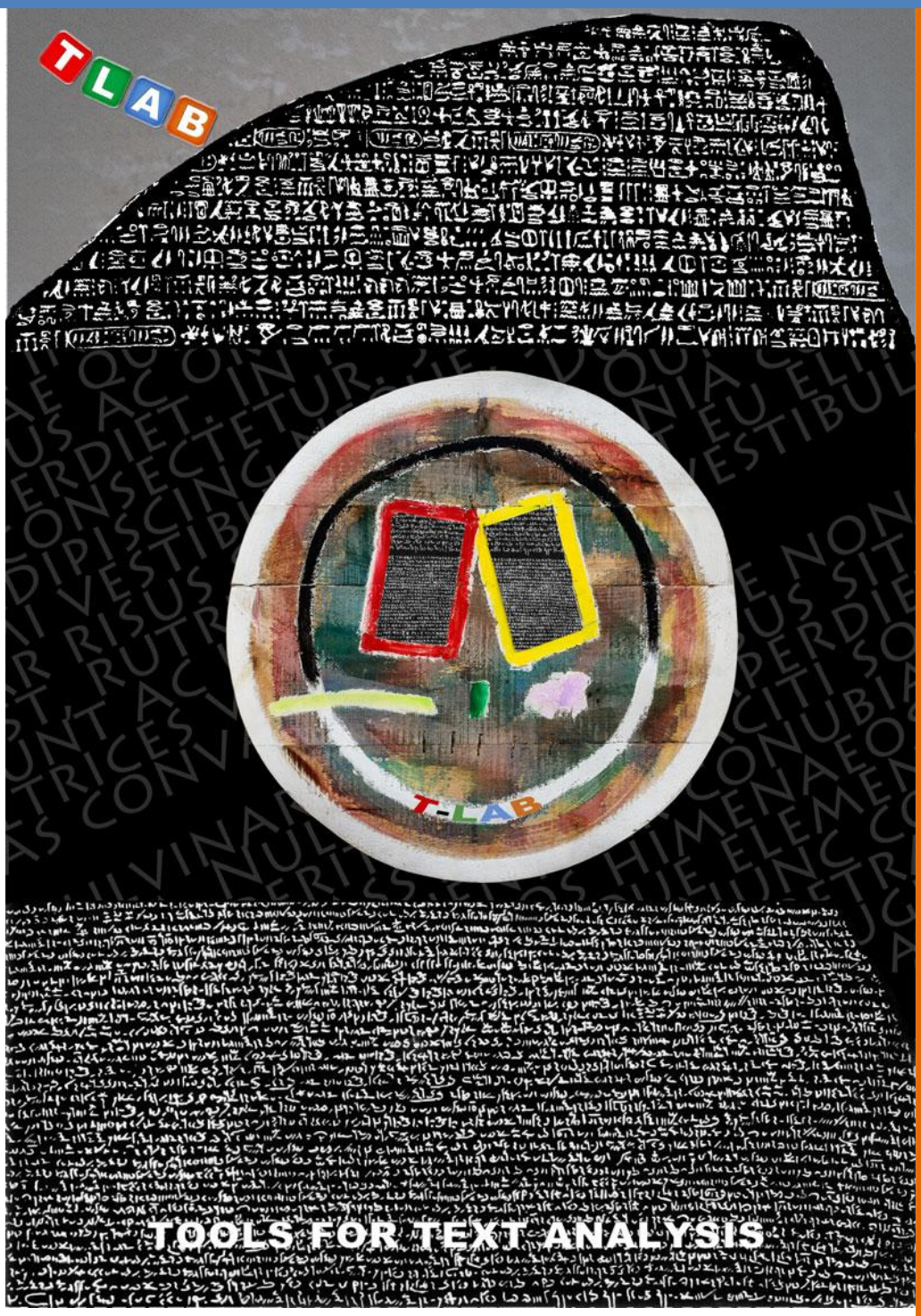
T-LAB 10

Manual del Usuario



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

Herramientas para el Análisis de Textos

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

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

T-LAB is a registered trademark

ÍNDICE

Instalación y requisitos del sistema....4
Qué hace y qué permite hacer....5
CONFIGURACIONES DE ANÁLISIS....37
Configuración Automática y Configuración Personalizada....38
Personalización del Diccionario....44
ANÁLISIS DE CO-OCURRENCIAS....47
Asociaciones de Palabras....48
Análisis de Co-Palabras y Mapas Conceptuales....59
Comparaciones entre Parejas de Palabras-Clave....70
Análisis de Secuencias y Análisis de Redes....76
Concordancias....90
ANÁLISIS TEMÁTICOS....93
Análisis Temático de Contextos Elementales....94
Modelización de Temas Emergentes....117
Clasificación Temática de Documentos....130
Clasificación basada en Diccionarios....134
Textos y Discursos como Sistemas Dinámicos....149
ANÁLISIS COMPARATIVOS....167
Especificidades....168
Análisis de Correspondencias....177
Análisis de Correspondencias Múltiples....185
Cluster Analysis....187
Descomposición de Valores Singulares (SVD)....194
PREPARACION DEL CORPUS....197
Preparación del Corpus....198
Criterios Estructurales....199
Criterios Formales....200
ARCHIVO....202
Importar un único archivo....203
Preparar un Corpus (Corpus Builder)....208
Abrir un Projecto ya existente....218
HERRAMIENTA LEXICO....219
Text Screening / Desambiguación de Palabras....220
Vocabulario del Corpus....223
Palabras Vacías....225
Multi-palabras....227
Segmentación de Palabras....229
OTRAS HERRAMIENTAS....231
Variable Manager....232
Búsqueda avanzada en el Corpus....235
Clasificación de nuevos documentos....237
Contextos Clave de Palabras Temáticas....239
Exportar Tablas Personalizadas....243
Editor....247
Importar-Exportar una Lista de Identificadores....248

GLOSARIO 250
Análisis de Correspondencias 251
Cadenas de Markov 252
Chi-cuadrado 253
Cluster Analysis 254
Codificación 255
Contextos Elementales 256
Corpus y Subconjuntos 258
Desambiguación 260
Diccionario 261
Documentos Primarios 261
Especificidad 262
Graph Maker 263
Homógrafos 265
IDnumber 266
Índices de Asociación 267
Isotopía 269
Lematización 270
Lexia y Lexicalización 271
MDS 272
Multiwords (Multipalabras) 273
N-gramas 274
Naïve Bayes Clasificador 275
Normalización 276
Núcleos Temáticos 277
Ocurrencias y Co-ocurrencias 277
Palabras clave 279
Palabras y Lemas 279
Perfil 280
Polos de Factores 280
Stop Word List 281
Tablas de datos 282
TF-IDF 283
Umbral de frecuencia 284
Unitad de Análisis 285
Unidad de Contexto 285
Unitad Lexical 285
Valor Test 286
Variables y Modalidades 287
BIBLIOGRAFIA BASICA 288

Instalación y requisitos del sistema

Configuración mínima requerida:

- Windows 7 o superior
- 4 Gb de RAM
- pantalla Full HD (tamaño recomendado 1920 x 1080)

Instalación:

- Hacer doble clic en SETUP.EXE
- Seguir las instrucciones que aparecen en pantalla
- Salir del programa
- Esperar la respuesta para conseguir su llave de autenticación
- Para más información, ver https://www.mytlab.com/T-LAB_Installation.pdf

Qué hace y qué permite hacer

T-LAB es un software compuesto por un conjunto de **herramientas lingüísticas, estadísticas y gráficas para el análisis de los textos.** Dichas herramientas se pueden emplear en las siguientes prácticas de investigación: Análisis del Contenido, Sentiment Analisis, Análisis semántico, Análisis Temático, Minería de Textos, Mapas Perceptuales, Análisis del Discurso y Network Text Analysis.

En efecto, gracias a las herramientas de **T-LAB**, los investigadores pueden gestionar ágilmente actividades de análisis como las siguientes:

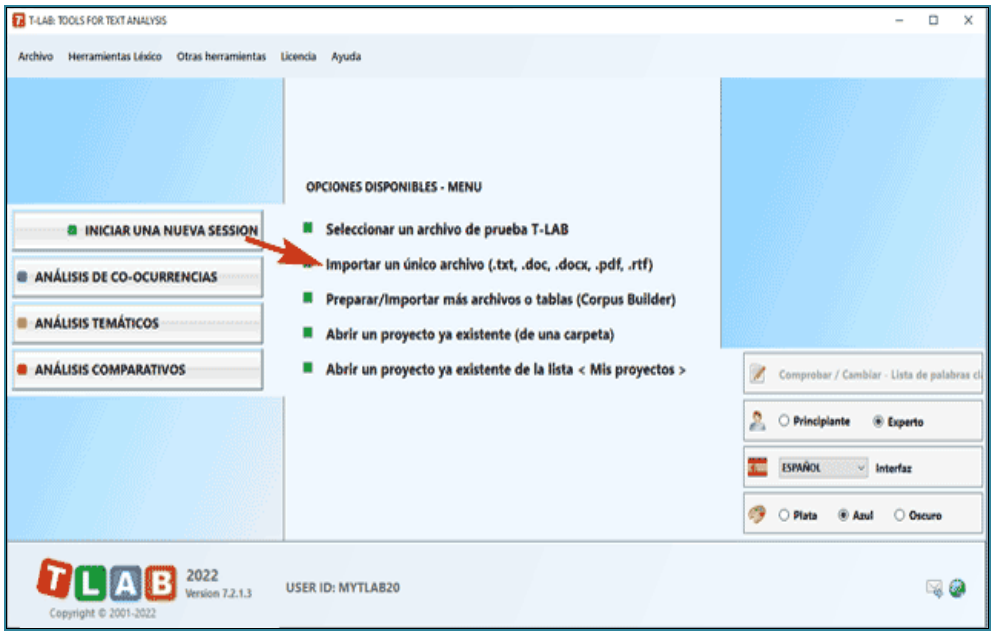
- Explorar, medir y mapear las **relaciones de co-ocurrencia entre palabras-clave**;
- Hacer una **clasificación automática** de las unidades de contexto y de los documentos, bien a través de una **metodología top-down** (es decir, mediante **categorías predefinidas**) o bien utilizando una **metodología bottom-up** (es decir, mediante el análisis de **temas emergentes**);
- Comprobar qué **unidades lexicales** (es decir, palabras o lemas), qué **unidades de contexto** (es decir, frases o párrafos) y qué **temas** son 'típicos' de subconjuntos específicos de determinados textos (p. ej.: discursos de líderes políticos, entrevistas con categorías específicas de personas, etc.);
- Aplicar categorías para la **sentiment analysis**;
- Ejecutar diferentes tipos de **análisis de las correspondencias** y de **análisis de los clústeres**;
- Generar **mapas semánticos** que representen **aspectos dinámicos del discurso** (es decir, relaciones secuenciales entre palabras o temas);
- Representar y analizar cualquier texto como si fuera una **red de relaciones**;
- Obtener medidas y representaciones gráficas sobre **textos y discursos tratados como sistemas dinámicos**;
- Personalizar y aplicar, tanto al análisis lexical como al análisis de contenido, **diferentes tipos de diccionarios**;
- Analizar todo el **corpus** o sólo algunos de sus **subconjuntos** (p. ej.: grupos de documentos) utilizando diferentes listas de palabras-clave
- Verificar los contextos de ocurrencia (p. ej.: **concordancias**) de palabras y lemas;
- Crear, explorar y exportar diferentes tipos de **tablas de contingencia** y **matrices de co-ocurrencias**.

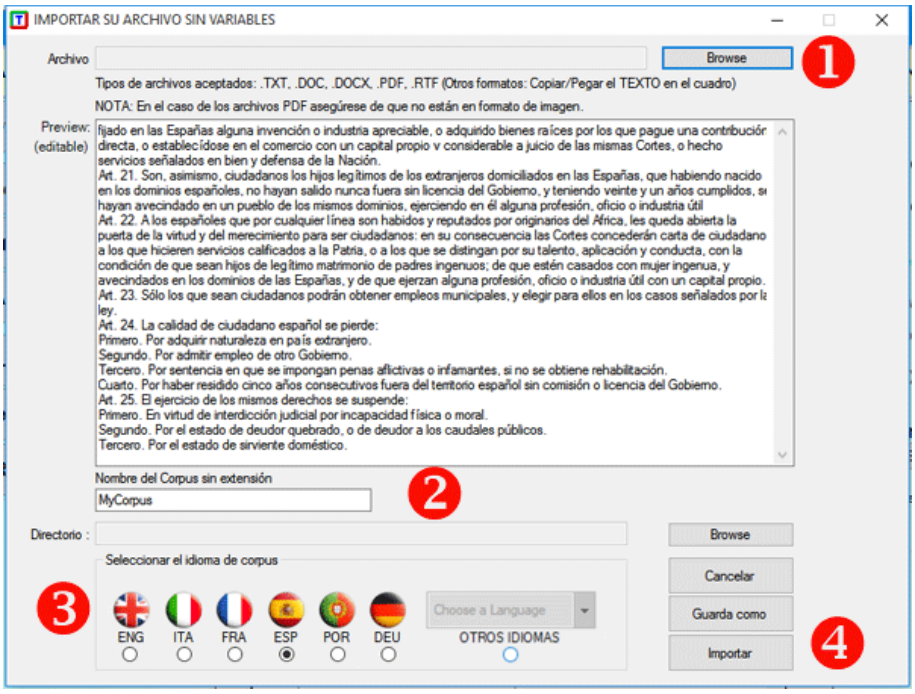
La interfaz de **T-LAB** es muy fácil de utilizar y los textos a analizar pueden ser de varios tipos:

- un único texto (ej. una entrevista, un libro, etc.);
- un conjunto de textos (ej. más entrevistas, páginas web, artículos de periódicos, respuestas a preguntas abiertas, mensajes Twitter, etc.).

