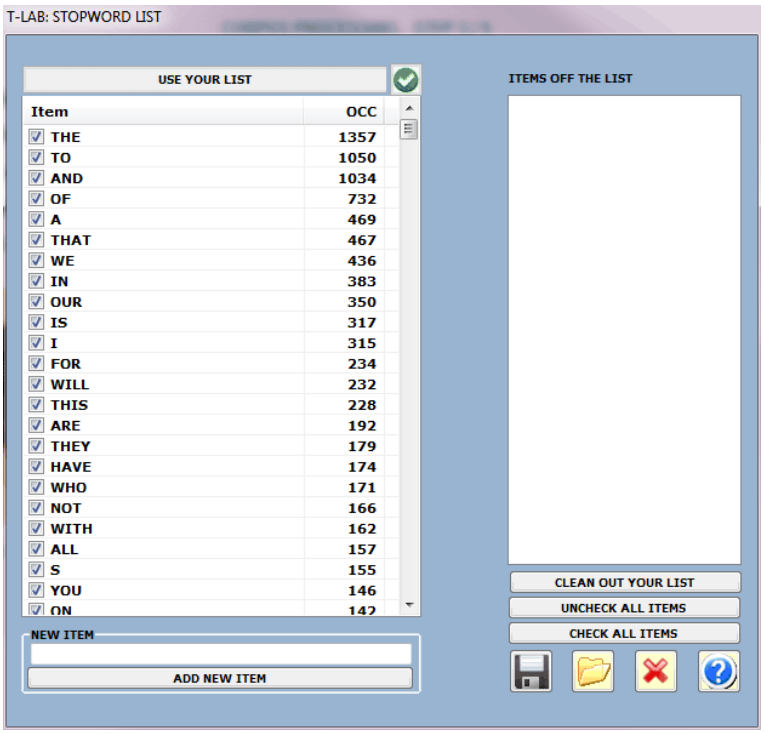
3 - MULTI-WORD CHECK

The "**Basic**" option activates the automatic use of T-LAB **multi-word** list.
Whereas the "**Advanced**" option, enabled with automatic lemmatization only, allows the user:
- to verify and modify the list of multi-words not included in the T-LAB database;
- to import and use **customized lists** (Multiwords.txt files).



4 - STOP-WORD CHECK

The "**Basic**" option activates the automatic use of T-LAB **stop-word** list.
Differently the "**Advanced**" option allows the user:
- to verify and modify the list of stop-words within the corpus;
- to import and use **customized lists** (StopWords.txt files).

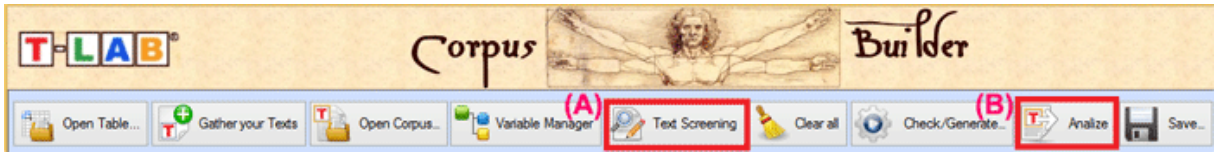


5 – KEY-TERM SELECTION

Available options allow us to choose the selection method (**TF-IDF** or **Chi-Square**) and the maximum number of **lexical units** to be included in a list used by **T-LAB** for analysing texts with **automatic settings**.
N.B.: When the importation process is over, by using the **customized settings**, the user can review the key-term selection and build various **lists** to be applied.

Prepare a Corpus (Corpus Builder)

N.B.: The pictures shown in this section have been obtained by using a previous version of **T-LAB**. In **T-LAB 10** this tool includes two additional buttons: a) one, named **Text Screening**, which becomes enabled when the corpus size is up to 20 MB; b) the other which allows the user to immediately proceed with the import of selected textual materials (see the below picture).



This software tool is intended to **simplify** and **speed up** any transformation of documents and textual materials into a **corpus** file ready to be processed by **T-LAB**.

More specifically, such a tool allows the following operations:

1. Automatically **import** various types of files;
2. **Edit** and **tag** them by using categorical variables;
3. **Save** the result as a corpus file ready to be imported by **T-LAB**;
4. **Check** and **modify** any corpus file which corresponds to the **T-LAB** format.

While the way that files are imported (see '1' above) varies according to their format, all the other operations follow the same logic.

Below is a short description of how to import the various files.

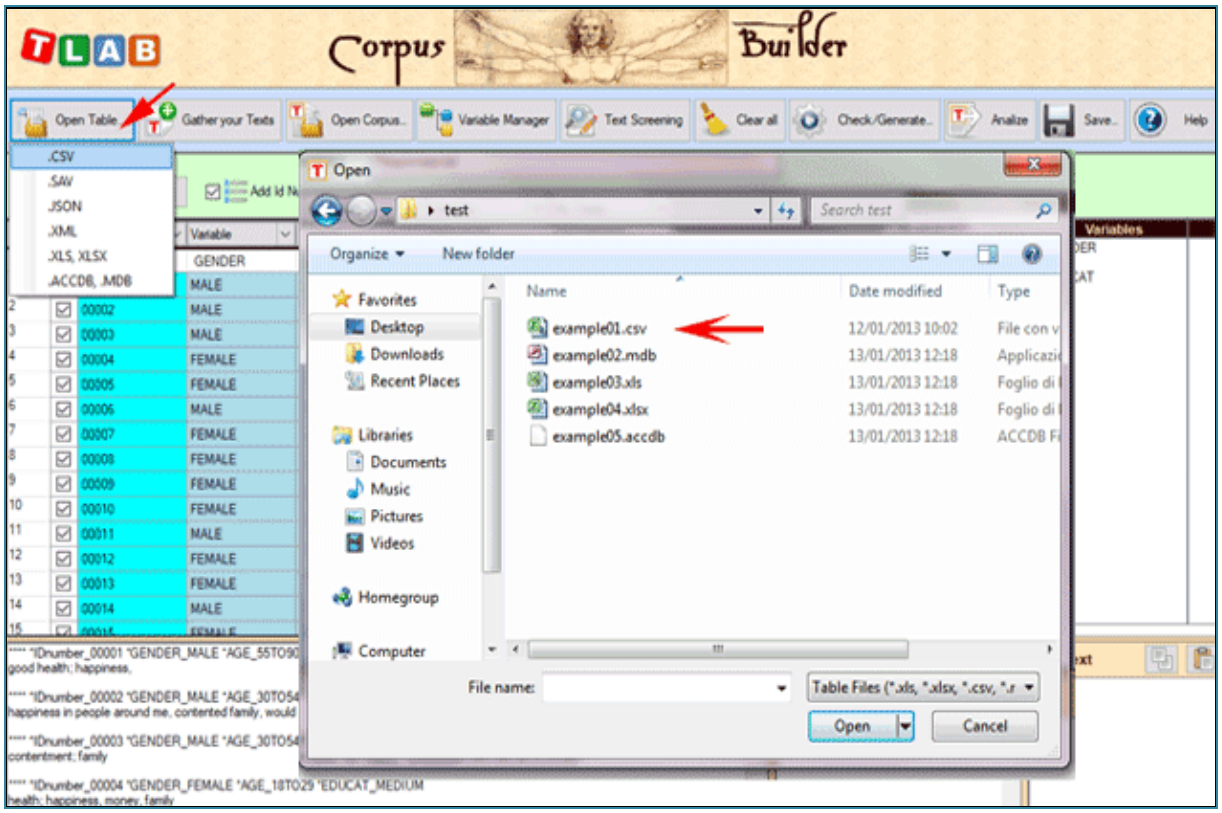
A - **Importing files in tabular or spreadsheet format** (CSV, .SAV, .JSON, .XML, .XLS, XLSX, .MDB, .ACCDB).

A **single file** which includes up to 30,000 records can be imported by the 'Open Table' option or by a simple drag and drop method (N.B.: When texts in each line do not exceed 2.000 characters, up to 99.999 records can be imported).

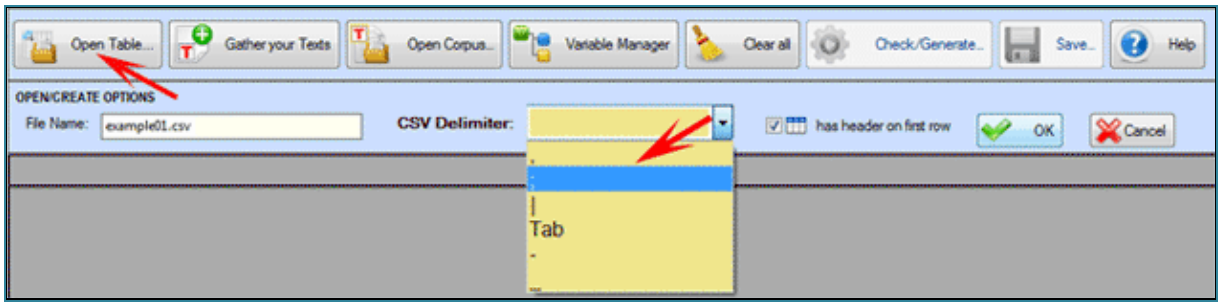
The fields/columns of such a file can contain the following data:

- Categorical Variables (one for each column, up to a 50 columns)
- Texts to be analysed (only one column);
- IDnumbers, i.e. identifiers of subjects (e.g. in the case of answers to open-ended questions) or of context units in which the corpus to be imported is subdivided.

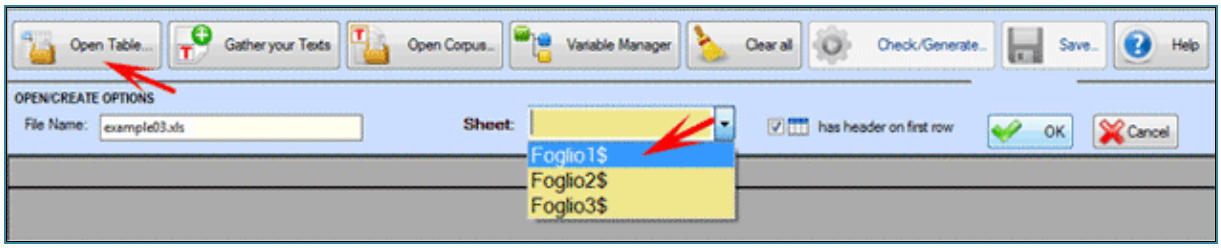
N.B.:
- While the presence of Categorical Variables and IDnumbers is optional, the presence of at least one column containing the texts to be analysed is mandatory.



When importing a .CSV file, the corresponding delimiter must be selected (see below).



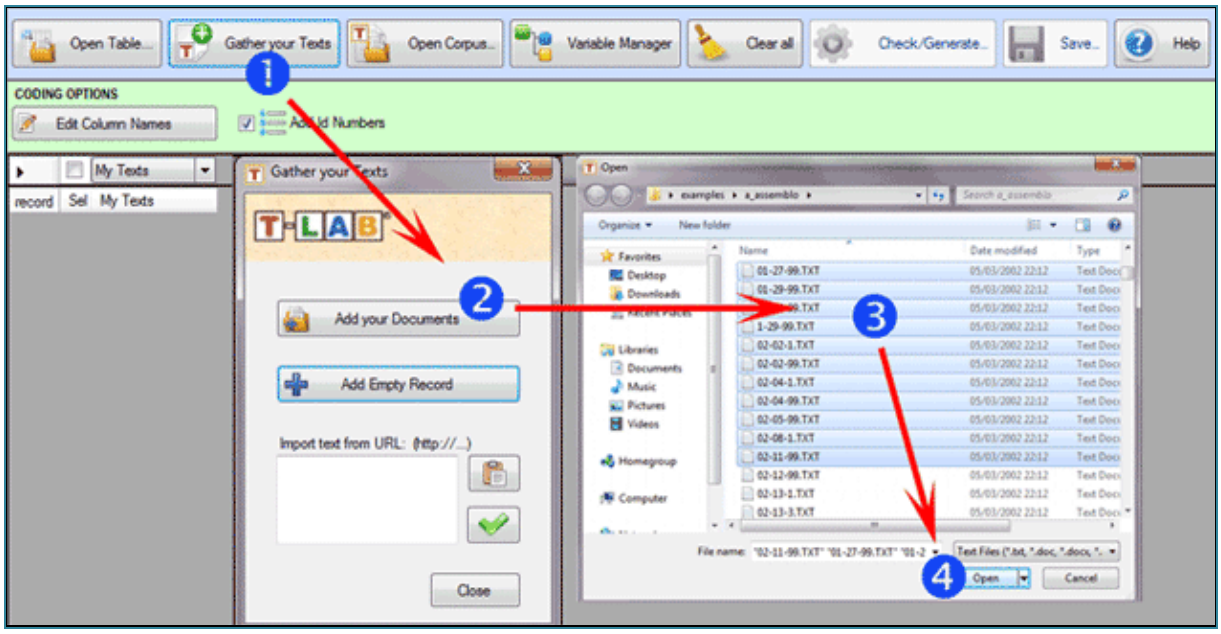
When importing Excel or Access files, only one table can be selected (see below).



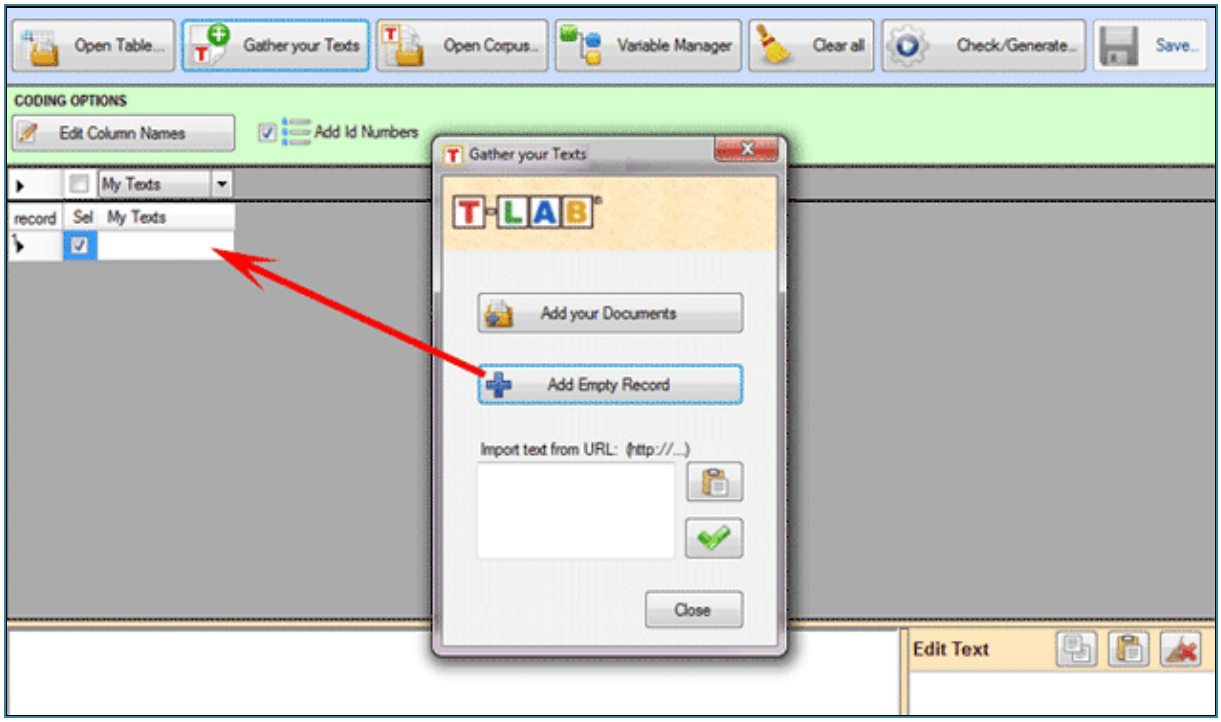
B - **Importing document files of various formats**

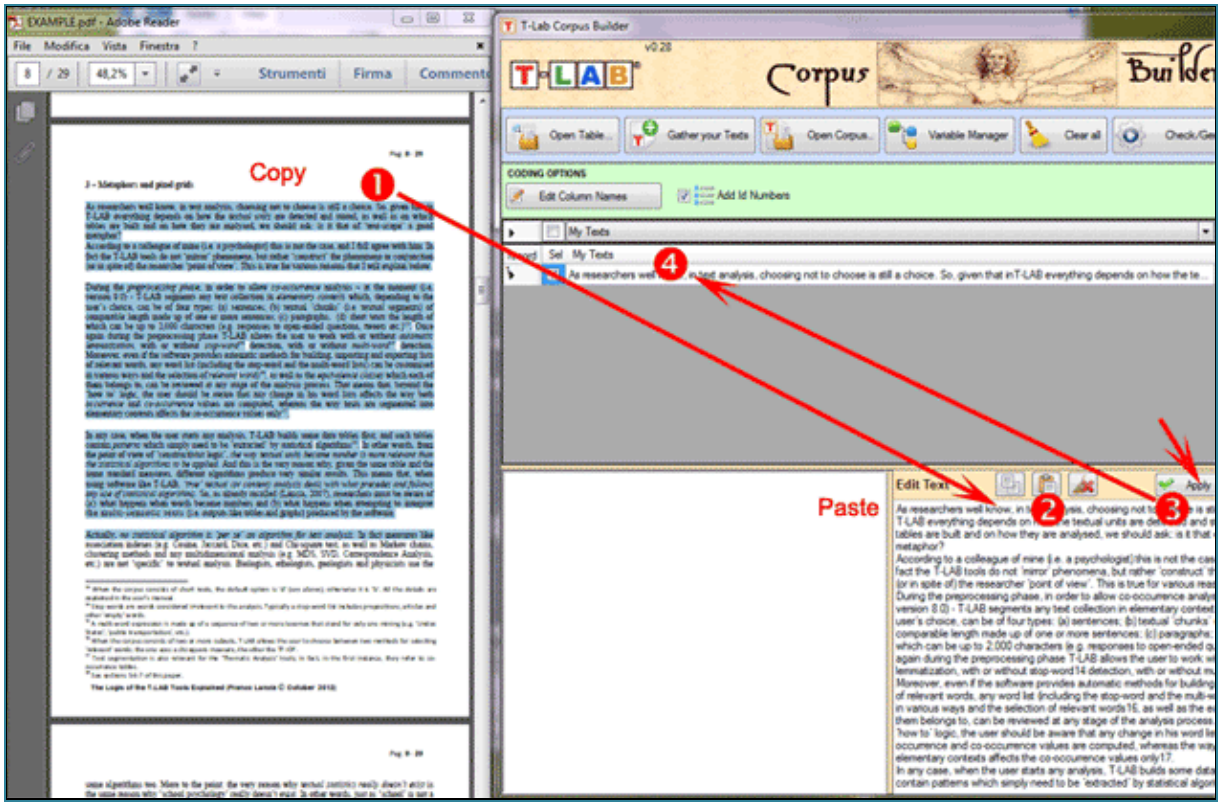
Up to 30,000 documents can be imported, either one by one or by multiple selection, through the 'Gather your Texts' option (see below). Three methods are available:

The **first method** ('Add your Documents') allows the automatic importation of .TXT, .DOC, .DOCX, .PDF and .RTF files.



The **second method** ('Add EmptyRecord') allows the user to copy/paste any type of text (see below).





The **third method** ('Import Text from URL') allows downloading HTML files from Internet, as well as editing their content before the importation (see below).

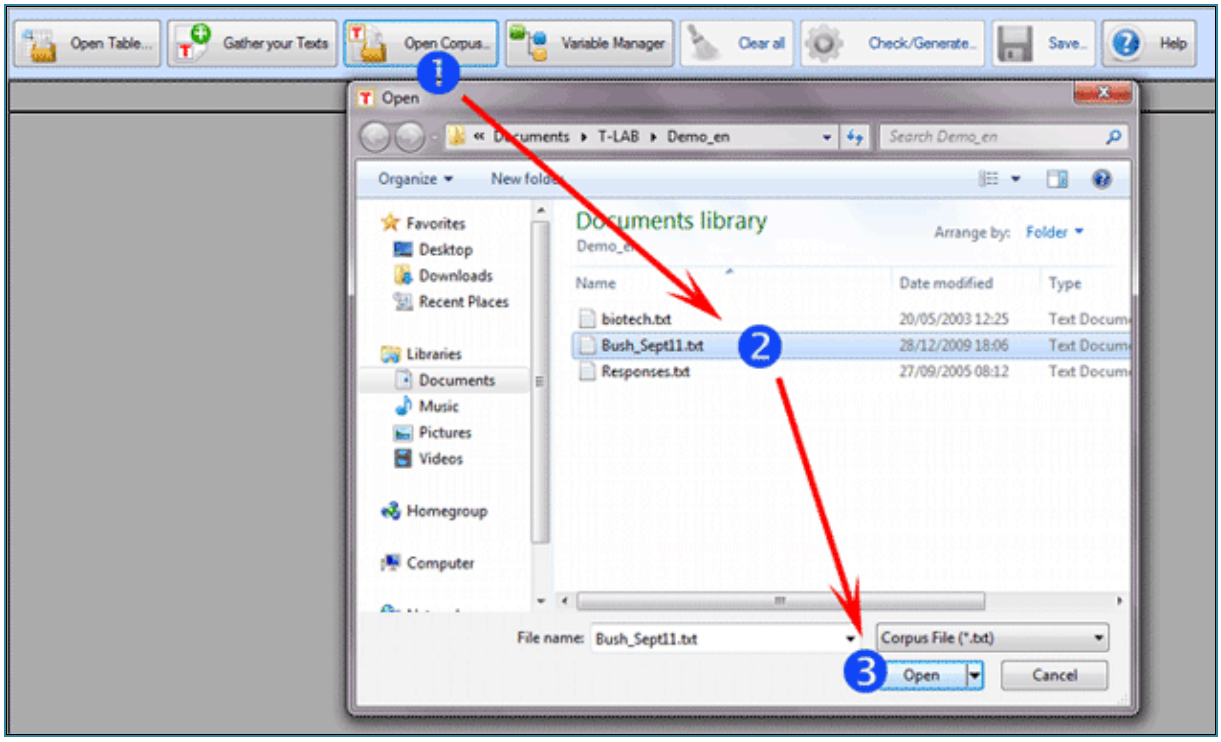
C - **Importing a corpus** file already encoded according to the **T-LAB** specifications**.**

The use of the 'Open Corpus' option is advisable in the following three cases:

1 – the user intends to modify the structure of a corpus already encoded (e.g. add further texts by means of methods explained in the previous 'B' section, modify the labels of variables and values, etc.);
2 – the user intends to check/fix errors of his manual coding that had been possibly done without the aid of the Corpus Builder module;
3 – the user intends to import a corpus file with a 'raw' coding, that is a corpus the sections of which (i.e. documents or records) are preceded by a coding line with four asterisks ('**** ')., just followed by a blank space and a return key (see below).

```
**** ¶
Much has been written about how to facilitate an effective
meeting, but apparently not every meeting facilitator has read
the literature because every occupational health nurse has
endured a "bad" meeting. Individuals who chair meetings have a
responsibility to create meetings that are worthwhile to the
attendees; attendees have a responsibility to be prepared for
meetings so meetings are productive. This article reviews key
meeting strategies, providing readers with ways to improve
meetings they attend or facilitate.¶
¶
**** ¶
Population health-based chronic care models of care are useful
in improving the health of a population while decreasing the
health care dollars spent on the population. Diabetes is a
disease that can be evaluated and treated using these models of
care. The Metro Nashville Public Schools Diabetes Health
Management Program has been shown to be beneficial to both
clients and their insurance trust in improving the health of
this population of individuals and decreasing the dollars spent
on this disease.¶
¶
**** ¶
Worker health is influenced by workplace, work processes, and
workmates. This case study shows it is possible to create health
```

In all the above cases (1,2,3) it is sufficient to select a single file by means of the 'Open Corpus' option (see below) or use the drag and drop method.

