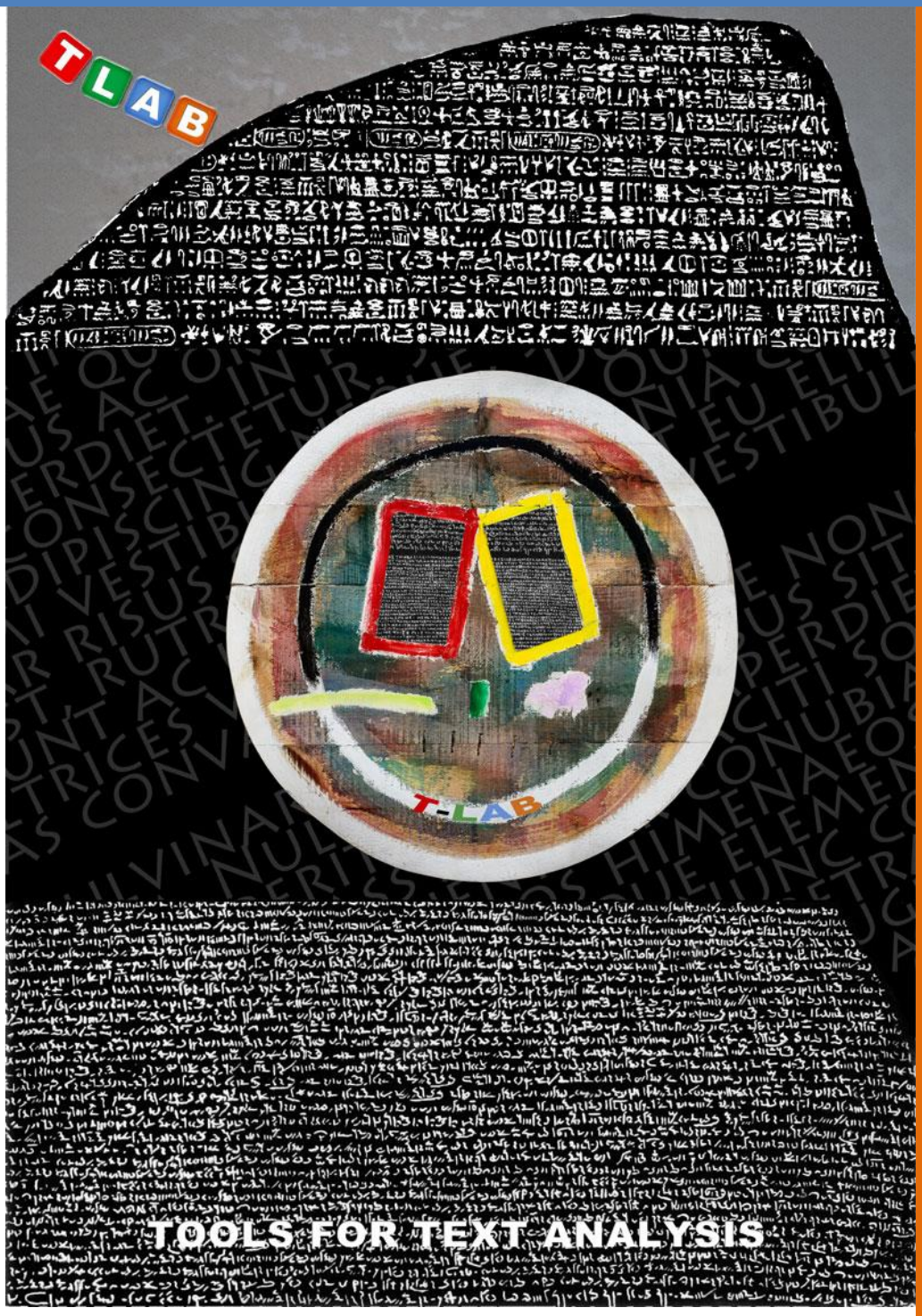
T-LAB 10

Quick Introduction



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

Strumenti per l'Analisi dei Testi

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

T-LAB: cosa fa e cosa consente di fare

(dal Manuale dell'Utilizzatore)

T-LAB è un software costituito da un insieme di **strumenti linguistici**, **statistici e grafici per l'analisi dei testi** che possono essere utilizzati nelle seguenti pratiche di ricerca: Analisi di Contenuto, Sentiment Analysis, Analisi Semantica, Analisi Tematica, Text Mining, Perceptual Mapping, Analisi del Discorso, Network Text Analysis, Document Clustering, Text Summarization.

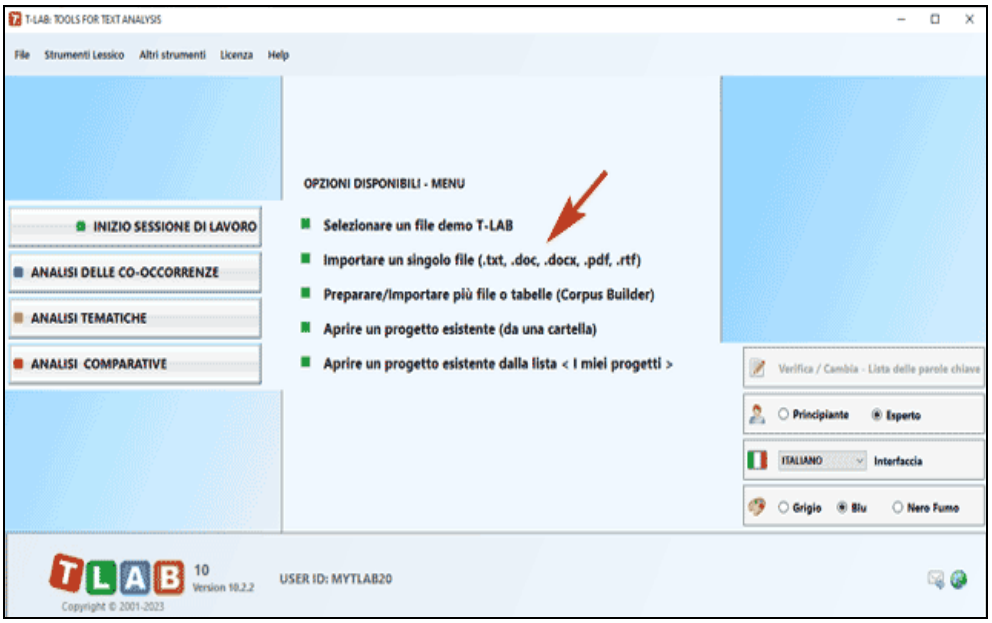
In effetti, tramite gli strumenti **T-LAB** i ricercatori possono gestire agevolmente attività di analisi come le seguenti:

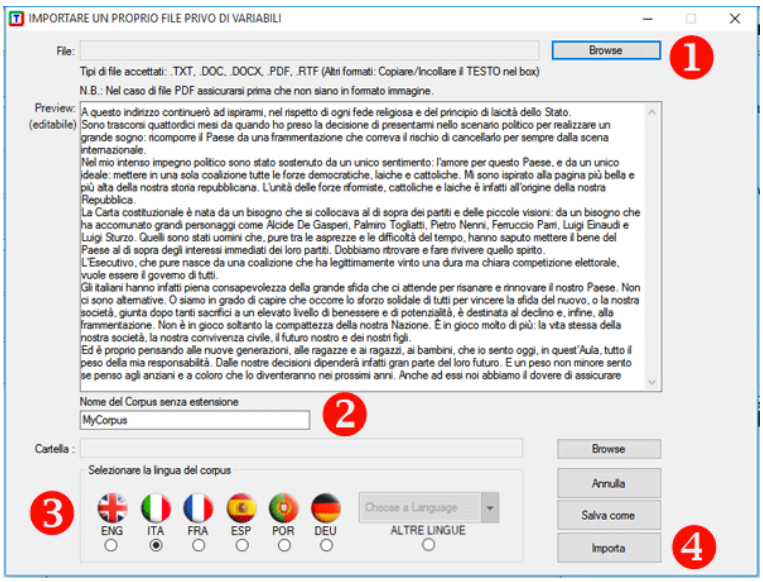
- esplorare, misurare e mappare la **relazioni di co-occorrenza** tra parole-chiave;
- realizzare una **classificazione automatica** di unità testuali o documenti, sia tramite un approccio **bottom-up** (cioè che tramite l'analisi dei **temi emergenti**), sia tramite un approccio **top-down** (cioè tramite l'uso di **categorie predefinite**);
- verificare quali **unità lessicali** (cioè parole o lemmi), quali **unità di contesto** (cioè frasi o paragrafi) e quali **temi** sono 'caratteristici' di specifici sottoinsiemi di testi (ad es., discorsi di specifici leader politici, interviste con specifiche categorie di persone, etc.);
- applicare categorie per la **sentiment analysis**;
- eseguire vari tipi di **analisi delle corrispondenze** e **cluster analysis**;
- creare **mappe semantiche** che rappresentano **aspetti dinamici** del discorso (cioè relazioni sequenziali tra parole o temi);
- rappresentare ed esplorare un qualsiasi testo come una **rete** di relazioni;
- ottenere misure e rappresentazioni grafiche relative a **testi e discorsi** trattati come **sistemi dinamici**;
- personalizzare e applicare **vari tipi di dizionari**, sia per l'analisi lessicale che per l'analisi di contenuto;
- verificare i contesti di occorrenza (ad es., **concordanze**) di parole e lemmi;
- analizzare tutto il **corpus** o solo alcuni dei suoi **sottoinsiemi** (ad esempio gruppi di documenti) utilizzando varie liste di parole-chiave;
- creare, esplorare ed esportare vari tipi di **tabelle di contingenza** e **matrici di co-occorrenze**.

L'interfaccia del software è particolarmente **user friendly** e i testi analizzabili possono essere i più vari:
- un singolo testo (es. un'intervista, un libro, etc.);
- un insieme di testi (es. più interviste, pagine web, articoli di giornale, risposte a domande aperte, messaggi Twitter etc.).

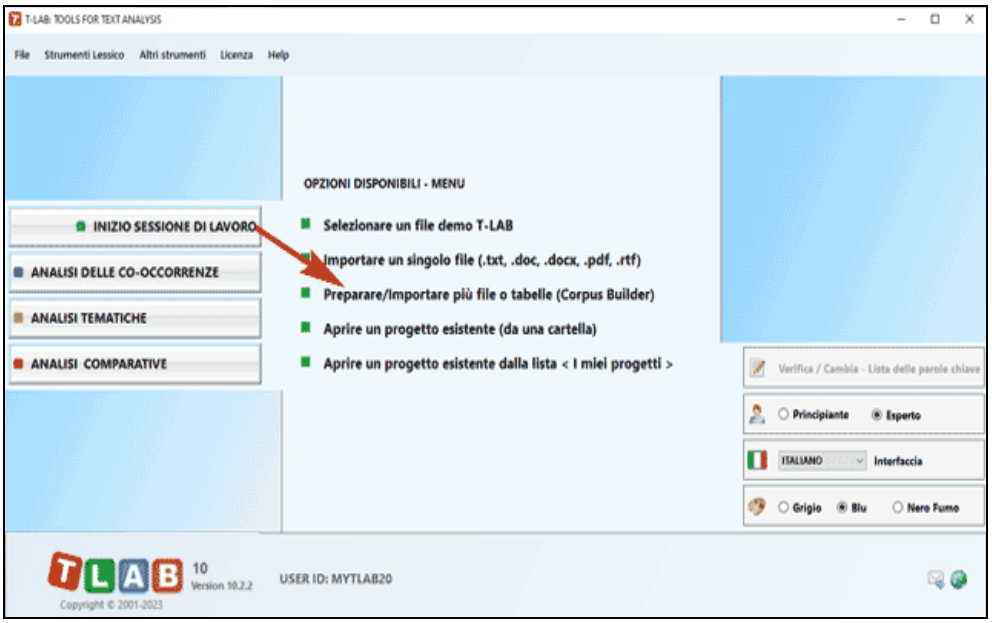
Tutti i testi possono essere codificati con variabili categoriali e/o con un identificativo (**Unique Identifier**) che corrisponde a unità di contesto o a casi (es. risposte a domande aperte).

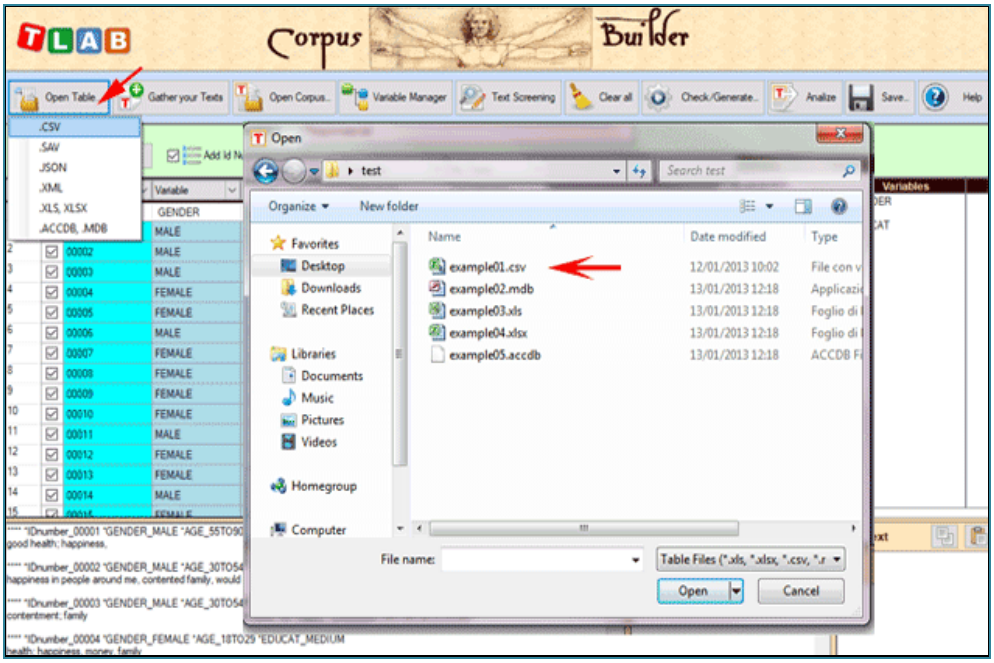
Nel caso di un singolo documento (o di un corpus trattato come unico testo) **T-LAB** non richiede ulteriori accorgimenti: basta selezionare l'opzione 'Importare un singolo file…' e procedere (vedi sotto).

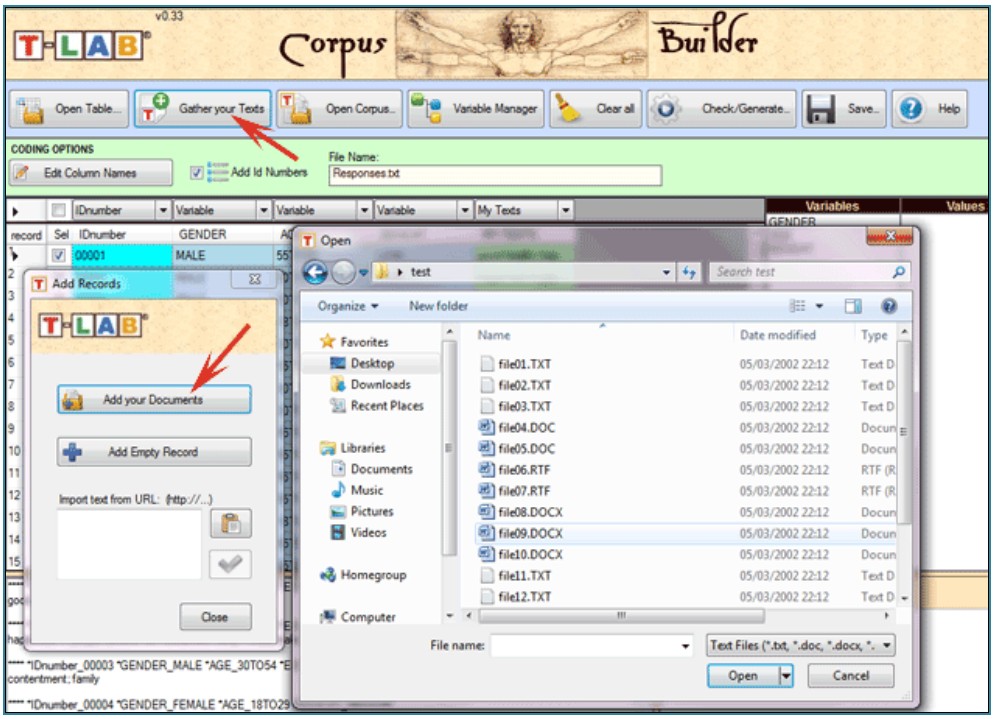




Diversamente, negli altri casi va usato il modulo **Corpus Builder** che – in modo automatico - facilita la trasformazione di vari tipi di materiali testuali e vari tipi di file in un **corpus** pronto per essere importato da **T-LAB** (vedi sotto)**.**