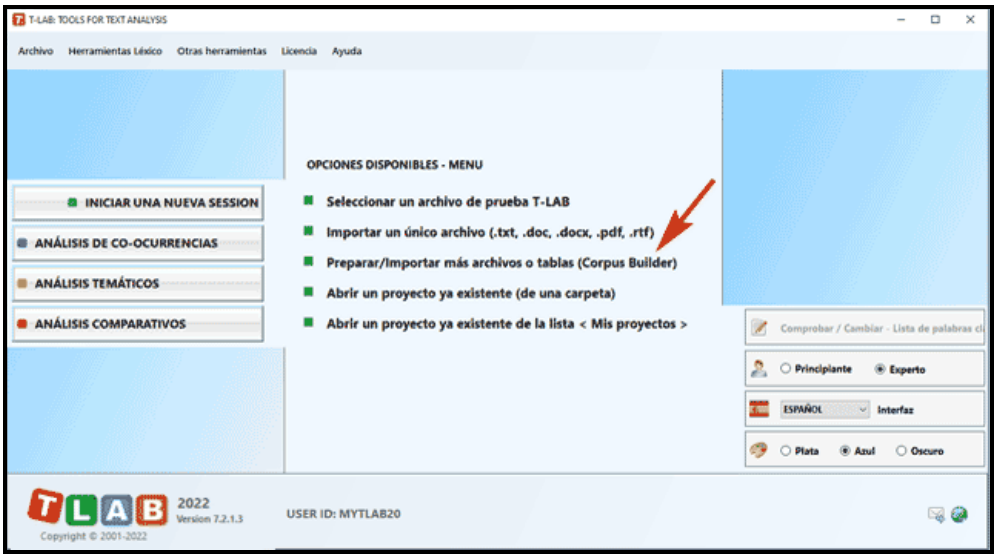
Todos los textos a analizar pueden ser codificados con **variables** categoriales y pueden incluir un identificativo (**Unique Identifier**) que corresponde a unidades de contexto o casos (ej. respuestas a preguntas abiertas).

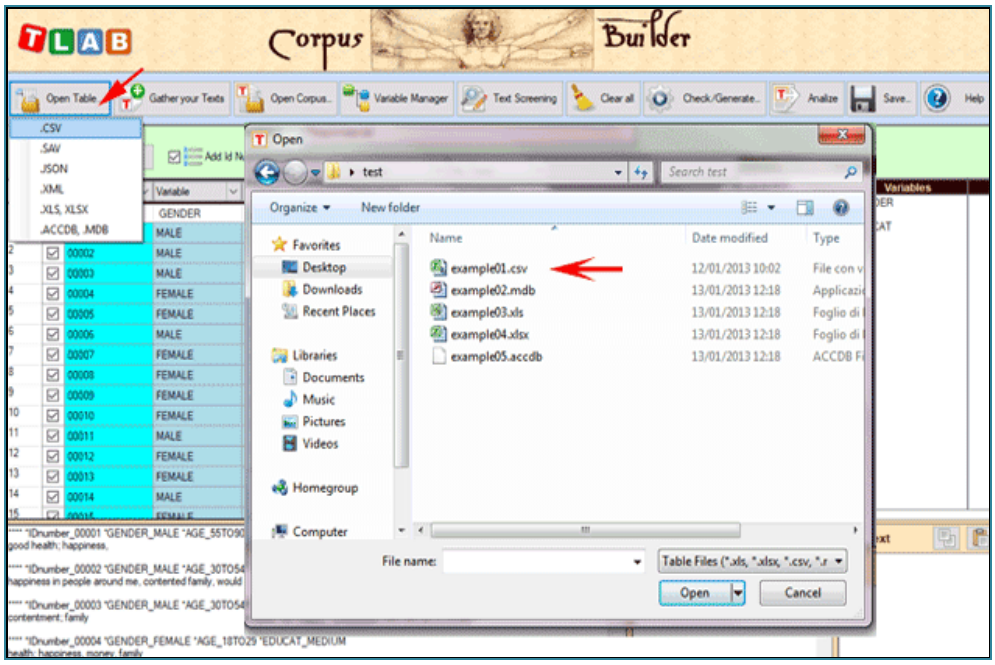
En el caso de un **único documento** (o corpus considerado como único texto), **T-LAB** no necesita nada más: es suficiente seleccionar la opción 'Importar un único archivo...'' (véase abajo).

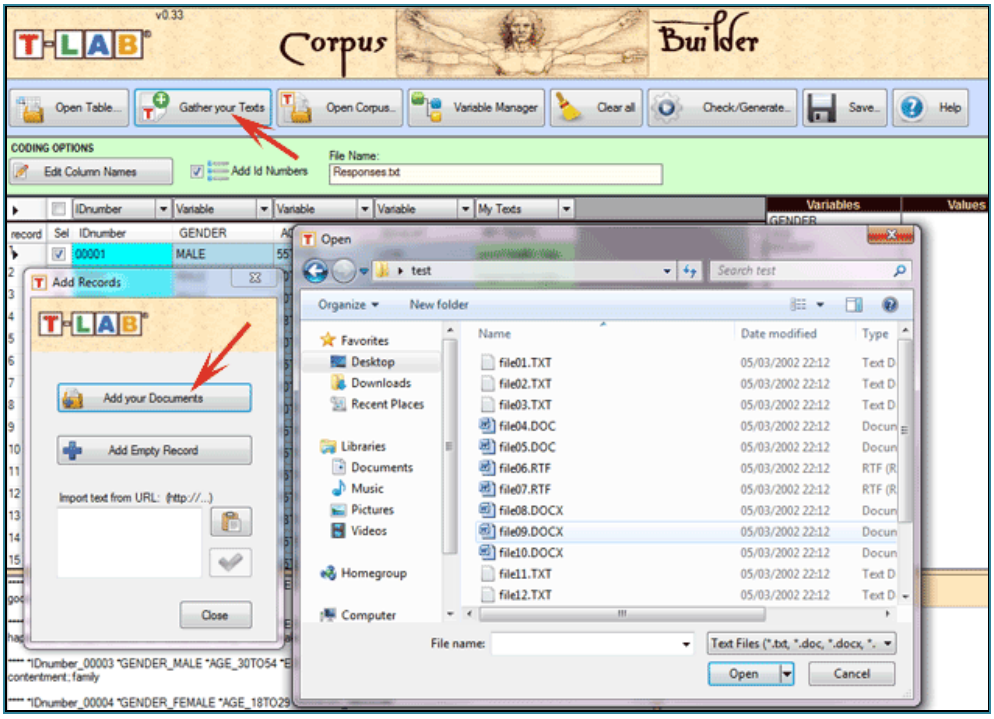




Cuando, en cambio, el corpus está compuesto por **más textos** y/o cuando se utilizan codificaciones que remiten al uso de alguna **variable**, la preparación del corpus requiere el uso del módulo **Corpus Builder** (véase abajo) que permite transformar los textos a analizar en un corpus codificado y listo para la importación.



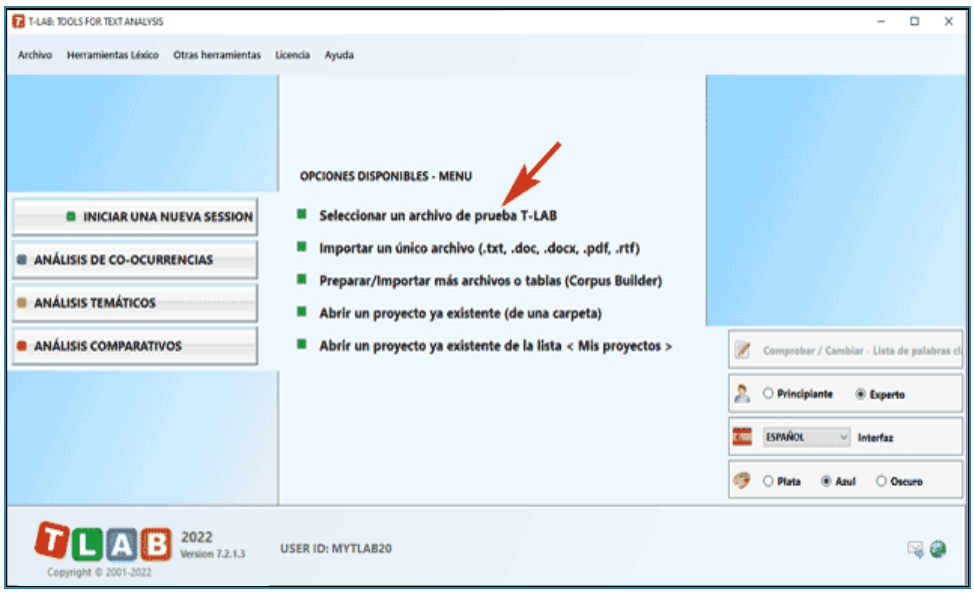




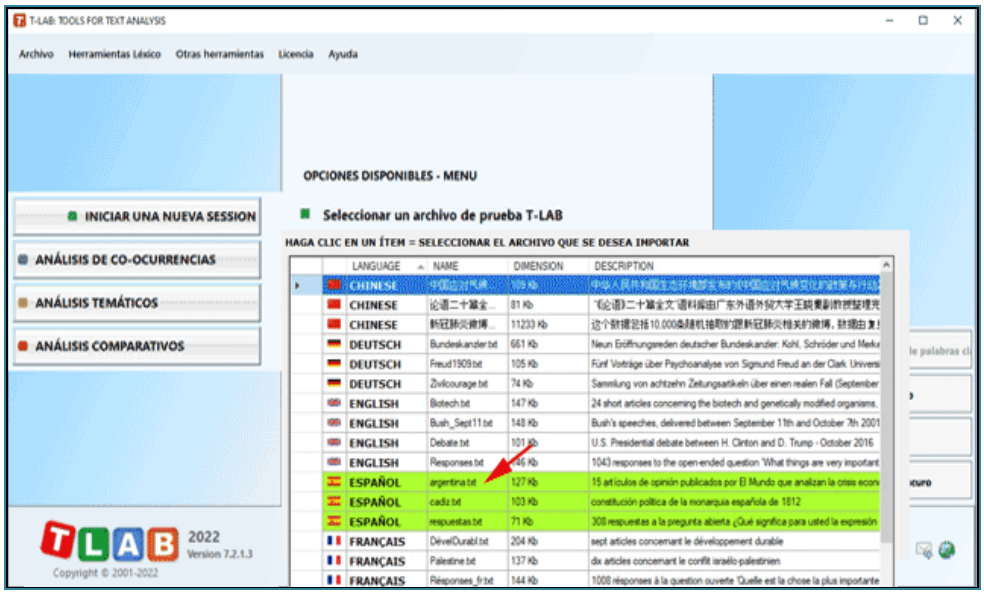
Nota: En el estado actual, **T-LAB** puede analizar archivos/corpus de hasta 90 MB de tamaño (es decir, aprox. 55.000 páginas en formato texto), garantizando el uso integrado de las distintas herramientas. Para más informaciones, véase la sección 'Requisitos y Prestaciones' del Help/Manual.

