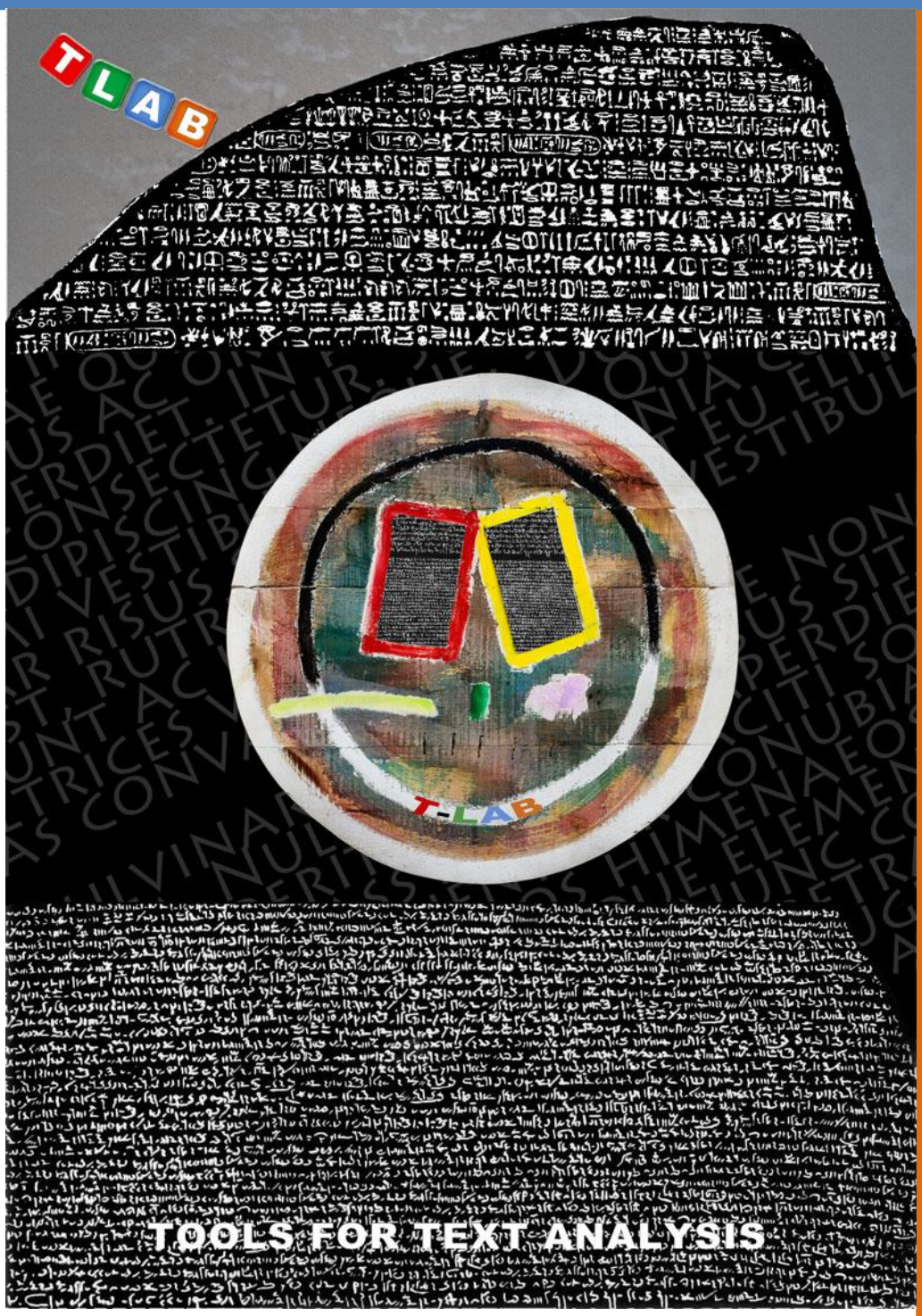
T-LAB 10

Manuel de l'Utilisateur



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

Outils pour l'Analyse de Textes

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

SOMMAIRE

Installation et configuration requise

Configuration minimum requise:

- Windows 7 ou postérieur
- minimum de 4 Gb RAM
- Résolution d'écran Full HD (1920 x 1080 recommandée)

Installation:

- Cliquez sur Setup.exe
- Suivez les instructions à l'écran
- Quittez **T-LAB**
- Attendez la réponse avec votre clef d'activation
- Pour plus d'informations voir https://www.mytlab.com/T-LAB_Installation.pdf

Ce qu'il fait et ce qu'il vous permet de faire

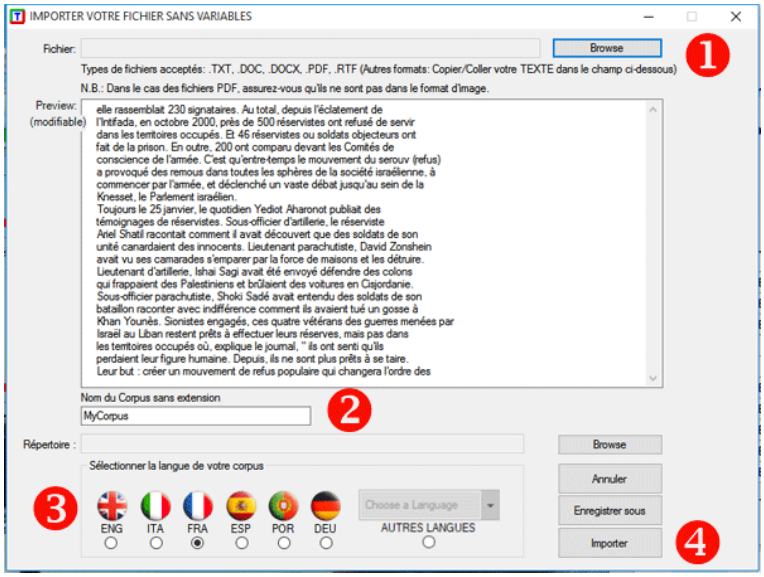
T-LAB est un logiciel composé par un ensemble d'**outils linguistiques, statistiques et graphiques pour l'analyse des textes** qui peuvent être utilisés dans les pratiques de recherche suivantes: Analyse du Contenu, Sentiment Analysis, Analyse Sémantique, Analyse Thématique, Text Mining, Perceptual Mapping, Analyse du Discours, Network Text Analysis.

En fait, au moyen des outils **T-LAB** les chercheurs peuvent facilement gérer les activités d'analyse suivantes:

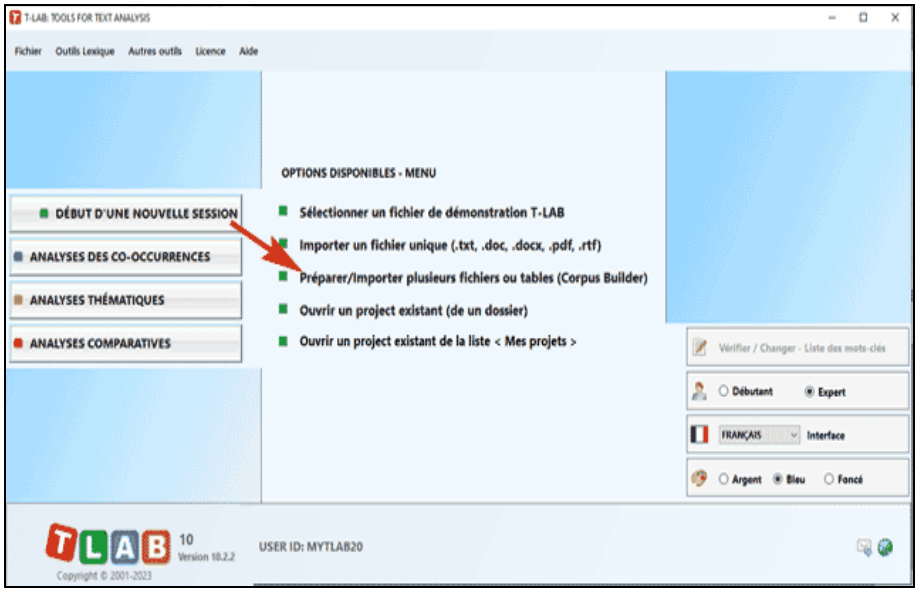
- explorer, mesurer et topographier les **relations de co-occurrence** entre mots-clés;
- réaliser une **classification automatique** d'unités de contexte et de documents, soit à travers une **approche bottom-up** (c'est-à-dire à travers l'analyse des thèmes émergents) soit à travers une **approche top-down** (c'est-à-dire à travers l'utilisation de catégories prédéfinies);
- vérifier quelles **unités lexicales** (c'est-à-dire mots ou lemmes), quelles **unités de contexte** (c'est-à-dire phrases ou paragraphes) et quels **thèmes** sont «typiques» de sous-ensembles de textes spécifiques (par exemple, les discours de certains leaders politiques, les interviews avec certaines catégories de personnes, etc.);
- appliquer des catégories pour la **sentiment analysis**;
- effectuer différents types d'**analyse des correspondances** et de **clusters analysis**;
- créer des **cartes sémantiques** qui représentent des **aspects dynamiques du discours** (c'est-à-dire des relations séquentielles entre les mots ou les thèmes);
- représenter et explorer un texte quelconque comme un **réseau** de relations;
- obtenir des mesures et des représentations graphiques concernant les **textes et discours** traités **comme des systèmes dynamiques**;
- personnaliser et appliquer **différents types de dictionnaires**, aussi bien pour l'analyse lexicale que pour l'analyse du contenu;
- vérifier les contextes d'occurrence (par ex., **concordances**) de mots et de lemmes;
- analyser tout le **corpus** ou seulement certains de ses **sous-ensembles** (par ex. des groupes de documents) en utilisant différentes listes de mots-clés
- créer, explorer et exporter différents types de **tableaux de contingence** et de **matrices de co-occurrences**.

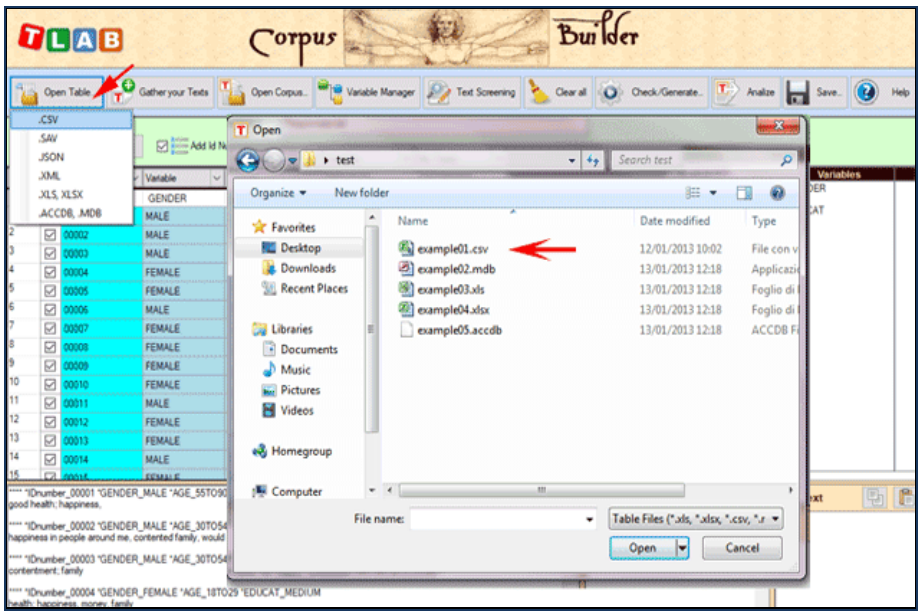
L'interface utilisateur est **très conviviale** et les textes à analyser peuvent être des plus variés:
- un seul texte (ex. une interview, un livre, etc.);
- un ensemble de textes (ex. diverses interviews, pages web, articles de journal, réponses à des questions ouvertes, messages Twitter, etc.).

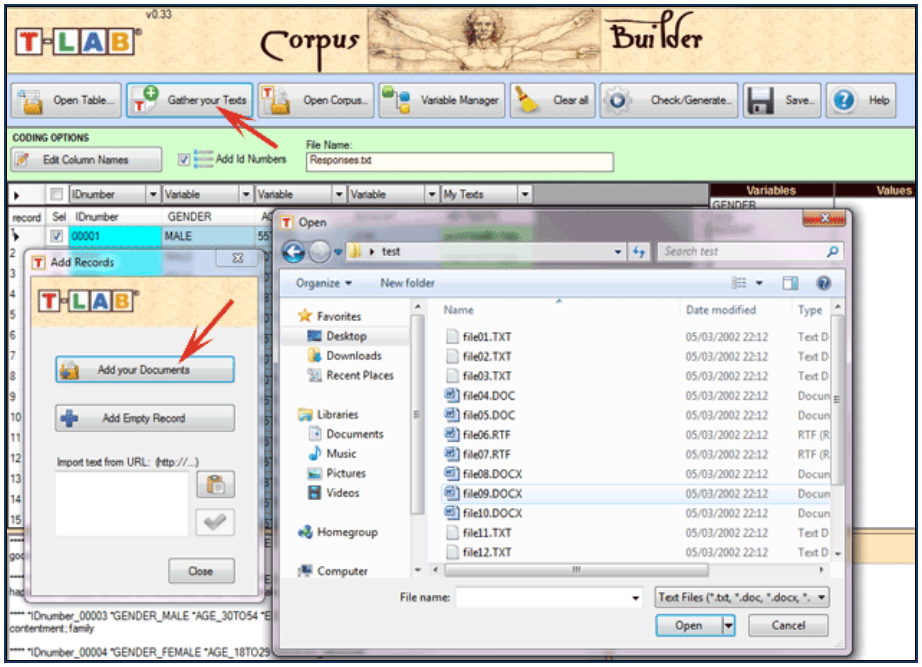
Tous les textes peuvent être codifiés avec des **variables** catégorielles et peuvent inclure un identificateur (**Unique Identifier**) qui correspond à des unités de contexte ou à des cas (ex. réponses à des questions ouvertes).
Dans le cas d'un seul document (ou un corpus considéré comme un texte unique) **T-LAB** nécessite pas de travail supplémentaire: il vous suffit de sélectionner l'option 'Importer un fichier unique (voir ci-dessous).



Différemment, dans les autres cas il faut utiliser le module **Corpus Builder** (voir ci-dessous) qui transforme automatiquement des documents textuels et différents types de fichiers (c'est-à-dire jusqu'à dix différents formats) dans un corpus prêt à être importé par **T-LAB**.



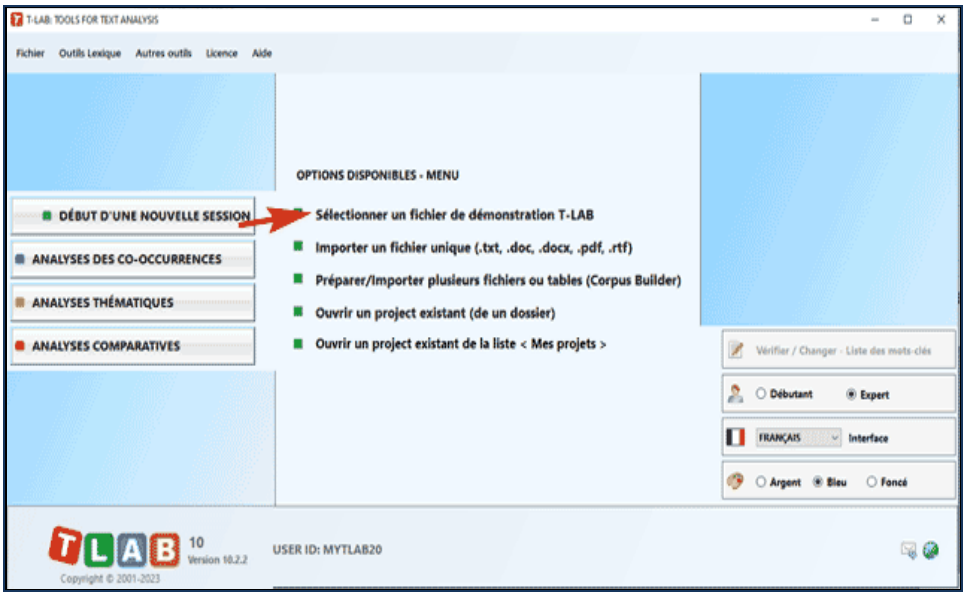




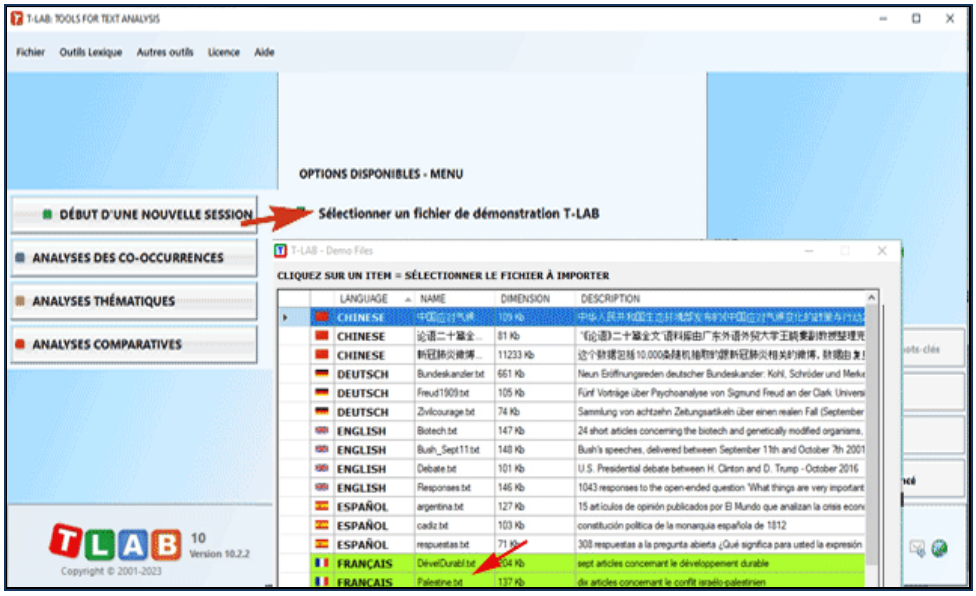
N.B. : En ce moment, afin d'assurer l'utilisation intégrée des différents outils, chaque fichier/corpus à analyser ne devrait pas dépasser 90 Mo (c'est-à-dire environ 55.000 pages au format .txt). Pour plus d'informations, voir la section 'Conditions et performances' du Manuel / Help.

Six étapes suffisent pour explorer rapidement les fonctions du logiciel:

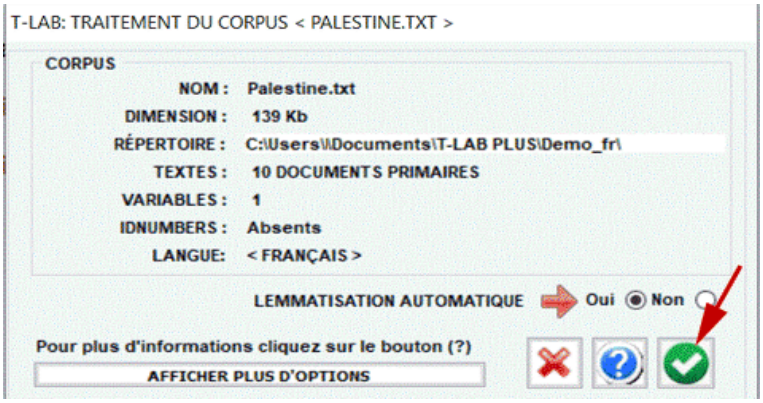
1 - Cliquer l'option 'Sélectionner un fichier de démonstration..'



2 - Sélectionner un corpus à analyser

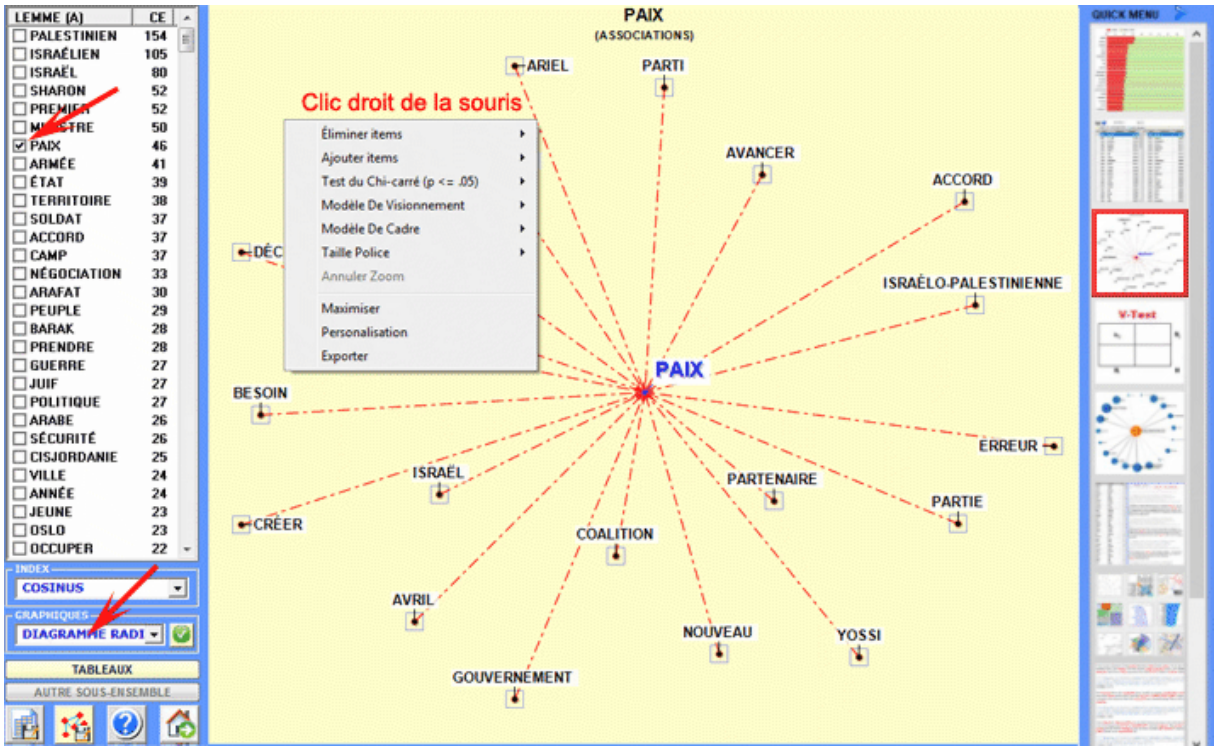


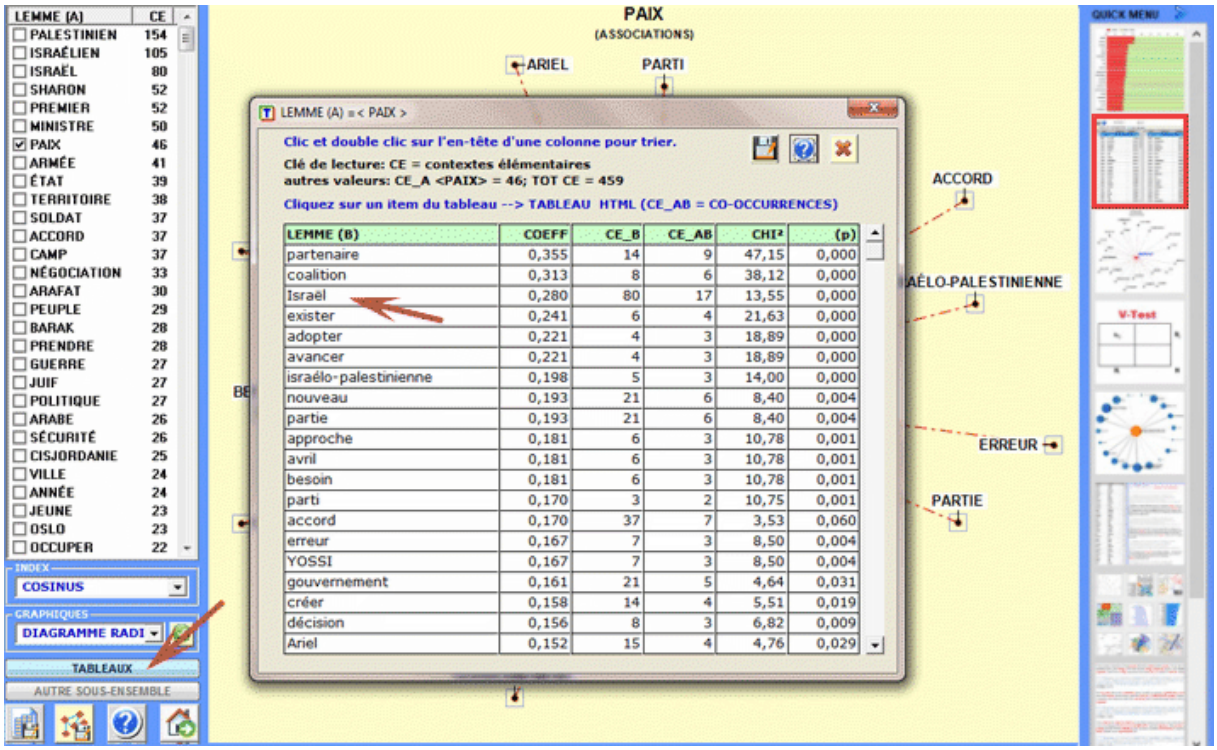
3 - Cliquer sur "ok" dans la première fenêtre de configuration



4 - Choisir un outil à l'intérieur d'un des sous-menus "Analyse"

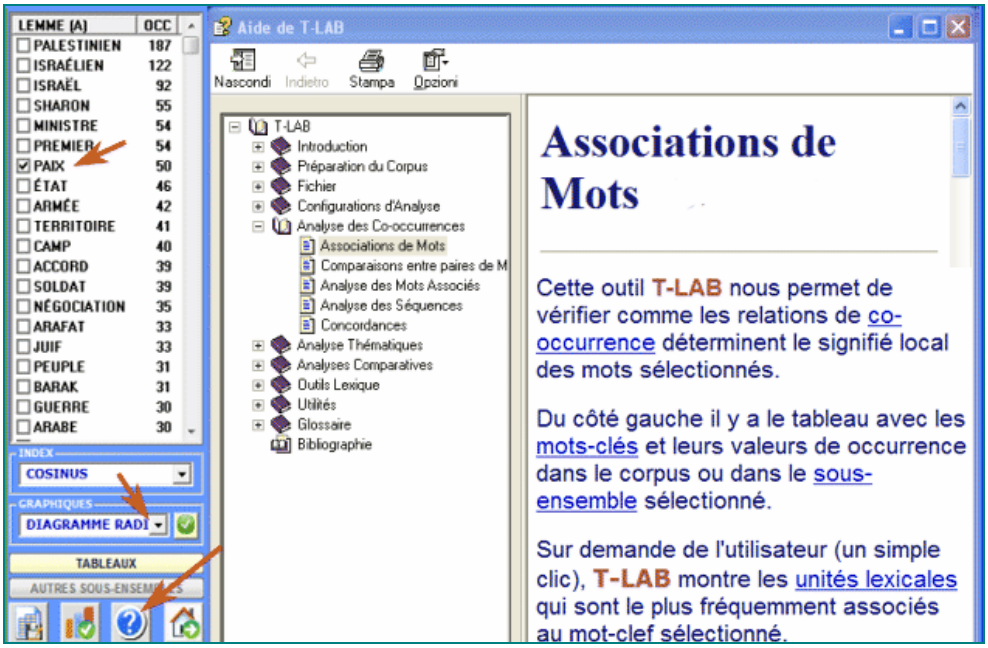
5 - Examiner les résultats





LEMME (B)	COEFF	CE_B	CE_AB	CHI²	(p)
partenaire	0,355	14	9	47,15	0,000
coalition	0,313	8	6	38,12	0,000
Israël	0,280	80	17	13,55	0,000
exister	0,241	6	4	21,63	0,000
adopter	0,221	4	3	18,89	0,000
avancer	0,221	4	3	18,89	0,000
israélo-palestinienne	0,198	5	3	14,00	0,000
nouveau	0,193	21	6	8,40	0,004
partie	0,193	21	6	8,40	0,004
approche	0,181	6	3	10,78	0,001
avril	0,181	6	3	10,78	0,001
besoin	0,181	6	3	10,78	0,001
parti	0,170	3	2	10,75	0,001
accord	0,170	37	7	3,53	0,060
erreur	0,167	7	3	8,50	0,004
YOSSI	0,167	7	3	8,50	0,004
gouvernement	0,161	21	5	4,64	0,031
créer	0,158	14	4	5,51	0,019
décision	0,156	8	3	6,82	0,009
Ariel	0,152	15	4	4,76	0,029

6 - Utiliser l'aide contextuelle pour interpréter les graphiques et les tableaux.



Cette section introductive fournit les informations essentielles afin de mieux comprendre ce que **T-LAB** fait et comment il peut être utilisé.

Du point de vue externe, l'utilisation du logiciel est organisée par l'**interface**, c'est-à-dire par le **menu principal**, par les **sous-menus** et les **fonctions** qui les composent.

D'un point de vue logique, en plus de l'interface usager, le système **T-LAB** est organisé par deux composantes principales:

- le **database**, c'est-à-dire le lieu informatique dans lequel le **corpus** en input (soit le texte ou l'ensemble des textes à analyser) est représenté comme un ensemble de **tableaux** dans lesquels sont enregistrées les **unités d'analyse**, leurs caractéristiques et leur relations réciproques.
- les **algorithmes**, c'est-à-dire des sous-ensembles d'**instructions** qui permettent d'utiliser l'interface usager, de consulter et modifier le database, de construire d'ultérieurs tableaux avec les données contenues dans ce dernier, d'effectuer des **calculs statistiques** et de produire des **outputs** qui représentent les relations entre les données analysées.

Pour comprendre comment **T-LAB** fonctionne et comment il peut être utilisé, il est fondamental de savoir clairement quelles unités d'analyse sont archivées dans son database et quels algorithmes statistiques sont utilisés dans les diverses analyses. En effet, les tableaux de données analysées sont toujours constitués de lignes et de colonnes dont les titres correspondent aux unités d'analyse archivées dans le database, alors que les algorithmes règlent les processus qui permettent de repérer des relations significatives entre les données et d'extraire des informations utiles.

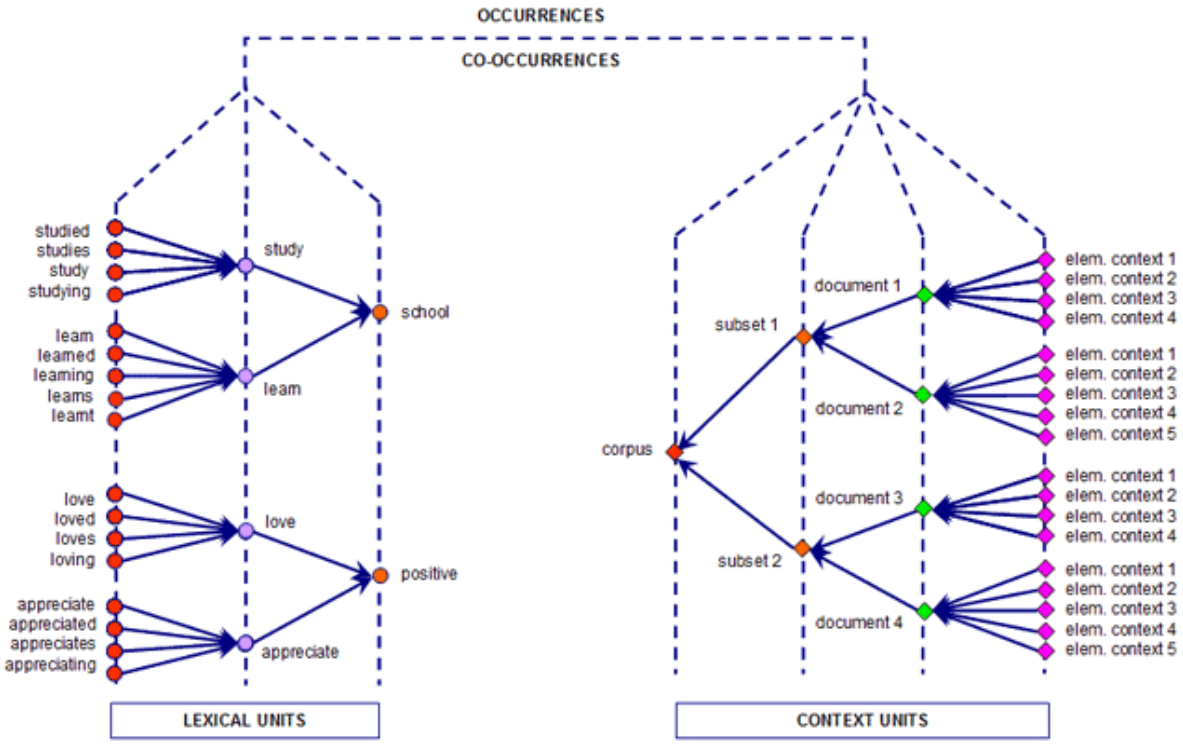
Les **unités d'analyse** de T-**LAB** sont de deux types: **unités lexicales** et **unités de contexte**.

A - les **UNITÉS LEXICALES** sont des mots, simples ou multiples, archivés et classifiés sur la base d'un critère. Plus précisément, dans le database **T-LAB** chaque unité lexicale constitue un record classifié avec deux champs: **mot et lemme**. Dans le premier champ, appelé **mot**, sont listés les mots ainsi qu'ils apparaissent dans le corpus, alors que dans le second, appelé **lemme**, sont listés les labels attribués à des groupes d'unités lexicales classifiées selon des critères linguistiques (ex. **lemmatisation**) ou au moyen de dictionnaires et de grilles sémantiques définies par l'utilisateur.

B - les **UNITÉS DE CONTEXTE** sont des portions de texte dans lesquelles le corpus peut être subdivisé. Plus exactement, dans la logique **T-LAB**, les unités de contexte peuvent être de trois types:

B.1 documents primaires, correspondants à la subdivision "naturelle" du corpus (ex. interviews, articles, réponses à des questions ouvertes, etc.), ou bien aux contextes initiaux définis par l'utilisateur;
B.2 contextes élémentaires, correspondants à des unités syntagmatiques (ex. fragments de texte, phrases, paragraphes) dans lesquelles chaque document primaire peut être subdivisé;
B.3 sous-ensembles du corpus, correspondants à des groupes de documents primaires reconductibles à la même catégorie (ex. interviews d' "hommes" ou de "femmes", articles d'une année particulière ou d'un titre particulier, et ainsi de suite), ou à clusters thématiques obtenus avec des instruments spécifiques de **T-LAB**.

Le diagramme suivant illustre les relations possibles entre les unités lexicales et les unités de contexte que **T-LAB** nous permet d'analyser.

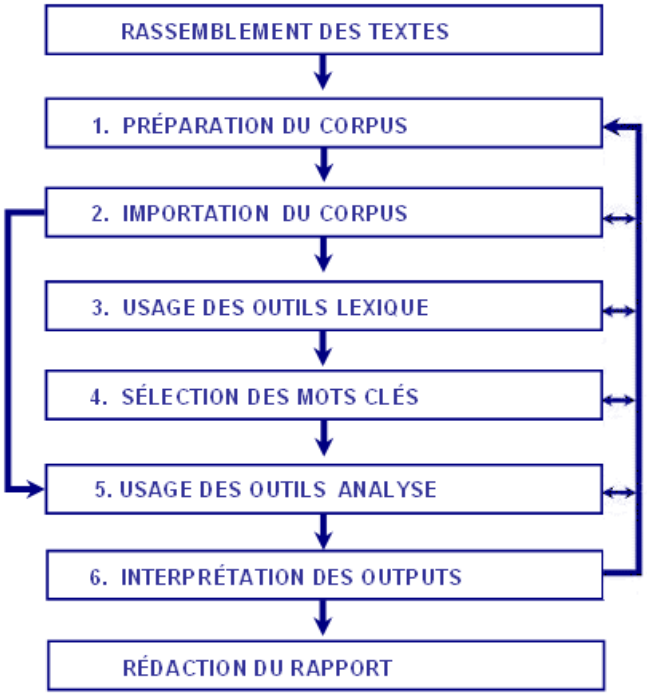


À partir de cette organisation du database, **T-LAB** permet - de façon automatique - d'explorer et d'analyser les relations entre les unités d'analyse de **tout le corpus** ou de ses **sous-ensembles**.

Dans **T-LAB**, la sélection d'un quelconque instrument d'analyse (clic de la souris) active toujours un processus semi-automatique qui, grâce à quelques simples opérations, génère un tableau input, applique un algorithme de type statistique et produit quelques outputs.

Un **projet** de travail "typique" dans lequel est utilisé **T-LAB** est constitué de l'ensemble des activités analytiques (opérations) qui ont pour objet le même **corpus** et est organisé par une **stratégie** et par un **plan** de l'utilisateur. Ainsi, il commence par le **rassemblement des textes** à analyser et s'achève par la **rédaction d'un rapport**.

La succession des diverses phases est illustrée dans le diagramme suivant:



N.B.:
- Les six phases énumérées, de la préparation du corpus à l'interprétation des outputs, sont supportées par des instruments **T-LAB** et sont toujours réversibles;
- Grâce aux **configurations automatiques T-LAB** il est possible d'éviter deux phases (3 et 4); toutefois, aux fins de la **qualité** des résultats, leur réalisation est fortement recommandée.

1 - La **PRÉPARATION DU CORPUS** consiste en la transformation des textes à analyser dans un fichier (**corpus**) qui peut être élaboré par le logiciel.

Dans le cas de textes uniques (ou corpus considéré comme texte unique) on n'a pas besoin d' autre travail. Autrement, si le corpus se compose de plusieurs documents primaires codifiés (**variables et modalités**), dans la phase de préparation on doit utiliser l'outil **Corpus Builder**, qui transforme automatiquement tout matériel textuel et divers types de fichiers (c.-à-d. jusqu'à dix formats différents) dans un fichier corpus prêt à être importé par **T-LAB.**

N.B.:

- au terme de la phase de préparation du corpus on recommande de créer un nouveau dossier de travail avec à l'intérieur le fichier corpus à importer ;

- durant les analyses il est recommandé de garder le corpus et le dossier de travail relatif sur un disque dur de l'ordinateur où **T-LAB** est installé. Dans le cas contraire, l'exécution des diverses procédures pourrait être ralentie et le logiciel pourrait signaler des erreurs.

2 - **L'IMPORTATION DU CORPUS** consiste en une série de **processus automatiques** qui transforment le corpus en un ensemble de tableaux intégrés dans le **database T-LAB**.

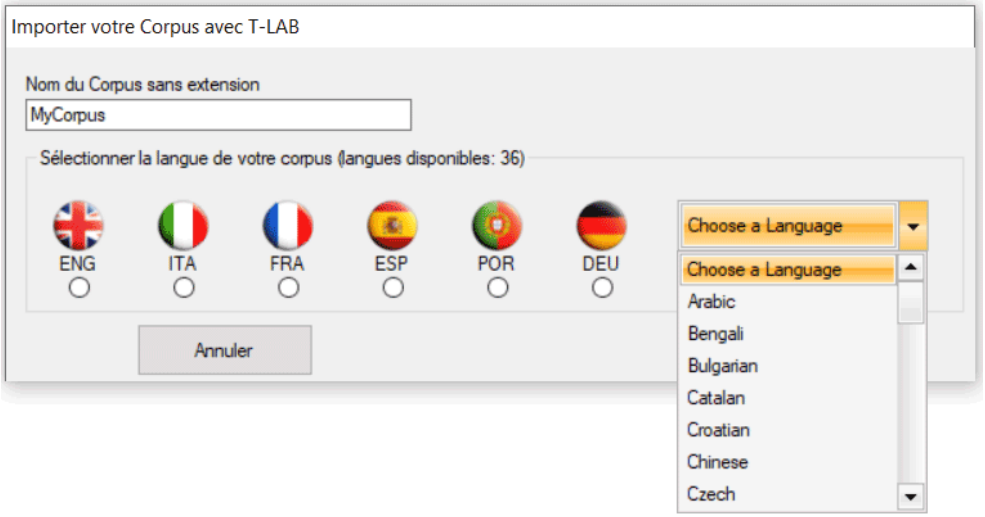
Pendant la phase d'importation du corpus, **T-LAB** effectue les traitements suivants: **normalisation** du corpus; détection des **multi-words** et des **stop-words**; segmentation des **contextes élémentaires**; **lemmatisation** automatique ou **stemming**; construction du **vocabulaire**; sélection des **mots-clés**.

De suite la liste complète des trente langues pour lesquelles la lemmatisation automatique ou bien le processus de stemming sont supportés par **T-LAB**.

LEMMATISATION: allemand, anglais, catalan, croate, espagnol, français, italien, latín, polonais, portugais, roumain, russe, serbe, slovaque, suédois, ukrainien.

STEMMING: arabe, bengali, bulgare, danois, hollandais, finlandais, grec, hindi, hongrois, indonésien, marathi, norvégien, persan, tchèque, turc.

En tout les cas, sans lemmatisation automatique et / ou en utilisant des dictionnaires personnalisés, l'utilisateur peut analyser textes dans **toutes les langues**, à condition que les mots soient séparés par des espaces et/ou des signes de ponctuation.

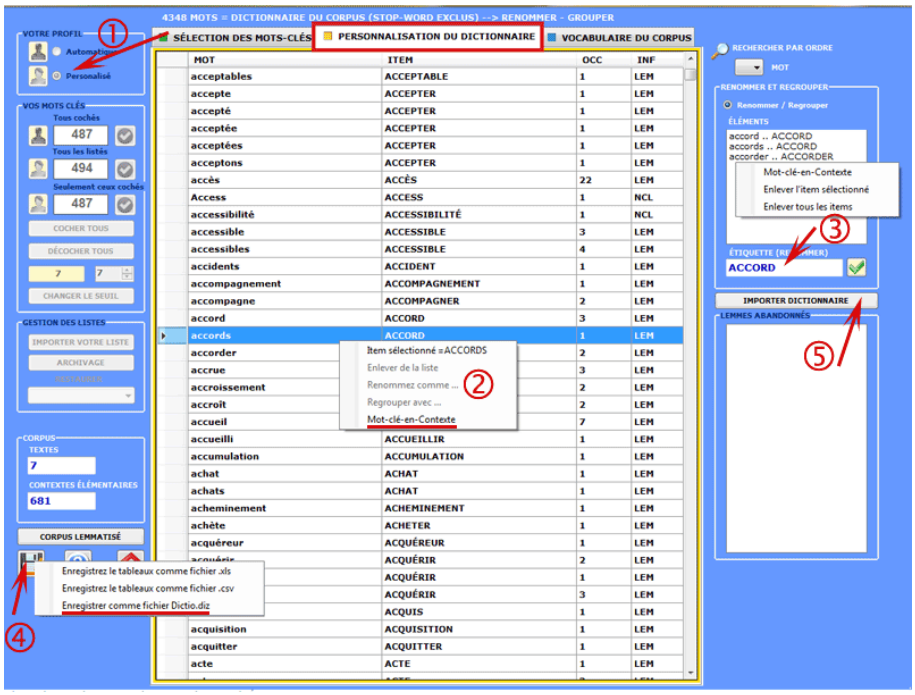


À partir de la sélection de la langue, l'intervention de l'utilisateur (options avancées) est requise afin de définir les choix indiqués dans la fenêtre suivante.

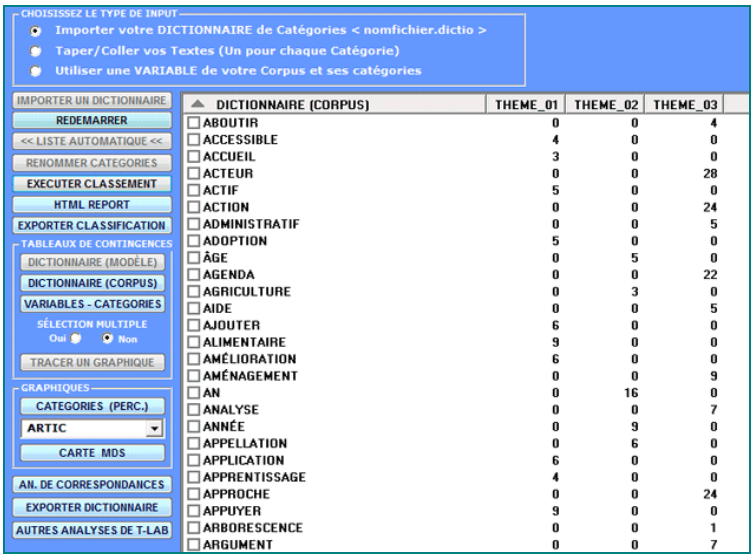
N.B. : Puisque les options de prétraitement déterminent le type et la quantité d'unités d'analyse (c.-à-d. des unités de contexte et des unités lexicales), les différents choix de l'utilisateur déterminent différents résultats de l'analyse. Pour cette raison, tous les outputs de **T-LAB** (c.-à-d. graphiques et tableaux) montrés dans le manuel et dans l'aide en ligne sont simplement indicatifs.

3 - L'UTILISATION DES OUTILS LEXIQUE est finalisée à la vérification de la correcte **reconnaissance** des unités lexicales et à personnaliser leur **classification**, c'est-à-dire à vérifier et à modifier les choix automatiques faits par **T-LAB**.

Les modalités des diverses interventions sont illustrées dans les rubriques de l'aide (et du manuel) correspondantes. En particulier on renvoie à la rubrique de l'aide (et du manuel) correspondante pour une description détaillée du processus **Personnalisation du Dictionnaire**. En effet, n'importe quel changement relatif aux voix du dictionnaire (par ex., le regroupement de deux ou plusieurs items) influe aussi bien sur le calcul des occurrences que sur celui des co-occurrences.

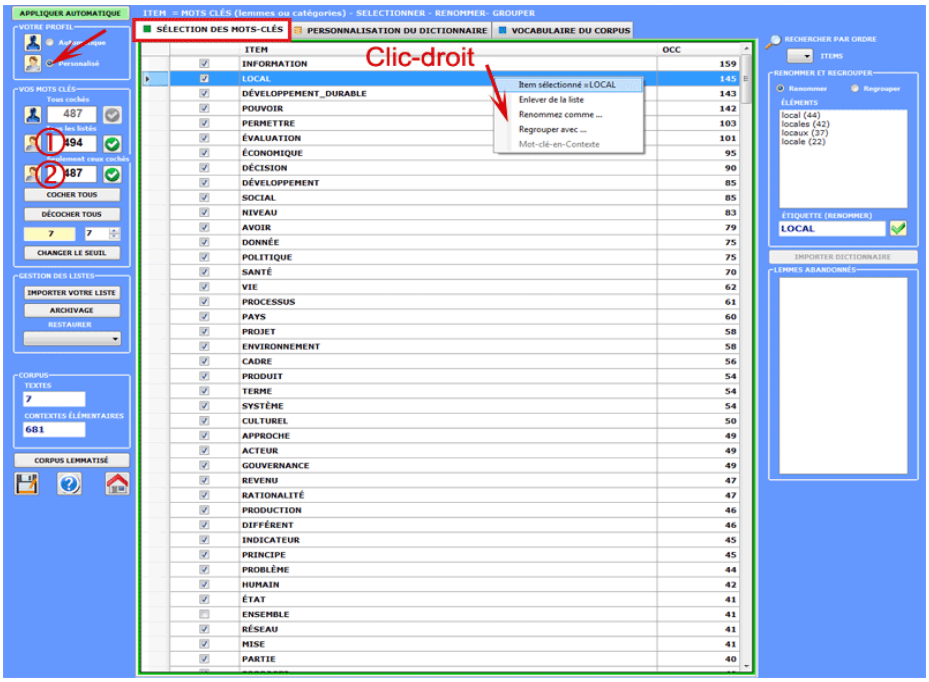


NB: Lorsque l'utilisateur, sans perdre aucune information lexicale, a l'intention d'appliquer des schémas de codage qui regroupent plusieurs mots ou lemmes dans peu de catégories (de 2 à 50), il est conseillé d'utiliser l'outil **Classification Basée sur des Dictionnaires** inclus dans le sous-menu **Analyse Thématique** (voir ci-dessous).



4 - LA SÉLECTION DES MOTS-CLÉS consiste en la prédisposition d'un ou de plusieurs listes d'unités lexicales (mots, lemmes ou catégories) à utiliser pour construire les tableaux données à analyser.

L'option **configurations automatiques** rend disponible des listes de **mots-clés** sélectionnés par **T-LAB**; toutefois, puisque le choix des unités d'analyse est extrêmement important aux fins des élaborations successives, on conseille vivement l'utilisation des **configurations personnalisées**.
De cette façon l'utilisateur pourra choisir de modifier la liste suggérée par **T-LAB** et/ou de construire des listes qui correspondent mieux à ses objectifs de recherche.



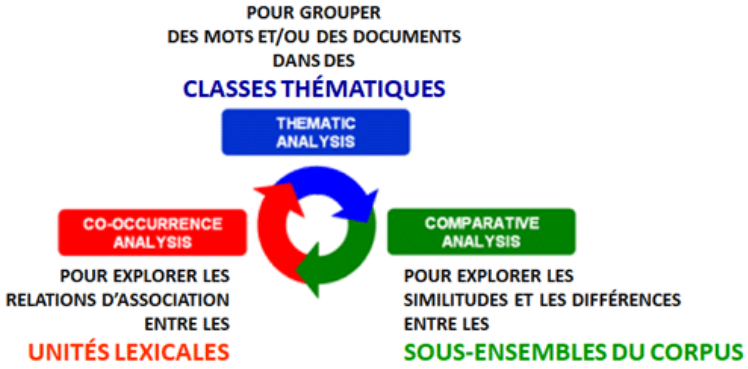
De toute façon, dans la construction de ces listes, valent les critères suivants:

- vérifier l'**importance** quantitative (total des occurrences) et qualitative (non banalité du sens) des divers items;
- vérifier les **limitations** (voir note à la fin de cette section) des instruments analytiques que l'on entend utiliser;
- vérifier si l'ensemble des items est compatible avec la propre **stratégie** de recherche (voir point suivant: 5).

5 - L'UTILISATION DES OUTILS D'ANALYSE est finalisée à la production d'outputs (tableaux et graphiques) qui représentent des **relations significatives** entre les unités d'analyse et qui permettent de faire des **inférences**.

Au moment actuel **T-LAB** inclut vingt différents outils d'analyse et chacun d'eux a sa propre logique; c'est-à-dire, chacun d'eux emploie des algorithmes spécifiques et produit des outputs spécifiques.

Par conséquent, selon le type de textes qu'il a l'intention d'analyser et des objectifs qu'il veut poursuivre, l'utilisateur doit décider de fois en fois quels sont les outils les plus appropriés pour sa stratégie d'analyse.

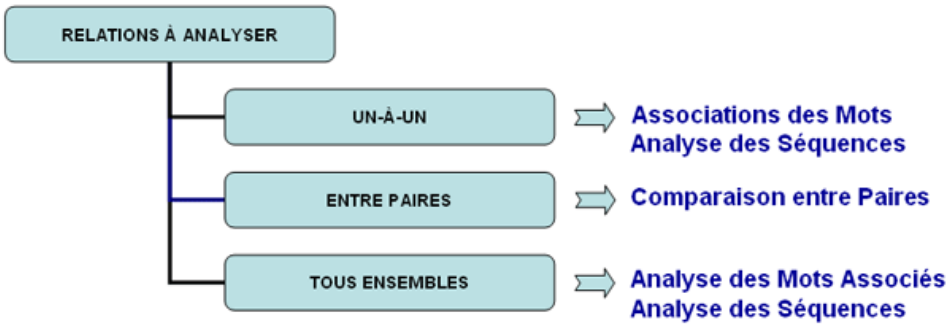


À cette fin, outre la distinction entre outils pour l'**analyse des cooccurrences**, pour l'**analyse comparative** et pour l'**analyse thématique**, il est utile de considérer que certains de ces derniers instruments permettent d'obtenir d'ultérieurs **sous-ensembles** fondés sur la similarité des contenus qui peuvent être inclus dans d'autres étapes de l'analyse.

Toutefois, compte tenu du fait que l'utilisation des outils **T-LAB** peut être circulaire et réversible, nous pouvons identifier trois points de démarrage (start points) qui correspondent aux trois sous-menus ANALYSE.

A : OUTILS POUR LES ANALYSES DE CO-OCCURRENCES

Ces outils nous permettent d'analyser différentes typologies de relations entre les mots.

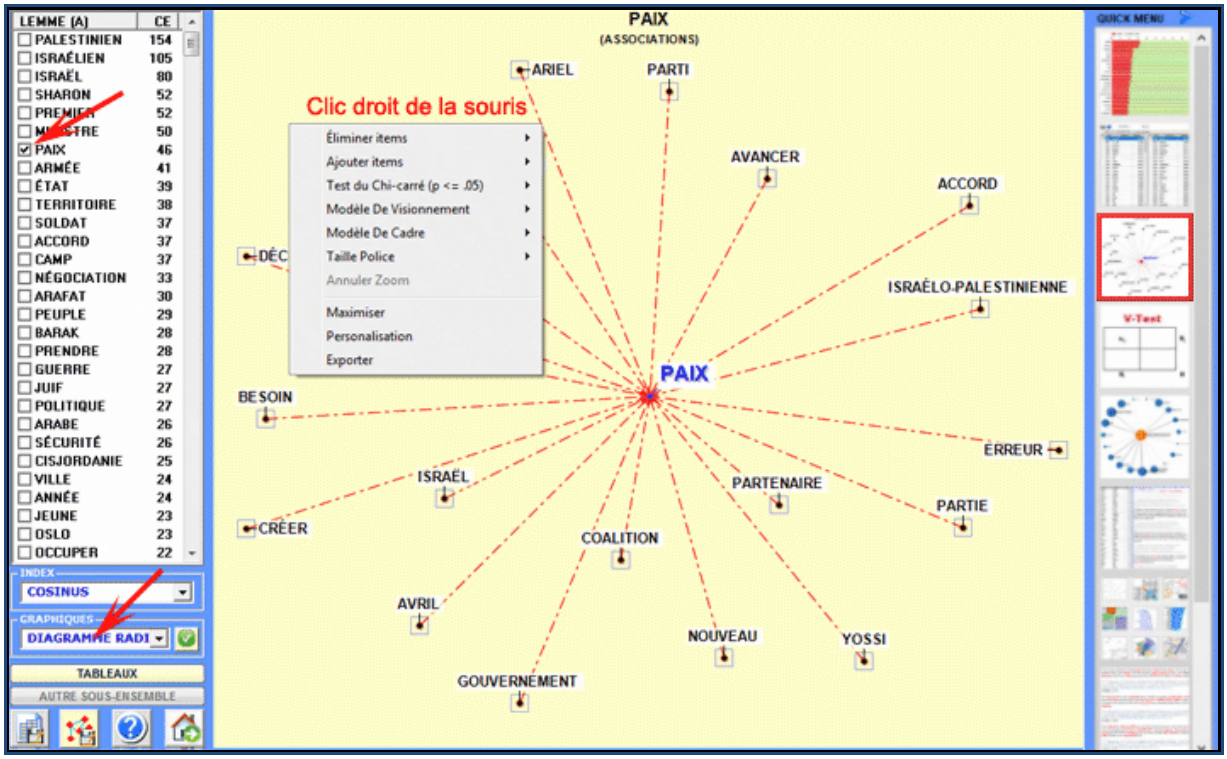


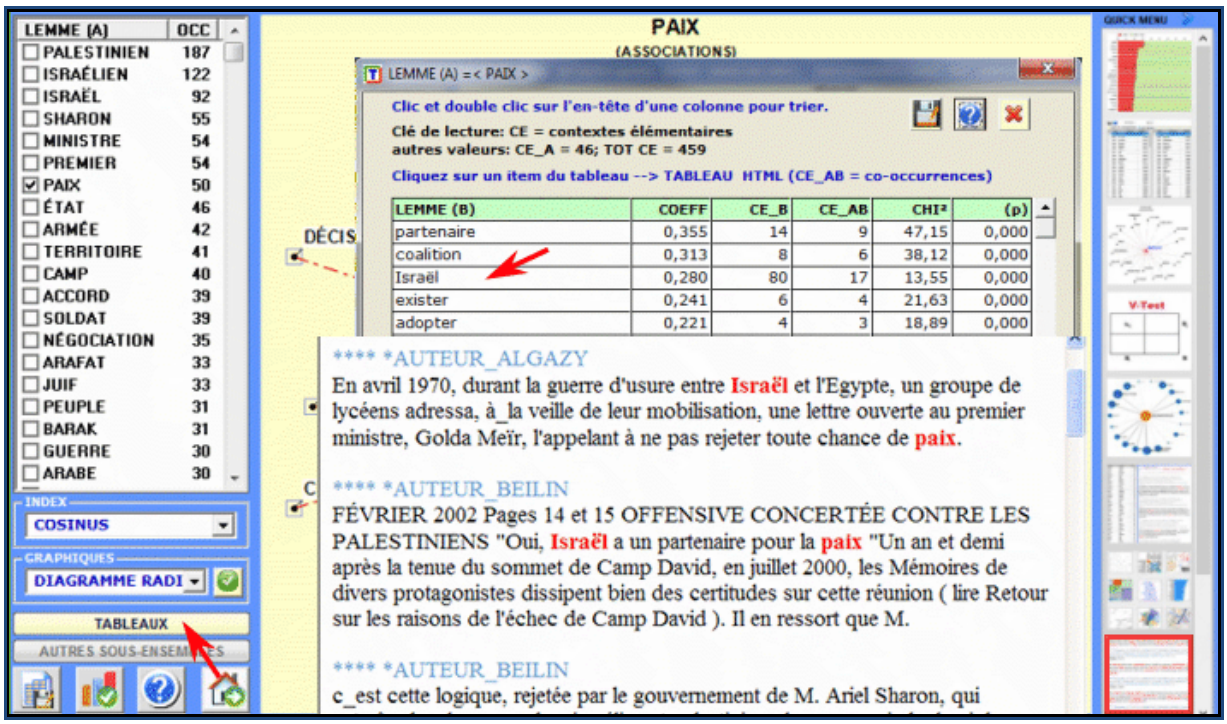
Selon les types de relations à analyser, les fonctions **T-LAB** indiquées dans ce diagramme (box colorés) utilisent un ou plusieurs des instruments statistiques suivants: **Indices d'Association**, **Test du Chi-Deux**, **Cluster Analysis**, **Multidimensional Scaling, Principal Component Analysis, t-SNE** et **Chaînes Markoviennes.**

Voici quelques exemples (N.B. : pour plus d'informations sur l'interprétation des outputs, veuillez vous référer aux sections correspondantes du guide / manuel):

- Associations des Mots

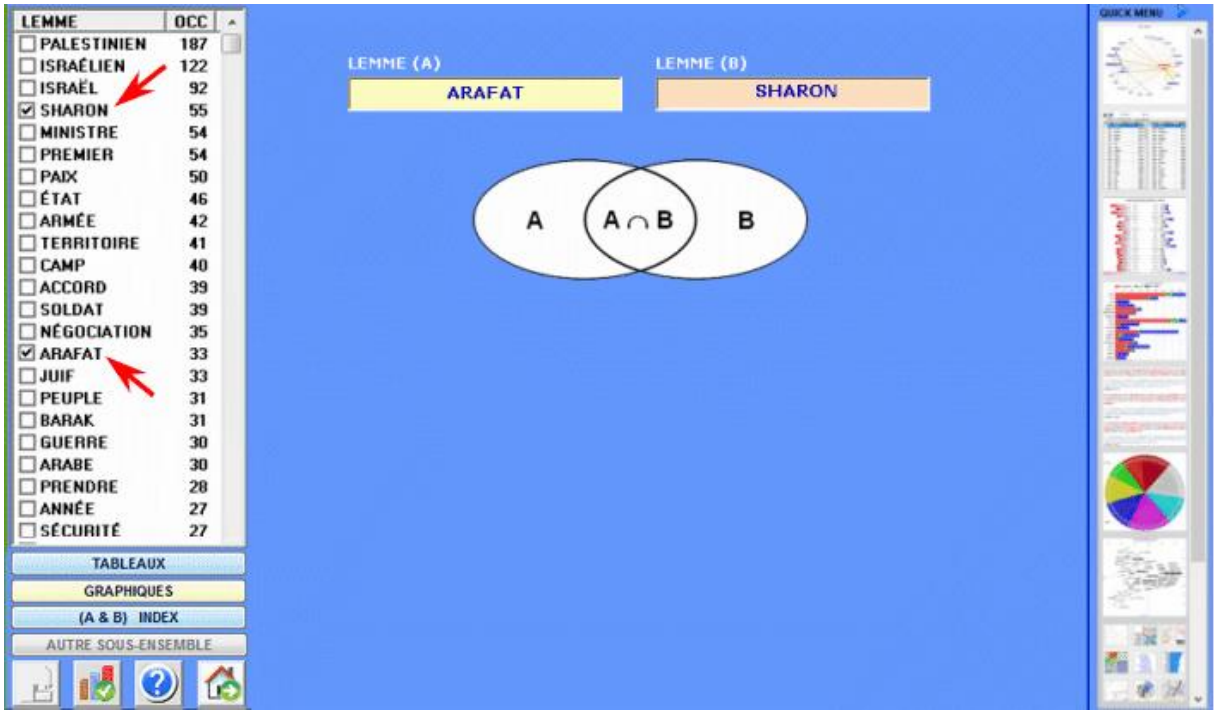
Cet outil **T-LAB** nous permet de vérifier comment les relations de **co-occurrence** déterminent le signifié local des mots sélectionnés.

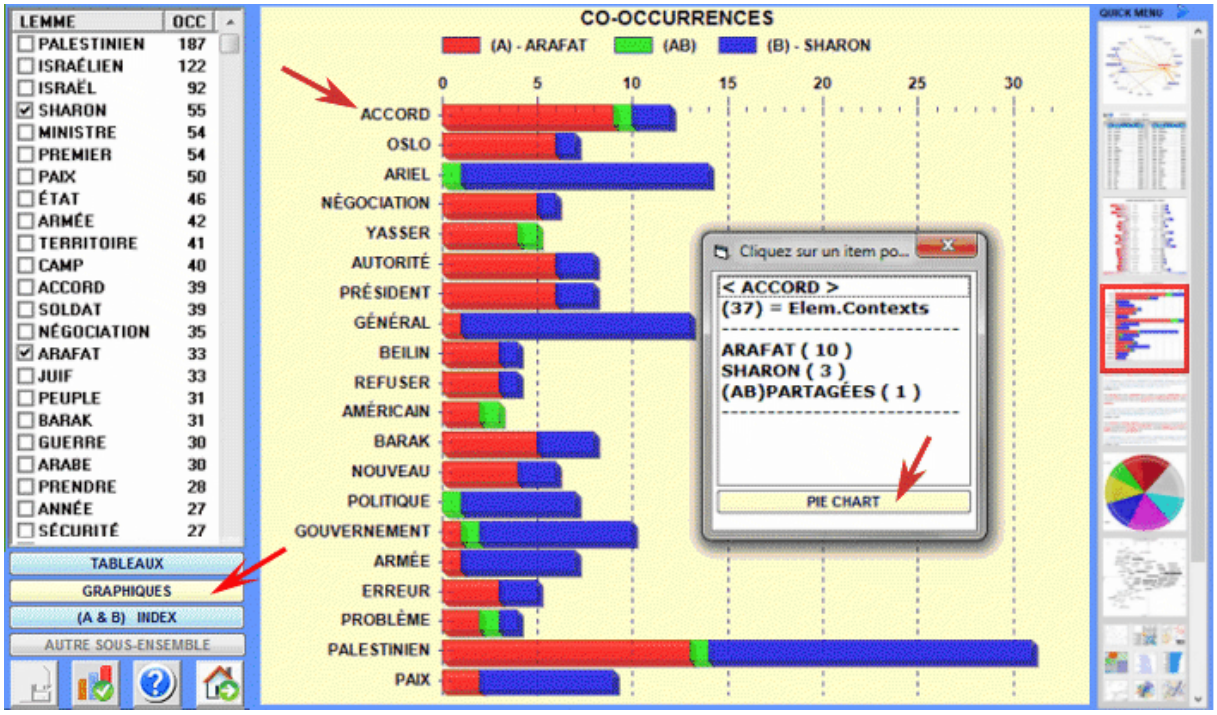




- Comparaison entre Paires

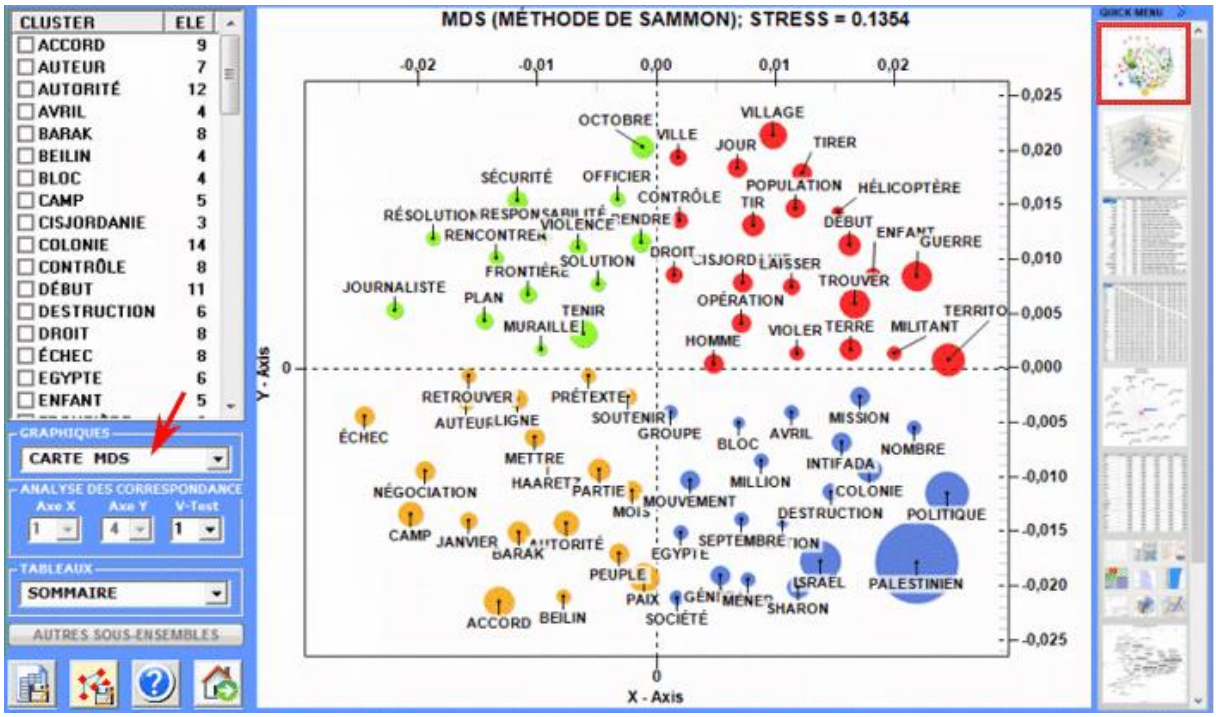
Cet outil **T-LAB** nous permet de comparer des ensembles de **contextes élémentaires** (c.-à-d. contextes de co-occurrence) dans lesquels sont présents les éléments d'une paire de **mots-clés**.

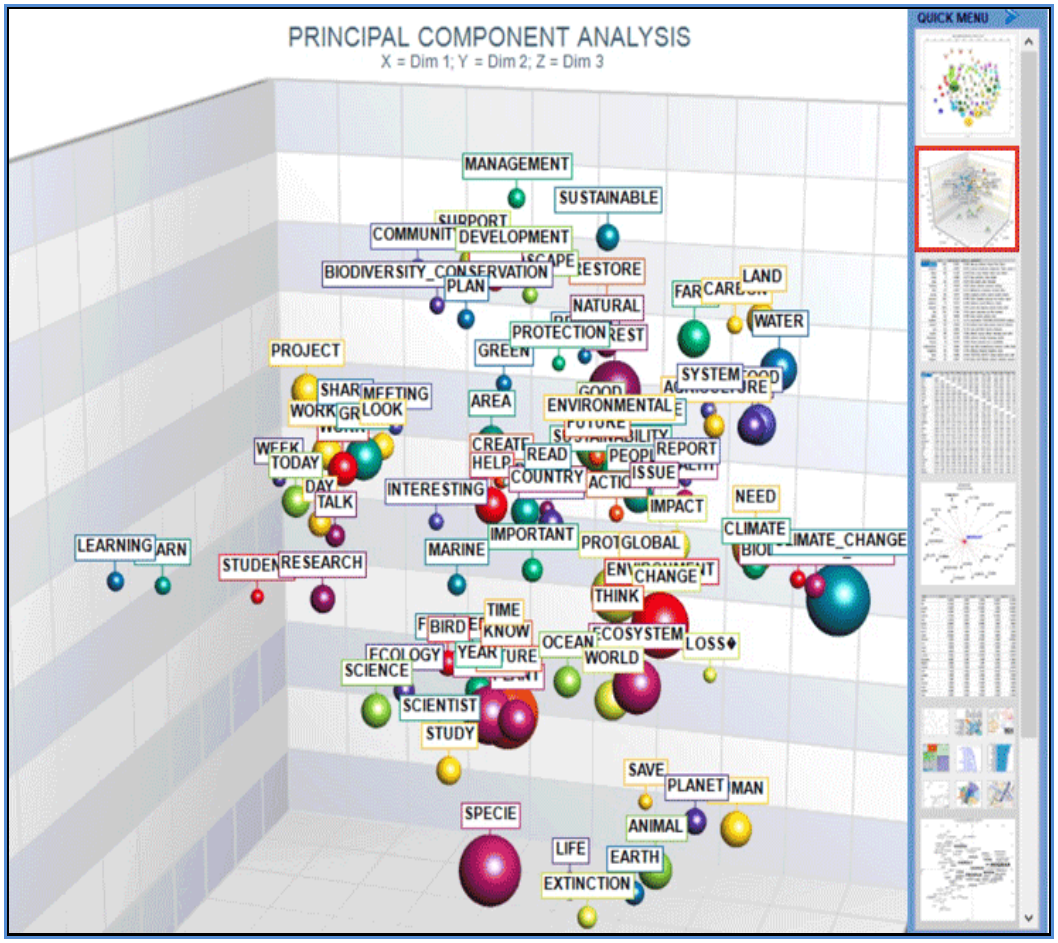




- Analyse des Mots Associés

Cet outil **T-LAB** nous permet de cartographier les relations de co-occurrence entre les **ensembles de mots-clés**.



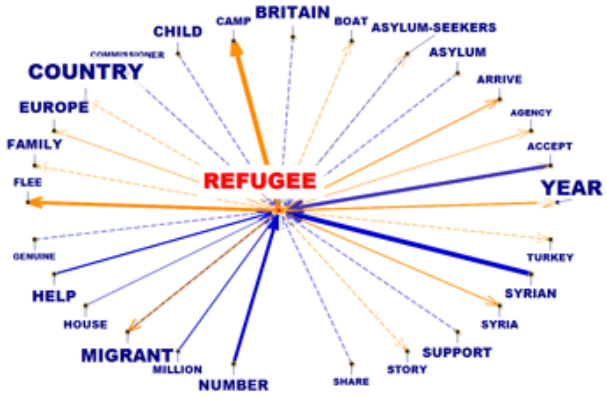


- Analyse des Séquences et Analyse des Réseaux

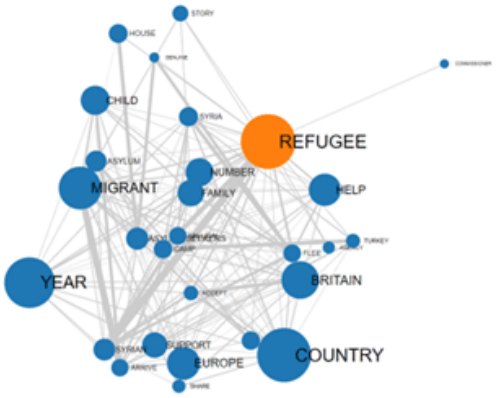
Cet outil **T-LAB** tient compte des **positions** des différentes unités lexicales à l'intérieur des phrases et il nous permet de représenter et d'explorer n' importe quel texte comme un **réseau de relations**.

Ceci signifie, après avoir exécuté ce type d'analyse, que l'utilisateur peut vérifier les relations entre les nœuds du réseau (c'est-à-dire les mots-clés) à plusieurs niveaux: a) en relations du type un-à-un; b) à l'intérieur d' «ego network»; c) à l'intérieur des «communautés» auxquelles ils appartiennent; d) à l'intérieur du réseau entier constitué par le texte en analyse.

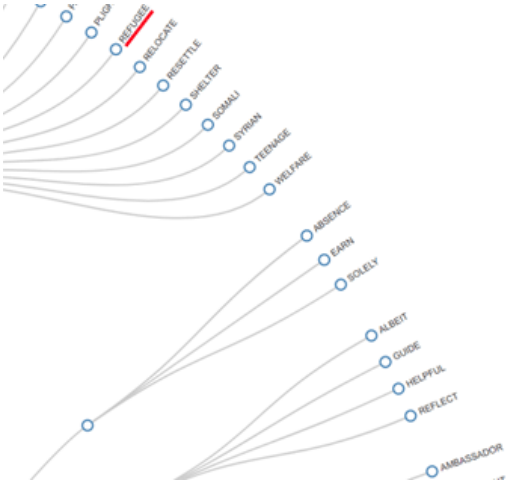
RELATIONS DU TYPE UN-A-UN



EGO-NETWORK

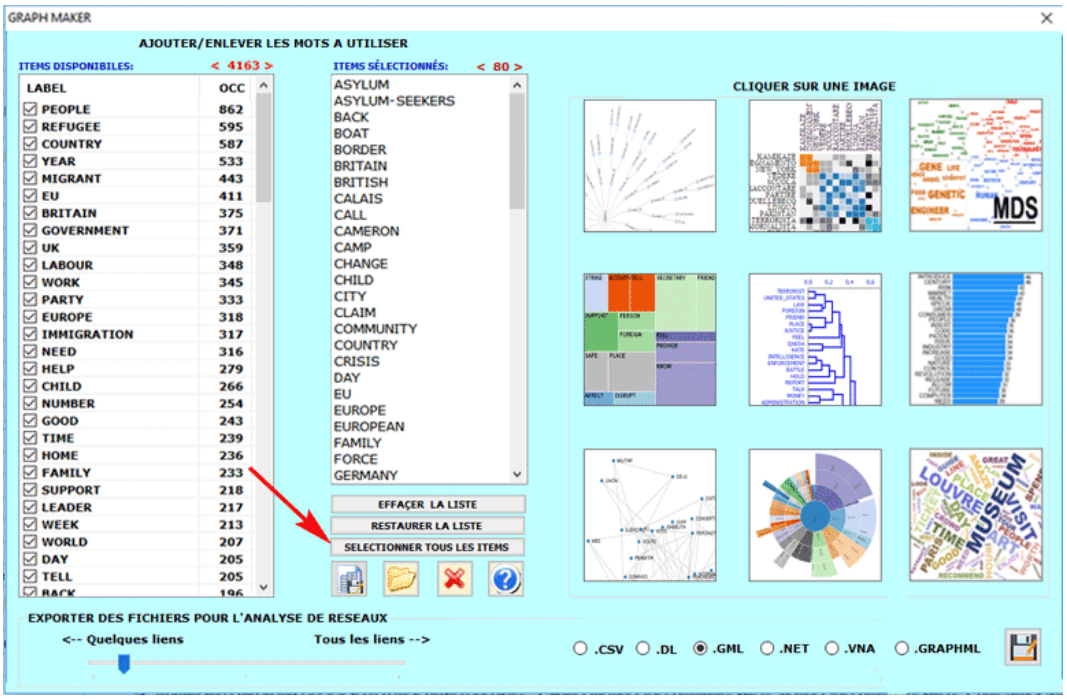


COMMUNAUTES



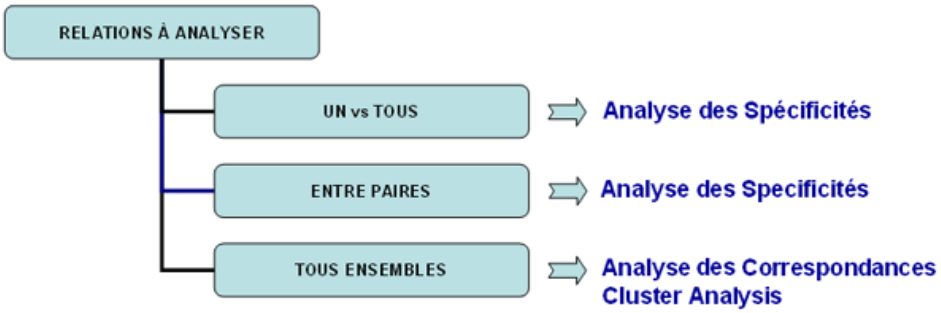
RESEAU ENTIER

De plus, à n'importe quel moment, en faisant clic sur l'option GRAPH MAKER, l'utilisateur peut créer des différents types de graphiques en utilisant des listes personnalisées de mots-clés, (voir ci-dessous).

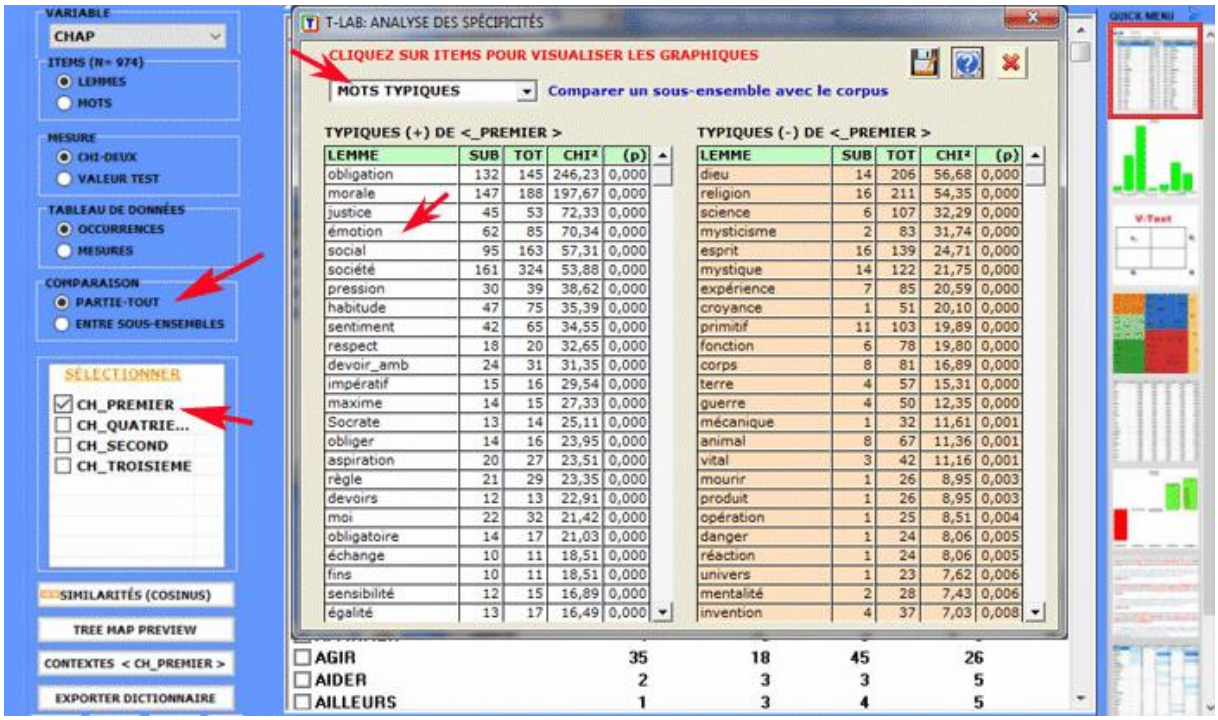


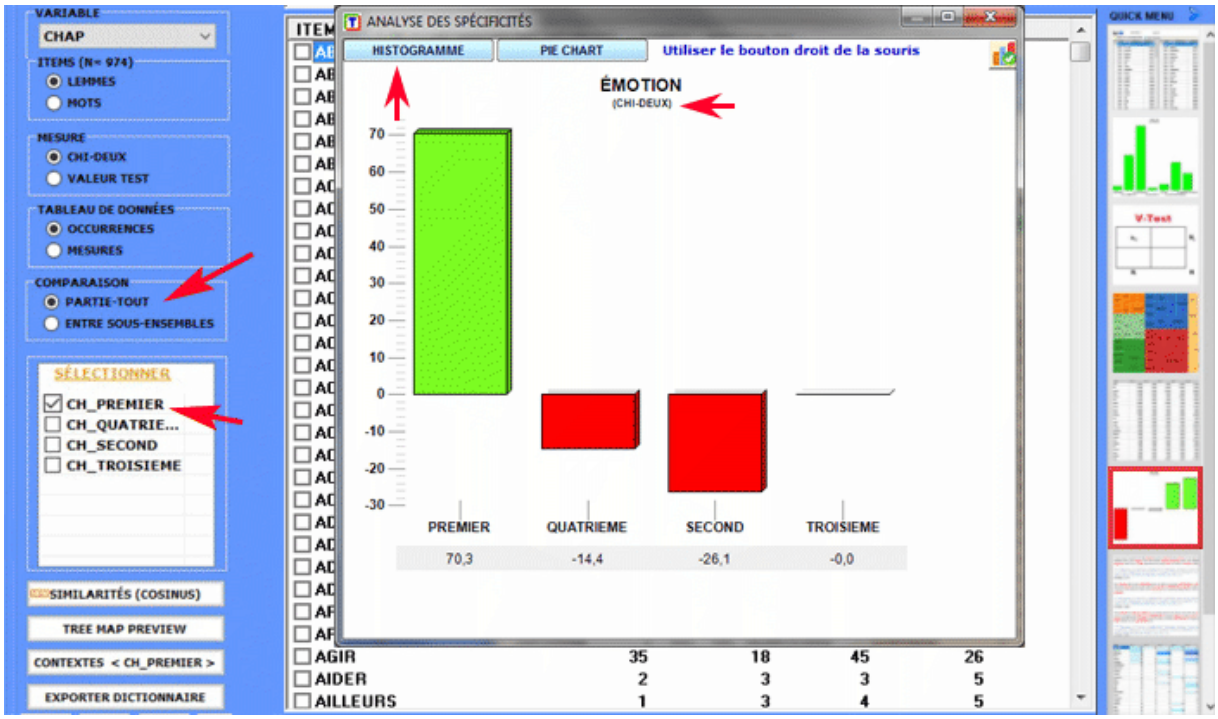
B : OUTILS POUR LES ANALYSES COMPARATIVES

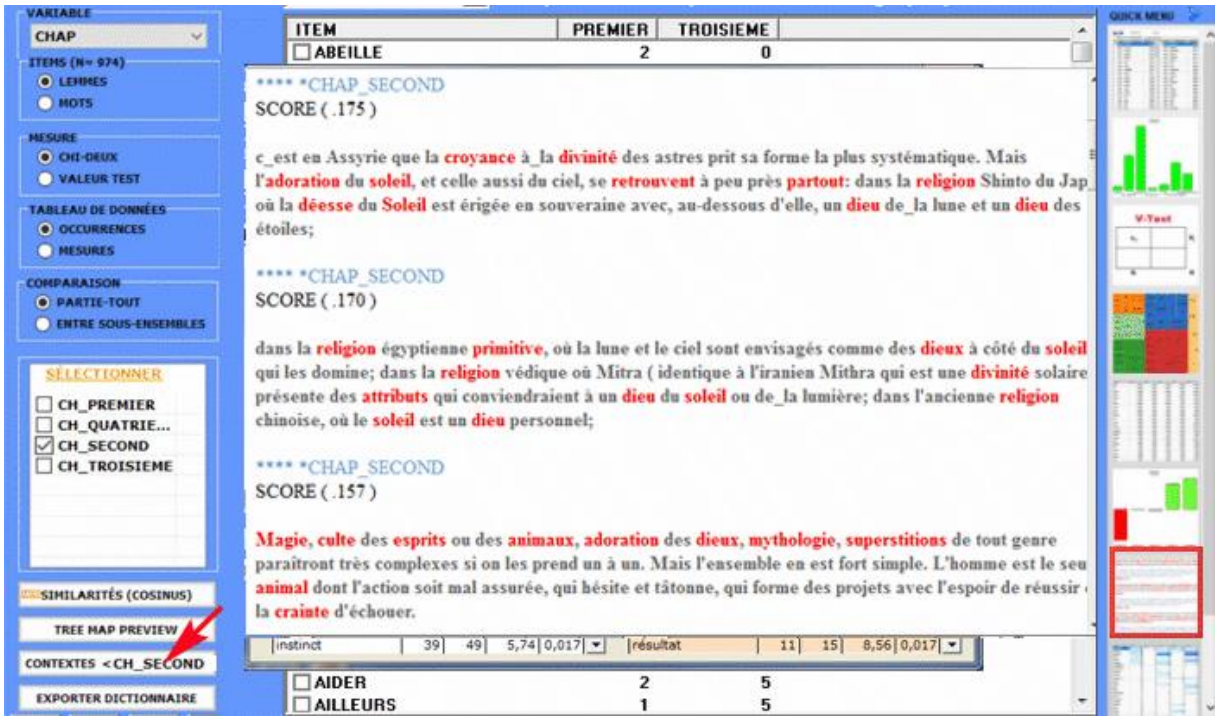
Ces outils nous permettent d'analyser différentes typologies de relations entre les unités de contexte.



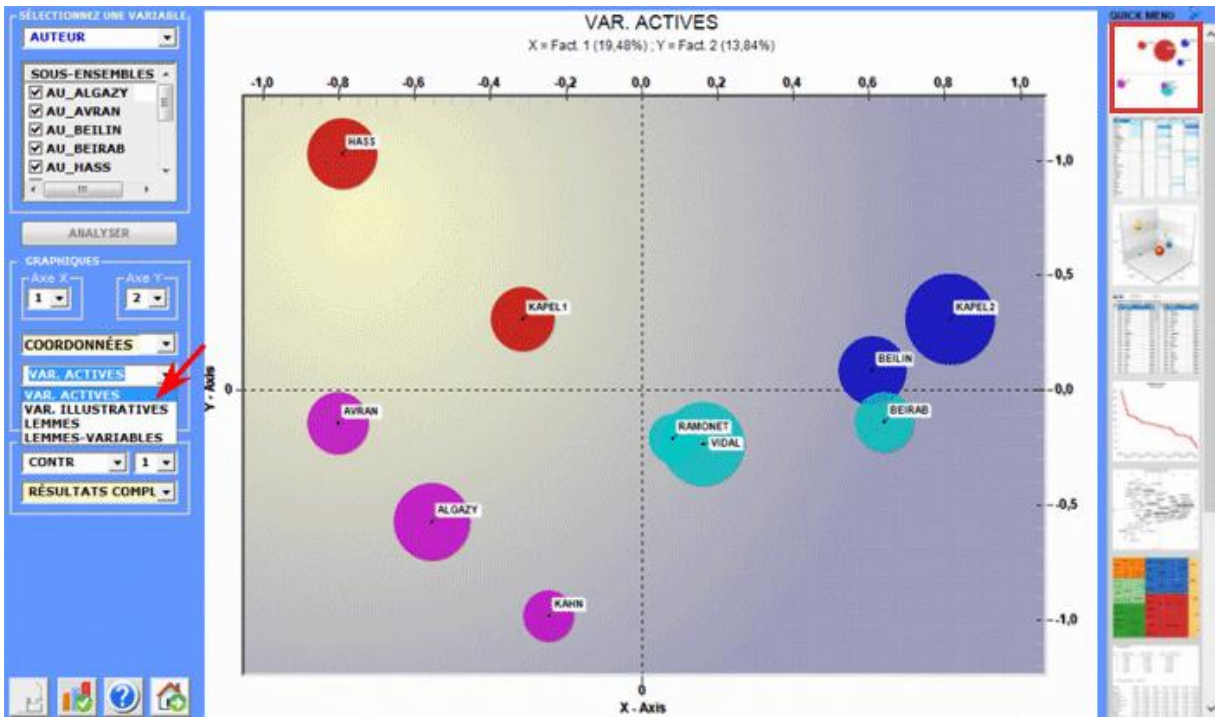
L'**Analyse des Spécificités** permet de vérifier quels mots sont “typiques” ou “exclusifs” de chaque sous-ensemble du corpus. En outre il nous permet d'extraire les contextes typiques, c'est-à-dire les contextes élémentaires caractéristiques, de chacun des sous-ensembles analysés (par exemple, les phrases "typiques" utilisées par certains leaders politiques).

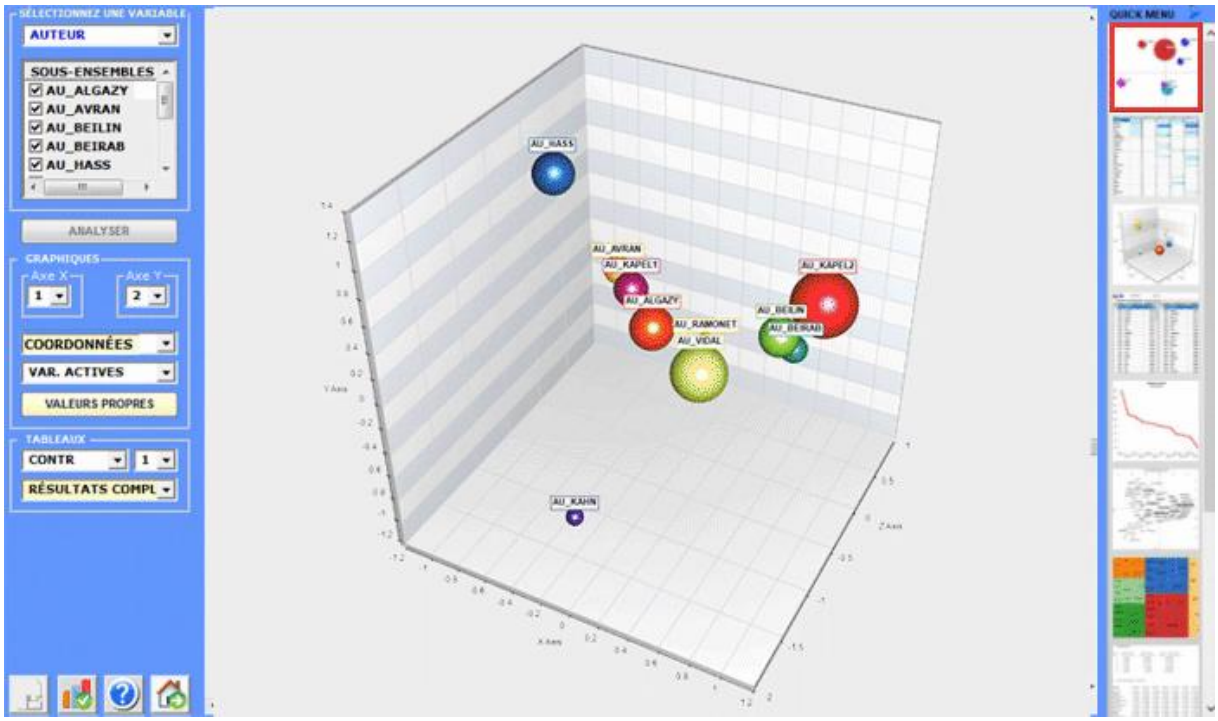

VARIABLE
CHAP
ITEMS (N= 974)
LEMMES
MOTS
MESURE
CHI-DEUX
VALEUR TEST
TABLEAU DE DONNÉES
OCCURRENCES
MESURES
COMPARAISON
PARTIE-TOUT
ENTRE SOUS-ENSEMBLES
SÉLECTIONNER
CH_PREMIER
CH_QUATRIE...
CH_SECOND
CH_TROISIEME
SIMILARITÉS (COSINUS)
TREE MAP PREVIEW
CONTEXTES < CH_PREMIER >
EXPORTER DICTIONNAIRE
T-LAB: ANALYSE DES SPÉCIFICITÉS
CLIQUEZ SUR ITEMS POUR VISUALISER LES GRAPHIQUES
MOTS TYPIQUES
Comparer un sous-ensemble avec le corpus
TYPIQUES (+) DE <_PREMIER >
TYPIQUES (-) DE <_PREMIER >
QUICK MENU
AGIR 35 18 45 26
AIDER 2 3 3 5
AILLEURS 1 3 4 5


VARIABLE
CHAP
ITEMS (N= 974)
LEMMES
MOTS
MESURE
CHI-DEUX
VALEUR TEST
TABLEAU DE DONNÉES
OCCURRENCES
MESURES
COMPARAISON
PARTIE-TOUT
ENTRE SOUS-ENSEMBLES
SÉLECTIONNER
CH_PREMIER
CH_QUATRIE...
CH_SECOND
CH_TROISIEME
SIMILARITÉS (COSINUS)
TREE MAP PREVIEW
CONTEXTES < CH_PREMIER >
EXPORTER DICTIONNAIRE
ANALYSE DES SPÉCIFICITÉS
HISTOGRAMME
PIE CHART
Utiliser le bouton droit de la souris
ÉMOTION
(CHI-DEUX)
PREMIER
QUATRIEME
SECOND
TROISIEME
70,3
-14,4
-26,1
-0,0
QUICK MENU
AGIR 35 18 45 26
AIDER 2 3 3 5
AILLEURS 1 3 4 5

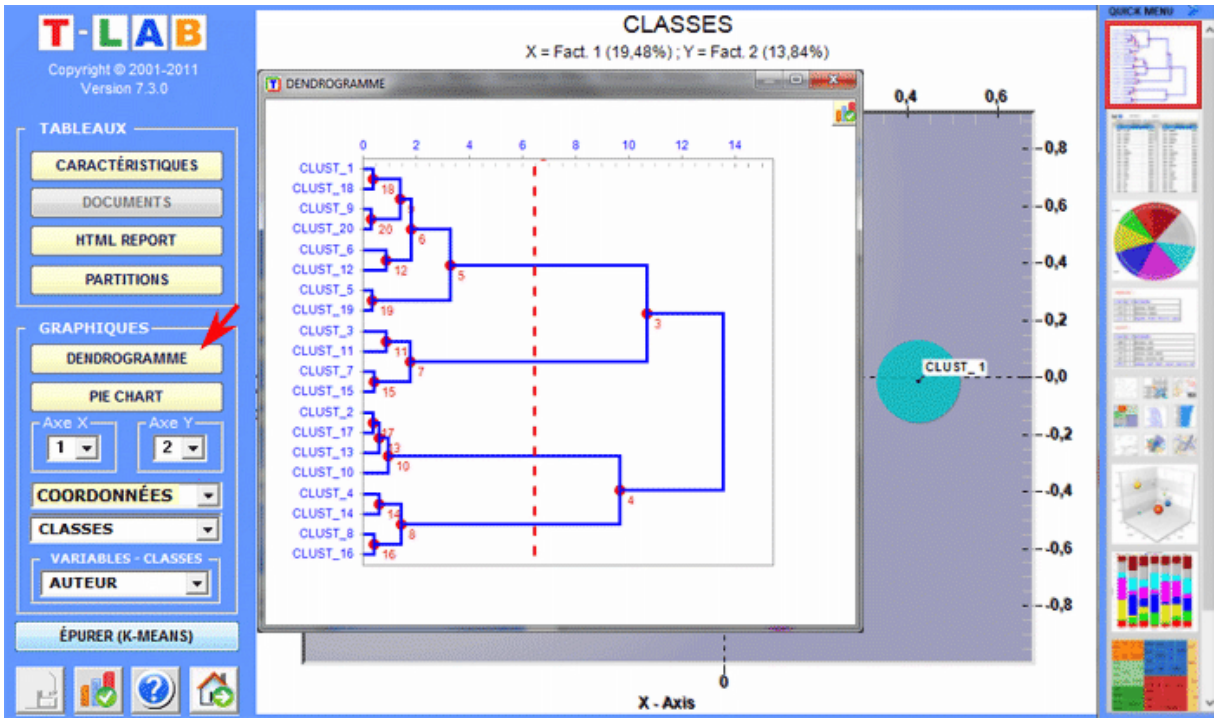


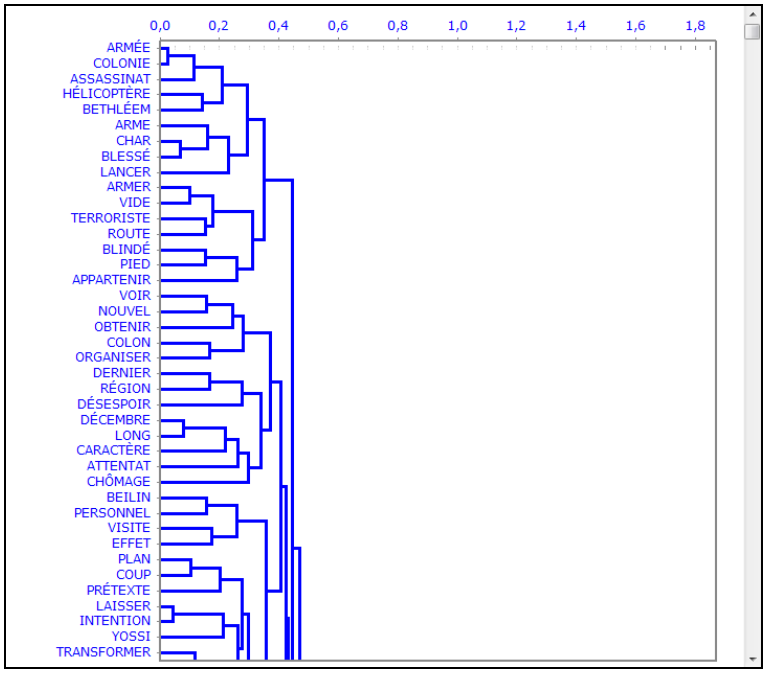
L'**Analyse des Correspondances** permet d'explorer différentes typologies de relations (différences et ressemblances) entre les unités de contexte.





La **Cluster Analysis**, qui peut être réalisée avec différentes techniques, permet d'identifier des groupes d'unités textuelles qui aient deux caractéristiques complémentaires: maximum d'homogénéité dans leur interne et maximum d'hétérogénéité entre eux deux et les autres clusters.





C : OUTILS POUR LES ANALYSES THÉMATIQUES

Ces outils permettent de repérer, examiner et cartographier les "thèmes" présents dans les textes analysés.

Puisque "thème" est un mot polysémique, dans ce cas il est utile se référer à des définitions opérationnelles. En fait, dans ces outils de **T-LAB**, le mot "thème" est un label utilisé pour indiquer quatre entités différentes :

1- un **cluster thématique** d'unité de contexte caractérisé par les mêmes patterns de mots-clés (voir **Analyse Thématique des Contextes Elémentaires** et **Classification thématique des Documents)**;

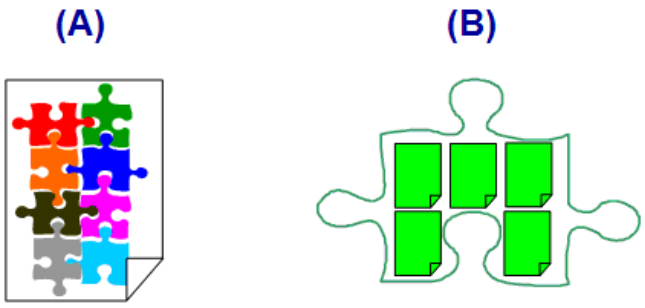
2- un **groupe thématique de mots-clés** classés comme appartenant à la même catégorie (voir l'outil **Classification basée sur des Dictionnaires**);

3- un **élément d'un modèle probabiliste** qui représente chaque unité de contexte (soit un contexte élémentaire, soit un document), comme généré par un mélange de "thèmes" ou "topics" (voir les outils **Modélisation des Thèmes émergentes** et **Textes et Discours comme Systèmes Dynamiques**);

4- un **mot-clé** ("thématique") **spécifique** utilisé pour extraire un ensemble de contextes élémentaires dans lesquels ce mot est associé à un groupe de mots spécifique présélectionnés par l'utilisateur (voir **Contextes-Clé de Mots Thématiques**).

Par exemple, selon le type d'outil que nous sommes en train d'utiliser, un document spécifique peut être analysé comme étant composé de différents « thèmes » (voir «A» ci-dessous) ou bien comme appartenant à un ensemble de documents concernant le même «thème» (voir « B » ci-dessous). En effet, dans le cas «A» chaque thème peut correspondre à

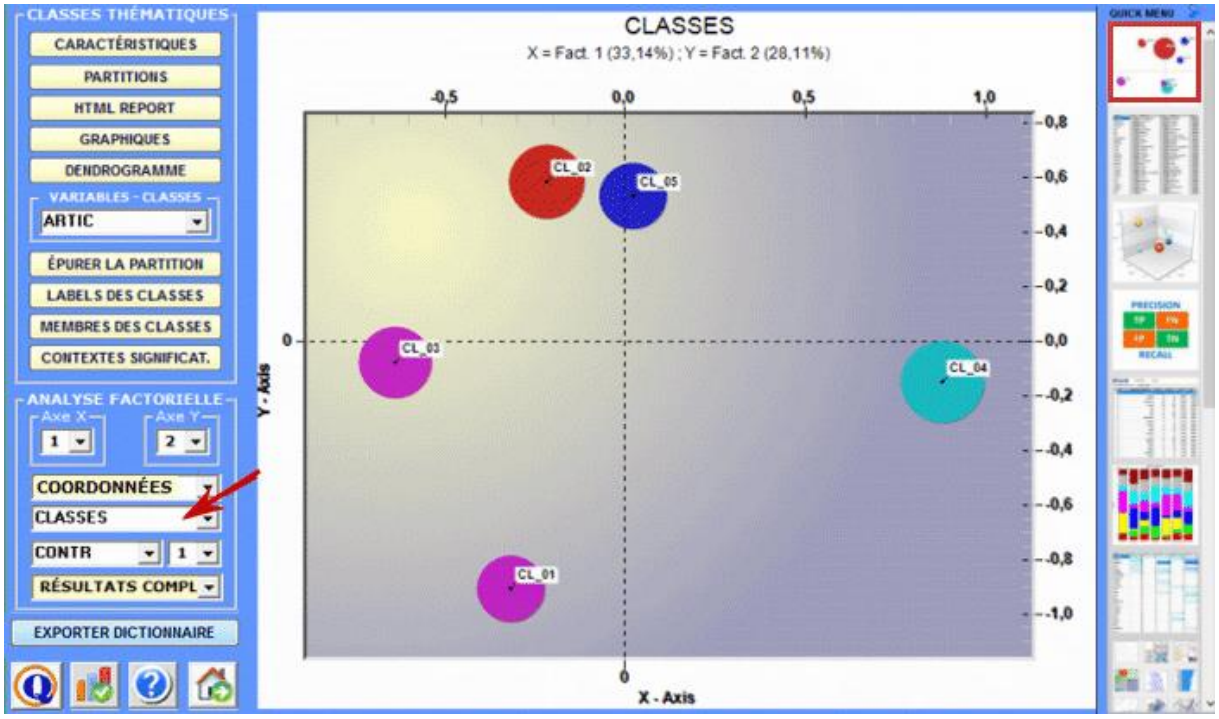
un mot ou bien à une phrase, tandis que dans le cas «B» un thème peut être une étiquette attribuée à un groupe de documents caractérisés par les mêmes patterns de mots-clés.

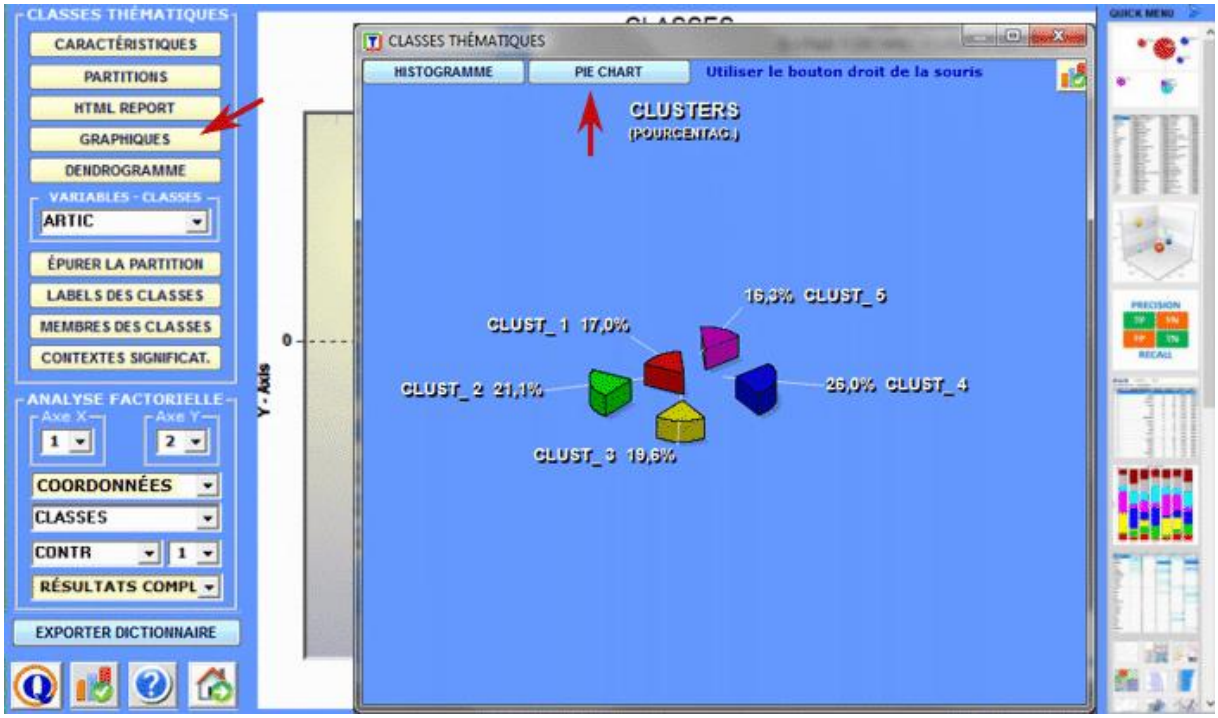


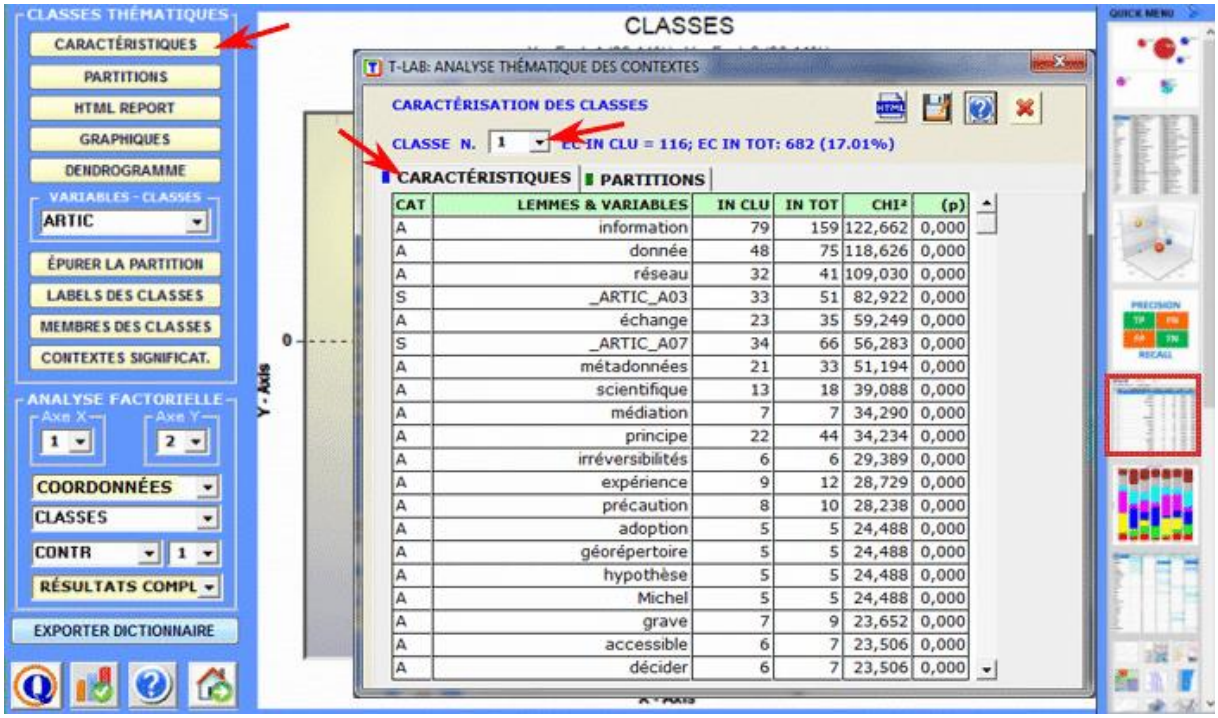
En détail, les façons dont **T-LAB** extrait les thèmes sont les suivantes:

1 - soit l'outil **Analyse Thématiques des Contextes Elémentaires**, soit l'outil **Classification Thématique des Documents** fonctionnent de manière suivante :

a- ils réalisent une **analyse des cooccurrences** pour identifier les classes thématiques de unités de contexte;
b- ils réalisent une **analyse comparative** pour confronter les profils des différentes classes;
c- ils produisent différents types de graphiques et de tableaux (voir ci-après);
d- ils permettent d'archiver les **nouvelles variables** obtenues (classes thématiques) et de les utiliser dans d'autres analyses.

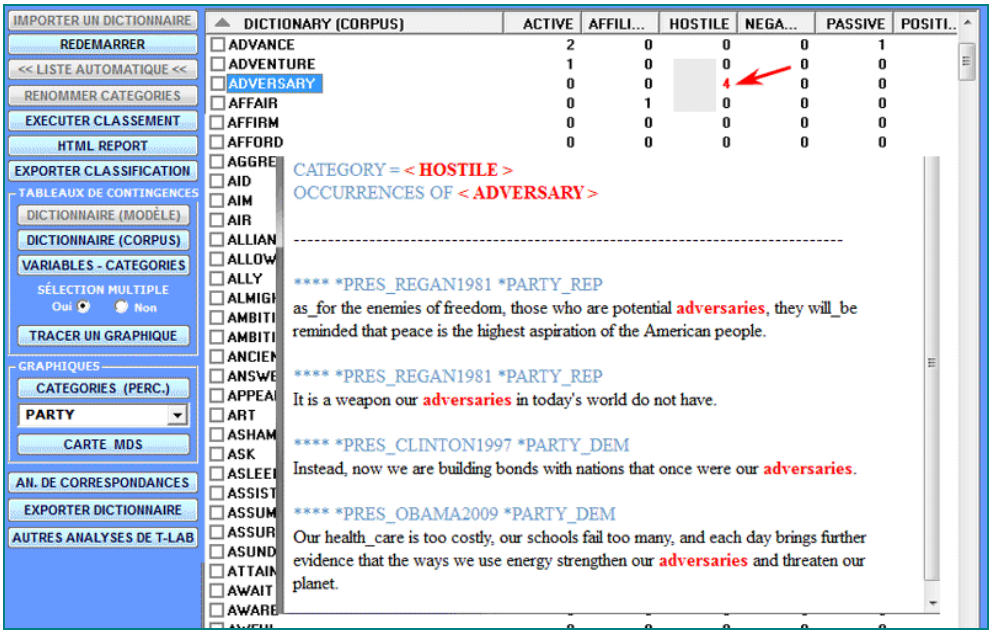


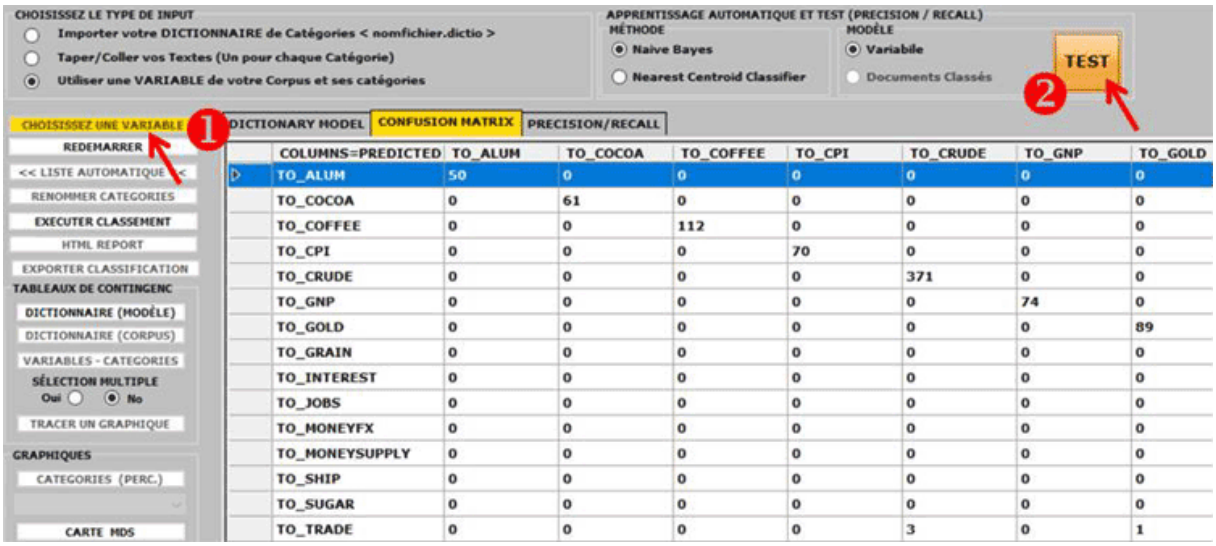




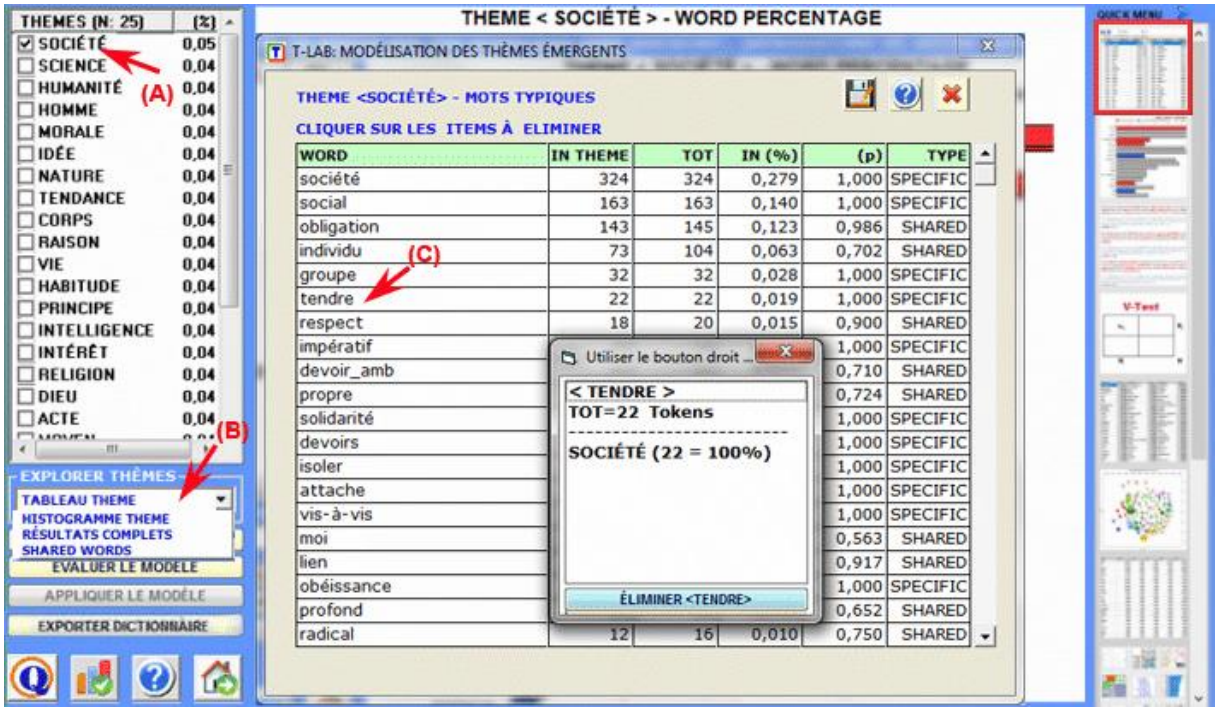
CAT	LEMMES & VARIABLES	IN CLU	IN TOT	CHI²	(p)
A	information	79	159	122,662	0,000
A	donnée	48	75	118,626	0,000
A	réseau	32	41	109,030	0,000
S	_ARTIC_A03	33	51	82,922	0,000
A	échange	23	35	59,249	0,000
S	_ARTIC_A07	34	66	56,283	0,000
A	métadonnées	21	33	51,194	0,000
A	scientifique	13	18	39,088	0,000
A	médiation	7	7	34,290	0,000
A	principe	22	44	34,234	0,000
A	irréversibilités	6	6	29,389	0,000
A	expérience	9	12	28,729	0,000
A	précaution	8	10	28,238	0,000
A	adoption	5	5	24,488	0,000
A	géorépertoire	5	5	24,488	0,000
A	hypothèse	5	5	24,488	0,000
A	Michel	5	5	24,488	0,000
A	grave	7	9	23,652	0,000
A	accessible	6	7	23,506	0,000
A	décider	6	7	23,506	0,000

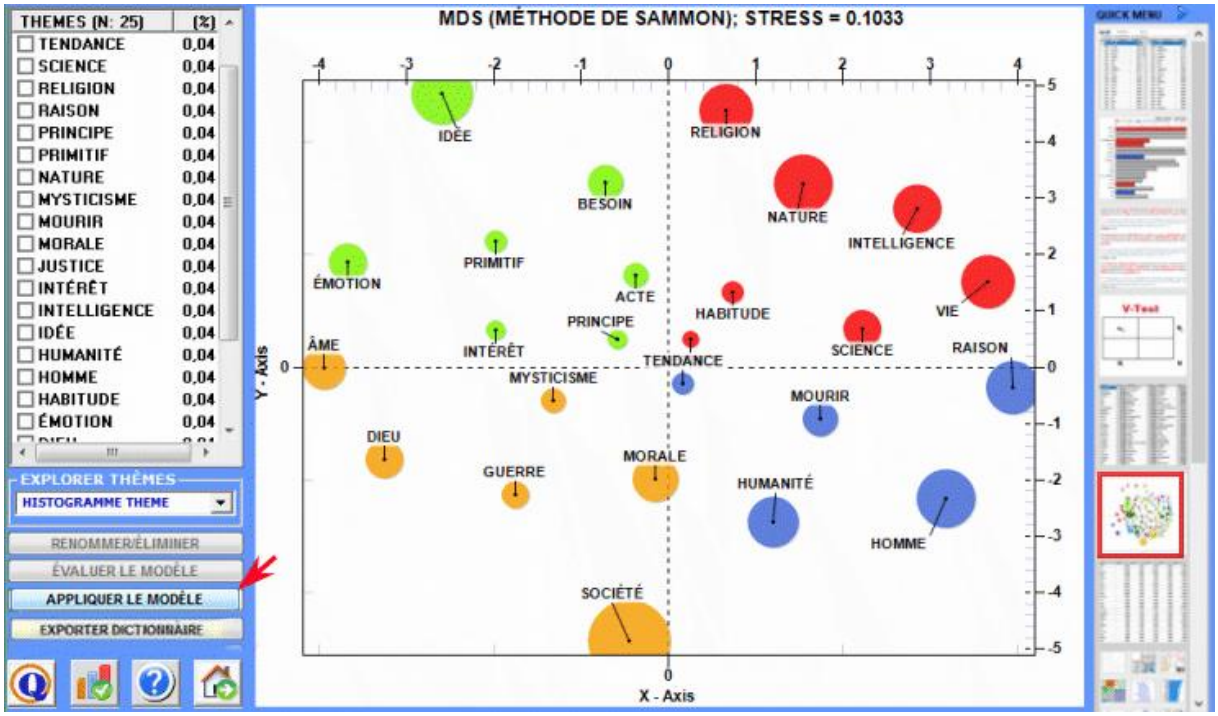
2 - à l'aide de l'outil **Classification Basée sur des Dictionnaires** nous pouvons facilement construire / tester / appliquer des modèles (par ex. des dictionnaires de catégories) soit pour l'analyse classique du contenu soit pour la sentiment analysis. En effet cet outil nous permet d'effectuer une classification automatique de type top-down aussi bien des unités lexicales (c'est-à-dire mots et lemmes) que des unités de contexte (c'est-à-dire phrases, paragraphes et documents courts).