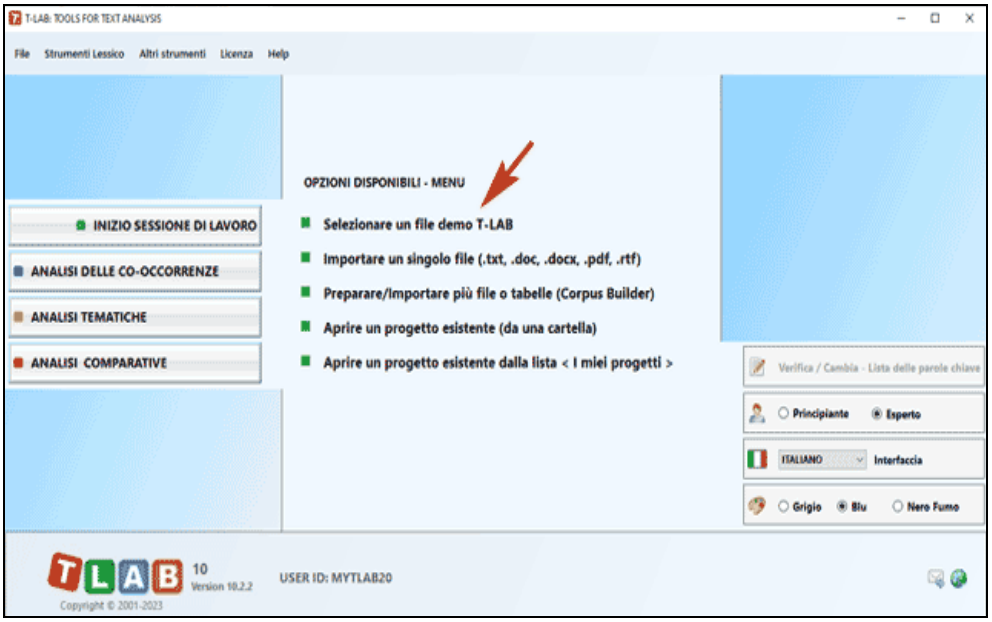




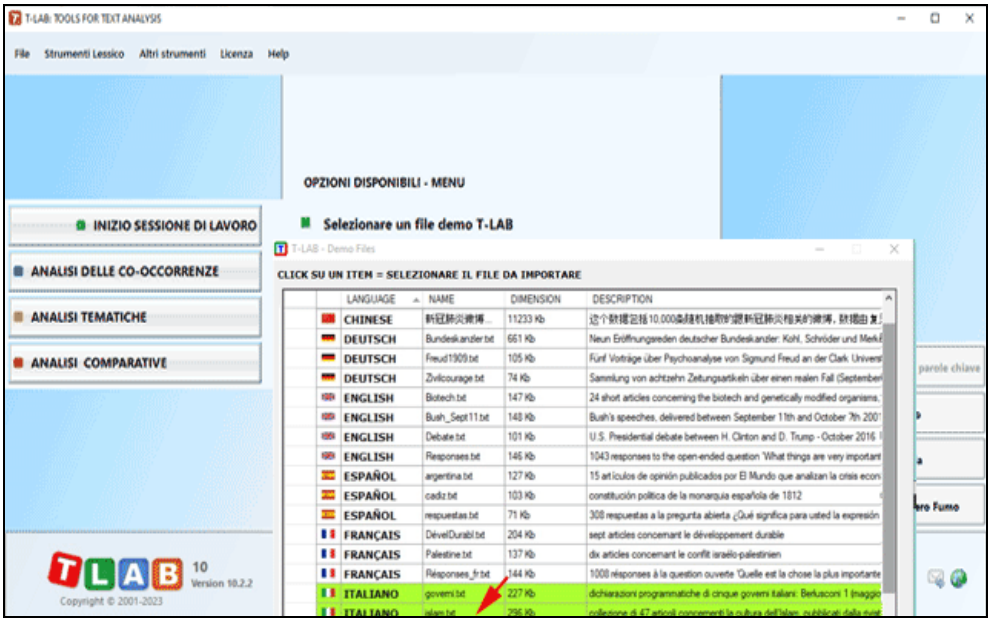
N.B.: Al momento - per garantire l'uso integrato dei vari strumenti - ogni file/corpus da analizzare non deve superare i 90 Mb (cioè circa 55.000 pagine in formato testo). Per ulteriori informazioni, vedere la sezione 'Requisiti e prestazioni' dell'Help / Manuale.

Per verificare rapidamente le funzionalità del software sono sufficienti i seguenti passi:

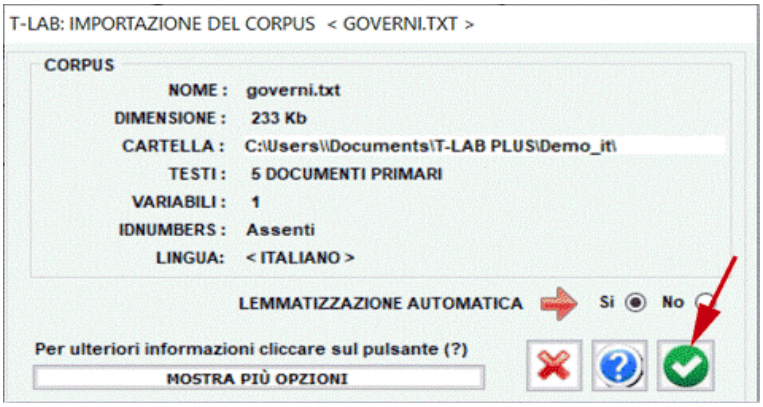
1 - **Selezionare l'opzione 'Selezionare un file demo T-LAB"**



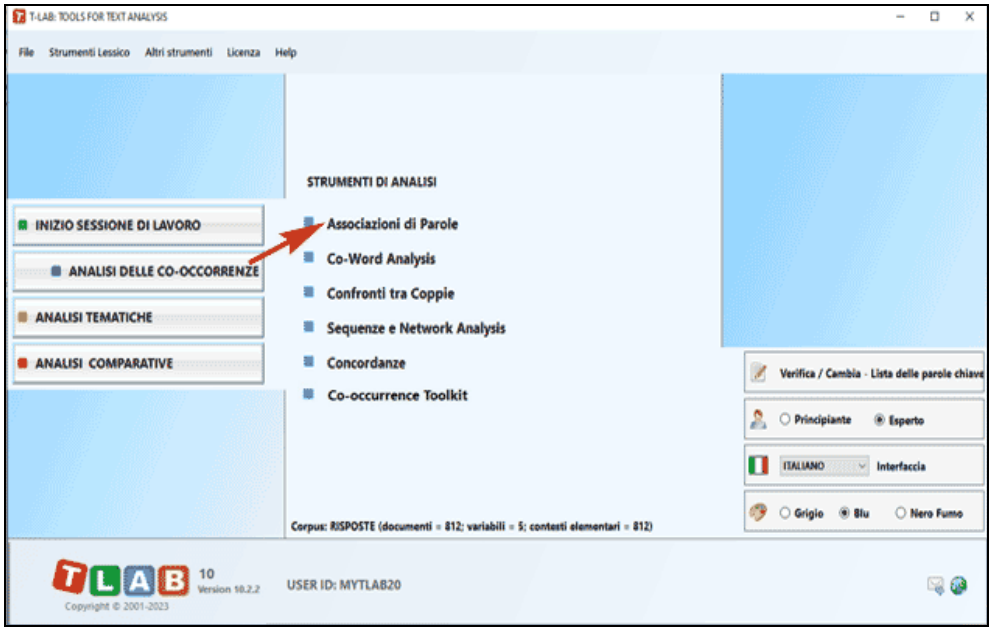
2 - Selezionare un corpus da analizzare



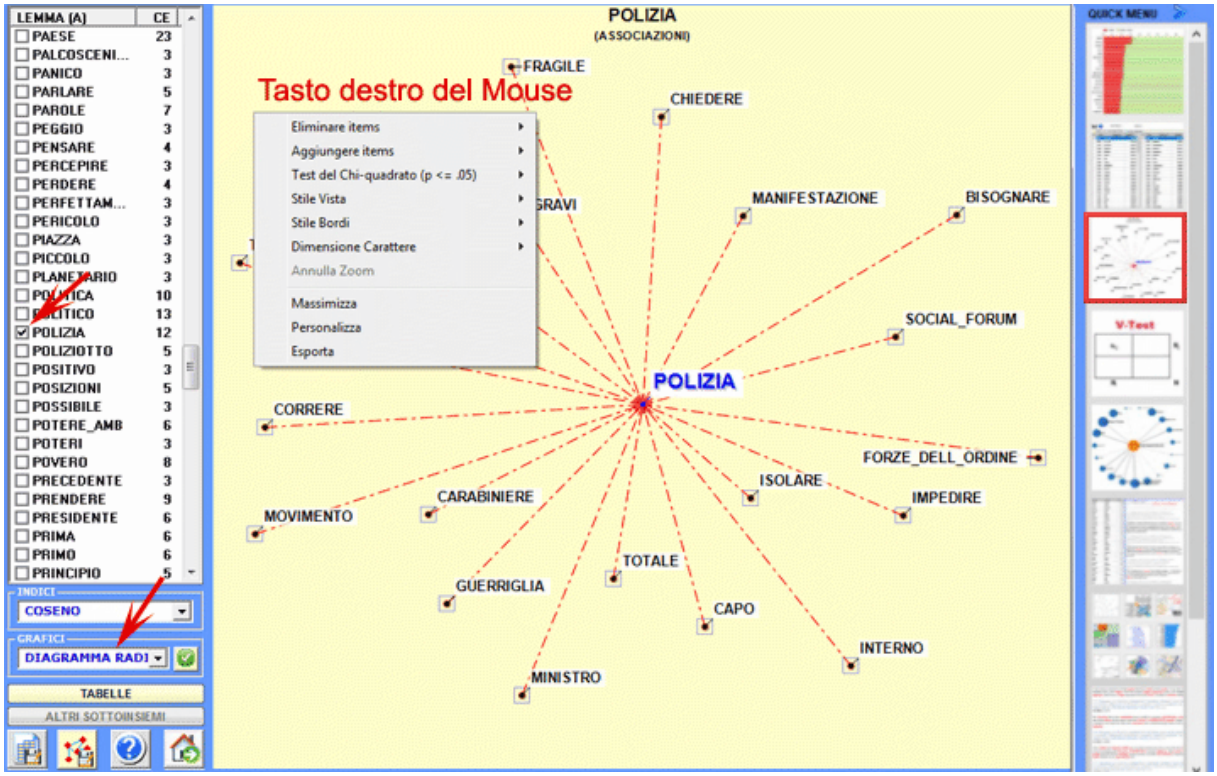
3 - Cliccare su "ok" nella prima finestra di Setup

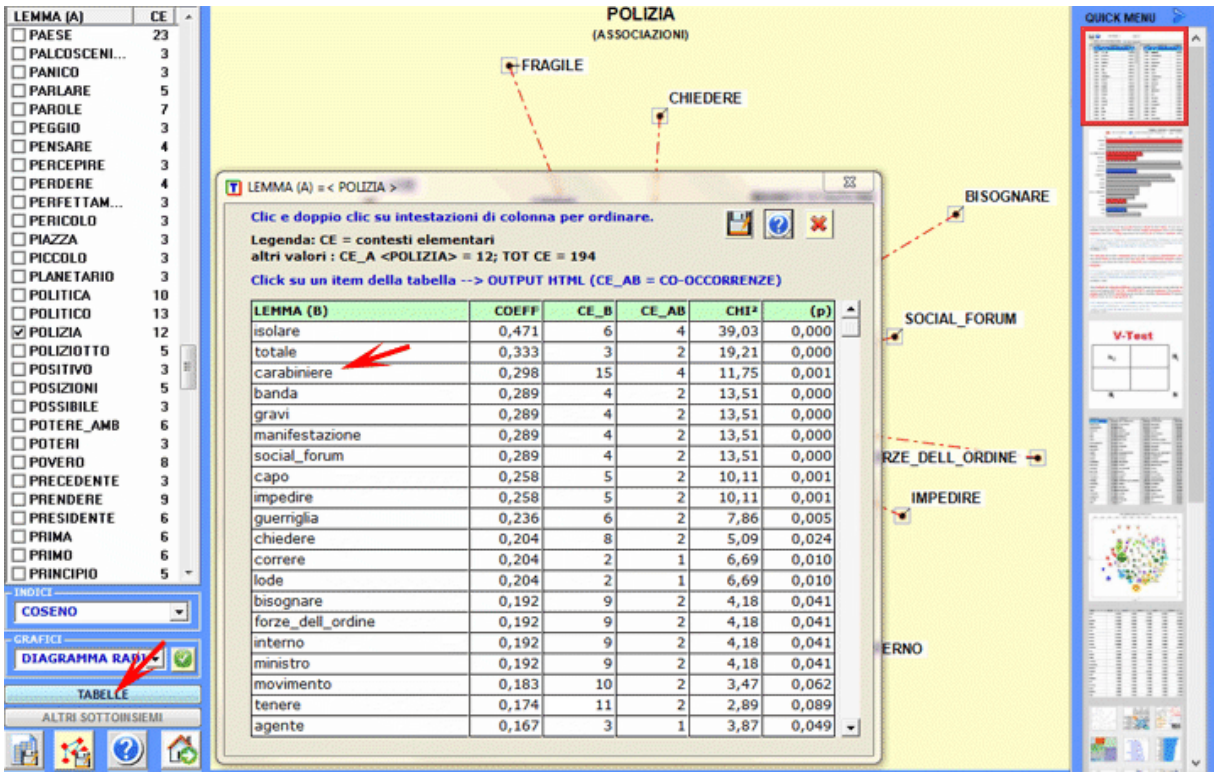


4 - Scegliere uno strumento all'interno di uno dei sub-menu ‘Analisi’



5 - **Verificare i risultati**





6 - **Utilizzare l'help contestuale per interpretare grafici e tabelle**

Tutti i grafici possono essere personalizzati e salvati in vari formati.

Di seguito vengono fornite le informazioni essenziali per capire cosa **T-LAB** fa e come può essere usato.

Dal punto di vista esterno, l'uso del software è organizzato dall'**interfaccia**, cioè dal **menu principale**, dai **sub-menu** e dalle **funzioni** (strumenti) che li compongono.

Da un punto di vista logico, oltre che dall'interfaccia utente, il sistema **T-LAB** è organizzato da due componenti principali:

- il **database**, cioè è il "luogo" informatico in cui il corpus in input (cioè il testo o l'insieme dei testi da analizzare) è rappresentato come un insieme di **tabelle** in cui sono registrate le **unità di analisi**, le loro caratteristiche e le loro reciproche relazioni;
- gli **algoritmi**, cioè sottoinsiemi di **istruzioni** che consentono di usare l'interfaccia utente, di consultare e modificare il database, di costruire ulteriori tabelle con in dati in esso contenuti, di effettuare **calcoli statistici** e di produrre **output** che rappresentano le relazioni tra i dati analizzati.

Per capire come **T-LAB** funziona e come può essere usato, è di fondamentale importanza aver chiaro quali unità di analisi sono archiviate nel suo database e quali algoritmi statistici vengono usati nelle varie analisi. Infatti, le tabelle dati analizzate sono sempre costituite da righe e colonne le cui intestazioni corrispondono alle unità di analisi archiviate nel database, mentre gli algoritmi regolano i processi che consentono di individuare relazioni significative tra i dati e di estrarre utili informazioni.

Le **unità di analisi** di **T-LAB** sono di due tipi: **unità lessicali** e **unità di contesto**.

A - le **UNITA' LESSICALI** sono parole, singole o multiple, archiviate e classificate in base a un qualche criterio. Più precisamente, nel database **T-LAB** ogni unità lessicale costituisce un record classificato con due campi: forma e lemma. Nel primo campo, denominato **forma**, sono elencate le parole così come compaiono nel corpus, mentre nel secondo, denominato **lemma**, sono elencate le label attribuite a gruppi di unità lessicali classificate secondo criteri linguistici (es. lemmatizzazione) o tramite dizionari e griglie semantiche definite dall'utilizzatore.

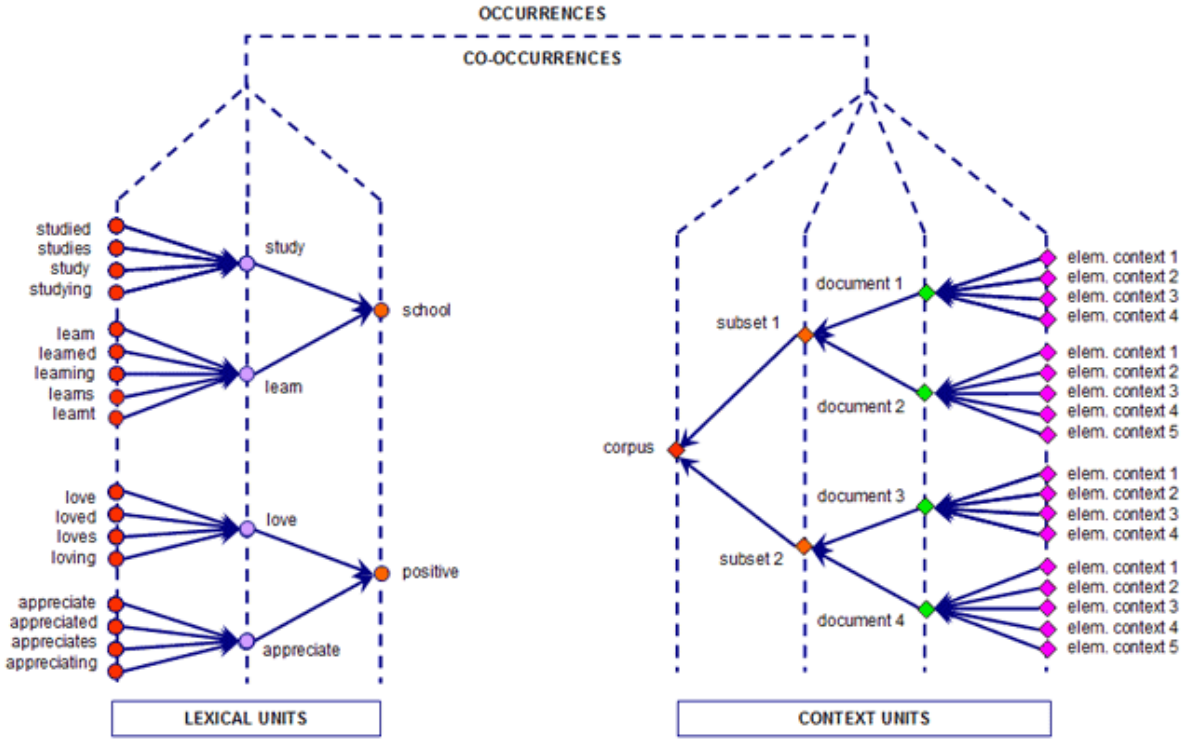
B - le **UNITA' DI CONTESTO** sono porzioni di testo in cui può essere suddiviso il corpus. Più esattamente, nella logica **T-LAB**, le unità di contesto possono essere di tre tipi:

B.1 **documenti primari**, corrispondenti alla suddivisione "naturale" del corpus (es. interviste, articoli, risposte a domande aperte, etc.), ovvero ai **contesti iniziali** definiti dall'utilizzatore;

B.2 **contesti elementari**, corrispondenti alle unità sintagmatiche (frammenti di testo, frasi, paragrafi) in cui può essere suddiviso ogni contesto iniziale;

B.3 **sottoinsiemi del corpus**, corrispondenti a gruppi di documenti primari riconducibili alla stessa "categoria" (es. interviste di "uomini" o di "donne", articoli di un particolare anno o di una particolare testata, etc.) o a cluster tematici ottenuti con specifici strumenti **T-LAB**.