Para verificar rápidamente las funciones del software son suficientes seis pasos:

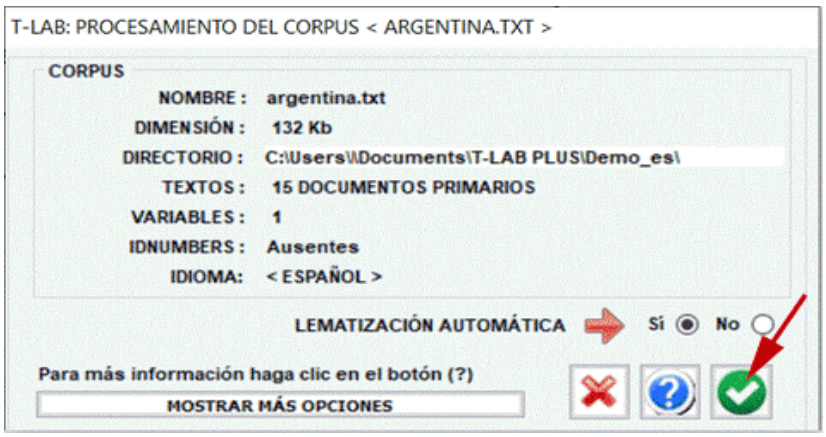
1 - Pulsar la opción 'Seleccionar un archivo de prueba T-LAB'



2 - Seleccionar un corpus a analizar

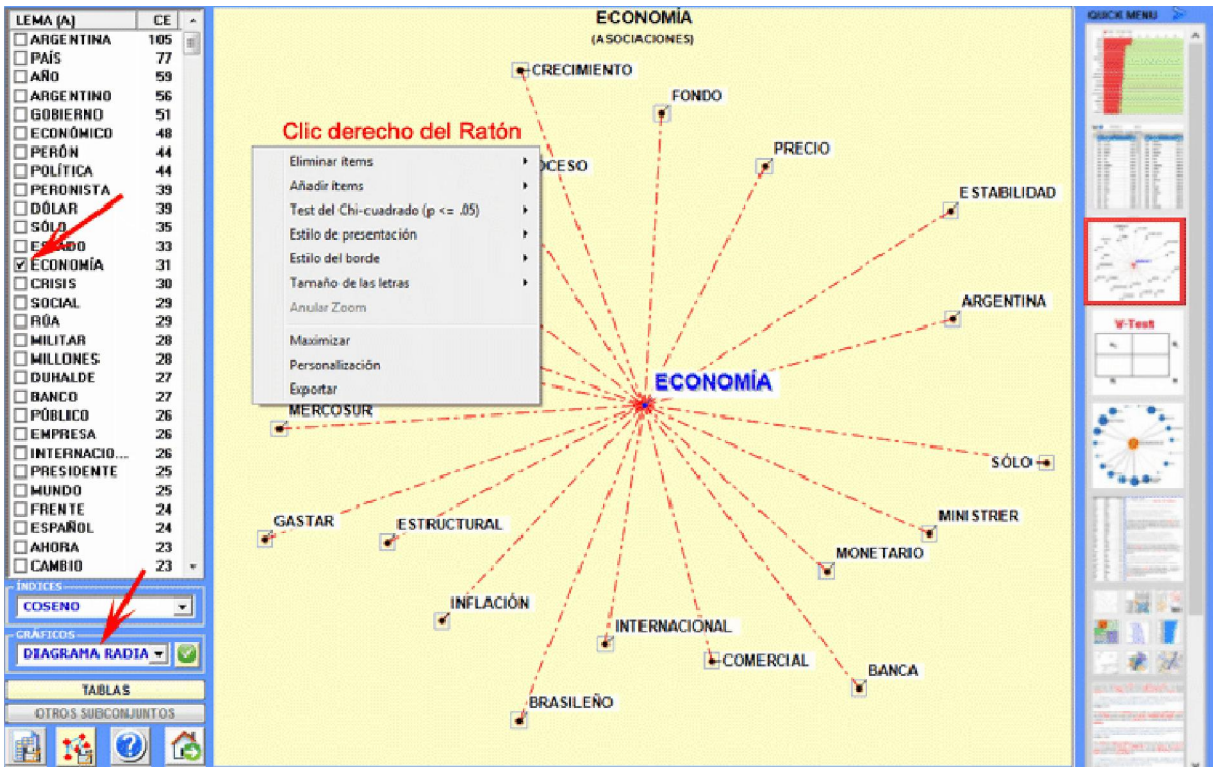


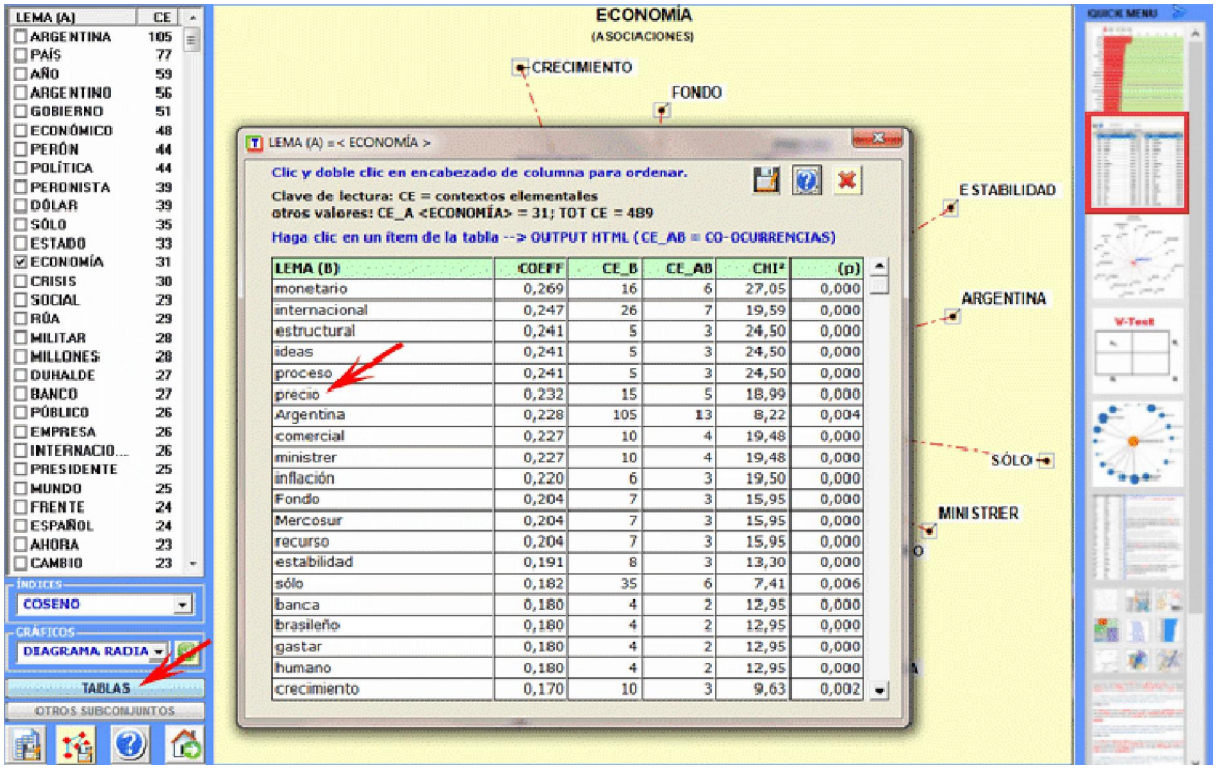
3 - Pulsar "ok" en la ventana de Configuración



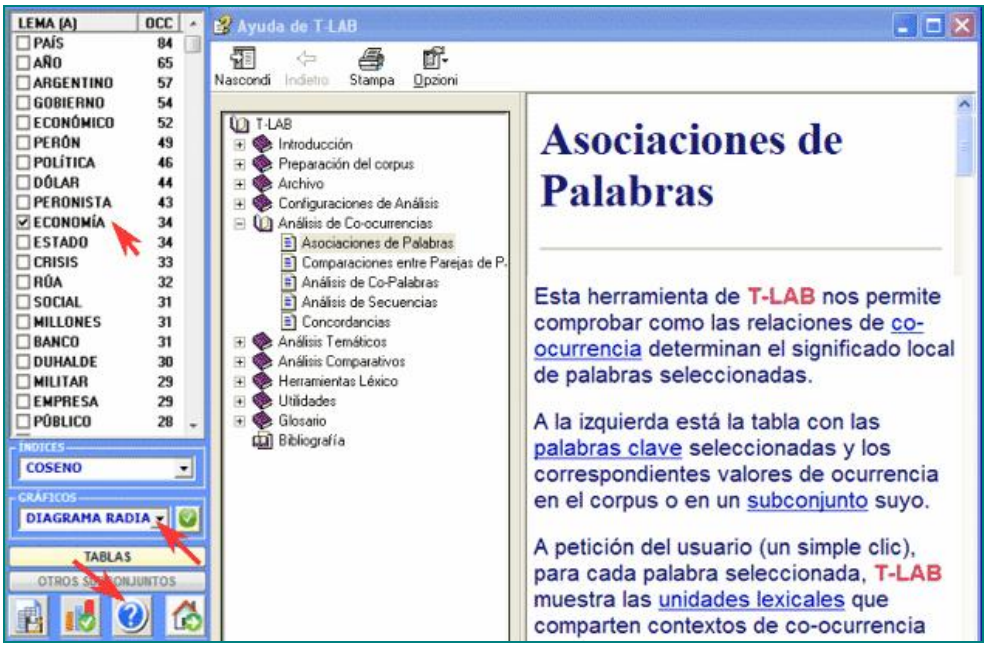
4 - Seleccionar una herramienta en uno de los submenús de "Análisis"

5 - Verificar los resultados





6 - Utilizar la ayuda contextual para interpretar gráficos y tablas.



Esta sección introductoria proporciona las informaciones básicas para entender cómo funciona **T-LAB** y cómo se puede utilizar.

Desde el punto de vista externo, el uso del software está organizado por la **interfaz**, es decir por el **menú principal**, los submenús y las **funciones** que lo componen.

Desde el punto de vista lógico, además de la interfaz, el sistema **T-LAB** está organizado por dos componentes principales:

- la **base de datos**, es decir el lugar informático en el que el **corpus** en input (es decir el texto o el conjunto de textos a analizar) está representado como un conjunto de **tablas** en las que se registran las **unidades de análisis**, sus características y sus relaciones recíprocas;

- los **algoritmos**, o sea subconjuntos de **instrucciones** que permiten usar la interfaz del usuario, consultar y modificar la base de datos, crear ulteriores tablas con datos en ésta contenidos, efectuar **cálculos estadísticos** y producir **outputs** que representen las relaciones entre los datos analizados.

Para entender cómo funciona **T-LAB** y cómo puede usarse es muy importante tener claro qué unidades de análisis se archivan en su base de datos y cuáles algoritmos estadísticos se usan en los distintos análisis. En efecto, las tablas de datos analizadas están siempre constituidas por filas y columnas cuyos membretes corresponden a las unidades de análisis archivadas en la base de datos, mientras que los algoritmos regulan los procesos que permiten descubrir relaciones significativas entre los datos y extraer informaciones útiles.

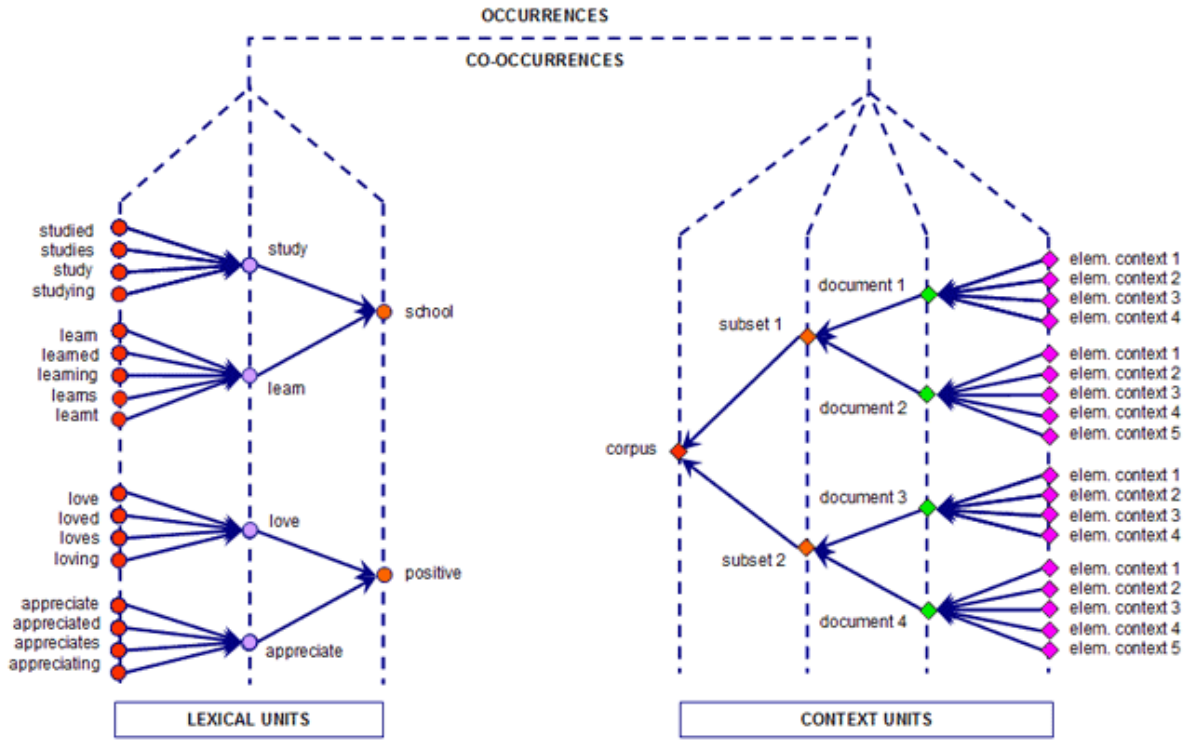
Las **unidades de análisis** de **T-LAB** son de dos tipos: **unidades lexicales** y **unidades de contexto**.