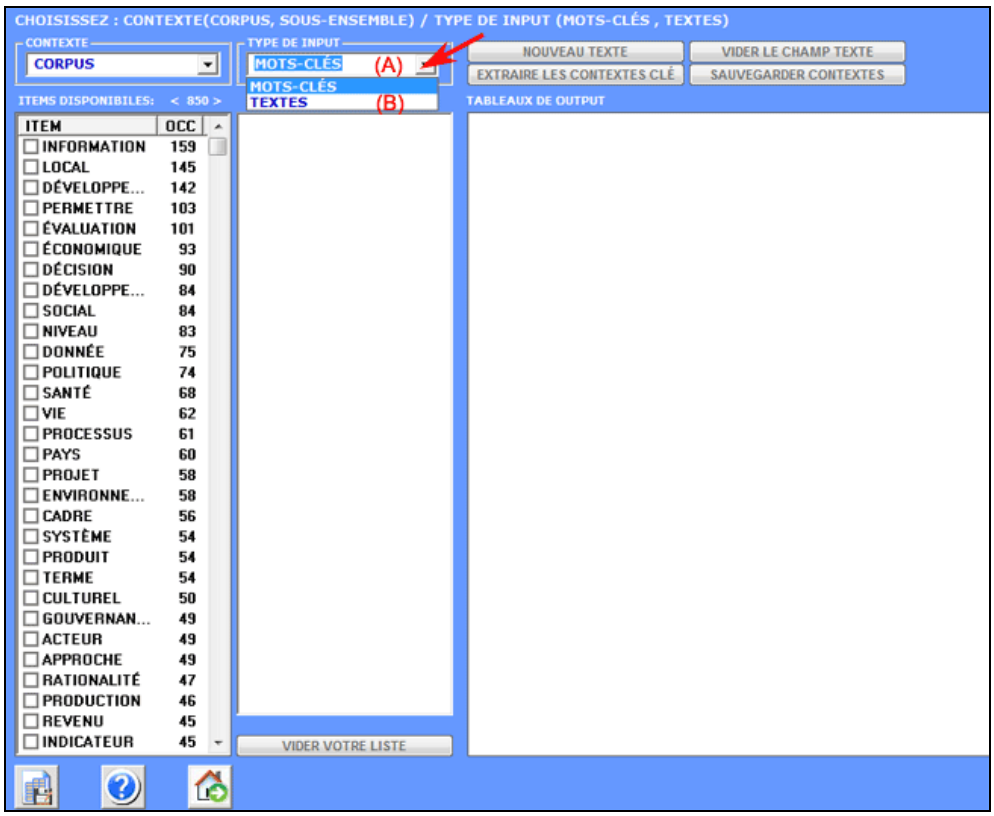


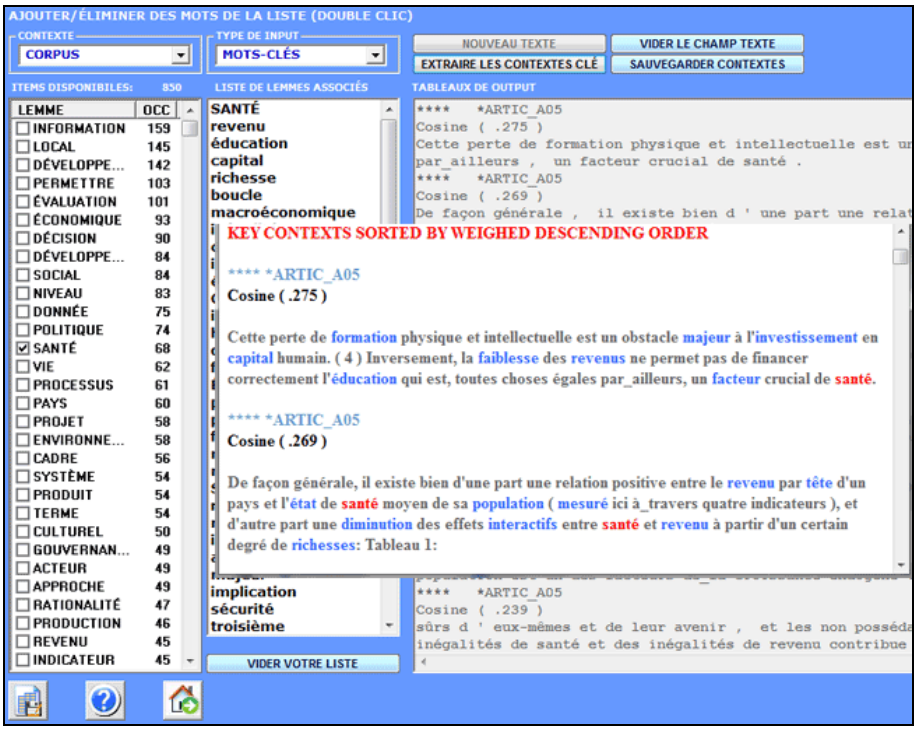
3 - grâce à l'outil **Modélisation des Thèmes Émergents** (voir ci-dessous) les composants du «mélange» thématique peuvent être décrits par leur vocabulaire caractéristique et peuvent être utilisés pour la construction de grilles pour l'analyse qualitative et / ou pour la classification automatique des unités de contexte (c'est-à-dire contextes élémentaires ou documents).





4 - l'outil **Contextes Clé des Mots Thématiques** (voir ci-dessous) peut être utilisé pour deux buts différents: (a) extraire des listes d'unités de contexte (c'est-à-dire contextes élémentaires) qui permettent d'approfondir la valeur thématique de **mots-clés** spécifiques, (b) extraire des groupes d'unités de contexte qui sont semblables à n'importe quel **texte** « exemple » choisi par l'utilisateur.

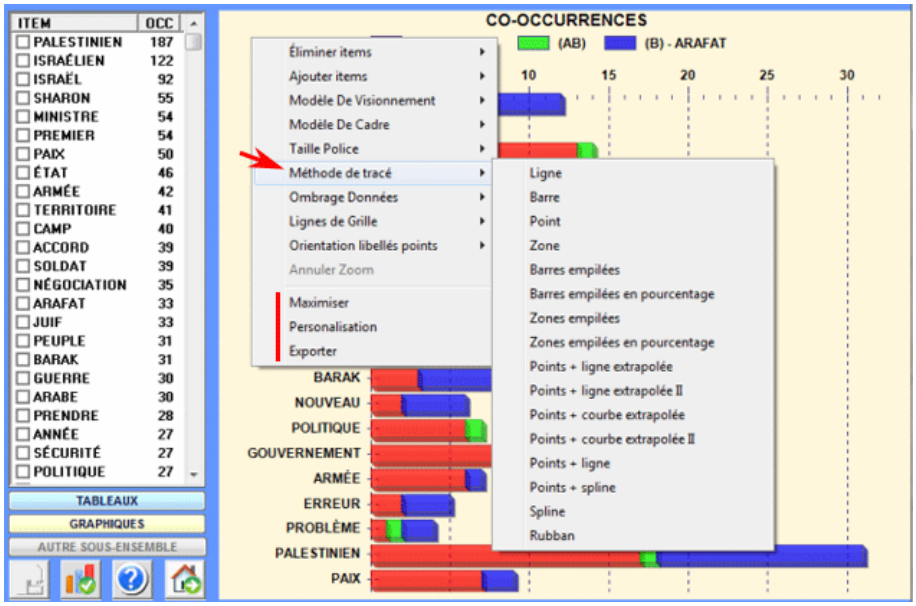


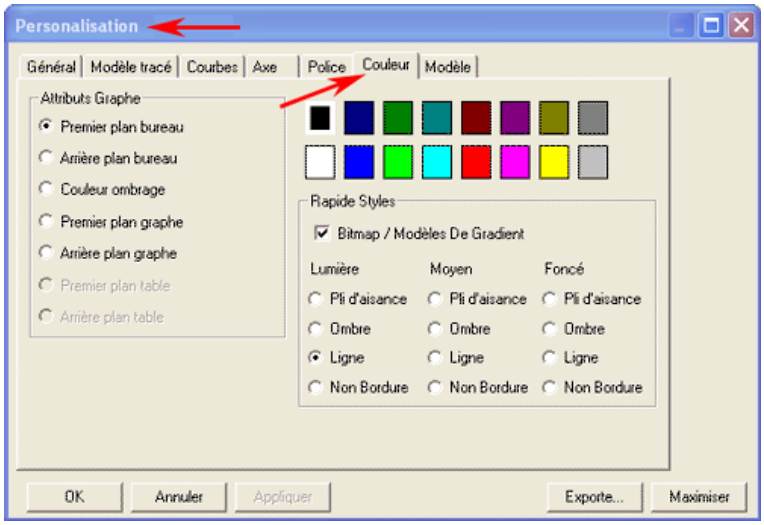


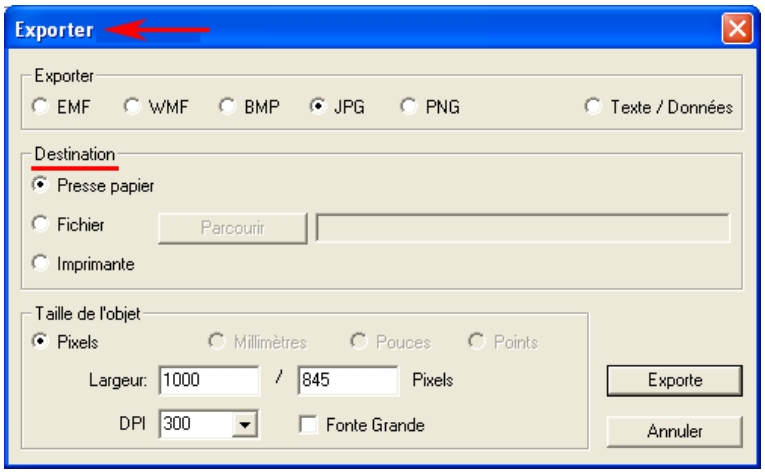
6 - L' INTERPRÉTATION DES OUTPUTS consiste en la consultation des tableaux et des graphiques produits par **T-LAB**, en l'éventuelle personnalisation de leur format et dans le fait de faire des inférences sur la signification des relations représentées.

Dans le cas des **tableaux**, selon les cas, **T-LAB** permet de les exporter dans des fichiers avec les extensions suivantes: **.DAT**, **.TXT**, **.CSV**, **.XLXS**, **.HTML**. Ceci signifie que, en se servant de n'importe quel éditeur de textes et/ou d'un applicatif de la suite Microsoft Office, l'utilisateur peut facilement les importer et les réélaborer.

Dans le cas des **graphiques**, les sous-menus appropriés activés avec le clic droit de la souris permettent d'effectuer plusieurs opérations: zoom (clic gauche et glisser), maximisation, personnalisation et exportation des outputs en plusieurs formats.





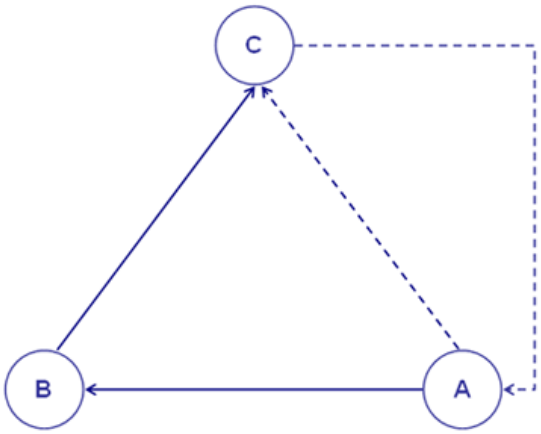


Certains critères généraux pour l'interprétation des outputs **T-LAB** sont illustrés dans un papier cité dans la **Bibliographie** et disponible sur le site https://www.tlab.it (Lancia F.: 2007). Dans ce dernier on propose l'hypothèse que les outputs des élaborations statistiques (tableaux et graphiques) sont un type particulier de textes, c'est-à-dire des objets multi-sémiotiques caractérisés par le fait que les relations entre les signes et les symboles sont ordonnées par des mesures qui renvoient à des **codes** spécifiques.

Dans d'autres termes, aussi bien dans le cas des textes écrits dans le langage naturel que dans ceux écrits dans le langage de la statistique, la possibilité de faire des inférences sur les relations qui organisent les **formes du contenu** est garantie par le fait que les relations entre les **formes de l'expression** ne sont pas casuelles (random); en effet, dans le premier cas (langage naturel) les unités signifiantes se succèdent ordonnées de façon linéaire (l'une après l'autre dans le chaîne du discours), alors que dans le second cas (tableaux et graphiques) les principes d'ordonnance sont constitués par les mesures qui déterminent l'organisation des **espaces sémantiques** multidimensionnels.

Même si les espaces sémantiques représentés dans les cartes **T-LAB** sont très variés, et chacun d'eux requiert des procédures interprétatives spécifiques, nous pouvons faire l'hypothèse que - en général - la logique du processus inférentiel est la suivante:

A - relever une relation significative entre les unités "présentes" sur le plan de l'expression (par ex. entre "données" des tableaux et/ou entre "labels" des graphiques);
B - explorer et confronter les traits sémantiques des mêmes unités et les contextes auxquels elles sont mentalement et culturellement associées (plan du contenu);
C - construire une hypothèse ou une catégorie d'analyse qui, dans le contexte défini par le corpus, rendent raison des relations entre formes de l'expression et formes du contenu.



Actuellement les options de **T-LAB** ont les **limitations** suivantes:

- dimension du corpus: max 90Mo correspondant à environ 55.000 pages de format .txt;
- documents primaires : max 30.000 (N.B.: Lorsque aucun des textes dépasse 2.000 caractères, la limite est de 99.999 documents);
- variables catégorielles: maximum 50, chacune avec un maximum de 150 modalités;
- modélisation des thèmes émergents : max 5.000 unités lexicales (*) pour 5.000.000 occurrences;
- analyse thématique des contextes élémentaires: max 300.000 lignes (unités de contexte) par 5.000 colonnes (unités lexicales);
- classification thématique des documents: max 99.999 lignes (unités de contexte) par 5.000 colonnes (unités lexicales);
- analyse des spécificités (unités lexicales x modalités d'une variable): max 10.000 lignes x 150 colonnes;
- analyse des correspondances (unités lexicales x modalités d'une variable): max 10.000 lignes x 150 colonnes;
- analyse des correspondances (unités de contexte x unités lexicales): max 10.000 lignes x 5.000 colonnes;
- analyse des correspondances multiples (contextes élémentaires x modalités des variables): max 150.000 lignes x 250 colonnes;
- décomposition en valeurs Singulières (SVD) : max 300.000 lignes par 5.000 colonnes;
- classification (cluster analysis) qui emploie les résultats d'une précédente analyse des correspondances (ou SVD): max10.000 lignes (unités lexicales ou contextes élémentaires);
- association des mots, comparaisons entre paires de mots-clés: max 5.000 unités lexicales;
- analyse des mots associés et cartes conceptuelles: max 5.000 unités lexicales;
- analyse de séquences: maximum 5.000 unités lexicales (ou catégories) pour 3.000.000 occurrences.

(*) Dans **T-LAB**, les 'unités lexicales' sont mots, multi-mots, lemmes et catégories sémantiques. Ainsi, lorsque la lemmatisation automatique est appliquée, 5.000 unités lexicales correspondent à environ 12.000 mots.

CONFIGURATIONS D'ANALYSE

Configuration Automatique et Personnalisée

Le choix de la configuration **automatique** (A) ou **personnalisée** (B) concerne la liste des **mots-clés** utilisés dans toutes les analyses effectuées au moyen de **T-LAB**. Ce choix est réversible jusqu'à ce que l'utilisateur n' effectue pas des opérations qui modifient le Dictionnaire du corpus.

A) **CONFIGURATION AUTOMATIQUE**

Le choix de la **configuration automatique** signifie que la liste des mots-clés comprenne **jusqu'à un maximum de 5000 unités lexicales** sélectionnées automatiquement par **T-LAB** et appartenant aux catégories grammaticales qui sont plus denses de sens: noms, verbes, adjectifs et adverbes.

Le critère de sélection change selon le genre de fichier analysé.

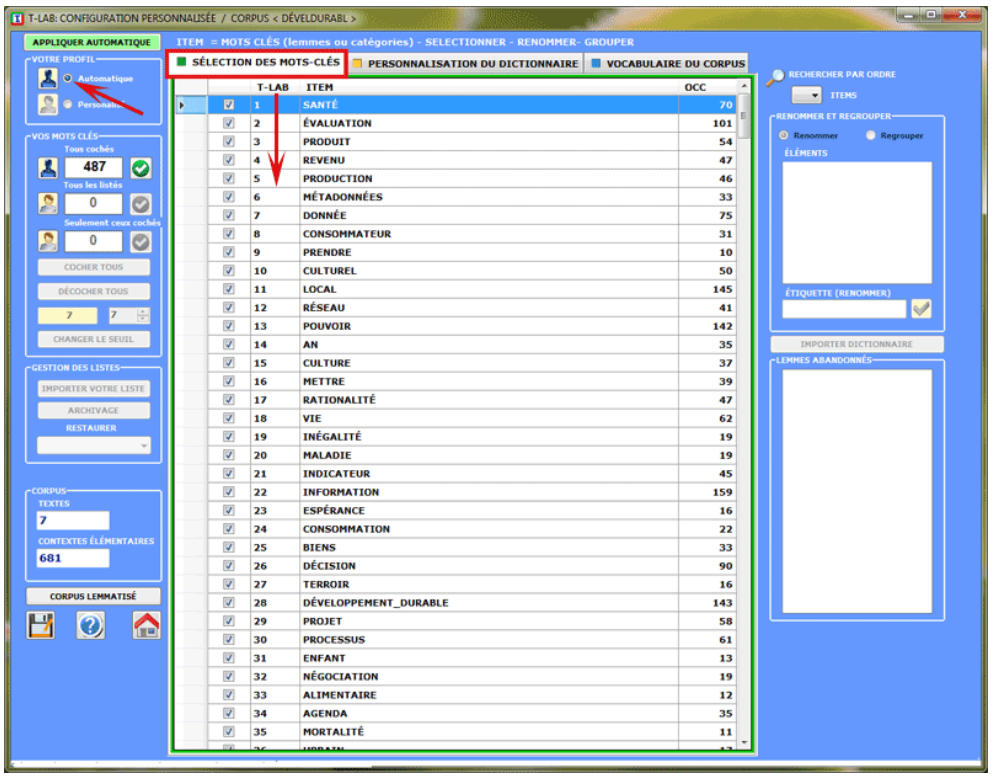
Si le corpus est un texte unique **T-LAB** choisit simplement les mots avec les valeurs d'**occurrence** les plus élevées.

Quand le corpus se compose de deux textes ou plus **T-LAB** emploie l'algorithme suivant:

- il choisit les mots avec les valeurs d'occurrence plus élevées que le seuil minimum;
- il applique le TF-IDF ou le test du chi-deux à toutes les croix de chaque mot pour tous les textes analysés (N.B. : Dans le cas du chi-deux, les textes analysés doivent être maximum 500);
- il choisit les mots avec les valeurs du TF-IDF ou du chi-deux les plus élevées, c'est-à-dire les mots qui, dans le corpus, font la différence.

N.B.. :
Dans le cas où le corpus est constitué par deux ou plusieurs textes, l'utilisateur peut choisir le critère de sélection (chi-carré ou bien TF-IDF) dans le stade de l'importation (voir ci-dessous).

- Lorsque l' option pour la **configuration automatique** est activée, le tableau avec la liste des **Mots-clés** inclut une colonne 'T-LAB' qui indique l' importance de chaque élément selon le critère sélectionné (voir ci-dessous).



B) CONFIGURATION PERSONNALISEE

Le choix de la **configuration personnalisée** permet à l' utilisateur de sélectionner, renommer et regrouper les unités lexicales (mots, lemmes ou catégories) à inclure dans les analyses successives de **T-LAB**.

Dans le tableau est reportée la liste (liste 1) des unités lexicales avec des valeurs d'occurrence égales ou supérieures au **seuil** préfixé. Certaines d'entre elles, celles indiquées avec un "☑", font partie d'une sous-liste (liste 2) proposée par **T-LAB** (voir Configuration Automatique).

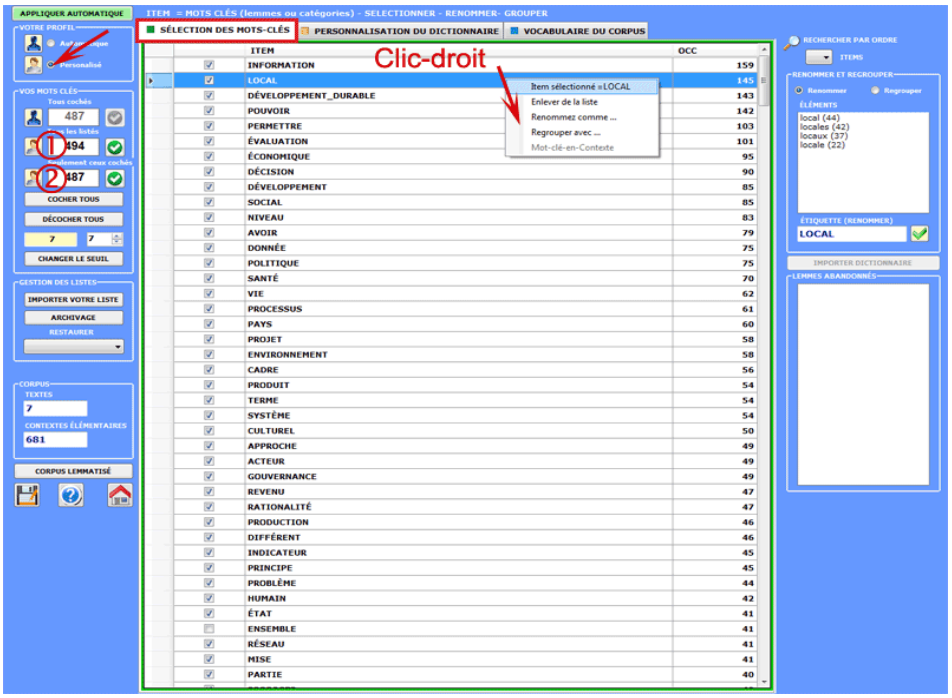
En fonction des analyses qu'il entend effectuer, l'utilisateur peut décider d'utiliser/modifier la liste (1) ou la liste (2).

Dans les deux cas les opérations possibles sont les suivantes:

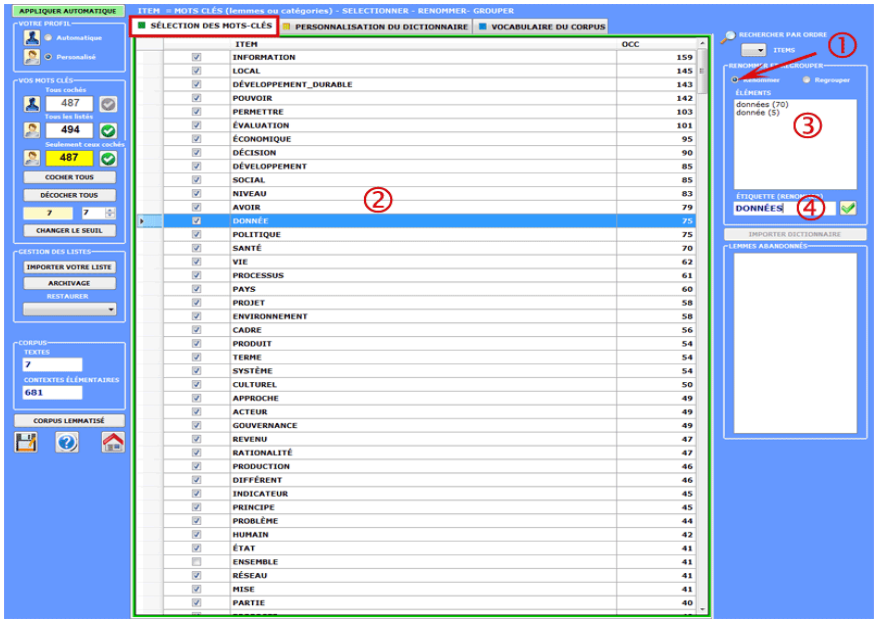
- **modifier** la valeur de seuil ;
- **sélectionner** les lemmes à rejeter de l'analyse ;
- **rétablir** l'utilisation d'un lemme ou plus ;
- **sélectionner/ de-sélectionner** les items à utiliser.

Un clic sur le bouton "utiliser liste (1)" ou le bouton "utiliser liste (2)" rend active l'option "personnalisée" des configurations d'analyse.

Les options qui concernent les interventions sur les lemmes individuels sont accessibles par le bouton droit de la souris en sélectionnant un élément quelconque du tableau (voir ci-dessous).

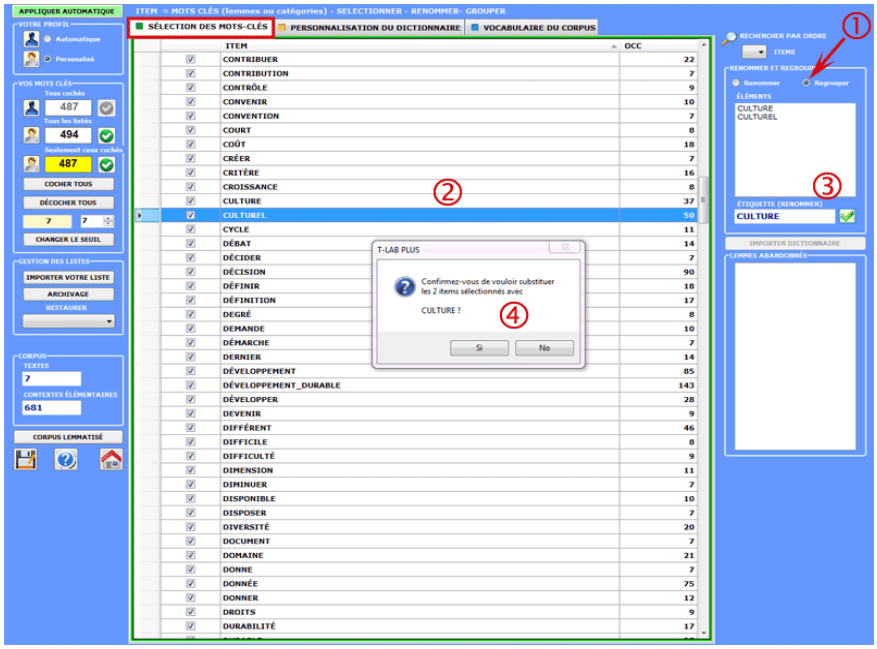


Pour **renommer** chaque lemme agir comme suit :
1. s' assurer que l' option "renommer" soit active;
2. cliquer sur un item du tableau;
3. choisir un des mots ou bien frapper un label à votre choix;
4. cliquer sur "remplacer".



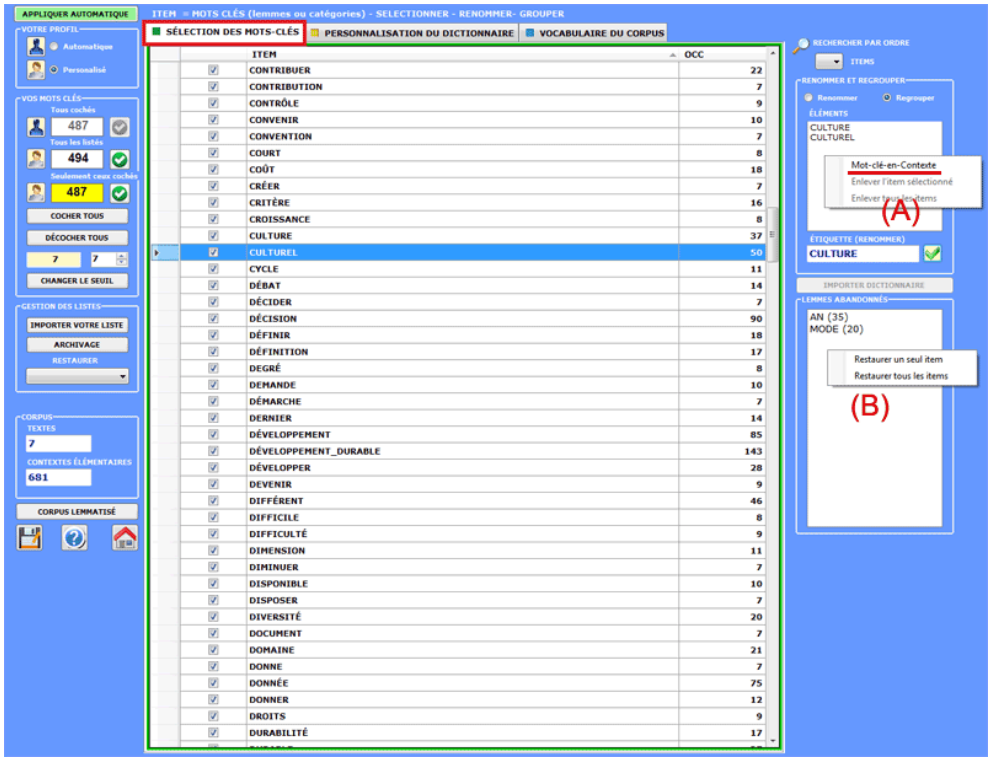
Les **regroupements** de deux ou plus lemmes doivent être effectués comme suit :

1. sélectionner la modalité "regroupements";
2. cliquer sur deux ou plus items du tableau;
3. choisir (à l' aide d' un clic) un des mots ou bien frapper un label à votre choix;
4. cliquer sur "remplacer".

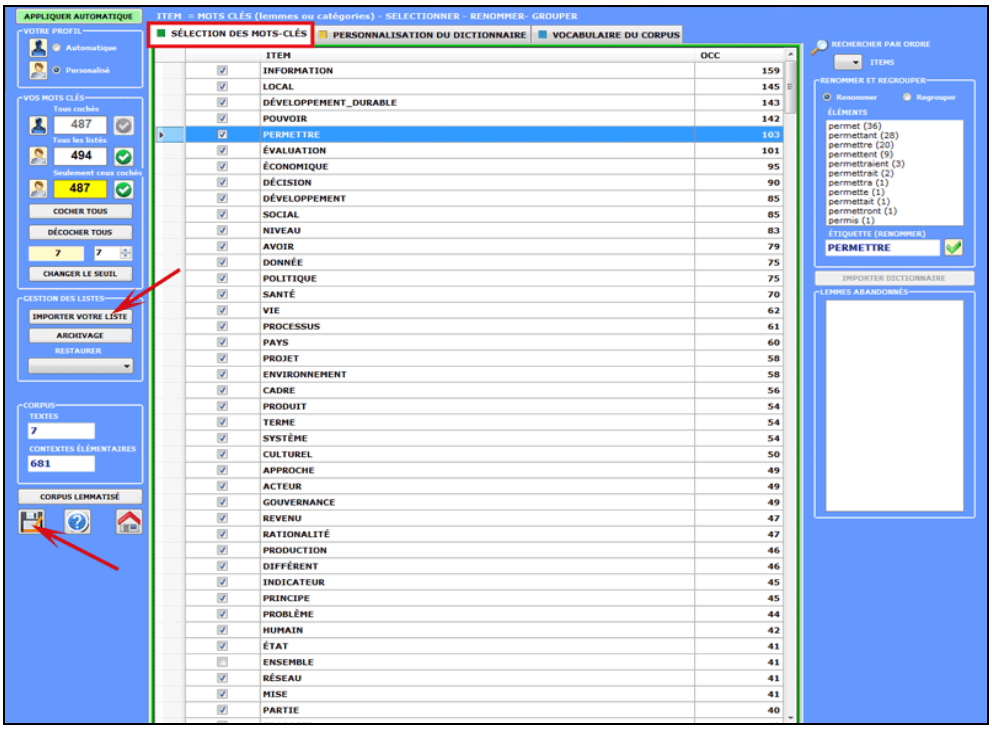


Des options supplémentaires peuvent être activées à l'aide du **bouton à droite** dans la case avec les éléments à renommer / regrouper (A) ou dans la case avec les « lemmes supprimés » (B).

En particulier, lorsque - dans le cas (A) – on sélectionne l' option «key-word-in-Context », vous pouvez accéder automatiquement à l'instrument **Concordances** et vérifier les contextes d'occurrence des différents éléments (voir ci-dessous).



Un bouton spécifique (voir ci-dessous) vous permet d'**importer des listes personnalisées de mots-clés**.
Chaque liste à importer (ficher nommé MyList.diz), peut contenir jusqu'à un maximum de 10.000 lignes (min = 20). Chaque ligne de votre liste doit être un mot sans espaces vides ni signes de ponctuation.
Un modèle de fichier MyList.diz est automatiquement créé par T-LAB lors de l'enregistrement de votre liste des mots clé (voir le bouton approprié en bas à gauche).



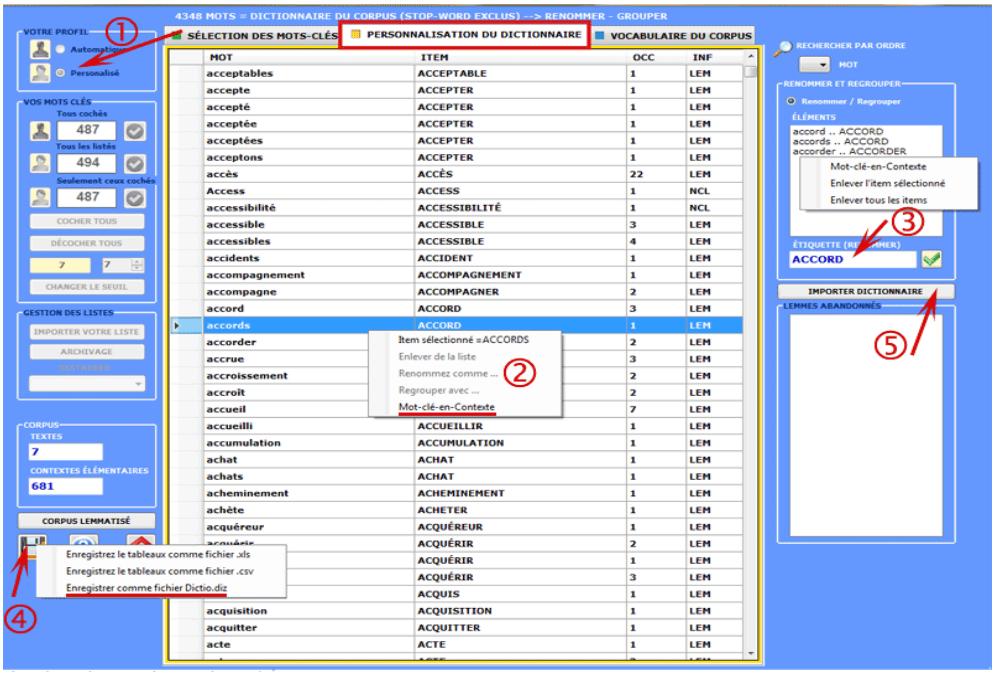
Le **paramétrage des analyses**, à savoir aussi bien la sélection des lemmes (ou catégories) que le **dictionnaire** utilisé, peut être sauvegardé et récupéré à travers les boutons réservés à cet usage. Ceci signifie que le même corpus - sans avoir besoin d'être importé à nouveau - peut être analysé avec différents dictionnaires et avec des sélections différentes de mots-clés. Ceci pour un maximum de 10 différents paramétrages.

En outre, dans **T-LAB** il est prévu que les paramétrages puissent être revus en plusieurs sessions, même en utilisant plusieurs fois la fonction **Personnalisation du Dictionnaire.**

Personnalisation du Dictionnaire

L'option **Personnalisation du Dictionnaire** ouvre une fenêtre par laquelle l'utilisateur peut intervenir sur le dictionnaire du corpus, c'est-à-dire qu'il peut renommer ou grouper les **lemmes** (voir l'option '3' ci-dessous).exporter son dictionnaire (voir l'option '4' ci-dessous) ou importer un **dictionnaire extérieur** (voir l'option '5' ci-dessous).

Le point de départ est un tableau (le **Dictionnaire du Corpus**) avec toutes les correspondances forme/lemme, les occurrences respectives dans le corpus et quelques étiquettes qui se rapportent à la **lemmatisation automatique** (colonne **INF**).



Avant chaque intervention, en sélectionnant une forme spécifique et en utilisant le bouton droit de la souris, vous pouvez vérifier les **concordances** (Key-Word-in-Context) qui vous intéressent.

En tout cas, avant toute intervention, après avoir cliqué sur le "sélection des mots clés", doit être activé la configuation personnalisée (voir ci-dessus l'option «1»).

Les **interventions possibles**, quoique différentes dans leurs buts (révision des lemmatisations et/ou applications des grilles pour l'analyse du contenu), donnent toutes une réorganisation de la base de données de **T-LAB**, c.-à-d. créent différents tableaux pour les analyses suivantes. Par conséquent toutes les opérations doivent être faites seulement sur les mots (lemmes ou

catégories) considérés intéressants pour les analyses suivantes. **T-LAB**, en fait, rend disponible une autre option, **Configuration Personnalisée** (voir **Sélection des Mots-Clés**), avec laquelle l'utilisateur peut décider quels lemmes retenir et lesquels abandonner (exclus de l'analyse).

Les deux fonctions (**Personnalisation du dictionnaire** et **Configuration Personnalisée**) sont étroitement reliées entre elles et l'utilisateur peut facilement se déplacer de l'une à l'autre afin de changer ses choix.

En **Personnalisation du Dictionnaire,** pour changer les labels (ou “lemmes”) attribués aux mots, deux modalités d'intervention sont prévues :

- une qui prévoit la possibilité de déplacer des mots choisis dans la boîte du côté droit et, après, de les renommer avec l'option "remplacer" (N.B.: dans ce cas, la nouvelle étiquette peut être définie à l'aide d'un des lemmes sélectionnéou en tapant dans le champ "étiquette");
- une qui prévoit l'importation d'un dictionnaire personnalisé, réservée aux utilisateurs experts qui disposent de leurs propres listes pour classifier les mots présents en un ou plus corpus.

N.B.: L’ utilisation du bouton droit dans la case Renommer / Regrouper habilite un menu contextuel qui permet trois opérations: a) vérifier les concordances (KeyWord-in-Context) de l'élément sélectionné; b) supprimer de la case l’élément sélectionné ; c) supprimer de la case tous les éléments sélectionnés.

Pour **importer un dictionnaire personnalisé** il faut que l'utilisateur ait préparé un fichier **Dictio.diz** ou bien un fichier **Dizionary.diz** qui peuvent se composer par "n" lignes, chacune avec un couple de chaînes séparées par le caractère ";".
La longueur maximum d'une chaîne (mot, lemme ou catégorie) est de 50 caractères: aucun espace vide ou apostrophe ne doit être inclus.
Pour chaque couple, la première chaîne - du côté gauche - indique l'étiquette (lemme ou catégorie) défini par l'utilisateur, la seconde indique le mot (cas **Dictio.diz**) ou le lemme (cas **Dictionary.diz**) correspondant déjà présent dans le dictionnaire de **T-LAB**.

Voici quelques exemples:

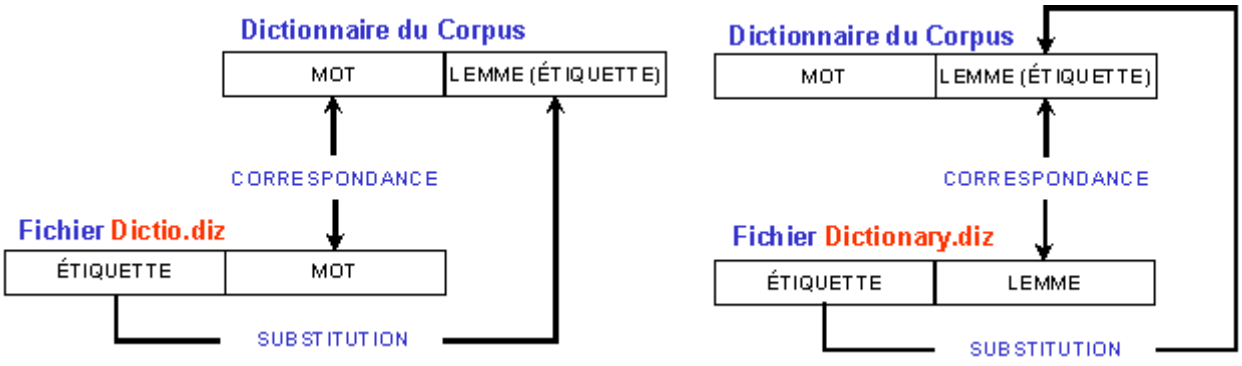
(Fichier **Dictio.diz**)
EXAMINER;examina
EXAMINER;examinai
EXAMINER;examinaient
EXAMINER;examinais

BEAU;beau
BEAU;beaux
BEAU;belle
BEAU;belles

(Fichier **Dictionary.diz**)
ACCUEIL;accueil
ACCUEIL;accueillir
ACCUEIL;accueillant

PENSÉE_ABSTRAITE;conceptualiser
PENSÉE_ABSTRAITE;analyse
PENSÉE_ABSTRAITE;analyser
PENSÉE_ABSTRAITE;interpréter

Selon le type de fichier que vous importez, les changements seron comme suit:



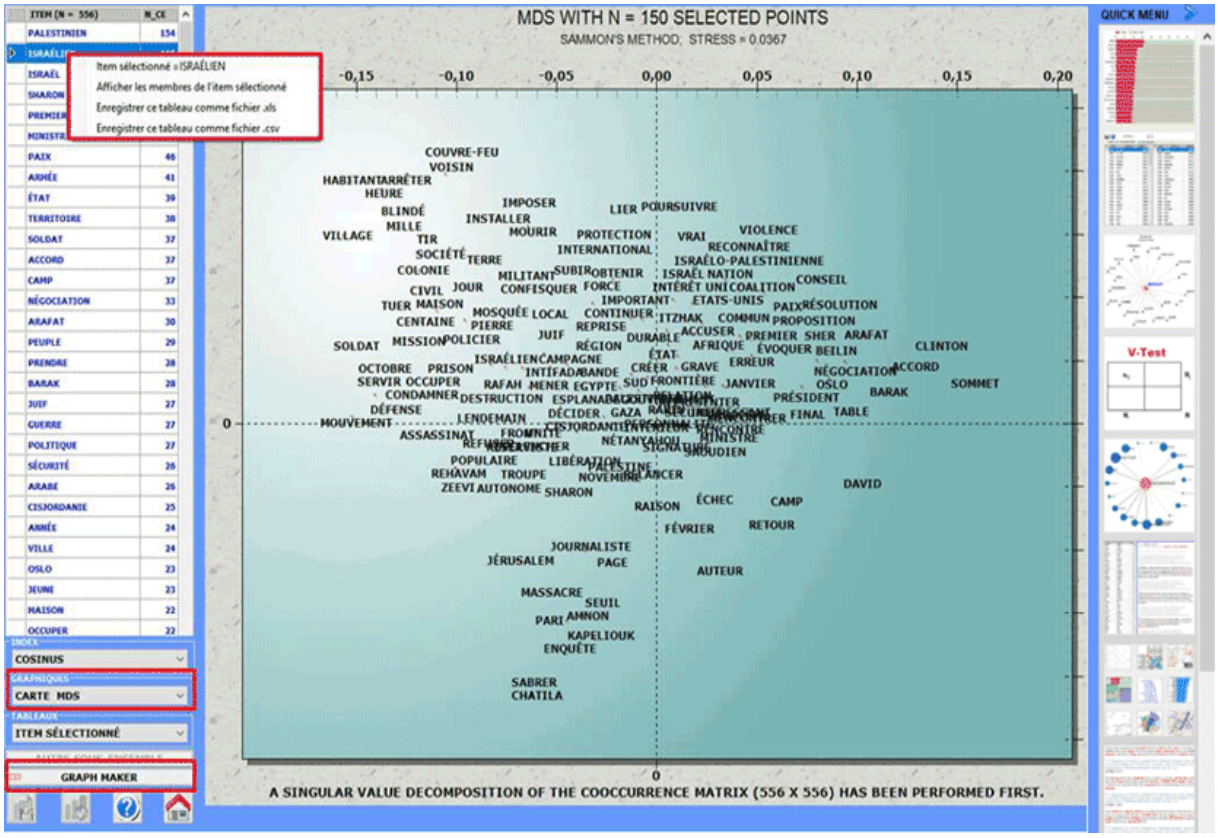
N.B. :

- en utilisant l'option **corpus lemmatisé** il est possible d'exporter une copie du corpus (fichier .txt) dans la quelle chaque mot sera remplacé par le lemme correspondant ;
- quand l'utilisateur modifie le dictionnaire d'un corpus, il ne peut plus utiliser l'option configuration automatique pour analyser le même corpus.

ANALYSES DES CO-OCCURRENCES

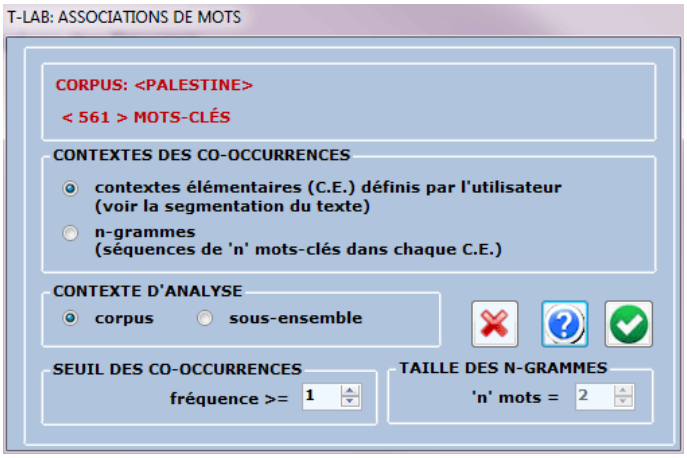
Associations de Mots

N.B. Les images de cette section font référence à une version précédente de T-LAB. En **T-LAB 10**, l'aspect est légèrement différent. En outre : a) Il y a une nouvelle option qui permet à l'utilisateur de tracer une **Map Overview** avec les mots les plus pertinents; b) un nouveau bouton (GRAPH MAKER) permet à l'utilisateur de créer et d'exporter plusieurs graphiques dynamiques au format HTML; c) le **bouton droit** sur les tableaux avec les mots-clés rend disponibles des options supplémentaires; d) une galerie d'images à accès rapide qui fonctionne comme un menu supplémentaire permet de basculer entre les différentes sorties en un seul clic.
Certaines de ces nouvelles fonctionnalités sont mises en évidence dans l'image ci-dessous.



Cet outil **T-LAB** nous permet de vérifier comment les relations de **co-occurrence** et de similarité qui, à l'intérieur du corpus ou d'un de son sous-ensemble, déterminent le sens local de mots clé sélectionnés par l'utilisateur.

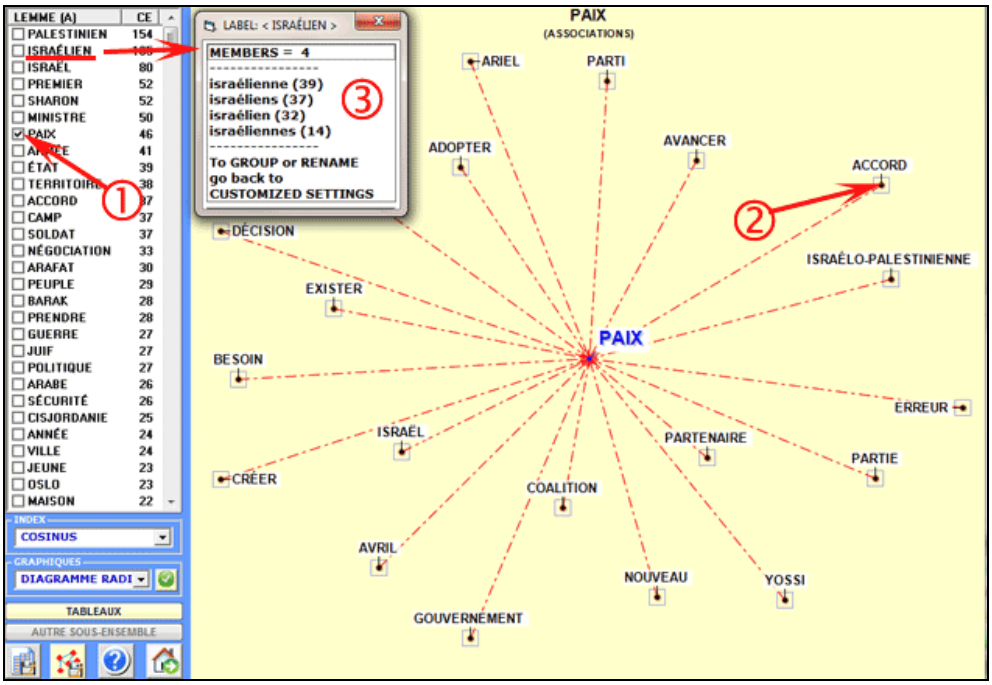
Cette vérification peut être faite au moyen d' options prédéfinies (A) ou à travers des options sélectionnées par l'utilisateur (B).



Dans le premier cas (**A**: options prédéfinies) les **cooccurrences** des mots sont calculées à l' intérieur des **contextes élémentaires** sélectionnés en phase d'importation du corpus (ex. phrases, fragments, paragraphes, etc.); différemment, dans le second cas (**B**: options sélectionnées par l'utilisateur) les cooccurrences peuvent aussi être calculées à l' intérieur de séquences de mots de longueur variable (c'est-à-dire **n-grammes**, voir section du glossaire correspondante) et il est aussi possible de décider le seuil minimum (c'est-à-dire la fréquence) des cooccurrences à considérer.

La fenêtre de travail (voir ci-dessous) devient disponible tout de suite après avoir effectué le calcul des cooccurrences entre tous les mots inclus dans la liste sélectionnée par l'utilisateur.

À gauche de cette fenêtre il y a un tableau avec la liste des mots et les valeurs numériques qui indiquent la quantité de contextes élémentaires ou de n-grammes dans lesquels chaque mot résulte présent.



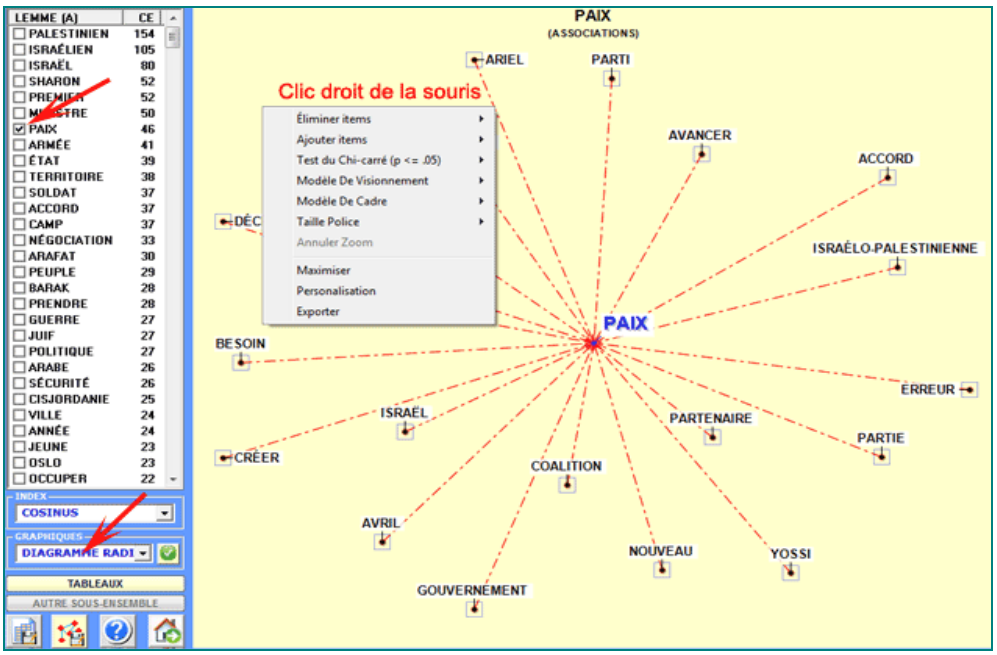
Un simple clic sur les mots du tableau (option "1") ou sur les points des graphiques (option "2") permet de vérifier les associations relatives à chaque mot cible. Différemment, un click

sur les labels inclus dans le tableau (option “3”) permet de vérifier les items inclus dans chaque lemme.

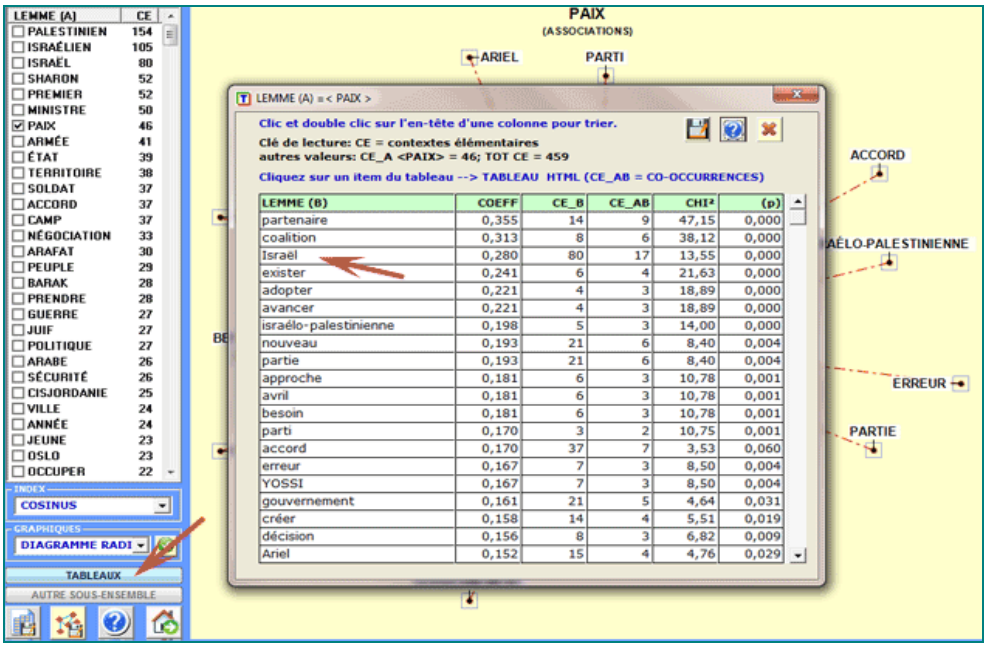
De fois en fois, la sélection des mots associés est effectuée à travers le calcul d’un **Index d’Association** (voir section correspondante du glossaire) ou par un index de ressemblance du deuxième ordre (voir explication à la fin de cette section). Dans le premier cas les index disponibles sont six (Cosinus, Dice, Jaccard, Équivalence, Inclusion et Informartion Mutuelle) et leur calcul est plutôt rapide ; différemment, dans le cas des index du deuxième ordre – et surtout lorsque le corpus est de dimensions considérables – l’analyse des données peut demander plusieurs minutes. En outre, il faut tenir compte du fait que, dans le cas des index du deuxième ordre, les résultats sont aussi plus fiables que plus nombreux sont les mots inclus dans la liste.

Pour chaque interrogation, **T-LAB** produit graphiques et tableaux.
Soit les graphiques soit les tableaux peuvent être sauvés utilisant les boutons appropriés.

Dans le **diagramme radial**, le lemme choisi est placé au centre. Les autres sont distribués autour de lui, chacun à la distance proportionnelle à son degré d'association. Les rapports significatifs sont donc du type "un à un" entre le lemme central et chacun des autres.
Chaque clic sur un item produit un nouveau diagramme et, en employant le bouton droit de la souris, il est possible d'ouvrir une fenêtre de dialogue qui permet plusieurs personnalisations des graphiques.



Les **tableaux** contiennent des données qui permettent de vérifier les relations entre occurrences et cooccurrences des mots (Max. 50) qui résultent les plus associés à celui sélectionné.

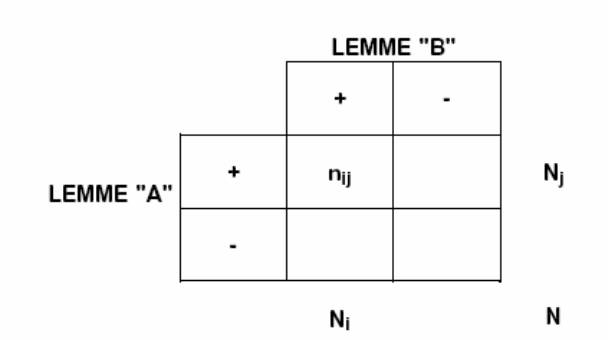


LEMME (B)	COEFF	CE_B	CE_AB	CHI²	(p)
partenaire	0,355	14	9	47,15	0,000
coalition	0,313	8	6	38,12	0,000
Israël	0,280	80	17	13,55	0,000
exister	0,241	6	4	21,63	0,000
adopter	0,221	4	3	18,89	0,000
avancer	0,221	4	3	18,89	0,000
israélo-palestinienne	0,198	5	3	14,00	0,000
nouveau	0,193	21	6	8,40	0,004
partie	0,193	21	6	8,40	0,004
approche	0,181	6	3	10,78	0,001
avril	0,181	6	3	10,78	0,001
besoin	0,181	6	3	10,78	0,001
parti	0,170	3	2	10,75	0,001
accord	0,170	37	7	3,53	0,060
erreur	0,167	7	3	8,50	0,004
YOSSI	0,167	7	3	8,50	0,004
gouvernement	0,161	21	5	4,64	0,031
créer	0,158	14	4	5,51	0,019
décision	0,156	8	3	6,82	0,009
Ariel	0,152	15	4	4,76	0,029

Les clés de lecture sont les suivantes:

- **LEMME (A)** = lemme sélectionné;
- **LEMME (B)** = lemmes associés avec le LEMME (A);
- **COEFF** = valeur de l'index d'association sélectionné;
- **TOT CE** = total des contextes élémentaires (CE) ou des n-grammes analysés;
- **CE_A** = total des CE dans lesquels le lemme sélectionné (A) est présent;
- **CE_B** = total des CE dans lesquels chaque lemme associé (B) est présent;
- **CE_AB** = total des CE dans lesquels les lemmes "A" e "B" sont associés (co-occurrences);
- **CHI2** = valeur du Chi Deux (signification des co-occurrences) ;
- **(p) =** probabilité associée à la valeur du chi-deux (def=1).

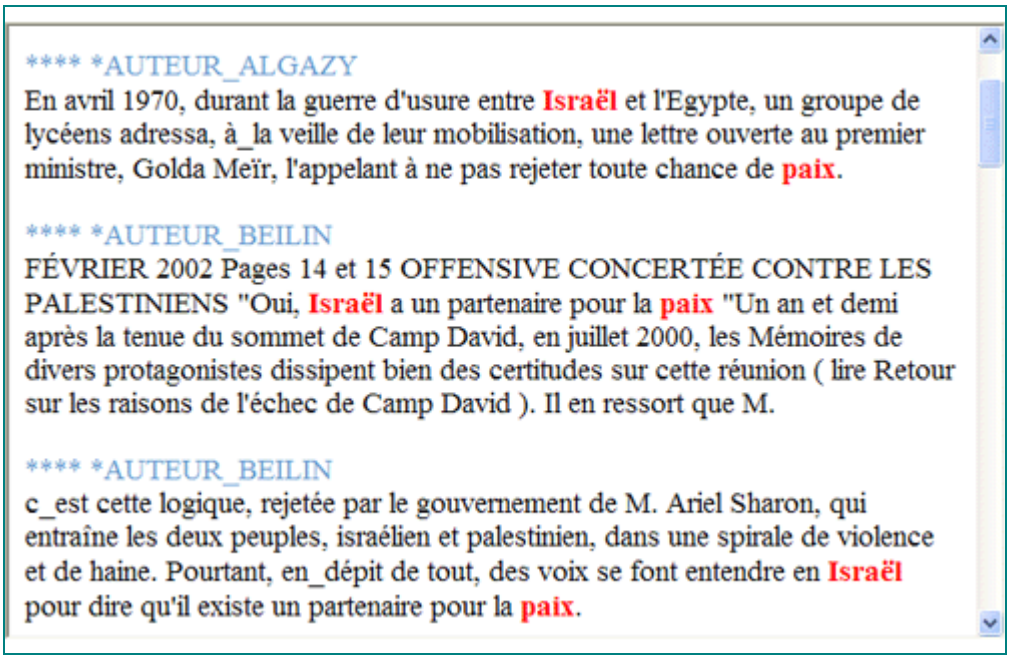
Dans le cas du **Chi Deux**, pour chaque couple de lemmes ("A" e B") la structure du tableau analysé est la suivante.



Avec: nij = CE_AB; Nj = CE_A; Ni = CE_B; N = TOT CE.

Un clic sur chaque étiquette (par exemple "Israël") permet de sauvegarder un fichier avec tous

les contextes élémentaires (c.-à-d. phrases ou paragraphes) où il est en couple avec le mot choisi (co-occurrences de "paix" et "Israël").

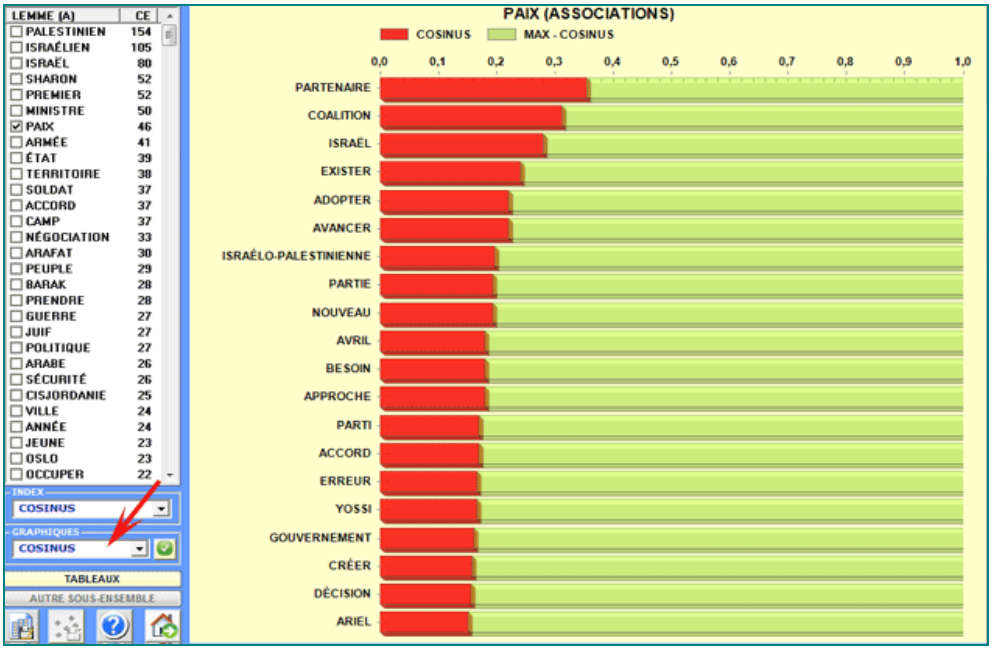


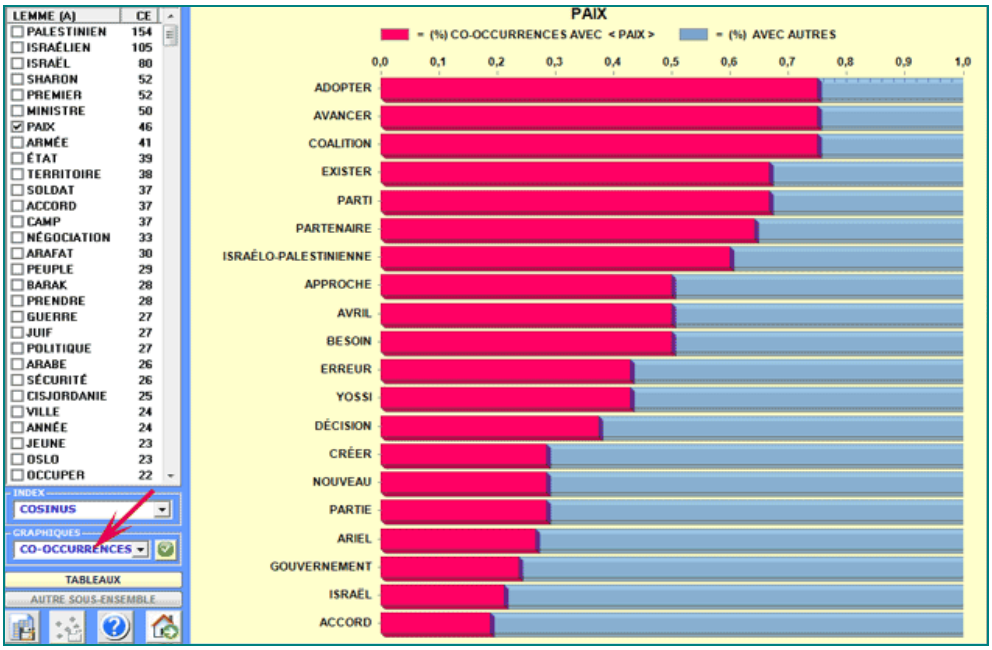
**** *AUTEUR_ALGAZY
En avril 1970, durant la guerre d'usure entre **Israël** et l'Egypte, un groupe de lycéens adressa, à_la veille de leur mobilisation, une lettre ouverte au premier ministre, Golda Meïr, l'appelant à ne pas rejeter toute chance de **paix**.

**** *AUTEUR_BEILIN
FÉVRIER 2002 Pages 14 et 15 OFFENSIVE CONCERTÉE CONTRE LES PALESTINIENS "Oui, **Israël** a un partenaire pour la **paix** "Un an et demi après la tenue du sommet de Camp David, en juillet 2000, les Mémoires de divers protagonistes dissipent bien des certitudes sur cette réunion (lire Retour sur les raisons de l'échec de Camp David). Il en ressort que M.

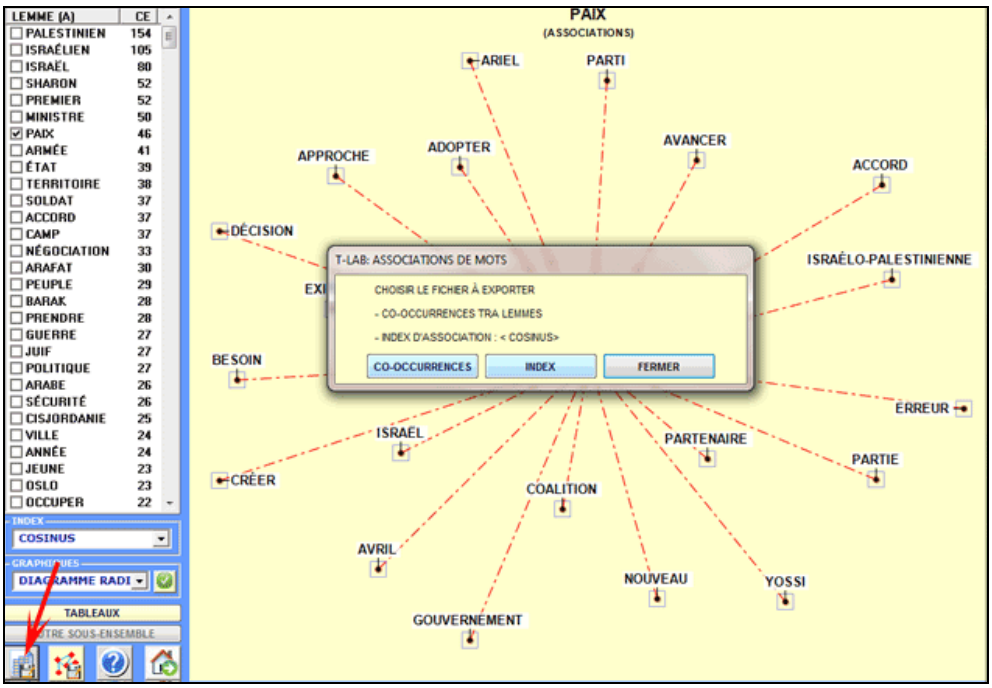
**** *AUTEUR_BEILIN
c_est cette logique, rejetée par le gouvernement de M. Ariel Sharon, qui entraîne les deux peuples, israélien et palestinien, dans une spirale de violence et de haine. Pourtant, en_dépit de tout, des voix se font entendre en **Israël** pour dire qu'il existe un partenaire pour la **paix**.

D'autres graphiques (Histogrammes) nous permettent d'apprécier les valeurs du **coefficient** utilisé et les **pourcentages** des co-occurrences.



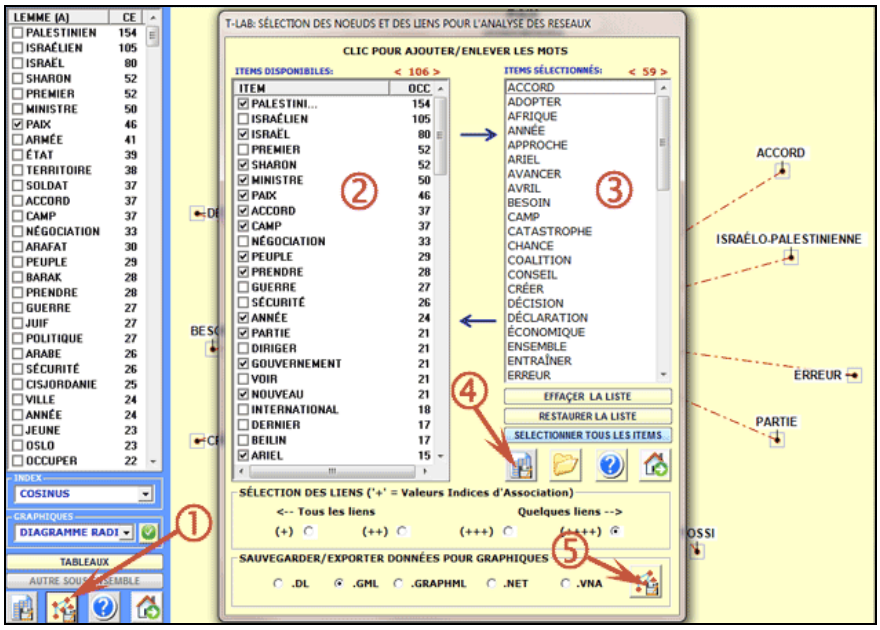


En cliquant sur le bouton en bas à gauche, l'utilisateur peut **exporter différents types de tableaux** (voir l'image ci-dessous).

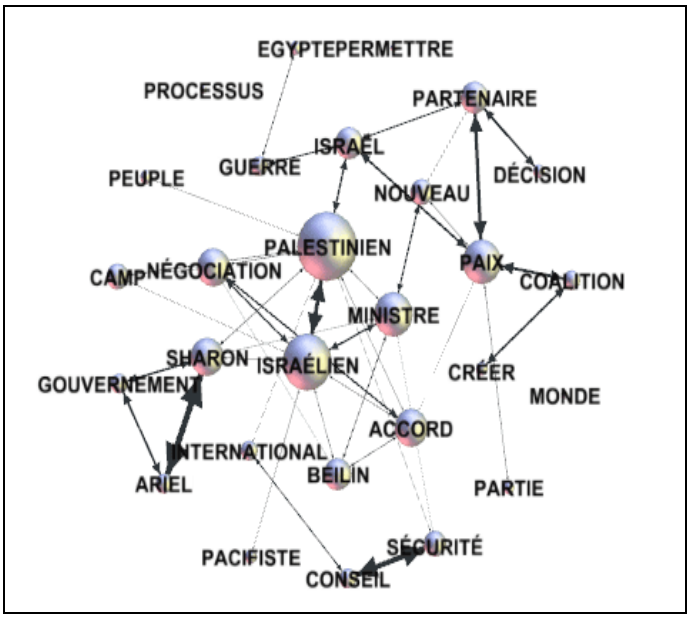


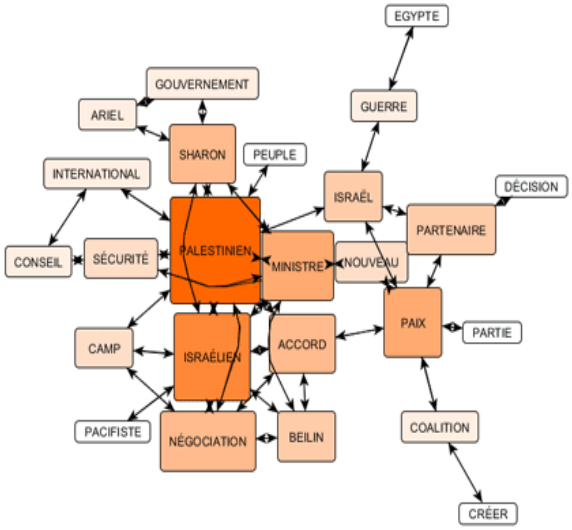
Une autre fenêtre **T-LAB** (voir image suivante, étape 1) permet de créer des fichiers graphiques qui peuvent être édités avec un logiciel pour le network analysis comme Gephi, Pajek, Ucinet, yEd et d'autres encore. Dans ce cas, les **nœuds** du réseau sont constitués par les mots associés au mot cible. Les options disponibles sont les suivantes : sélectionner les items (c'est-à-dire les " nœuds") à insérer dans les graphiques (voir ci-dessous, étapes 3 et 4), exporter le type de fichier graphique sélectionné (voir ci-dessous, étape 5).

N.B.: En **T-LAB 10** la fenêtre suivante a été remplacée par l'outil **GRAPH MAKER**.



Par exemple, des fichiers .gml exportés par **T-LAB** peuvent permettre de réaliser des graphiques comme les suivantes.





N.B.: Le premier graphique a été créé au moyen de (https://gephi.org/), le second au moyen de yEd (http://www.yworks.com/en/products_yed_download.html/), deux logiciels disponibles en téléchargement gratuit.

Les modalités de calcul des différents index d'association (ou proximité) sont illustrées dans la section correspondante du Manuel/Aide (voir glossaire). Comme on pourra vérifier, tous ces index sont obtenus à travers une normalisation des valeurs de cooccurrence qui concernent des couples de mots; donc – dans les calculs du **premier ordre** – deux mots jamais co-occurrents ont un index d'association égal à "0". Différemment, les index du **deuxième ordre** soulignent des phénomènes de "similarité" concernant l'usage (et donc le sens) des mots qui ne dépendent pas directement de leurs cooccurrences; en effet, dans ce cas, deux mots jamais co-occurrents peuvent avoir un index d'association même très élevé.

En utilisant certains concepts de la linguistique structurelle, nous pouvons affirmer que, pendant que les index du "premier ordre" relèvent des phénomènes qui concernent l' axe syntagmatique (combinaison et proximité "in praesentia", c'est-à-dire des mots "l'un à coté de l'autre" dans une phrase spécifique) les index du "deuxième ordre" relèvent des phénomènes qui concernent l'axe paradigmatique (association et similarité "in absentia", c'est-à-dire des relations de quasi synonymie entre deux ou plusieurs termes utilisés par le même auteur.

Pour comprendre la façon dont **T-LAB** calcule les index du "deuxième ordre", il est utile de rappeler que les index du "premier ordre" peuvent être utilisés pour construire des matrices de proximité comme la suivante (**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

Matrice "A" : similarité du premier ordre.

Dans cette matrice symétrique (A) la valeur 0.373 (en jaune) correspond à l'index le plus élevé du "premier ordre" et il indique l'association entre les mots "w_03" et "w_10". Plus précisément, il s'agit d'un index d'équivalence obtenu en divisant le carré de leurs cooccurrences par le produit de leurs occurrences (360^2/267*553).

À partir de la matrice ci-dessus (A), **T-LAB** construit une deuxième matrice (B) obtenue en calculant les cosinus résultants de la comparaison de toutes les colonnes qui contiennent les index du premier ordre (voir matrice 'A'). Comme on peut vérifier, dans le tableau 'B' suivant , la valeur de "similarité" plus élevée concerne la relation entre les mots "w_06" et "w_08". Ceci signifie que les vecteurs respectifs (voir les deux colonnes soulignées en vert dans la matrice 'A') résultent être entre eux très semblables (cosinus =0.905), même si l'association du "premier ordre" entre les deux mots en question résulte plutôt basse (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

Matrice "B" : similarité du deuxième ordre.

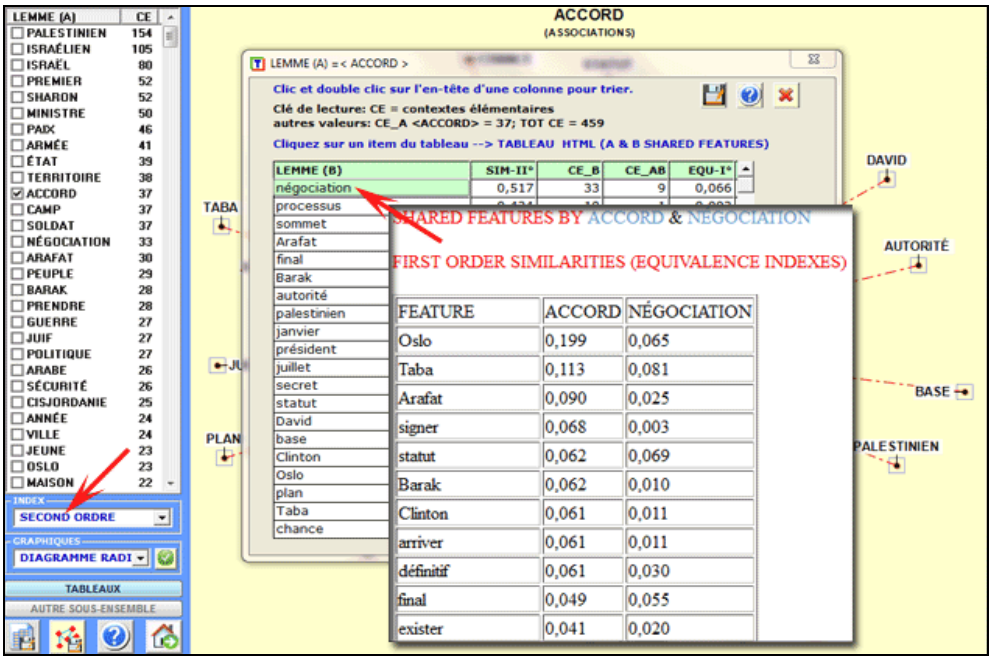
Autrement dit, un index du "premier ordre" est obtenu en appliquant une formule qui inclut des valeurs de cooccurrence et occurrence, pendant qu'un index du "deuxième ordre" est obtenu en multipliant deux vecteurs normalisés.

Au-delà des modalités de calcul, il faut souligner le fait que dans les deux cas ("A" et "B") deux différents phénomènes sont relevés. Dans le premier cas ("A"), en effet, le focus est sur les cooccurrences; différemment, dans le second cas ("B") - et indépendamment de leurs cooccurrences - le focus est sur les ressemblances entre "profils" dont les données font référence à l'usage des mots de la part des auteurs des textes analysés.

Juste pour faire un exemple, dans l'analyse de Pinocchio du premier ordre le terme "fée" résulte généralement associé (voir cooccurrences) avec "gentille" et "cheveux bleus"; différemment, dans l'analyse du second ordre, le terme qui résulte le plus semblable à "fée" est "maman" , même si les cooccurrences entre ces deux termes ("fée" et "maman") sont - à l'intérieur du conte de fées de Collodi - presque insignifiantes (c'est-à-dire seulement 3).

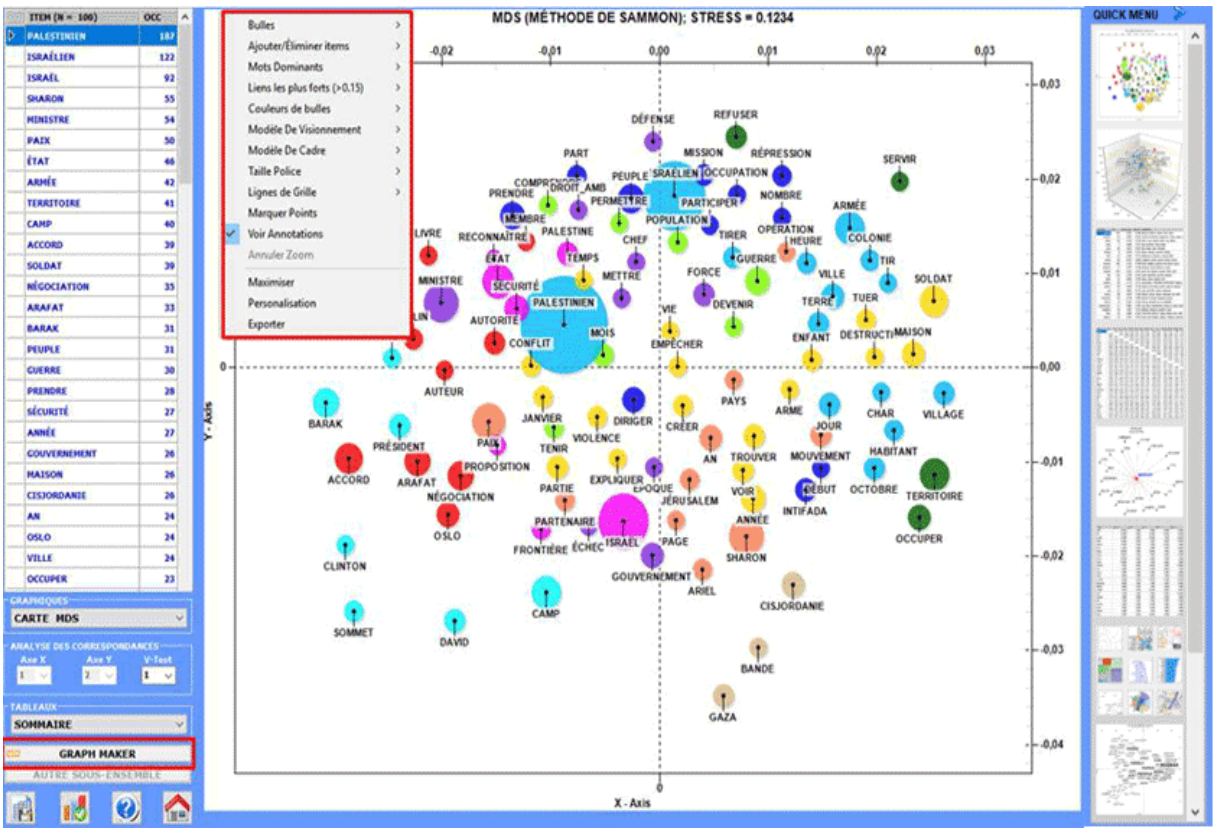
Les tableaux visualisés par **T-LAB** permettent de vérifier soit les similarités du deuxième ordre (voir sous colonne SIM-II°) soit les index du premier ordre (EQU-I°, c'est-à-dire index d'équivalence).

En outre, en cliquant sur chaque item de ce tableau, il est possible de visualiser des fichiers HTML qui permettent de vérifier quelles caractéristiques ("features") déterminent la ressemblance entre chaque couple de mots. Par exemple, le tableau suivant montre que la similarité du deuxième ordre entre "accord" et "négociation" est en premier lieu déterminée par des caractéristiques partagées telles que "Oslo", "Taba", "Arafat", etc.


ACCORD
(ASSOCIATIONS)
LEMME (A) =< ACCORD >
Clic et double clic sur l'en-tête d'une colonne pour trier.
Clé de lecture: CE = contextes élémentaires
autres valeurs: CE_A <ACCORD> = 37; TOT CE = 459
Cliquez sur un item du tableau --> TABLEAU HTML (A & B SHARED FEATURES)
LEMME (B) SIM-II° CE_B CE_AB EQU-I°
négociation 0,517 33 9 0,066
processus
sommet
Arafat
final
Barak
autorité
palestinien
janvier
président
juillet
secret
statut
David
base
Clinton
Oslo
plan
Taba
chance
SHARED FEATURES BY ACCORD & NEGOCIATION
FIRST ORDER SIMILARITIES (EQUIVALENCE INDEXES)
FEATURE ACCORD NÉGOCIATION
Oslo 0,199 0,065
Taba 0,113 0,081
Arafat 0,090 0,025
signer 0,068 0,003
statut 0,062 0,069
Barak 0,062 0,010
Clinton 0,061 0,011
arriver 0,061 0,011
définitif 0,061 0,030
final 0,049 0,055
exister 0,041 0,020
LEMME (A) CE
PALESTINIEN 154
ISRAÉLIEN 105
ISRAËL 80
PREMIER 52
SHARON 52
MINISTRE 50
PAIX 46
ARMÉE 41
ÉTAT 39
TERRITOIRE 38
ACCORD 37
CAMP 37
SOLDAT 37
NÉGOCIATION 33
ARAFAT 30
PEUPLE 29
BARAK 28
PRENDRE 28
GUERRE 27
JUIF 27
POLITIQUE 27
ARABE 26
SÉCURITÉ 26
CISJORDANIE 25
ANNÉE 24
VILLE 24
JEUNE 23
OSLO 23
MAISON 22
INDEX
SECOND ORDRE
GRAPHIQUES
DIAGRAMME RADI
TABLEAUX
AUTRE SOUS-ENSEMBLE
TABA
PLAN
DAVID
AUTORITÉ
BASE
PALESTINIEN

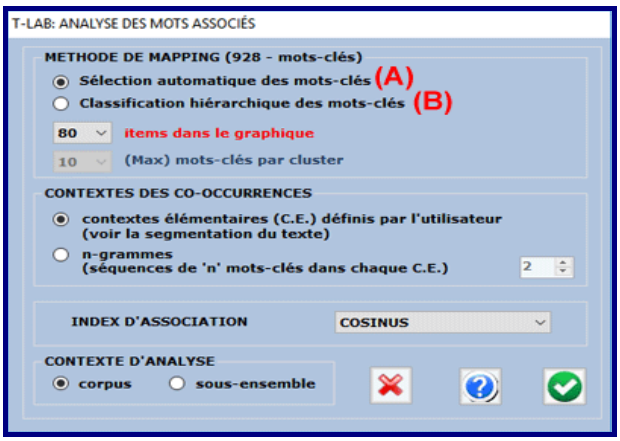
Analyse des Mots Associés et Cartes Conceptuelles

N.B.: Les images de cette section font référence à une version précédente de T-LAB. En **T-LAB 10**, l'aspect est légèrement différent. En outre: a) quand la 'sélection automatique des mots-clés' est sélectionnée, dans la carte **MDS** des différentes couleurs sont utilisées pour indiquer différents clusters d'éléments; b) la technique de visualisation appelée t-SNE (t-Distributed Stochastic Neighbor Embedding) a été ajoutée; c) un nouveau bouton (**Graph Maker**) qui permet à l'utilisateur de créer plusieurs graphiques dynamiques en format HTML est disponible; d) Le **bouton droit** sur les graphiques ou sur les tableaux avec les mots-clés rend disponible certaines options additionnelles; e) une galerie d'images à accès rapide qui fonctionne comme un menu supplémentaire permet de basculer entre les différentes sorties en un seul clic.
Certaines de ces nouvelles fonctionnalités sont mises en évidence dans l'image ci-dessous.



Cet outil **T-LAB** nous permet d'analyser deux types de relations concernant les **cooccurrences** des mots:

A - entre les **mots-clés** (lemmes ou catégories) sélectionnés, si leur quantité n'excède pas 500 éléments (minimum 10) ;
B - entre (et à l'intérieur de) **clusters** (c.-à-d. **Noyaux Thématiques**), si la quantité des **mots-clés** sélectionnés excède 100 éléments (maximum 3.000).

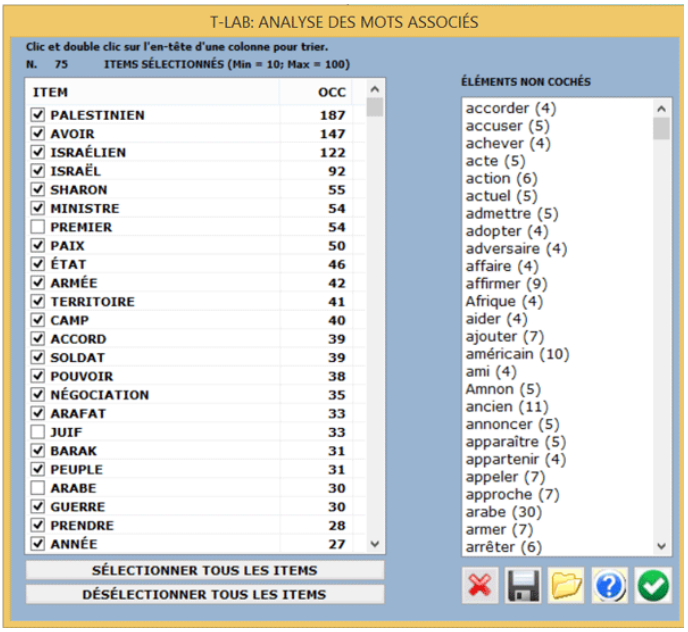


L'utilisateur peut choisir l'**index d'association** à employer et, seulement pour l'option B, aussi bien la quantité maximum de clusters à obtenir (de 50 a 100) que la quantité maximum de mots-clés par cluster.

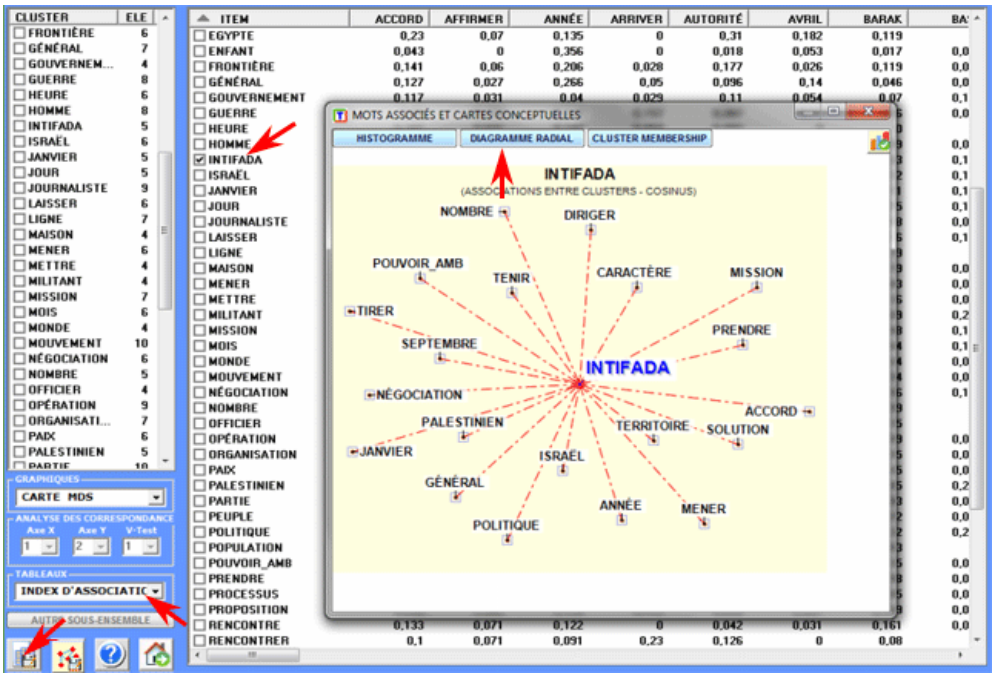
Le processus de calcul inclut les étapes suivantes:

1- construction d'une matrice des cooccurrences (mot x mot);
2- calcul des index d'association sélectionnés (Cosinus, Dice, Jaccard, Equivalence, Inclusion, Information Mutuelle)
3- classification hiérarchique;
4- construction d'une deuxième matrice des cooccurrences (cluster x cluster);
5- représentation de graphique par Muldimensional Scaling et Analyse de Correspondances.

N.B:
- dans le cas «A» (voir ci-dessus), l'utilisateur peut revoir et personnaliser la sélection des mots-clés (voir l'image suivante) et **T-LAB** n'effectue pas les passages 3 et 4;

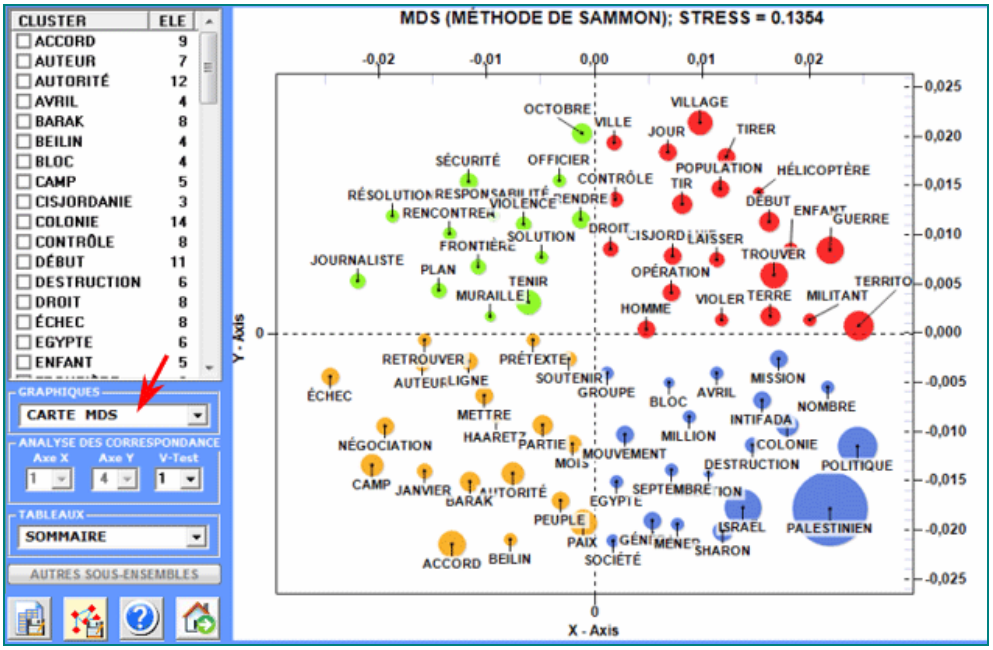


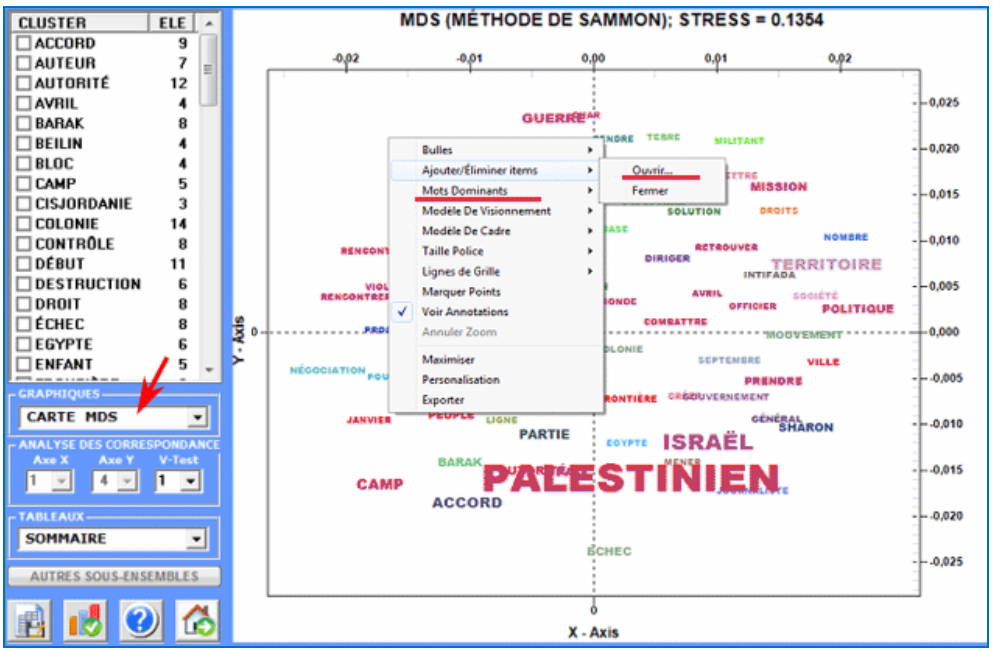
- la qualité des résultats dépend d'un bon **choix des mots clés**;
- puisque les lexies (multi-words) non classifiées par T-LAB sont des cas spécifiques de cooccurrence et l'option "B" les traite comme des petits faisceaux (ex. "Twin" + "Towers"), il est recommandé de résoudre ces cas pendant la **phase de prétraitement**. Dans tous les cas, sans répéter l'importation de corpus, il est possible de faire des changements au moyen de l'outil **Personnalisation du Dictionnaire** (par exemple en assignant l'étiquette "Twin_Towers" aux deux items différents "Twin" + "Towers") ;
- tous les tableaux de données peuvent être vérifiés en cliquant sur les boutons appropriés (voir ci-dessous).



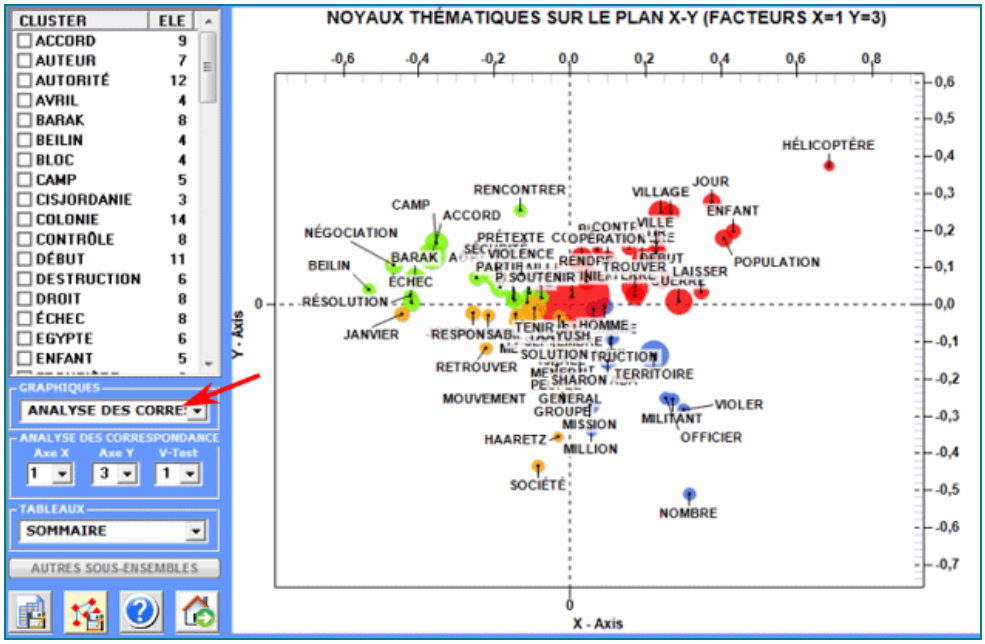
Après l'analyse automatique, l'utilisateur peut visualiser trois types de diagrammes et - en employant le bouton droit de la souris - ouvrir une fenêtre de dialogue qui permet plusieurs personnalisations.

1 - Carte MDS

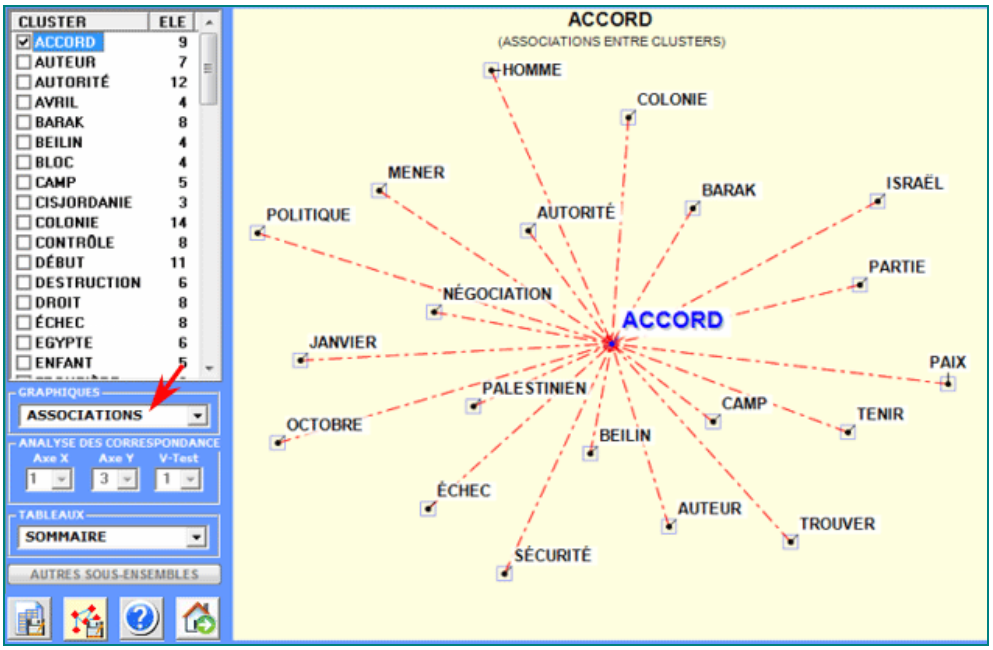

MDS (MÉTHODE DE SAMMON); STRESS = 0.1354
CLUSTER
ELE
ACCORD 9
AUTEUR 7
AUTORITÉ 12
AVRIL 4
BARAK 8
BEILIN 4
BLOC 4
CAMP 5
CISJORDANIE 3
COLONIE 14
CONTRÔLE 8
DÉBUT 11
DESTRUCTION 6
DROIT 8
ÉCHEC 8
EGYPTE 6
ENFANT 5
GRAPHIQUES
CARTE MDS
ANALYSE DES CORRESPONDANCE
Axe X
Axe Y
V-Test
TABLEAUX
SOMMAIRE
AUTRES SOUS-ENSEMBLES
X - Axis
Y - Axis


MDS (MÉTHODE DE SAMMON); STRESS = 0.1354
Bulles
Ajouter/Éliminer items
Mots Dominants
Modèle De Visionnement
Modèle De Cadre
Taille Police
Lignes de Grille
Marquer Points
Voir Annotations
Annuler Zoom
Maximiser
Personalisation
Exporter
Ouvrir...
Fermer
PALESTINIEN
ISRAËL
TERRITOIRE
CAMP
ACCORD
PARTIE
X - Axis
Y - Axis

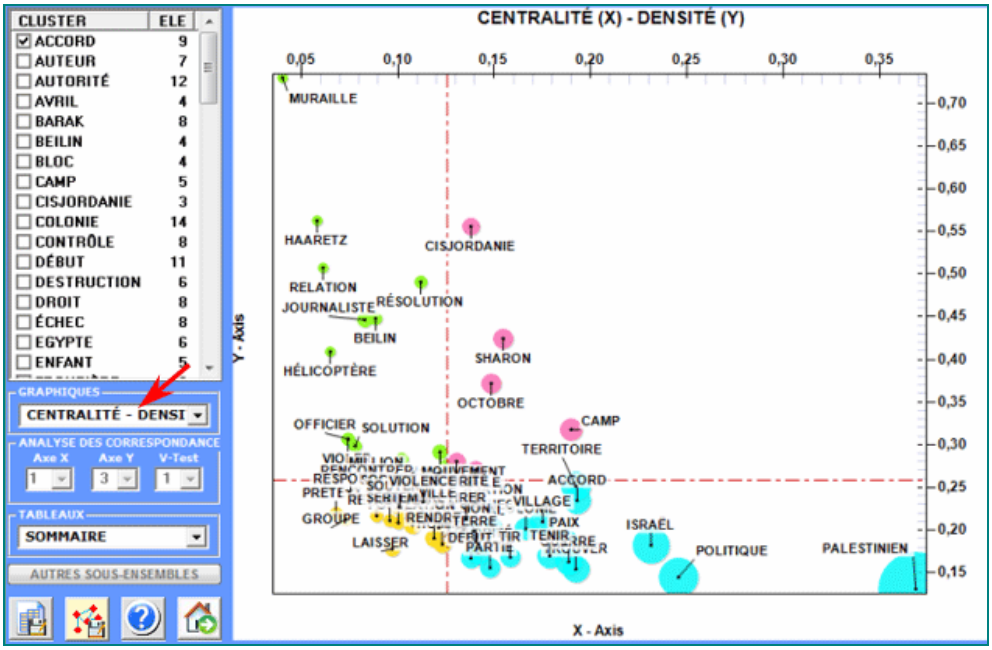
2 - **Analyse Factorielle des Correspondances**



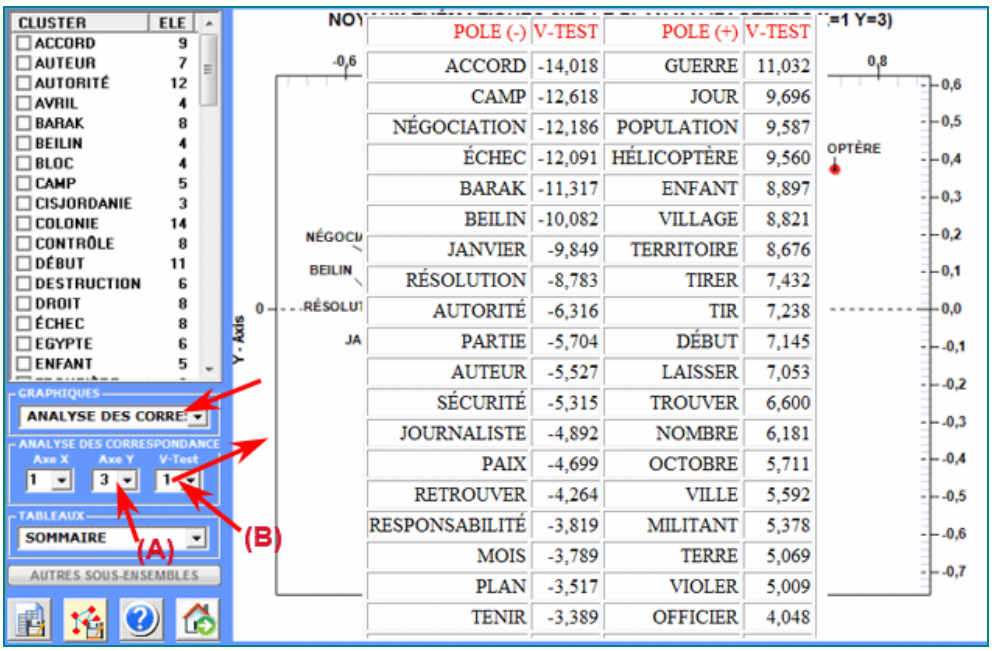
3 - **Diagramme des Associations**



4 - Carte avec les mesures de **Centralité et Densité** (seulement après une cluster analysis)



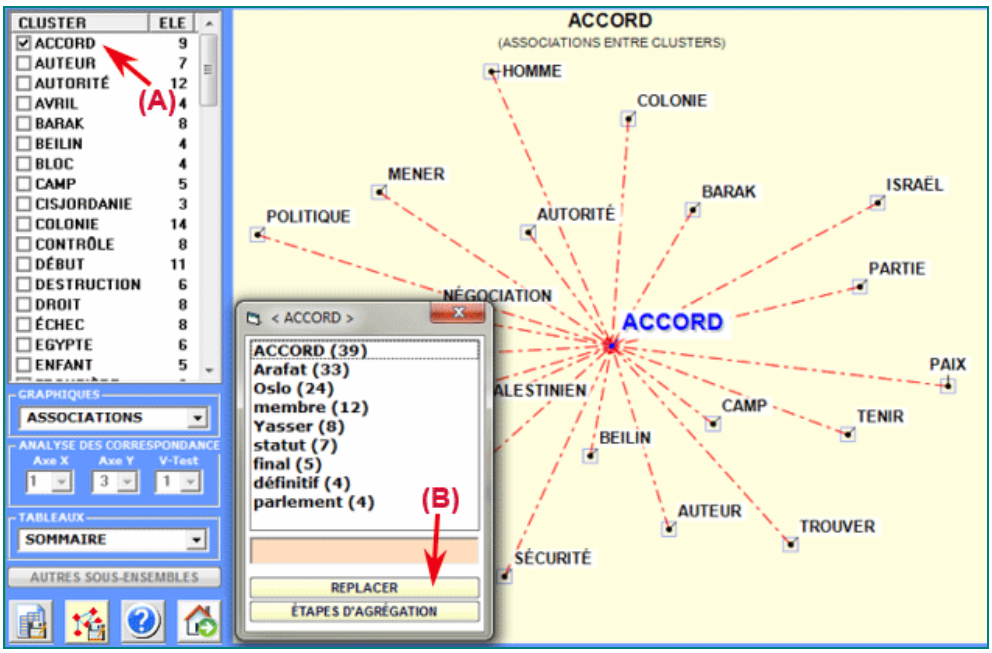
En particulier, les résultats obtenus par l'**Analyse des Correspondances** peuvent être représentés en utilisant les coordonnées des dix premiers axes (voir "A" ci-dessous). Puisque **T-LAB** nous permet de vérifier les **Valeurs Test** de chaque facteur (voir "B" ci-dessous), ce genre de output peut être employé pour une interprétation attentive des rapports entre les clusters et/ou entre les mots-clés.

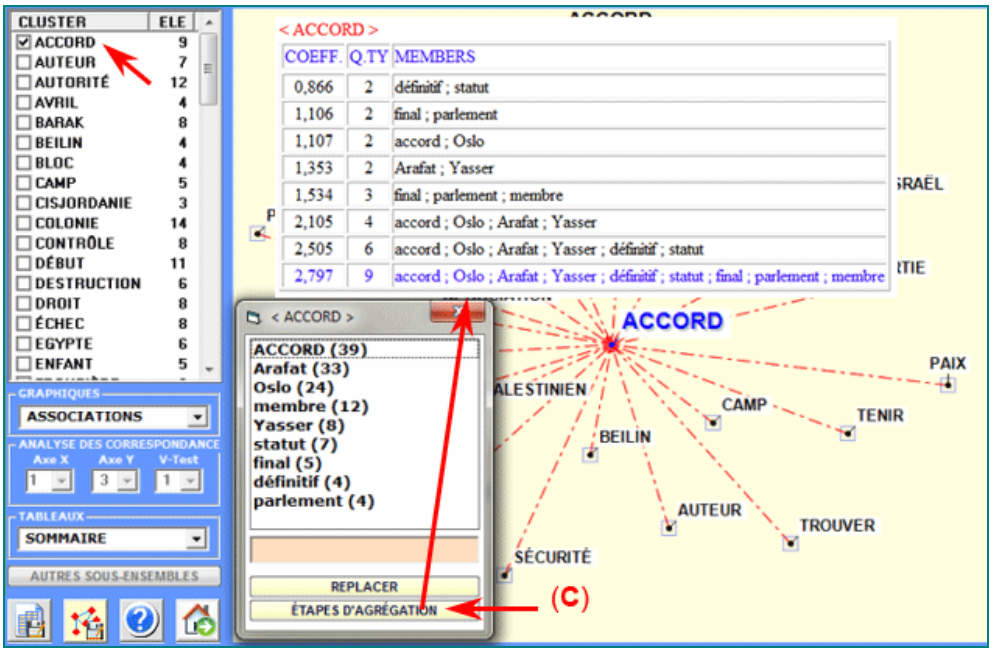


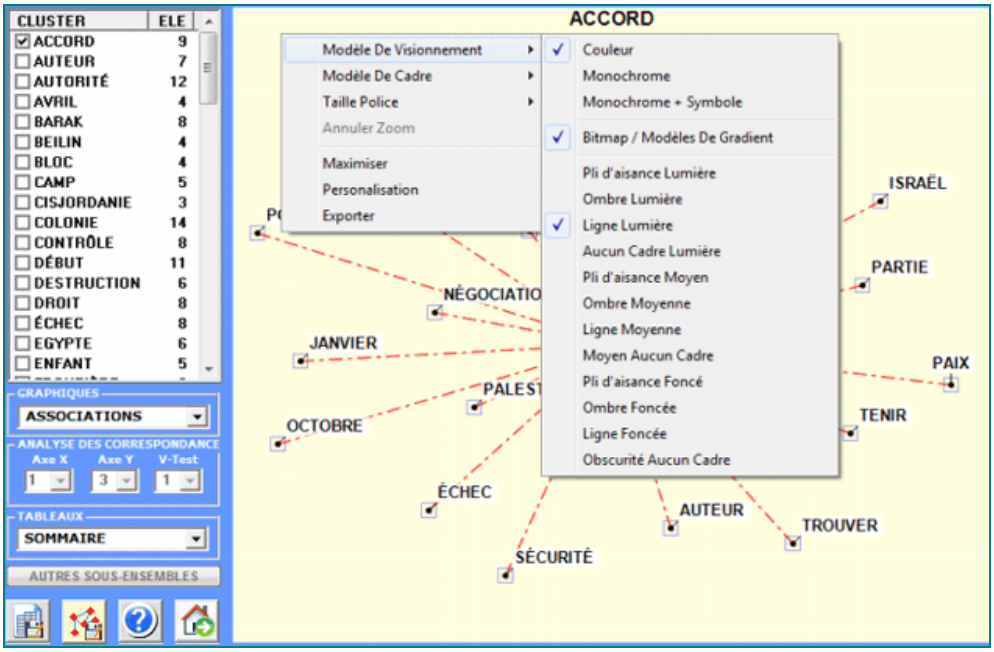
POLE (-)	V-TEST	POLE (+)	V-TEST
ACCORD	-14,018	GUERRE	11,032
CAMP	-12,618	JOUR	9,696
NÉGOCIATION	-12,186	POPULATION	9,587
ÉCHEC	-12,091	HÉLICOPTÈRE	9,560
BARAK	-11,317	ENFANT	8,897
BEILIN	-10,082	VILLAGE	8,821
JANVIER	-9,849	TERRITOIRE	8,676
RÉSOLUTION	-8,783	TIRER	7,432
AUTORITÉ	-6,316	TIR	7,238
PARTIE	-5,704	DÉBUT	7,145
AUTEUR	-5,527	LAISSER	7,053
SÉCURITÉ	-5,315	TROUVER	6,600
JOURNALISTE	-4,892	NOMBRE	6,181
PAIX	-4,699	OCTOBRE	5,711
RETROUVER	-4,264	VILLE	5,592
RESPONSABILITÉ	-3,819	MILITANT	5,378
MOIS	-3,789	TERRE	5,069
PLAN	-3,517	VIOLER	5,009
TENIR	-3,389	OFFICIER	4,048

Les diagrammes peuvent être explorés et personnalisés de manières suivantes:

ACTION	RÉSULTAT
clic sur un item du tableau ou sur un point du graphique	diagramme des associations
double clic sur une étiquette de la colonne "CLUSTER" (voir "A" ci-dessous)	liste avec les éléments du cluster
clic sur le bouton "Replacer" (voir "B" ci-dessous)	nouvelle étiquette assignée au cluster
clic sur le bouton "étapes d'agrégation" (voir "C" ci-dessous)	étapes d'agrégation dans le cluster
bouton droit de la souris	personnalisations des graphiques

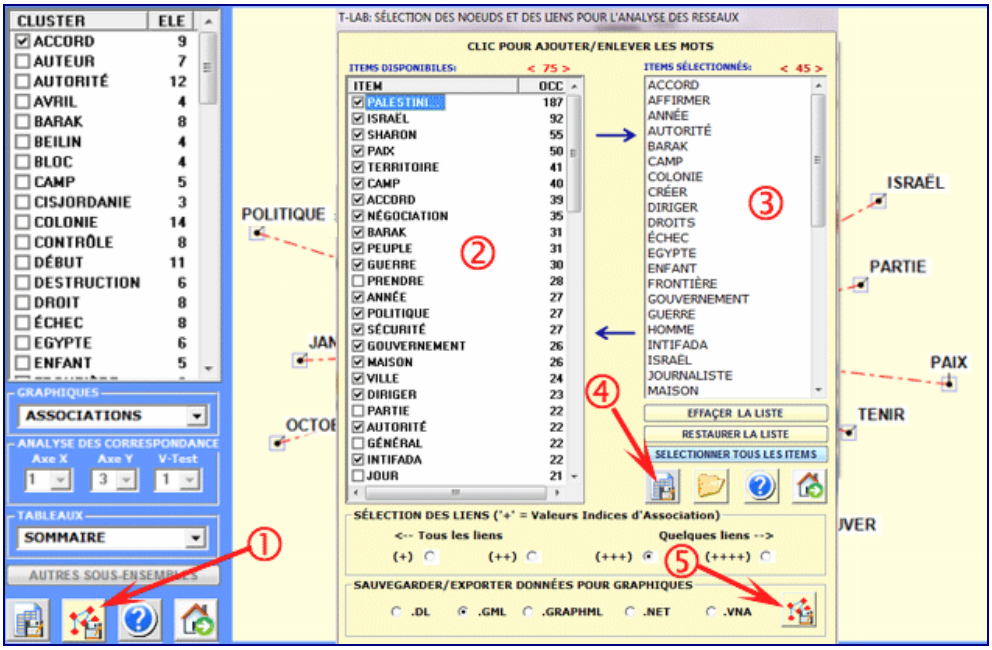




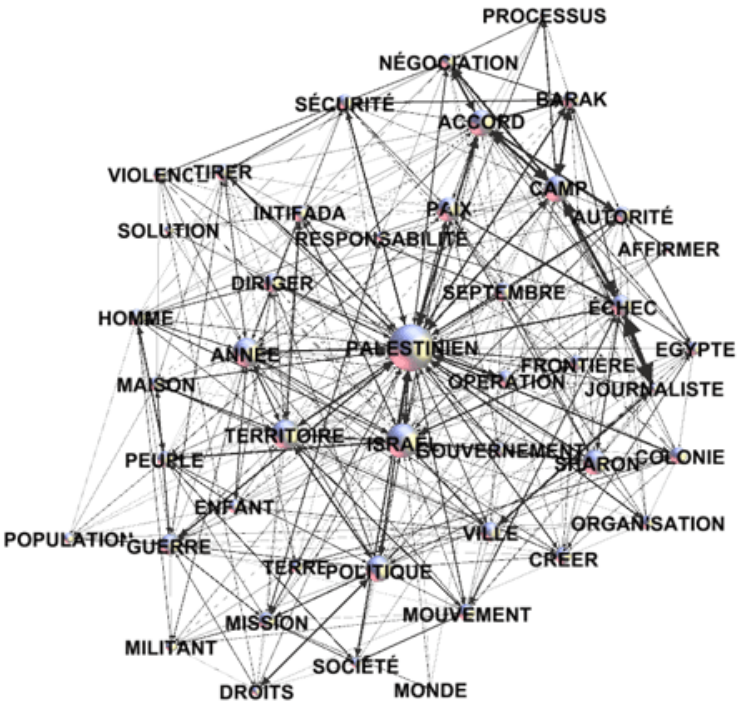


Une autre fenêtre **T-LAB** (voir image suivante, étape 1) permet de créer des fichiers graphiques qui peuvent être édités avec un logiciel pour le **network analysis** tel que Gephi, Pajek, Ucinet, yEd et d'autres. Dans ce cas, les options disponibles sont les suivantes: sélectionner les items (c'est-à-dire les "nœuds") à insérer dans les graphiques (voir ci-dessous, étapes 2 et 3), exporter la matrice correspondante de proximité (voir ci-dessous, étape 4), sélectionner les liens en base à leurs index d'association et exporter le type de fichier choisi (voir ci- dessous, étape 5).

N.B.: En **T-LAB 10** la fenêtre suivante a été remplacée par l'outil **GRAPH MAKER**.

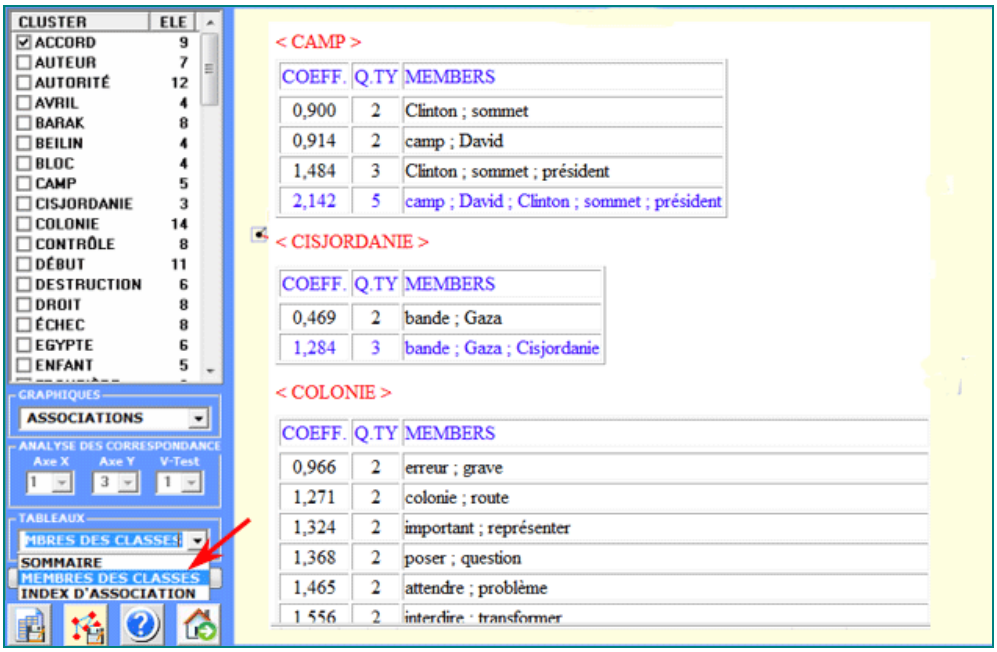


Par exemple, un fichier .gml exporté par **T-LAB** peut permettre de réaliser un graphique comme le suivant.



Les tableaux exportables avec cet instrument T-LAB sont de trois types:

1 - le tableau "**Membres des classes**" (voir ci-dessous) concerne l'agrégation hiérarchique des mots dans chaque cluster;



2 - le tableau "**Sommaire**" (voir ci-dessous) inclut les mesures suivantes:

- ECQ = quantité de contextes élémentaires dans lesquels deux mots (ou plus) de chaque cluster sont co-occurrentes;
- Centrality = moyenne des index d'association concernant les rapports entre clusters;
- Density = moyenne des index d'association des mots dans chaque cluster.