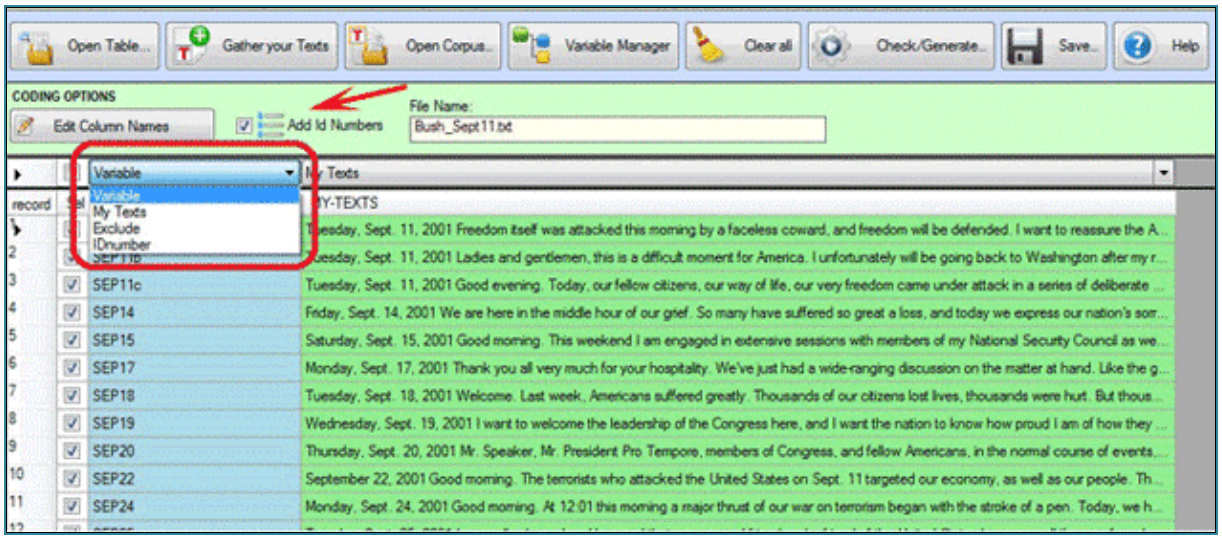


Operations that follow any file importation

At the end of the phase with which the files have been imported in **Corpus Builder**, either when the user is 'not' interested in the use of variables or when the encoding operations have already been carried out, he may proceed with the 'Check / Generate' option and afterwards with the exportation of the corpus to be imported in **T-LAB**.

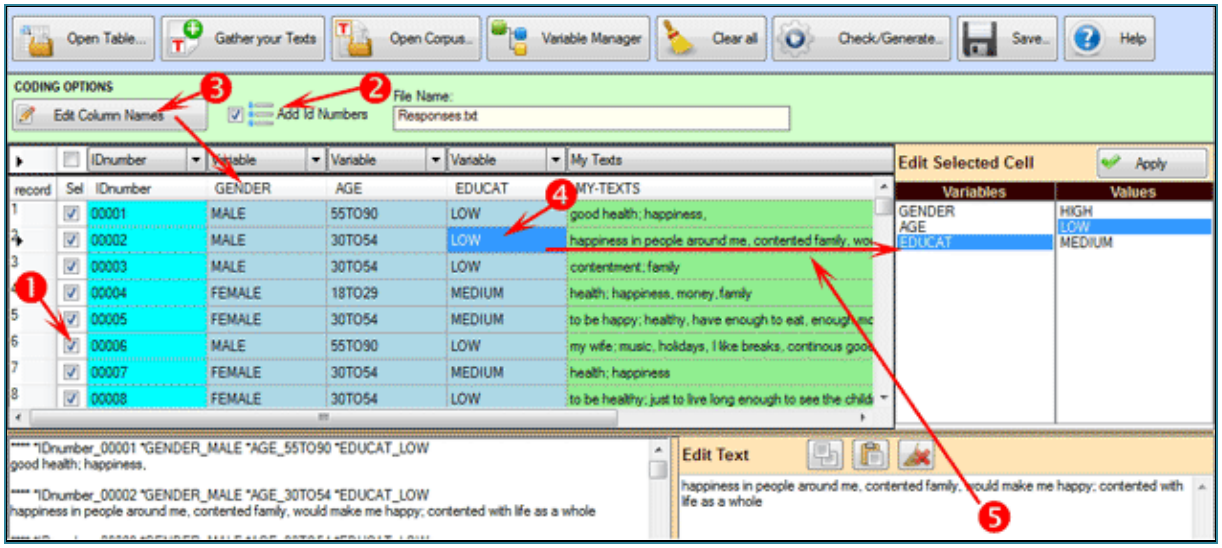
When the corpus is encoded it should be recalled that in all three types of importation mentioned in the preceding sections of this document ('A', 'B', 'C') data are displayed in various columns, the headers of which can be the following:

- 'Variable', i.e. categorical variables, the use of which is requested when analysing the characteristics and the reciprocal relationships of various corpus subsets;
- 'IDnumber', i.e. identifiers of cases/records, the use of which is optional;
- 'My Texts', i.e. the texts to be analysed, the use of which is mandatory and is allowed in a single column only.
- 'Exclude', the use of which indicates that data in the corresponding column(s) must not be saved by the Corpus Builder module.



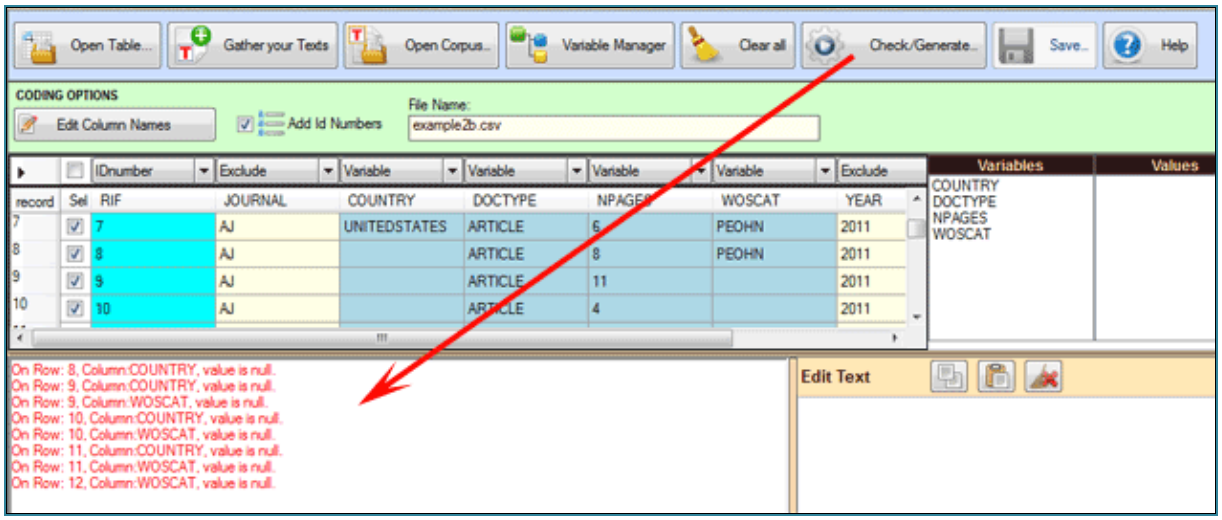
In **all cases** it must be remembered that:

- each record can be selected or deselected (see below '1'),
- IDnumbers can be automatically added (see below '2'),
- each variable name can be edited and changed (see below '3');
- each variable value can be edited and changed (see below '4')
- each 'My Text' field can be edited and changed too (see below '5').



Further information:

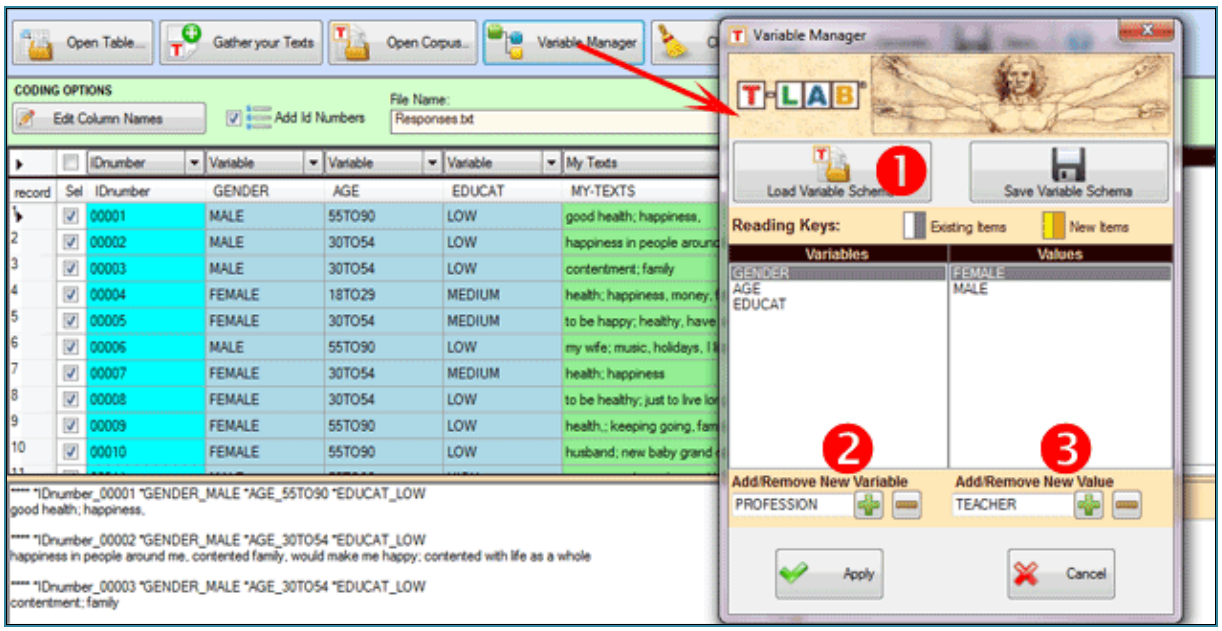
- the number of columns with categorical variables must not exceed 50;
- each variable can have a maximum of 150 values;
- the IDnumber values, if used, must be progressive starting from 1 (e.g. 1, 2, 3, etc.);
- each label, both for variables and values, must not exceed the length of **25** alphanumeric characters (at least 2) and must not be interrupted by blank spaces;
- when doing any operation, all detected errors are visualized in the bottom-left window (see below).



Use of the Variable Manager tool

The 'Variable Manager' tool allows the user to build, edit, load, save and change any coding scheme, even from a different corpus.

Each coding scheme includes the list of variables and that of their values (see below) .



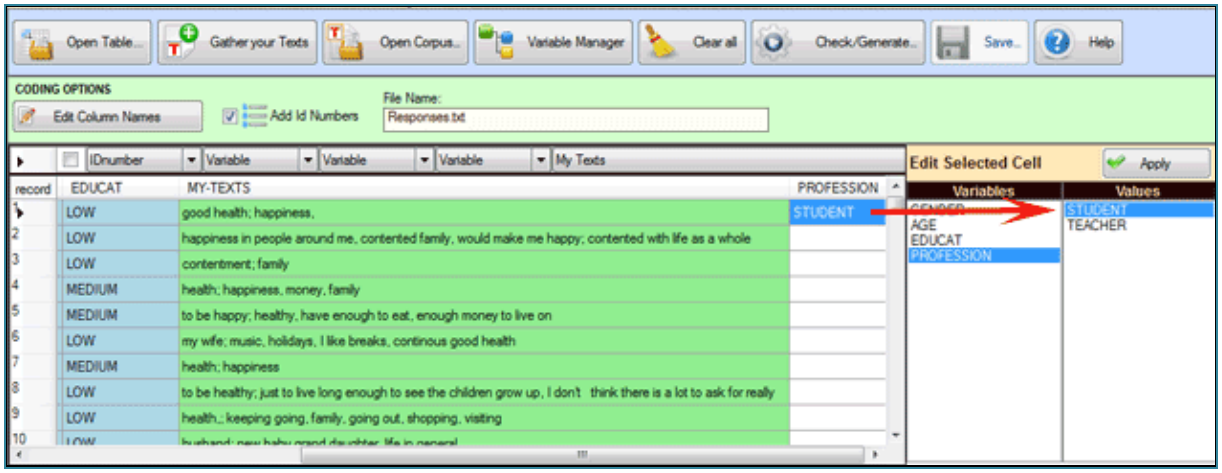
In order to add variables from a different corpus or from a previously saved scheme the 'Load Variable Scheme' option must be selected (see above '1'). Otherwise, to manually add variables and their values, the sequence of the above '2' and '3' options must be used.

Any value assigned to individual records must be added manually (see below) and in a single work session, this is because when saving coding schemes the values of each record are not recorded. Consequently, when the user is dealing with a corpus that includes a considerable number of records and / or his job requires more than one session, it is recommended to proceed as follows:

1 - import the amount of files / records that can be encoded in a single work session;

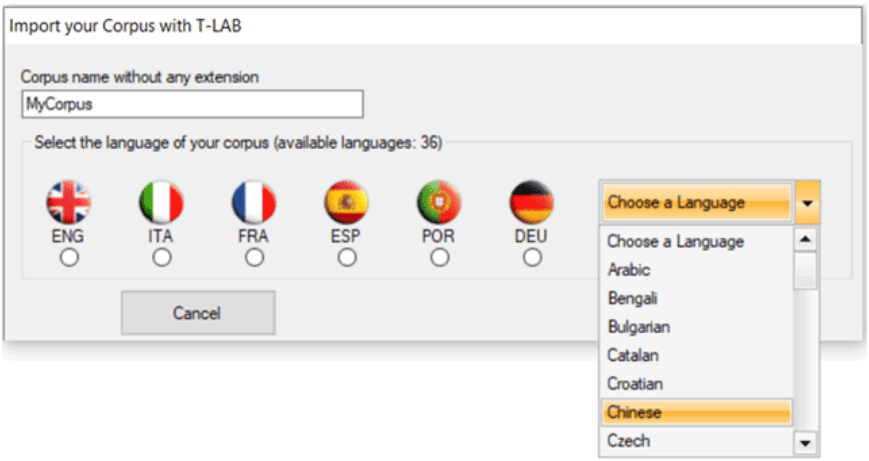
2 - save one's work as a corpus (see the 'Save' option in the Corpus Builder menu);

3- then, in the subsequent session, re-import the corpus previously saved (see above, point '2 '), add further records / files to encode and continue.

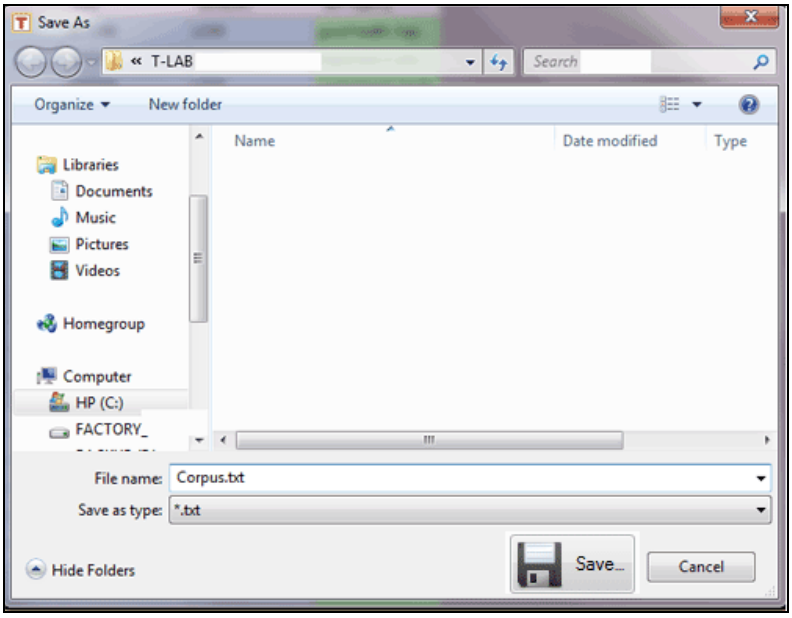


When the basic operations have been carried out (i.e., two or more texts have been gathered), by clicking the 'Check/Generate' button the user can verify the correctness of his work and export (A) or save (B) a corpus ready to be imported by **T-LAB**.

In the first case (A – see below) Corpus Builder creates a new folder under the directory '..\My Documents\T-LAB PLUS\" and automatically starts the importation procedure.
N.B.: In this case the new folder has the same name of the corpus file.



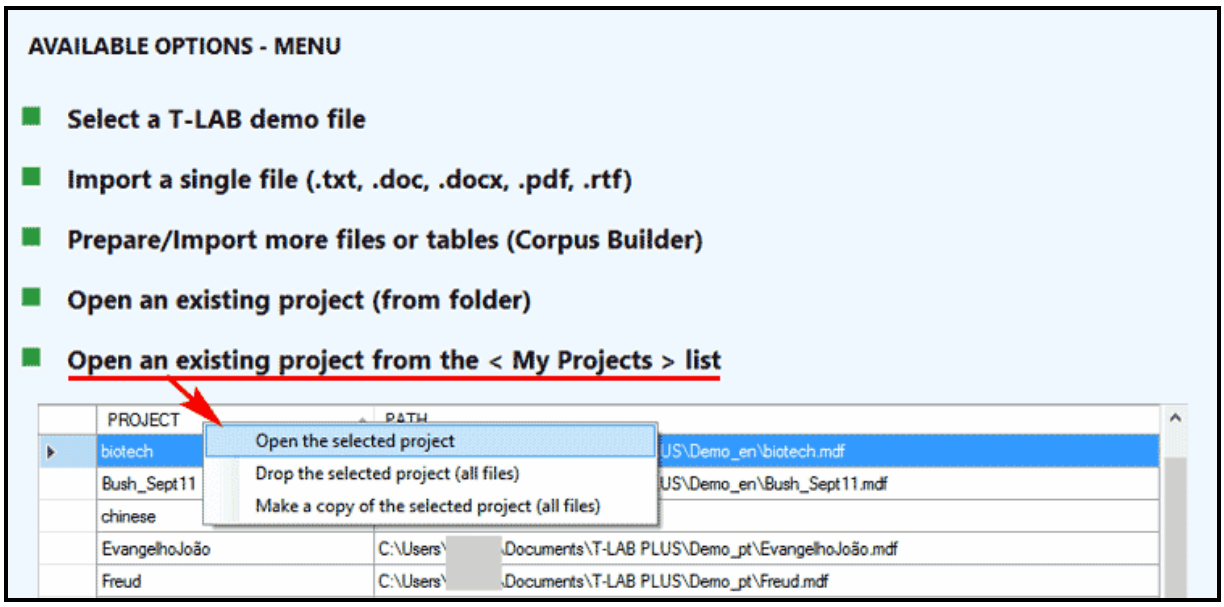
In the second case (B – see below) the user is enabled to save his corpus in whatever folder he wishes and aftwerwards he has to use the 'Import a corpus' option of **T-LAB**.
N.B.: In this case it is recommended that a new folder be created which should contain only the corpus to be imported.


Save As
« T-LAB
Search
Organize
New folder
Name
Date modified
Type
Libraries
Documents
Music
Pictures
Videos
Homegroup
Computer
HP (C:)
FACTORY_
File name:
Corpus.txt
Save as type:
*.txt
Hide Folders
Save...
Cancel

Open an existing project

Through this option the user can go back to work on a project already started, either by selecting the files from an existing folder or from the list proposed by **T-LAB**.

Also, when selecting an item from the list proposed by **T-LAB**, the right click of the mouse allows the user to delete its files or back them up to another folder (or to another device).