A - las **UNIDADES LEXICALES** son palabras, simples o múltiples, archivadas y clasificadas en base a un cierto criterio. En particular, en la base de datos **T-LAB**, cada unidad lexical constituye un registro clasificado con dos campos: **palabra** y **lema**. En el primer campo (**palabra**) se enumeran las palabras así como aparecen en el corpus, mientras que en el segundo (**lema**), se enumeran las etiquetas atribuidas a grupos de unidades lexicales clasificadas según criterios lingüísticos (ej. **lematización**) o a través de diccionarios y plantillas semánticas definidas por el usuario.

B - las **UNIDADES DE CONTEXTO** son porciones de texto en las que se puede dividir el corpus. En particular, en la lógica **T-LAB**, las unidades de contexto pueden ser de tres tipos:

B.1 documentos primarios correspondientes a la subdivisión "natural" del corpus (ej. entrevistas, artículos, respuestas a preguntas abiertas, etc.), o sea a los **contextos iniciales** definidos por el usuario;
B.2 contextos elementales, correspondientes a las unidades sintagmáticas (ej. fragmentos de texto, frases, párrafos) en las cuales cada documento primario puede ser subdividido;
B.3 subconjuntos del corpus que corresponden a grupos de documentos primarios atribuibles a la misma categoría (es. entrevistas de "hombres" o de "mujeres", artículos de un determinado año o de un determinado periódico, y así sucesivamente) o a clústers temáticos conseguidos a través de específicos instrumentos **T-LAB**.

El siguiente diagrama muestra las posibles relaciones que **T-LAB** nos permite analizar entre unidades lexicales y unidades de contexto.

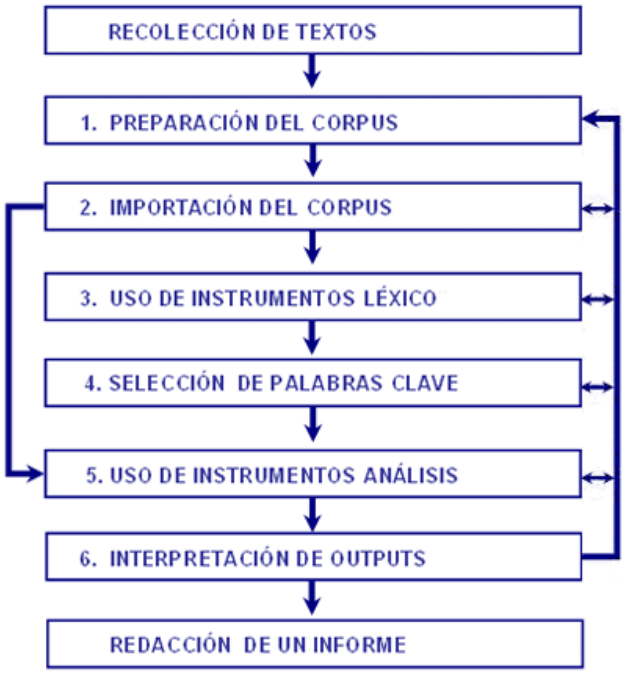


A partir de esta organización de la base de datos, **T-LAB** permite - automáticamente - explorar y analizar las relaciones entre las unidades de análisis de **todo el corpus** o de sus **subconjuntos**.

En **T-LAB**, la elección de cualquier instrumento de análisis (clic del ratón) activa siempre un proceso semi-automático que, con pocas o simples operaciones, genera algunas tablas input, aplica algún algoritmo de tipo estadístico y crea algunos outputs (ver diagrama siguiente).

Hipotéticamente, cada **proyecto** de trabajo en el que se usa **T-LAB** está constituido por el conjunto de actividades analíticas (operaciones) que tienen por objeto el mismo **corpus** y está organizado por una **estrategia** y por un **plan** del usuario. Por lo tanto, inicia con la **recolección de textos** a analizar y termina con la **redacción de un informe**.

La sucesión de las distintas fases está ilustrada en el siguiente diagrama:



NOTA:
- Las seis fases numeradas, desde la preparación del corpus a la interpretación de los output, tienen el soporte de los instrumentos **T-LAB** y son siempre reversibles;
- Por medio de las **configuraciones automáticas T-LAB** se pueden evitar dos fases (3-4); sin embargo, a los fines de la **calidad** de los resultados se recomienda la ejecución de las mismas.

Intentamos ahora comentar, una tras otra, las distintas fases.

1 - La **PREPARACIÓN DEL CORPUS** consiste en la transformación de los textos a analizar en un archivo (**corpus**) que puede ser elaborado por el software.

En el caso de un único texto (o corpus considerado como único texto), **T-LAB** no necesita nada más.

Cuando, en cambio, el corpus está formado por varios textos y hay **códigos** que hacen referencia a algunas **variables** se hace necesario utilizar el modulo **Corpus Builder,** que permite transformar los textos a analizar en un corpus codificado y listo para la importación.

NOTA:

- Al término de la fase de preparación se recomienda crear una nueva carpeta de trabajo en cuyo interior sólo se encuentre el archivo corpus a importar;

- Se recomienda, durante los análisis, mantener el archivo del corpus y la carpeta de trabajo relativa en un disco duro del ordenador donde se ha instalado **T-LAB**. De no ser así, podría ralentizarse la ejecución de los diferentes procedimientos y el programa podría llegar a mostrar errores.

2 - **LA IMPORTACIÓN DEL CORPUS** consiste en una serie de procesos automáticos que transforman el corpus en un conjunto de tablas integradas en la **base de datos T-LAB**.

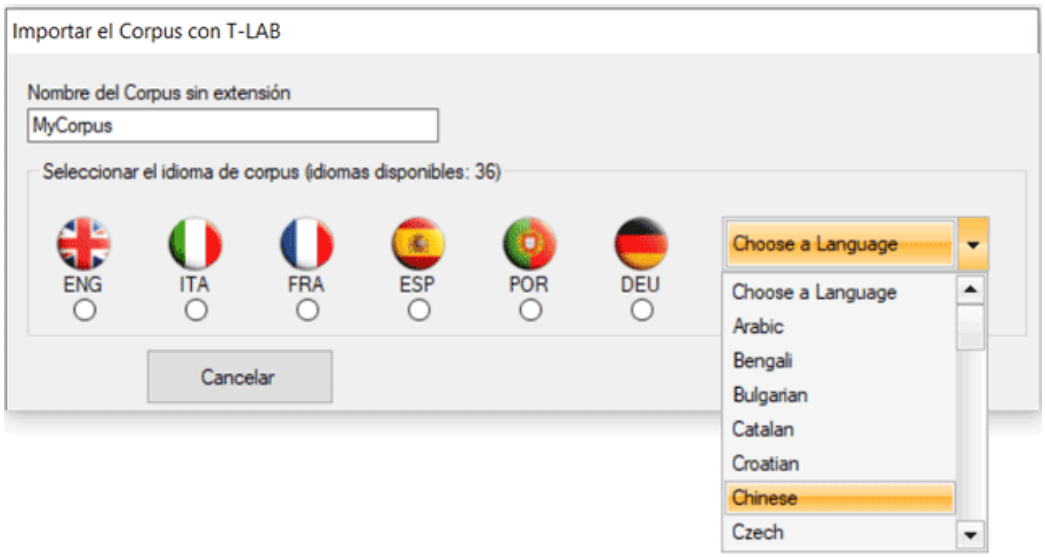
Durante el proceso de importación, **T-LAB** realiza los tratamientos siguientes: **normalización** del corpus; detección de **multi-palabras** y **palabras vacías**; segmentación en **contextos elementales**; **lematización** automática o **stemming**; construcción del **vocabulario**; selección de las **palabras clave**.

A continuación, se presenta el listado completo de los 30 idiomas para los cuales **T-LAB** prevé la posibilidad de implementar procesos de lematización automática y de stemming.

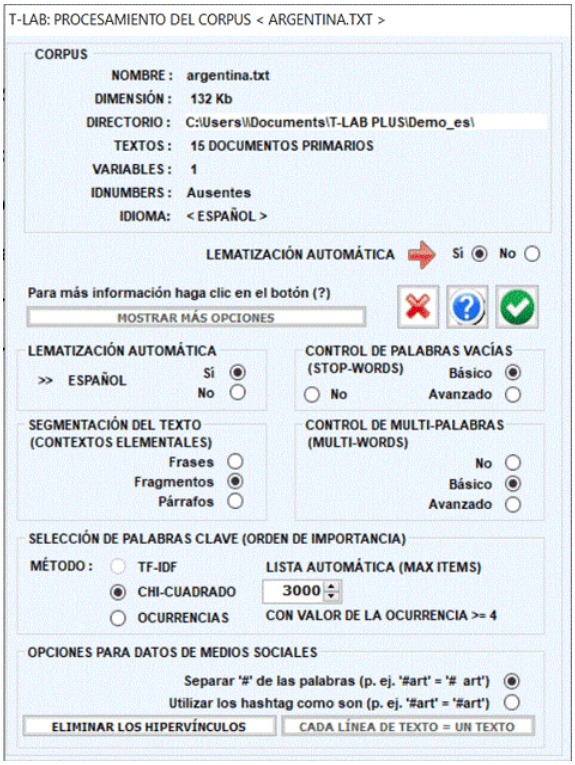
LEMATIZACIÓN: alemán, catalán, croata, eslovaco, español, francés, inglés, italiano, latín, polaco, portugués, rumano, ruso, serbo, sueco y ucraniano.

STEMMING: árabe, bengalí, búlgaro, checo, danés, finlandés, griego, hindi, húngaro, indonesio, marathi, noruego, persa y turco.

En cualquier caso, sin lematización automática y / o mediante diccionarios personalizados, el usuario puede analizar textos en **todos los idiomas**. Lo importante es que las palabras estén separadas por espacios y/o signos de puntuación.



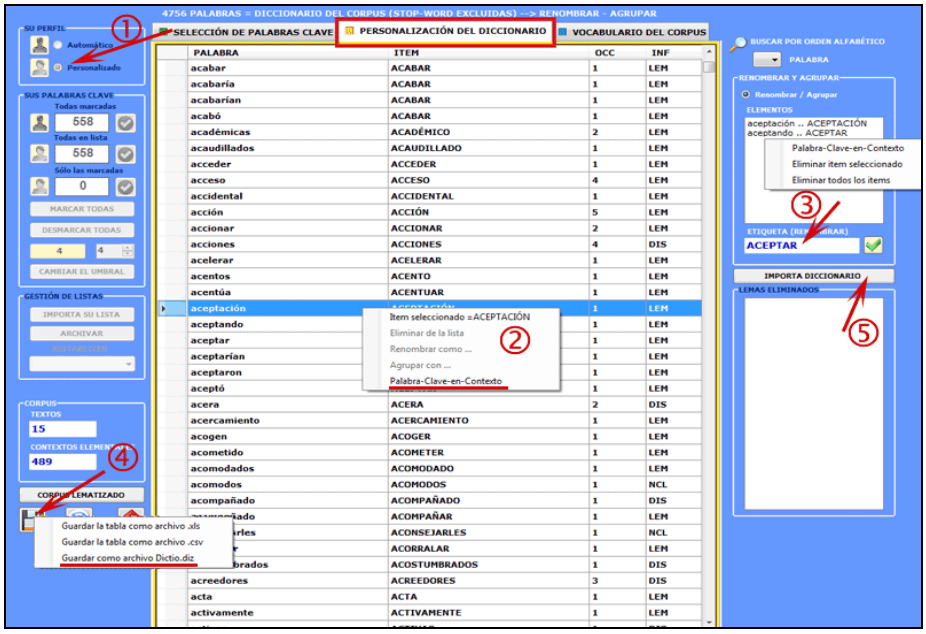
Una vez seleccionado el idioma, la intervención del usuario será necesaria para definir las opciones indicadas en la ventana siguiente:



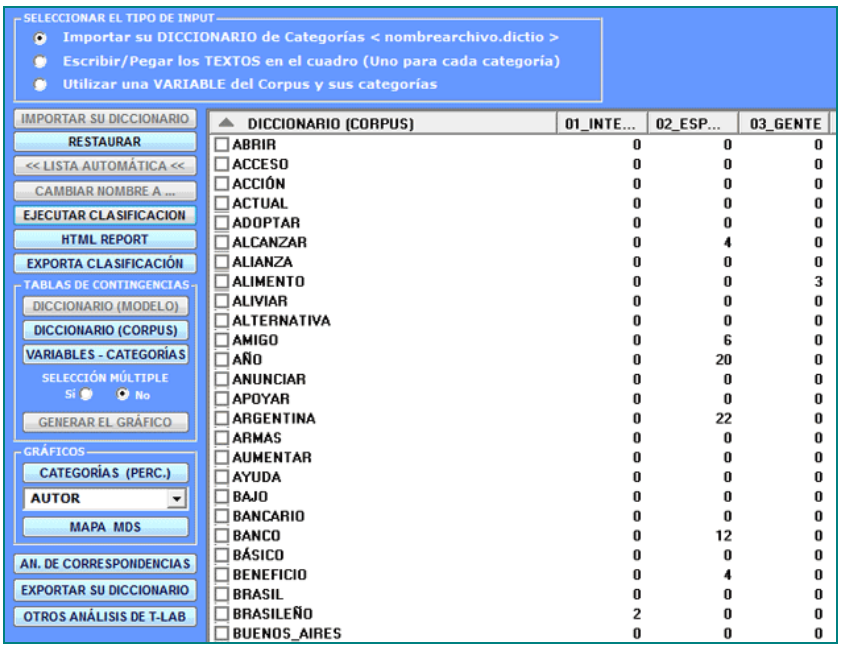
NOTA: Puesto que las diferentes opciones determinan el tipo y la cantidad de unidades de análisis (es decir las unidades de contexto y las unidades lexicales), diversas opciones determinan diversos resultados de análisis. Por esta razón, todos los outputs de **T-LAB** (es decir gráficos y tablas) utilizados en el manual del usuario y en la ayuda en red son solo indicativos.