CLUSTER	ECQ	CENTRALITY	DENSITY	MEMBERS
ACCORD	78	0,193	0,234	ACCORD; ARAFAT; DÉFINITIF; FINAL; MEMBRE; OSLO; PARLEMENT; STATUT; YASSER
AUTEUR	28	0,109	0,260	AUTEUR; COMPRENDRE; CONSACRER; ÉLABORER; INTÉRESSANT; PERSONNEL; PLACE
AUTORITÉ	44	0,141	0,172	AFFIRMER; AUTORITÉ; COMMUN; CONCESSION; DISPOSER; FONDER; HISTORIQUE; NÉCESSAIRE; NÉGOCIER; PARVENIR; PRINCIPE; RETRAIT
AVRIL	8	0,100	0,209	ACHEVER; ACTE; AVRIL; DERNIER
BARAK	43	0,126	0,211	BARAK; DIKTAT; ÉCRIRE; EFFET; EHOUD; ENTRAÎNER; ÉVOQUER; SHER
BEILIN	20	0,089	0,447	BEILIN; JUSTICE; SECRET; YOSSI
BLOC	6	0,075	0,243	BLOC; CONCLURE; CONDUIRE; ÉGYPTIEN
CAMP	58	0,190	0,318	CAMP; CLINTON; DAVID; PRÉSIDENT; SOMMET
CISJORDANIE	31	0,138	0,556	BANDE; CISJORDANIE; GAZA
COLONIE	69	0,167	0,201	ATTENDRE; COLONIE; CONSTRUCTION; ERREUR; GRAVE; IMPORTANT; INTERDIRE; POSER; POUVOIR_AMB; PROBLÈME; QUESTION; REPRÉSENTER; ROUTE; TRANSFORMER
CONTRÔLE	24	0,100	0,239	ACCORDER; AUTORISATION; CIRCULATION; CONTRÔLE; DIFFICILE; ÉLIRE; ENTRER; OUVRIR

3 - le tableau "**Index d'Association**" (voir ci-dessous) inclut des mesures concernant les rapports entre (between) et dans (within) les clusters.

Between	Within

< NÉGOCIATION >

CLUSTER	INDEX
accord	0,550
retrouver	0,528
camp	0,510
mettre	0,456
palestinien	0,386
janvier	0,375
BEILIN	0,363
Barak	0,316
violence	0,313
mener	0,290
Egypte	0,268
échec	0,266
rendre	0,228
Intifada	0,222
soutenir	0,213
ligne	0,212
responsabilité	0,207
tirer	0,201
septembre	0,193

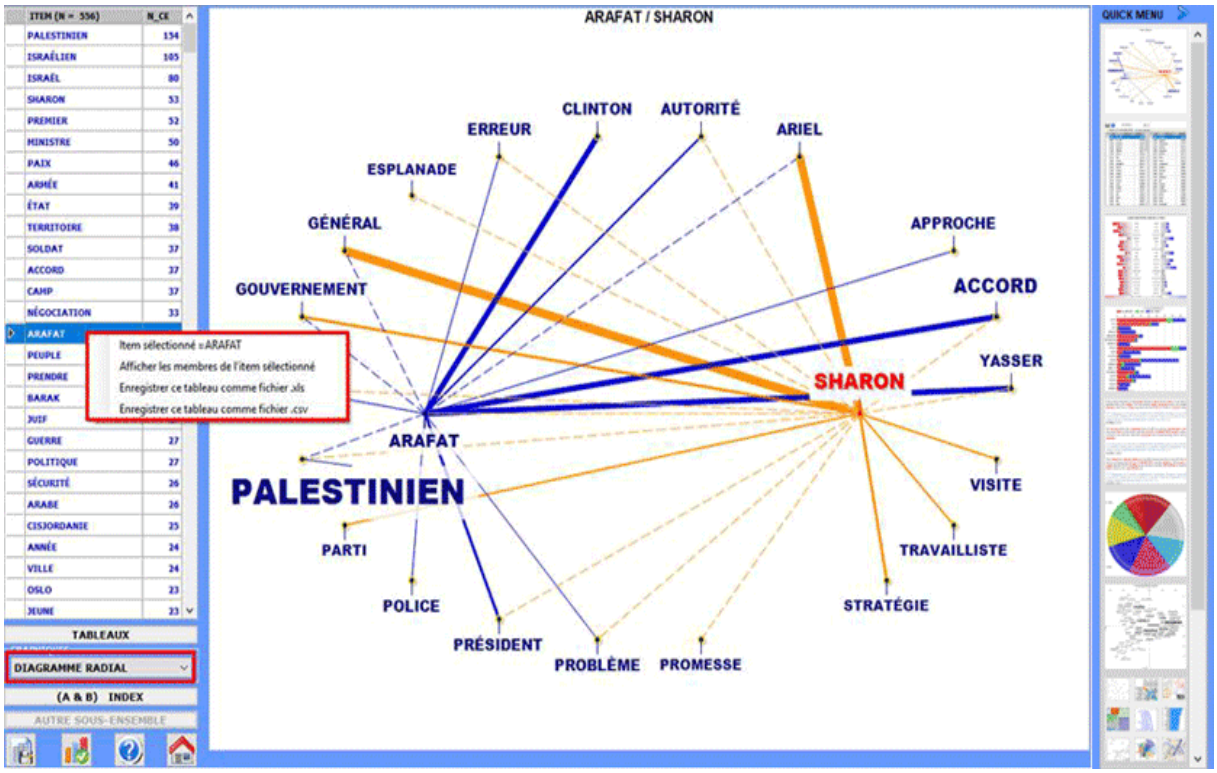
< NÉGOCIATION >

LEMMA_A	LEMMA_B	INDEX
document	revenir	0,447
pourparlers	reprise	0,408
négociation	table	0,395
reprise	table	0,378
revenir	table	0,338
pourparlers	table	0,309
négociation	reprise	0,261
document	reprise	0,250
reprise	revenir	0,224
document	pourparlers	0,204
document	table	0,189
pourparlers	revenir	0,183
document	négociation	0,174
négociation	revenir	0,156
négociation	pourparlers	0,142

N.B.:
- Quand la cluster analysis n'a pas été réalisée, le tableau "Membres des classes" n'est pas disponible, le tableau "Sommaire" est simplifié et le tableau "Index d'Association" concerne seulement les cooccurrences des mots ;
- Lorsqu'on quitte cette analyse, le dictionnaire des noyaux thématiques (c.-à-d. la liste des étiquettes assignées à chaque faisceau de mots) peut être exporté et, après une attentive révision, peut être importé en utilisant l'outil **Personnalisation du Dictionnaire**. De cette façon l'utilisateur pourra réaliser quelques analyses du deuxième ordre (c.-à-d. analyses qui concernent "thèmes" ou "concepts").

Comparaisons entre paires de Mots-Clés

N.B.: Les images de cette section font référence à une version précédente de T-LAB. En **T-LAB 10**, l'aspect est légèrement différent. En outre, le **bouton droit** sur les tableaux avec les mots-clés rend disponibles des options supplémentaires. Un nouveau **diagramme radial** est disponible qui permet de vérifier rapidement les différences entre les associations de mots est aussi disponible. Une galerie d'images à accès rapide qui fonctionne comme un menu supplémentaire permet de basculer entre les différentes sorties en un seul clic.
Certaines de ces nouvelles fonctionnalités sont mises en évidence dans l'image ci-dessous.

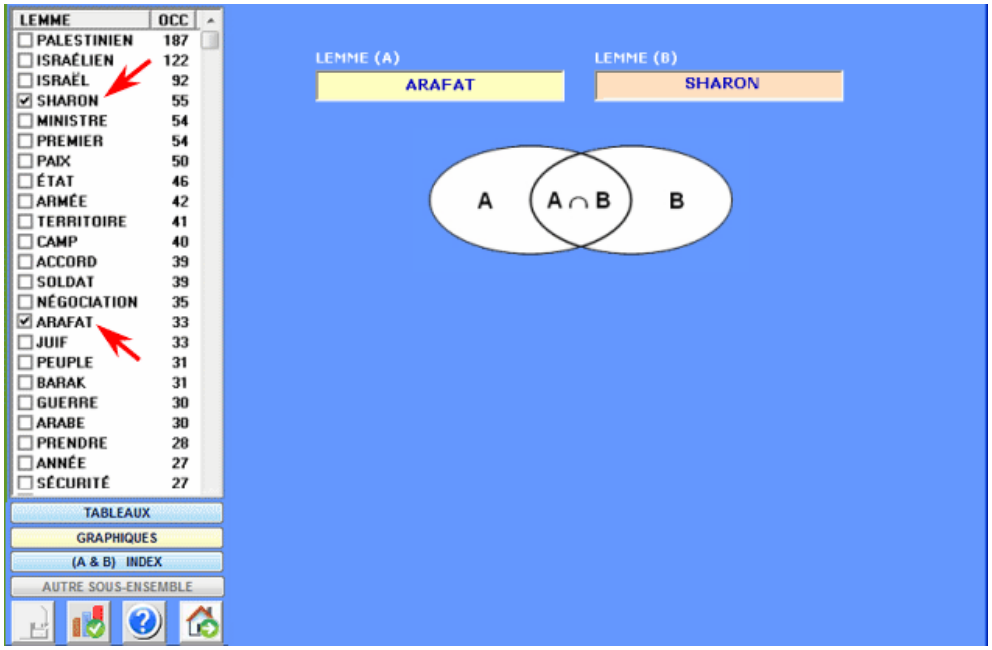


Cet outil **T-LAB** nous permet de comparer des ensembles de **contextes élémentaires** (c.-à-d. contextes de co-occurrence) dans lesquels sont présents les éléments d'une paire de **mots-clés**.

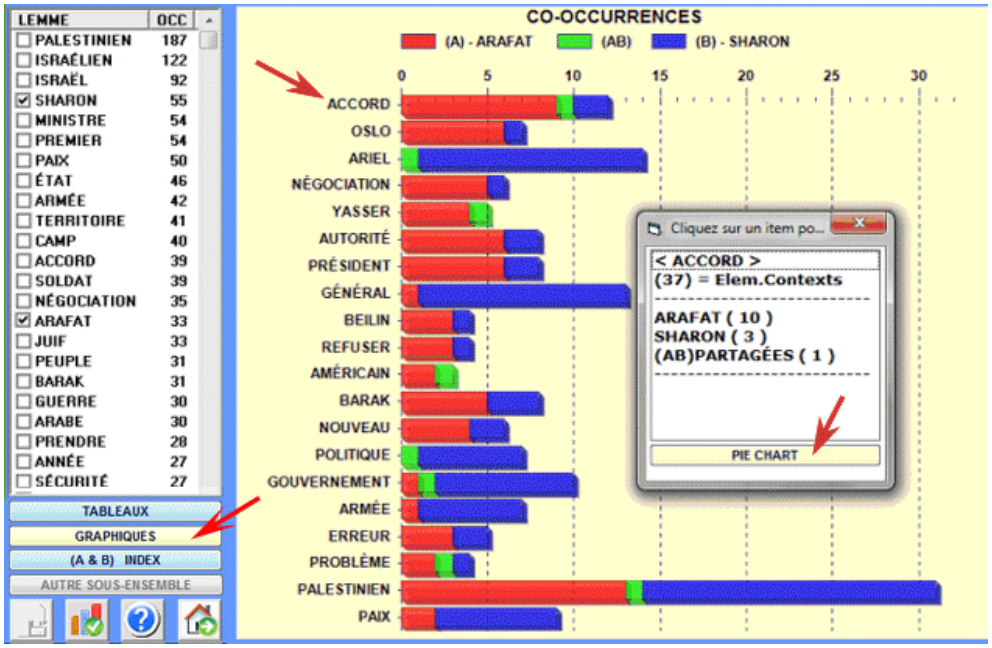
A gauche on trouve le tableau avec les **lemmes** sélectionnés et leur valeurs d'**occurrence** dans tout le **corpus** ou dans un de ses **sous-ensembles**.

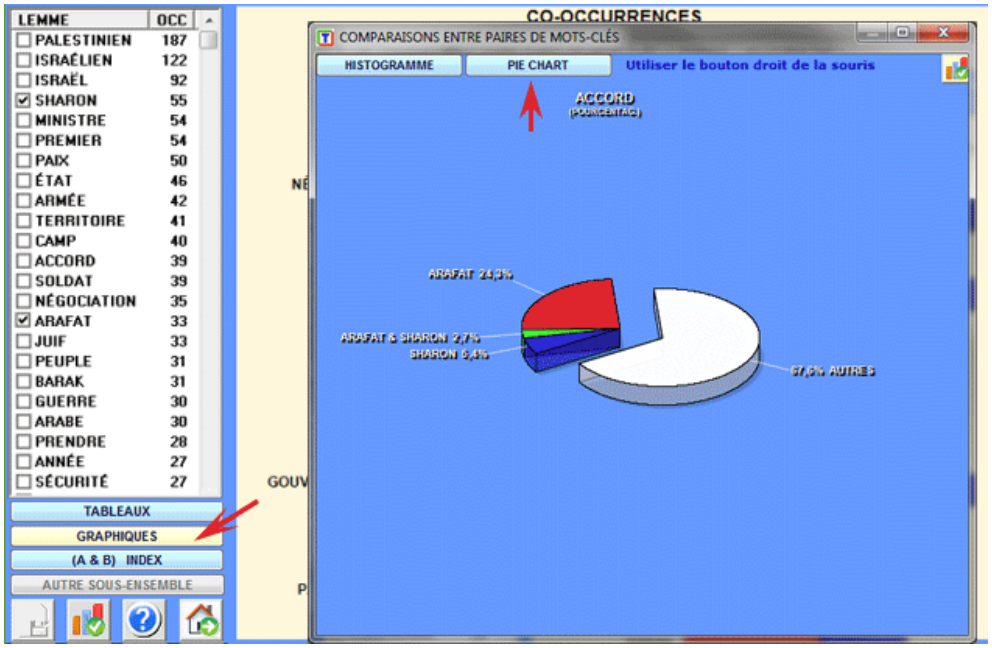
L'utilisateur est invité à sélectionner l'un après l'autre, avec un clic, deux d'entre eux (une "paire").

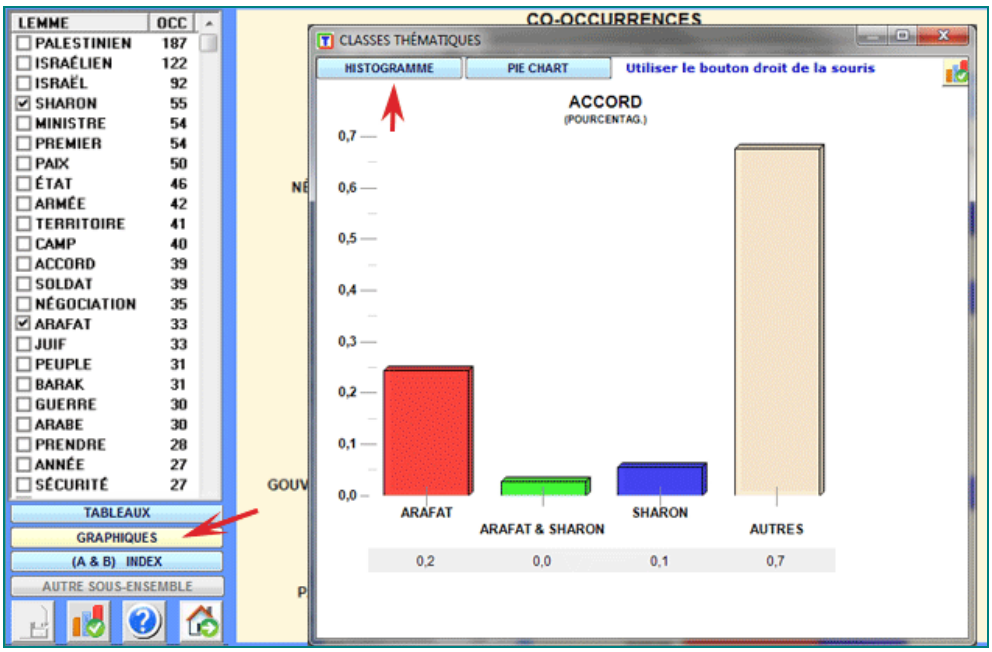
.



Un diagramme à barres (voir ci-dessous) nous permet d'apprécier le nombre de contextes élémentaires dans lesquels chaque lemme est en relation de co-occurrence avec le mot-clé "A" (couleur rouge), avec le mot-clé "B" (couleur bleue) et avec tous les deux (AB: couleur verte). Avec un double-clic sur chaque étiquette du diagramme il est également possible de vérifier les valeurs correspondantes.





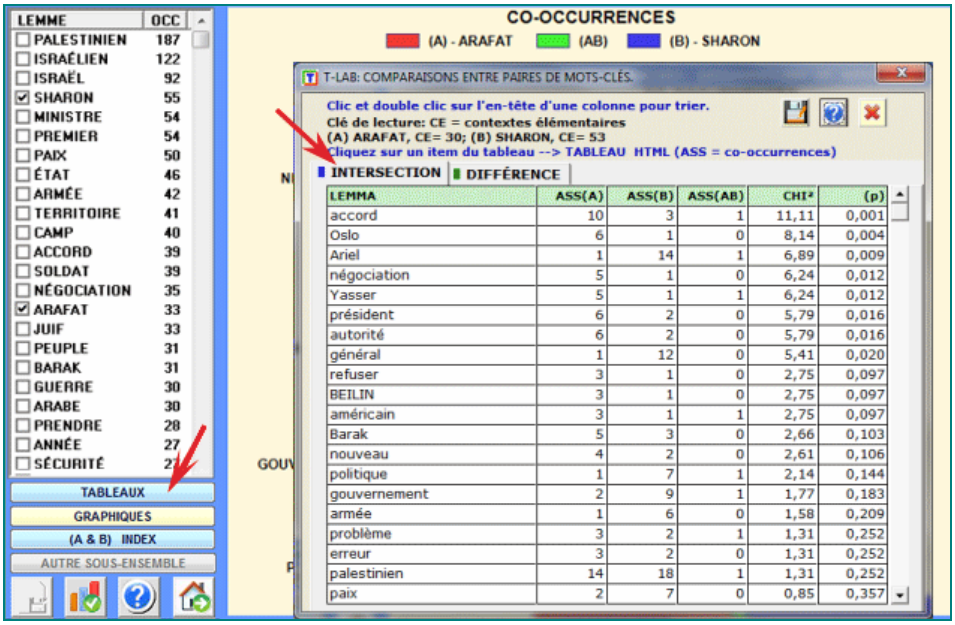


Les comparaisons proposées par **T-LAB** concernent les relations entre les éléments de la "paire" et chacun des mots contenus dans le tableau (voir ci-dessus).

Soit:
A = ensemble des **contextes élémentaires** (TOT C.E. = 30) où le premier mot de la paire (par ex. "Arafat") est présent;
B = ensemble des contextes élémentaires (TOT C.E. = 53) où le deuxième mot de la paire (par ex. "Sharon") est présent.

Le premier type de comparaison concerne les **associations partagées** (voir bouton **intersection**), c.-à-d. les mots qui sont présents tant dans "A" que dans "B".

Dans le tableau chaque ligne montre les valeurs correspondantes aux comparaisons de chaque lemme.



Les clés de lecture sont les suivantes:

- **ASS (A)** = nombre de contextes élémentaires dans lesquels chaque lemme est en relation de co-occurrence avec (A);
- **ASS (B)** = nombre de contextes élémentaires dans lesquels chaque lemme est en relation de co-occurrence avec (B);
- **ASS (AB)** = nombre de contextes élémentaires dans lesquels chaque lemme est en relation de co-occurrence avec (A) et (B) ;
- **CHI2** = CHI-deux ;
- **(p)** = probabilité associée à la valeur du chi-deux (def=1).

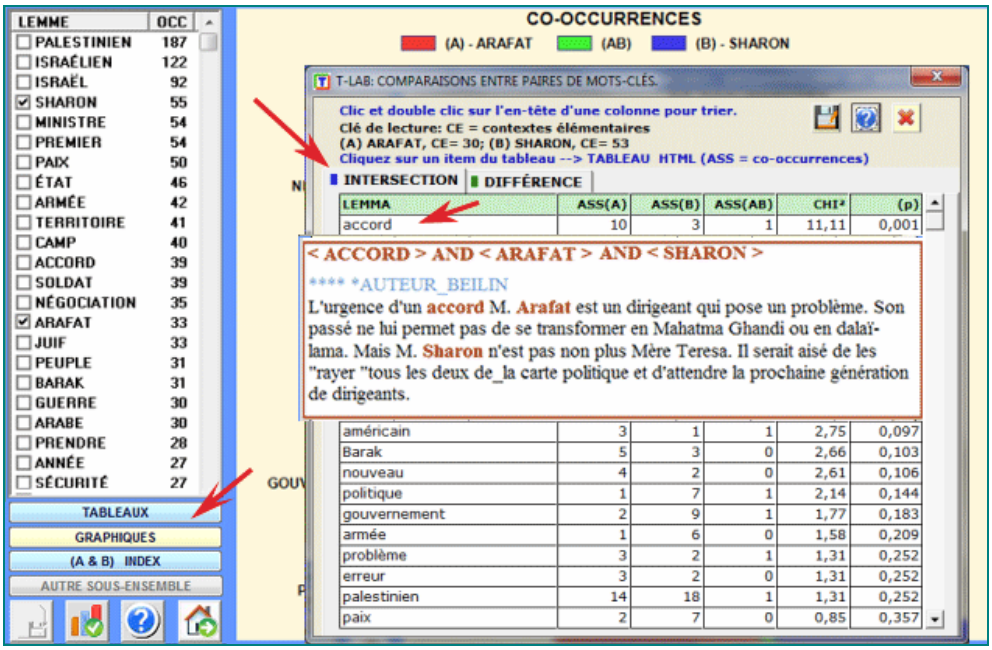
Dans ce cas-ci, pour chaque mot-clé (par ex. "accord") T-LAB construit un tableau comme suit et il y applique le test du **CHI Deux**:

	ASSOC.	NON ASSOC.	TOT.
A	10	20	30
B	3	50	53
	13	50	309

Dans la ligne (A) on indique le nombre de contextes élémentaires dans lesquels le mot "accord" est présent (10) ou absent (20) par rapport au total des contextes (30) propres du premier mot de la paire ("Arafat").
Dans la ligne (B) on indique le nombre de contextes élémentaires dans lesquels le mot "accord" est présent (3) ou absent (50) par rapport au total des contextes (53) propres du deuxième mot de la paire ("Sharon").

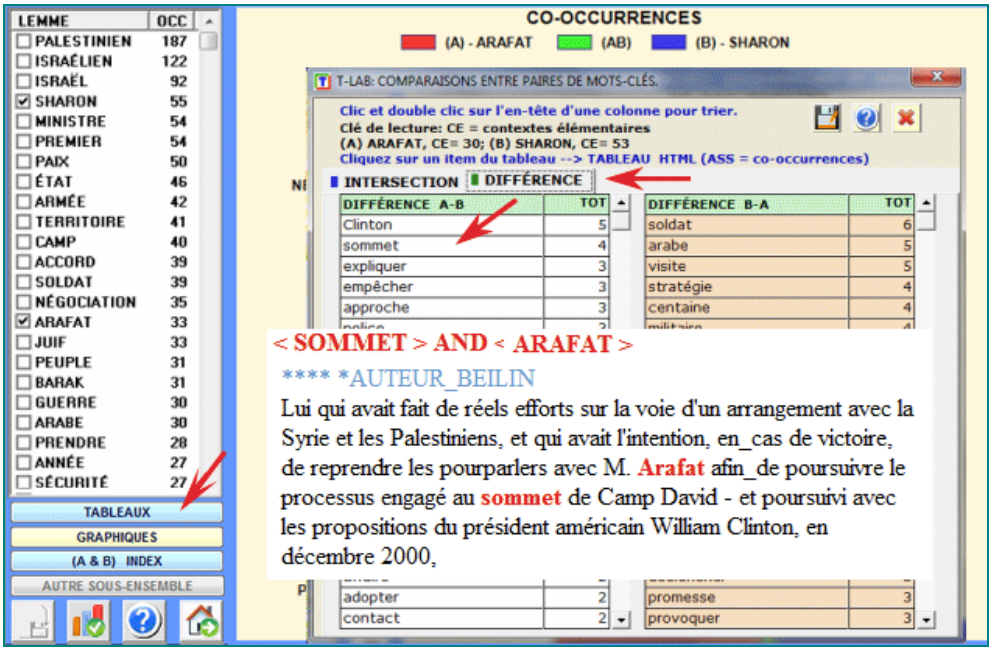
N.B.: dans ce cas, la valeur du CHI Deux correspond à 11,106.

D'ailleurs un double-clic sur chaque item de la table nous permet de sauver un dossier de HTML avec le nombre de contextes élémentaires dans la colonne correspondante.

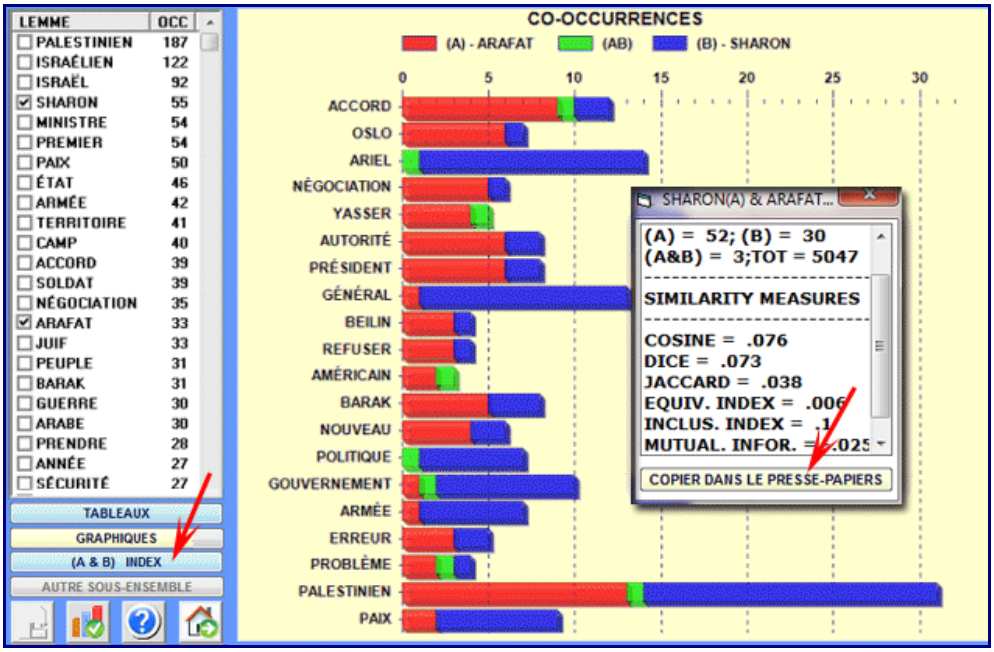


Le deuxième type de comparaison concerne les **différences** entre A et B (A - B e B - A).

Dans ce cas **T-LAB** propose deux tableaux avec les Mots-Clés qui sont associés au premier terme de la paire ou au deuxième de façon exclusive. Dans chacun des deux tableaux, la colonne "TOT" indique le nombre de contextes élémentaires pour lesquels chaque lemme n'est associé qu'avec un seul terme de la paire.



Enfin, en cliquant sur le bouton approprié, (voir l'image suivante) il est possible de vérifier et d'exporter tous les index de similarité qui concernent le couple de mots en examen.


LEMME OCC
PALESTINIEN 187
ISRAÉLIEN 122
ISRAËL 92
SHARON 55
MINISTRE 54
PREMIER 54
PAIX 50
ÉTAT 46
ARMÉE 42
TERRITOIRE 41
CAMP 40
ACCORD 39
SOLDAT 39
NÉGOCIATION 35
ARAFAT 33
JUIF 33
PEUPLE 31
BARAK 31
GUERRE 30
ARABE 30
PRENDRE 28
ANNÉE 27
SÉCURITÉ 27
TABLEAUX
GRAPHIQUES
(A & B) INDEX
AUTRE SOUS-ENSEMBLE
CO-OCCURRENCES
(A) - ARAFAT
(AB)
(B) - SHARON
0 5 10 15 20 25 30
ACCORD
OSLO
ARIEL
NÉGOCIATION
YASSER
AUTORITÉ
PRÉSIDENT
GÉNÉRAL
BEILIN
REFUSER
AMÉRICAIN
BARAK
NOUVEAU
POLITIQUE
GOUVERNEMENT
ARMÉE
ERREUR
PROBLÈME
PALESTINIEN
PAIX
SHARON(A) & ARAFAT...
(A) = 52; (B) = 30
(A&B) = 3;TOT = 5047
SIMILARITY MEASURES
COSINE = .076
DICE = .073
JACCARD = .038
EQUIV. INDEX = .006
INCLUS. INDEX = .1
MUTUAL. INFOR. = .025
COPIER DANS LE PRESSE-PAPIERS

Analyse des Séquences et Analyse des Réseaux

Cet outil **T-LAB** tient compte des **positions** des différentes unités lexicales à l'intérieur des phrases et il nous permet de représenter et d'explorer n' importe quel texte comme un **réseau** de relations.

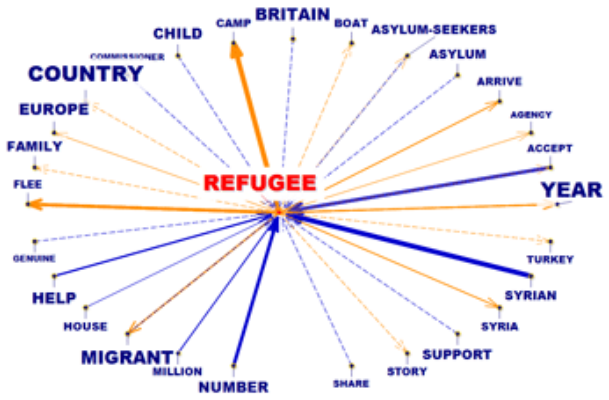
Les différentes options disponibles peuvent être utilisées pour des buts tels qu'analyses Co-Word, analyses thématiques et désambiguïsations.

En effet, après avoir construit deux matrices dans lesquelles tous les couples de prédécesseurs et successeurs sont enregistrés, **T-LAB** calcule les **probabilités de transition** (chaînes de Markov) et il fournit différents outputs qui concernent les mots cible.

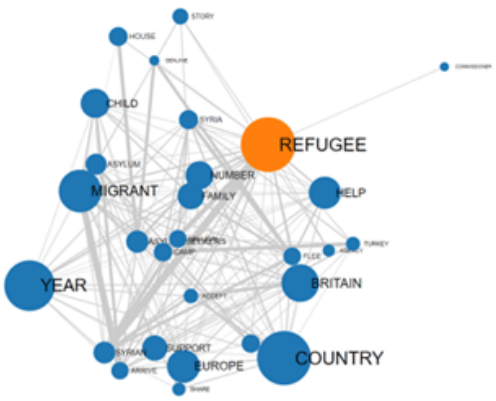
En outre, il est possible d'exécuter un **cluster analysis** et d'explorer les relations sémantiques entre les mots soit à l'intérieur du réseau entier qu'à l'intérieur de «clusters thématiques» (N.B: Dans ce cas-ci, l'algorithme de clustérisation est constitué par la «méthode Louvain» développée par Blondel V.D., Guillame J.-L, Lambiotte R., Lefebre E., 2008).

Ceci signifie, après avoir exécuté ce type d'analyse, que l'utilisateur peut vérifier les relations entre les nœuds du réseau (c'est-à-dire les mots-clés) à plusieurs niveaux: a) en relations du type un-à-un; b) à l'intérieur d' «ego network»; c) à l'intérieur des «communautés» auxquelles ils appartiennent; d) à l'intérieur du réseau entier constitué par le texte en analyse.

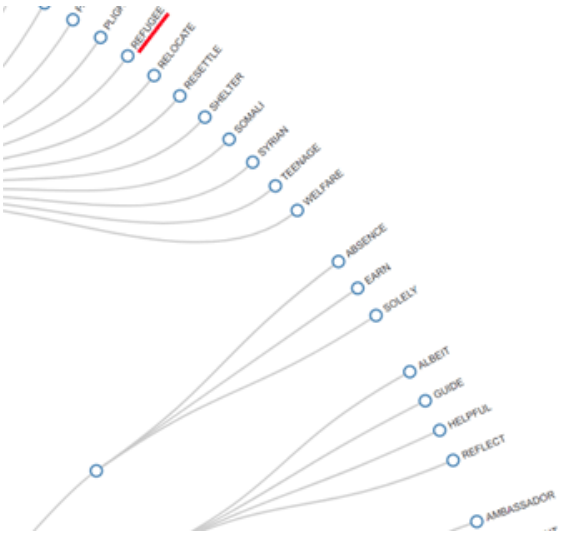
RELATIONS UN-À-UN



EGO-NETWORK



COMMUNAUTES



RÉSEAU ENTIER

Les renseignements sur l'utilisation des différentes options d'analyse sont organisés en trois sections:

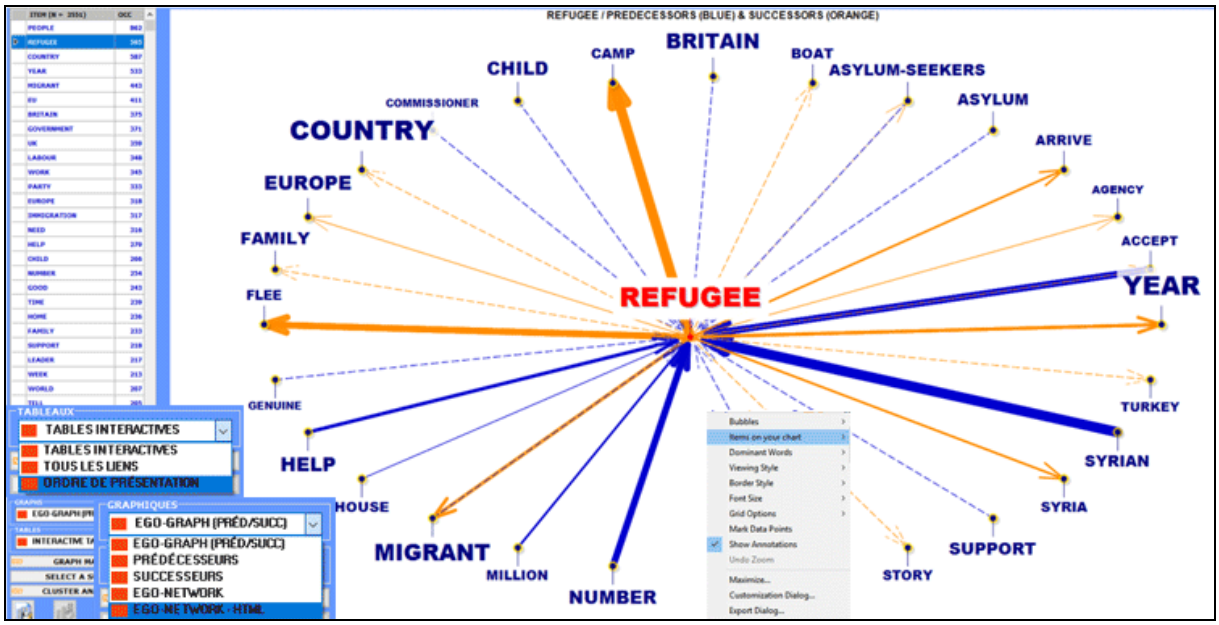
A - Explorer les connexions du type un-à-un et les "ego network";
B - Explorer les «communautés» (c'est-à-dire les clusters thématiques) et le réseau entier;
C - Certains détails techniques.

N.B.: Pour motifs d'édition, cette page inclut des exemples d'analyse tirés d'un corpus dont les textes sont en anglais.

A - EXPLORER LES CONNEXIONS DU TYPE UN-À-UN ET LES "EGO NETWORK"

Quand l'analyse automatique est terminée, divers graphiques et tableaux qui permettent de vérifier les relations et les données qui concernent les mots-clés sélectionnés sont disponibles (N.B: à ce but il est suffisant de cliquer sur un item des tableaux ou sur un point quelconque montré dans les graphiques).

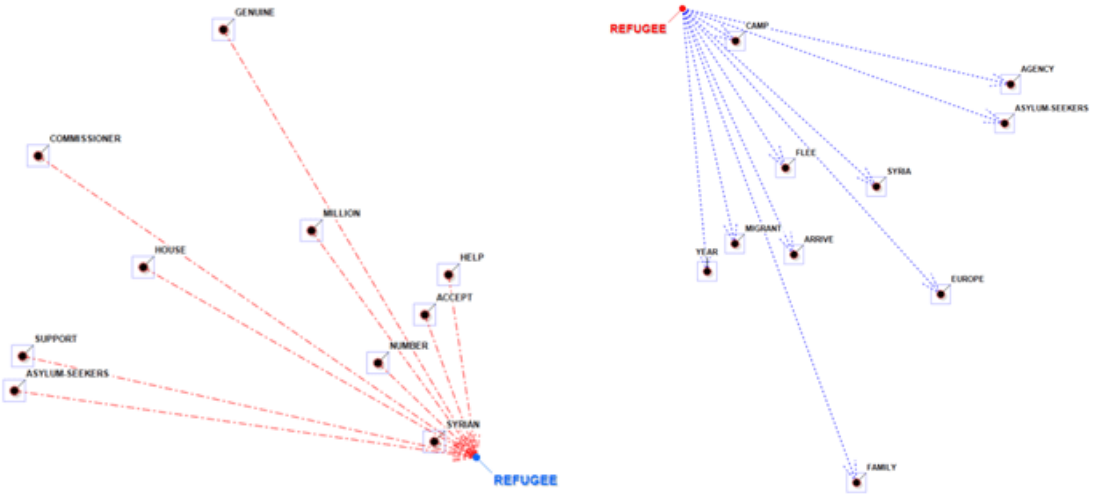
Tous les **graphiques** peuvent être personnalisés et exportés en divers formats (utiliser le bouton droit de la souris).



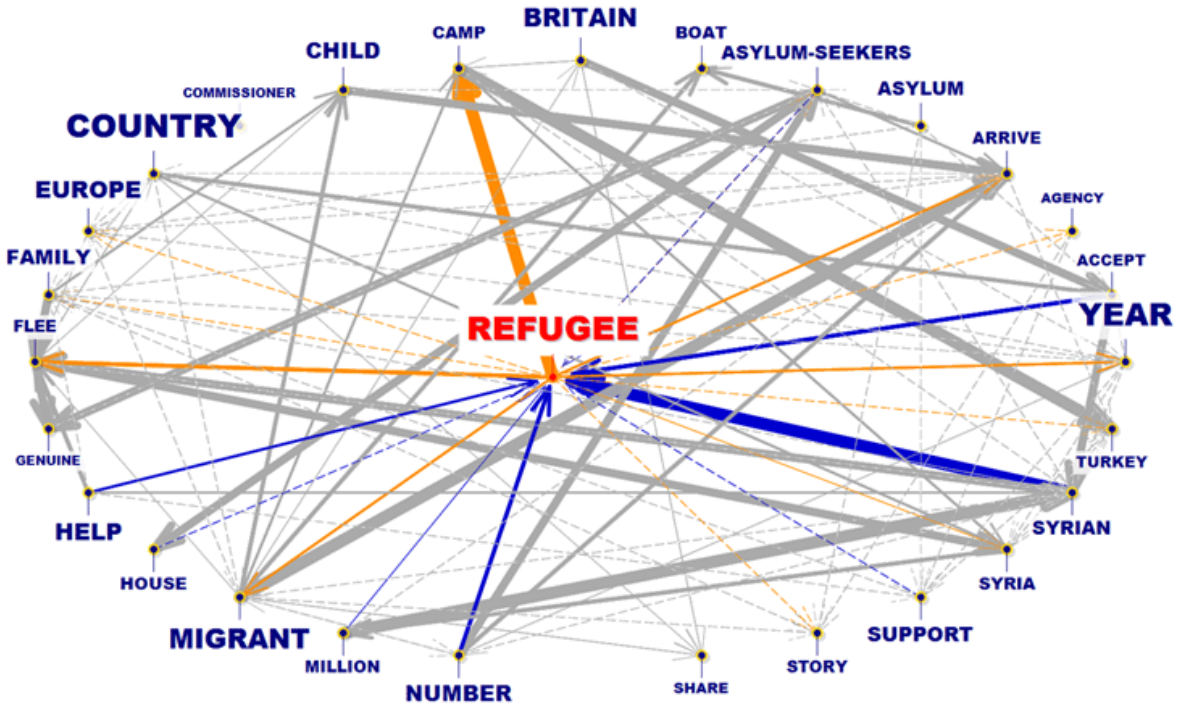
En deux des graphiques, les éléments les plus voisins à ceux sélectionnés sont ceux qui ont les probabilités les plus élevées de venir devant (prédécesseurs) ou après (successeurs) de ceux-ci.

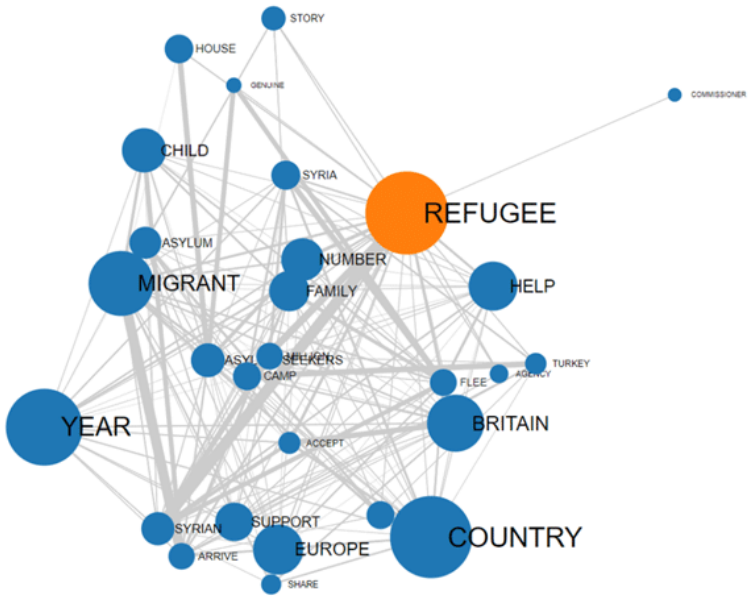
.

PREDECESSEURS **SUCCESSEURS**



Dans les autres cas, la proximité entre les termes-clés est représentée par les différentes épaisseurs des flèches qui les joignent (voir ci-dessous).



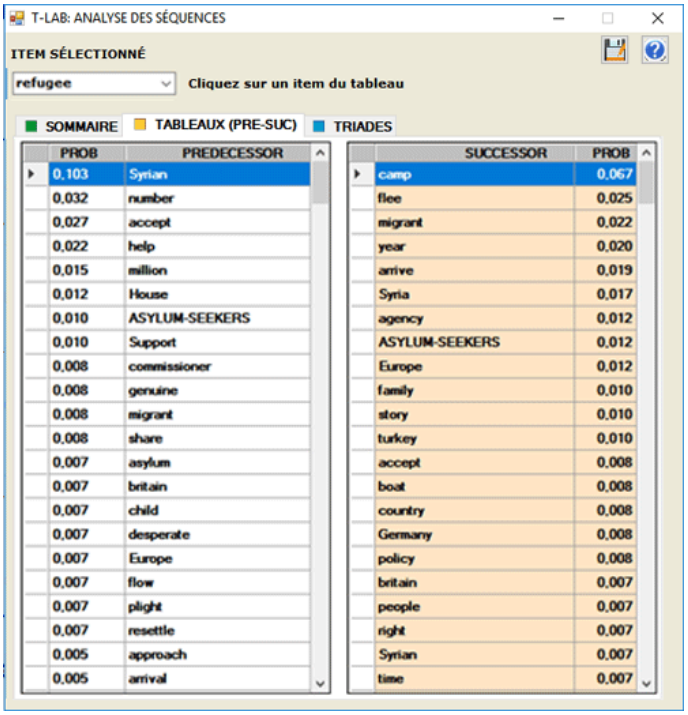


Toutes les données peuvent être vérifiées au moyen de plusieurs types de **tableaux**.

En détail:

Les **TABLEAUX INTERACTIFS** montrent les listes des prédécesseurs et des successeurs associés avec les mots-clés sélectionnés.

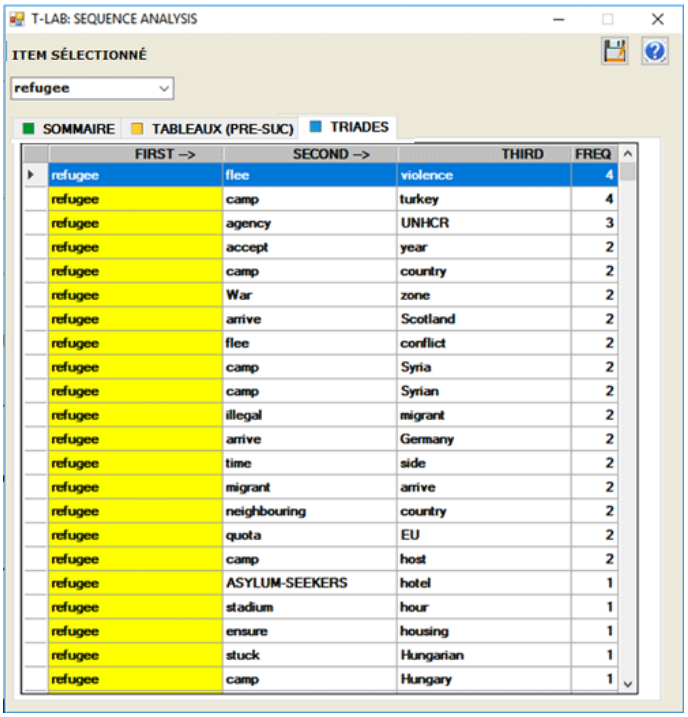
La liste est en ordre décroissant selon les valeurs de probabilité ("PROB"). Par exemple, dans le tableau suivant, la probabilité que "camp" suive "refugee" est égale à 0.067, c.-à-d. le 6.7%.



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

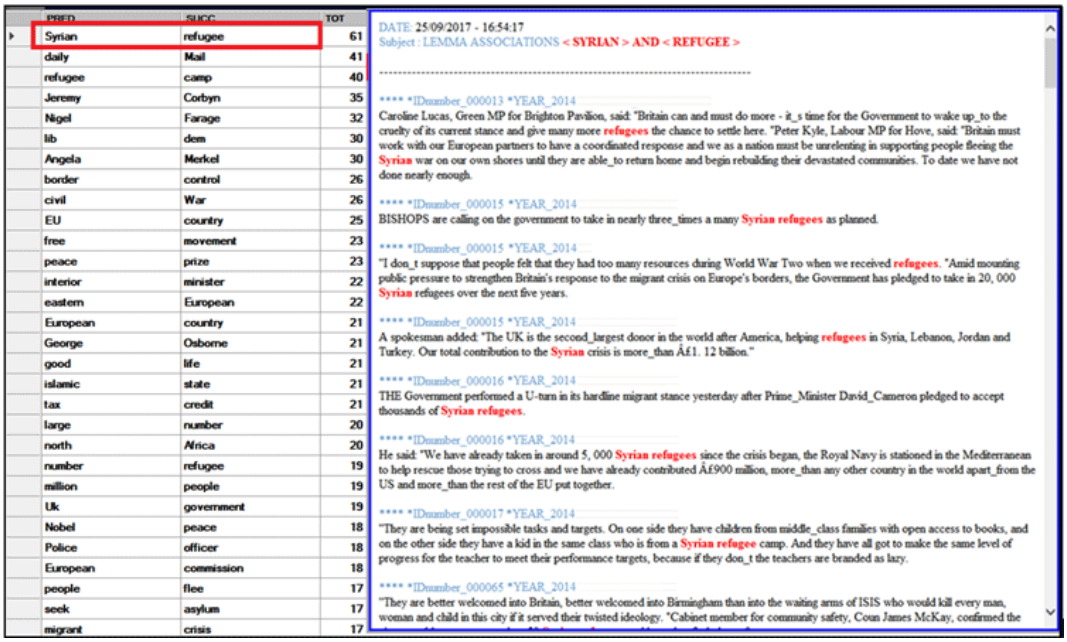
L'option "triades" nous permet de visualiser quelques tables avec des séquences de trois éléments dans lesquels, selon le choix de l'utilisateur, le mot choisi est dans la première, dans la deuxième ou dans la troisième position. Pour chaque triade **T-LAB** montre les valeurs d'occurrence correspondantes. (N.B.: Dans les triades les **mots vides** ne sont pas inclus).



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

Le tableau **TOUS LES LIENS** (voir ci-dessous), qui est particulièrement utile pour désambiguïser les sens des mots, contient tous les couples de prédécesseurs et de successeurs, et aussi les occurrences respectives. En faisant clic sur une ligne de ce tableau, tous les segments de texte, (c'est-à-dire les contextes élémentaires) dans lesquels les deux membres de chaque couple sont présents en même temps (c'est-à-dire co-occurrences), seront visualisés en format HTML sur le côté droit du tableau.

.



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.

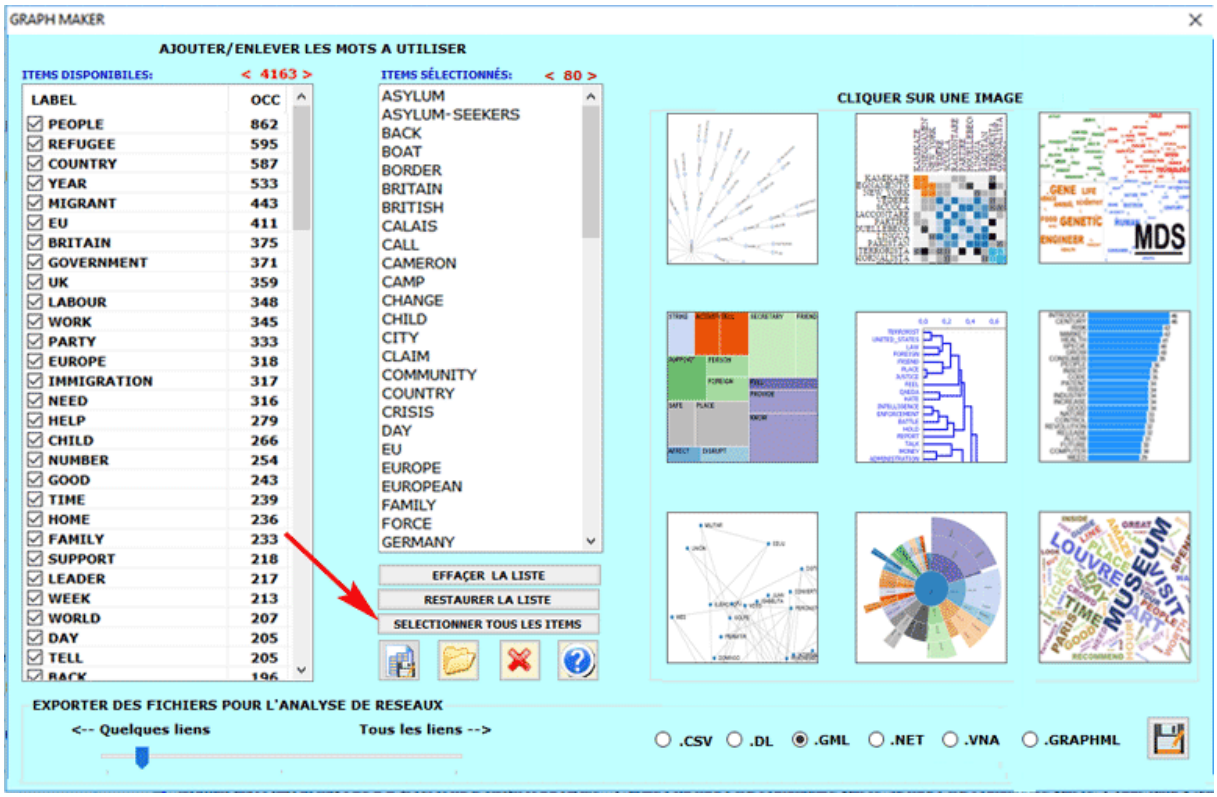
**** *IDnumber_000016 *YEAR_2014
He said: "We have already taken in around 5, 000 Syrian refugees since the crisis began, the Royal Navy is stationed in the Mediterranean to help rescue those trying to cross and we have already contributed Â£900 million, more_than any other country in the world apart_from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

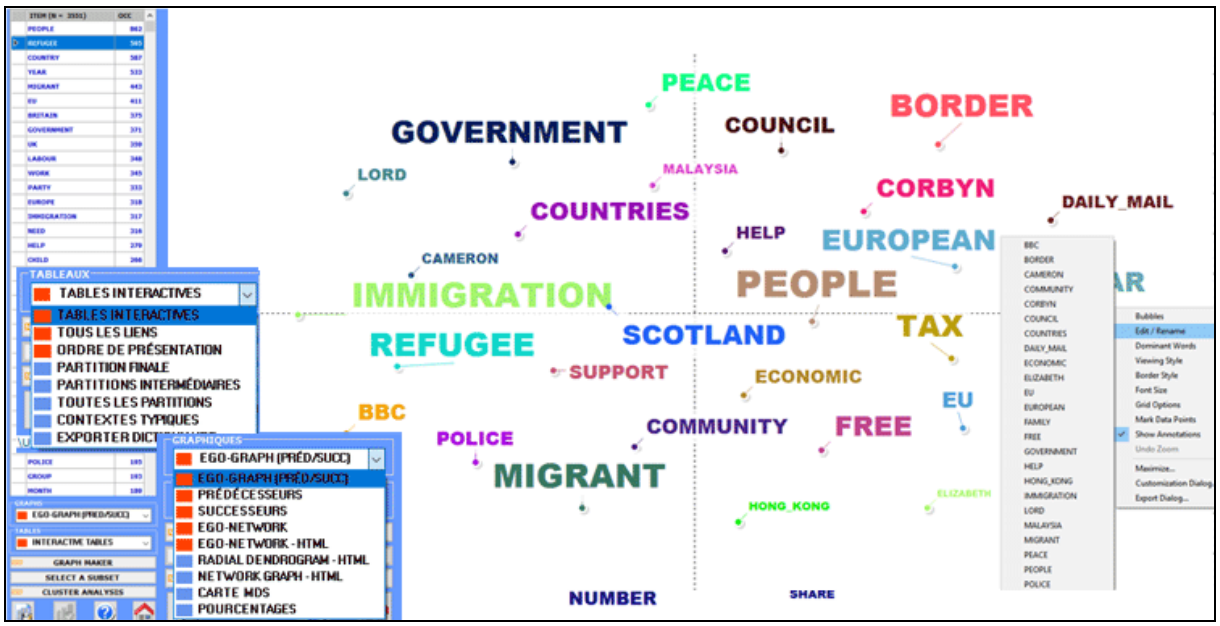
Le tableau **RANG D'APPARITION**, avec la fréquence et l'ordre moyen d'apparition (ou d'évocation) de chaque terme à l'intérieur des segments de texte, est visible seulement quand le corpus est constitué par des textes brefs, par exemple des réponses à des questions ouvertes.

À n'importe quel moment, en faisant clic sur l'option **GRAPH MAKER,** l'utilisateur peut créer des différents types de graphiques en utilisant des listes personnalisées de mots-clés, (voir ci-dessous)
N.B.: Les utilisateurs experts intéressés à exporter des fichiers en formats divers (par exemple .dl .gml etc.) avec les données relatives à tous les links, peuvent faire clic sur le bouton «SÉLECTIONNER TOUS LES ITEMS».



B - EXPLORER LES « COMMUNAUTÉS » (C'EST-À-DIRE LES CLUSTERS THÉMATIQUES) ET LE RÉSEAU ENTIER

Quand on fait une analyse cluster, d' **autres graphiques et tableaux** sont disponibles. Ils sont tous marqués avec un petit rectangle bleu (voir ci-dessous).

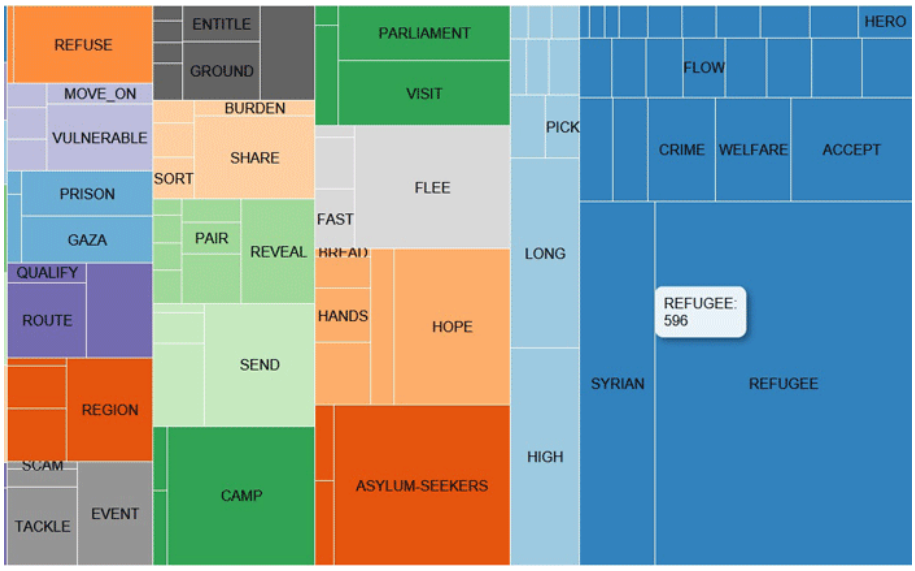


Un premier tableau résume les caractéristiques (c'est-à-dire les termes-clés), de la **PARTITION FINALE** obtenue par l'algorithme de clustérisation.
Dans ce tableau, les caractéristiques de chaque cluster thématique sont ordonnées par la valeur relative **TF-IDF** (voir ci-dessous).

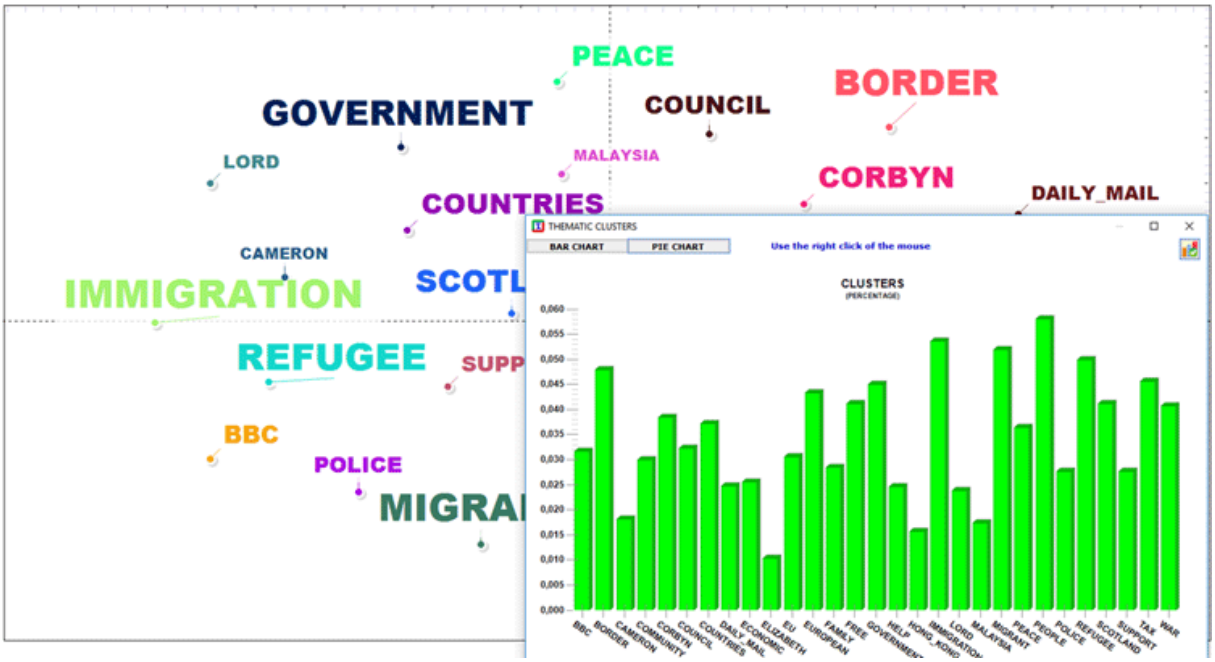
N.B : Lorsqu'un cluster de la partition finale comprend seulement deux mots, habituellement cela signifie qu'un cas de multiword n'a pas été résolu pendant la phase de pré-traitement

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

En cliquant sur n'importe quel mot dans le tableau ci-dessus (ainsi que dans le tableau **TOUTES LES PARTITIONS**), un TreeMap nous permet de vérifier les communautés auxquelles il appartient (voir ci-dessous).

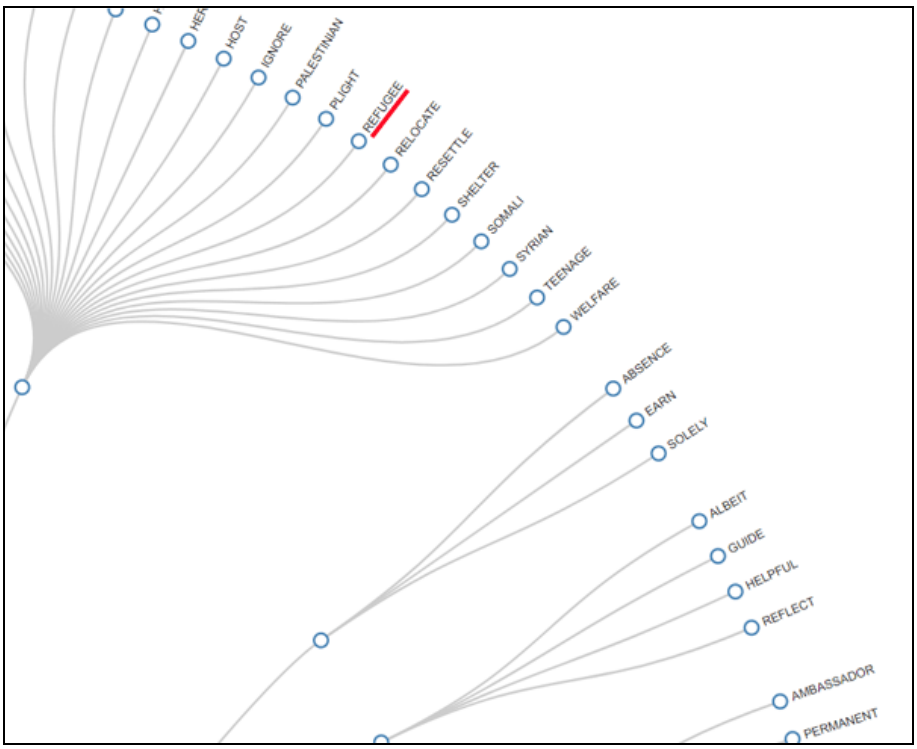


La **CARTE MDS** et le graphique **POURCENTAGES** nous permettent de vérifier le «poids» de chaque cluster ainsi que les relations entre les différents clusters à l'intérieur de la partition finale (voir ci-dessous).

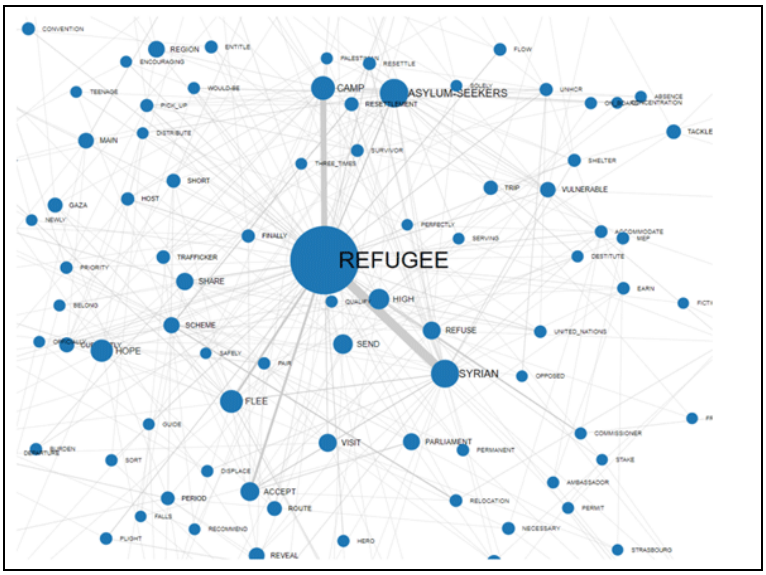


Selon le numéro de mots-clés, deux graphiques en format HTML nous permettent de vérifier leurs relations soit à l'intérieur du réseau entier qu'à l'intérieur du cluster auquel ils appartiennent (voir ci-dessous).

RADIAL DENDROGRAM



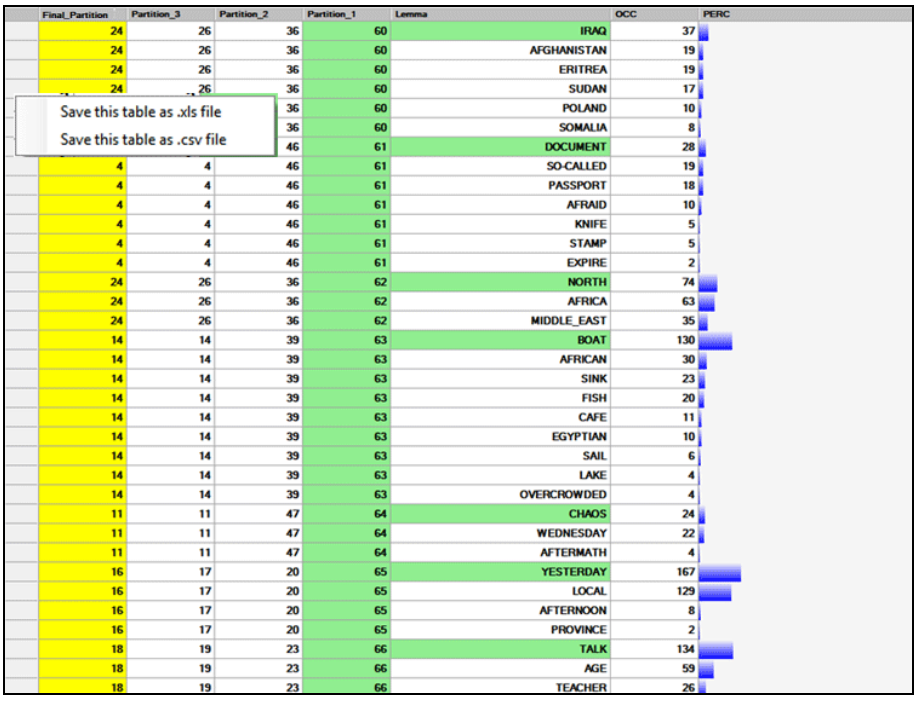
NETWORK (FORCE-DIRECTED GRAPH)



Trois autres tableaux nous fournissent d'autres renseignements obtenus par l'analyse cluster.

En particulier:

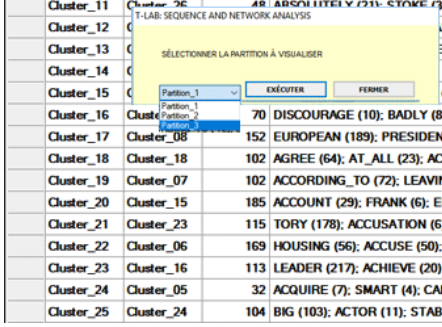
Le tableau **TOUTES LES PARTITIONS** permet de vérifier comment les mots-clés ont été groupés dans chaque partition de l' analyse des clusters (voir ci-dessous).
N.B.: Pour réglage prédéfini, ce tableau est présenté ordonné sur la première partition (c'est-à-dire celle avec le plus grand numéro de clusters), et chaque passage d'un petit cluster à l'autre est marqué en soulignant en vert le premier mot qui lui appartient.



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	
		36	60	POLAND	10	
		36	60	SOMALIA	8	
		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	

Le tableau **PARTITIONS INTERMEDIAIRES** nous permet de vérifier comment les mots-clés ont été groupés dans la partition sélectionnée.
Dans cet tableau, les caractéristiques de chaque groupe thématique sont triées par leur valeurs d'occurrence (voir ci-dessous).

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); EM
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); C
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); T
Cluster_12			BUSE (25); DOMESTIC (17); SLAVERY (7); HORRENDOUS (3); PAIN (21); ADDITIONAL (13); NE
Cluster_13			(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	Cluste	70	DISCOURAGE (10); BADLY (8); ACCEPTABLE (2); ACTIVIST (39); MINUTE (28); SWEAR (2); WESTERN (24); MANIFESTO (20
Cluster_17	Cluster_08	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



Le tableau **CONTEXTES TYPIQUES** nous permet de contrôler les segments de texte qui ont le plus haut score d'association avec les différents clusters de la meilleure partition. Dans ce tableau le "score" se réfère à la ressemblance (index cosinus) entre le vecteur des caractéristiques de chaque cluster et le vecteur dans lequel chaque segment de texte est représenté.

N.B. Le segment de texte plus significatif de chaque cluster est marqué en jaune.

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
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
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
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 member
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 issuec
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 pled
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
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 '
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

Comme en d'autres cas d'analyse thématique, **T-LAB** permet d'**exporter le dictionnaire** de la partition meilleure qui peut être utilisé pour d'autres analyses.

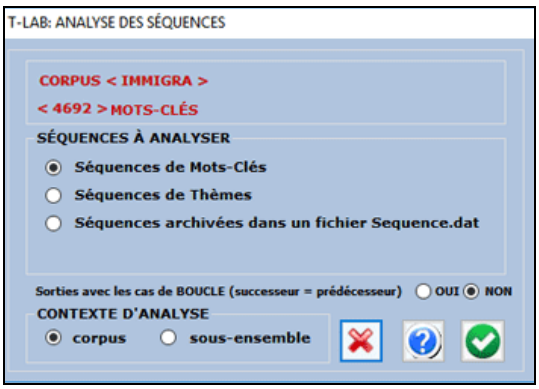
C - QUELQUES DÉTAILS TECHNIQUES

Les types de séquences que cet outil **T-LAB** nous permet d'analyser sont les suivants:

1- **Séquences de mots-clés**, dont les éléments sont des unités lexicales (c'est-à-dire mots ou lemmes) présentes dans le corpus ou dans un de ses sous-ensembles. Dans ce cas, le nombre maximum des 'nœuds' (à savoir les " types" d'unités lexicales) est de 5.000;
N.B.: Lorsque la lemmatisation automatique est appliquée, 5.000 unités lexicales correspondent à environ 12.000 mots.

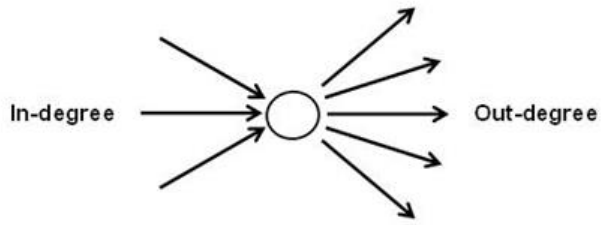
2- **Séquences de Thèmes**, dont les éléments sont des unités de contexte (c'est-à-dire des contextes élémentaires) classifiées par un outil **T-LAB** pour l'analyse thématique.
N.B.: Dans ce cas, étant donné que la séquence des contextes élémentaires (phrases ou paragraphes) caractérise toute la "chaîne" (prédécesseurs et successeurs) du corpus, **T-LAB** produit une forme spécifique de l'analyse du discours, dans laquelle les nœuds (c'est-à-dire les "thèmes") peuvent varier d'un minimum de 5 à un maximum de 50.

3 - **Séquences archivées dans un fichier Sequence.dat**, préparé par l'utilisateur (voir les explications relatives à la fin de cette section). Dans ce cas le nombre maximum de records est de 50.000 et le nombre de " types " (c'est-à-dire de nœuds) ne doit pas dépasser 5.000.

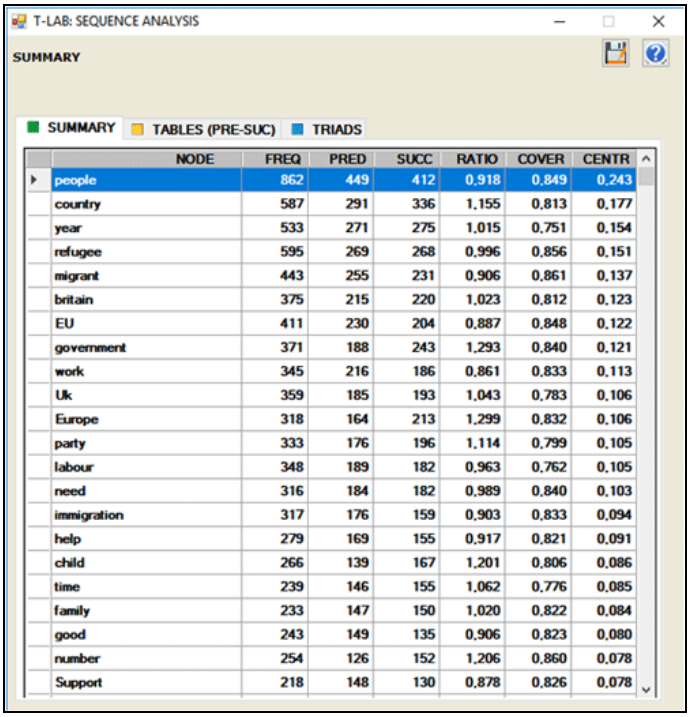


Les informations suivantes sont fournies pour aider l'utilisateur à mieux comprendre les données dans le tableau **SOMMAIRE**.

Selon la théorie des graphiques, les prédécesseurs et les successeurs de chaque **nœud** (dans ce cas-ci, chaque unité lexicale) peuvent être représentés au moyen de flèches (arcs) en entrée (in-degree = types de prédécesseurs) ou en sortie (out-degree = types de successeurs).



Par exemple, dans la table suivante "people" a 412 types de successeurs et 449 types de prédécesseurs.
Et son degré de centralité est 0.243.



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

Selon leur rapport (successeurs/prédécesseurs), il est possible de vérifier la variété sémantique engendrée par chaque noeud:

- si le rapport est plus grand que 1, le nœud est définie "source";
- si le rapport est égal à 1, le nœud est défini "relais";
- si le rapport est inférieur à 1, le nœud est défini "puits".

Dans le même tableau, pour chaque unité lexicale, la colonne "cover" (couverture) indique le pourcentage de ses occurrences précédées ou suivies des unités lexicales incluses dans la liste de l'utilisateur.

Quand les unités analysées "couvrent" la totalité de celles présentes dans le corpus, la valeur de "cover" est égale à 1; autrement, c'est une valeur inferieure. D'ailleurs: quand la valeur de "cover" est égale à 1, également les totaux des probabilités (des prédécesseurs et des successeurs) sont égales à 1; autrement, ils sont des valeurs inferieures. Dans les deux cas, le pourcentage "résiduel" est déterminé par le fait qu'il y a des prédécesseurs et des successeurs non inclus dans l'analyse.

Par exemple, la séquence représentée dans l'image suivante est constituée par 39 événements. De ces derniers, seulement 16 (les hypothétiques unités analysées) sont "couverts" (boîtes grises); ceci parce que certains d'eux, par exemple ceux correspondants aux occurrences de l'unité lexicale "A", ont des prédécesseurs et des successeurs non inclus dans l'analyse (boîtes blanches).

Différemment, quand l'utilisateur analyse **séquences de thèmes** ou un **fichier externe** tout les événements sont "couverts".

N.B.: Pour analyser un fichier externe, l'utilisateur doit préparer le fichier « Sequence.dat » correspondant; puis, après avoir ouvert un projet déjà existant, il doit sélectionner l'option "Séquences enregistrées dans un fichier Sequence.dat».

La méthode de calcul, les graphiques et les tables sont analogues à ceux déjà décrites (voir ci-dessus).

Le fichier Sequence.dat, qui peut contenir chaque genre d'étiquettes (par exemple les noms des parleurs dans une conversation, des catégories obtenues par analyse du contenu, des séquences d'événements, etc.), doit se composer par "N" lignes (minimum 50 maximum 50.000), chacune avec une étiquette d'un maximum de 50 caractères, sans signes de ponctuation ni espaces vides.

Les types d'étiquettes doivent être maximum 5.000.

Voici quelques exemples de fichier Sequence.dat dans le format correct

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
... ...	... ...	... ...

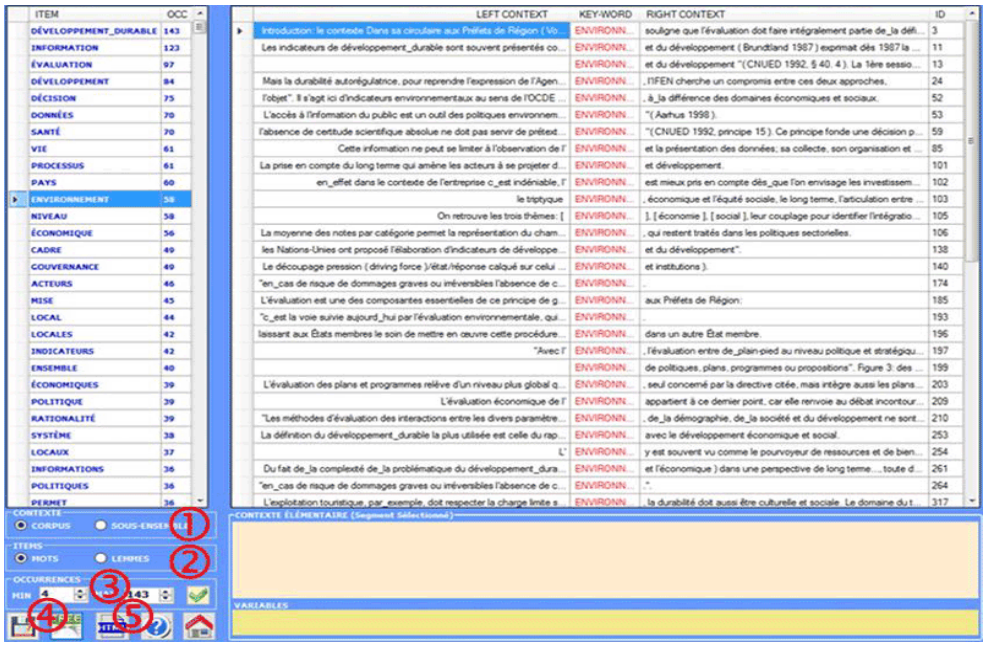
Aussi bien après l'analyse des séquences (syntagmes) du corpus qu'après l'analyse d'un fichier externe (Sequence.dat), **T-LAB** produit des tableaux dans le dossier MY-OUTPUT.

Concordances

Cet outil **T-LAB** nous permet de vérifier les contextes d'occurrence de chaque unité lexicale.

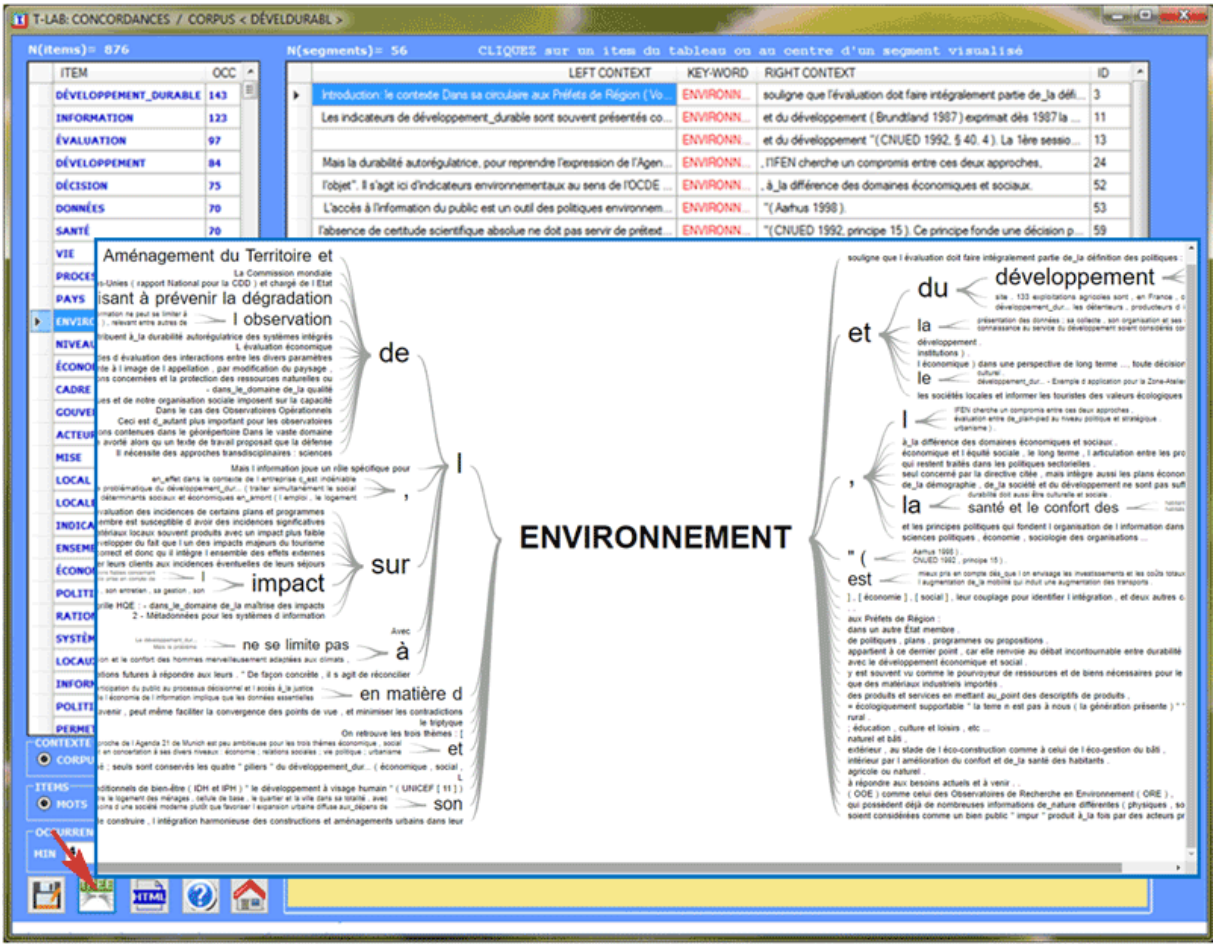
Les recherches de genre KWIC (Key-Word-in-Context) peuvent être faites par **formes** ou par **lemmes** (voir l'option '2' ci-dessous), tant au sein du **corpus** entier qu' à l'intérieur d'un de ses **sous-ensembles** (voir l'option '1' ci-dessous).

Pour chaque unité lexicale (**mot** ou **lemme**) du corpus, avec un simple clic, il est possible de vérifier quels sont ses contextes d'occurrence (les **contextes élémentaires**).



De plus, il est possible de créer un 'Word Tree' dynamique (voir l'option ci-dessus '4') ou sauvegarder un fichier en mode **HTML**.

.



DATE: 20/06/2015 - 09:33:59
CONCORDANCES **ENVIRONNEMENT**

**** *ARTIC_A01
Introduction: le contexte Dans sa circulaire aux Préfets de Région (Voynet 1999), Dominique Voynet, la Ministre de l'Aménagement du Territoire et de l'**ENVIRONNEMENT** souligne que l'évaluation doit faire intégralement partie de_la définition des politiques:

**** *ARTIC_A01
Les indicateurs de développement_durable sont souvent présentés comme une base rationnelle à_la décision, dépassant les limites de_la rationalité économique. La Commission mondiale de l'**ENVIRONNEMENT** et du développement (Brundtland 1987) exprimait dès 1987 la nécessité d'adopter de nouvelles méthodes de mesure et d'évaluation des progrès en faveur du développement_durable.

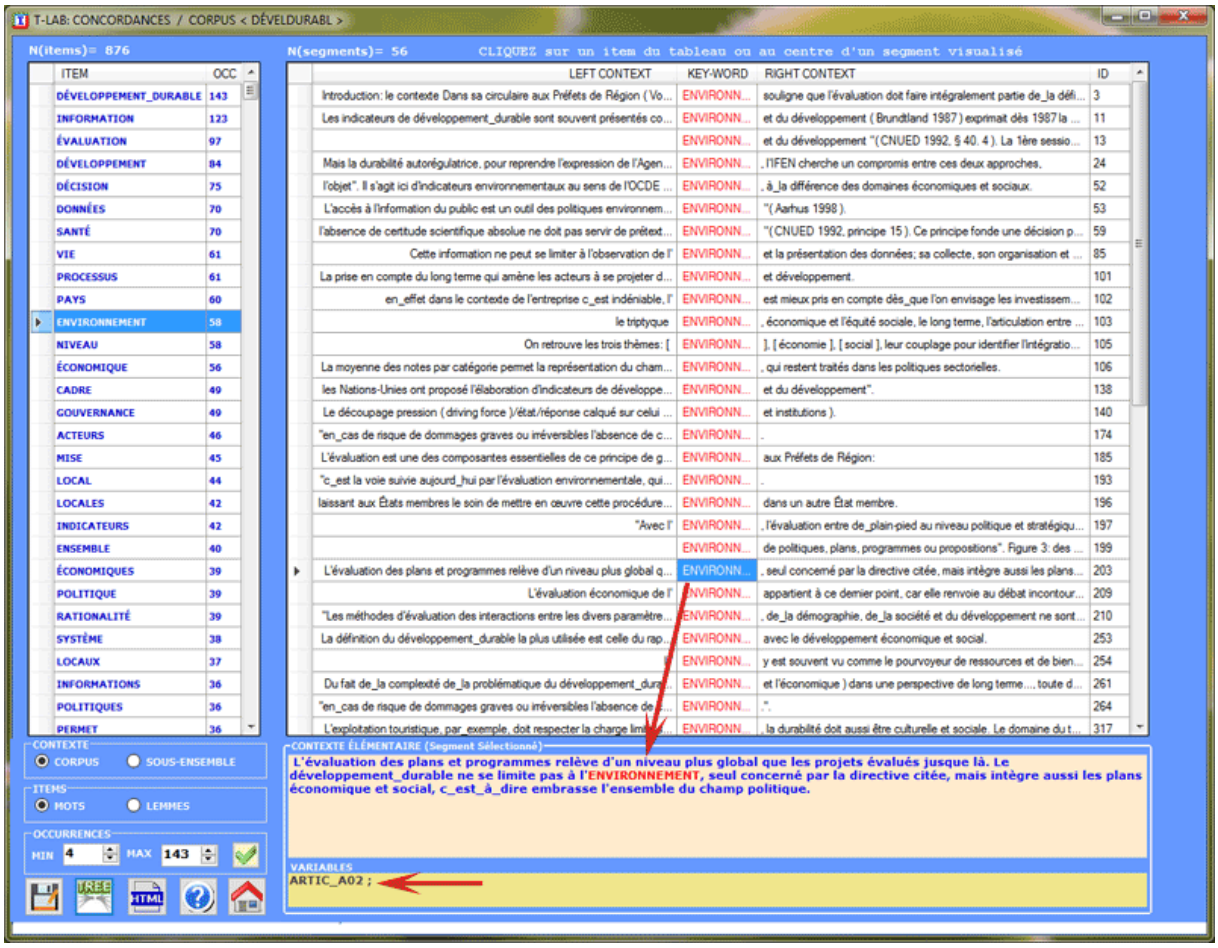
**** *ARTIC_A01
ENVIRONNEMENT et du développement "(CNUED 1992, § 40. 4). La 1ère session de_la Commission du développement_durable de 1993, proposait de "développer des indicateurs réalistes, utilisables et compréhensibles "(CSD1 1993).

**** *ARTIC_A01
Mais la durabilité autorégulatrice, pour reprendre l'expression de l'Agenda 21, nécessiterait la mise en_place d'indicateurs élaborés au cœur même du processus de décision. Bien qu'impliqué dans l'évaluation pour les Nations-Unies (rapport National pour la CDD) et chargé de l'Etat de l'**ENVIRONNEMENT**, l'IFEN cherche un compromis entre ces deux approches,

**** *ARTIC_A01
l'objet". Il s'agit ici d'indicateurs environnementaux au sens de l'OCDE pression, état, réponse. Les autres volets de_la loi n'évoquent pas les indicateurs de développement_durable. Mais l'information joue un rôle spécifique pour l'**ENVIRONNEMENT**, à_la différence des domaines économiques et sociaux.

**** *ARTIC_A01
L'accès à l'information du public est un outil des politiques environnementales. Ce principe est inscrit dans des textes comme la "Convention sur l'accès à l'information, la participation du public au processus décisionnel et l'accès à la

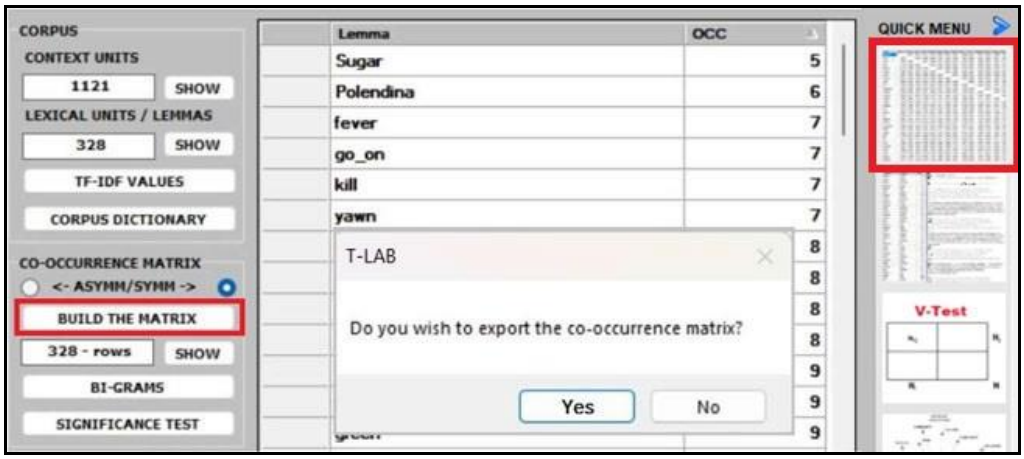
D'ailleurs, en cliquetant le centre d'un segment montré il est possible de visualiser tout son contenu et de vérifier les catégories utilisées dans ses lignes de codage (voir ci-dessous).



Co-occurrence Toolkit

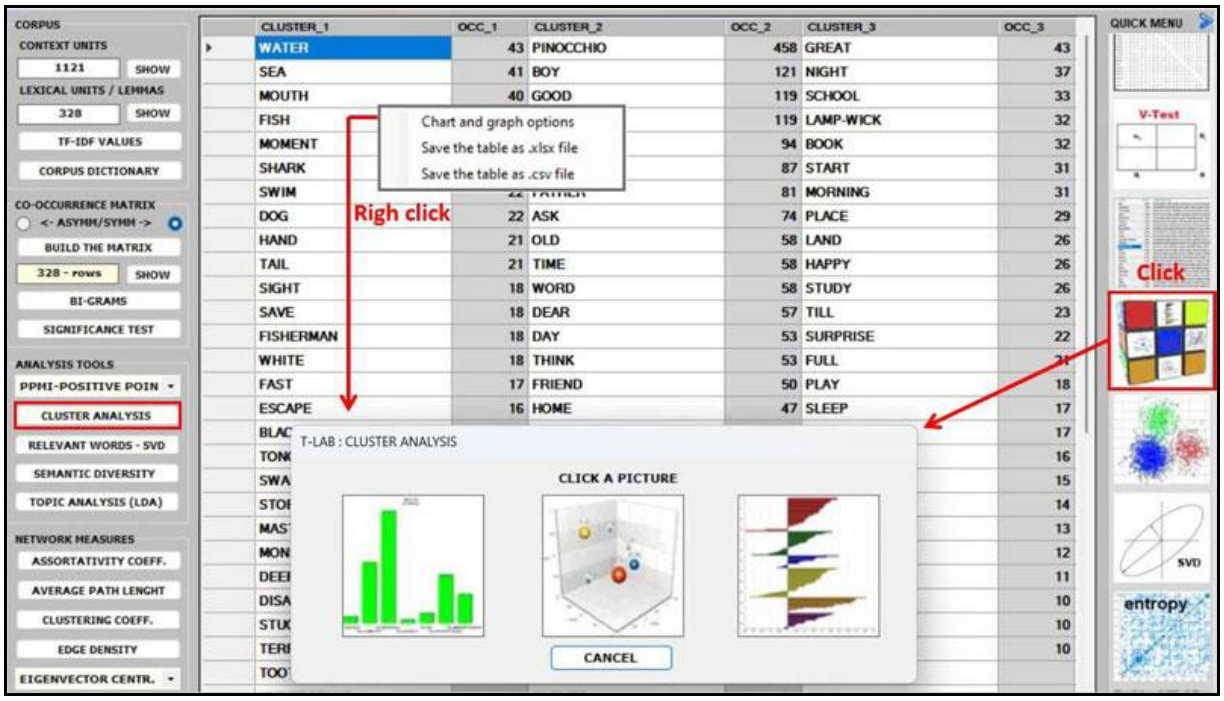
N.B .: Cette section est uniquement disponible en anglais.

This tool, which can be used for a variety of tasks, offers a set of techniques for building and analysing **word co-occurrence matrices** with up to 5,000 columns.
The matrices to be built can be both **symmetric** and **asymmetric**, and they can represent the co-occurrences of the words either within the whole **corpus** or within a **subset** of it.



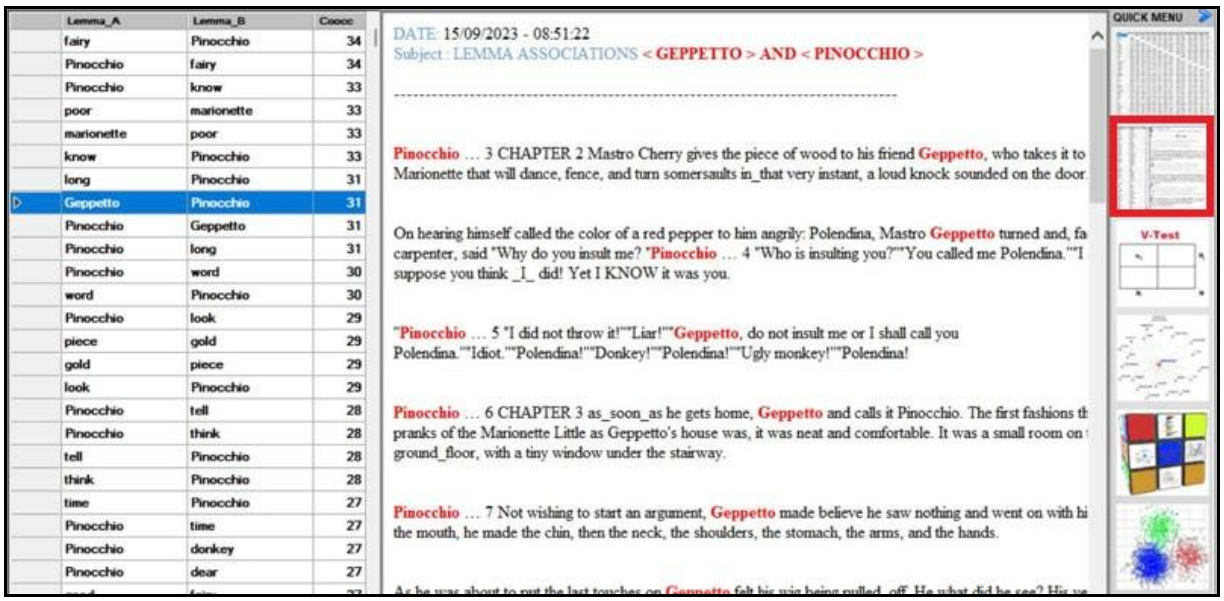
N.B.: In the case of word co-occurrences, the difference between symmetric and asymmetric matrices is that symmetric matrices assume that the order of words does not matter (i.e., they are represented as undirected graphs where the values in a row and a column are the same), while asymmetric matrices take into account the direction of co-occurrence and, for this reason, are represented as directed graph where the values in a row (i.e., successor) and a column (i.e., predecessor) are not necessarily the same.

Whichever tool you are using, the way to export tables and graphs is very simple (see picture below).



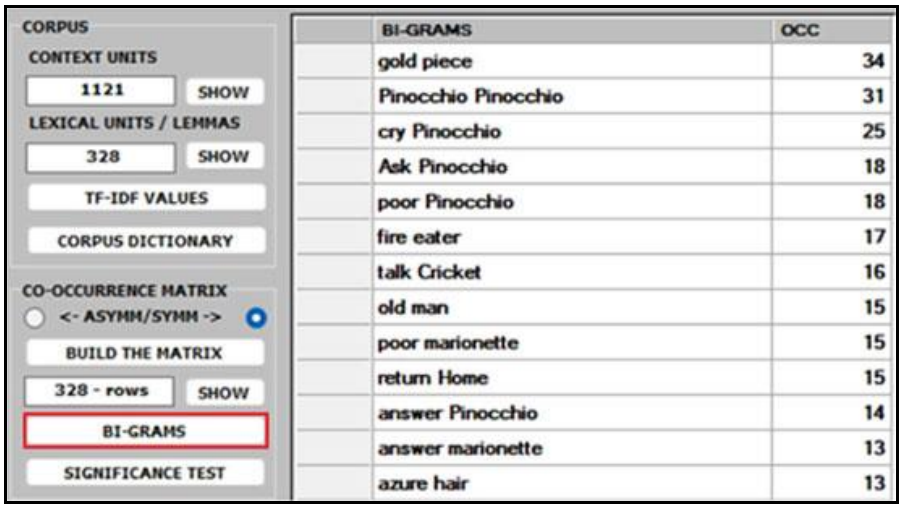
After building any co-occurrence matrix, the user is allowed to extract the relevant information by using about fifteen options listed on the left menu (see the above picture).

N.B.:
- all the below pictures have been obtained by analysing the English version of "The Adventures of Pinocchio" (by Carlo Collodi) and its symmetric word co-occurrence matrix.
- all items in the tables are 'lemmas' because a **T-LAB** lemmatization has been performed on the Pinocchio corpus first.
- whatever matrix you are analysing, it is always possible to check the text segments in which pairs of words co-occur (see picture below).

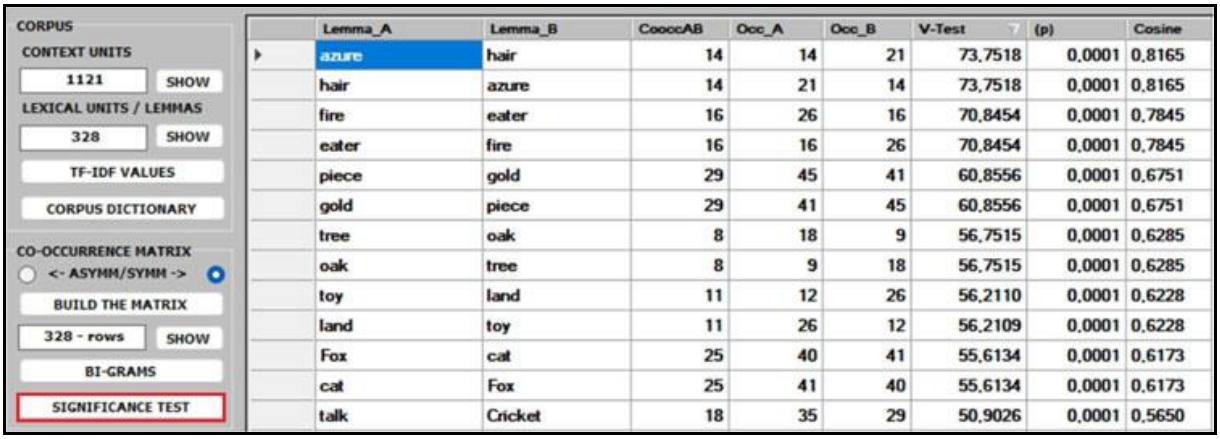


Below are the descriptions of the various analysis options:

- both the **BI-GRAMS** and the **SIGNIFICANCE TEST** extract pairs of words (e.g., collocations) which can be relevant for customizing the corpus dictionary an also for detecting small groups of related words which can affect any cluster analysis (see pictures below).

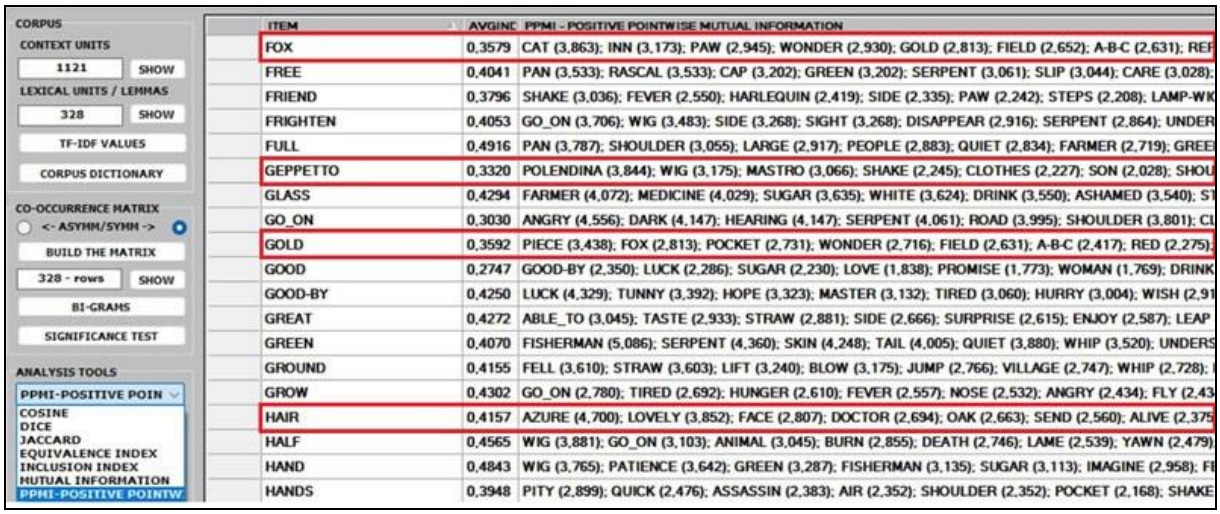


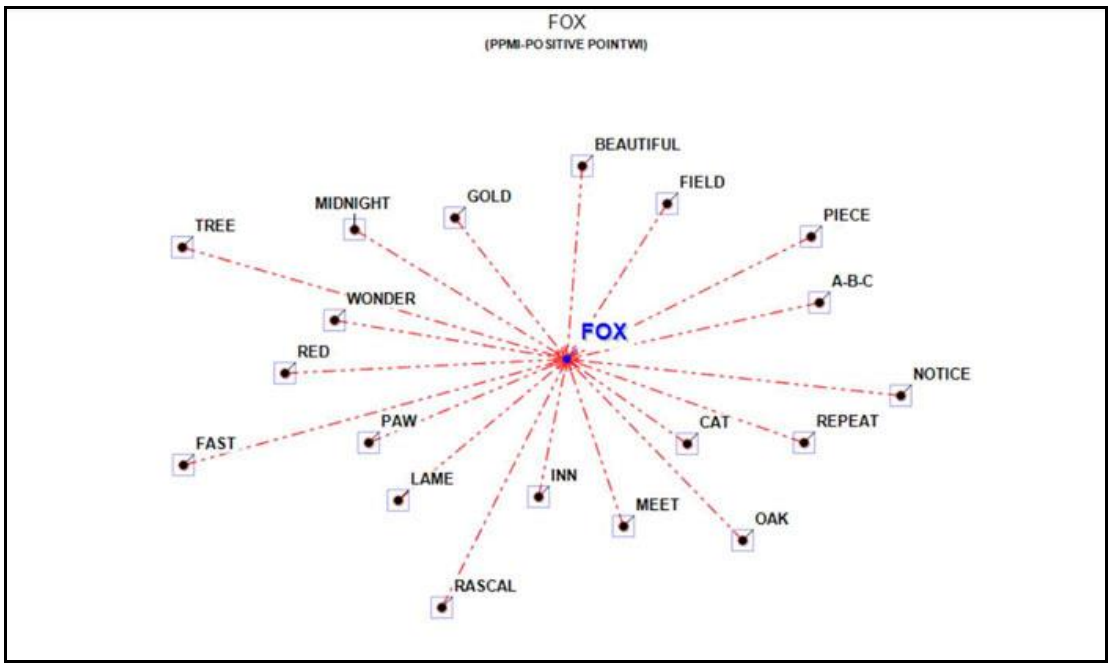
BI-GRAMS	OCC
gold piece	34
Pinocchio Pinocchio	31
cry Pinocchio	25
Ask Pinocchio	18
poor Pinocchio	18
fire eater	17
talk Cricket	16
old man	15
poor marionette	15
return Home	15
answer Pinocchio	14
answer marionette	13
azure hair	13



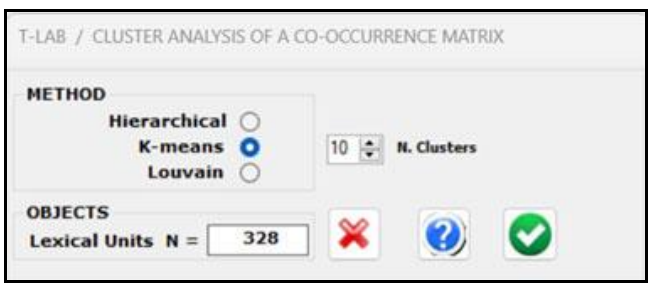
Lemma_A	Lemma_B	CooccAB	Occ_A	Occ_B	V-Test	(p)	Cosine
azure	hair	14	14	21	73,7518	0,0001	0,8165
hair	azure	14	21	14	73,7518	0,0001	0,8165
fire	eater	16	26	16	70,8454	0,0001	0,7845
eater	fire	16	16	26	70,8454	0,0001	0,7845
piece	gold	29	45	41	60,8556	0,0001	0,6751
gold	piece	29	41	45	60,8556	0,0001	0,6751
tree	oak	8	18	9	56,7515	0,0001	0,6285
oak	tree	8	9	18	56,7515	0,0001	0,6285
toy	land	11	12	26	56,2110	0,0001	0,6228
land	toy	11	26	12	56,2109	0,0001	0,6228
Fox	cat	25	40	41	55,6134	0,0001	0,6173
cat	Fox	25	41	40	55,6134	0,0001	0,6173
talk	Cricket	18	35	29	50,9026	0,0001	0,5650

- the **ASSOCIATIONS** option, in addition to the indexes used by other **T-LAB** tools (see Word Associations and Co-Word Analysis), includes the **PPMI** (i.e., Positive Pointwise Mutual Information), which is a measure of how much more likely two words are to co-occur than by chance, based on their probabilities in a text corpus. It can be used to distinguish between words that are simply co-occurring by chance and words that are semantically related. It can also reduce the effect of high-frequency words that co-occur with many other words by chance. Moreover, unlike other indexes (e.g., Cosine, Dice, Jaccard etc.) its maximum value is not '1' and its upper bound can vary.



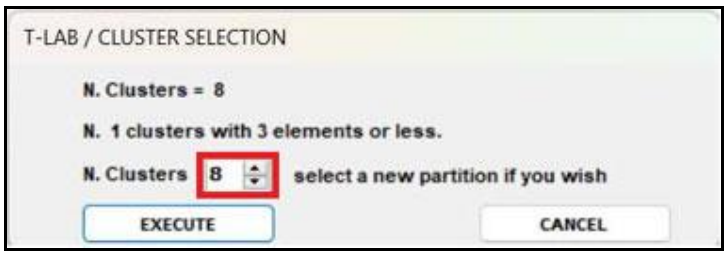


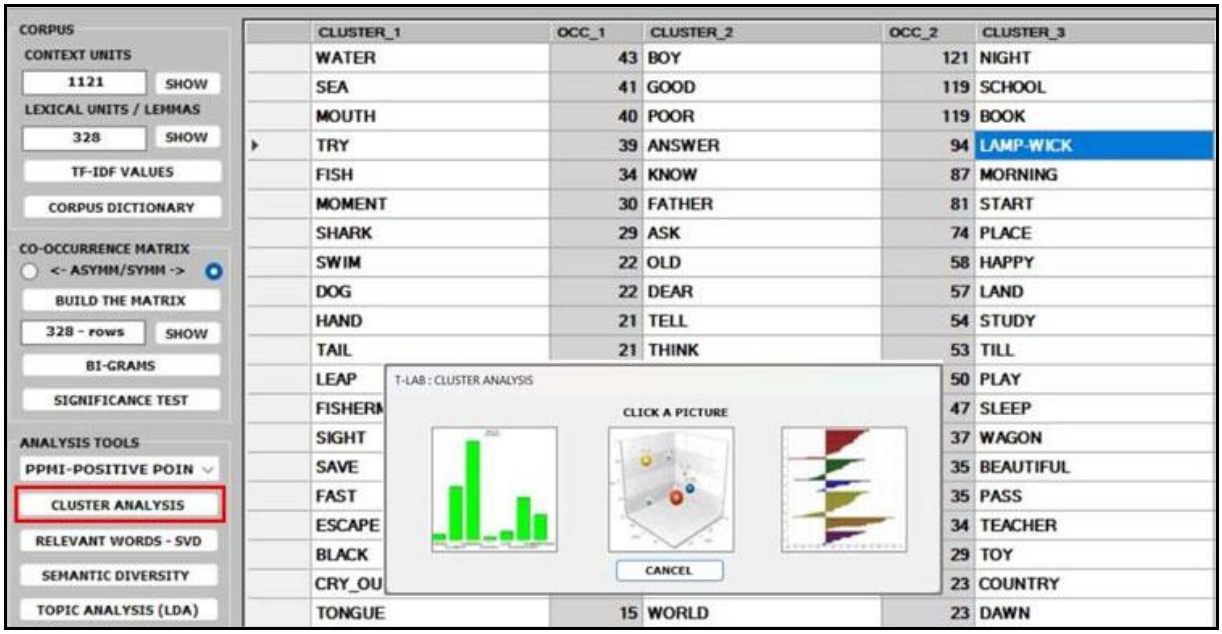
- the **CLUSTER ANALYSIS** offers three methods for analysing a word co-occurrence matrix: **Hierarchical**, **K-means** and **Louvain**.



All the above three methods use vectors which are normalized by the cosine coefficient, and one of them (i.e., the K-means) performs the clustering on the first 10 dimensions obtained by a SVD (i.e., Singular Value Decomposition) of the normalized word co-occurrence matrix.
To evaluate the quality of clustering results, **T-LAB** provides the **Silhouette** scores for each data point. Moreover, when clicking the '**Q**' button located at the bottom left corner of the screen, the user is allowed to obtain three different quality indices (i.e.: Calinski-Harabasz, Dunn and ICC-rho).

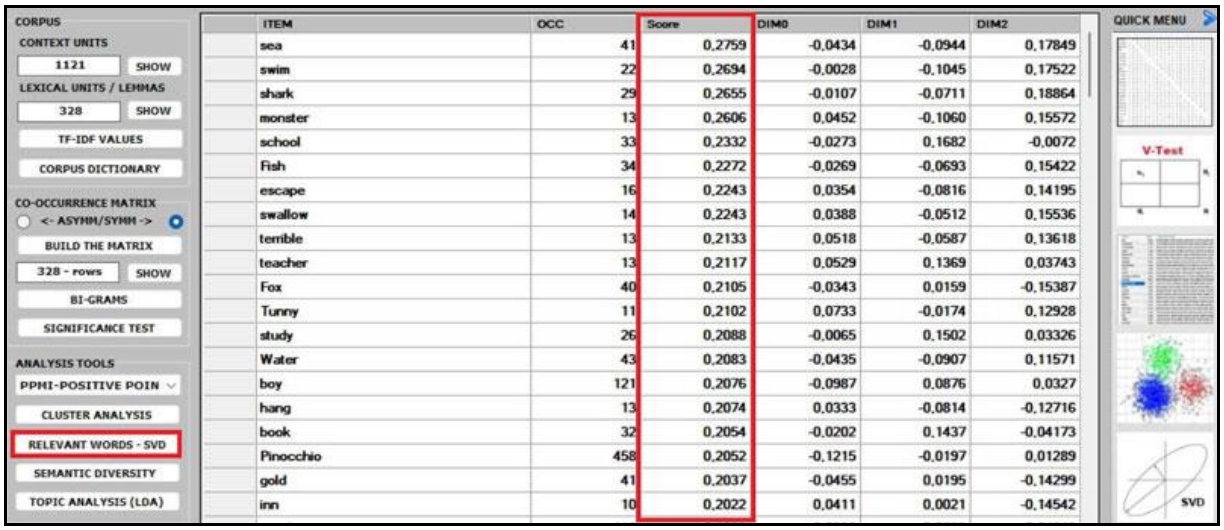
N.B.:
- Depending on the clustering method, the **relationships between words within each cluster** can be visualized through different types of charts and graphs.
- When performing a hierarchical clustering, the user is allowed to change the number of clusters (i.e., the cluster partition) within a range from 3 to 20.



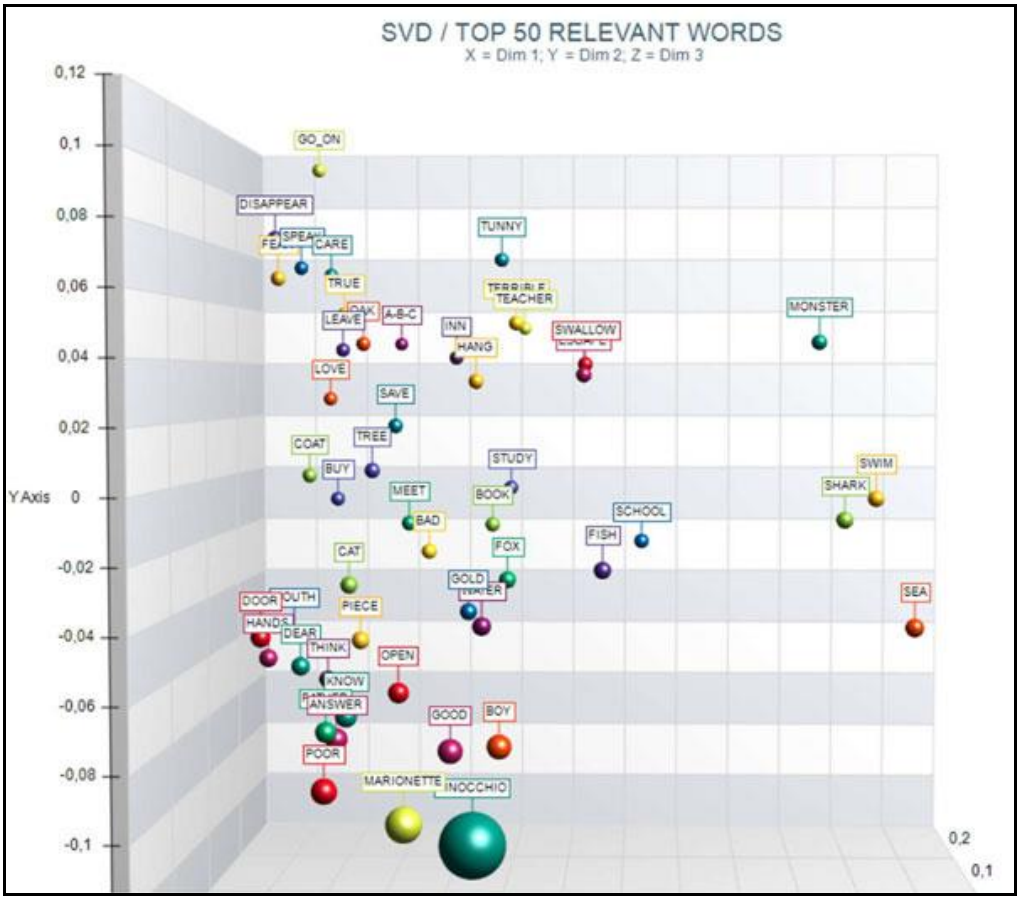


- the **RELEVANT WORDS - SVD** provides a relevance score for each word, which is computed by summing the square of its first 3 dimensions (i.e., the eigenvectors), each one multiplied by its corresponding singular value, and then by computing the square root of that sum.

This means that the words with the higher scores are the farthest from the point of origin, which is the point where the horizontal axis (x-axis) and the vertical axis (y-axis) intersect. And, for this reason, they are the words that most contribute to organizing semantic polarizations, which can also have emotional connotations.
N.B.: In this case, the SVD is performed on a centered matrix and therefore it is equivalent to PCA.

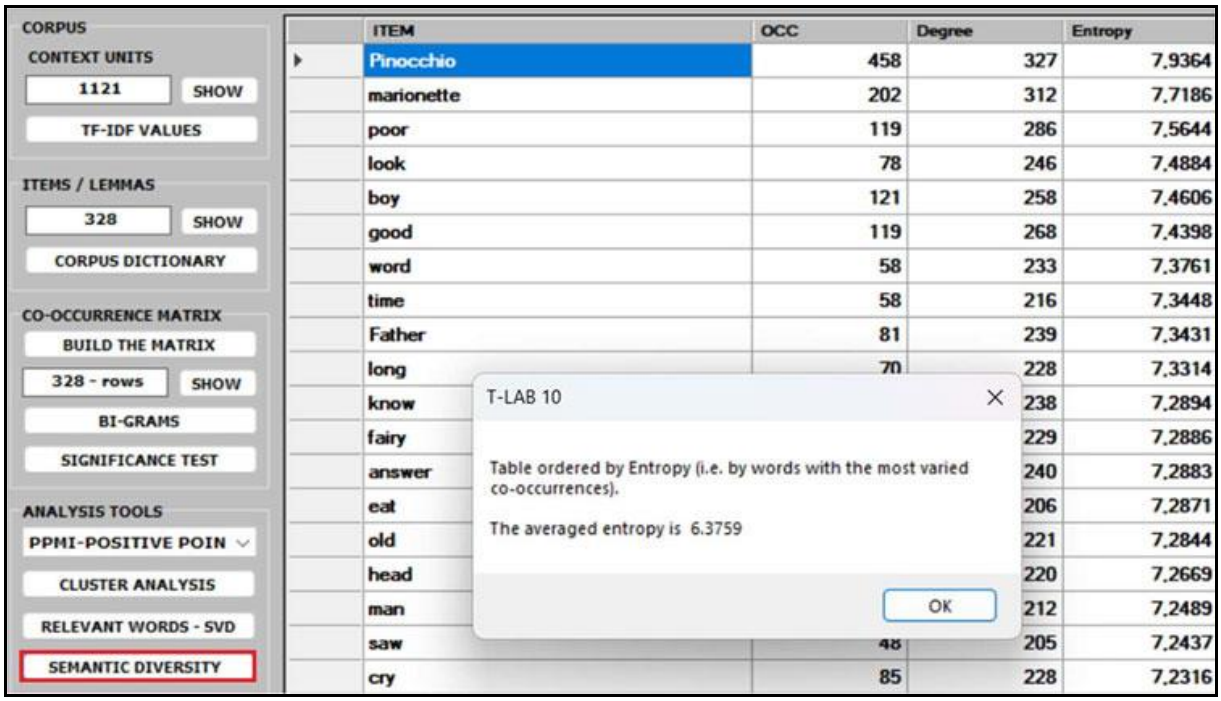


ITEM	OCC	Score	DIM0	DIM1	DIM2
sea	41	0,2759	-0,0434	-0,0944	0,17849
swim	22	0,2694	-0,0028	-0,1045	0,17522
shark	29	0,2655	-0,0107	-0,0711	0,18864
monster	13	0,2606	0,0452	-0,1060	0,15572
school	33	0,2332	-0,0273	0,1682	-0,0072
Fish	34	0,2272	-0,0269	-0,0693	0,15422
escape	16	0,2243	0,0354	-0,0816	0,14195
swallow	14	0,2243	0,0388	-0,0512	0,15536
terrible	13	0,2133	0,0518	-0,0587	0,13618
teacher	13	0,2117	0,0529	0,1369	0,03743
Fox	40	0,2105	-0,0343	0,0159	-0,15387
Tunny	11	0,2102	0,0733	-0,0174	0,12928
study	26	0,2088	-0,0065	0,1502	0,03326
Water	43	0,2083	-0,0435	-0,0907	0,11571
boy	121	0,2076	-0,0987	0,0876	0,0327
hang	13	0,2074	0,0333	-0,0814	-0,12716
book	32	0,2054	-0,0202	0,1437	-0,04173
Pinocchio	458	0,2052	-0,1215	-0,0197	0,01289
gold	41	0,2037	-0,0455	0,0195	-0,14299
inn	10	0,2022	0,0411	0,0021	-0,14542



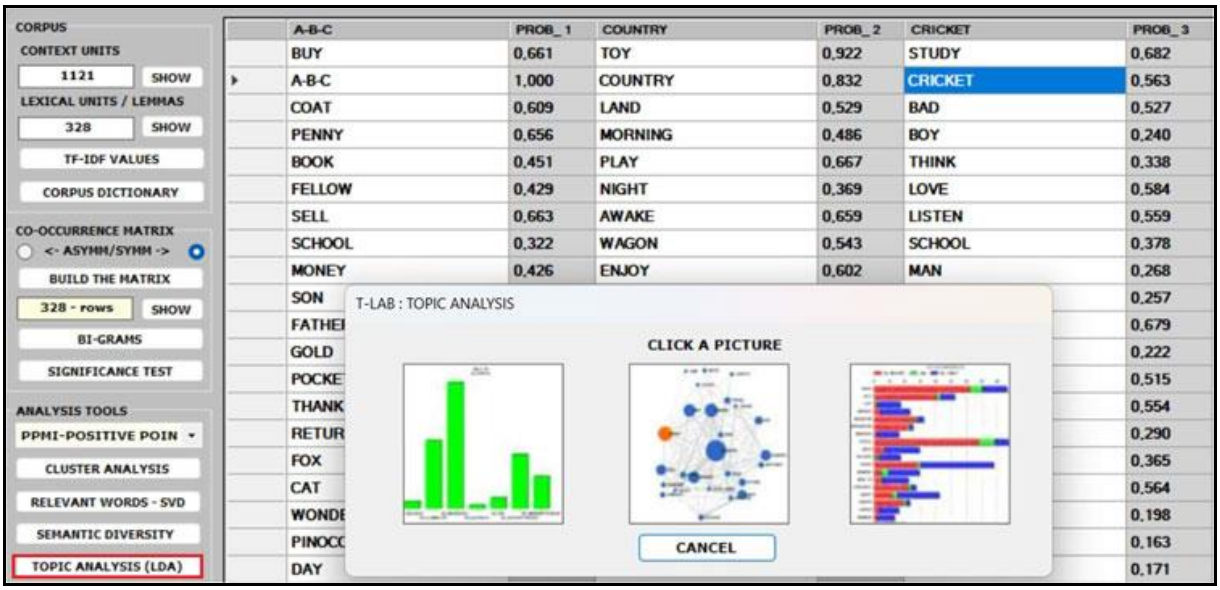
- the **SEMANTIC DIVERSITY** of each word (i.e., its ability to have links with many other words) is measured by means of the **entropy** index.