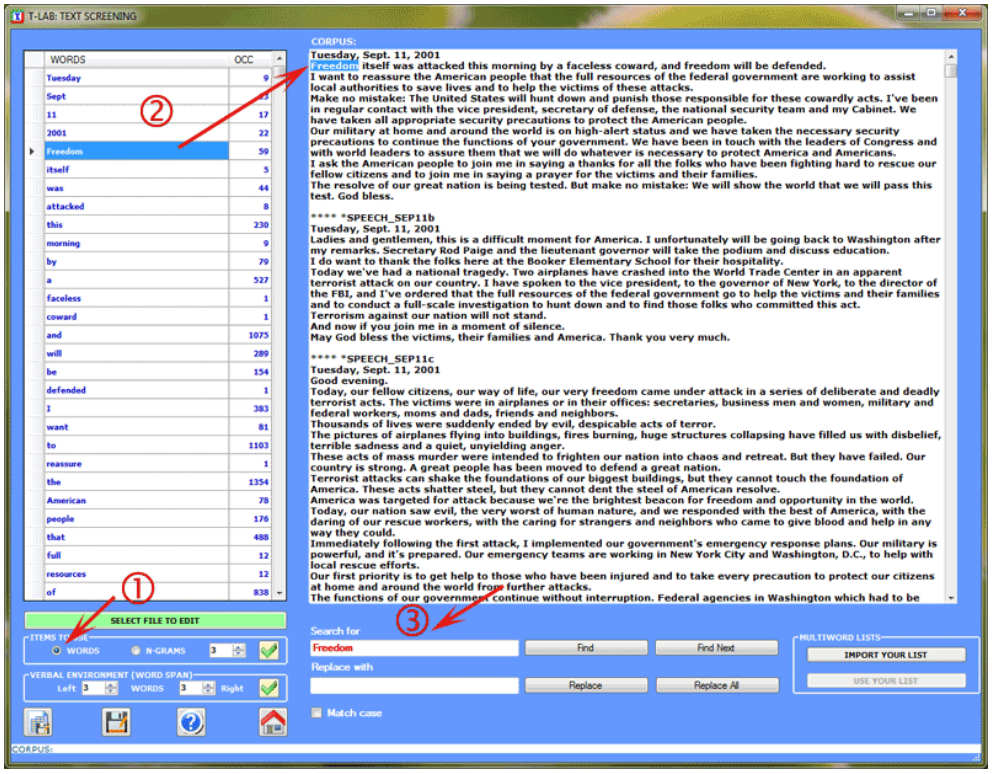
LEXICAL TOOLS

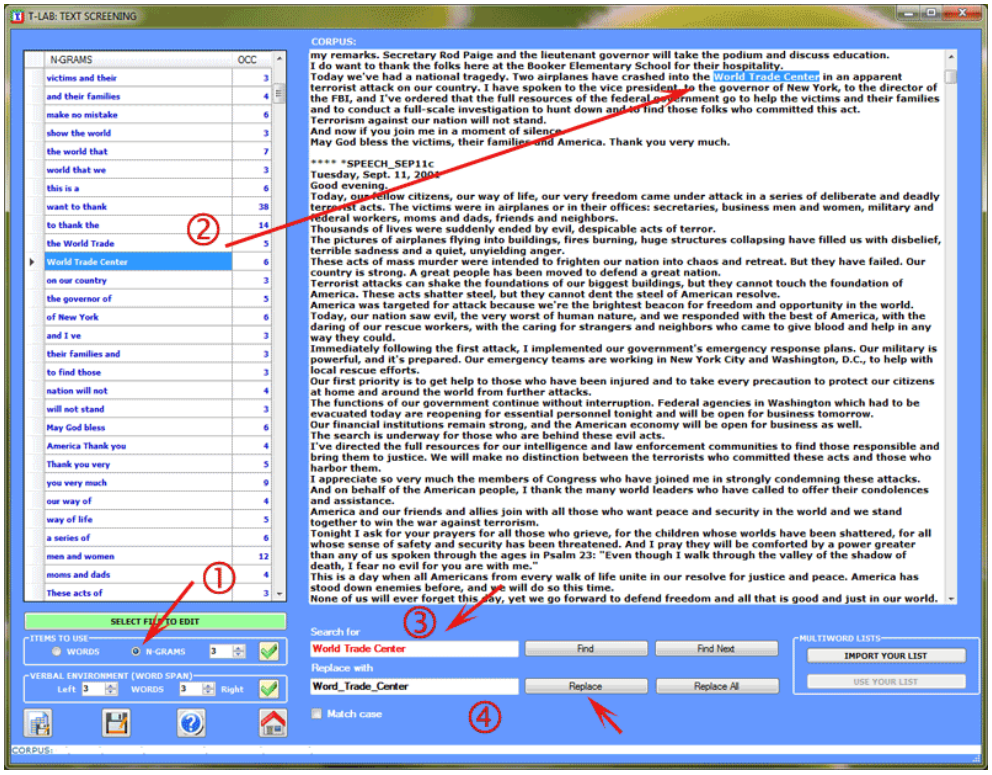
Text Screening / Disambiguations

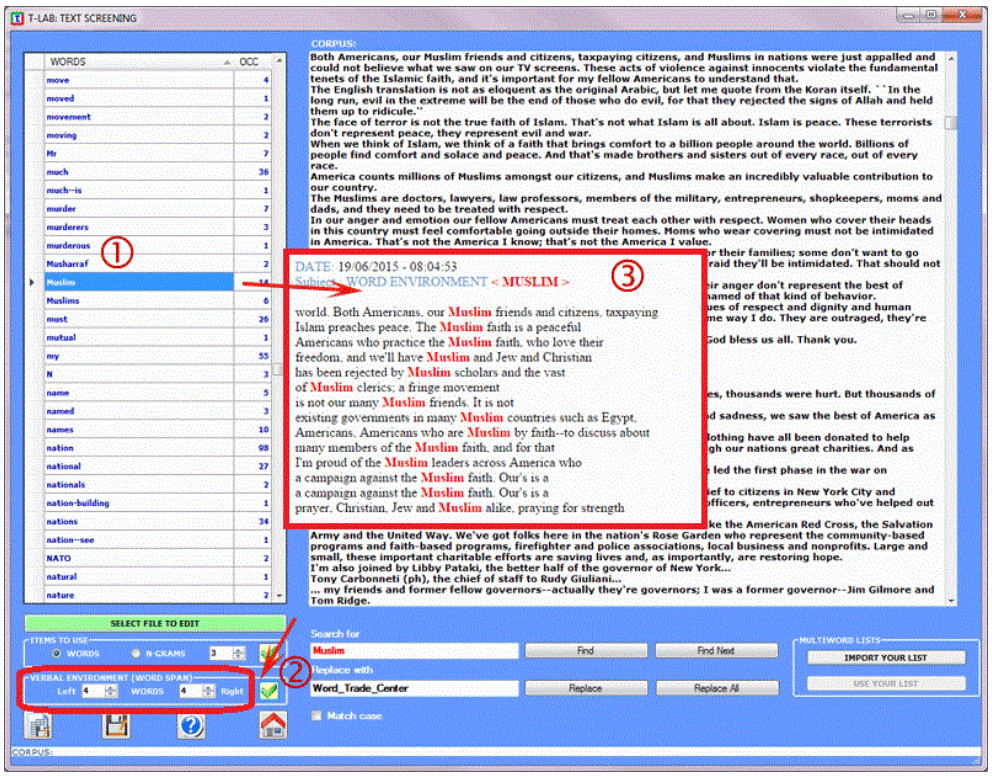
This **T-LAB** tool allows you to edit any corpus file (up to 30 Mb in size) and to perform useful operations either for a first **exploration** of contents or for the **disambiguation** of specific lexical units.

In particular, this tool automatically (and quickly) produces various **lists** and allows the user to perform operations such as **search** / **replace**.
The lists which can be obtained are the following:
a- **words** with their occurrences;
b- word **n-grams** with their occurrences;
c- customizable **word spans** of the selected keywords.

The below images show the possible operations in the three cases (a-b-c)



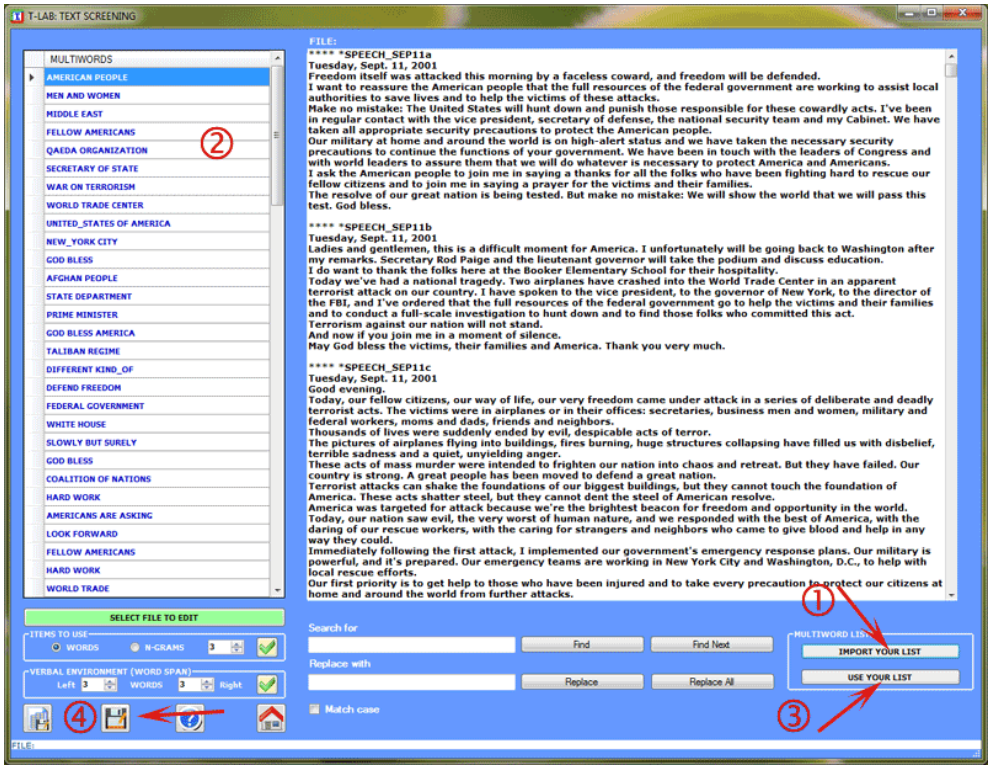




N.B.: A click on the bottom left button allows you to export both ‘a’ and ‘b’ lists as Excel files, while the lists 'c' lists are automatically exported as .html files.

-

It is also possible to import customized lists of Multiwords and possibly apply them to the displayed corpus (see the below picture).



-

When finished, if the user has edited the text and wants to save it, **T-LAB** allows them to create a new file (corpus_dis.txt) which, properly renamed, can be imported and analysed (see the above option ‘4’).

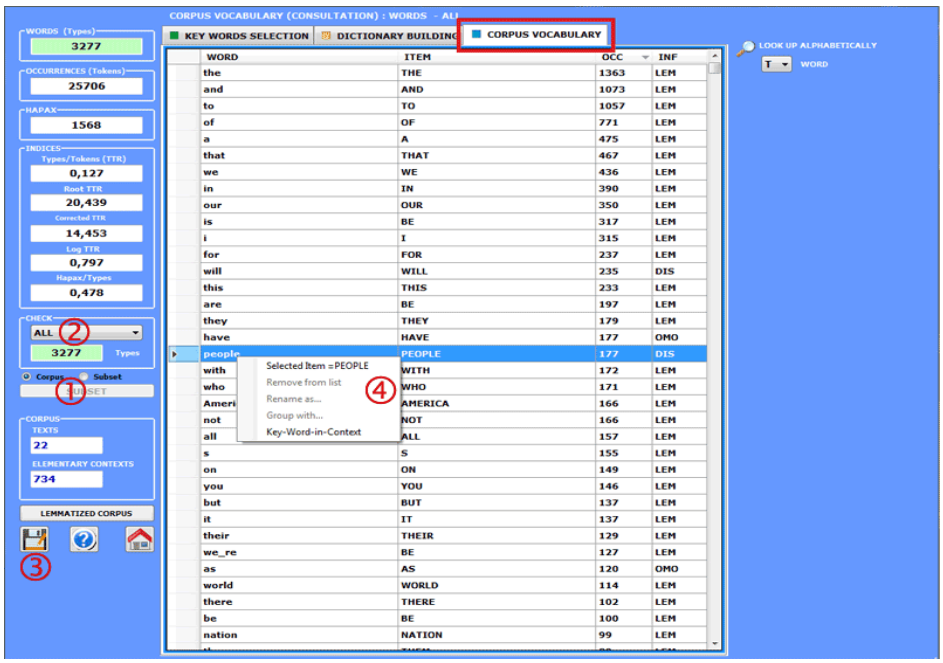
Corpus Vocabulary

This **T-LAB** tool allows us to check the **Vocabulary** of the corpus and its subsets (see option '1' below).
Moreover some measures of **lexical richness** are provided.

The Vocabulary table is a list including all distinct words (i.e. word types), the frequency of their occurrences (i.e. word tokens), their corresponding lemma (or label) and some categories used by **T-LAB** (see Glossary/Lemmatization).

The user can select (see option '2' below) the lexical units which belong to each category, view the corresponding table and save it as a .xls file (see option '3' below).

In addition, by right clicking any item, you can check its concordances (Key-Word-in-Context) (see option '4' below).



The measures of lexical richness are five:

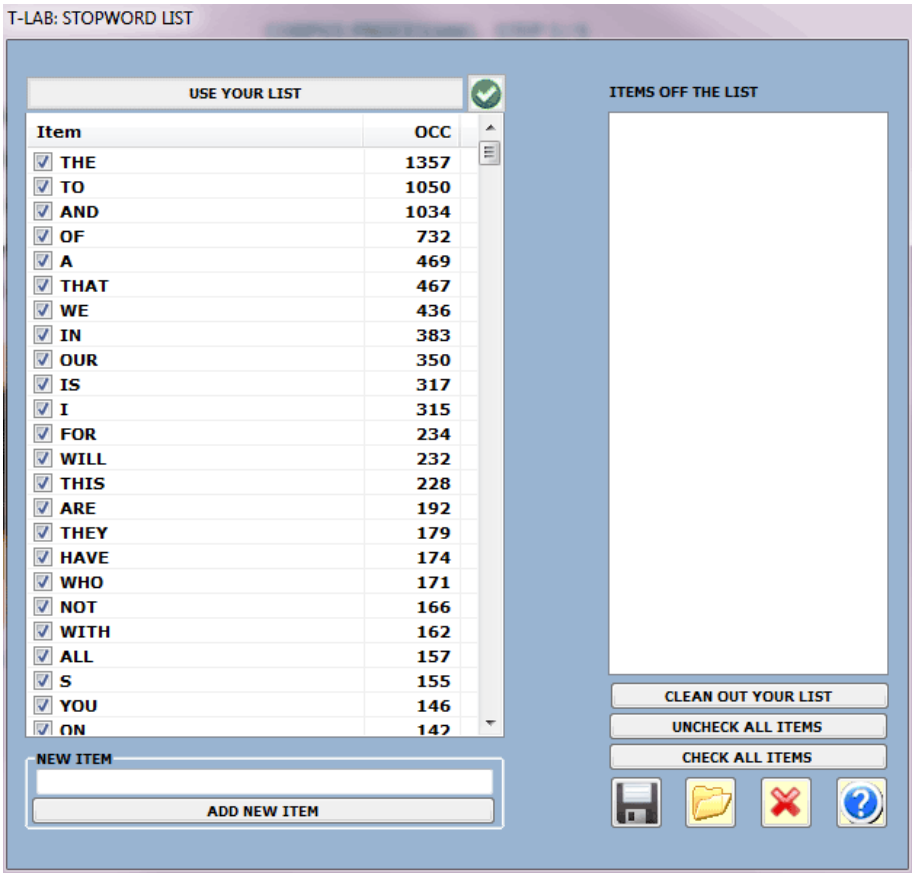
Type/Token ratio (i.e. TTR);
Root TTR (Guiraud, 1960), obtained by dividing the number of types by the square root of the

number of the tokens;
Corrected TTR (Carroll, 1964), obtained by dividing the number of types by the square root of twice the number of the tokens;
Log TTR (Herdan, 1960), obtained by dividing the logarithm of the number of types by the logarithm of the the number of the tokens;
Hapax/Types ratio.

N.B.:
- Hapax (i.e. Hapax Legomena) are words which occur only once in a corpus;
- When analysing a corpus subset, all measures of lexical richness do not include stop words (e.g. articles and prepositions).

Stop-Word list

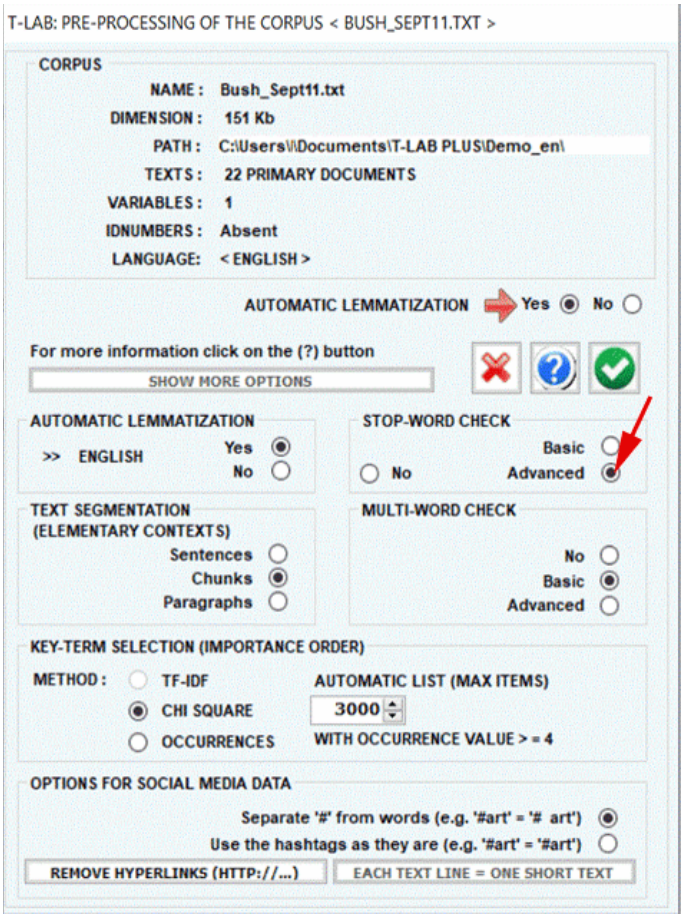
This option allows the user to create/modify **StopWord** lists within the following form:



In the StopWords.txt file arranged by the user the following rules must be respected:

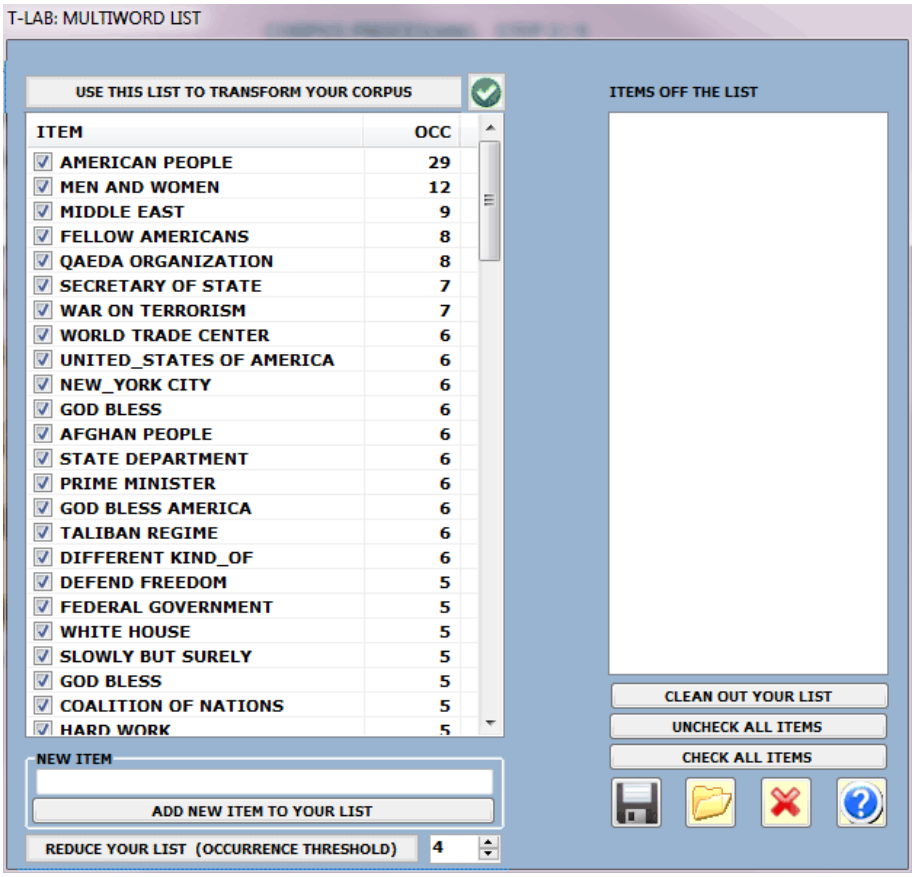
- the maximum length of a word string is 50 characters;
- neither blank spaces nor punctuation marks must be included.

In any case, to verify/use StopWord lists during the importation of a **new corpus** just select the "**Advanced**" option in the following form:


T-LAB: PRE-PROCESSING OF THE CORPUS < BUSH_SEPT11.TXT >
CORPUS
NAME : Bush_Sept11.txt
DIMENSION : 151 Kb
PATH : C:\Users\\Documents\T-LAB PLUS\Demo_en\
TEXTS : 22 PRIMARY DOCUMENTS
VARIABLES : 1
IDNUMBERS : Absent
LANGUAGE: < ENGLISH >
AUTOMATIC LEMMATIZATION Yes No
For more information click on the (?) button
SHOW MORE OPTIONS
AUTOMATIC LEMMATIZATION
>> ENGLISH Yes No
STOP-WORD CHECK
Basic
No Advanced
TEXT SEGMENTATION (ELEMENTARY CONTEXTS)
Sentences
Chunks
Paragraphs
MULTI-WORD CHECK
No
Basic
Advanced
KEY-TERM SELECTION (IMPORTANCE ORDER)
METHOD : TF-IDF
CHI SQUARE
OCCURRENCES
AUTOMATIC LIST (MAX ITEMS)
3000
WITH OCCURRENCE VALUE > = 4
OPTIONS FOR SOCIAL MEDIA DATA
Separate '#' from words (e.g. '#art' = '# art')
Use the hashtags as they are (e.g. '#art' = '#art')
REMOVE HYPERLINKS (HTTP://...)
EACH TEXT LINE = ONE SHORT TEXT

Multi-Word list

This option allows the user to create/modify **Multi-Word** lists within the following form.



Each Multiwords.txt file can be made up by "N" lines (max 5,000), each with a multiple word of max 50 characters, without punctuation marks.

Here are some lines of Multiwords.txt in the correct format:

Seattle people
Chamber of commerce
National Health Service
America's greatest traditions

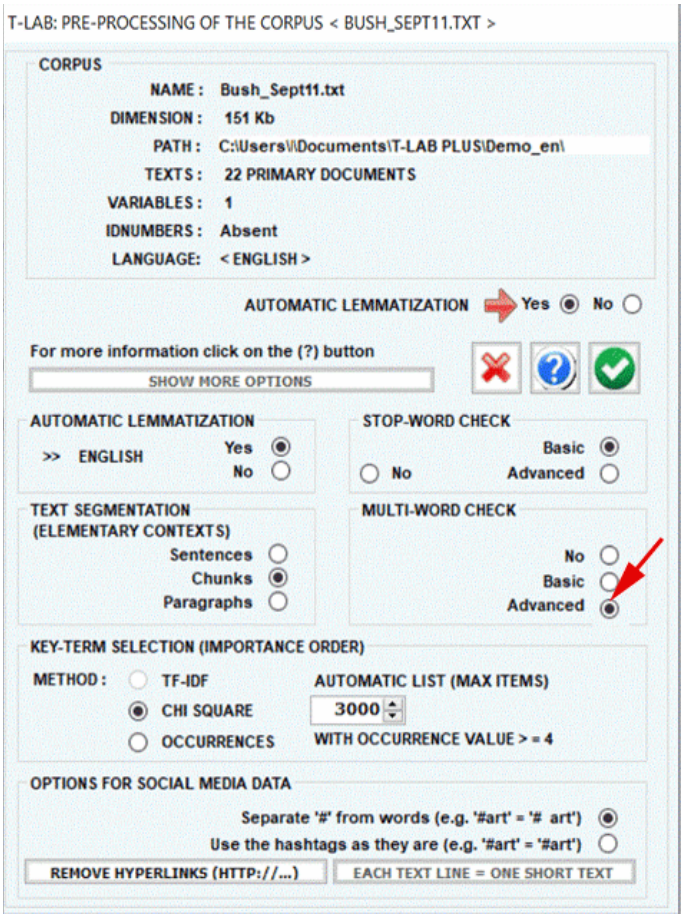
etc etc

By clicking on the "**Use this list…**" button, the user can produce an automatic and quick transformation of the multi-words present in a corpus in single strings that can be recognized

and classified by **T-LAB** (e.g. "secretary of state" turns into "secretary_of_state")

After running, this option generates a new file (**New_Corpus.txt**) which, properly renamed, can be analysed with **T-LAB**.