3 - EL USO DE LAS HERRAMIENTAS LÉXICO está destinado a verificar el correcto **reconocimiento** de las unidades lexicales y a personalizar su **clasificación**, es decir a verificar y modificar las selecciones automáticas hechas por **T-LAB**.

Las modalidades de las diversas intervenciones están descritas en las correspondientes voces de la ayuda (y del manual).
En particular se redirecciona a la correspondiente voz de la ayuda (y del manual) para una descripción detallada del proceso **Personalización del Diccionario**. De hecho, cualquier modificación a las voces del diccionario (p. ej.: agrupación de dos o más ítems) incide tanto en el cálculo de las ocurrencias como en el cálculo de las co-ocurrencias)

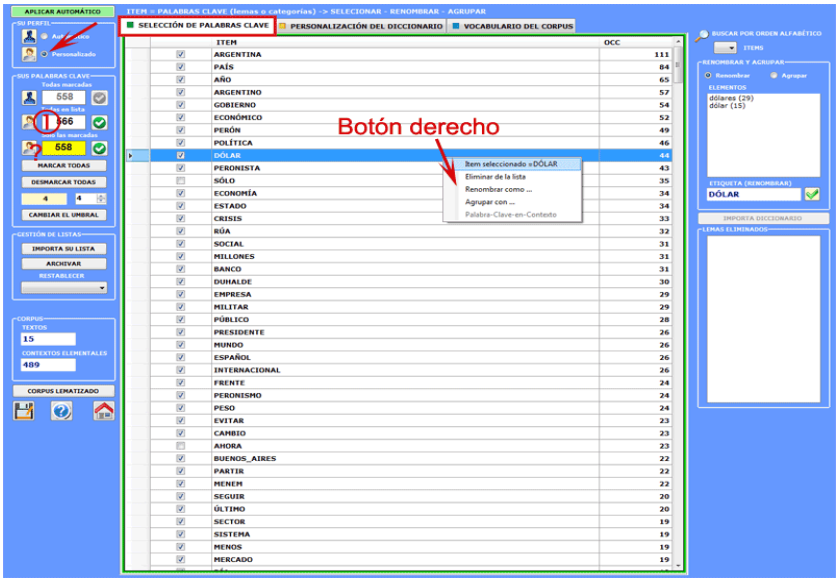


Nota: Cuando el usuario quiere aplicar esquemas de codificación que agrupen diferentes palabras o lemas en unas pocas categorías (de 2 a 50), y sin perder ninguna información lexical, se aconseja utilizar la herramienta **Clasificación basada en los Diccionarios** incluido en el submenú **Análisis Temáticos** (véase abajo).



4 - LA SELECCIÓN DE LAS PALABRAS-CLAVE consiste en la predisposición de una o más listas de unidades lexicales (palabras, lemas o categorías) a utilizar para crear las tablas de datos a analizar.

La opción **configuración automática** pone a disposición listas de **palabras clave** seleccionadas por **T-LAB**; sin embargo, dado que la elección de las unidades de análisis es muy relevante en relación a las sucesivas elaboraciones, se aconseja vivamente el uso de la **configuración personalizada**. De este modo el usuario podrá elegir la modificación de la lista sugerida por **T-LAB** y/o crear listas que correspondan mejor con sus objetivos de investigación.



En la creación de estas listas, son válidos los siguientes criterios:

- verificar la **relevancia** cuantitativa (total de las ocurrencias) y cualitativa (no banalidad del significado) de los distintos términos;
- verificar las **limitaciones** de los instrumentos analíticos que se desean utilizar;
- verificar si el conjunto de los términos es compatible con la propia **estrategia** de investigación (ver punto siguiente: 5).

5 - **EL USO DE LOS INSTRUMENTOS DE ANÁLISIS** está destinado a la producción de outputs (tablas y gráficos) que representan **relaciones significativas** entre las unidades de análisis y que permiten hacer **inferencias**.

Actualmente **T-LAB** incluye quince diversas herramientas de análisis y cada una de ellas tiene su propia lógica; es decir, cada herramienta utiliza algoritmos específicos y produce output específicos.

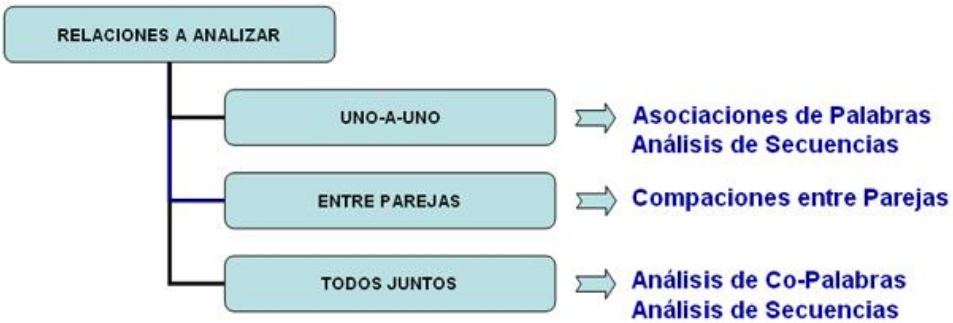
Consecuentemente, dependiendo de la tipología de textos que quiera analizar y de los objetivos que quiera alcanzar, el usuario debe decidir, cada vez que implemente una, qué instrumentos son más apropiados para su estrategia de análisis.

Para este propósito, además de la distinción entre instrumentos para **análisis de co-ocurrencias**, **análisis comparativos** y **análisis temáticos**, puede ser útil tomar en cuenta que algunos de estos nos permiten obtener nuevas **unidades del análisis** que se pueden incluir en otros procesos.

Sin embargo, teniendo en cuenta que el uso de las herramientas **T-LAB** puede ser circular y reversible, podríamos escoger tres puntos de inicio (start points) que corresponden a los tres sub-menús de ANÁLISIS:

A : INSTRUMENTOS PARA ANÁLISIS DE CO-OCCURRENCIAS

Estos instrumentos permiten analizar varios tipos de relaciones entre las palabras clave.

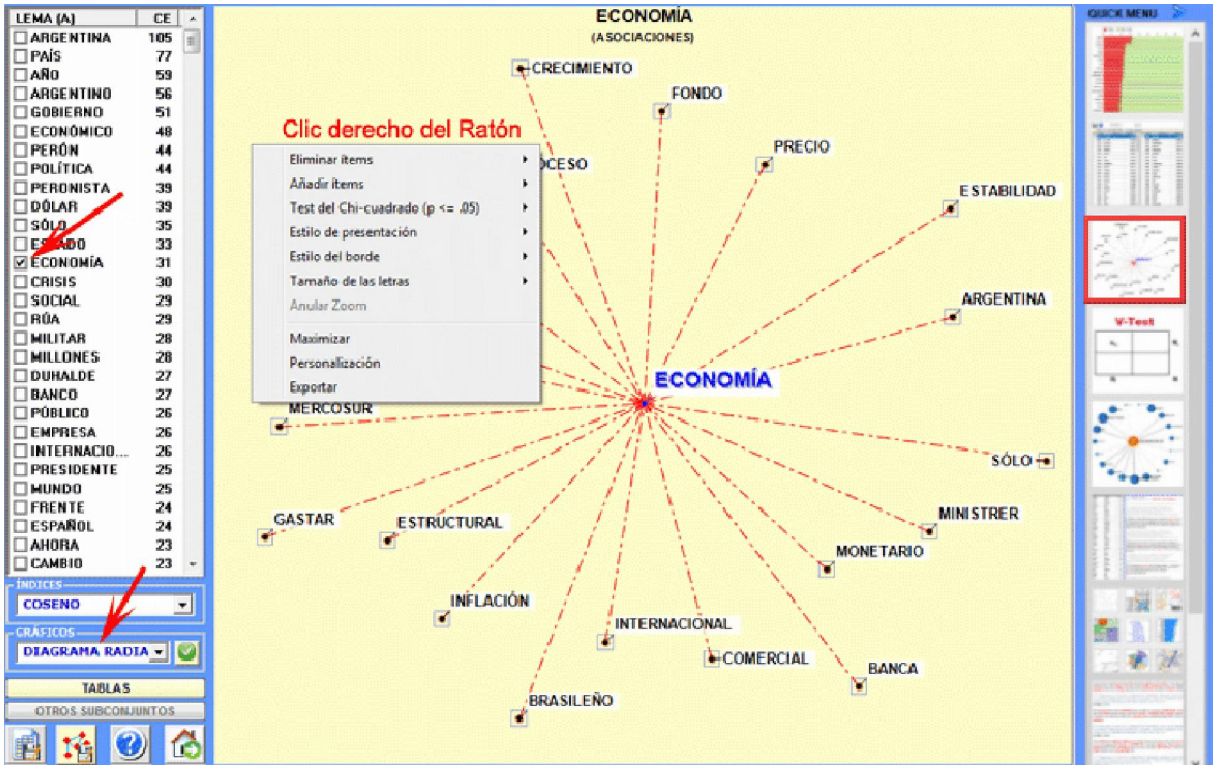


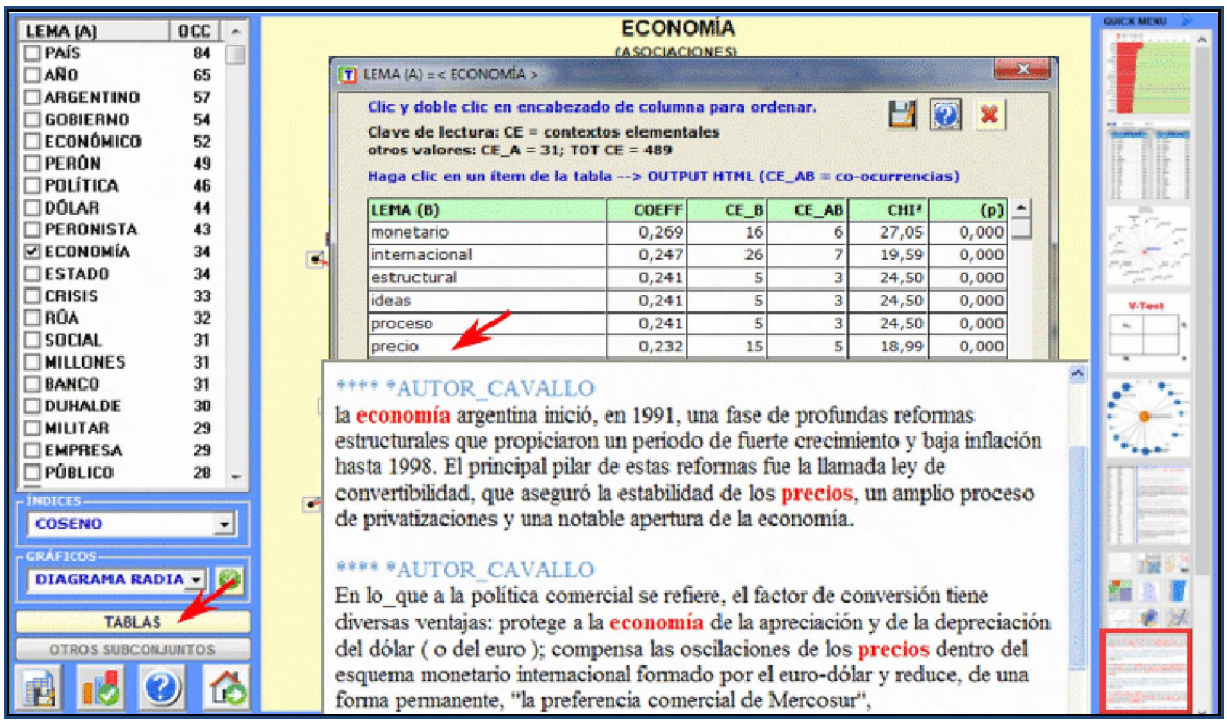
Según los tipos de relaciones a analizar, las funciones **T-LAB** indicadas en este diagrama (casillas coloradas) usan uno o más de los siguientes instrumentos estadísticos: **Índices de Asociación**, **Test del Chi Cuadrado**, **Cluster Analysis, Multidimensional Scaling**, **Principal Component Analysis, t-SNE** y **Cadenas Markovianas.**