N.B.: The average entropy of the word co-occurrence matrix can be used to quantify the 'complexity' of a text, since more complex texts (i.e., texts in which many words cooccur with a variety of other words) tend to have higher entropy than simpler texts (i.e., texts in which many words cooccur with only a few other words and – for that reason – are more predictable). And, since high entropy corresponds to low predictability, it may be also interesting to check which words in a text have higher predictability values (i.e., low entropy).



- the **TOPIC ANALYSIS** of the word co-occurrence matrix uses the same algorithm of the **T-LAB** Modeling of Emerging Themes tool (i.e., Latent Dirichlet Allocation and the Gibbs Sampling); however, in this case, both the indexes of the matrix (i.e., the 'i' and the 'j') refer to the same words and the values correspond to their co-occurrences. As can be verified, the results of this approach are quite interesting and consistent.

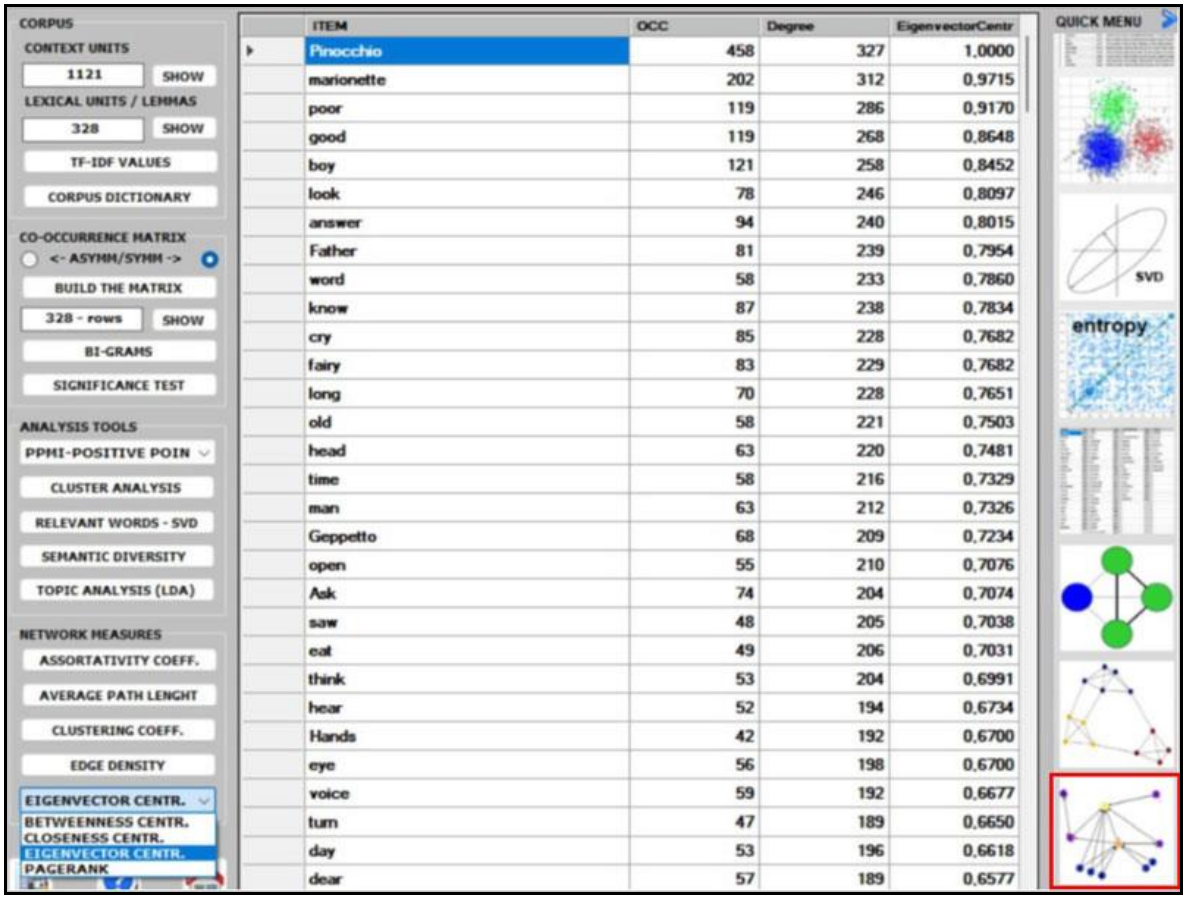
  N.B.: In the table below, the words are ordered by their frequency within each topic.



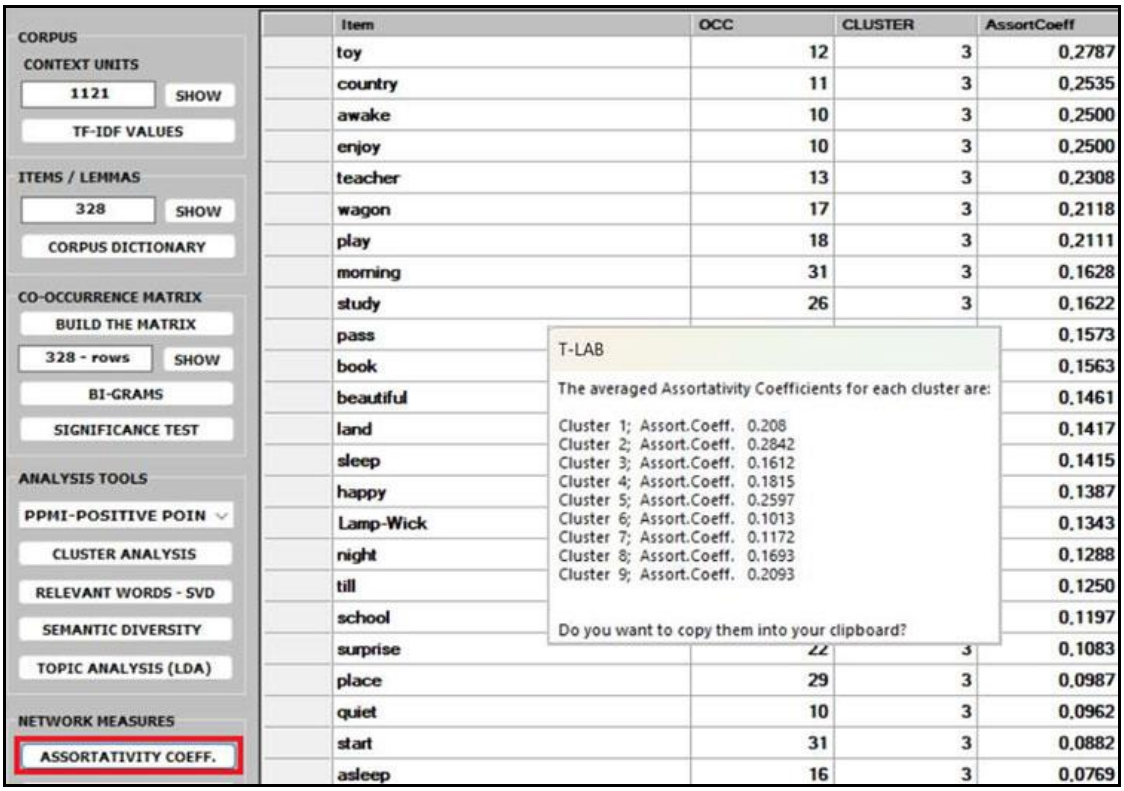
- Regarding the five **CENTRALITY MEASURES** (i.e., Betweenness centrality, Closeness centrality, Eigenvector centrality, Katz centrality and PageRank centrality) we observe

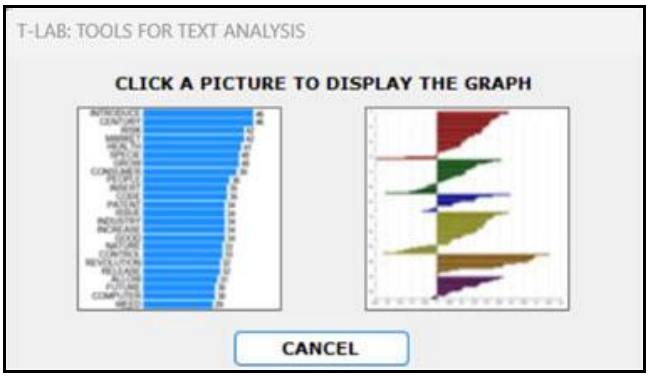
that, especially in the case of a symmetric word co-occurrence matrix, they are closely related to each other. Moreover, they usually rank more highly the words with higher occurrence values. The only exception seems to be the Betweenness centrality. In fact, it is possible for a vertex to have high betweenness centrality (i.e., to be able to connect important parts of the network) without having high indegree or high outdegree.

N.B.:
- All definitions of centrality measures, as well as their algorithms, can be easily checked on Wikipedia.
- In **T-LAB**, all the results of centrality measures are normalized to the maximum value. This means that all the results are between 0 and 1, which makes them easier to compare.

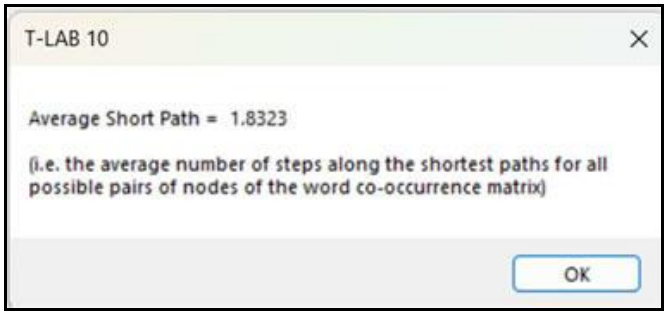


- the **ASSORTATIVITY COEFFICIENT** is a measure of how likely nodes of a certain type are to be connected to other nodes of the same type (i.e., 'similar' in some respects). In the case of **T-LAB**, the types refer to the results of a previous cluster analysis. Therefore, (a) if– for any 'i' node – the assortativity coefficient is positive and high, then it indicates that the node is strongly connected with other nodes of the same cluster; (b) if – for any 'k' cluster - the average assortativity coefficient is positive and high, then it indicates that the nodes which belong to the cluster are strongly connected with each other; (c) a global average high positive assortativity coefficient indicates that the clustering algorithm has successfully grouped nodes based on their links within the cluster they belong to. This means that nodes within the same cluster are more likely to be connected to each other than nodes from different clusters.



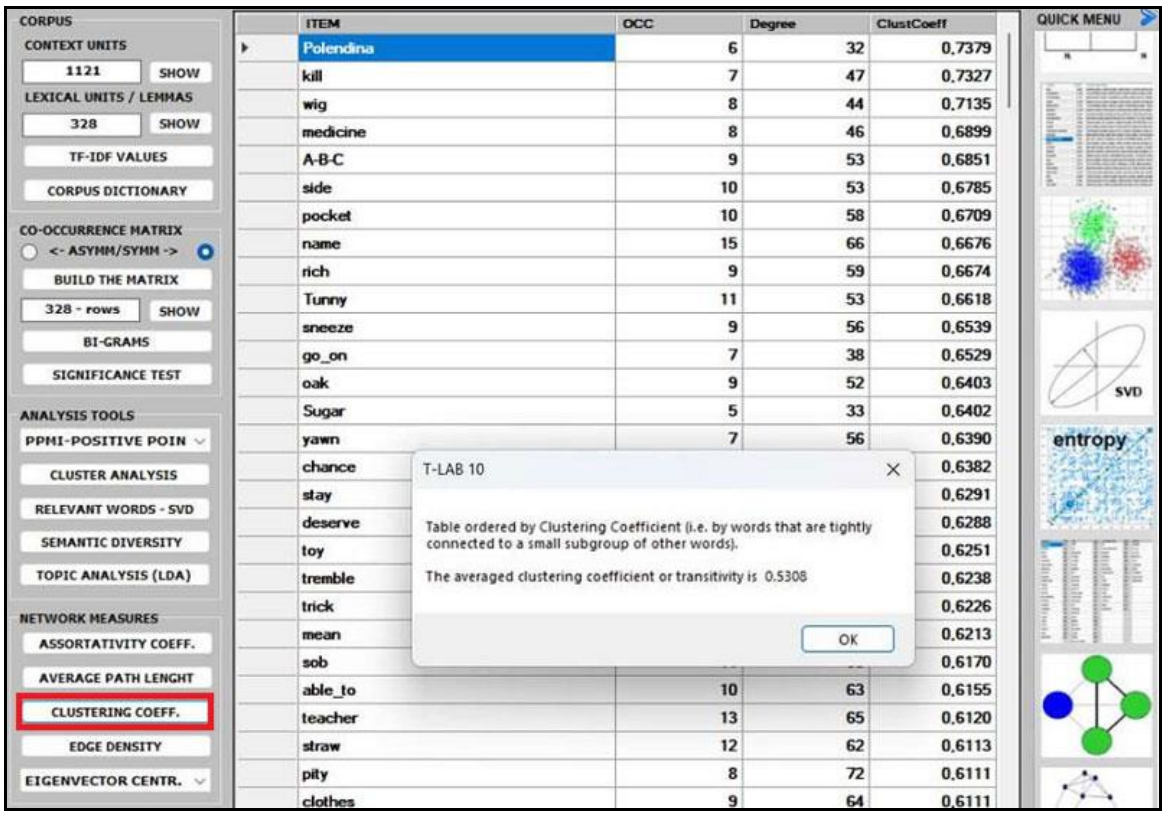


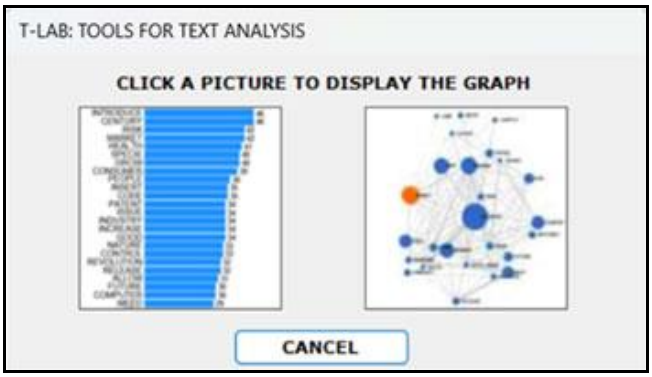
- the **AVERAGE PATH LENGTH** (or average short path), in this case, is defined as the average number of steps along the shortest paths for all possible pairs of nodes of the word co-occurrence matrix.



- the **CLUSTERING COEFFICIENT** deserves special attention. In fact, the 'local' clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together and to pair up with each other (i.e., something like 'The friend of my friend is my friend.'). In other words, the clustering coefficient of a node (i.e., word) quantifies how close its neighbours (i.e., other words) are to being a tightly connected subgroup (i.e., a clique). It is computed as the proportion of the 'actual' connections among its neighbours compared with the number of all its 'possible' connections. Its maximum value is '1', and the average clustering coefficient of all nodes it is also known as 'transitivity' of the network.

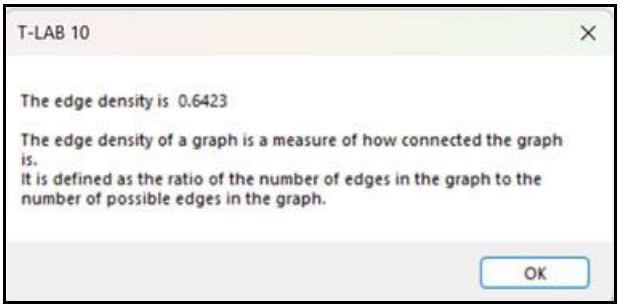
  N.B.:
  - When a network has a large clustering coefficient and a small average path length it can be considered a 'small world' (see Wikipedia).





- the **EDGE DENSITY** is a measure of how connected the graph is. It is defined as the ratio of the actual number of edges in the graph to the possible number of edges in the graph.
  A high edge density indicates that the nodes in the graph are more likely to be connected to each other. This means that there are many paths between any two nodes in the graph. A low edge density indicates that the nodes in the graph are more likely to be disconnected from each other. This means that there are few paths between any two nodes in the graph.

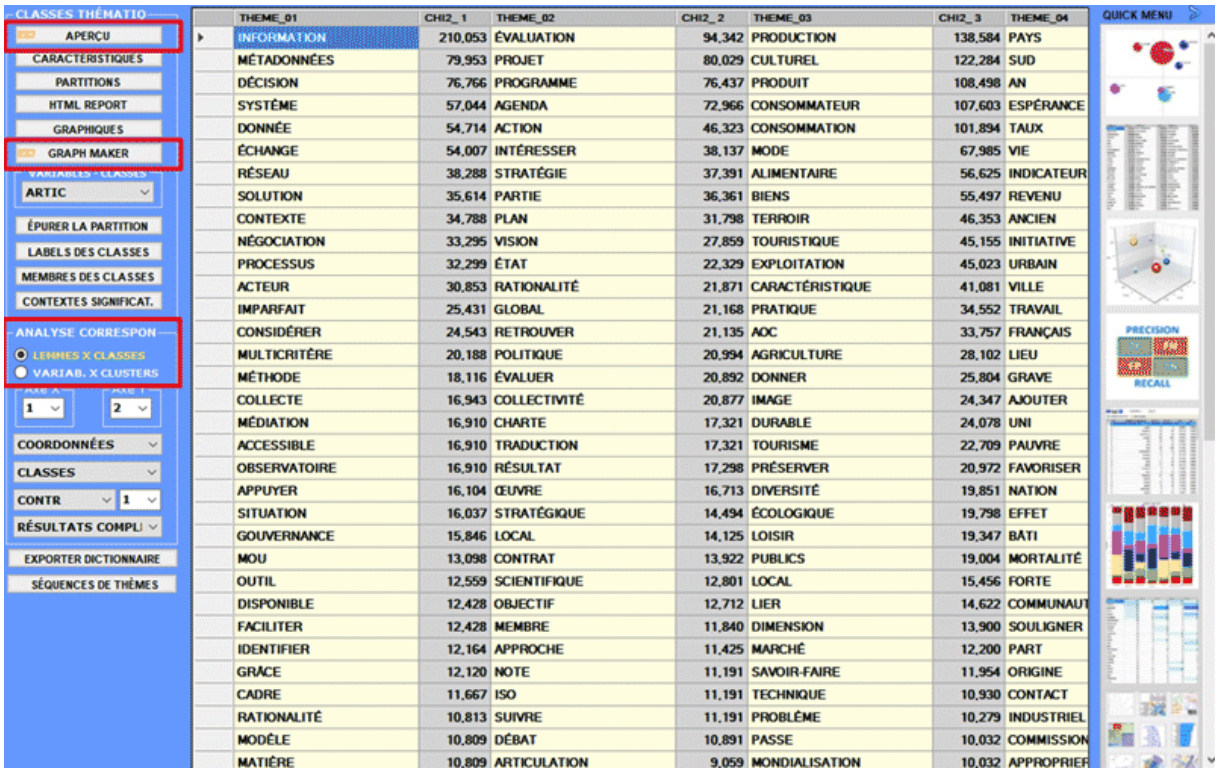
N.B.: It appears that there is a positive correlation between edge density and clustering coefficient. In fact, both measures refer to the connectivity of a graph and can be used to compare the properties of different graphs (i.e., in this case, the properties of different co-occurrence matrices).



ANALYSES THEMATIQUES

Analyse Thématique des Contextes Élémentaires

N.B.: Les images de cette section font référence à une version précédente de **T-LAB**. En **T-LAB 10**, l'aspect est légèrement différent. En outre: a) il y a un nouveau bouton (**TREE MAP PREVIEW**) qui permet à l'utilisateur de créer plusieurs graphiques dynamiques au format HTML; b) le bouton DENDROGRAMME a été remplacé par l'outil **GRAPH MAKER**; c) un autre tableau qui montre en colonnes différentes les mots typiques de chaque cluster est disponible; d) on peut effectuer d'autres analyses des correspondances entre les clusters thématiques et chaque variable disponible; e) une galerie d'images à accès rapide qui fonctionne comme un menu supplémentaire permet de basculer entre les différentes sorties en un seul clic.
Certaines de ces nouvelles fonctionnalités sont mises en évidence dans l'image ci-dessous.



THEME_01	CHI2_1	THEME_02	CHI2_2	THEME_03	CHI2_3	THEME_04
INFORMATION	210,053	ÉVALUATION	94,342	PRODUCTION	138,584	PAYS
MÉTADONNÉES	79,953	PROJET	80,029	CULTUREL	122,284	SUD
DÉCISION	76,766	PROGRAMME	76,437	PRODUIT	108,498	AN
SYSTÈME	57,044	AGENDA	72,966	CONSOMMATEUR	107,603	ESPÉRANCE
DONNÉE	54,714	ACTION	46,323	CONSOMMATION	101,894	TAUX
ÉCHANGE	54,007	INTÉRESSER	38,137	MODE	67,985	VIE
RÉSEAU	38,288	STRATÉGIE	37,391	ALIMENTAIRE	56,625	INDICATEUR
SOLUTION	35,614	PARTIE	36,361	BIENS	55,497	REVENU
CONTEXTE	34,788	PLAN	31,798	TERROIR	46,353	ANCIEN
NÉGOCIATION	33,295	VISION	27,859	TOURISTIQUE	45,155	INITIATIVE
PROCESSUS	32,299	ÉTAT	22,329	EXPLOITATION	45,023	URBAIN
ACTEUR	30,853	RATIONALITÉ	21,871	CARACTÉRISTIQUE	41,081	VILLE
IMPARFAIT	25,431	GLOBAL	21,168	PRATIQUE	34,552	TRAVAIL
CONSIDÉRER	24,543	RETROUVER	21,135	AOC	33,757	FRANÇAIS
MULTICRITÈRE	20,188	POLITIQUE	20,994	AGRICULTURE	28,102	LIEU
MÉTHODE	18,116	ÉVALUER	20,892	DONNER	25,804	GRAVE
COLLECTE	16,943	COLLECTIVITÉ	20,877	IMAGE	24,347	AJOUTER
MÉDIATION	16,910	CHARTE	17,321	DURABLE	24,078	UNI
ACCESSIBLE	16,910	TRADUCTION	17,321	TOURISME	22,709	PAUVRE
OBSERVATOIRE	16,910	RÉSULTAT	17,298	PRÉSERVER	20,972	FAVORISER
APPUYER	16,104	ŒUVRE	16,713	DIVERSITÉ	19,851	NATION
SITUATION	16,037	STRATÉGIQUE	14,494	ÉCOLOGIQUE	19,798	EFFET
GOUVERNANCE	15,846	LOCAL	14,125	LOISIR	19,347	BÂTI
MOU	13,098	CONTRAT	13,922	PUBLICS	19,004	MORTALITÉ
OUTIL	12,559	SCIENTIFIQUE	12,801	LOCAL	15,456	FORTE
DISPONIBLE	12,428	OBJECTIF	12,712	LIER	14,622	COMMUNAUT
FACILITER	12,428	MEMBRE	11,840	DIMENSION	13,900	SOULIGNER
IDENTIFIER	12,164	APPROCHE	11,425	MARCHÉ	12,200	PART
GRÂCE	12,120	NOTE	11,191	SAVOIR-FAIRE	11,954	ORIGINE
CADRE	11,667	ISO	11,191	TECHNIQUE	10,930	CONTACT
RATIONALITÉ	10,813	SUIVRE	11,191	PROBLÈME	10,279	INDUSTRIEL
MODÈLE	10,809	DÉBAT	10,891	PASSE	10,032	COMMISSION
MATIÈRE	10,809	ARTICULATION	9,059	MONDIALISATION	10,032	APPROPRIER

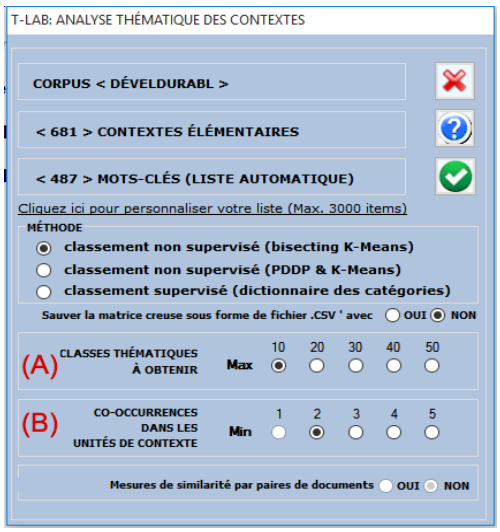
Cet outil **T-LAB** permet d'obtenir et d'explorer une **représentation des contenus du corpus** à travers un nombre restreint et significatif de **classes thématiques** (de 3 à 50), dont chacune:

a) est formée par un ensemble de **contextes élémentaires** (phrases, paragraphes, fragments de texte, réponses à des questions ouvertes) caractérisés par les mêmes patterns de mots-clés;

b) peut être décrite à travers les **unités lexicales** (mots, lemmes ou catégories) et les **variables** (si elles sont présentes) qui caractérisent les unités de contexte dont elle est composée.

A plusieurs égards, on peut affirmer que le résultat de l'analyse propose une carte des **isotopies** (iso = égal; topoi = lieux), dont chacune correspond a un thème "générique" ou "spécifique" (Rastier, 2002: 204) caractérisé par la co-occurrence de traits sémantiques.

Le processus d'analyse peut être effectué au moyen d'une méthode de **clustering non supervisée** (dans le cas particulier, un algorithme bisecting K-Means) ou bien à travers une **classification supervisée** (c'est-à-dire une approche top-down). Lorsqu'on choisit la deuxième (c'est-à-dire la classification supervisée), on vous demande d'importer un dictionnaire des catégories, qu'il soit aussi bien créé à travers une précédente analyse **T-LAB** que construit par l'utilisateur.



Une boîte de dialogue (voir ci-dessus) nous permet de fixer quelques paramètres de l'analyse.

En particulier:

- le paramètre (A) nous permet de fixer le nombre maximum de classes à inclure dans les outputs **T-LAB**.
- le paramètre (B) nous permet d'exclure de l'analyse les unités de contexte qui ne contiennent pas un nombre minimum de mots-clés inclus dans la liste utilisée.

N.B.:
- Lorsqu'on sélectionne l'option «classification supervisée», puisque le numéro de clusters à obtenir coïncide avec le numéro de catégories présentes dans le dictionnaire, le paramètre « A » n'est pas disponible
- Les deux paramètres ci-dessus produisent des changements significatifs des résultats seulement quand le nombre des unités de contexte est très grand et/ou quand il s'agit de textes courts.

Dans le cas de **classification non supervisée** (option de default), la procédure d'analyse est constituée par les étapes suivantes:

a - construction d'un tableau unités de contexte x unités lexicales (jusqu'à 300.000 lignes x 3.000 colonnes) avec valeurs du type présence/absence;
b - calcul du poids **TF-IDF** et usage de la norme euclidienne (longueur des vecteurs = 1);
c - classification des unités de contexte (mesure de similitude: coefficient du cosinus; méthode

de classification: bisecting K-means; références: Steinbach, Karypis, & Kumar, 2000; Savaresi, Booley, 2001);
d - archivage des partitions obtenues et, pour chacune d'entre elles:
e - construction d'un tableau de contingence unités lexicales x classes (n x k);
f - test du Chi-Deux appliqué à tous les croisements unités lexicales x classes.
g - analyse des correspondances du tableau de contingence unités lexicales x classes (références: Benzécri, 1984; Greenacre, 1984; Lebart, Salem, 1994).

N.B. : A partir de T-LAB 2016, la clusterisation des unités de contexte (voir l'étape «c» ci-dessus) peut être obtenue soit en utilisant l'algorithme bisecting K-means algorithm (1), soit en utilisant une version non centrée de l'algorithme PDDP (Principal Direction Divisive Partitioning) proposé par D. Booley (1998) pour sélectionner les centroïdes des de chaque bisection K-means.
La principale différence entre les deux algorithmes reste dans la méthode à travers la quelle les deux centroïdes de chaque bisection sont obtenus; en effet, dans le premier cas (1) ils sont le résultat d'une réitération, pendant que dans le second cas (2) ils sont obtenus par SVD (i.e. Singular Value Decomposition), c'est-à-dire par un algorithme 'one-shot' (voir Savaresi, S.M., & Boley, D.L.,2004).

Ainsi donc, cette procédure effectue un type d'**analyse des co-occurrences** (étape a-b-c) et ensuite un type d'**analyse comparative** (e-f-g). En particulier, l'analyse comparative utilise comme colonnes du tableau de contingence les modalités de la "nouvelle variable" obtenue par l'analyse des co-occurrences (modalités de la nouvelle variable = classes thématiques).

Dans le cas de **classification supervisée,** les phases de l'analyse comparative sont les mêmes (voir ci-dessus e-f-g), tandis que l'analyse des co-occurrences est réalisée comme suit:

a) normalisation des seed vectors (c'est-à-dire des profils des co-occurrences) correspondant aux catégories "k" du dictionnaire importé;

b) calcul des indices du cosinus et des distances euclidiennes entre chaque unité de contexte "i", et chaque vecteur "germe" "k";

c) attribution de chaque unité de contexte "i" à la classe ou à la catégorie "k" pour laquelle le germe correspondant est le plus proche (dans ce cas, la similitude maximale du cosinus et la distance euclidienne minimale doivent coïncider, autrement **T-LAB** considère l'unité de contexte "i" comme non classifiée).

N.B.: Lorsque l'utilisateur décide de **répéter / appliquer** les résultats d'une analyse précédente (c'est-à-dire **Analyse Thématique des Contextes Élémentaires** ou une **Modélisation des Thèmes Émergents**), T-LAB effectue uniquement une analyse comparative (étapes e-f-g).

À la fin de l'analyse, l'utilisateur peut effectuer aisément les opérations suivantes:

1 - explorer les caractéristiques des classes;
2 - explorer les relations entre classes;
3 - explorer les relations entre classes et variables;
4 - explorer les différentes partitions des classes;
5 - raffiner les résultats de la partition choisie et, au besoin, répéter quelques-unes des étapes ci-dessus (1,2,3);
6 - attribuer des étiquettes aux classes;
7 - vérifier quels sont les contextes élémentaires qui appartiennent à chaque classe;

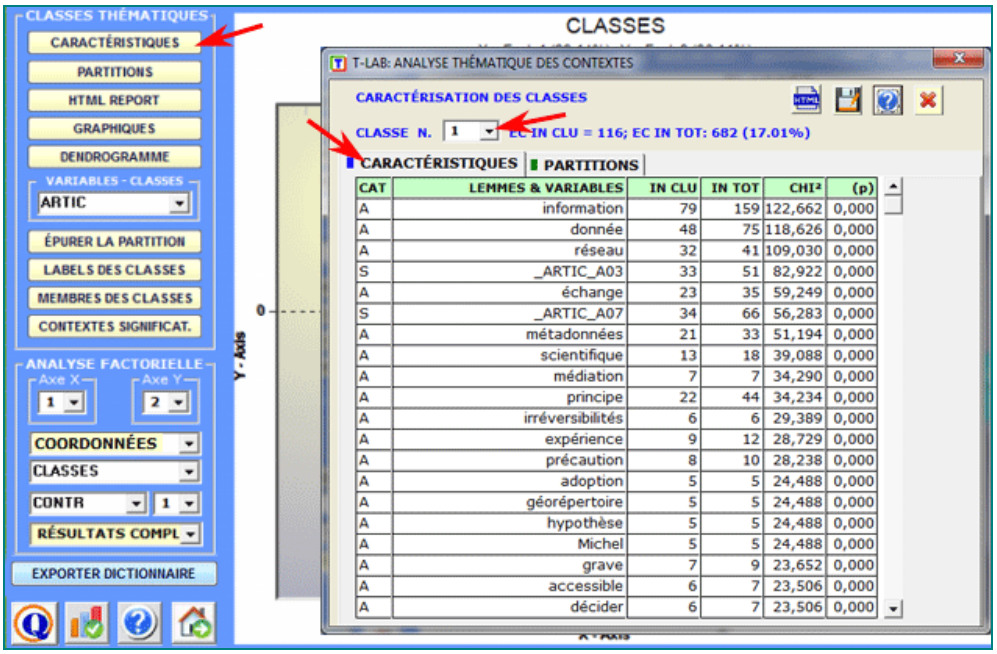
8- vérifier le "poids" de chaque contexte élémentaire au sein de la classe à la quelle il appartient;
9 - obtenir une classification thématique de documents (fournie seulement quand le corpus se compose au moins de 2 documents primaires qui ne sont pas des textes courts comme les réponses aux questions ouvertes);
10 - archiver la partition sélectionnée pour l'explorer avec d'autres outils **T-LAB** ;

11 - exporter un dictionnaire des catégories ;

12 – vérifier la qualité de la partition choisie et la cohérence sémantique des différents thèmes;
13 - en outre, lorsque le corpus est structuré comme un discours ou une conversation, c'est-à-dire lorsque les unités de contexte se succèdent selon un ordre temporel précis, il est possible d'explorer de façon dynamique les **séquences de thèmes** (voir ci-dessous partie finale de cette section).

Dans le détail:

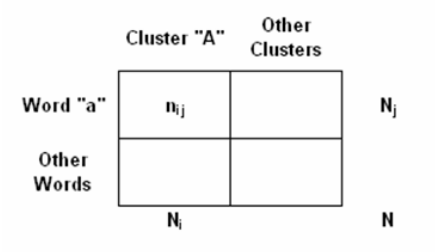
1 - Explorer les caractéristiques des classes



CAT	LEMMES & VARIABLES	IN CLU	IN TOT	CHI²	(p)
A	information	79	159	122,662	0,000
A	donnée	48	75	118,626	0,000
A	réseau	32	41	109,030	0,000
S	_ARTIC_A03	33	51	82,922	0,000
A	échange	23	35	59,249	0,000
S	_ARTIC_A07	34	66	56,283	0,000
A	métadonnées	21	33	51,194	0,000
A	scientifique	13	18	39,088	0,000
A	médiation	7	7	34,290	0,000
A	principe	22	44	34,234	0,000
A	irréversibilités	6	6	29,389	0,000
A	expérience	9	12	28,729	0,000
A	précaution	8	10	28,238	0,000
A	adoption	5	5	24,488	0,000
A	géorépertoire	5	5	24,488	0,000
A	hypothèse	5	5	24,488	0,000
A	Michel	5	5	24,488	0,000
A	grave	7	9	23,652	0,000
A	accessible	6	7	23,506	0,000
A	décider	6	7	23,506	0,000

En cliquant sur le bouton **Caractéristiques**, pour chaque classe apparaissent les unités lexicales et les variables qui la caractérisent; et, pour chacune d'entre elles (unités lexicales ou variables), sont indiquées les valeurs du chi deux et les sommes des contextes élémentaires où elles se trouvent, tant à l'intérieur de la classe sélectionnée (" IN_CLUST ") qu'à l'intérieur de l'ensemble analysé ("IN_TOT"). En outre, dans la colonne "CAT", on indique si la caractéristique a été sélectionnée par l'utilisateur dans la fonction **Configuration Personnalisée** ("A") ou bien si elle a été proposée par **T-LAB** comme description "supplémentaire" ("S")

Dans le cas du chi-deux la structure de la table analysée est la suivante:



Où:
n_{ij} se réfère aux occurrences du mot (a) dans la classe sélectionnée (A)
N_j se réfère à toutes les occurrences du mot (a) dans le corpus (ou le sous-ensemble) analysé ;
N_i se réfère à toutes les occurrences dans la classe sélectionnée (A);
N se réfère à toutes les occurrences du tableau de contingence mots x classes.

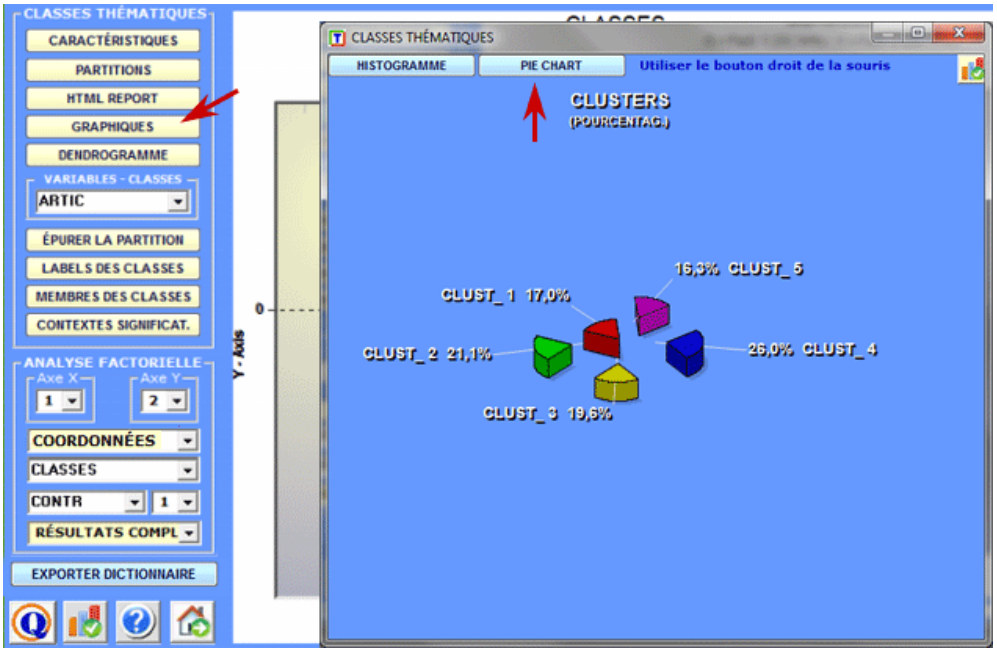
Un tableau HTML (voir ci-après) permet de vérifier dans les détails les caractéristiques des classes.
Il contient la liste des mots et des contextes élémentaires qui caractérisent la classe examinée: les premiers ordonnés par rapport au CHI2, les deuxièmes ordonnés par rapport à leur poids (score).

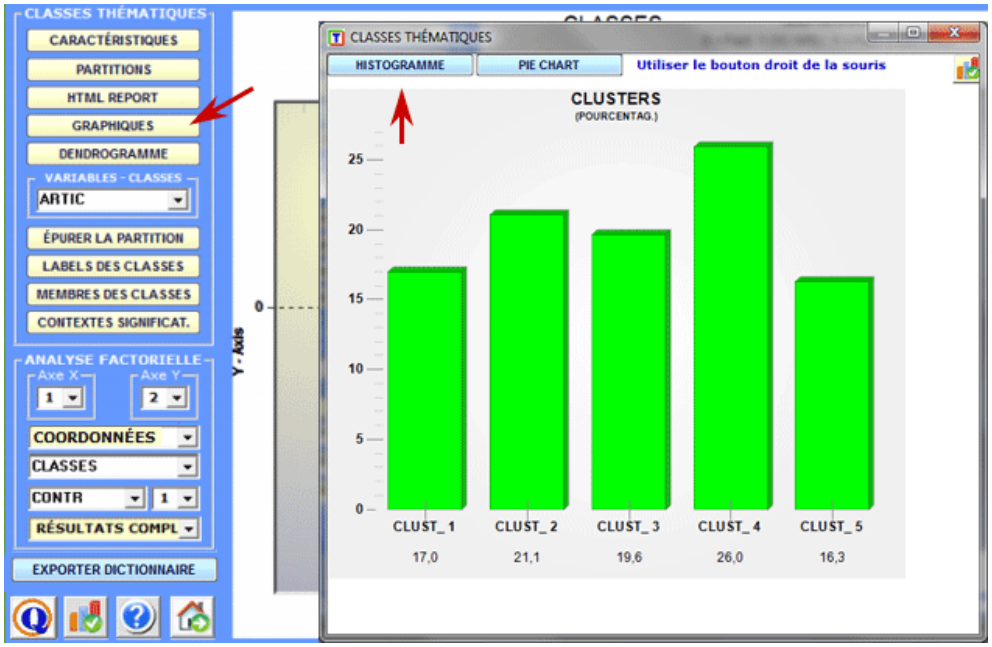
LEMMA	CHI SQUARE	WORD	OCC
information	122.662	information	56
information	122.662	informations	23
donnée	118.626	donnée	2
donnée	118.626	données	46
réseau	109.03	réseau	19
réseau	109.03	réseaux	13
échange	59.249	échange	10
échange	59.249	échanges	13
métadonnées	51.194	métadonnées	21

SCORE (63.759)

La définition de_la structure de ces métadonnées est menée non pas en inventoriant toutes les données existantes ou futures mais en identifiant les divers types de données. Ainsi les métadonnées structurées à_partir_de cette information la plus diversifiée possible doivent représenter au mieux la variété de cette information.

Des graphiques en secteurs (**pie charts**) et des histogrammes permettent de vérifier le pourcentage des unités de contextes appartenantes à chaque classe.

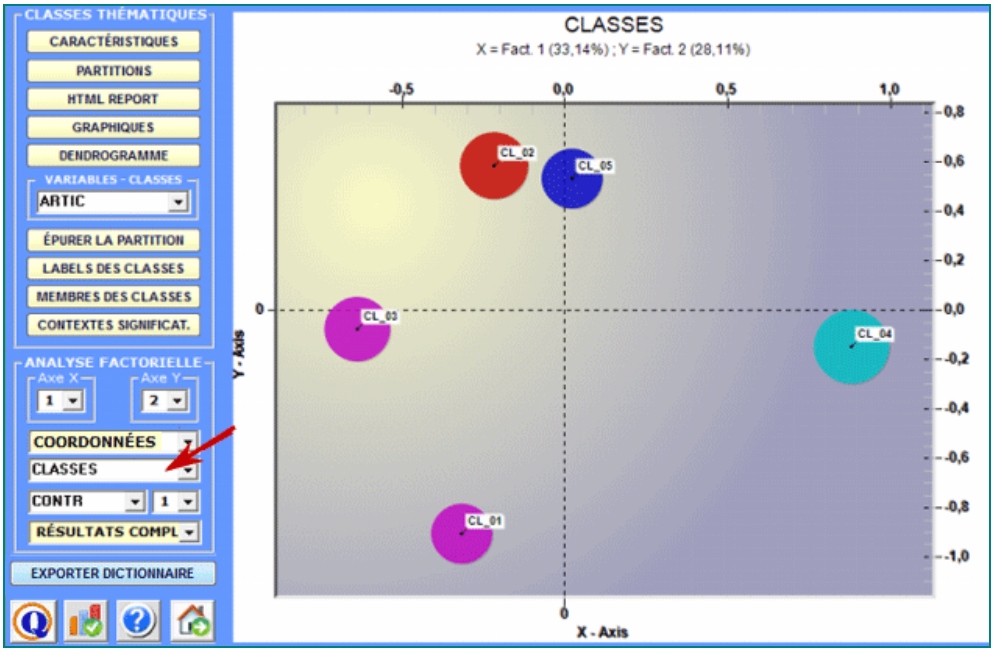




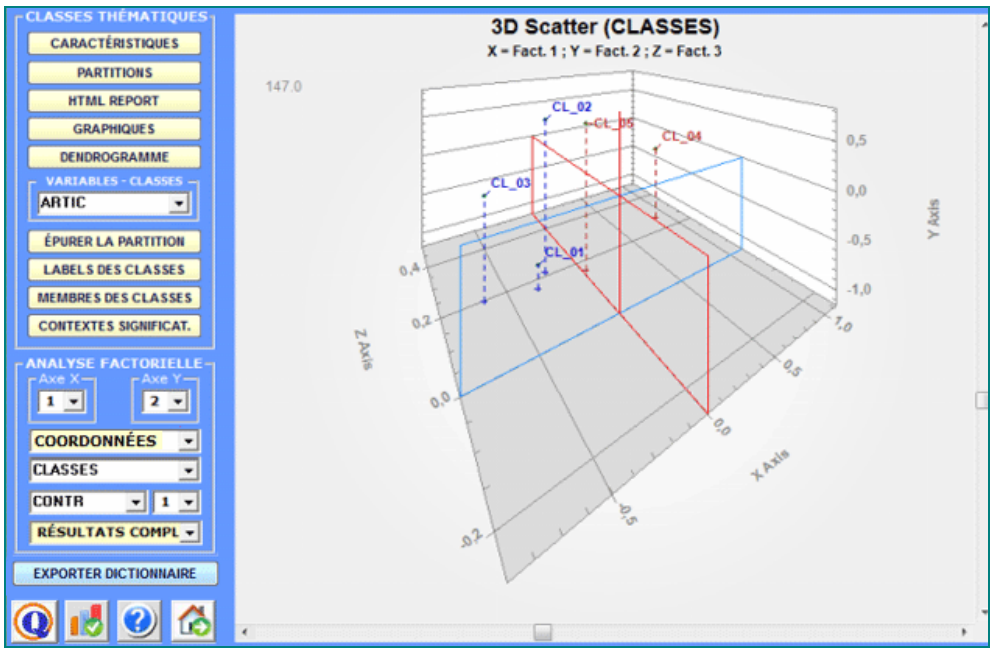
2 - Explorer les relations entre classes

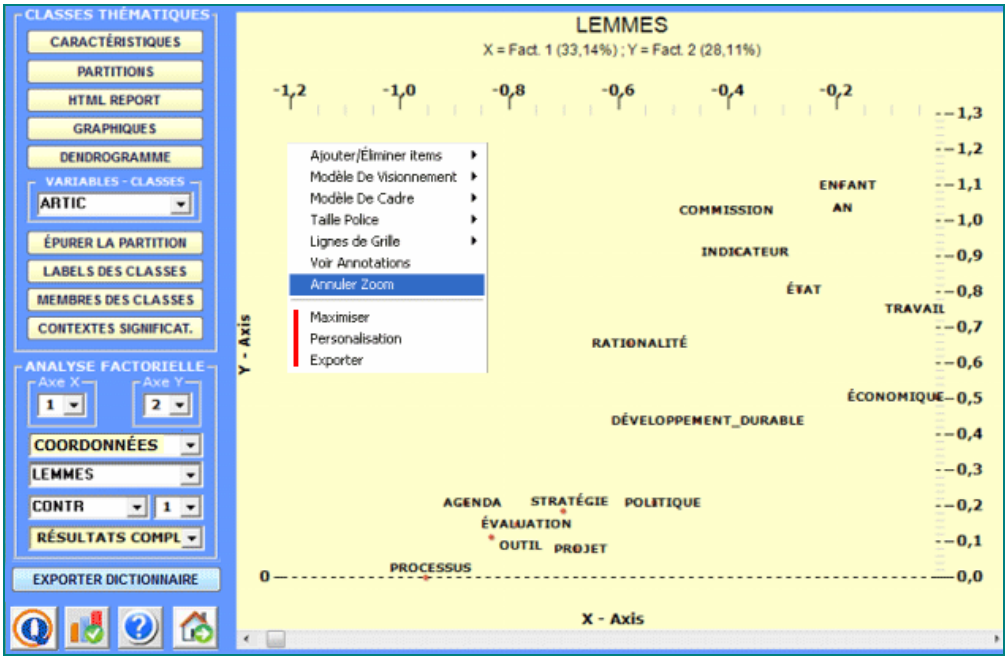
Certains graphiques, obtenus ou moyen de l'**Analyse des Correspondances**, permettent d'explorer les relations entre les classes à l'intérieur d'espaces bidimensionnels.

Plus spécifiquement :
- pour explorer les différentes combinaisons des axes factoriels, il suffit de les sélectionner dans les boîtes appropriées ("Axe X", "Axe Y") ;
- pour chacune des combinaisons (X-Y), on peut afficher différents types d'éléments (classe, lemmes et variables).

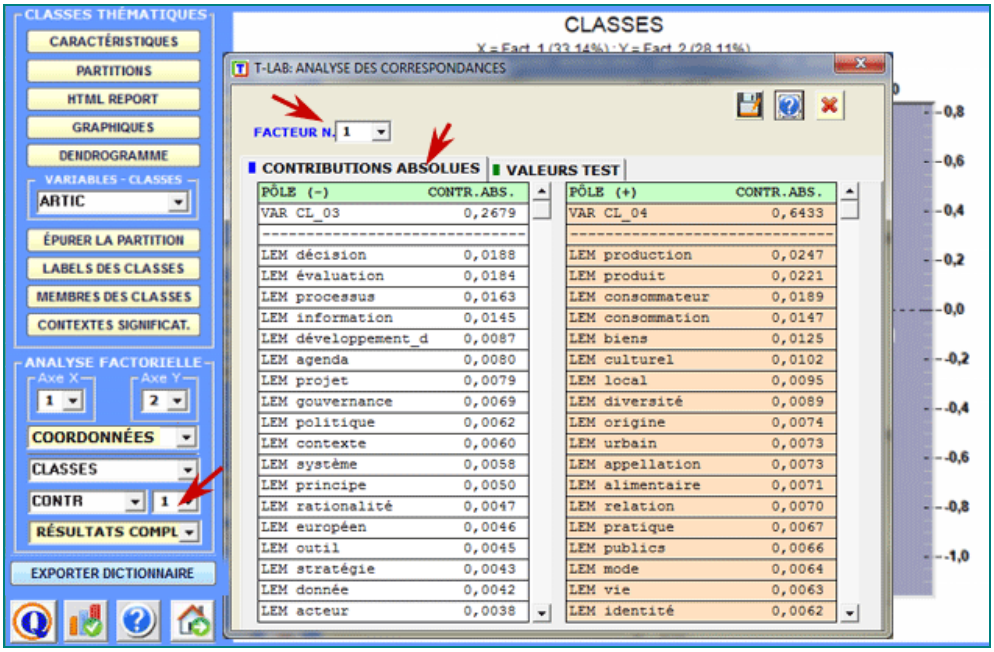


Tous les graphiques peuvent être personnalisés à travers l'utilisation de la fenêtre de dialogue appropriée (à l'aide du clic droit de la souris). De plus, lorsqu'il y a plus que trois clusters thématiques, leurs relations peuvent être explorées à travers les graphiques 3D (voir ci dessous).

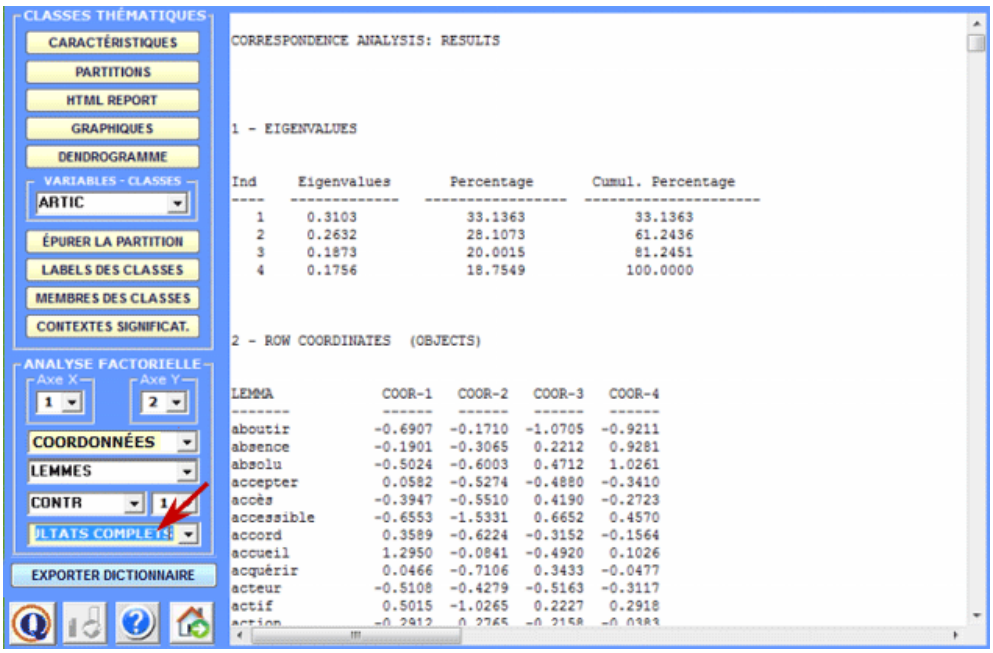




Pour chaque axe factoriel **T-LAB** fournit deux tableaux qui aident à l'interprétation.



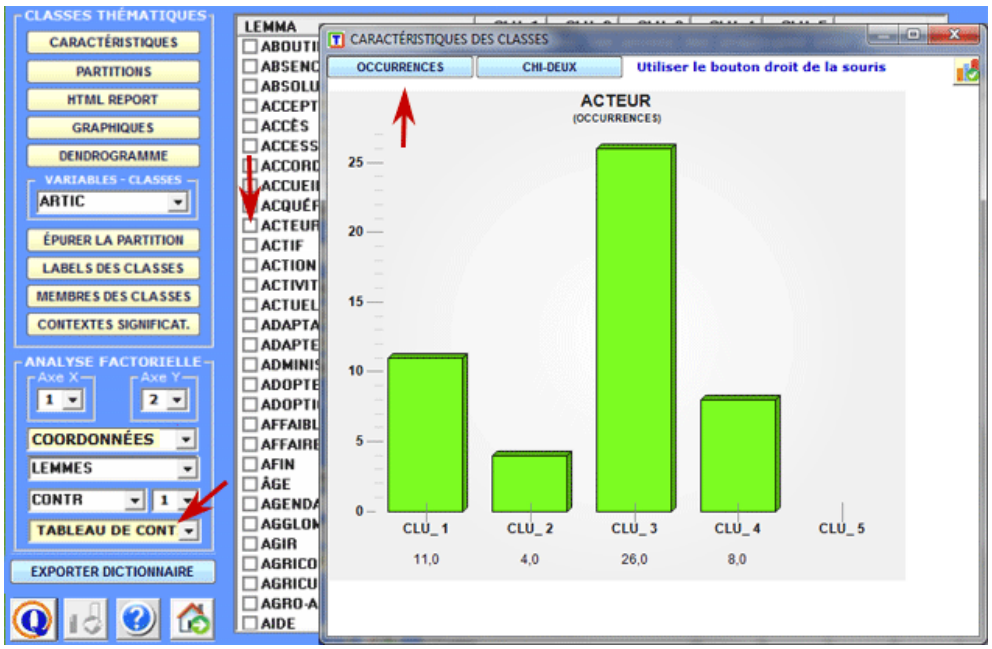
Un clic sur le bouton **Tableau des Résultats** nous permet de visualiser et de sauvegarder le fichier qui contient tous les résultats de l'analyse: valeurs propres, coordonnées, contributions absolues et relatives, valeurs test.

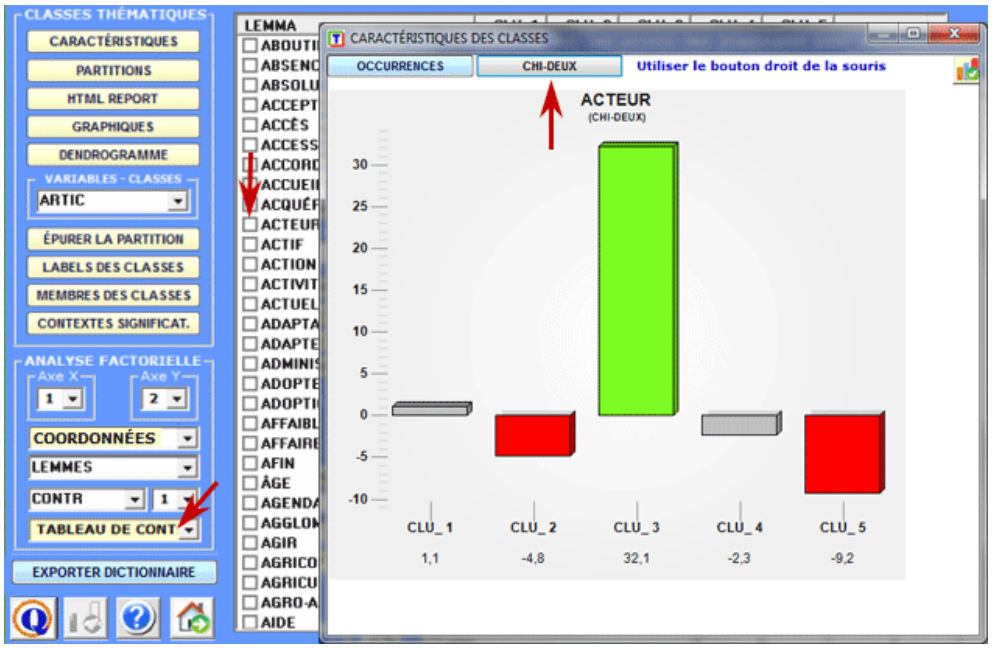


Une option spécifique (voir ci-dessous) nous permet de visualiser / exporter le tableau de contingences et de créer des graphiques montrant la répartition de chaque mot au sein des clusters.

De plus, en cliquant sur cellules spécifiques du tableau, il est possible de créer un fichier HTML montrant tous les contextes élémentaires où le mot en ligne est présent dans le cluster correspondant (colonne).

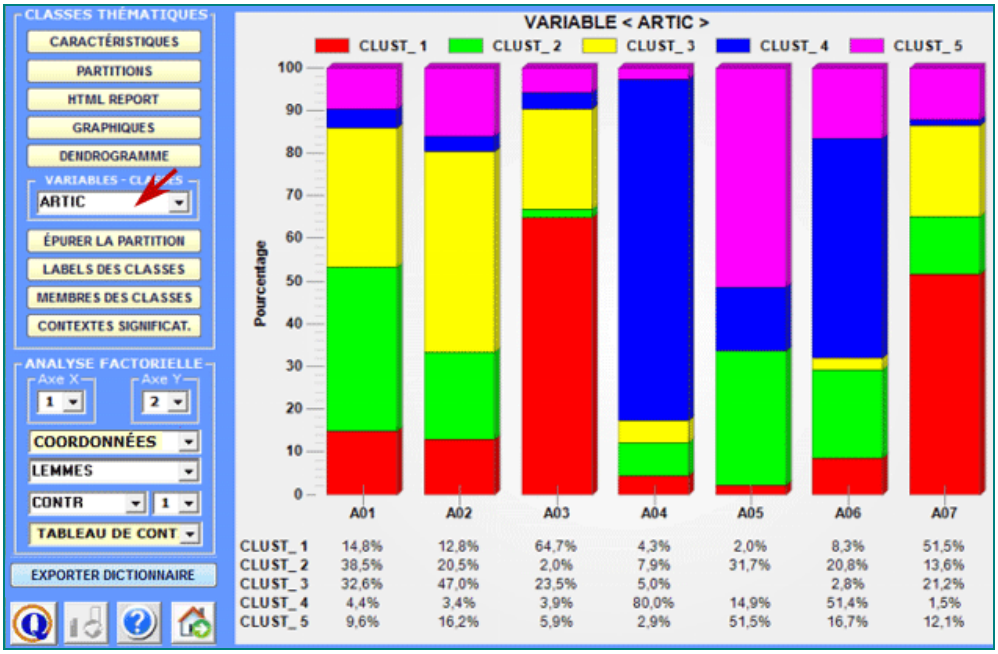
N.B. : Ce tableau comprend soit les mots-clés actives ('A') soit les mots-clés supplémentaires ('S').





3 - Explorer les relations entre classes et variables

Des **histogrammes** vous permettent de vérifier les rapports entre les classes et les variables.



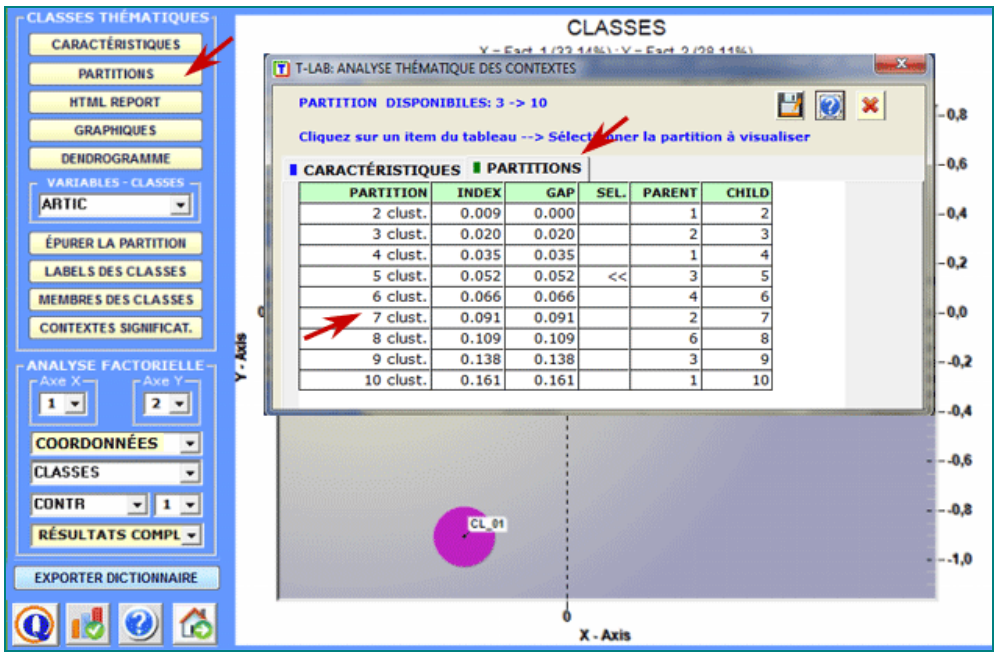
D'autres relations entre classes et variables peuvent être explorées à l'aide des options disponibles dans la section "Analyse factorielle " (voir ci-dessus).

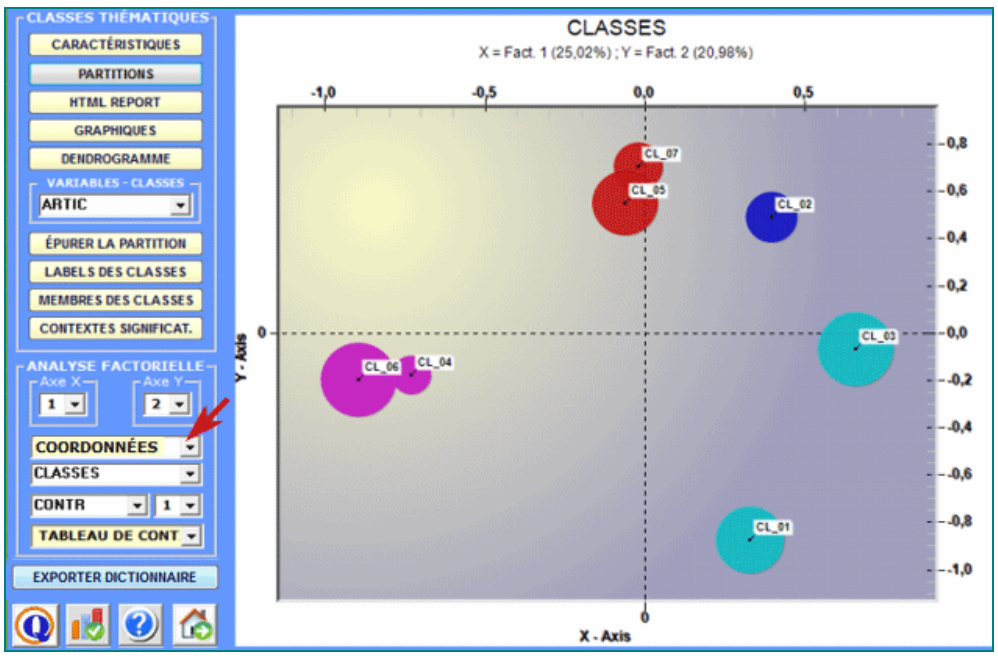
4 - **Explorer les différentes partitions des classes**

Puisque l'algorithme utilisé produit une classification hiérarchique, l'utilisateur peut facilement explorer plusieurs solutions de l'analyse: partitions composées de 3 à 50 classes.

Pour chaque partition obtenue, un tableau approprié (voir ci-après) montre les valeurs suivantes:
- "Index" correspond au rapport entre variance inter-classes et variance totale;
- "Gap" indique la différence entre la valeur de l'Index et celle de la partition immédiatement précédente:
- Nombre de la classe "fils" (child) obtenue à l'aide de la bi-section du "parent" correspondant.

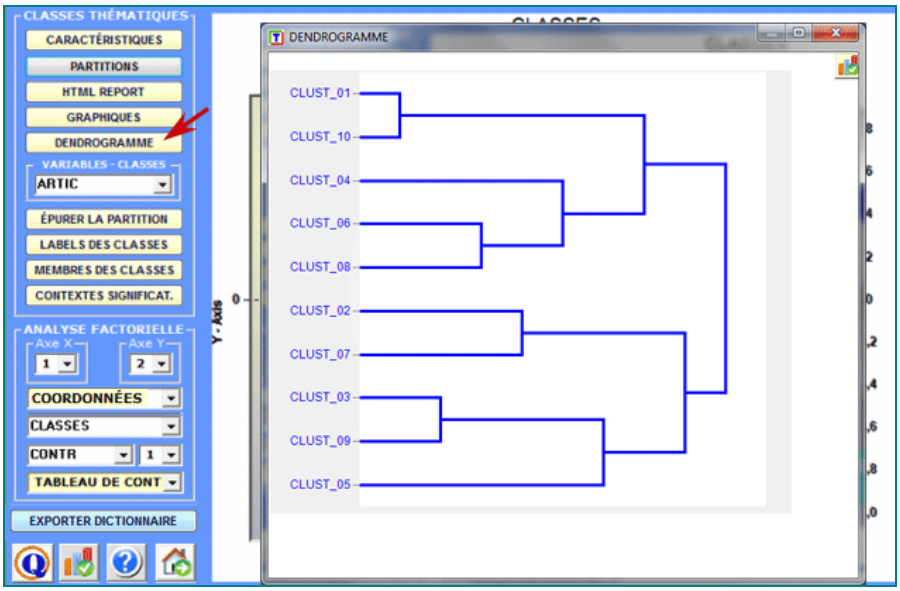
L'option **Partitions** (voir ci-après) vous permet d'explorer les caractéristiques des différentes solutions.



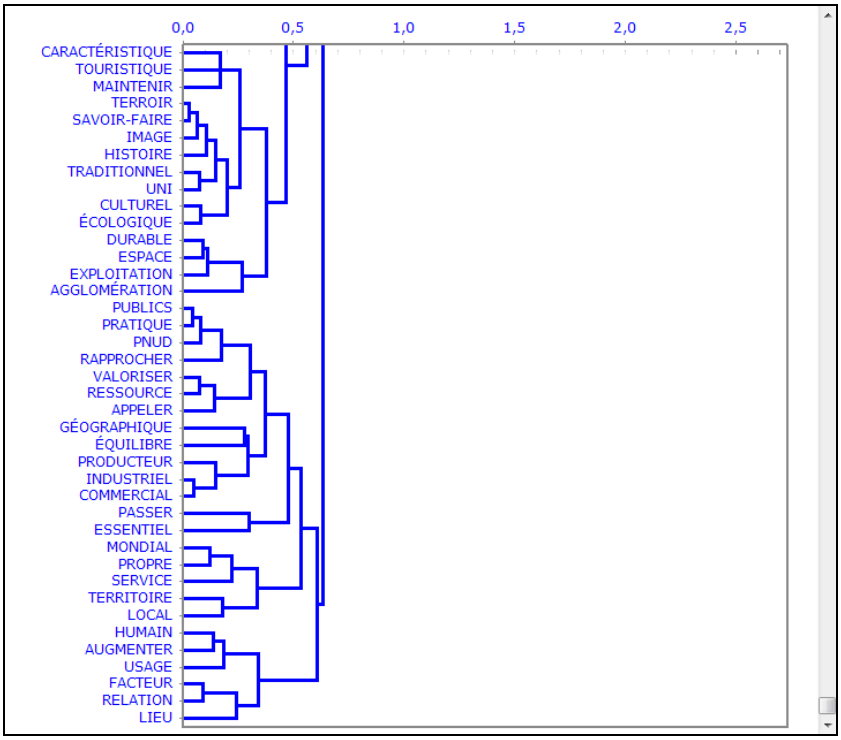


En outre, l'option dendrogramme (voir ci-dessous) permet deux possibilités:

A) vérifier l'arbre des différentes bi-sections de clusters



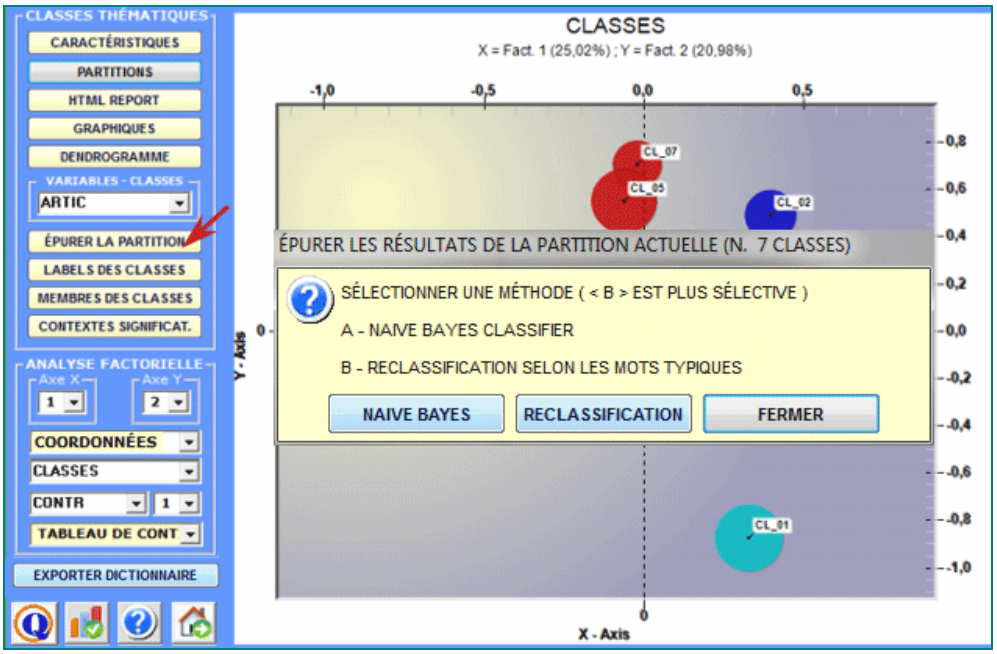
B) vérifier l'arbre des mots caractéristiques à chaque cluster.



5 - Raffiner les résultats de la partition choisie

Ensuite ayant exploré différentes solutions, l'utilisateur peut raffiner les résultats de la partition choisie et, au besoin, répéter certaines des étapes illustrées ci-dessus (1,2,3).

À cet effet, deux méthodes sont disponibles (voir image suivante).



Quand on choisit la méthode «**A**» (c'est-à-dire **Naïve Bayes Classifier**), cette option **T-LAB** nous permet de supprimer de l'analyse toutes les unités de contexte dont l'appartenance à une classe ne respecte pas les critères suivants:
a) pour chaque unité de contexte, l'appartenance à une classe, soit déterminée par le bisecting K-Means (unsupervised clustering) soit par le classificateur Naïve Bayes (supervised clustering), doit être la même;
b) la valeur maximum de la probabilité a posteriori, correspondante à l'appartenance de la j-unité de contexte à la k-classe, doit être au moins 50% plus haute que ses valeurs restantes (c.-à-d. les valeurs de la probabilité a posteriori dans les autres classes).

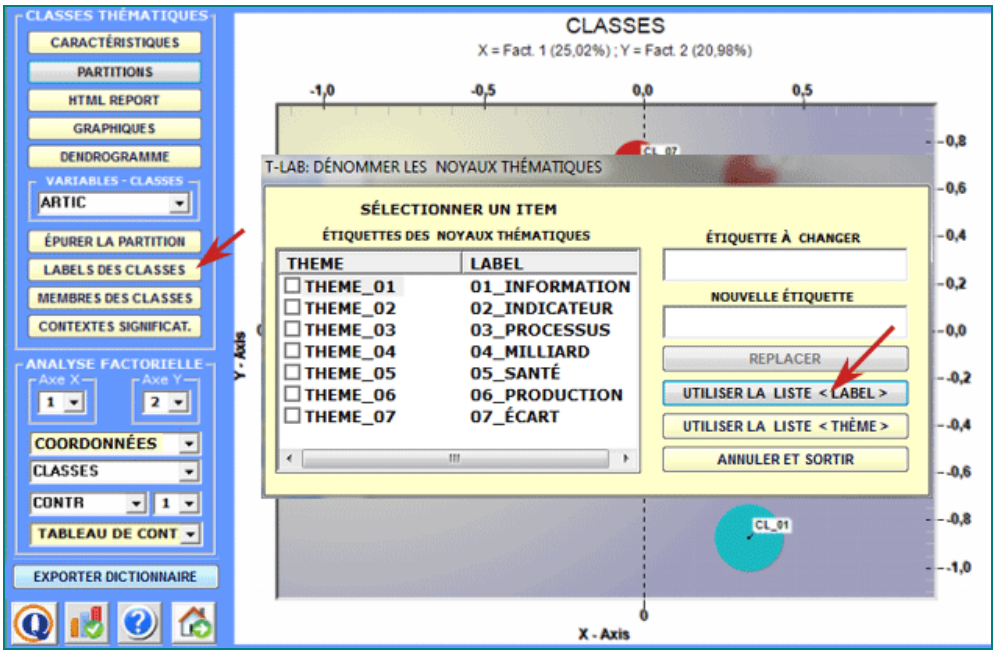
Autrement, dans le cas de la méthode "**B**" (c'est-à-dire **Reclassement selon les Mots Typiques**) **T-LAB** considère les caractéristiques des classes, c'est-à-dire les mots avec une valeur significative de Chi-Deux, comme items d'un dictionnaire des catégories et effectue les trois étapes de la "classification supervisée" décrites au début de cette section. Donc, lorsque l'utilisateur est intéressé à réappliquer des dictionnaires et à en comparer les résultats relatifs, il est vivement conseillé l'utilisation de cette méthode

Tous les résultats de ce calcul sont dans un tableau exporté par **T-LAB** (voir ci-dessous) où les valeurs de les probabilités a posteriori sont converties en format pourcentage.

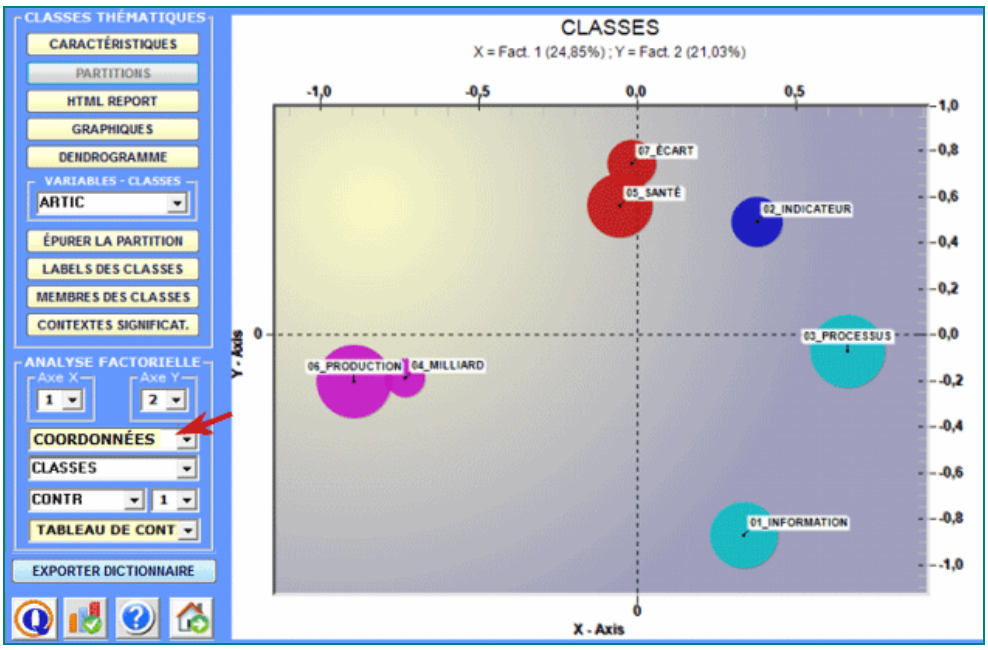
Context_ID	OLD	NEW	MATCH	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CONTEXT
'00001000001	2	2	YES	0	0,919	0,081	0	0	0	0	Les contrats de plan en l ' Etat , les Régions et l
'00001000002	2	2	YES	0	1	0	0	0	0	0	Nous nous interrogerons dans cet article sur le rô
'00001000003	2	2	YES	0	1	0	0	0	0	0	Introduction : le contexte Dans sa circulaire aux
'00001000004	5	5	YES	0	0	0	0	1	0	0	' ' L ' évaluation a_priori et a_posteriori des proje
'00001000005	2	2	YES	0	1	0	0	0	0	0	L ' évaluation n ' est pas l ' annexe d ' une politiqu
'00001000006	2	2	YES	0	1	0	0	0	0	0	' ' A l ' appui de sa proposition , la ministre donn
'00001000007	3	3	YES	0	0,012	0,988	0	0	0	0	Le couplage des contrats de plan et des procédur
'00001000008	2	2	YES	0	1	0	0	0	0	0	Enfin au niveau mondial , les indicateurs de dével
'00001000009	2	2	YES	0	1	0	0	0	0	0	feront de_plus_en_plus partie des outils de_la nég
'00001000010	2	2	YES	0	1	0	0	0	0	0	Une initiative internationale Comme le PNB et les
'00001000011	2	2	YES	0	1	0	0	0	0	0	Les indicateurs de développement_durable sont s
'00001000012	3	3	YES	0	0	1	0	0	0	0	Malgré cela le thème des indicateurs de développ
'00001000013	2	2	YES	0	1	0	0	0	0	0	environnement et du développement ' ' (CNUE
'00001000014	2	2	YES	0	1	0	0	0	0	0	Les indicateurs de développement_durable ont fait
'00001000015	2	2	YES	0	1	0	0	0	0	0	Il a fallu attendre en fait la 3ème session de_la Co
'00001000016	2	2	YES	0	1	0	0	0	0	0	Une liste d ' indicateurs a_été ensuite proposée p
'00001000017	2	2	YES	0	1	0	0	0	0	0	en_effet bien qu ' ils soient aussi supposés s ' adr
'00001000018	5	5	YES	0	0	0	0	1	0	0	Le PNUD avait déjà initié ce type d ' approche ave
'00001000019	3	3	YES	0	0	1	0	0	0	0	un groupe international de spécialistes de l ' évalu
'00001000020	5	5	YES	0	0	0	0	1	0	0	Une perspective holistique doit permettre la prise
'00001000021	1	1	YES	1	0	0	0	0	0	0	Sur le plan de_la méthode , l ' évaluation doit avo
'00001000022	3	3	YES	0	0	1	0	0	0	0	Il s ' agit d ' un processus capable d ' adaptation ,
'00001000023	5	5	YES	0	0	0	0	1	0	0	SCOPE publiera de son côté en 1997 un ouvrage
'00001000024	2	2	YES	0	1	0	0	0	0	0	Mais la durabilité autorégulatrice , pour reprendre
'00001000025	2	2	YES	0	1	0	0	0	0	0	et travaille à une ' ' boîte à outil ' ' pour organ
'00001000026	4	4	YES	0	0	0	1	0	0	0	) . Ils doivent permettre d ' apprécier les question
'00001000027	6	6	YES	0	0	0	0	0	1	0	Les contraintes écologiques et l ' évolution des te
'00001000028	2	2	YES	0	1	0	0	0	0	0	L ' évaluation doit s ' envisager dans le cadre de_l
'00001000029	1	1	YES	1	0	0	0	0	0	0	Il ne peut pas être question de dégager , de l ' ex
'00001000030	2	2	YES	0	0,998	0,002	0	0	0	0	de préciser le cadre de négociation dans le cadre

6 - Attribuer des étiquettes aux classes

Une fonction particulière de **T-LAB** permet d'attribuer des étiquettes aux classes.
(N.B: Lors de la première utilisation, certaines étiquettes sont proposées automatiquement par le logiciel).



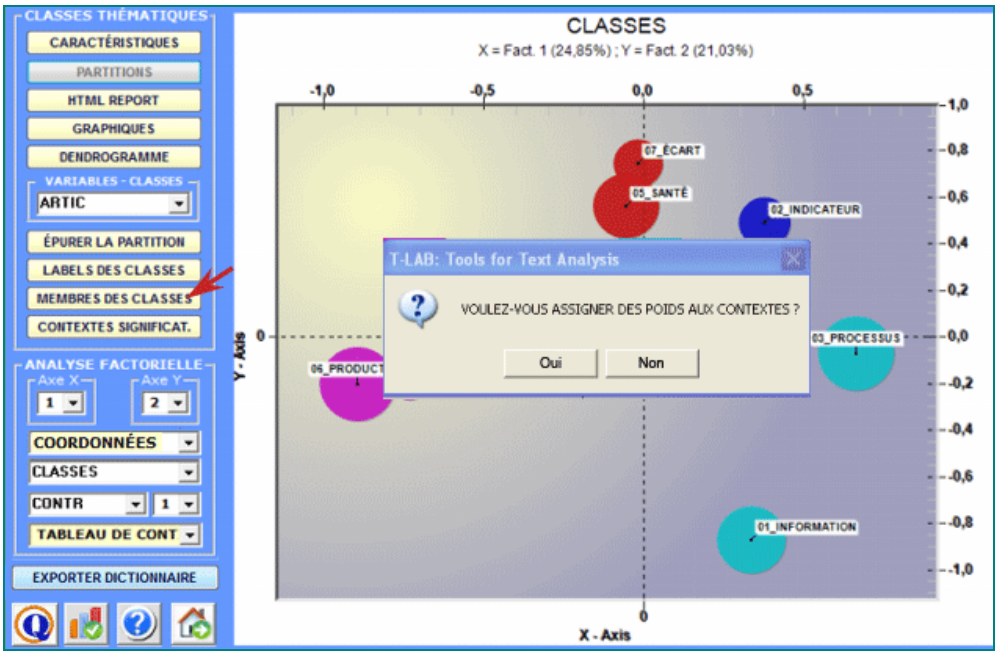
Les étiquettes attribuées aux différentes classes peuvent être affichées dans les différents graphiques disponibles (voir ci-après).



7- Vérifier quels sont les contextes élémentaires qui appartiennent à chaque classe ;

8 - Vérifier le poids de chaque contexte élémentaire au sein du cluster auquel il appartient ;

9 – Obtenir une classification des documents.



Le bouton **Membres des Classes** permet d'exporter trois types de tableaux (voir ci-après) sous format MS Excel:

a - " Cluster_Partitions.xls " avec toutes les correspondances unités de contexte x classe à l'intérieur des différentes partitions ;

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

b - "Themes_Contexts.xls" avec les correspondances unités de contexte x classe à l'intérieur de la partition sélectionnée.

(IDNUMBER)	THEME	SCORE	CONTEXT
'00001000001	02_INDICATEUR	8,9	Les contrats de plan en l ' Etat , les Régions et les autres collectivités locales d
'00001000002	02_INDICATEUR	13,3	Nous nous interrogerons dans cet article sur le rôle des indicateurs de développe
'00001000003	02_INDICATEUR	2,74	Introduction : le contexte Dans sa circulaire aux Préfets de Région (Voynet 19
'00001000004	05_SANTÉ	0,76	' ' L ' évaluation a_priori et a_posteriori des projets et réalisations doit être mise
'00001000005	02_INDICATEUR	12,07	L ' évaluation n ' est pas l ' annexe d ' une politique , elle en fait intégralement pa
'00001000006	02_INDICATEUR	22,82	' ' A l ' appui de sa proposition , la ministre donne une liste de critères en_vue d
'00001000007	03_PROCESSUS	18,41	Le couplage des contrats de plan et des procédures européennes , qui implique
'00001000008	02_INDICATEUR	61,81	Enfin au niveau mondial , les indicateurs de développement_durable , élaborés
'00001000009	02_INDICATEUR	68,55	feront de_plus_en_plus partie des outils de_la négociation entre les différents niv
'00001000010	02_INDICATEUR	47,19	Une initiative internationale Comme le PNB et les agrégats économiques ne peu
'00001000011	02_INDICATEUR	49,45	Les indicateurs de développement_durable sont souvent présentés comme une b
'00001000012	03_PROCESSUS	28,84	Malgré cela le thème des indicateurs de développement_durable est resté peu p
'00001000013	02_INDICATEUR	17,74	environnement et du développement ' ' (CNUED 1992 , § 40 . 4) . La 1ère
'00001000014	02_INDICATEUR	68,53	Les indicateurs de développement_durable ont fait l ' objet d ' un projet SCOPE ,
'00001000015	02_INDICATEUR	88,35	Il a fallu attendre en fait la 3ème session de_la Commission du développement_d
'00001000016	02_INDICATEUR	30,11	Une liste d ' indicateurs a_été ensuite proposée permettant la comparaison entre
'00001000017	02_INDICATEUR	49,75	en_effet bien qu ' ils soient aussi supposés s ' adresser aux décideurs des nivea
'00001000018	05_SANTÉ	2,06	Le PNUD avait déjà initié ce type d ' approche avec les indicateurs de développe
'00001000019	03_PROCESSUS	64,21	un groupe international de spécialistes de l ' évaluation et de chercheurs a propo
'00001000020	05_SANTÉ	10,07	Une perspective holistique doit permettre la prise en compte des éléments du tri
'00001000021	01_INFORMATION	12,83	Sur le plan de_la méthode , l ' évaluation doit avoir un horizon temporel étendu e
'00001000022	03_PROCESSUS	17,62	Il s ' agit d ' un processus capable d ' adaptation , intégré dans le processus de
'00001000023	05_SANTÉ	1,64	SCOPE publiera de son côté en 1997 un ouvrage de synthèse qui montre la varie
'00001000024	02_INDICATEUR	36,88	Mais la durabilité autorégulatrice , pour reprendre l ' expression de l ' Agenda 21
'00001000025	02_INDICATEUR	27,84	et travaille à une ' ' boîte à outil ' ' pour organiser des systèmes d ' indicateu
'00001000026	04_MILLIARD	0,92	) . Ils doivent permettre d ' apprécier les questions d ' équité inter et intra-génér
'00001000027	06_PRODUCTION	0,81	Les contraintes écologiques et l ' évolution des technologies , ainsi_que les que
'00001000028	02_INDICATEUR	32,61	L ' évaluation doit s ' envisager dans le cadre de_la gouvernance Mais , aussi rat
'00001000029	01_INFORMATION	2,26	Il ne peut pas être question de dégager , de l ' extérieur ou d ' en_haut , par une
'00001000030	02_INDICATEUR	16,74	de préciser le cadre de négociation dans le cadre d ' une rationalité procédurale

En particulier, la valeur d'importance (score) assignée à chaque contexte élémentaire (j) appartenant à la classe (k) vient de la formule suivante :

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

Où:
Scorej = valeur d'importance assignée au contexte élémentaire (**j**);
ΣX_{ij} = somme des valeurs du Chi-deux correspondantes aux mots clés (**i**) trouvés dans le contexte élémentaire (**j**) et qui sont typiques de la classe (**k**);
nj = nombre de mots clés, typiques de la classe (**k**), trouvés dans le contexte élémentaire (**j**);
N = nombre de mots clés typiques de la classe (**k**).

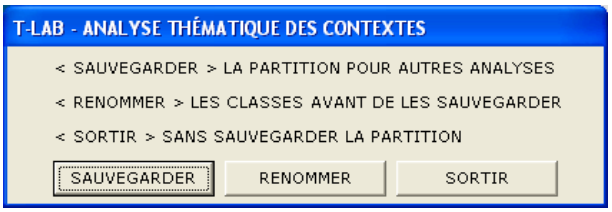
c - " Ec_Document_Classification.xls " (output fourni seulement quand le corpus se compose au moins de 2 documents primaires qui ne sont pas des textes courts comme les réponses aux questions ouvertes) énumérant les appartenances mélangées de chaque document (voir ci-dessous).

DOC_ID	VAR_01	BEST_CL	CLUST_ 1	CLUST_ 2	CLUST_ 3	CLUST_ 4	CLUST_ 5	CLUST_ 6	CLUST_ 7
1	AR_A01	2	0,119	0,446	0,355	0	0,022	0,008	0,049
2	AR_A02	3	0,111	0,141	0,672	0,005	0,042	0,003	0,027
3	AR_A03	1	0,738	0,009	0,161	0,025	0,068	0	0
4	AR_A04	6	0,01	0,067	0,046	0,189	0,028	0,652	0,008
5	AR_A05	5	0,002	0,016	0	0,006	0,598	0,014	0,365
6	AR_A06	6	0,025	0,034	0,023	0,126	0,167	0,569	0,057
7	AR_A07	1	0,698	0,056	0,208	0,009	0,027	0	0,003

Dans ce cas-ci les valeurs viennent de la formule ci-dessus (voir le " b") en additionnant les scores des contextes élémentaires appartenant à chaque document et en appliquant un calcul des pourcentages.

10 - Archiver la partition sélectionnée pour l'explorer avec d'autres outils T-LAB

Lorsqu'on quitte la fonction **Analyse thématique des Contextes Elémentaires**, des messages rappellent qu'il est possible d'explorer les classes obtenues avec d'autres outils **T-LAB**.



Si on choisit l'option **Sauvegarder**, la variable < **CONT_CLUST** > (classes de contextes élémentaires) demeure disponible uniquement dans certains types d'analyse (par exemple, Séquences de Thèmes, Associations de Mots, Comparaison entre Paires, Analyse des Mots Associés) et jusqu' au moment où l'utilisateur modifie sa liste de mots clés.

11 - Exporter un dictionnaire des catégories.

Lorsque cette option est sélectionnée, **T-LAB** crée deux fichiers :

a) un fichier dictionnaire avec l'extension '.**dictio**' prêt à être importé par l'intermédiaire d'un des outils pour l'analyse thématique. Dans ce dictionnaire chaque cluster correspond à une catégorie décrite au moyen de ses mots caractéristiques, c'est-à-dire par tous les mots avec une valeur significative du chi-deux à son interne ;

b) un fichier **MyList.diz** prêt à être importé par la fonction 'Configuration Personnalisée. Etant donné que ce fichier contient la liste alphabétique de tous les mots avec une valeur significative du chi-carré, c'est-à-dire tous les mots qui déterminent la différence entre les clusters thématiques, son utilisation peut permettre de répéter certaines analyses avec une modalité plus sélective et discriminante.

12 – Vérifier la qualité de la partition choisie et la cohérence sémantique des différents thèmes

Lorsque vous cliquez sur le bouton 'Index de Qualité', **T-LAB** crée un fichier HTML qui contient diverses mesures.

Les premières de celles-ci se réfèrent à la qualité de la partition en 'k' classes, c'est-à-dire, par exemple, au rapport entre la variance intérieure et extérieure.

Les deuxièmes se réfèrent à la 'cohérence sémantique' de chaque cluster et plus en détail aux similitudes entre les premiers dix mots caractéristiques de chaque thème. En détail :

- les 10 premiers mots sont ceux qui ont la plus grande valeur du chi-carré ;
- les mesures de similarité sont calculées en utilisant le coefficient du cosinus ;
- comme dans le cas de l' outil 'Associations de mots', le coefficient du cosinus est calculé en vérifiant les cooccurrences de chaque paire de mots à l' intérieur des segments de texte définis en tant que contextes élémentaires.

.

13 - Explorer les Séquences de Thèmes.

Contrairement à l'outil Séquences de Thèmes inclus dans un sous-menu de T-LAB pour l'analyse des cooccurrences, cette option a été spécialement conçue pour intégrer l'analyse thématique des contextes élémentaires. Plus précisément : son usage a sens seulement lorsque le corpus entier peut être considéré comme un discours et/ou lorsque ses différentes sections (par exemple : chapitres d'un livre, parties d'une entrevue, interventions de différents participants à une conversation ou à une discussion, etc.) se succèdent dans un ordre temporel précis.
Dans ce cas, les relations analysées sont celles entre les contextes élémentaires (jusqu'à un maximum de 100. 000), le long de la chaîne linéaire du corpus, et chacun d'entre eux –soit comme « prédécesseur » ou comme « successeur » – est traité comme une unité d'analyse appartenant à un cluster thématique (ou comme non classé).
Tous les output fournis permettent à l'utilisateur d'explorer les relations séquentielles entre « thèmes », aussi bien de façon « statique » que de façon «dynamique». En particulier, au moyen de certains graphiques animés qui permettent d'apprécier la dynamique temporelle des séquences, l'utilisateur peut vérifier lorsque les gens sont engagés sur des thèmes particuliers

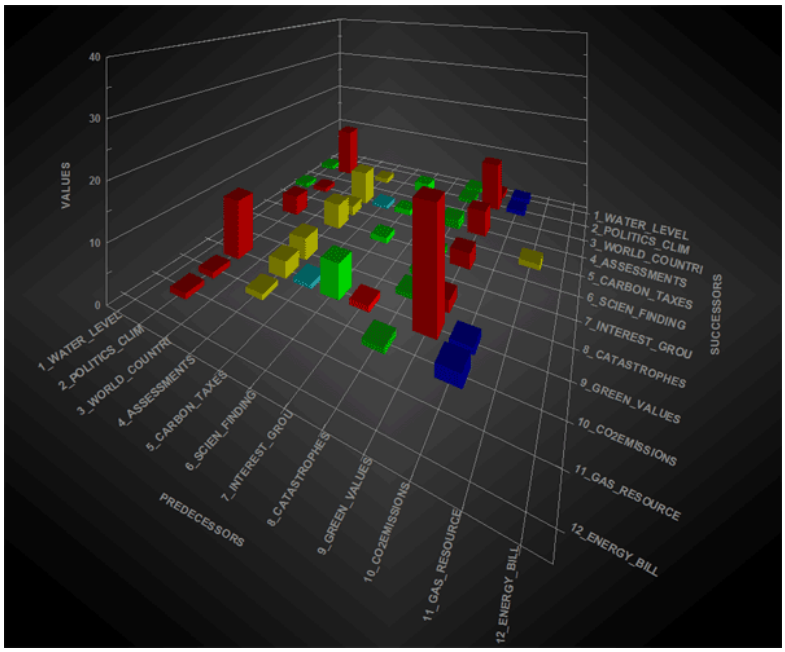
(voir, par exemple, les points sur la diagonale des matrices dans les images suivantes) et lorsqu' ils passent d' un thème dominant à l'autre.

Étape par étape, de suite on fournit une brève description des différentes options disponibles.

(N.B.: tous les output de l'exemple ont été obtenus à travers une analyse thématique du livre « The Politics of Climate Change " d' Anthony Giddens publiée sur le site de T-LAB).
Lorsque le bouton « Séquences de Thèmes » est activé, en cliquant dessus, le « player »suivant devient visible et actif.

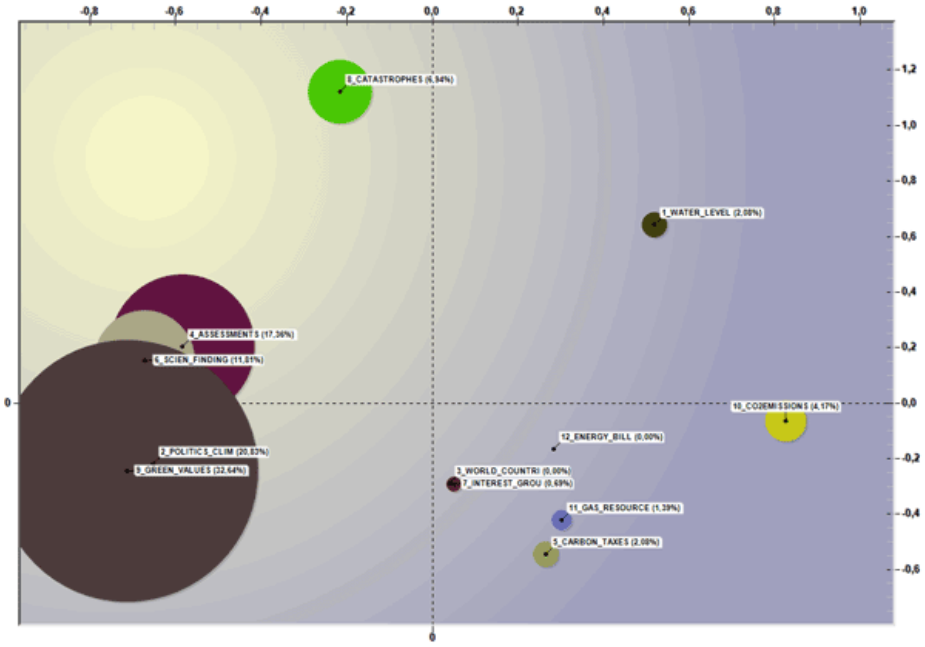
L' option "1" (voir ci-dessus) désigne le type de graphique choisi pour la visualisation des séquences, aussi bien à l' intérieur du corpus entier qu' à l' intérieur d' une partie de celui-ci (voir ci-dessus option "2").

L'option « matrice » fournit un graphique 3D qui résume les relations entre les prédécesseurs et les successeurs à l'aide de barres colorées placées aux croisements respectifs. Dans ce cas, lorsque des graphiques 3D animés sont visualisés, l' accroissement en hauteur des différentes barres indique l'augmentation des occurrences des séquences respectives (voir relations binaires entre « prédécesseurs » et « successeurs » dans le graphique suivant).



L'option «espace» fournit un graphique 2d dont les dimensions (c' est -à-dire les pourcentages) et les relations entre groupes thématiques sont représentées sur un plan organisé par deux axes factoriels sélectionnés par l'utilisateur. Dans ce cas, lorsque des graphiques animés sont affichés, les tailles des « bulles » - qui sont continuellement

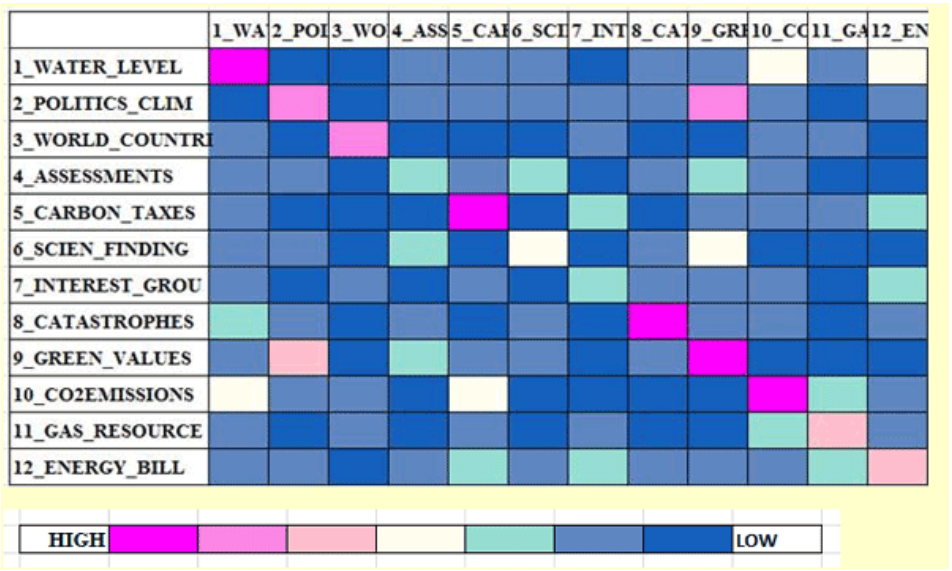
réadaptées à un total égal à 100 % - indiquent comment les pourcentages des éléments qui appartiennent à chaque cluster thématique varient avec le temps et, simultanément, le mouvement des flèches indique la direction dans laquelle les thèmes se suivent.



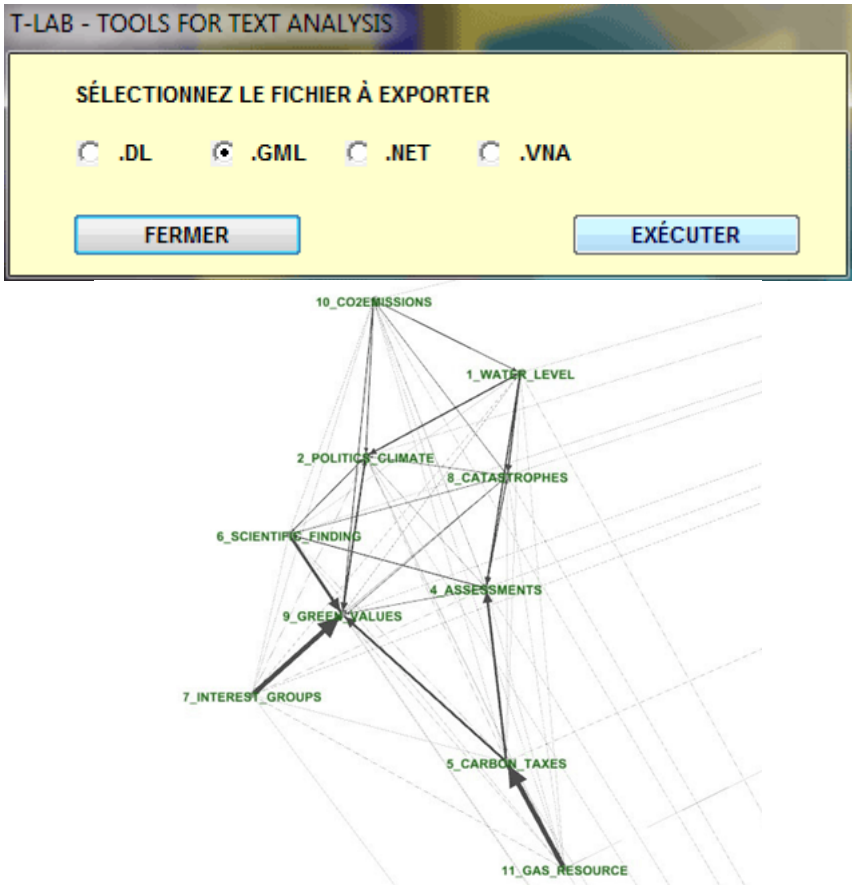
Dans les deux cas à peine décris, après l'arrêt de l' image (voir le bouton « pause »),on peut voir deux autres output :

A- des tableaux html qui résument les relations entre les prédécesseurs et les successeurs (voir ci-dessous) ;

	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_COUNTR	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- des fichiers graphiques qui peuvent être importés à partir d'un logiciel pour l'analyse de réseau.

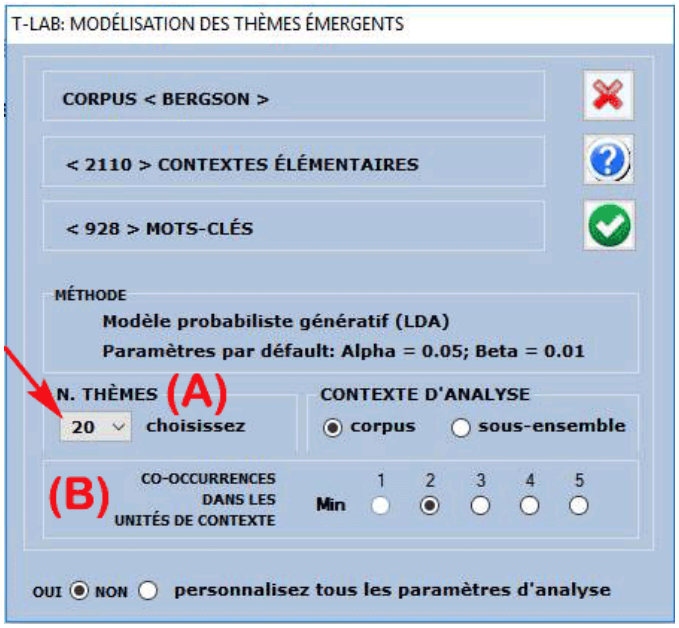


N.B. : le graphique précédent, qui fait référence au troisième chapitre du livre de Giddens, a été créé au moyen du logiciel Gephi (voir https://gephi.org/)

Modélisation des Thèmes Émergents

Cet outil **T-LAB** vous permet de **repérer, examiner et modeler les principaux thèmes qui émergent des textes**, aussi pour ensuite les utiliser dans des analyses qualitatives (ex. faire des grilles pour l'analyse de contenu) et quantitatives ultérieures.

Les thèmes émergents, qui sont décrits à travers leur vocabulaire caractéristique, c'est-à-dire à travers des ensembles de mots-clés (lemmes ou catégories) co-occurrents dans les unités de texte analysées, peuvent être en fait utilisés pour **classifier** ces dernières (ex. des documents ou des contextes élémentaires) et **obtenir de nouvelles variables** qui peuvent être utilisées dans des analyses ultérieures.



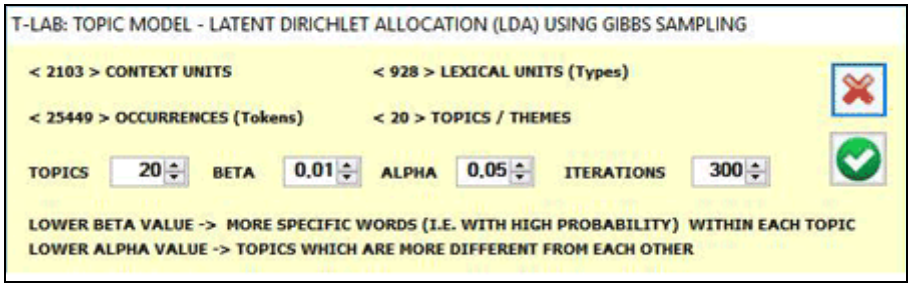
Un boîte de dialogue **T-LAB** (voir ci-dessus) vous permet de définir deux paramètres d'analyse.

En particulier:

- le paramètre (A) permet de définir le nombre de thèmes à obtenir. (Il convient de noter que plus ce nombre sera élevé plus cohérentes seront les relations de co-occurrence dans chaque theme; en outre, si nécessaire, certains themes - par exemple, ceux qui sont redondants ou difficiles à interpreter - peuvent être éliminés plus tard);

- le paramètre (B) vous permet d'exclure de l'analyse tous les unités de contexte qui ne contient pas un nombre minimum de mots-clés inclus dans la liste utilisée.

Seulement lorsque vous choisissez de personnaliser tous les paramètres d'analyse (voir l'option «Oui» ci-dessus), la fenêtre suivante sera affichée et plus d'options seront disponibles. (Notez que dans l'image suivante, le nombre d'unités de contexte est déterminé par le paramètre "B" dejà mentionné).



La **procédure automatique d'analyse** consiste en les étapes suivantes:

a – construction d'une matrice documents pour mots-clés, où les documents sont toujours des contextes élémentaires correspondant aux unités de contexte (c.-à-d. fragments, phrases, paragraphes) dans lesquelles le corpus a été subdivisé ;

b – analyse des données avec un modèle probabiliste qui utilise la "Latent Dirichlet Allocation" et le "Gibbs Sampling" (pour plus d'informations consulter les articles correspondants sur Wikipedia: http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation; http://en.wikipedia.org/wiki/Gibbs_sampling) ;

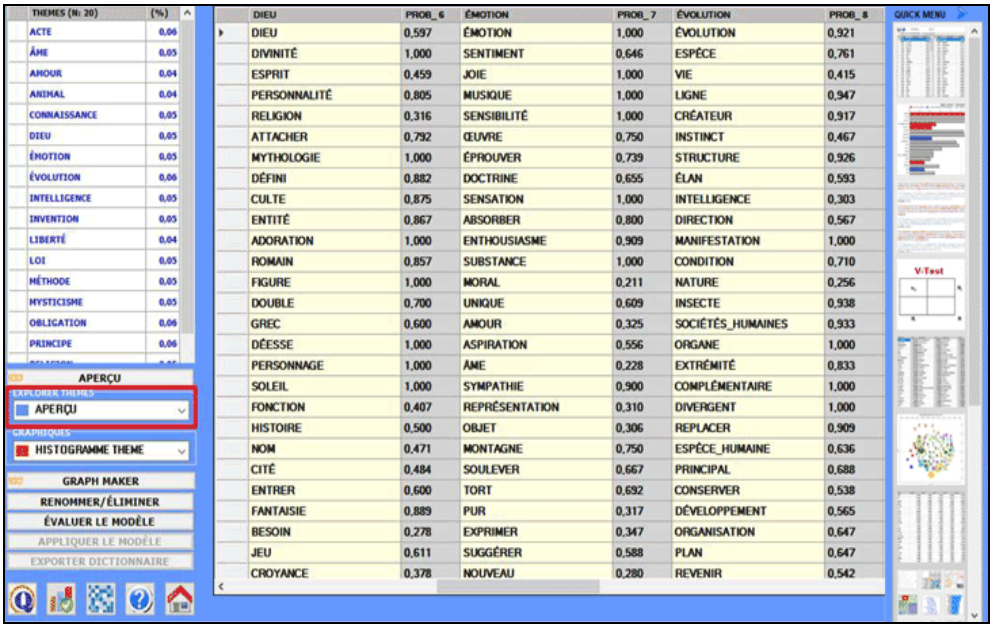
c – description de chaque thème à travers les valeurs de probabilité associées à leur mots caractéristiques, que ce soit spécifiques ou partagés par deux ou plusieurs thèmes.

A la fin du processus d'analyse, l'utilisateur pourra aisément effectuer les opérations suivantes :
1 – explorer les caractéristiques de chaque thème
2 – explorer les relations entre les différents thèmes
3 – renommer ou éliminer des thèmes spécifiques
4- vérifier la cohérence sémantique des différents thèmes;
5 – tester le modèle et assigner les thèmes aux unités de contexte, que ce soient documents ou contextes élémentaires.
6 – appliquer le modèle et créer une nouvelle variable thématique;
7 - exporter un dictionnaire des catégories.

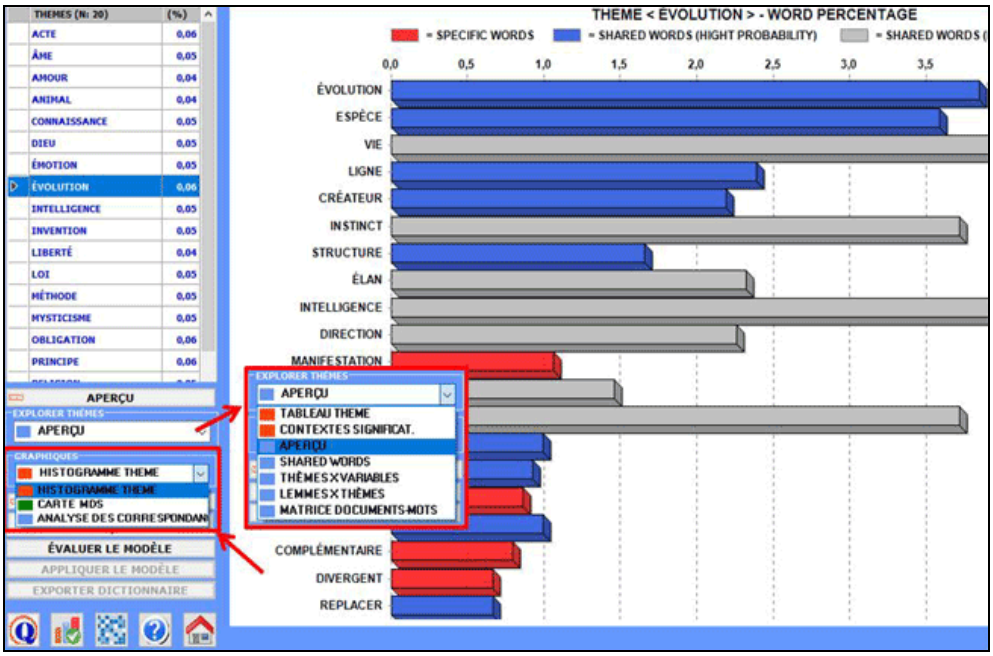
Dans le détail :

1 – Explorer les caractéristiques de chaque thème

La première sortie à être visualisée (et qui vous pouvez enregistrer) est un tableau avec un aperçu de tous les thèmes. Et, quand vous le souhaitez, le même tableau peut être affiché en utilisant le bouton "Aperçu" (voir ci-dessous).



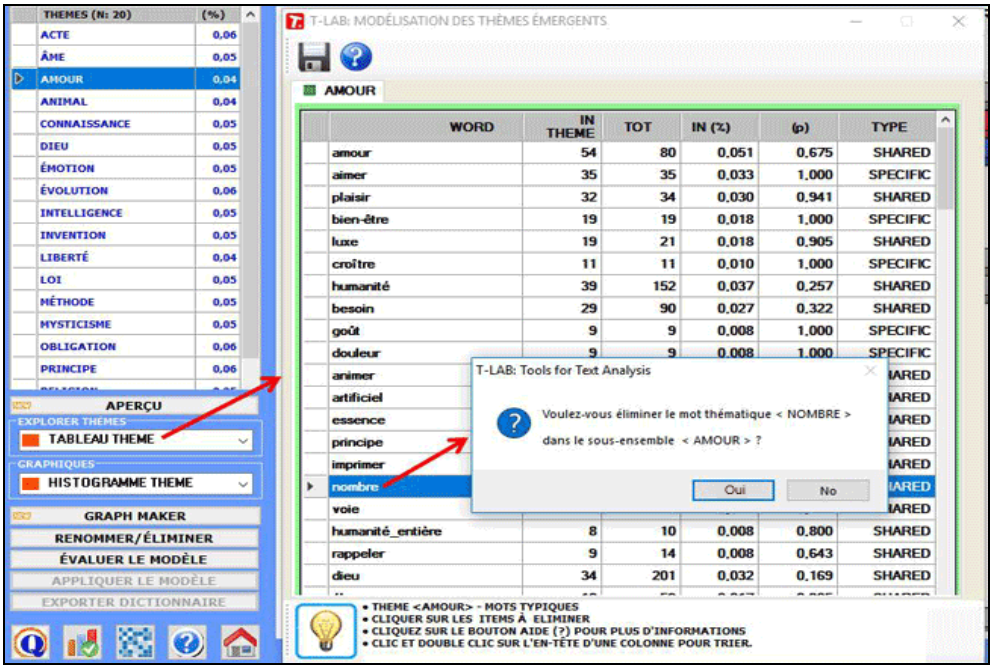
D'autres types de sortie sont accessibles en sélectionnant l'une des options mises en surbrillance dans l'image suivante.



N.B. : Dans ce graphique “hight probability” indique une probabilité >=.75.

Lorsqu'un thème est sélectionné, en cliquant sur l'option "Tableau Thème", vous pouvez vérifier ses caractéristiques. Aussi - en cliquant sur un mot du tableau - une autre option

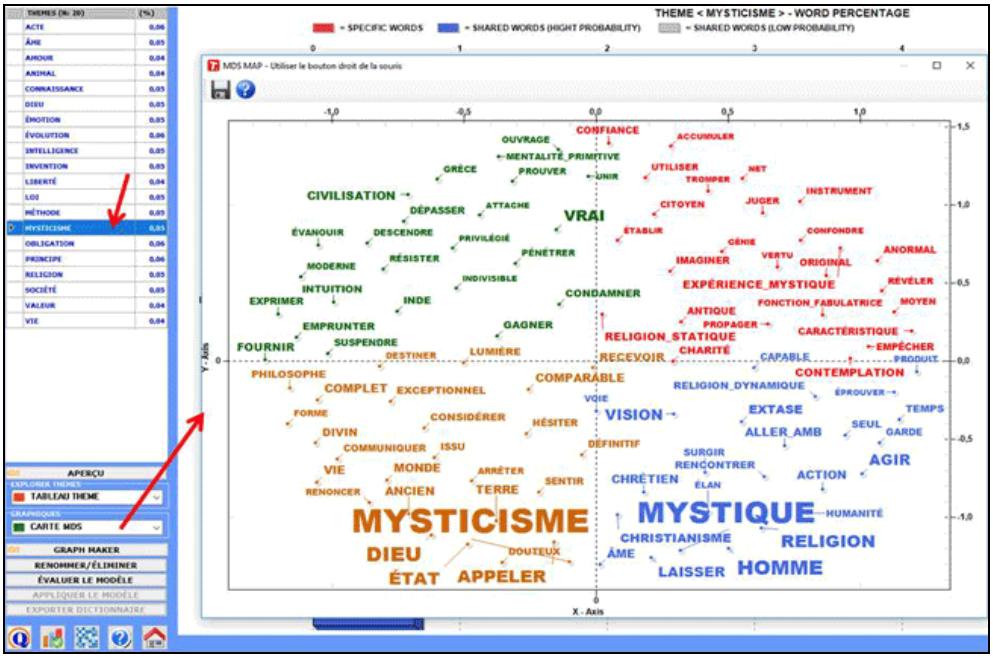
devient disponible qui vous permet de "supprimer" l'élément sélectionné (voir l'image ci-dessous).



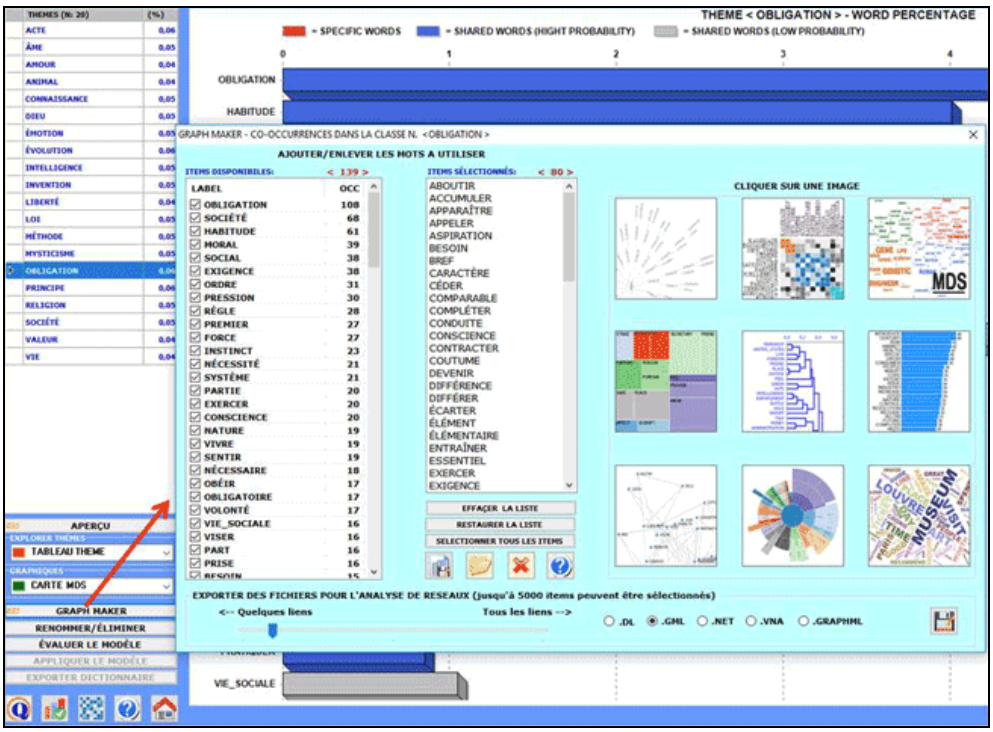
Les abréviations de ce tableau sont les suivantes:

IN THEME = occurrence (tokens) de chaque mot à l'intérieur du thème sélectionné
TOT = occurrence (tokens) de chaque mot à l'intérieur du corpus ou sous-ensemble analysé
IN (%) = poids en pourcentage de chaque mot à l'intérieur du thème sélectionné
(p)= valeur de probabilité associée à chaque relation mot/thème
TYPE = est marqué ***specific*** quand le mot (avec p=1) appartient seulement au thème sélectionné, et devient ***shared*** dans les autres cas (c'est-à-dire quand le mot est présent dans plus d'un thème)

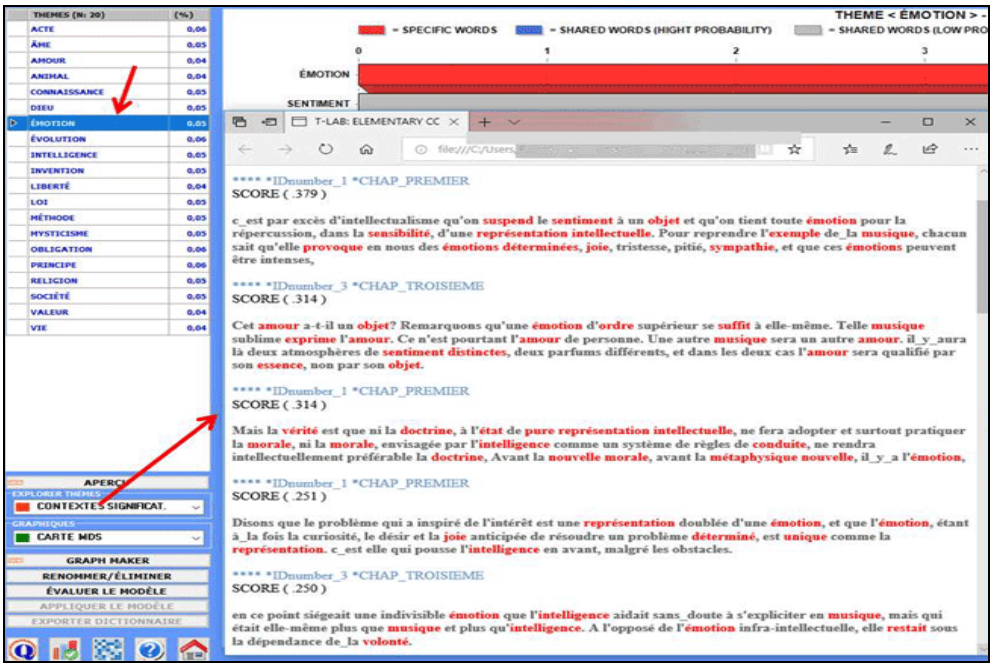
Lorsqu'un sujet est sélectionné, en cliquant sur l'option «Carte MDS» vous pouvez facilement explorer les relations sémantiques entre les mots qui sont plus caractéristiques (voir image ci-dessous).