Il diagramma seguente illustra le possibili relazioni tra unità lessicali e unità di contesto che **T-LAB** ci permette di analizzare.

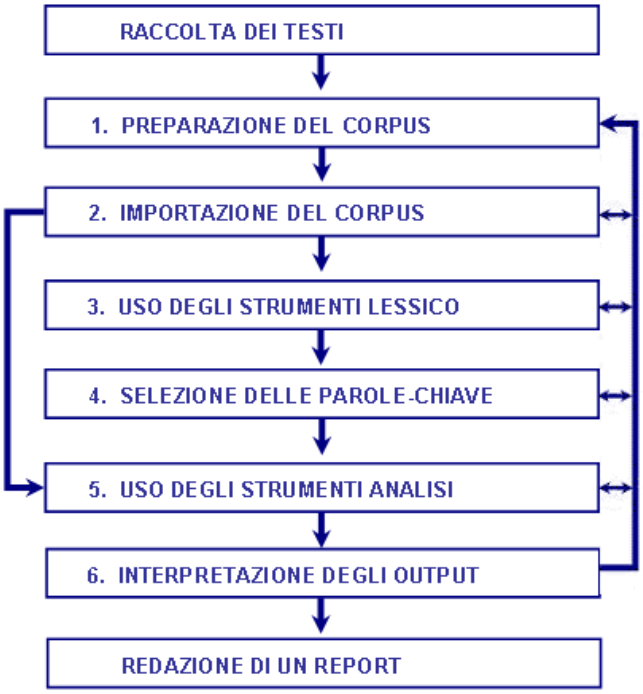


A partire da questa organizzazione del database, **T-LAB** consente - in modo automatico - di esplorare e di analizzare le relazioni tra le unità di analisi di **tutto il corpus** o di suoi **sottoinsiemi**.

In **T-LAB**, la selezione di un qualsivoglia strumento di analisi (click del mouse) attiva sempre un processo semiautomatico che, con poche e semplici operazioni, genera qualche tabella input, applica qualche algoritmo di tipo statistico e produce alcuni output.

In ipotesi, un tipico **progetto** di lavoro in cui viene usato **T-LAB** è costituito dall'insieme delle attività analitiche (operazioni) che hanno per oggetto il medesimo **corpus** ed è organizzato da una **strategia** e da un **piano** dell'utilizzatore. Quindi, inizia con la **raccolta dei testi** da analizzare e termina con la **redazione di un report**.

La successione delle varie fasi è illustrata nel diagramma seguente:



N.B.:
- Le sei fasi numerate, dalla preparazione del corpus all'interpretazione degli output, sono supportate da strumenti **T-LAB** e sono sempre reversibili;

- Tramite le impostazioni automatiche è possibile evitare due fasi (3 e 4); tuttavia, ai fini della **qualità** dei risultati, si raccomanda l'uso delle funzioni **Personalizzazione del Dizionario** (strumento del menu "Lessico") e **Impostazioni Personalizzate** (cioè selezione delle parole-chiave).

Proviamo ora a commentare le varie fasi una dopo l'altra:

1 - La **PREPARAZIONE DEL CORPUS** consiste nella trasformazione dei testi da analizzare in un file (**corpus**) che può essere elaborato dal software.

Nel caso di un unico testo (o di un corpus trattato come unico testo) **T-LAB** non richiede ulteriori accorgimenti.

Quando, invece, il corpus è costituito da più testi e vengono utilizzate **codifiche** che rinviano all'uso di qualche **variabile**, nella fase di preparazione bisogna utilizzare il modulo **Corpus Builder** che – in maniera automatica – procede alla trasformazione di vari materiali testuali in un file corpus pronto per essere importato da **T-LAB**.

N.B.:

- Al termine della fase di preparazione si raccomanda di creare una nuova cartella di lavoro con al suo interno il solo file corpus da importare.

- Durante le analisi, si raccomanda di tenere il file corpus e la relativa cartella di lavoro su un hard disk dello stesso computer su in cui è installato **T-LAB**. Diversamente, l'esecuzione delle varie procedure potrebbe risultare rallentata e il software potrebbe segnalare degli errori.

2 - **L'IMPORTAZIONE DEL CORPUS** consiste in una serie di **processi automatici** che trasformano il corpus in un insieme di tabelle integrate nel **database T-LAB**.

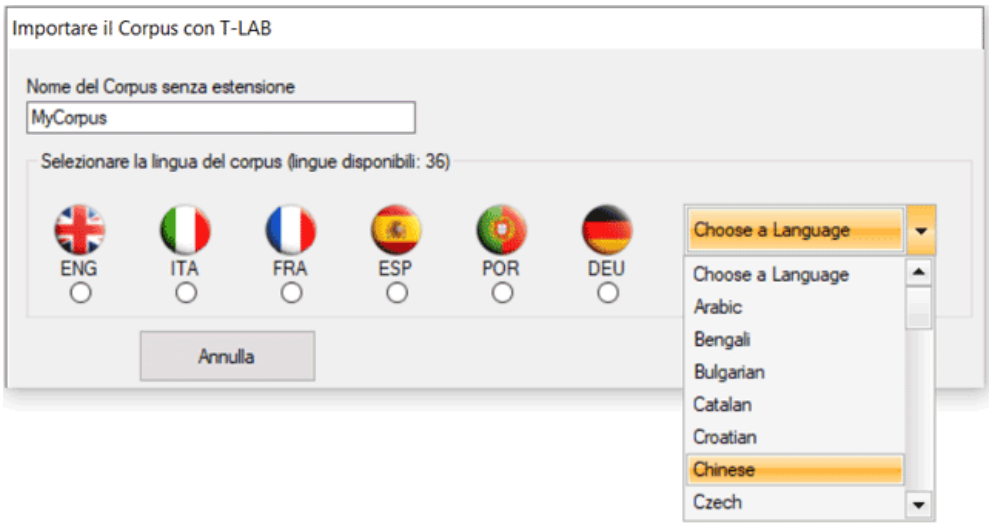
Nella fase di **pre-processing T-LAB** realizza i seguenti trattamenti: **normalizzazione** del testo; riconoscimento di **multi-words** e **stop-words**; **segmentazione** in contesti elementari; **lemmatizzazione** automatica o **stemming**; costruzione del **vocabolario** del corpus; selezione delle **parole chiave**.

Di seguito la lista complete delle trenta (30) lingue per le quali **T-LAB** supporta la lemmatizzazione automatica o lo stemming.

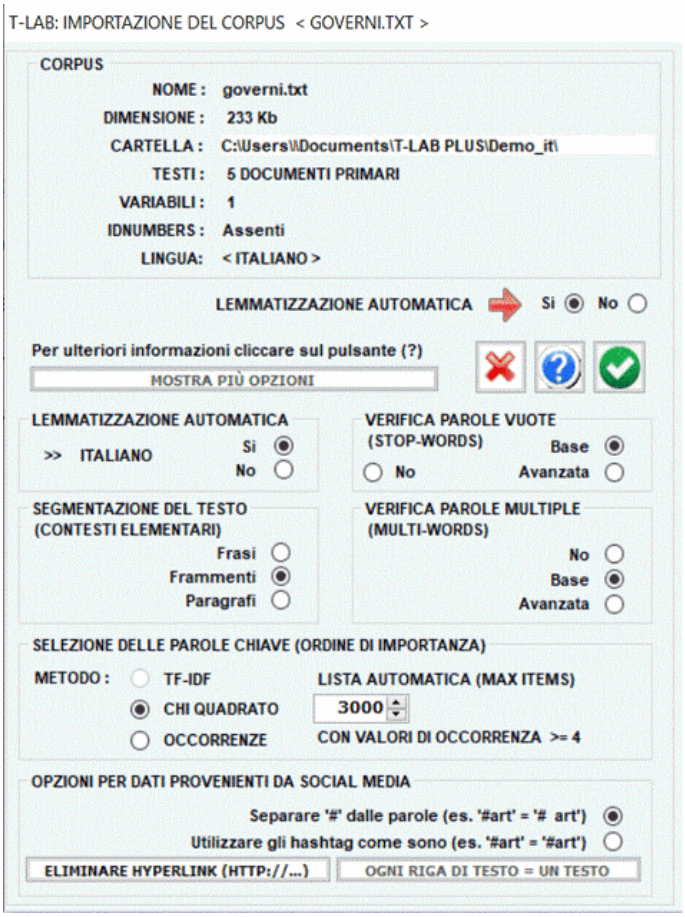
LEMMATIZZAZIONE: catalano, croato, francese, inglese, italiano, latino, polacco, portoghese, rumeno, russo, serbo, slovacco, spagnolo, svedese, tedesco, ucraino.

STEMMING: arabo, bengali, bulgaro, ceco, danese, finlandese, greco, hindi, indonesiano, marathi, norvegese, olandese, persiano, turco, ungherese.

In ogni caso, senza lemmatizzazione automatica e/o usando dizionari personalizzati, possono essere analizzati testi in **tutte le lingue** le cui parole siano separate da spazi e/o da punteggiatura.



A partire dalla selezione della lingua, l'intervento dell'utilizzatore è richiesto per definire le scelte indicate nella finestra seguente:

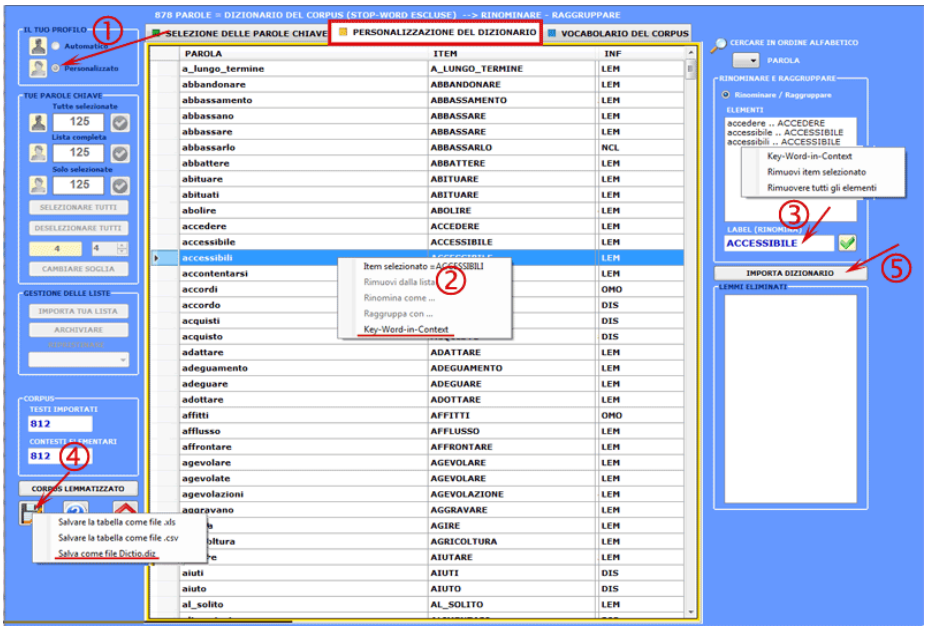


N.B.: Poiché i trattamenti preliminari determinano il tipo e la quantità delle unità di analisi (cioè quali e quante unità di contesto e quali e quante unità lessicali), scelte diverse in questa fase comportano risultati diversi delle successive analisi. Per questa ragione, tutti gli output **T-LAB** mostrati nel manuale e nell'help hanno solo valore indicativo.

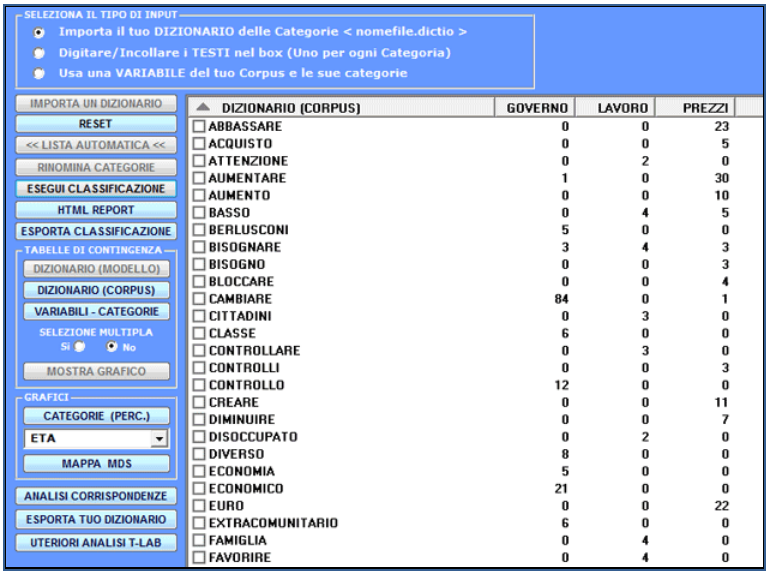
3 - L'USO DEGLI STRUMENTI LESSICO è finalizzato a verificare il corretto **riconoscimento** delle unità lessicali e a personalizzare la loro **classificazione**, cioè a verificare e a modificare le scelte automatiche fatte da **T-LAB**.

Le modalità dei vari interventi sono illustrate nelle corrispondenti voci dell'help (e del manuale).

In particolare si rinvia alla corrispondente voce dell'help (e del manuale) per una dettagliata descrizione del processo **Personalizzazione del Dizionario** (vedi sotto). Infatti, qualsiasi modifica relativa alle voci del dizionario (es., raggruppamento di due o più item) incide sia sul calcolo delle occorrenze che su quello delle co-occorrenze.

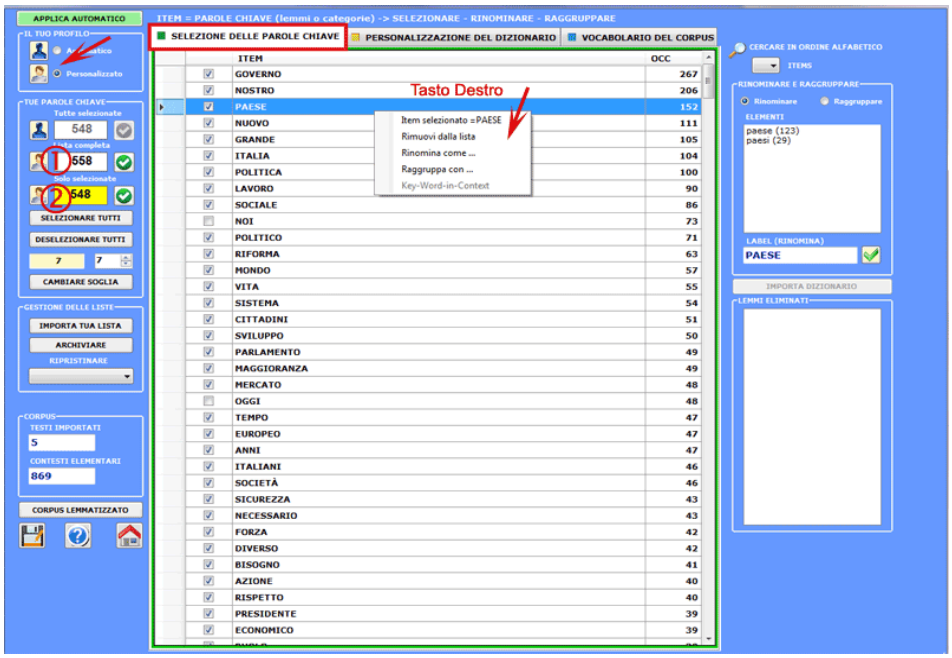


N.B.: Quando l'utilizzatore, senza perdere alcuna informazione lessicale, intende applicare schemi di codifica che raggruppano più parole o lemmi in poche categorie (da 2 a 50) è consigliabile utilizzare lo strumento **Classificazione Basata su Dizionari** incluso nel sottomenu **Analisi Tematiche** (vedi sotto).



4 - LA SELEZIONE DELLE PAROLE-CHIAVE consiste nella predisposizione di una o più liste di unità lessicali (parole, lemmi o categorie) da utilizzare per costruire le tabelle dati da analizzare.

L'opzione **impostazioni automatiche** rende disponibile liste di **parole chiave** selezionate da **T-LAB**; tuttavia, poiché la scelta delle unità di analisi è estremamente rilevante ai fini delle successive elaborazioni, si consiglia vivamente l'uso delle **impostazioni personalizzate**. In questo modo l'utilizzatore potrà scegliere di modificare la lista suggerita da **T-LAB** e/o di costruire liste che meglio corrispondono ai suoi obiettivi di indagine.



In ogni caso, nella costruzione di queste liste, valgono i seguenti criteri:

- verificare la **rilevanza** quantitativa (totale delle occorrenze) e qualitativa (non banalità del significato) dei vari item;

- verificare le **limitazioni** degli strumenti analitici che si intendono utilizzare (vedi nota a fine di questo capitolo);

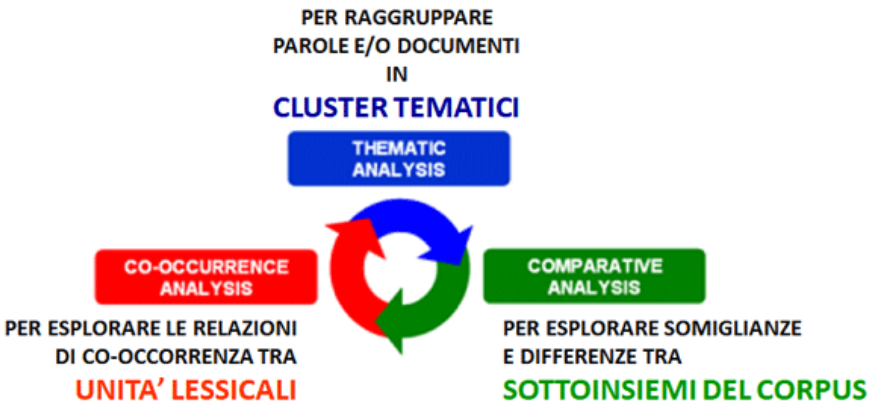
- verificare se l'insieme degli item è compatibile con la propria **strategia** di indagine (vedi punto seguente: 5).