To verify/use Multiword lists during the **importation of a new corpus** the user has to select the "**Advanced**" option in the following form:

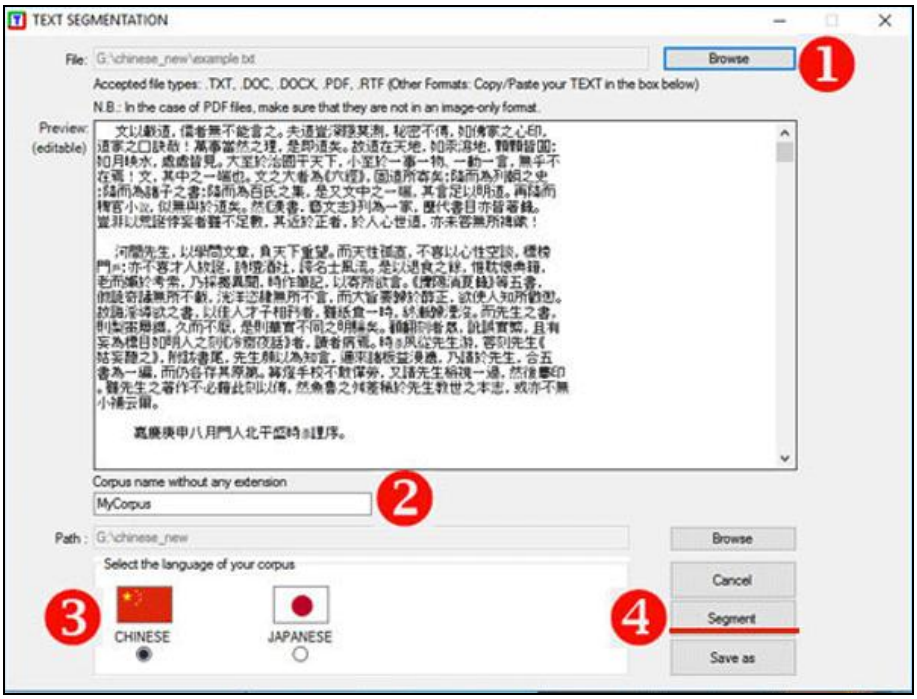


Word Segmentation

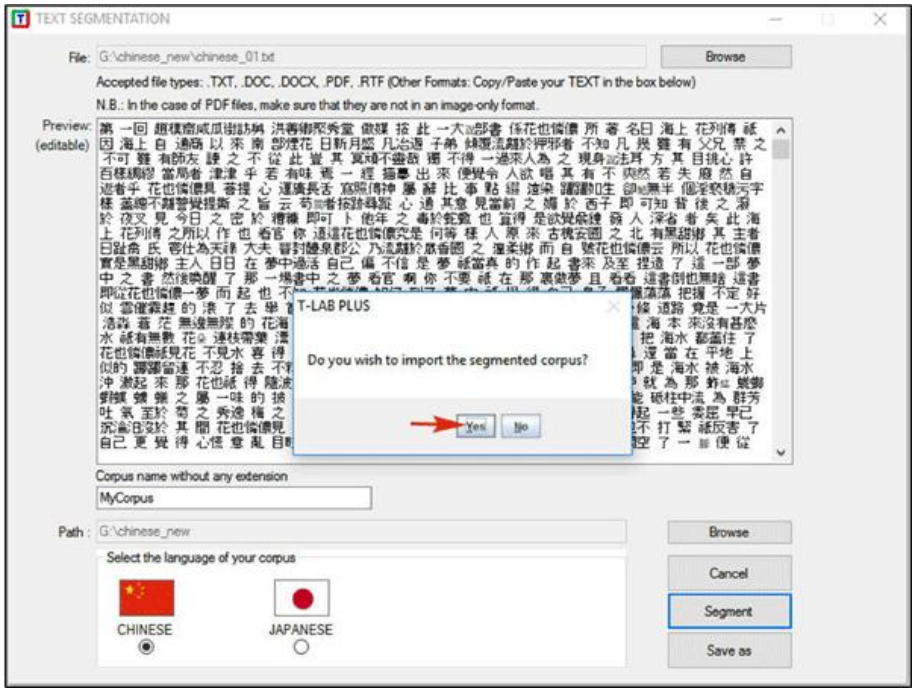
This **T-LAB** tool can be used before importing any Chinese or Japanese text (*) which has no delimiters (i.e. blank spaces and/or punctuation marks) between words.
(*) Either a single document or a corpus made up of various texts which include variable values can be processed.

Its use is very simple (see the below picture):

(1) select any file;
(2) choose the project name;
(3) select the language of your text;
(4) click on 'Segment'.
As a result, blank spaces will be added between words.



Subsequently, if you want to proceed with the importation process, just answer 'yes' to the question "Do you wish to import the segmented corpus?" (see the below picture).



N.B.: When you need to prepare a corpus made up of various texts which include coding lines (i.e. categorical variables), we recommend you to proceed as follows:

1- 'Gather' the unsegmented texts (*) through the Corpus Builder tool and then 'Save' your corpus file;

2 - Import the corpus just created through the Word Segmentation tool, then proceed as explained above.

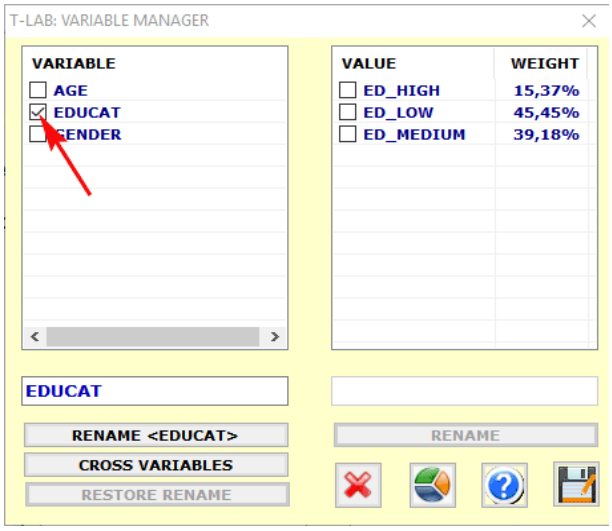
(*) This means that, in order to prepare your corpus, you don't need to segment each single file in advance.

OTHER TOOLS

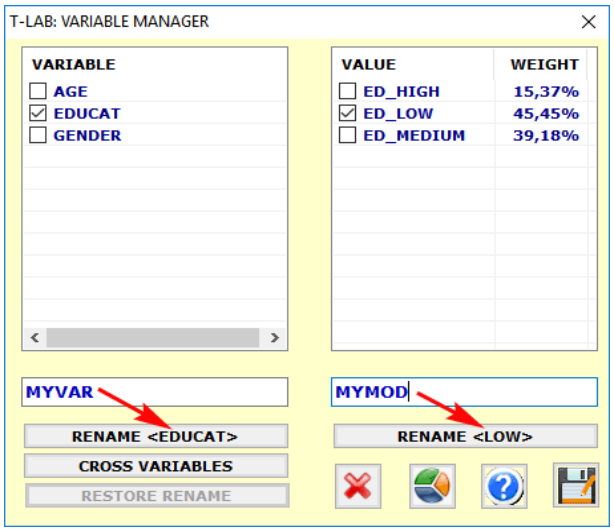
Variable Manager

This option, which is enabled only when the corpus includes partitions (i.e. variables and categories) allows six kinds of operations:

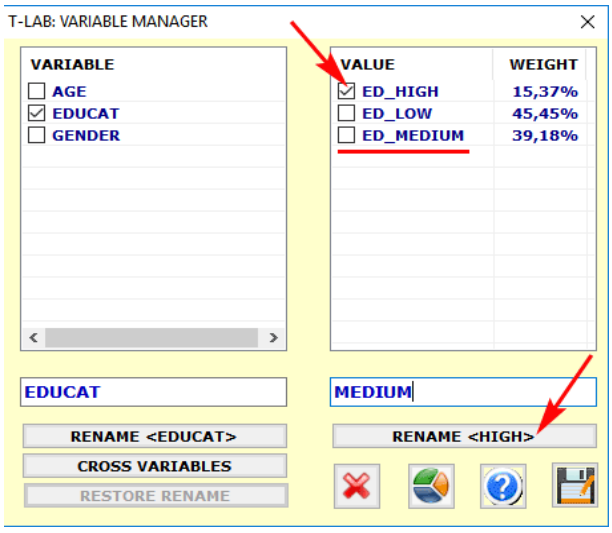
a) **verify** the categories of each variable;



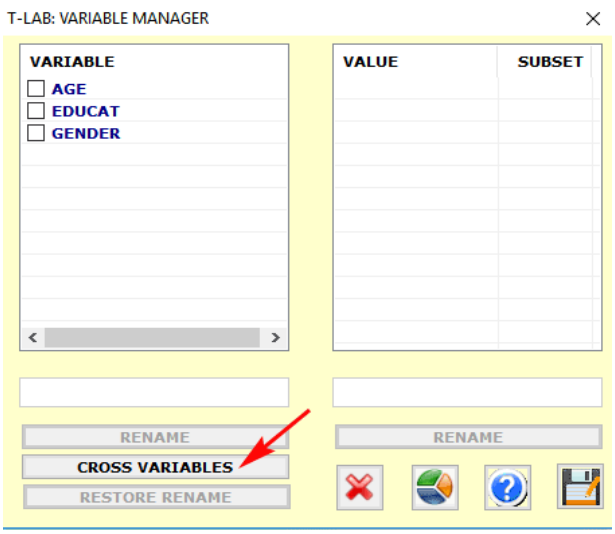
b) **rename** variables and categories;

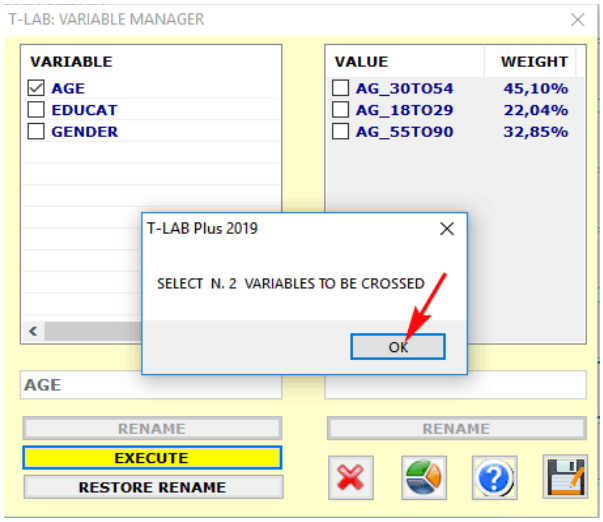


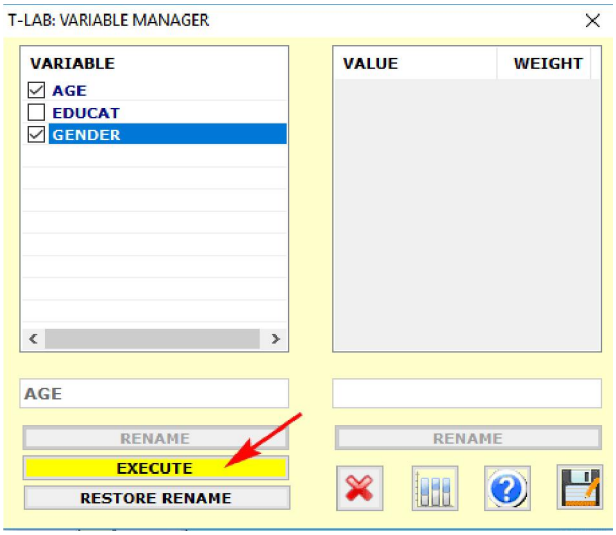
c) **group** two or more categories by assigning them the same label;

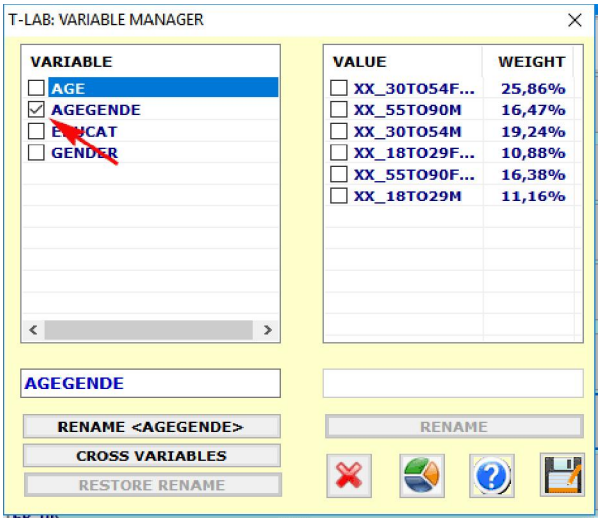


d) create a **crossed variable** to be available for further analyses.

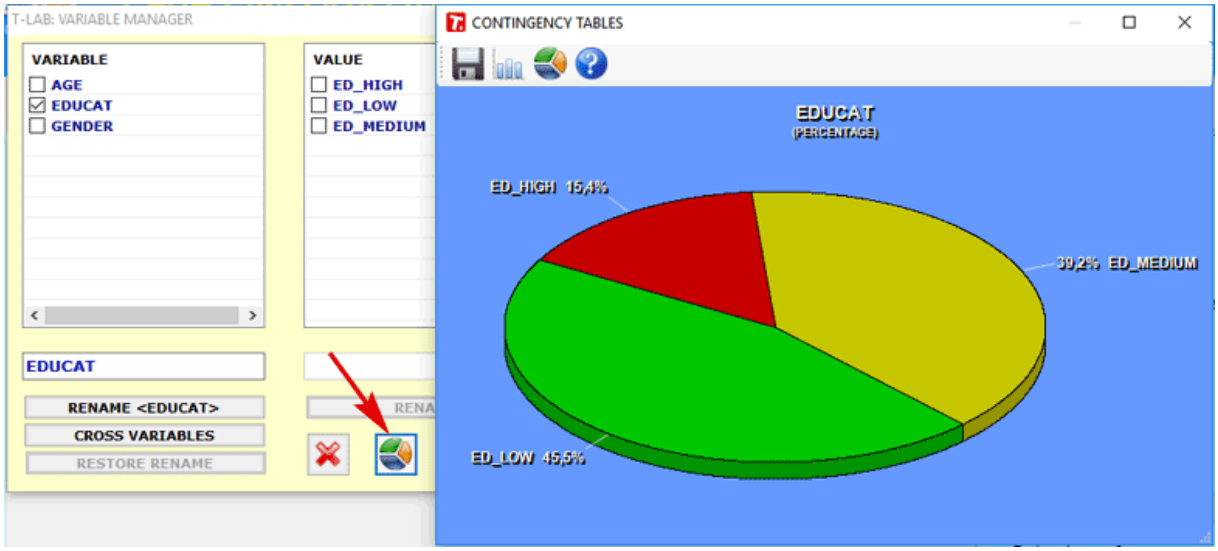


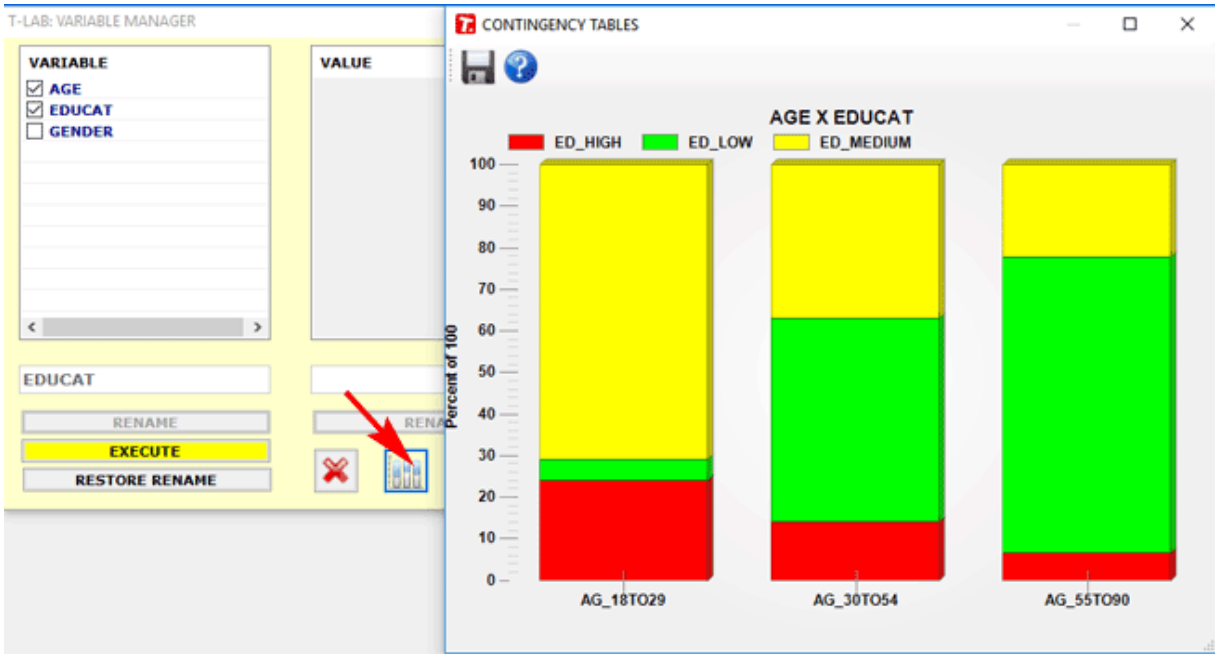






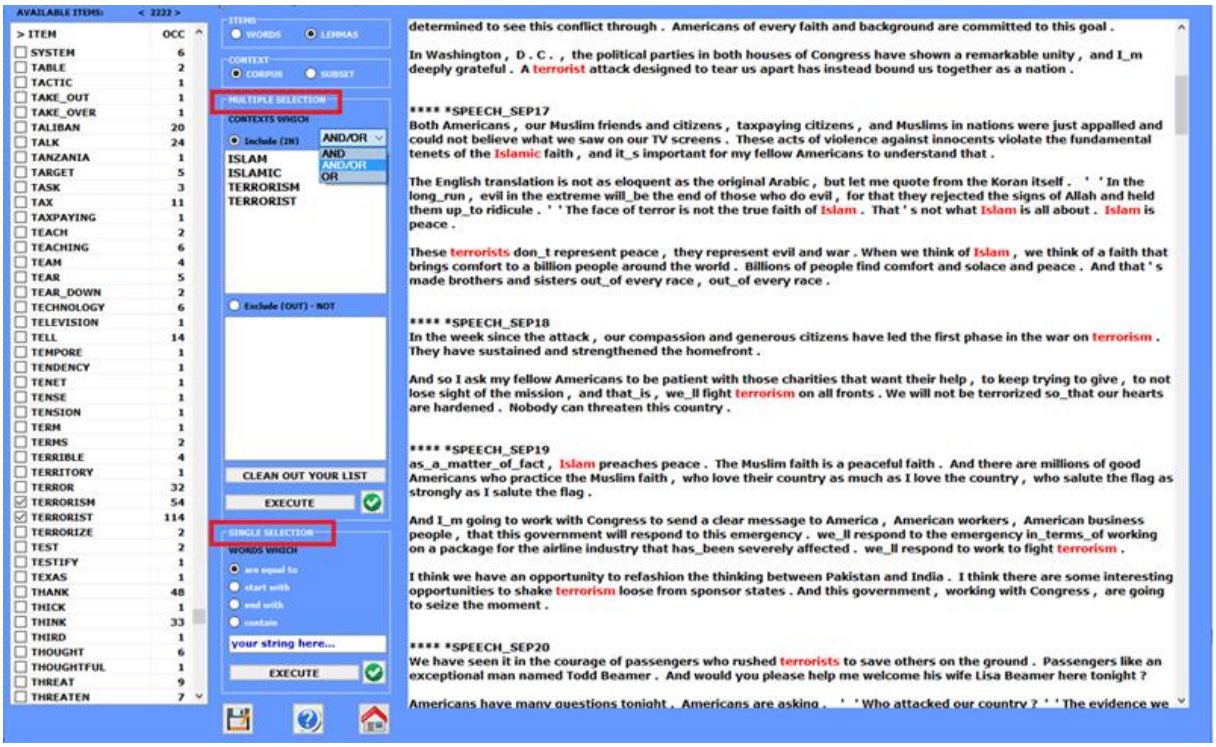
e) create some **charts**.





Advanced Corpus Search

This **T-LAB** tool allows us to extract and export all text segments (i.e. sentences or paragraphs) which match single or multiple selections of words, this either within the entire corpus or within a subset of it.



Its use is very intuitive: just select the desired options in each box (see below).

In the case of ‘multiple’ selections, each word is added by clicking the corresponding item of the table.

In the case of a ‘single’ selection, the target string must be typed in the appropriate box.

After clicking ‘execute’, the results are displayed on the right side of the window and can be saved as an .rtf file.

As the file created by **T-LAB** includes all the user’s coding, it can be also imported and analysed as a sub-corpus consisting of ‘n’ selected sentences or paragraphs.

N.B.: This option is enabled only when working on a corpus which has already been imported and a list of key words has been selected (see Analysis Settings).