De plus, en utilisant l'outil 'Graph Maker', des options graphiques supplémentaires deviennent disponibles (voir les images suivantes).



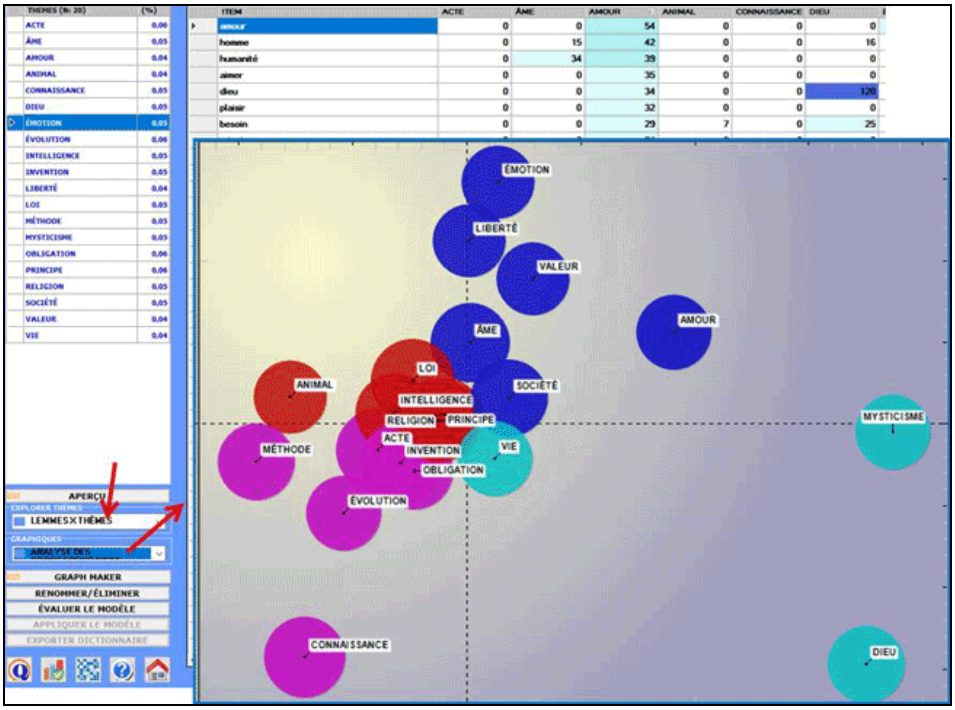
Lorsqu'un sujet est sélectionné, en cliquant sur l'option «contextes significatifs», un fichier HTML est créé où les 20 principaux segments de texte (qui correspondent le mieux aux caractéristiques du sujet) sont affichés (voir l'image ci-dessous).



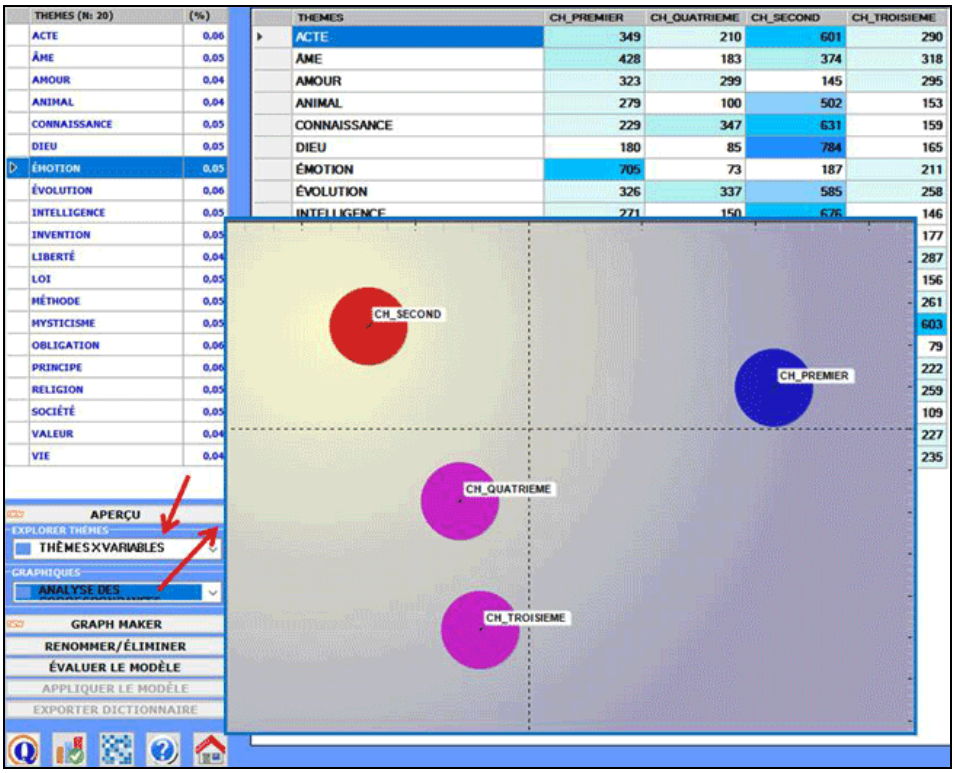
2 – Explorer les relations entre les différents thèmes

Avec l'outil 'Analyse de Correspondance', vous pouvez créer et explorer deux types de tableaux de contingence:

2.1) un tableau mots par thèmes (voir ci-dessous)

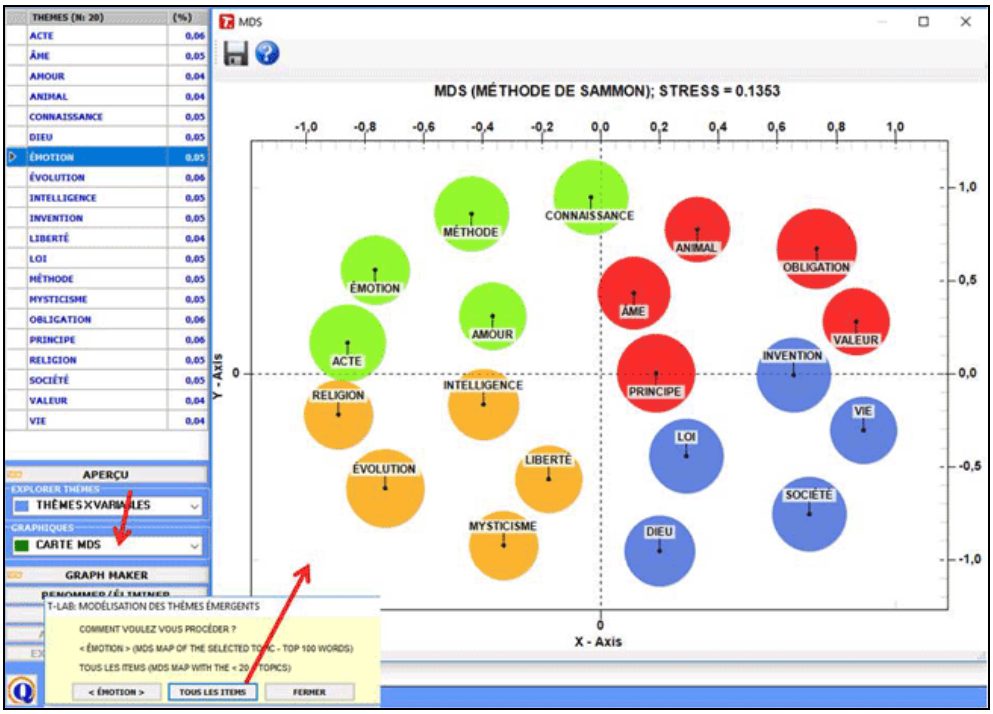


2.2) un tableau qui croise les thèmes avec les modalités de la variable sélectionnée

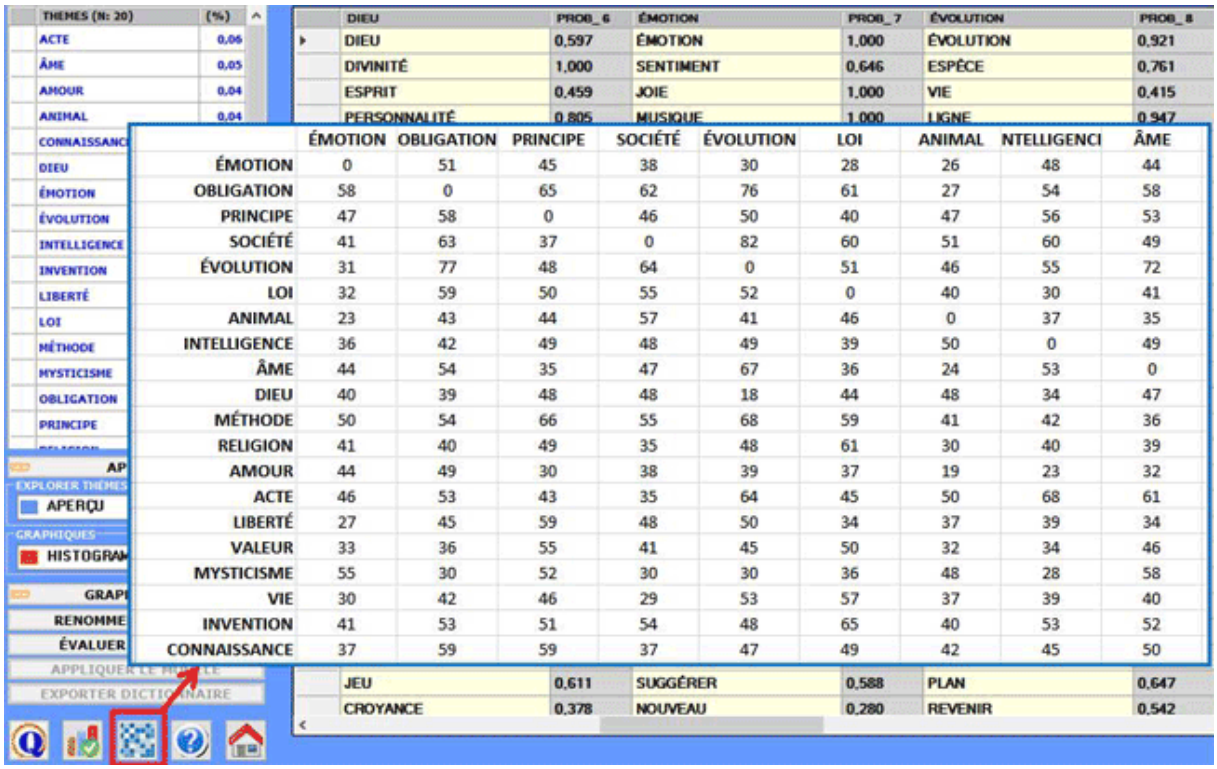


Il y a également deux autres options graphiques disponibles qui nous permettent de cartographier les relations entre les différents thèmes / topics.

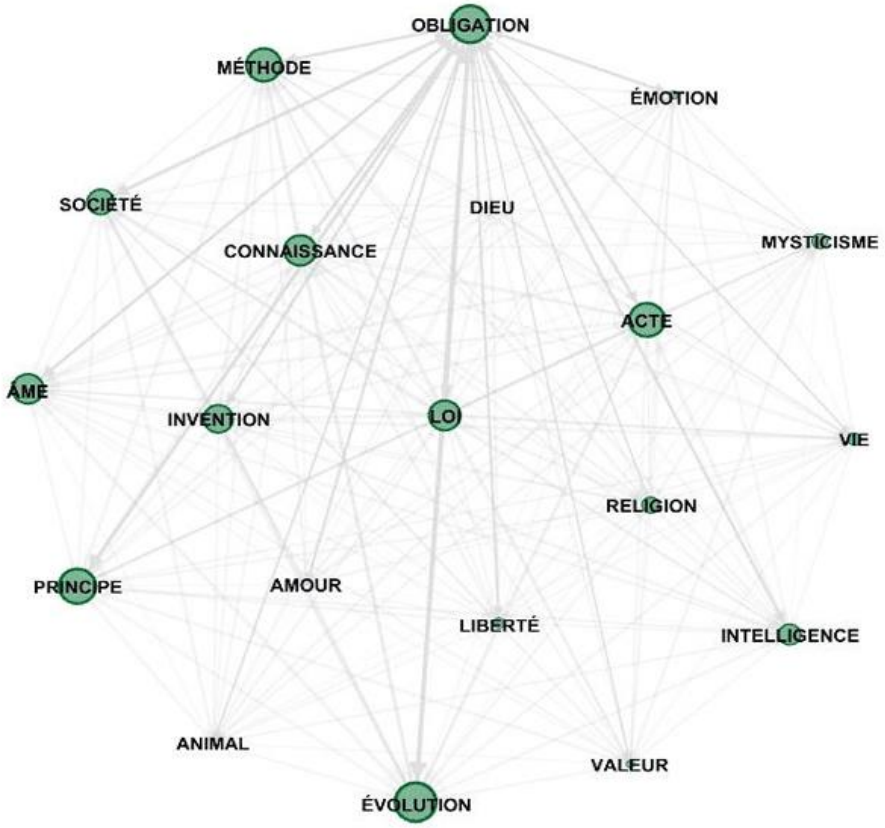
2.3) une carte MDS



2.4) certains diagrammes de réseau obtenu par l'exportation / importation du tableau d'adjacence créé par **T-LAB** (voir ci-dessous)



	ÉMOTION	OBLIGATION	PRINCIPE	SOCIÉTÉ	ÉVOLUTION	LOI	ANIMAL	NTELLIGENCI	ÂME
ÉMOTION	0	51	45	38	30	28	26	48	44
OBLIGATION	58	0	65	62	76	61	27	54	58
PRINCIPE	47	58	0	46	50	40	47	56	53
SOCIÉTÉ	41	63	37	0	82	60	51	60	49
ÉVOLUTION	31	77	48	64	0	51	46	55	72
LOI	32	59	50	55	52	0	40	30	41
ANIMAL	23	43	44	57	41	46	0	37	35
INTELLIGENCE	36	42	49	48	49	39	50	0	49
ÂME	44	54	35	47	67	36	24	53	0
DIEU	40	39	48	48	18	44	48	34	47
MÉTHODE	50	54	66	55	68	59	41	42	36
RELIGION	41	40	49	35	48	61	30	40	39
AMOUR	44	49	30	38	39	37	19	23	32
ACTE	46	53	43	35	64	45	50	68	61
LIBERTÉ	27	45	59	48	50	34	37	39	34
VALEUR	33	36	55	41	45	50	32	34	46
MYSTICISME	55	30	52	30	30	36	48	28	58
VIE	30	42	46	29	53	57	37	39	40
INVENTION	41	53	51	54	48	65	40	53	52
CONNAISSANCE	37	59	59	37	47	49	42	45	50

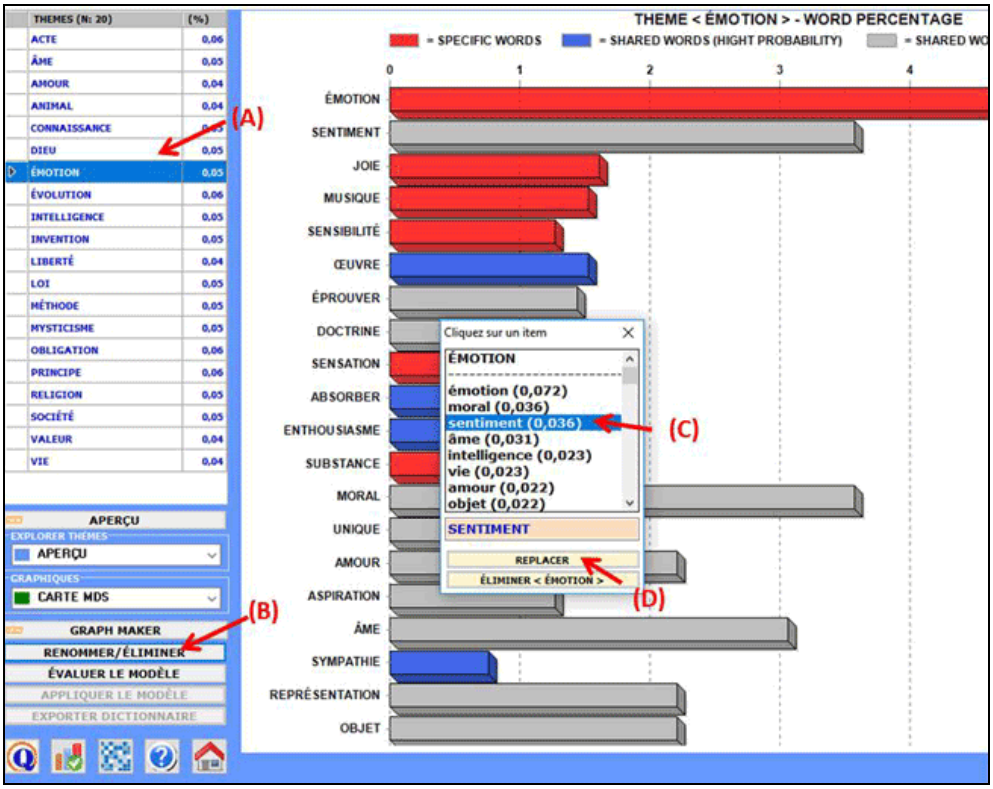


N.B. : le graphique précédent a été créé au moyen du logiciel Gephi (voir https://gephi.org/), après avoir importé un fichier **T-LAB**.

3- Renommer ou éliminer des thèmes spécifiques

Pour renommer ou éliminer des thèmes spécifiques il suffit de sélectionner les items correspondants (voir 'A' ci-dessous) et cliquer sur le bouton « renommer/éliminer » (voir 'B' ci-dessous).

Quand la boîte à options apparaît (voir ci-dessous), selon son propre objectif, l'utilisateur peut changer la désignation du thème (cela en choisissant parmi les mots disponibles ou en prenant un nouveau, voir 'C' ci-dessus) ou bien éliminer le thème sélectionné avec un clic sur le bouton correspondant (voir 'D' ci-dessous).



4 – Vérifier la cohérence sémantique des différents thèmes

Lorsqu' on clique sur le bouton 'Index de Qualité', **T-LAB** calcule les similarités entre les dix premiers mots (top 10) caractéristiques de chaque thème. Plus précisément:

- les 10 premiers mots sont ceux qui ont la plus grande valeur de probabilité;
- les mesures de similarité sont calculées en utilisant le coefficient du cosinus;
- comme dans le cas de l' outil '**Associations de mots'**, le coefficient du cosinus est calculé en vérifiant les cooccurrences de chaque paire de mots à l' intérieur des segments de texte définis en tant que contextes élémentaires.

En résultat, **T-LAB** crée un fichier HTML où les 'k' thèmes sont énumérés avec leur index correspondant de 'cohérence sémantique'.

N.B. : étant donné que les mesures de similitude varient suivant le changement des mots sélectionnés, il est recommandé de répéter la procédure chaque fois que l'un des dix premiers mots d' un thème est supprimé par l'utilisateur.

5 – Tester le modèle

A la fin de l'analyse des données (voir ci-dessus les points "A" et "B"), chaque unité de contexte (ex. un document ou un contexte élémentaire) est constituée d'un mélange de thèmes (ou sujets). Par ailleurs le procès de classification utilisé pour tester/appliquer le modèle assigne chaque unité de contexte au thème qui le caractérise le plus. Il en résulte que, à ce point, chaque thème devient de fait un cluster d'unités de contexte.
Pour cette raison quand l'option "tester le modèle" est sélectionnée, **T-LAB** crée deux fichiers XLS (voir ci-dessus) qui permettent à l'utilisateur de vérifier l'appartenance de chaque unité de contexte à un thème spécifique.

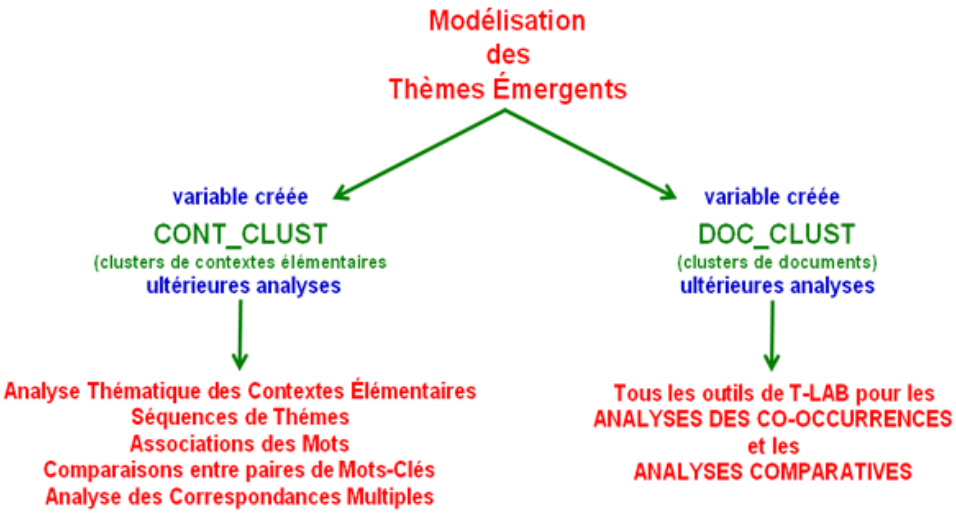
ID_DOC	BEST	ACTE	ÂME	AMOUR	ANIMAL	NNAISSAN	DIEU	ÉMOTION	ÉVOLUTION	ITELLIGEN	NVENTION	LIBERTÉ	LOI	MÉTHODE	MYSTICISME	OBLIGATION
1	15	4,131	5,529	5,762	4,433	2,644	2,692	13,907	6,429	4,109	3,290	5,785	5,354	4,463	3,203	19,362
2	6	7,131	4,551	2,020	7,607	7,553	19,503	2,922	11,491	11,310	7,256	4,271	6,924	5,245	2,776	3,863
3	14	3,356	3,916	5,519	2,270	1,580	4,482	3,787	4,946	2,149	1,966	4,349	1,634	3,611	14,870	0,861
4	8	2,230	2,209	4,385	1,544	3,756	1,323	0,954	5,863	1,777	3,858	2,221	2,067	2,778	3,441	2,567

N.B.: dans le tableau ci-dessus, chaque document a une valeur de probabilité associée à chaque theme.

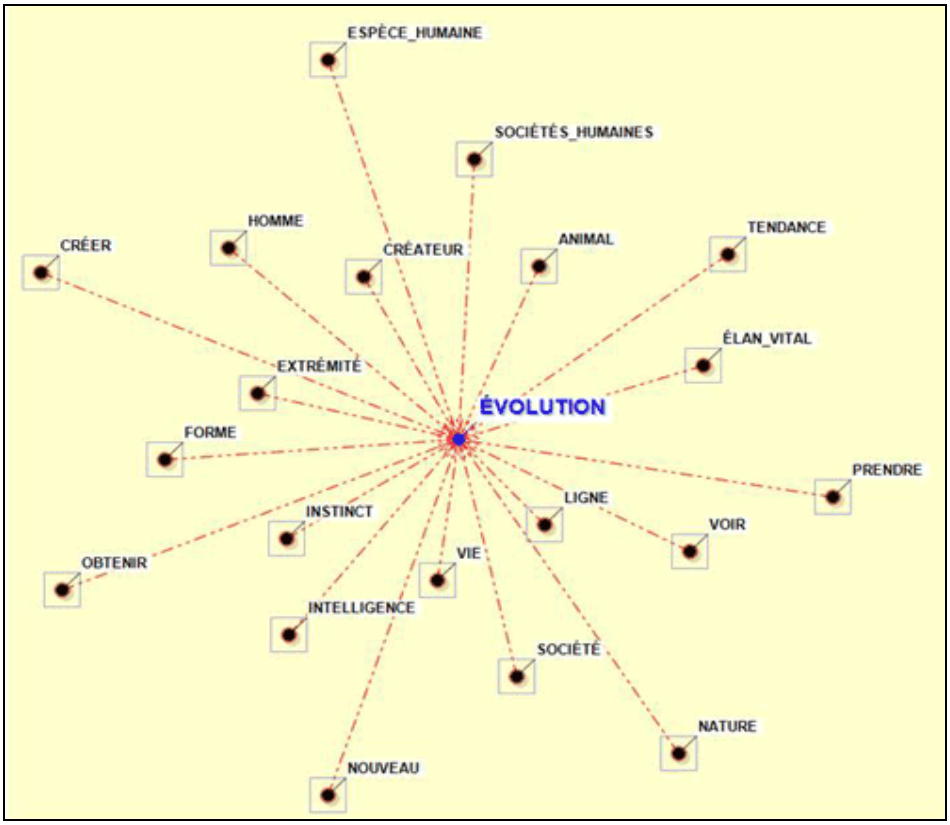
IdDoc	IdSeg	Topic	Score	Segm
2	1292	SOCIÉTÉ	0,612	A vrai dire , individu et société s ' impliquent réciproquement : les individus constitue
2	1219	DIEU	0,449	les dieux ne paraissent que plus tard , quand la substantialité pure et simple qu ' avaie
2	1223	DIEU	0,446	Ainsi , chez les Égyptiens , le dieu solaire Râ , d_abord objet d ' adoration suprême , a
3	1516	AMOUR	0,432	Il ne s ' empêcherait pas plus de_la répandre que le soleil de déverser sa lumière . Seu
1	630	SOCIÉTÉ	0,421	il faudra c et dans cet organisme plus ou moins artificiel l ' habitude joue le même rôl
1	628	SOCIÉTÉ	0,39	On se plaît à dire que la société existe , que dès lors elle exerce nécessairement sur se
4	1922	ÉVOLUTION	0,39	Nous en dirions autant de l ' instinct et de l ' intelligence , de_la vie animale et de_la v
1	143	OBLIGATION	0,381	Mais il est nécessaire qu ' il y ait des obligations
1	220	ÉMOTION	0,379	c_est par excès d ' intellectualisme qu ' on suspend le sentiment à un objet et qu ' on ti
4	2035	MYSTICISME	0,374	Mais , remontant des profondeurs obscures de l ' âme à_la surface de_la conscience ,
3	1581	DIEU	0,369	Par là , le et dès lors toutes se tiennent , elles forment un bloc . Beaucoup seraient d
2	751	SOCIÉTÉ	0,365	Si , dans c et ce tout , qui doit d ' être ce qu ' il est à l ' apport de ses parties , confère
2	1248	DIEU	0,364	c_est en Assyrie que la croyance à_la divinité des astres prit sa forme la plus systémati
1	590	SOCIÉTÉ	0,361	La vie aurait d_ailleurs pu s ' en tenir là , et ne rien faire de plus que de constituer des
1	111	OBLIGATION	0,359	Représentez-vous l ' obligation comme pesant sur la volonté à_la manière d ' une habit
2	1291	SOCIÉTÉ	0,352	Nous ne voulons pas dire non plus que la religion ait jamais été d ' essence sociale plut
1	468	VALEUR	0,349	Laissons de côté Platon , qui certainement comprend parmi les Idées suprasensibles ce
2	1258	DIEU	0,349	Celles-ci pouvaient d_ailleurs se réconcilier , les dieux du peuple subjugué entrant alo
1	401	VALEUR	0,344	quand des hommes se trouvent rapprochés les uns des autres par quelque marque dist
3	1562	MYSTICISME	0,343	bué autan elle oblige , mais ne nécessite pas . Celle-là est au_contraire inéluctable ,
1	133	OBLIGATION	0,342	Il en résul mais il s ' en faut que la distinction soit aussi nette pour la_plupart des hom
1	67	SOCIÉTÉ	0,331	nous appartenons à notre commune , a notre arrondissement , à notre département
2	1133	DIEU	0,315	On comprend d_ailleurs que chacune des deux continue à hanter l ' autre , qu ' il subsi
1	275	ÉMOTION	0,314	Mais la vérité est que ni la doctrine , à l ' état de pure représentation intellectuelle , n

6 – **Appliquer le modèle**

Après avoir appliqué et sauvegardé le modèle, tenant compte du fait que les thèmes sont archivés par **T-LAB** sous deux nouvelles variables qui se réfèrent à clusters de contextes élémentaires (CONT_CLUST) et/ou à clusters de documents (DOC_CLUST), les relations entre ces mêmes thèmes et/ou entre leurs caractéristiques pourront être ultérieurement explorées avec divers instruments d'analyse (voir ci-dessus).



Par exemple, en utilisant l'outil **Associations de Mots** et en sélectionnant le sous-ensemble (c.-à-d. le thème) « Evolution », vous pouvez créer le graphique ci-dessous.



7- **Exporter un dictionnaire des catégories**.
Lorsque cette option est sélectionnée, **T-LAB** crée un fichier dictionnaire avec l'extension **.dictio** prêt à être importé par l'intermédiaire d'un des outils pour l'analyse thématique. Dans ce dictionnaire chaque catégorie est décrite à travers ses mots caractéristiques.

Classification Thématique des Documents

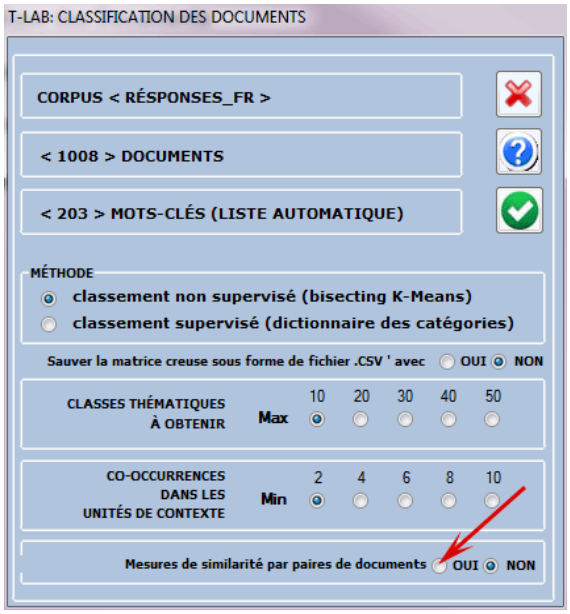
Cette fonction est activée uniquement lorsque le corpus que l'on analyse comprend d'un minimum de 20 à un maximum de 99.999 documents primaires.

Le processus d'analyse peut être effectué au moyen d'une méthode de clustering "non supervisée" (dans le cas particulier, un algorithme de bisecting K-Means) ou à travers une classification supervisée (c'est-à-dire une approche top-down). Lorsqu'on choisit la deuxième (c'est-à-dire la classification supervisée), on vous demande d'importer un dictionnaire des catégories, soit qu'il soit créé par une analyse précédente de **T-LAB** que construit par l'utilisateur.

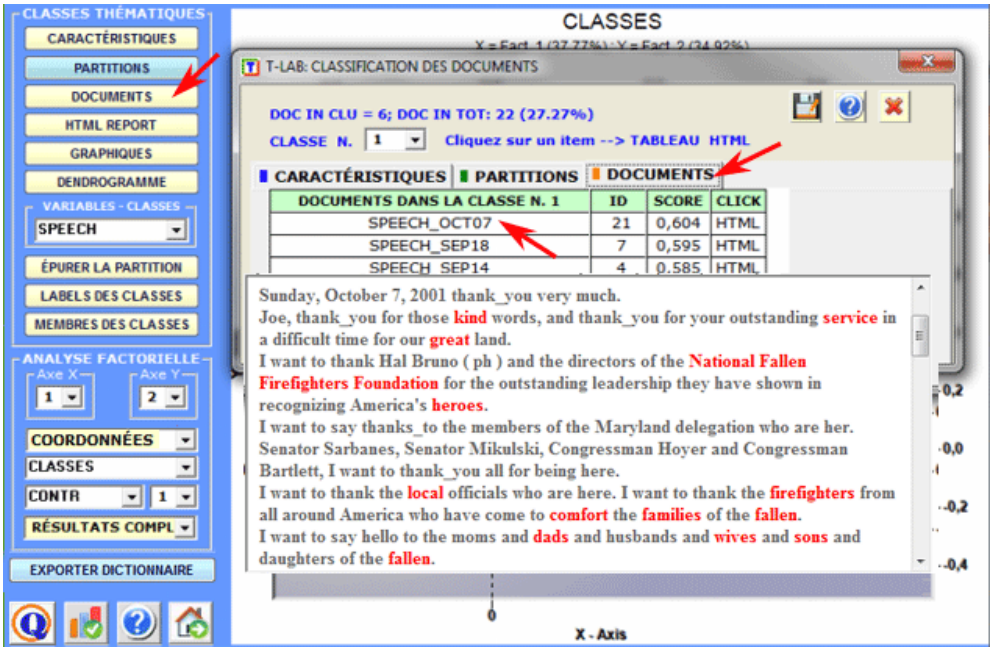
Son utilisation permet de construire des classes de documents et d'explorer leurs caractéristiques à l'aide d'opérations/options semblables à celles qui sont décrites dans la section du manuel dédiée à l'**Analyse Thématique des Contextes Elémentaires**.

Sa spécificité consiste dans le fait que le tableau analysé est formé par un nombre de lignes égal à celui des documents du corpus, chacun desquels est représenté comme un vecteur de valeurs indiquant les occurrences des mots qu'il contient.

En outre, lorsque les documents analysés ne dépassent pas les 3000, on peut obtenir des mesures de similarité (index du cosinus) entre chacun d'eux et tous les autres (voir ci-dessous). N.B. : dans ce cas le seuil minimum de l' 'index de similarité est fixé à 0,05.



D'ailleurs les outputs suivants sont différents:



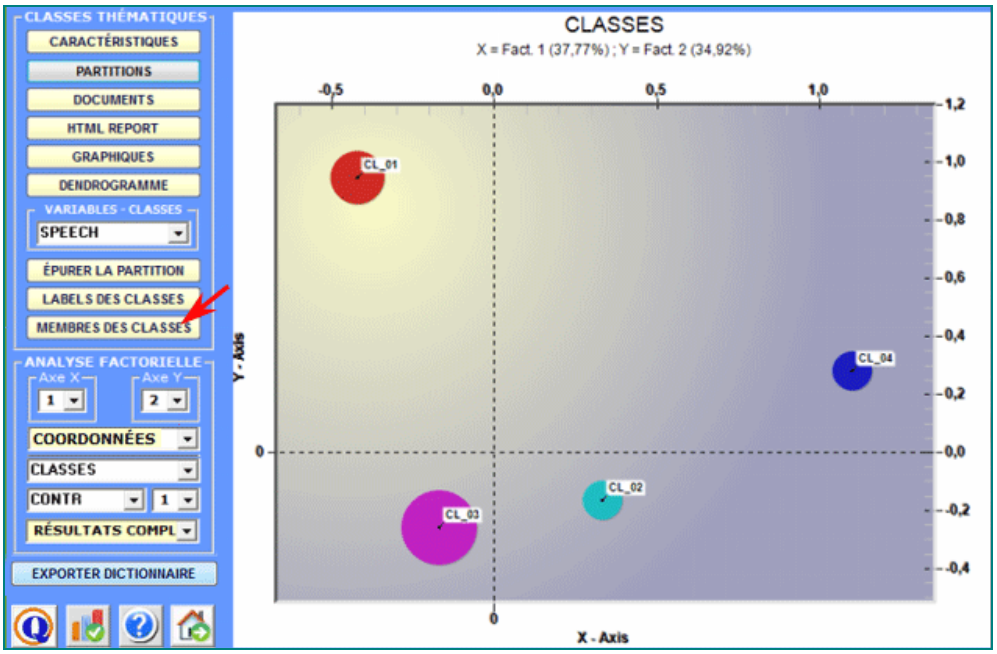
Les documents appartenants à chaque classe sont ordonnés par la valeur décroissante de leur importance et peuvent être explorés dans le format HTML.

Dans ce cas-ci la valeur d'importance (score) assignée à chaque document (i) de la classe (k) est obtenue en appliquant la formule suivante:

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

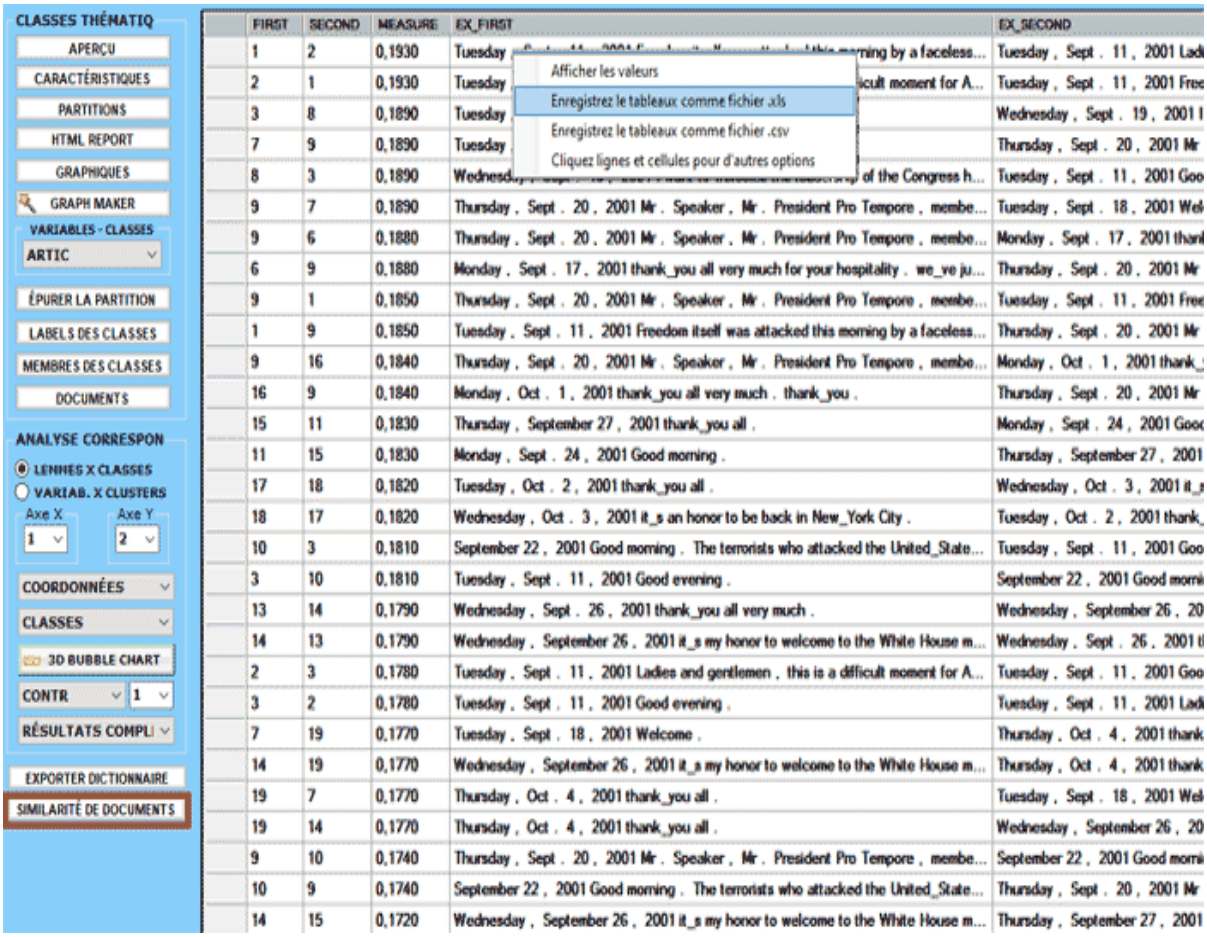
Où:
i - se réfère au document **i**;
k - se réfère à la classe **k**;
cos - est le symbole du cosinus;
$\mathbf{d_i}$ - est le vecteur normalisé du $\mathbf{TF_{j,i}\,IDF_j}$, où **j** se réfère à un mot du document **i** ;
$\mathbf{c_k}$ - est le vecteur normalisé du de $\mathbf{TF_{j,k}\,IDF_j}$, où **j** se réfère à un mot de la classe **k**.

En employant les scores obtenus par la formule ci-dessus, qui sont transformés en pourcentages, **T-LAB** rend disponible le fichier "Document_Membership_Degree.xls " (voir ci-dessous) contenant les classes auxquelles les documents sont assignés, soit par le bisecting K-Means (appartenance exclusive de chaque document à un classe) soit par le TF-IDF (appartenance mélangée de chaque document aux différentes classes).



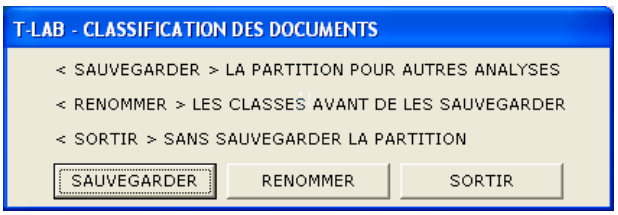
DOC_ID	VAR_01	CLUST_K-	BEST_TF-	MATCHIN(	CLUST-1	CLUST-2	CLUST-3	CLUST-4
1	BU_SEP11a	4	4	1	0,323	0,148	0,177	0,352
2	BU_SEP11b	1	1	1	0,478	0,107	0,159	0,256
3	BU_SEP11c	4	4	1	0,246	0,223	0,165	0,366
4	BU_SEP14	4	4	1	0,27	0,164	0,079	0,487
5	BU_SEP15	4	4	1	0,251	0,188	0,157	0,404
6	BU_SEP17	2	2	1	0,145	0,539	0,125	0,191
7	BU_SEP18	1	1	1	0,582	0,114	0,119	0,185
8	BU_SEP19	2	2	1	0,113	0,445	0,241	0,202
9	BU_SEP20	4	4	1	0,182	0,231	0,181	0,406
10	BU_SEP22	3	3	1	0,142	0,141	0,548	0,17
11	BU_SEP24	4	4	1	0,134	0,19	0,275	0,401
12	BU_SEP25	3	3	1	0,067	0,152	0,555	0,226
13	BU_SEP26a	2	2	1	0,117	0,536	0,165	0,183
14	BU_SEP26b	2	2	1	0,093	0,556	0,147	0,204
15	BU_SEP27	4	4	1	0,182	0,185	0,244	0,389
16	BU_OCT01	4	4	1	0,136	0,205	0,158	0,502
17	BU_OCT02	1	1	1	0,504	0,138	0,183	0,175
18	BU_OCT03	3	3	1	0,129	0,153	0,579	0,139
19	BU_OCT04	4	4	1	0,191	0,218	0,181	0,41
20	BU_OCT06	4	4	1	0,117	0,143	0,228	0,512
21	BU_OCT07	1	1	1	0,687	0,084	0,078	0,151
22	BU_OCT07	4	4	1	0,153	0,189	0,187	0,471

Lorsque le bouton **Similarité de Documents** est activé, en cliquant dessus on peut vérifier dans quelle mesure chaque document est similaire à chacun des autres. Dans ce cas, la mesure de similarité est le coefficient du cosinus et sa valeur varie en fonction de combien de mots ont été utilisés pour le classement thématique. L'image suivante décrit les options disponibles pour ce genre de vérification.



	FIRST	SECOND	MEASURE	EX_FIRST	EX_SECOND
	1	2	0,1930	Tuesday	Tuesday , Sept . 11 , 2001 Lad
	2	1	0,1930	Tuesday	Tuesday , Sept . 11 , 2001 Free
	3	8	0,1890	Tuesday	Wednesday , Sept . 19 , 2001 I
	7	9	0,1890	Tuesday	Thursday , Sept . 20 , 2001 Mr
	8	3	0,1890	Wednesd	Tuesday , Sept . 11 , 2001 Goo
	9	7	0,1890	Thursday , Sept . 20 , 2001 Mr . Speaker , Mr . President Pro Tempore , membe...	Tuesday , Sept . 18 , 2001 Wel
	9	6	0,1880	Thursday , Sept . 20 , 2001 Mr . Speaker , Mr . President Pro Tempore , membe...	Monday , Sept . 17 , 2001 thank
	6	9	0,1880	Monday , Sept . 17 , 2001 thank_you all very much for your hospitality . we_ve ju...	Thursday , Sept . 20 , 2001 Mr
	9	1	0,1850	Thursday , Sept . 20 , 2001 Mr . Speaker , Mr . President Pro Tempore , membe...	Tuesday , Sept . 11 , 2001 Free
	1	9	0,1850	Tuesday , Sept . 11 , 2001 Freedom itself was attacked this morning by a faceless...	Thursday , Sept . 20 , 2001 Mr
	9	16	0,1840	Thursday , Sept . 20 , 2001 Mr . Speaker , Mr . President Pro Tempore , membe...	Monday , Oct . 1 , 2001 thank_
	16	9	0,1840	Monday , Oct . 1 , 2001 thank_you all very much . thank_you .	Thursday , Sept . 20 , 2001 Mr
	15	11	0,1830	Thursday , September 27 , 2001 thank_you all .	Monday , Sept . 24 , 2001 Good
	11	15	0,1830	Monday , Sept . 24 , 2001 Good morning .	Thursday , September 27 , 2001
	17	18	0,1820	Tuesday , Oct . 2 , 2001 thank_you all .	Wednesday , Oct . 3 , 2001 it_
	18	17	0,1820	Wednesday , Oct . 3 , 2001 it_s an honor to be back in New_York City .	Tuesday , Oct . 2 , 2001 thank_
	10	3	0,1810	September 22 , 2001 Good morning . The terrorists who attacked the United_State...	Tuesday , Sept . 11 , 2001 Goo
	3	10	0,1810	Tuesday , Sept . 11 , 2001 Good evening .	September 22 , 2001 Good morni
	13	14	0,1790	Wednesday , Sept . 26 , 2001 thank_you all very much .	Wednesday , September 26 , 20
	14	13	0,1790	Wednesday , September 26 , 2001 it_s my honor to welcome to the White House m...	Wednesday , Sept . 26 , 2001 t
	2	3	0,1780	Tuesday , Sept . 11 , 2001 Ladies and gentlemen , this is a difficult moment for A...	Tuesday , Sept . 11 , 2001 Goo
	3	2	0,1780	Tuesday , Sept . 11 , 2001 Good evening .	Tuesday , Sept . 11 , 2001 Lad
	7	19	0,1770	Tuesday , Sept . 18 , 2001 Welcome .	Thursday , Oct . 4 , 2001 thank
	14	19	0,1770	Wednesday , September 26 , 2001 it_s my honor to welcome to the White House m...	Thursday , Oct . 4 , 2001 thank
	19	7	0,1770	Thursday , Oct . 4 , 2001 thank_you all .	Tuesday , Sept . 18 , 2001 Wel
	19	14	0,1770	Thursday , Oct . 4 , 2001 thank_you all .	Wednesday , September 26 , 20
	9	10	0,1740	Thursday , Sept . 20 , 2001 Mr . Speaker , Mr . President Pro Tempore , membe...	September 22 , 2001 Good morni
	10	9	0,1740	September 22 , 2001 Good morning . The terrorists who attacked the United_State...	Thursday , Sept . 20 , 2001 Mr
	14	15	0,1720	Wednesday , September 26 , 2001 it_s my honor to welcome to the White House m...	Thursday , September 27 , 2001

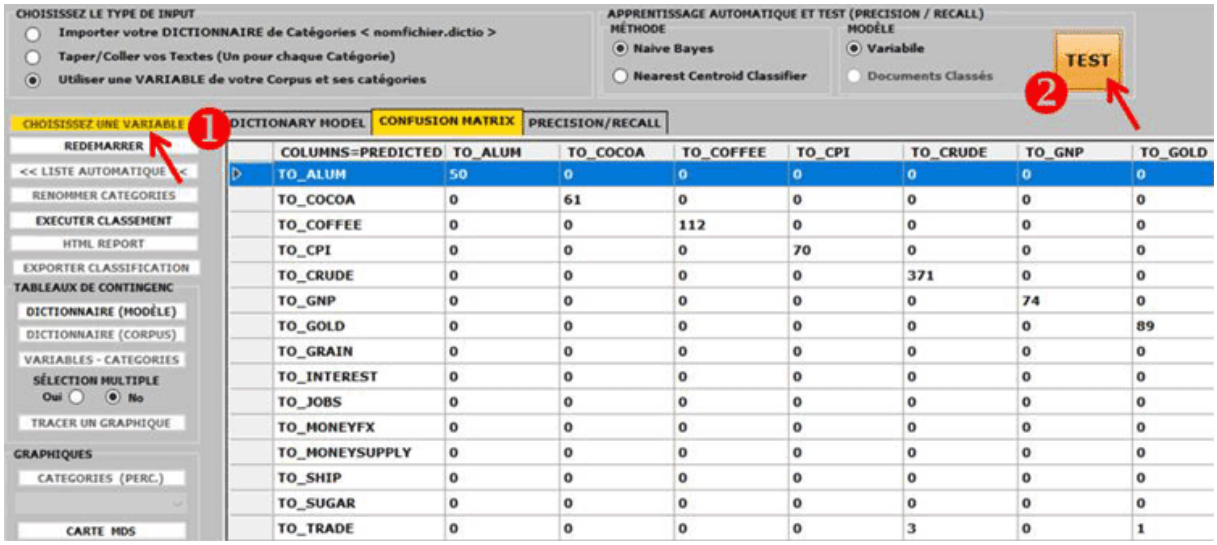
Lorsqu'on quitte cette fonction, des messages rappellent qu'il est possible d'explorer les classes obtenues avec d'autres outils **T-LAB**.



Si on choisit l'option "Sauvegarder", la variable < **DOC_CLUST** > (classes de documents) demeure disponible pour toutes les analyses suivantes du même corpus effectuées avec d'autres outils **T-LAB.**

Classification basée sur des Dictionnaires

N.B. : Les images de cette section font référence à une version précédente de T-LAB. En **T-LAB 10**, l'aspect est légèrement différent. En particulier, une nouvelle fonctionnalité permet de tester facilement n'importe quel modèle sur des données étiquetées (par exemple des données qui incluent des thèmes obtenus à partir d'une analyse qualitative précédente) et d'obtenir des résultats comme des matrices de confusion et des métriques de précision / rappel (voir image ci-dessous).



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

Cet outil **T-LAB** vous permet d'effectuer un **classement automatique** des **unités lexicales** (c' est-à-dire des mots et des lemmes, multiwords compris) ou des **unités de contexte** (c' est-à-dire des phrases, des paragraphes ou des documents courts) qui se trouvent dans un corpus en appliquant un ensemble de catégories prédéfinies ou bien choisies par l' utilisateur.

Selon le type de catégories utilisées, lesquelles peuvent être contenues dans un dictionnaire opportunément importé ou produites par **T-LAB**, ce classement peut être considéré un type d'**analyse du contenu** ou bien un type de **sentiment analysis**.

Puisque le processus d'analyse permet de créer de nouvelles variables et d' autres dictionnaires qui peuvent être exportés et importés dans d' autres projets d'analyse, cet outil peut aussi être utilisé soit pour explorer le même corpus à partir de points de vue différents, soit pour analyser deux ou plusieurs ensembles de textes en appliquant les mêmes modèles.

Parmi les **utilisations possibles** de cet outil, nous signalons les suivantes:

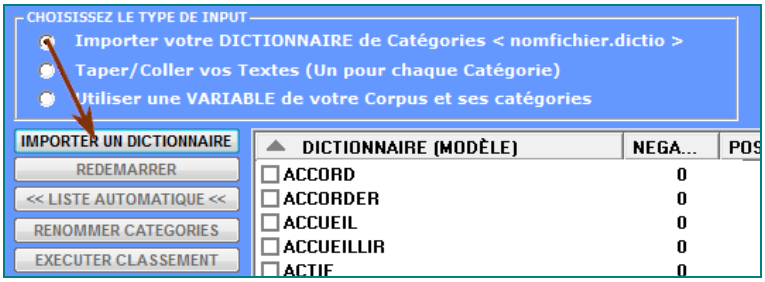
- Codage automatique de réponses à questions ouvertes;
- Analyse top-down des discours politique ;
- Sentiment Analysis de commentaires concernant des produits spécifiques;
- Vérification du processus psychothérapeutique;
- Validation de méthodes pour l' analyse qualitative.

De suite on trouve une brève description des quatre étapes principales du processus d'analyse, lesquelles - cependant - doivent être considérées indépendantes les unes des autres. En effet, le chercheur peut utiliser cet outil seulement pour personnaliser ses dictionnaires ou pour explorer son set de données.

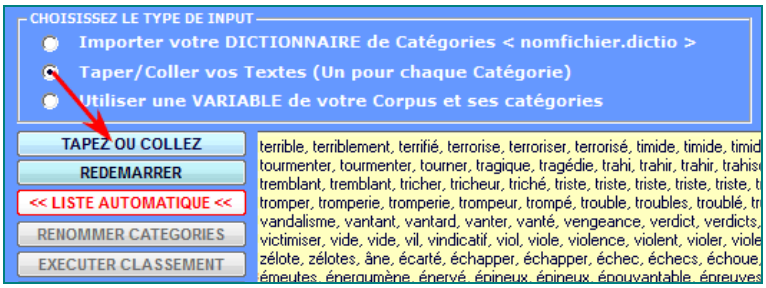
A) - PHASE DE PRÉ-PROCESSING

Les points de départ et les types correspondants de l'**input** de la phase de pré-processing peuvent être trois:

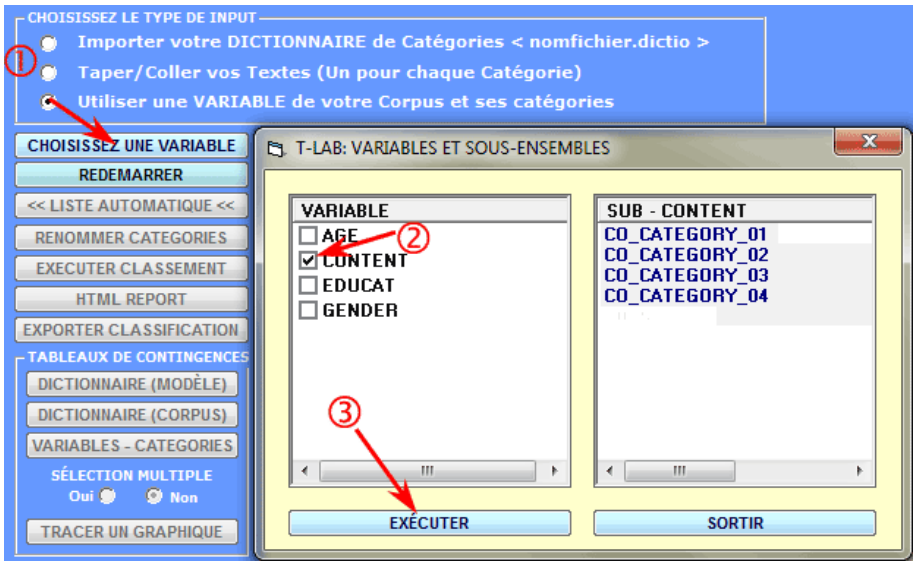
1 - un **dictionnaire** des catégories dans le format approprié est déjà disponible (voir les informations correspondantes dans la section 'E' de ce document). Dans ce cas, il suffit de cliquer l'option '**Importer un dictionnaire**' (voir ci-dessous);



2 - un dictionnaire des catégories doit dériver d'**exemples** de texte ou des **listes de mots** fournies par l'utilisateur. Dans ce cas, il suffit de taper ou copier / coller les textes dans la case appropriée (un exemple pour chaque catégorie, un après l'autre, maximum 100.000 caractères chacun);

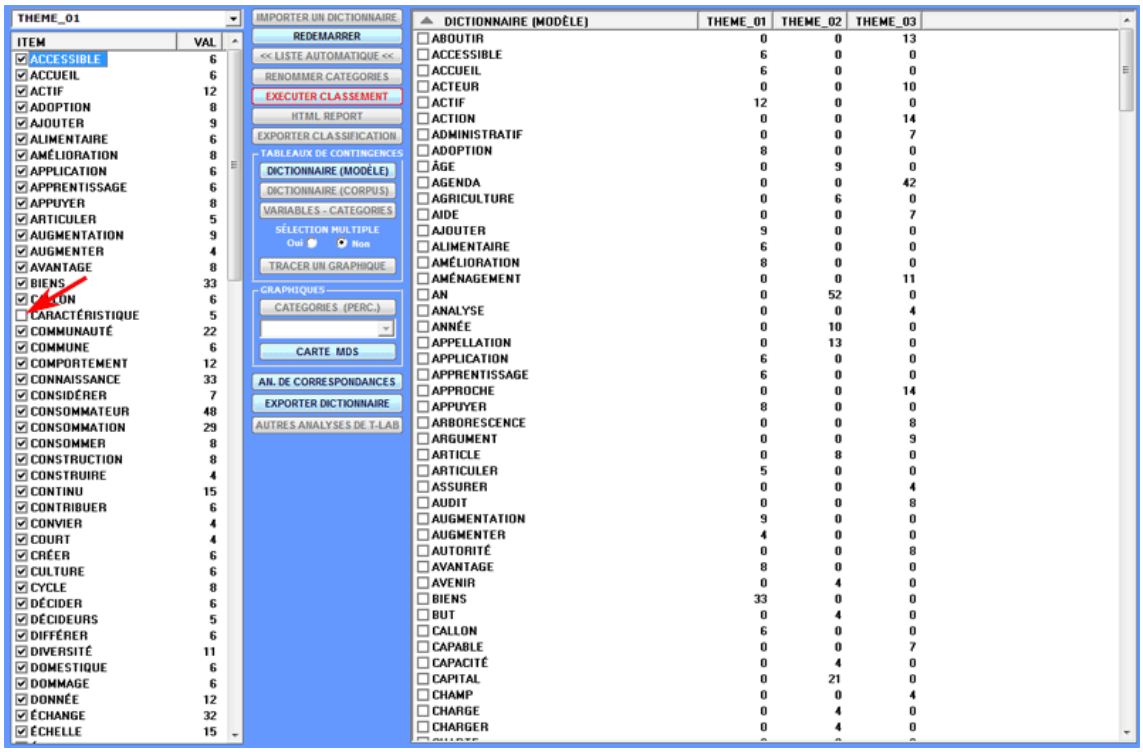


3 - un dictionnaire des catégories doit dériver d'une **variable** dérivant d'une d'analyse du contenu préalable. Dans ce cas, il suffit de cliquer sur l' option '**Choisissez une Variable**' et effectuer les choix appropriés (voir ci-dessous).



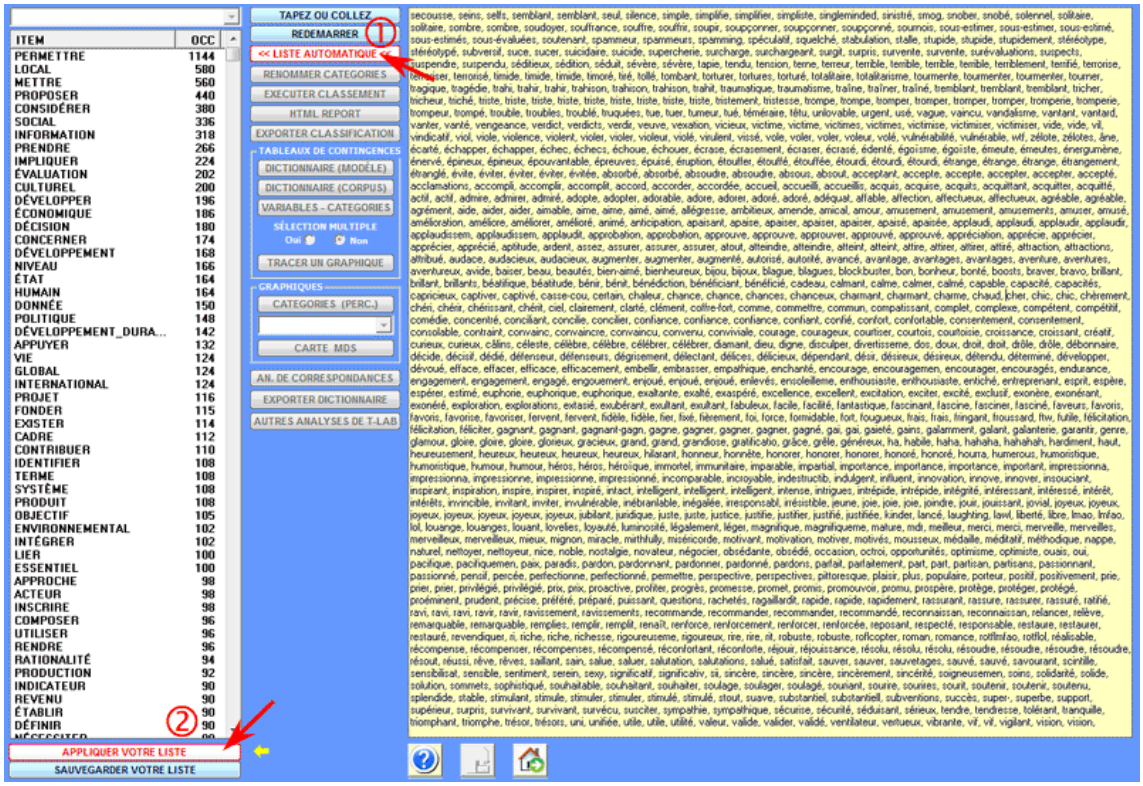
Selon un des trois cas énumérés ci-dessus, avant d'activer l'option 'Exécuter Classement', **T-LAB** fonctionne comme suit:

1 - le dictionnaire importé est transformé en un tableau de contingence que l'utilisateur peut explorer de diverses façons (voir la section 'C' du présent document) ; en outre, en sélectionnant chaque catégorie, un ou plusieurs des éléments correspondants peuvent être éliminés (voir image ci-dessous).



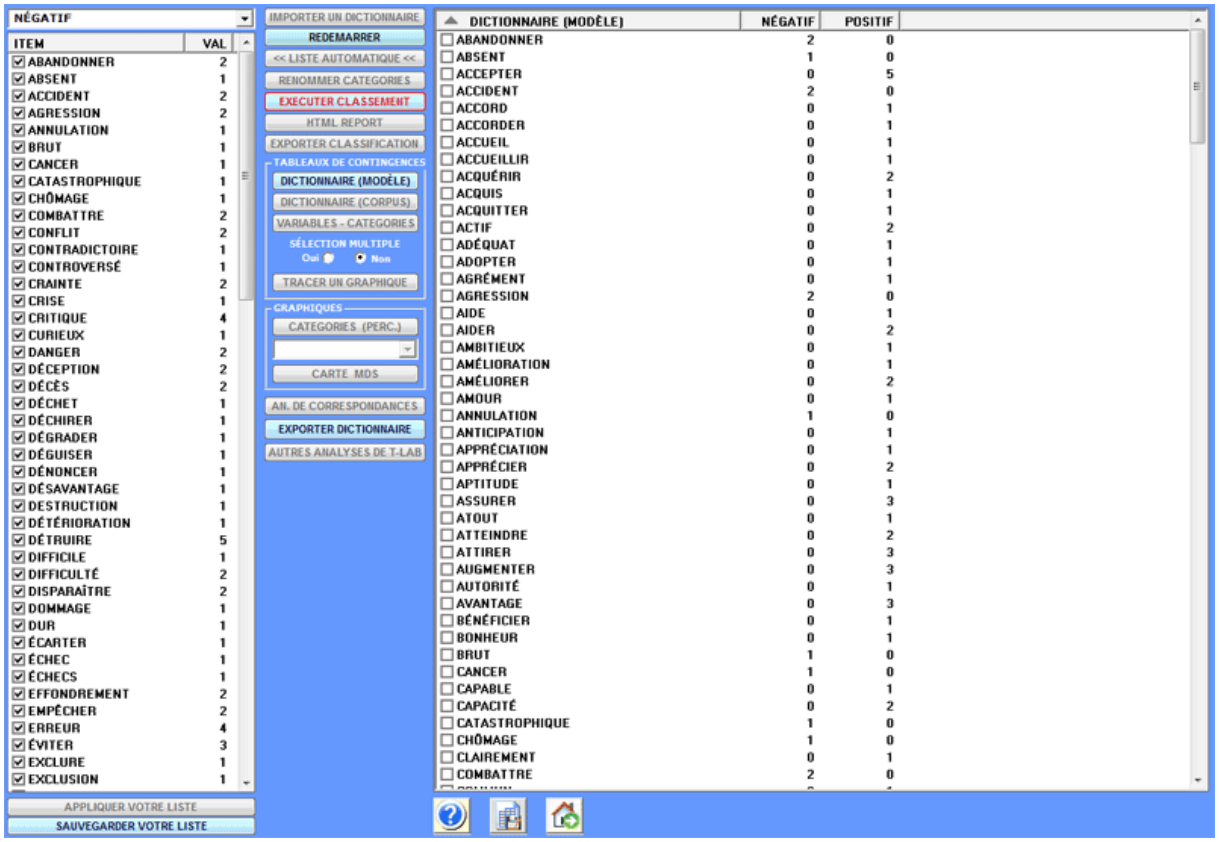
2 - lorsque les textes de l' exemple sont insérés dans la case correspondante, après avoir cliqué sur le bouton '**Liste Automatique**' (voir ci-dessous), **T-LAB**effectue un type spécifique de lemmatisation qui utilise seulement le vocabulaire du corpus sélectionné (voir la liste des mots dans la zone gauche de l'image suivante), puis, convertit chaque texte dans une liste dont

les éléments peuvent être sélectionnés et désélectionnés. Ensuite, pour valider chaque liste de mots (c'est-à-dire chaque catégorie du dictionnaire), il faut cliquer sur l' option '**Appliquer votre liste**' (voir ci-dessous). Toutes les opérations mentionnées doivent être répétées pour chaque catégorie du dictionnaire, ensuite l'utilisateur est autorisé à effectuer les opérations décrites dans la section 'C' de ce document.

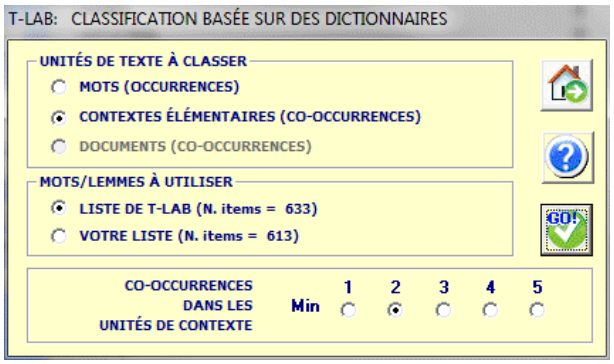


3 - lorsque vous sélectionnez une variable résultant d'une précédente analyse du contenu, **T-LAB** visualise le tableau relatif de contingence des mots par catégories et l'utilisateur peut effectuer toutes les opérations d' exploration des données (voir la section 'C' du document présent).

B) - PROCESSUS DE CLASSIFICATION



Après avoir cliqué sur l' option '**Exécuter classement**' (voir ci-dessus), selon le type de l' analyse de corpus , l' utilisateur peut effectuer les choix suivants:



À ce stade, si l'utilisateur décide de **classer les mots**, d'autres choix ne sont pas disponibles ; en effet, dans ce cas, les occurrences de chaque mot (c'est-à-dire les words tokens) sont tout simplement comptées comme les occurrences de la catégorie correspondante. Par exemple, si une catégorie de notre dictionnaire est 'religion' et celle-ci inclut des mots comme 'foi' et 'prière', lorsque l' on analyse un document qui contient les deux mots en question, **T-LAB** se limite à regrouper leurs occurrences. Par exemple, 2 occurrences de 'foi' et 3 occurrences de 'prière' deviennent 5 occurrences de 'religion'.

Sinon, si l'utilisateur décide de **classer les unités de contexte** (c' est-à-dire 'contextes élémentaires' comme des phrases et des paragraphes ou des 'documents '), **T-LAB** considère aussi bien les catégories dictionnaire que les unités de contexte à classer comme des profils de

co-occurrences (c'est-à-dire term vectors) et calcule leurs mesures de similarité. A cet effet, les profils de co-occurrences peuvent être filtrés à travers une 'Liste de T-LAB' (c' est-à-dire à partir d'une liste qui comprend tous les mots-clés avec les valeurs d'occurrence supérieures ou égales au seuil minimum de 4) ou à travers une liste personnalisée (c' est-à-dire une liste qui comprend tous les mots-clés dérivant d' un choix de la part de l'utilisateur), listes qui , toutefois, peuvent même parfois résulter égales. En outre , dans ces cas , **T-LAB** permet d'exclure de l'analyse des unités de contexte qui ne contiennent pas un nombre minimum de mots-clés en leur intérieur (voir ci-dessus le paramètre 'co-occurrences dans les unités de contexte').

Lorsque, comme dans le cas que l' on vient de décrire, les 'objets' à classer sont les unités de contexte, **T-LAB** se déroule comme suit:

a) il normalise les vecteurs correspondant aux catégories 'k' du dictionnaire utilisé, c' est-à-dire les profilés de colonne relatifs ;
b) il normalise les vecteurs correspondant aux unités de contexte à analyser;
c) il calcule des mesures de similarité (cosinus) et de différence (distance euclidienne) entre chaque vecteur 'i' correspondant à une unité de contexte et chaque vecteur 'k'
correspondant à une catégorie du dictionnaire utilisé ;
d) il attribue chaque unité de contexte ('i') à la classe ou à la catégorie ('k') avec laquelle il a une relation de similitude plus élevée. (Note: dans tous les cas, pour chaque paire 'unités de contexte / 'catégorie' il doit y avoir une correspondance entre la valeur maximale du cosinus et la valeur minimale de la distance euclidienne, sinon **T-LAB** considère l' unité de contexte 'i' comme 'non classifiée'.

Autrement dit, dans le cas que l' on vient de décrire, **T-LAB** utilise une sorte de méthode K-means où les centroïdes 'K' sont définis a priori et ils ne sont pas mis à jour pendant le processus d' analyse.

Étant donné que dans ce cas, la classification est de genre top-down , la qualité des résultats obtenus dépend essentiellement de deux facteurs :
1 - la 'pertinence' du dictionnaire utilisé (voir la relation entre le lexique du corpus et le dictionnaire des catégories) ;
2 - la capacité 'discriminante' de chacune des catégories (voir la relation entre les différentes catégories du dictionnaire).
En effet, lorsque ces deux facteurs sont optimaux, les deux paramètres de 'précision' et 'recall' (voir http://en.wikipedia.org/wiki/Precision_and_recall) ont des valeurs comprises entre 80% et 95%.

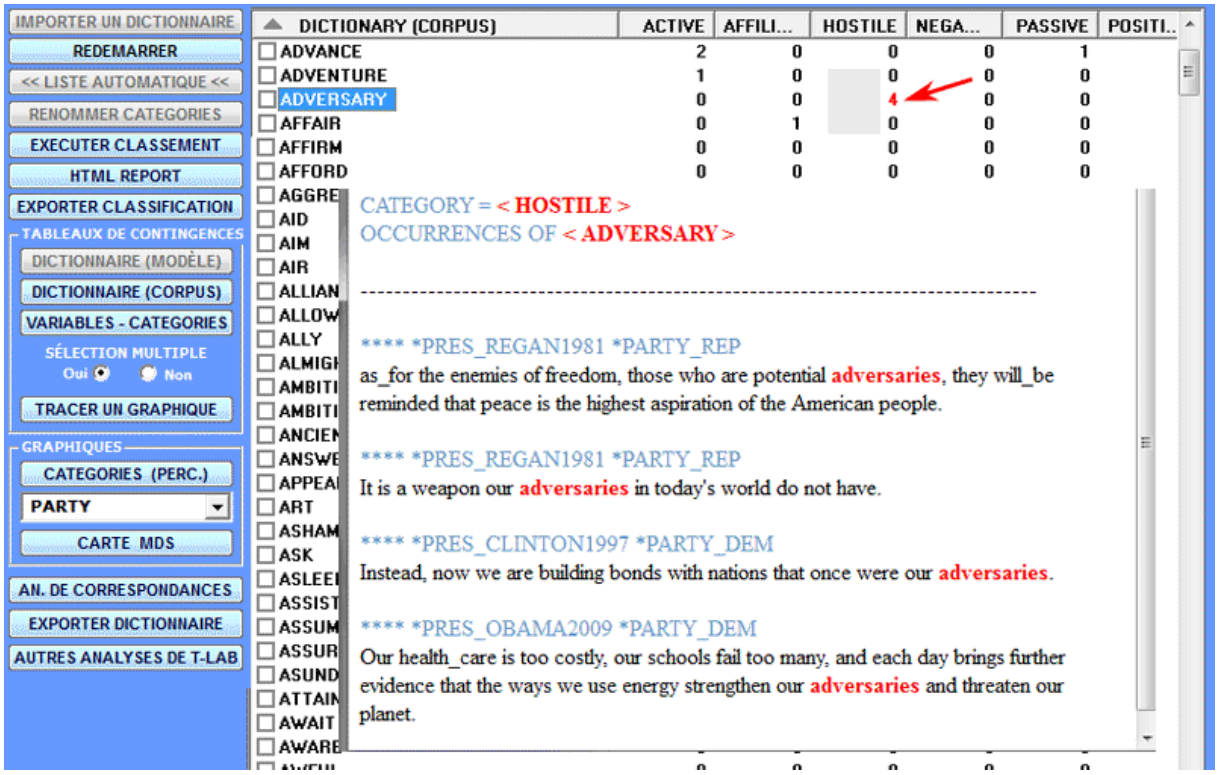
On rappelle qu' en ce moment **T-LAB** ne tient pas compte des formules de la négation, par conséquent, en effectuant une sentiment analysis, une phrase comme 'N' hais pas ton ennemi' peut être classée comme une tonalité 'négative' . Les utilisateurs experts peuvent gérer ce problème durant l'importation corpus (voir l'utilisation de listes pour les stop-words et multi-words). Par exemple, l'expression "N' hais pas" peut être transformée en "N_hais_pas" et, si on le retient approprié, elle peut être incluse dans la catégorie 'positif'.

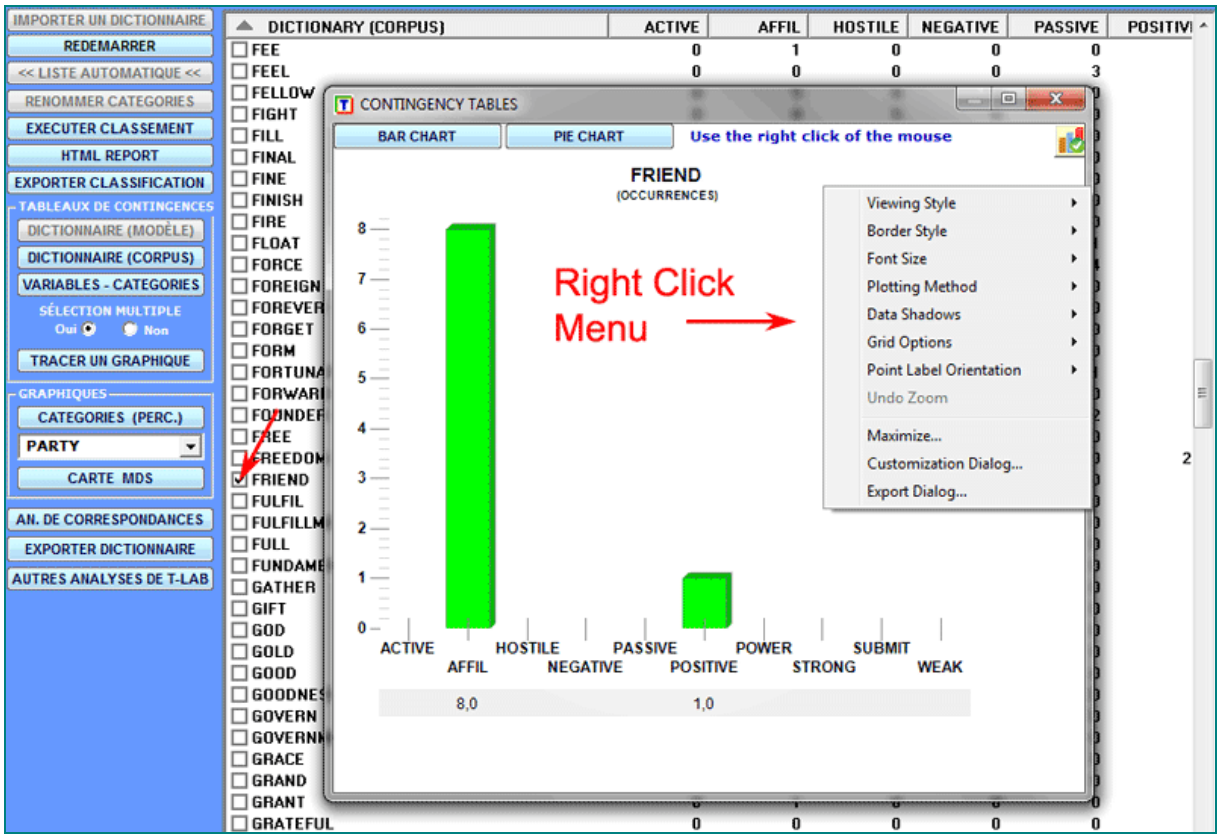
C) - EXPLORATION DES DONNÉES

Dans l'utilisation de cet instrument toute activité d'exploration se réfère à des **tableaux de contingence** dans lesquels, selon les cas, peuvent être représentées soit les données en input (par exemple un dictionnaire de catégories) que les données en output (par exemple les résultats du processus de classification).

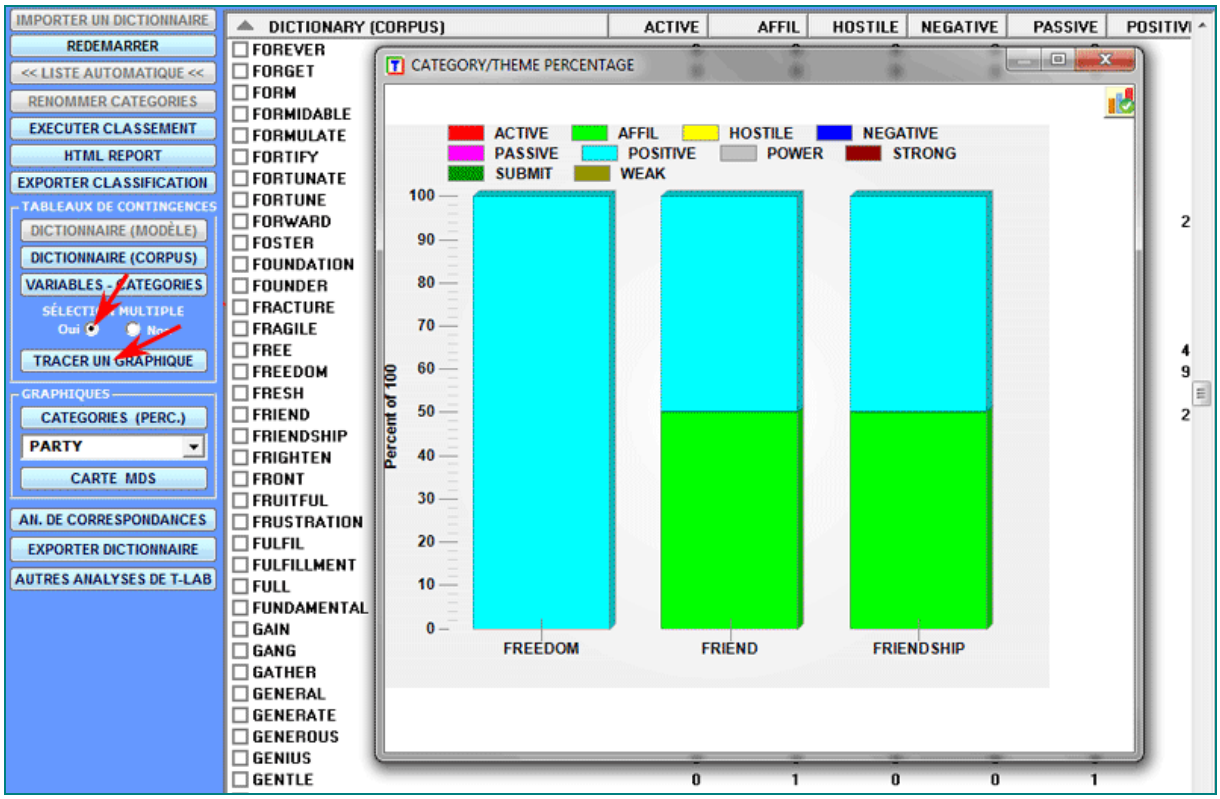
En particulier, en ce qui concerne les résultats de l'analyse, en fonction des unités textuelles classées - respectivement (a) 'mots', (b) 'contextes élémentaires' ou (c) 'documents' - les cellules des tableaux affichés contiennent les valeurs suivantes:
a) total des occurrences de chaque mot qui, dans le corpus analysé ou d' un de son sous-ensemble, a été classé comme appartenant à une catégorie prédéfinie (c' est-à-dire à la colonne ' j ' du tableau de contingence respectif). À noter que dans ce type de classification les mots appartenant simultanément à deux ou plusieurs catégories ont les mêmes valeurs répétées dans les colonnes correspondantes ;
b) total des contextes élémentaires affectés à une catégorie particulière (soit la colonne ' j ' dans laquelle est présent le mot dans la ligne (' i ') correspondante ;
c) total des occurrences de chaque mot (voir les lignes du tableau de contingence relatif) dans les documents attribués à chaque catégorie (voir les colonnes du tableau de contingence.

En cliquant sur les check-box correspondant aux différents items en ligne on peut obtenir des graphiques qui peuvent être personnalisés de différentes manières ; en outre, mais seulement en cas de classification de type 'b' (voir ci-dessus), en cliquant sur les valeurs contenues dans les cellules il est possible de visualiser les contextes d'occurrence de chaque mot.
Ci-dessous on trouve quelques outputs résultant d'un processus d'analyse dans lequel certaines catégories d'un dictionnaire 'classique' pour l'analyse du contenu (Harvard IV-4) ont été appliquées aux discours inauguraux des présidents des États-Unis.

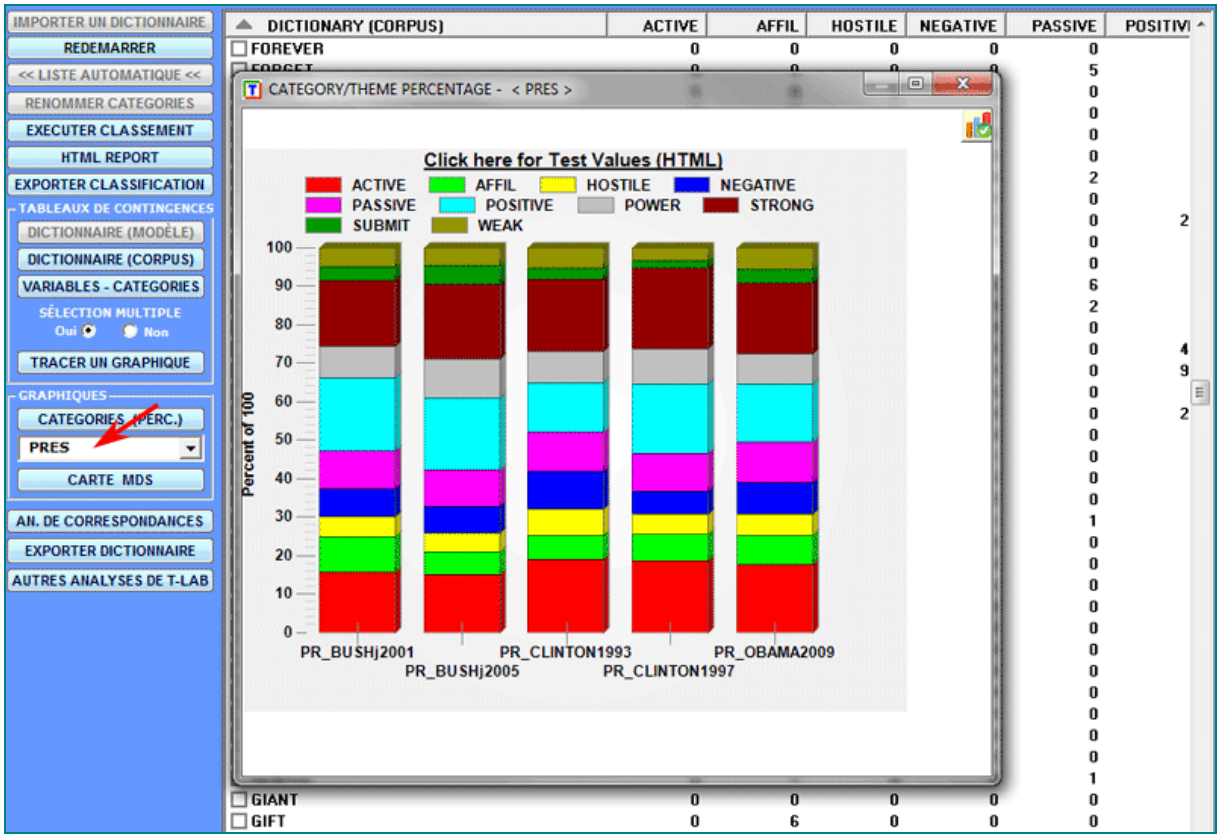




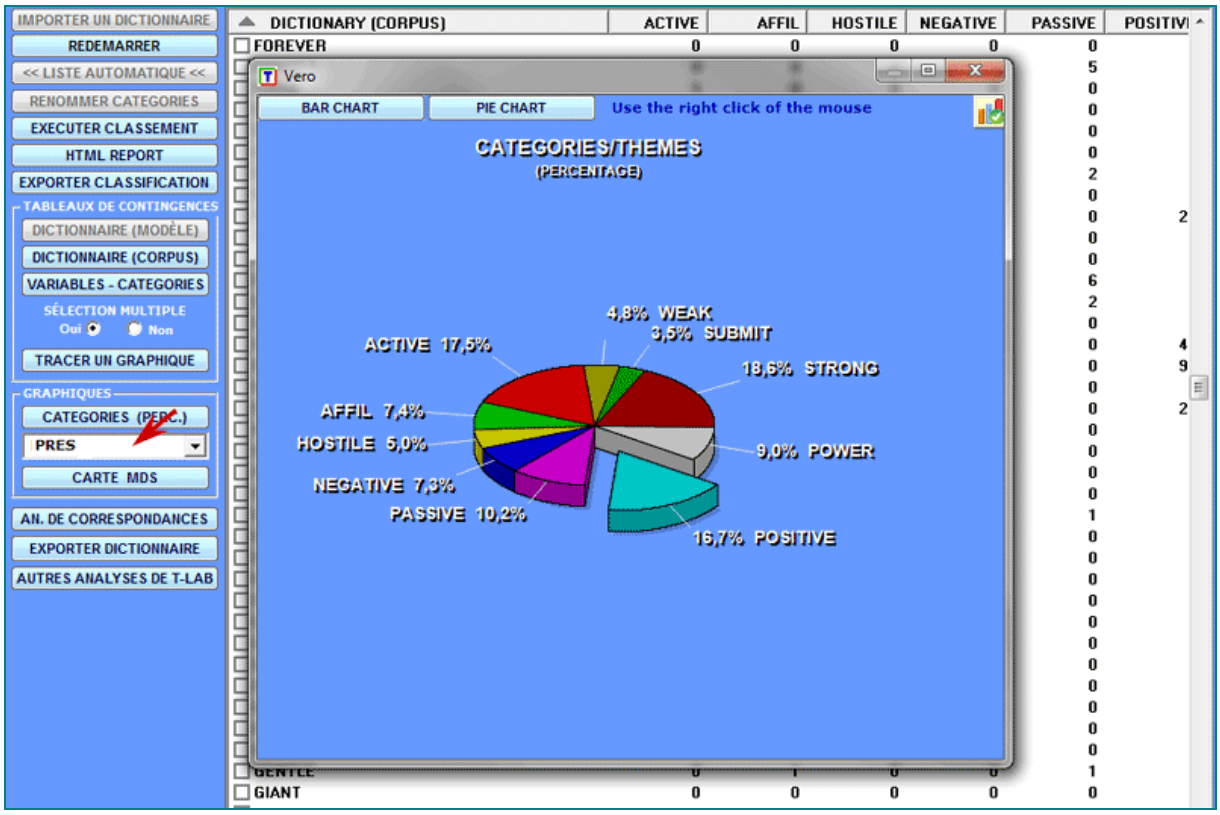
Pour créer des graphiques avec plus d' ensembles de données correspondant à plusieurs lignes des tableaux de contingence, il suffit de choisir '**Sélection multiple**' (option 'Oui'), de sélectionner jusqu'à 20 éléments et de cliquer sur '**Tracer le Graphique**' (voir ci-dessous).



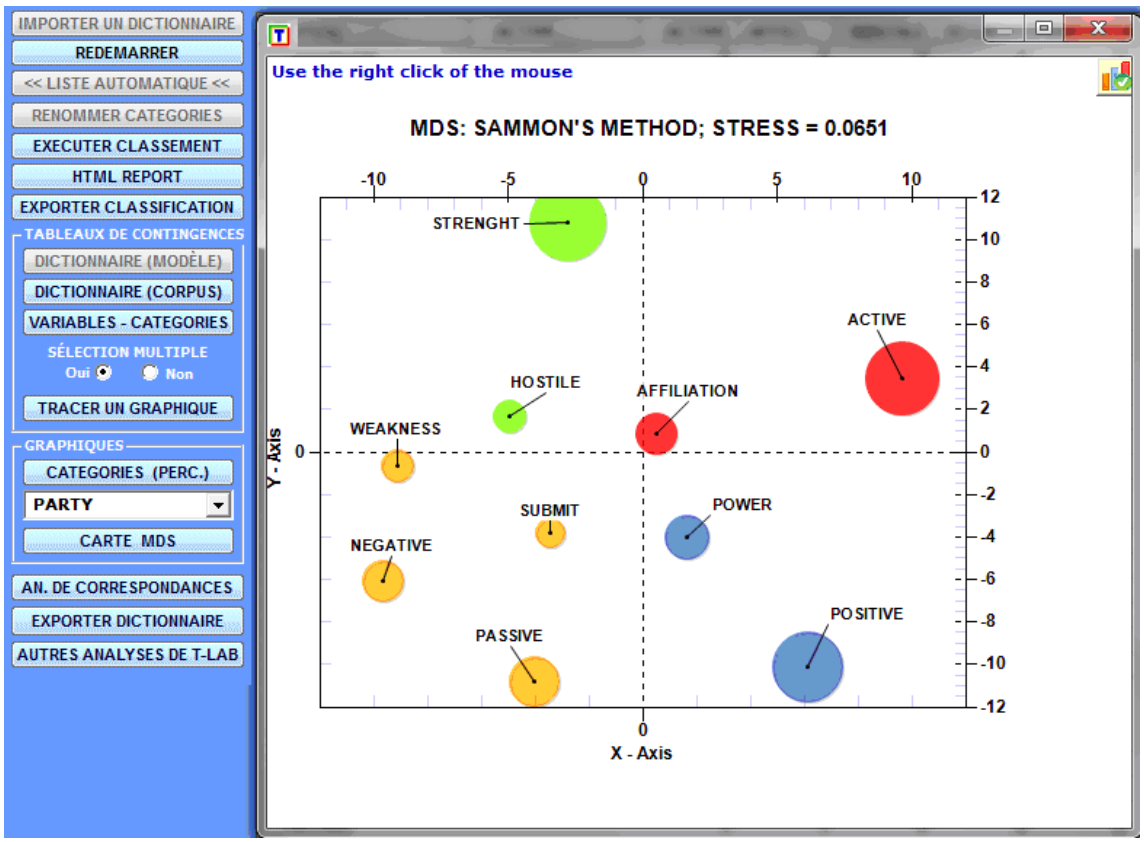
Les deux options ci-dessus sont également disponibles pour les tableaux avec les valeurs des variables.

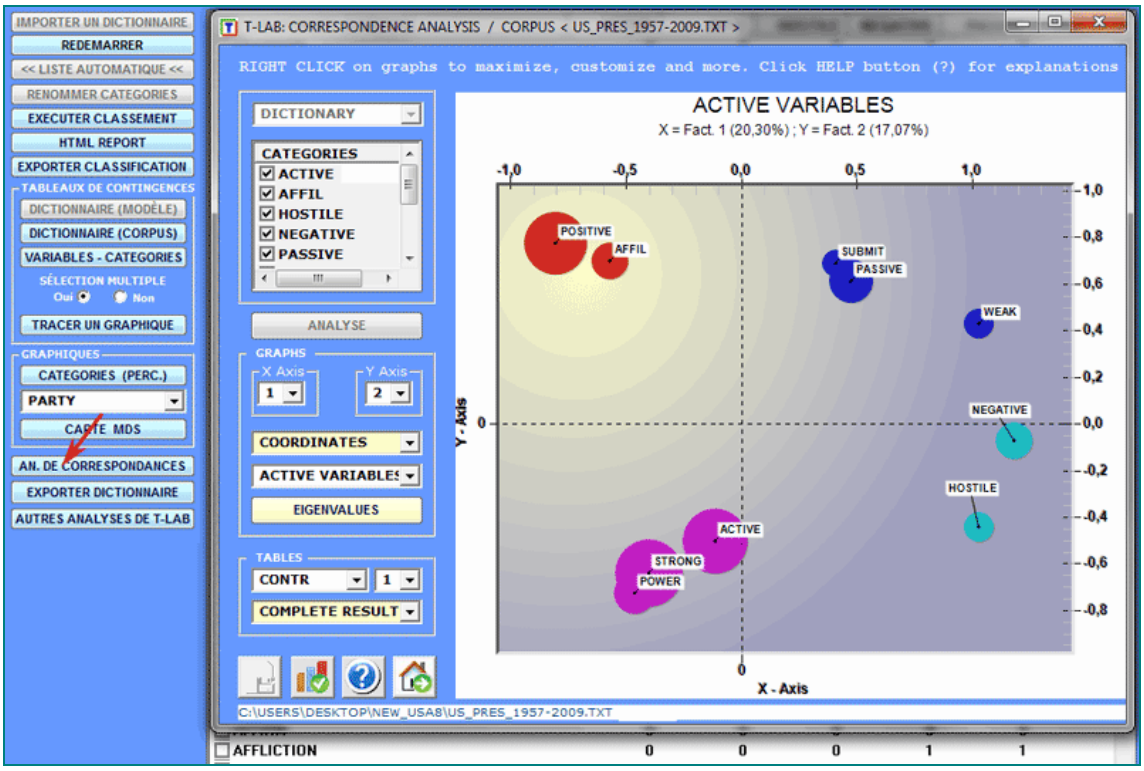


Les pourcentages des catégories peuvent être vérifiés de diverses façons (voir ci-dessous)



Pour explorer la structure totale des données contenues dans les tableaux de contingence, vous pouvez utiliser soit l' option '**MDS**' que l' option '**Analyse des Correspondances**' (voir ci-dessous).





Seulement dans le cas que des unités de contexte ont été classées il est possible d' afficher et d' exporter d' autres outputs avec les données correspondantes ; en outre, en ce cas aussi, il est possible d' enregistrer les résultats de l' analyse dans une nouvelle variable et continuer l' exploration avec d' autres outils du menu **T-LAB**.

En détail, en cliquant sur le bouton '**HTML Report**' vous pouvez voir certains résultats du processus de classification où un score de similarité (cosinus) est attribué à tous les 'contextes élémentaires' ou 'documents' appartenant aux différentes catégories (NB: les images qui suivent sont relatives à un corpus de documents contenant des brèves descriptions de sociétés).

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 ...

Des données similaires peuvent être exportées dans des fichiers XLS (voir ci-dessous) qui contiennent toutes les informations concernant les contextes élémentaires ('Context_Classification.xls') ou bien des documents ('Document_Classification.xls') correctement classés;

(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 , ir
'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 us
'0000500001	SEMICONDUCTOR	0,018	What We Build 3PAR® (NYSE Arca : PAR) is the leading global provider of utility storage , a ca
'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 , de

(2) - Document_Classification.xls

	'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

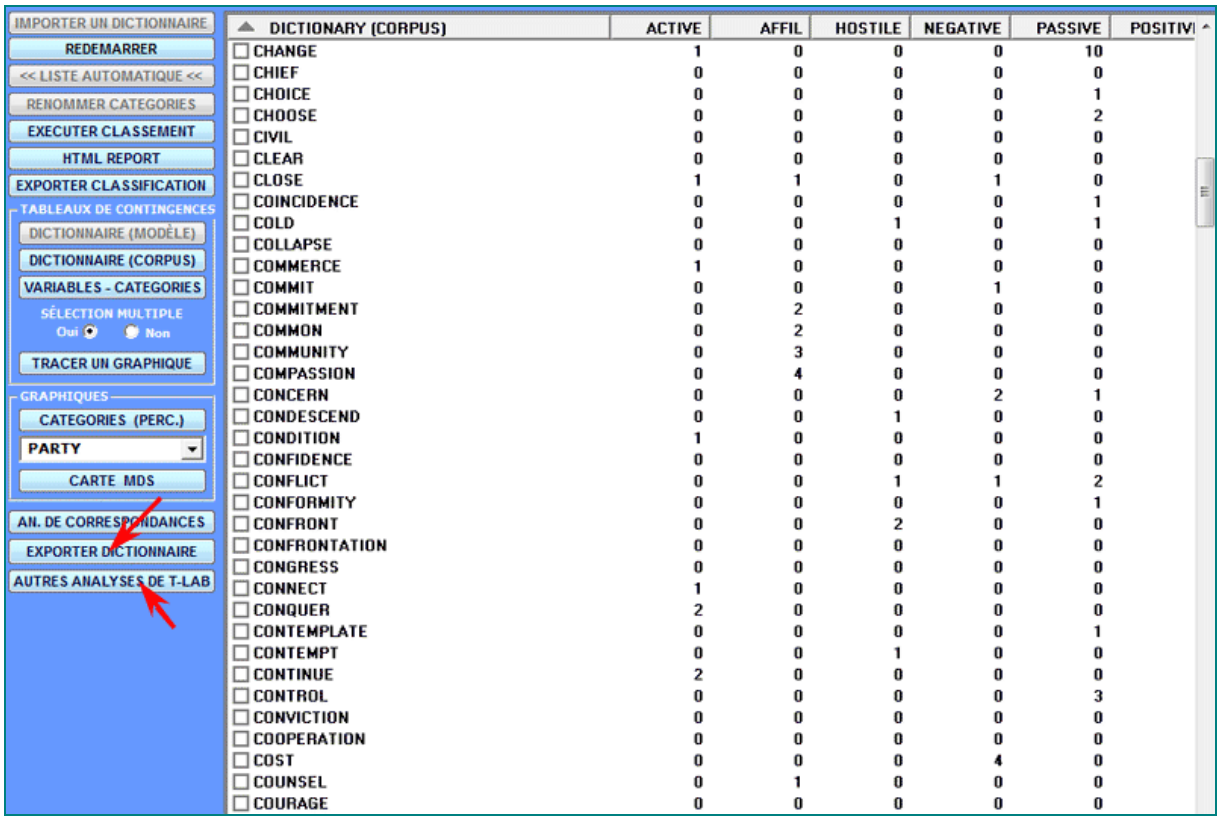
D) - AUTRES PHASES DU PROCESSUS D' ANALYSE

Lorsque le processus de classification a produit ses outputs, deux autres options sont disponibles:

- '**Exporter Dictionnaire**', qui crée un dictionnaire prêt à être importé et utilisé avec d' autres outils T-LAB pour les analyses thématiques ;

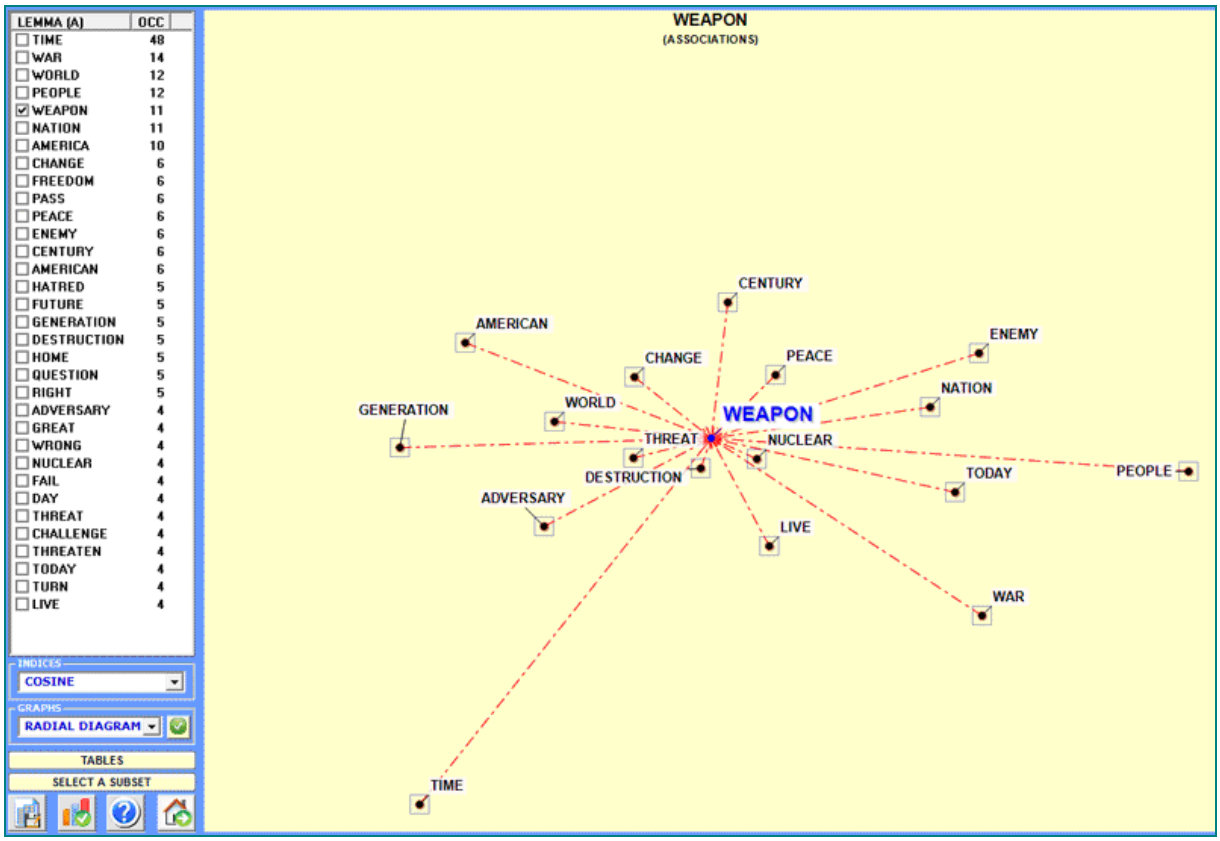
- '**Autres analyses T-LAB**', qui, en fonction de la structure du corpus analysé, du type de classement effectué et du nombre de catégories appliquées, génère une nouvelle variable qui peut être utilisée par d' autres outils **T-LAB** (voir ci-dessous).

.



DICTIONARY (CORPUS)	ACTIVE	AFFIL	HOSTILE	NEGATIVE	PASSIVE	POSITIV
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	

Ci-dessous vous trouvez un exemple obtenu par l'analyse d'un 'sous-ensemble' des contextes classés à l'aide de l' outil **Associations de Mots** (voir le menu principal de **T-LAB**).



E) - FORMAT INPUT/OUTPUT DES DICTIONNAIRES T-LAB

Ci-dessous sont rapportées toutes les informations sur le format des dictionnaires qui peuvent être importés par cet outil de **T-LAB**:

- tous les dictionnaires doivent être des fichiers texte (ASCII / ANSI) avec l' extension 'dictio' (ex. Mycategories.dictio) ;
- tous les dictionnaires créés par des outils **T-LAB** pour les analyses thématiques, y compris ceux créés par l'outil '**Classification Basée sur des Dictionnaires**', sont prêts à être importés sans autres interventions de la part de l' utilisateur ;
- d' autres dictionnaires, aussi bien 'standard' que personnalisés , doivent être produits en suivant les indications rapportées ci-dessous :

1 - chaque dictionnaire se compose de 'n' lignes et ne peut pas dépasser la limite de 100.000 records ;

2 - chaque ligne du dictionnaire comprend deux ou trois 'chaînes' séparées par un point-virgule (ex. : économique;crédit) ;

3 - pour chaque ligne, la première chaîne doit être une 'catégorie', la seconde un 'mot' (ou lemme), la troisième - si présente - doit être un nombre réel positif (c' est-à-dire un numéro entier) de '1' à '999' qui représente le 'poids' de chaque mot dans la catégorie correspondante ;

4 - la longueur maximale d'une chaîne (mot, lemme ou catégorie) est de 50 caractères et ne doit pas contenir ni espaces vides ni apostrophes ;

5 - lorsque le dictionnaire contient des multi-words (ex. Gouvernement Fédéral), les espaces vides doivent être remplacés par le caractère '_ ' (ex. Gouvernement _Fédéral);

6 - dans chaque dictionnaire, le numéro des catégories utilisées peut varier entre un minimum de deux à un maximum de 50. Lorsque le nombre de catégories est supérieur à 50 il est recommandé d'utiliser un dictionnaire en un format différent et de l' importer en utilisant l'outil **Personnalisation du Dictionnaire**. Dans ce cas, on vous rappelle que chaque mot doit être en correspondance univoque avec une (seule) catégorie.

De suite vous trouvez deux extraits de fichiers .dictio, respectivement avec deux et trois chaînes par ligne:

a) cas avec deux chaînes (c' est-à-dire 'paires' de catégories et de mots)

```
...
négatif;catastrophique
négatif;nuisible
...
positif;fantastique
positif;satisfait
...
```

b) cas avec trois chaînes (c'est-à-dire des catégories, des mots et des numéros)

```
...
négatif;catastrophique;10
négatif;nuisible;8
...
positif;fantastique;9
positif;satisfait;7
```

Textes et Discours comme Systèmes Dynamiques

N.B .: Cette section est uniquement disponible en anglais.

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.

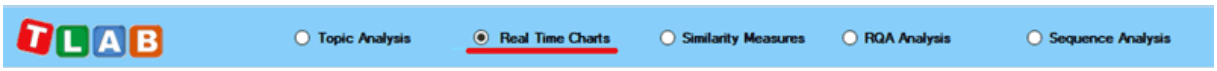
(**) 'Topic' and 'Theme' will be hereafter treated as synonymous terms.

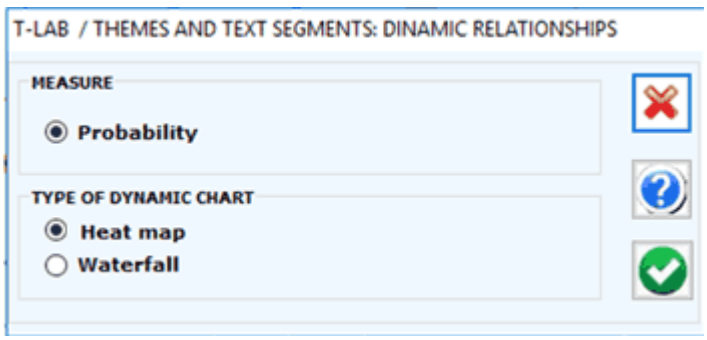
Please note that the main goal of the above automatic steps is to extract 'k' latent dimensions (where 'k' varies from 20 to 30) which determine the content structure of the analysed text and which – like a mixture model - can be used for exploring both text dynamics and similarities between text segments. 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.

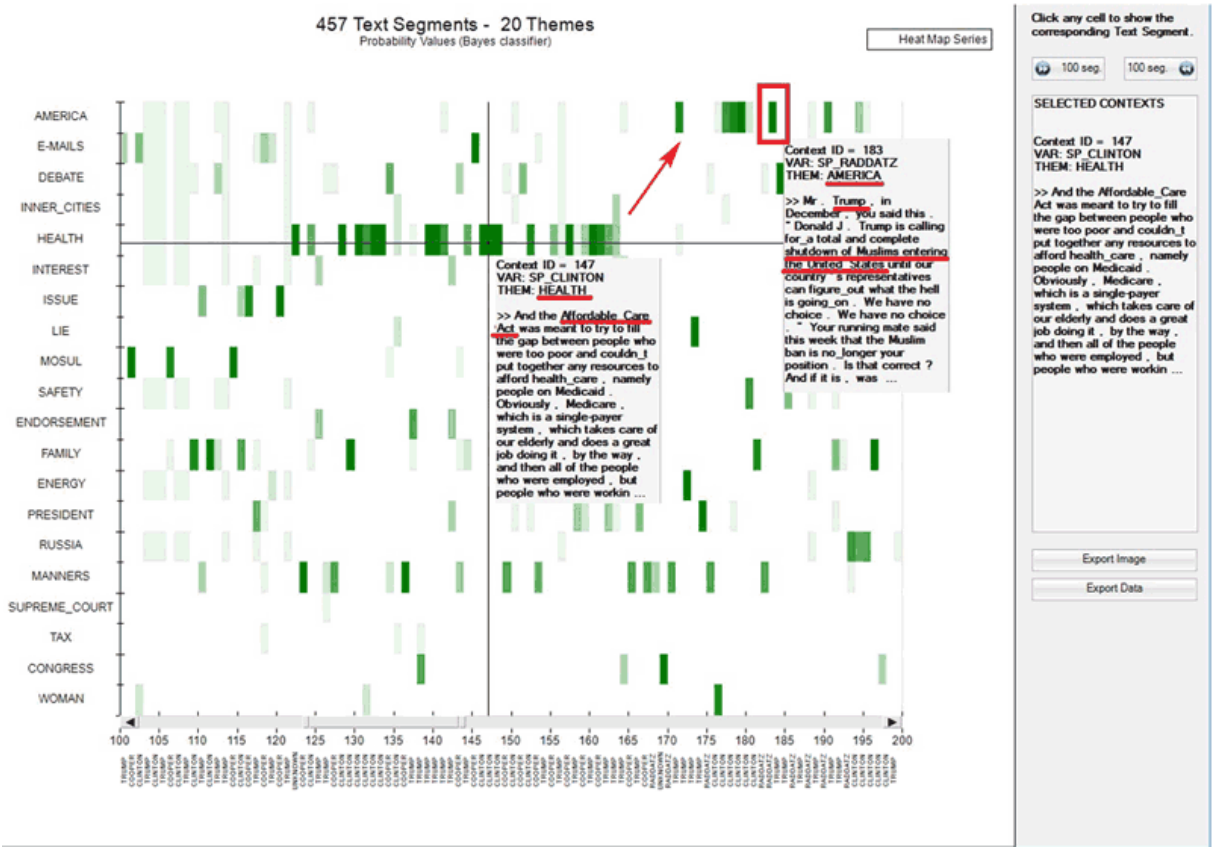
Two complementary charts allows us to easily appreciate various types of events, including the **strong recurrences** of some themes or the **shifts** from a theme to another (see the below pictures, obtained by analysing a presidential debate between Hillary Clinton and Donald Trump which took place 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.

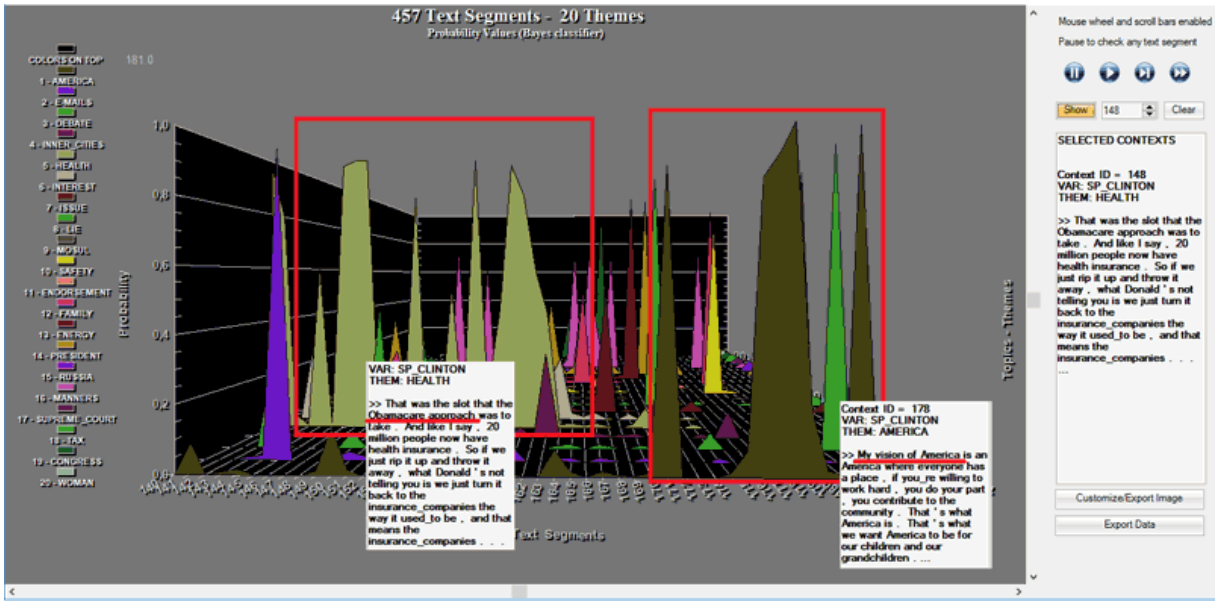
As the information summarized by these types of charts mainly refers to formal aspects of text contents, the same charts may be regarded as some sort of musical scores where the sequence of themes and their 'intensity' (i.e. probability) vary in time.

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

A.1 - Heat map



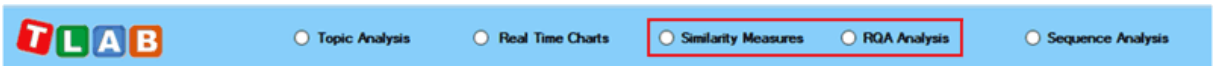
A.2 - Waterfall



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.

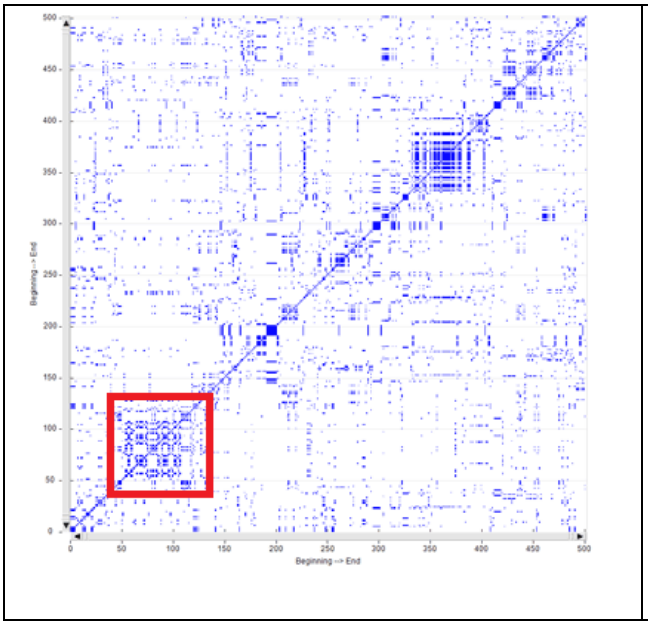


Fig. 1 - Time series

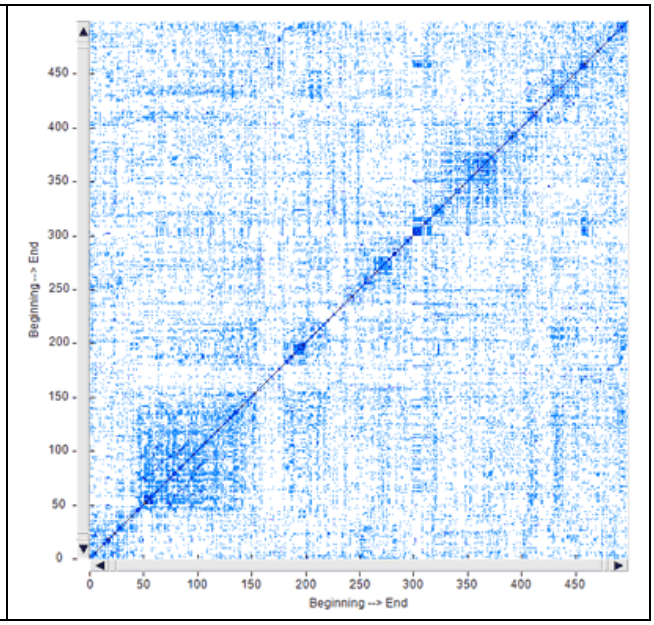
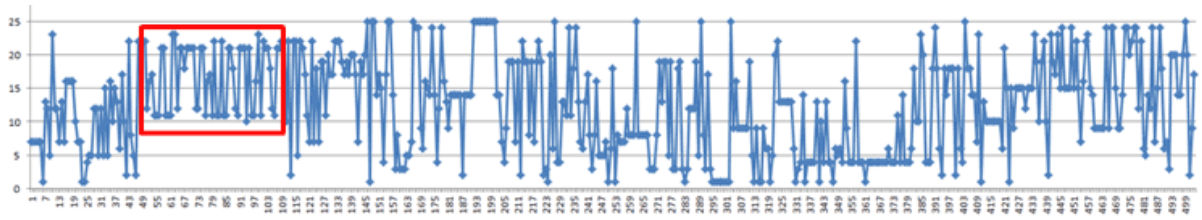
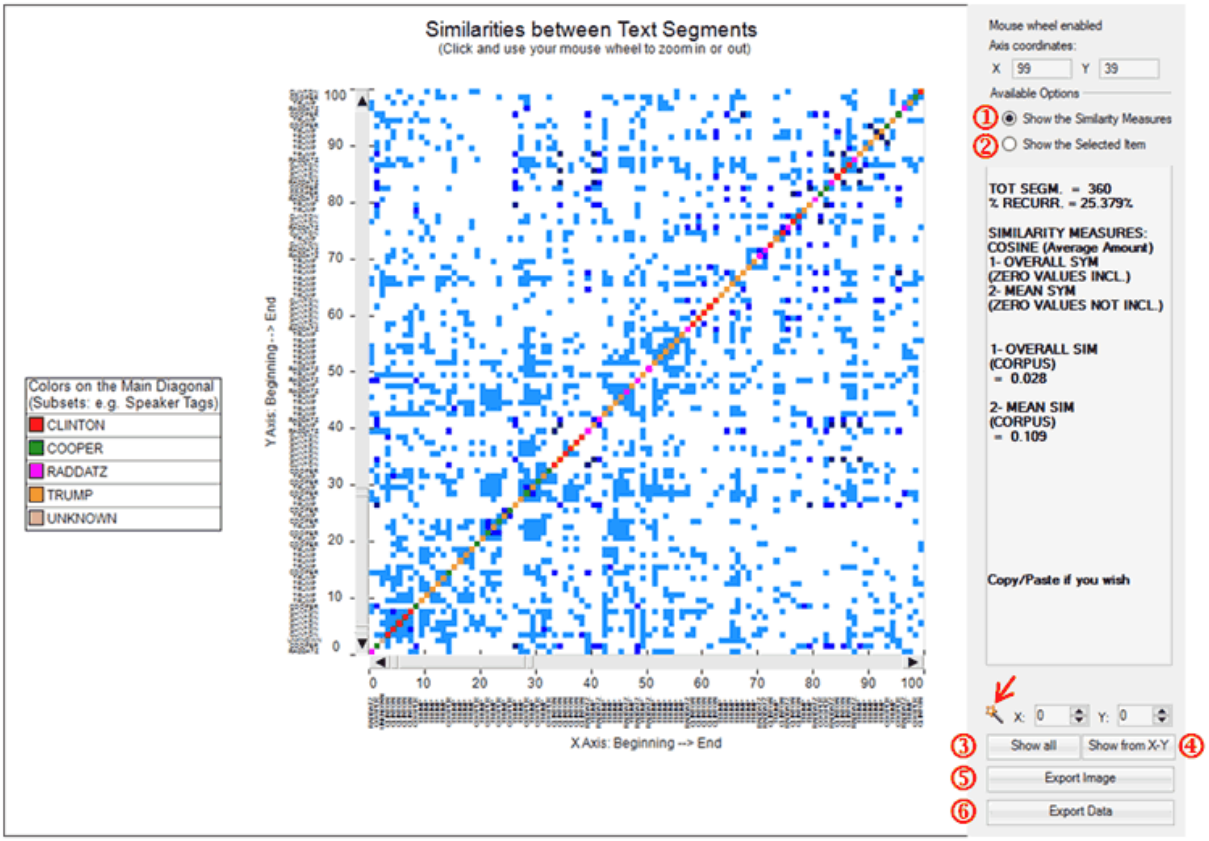


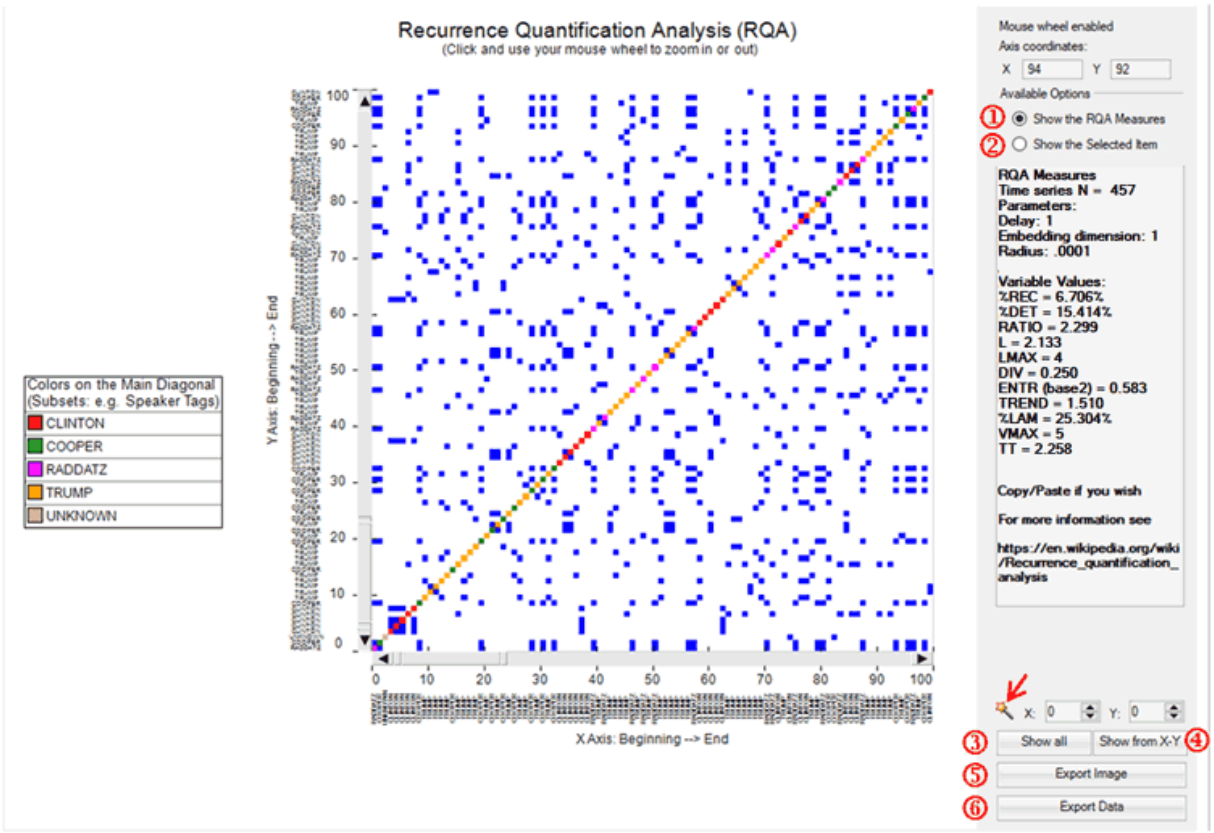
Fig2 - Similarities

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).