Aquí están algunos ejemplos (Nota: para más información sobre la interpretación de los resultados, véanse las secciones correspondientes en la guía/manual):

- Asociaciones de Palabras

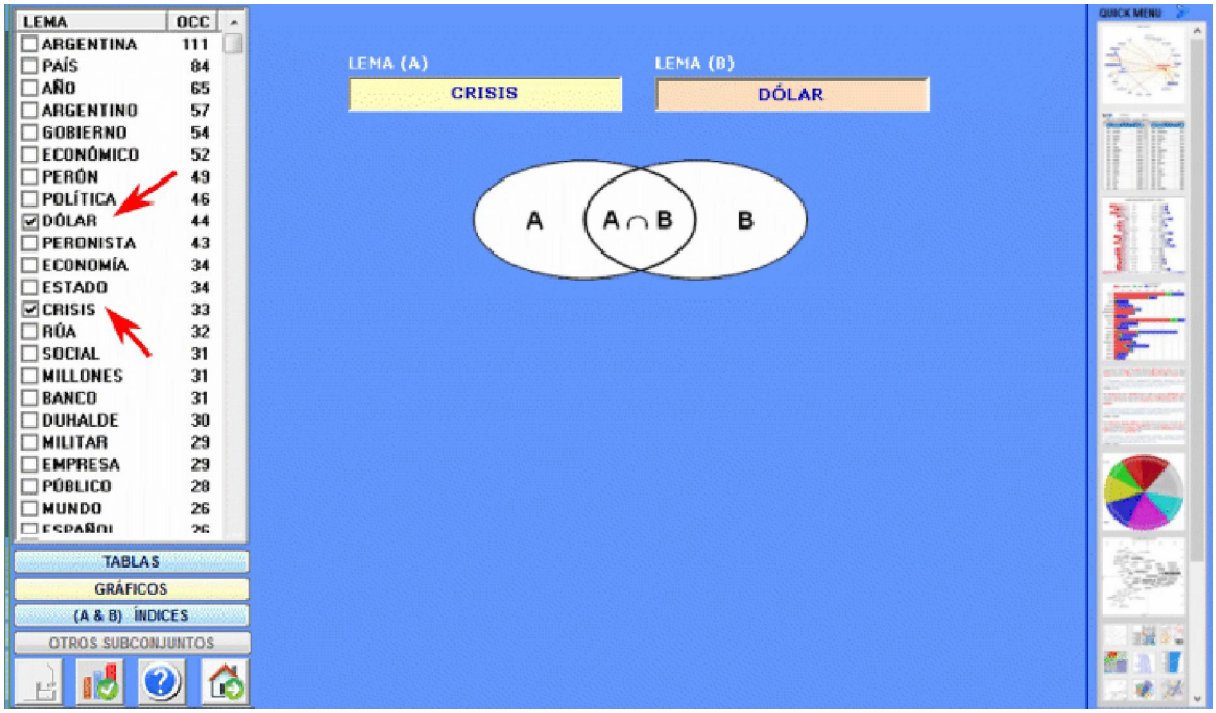
Esta herramienta de **T-LAB** nos permite comprobar como las relaciones de co-ocurrencia determinan el significado local de palabras seleccionadas.

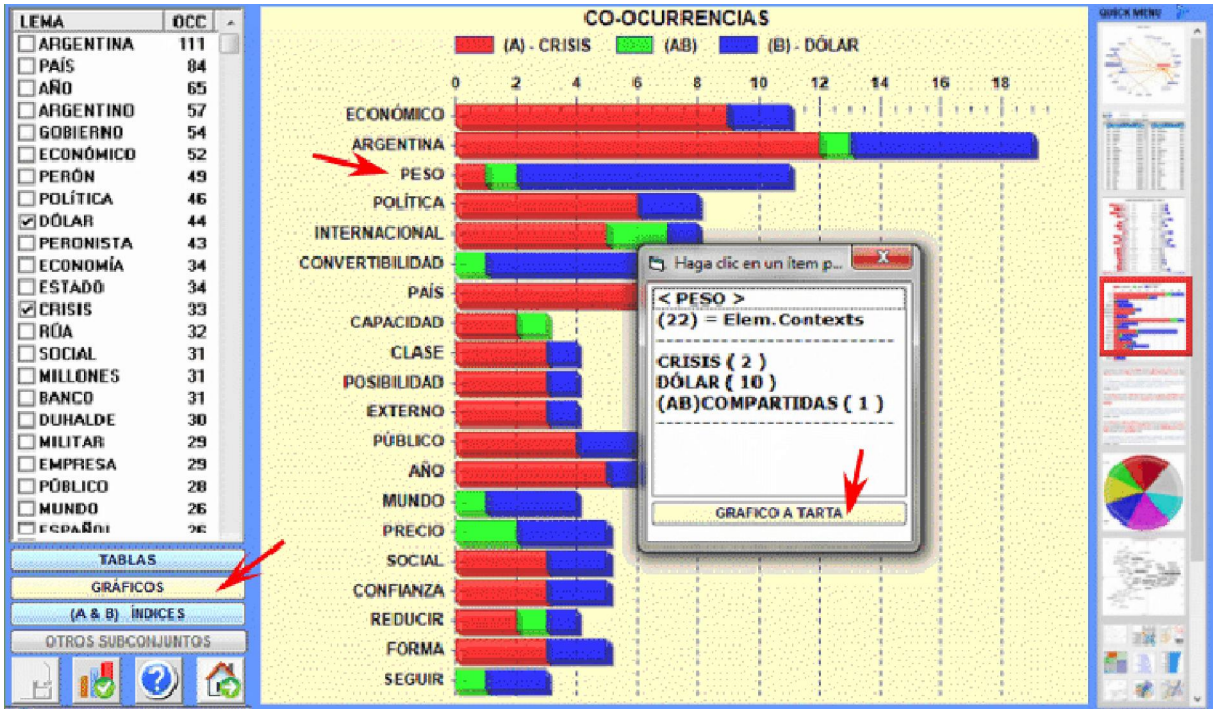




- Comparaciones entre Parejas

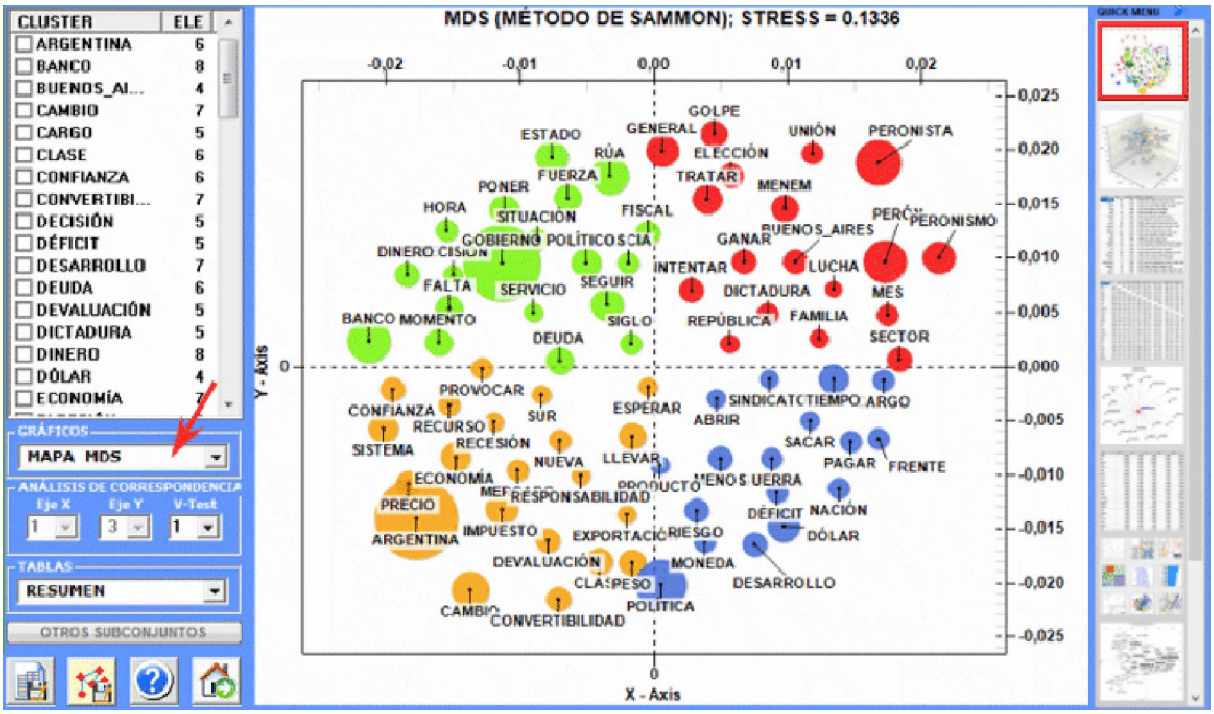
Esta herramienta de **T-LAB** nos permite comparar los conjuntos de contextos elementales (es decir contextos de co-ocurrencia) en los cuales los miembros de una pareja de palabras-clave están presentes.

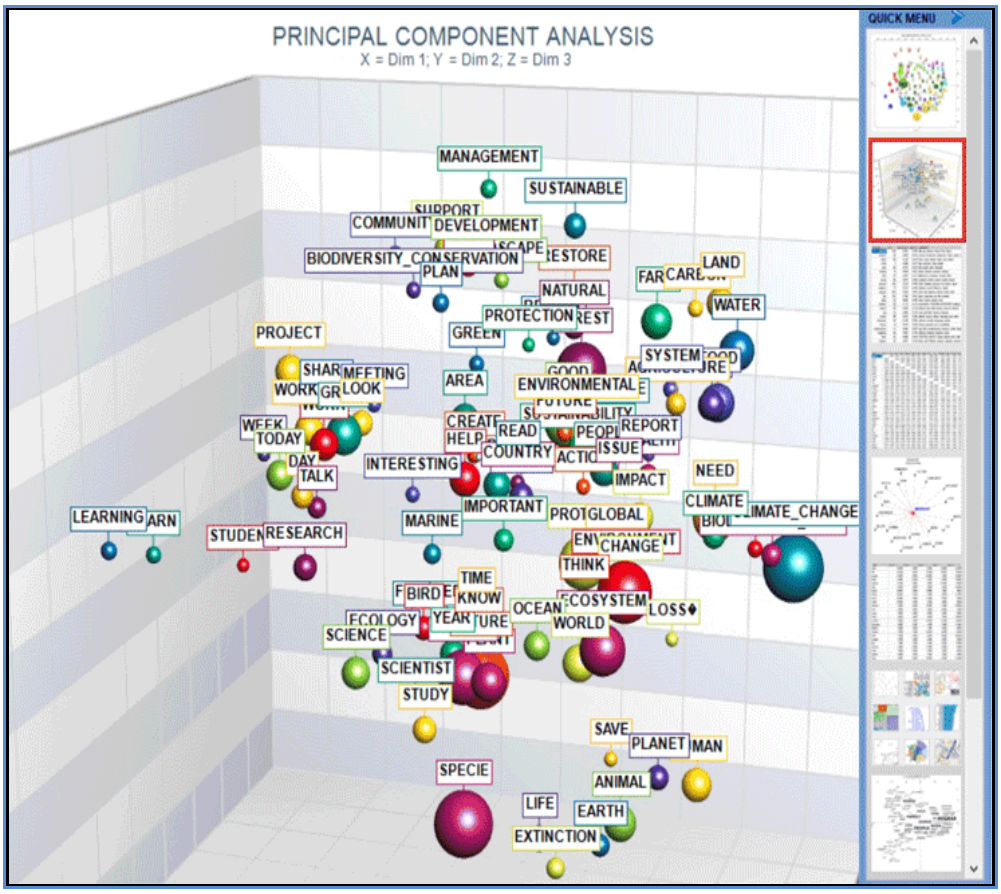




- Análisis de Co-Palabras

Esta herramienta de **T-LAB** nos permite trazar mapas de co-ocurrencias entre conjuntos de palabras clave.