5 - L'USO DEGLI STRUMENTI D'ANALISI è finalizzato alla produzione di output (tabelle e grafici) che rappresentano **relazioni significative** tra le unità di analisi e che consentono di fare **inferenze**.

Attualmente **T-LAB** include venti diversi strumenti di analisi, ciascuno dei quali funziona con una sua specifica logica; cioè, usa specifici algoritmi e produce specifici output.

Di conseguenza, a seconda della tipologia di testi che intende analizzare e degli obiettivi che intende perseguire, l'utilizzatore deve di volta in volta decidere quali strumenti sono più appropriati per la sua **strategia di analisi**.

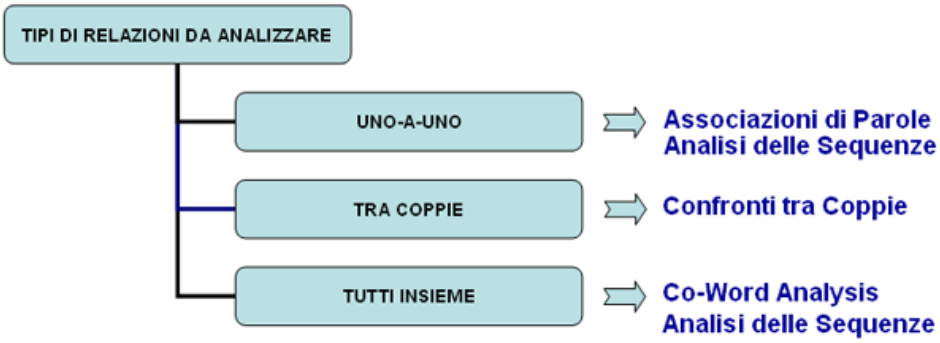
A questo proposito, oltre alla distinzione fra strumenti per **analisi delle co-occorrenze**, per **analisi comparative** e per **analisi tematiche**, è utile considerare che alcuni di questi ultimi consentono di ottenere ulteriori sottoinsiemi del corpus basati su similarità di contenuto.



In generale, anche se l'uso degli strumenti **T-LAB** può essere circolare e reversibile, possiamo individuare tre punti di avvio (start points) che corrispondono ai tre sub-menu ANALISI:

A : **STRUMENTI PER ANALISI DELLE CO-OCCORRENZE**

Questi strumenti consentono di analizzare vari tipi di relazioni tra le unità lessicali (parole, lemmi o categorie).

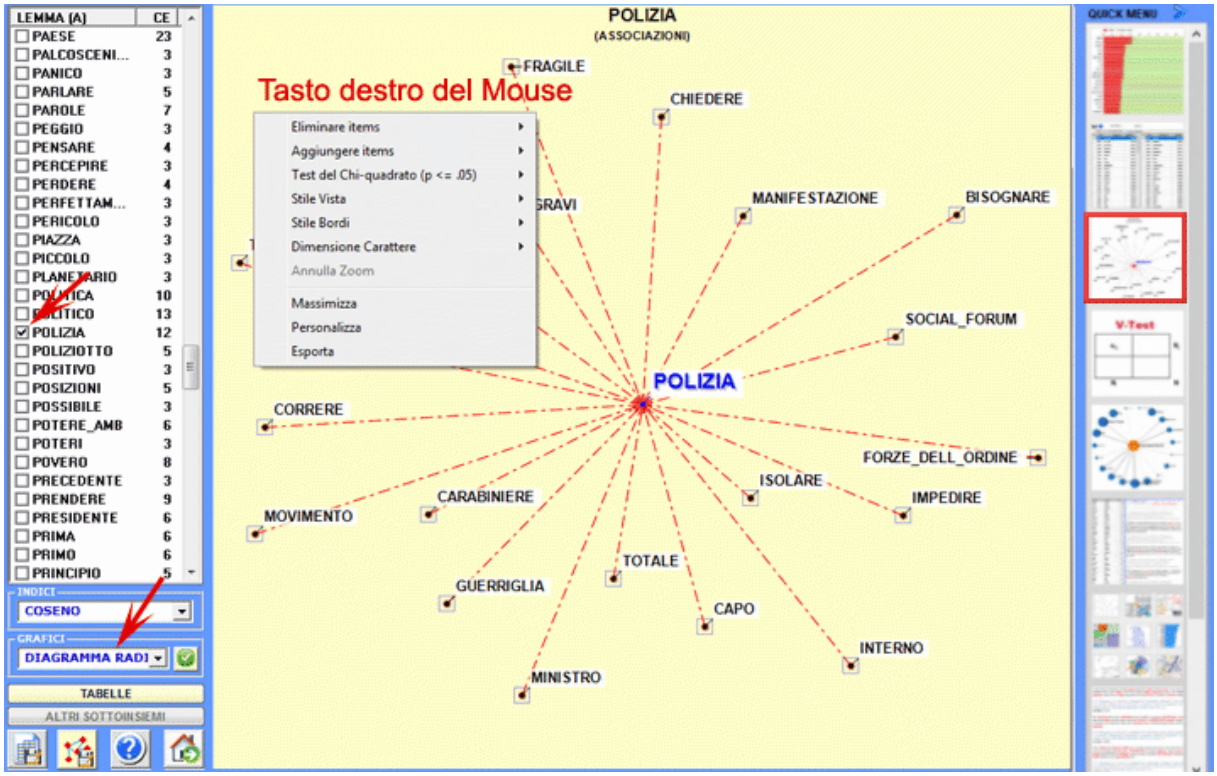


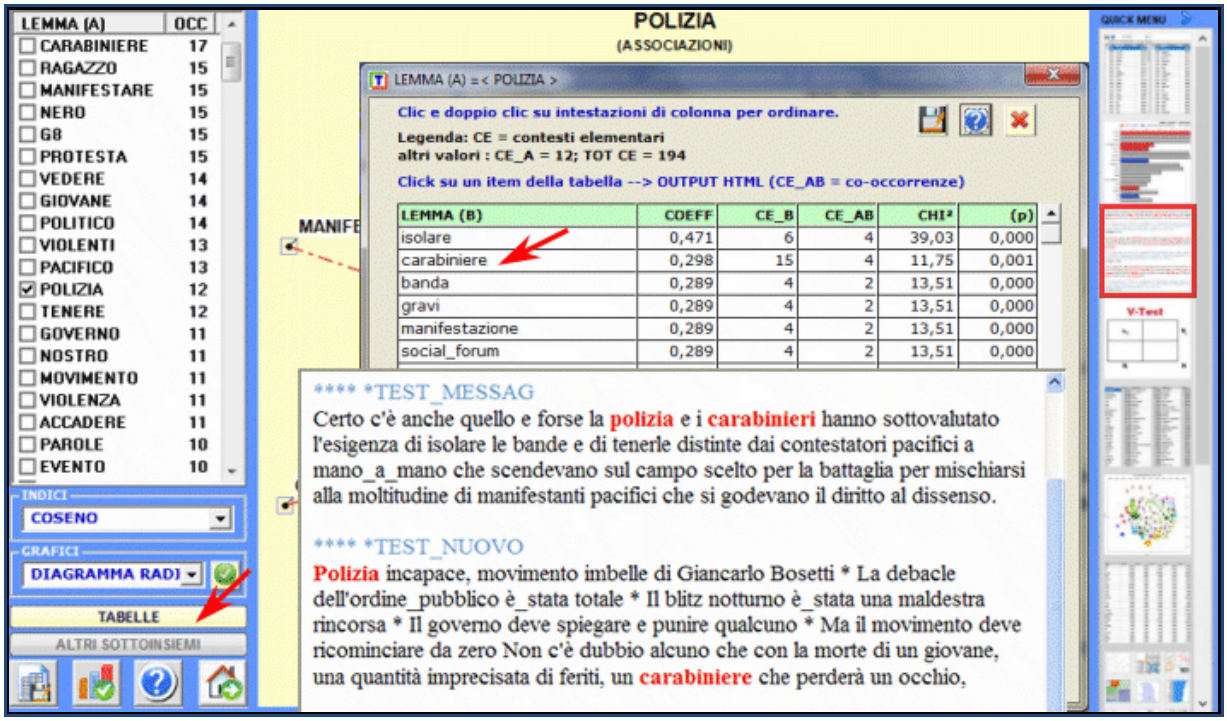
A seconda dei tipi di relazioni da analizzare, le funzioni **T-LAB** indicate in questo diagramma (box colorati) usano uno o più dei seguenti strumenti statistici: **Indici di Associazione**, **Test del Chi Quadro**, **Cluster Analysis**, **Multidimensional Scaling, Analisi delle Componenti Principlali, t-SNE** e **Catene Markoviane**.

Ecco alcuni esempi di output (N.B.: per ulteriori informazioni sulla interpretazione degli output si rimanda alle corrispondenti sezioni della guida / manuale):

- Associazioni di Parole

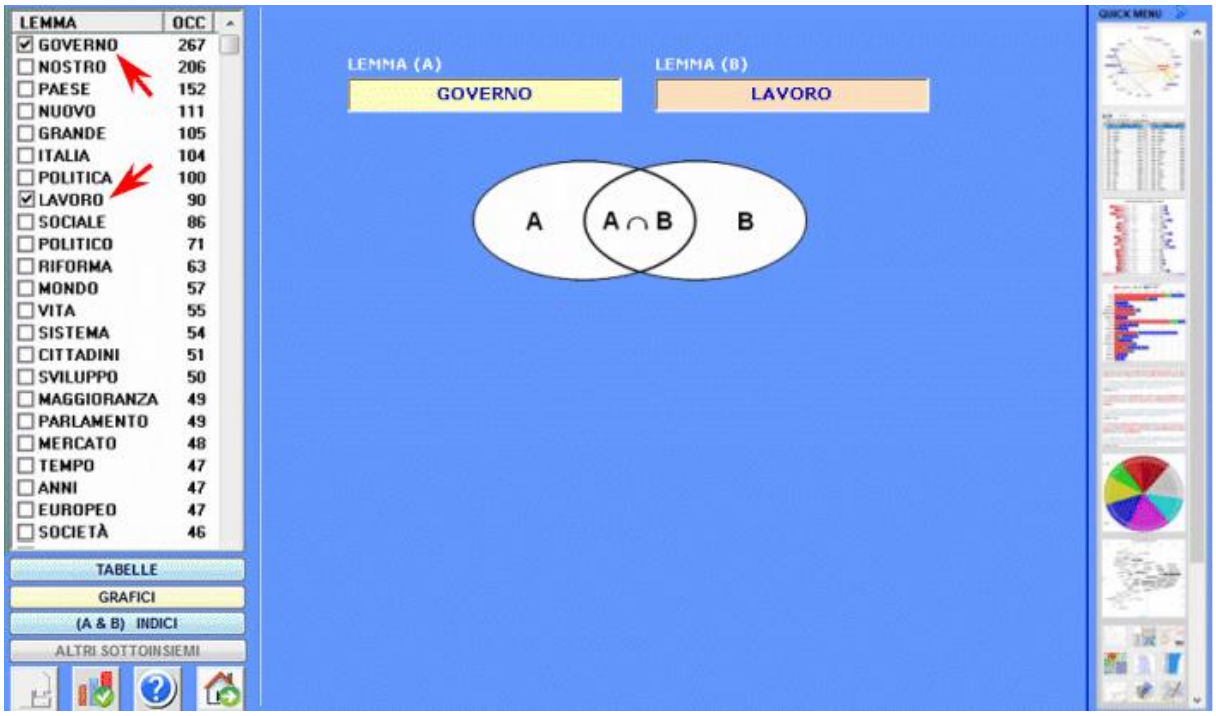
Questo strumento **T-LAB** ci consente di verificare come i contesti di **co-occorrenza** determinano il significato locale delle **parole chiave**.

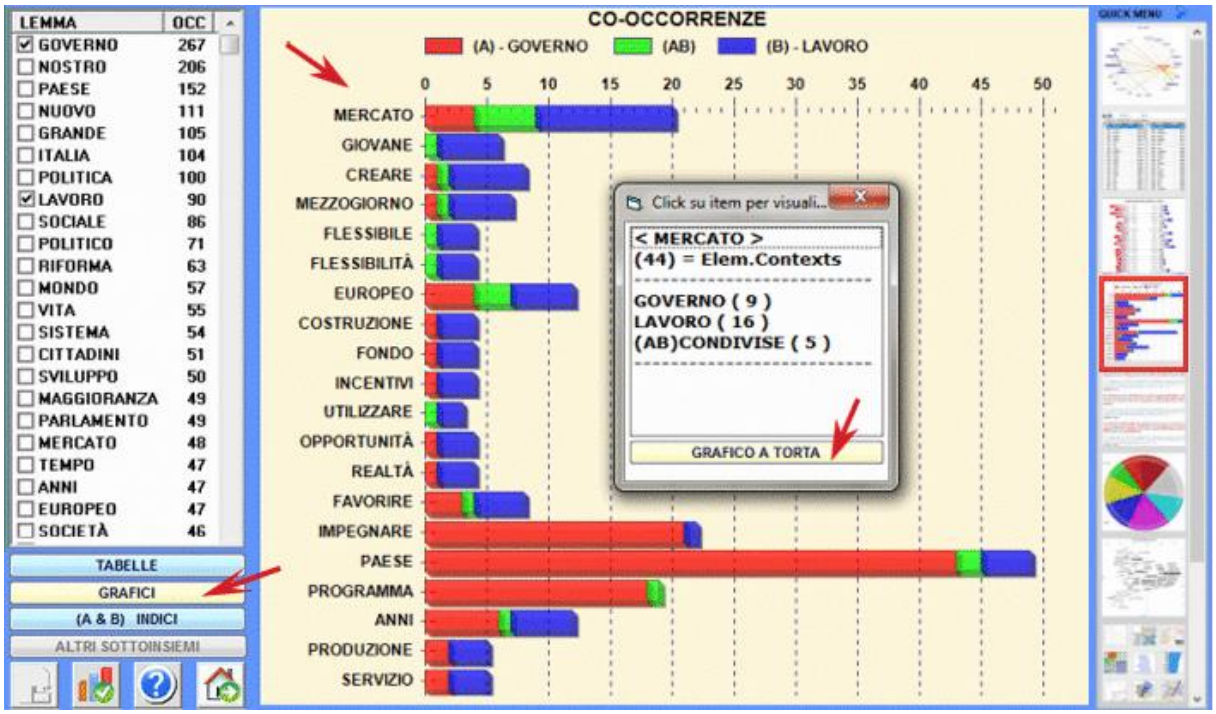




- Confronti tra Coppie

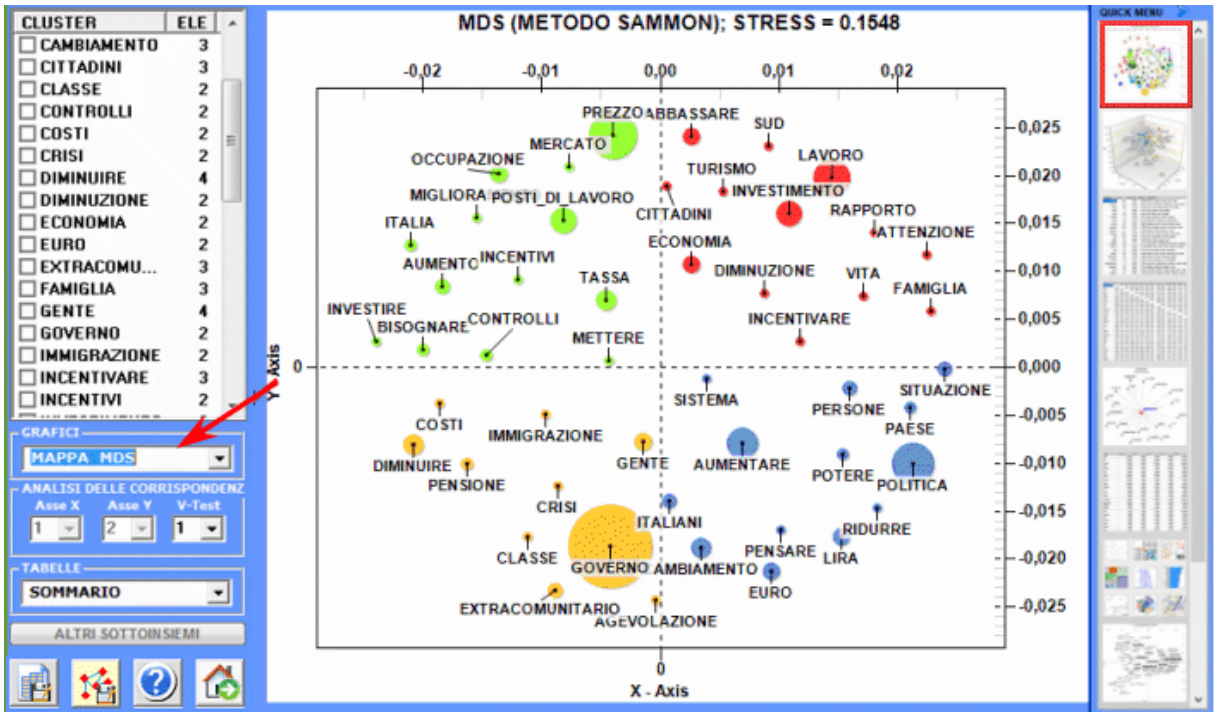
Questo strumento **T-LAB** consente di confrontare insiemi di **contesti elementari** (cioè contesti di co-occorrenza) in cui sono presenti gli elementi di una coppia di **parole chiave**.