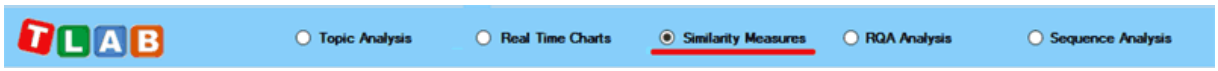


In particular:
-options '1' and '2' allow us to visualize the general measures ('1') or the transcript of the selected segment ('2');
-options '3' and '4' allow us to visualize the complete recurrence plot ('3') or a subsection of it ('4');
-options '5' and '6' allow us to export the image in different 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).

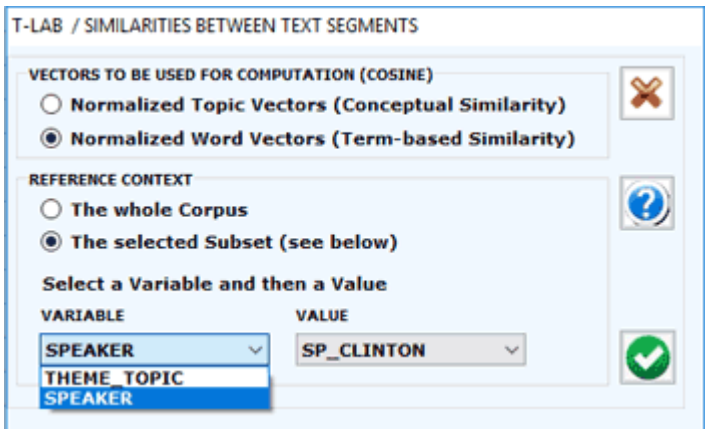
FIRST	SECOND	Cosine
SP_CLINTON	SP_CLINTON	0.0961
SP_CLINTON	SP_COOPER	0.1099
SP_CLINTON	SP_RADDATZ	0.1025
SP_CLINTON	SP_TRUMP	0.0847
SP_CLINTON	SP_UNKNOWN	0.1087
SP_COOPER	SP_CLINTON	0.1099
SP_COOPER	SP_COOPER	0.3106
SP_COOPER	SP_RADDATZ	0.2359
SP_COOPER	SP_TRUMP	0.1432
SP_COOPER	SP_UNKNOWN	0.1446
SP_RADDATZ	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

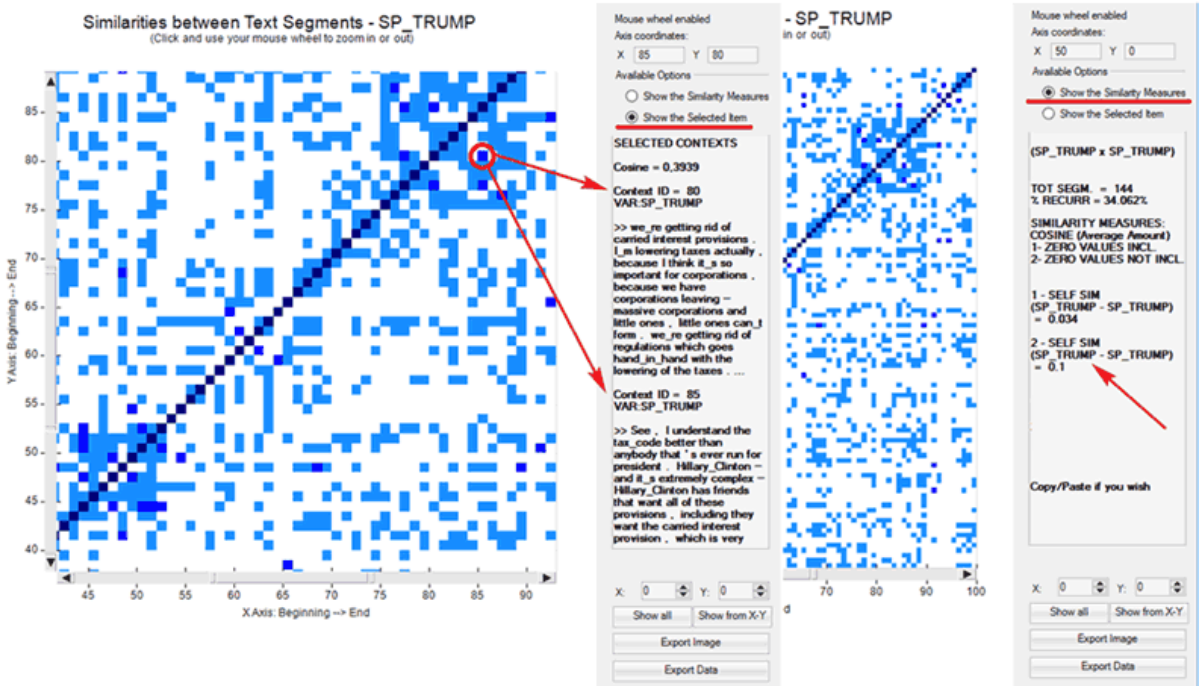


When choosing 'Similarity Measures', several options are made available (see picture below) which allow the user to select both the vectors to be used for the similarity computation and the reference context to be analysed (i.e. either the entire corpus or a subset of it).

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.

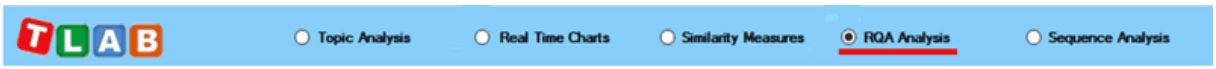


According to the design of the user interface, in this case - like in the RQA analysis (see section 'D' below) - the user can choose between visualizing the global measures or the transcripts of recurrent segments (see picture below). Moreover, when a corpus subset is selected, two further measures are 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).

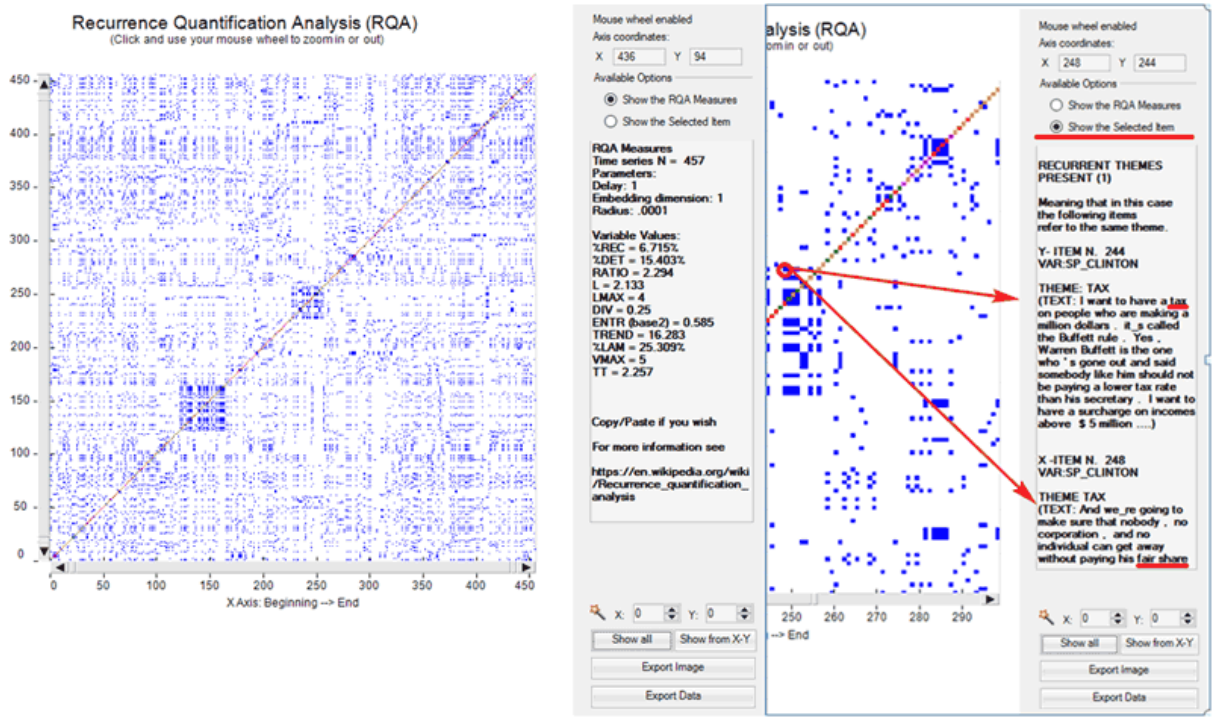
In this **T-LAB** tool, both in the case of the RQA Analysis and in the case of the Sequence Analysis (i.e. Markovian Analysis), a time series is represented by a categorical vector where each element is an integer which corresponds to the topic assigned to the 'i' text segment. However only in the case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

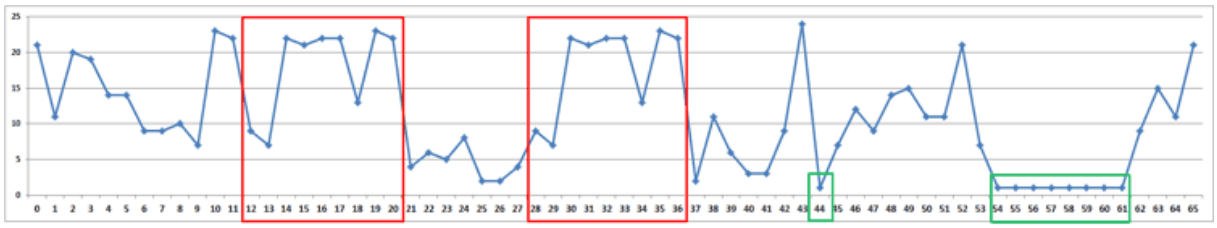
(**) For more information about the RQA measures see section 'E' below.



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.

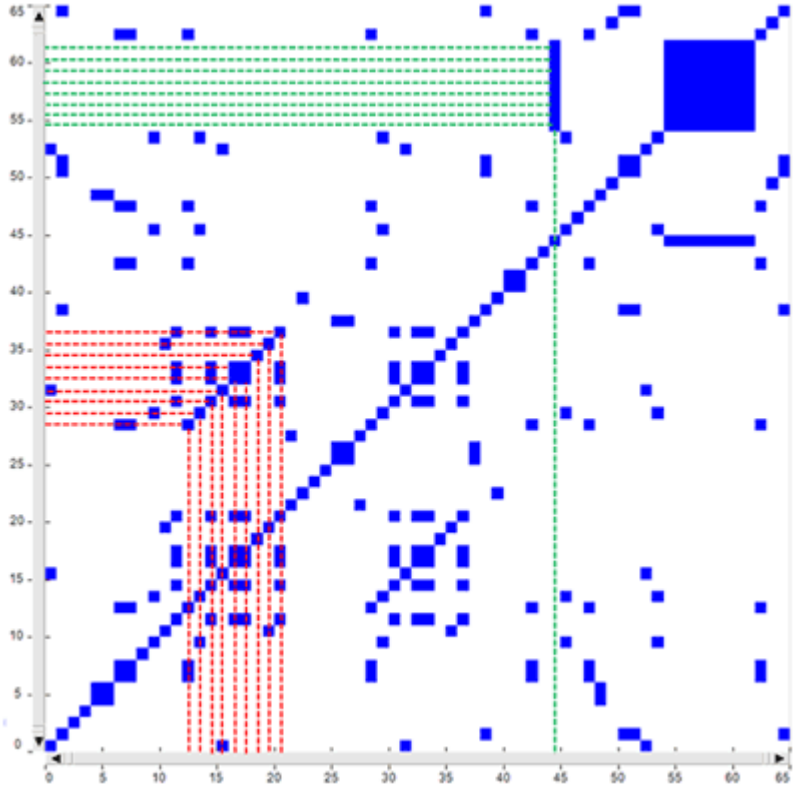
In particular, the distribution of diagonal lines allows for the investigation of **determinism** (i.e. the predictability of the system) and the distribution of vertical lines allows for the investigation of **intermittency** (i.e. the sequences which are interspersed by erratic breaks).



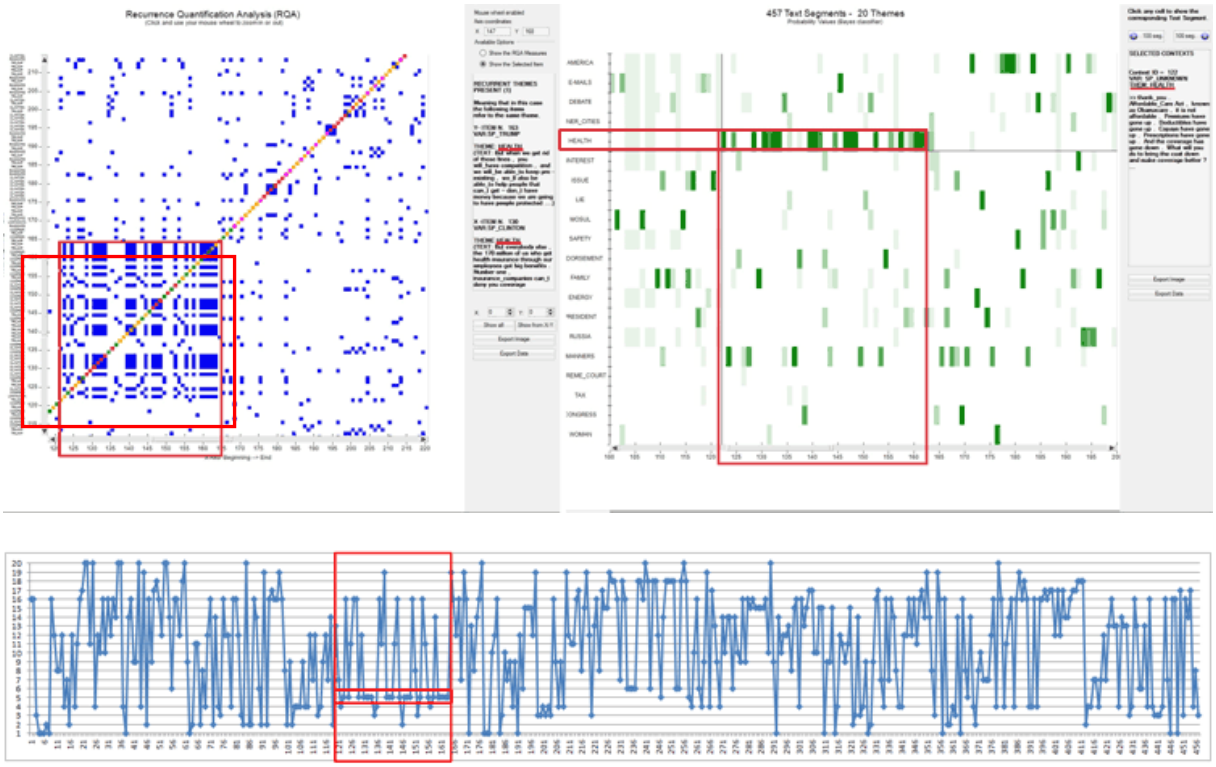
As an example, just consider the above fictitious time series. In it the same sequence of nine points/themes is repeated two times in different time spans (see the above red rectangles), respectively from t-12 to t-20 and from t-28 to t-36, where each 't' stands for a different text segment. In the same 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.

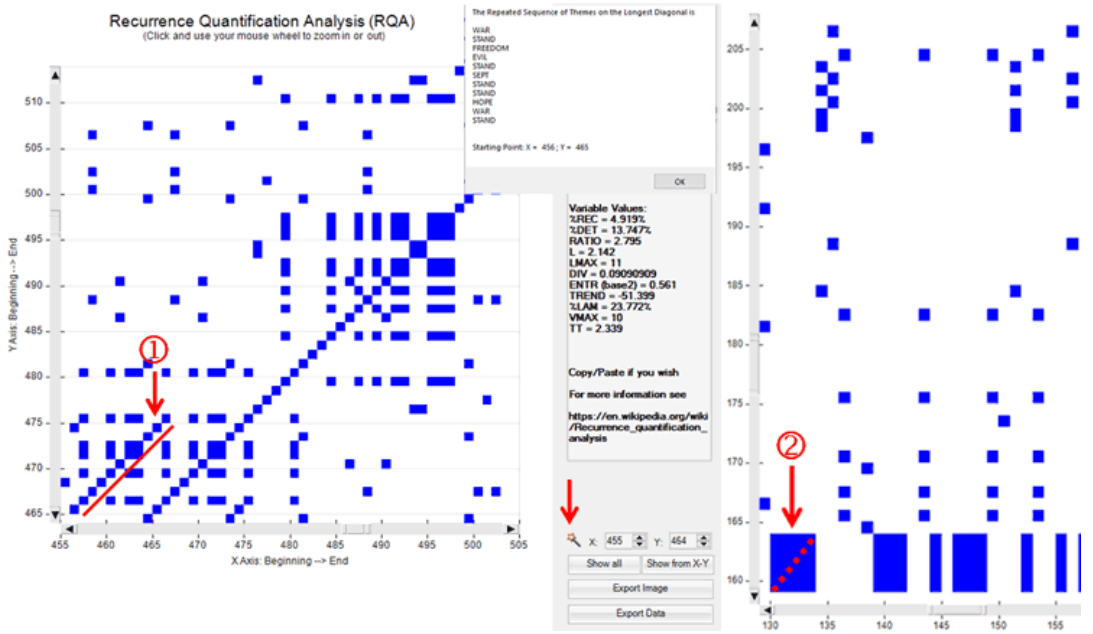


Regarding the **rectangular block** structures – which actually include both vertical and diagonal lines - they can be seen as referring to recurrences of the same topics in sub sections of the time series, i.e. to groups of overall similar feature vectors. In fact each dot in the graph represents a 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'.

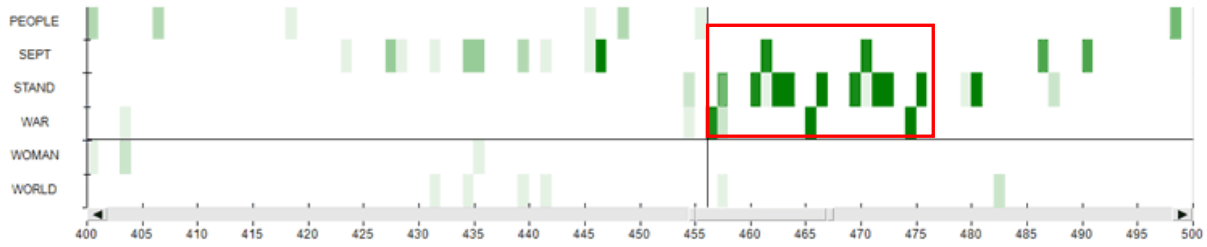


As stated above, in the RQA outputs the **longest diagonals** parallel to the main diagonal allow us to detect interesting repetitions of the same thematic sequence. However their shapes are not so evident as the rectangular block structures, also because sometimes they can be hidden inside one of them (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.

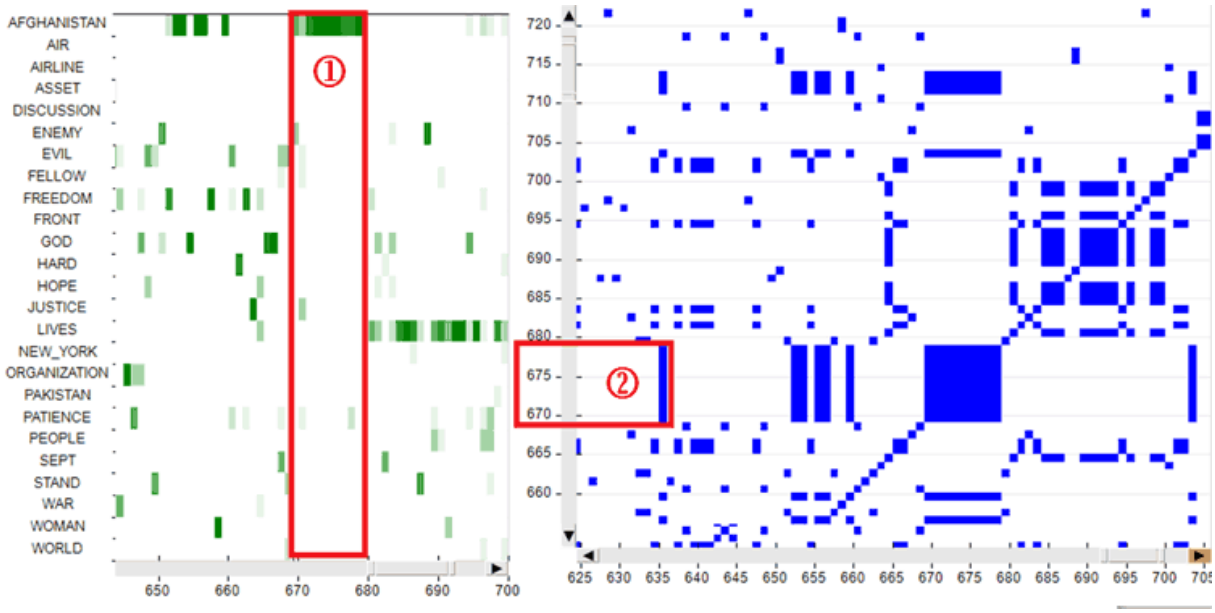
N.B.: Soon after the longest diagonal is detected **T-LAB** allows the user to export a file with the most frequent **repeated sequences**, each one of them including at least three concatenated themes. Such a file can be considered a sort of summary of the main themes - and of the corresponding variations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

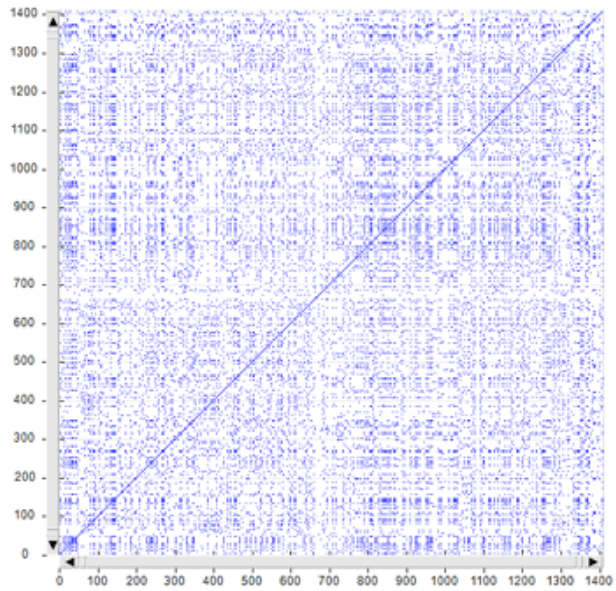
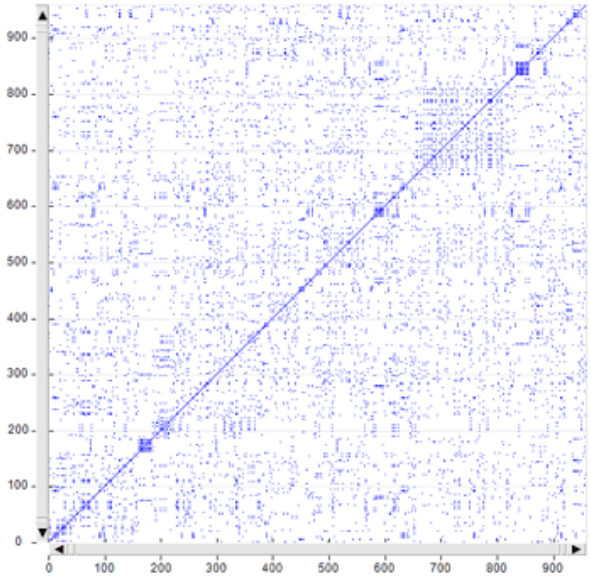
Measure	**Definition**
%REC - Recurrence Rate	The percentage of recurrence points in a Recurrence Plot which fall within a specified radius.
%DET - Determinism	The percentage of recurrence points which form diagonal line structures, main diagonal not included (N.B.: In RQA the main diagonal is also called 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

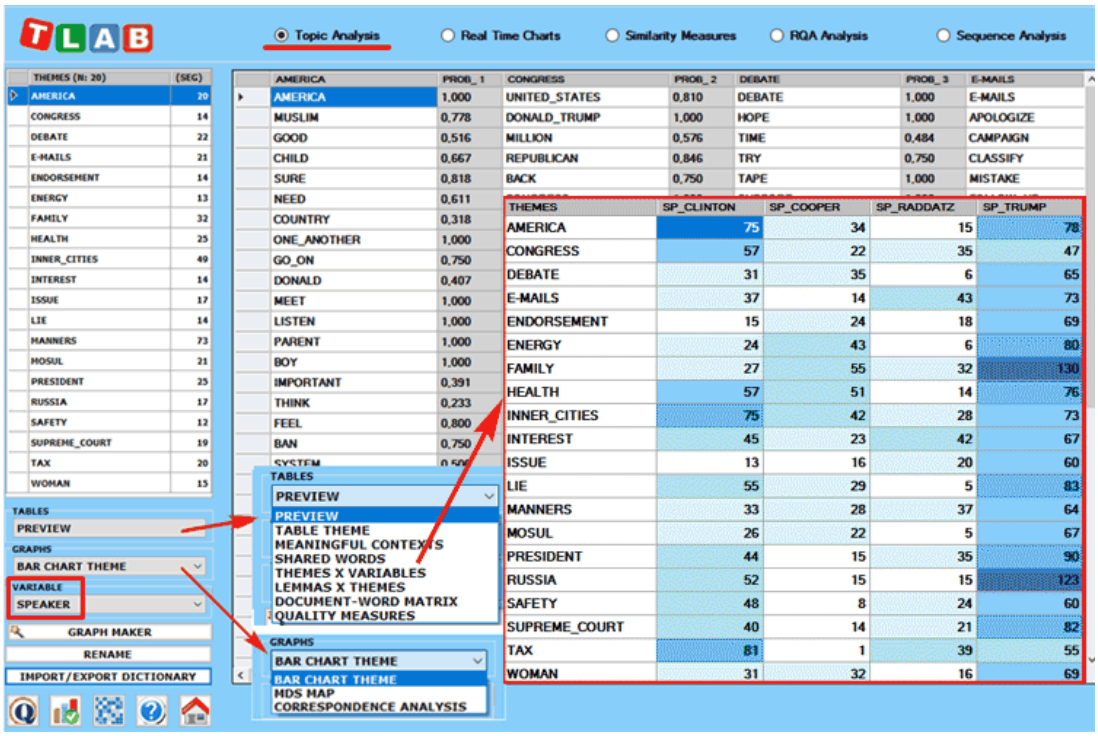
Here are the two corresponding recurrence plots.

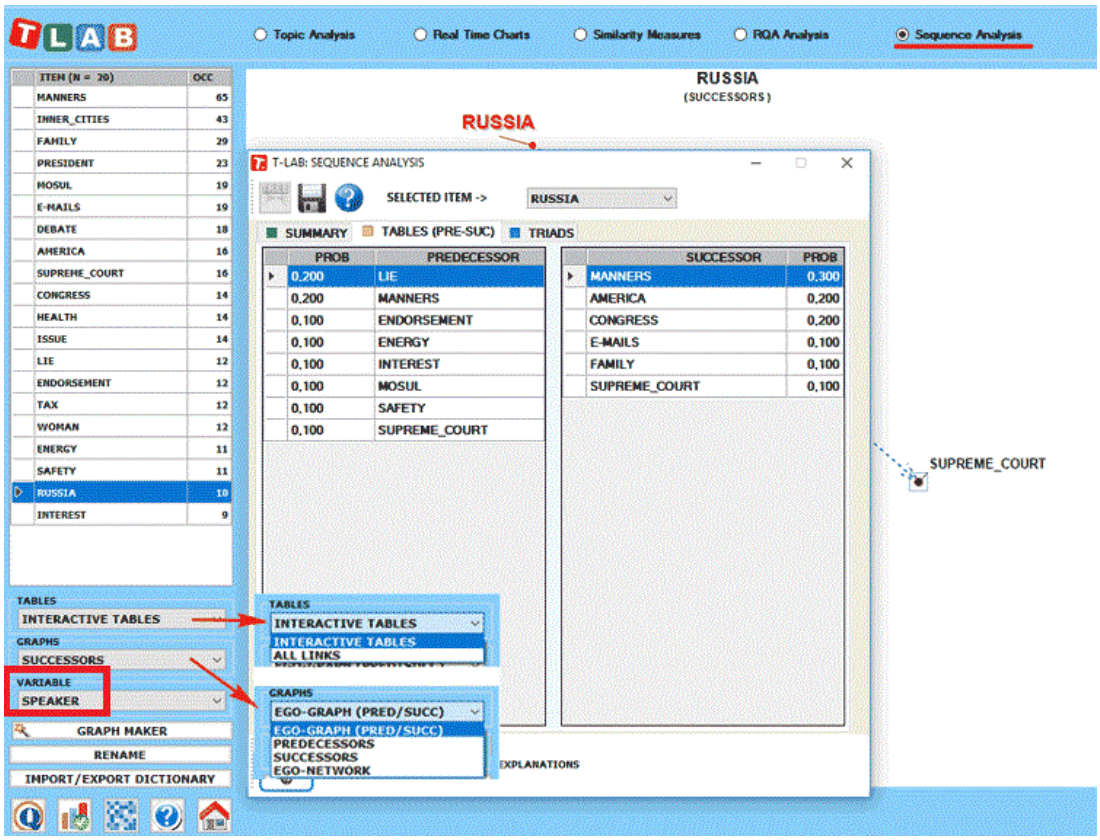
On the Origin of Species (a)	The adventures of Pinocchio (b)
	

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

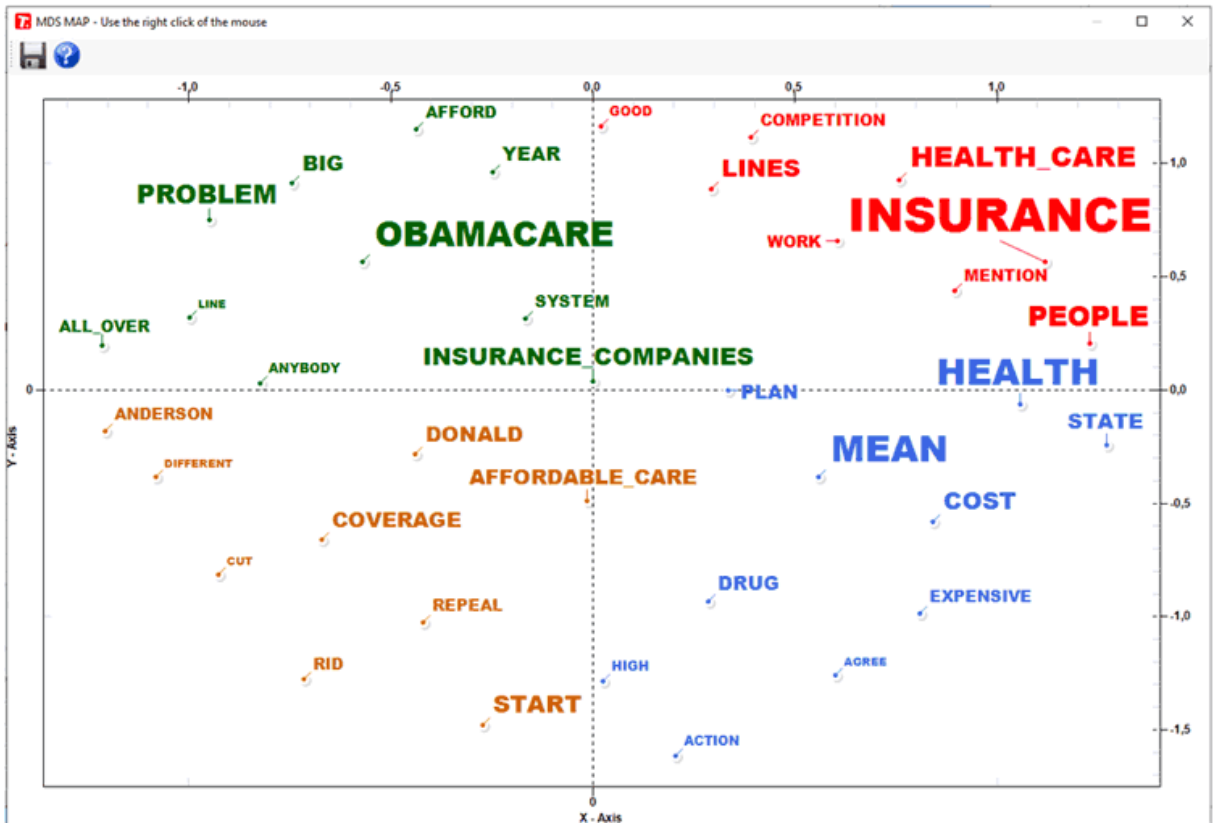
The below pictures summarize the main options of two tools already present in the **T-LAB** menu, which are integrated with the new ones and which are explained in the corresponding sections of this manual/help, i.e. 'Modeling of Emerging Themes' and 'Sequence and Network Analysis'.





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).

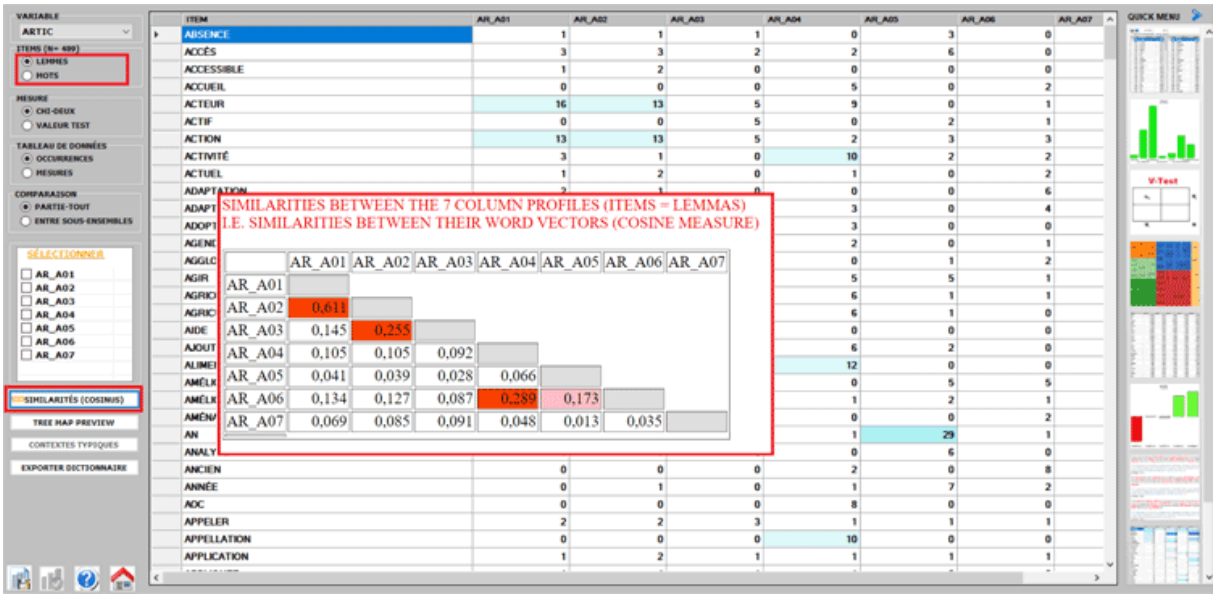

Topic Analysis
Real Time Charts
Similarity Measures
RQA Analysis
Sequence Analysis
GRAPH MAKER - CO-OCCURRENCES WITHIN THE CLUSTER N. <HEALTH >
ADD/REMOVE ITEMS TO BE USED
CLICK A PICTURE TO DISPLAY THE GRAPH
MDS
EXPORT DATA FILES FOR NETWORK ANALYSIS (up to 5.000 items can be selected at a time)

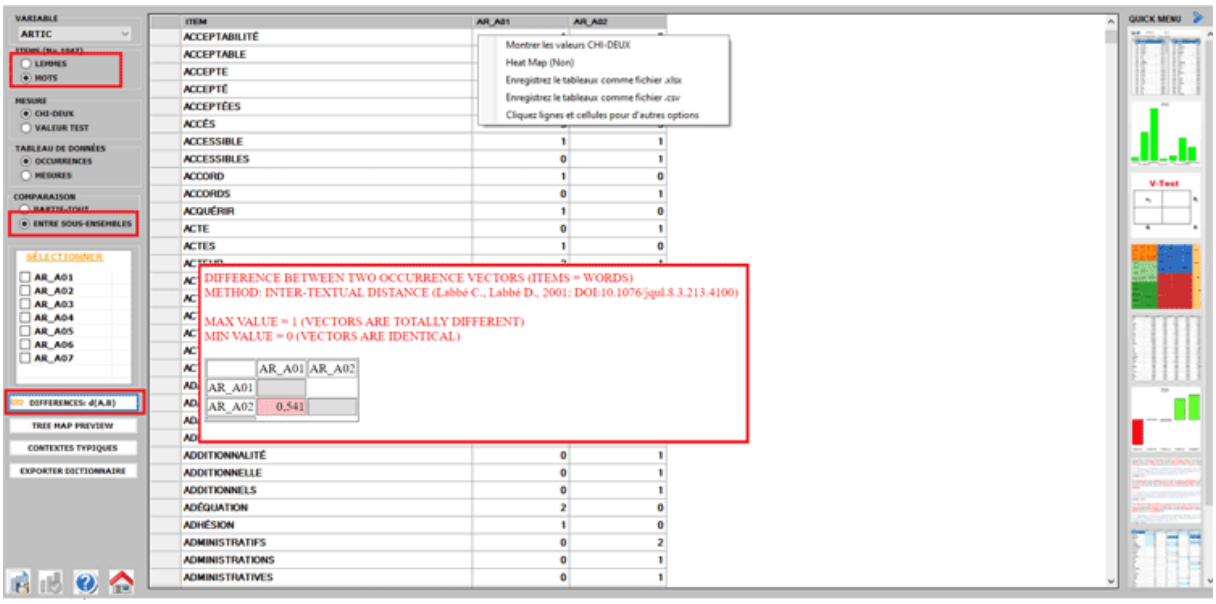

MDS MAP - Use the right click of the mouse
AFFORD
GOOD
COMPETITION
BIG
YEAR
LINES
HEALTH_CARE
PROBLEM
INSURANCE
OBAMACARE
WORK
MENTION
LINE
SYSTEM
PEOPLE
ALL_OVER
INSURANCE_COMPANIES
HEALTH
ANYBODY
PLAN
ANDERSON
STATE
DONALD
MEAN
DIFFERENT
AFFORDABLE_CARE
COST
COVERAGE
CUT
DRUG
EXPENSIVE
REPEAL
RID
HIGH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

ANALYSES COMPARATIVES

Analyse des Spécificités

N.B.: Les images de cette section font référence à une version précédente de T-LAB. En **T-LAB 10**, l'aspect est légèrement différent. En particulier, une galerie d'images à accès rapide qui fonctionne comme un menu supplémentaire permet de basculer entre les différentes sorties en un seul clic De plus l'utilisateur est autorisé à évaluer facilement les similitudes (ex. -Distance textuelle) entre les sous-ensembles du corpus (de 2 à 150), et donc aussi pour détecter les documents quasi-dupliqués et quasi-dupliqués (voir les images ci-dessous).





Cet outil **T-LAB** permet de vérifier quelles **unités lexicales** (c'est-à-dire mots, lemmes ou catégories) sont **typiques** ou **exclusives** dans un texte ou un **sous-ensemble du corpus** défini par une variable catégorielle; en outre il permet aussi d'identifier les **unités de contexte caractéristiques** des différents sous-ensembles en examen (par exemple les phrases "typiques" qui mieux différencient les discours des divers leaders politiques).

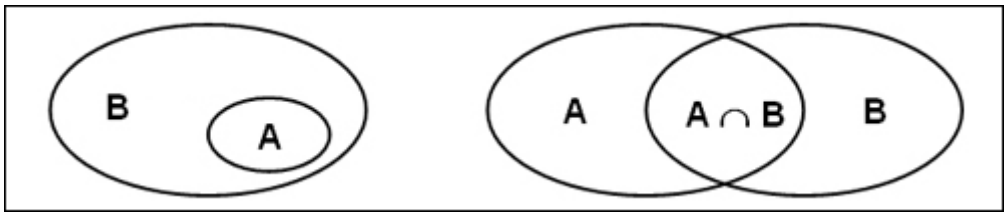
Les **unités lexicales typiques**, définies par la proportion des occurrences respectives (c'est-à-dire par leur sur / sous- utilisation), sont déterminées par le calcul **Chi-Carré** ou par la **Valeur Test**.

Les **contextes élémentaires caractéristiques** sont identifiés en calculant et en additionnant les valeurs **TF-IDF** normalisées assignées aux mots dont chaque phrase ou chaque paragraphe est constitué.

L'analyse de spécificités nous permet d'effectuer deux types de comparaisons:

1- entre une **partie** (ex. le sous-ensemble “A”) et le **tout** (ex. le corpus entier “B”);

2- entre des couples de **sous-ensembles** (“A” e “B”).



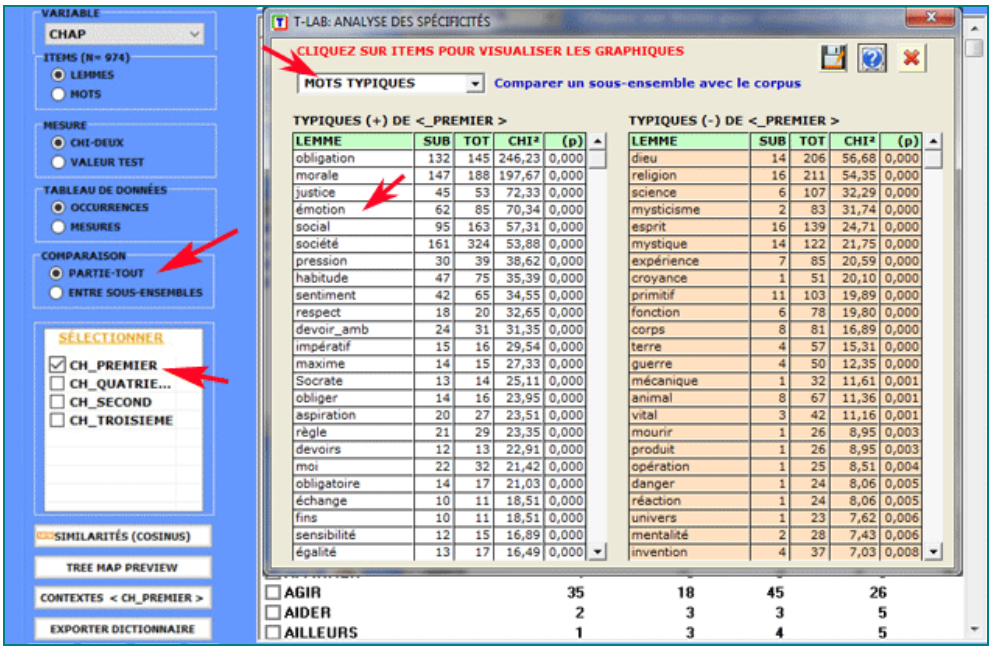
Dans chacun des cas on peut aussi bien analyser les Spécificités relatives aux **intersections** que celles relatives aux **différences**.

Les modalités du calcul utilisé sont montrées dans l'entrée correspondante du **glossaire**.

Les unités lexicales considérées peuvent être toutes (configuration automatique) ou seulement celles choisies par l'utilisateur (configuration personnalisée).

En succession, les quatre types de comparaisons possibles sont les suivantes:

1.1 - **partie /tout**: unités lexicales " typiques"



LEMME	SUB	TOT	CHI²	(p)
obligation	132	145	246,23	0,000
morale	147	188	197,67	0,000
justice	45	53	72,33	0,000
émotion	62	85	70,34	0,000
social	95	163	57,31	0,000
société	161	324	53,88	0,000
pression	30	39	38,62	0,000
habitude	47	75	35,39	0,000
sentiment	42	65	34,55	0,000
respect	18	20	32,65	0,000
devoir_amb	24	31	31,35	0,000
impératif	15	16	29,54	0,000
maxime	14	15	27,33	0,000
Socrate	13	14	25,11	0,000
obliger	14	16	23,95	0,000
aspiration	20	27	23,51	0,000
règle	21	29	23,35	0,000
devoirs	12	13	22,91	0,000
moi	22	32	21,42	0,000
obligatoire	14	17	21,03	0,000
échange	10	11	18,51	0,000
fins	10	11	18,51	0,000
sensibilité	12	15	16,89	0,000
égalité	13	17	16,49	0,000

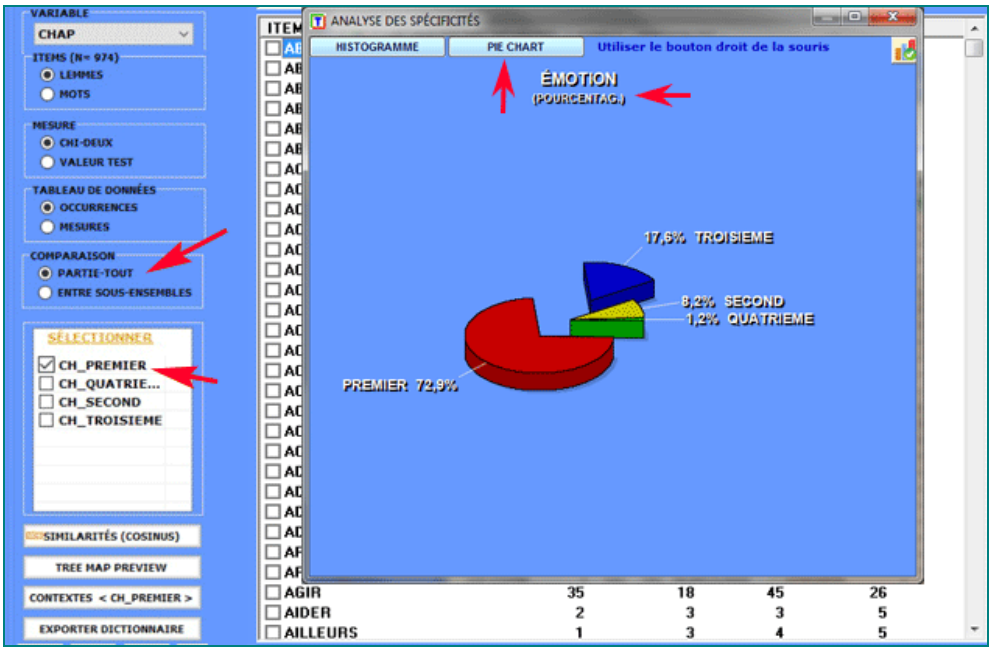
LEMME	SUB	TOT	CHI²	(p)
dieu	14	206	56,68	0,000
religion	16	211	54,35	0,000
science	6	107	32,29	0,000
mysticisme	2	83	31,74	0,000
esprit	16	139	24,71	0,000
mystique	14	122	21,75	0,000
expérience	7	85	20,59	0,000
croyance	1	51	20,10	0,000
primitif	11	103	19,89	0,000
fonction	6	78	19,80	0,000
corps	8	81	16,89	0,000
terre	4	57	15,31	0,000
guerre	4	50	12,35	0,000
mécanique	1	32	11,61	0,001
animal	8	67	11,36	0,001
vital	3	42	11,16	0,001
mourir	1	26	8,95	0,003
produit	1	26	8,95	0,003
opération	1	25	8,51	0,004
danger	1	24	8,06	0,005
réaction	1	24	8,06	0,005
univers	1	23	7,62	0,006
mentalité	2	28	7,43	0,006
invention	4	37	7,03	0,008

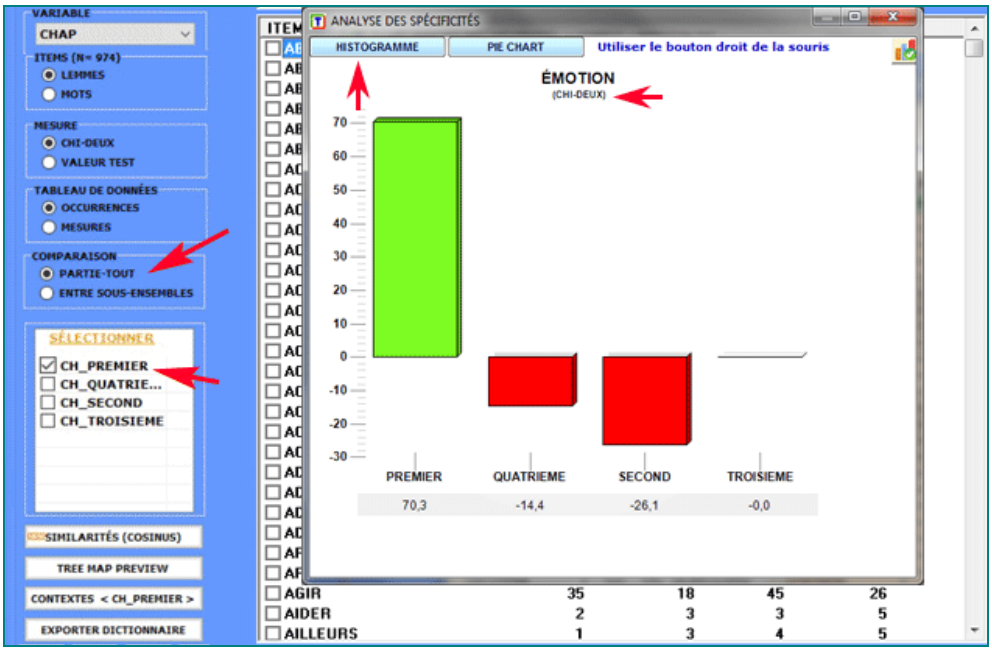
AGIR	35	18	45	26
AIDER	2	3	3	5
AILLEURS	1	3	4	5

Colonne par colonne les clés de lecture du tableau sont les suivantes:

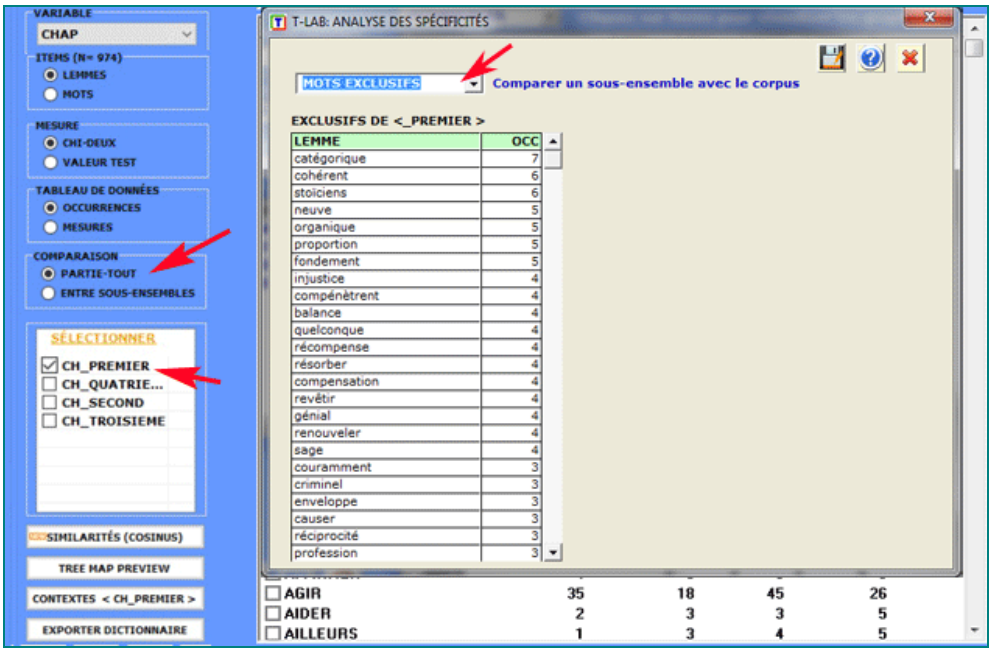
- LEMME = unités lexicales "spécifiques" (sur-utilisées ou sous-utilisées);
- SUB = occurrences de chaque LEMME dans le sous-ensemble examiné;
- TOT = occurrences de chaque LEMME dans le corpus ou dans les deux sous-ensembles examinés (voir 2.1). ;
- CHI2 = valeur du CHI deux (ou VTEST = Valeur Test) ;
- (p) = probabilité associée à la valeur du chi-deux (def=1).

En cliquant sur les éléments des tableaux, il est possible de créer différents types des graphiques.

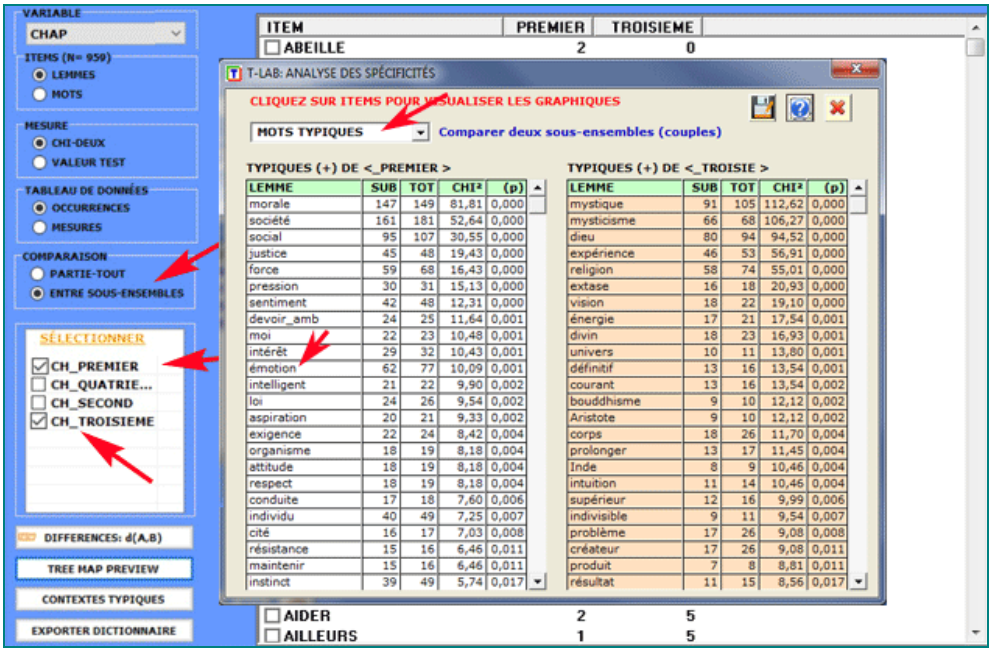


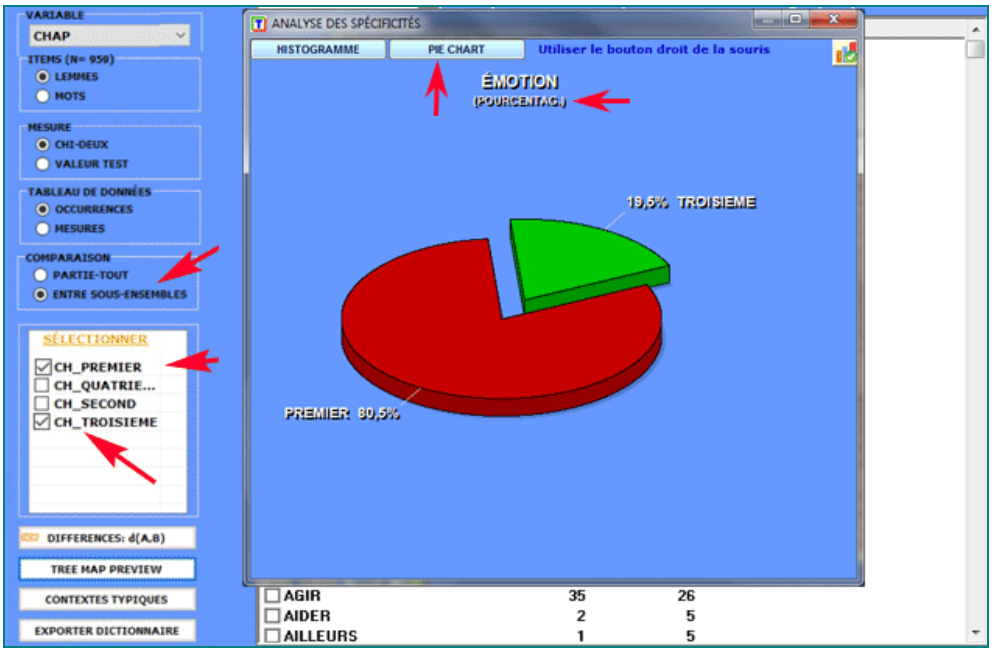


1.2 - **partie/tout**: unités lexicales "exclusives"

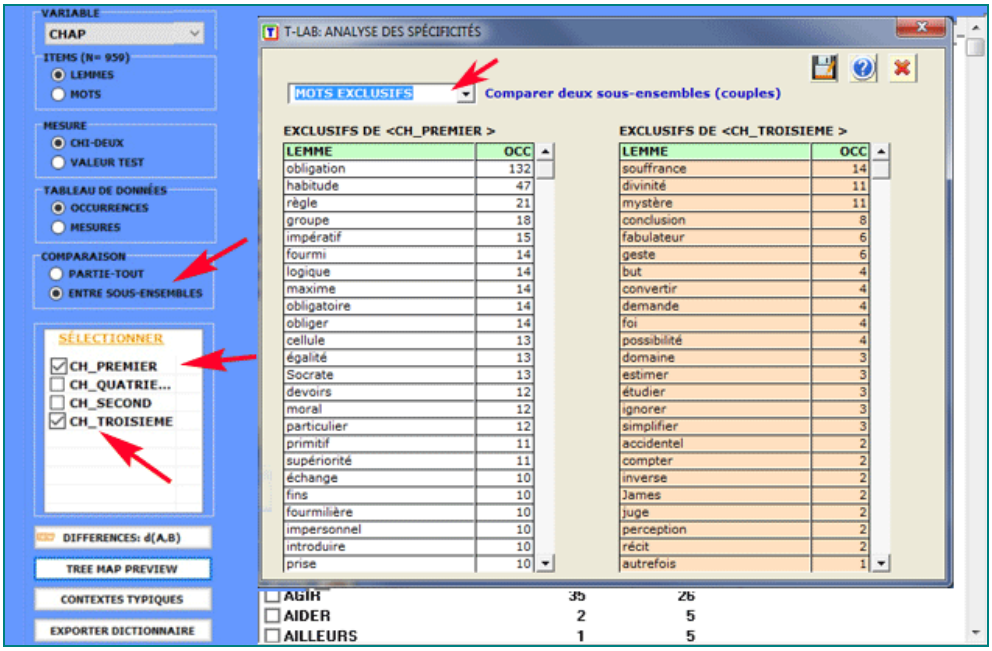


2.1 - **sous-ensemble/sous-ensemble**: unités lexicales "typiques"

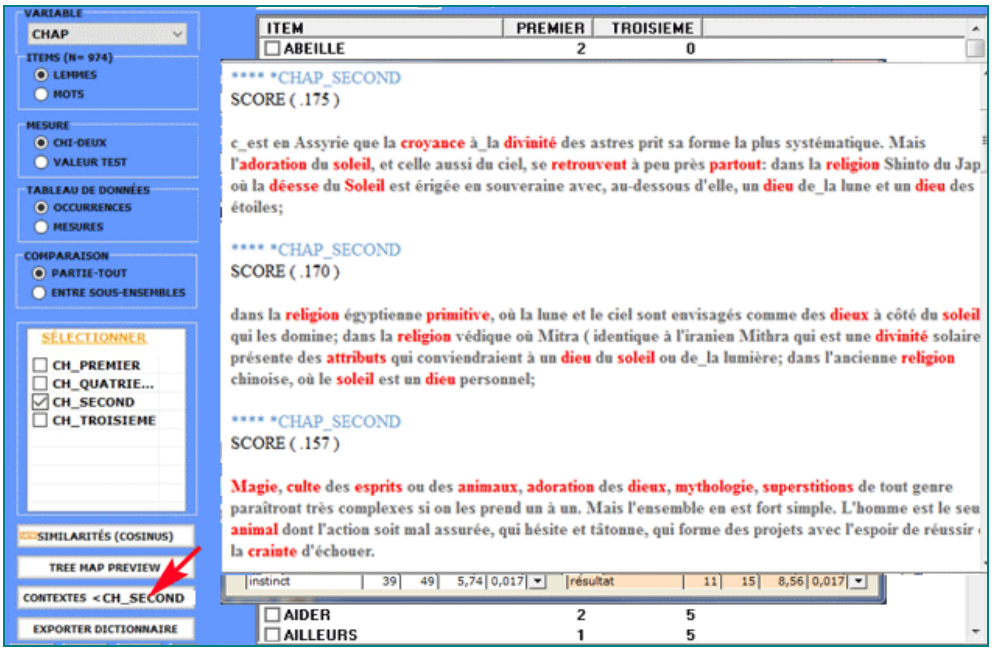




2.2 - **sous-ensemble/sous-ensemble**: unités lexicales "exclusives"

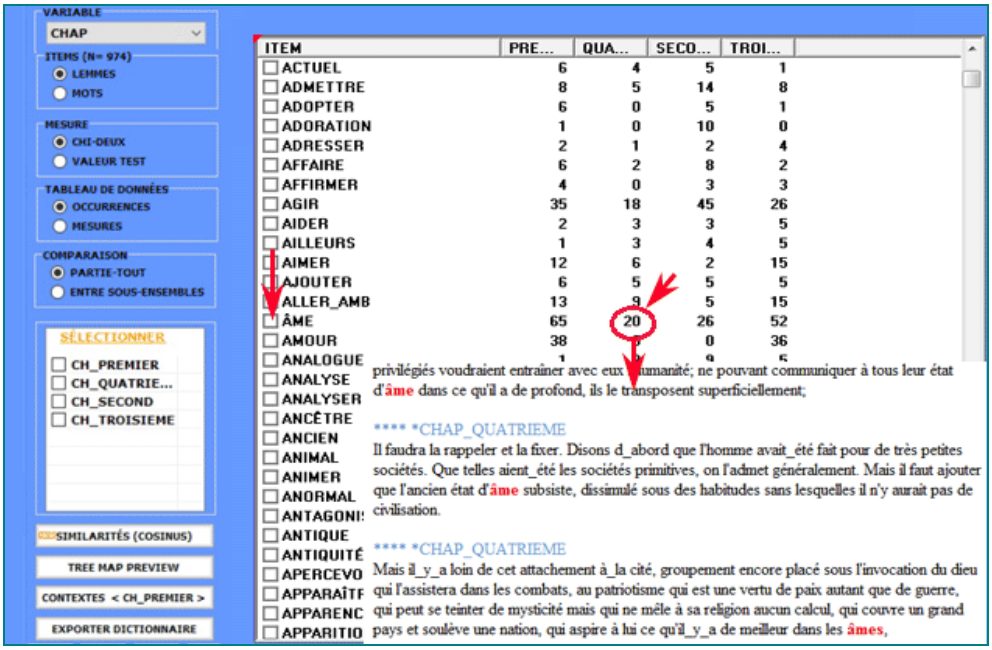


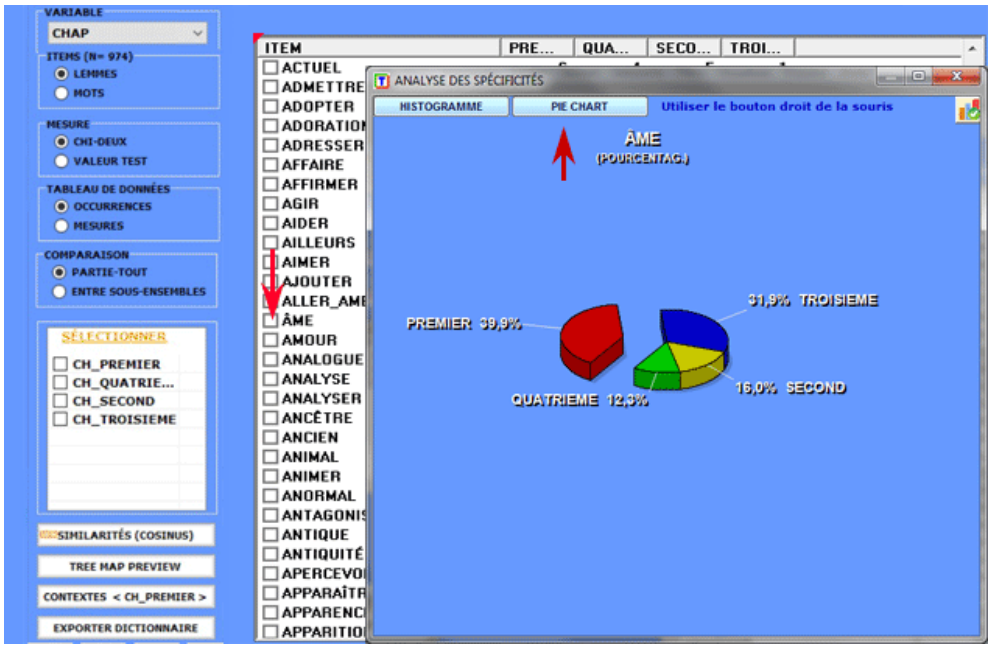
Pour chaque sous-ensemble analysé il est aussi possible de vérifier les contextes élémentaires (c'est-à-dire phrases ou paragraphes) qui mieux le distinguent des autres. Dans ce cas, la **spécificité** résulte du calcul de valeurs **TF-IDF** normalisées; plus particulièrement, le "score" attribué à chaque contexte élémentaire (voir l'image ci-dessous) est le résultat de la somme des valeurs TF-IDF assignées aux mots qui le composent.

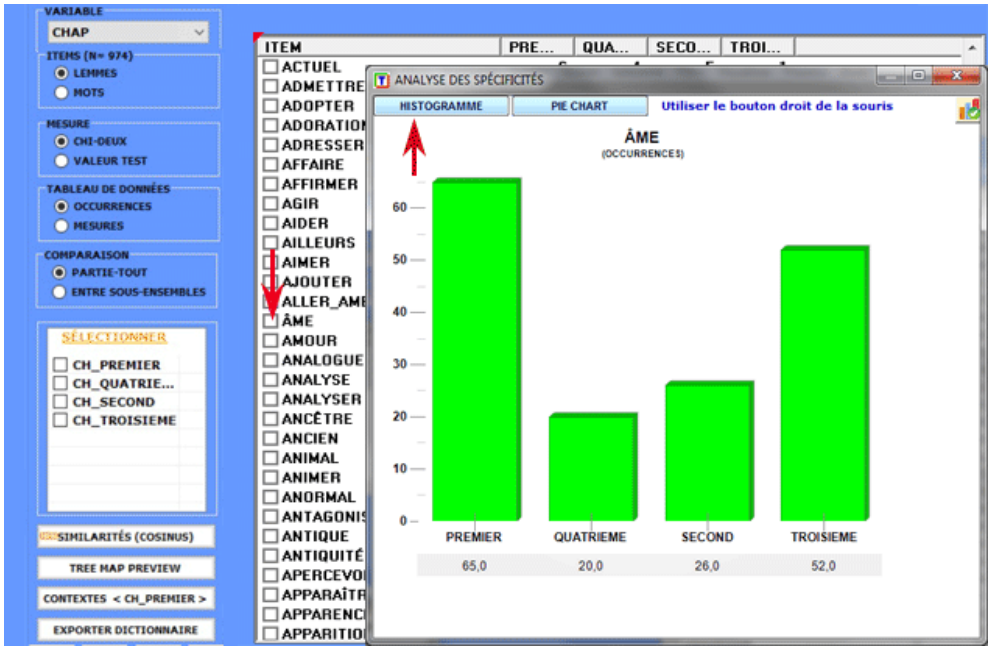


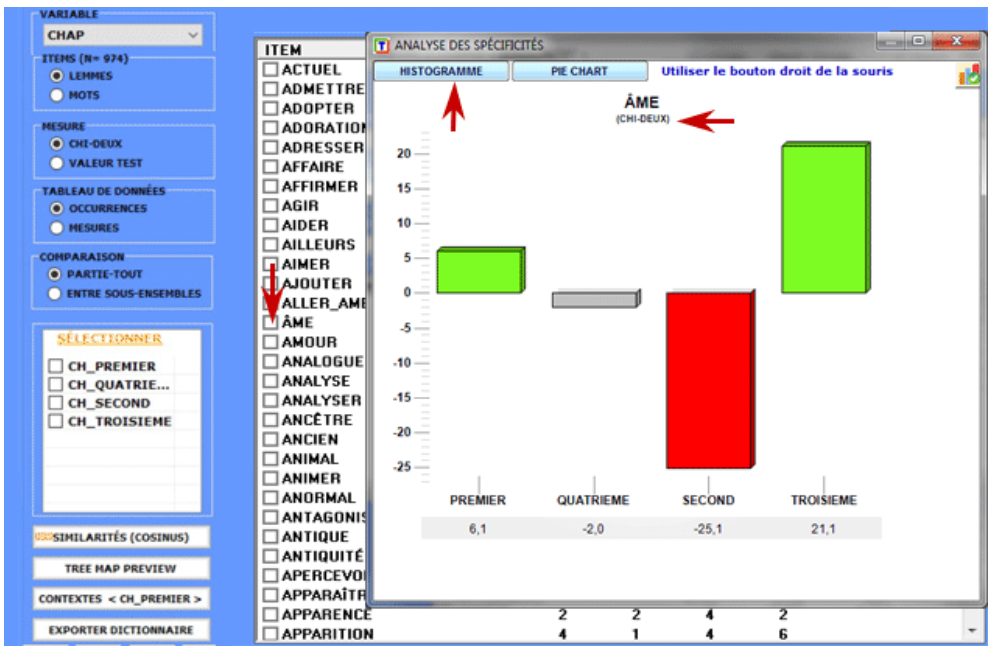
Tous les tableaux de contingence peuvent être facilement explorés et nous permettent de créer différents types des graphiques. De plus, en cliquant sur cellules spécifiques du tableau (voir

ci-dessous), il est possible de créer un fichier HTML montrant tous les contextes élémentaires où le mot en ligne est présent dans le sous-ensemble correspondant.

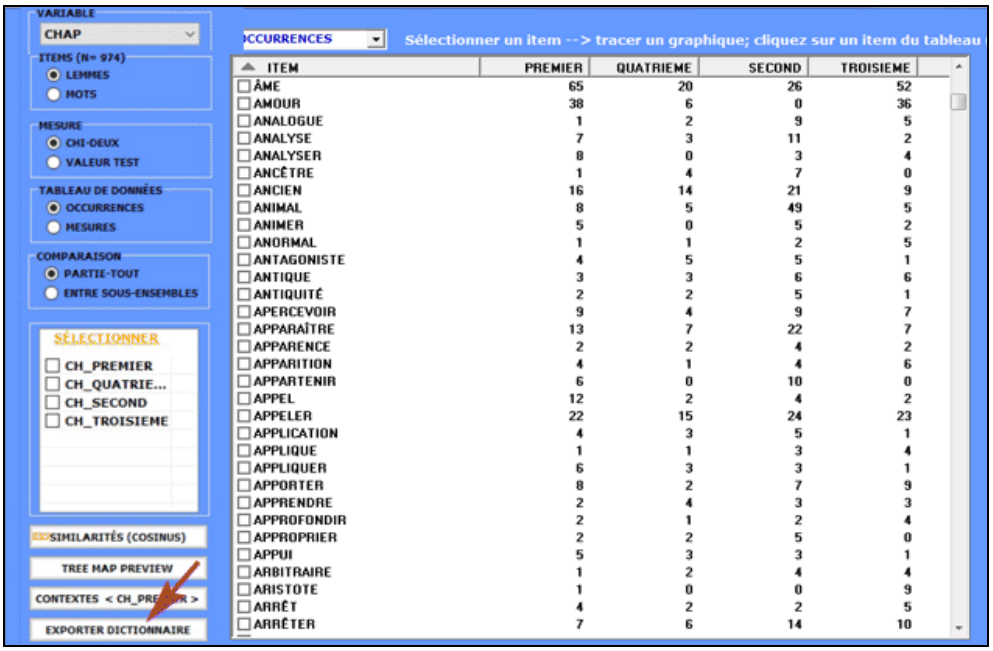





VARIABLE
CHAP
ITEMS (N= 974)
LEMMES
MOTS
MESURE
CHI-DEUX
VALEUR TEST
TABLEAU DE DONNÉES
OCCURRENCES
MESURES
COMPARAISON
PARTIE-TOUT
ENTRE SOUS-ENSEMBLES
SÉLECTIONNER
CH_PREMIER
CH_QUATRIE...
CH_SECOND
CH_TROISIEME
SIMILARITÉS (COSINUS)
TREE MAP PREVIEW
CONTEXTES < CH_PREMIER >
EXPORTER DICTIONNAIRE
ITEM
PRE...
QUA...
SECO...
TROI...
ACTUEL
ADMETTRE
ADOPTER
ADORATION
ADRESSER
AFFAIRE
AFFIRMER
AGIR
AIDER
AILLEURS
AIMER
AJOUTER
ALLER_AME
ÂME
AMOUR
ANALOGUE
ANALYSE
ANALYSER
ANCÊTRE
ANCIEN
ANIMAL
ANIMER
ANORMAL
ANTAGONIS
ANTIQUE
ANTIQUITÉ
APERCEVOI
APPARAÎTR
APPARENC
APPARITIO
ANALYSE DES SPÉCIFICITÉS
HISTOGRAMME
PIE CHART
Utiliser le bouton droit de la souris
ÂME
(OCCURRENCES)
60
50
40
30
20
10
0
PREMIER
QUATRIEME
SECOND
TROISIEME
65,0
20,0
26,0
52,0


VARIABLE
CHAP
ITEMS (N= 974)
LEMMES
MOTS
MESURE
CHI-DEUX
VALEUR TEST
TABLEAU DE DONNÉES
OCCURRENCES
MESURES
COMPARAISON
PARTIE-TOUT
ENTRE SOUS-ENSEMBLES
SÉLECTIONNER
CH_PREMIER
CH_QUATRIE...
CH_SECOND
CH_TROISIEME
SIMILARITÉS (COSINUS)
TREE MAP PREVIEW
CONTEXTES < CH_PREMIER >
EXPORTER DICTIONNAIRE
ITEM
ACTUEL
ADMETTRE
ADOPTER
ADORATION
ADRESSER
AFFAIRE
AFFIRMER
AGIR
AIDER
AILLEURS
AIMER
AJOUTER
ALLER_AME
ÂME
AMOUR
ANALOGUE
ANALYSE
ANALYSER
ANCÊTRE
ANCIEN
ANIMAL
ANIMER
ANORMAL
ANTAGONIS
ANTIQUE
ANTIQUITÉ
APERCEVOI
APPARAÎTR
APPARENCE
APPARITION
ANALYSE DES SPÉCIFICITÉS
HISTOGRAMME
PIE CHART
Utiliser le bouton droit de la souris
ÂME
(CHI-DEUX)
20
15
10
5
0
-5
-10
-15
-20
-25
PREMIER
QUATRIEME
SECOND
TROISIEME
6,1
-2,0
-25,1
21,1

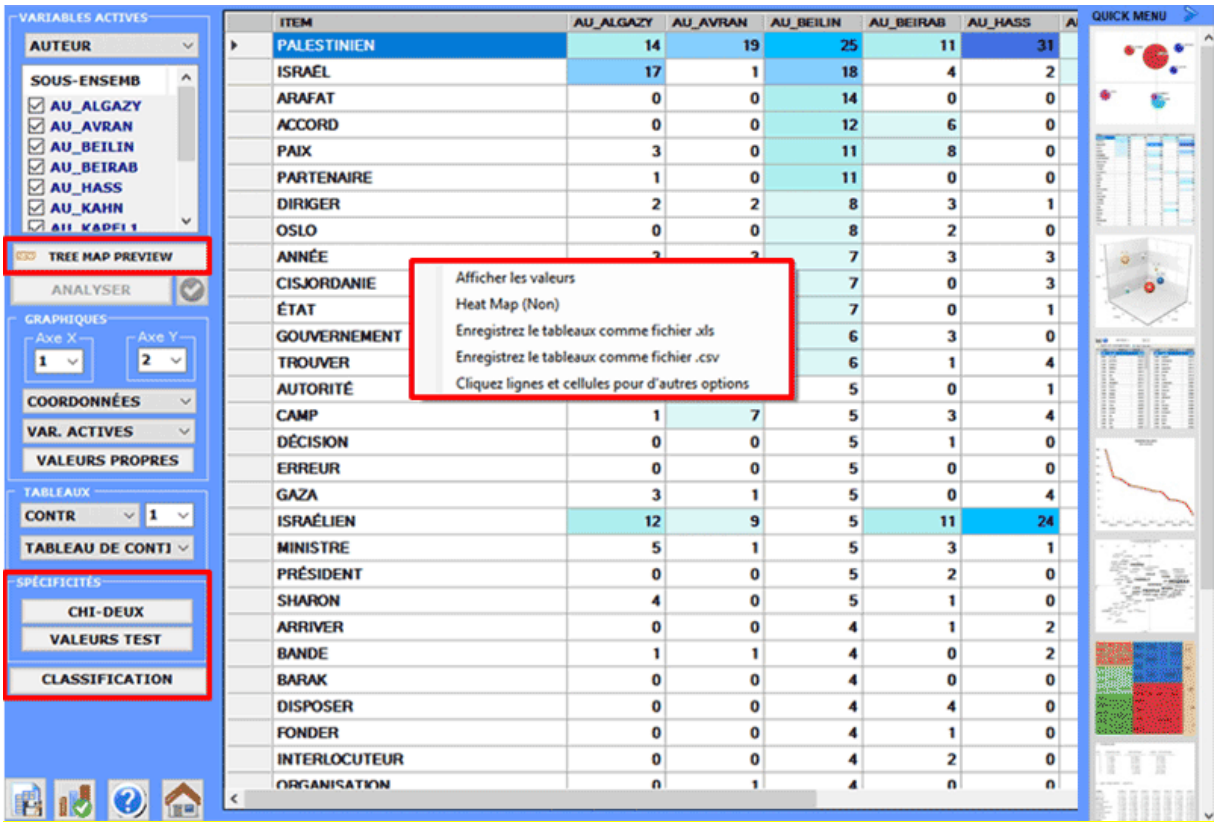
Finalement, en cliquant sur l'option appropriée (voir ci-dessous), un fichier **dictionnaire** avec l'extension .dictio est créé, qui est prêt à être importé par les outils **T-LAB** pour l'**analyse thématique**. Ce dictionnaire comprend tous les mots typiques de la variable catégorielle sélectionnée.



ITEM	PREMIER	QUATRIEME	SECOND	TROISIEME
ÂME	65	20	26	52
AMOUR	38	6	0	36
ANALOGUE	1	2	9	5
ANALYSE	7	3	11	2
ANALYSER	8	0	3	4
ANCÊTRE	1	4	7	0
ANCIEN	16	14	21	9
ANIMAL	8	5	49	5
ANIMER	5	0	5	2
ANORMAL	1	1	2	5
ANTAGONISTE	4	5	5	1
ANTIQUE	3	3	6	6
ANTIQUITÉ	2	2	5	1
APERCEVOIR	9	4	9	7
APPARAÎTRE	13	7	22	7
APPARENCE	2	2	4	2
APPARITION	4	1	4	6
APPARTENIR	6	0	10	0
APPEL	12	2	4	2
APPELER	22	15	24	23
APPLICATION	4	3	5	1
APPLIQUE	1	1	3	4
APPLIQUER	6	3	3	1
APPORTER	8	2	7	9
APPRENDRE	2	4	3	3
APPROFONDIR	2	1	2	4
APPROPRIER	2	2	5	0
APPUI	5	3	3	1
ARBITRAIRE	1	2	4	4
ARISTOTE	1	0	0	9
ARRÊT	4	2	2	5
ARRÊTER	7	6	14	10

Analyse des Correspondances

N.B.: Les images de cette section font référence à une version précédente de T-LAB. En **T-LAB 10**, l'aspect est légèrement différent. En outre : a) le **bouton droit** sur les tableaux avec les mots-clés rend disponibles des options supplémentaires ; b) il y a un nouveau bouton ('TREE MAP PREVIEW) qui permet à l'utilisateur de créer plusieurs graphiques dynamiques au format HTML; c) deux nouveaux boutons nous permettent de vérifier les spécificités de chaque variable en utilisant le test du Khi-deux ou la valeur test; d) il y a un bouton qui permet de réaliser une **analyse de clusters** qui utilise les coordonnées des objets (selon les cas, d' unités lexicales ou d' unités de contexte) sur les premiers axes factoriels (jusqu'à un maximum de 10) ; e) une galerie d'images à accès rapide qui fonctionne comme un menu supplémentaire permet de basculer entre les différentes sorties en un seul clic.
Certaines de ces nouvelles fonctionnalités sont mises en évidence dans l'image ci-dessous.



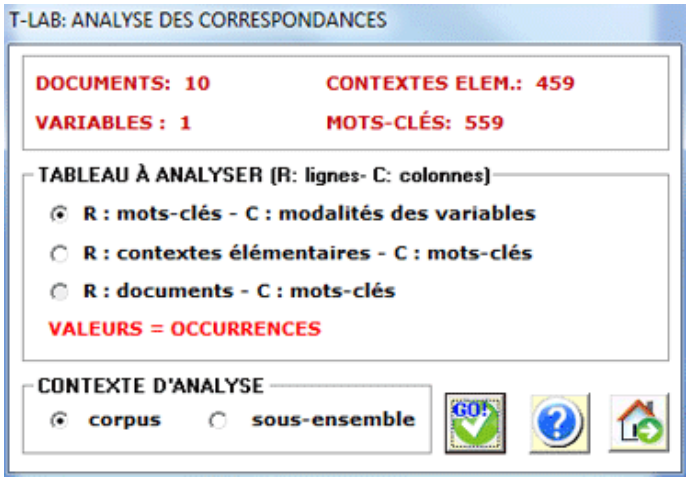
Cet outil **T-LAB** a le but de mettre en évidence les **similitudes** et les **différences** entre les unités de contexte.

Dans **T-LAB,** l'analyse des correspondances nous permet d'analyser trois genres de tableaux:

(**A**) tableaux mots par variable, avec les valeurs des **occurrences**;

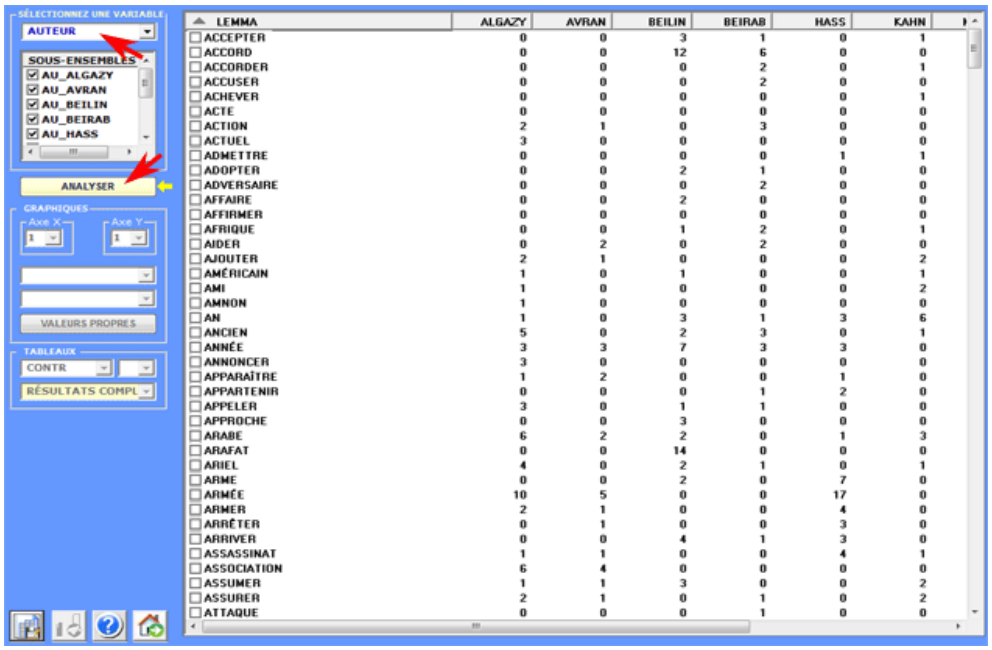
(**B**) tableaux contextes élémentaires par mots, avec les valeurs des **co-occurrences ;**

(**D**) tableaux documents par mots, avec les valeurs des **occurrences.**

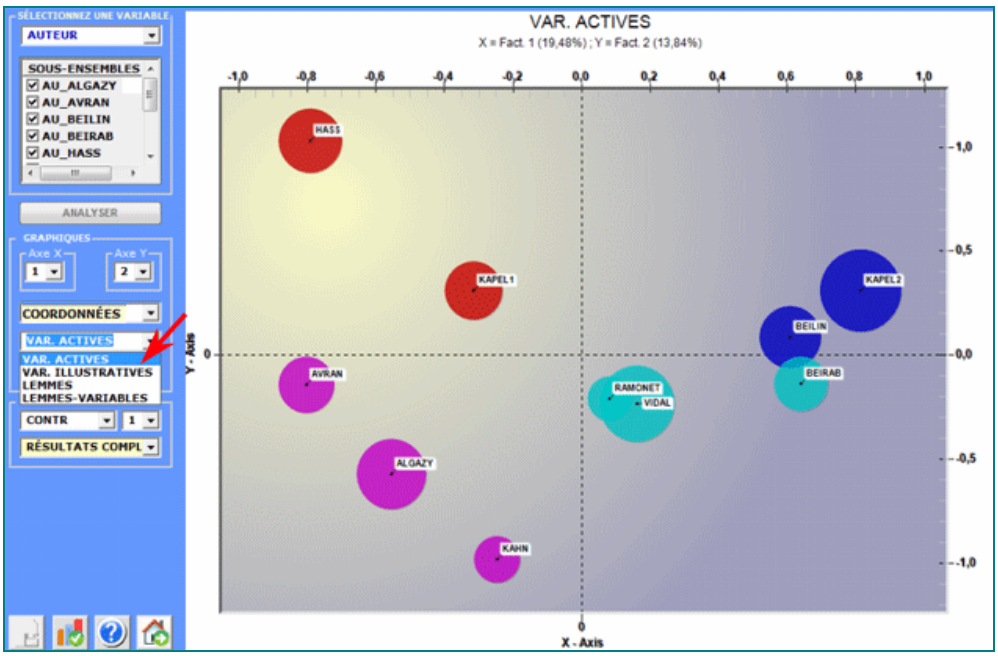


Pour analyser les tableaux (**A**) lemmes (ou mots) par variables, le corpus doit se composer de trois textes au minimum ou être codifié avec quelques variables (pas moins de trois modalités).

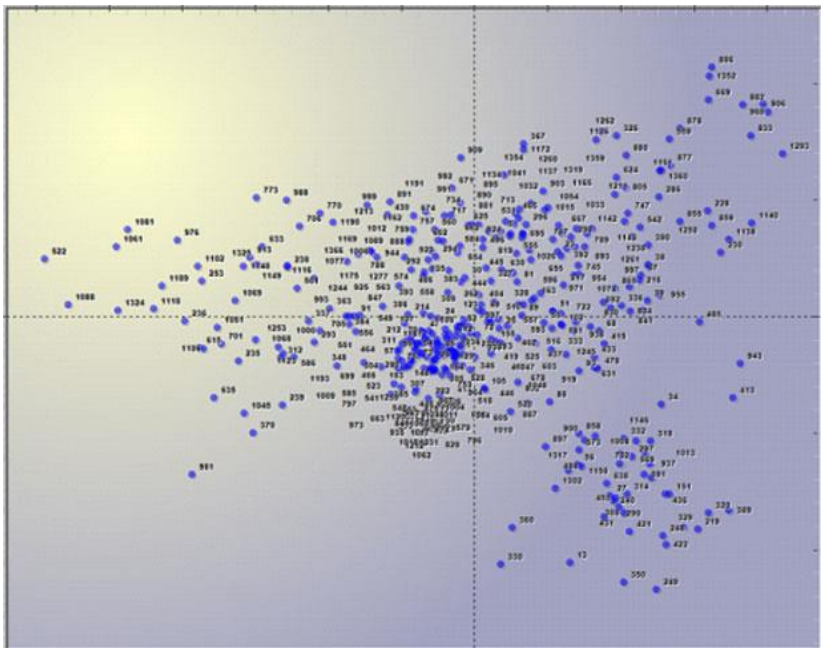
Les variables sont énumérées dans une boîte appropriée et peuvent être employées une à la fois. Après chaque choix, **T-LAB** montre le tableau de contingence correspondant et vous êtes invités à cliquer le bouton **analyser** (voir ci-dessous).



Le résultat de l'analyse se compose de tableaux à partir desquels **T-LAB** produit des diagrammes où sont représentés les rapports entre les sous-ensembles du corpus et entre les unités lexicales dont ils font partie.
Plus précisément, selon les cas, les types de graphiques disponibles montrent les relations entre **variables actives**, entre **variables illustratives**, entre lemmes, entre lemmes et variables.

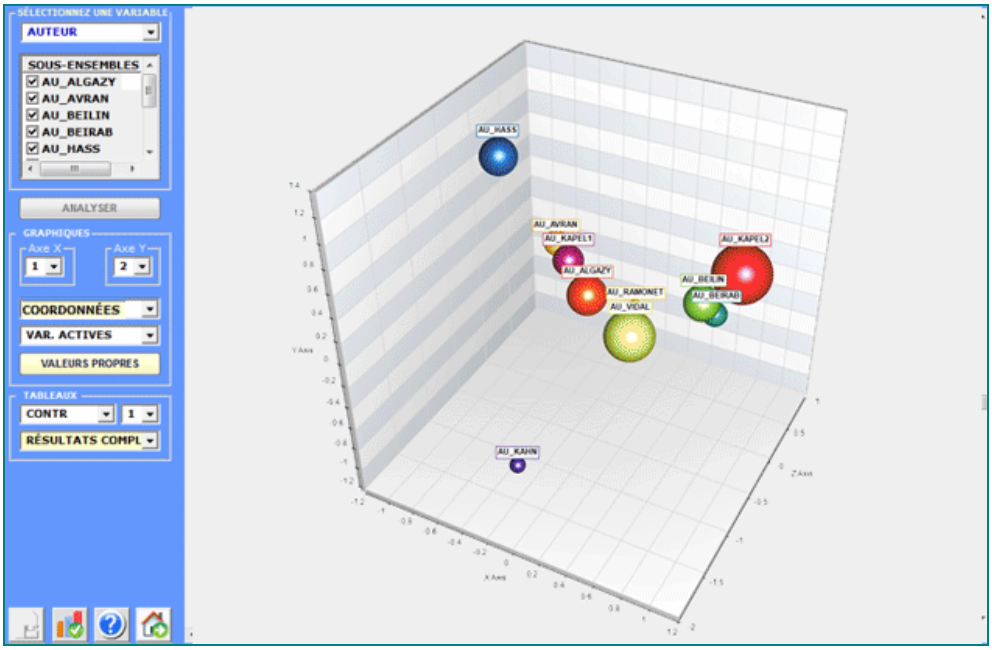


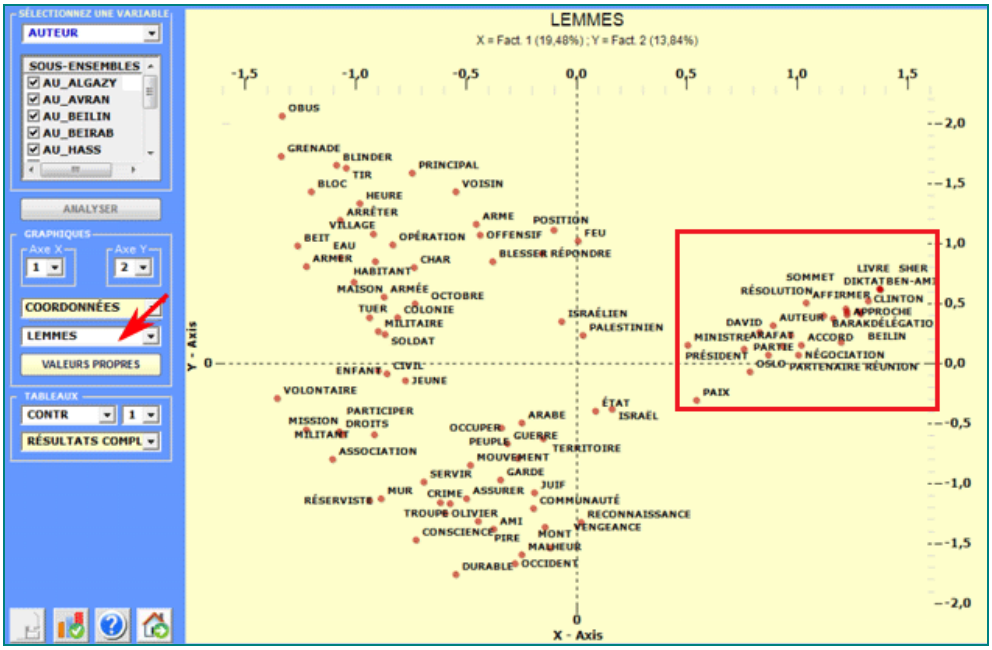
En outre, lorsque le tableau analysé est du genre documents pour mots, on peut voir les points (Max 3000) qui correspondent à chaque document.

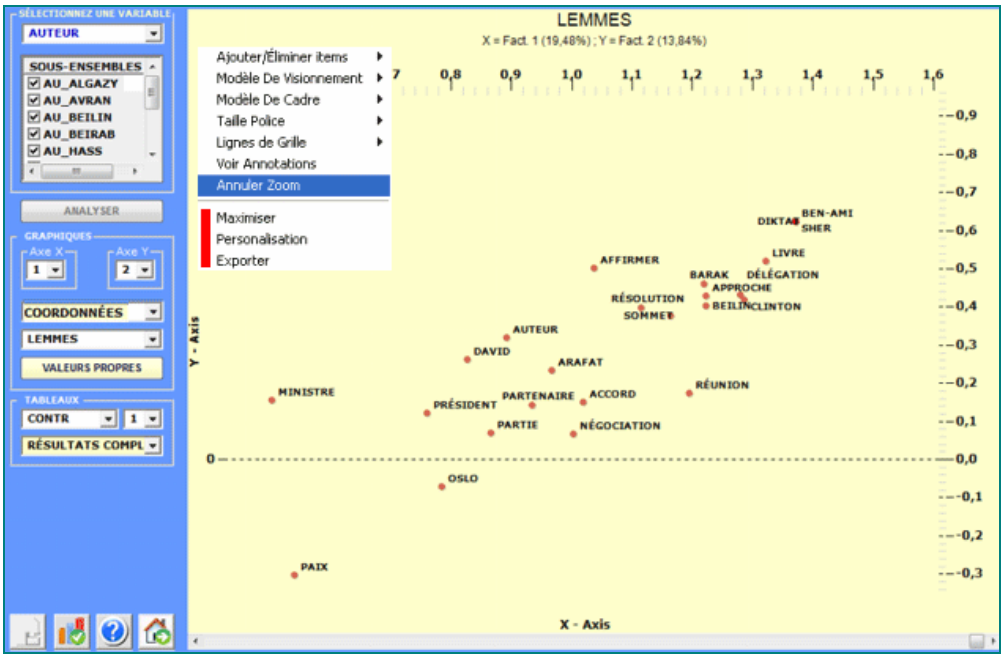


Tous les graphiques peuvent être maximisés et personnalisés en employant la boîte de

dialogue appropriée (utiliser le bouton droit de la souris). D'ailleurs, quand les catégories variables sont 3 ou plus, leurs rapports peuvent être explorés en **3d** (voir ci-dessous).

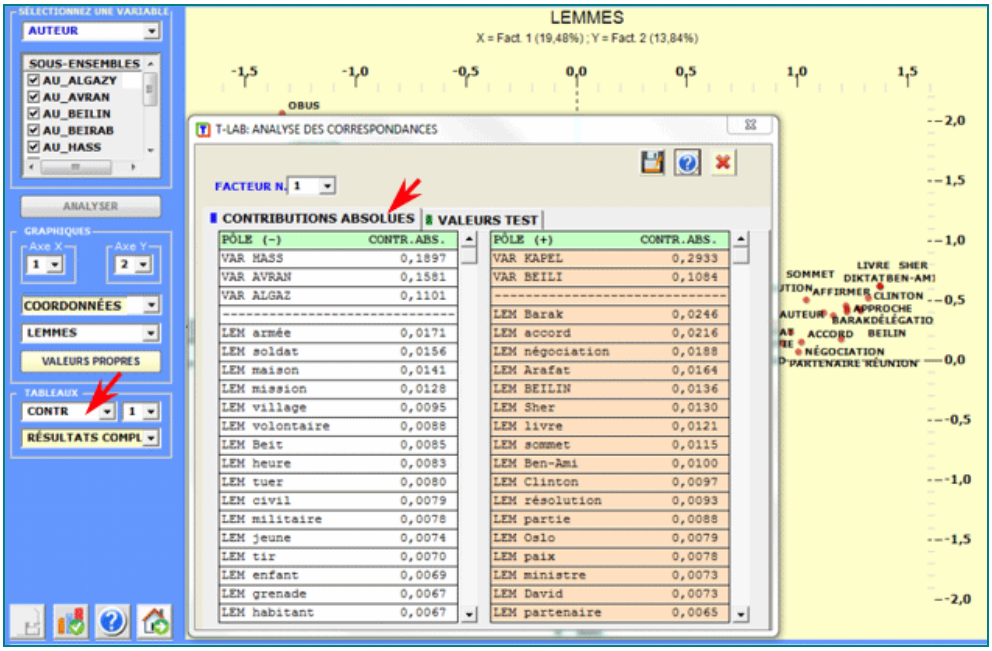




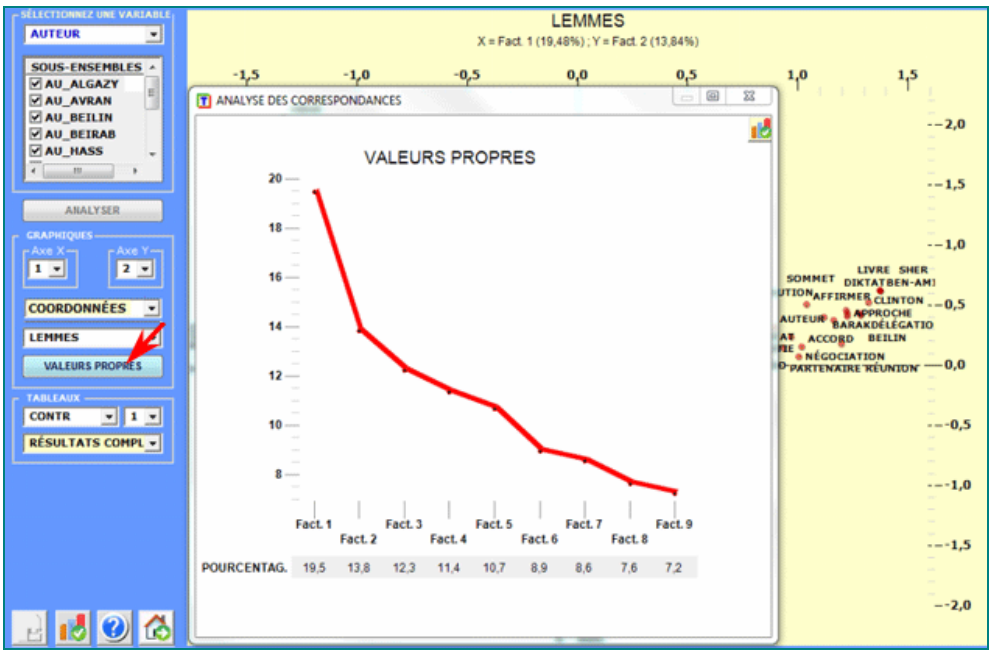


Pour explorer les diverses combinaisons des axes factoriels il suffit de les sélectionner dans les boîtes appropriées ("Axe X", "Axe Y").

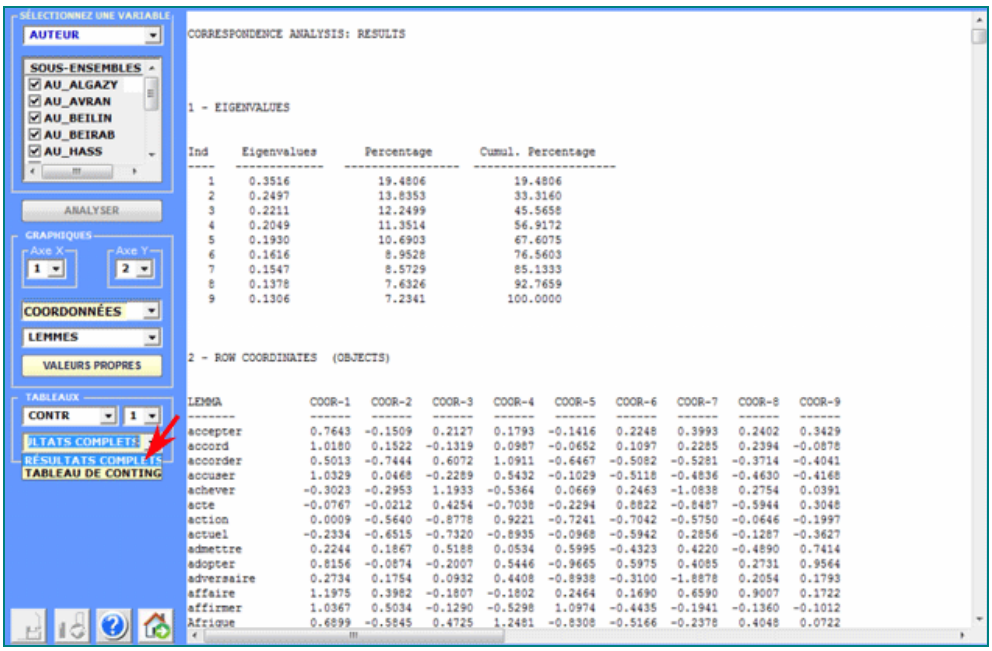
Pour chaque axe factoriel T-LAB fournit deux tableaux qui aident à l'interprétation: celles avec les **Contributions Absolues** et celles avec les **Valeurs Test**.



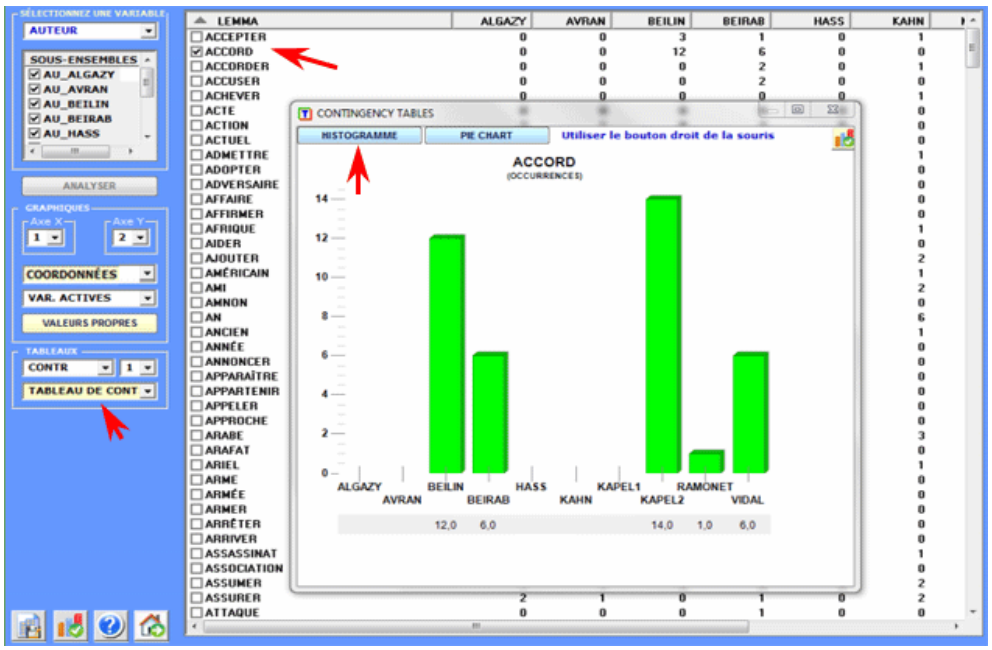
En utilisant l'histogramme des valeurs-propres il est possible d'apprécier l'importance relative de chaque facteur c.-à-d. le pourcentage d'inertie qu'ils déploient.

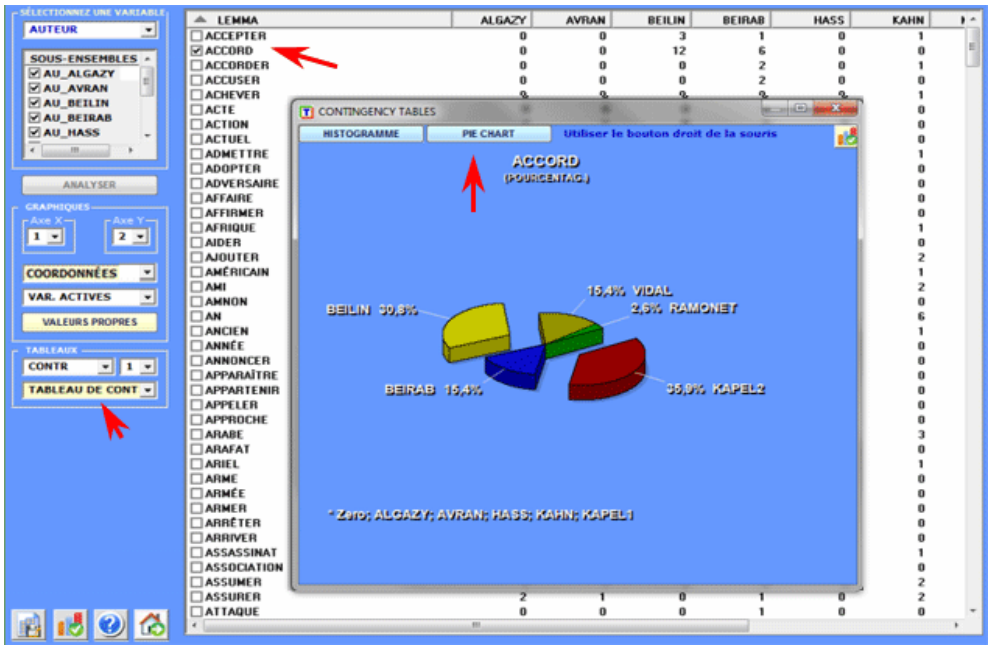


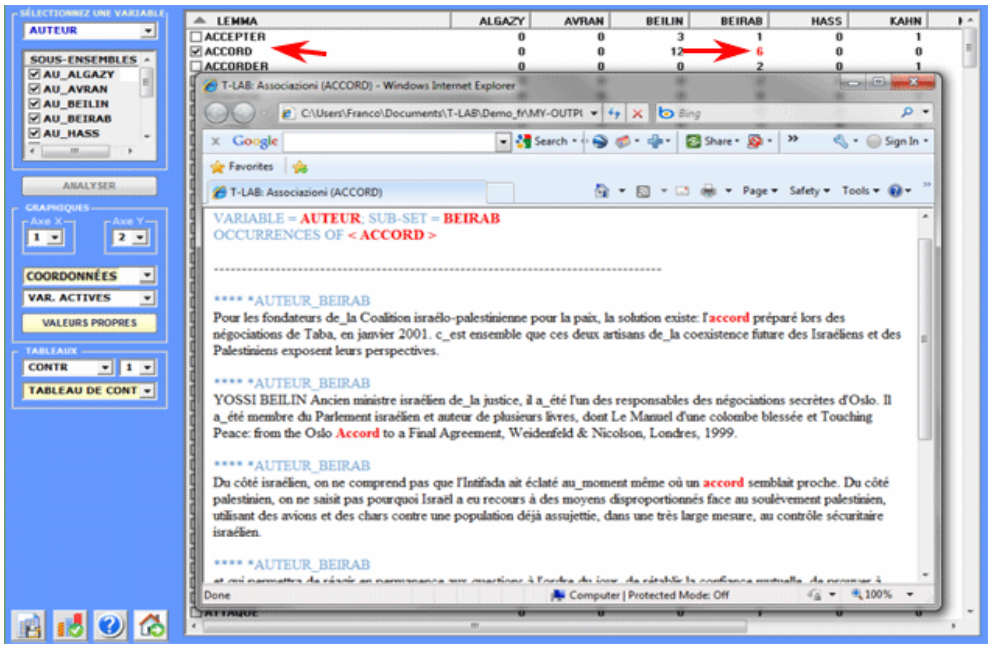
Un clic sur le bouton **Résultats Complets** vous permet de visualiser et de sauvegarder le fichier qui contient tous les résultats de l'analyse: valeurs propres, coordonnées, contributions absolues et relatives, valeurs test.



Tous les tableaux de contingence peuvent être facilement explorés et nous permettent de créer différents types des graphiques. De plus, en cliquant sur cellules spécifiques du tableau (voir ci-dessous), il est possible de créer un fichier HTML montrant tous les contextes élémentaires où le mot en ligne est présent dans le sous-ensemble correspondant.


AUTEUR
SOUS-ENSEMBLES
AU_ALGAZY
AU_AVRAN
AU_BEILIN
AU_BEIRAB
AU_HASS
ANALYSER
GRAPHIQUES
COORDONNÉES
VAR. ACTIVES
VALEURS PROPRES
TABLEAUX
TABLEAU DE CONT
LEMMA
ALGAZY
AVRAN
BEILIN
BEIRAB
HASS
KAHN
CONTINGENCY TABLES
HISTOGRAMME
PIE CHART
Utiliser le bouton droit de la souris
ACCORD
(OCCURRENCES)
ALGAZY
AVRAN
BEILIN
BEIRAB
HASS
KAHN
KAPEL1
KAPEL2
RAMONET
VIDAL
12,0
6,0
14,0
1,0
6,0


AUTEUR
SOUS-ENSEMBLES
AU_ALGAZY
AU_AVRAN
AU_BEILIN
AU_BEIRAB
AU_HASS
ANALYSER
GRAPHIQUES
COORDONNÉES
VAR. ACTIVES
VALEURS PROPRES
TABLEAUX
TABLEAU DE CONT
LEMMA
CONTINGENCY TABLES
HISTOGRAMME
PIE CHART
Utiliser le bouton droit de la souris
ACCORD
(POURCENTAGE)
15,4% VIDAL
2,6% RAMONET
BEILIN 30,8%
BEIRAB 15,4%
35,8% KAPEL2
* Zero: ALGAZY; AVRAN; HASS; KAHN; KAPEL1



Dans le cas des tableaux (**B**) et (**C**), elles sont constituées par autant de lignes que sont les unités de contextes (max 10.000) et autant de colonnes que sont les Mots-Clés sélectionnés (max 1.500).

L'algorithme de calcul et les résultats sont semblables à ceux de l'analyse unités lexicales par variables, sauf que dans ce cas, pour limiter le temps d'élaboration, **T-LAB** se limite à extraire les premiers 10 facteurs: un nombre plus que suffisant pour condenser la variabilité des données.

En outre, on peut par la suite effectuer deux types de **classification** dont les "objets" sont aussi bien constitués par des Mots-Clés que par des segments (contextes élémentaires).

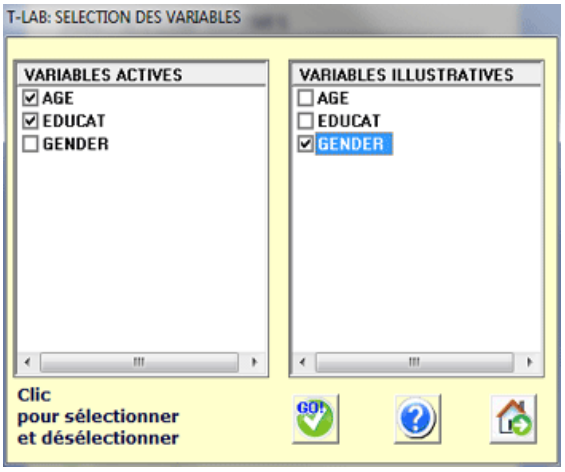
Analyse des Correspondances Multiples

N.B.: Les images de cette section font référence à une version précédente de T-LAB. En **T-LAB 10**, l'aspect est légèrement différent. Il y a aussi un bouton qui vous permet d'effectuer une **analyse de clusters** qui utilise les coordonnées des objets sur les premiers axes factoriels (jusqu'à un maximum de 10).

L'Analyse des Correspondances Multiples, qui peut être considérée une extension de l'Analyse des Correspondances simple (voir ci-dessus), nous permet d'analyser les rapports entre deux ou plus variables catégorielles.

Dans **T-LAB**, les limitations de ce type d'analyse sont les suivantes:
- 150.000 contextes élémentaires (lignes);
- 250 catégories (colonnes);
- 3.000 mots clés, analysés comme colonnes supplémentaires (voir Lebart L., Salem A., 1994).

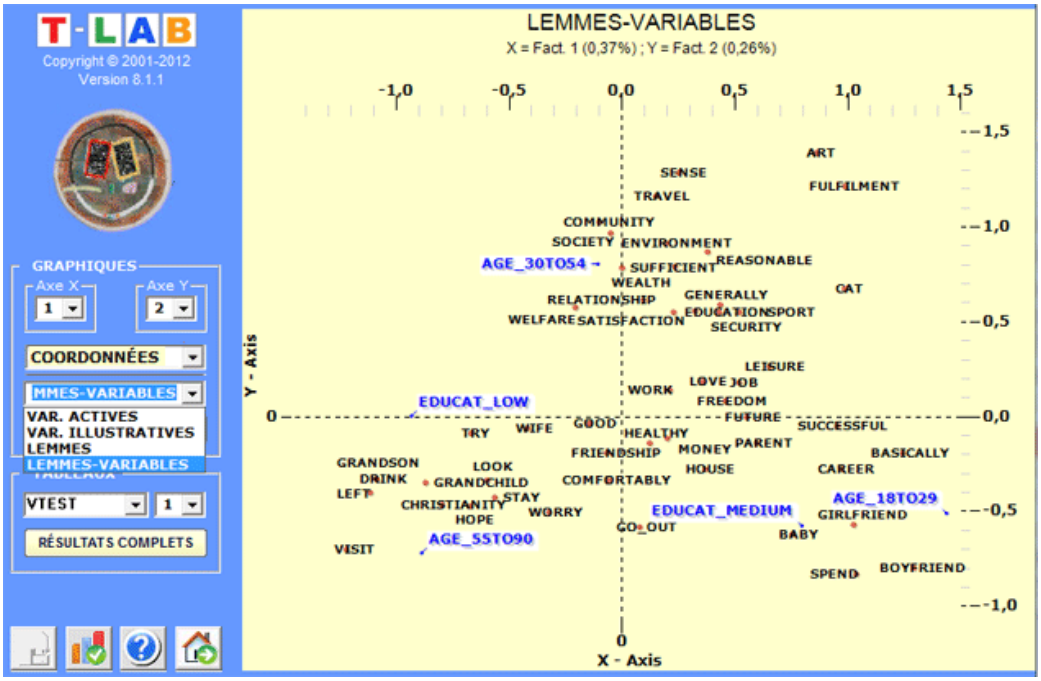
L'Analyse de Correspondances Multiples, disponible seulement si le corpus inclut au moins deux variables, exige que l'utilisateur choisisse ses options dans la fenêtre suivante:

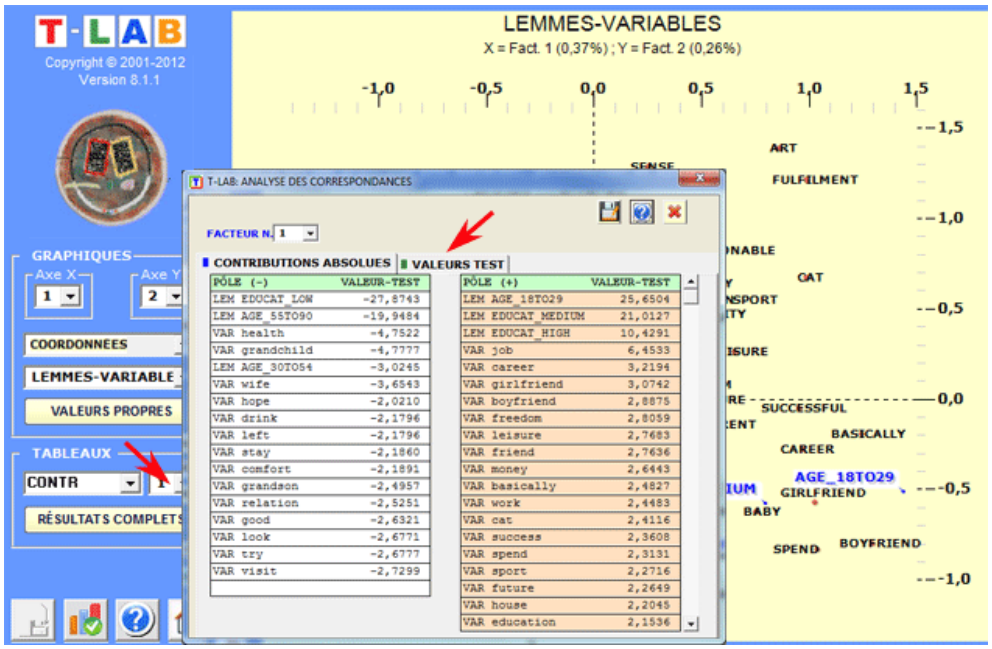


À la fin de l'analyse:

- les outputs de **T-LAB** sont analogues à ceux de l'Analyse de Correspondance (voir ci-dessous) complétés avec une tableau de Burt (Burt_Table.xls) qui inclut les valeurs croisées des variables utilisées;

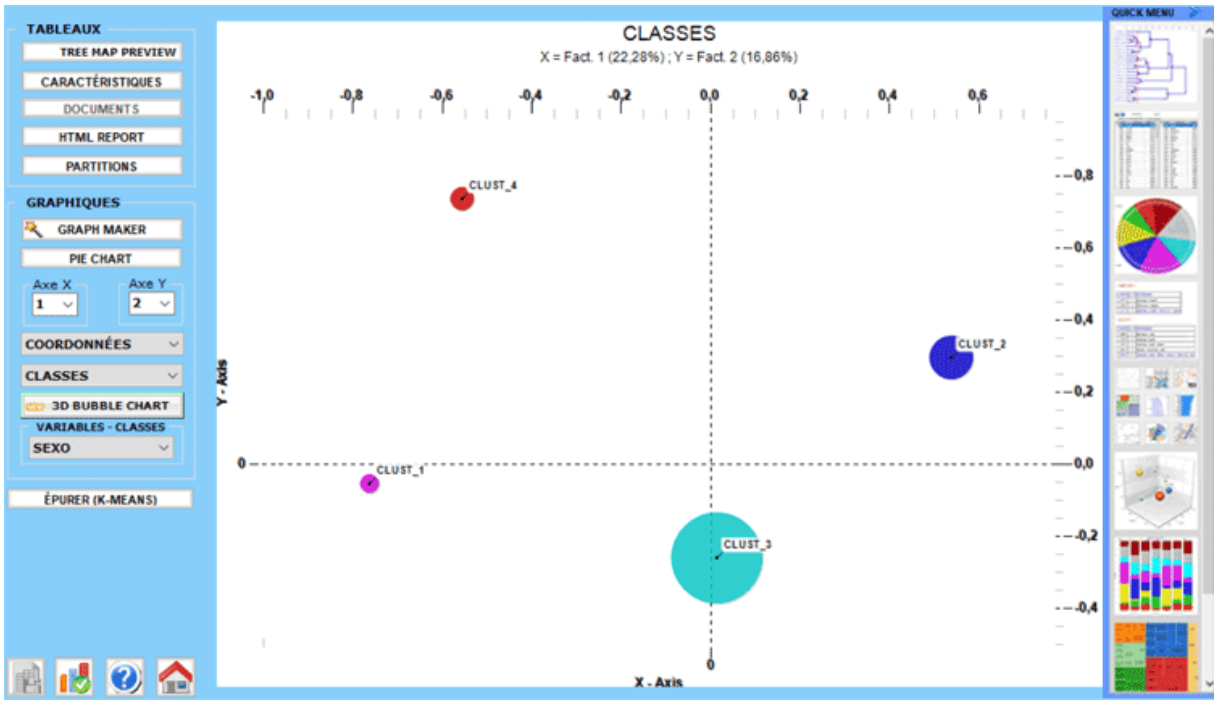
- seulement quand les contextes élémentaires correspondent aux documents primaires (par exemple réponses aux questions ouvertes) il est possible d'effectuer une **cluster analysis**.