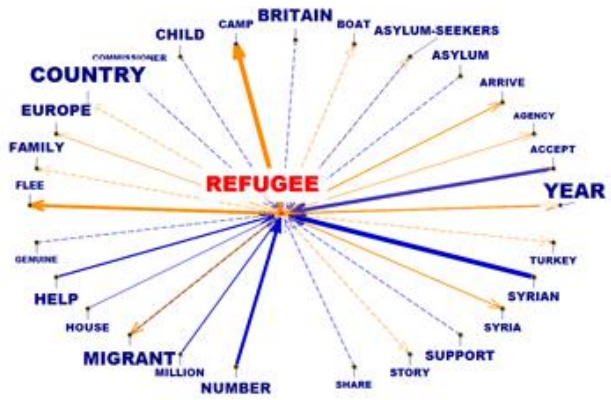


- **Análisis de Secuencias y Análisis de Redes**

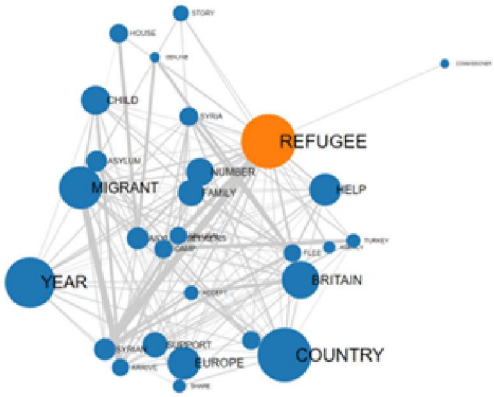
Esta herramienta de **T-LAB** tiene en cuenta las posiciones de las diferentes unidades lexicales dentro de las frases, permitiendonos así analizar y representar cualquier texto como si fuera una **red** de relaciones.

Una vez implementado este tipo de analisis, el usuario podrá explorar las relaciones que existen entre nodos de la red (esto es, las palabras clave) a diferentes niveles: a) dentro de las relaciones del tipo uno-a-uno; b) dentro de un "ego network"; c) dentro de las comunidades a las que pertenecen; d) dentro de la red formada por el texto analizado.

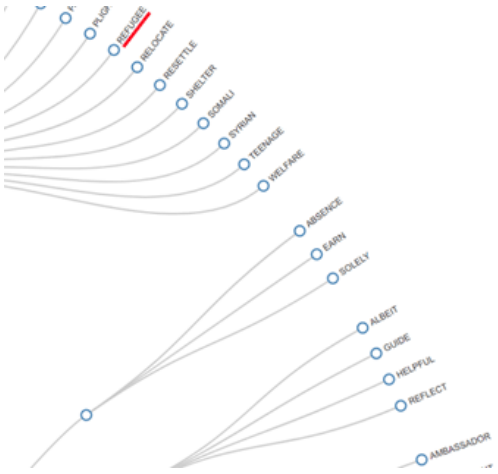
RELACIONES DEL TIPO UNO-AD-UNO



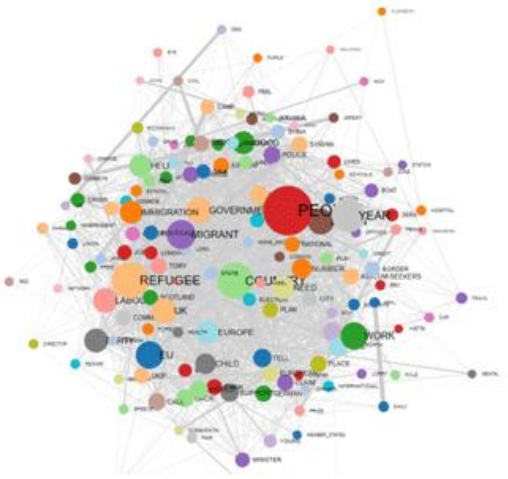
EGO-NETWORK



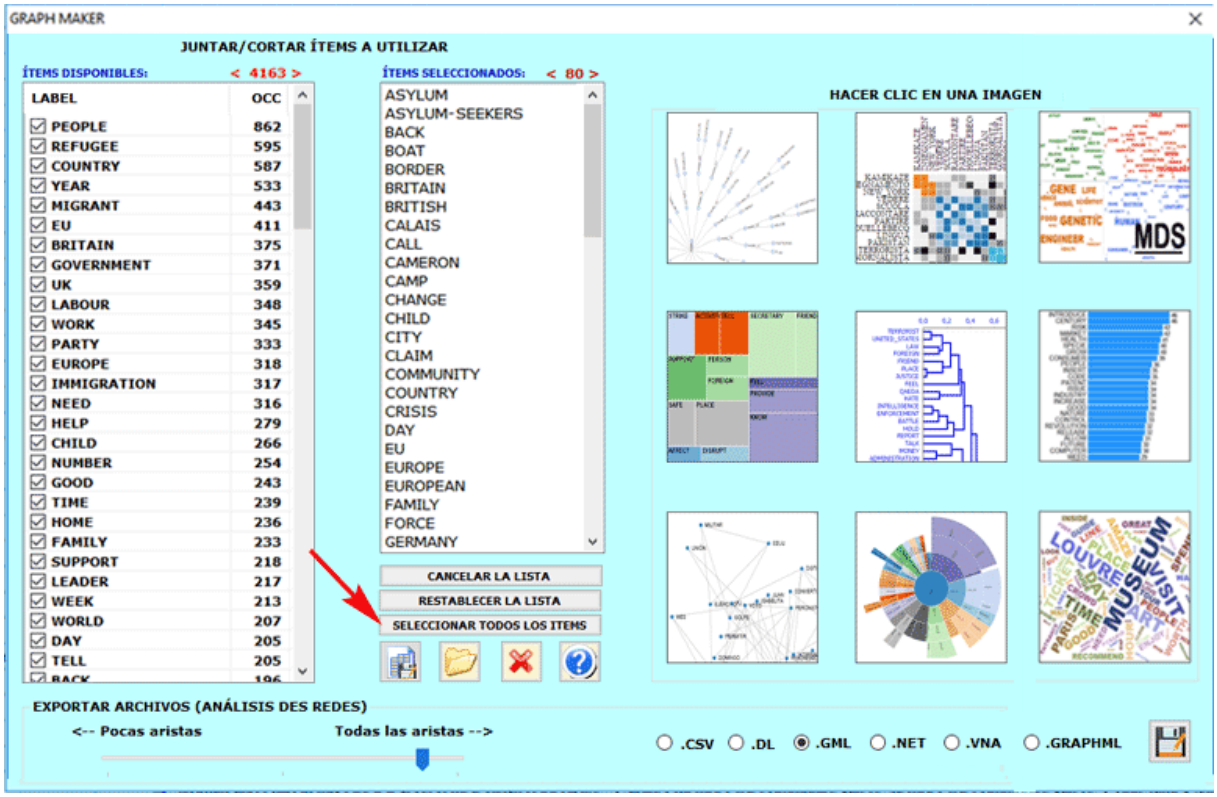
COMUNIDAD



RED ENTERA



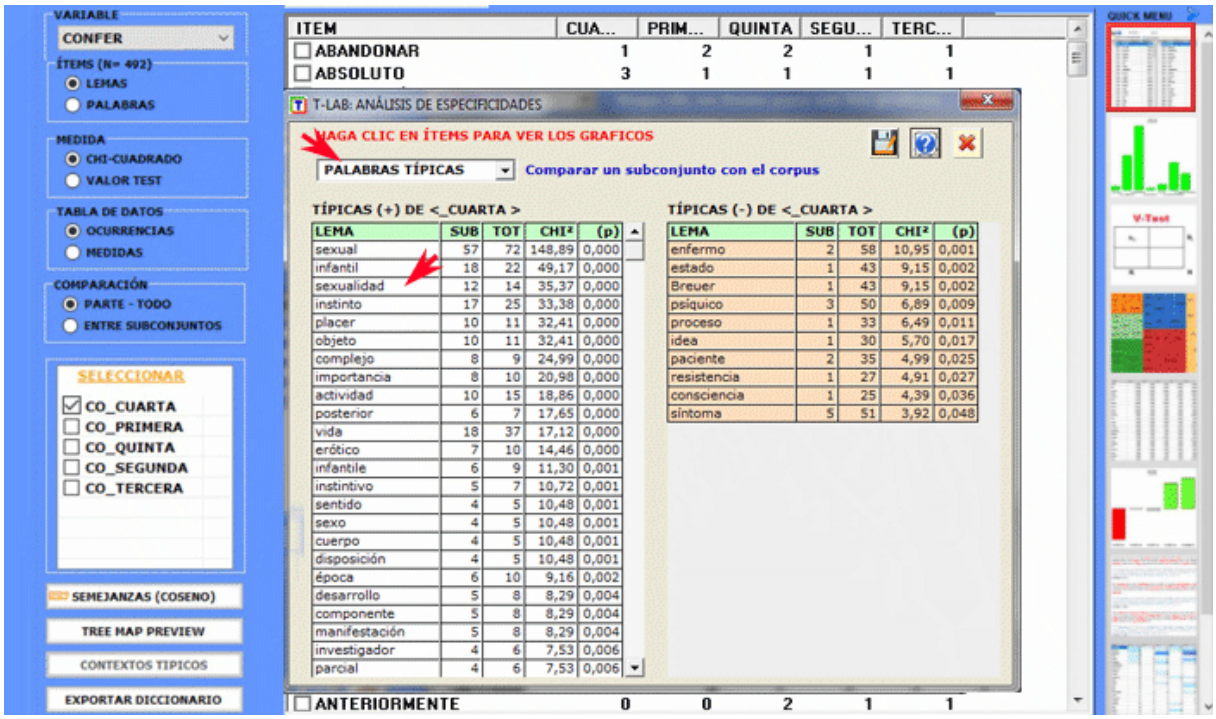
Además, cliqueando en la opción GRAPH MAKER el usuario podrá crear, en todo momento y a partir de los listados personalizados de palabras clave, diferentes tipos de graficos (véase abajo).