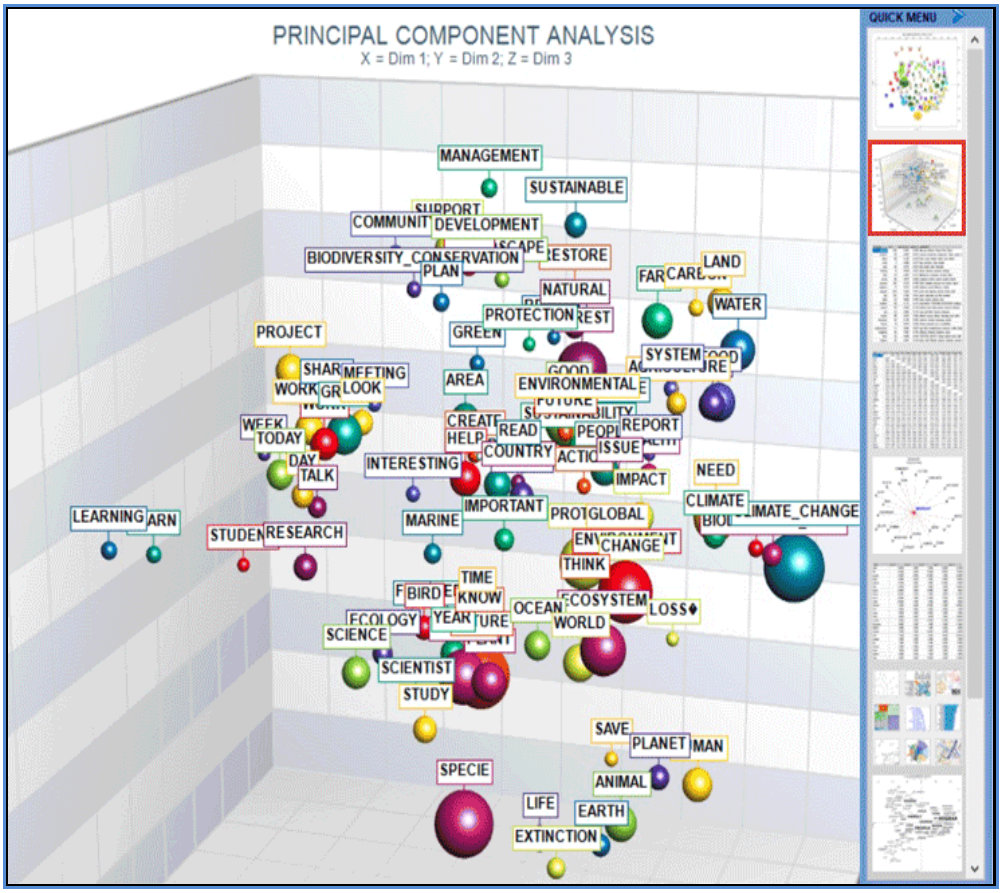




- Co-Word Analysis

L'uso di questa funzione **T-LAB** consente di analizzare le relazioni di **co-occorrenza** all'interno di gruppi di parole chiave.





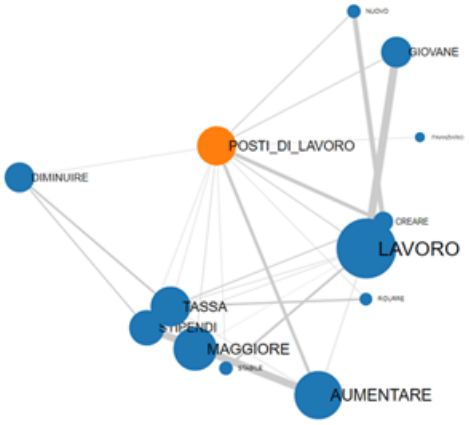
- Analisi delle Sequenze e Network Analysis

Questo strumento **T-LAB** tiene conto delle **posizioni** delle varie unità lessicali all'interno delle frasi e ci permette di rappresentare ed esplorare qualsiasi testo come una **rete** di relazioni.

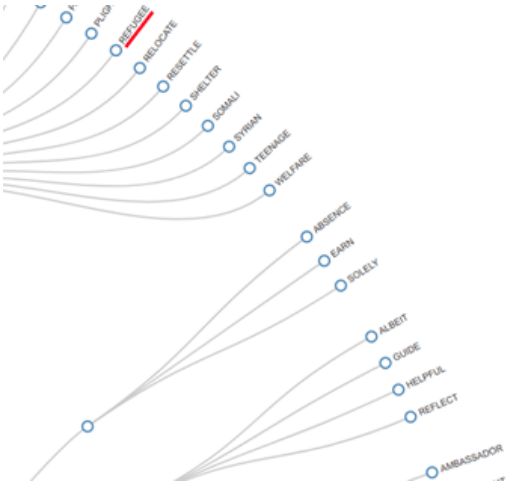
Ciò significa, dopo aver eseguito questo tipo di analisi, l'utilizzatore può verificare le relazioni tra i nodi della rete (cioè le parole chiave) a diversi livelli: a) in relazioni del tipo uno-a-uno; b) all'interno di 'ego network'; c) all'interno delle 'comunità' a cui appartengono; d) all'interno dell'intera rete costituita dal testo in analisi.

RELAZIONI DEL TIPO UNO-AD-UNO

EGO-NETWORK

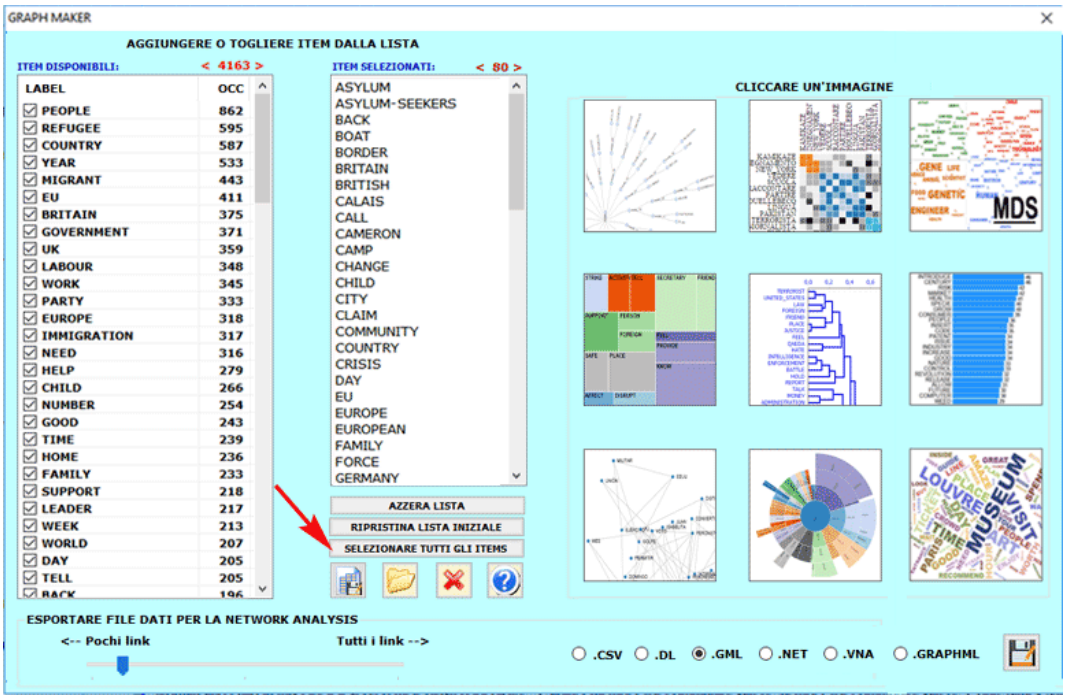


COMUNITA'



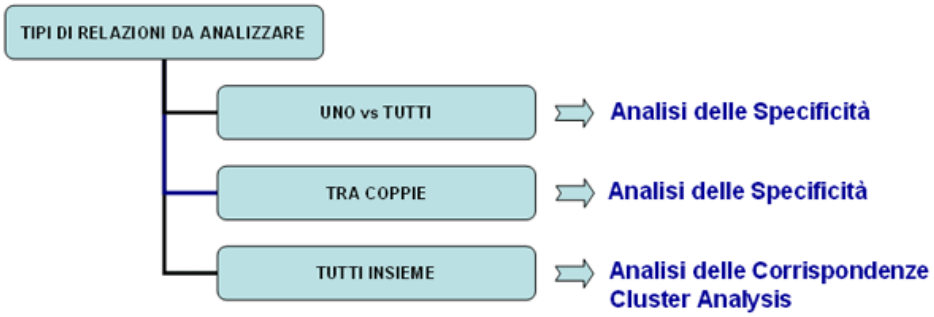
INTERA RETE

Inoltre, facendo clic sull'opzione **GRAPH MAKER**, l'utente può creare diversi tipi di grafici utilizzando elenchi personalizzati di parole chiave (vedi sotto).

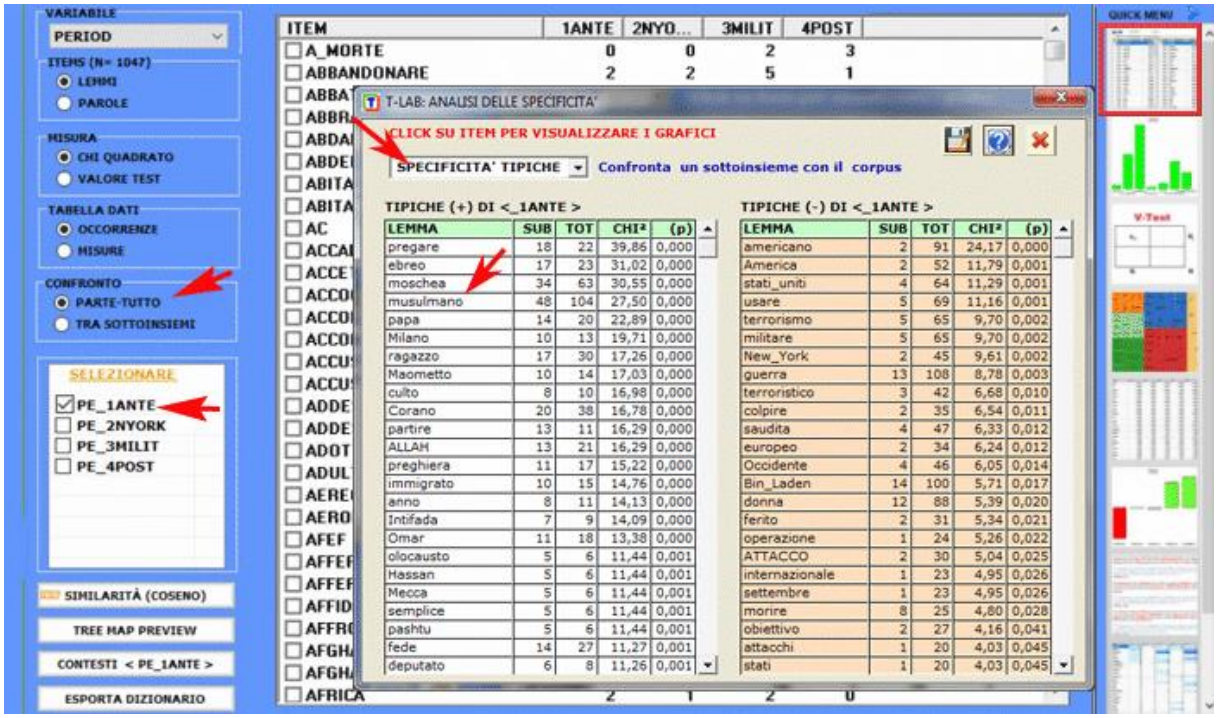


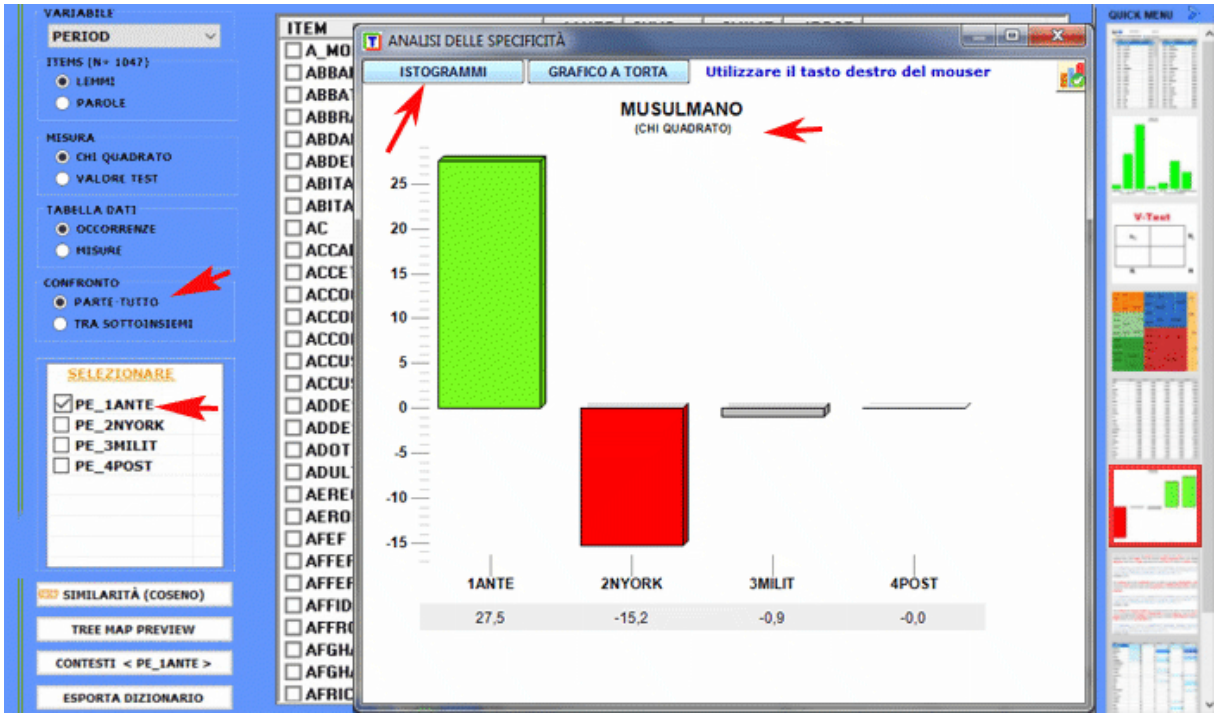
B : STRUMENTI PER ANALISI COMPARATIVE

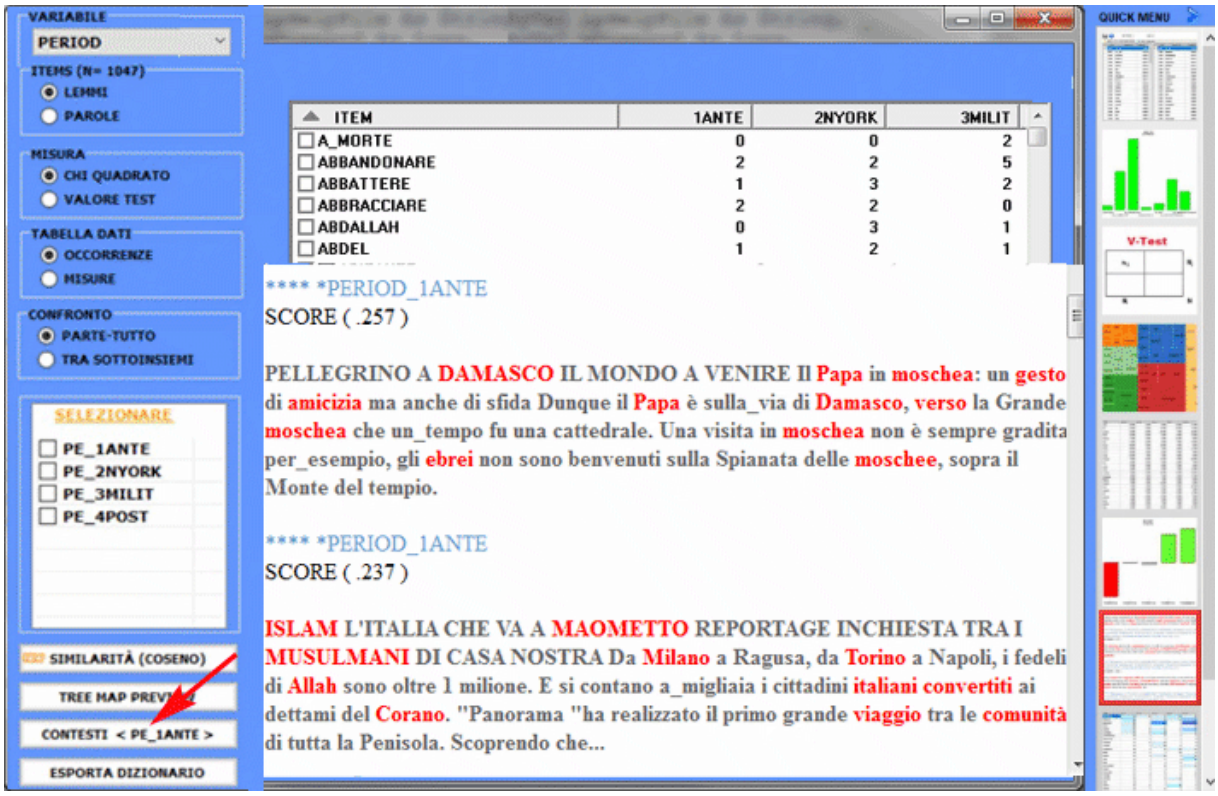
Questi strumenti consentono di analizzare vari tipi di relazioni tra le unità di contesto.