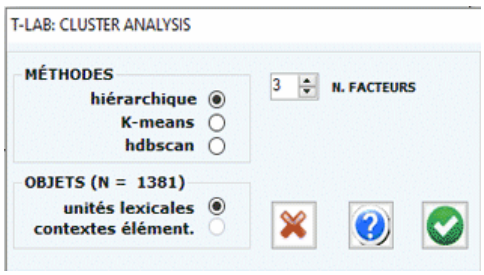


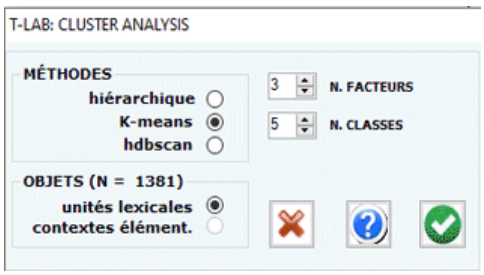
Cluster Analysis (Classification)

N.B.: Les images de cette section font référence à une version précédente de **T-LAB**.
En **T-LAB 10**, l'aspect est légèrement différent. En outre: a) il y a un nouveau bouton (TREE MAP PREVIEW) qui permet à l'utilisateur de créer plusieurs graphiques dynamiques au format HTML; b) le bouton DENDROGRAMME a été remplacé par l'outil GRAPH MAKER ; c) une galerie d'images à accès rapide qui fonctionne comme un menu supplémentaire permet de basculer entre les différentes sorties en un seul clic (voir l'image ci-dessous).



Cette option met en marche un calcul qui emploie les résultats d'une précédente **Analyse des Correspondances**; en particulier, le calcul emploie les coordonnées des objets (unités lexicales ou unités de contextes) sur les premiers axes factoriels (pour un maximum de 10).





Selon les cas, l'utilisateur peut choisir entre trois techniques de classification:

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

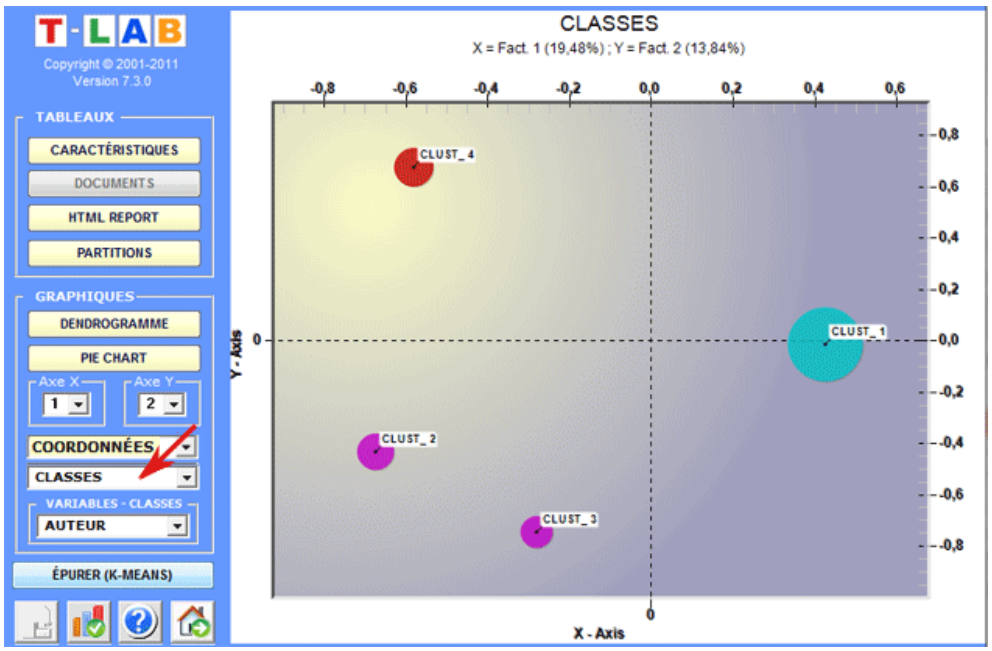
Les deux premières (a, b) permettent d'explorer (tableaux et graphiques) des solutions de 3 à 20 cluster; alors que la troisième (c), qui nécessite un paramètre supplémentaire (c'est-à-dire le nombre minimum de mots dans un cluster), permet à l'utilisateur d'explorer une seule solution.

N.B. : Lorsque la méthode 'hiérarchique' est sélectionnée, **T-LAB** rend disponible une option (voir le bouton 'épurer' ci-dessous) qui permet à l'utilisateur de combiner les méthodes de Ward et K-Means.

Une brève description des trois techniques se trouve dans le **glossaire** de ce manuel.

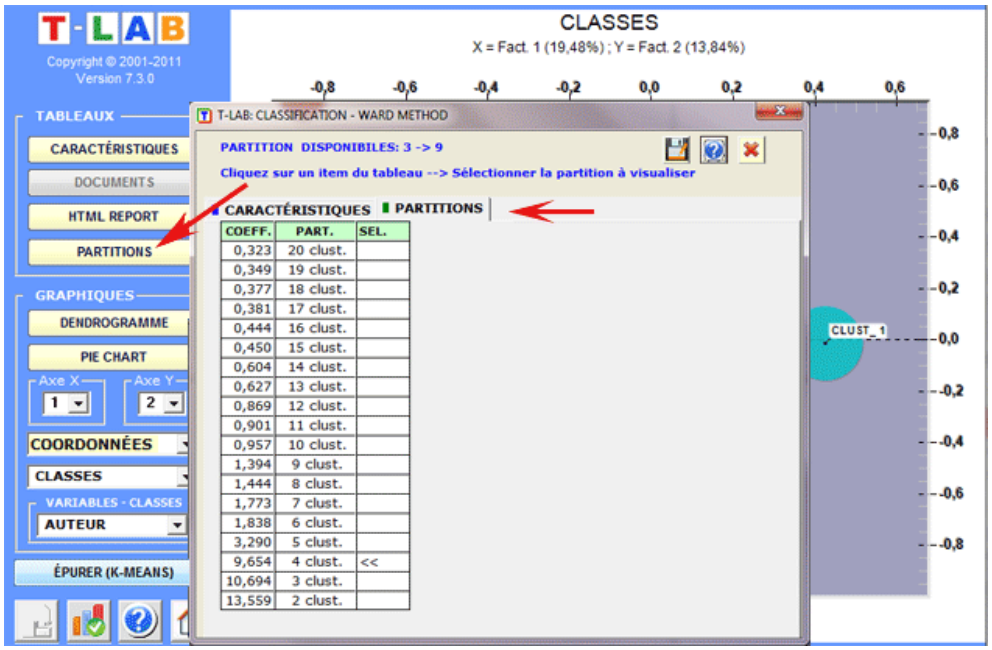
À la fin du traitement, **T-LAB** montre des graphiques et des tableaux.

Les graphiques représentent les classes dans l'espace détecté par la précédente Analyse des Correspondances.

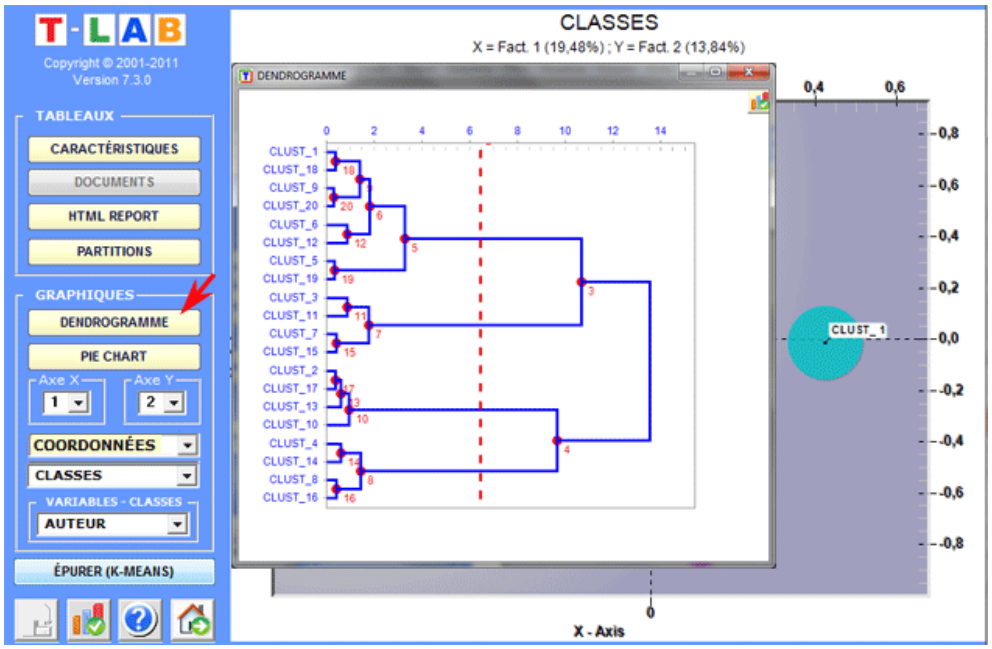


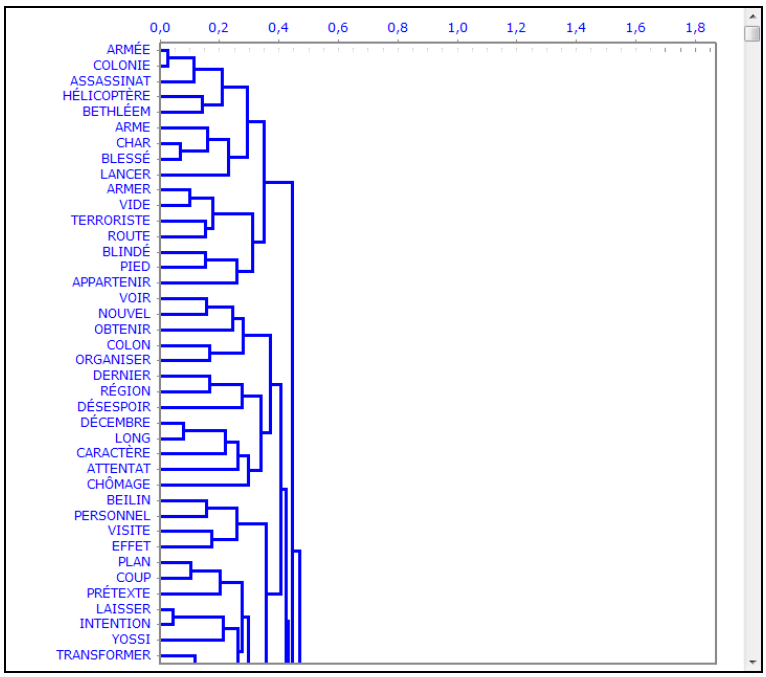
N.B.: Pour explorer les diverses combinaisons des axes factoriels il suffit de les sélectionner dans les boîtes appropriées ("Axe X", "Axe Y").

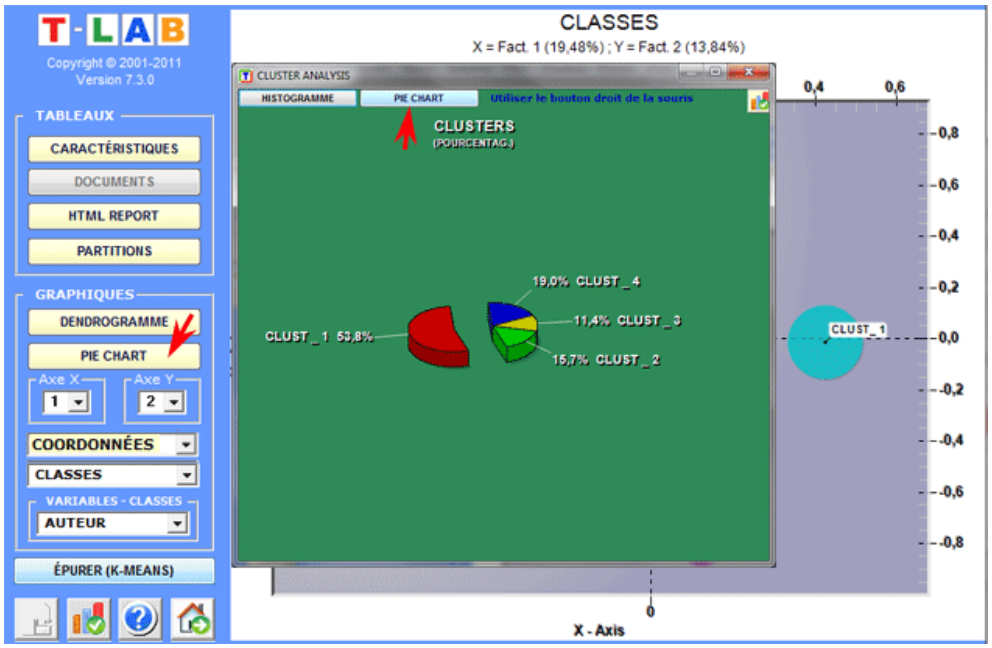
Dans le cas de classification hiérarchique l'utilisateur peut facilement explorer (graphiques et tableaux) les partitions différentes.

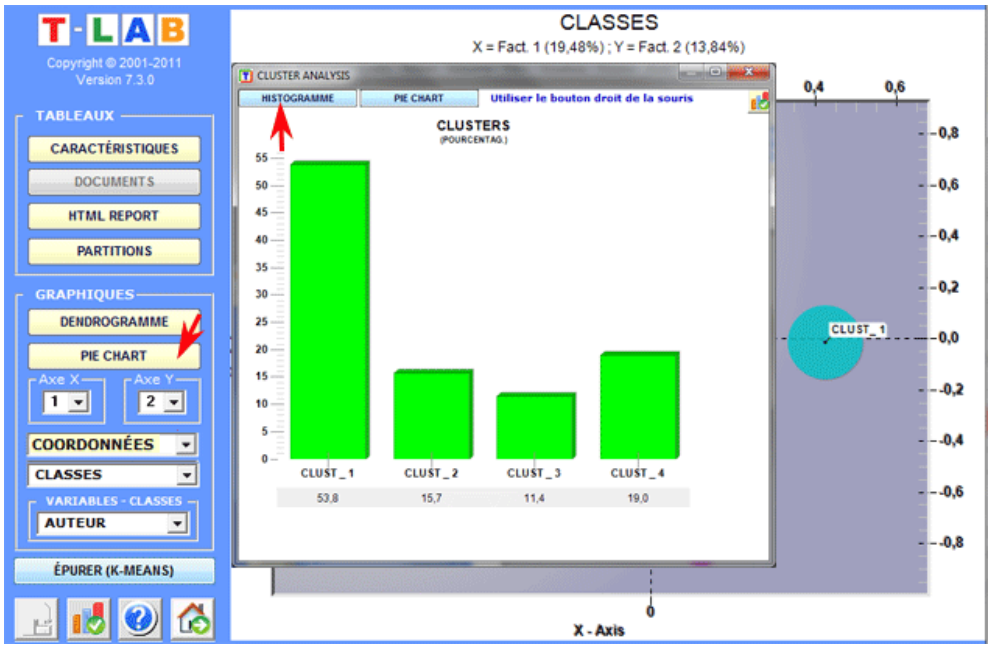


Dendrogrammes, graphiques circulaires (pie charts) et histogrammes montrent les caractéristiques de chaque partition.

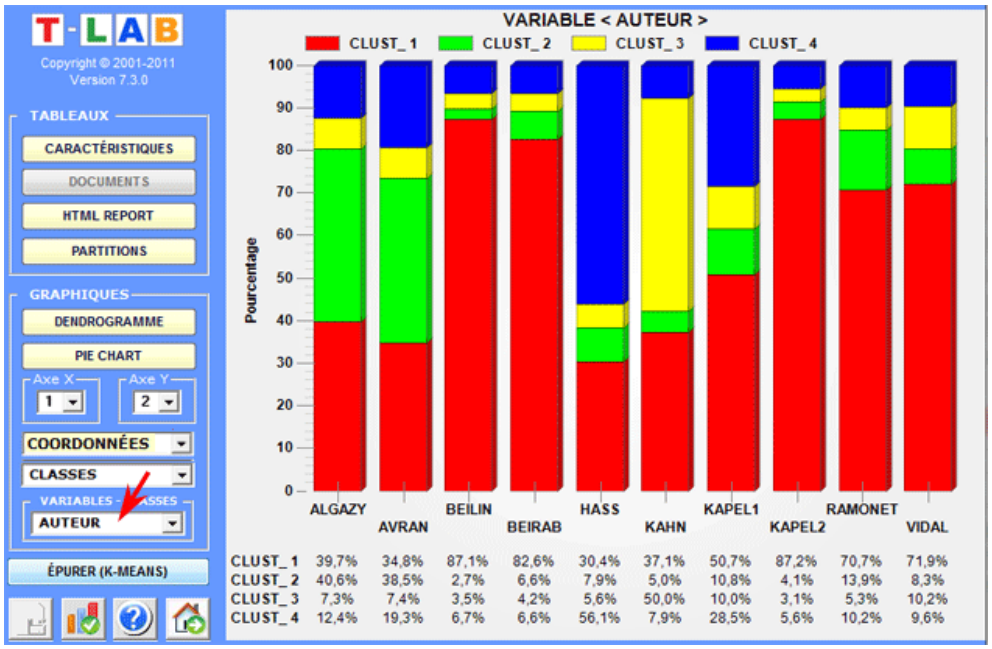



0,0
0,2
0,4
0,6
0,8
1,0
1,2
1,4
1,6
1,8
ARMÉE
COLONIE
ASSASSINAT
HÉLICOPTÈRE
BETHLÉEM
ARME
CHAR
BLESSÉ
LANCER
ARMER
VIDE
TERRORISTE
ROUTE
BLINDÉ
PIED
APPARTENIR
VOIR
NOUVEL
OBTENIR
COLON
ORGANISER
DERNIER
RÉGION
DÉSESPOIR
DÉCEMBRE
LONG
CARACTÈRE
ATTENTAT
CHÔMAGE
BEILIN
PERSONNEL
VISITE
EFFET
PLAN
COUP
PRÉTEXTE
LAISSER
INTENTION
YOSSI
TRANSFORMER


T-LAB
Copyright © 2001-2011
Version 7.3.0
TABLEAUX
CARACTÉRISTIQUES
DOCUMENTS
HTML REPORT
PARTITIONS
GRAPHIQUES
DENDROGRAMME
PIE CHART
Axe X
1
Axe Y
2
COORDONNÉES
CLASSES
VARIABLES - CLASSES
AUTEUR
ÉPURER (K-MEANS)
CLASSES
X = Fact. 1 (19,48%) ; Y = Fact. 2 (13,84%)
CLUSTER ANALYSIS
HISTOGRAMME
PIE CHART
Utiliser le bouton droit de la souris
CLUSTERS
(POURCENTAG.)
19,0% CLUST_4
11,4% CLUST_3
15,7% CLUST_2
CLUST_1 53,8%
CLUST_1
0,4
0,6
0,8
0,6
0,4
0,2
0,0
-0,2
-0,4
-0,6
-0,8
0
X - Axis

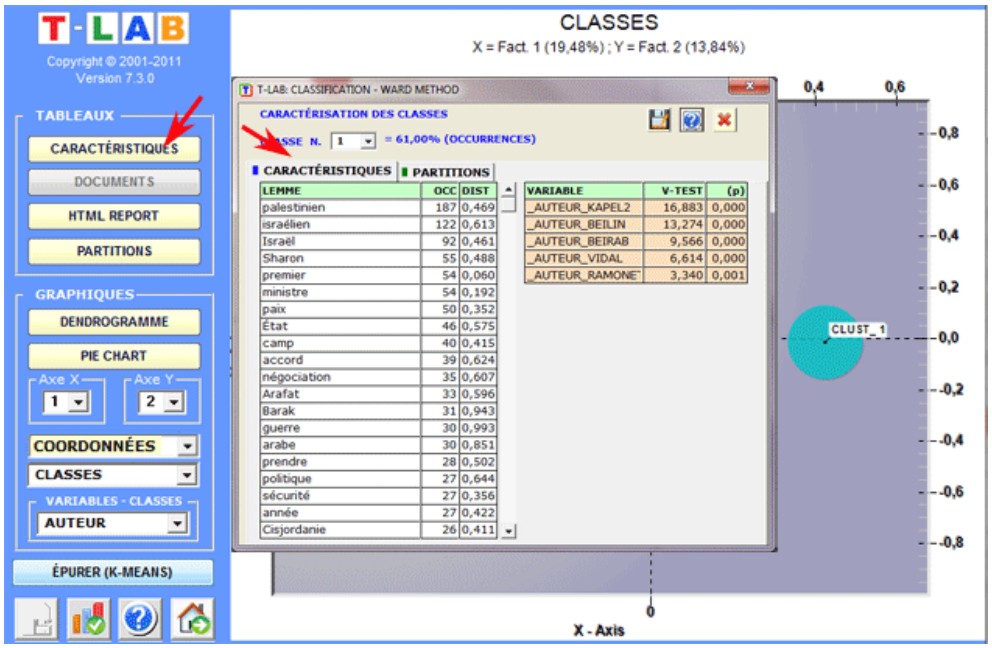


Des histogrammes nous permettent de vérifier les rapports entre les classes et les variables.

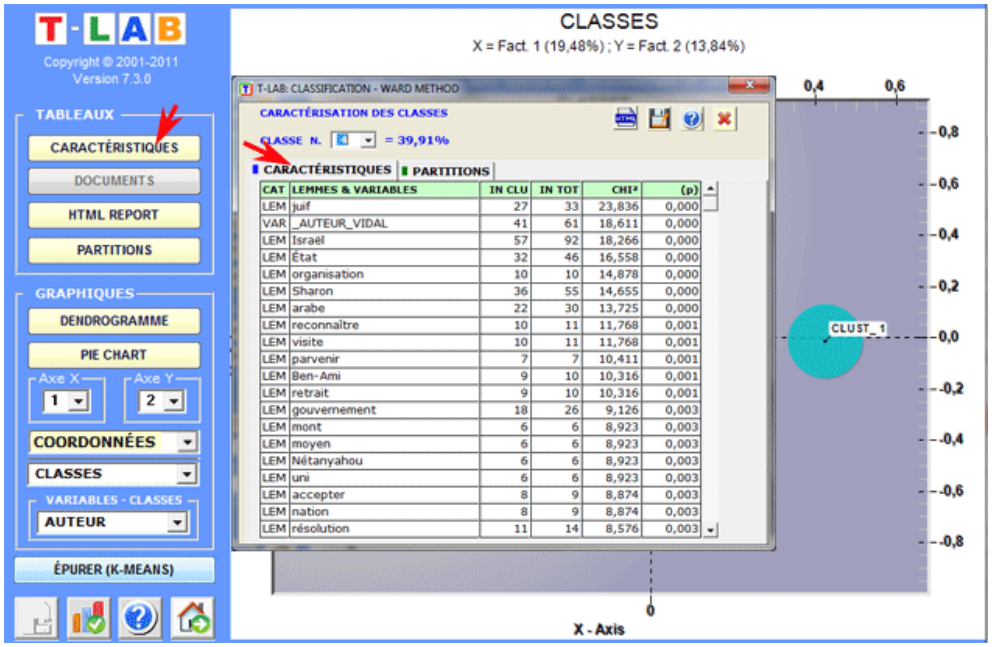


Les tableaux sont de deux types:

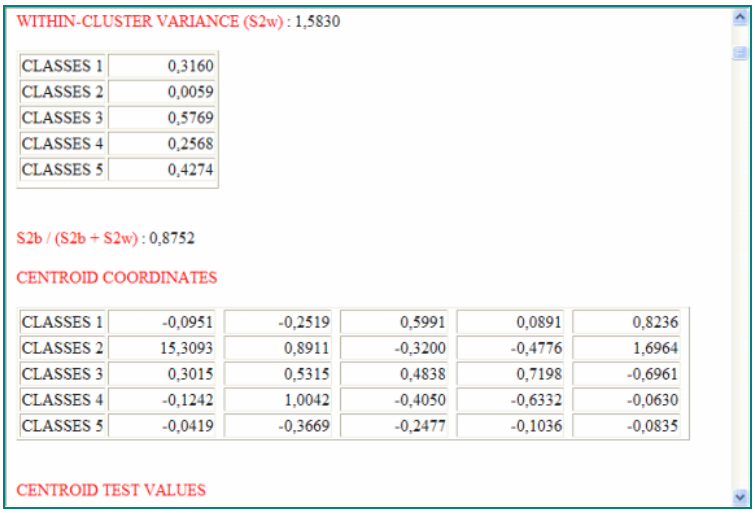
(**A**) si les objets classés sont les unités lexicales, pour chacune d'elles (et pour chaque classe) sont montrées les occurrences respectives ('OCC') et le distances ('DIST') au centroïde; aussi, pour chaque variable qui est sensiblement associée à la classe examinée, est montrée la Valeur Test.



(**B**) si les objets classés sont des contextes élémentaires, les caractéristiques de chaque classe (unités lexicales ou variables) sont décrites au moyen de la même méthode employée dans Analyse Thématique des Contextes Élémentaires. (voir ci-dessous).



Dans le cas d'analyses réalisées avec des méthodes hiérarchiques ou K-means **T-LAB** permet de visualiser et exporter un fichier (voir bouton "Output HTML") dans lequel les caractéristiques des clusters et certaines mesures concernant la qualité de la partition du partage en examen sont reportées.



WITHIN-CLUSTER VARIANCE (S2w) : 1,5830

CLASSES 1	0,3160
CLASSES 2	0,0059
CLASSES 3	0,5769
CLASSES 4	0,2568
CLASSES 5	0,4274

S2b / (S2b + S2w) : 0,8752

CENTROID COORDINATES

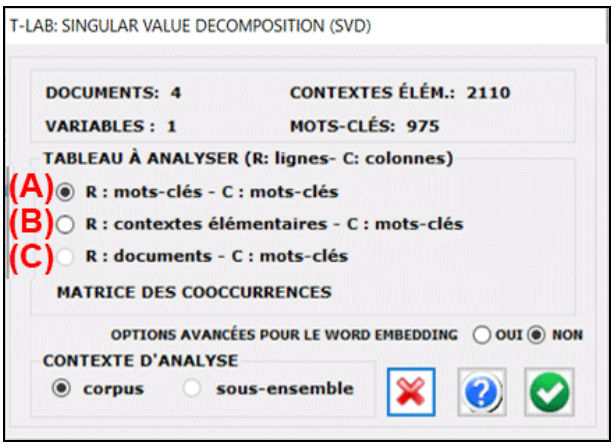
CLASSES 1	-0,0951	-0,2519	0,5991	0,0891	0,8236
CLASSES 2	15,3093	0,8911	-0,3200	-0,4776	1,6964
CLASSES 3	0,3015	0,5315	0,4838	0,7198	-0,6961
CLASSES 4	-0,1242	1,0042	-0,4050	-0,6332	-0,0630
CLASSES 5	-0,0419	-0,3669	-0,2477	-0,1036	-0,0835

CENTROID TEST VALUES

Décomposition en Valeurs Singulières (SVD)

La **Décomposition en Valeurs Singulières** (**SVD** – Singular Value Decomposition) est une technique de réduction des dimensions des données qui, dans le Text Mining, peut être utilisée pour découvrir les **dimensions latentes** (ou composants) qui déterminent les **similitudes sémantiques** entre les mots (c.-à-d. unités lexicales) ou entre documents (c.-à-d. unités de contexte).

T-LAB nous permet d'effectuer une SVD de **trois types de tableaux de données.** Dans le premier cas (voir 'A' ci-dessous), la table de données est une matrice de cooccurrences avec - en ligne et en colonne - les mots-clés sélectionnés. Dans le second cas (voir 'B' ci-dessous), le tableau de données contextes élémentaires X mots-clés contiendra des valeurs de présence / absence (c.-à-d. '1' et '0'). Dans le troisième cas (voir "C" ci-dessous), le tableau les données documents x mots-clés contiendra des valeurs d'occurrence.
N.B.: Veuillez noter que, lorsque vous analysez une matrice de cooccurrences dont les lignes et les colonnes sont des termes clés (voir «A» ci-dessous), **T-LAB** fournit des vecteurs denses de haute qualité (c'est-à-dire des word embeddings).



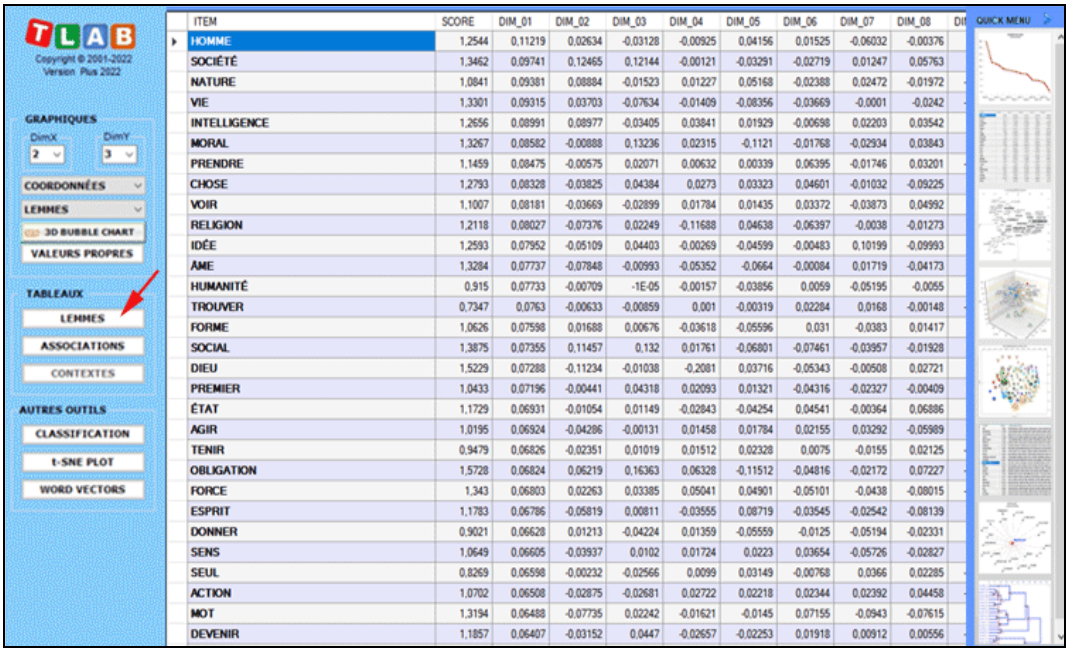
La procédure d'analyse comprend les étapes suivantes:
1 - construction du tableau de données à analyser (jusqu'à 300 000 lignes x 5 000 colonnes);
2 - normalisation TF-IDF et mise à l'échelle des vecteurs de lignes (norme euclidienne);
3 - extraction des 20 premières "dimensions latentes" à travers l'algorithme de Lanczos.
N.B.:
- Dans le cas des matrices de cooccurrence (voir «A» ci-dessus), la normalisation des données est obtenue en utilisant la mesure du cosinus;
- Lorsque les options avancées de word embedding sont sélectionnées, T-LAB calcule les valeurs PPMI (Positive Pointwise Mutual Information) et permet d'utiliser les 50 premières dimensions du SVD.

Les résultats de l'analyse sont affichés dans des **tableaux** et des **graphiques**.

En détail:

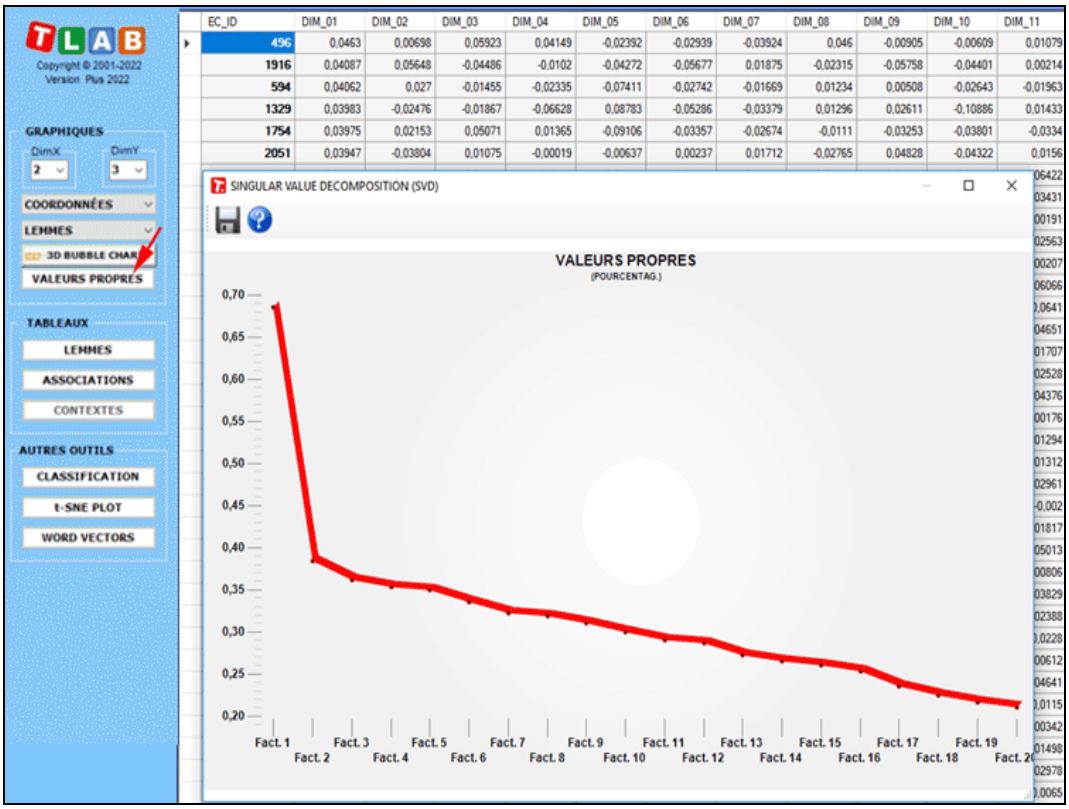
Deux tableaux - dont les lignes peuvent être des unités lexicales ou des unités de contexte - ont autant de colonnes que les dimensions extraites.

Dans le cas du tableau LEMMES (c'est-à-dire unités lexicales), une autre colonne s'affiche, dans laquelle les scores d'importance sont rapportés (voir ci-dessous).



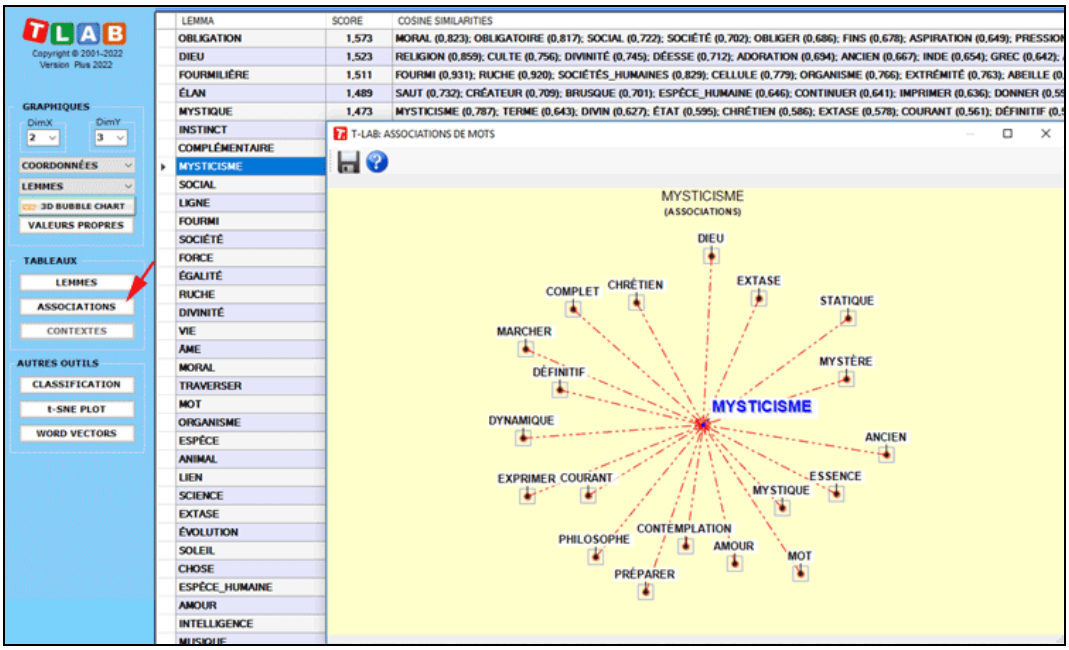
ITEM	SCORE	DIM_01	DIM_02	DIM_03	DIM_04	DIM_05	DIM_06	DIM_07	DIM_08
HOMME	1,2544	0,11219	0,02634	-0,03128	-0,00925	0,04156	0,01525	-0,06032	-0,00376
SOCIÉTÉ	1,3462	0,09741	0,12465	0,12144	-0,00121	-0,03291	-0,02719	0,01247	0,05763
NATURE	1,0841	0,09381	0,08884	-0,01523	0,01227	0,05168	-0,02388	0,02472	-0,01972
VIE	1,3301	0,09315	0,03703	-0,07634	-0,01409	-0,08356	-0,03669	-0,0001	-0,0242
INTELLIGENCE	1,2656	0,08991	0,08977	-0,03405	0,03841	0,01929	-0,00698	0,02203	0,03542
MORAL	1,3267	0,08582	-0,00888	0,13236	0,02315	-0,1121	-0,01768	-0,02934	0,03843
PRENDRE	1,1459	0,08475	-0,00575	0,02071	0,00632	0,00339	0,06395	-0,01746	0,03201
CHOSE	1,2793	0,08328	-0,03825	0,04384	0,0273	0,03323	0,04601	-0,01032	-0,09225
VOIR	1,1007	0,08181	-0,03669	-0,02899	0,01784	0,01435	0,03372	-0,03873	0,04992
RELIGION	1,2118	0,08027	-0,07376	0,02249	-0,11688	0,04638	-0,06397	-0,0038	-0,01273
IDÉE	1,2593	0,07952	-0,05109	0,04403	-0,00269	-0,04599	-0,00483	0,10199	-0,09993
ÂME	1,3284	0,07737	-0,07848	-0,00993	-0,05352	-0,0664	-0,00084	0,01719	-0,04173
HUMANITÉ	0,915	0,07733	-0,00709	-1E-05	-0,00157	-0,03856	0,0059	-0,05195	-0,0055
TROUVER	0,7347	0,0763	-0,00633	-0,00859	0,001	-0,00319	0,02284	0,0168	-0,00148
FORME	1,0626	0,07598	0,01688	0,00676	-0,03618	-0,05596	0,031	-0,0383	0,01417
SOCIAL	1,3875	0,07355	0,11457	0,132	0,01761	-0,06801	-0,07461	-0,03957	-0,01928
DIEU	1,5229	0,07288	-0,11234	-0,01038	-0,2081	0,03716	-0,05343	-0,00508	0,02721
PREMIER	1,0433	0,07196	-0,00441	0,04318	0,02093	0,01321	-0,04316	-0,02327	-0,00409
ÉTAT	1,1729	0,06931	-0,01054	0,01149	-0,02843	-0,04254	0,04541	-0,00364	0,06886
AGIR	1,0195	0,06924	-0,04286	-0,00131	0,01458	0,01784	0,02155	0,03292	-0,05989
TENIR	0,9479	0,06826	-0,02351	0,01019	0,01512	0,02328	0,0075	-0,0155	0,02125
OBLIGATION	1,5728	0,06824	0,06219	0,16363	0,06328	-0,11512	-0,04816	-0,02172	0,07227
FORCE	1,343	0,06803	0,02263	0,03385	0,05041	0,04901	-0,05101	-0,0438	-0,08015
ESPRIT	1,1783	0,06786	-0,05819	0,00811	-0,03555	0,08719	-0,03545	-0,02542	-0,08139
DONNER	0,9021	0,06628	0,01213	-0,04224	0,01359	-0,05559	-0,0125	-0,05194	-0,02331
SENS	1,0649	0,06605	-0,03937	0,0102	0,01724	0,0223	0,03654	-0,05726	-0,02827
SEUL	0,8269	0,06598	-0,00232	-0,02566	0,0099	0,03149	-0,00768	0,0366	0,02285
ACTION	1,0702	0,06508	-0,02875	-0,02681	0,02722	0,02218	0,02344	0,02392	0,04458
MOT	1,3194	0,06488	-0,07735	0,02242	-0,01621	-0,0145	0,07155	-0,0943	-0,07615
DEVENIR	1,1857	0,06407	-0,03152	0,0447	-0,02657	-0,02253	0,01918	0,00912	0,00556

N.B.: Le score d'**importance** de chaque lemme est calculé en additionnant les valeurs absolues de ses 20 premières coordonnées (c.-à-d. les vecteurs propres), chacune étant multipliée par la valeur propre correspondante.

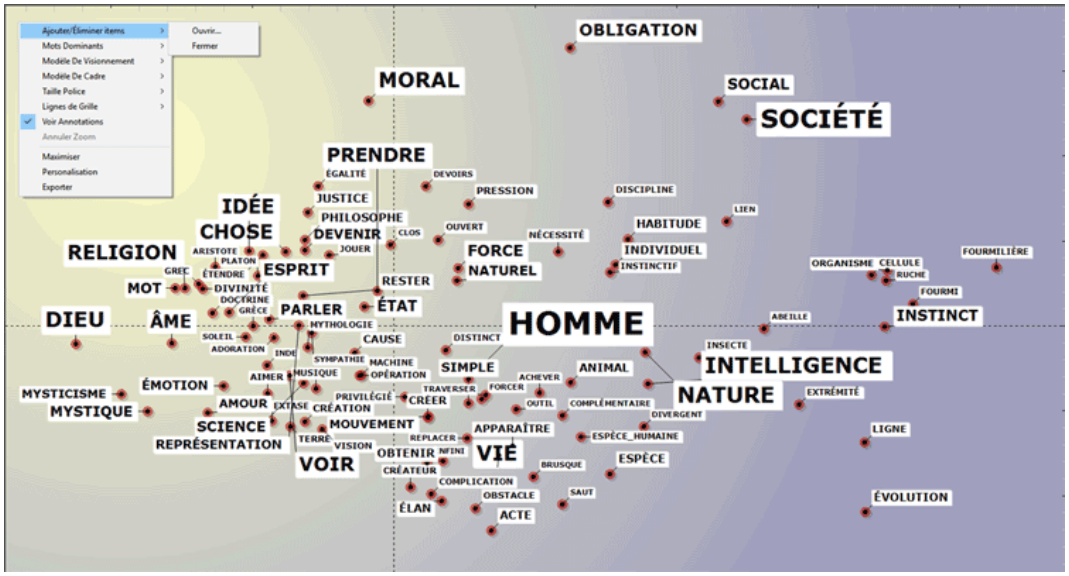


Tous les tableaux peuvent être triés par ordre croissant ou décroissant en cliquant sur un en-tête de colonne quelconque.
Pour **exporter** n'importe quelle tableau, utilisez simplement le clic droit de la souris lorsque les données sont affichées.
Veuillez noter que la première fois qu'un tableau est exporté, les valeurs propres sont également exportées. De cette façon, l'utilisateur peut évaluer le poids relatif de chaque dimension, c'est-à-dire le pourcentage de variance expliqué.

En cliquant sur le bouton **Associations**, une autre tableau s'affiche avec les mesures de similarité (c.-à-d. le cosinus) de chaque mot-clé. De plus, lorsque vous cliquez sur une ligne quelconque d'un tableau, un graphique s'affiche avec les données correspondantes.



Le **graphiques** principal montre les relations entre les mots-clés (c.-à-d. les lemmes) sur les dimensions sélectionnées (voir ci-dessous).



Par défaut, le tableau ci-dessus comprend les 100 lemmes les plus importants. Cependant, l'utilisateur peut personnaliser le nombre de lemmes et les caractéristiques du graphique.

PREPARATION DU CORPUS

Préparation du Corpus

Dans le cas de textes uniques (ou corpus considéré comme texte unique) on n'a pas besoin d' autre travail : il vous suffit de sélectionner l'option 'Importer un fichier unique..' (voir la section correspondant du manuel).

Autrement, si le corpus se compose de plusieurs documents primaires codifiés (**variables et modalités**), dans la phase de préparation on doit utiliser l'outil **Corpus Builder**, qui transforme automatiquement tout matériel textuel et divers types de fichiers (c.-à-d. jusqu'à dix formats différents) dans un fichier corpus prêt à être importé par **T-LAB.**

N.B. :
- dans tous les cas, nous conseillons un examen orthographique du matériel à analyser. D'ailleurs, si quelques acronymes importants sont ponctués (par exemple "N.U.") il est recommandé de les transformer en chaînes unitaires (par exemple "NU" ou "N_U"); ceci parce que, dans la phase de **normalisation**, **T-LAB** interprète les signes de ponctuation comme des séparateurs ;
- au terme de la phase de préparation on recommande de créer un nouveau dossier de travail avec à l'intérieur le fichier corpus à importer.

Critères structuraux

Les **critères structuraux** à respecter concernent la **taille** du **corpus** et sa subdivision en **parties**.

Quant à la taille, tous les outils **T-LAB** ont été testés avec un corpus 90Mo, correspondant à environ 55.000 **pages** de format .txt.

Les limites pour la **taille minimum** requièrent différents critères d'évaluation, parce que, sous un certain seuil, la taille du corpus peut compromettre la fiabilité de beaucoup d'analyses statistiques. À ce propos il suffit de suivre deux instructions: employer les corpus avec au moins 5.000 occurrences (approximativement 30 Ko); autrement, dans le cas des questions ouvertes, 50 réponses au minimum. En effet, dans ce dernier cas, chaque réponse constitue une différente unité de contexte.

Afin d'être traité, le corpus peut se composer de: un texte unique sans d'autres partitions; un texte subdivisé selon des critères établis par l'utilisateur (par exemple, un livre divisé en chapitres); un certain nombre de textes (par exemple, différents entretiens ou différents documents) classifiés par l'usage de **variables** ou **IDnumber**. Dans ces derniers cas, le corpus est subdivisé en parties qui doivent être codées par des **critères formels** précis.

Critères formels

Dans le cas d'un **corpus** composé d'un texte unique, et quand l'utilisateur ne fait pas recours aux **variables**, **il n'y a aucune autre opération à faire** et on peut procéder directement à l'**importation**.

Quand le corpus est constitué par **plusieurs textes** et/ou bien on utilise des **variables**, la préparation du corpus doit être réalisée par le module Corpus Builder qui, de façon automatique, respecte les critères suivants :

Chaque texte ou sous-ensemble (les "parties" définies par des variables et/ou IDnumber) doit être précédé par **une ligne de codage**.

Chaque ligne de codage a ce format:

- elle **commence** par **quatre astérisques** (****), suivis d'un espace blanc. **T-LAB** lit cette séquence ainsi: "ici commence un texte ou une unité de contexte défini par l'utilisateur";
- elle **continue** avec des chaînes composées par des **astérisques simples** et des étiquettes qui définissent les **sujets** (IDnumber), les **variables** et les **modalités** respectives;
- elle **finit** avec le retour à la ligne.

Voici quelques exemples.

La ligne suivante introduit un texte (ou un sous-ensemble du corpus) codifié avec trois variables - AGE, SEXE et MET (métier) - avec les respectives modalités (adul, fem, prof).

**** *AGE_adul *SEX_fem *MET_prof

La ligne suivante introduit un texte (ou un sous-ensemble du corpus) codifié avec les mêmes variables et l'etiquette **IDnumber**.

**** *IDnumber_0001 *AGE_adul *SEX_fem *MET_prof

La ligne suivante introduit un texte (ou un sous-ensemble du corpus) codifié avec deux variables: ANN (année) e MAG (magazine)

**** *ANN_98 *MAG_times

Dans **chaque ligne de codage**, les règles à respecter sont les suivantes :

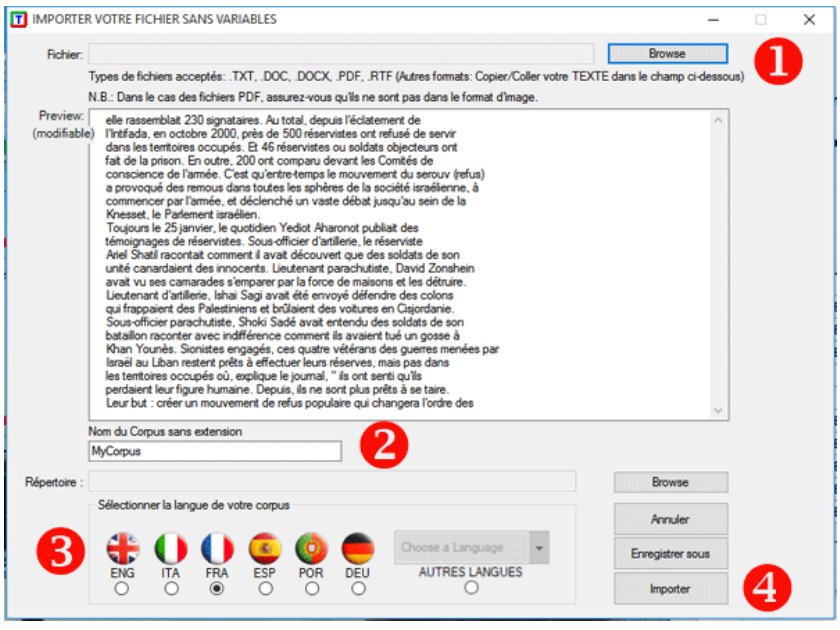
- chaque étiquette (IDnumber, variable, modalité) ne peut être entrecoupée par des espaces blancs.
- chaque étiquette - soit pour des variables, soit pour les modalités - ne peut être plus longue de 25 caractères (min. 2).
- chaque étiquette de variable doit être liée à la respective modalité avec le tiret bas ("_").
- entre deux différentes variables, c.-à-d. avant l'astérisque suivant, un espace blanc doit être inséré.
- chaque variable - avec les respectives modalités - doit être assignée pour chaque sous-ensemble du corpus.
- le variables utilisables sont maximum 50, chacune avec un maximum de 150 modalités.
- le numéro maximum d'IDnumber est fixé à 99.999 pour les textes brefs (Max. 2.000 caractères chacun, ex. réponses à questions ouvertes) et à 30.000 pour les autres cas.

FICHIER

Importer un fichier unique …

Dans le cas de textes uniques (ou corpus considéré comme texte unique) on n'a pas besoin d' autre travail : il vous suffit de sélectionner l'option 'Importer un fichier unique..' (voir l'image ci-dessous).

Ensuite, quatre étapes sont nécessaires (voir l'image ci-dessous) : (1) sélectionner un fichier ; (2) choisir le nom du projet ; (3) sélectionner la langue de votre texte; (4) cliquer sur "Importer" .



Ensuite une fenêtre apparaît (voir ci-dessous) dans laquelle l'utilisateur peut faire ses choix.

N.B.:
- Puisque les options de prétraitement déterminent le type et la quantité d'unités d'analyse (c.-à-d. des unités de contexte et des unités lexicales), les différents choix de l'utilisateur déterminent différents résultats de l'analyse. Pour cette raison, tous les outputs de **T-LAB** (c.-à-d. graphiques et tableaux) montrés dans le manuel et dans l'aide en ligne sont simplement indicatifs ;
- Toutes les étapes du prétraitement sont effectuées lors de l'importation de tout type de corpus.

1 - LEMMATISATION AUTOMATIQUE OU STEMMING

De suite la liste complète des trente langues pour lesquelles la lemmatisation automatique ou bien le processus de stemming sont supportés par **T-LAB**.

LEMMATISATION: allemand, anglais, catalan, croate, espagnol, français, italien, latín, polonais, portugais, roumain, russe, serbe, slovaque, suédois, ukrainien.

STEMMING: arabe, bengali, bulgare, danois, hollandais, finlandais, grec, hindi, hongrois, indonésien, marathi, norvégien, persan, tchèque, turc.

En tout les cas, sans lemmatisation automatique et / ou en utilisant des dictionnaires personnalisés, l'utilisateur peut analyser textes dans toutes les langues, à condition que les mots soient séparés par des espaces et/ou des signes de ponctuation.

Le résultat du processus de lemmatisation peut être vérifié avec la fonction **Vocabulaire** et peut être modifié avec la fonction **Personnalisation du Dictionnaire**.

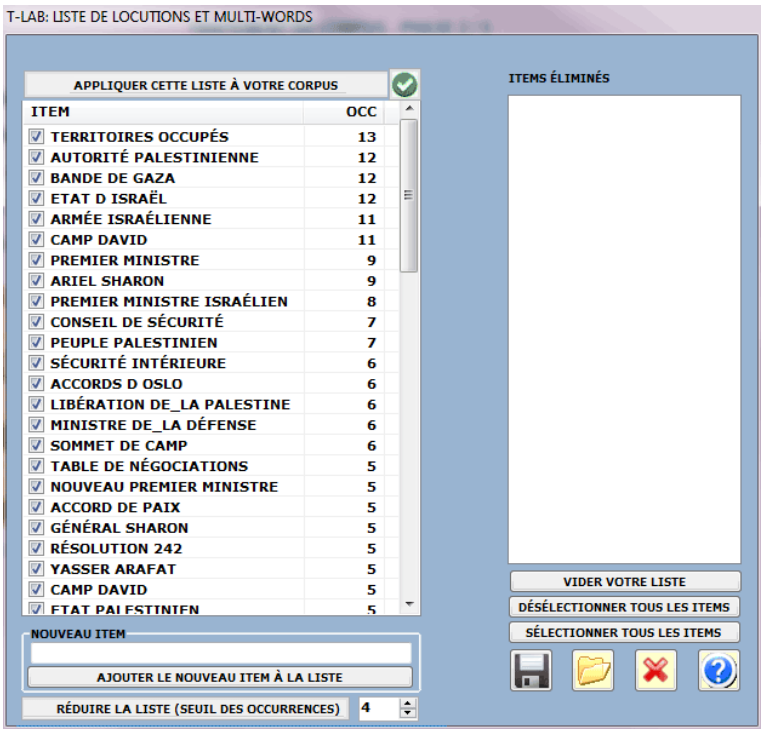
2 - SEGMENTATION DES TEXTES (CONTEXTES ÉLÉMENTAIRES)

Selon le choix de l'utilisateur, les types de **contextes élémentaires** utilisés pour le calcul des **co-occurrences** peuvent être les suivants: énoncés, fragments de longueur comparable, paragraphes ou textes courts (ex. réponses aux questions ouvertes).

Le fichier corpus_segments.dat permet à l'utilisateur de vérifier le résultat de la segmentation du corpus.

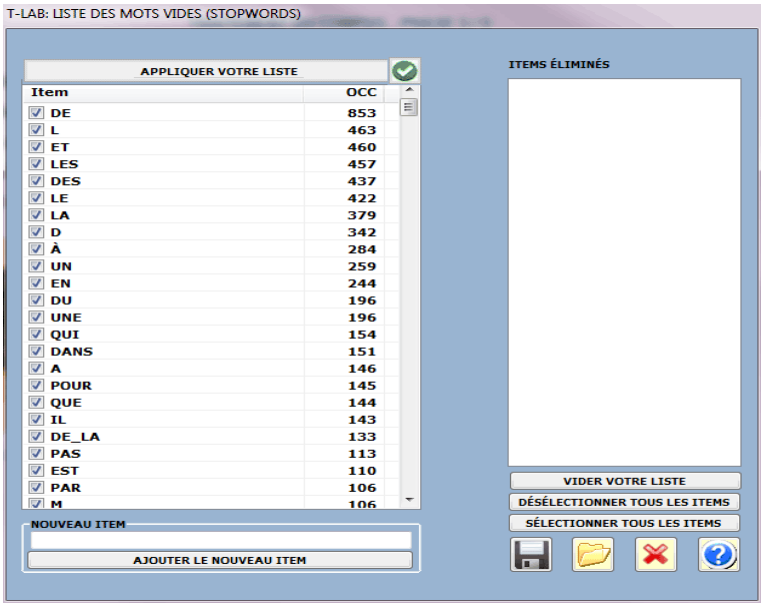
3 - EXAMEN DES MULTIWORDS

L'option "**Élémentaire**" active l'utilisation automatique de la liste **Multi-Words** de T-LAB. Différemment l'option "**Avancé**", habilitée seulement avec la lemmatisation automatique, permet à l'utilisateur de vérifier et de modifier la liste des Multi-Words non inclus dans le dictionnaire de T-LAB.
Il est aussi possible d'importer et d'employer d'**autres fichiers** Multiwords.txt.



4 - EXAMEN DES STOPWORDS

L'option "**Élémentaire**" active l'utilisation automatique de la liste **Stop-Words** de T-LAB. Différemment l'option "**Avancé**" permet à l'utilisateur de vérifier et modifier la liste des Stop-Words présentes dans le corpus à analyser.
Il est aussi possible d'importer et d'employer **autres fichiers** StopWords.txt.



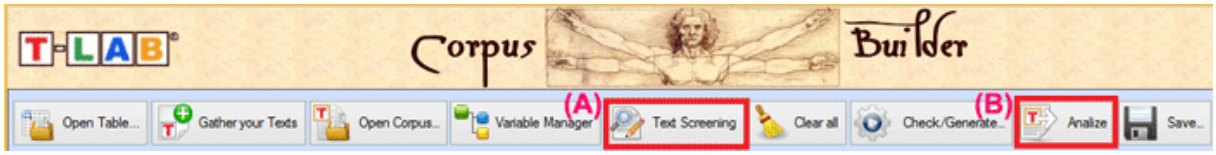
5 - SÉLECTION DES MOTS-CLÉS

Les options disponibles nous permettent de choisir la méthode de choix (**TF-IDF** ou **Chi-deux**) et la quantité maximum d'**unités lexicales** à inclure dans une liste employée par **T-LAB** pour analyser les textes avec les **configurations automatiques**.

N.B.: Lorsque la phase d'importation est terminée, en utilisant la **configuration personnalisée**, l'utilisateur peut vérifier la sélection des mots-clés et créer des **listes** différentes.

Préparer un Corpus (Corpus Builder)

N.B.: En **T-LAB 10**, cet outil comprend deux boutons supplémentaires: a) un qui, pour des corpus de dimensions non supérieures à 20 MB, active l' option **Text Screening**; b) l'autre qui permet de procéder immédiatement à l'**importation** des matériaux textuels sélectionnés (voir l'image ci- dessous).



Cet instrument logiciel a été projeté pour faciliter la préparation et la transformation de divers matériaux textuels dans un fichier **corpus** prêt à être importé par **T-LAB**. Plus spécifiquement, cet instrument permet d'exécuter rapidement les opérations suivantes:

.

1. **Importer** automatiquement divers types de fichiers;
2. **Éditer** et modifier les textes;
3. Gérer l'emploi de **variables catégorielles**;
4. **Sauver** le résultat du travail dans un fichier prêt à être importé par **T-LAB**;
5. **Vérifier et modifier** n'importe quel fichier corpus qui corresponde au format requis par **T-LAB**.

Pendant que la façon d'importer les fichiers (voir au-dessus ‘ 1’), se diffère selon leur format, toutes les autres opérations suivent la même logique.

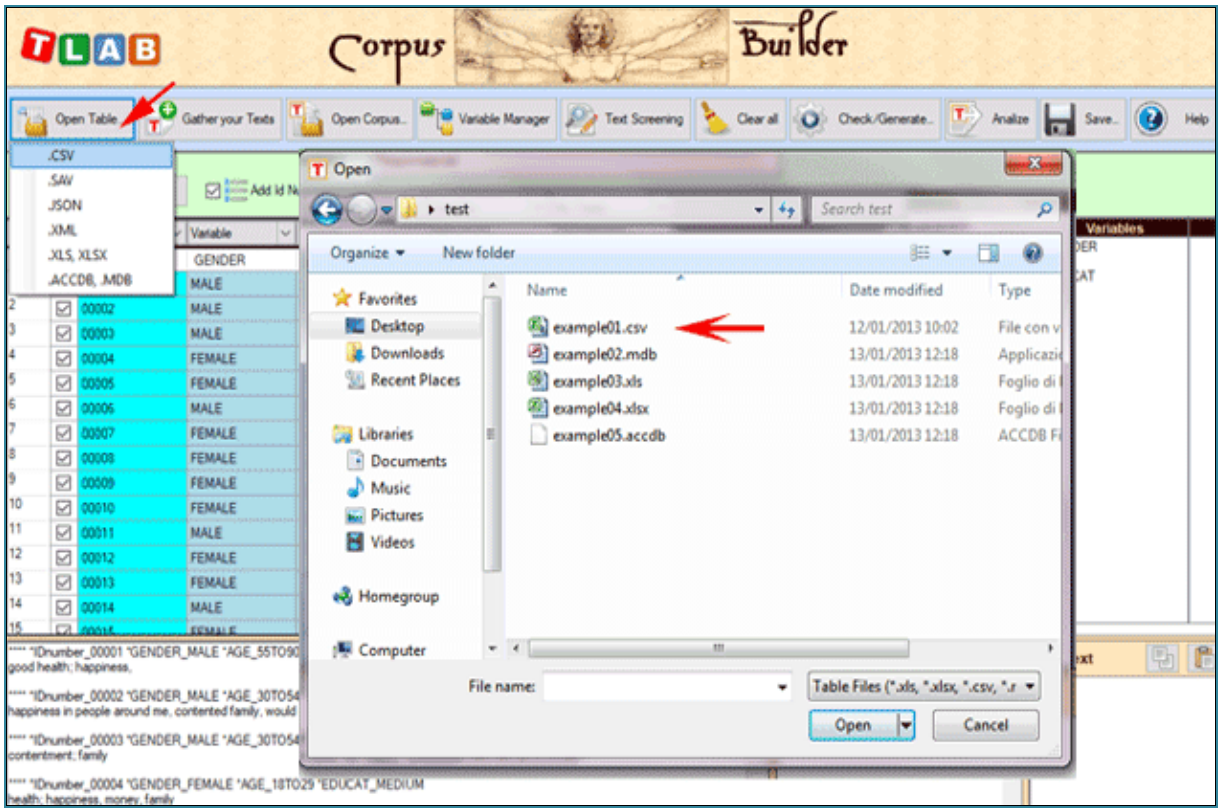
De suite une brève description des façons pour importer les différents types de fichiers.

A - **Importation de** fichiers **en format tabulaire** (CSV, .SAV, .JSON, .XML, .XLS, XLSX, .MDB, .ACCDB).

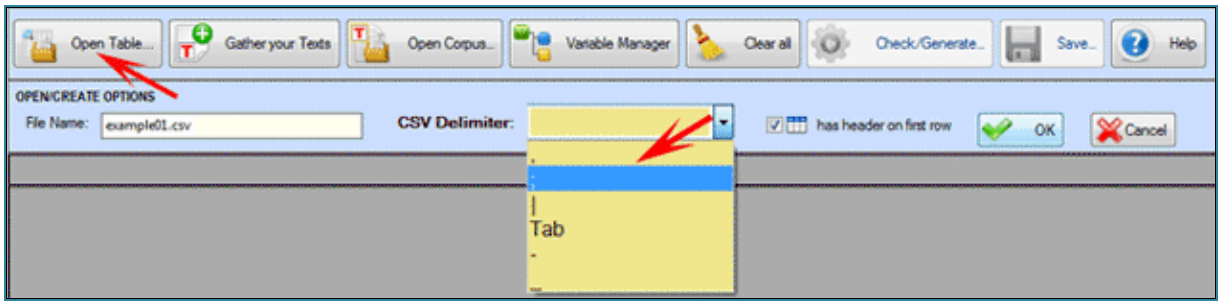
Un **seul fichier** qui inclue jusqu'à 30.000 records peut être importé en utilisant l'option “Open Table” ou bien par l’option drag and drop (NB: quand aucun des textes dépasse les 2.000 caractères, la limite des records à importer est étendue à 99.999).

Ce fichier peut être constitué de différentes colonnes contenant les données suivantes:
- variables catégorielles (une pour chaque colonne, jusqu'à un maximum de 50);
- textes à analyser (une seule colonne);
- IDnumbers, c'est-à-dire identificateurs des unités de contexte ou des cas.

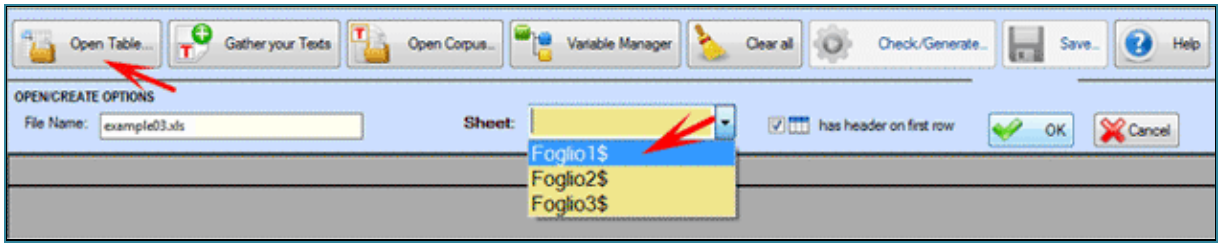
N.B. : Tandis que la présence de variables catégorielles et celle des IDnumbers est facultative, la présence d'au moins une colonne contenant les textes à analyser est obligatoire.



Quand un fichier .CSV est importé, on doit opportunément sélectionner le délimiteur employé (voir ci- dessous).



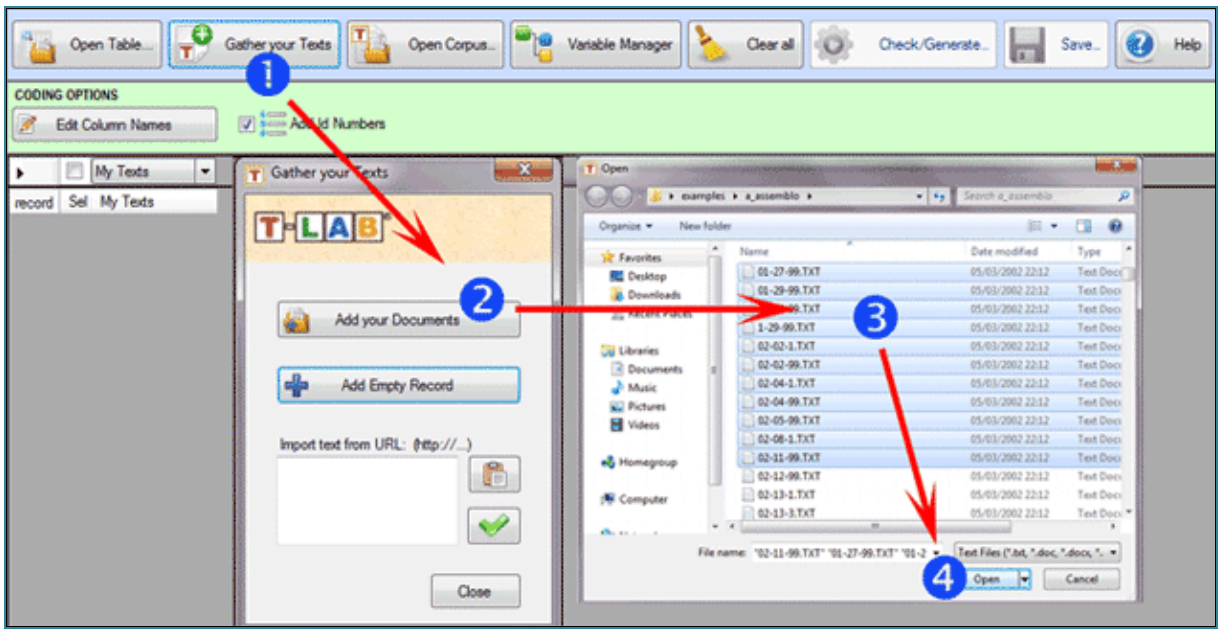
Quand des fichiers file Excel ou f Access sont importés, on peut sélectionner seulement un tableau (voir ci-dessous).



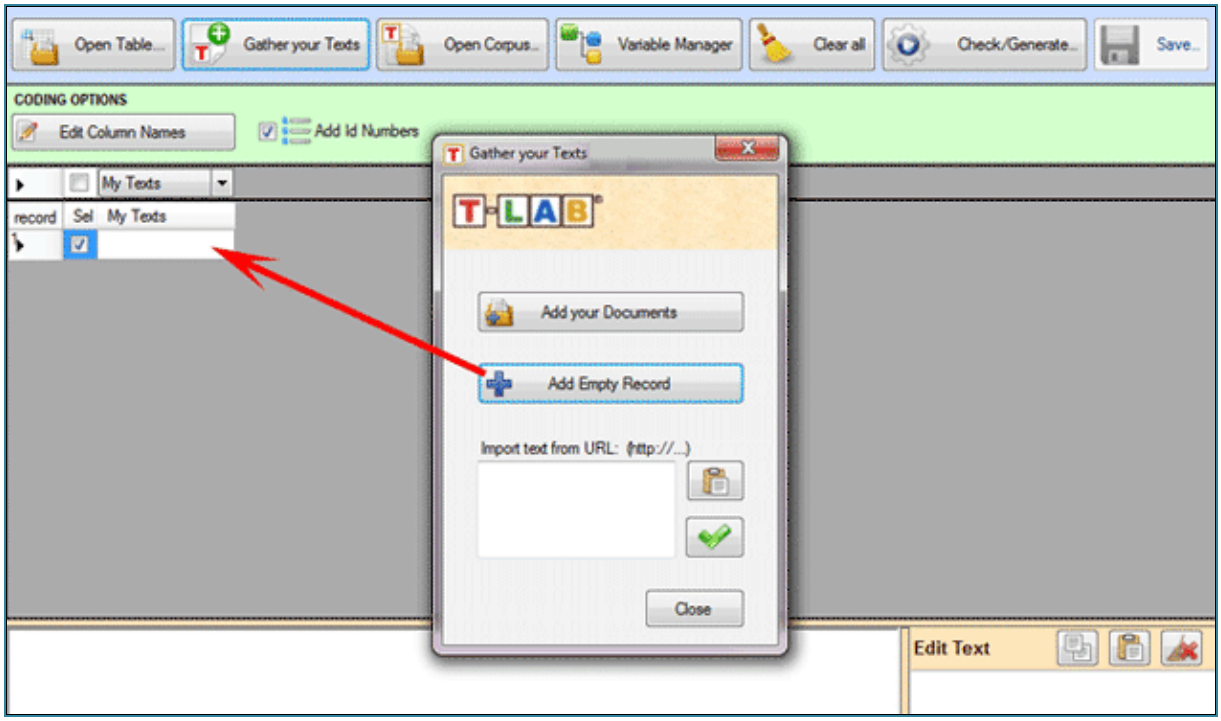
B - **Importation textes et documents**

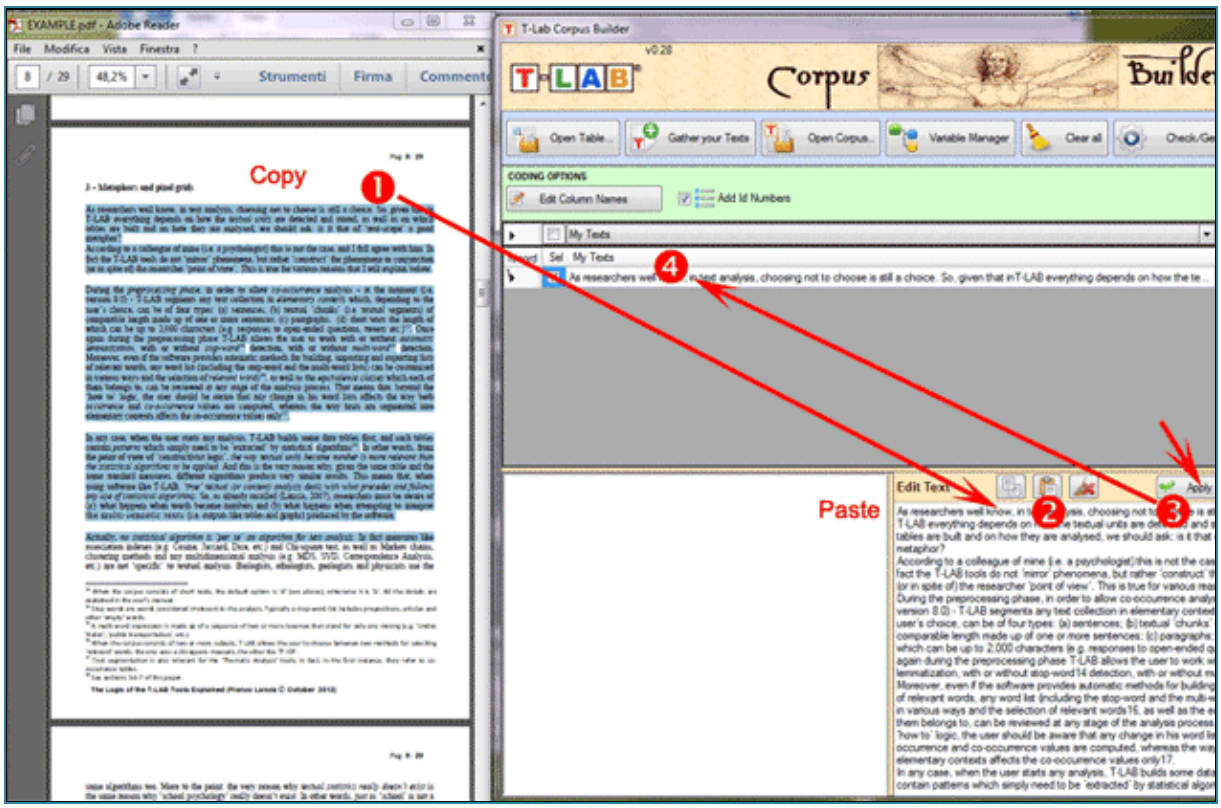
L'option “Gather your Texts” (voir ci-dessous) permet d'importer jusqu'à 30.000 documents, aussi bien un à la fois que par sélection multiple, avec **trois méthodes différentes**.

La **première méthode** (‘Add your Documents’) prévoit l'importation automatique de type de fichier .TXT, .DOC, .DOCX, .PDF, .RTF.



La **deuxième méthode** (‘Add EmptyRecord’) vous permet d'ajouter des enregistrements où vous pouvez copier/coller un texte (voir ci-dessous).





La **troisième méthode** (‘Import Text from URL’) vous permet de télécharger directement des fichiers HTML à partir d'internet, éditer le contenu pour d'éventuelles modifications et - ensuite - les importer (voir ci-dessous).

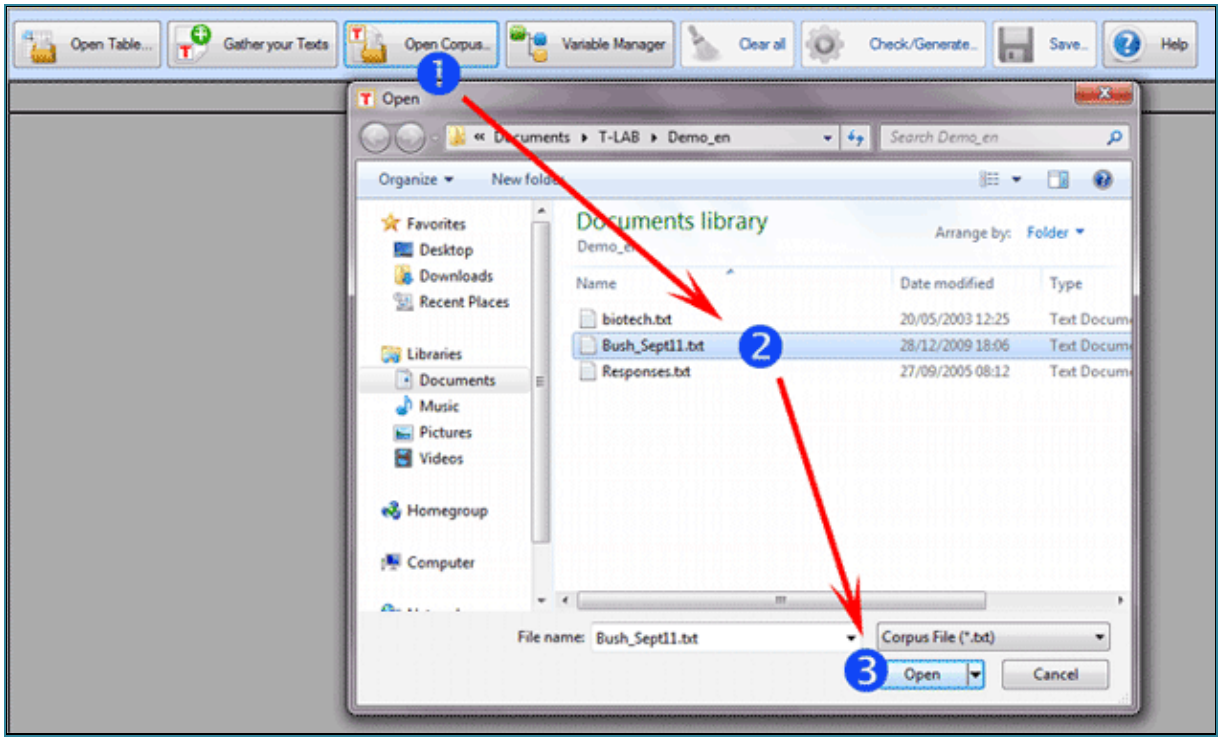
C - **Importation d'un corpus** déjà codifié selon les spécifiques de **T-LAB.**

Il est recommandé d'utiliser l'option 'Open Corpus' dans les trois cas suivants:

1 - l'utilisateur a l'intention de modifier la structure d'un fichier corpus déjà codifié (p. ex. , ajouter d'autres textes par les méthodes expliquées dans la section précédente "B", modifier les noms de variables et/ou de modalités, etc.)
2 - l'utilisateur a l'intention de vérifier/corriger les erreurs contenues dans un codage du corpus effectué manuellement et sans l'aide du module Corpus Builder;
3 - l'utilisateur a l'intention d'importer un fichier corpus avec un codage "brut" (voir l'image ci-dessous), c'est-à-dire un fichier corpus dont les pièces (documents ou fichiers) sont toutes précédées par une ligne avec quatre astérisques suivis d'un espace ('**** ').

```
**** ¶
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
```

Dans les trois cas mentionnés ci-dessus (1,2,3) il est suffisant de sélectionner un fichier individuel à travers l'option “Open Corpus” ou bien de le traîner avec la méthode drag and drop (voir ci-dessous).

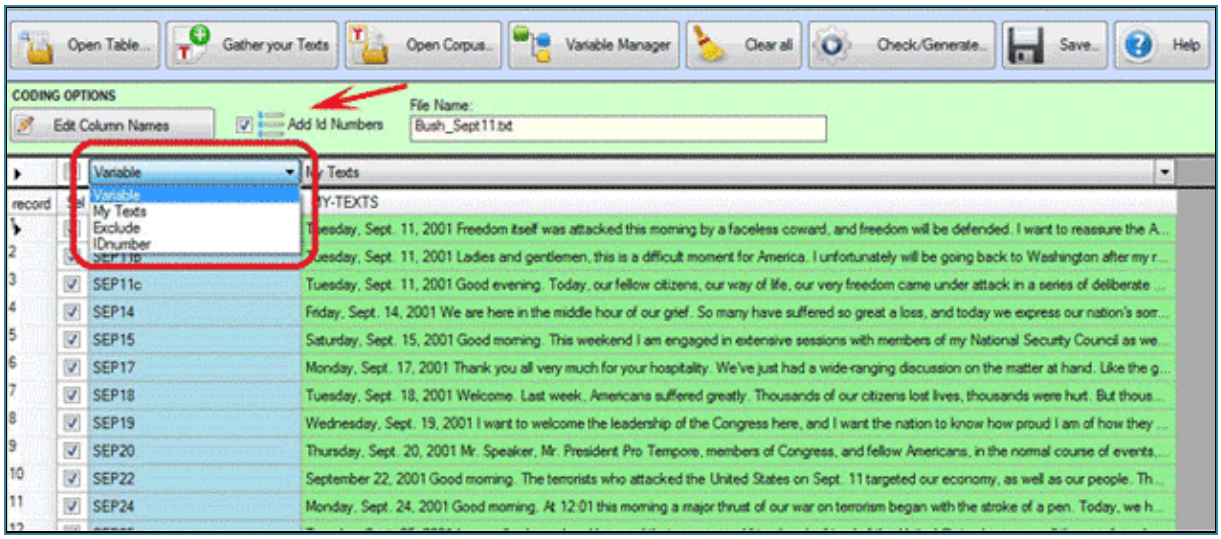


Opérations succesives à l'importation des fichiers

Après avoir importé les fichiers dans Corpus Builder, aussi bien dans le cas où on n'est pas intéressé à l'utilisation de variables, que dans le cas où les opérations de codage ont été effectuées, vous pouvez passer à l'option 'Check/Generate' et - après - à l'exportation du corpus à être importé par **T-LAB**.

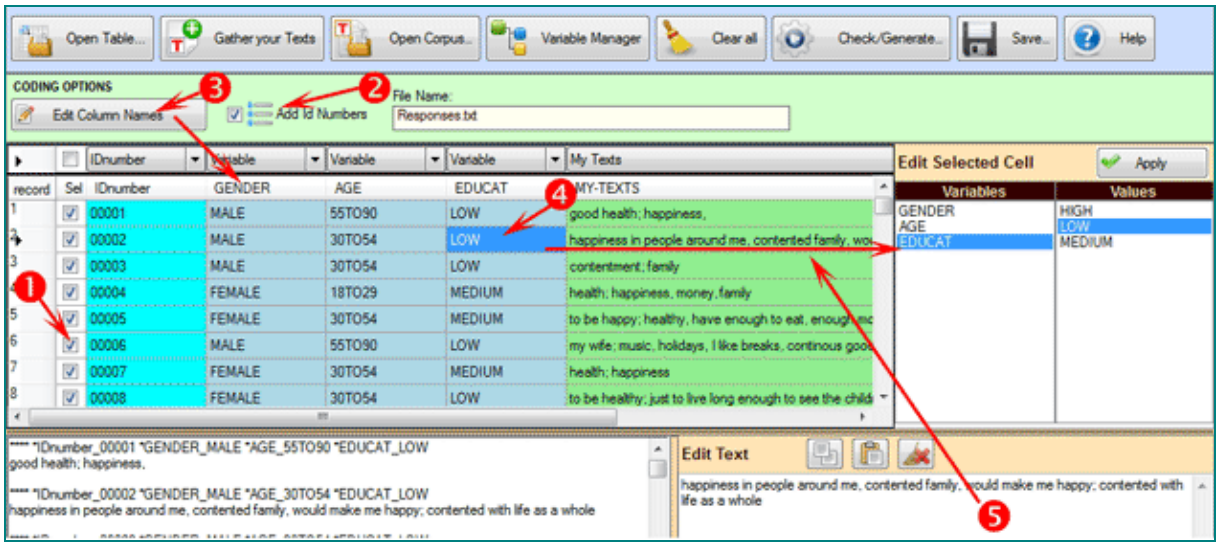
Lorsque le corpus contient des codages, il convient de rappeler que, dans les trois types d'importation mentionnés dans les sections précédentes de ce document (A', 'B', 'C'), les données sont visualisées en différentes colonnes, dont les en-têtes peuvent être les suivantes:

- 'Variable', (c'est-à-dire variables catégorielles), dont l'utilisation est nécessaire lorsque vous avez l'intention d'analyser les caractéristiques et les relations mutuelles de sous-ensembles du corpus;
- 'IDnumber' (c'est-à-dire identificateurs de cas/record), dont l'utilisation est facultative;
- 'My Texts', (c'est-à-dire les textes à analyser), dont l'utilisation est possible dans une colonne seulement et elle est obligatoire;
- 'Exclude', à être utilisé pour indiquer au Corpus Builder que les données contenues dans la colonne correspondante ne doivent pas être utilisées.



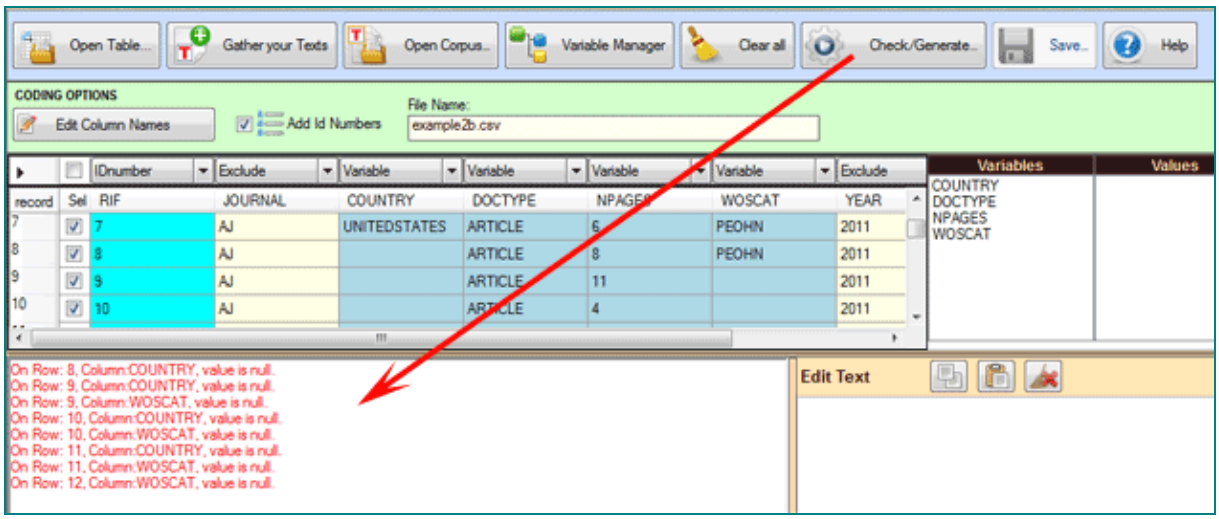
Les indications suivantes sont valables dans **tous les cas**:

- chaque record peut être sélectionné ou désélectionné (voir ci-dessous '1 ');
- les Idnumber peuvent être ajoutés automatiquement (voir ci-dessous '2');
- les noms des variables peuvent être édités et modifiés (voir ci-dessous '3');
- chaque valeur de variable peut être éditée et modifiée (voir ci-dessous '4');
- chaque champ "My Text" peur être édité et modifié (voir ci-dessous '5').



Il faut rappeler que:

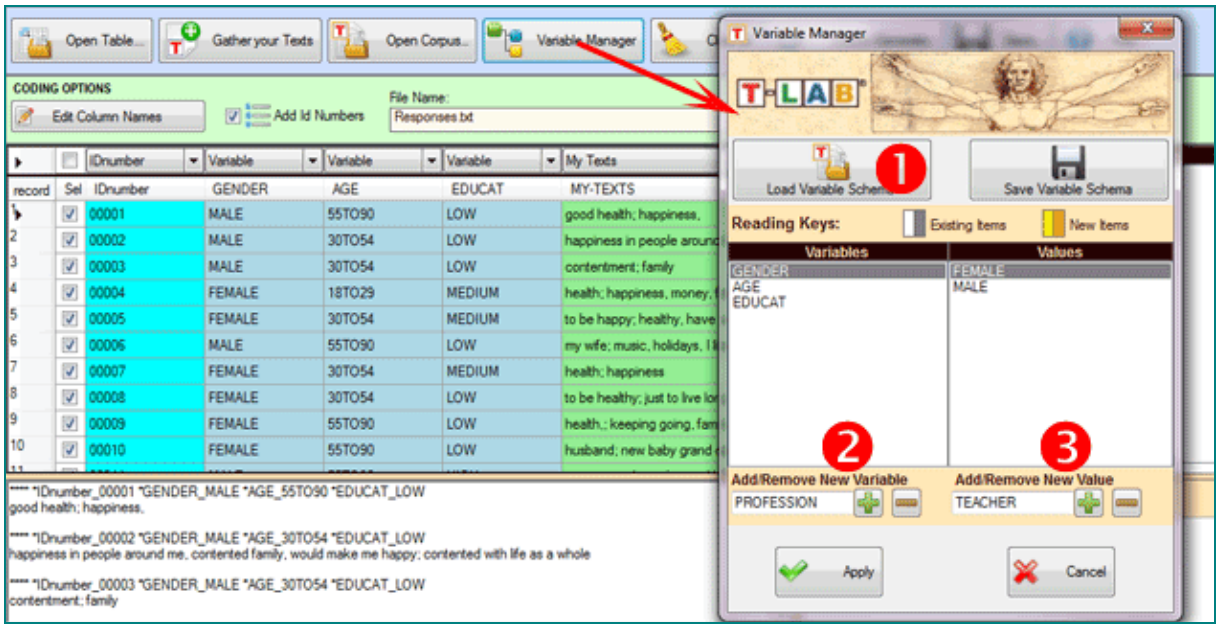
- Le numéro des colonnes avec des variables catégorielles ne doit pas dépasser les 50 et chacune d'elles doit avoir un minimum de 2 à un maximum de 150 valeurs;
- Les valeurs des IDnumber, si utilisées, doivent être progressives à partir de 1 (es., 1, 2, 3, etc.);
- Chaque étiquette - soit pour des variables, soit pour les modalités - ne peut être plus longue de 25 caractères (min. 2) et ne doit pas contenir espaces blancs;
- Toutes les fautes relevées par le logiciel sont visualisées dans la fenêtre en bas à gauche (voir ci-dessous).



Utilisation de l'outil Variable Manager

L'instrument "Variable Manager" permet d'éditer, de modifier et de sauver n'importe quel schéma de codage, provenant même d'un corpus différent (voir ci-dessous).

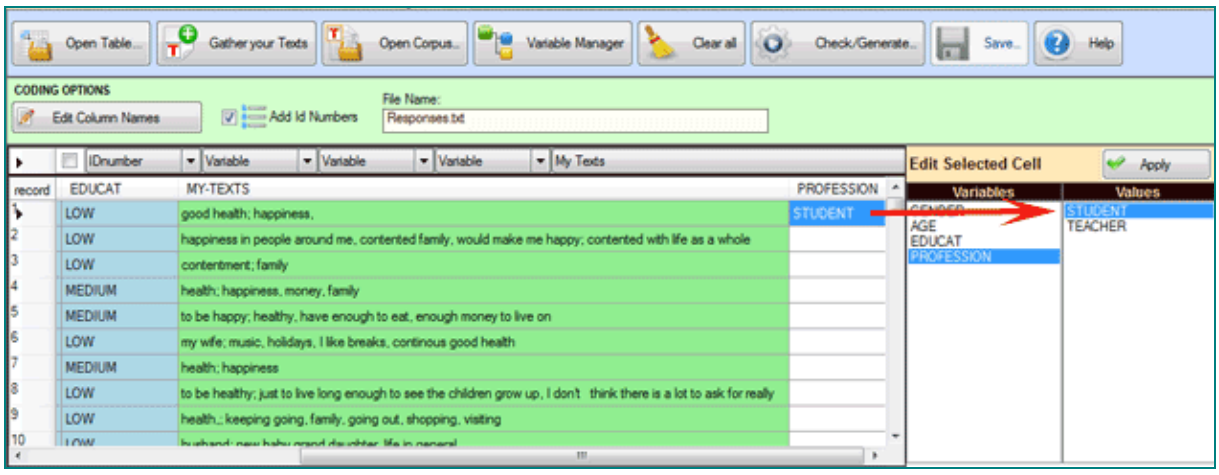
Chaque schéma inclut la liste des variables et leurs valeurs (voir ci-dessous).



Pour ajouter des variables provenant d'un autre corpus ou d'un schéma précédemment enregistré, vous devez sélectionner l'option '1' (voir ci-dessus). Differement, pour ajouter manuellement les variables et leurs valeurs, vous devez utiliser dans l'ordre les options '2' et ' 3' (voir ci-dessus).

L'ajout de valeurs des variables aux enregistrements individuels est à faire manuellement (voir ci-dessous) et en une seule session de travail; ceci parce que le sauvage du schéma n'inclut pas les codages attribués à chaque enregistrement. Par conséquent, dans le cas où l'utilisateur a l'intention de coder manuellement un corpus qui comprend un nombre considérable de documents et/ou il nécessite plus d'une session de travail, il est recommandé de procéder comme suit:
1 - importer la quantité de fichiers qu'on considère possible coder en une seule session de travail ;
2 - enregistrer le travail accompli comme un corpus (voir l'option 'Save' du menu Corpus Builder).

Puis, dans une session suivante, réimporter le corpus enregistré précédemment (voir ci-dessus, point '2 '), ajouter d'autres enregistrements/fichiers qu'on souhaite codifier et continuer.

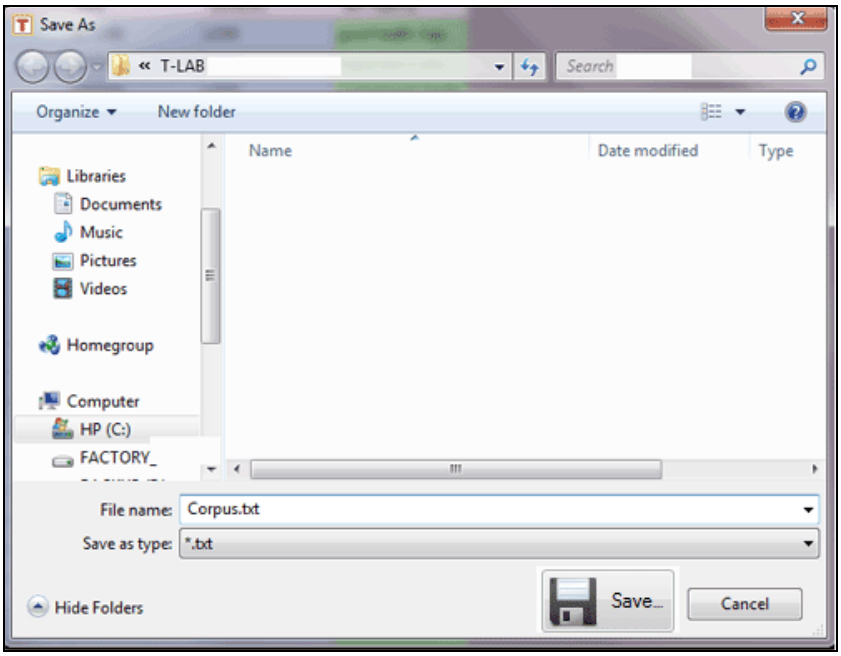


Lorsque l'utilisateur a terminé les opérations qu'il juge appropriées, l'option "Check/Generate" permet de vérifier leur exactitude et, si tout est ok, il est possible de exporter (A) ou sauver (B) un corpus qui est prêt à être importé de **T-LAB**.

Dans le premier cas (A - voir ci-dessous) Corpus Builder crée un nouveau dossier dans le répertoire ".. \ Mes Documents \ T-LAB PLUS\" et démarre automatiquement la procédure d'importation.
N.B.: Dans ce cas-ci, le nouveau dossier a le même nom du fichier corpus.

Dans le second cas (B - voir ci-dessous), l'utilisateur peut sauver son corpus dans le dossier qu'il souhaite et ensuite il doit utiliser la fonction "Importer un Corpus" de T-LAB.

N.B.: Dans ce cas-ci, il est recommandé de créer - chaque fois - un nouveau dossier de travail avec, en son intérieur, seulement le fichier corpus à importer.

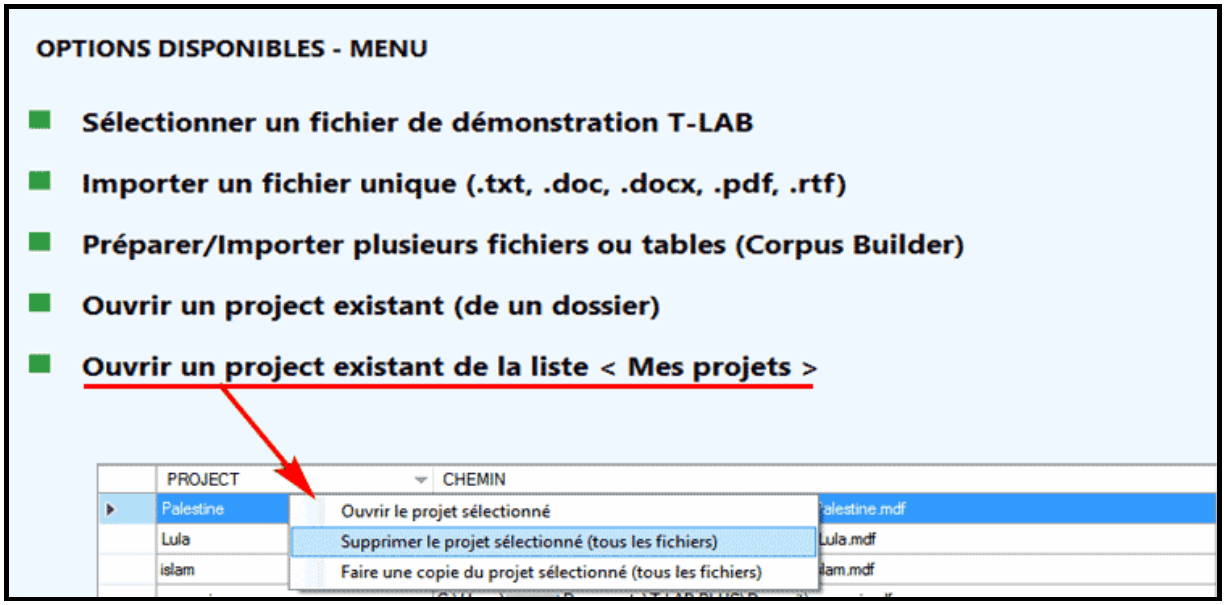


Ouvrir un Project Existant

À travers cette option l'utilisateur peut revenir travailler sur un projet déjà commencé, soit en sélectionnant le file d'un fichier existant ou bien d' une liste préparée par **T-LAB**.

En outre, quand on sélectionne un élément de la liste préparée par **T-LAB**, l'usage du bouton droit de la souris habilite l'utilisateur à éliminer les fichiers relatifs ou à en faire un sauvetage de secours dans un autre fichier.

.



OUTILS LEXIQUE

Text Screening /Désambiguïsations des Mots

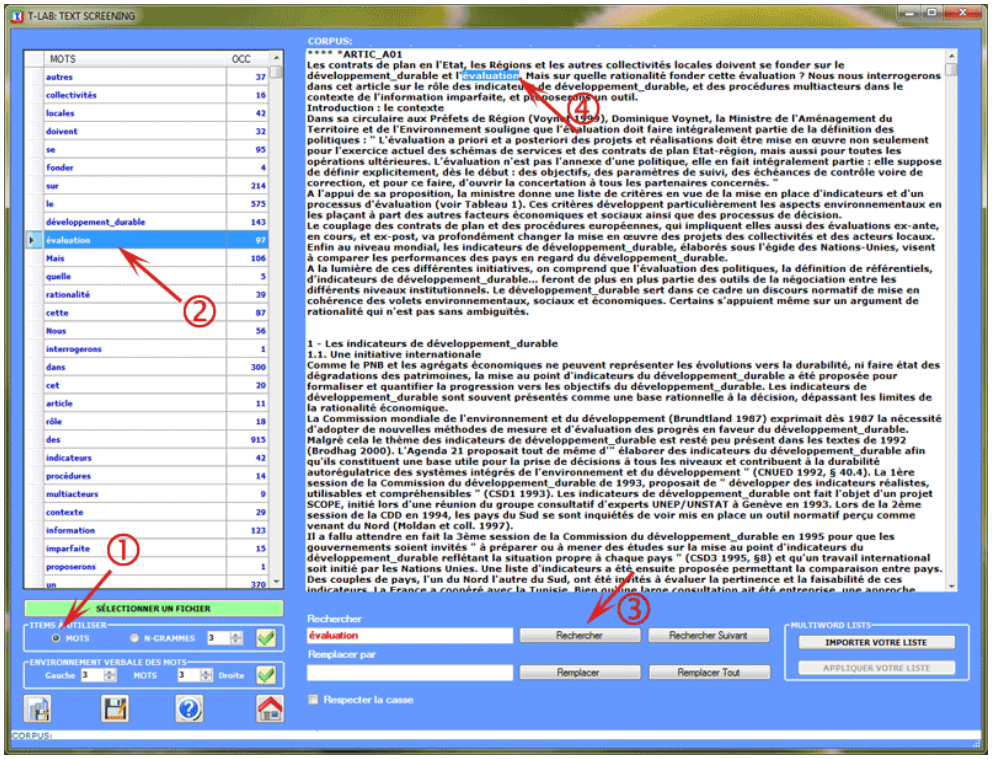
Cet outil **T-LAB** vous permet d'éditer n'importe quel fichier corpus (jusqu'à 30 Mb en taille) et d'effectuer **rapidement** une série d'opérations utiles aussi bien pour une première **exploration** de ses contenus que pour la **désambiguïsation** de ses unités lexicales spécifiques.

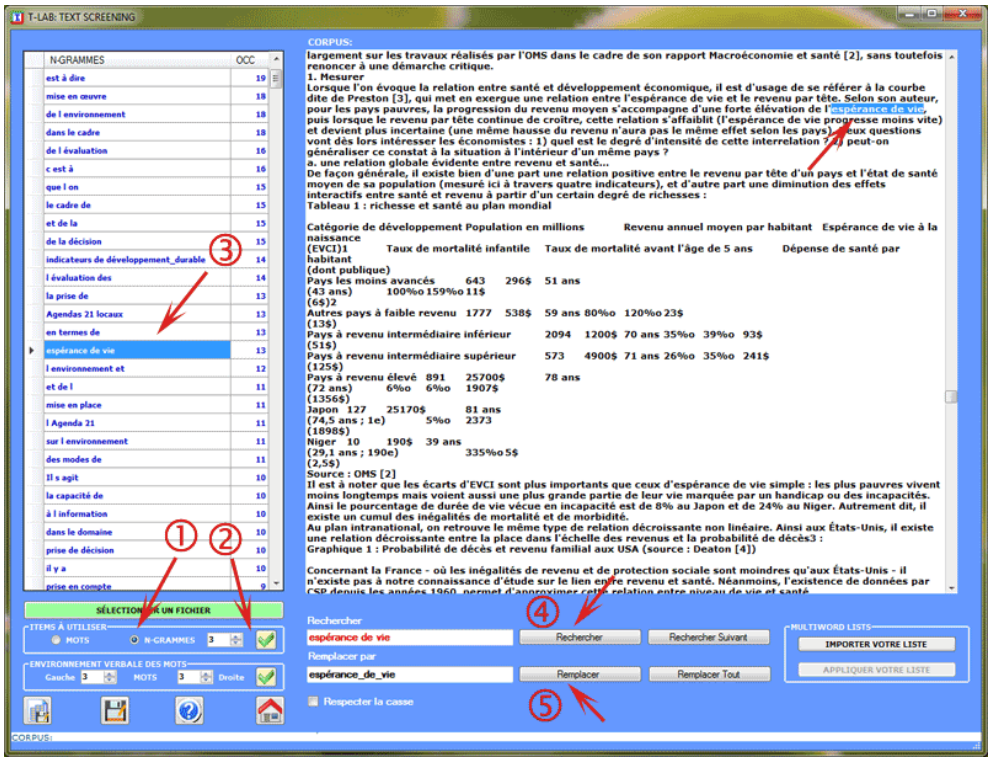
En particulier, cet outil produit rapidement une série de **listes** et permet des opérations du genre **recherche** / **remplace**.

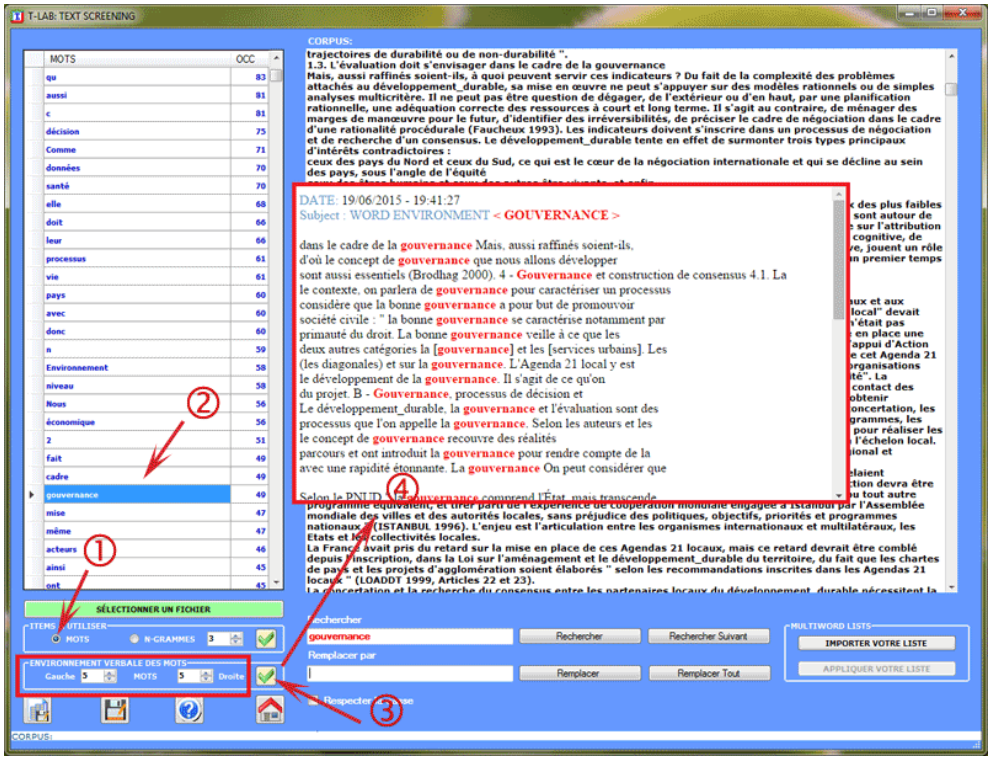
Les listes qui peuvent être obtenues sont les suivantes:

a- **mots** avec leurs occurrences;
b- **n-grammes** de mots avec leurs occurrences;
c- **environnements verbaux** des mots sélectionnés.

Les images ci-dessous montrent les opérations possibles dans les trois cas (a-b-c)

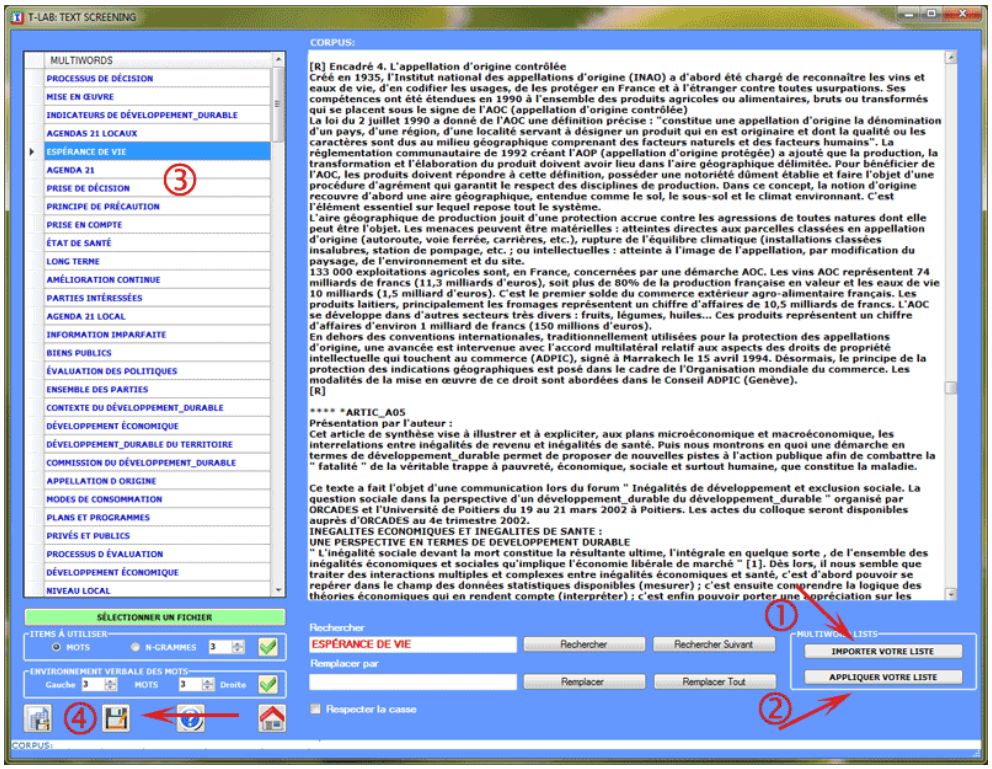






N.B.: Un clic sur le bouton en bas à gauche vous permet d'exporter les listes a-b en tant que fichiers Excel. Différemment, les listes 'c' sont automatiquement exportées en tant que fichiers .html.

On peut également importer des listes personnalisées de **Multiwords** et éventuellement les appliquer au corpus affiché (voir images ci-dessous).



À la fin des opérations, si l'utilisateur a modifié le texte et si il veut le sauver, **T-LAB** lui permet de créer un nouveau fichier (corpus_dis.txt) qui, renommé de façon appropriée, peut être importé et analysé (voir l'option 4 ci-dessous).

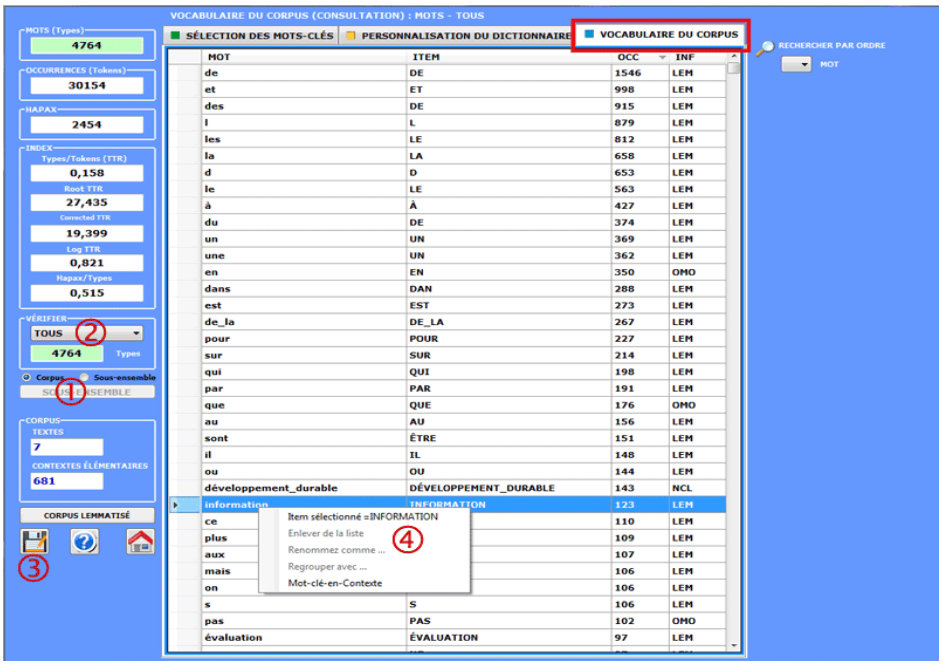
Vocabulaire du Corpus

Cet outil **T-LAB** nous permet de vérifier le vocabulaire du corpus de ses **sous-ensembles** (voir option "1" ci-dessous).;en outre, il nous fournit quelques mesures de la **richesse lexicale**.

Le tableau Vocabulaire est une liste comprenant tous les mots distincts (c.-à-d. "word types"), la quantité de leurs occurrences (c.-à-d. "word tokens"), leur lemmes correspondants et quelques catégories employées par **T-LAB** (voir le Glossaire/Lemmatisation).

L'utilisateur peut choisir (voir option "2" ci-dessous) les unités lexicales qui appartiennent à chaque catégorie, consulter le tableau correspondant et l'exporter comme fichier.xls. (voir option "3" ci-dessous).

En outre, en utilisant le bouton droit de la souris, il est possible de vérifier les concordances (Key-Word-in-Context) de chaque mot (voir option "4" ci-dessous).



Les mesures de richesse lexicale sont cinq:

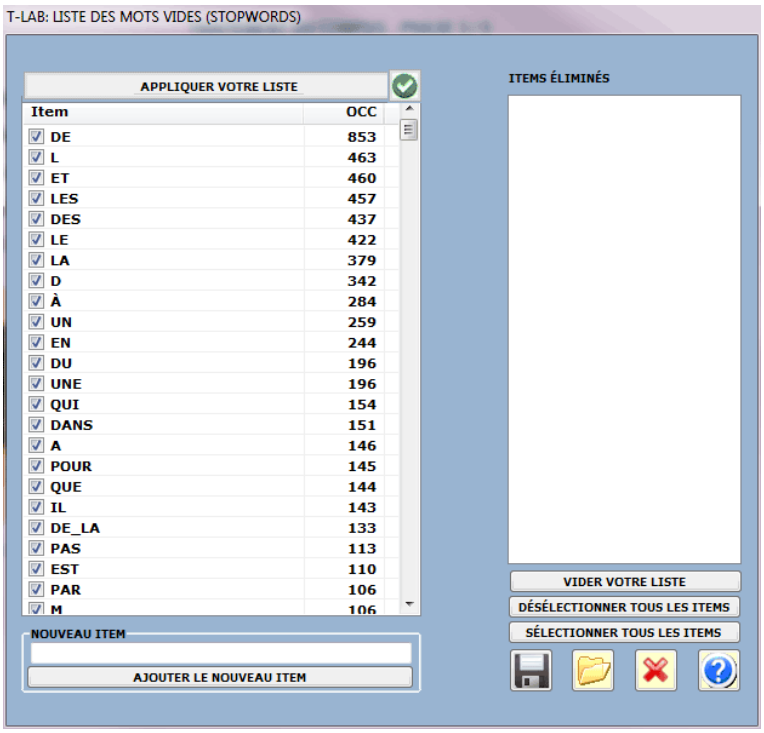
Type/Token ratio (TTR) ;
Root TTR (Guiraud, 1960), obtenue en divisant le nombre des "types" par la racine carrée du nombre des "tokens";
Corrected TTR (Carroll, 1964), obtenue en divisant le nombre des "types" par la racine carrée

de deux fois le nombre des "tokens";
Log TTR (Herdan, 1960), obtenue en divisant le logarithme du nombre des "types" par le logarithme du nombre des "tokens";
Hapax/Types ratio.

N.B.:
- Hapax (c.-à-d. Hapax Legomena) sont les mots utilisés une seule fois dans le corpus;
- quand on analyse un sous-ensemble de corpus, toutes les mesures de richesse lexicale n'incluent pas les mots vides (c.-à-d. articles, prépositions, etc.).

Mots Vides

Cette option permet de créer/modifier les fichiers des **Stop-Words** (Mots Vides).



Dans les fichiers StopWord.txt préparés par l'utilisateur les règles suivantes doivent être respectées:

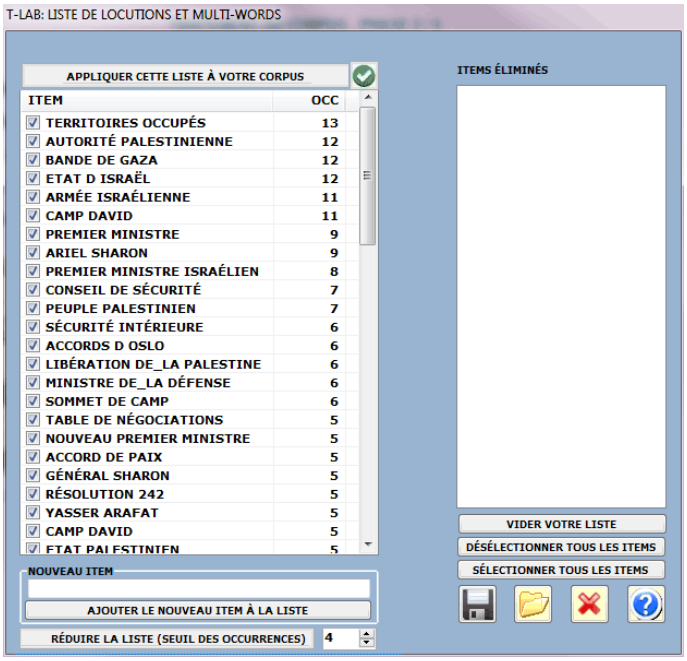
- la longueur d'un mot est fixée à max. 50 caractères;
- ni espaces vides ni signes de ponctuation doivent être inclus.

De toute façon, pour vérifier/utiliser les listes des StopWords pendant l'importation d'un **nouveau corpus** l'utilisateur doit choisir l'option "Avancé" dans la fenêtre suivante:


T-LAB: TRAITEMENT DU CORPUS < PALESTINE.TXT >
CORPUS
NOM : Palestine.txt
DIMENSION : 139 Kb
RÉPERTOIRE : C:\Users\\Documents\T-LAB PLUS\Demo_fr\
TEXTES : 10 DOCUMENTS PRIMAIRES
VARIABLES : 1
IDNUMBERS : Absents
LANGUE: < FRANÇAIS >
LEMMATISATION AUTOMATIQUE Oui Non
Pour plus d'informations cliquez sur le bouton (?)
AFFICHER PLUS D'OPTIONS
LEMMATISATION AUTOMATIQUE
>> FRANÇAIS
Oui
Non
EXAMEN DES STOP-WORDS
Élémentaire
Non
Avancé
SEGMENTATION DU TEXTE (CONTEXTES ÉLÉMENTAIRES)
Énoncés
Fragments
Paragraphes
EXAMEN DES MULTI-WORDS
Non
Élémentaire
Avancé
SELECTION DES MOTS-CLÉS (ORDRE D'IMPORTANCE)
MÉTHODE :
TF-IDF
CHI-DEUX
OCCURRENCES
LISTE AUTOMATIQUE (MAX ITEMS)
3000
AVEC LA VALEUR D'OCCURRENCE >= 4
OPTIONS POUR LES DONNÉES DES MÉDIAS SOCIAUX
Séparer '#' des mots (par ex. '#art' = '# art')
Utiliser les hashtags tels qu'ils sont (par ex. '#art' = '#art')
SUPPRIMER LES HYPERLIENS
CHAQUE LIGNE DE TEXTE = UN TEXTE

Listes de Locutions

Cette option permet de créer/modifier les fichiers des multi-mots (**Multi-Words**).