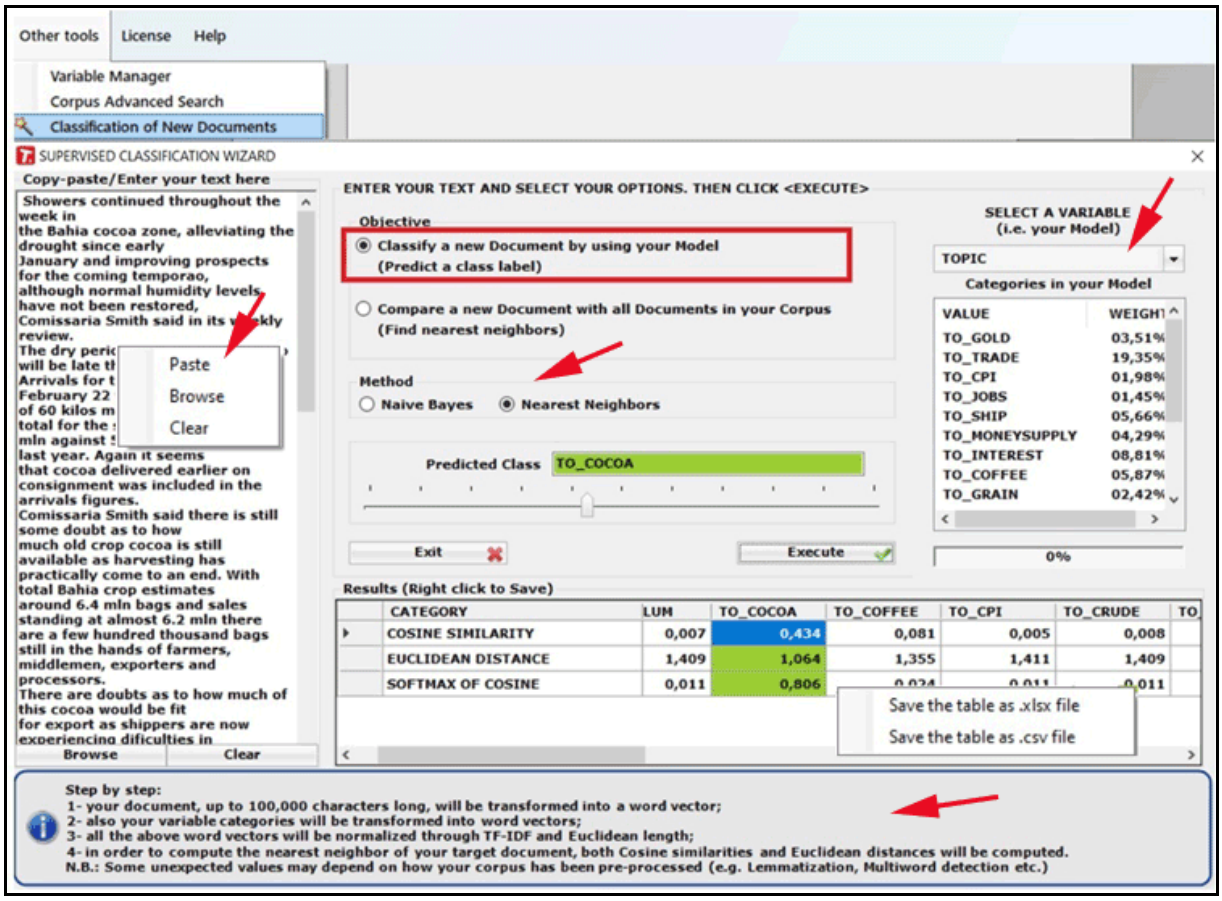
Classification of New Documents

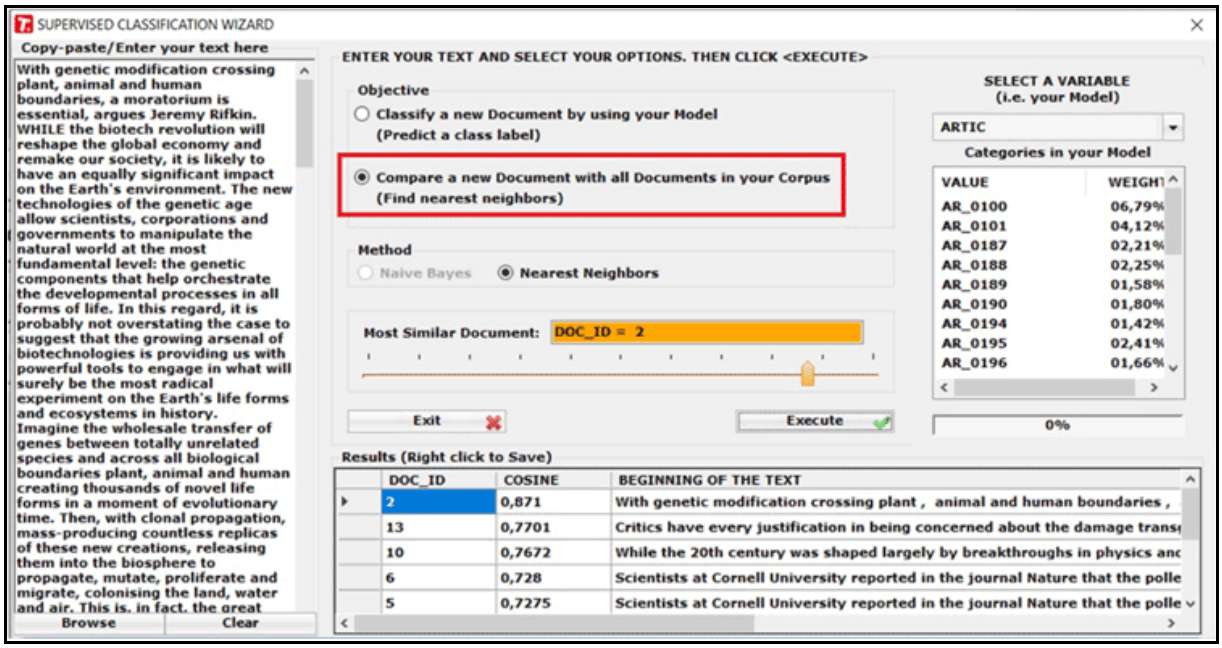
This tool, which is very easy to use, allows one to easily classify new documents according to a pre-existing model (i.e. any categorical variable) and also to compare any new document with all documents included in a corpus already analysed.

To this purpose, the following steps are required:

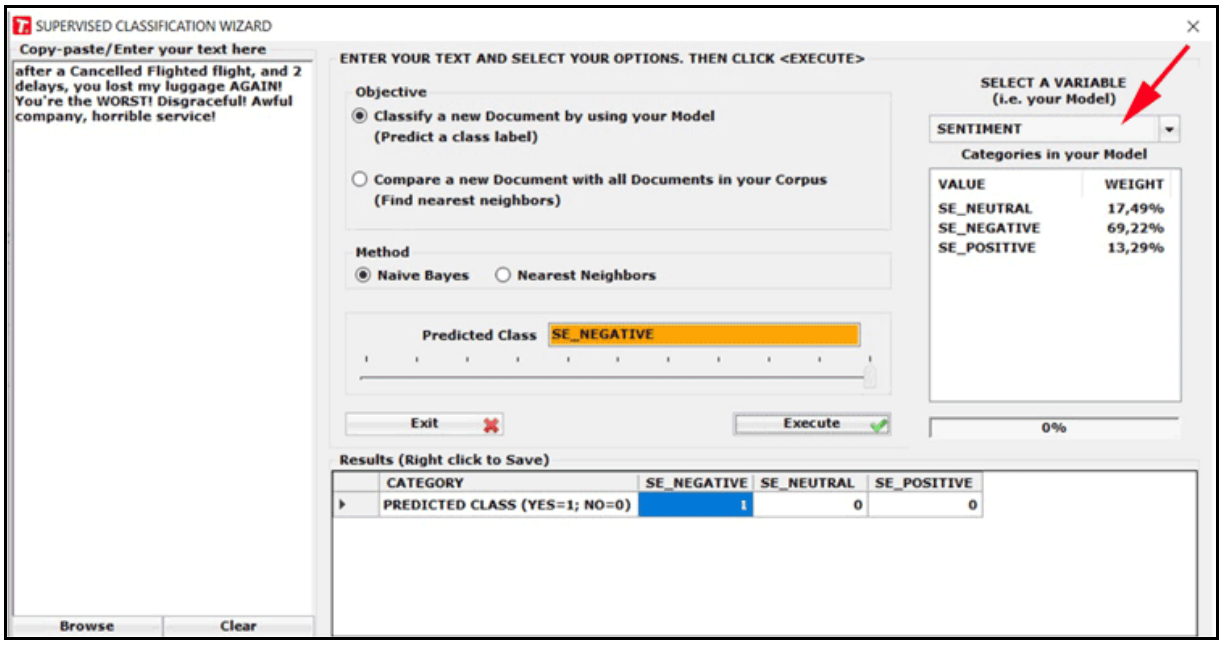
- enter a new document in the appropriate box;
- select a categorical variable to be used as a 'model';
- choose the desired 'objective' and a 'method';
- click 'execute'.

All results can be exported by using the right click options (see the below pictures).





When using this tool for sentiment analysis purpose, your corpus must include an appropriate categorical variable (see the below below).



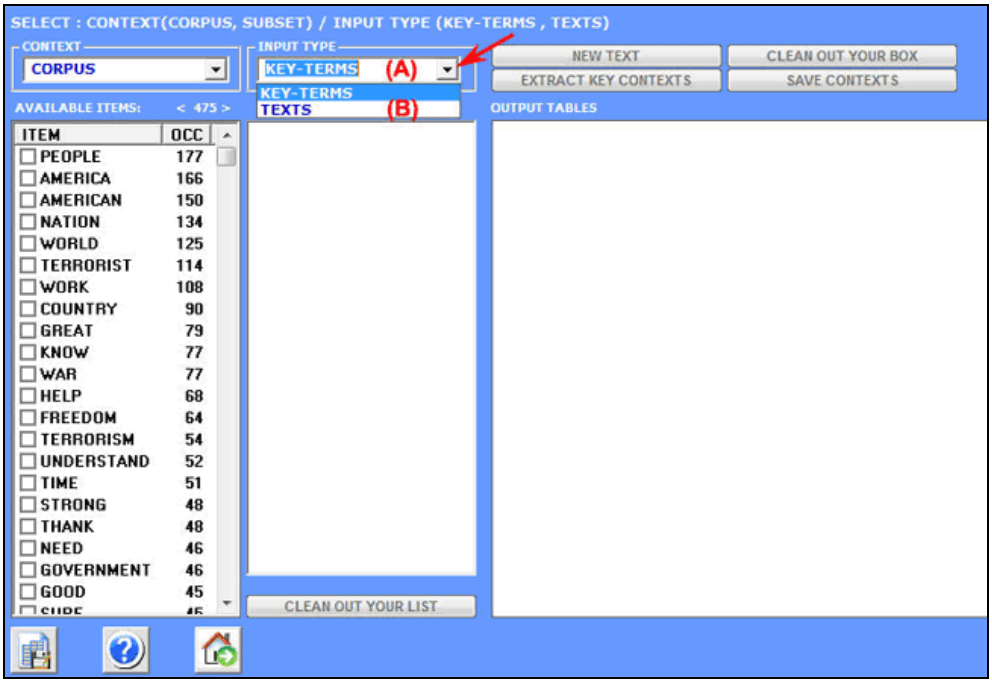
N.B.: When the user wishes to classify a dataset of new documents by using a supervised method, the dataset must be imported by T-LAB and then analysed by using a previously generated dictionary. To this purpose, the 'Thematic Document Classification' can be used, both for generating a dictionary of categories (i.e. unsupervised method) and for performing a supervised classification.

Key Contexts of Thematic Words

According to the type of input, this **T-LAB** can be used for two different purposes:

a) to extract lists of meaningful context units (i.e. elementary contexts or short documents) which allow us to deepen the thematic value of specific **key terms**;

b) to extract the context units which are the most similar to sample **texts** chosen by the user.



Here are some explanations for the two above cases.

Case (A);

Unlike **Concordances**, which allows the extraction of all elementary contexts in which the selected key words are just present (occurrences), and unlike **Word Associations**, which allows the extraction of all elementary contexts in which the selected key words are in matching pairs (co-occurrences), this tool allow us to extract the elementary contexts in which each selected key word is associated with other words (multiple co-occurrences) defining its thematic field.

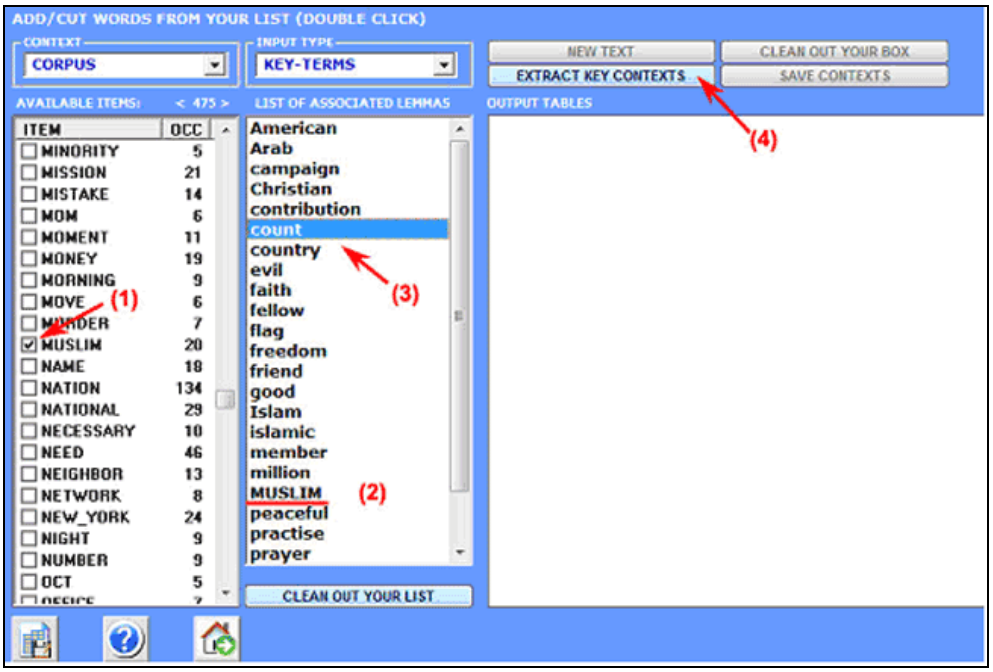
It works in the following way:

1- the user chooses a thematic word "X" (see "Muslim" below);

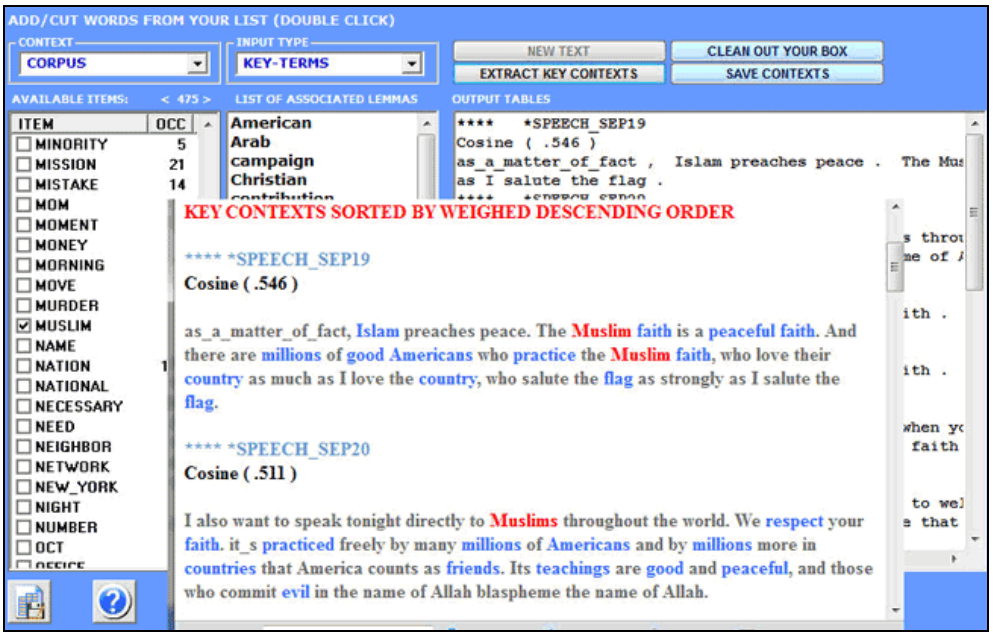
2- T-LAB proposes a list of words (max. 50) whose co-occurrence values with "X" are the most significant;

3- the user can remove irrelevant items from the list provided (just double click each item);

4- after clicking 'Extract Key Contexts' **T-LAB** assumes that the user list is a query vector and computes its **association indexes** (i.e. cosine coefficients) with all the elementary contexts of the corpus or of the selected corpus **subset**.



The output provided, both in **HTML** and TXT format, contain a list of the most significant key-contexts of "X", listed according to the descending order of their association indexes.



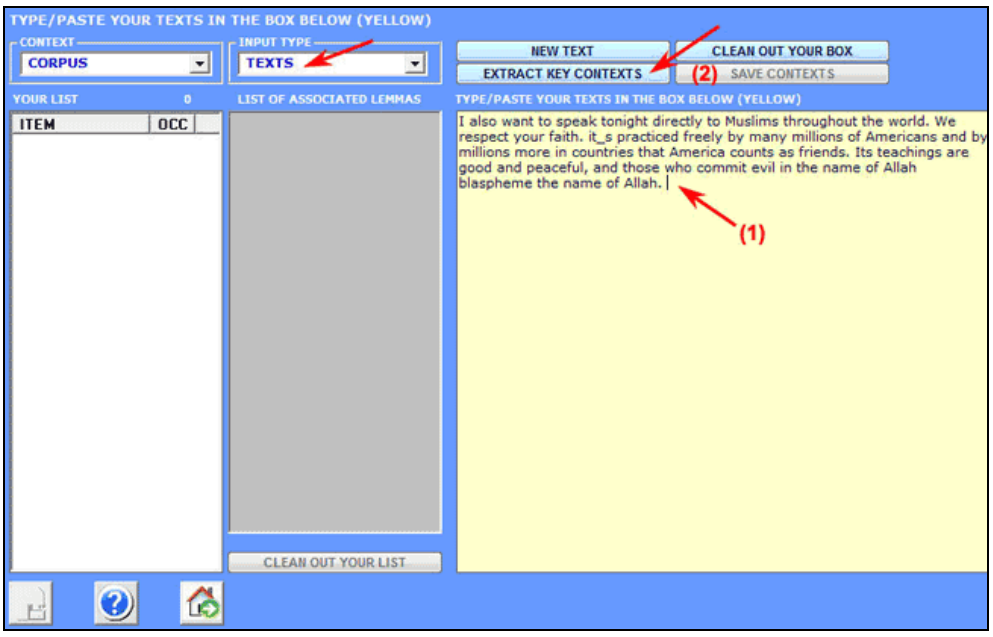
The above 1-4 steps can be reiterated for "n" sample words.

Case (B)

It works in the following way:

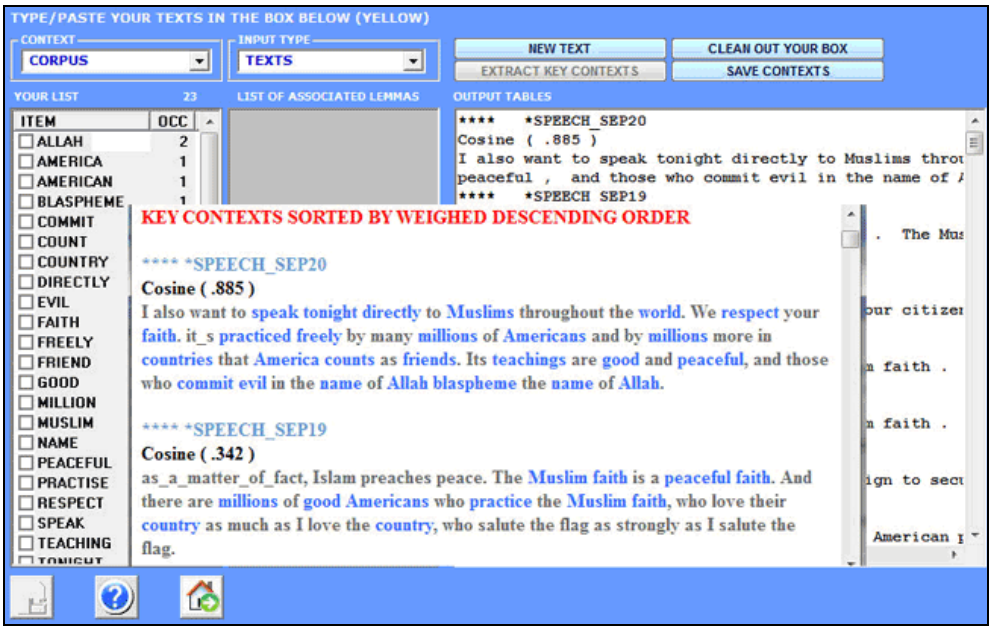
1- the user copy/paste a text (Max 5,000 characters) in the appropriate box;

2- after clicking the 'extract key contexts' button, **T-LAB** transforms the input text into a query vector and computes its **association indexes** (i.e. cosine coefficients) with all the elementary contexts of the corpus or of the selected corpus **subset**;



The output provided, both in **HTML** and TXT format, contain a list key-contexts which are the most similar to the input text.

N.B.: In such a case the similarity measure doesn't take into account multi-words the strings of which, either with or without the underscore ('_') character, do not correspond to the analysed text.



The above 1-2 steps can be reiterated for "n" sample texts.

Export Custom Tables

N.B.: The pictures shown in this section have been obtained by using a previous version of **T-LAB**. These pictures look slightly different in **T-LAB 10**, but the functionalities of the software are the same.

This option allows us to create, to explore and to export three kinds of tables:

a) those with the **occurrence** values of several lexical units within the corpus subsets defined by some variable (rectangular matrices);

b) those with the **co-occurrence** values of lexical units (square matrices) within the corpus or within the corpus subsets;

b) those with the **co-occurrence** values of lexical units units within all documents (sparse matrices using indexes).