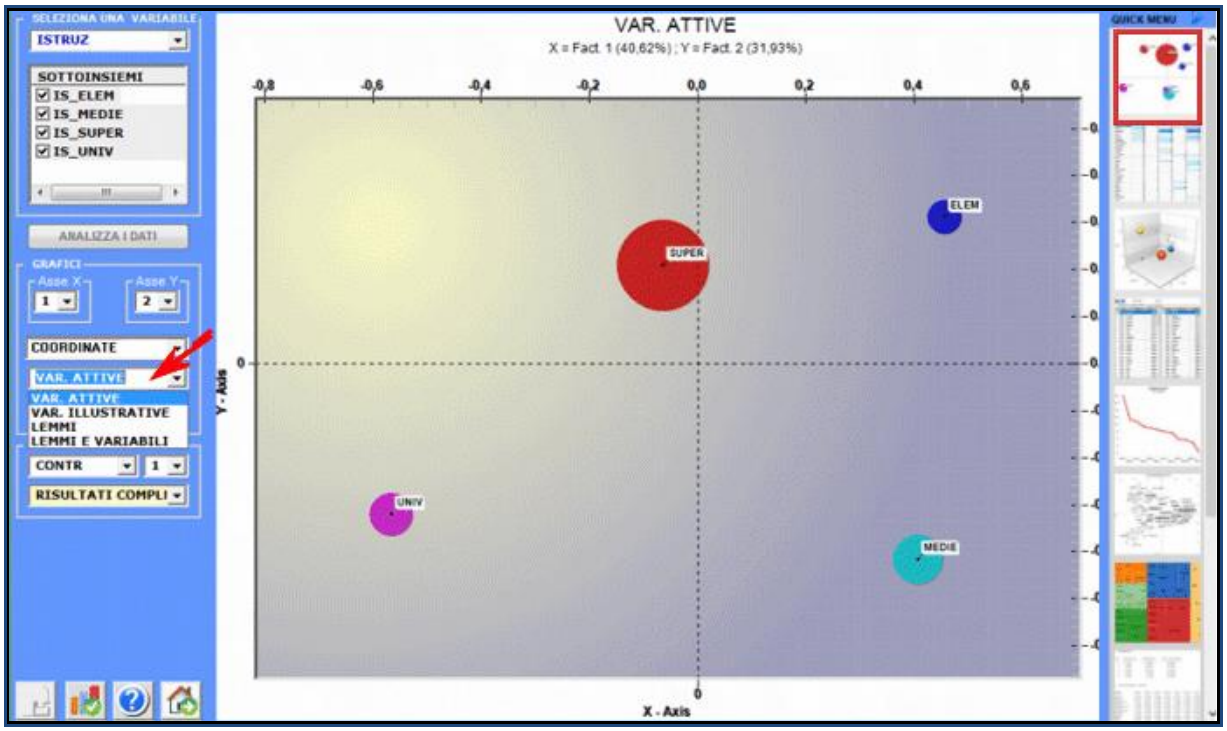
L'**Analisi delle Specificità** consente di verificare quali **parole** sono **tipiche** o **esclusive** di ogni specifico sottoinsieme del corpus. Inoltre permette di estrarre i **contesti tipici**, cioè i contesti elementari caratteristici, di ciascuno dei sottoinsiemi analizzati (ad esempio, le 'tipiche' frasi usate da specifiche leader politici).

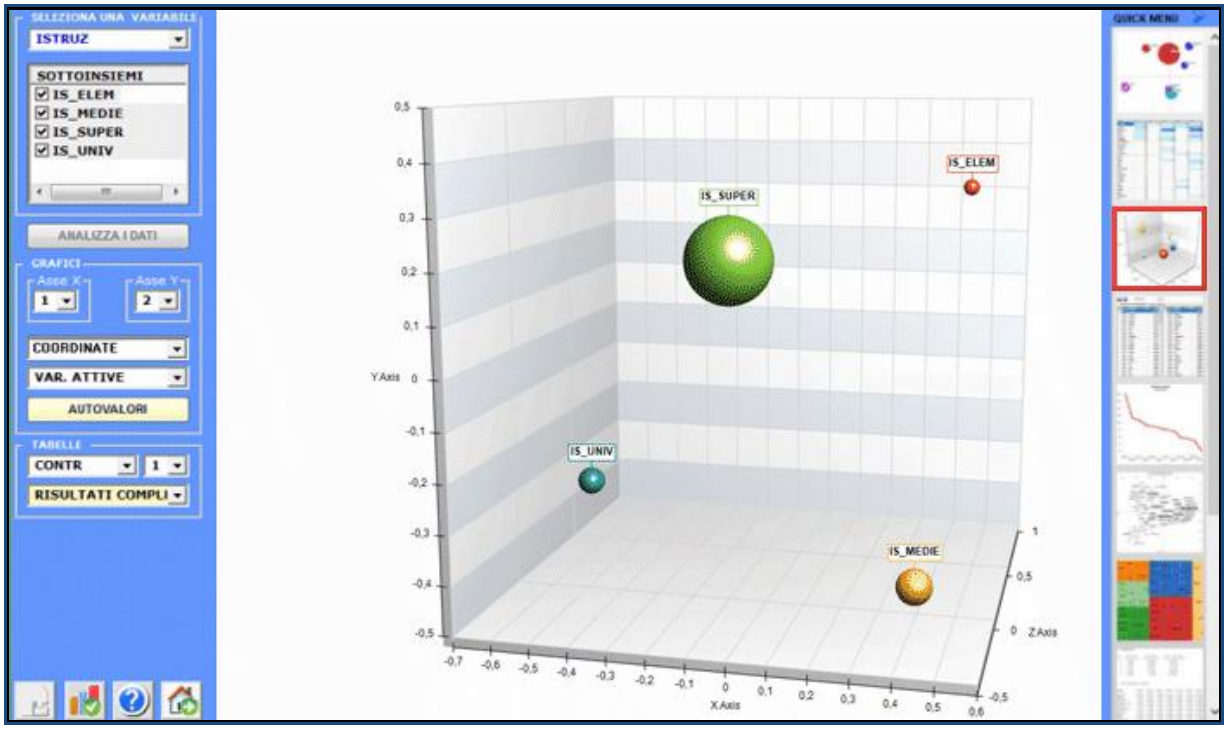

VARIABILE
PERIOD
ITEMS (N= 1047)
LEMMI
PAROLE
MISURA
CHI QUADRATO
VALORE TEST
TABELLA DATI
OCCORRENZE
MISURE
CONFRONTO
PARTE-TUTTO
TRA SOTTOINSIEMI
SELEZIONARE
PE_1ANTE
PE_2NYORK
PE_3MILIT
PE_4POST
SIMILARITÀ (COSENO)
TREE MAP PREVIEW
CONTESTI < PE_1ANTE >
ESPORTA DIZIONARIO
ITEM
1ANTE
2NYO...
3MILIT
4POST
A_MORTE 0 0 2 3
ABBANDONARE 2 2 5 1
T-LAB: ANALISI DELLE SPECIFICITA'
CLICK SU ITEM PER VISUALIZZARE I GRAFICI
SPECIFICITA' TIPICHE
Confronta un sottoinsieme con il corpus
TIPICHE (+) DI <_1ANTE >
LEMMA SUB TOT CHI² (p)
pregare 18 22 39,86 0,000
ebreo 17 23 31,02 0,000
moschea 34 63 30,55 0,000
musulmano 48 104 27,50 0,000
papa 14 20 22,89 0,000
Milano 10 13 19,71 0,000
ragazzo 17 30 17,26 0,000
Maometto 10 14 17,03 0,000
culto 8 10 16,98 0,000
Corano 20 38 16,78 0,000
partire 13 11 16,29 0,000
ALLAH 13 21 16,29 0,000
preghiera 11 17 15,22 0,000
immigrato 10 15 14,76 0,000
anno 8 11 14,13 0,000
Intifada 7 9 14,09 0,000
Omar 11 18 13,38 0,000
olocausto 5 6 11,44 0,001
Hassan 5 6 11,44 0,001
Mecca 5 6 11,44 0,001
semplice 5 6 11,44 0,001
pashtu 5 6 11,44 0,001
fede 14 27 11,27 0,001
deputato 6 8 11,26 0,001
TIPICHE (-) DI <_1ANTE >
LEMMA SUB TOT CHI² (p)
americano 2 91 24,17 0,000
America 2 52 11,79 0,001
stati_uniti 4 64 11,29 0,001
usare 5 69 11,16 0,001
terrorismo 5 65 9,70 0,002
militare 5 65 9,70 0,002
New_York 2 45 9,61 0,002
guerra 13 108 8,78 0,003
terroristico 3 42 6,68 0,010
colpire 2 35 6,54 0,011
saudita 4 47 6,33 0,012
europeo 2 34 6,24 0,012
Occidente 4 46 6,05 0,014
Bin_Laden 14 100 5,71 0,017
donna 12 88 5,39 0,020
ferito 2 31 5,34 0,021
operazione 1 24 5,26 0,022
ATTACCO 2 30 5,04 0,025
internazionale 1 23 4,95 0,026
settembre 1 23 4,95 0,026
morire 8 25 4,80 0,028
obiettivo 2 27 4,16 0,041
attacchi 1 20 4,03 0,045
stati 1 20 4,03 0,045
AFRICA 2 1 2 0
QUICK MENU


VARIABILE
PERIOD
ITEMS (N= 1047)
LEMMI
PAROLE
MISURA
CHI QUADRATO
VALORE TEST
TABELLA DATI
OCCORRENZE
MISURE
CONFRONTO
PARTE-TUTTO
TRA SOTTOINSIEMI
SELEZIONARE
PE_1ANTE
PE_2NYORK
PE_3MILIT
PE_4POST
SIMILARITÀ (COSENO)
TREE MAP PREVIEW
CONTESTI < PE_1ANTE >
ESPORTA DIZIONARIO
ITEM
ANALISI DELLE SPECIFICITÀ
ISTOGRAMMI
GRAFICO A TORTA
Utilizzare il tasto destro del mouser
MUSULMANO
(CHI QUADRATO)
25
20
15
10
5
0
-5
-10
-15
1ANTE
2NYORK
3MILIT
4POST
27,5
-15,2
-0,9
-0,0
QUICK MENU

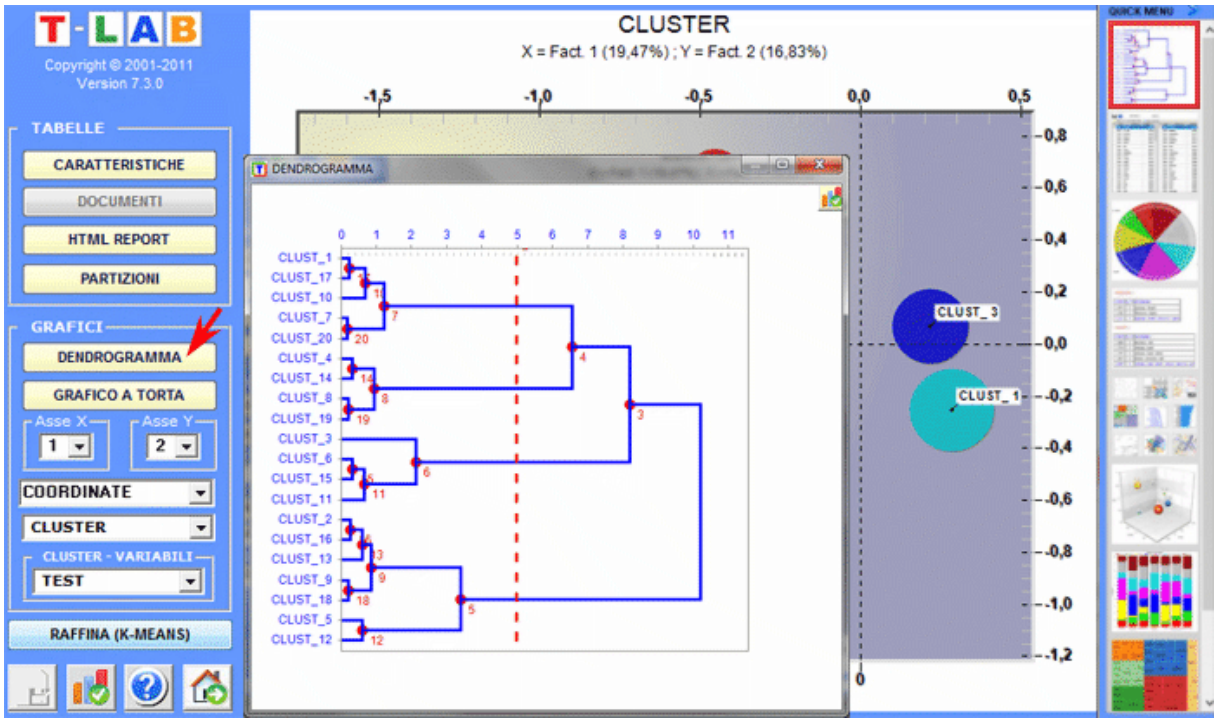


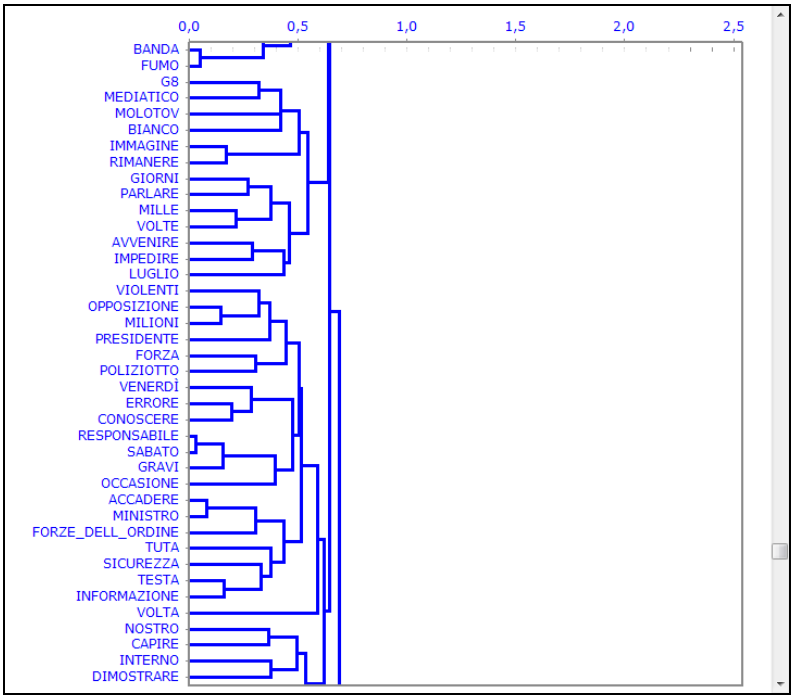
L'**Analisi delle Corrispondenze** consente di esplorare vari tipi di relazioni (somiglianze e differenze) tra gruppi di unità di contesto.





La **Cluster Analysis**, che può essere effettuata con varie tecniche, consente di individuare gruppi di unità di analisi che abbiano due caratteristiche complementari: massima omogeneità al loro interno e massima eterogeneità tra ciascuno di essi e gli altri.





C : STRUMENTI PER ANALISI TEMATICHE

Questi strumenti consentono di individuare, esaminare e mappare i “temi” presenti nei testi analizzati.

Poiché **tema** è una parola polisemica, in questo caso è utile far riferimento ad alcune definizione operative. Infatti, in questi strumenti **T-LAB**, “tema” è una label usata per indicare quattro diverse entità:

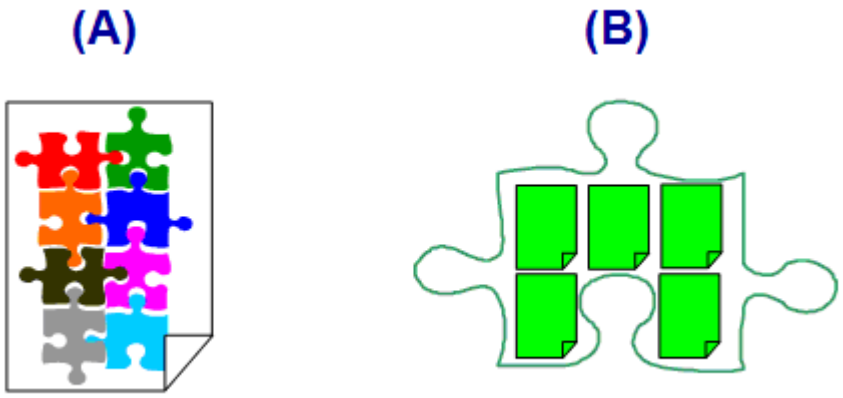
1 - un **cluster tematico di unità di contesto** caratterizzate dagli stessi pattern di parole chiave (vedi gli strumenti Analisi Tematica dei Contesti Elementari, Classificazione Tematica dei Documenti e Classificazione basata su Dizionari);

2 - un **gruppo tematico di parole-chiave** classificate come appartenenti alla stessa categoria (vedi lo strumento Classificazione Basata su Dizionari);

3 - una **componente di un modello probabilistico** che rappresenta ogni unità di contesto (sia essa un contesto elementare o un documento) come generato da una mistura di "temi" o "topics" (vedi gli strumenti Modellizzazione dei Temi Emergenti e Testi e Discorsi come Sistemi Dinamici);

4 - una **specifica parola chiave** usata per estrarre un insieme di contesti elementari in cui essa è associata con uno specifico gruppo di parole preselezionate dall'utilizzatore (vedi lo strumento Contesti Chiave di Parole Tematiche);

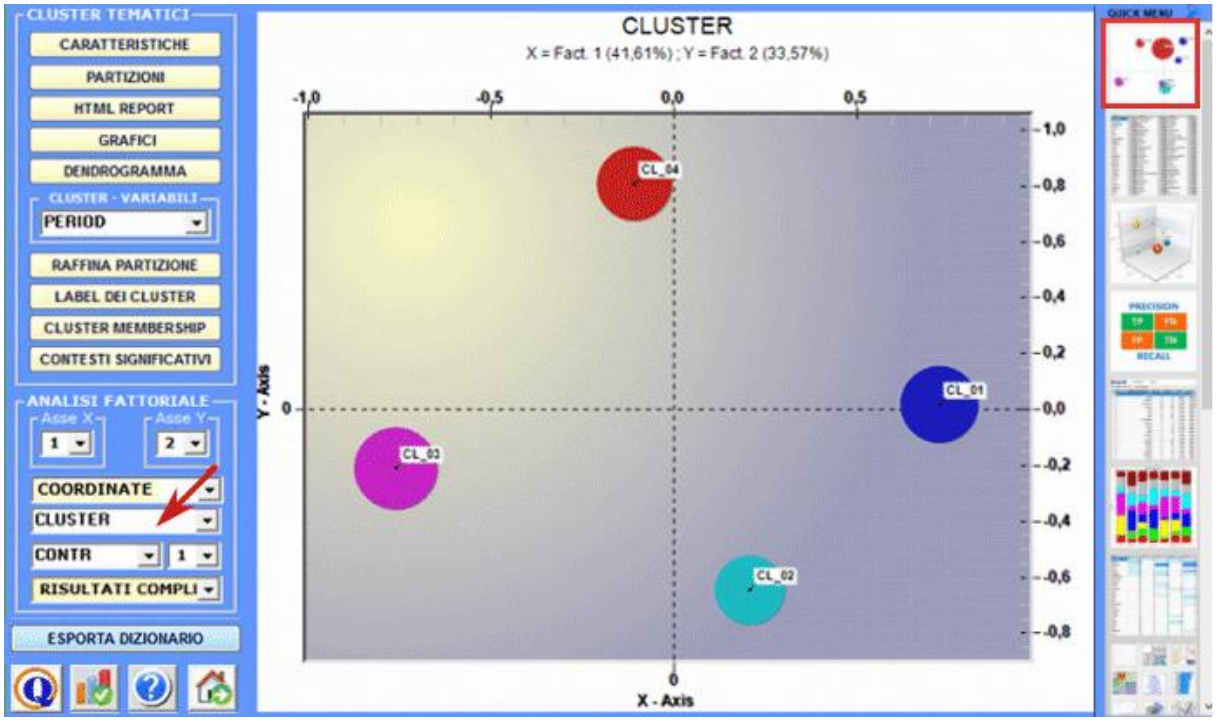
Per esempio, a seconda del tipo di strumento che stiamo usando, uno specifico documento può essere analizzato come composto da vari 'temi' (vedi 'A' sotto) o come appartenente a un insieme di documenti concernenti lo stesso 'tema' (vedi 'B' sotto). Infatti, nel caso 'A' ogni tema può corrispondere ad una parola o a una frase, mentre nel caso 'B' un tema può essere un'etichetta assegnata a un gruppo di documenti caratterizzati da gli stessi pattern di parole chiave.