Chaque fichier Multiwords.txt peut se composer par "N" lignes (maximum 5.000), chacune avec un mot multiple non excédant les 50 caractères et sans signes de ponctuation.

Voici quelques lignes du fichier Multiwords.txt dans le format correct:

chambre de commerce
Haute Cour de Justice
forces de l'ordre
etc etc

En cliquant sur le bouton "**Appliquer cette liste …**", l'utilisateur peut produire une transformation rapide des multi-mots présents dans un corpus en chaînes unitaires qui peuvent être identifiées et classifiées par **T-LAB** (par exemple "Autorité palestinienne" se transforme en "Autorité_palestinienne")

Après son fonctionnement, cette option produit un nouveau fichier (**New_Corpus.txt**) qui, opportunément retitré, peut être analysé avec **T-LAB**.

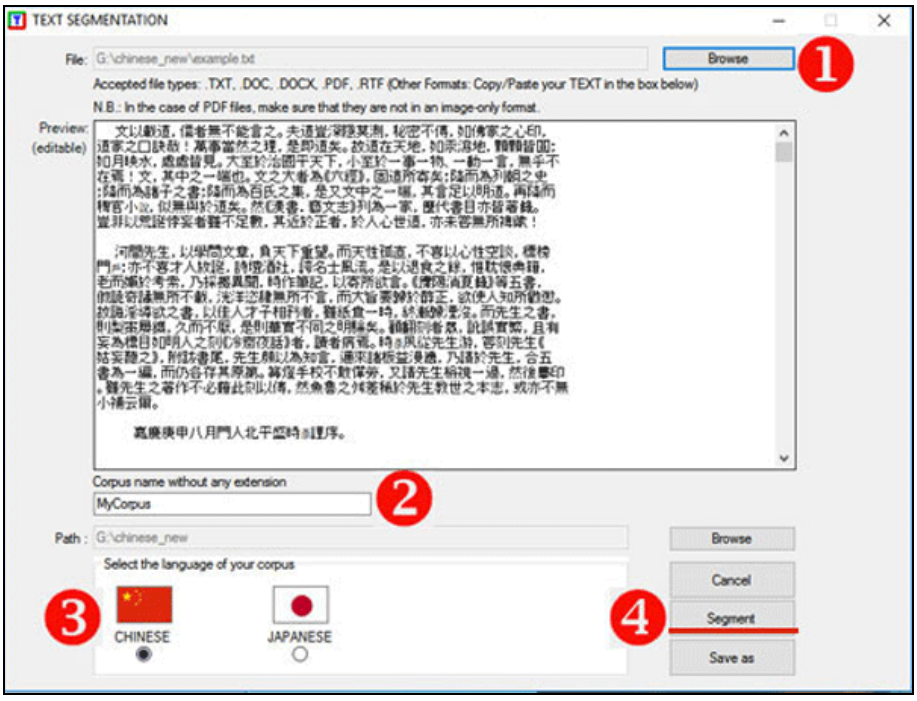
Pour vérifier/utiliser les listes des Multiwords pendant l'**importation d'un nouveau corpus** l'utilisateur doit choisir l'option "Avancé" dans la fenêtre suivante:

Segmentation de Mots

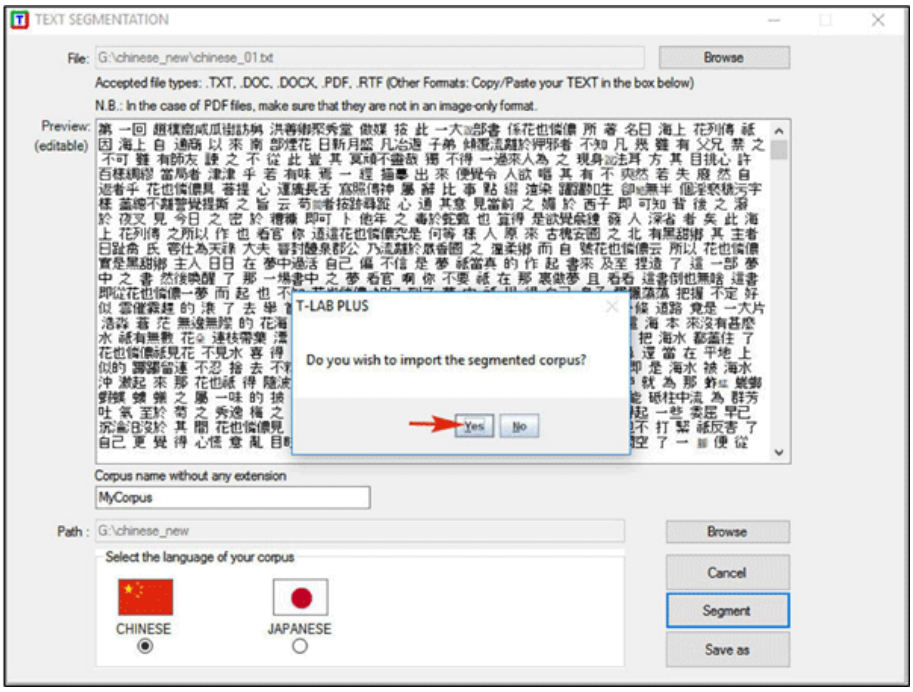
Cet outil **T-LAB** peut être utilisé avant d'importer n'importe quel texte (*) chinois ou japonais qui n'ait pas de délimiteurs, c'est-à-dire des espaces et / ou bien des signes de ponctuation, entre les mots.
(*) Le texte à traiter peut être constitué par un document unique ou par une collection de documents qui incluent des variables catégorielles.

Son utilisation est très simple (voir image ci-dessous):

(1) sélectionner un fichier quelconque;
(2) choisir le nom du projet;
(3) sélectionner la langue du texte;
(4) cliquer sur 'Segmenter'.
En résultat, des espaces vides seront ajoutés entre les mots.



Successivement, si on veut procéder à l' importation, il suffit de répondre 'OUI' à la question " Veux-tu importer le corpus segmenté?" (voir image ci-dessous).



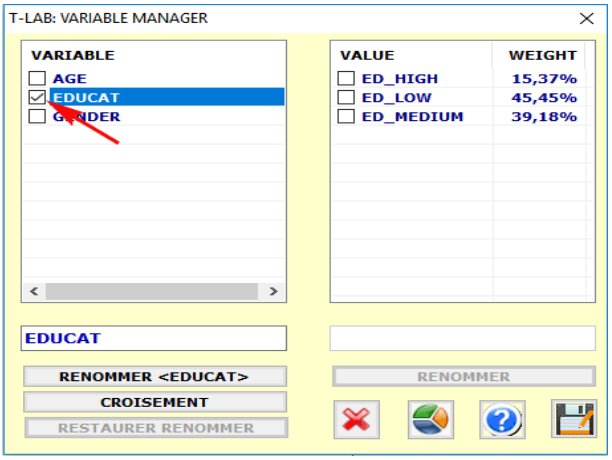
N.B.: Lorsqu' on veut préparer un corpus constitué par plusieurs textes qui comprennent les lignes de codification (c'est-à-dire des variables catégorielles) on conseille de procéder de la manière suivante:
1- 'Assembler' les textes non segmentés (*) au moyen de l' outil Corpus Builder et 'Sauver' le fichier corpus;
2 - Importer le corpus à peine créé au moyen de l' outil Text Segmenter; ensuite procéder comme expliqué auparavant.
(*) Ceci signifie que ,quand on prépare le corpus , il n'est pas nécessaire de segmenter chaque fichier à l' avance.

AUTRES OUTILS

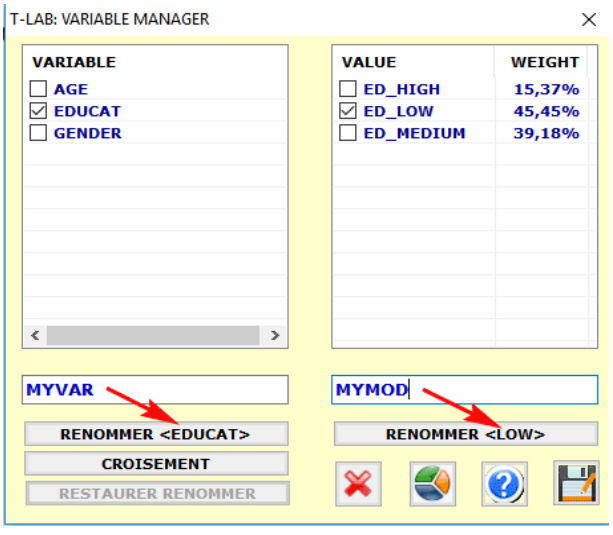
Variable Manager

Cette option, qui est disponible seulement quand le corpus inclut des partitions (variables et catégories), permet cinq genres d'opérations:

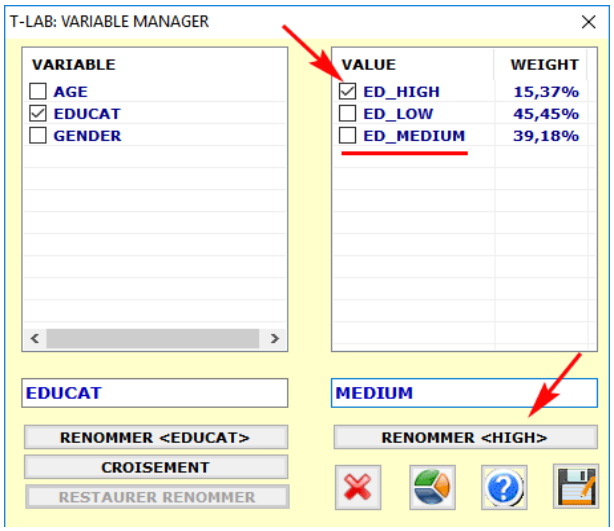
a) **vérifier** les catégories de chaque variable;



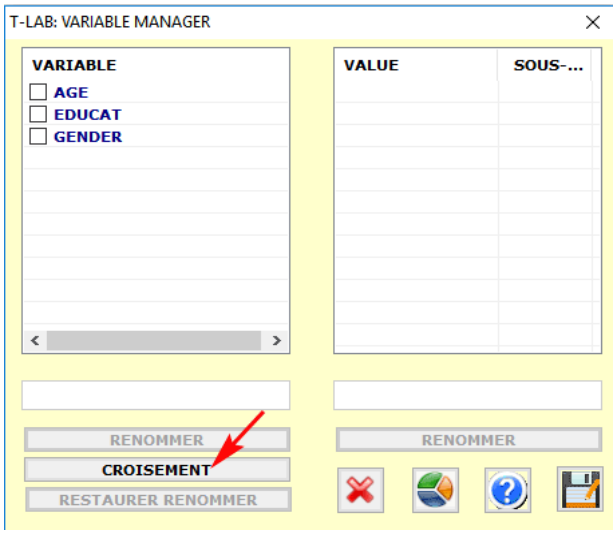
b) **renommer** variables et catégories;

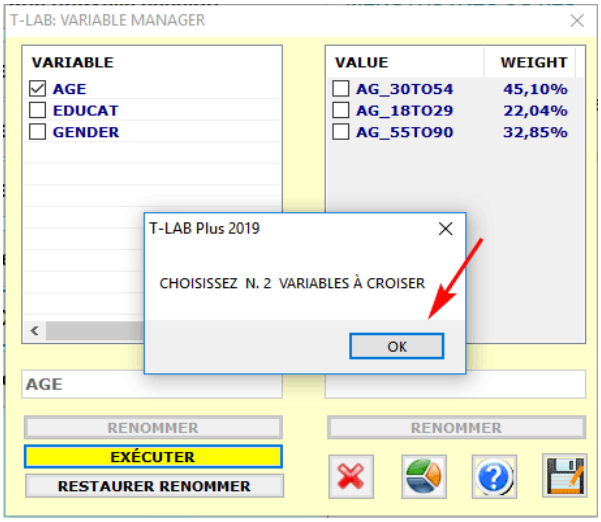


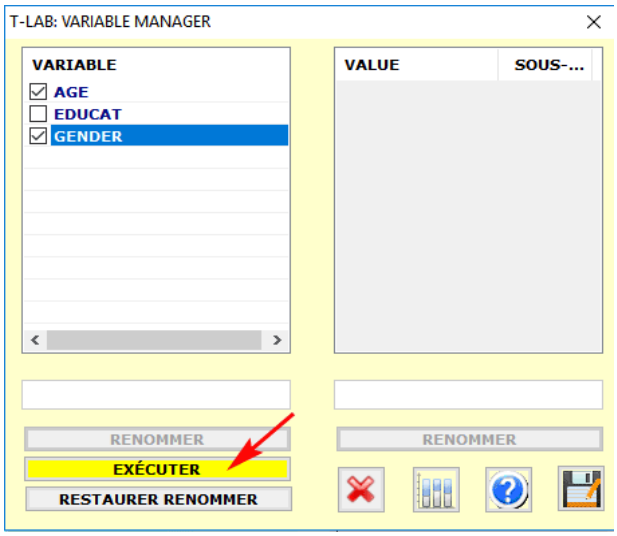
c) **grouper** deux ou plusieurs catégories en leur attribuant la même étiquette;

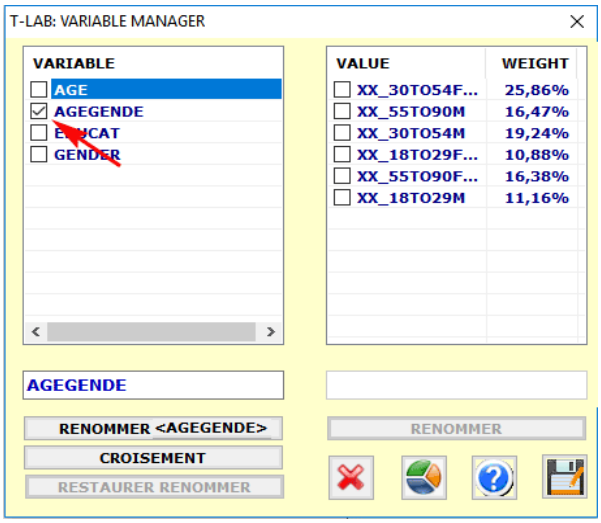


d) créer une **variable croisée**.

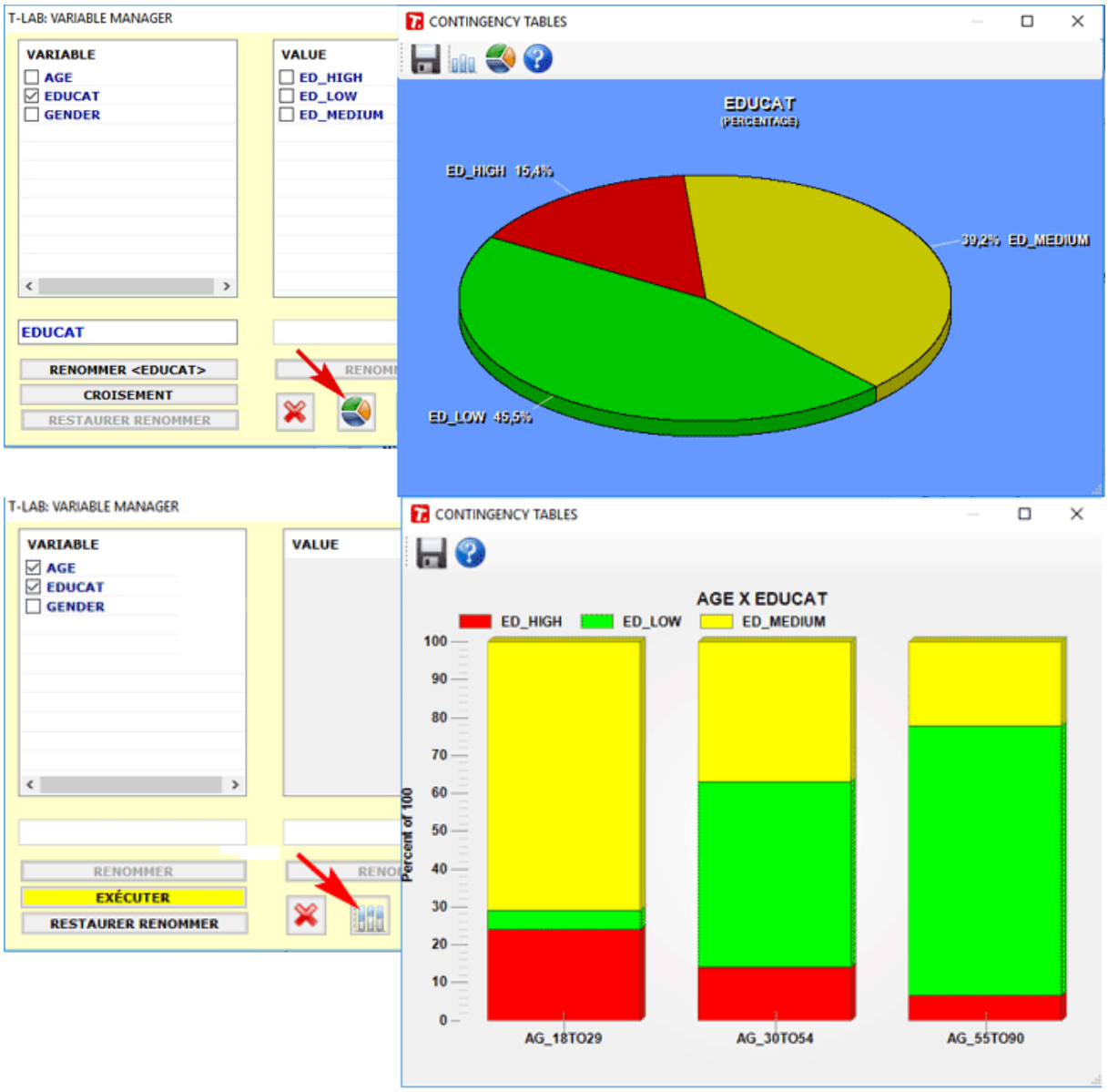






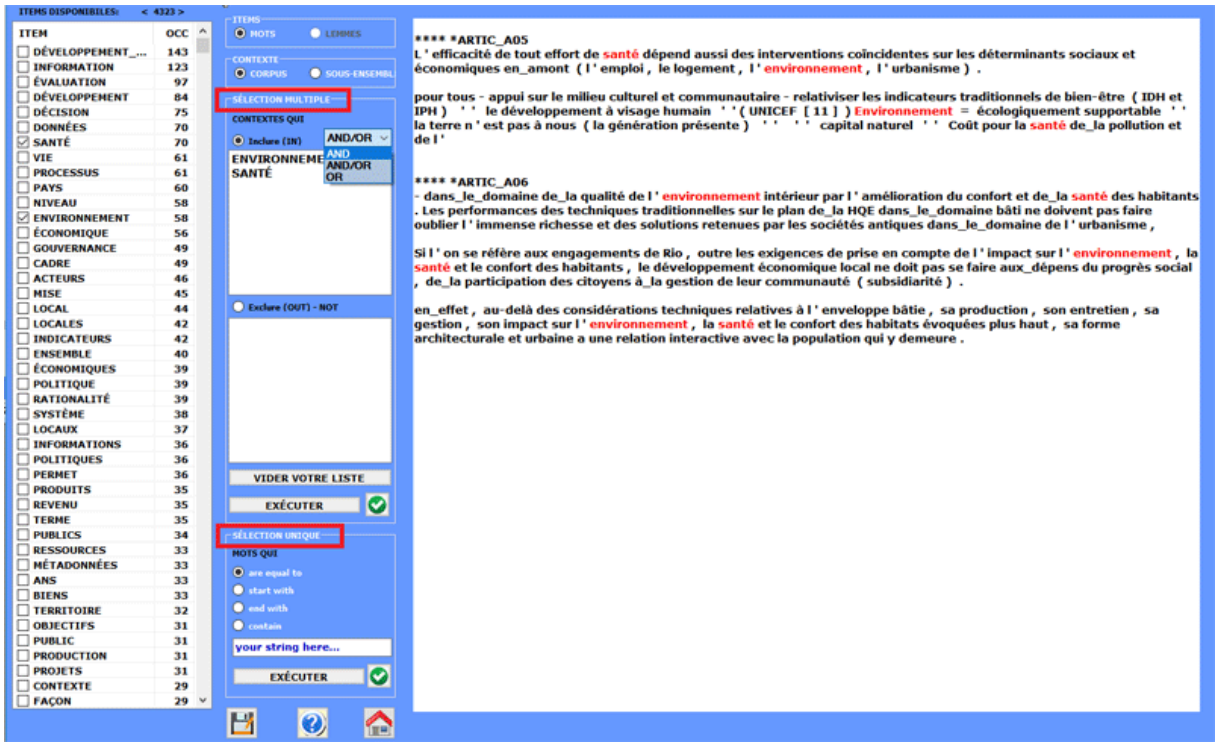


e) créer des **graphiques**.

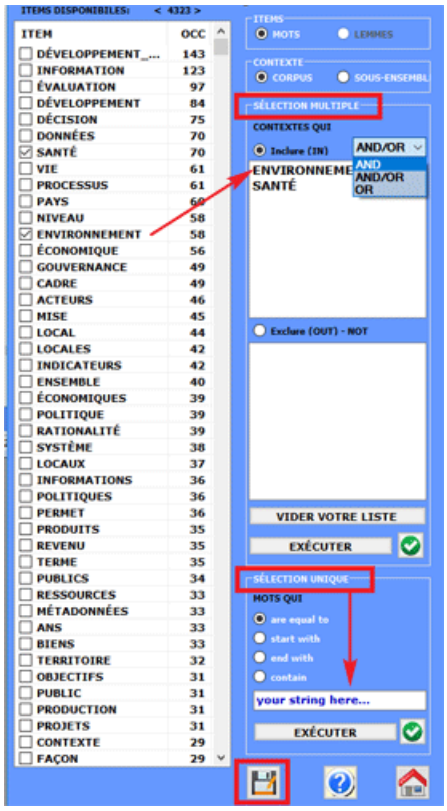

T-LAB: VARIABLE MANAGER
VARIABLE
AGE
EDUCAT
GENDER
VALUE
ED_HIGH
ED_LOW
ED_MEDIUM
EDUCAT
RENOMMER <EDUCAT>
CROISEMENT
RESTAURER RENOMMER
CONTINGENCY TABLES
EDUCAT
(PERCENTAGE)
ED_HIGH 15,4%
39,2% ED_MEDIUM
ED_LOW 45,5%
T-LAB: VARIABLE MANAGER
VARIABLE
AGE
EDUCAT
GENDER
VALUE
RENOMMER
EXÉCUTER
RESTAURER RENOMMER
CONTINGENCY TABLES
AGE X EDUCAT
ED_HIGH
ED_LOW
ED_MEDIUM
Percent of 100
100
90
80
70
60
50
40
30
20
10
0
AG_18TO29
AG_30TO54
AG_55TO90

Recherche Avancée à l'interieur du Corpus

Cet outil **T-LAB** nous permet d'extraire et d'exporter tous les fragments de texte (c'est-à-dire phrases ou paragraphes) qui correspondent à des queries avec des mots simples ou multiples. Ceci soit à l'intérieur du corpus qu'à l'intérieur de ses sous-ensembles.



Son usage est extrêmement intuitif: il suffit de sélectionner les options désirées à l'intérieur des boîtes correspondantes (voir ci-dessous).



En cas de sélections «multiples», les mots peuvent être sélectionnés/ajoutés en cliquant sur les items correspondants du tableau sur la gauche.

Dans le cas de sélections «simples», la chaîne à chercher doit être tapée dans la boîte appropriée.

Après avoir cliqué sur «exécuter», les résultats de la recherche sont visibles dans la boîte à droite et ils peuvent être sauvés dans un fichier.rtf.

Ce fichier, qui inclut tous les codages de **T-LAB**, peut aussi être importé et analysé en tant que sub-corpus.

N.B.: Cette option est activée uniquement lorsque vous travaillez sur un corpus déjà importé et qu'une liste de mots clés a été déjà sélectionnée.

Classification des Nouveaux Documents

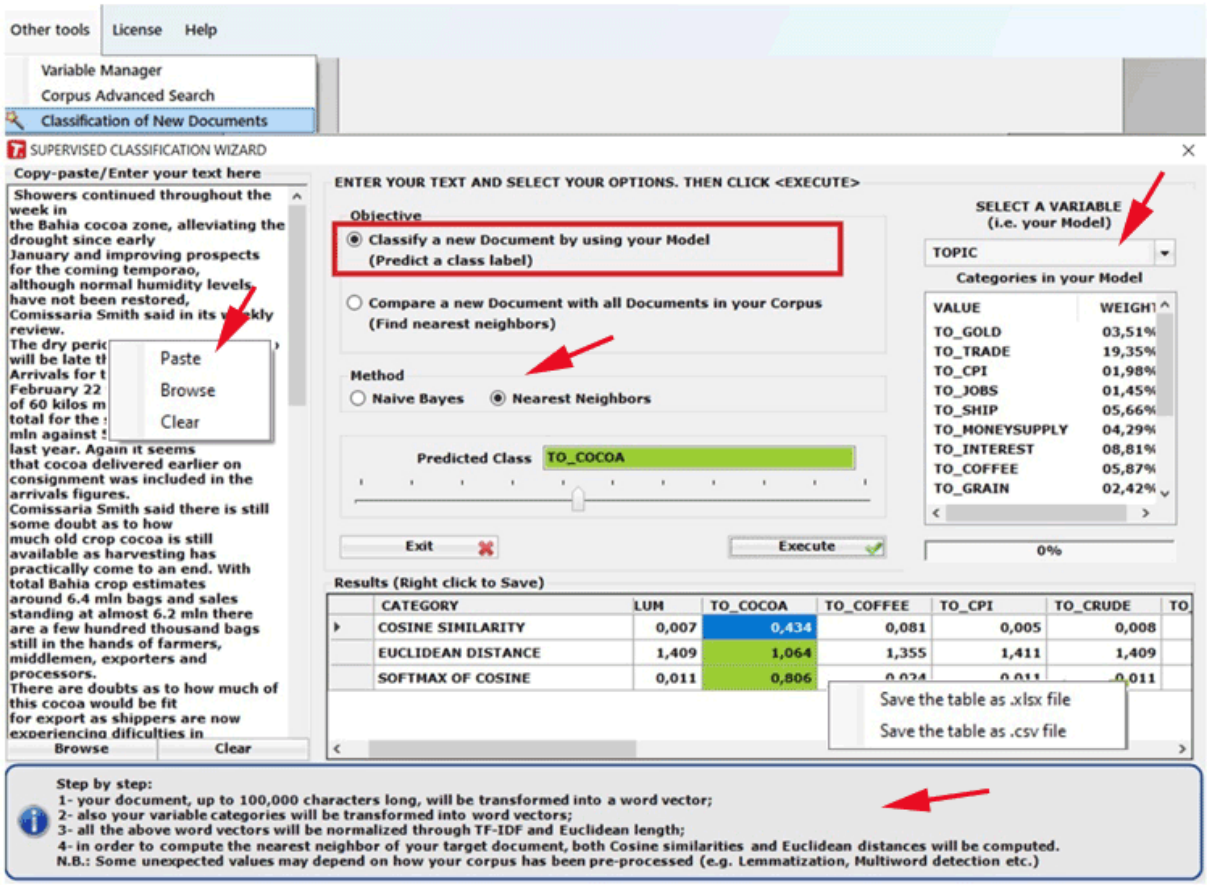
N.B .: Cette section est uniquement disponible en anglais.

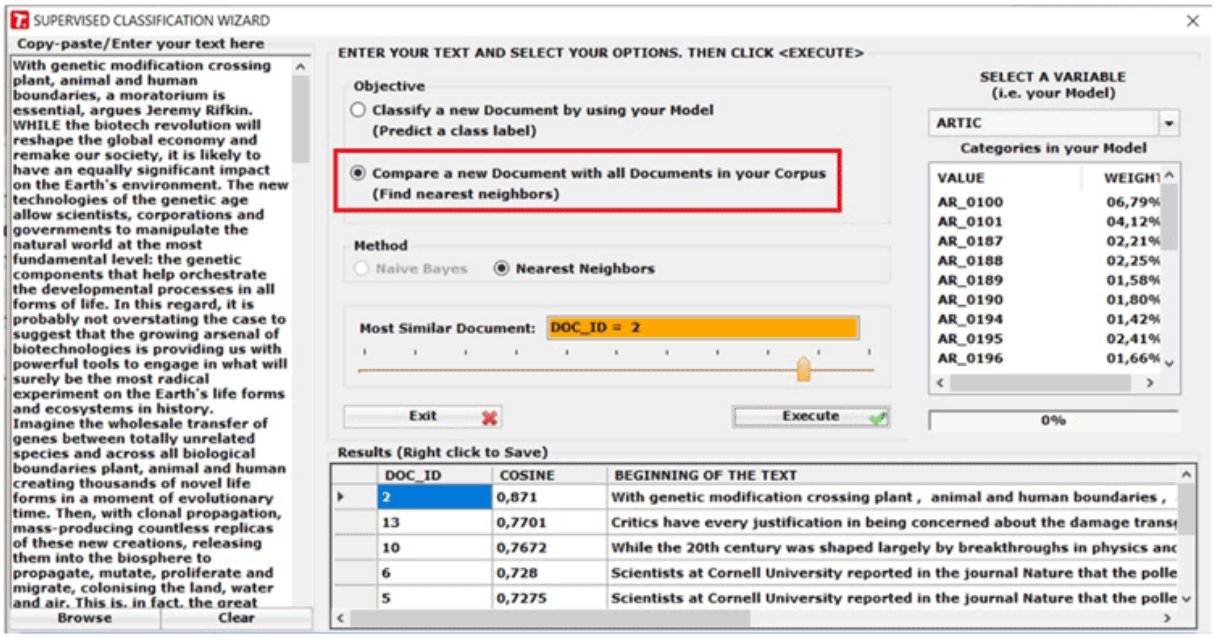
This tool, which is very easy to use, allows one to easily classify new documents according to a pre-existing model (i.e. any categorical variable) and also to compare any new document with all documents included in a corpus already analysed.

To this purpose, the following steps are required:

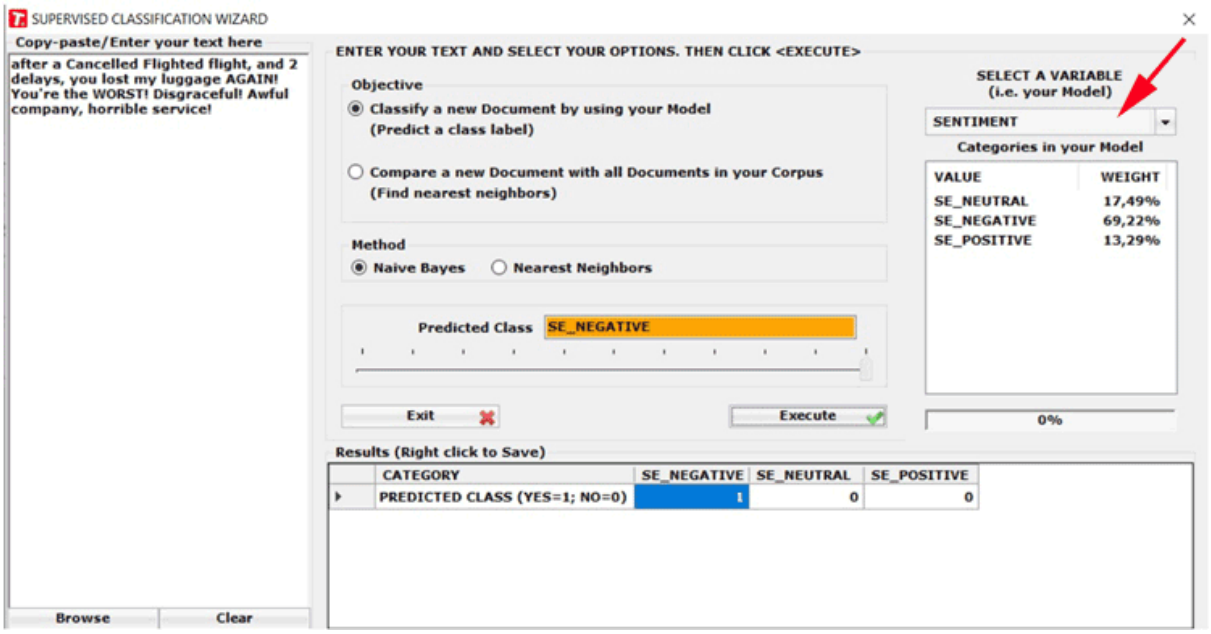
- enter a new document in the appropriate box;
- select a categorical variable to be used as a 'model';
- choose the desired 'objective' and a 'method';
- click 'execute'.

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).

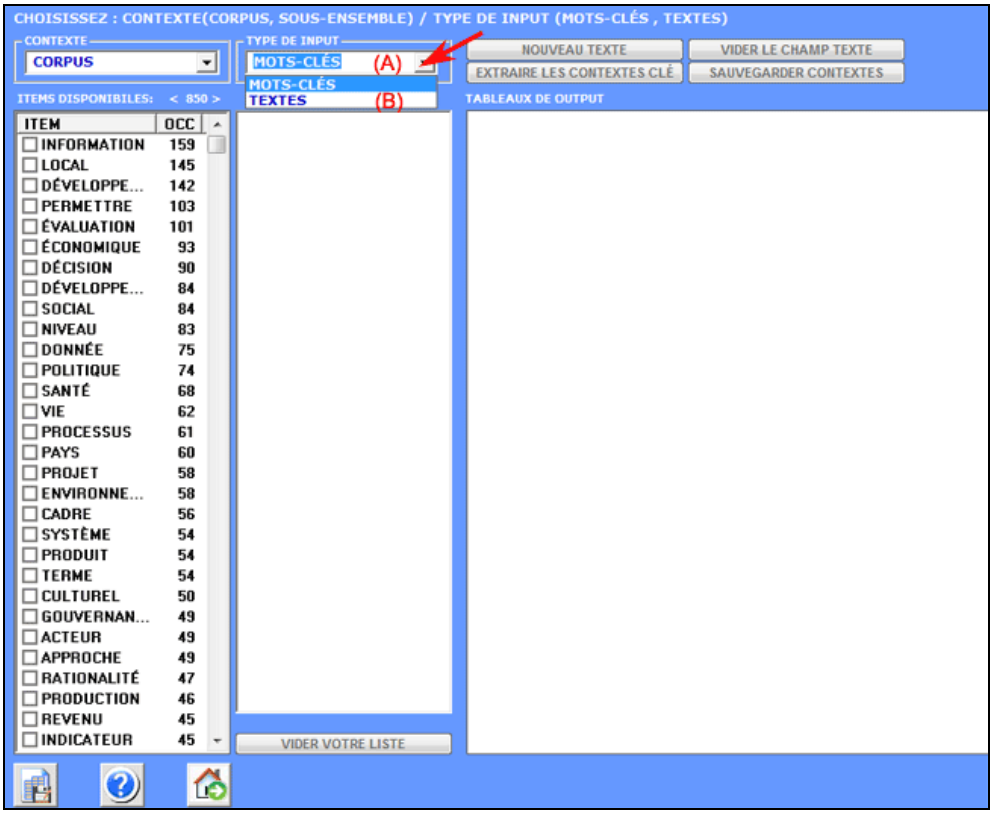


N.B.: When the user wishes to classify a dataset of new documents by using a supervised method, the dataset must be imported by T-LAB and then analysed by using a previously generated dictionary. To this purpose, the 'Thematic Document Classification' can be used, both for generating a dictionary of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

Contextes Clé de Mots Thématiques

Cet outil **T-LAB** peut être utilisé pour deux buts différents:

a) extraire des ensembles d'unités de contexte qui permettent d'approfondir la valeur thématique de **mots-clés spécifiques**;
b) extraire les unités de contexte qui résultent les plus semblables à des **textes échantillons proposés par l'utilisateur**.

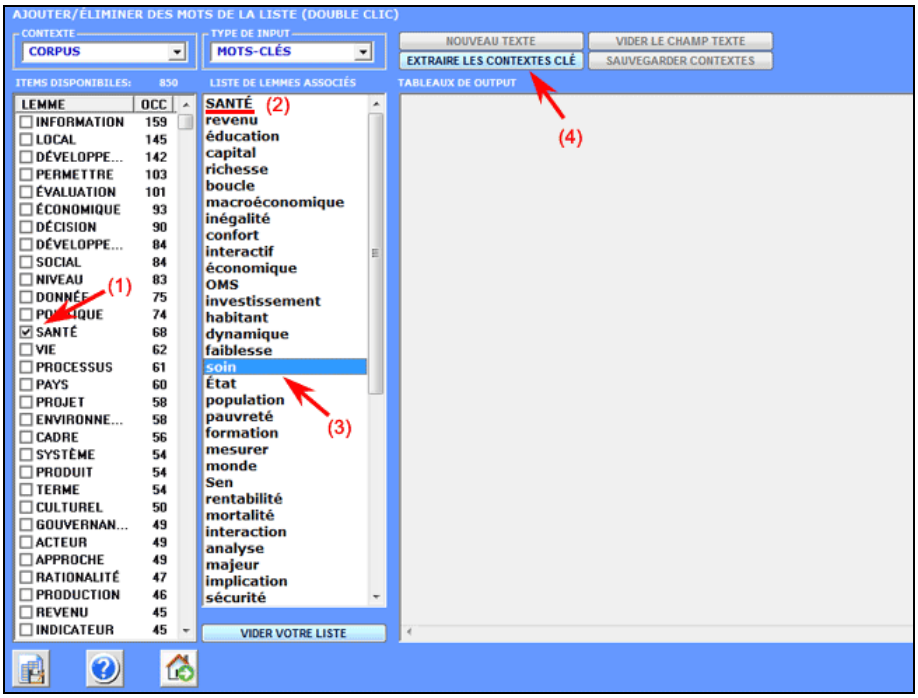


Étape par étape, les procédures respectives sont les suivantes:

Case (A)

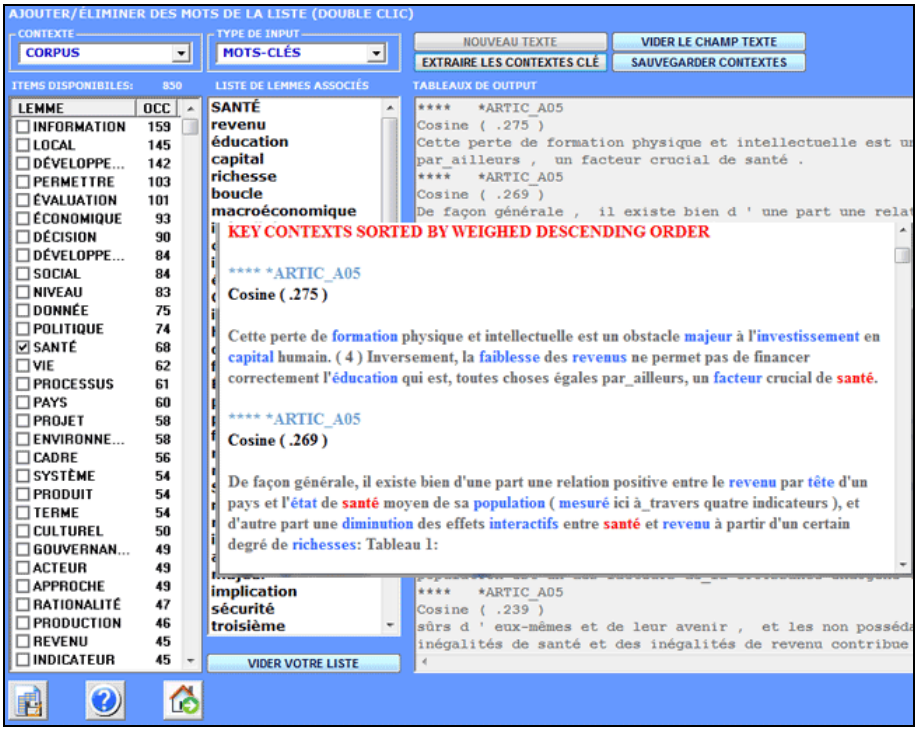
1- l'utilisateur choisit (double clic) un mot thématique "X" (voir ci-dessous);
2- T-LAB propose une liste de mots (maximum 50) dont les valeurs de co-occurrence avec "X" sont les plus significatifs;
3- l'utilisateur peut enlever les items non pertinents de la liste. **T-LAB** assume que la liste sélectionnée est un "query vector" et calcule ses index d'association (c.-à-d. les coefficients de cosinus) avec tous les contextes élémentaires du corpus ou du sous-ensemble sélectionné;

4- l'output est un fichier **HTML** qui contient une liste des contextes clé les plus significatifs de "X", énumérés par l'ordre décroissant de leurs index d'association;



À la différence de **Concordances**, qui permet l'extraction de tous les contextes élémentaires dans lesquels les mots clés sélectionnés sont présents (occurrences), et à la différence d' **Associations des Mots**, qui permet l'extraction de tous les contextes élémentaires dans lesquels les mots clés sélectionnés sont accouplés (co-occurrences), cet outil nous permet d'extraire les contextes élémentaires dans lesquels chaque mot clé est associé à d'autres mots (co-occurrences multiples) définissant son champ thématique.

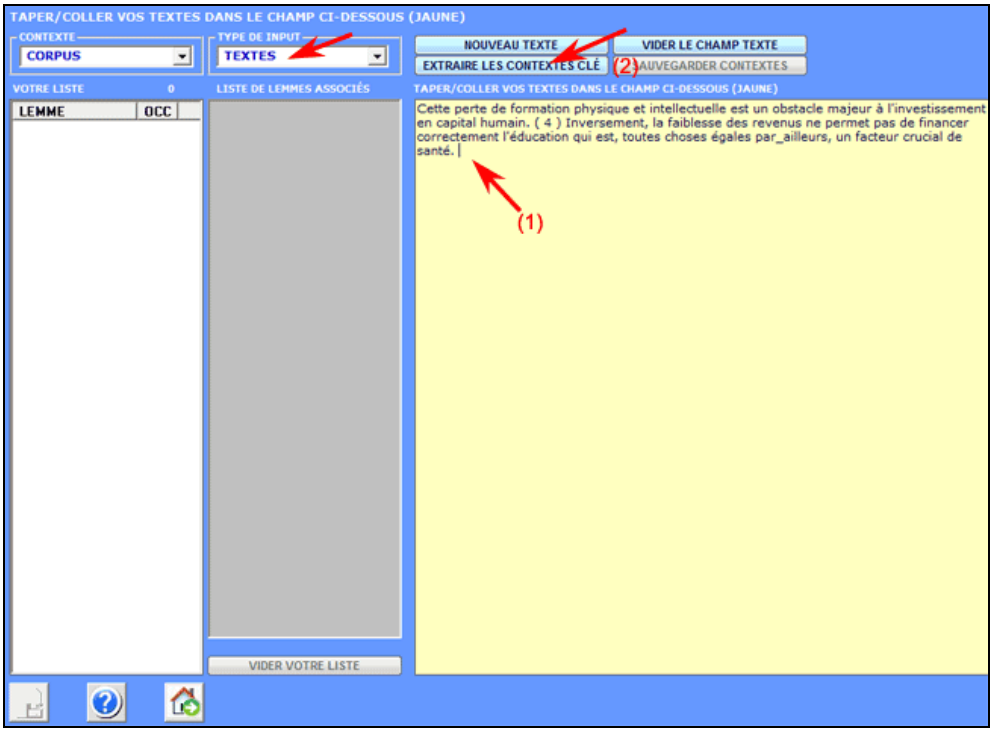
Les outputs, soit en format HTML que TXT contiennent une liste des contextes clé de " X " les plus significatifs, énumérés dans l'ordre décroissant de leurs indices d'association.



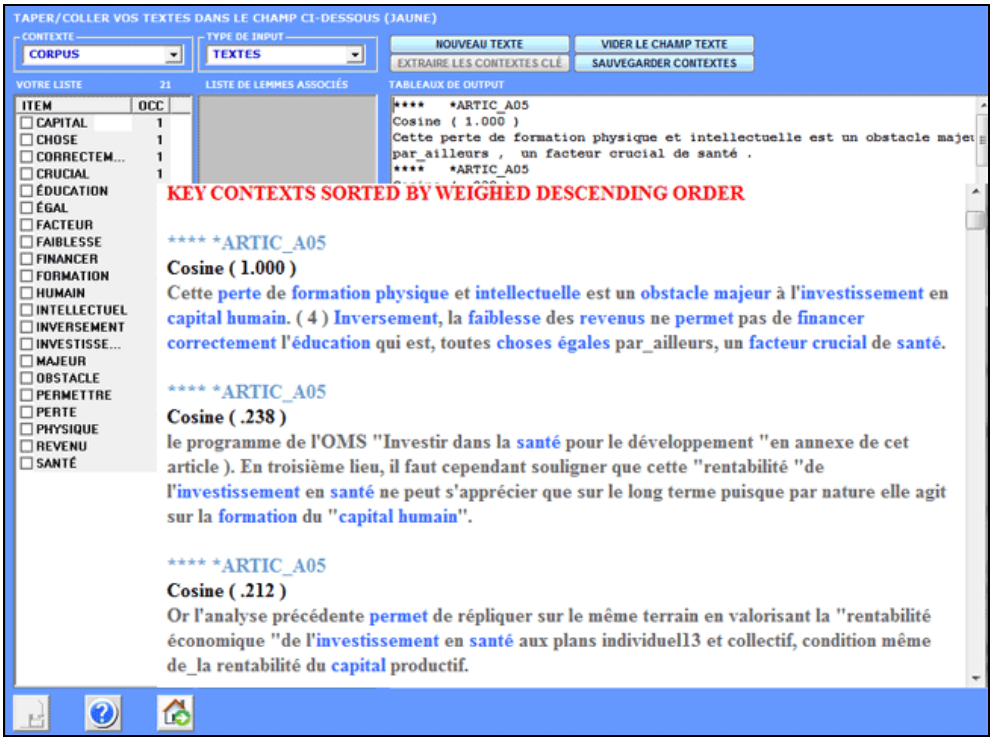
Les étapes 1-4 peuvent être réitérées pour "n" mots thématiques.

Case (B)

Il fonctionne comme suit:
1 - l'utilisateur copie / colle un texte «modèle» (max 5000 caractères) dans la case correspondante;
2 - après avoir cliqué l'option «extrait contextes clé », **T-LAB** transforme le texte introduit en un vecteur (query vector) et calcule les indices d'association relatifs (c'est-à-dire les coefficients cosinus) avec tous les contextes élémentaires du corpus ou du sous-ensemble sélectionné;



Les outputs, soit en format HTML qu'en format TXT contiennent une liste des contextes clé qui sont les plus proches au texte en input.
NB: Dans ce cas la mesure de similarité ne tient pas compte des mots multiples dont les chaînes, avec ou sans le trait underscore (« _ »), ne correspondent pas au texte analysé.


TAPER/COLLER VOS TEXTES DANS LE CHAMP CI-DESSOUS (JAUNE)
CONTEXTE
CORPUS
TYPE DE INPUT
TEXTES
NOUVEAU TEXTE
VIDER LE CHAMP TEXTE
EXTRAIRE LES CONTEXTES CLÉ
SAUVEGARDER CONTEXTES
VOTRE LISTE
21
LISTE DE LEMMES ASSOCIÉS
TABLEAUX DE OUTPUT
ITEM
OCC
CAPITAL 1
CHOSE 1
CORRECTEM... 1
CRUCIAL 1
ÉDUCATION
ÉGAL
FACTEUR
FAIBLESSE
FINANCER
FORMATION
HUMAIN
INTELLECTUEL
INVERSEMENT
INVESTISSE...
MAJEUR
OBSTACLE
PERMETTRE
PERTE
PHYSIQUE
REVENU
SANTÉ
**** *ARTIC_A05
Cosine (1.000)
Cette perte de formation physique et intellectuelle est un obstacle majeu
par_ailleurs , un facteur crucial de santé .
**** *ARTIC_A05
KEY CONTEXTS SORTED BY WEIGHED DESCENDING ORDER
**** *ARTIC_A05
Cosine (1.000)
Cette perte de formation physique et intellectuelle est un obstacle majeur à l'investissement en capital humain. (4) Inversement, la faiblesse des revenus ne permet pas de financer correctement l'éducation qui est, toutes choses égales par_ailleurs, un facteur crucial de santé.
**** *ARTIC_A05
Cosine (.238)
le programme de l'OMS "Investir dans la santé pour le développement "en annexe de cet article). En troisième lieu, il faut cependant souligner que cette "rentabilité "de l'investissement en santé ne peut s'apprécier que sur le long terme puisque par nature elle agit sur la formation du "capital humain".
**** *ARTIC_A05
Cosine (.212)
Or l'analyse précédente permet de répliquer sur le même terrain en valorisant la "rentabilité économique "de l'investissement en santé aux plans individuel13 et collectif, condition même de_la rentabilité du capital productif.

Exporter des Tables Personnalisées

N.B.: Les images de cette section font référence à une version précédente de T-LAB.
En **T-LAB 10**, l'aspect est légèrement différent, mais les fonctions sont les mêmes.

Cette option nous permet de créer, d'explorer et d'exporter trois types de tableaux:

a) ceux avec les valeurs d'**occurrences** des unités lexicales dans les sous-ensembles du corpus définis au moyen de quelque variable (matrices rectangulaires);

b) ceux avec les valeurs de **co-occurrences** des unités lexicales (matrices carrées) dans le corpus ou dans ses sous-ensembles.

c) ceux avec les **occurrences** des différentes unités lexicales à l'intérieur de tous les documents (matrices dispersées avec les index des éléments différents).

Les dimensions maximum de tels tableaux sont respectivement: a) 10.000 lignes pour 150 colonnes; b) 1.500 lignes de 1.500 colonnes; c) 30.000 documents pour 10.000 unités lexicales.

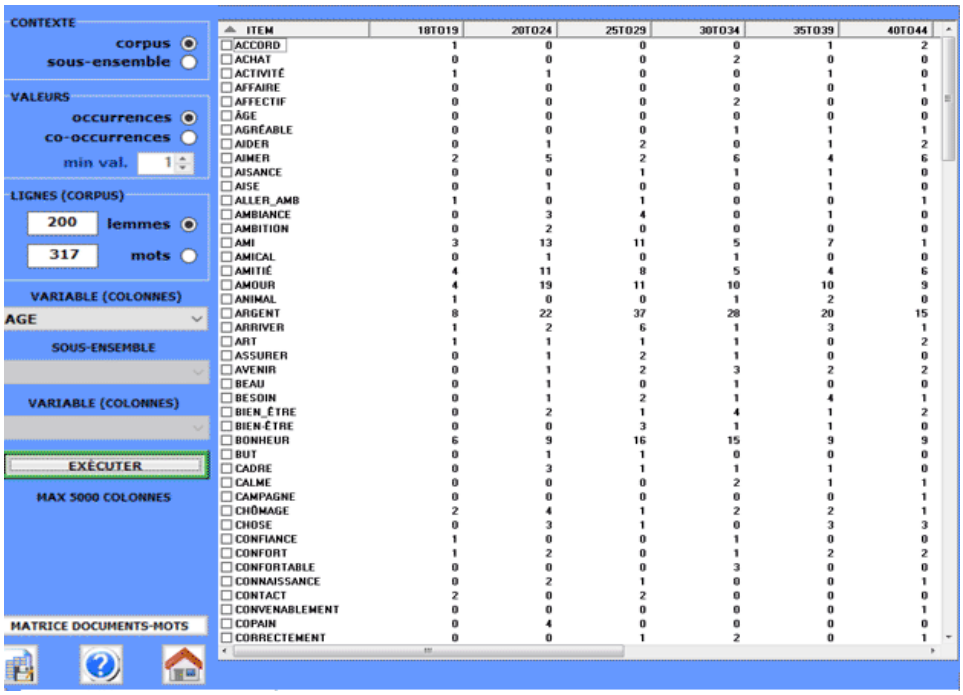
Les dimensions maximales de ces tableaux sont respectivement: a) 10.000 lignes par 150 colonnes, b) 5.000 lignes par 5.000 colonnes..

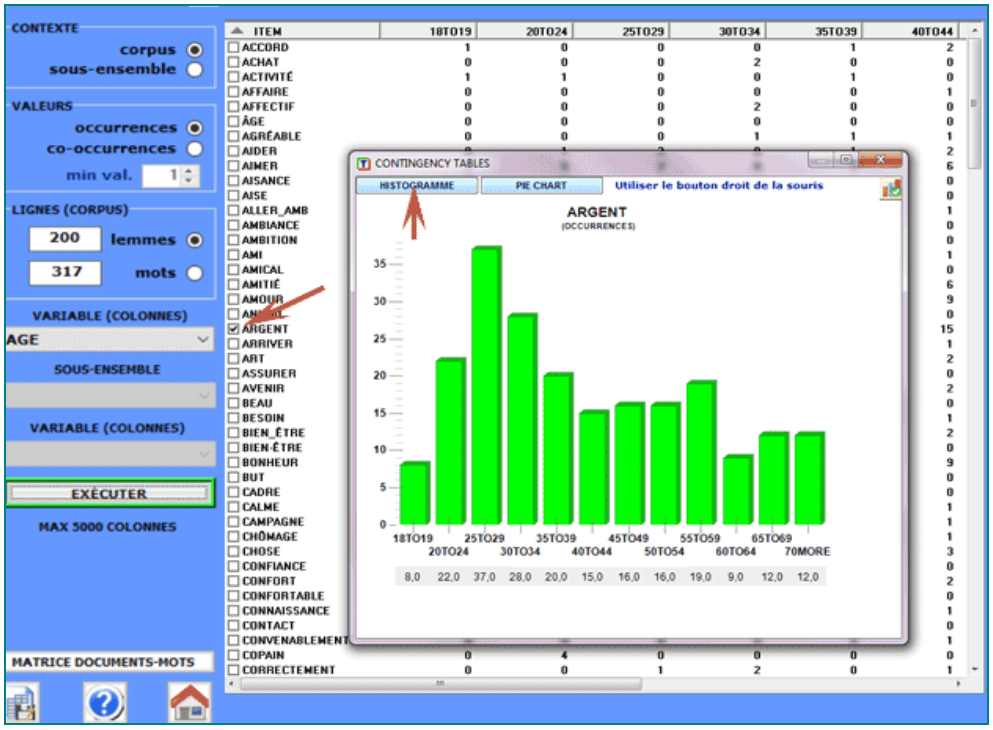
L'utilisation de cette fonction est très intuitive.

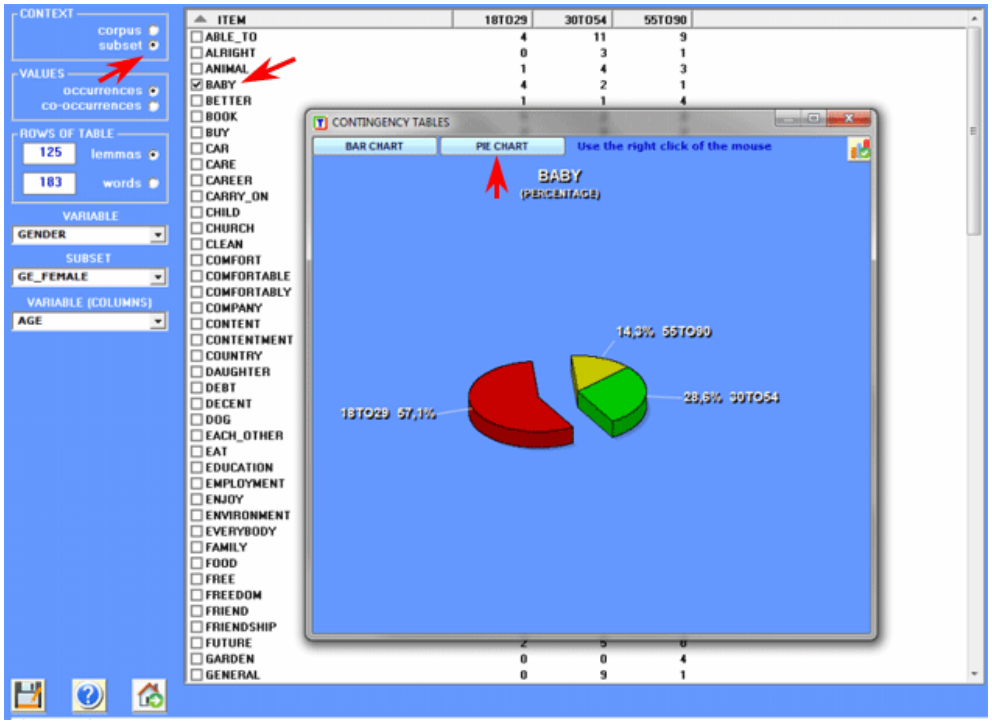
Dans les cas plus simples, l'utilisateur doit choisir la variable dont les modalités constitueront les colonnes du tableau.

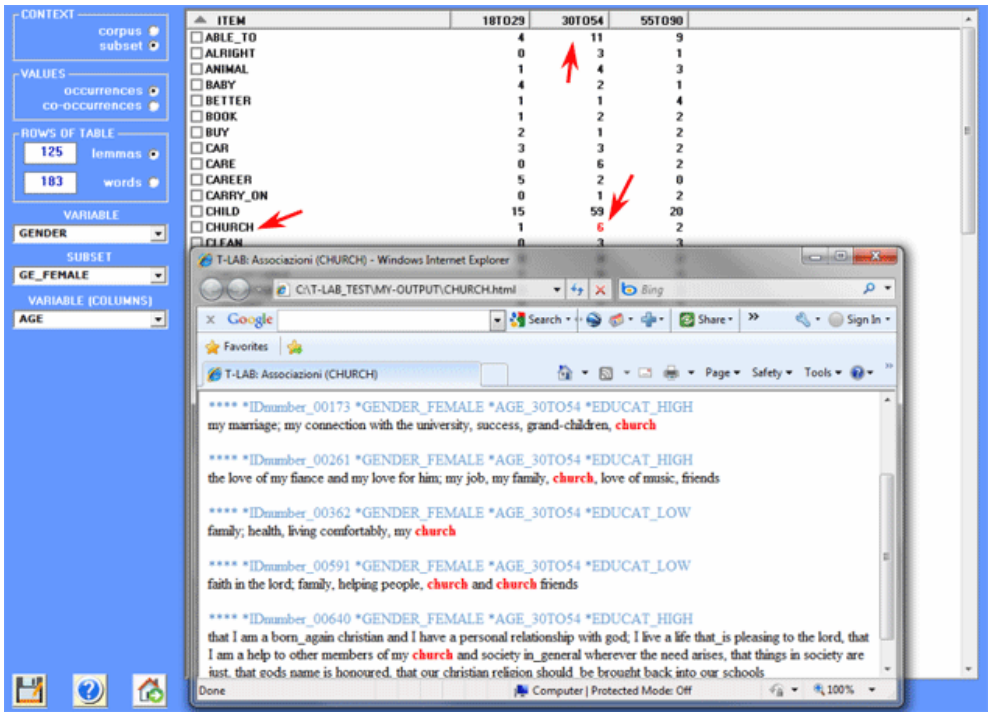
Dans les cas plus complexes, il doit choisir une variable et un sous-ensemble.

Tous les tableaux nous permettent de créer différents types des **graphiques**.
De plus, en cliquant sur cellules spécifiques d'un tableau, il est possible de créer un **fichier HTML** montrant tous les contextes élémentaires où le mot en ligne est présent dans le sous-ensemble correspondant (voir ci-dessous).


CONTEXTE
corpus
sous-ensemble
VALEURS
occurrences
co-occurrences
min val. 1
LIGNES (CORPUS)
200 lemmes
317 mots
VARIABLE (COLONNES)
AGE
SOUS-ENSEMBLE
VARIABLE (COLONNES)
EXÉCUTER
MAX 5000 COLONNES
MATRICE DOCUMENTS-MOTS
ITEM 18TO19 20TO24 25TO29 30TO34 35TO39 40TO44
ACCORD 1 0 0 0 1 2
ACHAT 0 0 0 2 0 0
ACTIVITÉ 1 1 0 0 1 0
AFFAIRE 0 0 0 0 0 1
AFFECTIF 0 0 0 2 0 0
ÂGE 0 0 0 0 0 0
AGRÉABLE 0 0 0 1 1 1
AIDER 0 1 2 0 1 2
AIMER 2 5 2 6 4 6
AISANCE 0 0 1 1 1 0
AISE 0 1 0 0 1 0
ALLER_AMB 1 0 0 1 0 0 1
AMBIANCE 0 3 4 0 1 0
AMBITION 0 2 0 0 0 0
AMI 3 13 11 5 7 1
AMICAL 0 1 0 1 0 0
AMITIÉ 4 11 8 5 4 6
AMOUR 4 19 11 10 10 9
ANIMAL 1 0 0 1 2 0
ARGENT 8 22 37 28 20 15
ARRIVER 1 2 6 1 3 1
ART 1 1 1 1 0 2
ASSURER 0 1 2 1 0 0
AVENIR 0 1 2 3 2 2
BEAU 0 1 0 1 0 0
BESOIN 0 1 2 1 4 1
BIEN_ÊTRE 0 2 1 4 1 2
BIEN-ÊTRE 0 0 3 1 1 0
BONHEUR 6 9 16 15 9 9
BUT 0 1 1 0 0 0
CADRE 0 3 1 1 1 0
CALME 0 0 0 2 1 1
CAMPAGNE 0 0 0 0 0 1
CHÔMAGE 2 4 1 2 2 1
CHOSE 0 3 1 0 3 3
CONFIANCE 1 0 0 1 0 0
CONFORT 1 2 0 1 2 2
CONFORTABLE 0 0 0 3 0 0
CONNAISSANCE 0 2 1 0 0 1
CONTACT 2 0 2 0 0 0
CONVENABLEMENT 0 0 0 0 0 1
COPAIN 0 4 0 0 0 0
CORRECTEMENT 0 0 1 2 0 1


CONTEXTE
corpus
sous-ensemble
VALEURS
occurrences
co-occurrences
min val. 1
LIGNES (CORPUS)
200 lemmes
317 mots
VARIABLE (COLONNES)
AGE
SOUS-ENSEMBLE
VARIABLE (COLONNES)
EXÉCUTER
MAX 5000 COLONNES
MATRICE DOCUMENTS-MOTS
CONTINGENCY TABLES
HISTOGRAMME
PIE CHART
Utiliser le bouton droit de la souris
ARGENT
(OCCURRENCES)
18TO19 20TO24 25TO29 30TO34 35TO39 40TO44 45TO49 50TO54 55TO59 60TO64 65TO69 70MORE
8,0 22,0 37,0 28,0 20,0 15,0 16,0 16,0 19,0 9,0 12,0 12,0





Pour exporter des matrices dispersées de type documents pour mots, il suffit d'appuyer sur le bouton approprié (voir ci-dessus).

Dans ce cas, les types d'output sont deux:

Le premier (Sparse_Matrix.csv) a le format suivant:

Doc_Index;Word_Index;Word_Occ

00001; 1; 12

00001; 2; 5

Le second (Word_Indexes.csv) a le format suivant :

Word_Index;Word_Label

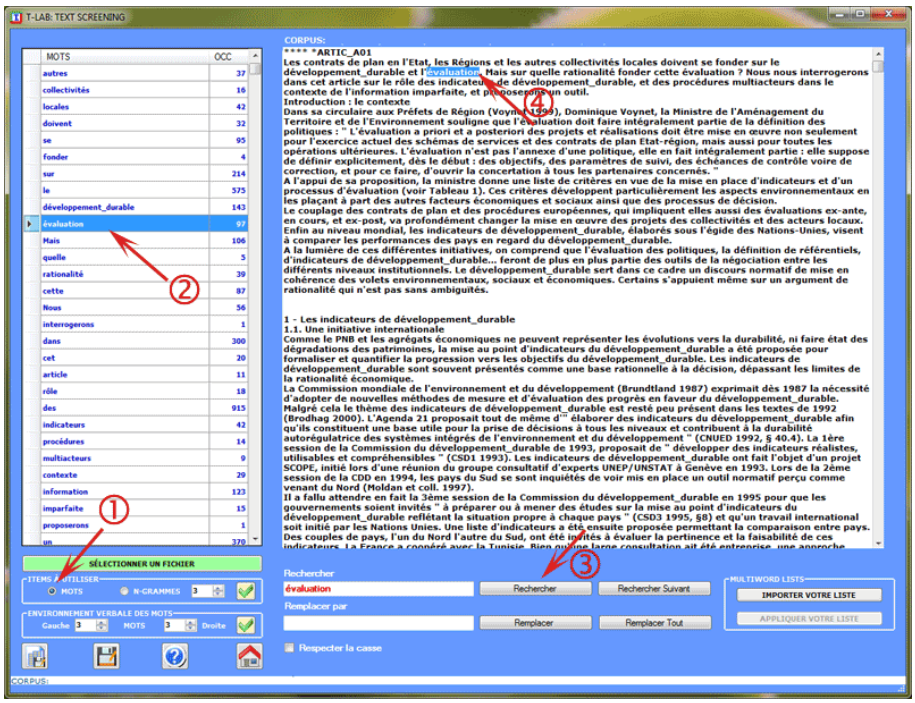
1; abolir

2; accepter

…

Editeur

N.B.: En **T-LAB 10**, les fonctions pour l'édition des fichiers en format texte sont inclues dans l' outil **Text Screening** (voir ci-dessous).



Importer-Exporter une Liste des Identificateurs

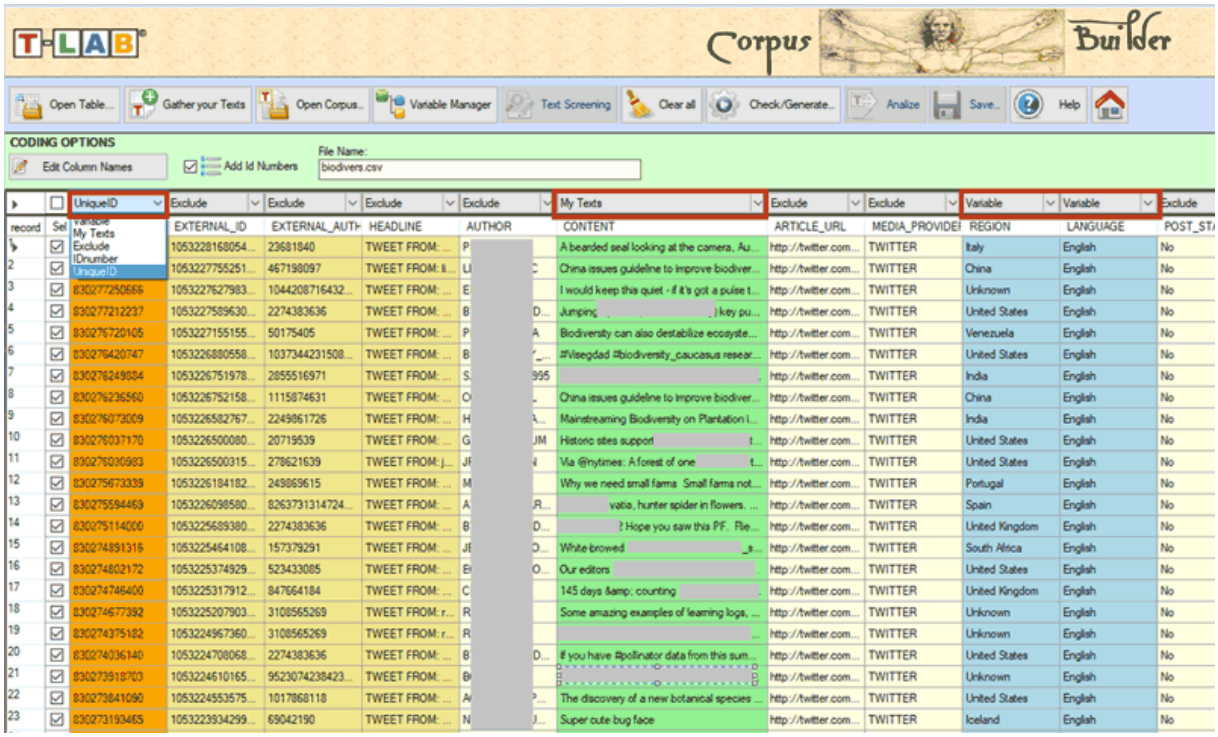
Dans **T-LAB**, un identificateur unique ('Unique Identifier') est une variable catégorique avec une valeur distincte pour chaque document (ou cas).

Une liste d'identificateurs uniques peut être constituée de n'importe quel type de chaînes alphanumériques (par exemple, numéros d'identification, noms propres, noms géographiques, noms de livres, etc.), d'une longueur maximale de 50 caractères et sans espaces blancs.

Puisque les identificateurs sont uniques, il est impossible d'effectuer une analyse des données à leur sujet. Ils sont seulement utilisés pour identifier les résultats dans les sorties du logiciel.

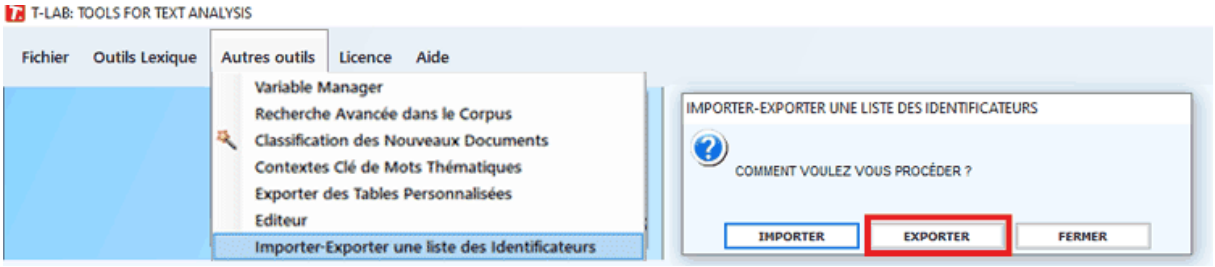
Dans **T-LAB**, en utilisant les options importer / exporter, toute liste d'identificateurs uniques peut être modifiée à tout moment.

Lors de l'importation de données sous forme de tableau les identifiants uniques doivent figurer dans la première colonne, comme dans l'exemple suivant concernant les messages Twitter.



Dans les autres cas (c'est-à-dire collections de documents qui ne sont pas au format tabulaire) la procédure recommandée est la suivante:

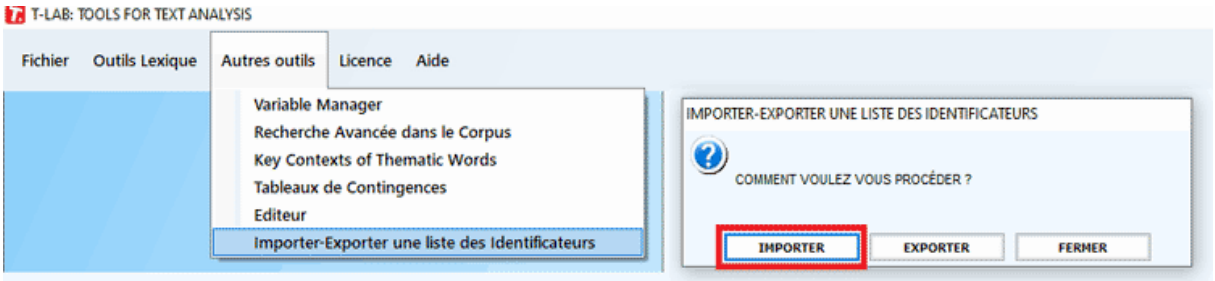
1- Importer d'abord votre corpus;
2- Exporter la liste des identificateurs crée automatiquement par **T-LAB.**



3- Editer et modifier le fichier CSV créé par **T-LAB** (c'est-à-dire modifier simplement les valeurs de 'MyIdentifier' en fonction de vos besoins. Voir l'image ci-dessous)

MyID	MyIdentifier
1	TOBEREPLACED00001
2	TOBEREPLACED00002
3	TOBEREPLACED00003
4	TOBEREPLACED00004
5	TOBEREPLACED00005
6	TOBEREPLACED00006
7	TOBEREPLACED00007
8	TOBEREPLACED00008
9	TOBEREPLACED00009
10	TOBEREPLACED00010
...	...

4- Importer le fichier CSV contenant vos identificateurs uniques revus.



GLOSSAIRE

Analyse des Correspondances

La **technique factorielle** nommée Analyse des Correspondances est appliquée à l'étude des **tableaux de données** dont les cellules contiennent des valeurs de fréquence (nombres positifs) ou des valeurs de présence-absence ("1" et "0").

Comme toutes les techniques factorielles, l'Analyse des Correspondances permet l'extraction de nouvelles variables - les **facteurs** - qui ont la propriété de récapituler d' une façon organisée l'information significative contenue dans les innombrables cellules des tableaux de données; en outre, cette technique d'analyse permet la représentation graphique - dans un ou plusieurs espaces - des points qui détectent les **objets** en lignes et colonnes, qui - dans notre cas - sont les entités linguistiques (mots, lemmes, segments de texte, textes, etc) avec les respectives caractéristiques de provenance.

En termes géométriques, chaque facteur organise une dimension spatiale qui peut être représentée comme ligne ou axe - dont le centre (ou barycentre) est la valeur "0", et qui se développe d'une manière bipolaire vers l'extrémité négative (-) et positive (+), de sorte que les objets mis sur les pôles opposés sont les plus différents, presque comme la "gauche " et la "droite " sur les axes de la politique.

Dans **T-LAB** les résultats d'analyse sont récapitulés par des graphiques qui permettent d'évaluer des rapports de proximité/distance - ou de similitude/différence - entre les objets considérés.

En outre, **T-LAB** montre des mesures, en particulier les **Contributions Absolues** et les **Valeurs Test**, qui aident à interpréter les **pôles factoriels** qui organisent les rapports de similitude/différence entre les objets considérés.

Chaînes de Markov

Une chaîne markovienne (du nom du mathématicien russe Andrei Andreiëvich Markov) est constituée d'une **succession** (ou séquence) d'évènements, généralement indiqués comme **états**, caractérisée par deux propriétés:

- l'ensemble des évènements et de leurs issues possibles est fini;
- l'issue de chaque évènement dépend seulement (ou au maximum) de l'évènement immédiatement précédent.

Avec la conséquence qu'une valeur de probabilité correspond à chaque transition d'un évènement à l'autre.

Dans le domaine scientifique, le modèle des chaînes markoviennes est utilisé pour analyser les successions d'évènements économiques, biologiques, physiques, etc. Dans le domaine des études linguistiques ses applications ont pour objet les combinaisons possibles des diverses unités d'analyses sur l'axe des relations syntagmatiques (l'une après l'autre).

Dans **T-LAB** l'analyse des chaînes markoviennes concerne deux types de **séquences**:

- celles concernant les relations entre unités lexicales (mots, lemmes ou catégories) présentes dans le corpus en analyse;
- celles présentes dans des fichiers externes préétablis par l'utilisateur.

Dans les deux cas, en premier lieu sont constitués des tableaux carrés dans lesquels sont reportées les occurrences des transitions, c'est-à-dire des quantités qui indiquent le nombre de fois qu'une unité d'analyses précède (ou suit) l'autre. Successivement, les occurrences des transitions sont transformées en valeurs de probabilité (voir image suivante).

	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

Pour plus d'informations voir **Analyse des Séquences**.

Chi-Deux

C'est un test statistique utilisé pour vérifier si les valeurs de fréquence obtenues par une enquête, et enregistrées dans un certain tableau croisé, sont sensiblement différentes de leurs valeurs théoriques.

En général **T-LAB** applique ce test à des tableaux (2 x 2); par conséquent la valeur seuil est 3.84 (df = 1; p. 0.05) ou 6.64 (df = 1; p. 0.01).

Par exemple, afin de vérifier la signification des occurrences d'un mot ("x") dans une unité de contexte ("A") le test est appliqué à une table comme suit:

	Context "A"	Other Contexts		
Word "x"	15	198	213	N_j
Other Words	572	2420	2992	
	587	2618	3205	N_{ij}
	N_i			

La formule du chi-deux, dans sa version simplifiée, est la suivante:

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

où "O" et "E" respectivement sont les fréquences observées et les fréquences théoriques.

Pour chaque cellule, les occurrences théoriques (E) sont calculées comme suit: (Ni x Nj)/Nij.

Par suite, dans l'exemple considéré la valeur du Chi-deux est égale à 19.38.

Puisqu'il est plus grand que la valeur critique, l'hypothèse nulle (absence de différence significative) peut être rejetée.

Classification (Cluster Analysis)

Ensemble de techniques statistiques qui ont le but de détecter des **groupes d'objets** avec deux caractéristiques complémentaires:

A - l'homogénéité interne la plus élevée (à l'intérieur de chaque classe);

B - l'hétérogénéité externe la plus élevée (parmi les différentes classes).

Dans le langage de la statistique, ces caractéristiques correspondent respectivement à la variance interne (within cluster variance) et à celle externe (between cluster variance).

En général, il y a deux genres de classification:

- **méthodes hiérarchiques**, dont les algorithmes reconstruisent la hiérarchie entière des objets sous l'analyse (le soi-disant "arbre"), soit dans un ordre ascendant (CAH) soit dans un ordre descendant (CDH);

- **méthodes de division**, où l'utilisateur définit précédemment les nombres de classe dans lesquels l'ensemble des objets doit être partitionné.

Dans **T-LAB** des algorithmes des deux types sont utilisés.

En particulier:

- la fonction **Analyse des Mots Associés et Cartes Conceptuelles** utilise une méthode hiérarchique;
- la fonction **Cluster Analysis** permet d'utiliser trois méthodes différentes: deux hiérarchiques et une à partitions;
- les fonctions **Analyse Thématique des Contextes Élémentaires** et **Classification Thématique des Documents** utilisent un algorithme du type bisecting K-means.

Certaines publications citées dans la **Bibliographie** permettent d'approfondir aussi bien les aspects généraux des diverses méthodes (Bolasco S., 1999; Lebart L., A. Morineau, M. Piron, 1995), que les aspects spécifiques concernant Hdbscan (Campello R. J. G. B., Moulavi D., Zimek A. & Sander J. , 2015) et la méthode bisecting K-means (Steinbach, M., G. Karypis, V. Kumar, 2000; Savaresi S.M., D.L. Boley, 2001).

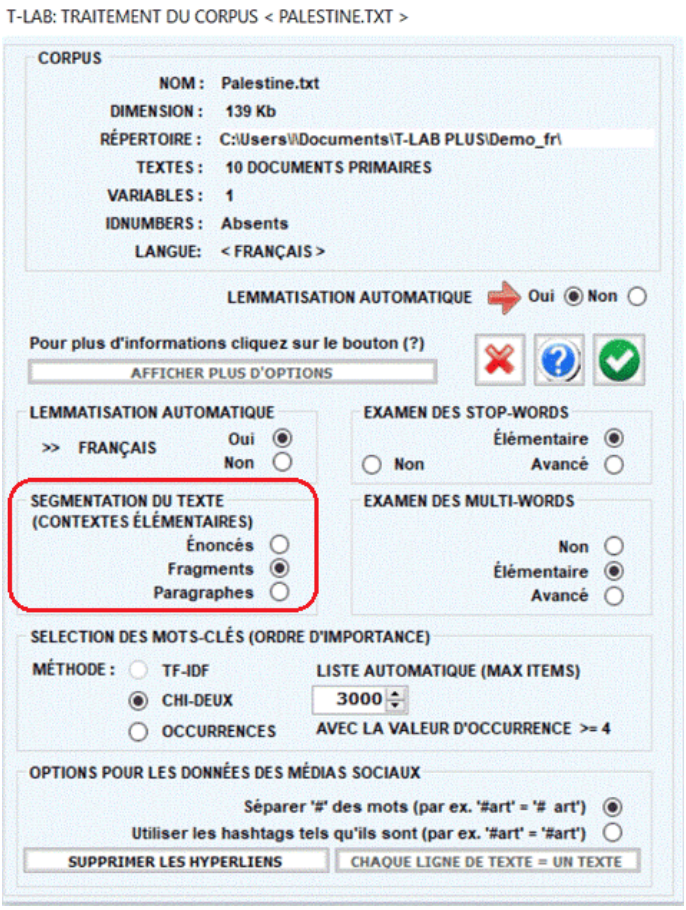
Codage

Avant l'importation du corpus, l'utilisateur peut insérer des lignes de codification au début de chaque **unité de contexte** qu'il souhaite classifier avec une ou plusieurs **variables**.

Normalement, les unités de contexte **classifiées** correspondent aux **documents primaires**.

Contextes élémentaires

Pendant la phase d'importation, **T-LAB** réalise une **segmentation** du **corpus** en **contextes élémentaires**: ceci pour faciliter les explorations de l'utilisateur et, surtout, pour rendre possibles les analyses qui requièrent le calcul des **co-occurrences**.



Selon la choix de l'utilisateur, les types de contextes élémentaires peuvent être les suivants:

1 - **Énoncés**

Contextes élémentaires marqués par ponctuation (.?!) et dont la longueur est inférieure à 1.000 caractères (minimum : 50 caractères).

2 - **Fragments**

Contextes élémentaires de longueur comparable composés d'un ou plusieurs énoncés.

Dans ce cas, les règles de segmentation utilisées par T-LAB sont les suivantes:

- considérer comme contexte élémentaire chaque séquence de mots interrompue par le point à la ligne et dont les dimensions sont inférieures à 400 caractères;
- dans le cas où, dans la longueur maximale, n'est présent aucun point à la ligne, chercher, dans l'ordre, d'autres signes de ponctuation (? ! ; : ,). S'il n'y en a pas, segmenter sur la base d'un critère statistique, mais sans tronquer les unités lexicales.

3 - **Paragraphes**

Contextes élémentaires marqués par ponctuation (.?!) et par le retour de chariot, dont la longueur maximale est 2.000 caractères.

4 - **Textes Courts**

Cette option est habilitée seulement quand la longueur maximale des textes n'excède pas les 2.000 caractères (ex. réponses aux questions ouvertes).

N.B.:

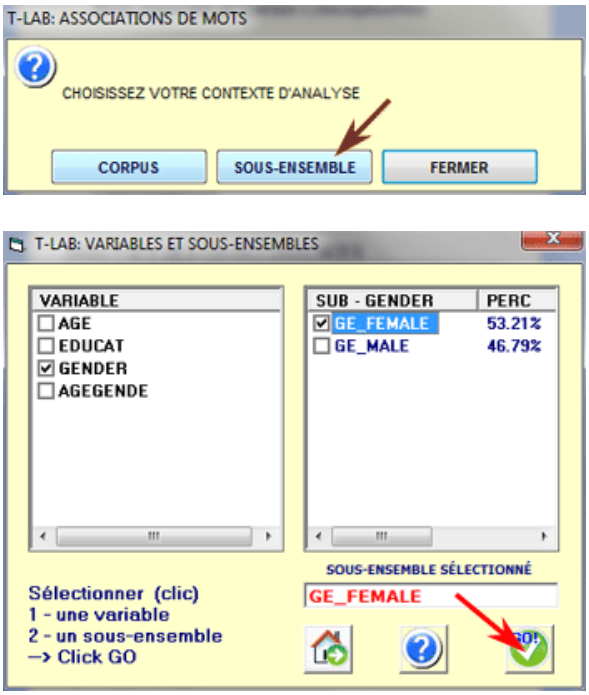
- le fichier **corpus_segments.dat** contient le résultat de la segmentation du corpus;
- l'option **Concordances** permet de vérifier les contextes élémentaires où chaque **mot** (ou **lemme**) est présent.

Corpus et Sous-ensembles

Le **corpus** est un ensemble des textes (un ou plus) rassemblés pour être analysés.

Chaque **sous-ensemble** du corpus est défini au moyen d'une modalité de quelque **variable**.

T-LAB permet d'explorer et d'analyser les relations entre les unités d'analyse de tout le **corpus** ou de ses **sous-ensembles**.



Quelques exemples de **corpus**:

- un texte ou un document qui traite un sujet quelconque;
- un ensemble d'articles de journaux qui traitent le même sujet;
- un ou plus entretiens effectués pour le même projet de recherche;
- un ensemble de réponses à une ou plusieurs questions ouvertes d'un questionnaire;
- une ou plusieurs transcriptions de focus-group.

Quelques exemples de **sous-ensemble**:

- un ou plusieurs chapitres d'un livre;
- un ou plusieurs articles de journal édités dans la même année;
- une ou plusieurs entrevues avec la même catégorie de personnes;
- un sous-ensemble de réponses à une question ouverte.

N.B.: D'autres sous-ensembles du corpus sont les "**classes thématiques**" des documents ou des contextes élémentaires obtenus en utilisant les outils correspondants de **T-LAB**.

Dans le cas d'un corpus composé de plus d'un texte, afin d'en faire un **ensemble correctement analysable**, il faut que toutes ses pièces aient deux caractéristiques qui les rendent comparables:

- une certaine homogénéité de leurs thèmes et/ou du contexte dans lequel ils ont été produits, ceci dans le but d'obtenir des données comparables entre elles;

- un rapport équilibré entre leurs dimensions, en termes d'occurrences ou en termes de K bytes, ceci dans le but de ne pas encourir dans des anomalies statistiques.

Dans la logique de **T-LAB**, le corpus est une **base de données** organisées en **entrées** (anglais : records) et en **champs**.
Avec plus de précision, les entrées se composent des entités enregistrées (textes, segments de texte, mots) et les champs se composent des variables employées pour classifier les différentes entités (les auteurs des textes, les contextes de référence, etc.).

Voir **La Préparation du Corpus**

Désambiguïsation

Opération par laquelle on résout les cas d'**ambiguïté** sémantique, en particulier les cas d'**homographie**, c'est-à-dire des mots qui ont la même **forme graphique** mais un sens différent.

 N.B.: En **T-LAB 10**, des fonctions spécifiques pour la désambiguïsation sont implémentées dans l' outil **TextScreening**; en outre, durant la phase d' importation , T-LAB reconnaît et « distingue » trois types d'objets linguistiques :

- les noms propres;
- les **multiwords** (c.-à-d. les mots composés et les locutions);
- les temps composés des verbes.

De toute façon, **T-LAB** emploie des listes de sa base de données, construites et testées pour limiter les cas les plus fréquents d'ambiguïté (critère d'**efficacité**) et pour modérer les durées des traitements (critère d'**efficience**).

Dictionnaire

Les dictionnaires **T-LAB** sont des tableaux ou des fichiers qui contiennent des schémas de classification des unités lexicales (c'est-à-dire des mots).

Les schémas de classification, et donc les dictionnaires, peuvent être soit **linguistiques** (a), soit **thématiques** (b). Les deux peuvent être exportés et personnalisés.

Dans le cas de «a» (c'est-à-dire renommer ou regrouper les éléments de la liste de mots-clés), l'utilisateur peut se référer à l'outil **Personnalisation du Dictionnaire**.

Dans le cas de «b» (c'est-à-dire exporter / utiliser un dictionnaire pour une classification supervisée), l'utilisateur peut se référer à n'importe quel outil **T-LAB** pour l'analyse thématique (par exemple, **Classification basée sur des Dictionnaires**, **Classification Thématique des Documents**, etc.).

Document Primaire

Les documents primaires sont des textes (ou parties du corpus) qui correspondent aux unités de contexte précédées d'une ligne de **codification**.

Selon les cas, il peut s'agir de: livres ou chapitres de livres, articles de quotidiens, transcriptions d'interviews, réponses à des questions ouvertes etc.

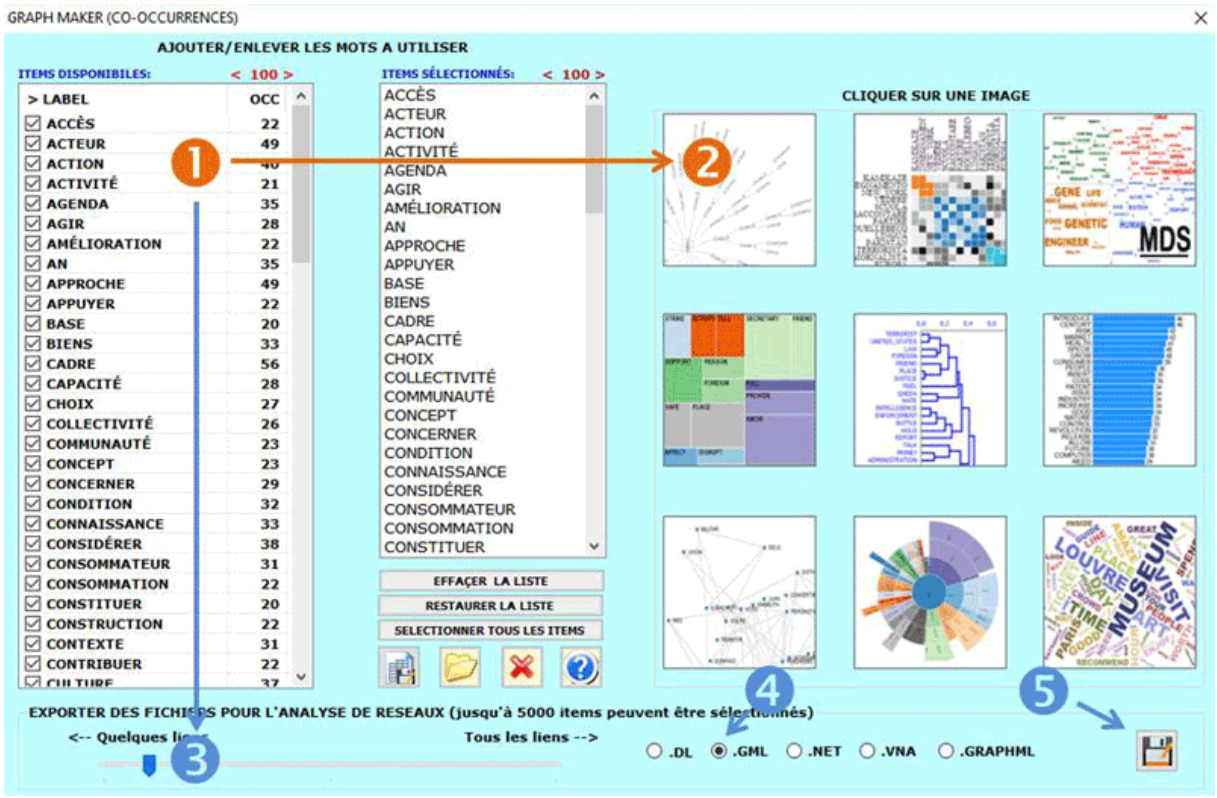
Graph Maker

L'outil **Graph Maker** permet à l'utilisateur de créer et d'exporter plusieurs graphiques dynamiques au format HTML qui peuvent être utilisés pour deux objectifs :

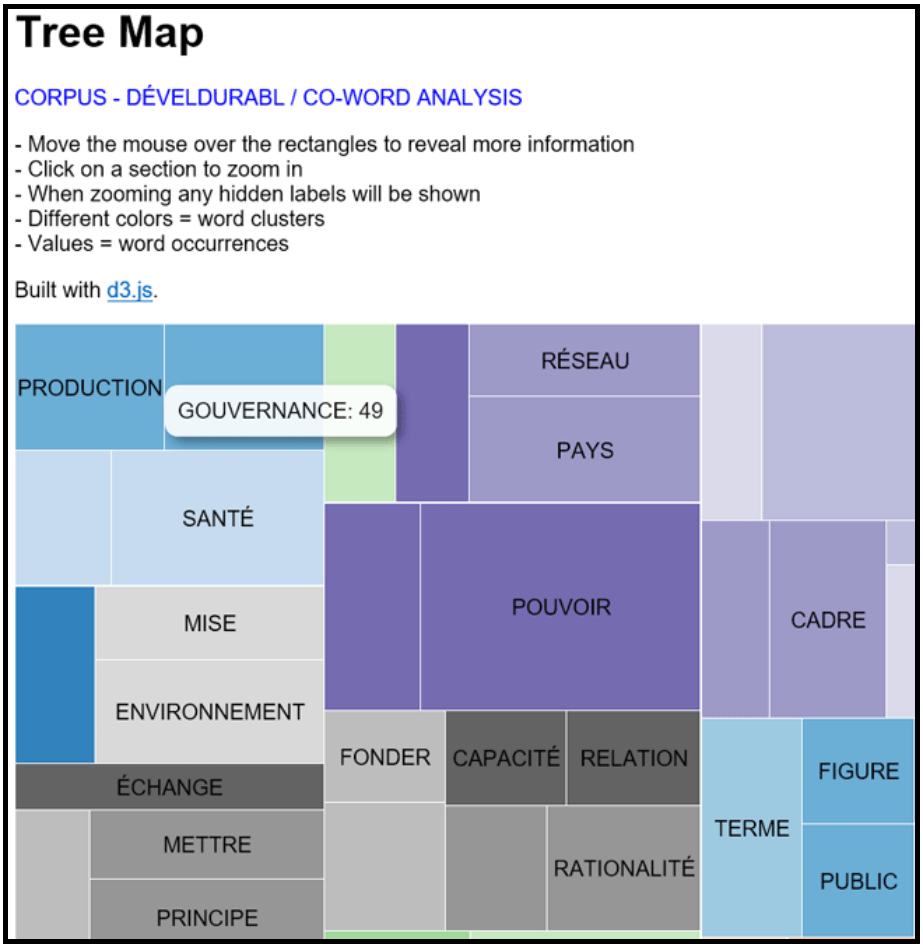
(a) explorer les relations de **co-occurences** entre des mots;
(b) effectuer un quelque type de **network analysis**.

Dans le cas (a) il faut seulement deux passages (voir l' image qui suit):
1- sélectionner les items (c'est-à-dire les mots-clés) à utiliser;
2 - cliquer une image quelconque pour visualiser le graphique correspondant .

Dans le cas (b), après la sélection des mots-clés (voir point '1' ci-dessous) , l'utilisateur peut filtrer les liens à utiliser (voir point '3' ci-dessous) , et donc il peut choisir le format de l'output, (voir point '4' ci-dessous), et cliquer sur le bouton 'sauve' (voir point '5' ci-dessous).



NB.: Chaque output en format HTML inclut des instructions faciles qui aident l'exploration (voir image suivante).


Tree Map
CORPUS - DÉVELDURABL / 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.
PRODUCTION
GOUVERNANCE: 49
RÉSEAU
PAYS
SANTÉ
POUVOIR
CADRE
MISE
ENVIRONNEMENT
FONDER
CAPACITÉ
RELATION
ÉCHANGE
METTRE
PRINCIPE
RATIONALITÉ
TERME
FIGURE
PUBLIC

Homographes

Deux mots ou plus sont **homographes** quand ils ont la même forme graphique (c'est-à-dire qu'ils sont écrits de la même manière) mais ont un sens différent.

Dans la langue italienne et celle française il y a des milliers d'homographes. Dans **T-LAB** sont implémentées des routines de **désambiguïsation** qui réduisent leur impact. En particulier, la normalisation des **multiwords** et des verbes composés.

Ainsi, par exemple, - dans la langue italienne -, la normalisation de la locution "il punto di vista" ("il_punto_di_vista") nous permet de distinguer les occurrences de "punto" et "vista" (deux homographes typiques). Ainsi - dans la langue française - la normalisation de la locution "aide-mémoire " nous permet de distinguer les occurrences de l'homographe "mémoire ".

IDnumber

IDnumber est un label qui peut être inséré dans les lignes de codification comme identificatif des sujets (ex dans le cas de réponses à des questions ouvertes) ou des unités de contexte dans lesquelles est subdivisé le corpus à importer (voir **Préparation du corpus**).

Dans **T-LAB**, chaque fois qu'il est utilisé, le label "IDnumber" doit être suivi d'un tiret bas ("_") et d'un numéro progressif de max 5 chiffres (voir exemple suivant).

**** *IDnumber *AGE_adul *SEX_fem *MET_prof

Suit le texte d'une réponse ou d'un document.
……

Chaque corpus peut inclure des numérations progressives (IDnumber) de max 30.000 sujets ou unités de contexte.

N.B.:
La première valeur de l'IDnumber doit être "1" (ex. IDnumber_00001).
Dans le cas où les textes recueillis par l'utilisateur sont en format MS Excel, dans le CD **T-LAB** une macro est disponible qui de façon automatique les transforme en un corpus codifié et prêt pour l'importation.

Index d'association

Dans **T-LAB** les indices d'association (ou de similarité) sont utilisés pour analyser les **cooccurrences** des **unités lexicales** (**LU**, lexical units) à l'intérieur des **contextes élémentaires** (**EC**, elementary contexts), c'est-à-dire des données binaires du type présence/absence.

Par exemple, étant donnés deux **LU** et dix **EC**, nous pouvons construire l'exemple suivant

	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

Les mêmes données peuvent être représentées de la façon suivante:

	LU_2		
LU_1	*Present*	*Absent*	*Total*
Present	*3*	*4*	*7*
Absent	*2*	*1*	*3*
Total	*5*	*5*	*10*

En généralisant et en utilisant les lettres de l'alphabet:

	LU_2		
LU_1	*Present*	*Absent*	*Total*
Present	*a*	*b*	*a + b*
Absent	*c*	*d*	*c + d*
Total	*a + c*	*b + d*	*n*

Les formules correspondantes aux trois indices d'associations utilisés par **T-LAB** sont les suivantes:

Jaccard

$$\frac{a}{a+b+c}$$

Dice

$$\frac{2a}{2a+b+c}$$

Cosinus

$$\frac{a}{\sqrt{(a+b)} \times \sqrt{(a+c)}}$$

Équivalence

$$\frac{a^2}{(a+b) \times (a+c)}$$

Inclusion

$$\frac{a}{\mathrm{Min}((a+b),(a+c))}$$

Information Mutuelle

$$\mathrm{Log}\ \frac{a/N}{(a+b) \times (a+c)}$$

En faisant l'hypothèse d'avoir obtenu des indices d'association des relations entre dix **LU**, nous pouvons construire un tableau comme le suivant:

	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	

De fait, **T-LAB** construit et analyse des tableaux analogues de dimensions N x N (où N peut correspondre à diverses centaines de colonnes), aussi bien à travers **Multidimensional Scaling** qu'à travers **Cluster Analysis**.

Les mêmes tableaux sont, en outre, utilisés pour calculer des index de **similarité du deuxième ordre** entre couples de mots clés (voir l'instrument **Associations de Mots**).

Isotopie

La notion d'Isotopie (ISO = même; TOPOI = endroit) se rapporte à une conception de signification comme "effet contextuel", et c'est quelque chose qui n'appartient pas aux mots considérés un à un, mais qui dérive de leurs rapports entre les textes ou les discours.

La fonction des Isotopies est celle de faciliter l'interprétation des discours (ou des textes); en fait, chaque isotopie détecte un contexte de référence partagé par un certain nombre de mots, ne résultant pas de leurs significations individuelles. Cela dans la logique que le tout est quelque chose de différent de l'addition de ses éléments. La détection d'une isotopie n'est donc pas une simple constatation d'une " donnée ", mais le résultat d'un travail d'interprétation (F. Rastier 1987).

La notion d'isotopie a été proposée par le sémiologue A.J. Greimas (1966) pour définir la récurrence, dans les mêmes unités syntagmatiques (énoncés ou textes), de mots avec des traits sémantiques (les sèmes) en commun.

Dans la logique de **T-LAB**, le relevé des isotopies dérive de l'analyse des occurrences et des cooccurrences.

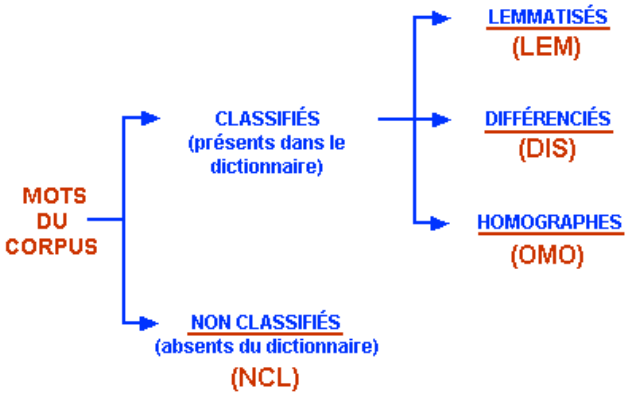
Lemmatisation

Dans les dictionnaires que nous pouvons consulter, chaque entrée correspond à un lemme qui - généralement - définit un ensemble de mots avec la même racine (ou lexème) et qui appartiennent à la même catégorie grammaticale (verbe, adjectif, etc.).

En général, la **lemmatisation** se fait de la manière suivante: les formes de verbes à l'infinitif, les noms au singulier, les adjectifs au masculin singulier et ainsi de suite.

Par exemple, les **formes fléchies** "parlait" et "parlassent", résultant d'une combinaison d'une **racine** (< parl- >) et deux différents suffixes (< - ait > et < - assent >), sont ramenées au même lemme (< parler >).
Il y a, cependant, certains cas pour lesquels la lemmatisation n'observe pas la règle de la racine, par exemple dans le cas de verbes irréguliers.

Pendant la phase d'importation du **corpus**, **T-LAB** consent d'effectuer un genre spécifique de lemmatisation automatique qui suit la logique de l' "arbre" suivant :



Évidemment, le dictionnaire de référence est celui de **T-LAB**.
Les abréviations des quatre catégories sont employées dans beaucoup de tableaux, toujours dans la colonne " INF ".
N.B. :
- la catégorie "DIS " ("à distinguer") signifie que **T-LAB** n'applique pas la lemmatisation standard, pour ne pas annuler les significations différentes au sein des différentes formes (par exemple : < bien > et <biens >) ;
- parfois, afin de différencier les homographes, **T-LAB** ajoute le caractère '_' (tiret bas) à leur lemme

Lexie et Lexicalisation

Selon Pottier (voir **Bibliographie**), la **lexie** est une expression constituée d'un ou de plusieurs mots qui se comportent comme une unité lexicale avec sens autonome.

Ses types fondamentaux sont trois: *simple*, correspondant au mot dans le sens commun du terme (ex. "*cheval*", "*mangeait*"); *composée*, constituée de deux ou plus de deux mots intégrés dans une unique forme (ex. "*biotechnologie*", "*mange-disque*"); *complexe*, constituée d'une séquence en voie de lexicalisation (ex. "*à mon avis*", "*complexe industriel*").

La **lexicalisation** est le processus linguistique à travers lequel un syntagme ou un groupe de mots devient une unité lexicale ou se comporte comme telle.

Dans **T-LAB** la fonction **Liste de Locutions** permet de construire une liste des lexies complexes présentes dans le corpus et de procéder à leur transformation en chaînes unitaires (lexicalisation).

MDS

Ensemble de techniques statistiques qui analysent les données de similitude dans le but de fournir une représentation visuelle de leurs rapports dans un espace de dimensions réduites.

Dans **T-LAB** un type de MDS (la méthode de Sammon) est employé afin de représenter les rapports parmi les unités lexicales et les rapports parmi les noyaux thématiques (voir **Analyse des Mots Associés** et **Modélisation des Thèmes Émergents**).

Les tableaux des données sont des matrices carrées qui contiennent des valeurs de proximité (dissimilarités) dérivées du calcul d'un index d'association.

Les résultats obtenus, comme ceux de l'analyse des correspondances, nous permettent d'interpréter les rapports parmi les "objets" et les dimensions qui organisent l'espace dans lequel ils sont représentés.

Le degré de correspondance entre les distances parmi les points de carte MDS et ceux de la matrice input est mesuré (inversement) par une fonction de Stress. Moins est la valeur du Stress (par ex. < 0,10), plus grande est la qualité de l'ajustement obtenu.

La formule du stress est la suivante:

$$S = \sum_{i \neq j} \frac{(d_{ij}^{*} - d_{ij})^2}{d_{ij}^{*}}$$

où d_{ij}^{*} sont les distances parmi les points (ij) de la matrice input et d_{ij} sont les distances parmi les mêmes points dans la carte MDS.

Mots-clés

Dans la logique **T-LAB** toutes les **unités lexicales** sont des Mots-clés (mots, lemmes, lexies, catégories) qui, d'une fois à l'autre, sont incluses dans les tableaux à analyser.

Opérativement, la sélection des mots-clés peut être effectuée selon deux modalités: **automatique** et **personnalisée**.
N.B.: Seulement la deuxième modalité permet de modifier la liste des unités lexicales et d'employer des **dictionnaires personnalisés**.

Mots et Lemmes

N'importe quel logiciel d'analyse des textes identifie avant tout les soi-disant **formes graphiques** (ou formes brutes), c'est-à-dire les chaînes de caractères séparées par les espaces vides.

Puis, s'accordant à leurs algorithmes spécifiques ou aux catégories employées par les spécialistes, les logiciels identifient les **lexèmes**, les **mots-clés**, etc.

Les tableaux **T-LAB**, pour toutes les unités lexicales présentes dans le database du corpus, reportent deux informations:

- la première, appelée **mot**, contient la transcription des unités lexicales (chaque mot, **lexie** ou multi-word) comme "chaînes" (en: strings) reconnues par le logiciel;

- la seconde, appelée **lemme**, contient le label avec lequel ont été regroupées et classifiées les unités lexicales.

Selon les cas, le lemme peut être:

- le résultat du processus de lemmatisation automatique;
- la rubrique d'un "dictionnaire personnalisé";
- une catégorie qui indique un groupe de synonymes;
- une catégorie d'analyses du contenu;
- etc.

Multiwords

Ensemble de deux mots ou plus qui, au niveau du signifié, fonctionnent comme **un seul mot**.

La catégorie de Multiwords, dont les bornes sont diffèrentes selon le modèle analytique utilisé, inclue des sous-ensembles aussi bien de **noms composés** (par exemple : "Ministère de la Justice") que de **locutions** (par exemple: "au fur et à mesure").

La liste des Multiwords utilisée par **T-LAB**, évidemment, n'est pas exhaustive. Elle est construite et testée selon deux critères:

a) limitation des cas d'ambiguïté les plus fréquents (critère d'**efficacité**);

b) modération des durées des traitements (critère d'**efficience**).

Dans **T-LAB** il est également possible d'employer une **liste personnalisée des Multiwords**.

N-Grammes

En **T-LAB** un n-gramme est une séquence de deux (bi-gramme) ou plus mots clés présents à l'intérieur du même **contexte élémentaire**.

Son usage est réservé au calcul des **cooccurrences** et, à l'intérieur du même contexte élémentaire, la contiguïté des mots considérés ne tient pas compte ni des "mots vides" (c'est-à-dire stop-words) ni de la ponctuation.

Prenons, par exemple, le contexte élémentaire suivant:

La loi favorise l'égal accès des femmes et des hommes aux mandats électoraux et fonctions électives, ainsi qu'aux responsabilités professionnelles et sociales.

En supposant que les treize items en rouge soient inclus dans notre liste de mots clés, la subdivision en bi-grammes produit les contextes suivants de cooccurrence:

loi & favoriser
favoriser & égal
égal & accès
etc. etc.

Différemment, dans les cas de trio-grammes le résultat serait le suivant:

loi & favoriser & égal
favoriser & égal & accès
égal & accès & femme
etc. etc.

Il est important de souligner que, dans le cas des contextes élémentaires, les cooccurrences sont basées sur la présence des mots dans le même "endroit" (par ex. phrase, paragraphe etc.); différemment, dans le cas des n-grammes, les cooccurrences sont basées sur une relation de contiguïté.

En **T-LAB** l'analyse des cooccurrences basées sur des n-grammes peut être réalisée avec l'outil **Associations de Mots**. En outre, l'analyse markovienne des bi-grammes peut être effectuée à l'aide de l'outil **Analyse des Séquences**.

Naïve Bayes

La formule du Classifieur Naive Bayes (NB) utilisée en **T-LAB** est la suivante:

$$v_{\mathbf{NB}} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j)$$

Avec:

argmax = la valeur maximum de la probabilité a posteriori;
$v_j \in V$ - se rapporte à le j-classe (v_j) de la partition (V);
$P(v_j)$ = probabilité a priori de chaque j-classe;
$\prod_i P(a_i \mid v_j)$ = produit des probabilités de chaque (a_i) mot à l'intérieur de chaque (v_j) classe.

Normalisation

Dans **T-LAB**, la normalisation du corpus a un double but:

a) la détection correcte de mots en tant que formes graphiques;

b) la solution de quelques cas d'ambiguïté.

Ceci signifie que **T-LAB**, en premier lieu, réalise un certain nombre de transformations du fichier à analyser: élimination des espaces blancs en plus, marquage des apostrophes, addition d'un espace blanc avant et après des signes de ponctuation, réduction des majuscules, etc...

Deuxièmement, **T-LAB** marque un ensemble de formes identifiées en tant que **noms propres**, convertit les formes identifiées comme **multiwords** dans des chaînes unitaires (par exemple "en quelque sorte " -> de "en_quelque_sorte"; "Ministère de la Justice" -> "Ministère_de_la_Justice").

Dans la routine de normalisation, afin d'avoir une identification correcte des formes graphiques, **T-LAB** emploie les **séparateurs** suivants:

, ; : . ! ? ' " ( ) < > + / = [ ] { }

Noyaux Thématiques

T-LAB emploie la locution **noyaux thématiques** (**N.T.**) dans des fonctions qui produisent des cartes de **mots-clés**.

Les **N.T.** sont des faisceaux de mots, **co-occurrents** dans les contextes élémentaires du corpus, qui - sur les cartes - sont représentés par des étiquettes qui peuvent être définies et changées par l'utilisateur.

Occurrences et Cooccurrences

Dans l'analyse des données textuelles les notions d'occurrence et de cooccurrence ont une importance fondamentale.

Les **occurrences** sont les quantités qui résultent du calcul de combien de fois (fréquences) une unité lexicale (**LU**, lexical unit) est présent dans un **corpus** ou dans les unités de contexte (**CU**, context units) qui le composent.

Leur distribution peut être représentée en tableaux de contingence tels que le suivant

	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

Les **cooccurrences** sont des quantités qui résultent du calcul de combien de fois deux unité lexicales sont présentes dans les mêmes contextes élémentaires (**EC**, elementary contexts).

Leur distribution peut être représentée en tableaux du type présence/absence tels que le suivant

(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

Avec une simple transformation, les tableaux du type “A” (rectangulaire) peuvent être transformés en tableaux du type “B” (carrés et symétriques) dans lesquels pour chaque couple d’unités lexicale est indiquée la quantité de leurs cooccurrences, c’est-à-dire le total de contextes élémentaires dans lesquels ils sont présents simultanément.

(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	...	

Dans **T-LAB**, la plupart des analyses des textes sont effectuées par l'étude des rapports entre des Occurrences et des Cooccurrences, par des **index d'association** spécifiques, ou par l'utilisation des techniques statistiques multidimensionnelles comme la Classification (**Cluster Analysis**) et l'**Analyse des Correspondances**.

Polarités factorielles

Dans l'**Analyse des Correspondances** chaque facteur organise une dimension spatiale qui peut être représentée par une ligne ou un axe - dont le centre (ou barycentre) est la valeur "0", et se développe d'une manière bipolaire vers l'extrémité négative (-) et positive (+), de sorte que les objets mis sur les pôles opposés sont les plus différents, presque comme la "gauche" et la "droite" sur les axes de la politique.

A ce propos il est utile de se rappeler ce que J.P. Benzecri, un mathématicien qui a donné des contributions importantes à ce genre de technique d'analyse, a écrit à ce sujet (1984):

"interpréter un axe, c'est trouver ce qu'il y a d'analogue d'une part entre tout ce qui est écrit à droite de l'origine, d'autre part entre tout ce qui s'écarte à gauche; et exprimer avec concision et exactitude, l'opposition entre les deux extrêmes" (1984, p. 302, voir la **bibliographie**).

N.B.: *Quand les graphiques factoriels sont bidimensionnels (ou tridimensionnels) les oppositions sont plus de deux: outre que l'opposition gauche-droite, il y a aussi l'opposition haut-bas. Cependant les critères d'interprétation sont les mêmes.*

Profil

Dans **T-LAB** le profil d'une **unité d'analyse** (unité lexicale ou unité de contexte) correspond au vecteur (ligne ou colonne) du tableau de données qui contient ses valeurs d'**occurrence** ou de **cooccurrence**.

N.B.:
Dans l'**Analyse des Correspondances** sont nommés **actifs** les profils (lignes ou colonnes) qui participent à la construction des axes factorielles; tandis que sont nommés **supplémentaires** ceux dont les valeurs sont calculées a posteriori.

Seuil de fréquence

Pendant la phase de prétraitement **T-LAB** calcule un seuil de fréquence pour choisir les mots (formes ou lemmes) à insérer dans la liste des **mots-clés**, utilisée dans les analyses à **configuration automatique**.

De toute façon, afin de garantir la fiabilité de tous les calculs statistiques, le seuil minimum **T-LAB** est fixé à la valeur 4.

*Pour ce calcul on emploie un algorithme documenté dans un des livres de la **bibliographie** (Bolasco, 1999).*

Il se déroule selon les étapes suivantes:

a) *détection de la gamme de basse fréquence qui, à partir de la fréquence minimum ("1") est définie par le premier "saut" dans les valeurs croissantes d'occurrences;*

b) *choix de valeur- seuil qui, selon des tailles du corpus, correspond à la valeur minimum dans le premier ou dans le deuxième décile de la gamme (10% ou 20%).*

Spécificités

Dans **T-LAB, Analyse des Spécificité**s est le nom d'un outil qui permet de vérifier les unités lexicales (c'est-à-dire : mots, lemmes ou catégories) et les contextes élémentaires (c'est-à-dire : phrases ou paragraphes) qui sont typiques (ou bien «caractéristiques») d'un texte ou d'un sous-ensemble du corpus défini par une variable catégorielle.

Les **unités lexicales** « typiques », définies par la proportion des occurrences respectives (c'est-à-dire par leur sur / sous- utilisation), sont déterminées par le calcul **Chi-Carré** ou par la **Valeur Test**.

Les **contextes élémentaires** « typiques » sont identifiés en calculant et en additionnant les valeurs **TF-IDF normalisées** assignées aux mots dont chaque phrase ou chaque paragraphe est constitué.

Stop word list (Liste des mots vides)

Dans la pratique d'analyse des textes, beaucoup de mots sont définis "vides" parce que - tout seuls - ils n'ont aucun contenu spécifique.

Un critère standard pour établir une liste de ces mots (**Stop word list**) n'existe pas.

Dans **T-LAB** la liste comporte les classes suivantes:

- prépositions
- articles
- adverbes et adjectifs indéfinis
- exclamations
- interjections
- pronoms (démonstratifs, indéfinis et relatifs)
- verbes auxiliaires et modaux.

Dans tous les cas, l'utilisateur peut importer des **listes personnalisées de StopWords**.

Tableaux de données

Les tableaux de données (ou **matrices**) se composent des lignes et des colonnes, et des valeurs enregistrées dans les cellules respectives. Elles nous permettent de synthétiser - d'une façon ordonnée - les observations à analyser (input), ou les résultats obtenus par les analyses (output).

Pour plus d'une raison, les statisticiens disent qu'une analyse réussie est due à la construction d'un "bon tableau".

T-LAB, selon les genres d'analyses, utilise trois différents types de tableaux, correspondants à autant de manières d'établir des croisements entre données en ligne et en colonne:

- lemmes (ou mots) en ligne et textes (ou **variables**) en colonne;
- textes ou segments des textes en ligne et lemmes (ou mots) en colonne;
- lemmes (ou mots) en ligne et en colonne (matrices carrées).

Tous ces types, de différentes manières, synthétisent les phénomènes d'**occurrence** et de **Cooccurrence**.

TF-IDF

Cette mesure, proposée par G. Salton (1989), permet d'attribuer un score d'importance à un terme (unité lexicale) dans un document (unité de contexte).

Sa formule est la suivante:

w *i,j* = ***tf*** *i,j x* ***idf*** *i* (*Term Frequency* x *Inverse Document Frequency*)

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Avec:

tf *i,j* = fréquence d'apparition de ***i*** (terme) dans ***j*** (document)
df *i* = nombre de documents du corpus contenant ***i***
N = nombre de documents du corpus

On peut normaliser la Fréquence du Terme (***tf*** *i,j*) dans la manière suivante:

tf *i,j* = ***tf*** *i,j* / Max (***f*** *i,j*)

où Max (***f*** *i,j*) est la fréquence maximale de **i** (un terme quelconque) dans **j** (document).

Unité d'Analyse

Les **unités d'analyse** de **T-LAB** sont de deux types: **unités lexicales** et **unités de contexte**.

A - les **unités lexicales** sont des **mots**, simples ou "multiples", archivés et classifiés sur la base d'un critère. Plus précisément, dans le database **T-LAB** chaque unité lexicale constitue un record classifié avec deux champs: **forme** et **lemme**. Dans le premier champ, appelé "forme", sont listés les mots ainsi qu'ils apparaissent dans le corpus, alors que dans le second, appelé "lemme", sont listés les labels attribués à des groupes d'unités lexicales classifiées selon des critères linguistiques (ex. **lemmatisation**) ou au moyen de **dictionnaires** et de **grilles sémantiques** définies par l'utilisateur.

B - les **unités de contexte** sont des portions de texte dans lesquelles le corpus peut être subdivisé. Plus exactement, dans la logique **T-LAB**, les unités de contexte peuvent être de trois types:

B.1 **documents primaires**, correspondants à la subdivision "naturelle" du corpus (ex. interviews, articles, réponses à des questions ouvertes, etc.), ou bien aux **contextes initiaux** définis par l'utilisateur;

B.2 **contextes élémentaires**, correspondants à des unités syntagmatiques d'une ou de plusieurs phrases et définis de façon automatique (ou semi-automatique) par **T-LAB**. Ainsi, dans le database **T-LAB** chaque document primaire se révèle être constitué d'un ou de plusieurs contextes élémentaires;

B.3 **sous-ensembles du corpus**, correspondants à des groupes de documents primaires reconductibles à la même "catégorie" (ex. interviews d' "hommes" ou de "femmes", articles d'une année particulière ou d'un titre particulier, et ainsi de suite);

Unité de Contexte

Voir **unité d'analyse**.

Unité Lexicale

Voir **unité d'analyse**.

Valeur-Test

Il s'agit d'une mesure statistique que **T-LAB** utilise pour mesurer et caractériser deux types de relations:

a) celles qui existent entre n'importe quelle unité lexicale et n'importe quelle catégorie de variables, dont les valeurs respectives d'occurrence sont reportées dans un tableau de contingence;
b) celles qui concernent n'importe quelle ligne ou n'importe quelle colonne d'un tableau de contingence avec les facteurs extraits au moyen d'une analyse des correspondances du même tableau.

Selon les relations analysées, les formules de la valeur test, extraites d'un des volumes de la bibliographie (Lebart L. Morineau A. Piron M., 1995, pp. 181-184), sont les suivantes:

a)

$$t_k(j) = \frac{n_{jk} - n_k \cdot \frac{n_j}{n}}{\sqrt{n_k \cdot \frac{n - n_k}{n - 1} \cdot \frac{n_j}{n} \cdot \left(1 - \frac{n_j}{n}\right)}}$$

où «njk» indique les occurrences à l'intérieur d'une cellule, tandis que « nj» et «nk » correspondent respectivement aux totaux marginaux de la ligne et de la colonne;

b)

$$t\alpha(j) = \sqrt{\mathrm{nj}\,\frac{\mathrm{n}\,\text{-}\,1}{\mathrm{n}\,\text{-}\mathrm{nj}}}\;\varphi\alpha\mathrm{j}$$

où « nj » et «φ α j » désignent respectivement les occurrences de l'objet j-ème et sa coordonnée sur le facteur α-ème.

La valeur de test possède deux propriétés importantes: une valeur de seuil (1,96), correspondant à la signification statistique plus couramment utilisée (p. 0,05), et un signe (- / +). Ceci veut dire que, en ordonnant les valeurs en ordre croissant ou décroissant, on peut rapidement identifier l'importance de chaque élément analysé.

T-LAB permet une consultation interactive des **tableaux avec les valeurs-test**.

Variables et Modalités

Dans **T-LAB** les **variables** sont des étiquettes (clés) employées pour identifier et classifier n'importe quel sous-ensemble du **corpus**: noms de caractéristiques identifiant toutes sortes de sujets, textes et contextes.

Chaque variable a deux **modalités** ou plus, chacune d'entre elles - de façon univoque - correspond à un élément de codage. Par exemple, la variable "sexe" a deux modalités (femelle et mâle).

Dans **T-LAB**, chaque texte ou segment (sous-ensemble du corpus) peut être classé pour **un maximum de 50 variables**. Évidemment, pour chacune d'elles, on doit définir les modalités respectives (max 150), suivant les instructions illustrées dans la **préparation du corpus**.

Afin d'obtenir davantage d'informations, voyez les exemples dans les dossiers démo.

BIBLIOGRAPHIE

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*

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)

Carroll J.B. (1964): *Language and Thought, Englewood Cliff NJ, Prentice Hall*

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., D.L. Boley (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)