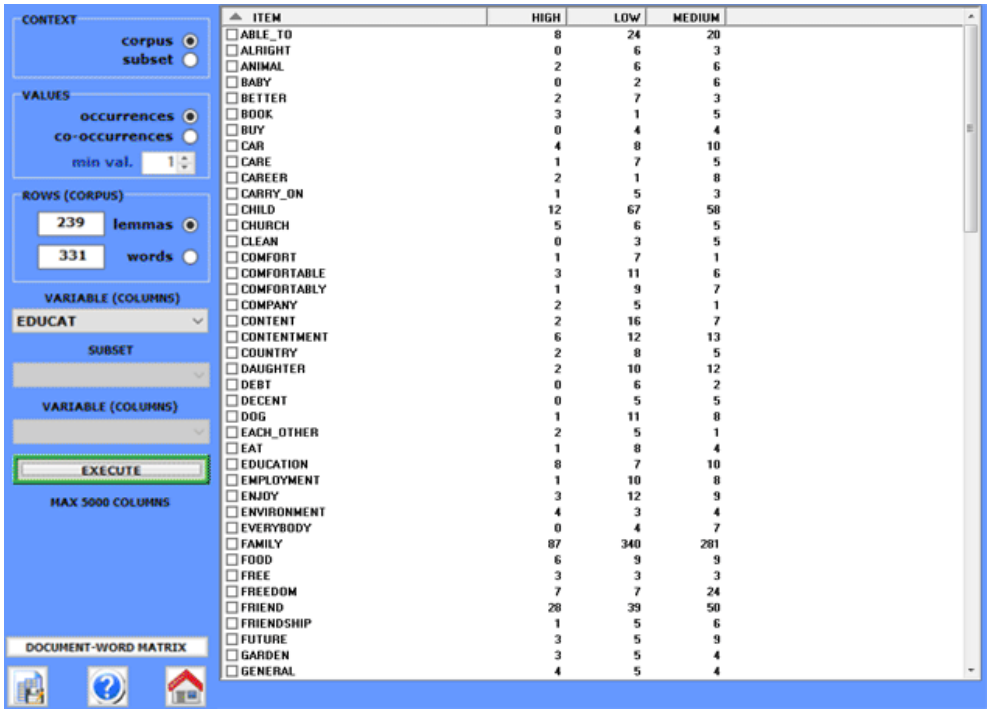
The maximum sizes of such tables are respectively: a) 10,000 rows by 150 columns; b) 5,000 rows by 5,000 columns; c) 30,000 documents by 10,000 lexical units.

The use of this function is very intuitive.

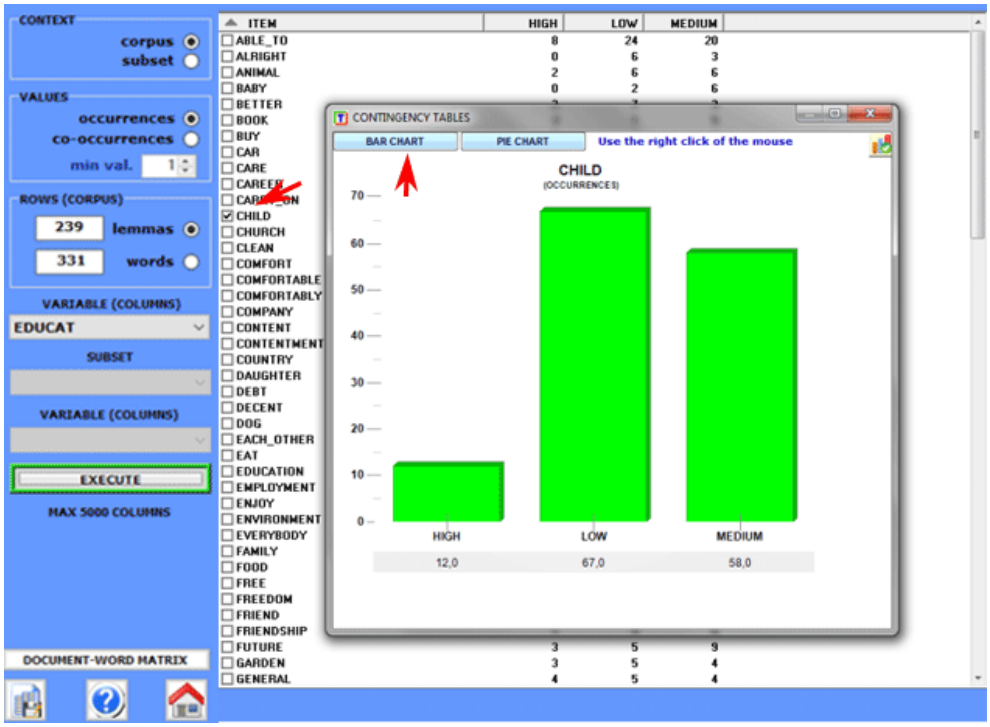
When dealing with simpler cases, you are asked to select the variable one whose categories will constitute the columns of the output table.

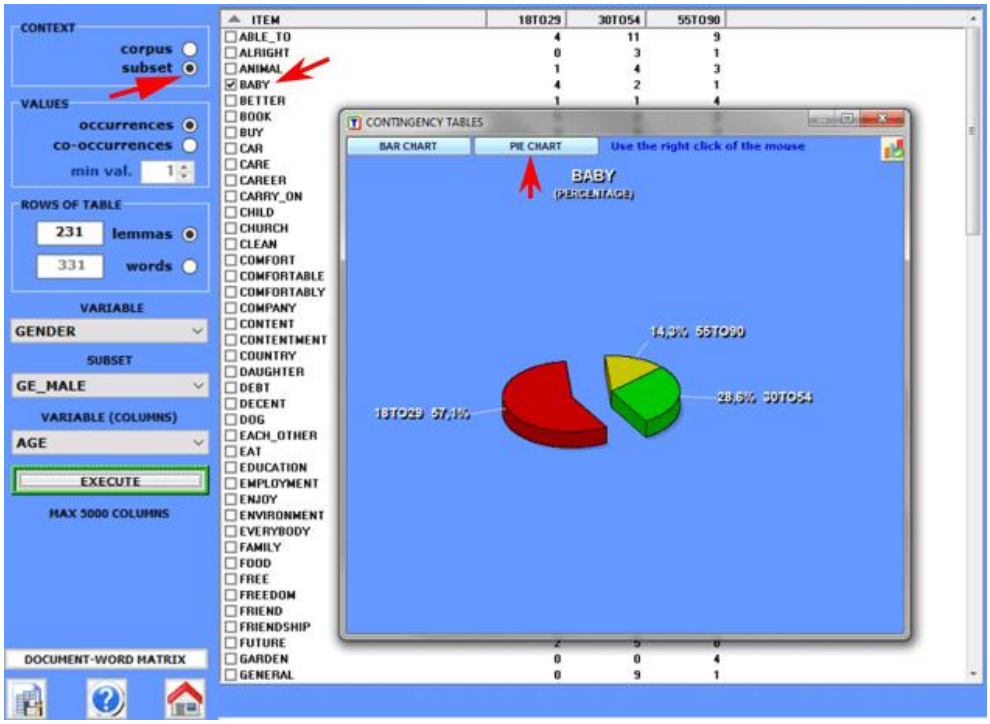
When dealing with more complex cases, you are asked to select one variable and a subset.

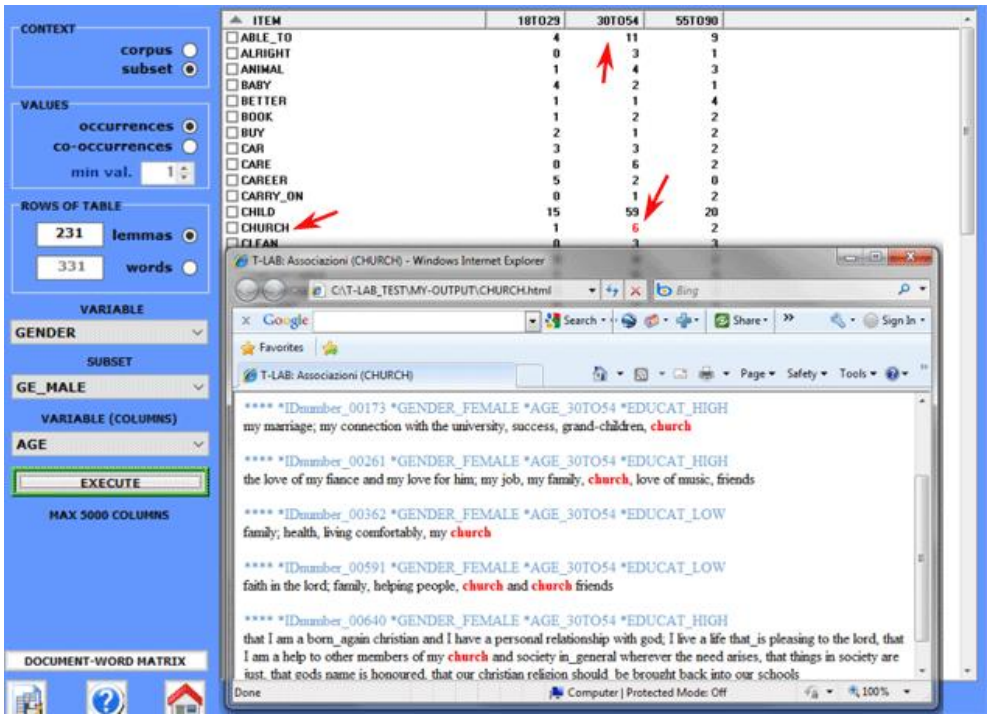
All contingency tables allow us to create various **charts**.
Moreover, by clicking on specific cells of a table, it is possible to create a **HTML file** including all elementary contexts where the word in row is present in the corresponding subset (see below).



ITEM	HIGH	LOW	MEDIUM
ABLE_TO	8	24	20
ALRIGHT	0	6	3
ANIMAL	2	6	6
BABY	0	2	6
BETTER	2	7	3
BOOK	3	1	5
BUY	0	4	4
CAR	4	8	10
CARE	1	7	5
CAREER	2	1	8
CARRY_ON	1	5	3
CHILD	12	67	58
CHURCH	5	6	5
CLEAN	0	3	5
COMFORT	1	7	1
COMFORTABLE	3	11	6
COMFORTABLY	1	9	7
COMPANY	2	5	1
CONTENT	2	16	7
CONTENTMENT	6	12	13
COUNTRY	2	8	5
DAUGHTER	2	10	12
DEBT	0	6	2
DECENT	0	5	5
DOG	1	11	8
EACH_OTHER	2	5	1
EAT	1	8	4
EDUCATION	8	7	10
EMPLOYMENT	1	10	8
ENJOY	3	12	9
ENVIRONMENT	4	3	4
EVERYBODY	0	4	7
FAMILY	87	340	281
FOOD	6	9	9
FREE	3	3	3
FREEDOM	7	7	24
FRIEND	28	39	50
FRIENDSHIP	1	5	6
FUTURE	3	5	9
GARDEN	3	5	4
GENERAL	4	5	4




CONTEXT
corpus
subset
VALUES
occurrences
co-occurrences
min val. 1
ROWS OF TABLE
231 lemmas
331 words
VARIABLE
GENDER
SUBSET
GE_MALE
VARIABLE (COLUMNS)
AGE
EXECUTE
MAX 5000 COLUMNS
DOCUMENT-WORD MATRIX
ITEM 18TO29 30TO54 55TO90
CONTINGENCY TABLES
BAR CHART
PIE CHART
Use the right click of the mouse
BABY
(PERCENTAGE)
14,3% 55TO90
28,6% 30TO54
18TO29 57,1%


CONTEXT
corpus
subset
VALUES
occurrences
co-occurrences
min val. 1
ROWS OF TABLE
231 lemmas
331 words
VARIABLE
GENDER
SUBSET
GE_MALE
VARIABLE (COLUMNS)
AGE
EXECUTE
MAX 5000 COLUMNS
DOCUMENT-WORD MATRIX
ITEM 18TO29 30TO54 55TO90
T-LAB: Associazioni (CHURCH) - Windows Internet Explorer
C:\T-LAB_TEST\MY-OUTPUT\CHURCH.html
T-LAB: Associazioni (CHURCH)
**** *IDnumber_00173 *GENDER_FEMALE *AGE_30TO54 *EDUCAT_HIGH
my marriage; my connection with the university, success, grand-children, church
**** *IDnumber_00261 *GENDER_FEMALE *AGE_30TO54 *EDUCAT_HIGH
the love of my fiance and my love for him; my job, my family, church, love of music, friends
**** *IDnumber_00362 *GENDER_FEMALE *AGE_30TO54 *EDUCAT_LOW
family; health, living comfortably, my church
**** *IDnumber_00591 *GENDER_FEMALE *AGE_30TO54 *EDUCAT_LOW
faith in the lord; family, helping people, church and church friends
**** *IDnumber_00640 *GENDER_FEMALE *AGE_30TO54 *EDUCAT_HIGH
that I am a born_again christian and I have a personal relationship with god; I live a life that_is pleasing to the lord, that I am a help to other members of my church and society in_general wherever the need arises, that things in society are just, that gods name is honoured, that our christian religion should be brought back into our schools

In order to export any sparse matrix document x word just click the appropriate button (i.e. 'Document-Word Matrix').

In this case, the available outputs are two:

The first (Sparse_Matrix.csv) has the following format:

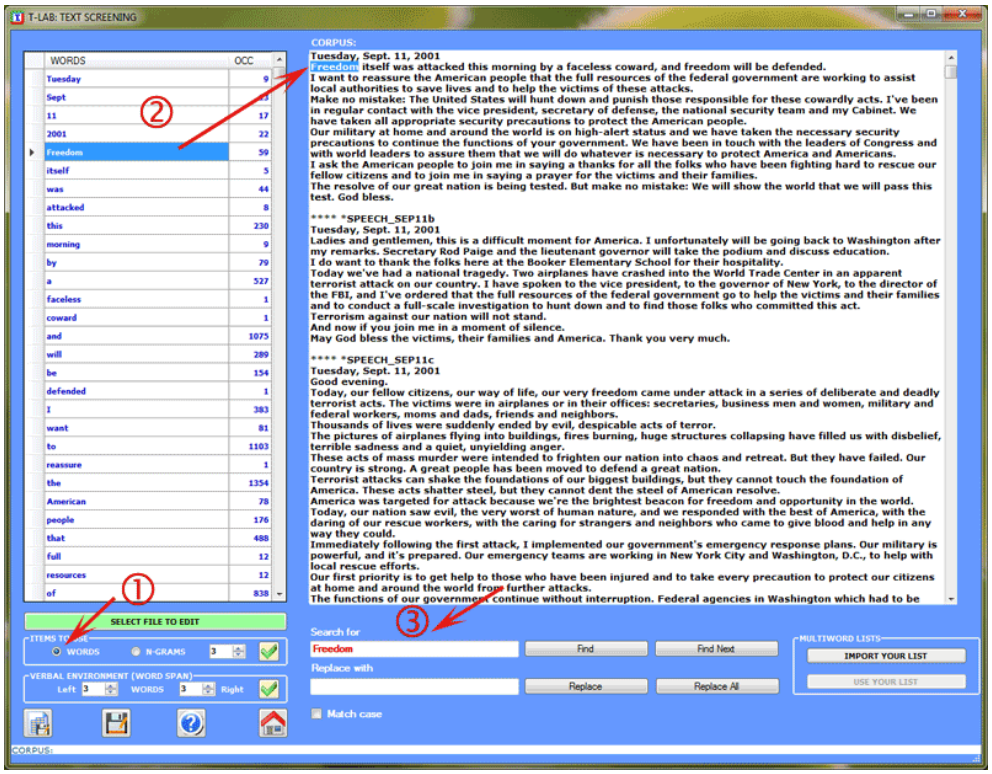
Doc_Index; Word_Index; Word_Occ
00001; 1; 12
00001; 2; 5
…..

The second (Word_Indexes.csv) has the following format:

Word_Index; Word_Label
1; abroad
2; accept
…..

Editor

In **T-LAB 10** some editing functions for .txt files are included in the **Text Screening** tool (see below).



Import-Export Identifiers List

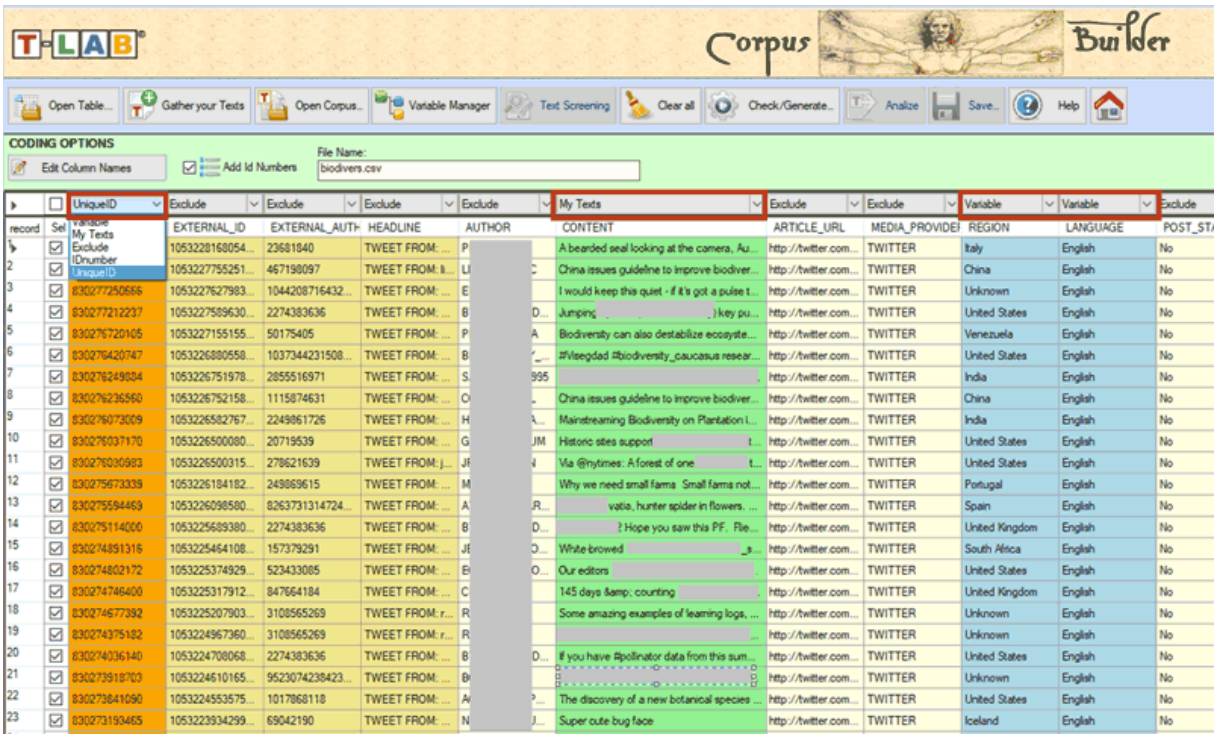
In **T-LAB** a unique identifier is a categorical variable with a distinct value for every document (or case).

A list of unique identifiers may consist of any type of alphanumerical strings (e.g. Interviewee ID numbers, proper names, geographical names, names of books etc.) up to 50 characters long and without blank spaces.

As unique identifiers are singular, it's impossible to perform any data analysis on them. Instead, they are used to identify results in the software outputs.

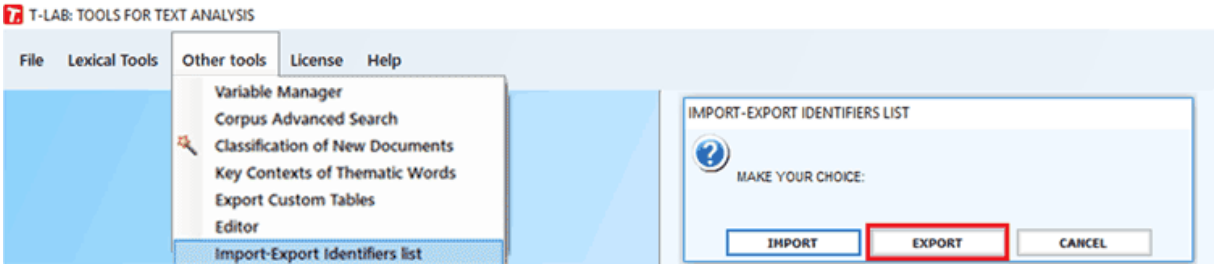
In **T-LAB**, through the import/export options, any unique identifier list can be modified at any moment.

When importing data in tabular format, the unique identifiers must be in the first column, like the following example concerning Twitter messages.



In the other cases (i.e. document collections which are not in tabular format) the recommend procedure is the following:

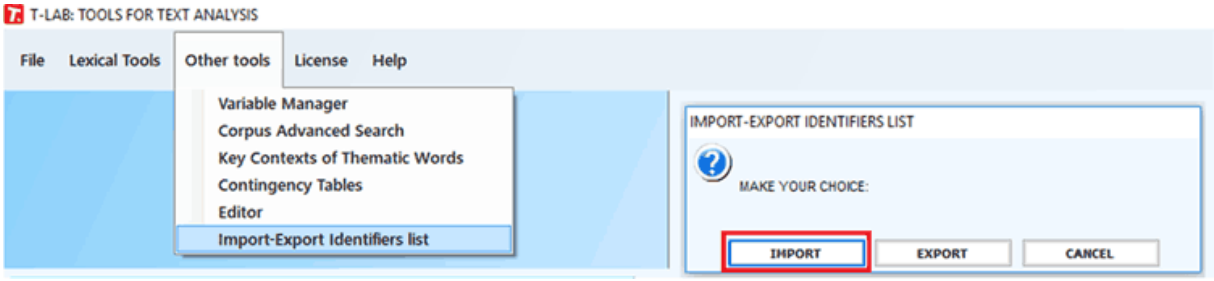
1-Import your corpus first;
2-Export the list of the identifiers automatically created by **T-LAB.**



3-Edit and modify the CSV file created by **T-LAB** (i.e. just modify the 'MyIdentifier' values according to your needs. See picture below).

MyID	MyIdentifier
1	TOBEREPLACED00001
2	TOBEREPLACED00002
3	TOBEREPLACED00003
4	TOBEREPLACED00004
5	TOBEREPLACED00005
6	TOBEREPLACED00006
7	TOBEREPLACED00007
8	TOBEREPLACED00008
9	TOBEREPLACED00009
10	TOBEREPLACED00010
...	...

4- Import the CSV file which includes your reviewed unique identifiers.



GLOSSARY

Analysis Unit

The **analysis units** used in **T-LAB** are of two types: **lexical units** and **context units**.

A. the **lexical units** are **words** and multi-words, filed and classified on the basis of a criterion. More precisely, in the **T-LAB** database each lexical unit consists in a classified record with two fields: **word** and **lemma**. In the first field ("word"), the words are listed as they appear in the corpus, while in the second ("lemma") the labels attributed to lexical unit groups are listed and classified according to linguistic criteria (eg. **Lemmatization**) or by **dictionaries** and **semantic grids** defined by the user.

B. the **context units** are portions of text that the corpus can be divided into. More precisely, according to **T-LAB** logic, there can be three types of context units:

- **B.1** **primary documents**, which correspond to the "natural" subdivision of the corpus (eg. interviews, articles, answers to open-ended questions, etc.), that is the initial context defined by the user;
- **B.2** **elementary contexts**, which correspond to syntagmatic units (i.e. fragments, sentences, paragraphs) in which each primary document can be subdivided;
- **B.3** **corpus subsets**, which correspond to groups of primary documents which lead to the same "category" (eg. interviews with "men" or "women", articles in a specific year or a particular magazine and so on).

Association Indexes

In **T-LAB** the association indexes (or similarity coefficients) are used to analyse the **co-occurrences** of the **lexical units** (**LU**) inside the **elementary contexts** (**EC**), that is to analyse binary data of the presence/absence type.