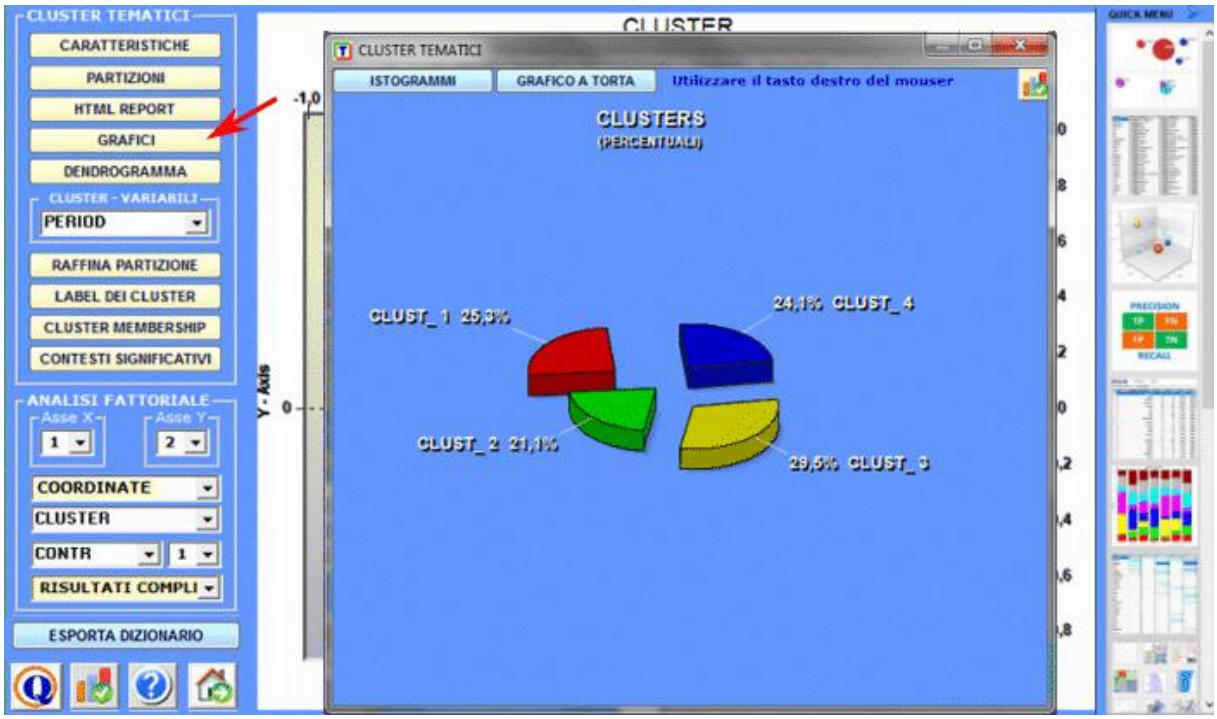


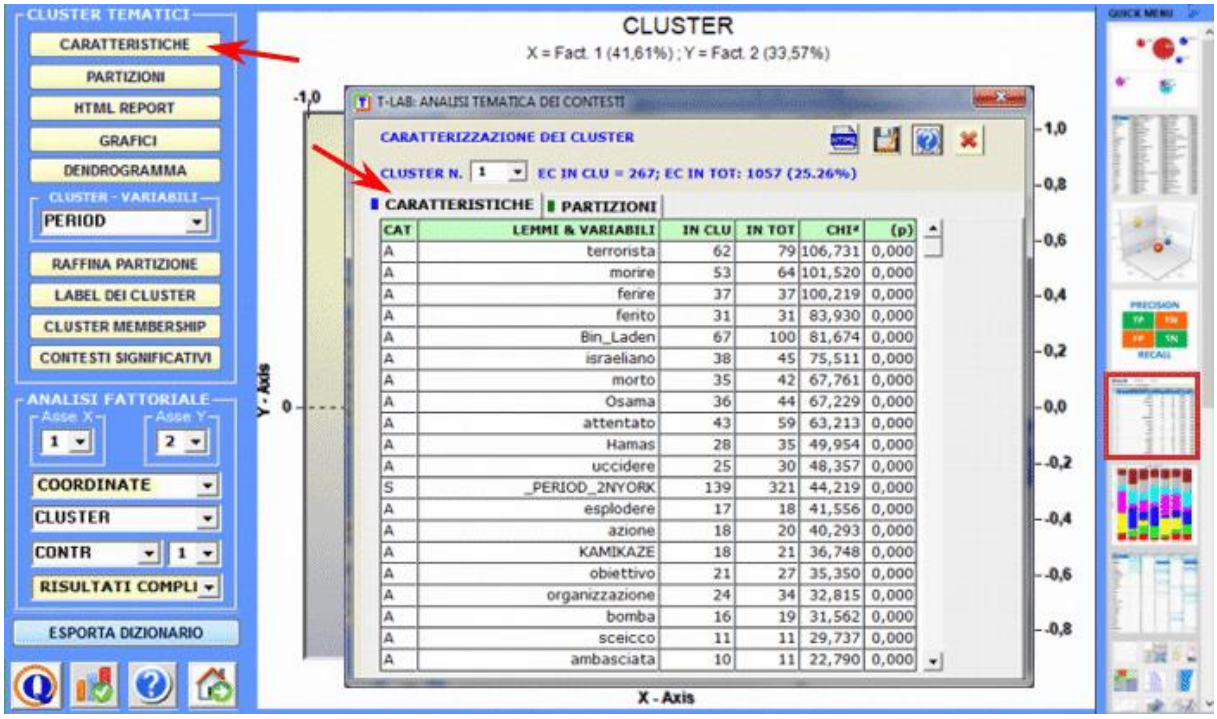
Nel dettaglio, i modi in cui **T-LAB** 'estrae' temi sono i seguenti:

1 - sia l' **Analisi Tematica dei Contesti Elementari** che la **Classificazione Tematica dei Documenti** funzionano nel modo seguente:

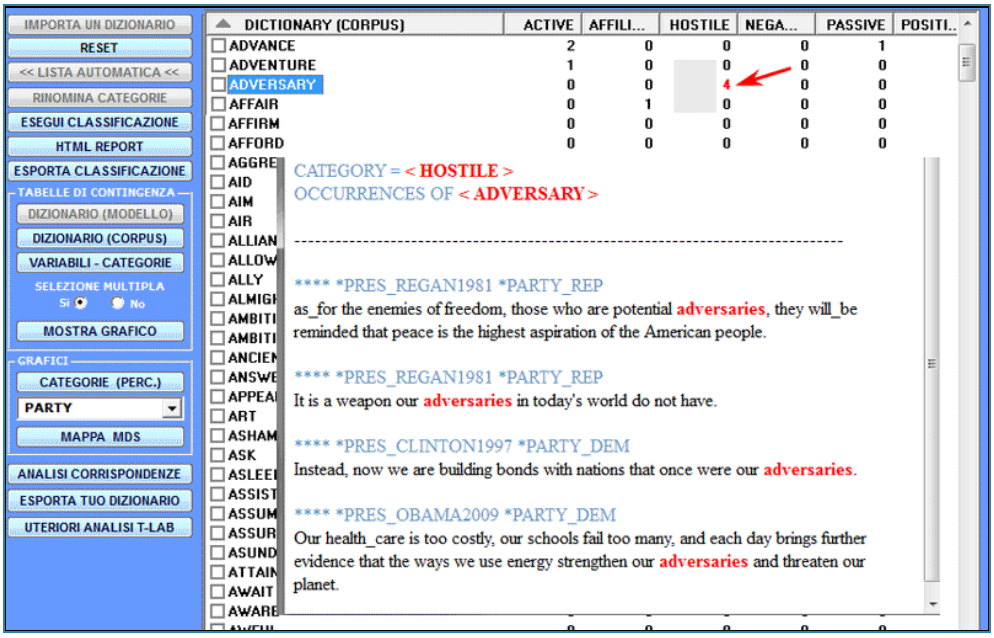
a- realizzano un'**analisi delle co-occorrenze** per individuare cluster tematici di unità di contesto;
b- realizzano un'**analisi comparativa** per confrontare i profili dei vari cluster;
c- producono vari tipi di grafici e tabelle (vedi sotto);
d- consentono di archiviare le **nuove variabili** ottenute (cluster tematici) e di utilizzarle in ulteriori analisi.

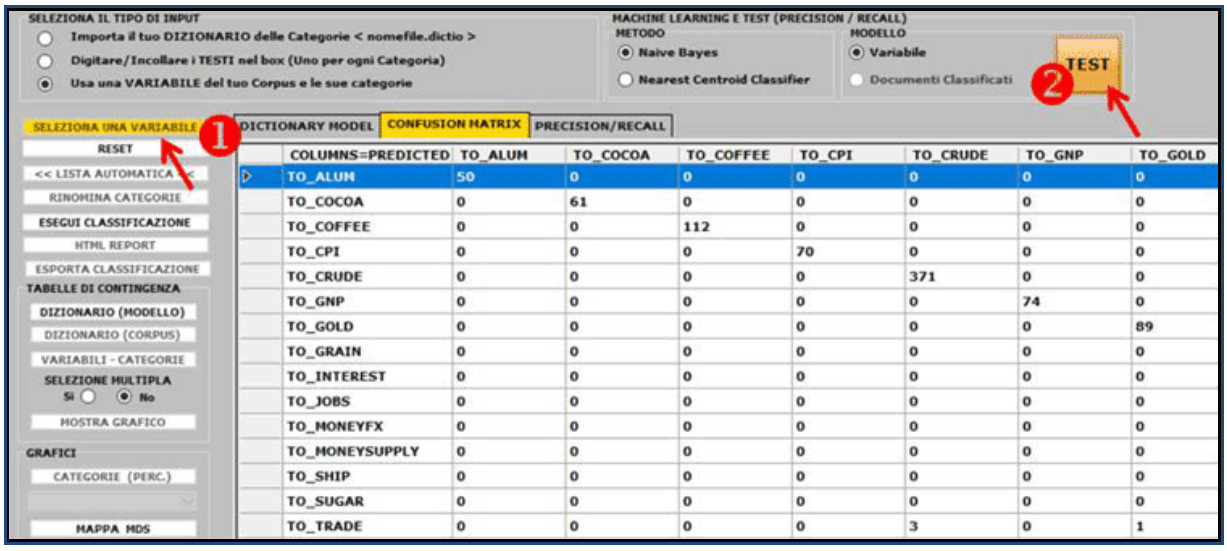




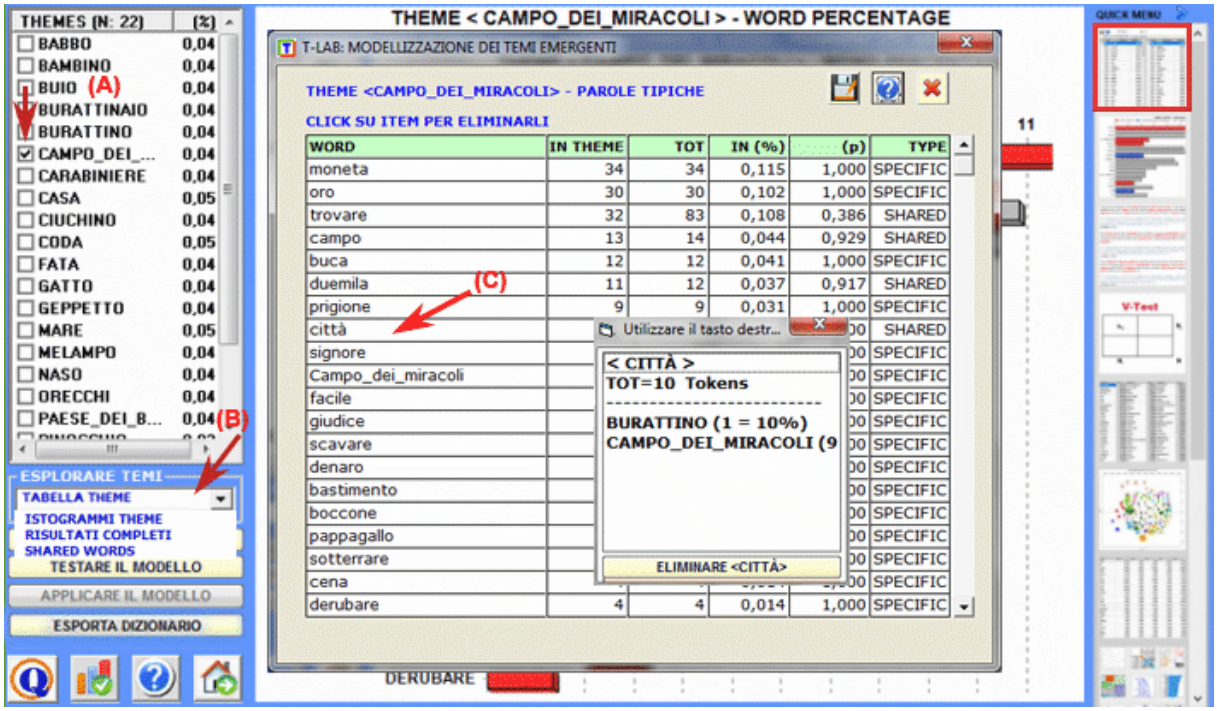


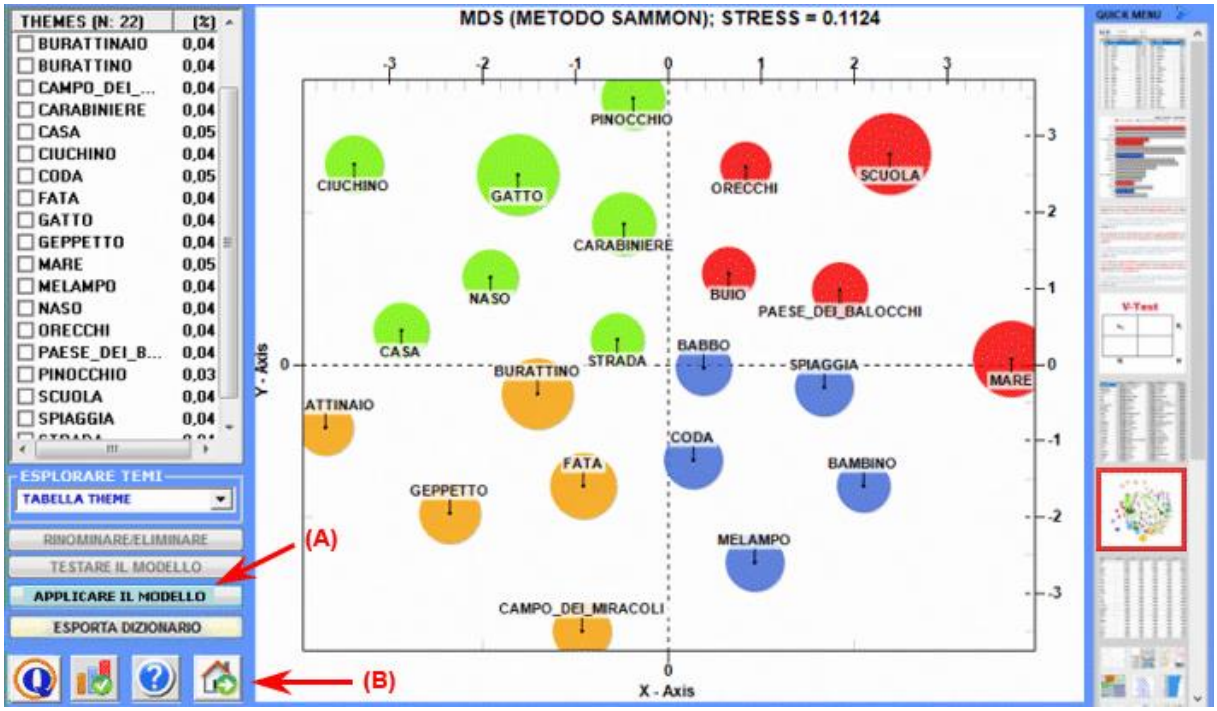
2 - tramite lo strumento **Classificazione Basata su Dizionari** possiamo facilmente costruire / testare / applicare modelli (ad esempio dizionari di categorie) sia per la classica analisi di contenuto che per la sentiment analysis. Infatti questo strumento ci permette di eseguire una classificazione automatica di tipo top-down sia delle unità lessicali (cioè parole e lemmi) che delle unità di contesto (cioè frasi, paragrafi e documenti brevi).