For instance, given two **LU** and ten **EC**, we can obtain the following example:

	EC_1	EC_2	EC_3	EC_4	EC_5	EC_6	EC_7	EC_8	EC_9	EC_10
LU_1	1	0	1	1	1	0	1	0	1	1
LU_2	0	1	0	1	0	0	1	1	0	1

The same data can be represented in the following way:

	LU_2		
LU_1	*Present*	*Absent*	*Total*
Present	3	4	7
Absent	2	1	3
Total	5	5	10

Generalizing and using the letters of the alphabet:

	LU_2		
LU_1	*Present*	*Absent*	*Total*
Present	a	b	$a + b$
Absent	c	d	$c + d$
Total	$a + c$	$b + d$	n

The formulas corresponding to the six association indexes used by **T-LAB** are the following:

Jaccard

$$\frac{a}{a+b+c}$$

Dice

$$\frac{2a}{2a+b+c}$$

Cosine

$$\frac{a}{\sqrt{(a+b)} \times \sqrt{(a+c)}}$$

Equivalence

$$\frac{a^2}{(a+b) \times (a+c)}$$

Inclusion

$$\frac{a}{\mathrm{Min}((a+b),(a+c))}$$

Mutual Information

$$\mathrm{Log}\,\frac{a/N}{(a+b) \times (a+c)}$$

If, for example, we have calculated association coefficients of co-occurrence relationships concerning ten **LU**, we can obtain a table like the following:

	LU_1	LU_2	LU_3	LU_4	LU_5	LU_6	LU_7	LU_8	LU_9	LU_10
LU_1		0,067	0,048	0,286	0,154	0,077	0,060	0,309	0,231	0,077
LU_2	0,067		0,269	0,134	0,000	0,072	0,056	0,072	0,072	0,072
LU_3	0,048	0,269		0,048	0,156	0,104	0,040	0,052	0,052	0,156
LU_4	0,286	0,134	0,048		0,077	0,000	0,060	0,154	0,000	0,077
LU_5	0,154	0,000	0,156	0,077		0,667	0,000	0,000	0,000	0,333
LU_6	0,077	0,072	0,104	0,000	0,667		0,000	0,000	0,000	0,417
LU_7	0,060	0,056	0,040	0,060	0,000	0,000		0,129	0,129	0,000
LU_8	0,309	0,072	0,052	0,154	0,000	0,000	0,129		0,167	0,083
LU_9	0,231	0,072	0,052	0,000	0,000	0,000	0,129	0,167		0,000
LU_10	0,077	0,072	0,156	0,077	0,333	0,417	0,000	0,083	0,000	

In effect **T-LAB** produces and analyses analogous tables of N x N dimensions (where N can correspond to hundreds of columns), both using **Multidimensional Scaling** and **Cluster Analysis.**

Similar tables are also used to calculate **second order similarities** between pairs of keywords (see the **Word Associations** tool).

Chi-Square

Chi-square is a statistical test to check if the frequency values obtained by a survey and recorded in some cross-table, are significantly different from the theoretical ones.

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).

For example, in order to verify the significance of a word ("x") occurrences within a context unit ("A") the test is applied to a table as follows:

	Context "A"	Other Contexts		
Word "x"	15	198	213	N_j
Other Words	572	2420	2992	
	587	2618	3205	N_{ij}
	N_i			

The chi-square formula, in its simplified version, is the following:

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

where "O" and "E" stand respectively for the observed frequencies and the expected ones.

For each cell, the expected (E) occurrences are calculated as follows: (Ni x Nj)/Nij.

Following the above example the CHI value is equal to 19.38.

Since it is greater than the critical value, the null hypothesis (absence of meaningful difference) can be rejected.

Cluster Analysis

Cluster analysis is a set of statistical techniques the aim of which is to detect groups of objects with two complementary features:

A - High internal (within cluster) homogeneity;

B - High external (between cluster) heterogeneity.

In statistical language, the characteristics “A” and “B” respectively correspond to the within and between cluster variance.

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

- **Hierarchical methods**, whose algorithms rebuild the whole hierarchy of the objects under analysis (the so called "tree"), whether in an ascending order or in a descending order;

- **Partitioning methods**, where the user defines beforehand the cluster numbers in which the set of objects under analysis is divided.

T-LAB uses both types of algorithms.

In particular:

- the **Co-Word Analysis and Concept Mapping** option uses a hierarchical method;
- the **Cluster Analysis** option allows the use of three different methods: two hierarchical and one partitioning;
- the **Thematic Analysis of Elementary Contexts** and **Thematic Document Classification** options use a bisecting K-means algorithm .

Some of the publications quoted in the **Bibliography** provide further information on the general aspects of the various methods (Bolasco S., 1999; Lebart L., A. Morineau, M. Piron, 1995), the specific aspects relating to the Hdbscan (Campello R. J. G. B., Moulavi D., Zimek A. & Sander J. , 2015) and the bisecting K-means method (Steinbach, M., G. Karypis, V. Kumar, 2000; Savaresi S.M., D.L. Boley, 2001).

Coding

Before importing the corpus, the user can insert coding lines at the beginning of each **context unit** that he wants to classify using one or more **variables**.

As a rule the **classified** context units correspond to the **primary documents**.

Context Unit

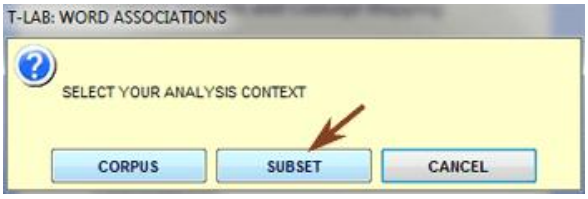
See **analysis unit**.

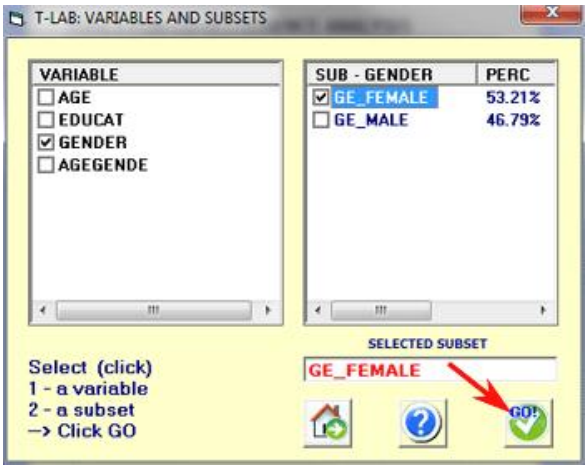
Corpus and Subsets

A **corpus** is a collection of one or more texts selected for analysis.

Each corpus **subset** is defined by means of a **category of a variable**.

T-LAB makes it possible to explore and to analyse the relationships between the analysis units of the whole **corpus** or its **subsets**.





Some **corpus** examples:

- a single text or document concerning any subject;
- a set of articles taken from the press, concerning the same subject;
- one or more interviews carried out inside the same research project;
- a set of answers to an open-ended question;
- one or more focus group transcripts.

Some **subset** examples:

- one or more chapters of a book;
- one or more newspaper articles published in the same year;
- one or more interviews with the same people category;
- a subset of answers to an open-ended question.

N.B.: Further corpus subsets are the "**thematic clusters**" of documents or elementary contexts obtained by using the corresponding **T-LAB** tools.

In the case of a corpus made up of more than one text, in order to make it **a set correctly analysable**, it is required that all of its parts have two features that make them comparable:

a) a thematic and/or contextual homogeneity of their content;

b) a balanced relationship between their dimensions, both in terms of occurrences and in terms of Kbytes.

In **T-LAB** logic, the corpus is a **database** set up in **records** and **fields**. More precisely, records are made up of recorded entities (texts, text segments, words) and fields are made up of variables used to classify the different entities (text authors, reference contexts, theme types, etc.).

See **Corpus Preparation**.

Correspondence Analysis

Correspondence Analysis is a factorial analysis technique applied to the study of **data tables** whose cells contain either frequency values (real positive numbers) or presence-absence values ("1" or "0").

Like all factorial analysis techniques, correspondence analysis allows the extraction of new variables - the **factors** - with the property of summarizing in a organized way the significant information contained in the countless data tables cells; furthermore, this analysis technique allows the creation of graphs showing - in one or more spaces - the points that detect the **objects** in rows and columns, that is - in our case - the linguistic entities (words, lemmas, texts segments and texts) with the respective source features.

In geometrical terms, each factor sets up a spatial dimension - that can be represented as an axis line - whose center (or barycentre) is the value "0", and that develops in a bipolar way towards the negative (-) and positive (+) end, so that the objects put on opposite poles are the most different, almost like the "left" wing and the "right" wing on the political axes.

In **T-LAB** the analysis results are summarized through graphs that allow the evaluation of the relationships of proximity/distance - or rather similarity/dissimilarity - between the considered objects.

Furthermore, **T-LAB** shows measures (i.e. **Absolute Contributions** and **Test Value)** that help to understand the **poles of factors** that set up similarities/dissimilarities between the considered objects.

Data Tables

Data tables (or **matrices**) are made up of rows and columns, and of values recorded in the respective cells. They allow us to synthesize - in an orderly manner - either the observations to be analysed (input), or the results obtained by the analyses (output).

For more than one reason, statisticians say that a successful analysis is obtained only through the building of a "good table".

In **T-LAB**, depending on the types of analyses, there are three different tables, corresponding to as many ways of building crossings among rows and columns:

- words (or lemmas) in rows and variable categories in columns;
- context units (i.e. documents or elementary contexts) in rows and words (or lemmas) in columns;
- words (or lemmas) in rows and columns.

Disambiguation

Disambiguation is an operation that tries to resolve word sense **ambiguity** cases, particularly the ones ascribed to **homographs**, that is words with the same **graphic form** but different meanings.

In **T-LAB 10** some disambiguation functions are implemented in the **Text Screening** tool. Moreover, during the import stage, **T-LAB** recognizes and distinguishes three kinds of linguistic objects:

- proper nouns;
- **multiwords** (compound words and idioms);
- compound tenses of verbs.

In any case, **T-LAB** uses lists in its database which have been built and tested to limit the most frequent cases of ambiguity (**effectiveness** criterion) and to moderate processing times (**efficiency** criterion).

Dictionary

T-LAB dictionaries are tables or files which contain classification schemes for **lexical units** (i.e. words).

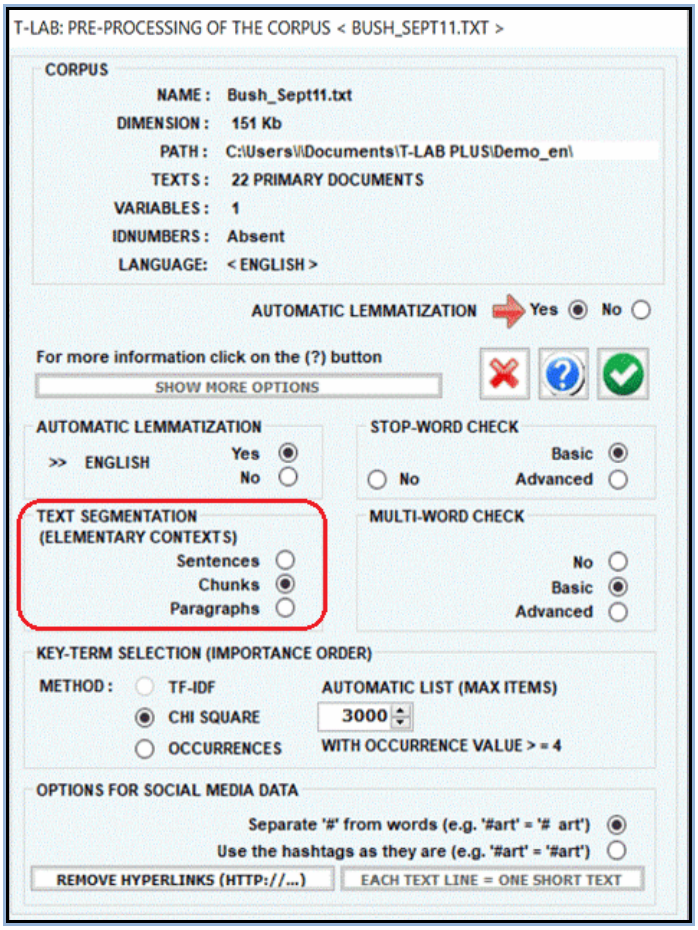
The classification schemes, and so the dictionaries, can be either **linguistic-based** (a) or **thematic-based** (b). Both can be exported and customized.

In the case of 'a' (e.g. rename or group the items of the key-word list) the user can refer to the **Dictionary Building** tool.

In the case of 'b' (e.g. export/use a dictionary for a supervised classification) the user can refer to any **T-LAB** tool for thematic analysis (i.e. **Dictionary-Based Classification**, **Thematic Document Classification** etc.).

Elementary Context

During the importation phase, **T-LAB** makes a corpus **segmentation** into **elementary contexts** in order to help user exploration and, above all, to make analyses that require the **co-occurrences** computation.



According to the user's choices, the elementary contexts can be:

1 - **Sentences**

Elementary contexts ending with punctuation marks (.? !), whose length range is 50-1,000 characters.

2 - **Chunks**

Elementary contexts of comparable length made up of one or more sentences.

More precisely:

- **T-LAB** considers an elementary context to be every sequence of words interrupted by full stop and carriage return, whose dimensions are inferior to 400 characters;

- in the case where, within the maximum length, a full stop is not present, it searches for other punctuation marks in the following order (? ! ; : ,). If none are found, it performs segmentation on the basis of a statistical criterion, but without cutting the lexical units.

3 - **Paragraphs**

Elementary contexts ending with punctuation marks (.? !) and the return key, whose maximum length is 2,000 characters.

4 - **Short Texts**

This option is enabled only when the maximum length of texts is 2,000 characters (e.g. responses to open-ended questions).

N.B.:

- the **corpus_segments.dat** file contains the result of corpus segmentation;

- In T-LAB, the **Concordances** option allows the checking of elementary contexts where each **word** (or **lemma**) is present.

Frequency Threshold

During the pre-processing phase **T-LAB** computes a minimum frequency threshold to select words (or lemmas) for the automatic **key-words** list.

In any case, in order to guarantee the reliability of all statistical computations, the minimum **T-LAB** threshold is 4.

For this computation an algorithm documented in one of the books in the Bibliography is used (Bolasco, 1999). It requires the following steps:

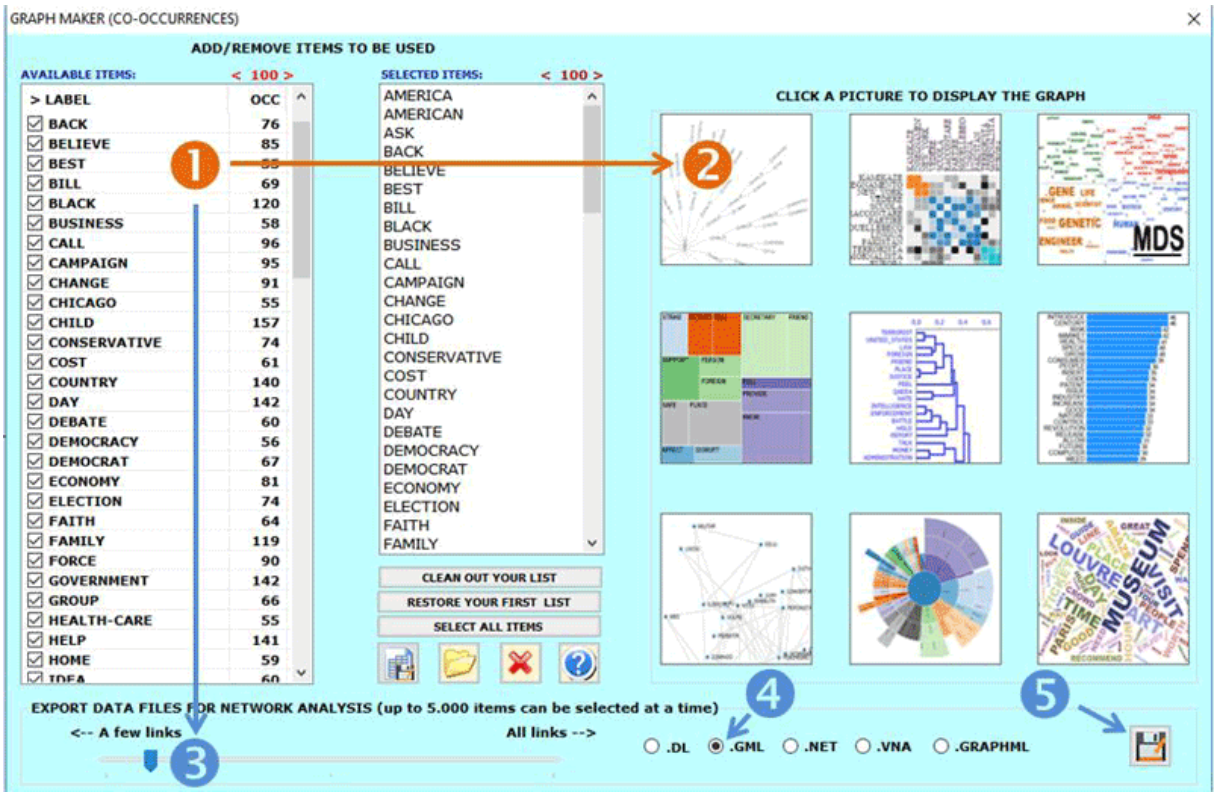
- *low frequency range detection; the range (starting from the minimum frequency -"1") is defined by the first "jump" in the growing occurrences values;*
- *threshold value choice. The threshold value, according to corpus sizes, corresponds to the minimum value in the first and in the second range decile (10% or 20%).*

Graph Maker

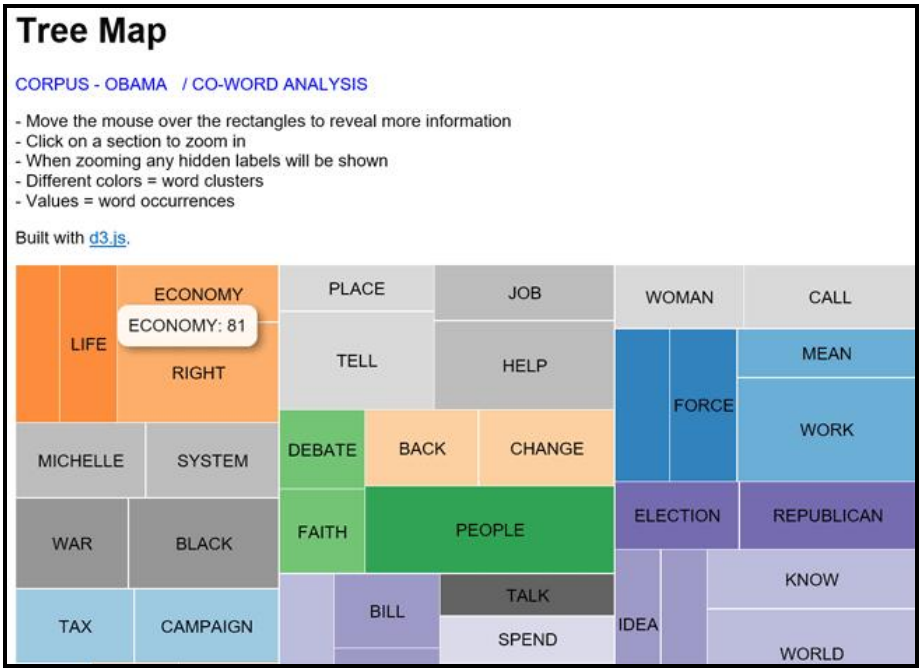
The **Graph Maker** tool allows the user to create several dynamic charts in HTML format which can be used for two purposes:
(a) for exploring **co-occurrences** between key words;
(b) for performing some sort of **network analysis**.

In the (a) case, two steps are required (see the below picture):
1- Select the items (i.e. key-terms) to be used;
2- Click any picture to display the corresponding graph.

In the (b) case, after the key-term selection (see '1' below), the user can filter the links to be mapped (see '3' below), then can choose the output format (see '4' below) and click the 'save' button (see '5' below).



N.B.: Each output in HTML format includes further instructions which are easy to understand (see the below picture).


Tree Map
CORPUS - OBAMA / CO-WORD ANALYSIS
- Move the mouse over the rectangles to reveal more information
- Click on a section to zoom in
- When zooming any hidden labels will be shown
- Different colors = word clusters
- Values = word occurrences
Built with d3.js.
ECONOMY
ECONOMY: 81
LIFE
RIGHT
PLACE
JOB
WOMAN
CALL
TELL
HELP
MEAN
FORCE
WORK
MICHELLE
SYSTEM
DEBATE
BACK
CHANGE
WAR
BLACK
FAITH
PEOPLE
ELECTION
REPUBLICAN
KNOW
TAX
CAMPAIGN
BILL
TALK
SPEND
IDEA
WORLD

Homographs

Two or more words are **homographs** when they have the same graphic form (they are written in the same way) but with different meanings.

In Italian and English there are thousands of homographs.

T-LAB implements **disambiguation** routines that reduce their impact. In particular, **multiword** normalization and compound verb tense identification.

For example, the normalization of the sequence "at present" (modified in "at_present") allows us to distinguish "present" as gift and “present” as time.

IDnumber

The **IDnumber** is a label that can be inserted in the coding line as a identifier of subjects (e.g. in the case of answers to open-ended questions) or of context units in which the corpus to be imported is subdivided (see **Corpus Preparation**).

In **T-LAB**, each time that the "IDnumber" label is used, it must be followed by a low dash ("_") and by a progressive number of max 5 digits (see the following example).

* * * * *IDNUMBER_0001 *AGE_ADUL *SEX_FEM *OCC_PROF
This is followed by the text of an answer or document.
……

Each corpus can include progressive IDnumbers up to a maximum of 30.000 subjects or context units.

N.B.:
The first IDnumber value must be "1" (e.g. IDnumber_00001).
In the case of texts collected by the user which are in MS Excel format, a macro is provided in the **T-LAB** installation package which automatically transforms them into an encoded corpus which is ready to be imported.

Isotopy

Isotopy (iso = same; topos = place) refers to a meaning conception as a "contextual effect", that is something that does not belong to words considered one by one, but as a result of their relationships within texts or speeches.

The Isotopies help in understanding speeches (or texts); in fact, each of the isotopies detects a reference context shared among a number of words, which however does not result from their specific meanings. That is because the whole is something different from the summation of its parts.

Isotopy detection, therefore, is not a simple "fact" observation, but the result of an interpretation process (F. Rastier 1987).

It was first proposed by the semiologist A.J. Greimas (1966) to define the recurrence, within syntagmatic units (sentences or texts), of words with the same semantic traits.

In **T-LAB** logic, the detection of isotopes derives from the analysis of occurrence and co-occurrence tables.

Key-Terms

In **T-LAB** logic, the Key-Terms (or Key-Words) refer to all the **lexical units** (words, lemmas, lexies, categories) which, each time, are included in the tables to be analysed.

Normally they belong to the category of **content words**: nouns, verbs, adjectives and adverbs.

Operationally, the selection of the key words can be made using two procedures: **automatic** and **customized**.

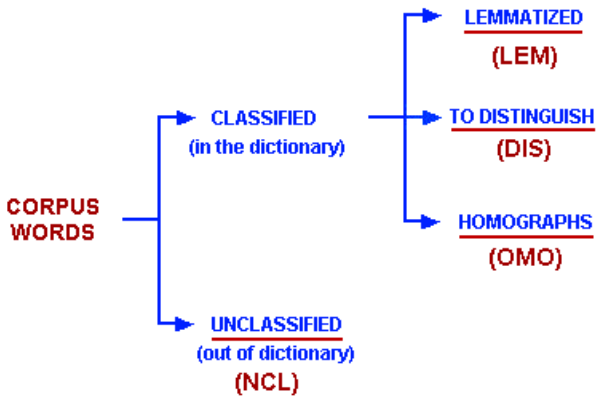
N.B.: The latter only allows us to modify the lexical unit lists and use **customized dictionaries**.

Lemmatization

Lemmatization involves the reduction of corpus words to their respective headwords (i.e. lemmas). In the linguistic dictionaries that we may consult, every entry corresponds to a lemma that - generally - defines a set of words with the same lexical root (or lexeme) and that belongs to the same grammatical category (verb, adjective, etc.).

As a rule, **lemmatization** entails that verb forms are taken back to the base form, nouns to the singular form, and so on. For example, the **inflected forms** "speaks" and "speaking", resulting from a combination of a sole **root** with two different suffixes ("-s" and "-ing"), are brought back to the same lemma "speak". There are, however, some cases in which the lemmatization doesn't observe the rule of the common root; particularly in the case of many irregular verbs.

During the corpus importation phase, **T-LAB** carries out a specific kind of automatic lemmatization, that follows the logic of the following "tree".



Obviously, the reference dictionary is the one implemented in **T-LAB**.

The abbreviations of the four-categories are used in many tables, always in the "INF" column (or field).

N.B:
-the "DIS" category ("to distinguish") means that **T-LAB** does not apply the standard lemmatization, in order to avoid annulling the significant meanings among the different forms.;
- sometimes, in order to differentiate homographs, **T-LAB** adds the underscore ('_') character to their lemma.

Lexical Unit

See **analysis unit**.

Lexie and Lexicalization

According to Pottier (see **Bibliography**), the **lexie** is an expression consisting in one or more words which behave as single lexical unit.

There are three main types: *simple*, which corresponds to the word in its most common meaning (e.g.. "horse", "eating"); *compound*, consisting in two or more words integrated in one single form (es. "biotechnology", "videoplayer"); *complex*, consisting in a sequence of words subject to lexicalization (e.g.. "in my opinion", "chamber of commerce").

Lexicalization is the linguistic process through which a syntagm (a sequence of words) becomes a lexical unit or behaves as such.

In **T-LAB** the **Multiwords List** option allows the user to produce a list of the complex lexies present in the corpus and to proceed with their transformation into unit strings (lexicalization).

Markov Chain

A Markov chain (from the name of the Russian mathematician Andrei Andreiëvich Markov) consists in a **succession** (or sequence) of events, generally suitable as **status**, characterized by two properties:

- the series of the events and their possible outcomes is a finite set;
- the outcome of each event depends only (or at the most) on the immediately precedent event.

With the consequence that a probability value corresponds to every transition from one event to the other.
In scientific domain, the Markovian chains model is used to analyse the succession of economic, biological, physical events etc. In the domain of linguistic studies its application concerns the possible combinations of the various analysis units on the syntagmatic axis (one item after the other).

In **T-LAB** the analysis of the Markovian chains relates to two types of **sequences**:

- those concerning the relationships between lexical units (words, lemmas or categories) present in the corpus under analysis;
- those present in external files prepared by the user.

In both cases, to start with, some square tables are constructed in which the occurrence of transitions is recorded, that is the quantity that indicates the number of times in which an analysis unit precedes (or follows) the other. Subsequently, the transition occurrences are transformed into probability values (see the following images):

	s_1	s_2	s_3	s_4	s_5	s_6	TOT
s_1	0	8	7	11	2	1	29
s_2	6	0	24	5	10	8	53
s_3	9	24	0	3	28	16	80
s_4	3	7	5	0	6	14	35
s_5	4	5	26	11	0	7	53
s_6	7	9	18	5	7	0	46

	s_1	s_2	s_3	s_4	s_5	s_6	TOT
s_1	0,00	0,28	0,24	0,38	0,07	0,03	1
s_2	0,11	0,00	0,45	0,09	0,19	0,15	1
s_3	0,11	0,30	0,00	0,04	0,35	0,20	1
s_4	0,09	0,20	0,14	0,00	0,17	0,40	1
s_5	0,08	0,09	0,49	0,21	0,00	0,13	1
s_6	0,15	0,20	0,39	0,11	0,15	0,00	1

For further information see **Sequence Analysis.**

MDS

MDS is a set of data analysis techniques that allow us to analyse similarity matrices in order to provide a visual representation of the relationships among the data within a space of reduced dimensions.

T-LAB uses a type of **MDS** (Sammon's method) in order to represent the relationships among the lexical units or among the thematic nuclei (see **Co-Word Analysis** and **Modeling of Emerging Themes**).

The input tables are constituted by square matrices which contain proximity values (dissimilarities) derived from the calculation of an association index.

The results obtained, like those of the correspondence analysis, allow us to interpret both the relationships between the "objects" and the dimensions that organize the space in which they are represented.

The degree of correspondence between the distances among points implied by the MDS map and the matrix input is measured (inversely) by a stress function. The lesser the stress value (e.g. < 0.10), the greater the goodness of the obtained adjustment.

The stress formula (Sammon's method) is the following:

$$S = \sum_{i \neq j} \frac{(d_{ij}^{*} - d_{ij})^2}{d_{ij}^{*}}$$

Where d_{ij}^{*} stands for the distance between two points (ij) within the input matrix and d_{ij} stands for the distance between the same points (ij) within Sammon's map.

Multiwords

A set of two or more words (multi-words) that stand for only one meaning.

The multiword category, which differs according to the analytical model used, includes subsets as **compound words** (for eg. "public transport" or "occupation level"), **phrasal verbs** (for eg. "get off" or "take away") and **idioms** (for eg.: "with respect of" or "out of touch with").

The multiword list implemented in **T-LAB**, obviously, is not exhaustive.
It is built and tested with two criteria:

- to limit the most frequent ambiguity cases (**effectiveness** criterion);
- to moderate the **normalization** processing times (**efficiency** criterion).

In **T-LAB** it is also possible to use a **customized Multi-Word list**.

N-Grams

In **T-LAB** an n-gram is a sequence of two (bi-gram) or more contiguous key words present within the same **elementary context** (i.e. sentence, text fragment or paragraph).

When used for computing word **co-occurrences**, n-gram segmentation overlooks both stop-words and punctuation marks.

Let's consider the following example:

The Citizens of each State shall be entitled to all Privileges and Immunities of Citizens in the several States.

By assuming that the seven items in red are included in our key-term list and that an automatic lemmatization has been applied, a bi-gram segmentation produces the following co-occurrence contexts:

citizen & state
state & entitle
entitle & privilege
privilege & immunity
immunity & citizen
citizen & state.

Differently, a three-gram segmentation produces the following co-occurrence contexts:

citizen & state & entitle
state & entitle & provilege
entitle & privilege & immunity
privilege & immunity & citizen
immunity & citizen & state
citizen & state.

It is worth recalling that, when segmenting texts into elementary contexts, co-occurrences depend on the presence (or absence) of key words; whereas, when using an n-gram segmentation, co-occurrences indicate a sequential relationship between words.

In **T-LAB** an n-gram based co-occurrence analysis can be performed with the advanced options of the **Word Association** tool; moreover, a Markovian analysis of bi-grams can be performed with the **Sequence Analysis** tool.

Naïve Bayes Classifier

Here is the formula of the Naïve Bayes Classifier (NB) used by **T-LAB**:

$$v_{\mathbf{NB}} = \underset{v_j \in V}{\arg\max}\, P(v_j) \prod_i P(a_i \mid v_j)$$

Where:

argmax – refers to the maximum posterior value;
$v_j \in V$ - refers to the j-cluster (v_j) of the partition (V);
$P(v_j)$ - refers to the prior probability of each j-cluster, that is to their relative frequencies;
$\prod_i P(a_i \mid v_j)$ - is the probability product of each (a_i) word within each (v_j) cluster, where each $P(a_i \mid v_j)$ is an element of a normalized vector of relative frequencies and each (a_i) word is present in the i-context unit to be reclassified.

Normalization

In **T-LAB**, corpus normalization has the double goal of:

a) allowing correct word detection as **raw forms**;

b) solving some **ambiguity** cases.

This means that **T-LAB**, in the first place, carries out a number of processes on the file under analysis: blank space in excess elimination, apostrophe marking, space addition after punctuation marks, capital letter reduction, etc.

Secondly, **T-LAB** marks a set of strings recognized as **proper nouns**; then converts the sequences of row forms recognized as **multiwords** in unitary strings, in order to use them in that form during the analysis process ("in terms of" and "point of view" become respectively "in_terms_of" and "point_of_view").

These operation parameters cannot be modified by the user.

In order to have a correct recognition of raw forms, in the normalization routine, **T-LAB** uses the following marks:

, ; : . ! ? ' " ( ) < > + / = [ ] { }

Occurrences and Co-occurrences

In text analysis these two concepts are of fundamental importance.

The **Occurrences**, in fact, are quantities which result from the computation of how many times (frequences) a single lexical unit (**LU**) occurs within a **corpus** or within the context units (**CU**) in which it is subdivided.

Their distribution can be represented in contingency tables as follows:

	CU_1	CU_2	CU_3	CU_4
LU_1	19	1	12	14
LU_2	17	0	1	8
LU_3	8	4	2	9
LU_4	101	0	13	0
LU_5	32	1	29	11
LU_6	4	3	0	30
LU_7	10	1	3	21
LU_8	5	1	1	34
LU_9	25	5	0	54

Co-occurrences, then, are quantities which result from a computation of how many times two or more lexical units are present together in the same elementary contexts (**EC**).

Their distribution can be represented in tables such as the following:

(A)

	LU_1	LU_2	LU_3	...	LU_n
EC_1	0	1	0	...	1
EC_2	1	0	0	...	0
EC_3	0	1	1	...	0
EC_4	0	0	0	...	0
EC_5	1	1	0	...	1
EC_6	0	0	0	...	0
EC_7	0	0	1	...	0
EC_8	1	0	0	...	0
EC_9	0	0	0	...	0
EC_10	0	1	0	...	0
EC_11	1	0	1	...	0
EC_12	0	0	0	...	1
EC_13	1	1	0	...	0
EC_14	0	0	1	...	0
EC_15	0	0	0	...	0
EC_16	0	1	0	...	1
EC_17	0	0	1	...	0
EC_18	0	0	0	...	0
EC_19	1	0	0	...	0
EC_20	0	0	0	...	1

With a simple transformation, the "A" type table (rectangular) can be transformed into "B" type (squared and symmetrical) in which for each pair of lexical units the quantity of their co-occurrences is indicated, that is the total number of the elementary contexts in which they are present together.

(B)

	LU_1	LU_2	LU_3	...	LU_n
LU_1		2	1	...	1
LU_2	2		1	...	3
LU_3	1	1		...	0
...	...	...	...		...
LU_n	1	3	0	...	

In **T-LAB** text analysis is mostly carried out by the study of relationships among occurrences and among co-occurrences, either through specific **association indexes**, or through the use of multidimensional statistical techniques like **cluster analysis** and **correspondence analysis**.

Poles of Factors

In **Correspondence Analysis** each factor sets up a spatial dimension - that can be represented as an axis line - whose centre (or barycentre) is the value "0", and that develops in a bipolar way towards the negative (-) and positive (+) end, so that the objects put on the opposite poles are the most different, almost like the "left" wing and "right" wing on the political axes.

It is useful to remember what J.P Benzecri, a mathematician and one of the most important contributors to this kind of analysis technique, wrote about it:

"Understanding a factorial axis means finding what is similar, on the one hand all that is on the right of the origin (barycentre), on the other all that is on the left of it, and then expressing concisely and exactly the opposition between the two extremes". (1984, p. 302, see **Bibliography**).

N.B. : When factorial graphs are bi-dimensional (or tri-dimensional) the oppositions are more than two: in addition to left and right, there is up and down. Nevertheless the interpretation criteria are the same.

Primary Document

The primary documents are texts (or portions of the corpus) that correspond to the context units preceded by a **coding line**.

Depending on the cases, they can be: books or book chapters, newspaper articles, interview transcripts, answers to open-ended questions etc.

Profile

In **T-LAB** the profile of an **analysis unit** (lexical unit or context unit) corresponds to the vector (row or column) of the data table that contains the distribution of its **occurrence** or **co-occurrence** values.

N.B.:

In the **Correspondences Analysis** the profiles (row or column vectors) that take part in the construction of the factorial axes are **active**; while **supplementary** profiles are the ones whose values are calculated a posteriori.

Specificity

In **T-LAB**, **Specificity Analysis** is the name of a tool that allows us to check the lexical units (i.e. words, lemmas or categories) and the elementary contexts (i.e. sentences or paragraphs) which are **typical** in a text or a corpus subset defined by a categorical **variable**.

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 normalized **TF-IDF** values assigned to the words which each sentence or paragraph consists of.

Stop Word List

In text analysis, many words are defined "empty" because - on their own - they don't have any specific and/or significant content.

A standard criterion doesn't exist in order to build a list of these words (**Stop Word List**).

In **T-LAB** the list is taken from the following classes:

- indefinite adjectives;
- articles;
- adverbs;
- exclamations;
- interjections ;
- prepositions ;
- pronouns (demonstrative, indefinite and relative);
- auxiliary verbs;
- modal verbs.

Moreover the user can import **customized Stop-Word lists**.

Test Value

This is a statistical measure which **T-LAB** uses to characterise two kinds of relationships:

a) those concerning lexical units with variable categories, the occurrences of which are summarized in contingency tables;
b) those concerning any row and column of a contingency table with the factors detected by a correspondence analysis of such a table.

Depending on the relationships analysed, the formulas of the test value, taken from one of the bibliography volumes (Lebart L. Morineau A. Piron M. , 1995, pp. 181-184), are the following:

a)

$$t_k(j) = \frac{n_{jk} - n_k \cdot \frac{n_j}{n}}{\sqrt{n_k \cdot \frac{n - n_k}{n-1} \cdot \frac{n_j}{n} \cdot \left(1 - \frac{n_j}{n}\right)}}$$

where 'n_{jk}' indicates the occurrences within a cell, while 'n_j' and 'n_k' correspond respectively to the marginal total of a row and of a column;

b)

$$t\alpha(j) = \sqrt{\mathrm{nj}\,\frac{\mathrm{n}-1}{\mathrm{n}-\mathrm{nj}}}\;\varphi\alpha\mathrm{j}$$

where "n_j" and "$\varphi\alpha_j$" indicate respectively the occurrences of the j-th object and its coordinate on the α-th factorial axis.

The test value has two significant properties: a threshold value (1.96), corresponding to the statistical significance most commonly used (p. 0.05), and a sign (-/+).

This means that, after sorting the values in an ascending or descending order, it is possible to quickly single out the relevance of each analysed item.

T-LAB allows a **Test Value table** consultation in an interactive way.

TF-IDF

This measure, proposed by G. Salton (1989), allows us to evaluate the weight of a term (lexical unit) within a document (context unit).

Its formula is the following:

w *i,j* = ***tf*** *i,j x* ***idf*** *i* (*Term Frequency* x *Inverse Document Frequency*)

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Where:

tf *i,j* = number of occurrences of ***i*** (term) in ***j*** (document)
df *i* = number of documents containing ***i***
N = total number of documents

*Term Frequency (**tf** i,j*) value can be normalized as follows:

tf *i,j* = ***tf*** *i,j* / Max (***f*** *i,j*)

where Max (***f*** *i,j*) is the maximum frequency of ***i*** (any term) in ***j*** (document).

Thematic Nucleus

T-LAB uses the locution **thematic nuclei** in some functions that produce **key-word** maps.

Thematic nucleus stands for little word clusters, **co-occurring** in corpus elementary contexts, which - on maps - are represented by by labels which can be definited and changed by the user.

Variables and Categories

In **T-LAB**, **variables** are labels used to identify and classify any corpus subset: names of features identifying subjects, texts and context types.

Every variable has two or more **categories**, each of them - univocally - corresponds to a coding value. For example, the "sex" variable has two categories (female and male).

In **T-LAB**, every text can be identified through **a maximum of 50 variables**.

Obviously, for each of them, the respective category must be stated (max 150 each), following the instructions contained in **Corpus Preparation**.

For more information, see the examples in the demo files.

Words and Lemmas

Any text analysis software first of all identifies the so called **raw forms**, that is the strings of letters separated by blank spaces. Then, according either to their specific algorithms or to the categories used by the specialists, the software recognizes **lexemes**, **key-words**, etc.

The **T-LAB** tables, for all the lexical units present in the corpus database, provide two types of information:

- the first one, named "**word**", contains the transcript of the lexical units (single words or multi-words) as "strings" which are recognized by the software;

- the second, named "**lemma**", contains the labels (or tags) used for grouping and classifying the lexical units.

According to the case, a **lemma** can be:

- the result of the automatic lemmatization process;
- an item of a customized dictionary;
- a category grouping synonyms;
- a content analysis category;
- etc.

BIBLIOGRAPHY

Bardin L. (1977):*L'analyse de contenu, Paris, P.U.F.*

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

Blei D.M. (2012):*Introduction to Probabilistic Topic Models, Communications of the ACM, Volume 55 Issue 4, April 2012 Pages 77-84*

Blondel V.D., Guillaume J.-L., Lambiotte R., Lefebvre E. (2008): *Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008 (12pp)*

Bolasco S. (1999):*Analisi Multidimensionale dei dati. Metodi, strategie e criteri di interpretazione, Roma, Carocci*

Boley D.L. (1998):*Principal direction divisive partitioning, Data Mining and Knowledge Discovery, 2(4), 325-344*

Carroll J.B. (1964): *Language and Thought, Englewood Cliff NJ, Prentice Hall*

Campello R. J. G. B., Moulavi D., Zimek A. & Sander J. (2015): *Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection*. ACM Trans. Knowl. Discov. Data 10, 1, Article 5 (July 2015)

De Mauro T. Mancini F. Vedovelli M. Voghera M. (1993):*Lessico di frequenza dell'italiano parlato (Fondazione IBM), Milano, Etas Libri*

Fernández A., Gómez S. (2008): *Solving Non-uniqueness in Agglomerative Hierarchical Clustering Using Multidendrograms, Journal of Classification, 25: 43-65*

Greenacre M.J. (1984):*Theory and Applications of Correspondance Analysis, New York, Academic Press*

Greimas A.J. (1966):*Sémantique structurale, Paris, Larousse*

Guiraud P. (1960): *Problèmes et méthodes de la statistique linguistique. Dordrecht, Reidel*

Herdan, G. (1960): *Quantitative Linguistics. London, Butterworth*

Kohonen T. (1989):*Self-Organization and Associative Memory, Berlin, Springer-Verlag*

Krippendorff K. (1980):*Content Analysis. An Introduction to its Methodology, London, Sage inc.*

Lancia F. (2004) :*Strumenti per l'analisi dei testi. Introduzione all'uso di **T-LAB**, Milano, FrancoAngeli*

Lancia F. (2005), *Word co-occurrence and Similarity in Meaning*, *www.tlab.it*

Lancia F. (2012) : *The Logic of the T-LAB Tools Explained, www.tlab.it*

Lebart L., Morineau A., Piron M. (1995):*Statistique exploratoire multidimensionnelle, Paris, Dunod*

Lebart L., Salem A. (1994):*Statistique textuelle, Paris, Dunod*

Maranda P. (1990):*DisCan: User's Manual, Québec, Nadeau Caron Informatique*

Marwan N., Romano M., Thiel M. & Kurths J. (2007): *Recurrence Plots for the Analysis of Complex Systems,* Phys. Rep. 438, 240-329.

Michelet B. (1988):*L'analyse des associations, Thèse de doctorat, Université Paris VII, Paris*

Miller M.M., Riechert B.P. (1994):*Identifying Themes via Concept Mapping: A New Method of Content Analysis, Paper presented at the Communication Theory and Methology Division of the Association for Education in Journalism and Mass Communication Annual Meeting,*

Atlanta
Pottier B.(1974) : *Linguistique générale, théorie et description*, *Paris, Klincksieck*
Rastier F. (1987):*Sémantique interprétative, Paris, PUF*
Rastier F., Cavazza M., Abeillé A. (2002):*Semantics for Descriptions, Stanford, CSLI*
Saussure (de) F. (1916), *Cours de Linguistique générale*, *Lusanne-Paris, Payot,*
Salton G. (1989):*Automatic text processing: the transformation, analysis, and retrieval of Information by Computer, Addison-Wesley, Reading, Massachussets*
Savaresi S.M., Boley D.L. (2001): *On the performance of bisecting K-means and PDDP*, *1st SIAM Conference on DATA MINING, Chicago, IL, USA, April 5-7, paper n.5, pp.1-14*
Savaresi S.M., Boley D.L. (2004): *A Comparative Analysis on the Bisecting k-means and the PDDP Clustering Algorithms*, *International Journal on Intelligent Data Analysis, 8(4): 345-362*
Steinbach M., Karypis G., Kumar V. (2000): *A comparison of Document Clustering Techniques. Proceedings of World Text Mining Conference, KDD2000, Boston*
Steyvers M., Griffiths T. (2007) : *Probabilistic Topic Models. In Landauer, T.; McNamara, D; Dennis, S.; et al. Handbook of Latent Semantic Analysis, Mahwak, NJ, Lawrence Erlbaum*
van der Maaten L.J.P., & G.E. Hinton (2008): *Visualizing High-Dimensional Data Using t-SNE*. *Journal of Machine Learning Research* 9(Nov):2579-2605, 2008
Webber C. L., & Zbilut J. P. (2005) : *Recurrence Quantification Analysis of Nonlinear Dynamical Systems.* In M. Riley, & G. Van Orden (Eds.), *Tutorials in Contemporary Nonlinear Methods for the Behavioral Sciences* (pp. 26-94)