IMPORTA UN DIZIONARIO
RESET
<< LISTA AUTOMATICA <<
RINOMINA CATEGORIE
ESEGUI CLASSIFICAZIONE
HTML REPORT
ESPORTA CLASSIFICAZIONE
TABELLE DI CONTINGENZA
DIZIONARIO (MODELLO)
DIZIONARIO (CORPUS)
VARIABILI - CATEGORIE
SELEZIONE MULTIPLA
Si No
MOSTRA GRAFICO
GRAFICI
CATEGORIE (PERC.)
PARTY
MAPPA MDS
ANALISI CORRISPONDENZE
ESPORTA TUO DIZIONARIO
UTERIORI ANALISI T-LAB
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.

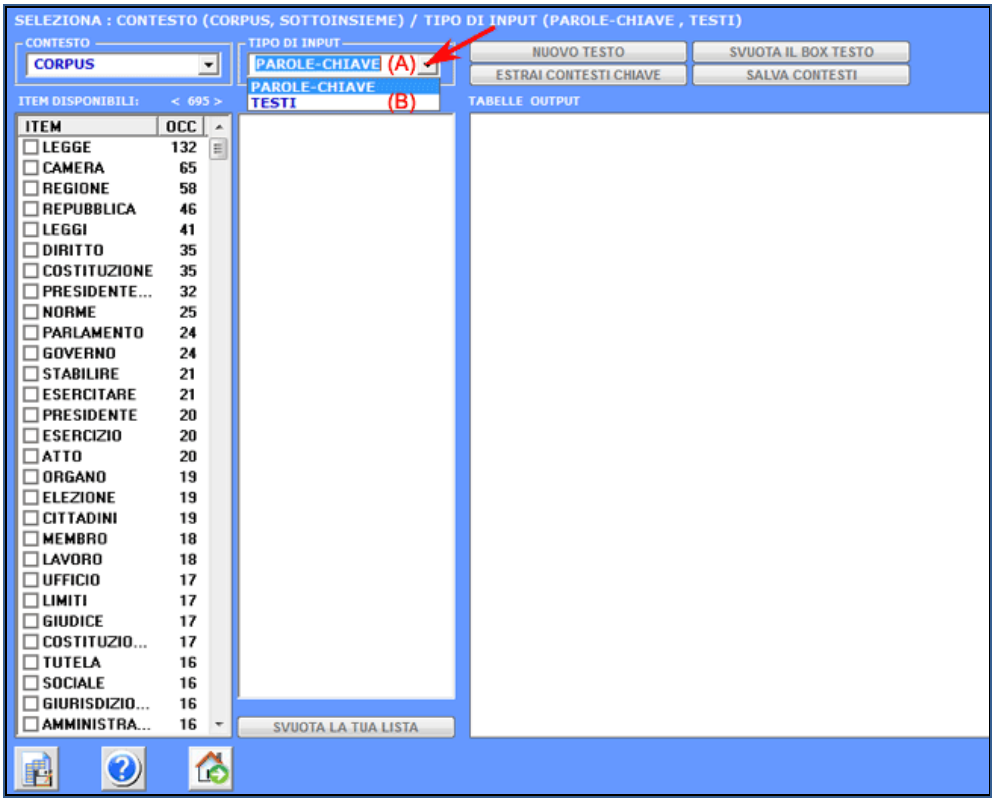

SELEZIONA IL TIPO DI INPUT
Importa il tuo DIZIONARIO delle Categorie < nomefile.dictio >
Digitare/Incollare i TESTI nel box (Uno per ogni Categoria)
Usa una VARIABILE del tuo Corpus e le sue categorie
MACHINE LEARNING E TEST (PRECISION / RECALL)
METODO
Naive Bayes
Nearest Centroid Classifier
MODELLO
Variabile
Documenti Classificati
TEST
SELEZIONA UNA VARIABILE
1
2
DICTIONARY MODEL
CONFUSION MATRIX
PRECISION/RECALL
RESET
<< LISTA AUTOMATICA <<
RINOMINA CATEGORIE
ESEGUI CLASSIFICAZIONE
HTML REPORT
ESPORTA CLASSIFICAZIONE
TABELLE DI CONTINGENZA
DIZIONARIO (MODELLO)
DIZIONARIO (CORPUS)
VARIABILI - CATEGORIE
SELEZIONE MULTIPLA
Si No
MOSTRA GRAFICO
GRAFICI
CATEGORIE (PERC.)
MAPPA MDS
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

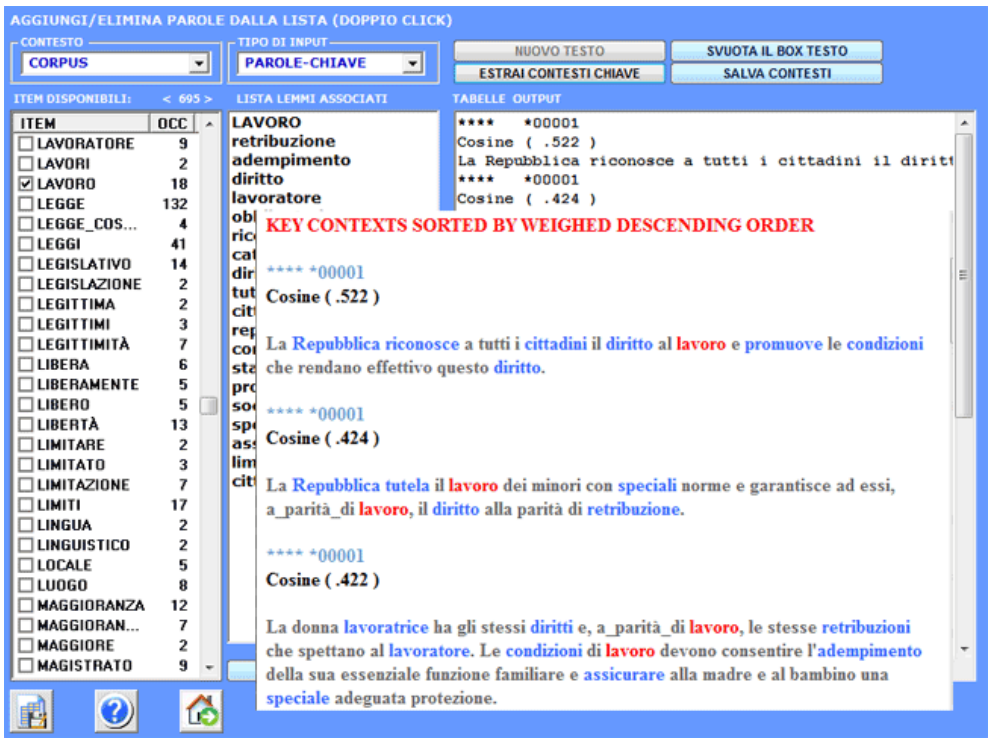
3 - tramite lo strumento **Modellizzazione dei Temi Emergenti** (vedi sotto) i componenti della 'mistura' tematica possono essere descritti attraverso il loro vocabolario caratteristico e possono essere utilizzati per la costruzione di griglie per l'analisi qualitativa e / o per la classificazione automatica delle unità di contesto (cioè contesti elementari o documenti).





4 - lo strumento **Contesti Chiave di Parole Tematiche** (vedi sotto) può essere utilizzato per due diversi scopi: (a) estrarre elenchi di unità di contesto (cioè contesti elementari) che permettono di approfondire il valore tematico di specifiche **parole chiave**; (b) estrarre gruppi di unità di contesto che risultano simili a una qualche **testo** 'esempio' scelto dall'utilizzatore.

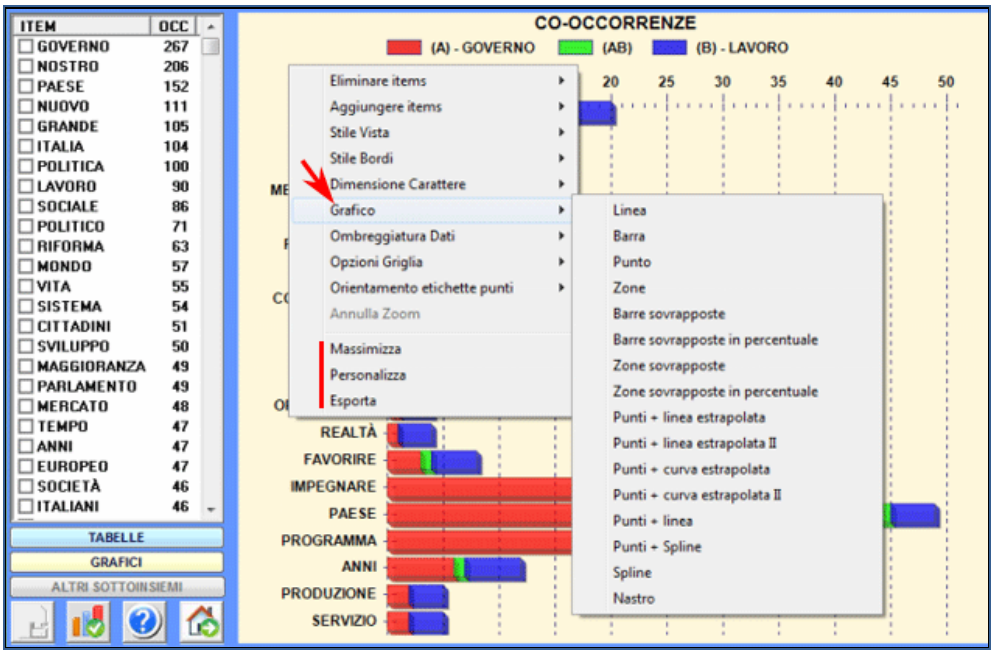


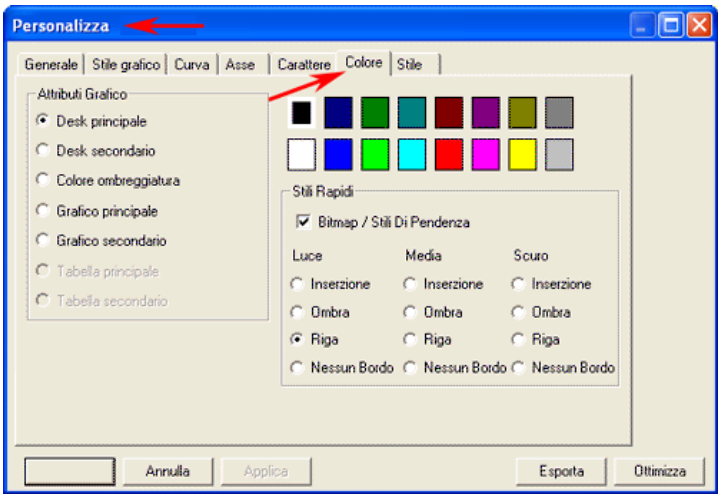


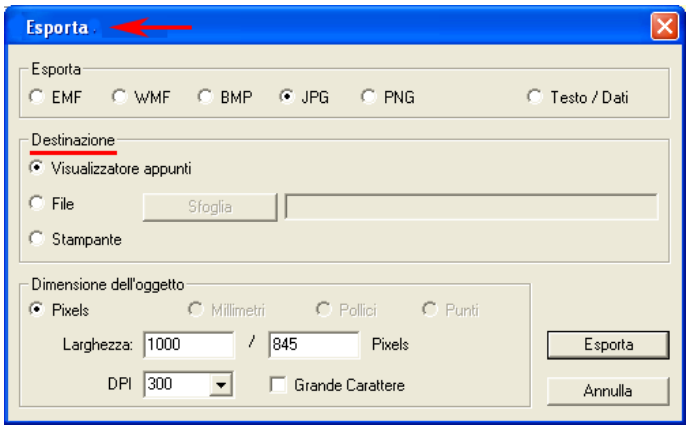
6 - L' INTERPRETAZIONE DEGLI OUTPUT consiste nella consultazione delle tabelle e dei grafici prodotti da **T-LAB**, nell'eventuale personalizzazione del loro formato e nel fare inferenze sul significato delle relazioni in essi rappresentate.

Nel caso delle **tabelle**, a seconda dei casi, **T-LAB** consente di esportarle in file con le seguenti estensioni: **.DAT**, **.TXT**, **.CSV**, **.XLXS**, **.HTML**. Ciò significa che, servendosi di qualunque editore di testi e/o di un qualche applicativo della suite Microsoft Office, l'utilizzatore può facilmente importarli e rielaborarli.

Nel caso dei **grafici**, appositi sub-menu attivati con il tasto destro del mouse consentono vari tipi di operazioni: zoom (clic con il tasto sinistro e selezionare un rettangolo), massimizzazione, personalizzazione ed esportazione degli output in diversi formati (vedi sotto, uso del tasto destro).







Alcuni criteri generali per l'interpretazione degli output **T-LAB** sono illustrati in un paper citato in Bibliografia e disponibile nel sito https://www.tlab.it (Lancia F.: 2007). In questo viene proposta l'ipotesi che gli output delle elaborazioni statistiche (tabelle e grafici) sono un tipo particolare di testi, cioè degli oggetti multi-semiotici caratterizzati dal fatto che le relazioni tra segni e simboli sono ordinate da misure che rinviano a specifici **codici**.

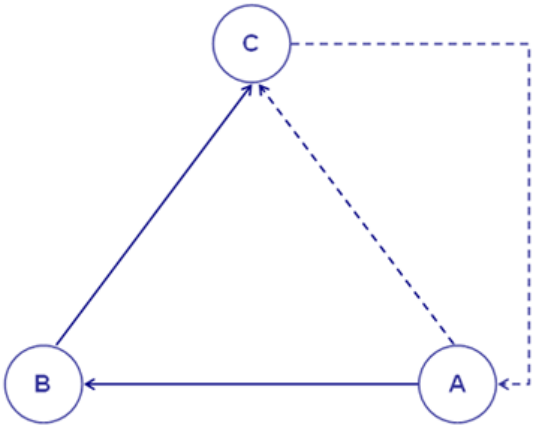
In altri termini, sia nel caso dei testi scritti in linguaggio naturale che in quelli scritti nel linguaggio della statistica, la possibilità di fare inferenze sulle relazioni che organizzano le **forme del contenuto** è fondata sul fatto che le relazioni tra le **forme dell'espressione** non sono casuali (random); infatti, nel primo caso (linguaggio naturale) le unità significanti si susseguono ordinate in modo lineare (una dopo l'altra nella catena del discorso), mentre nel secondo caso (tabelle e grafici) i principi di ordinamento sono costituiti dalle misure che determinano l'organizzazione degli **spazi semantici** multidimensionali.

Anche se gli spazi semantici rappresentati nelle mappe **T-LAB** sono molto vari, e ciascuno di essi richiede specifiche procedure interpretative, possiamo fare l'ipotesi che - in generale - la logica del processo inferenziale è la seguente:

A - rilevare una qualche relazione significativa tra le unità "presenti" sul piano dell'espressione (ad es. tra "dati" di tabelle e/o tra "label" di grafici);

B - esplorare e confrontare i tratti semantici delle stesse unità e i contesti a cui esse sono mentalmente e culturalmente associate (piano del contenuto);

C - costruire qualche ipotesi o qualche categoria di analisi che, nel contesto definito dal corpus, renda ragione delle relazioni tra forme dell'espressione e forme del contenuto.



Infine qualche informazione sui **vincoli attuali** delle opzioni **T-LAB**:

- dimensioni del corpus: max 90 Mb, pari a circa 55.000 pagine in formato .txt;
- documenti primari: max 30.000 (N.B.: quando nessuno dei testi supera i 2.000 caratteri, il limite è esteso a 99.999);
- variabili categoriali: max 50, ciascuna delle quali con max 150 modalità;
- modellizzazione dei temi emergenti: max 5.000 unità lessicali (*) per max 5.000.000 occorrenze;
- analisi tematica dei contesti elementari: max 300.000 righe (unità di contesto) x 5.000 colonne (unità lessicali);
- classificazione tematica dei documenti: max 99.999 righe (documenti) x 5.000 colonne (unità lessicali);
- analisi delle specificità (unità lessicali x categorie di una variabile): max 10.000 righe per 150 colonne;
- analisi delle corrispondenze (unità lessicali x categorie di una variabile): max 10.000 righe per 150 colonne;
- analisi delle corrispondenze (unità di contesto x unità lessicali): max 10.000 righe per 5.000 colonne;
- analisi delle corrispondenze multiple (contesti elementari x categorie di due più variabili): max 150.000 righe per 250 colonne;
- scomposizione in valori singolari (SVD): max 300.000 righe x 5.000 colonne;
- cluster analysis che utilizza i risultati di una precedente analisi delle corrispondenze (o SVD): max 10.000 righe (unità lessicali o contesti elementari);
- associazioni di parole, confronti tra coppie e co-word analysis: liste di max 5.000 unità lessicali;
- analisi delle sequenze: max 5.000 unità lessicali (o categorie) con testi di max 3.000.000 occorrenze.

(*) In **T-LAB**, 'unità lessicali' sono parole, multi-words, lemmi e categorie semantiche. Quindi, quando viene applicata la lemmatizzazione automatica, 5.000 unità lessicali corrispondono a circa 12.000 parole.