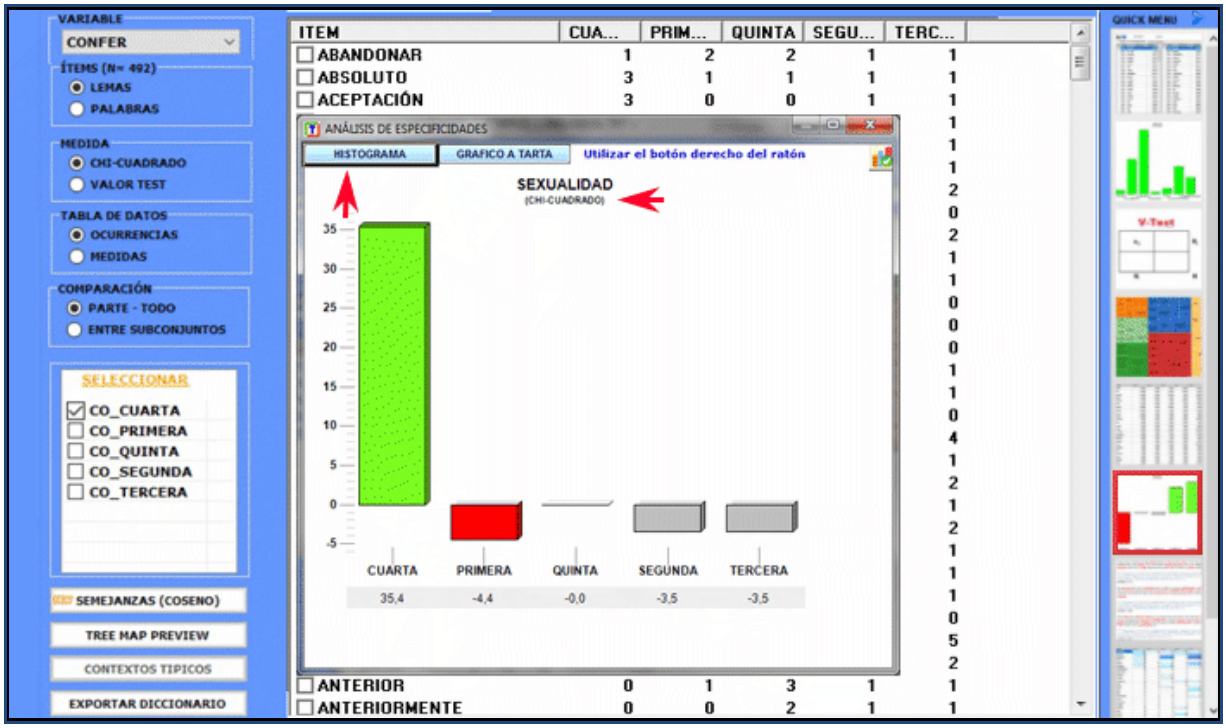


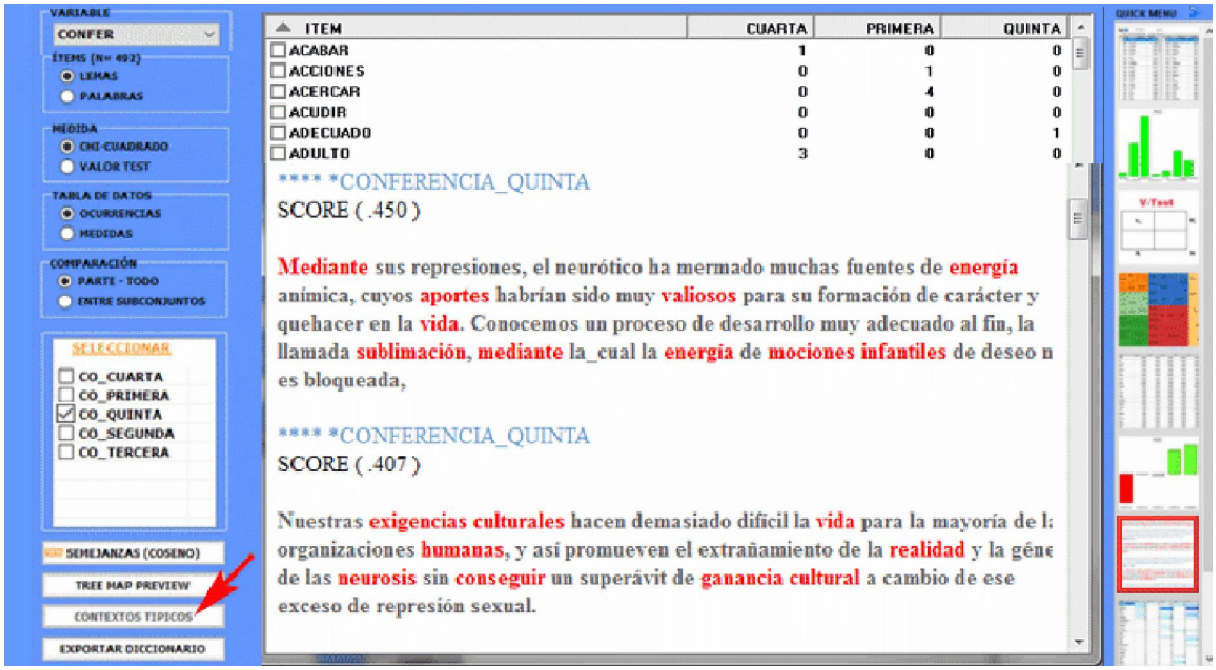
B : INSTRUMENTOS PARA ANÁLISIS COMPARATIVOS

Estos instrumentos permiten analizar varios tipos de relaciones entre las unidades de contexto.

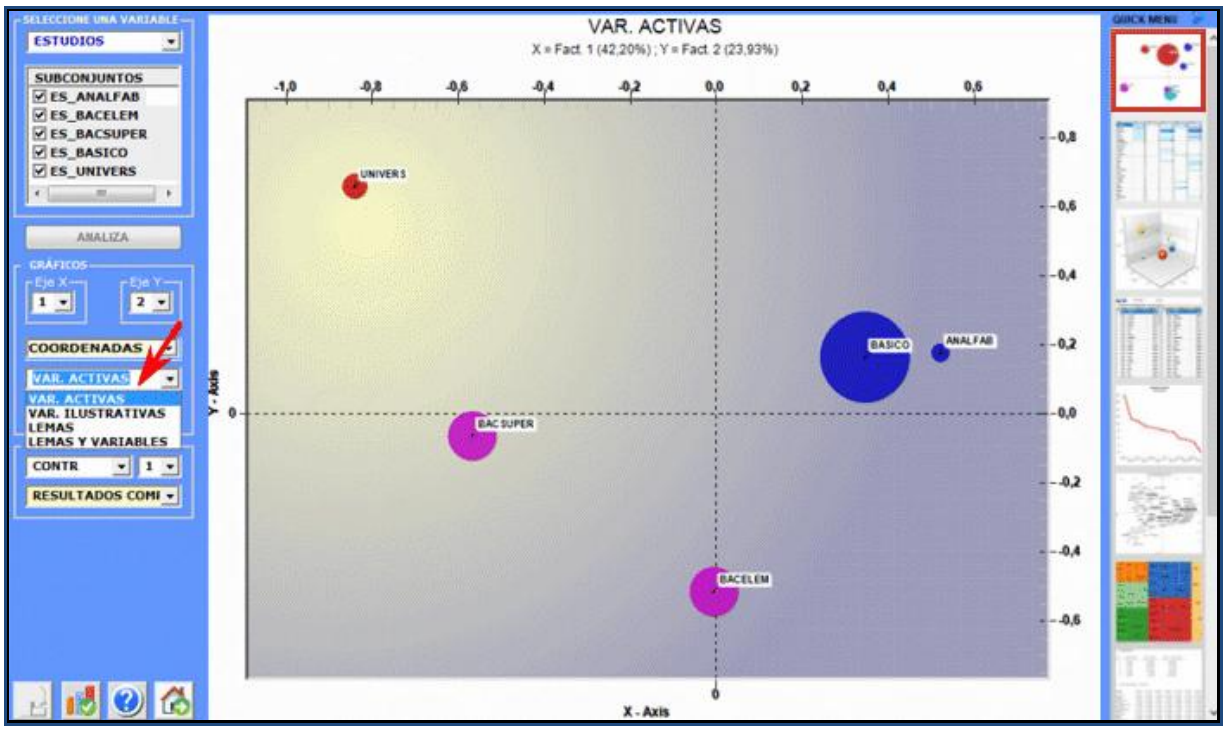
El **Análisis de las Especificidades** permite verificar cuáles palabras son "típicas" o "exclusivas" de cada subconjunto del corpus. Además, nos permite extraer los contextos típicos, es decir, los contextos elementales característicos, de cada uno de los subconjuntos analizados (p. ej.: las frases 'típicas' utilizadas por los líderes políticos).

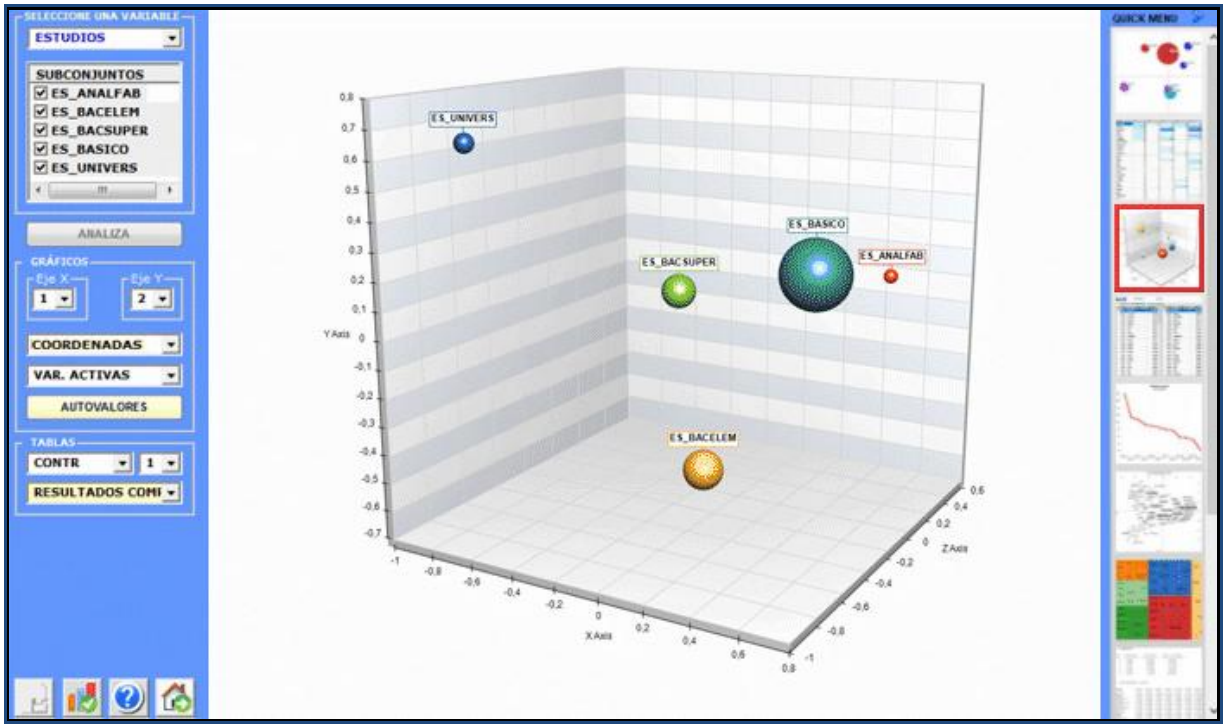

VARIABLE
CONFER
ÍTEMS (N= 492)
LEMAS
PALABRAS
MEDIDA
CHI-CUADRADO
VALOR TEST
TABLA DE DATOS
OCURRENCIAS
MEDIDAS
COMPARACIÓN
PARTE - TODO
ENTRE SUBCONJUNTOS
SELECCIONAR
CO_CUARTA
CO_PRIMERA
CO_QUINTA
CO_SEGUNDA
CO_TERCERA
SEMEJANZAS (COSENO)
TREE MAP PREVIEW
CONTEXTOS TIPICOS
EXPORTAR DICCIONARIO
ITEM
CUA...
PRIM...
QUINTA
SEGU...
TERC...
ABANDONAR 1 2 2 1 1
ABSOLUTO 3 1 1 1 1
T-LAB: ANÁLISIS DE ESPECIFICIDADES
HAGA CLIC EN ÍTEMS PARA VER LOS GRAFICOS
PALABRAS TÍPICAS
Comparar un subconjunto con el corpus
TÍPICAS (+) DE <_CUARTA >
LEMA SUB TOT CHI² (p)
sexual 57 72 148,89 0,000
infantil 18 22 49,17 0,000
sexualidad 12 14 35,37 0,000
instinto 17 25 33,38 0,000
placer 10 11 32,41 0,000
objeto 10 11 32,41 0,000
complejo 8 9 24,99 0,000
importancia 8 10 20,98 0,000
actividad 10 15 18,86 0,000
posterior 6 7 17,65 0,000
vida 18 37 17,12 0,000
erótico 7 10 14,46 0,000
infantile 6 9 11,30 0,001
instintivo 5 7 10,72 0,001
sentido 4 5 10,48 0,001
sexo 4 5 10,48 0,001
cuerpo 4 5 10,48 0,001
disposición 4 5 10,48 0,001
época 6 10 9,16 0,002
desarrollo 5 8 8,29 0,004
componente 5 8 8,29 0,004
manifestación 5 8 8,29 0,004
investigador 4 6 7,53 0,006
parcial 4 6 7,53 0,006
TÍPICAS (-) DE <_CUARTA >
LEMA SUB TOT CHI² (p)
enfermo 2 58 10,95 0,001
estado 1 43 9,15 0,002
Breuer 1 43 9,15 0,002
psíquico 3 50 6,89 0,009
proceso 1 33 6,49 0,011
idea 1 30 5,70 0,017
paciente 2 35 4,99 0,025
resistencia 1 27 4,91 0,027
consciencia 1 25 4,39 0,036
síntoma 5 51 3,92 0,048
ANTERIORMENTE 0 0 2 1 1
QUICK MENU


VARIABLE
CONFER
ÍTEMS (N= 492)
LEMAS
PALABRAS
MEDIDA
CHI-CUADRADO
VALOR TEST
TABLA DE DATOS
OCURRENCIAS
MEDIDAS
COMPARACIÓN
PARTE - TODO
ENTRE SUBCONJUNTOS
SELECCIONAR
CO_CUARTA
CO_PRIMERA
CO_QUINTA
CO_SEGUNDA
CO_TERCERA
SEMEJANZAS (COSENO)
TREE MAP PREVIEW
CONTEXTOS TIPICOS
EXPORTAR DICCIONARIO
ITEM
CUA...
PRIM...
QUINTA
SEGU...
TERC...
ABANDONAR 1 2 2 1 1
ABSOLUTO 3 1 1 1 1
ACEPTACIÓN 3 0 0 1 1
ANÁLISIS DE ESPECIFICIDADES
HISTOGRAMA
GRAFICO A TARTA
Utilizar el botón derecho del ratón
SEXUALIDAD
(CHI-CUADRADO)
35
30
25
20
15
10
5
0
-5
CUARTA
PRIMERA
QUINTA
SEGUNDA
TERCERA
35,4
-4,4
-0,0
-3,5
-3,5
ANTERIOR 0 1 3 1 1
ANTERIORMENTE 0 0 2 1 1
QUICK MENU

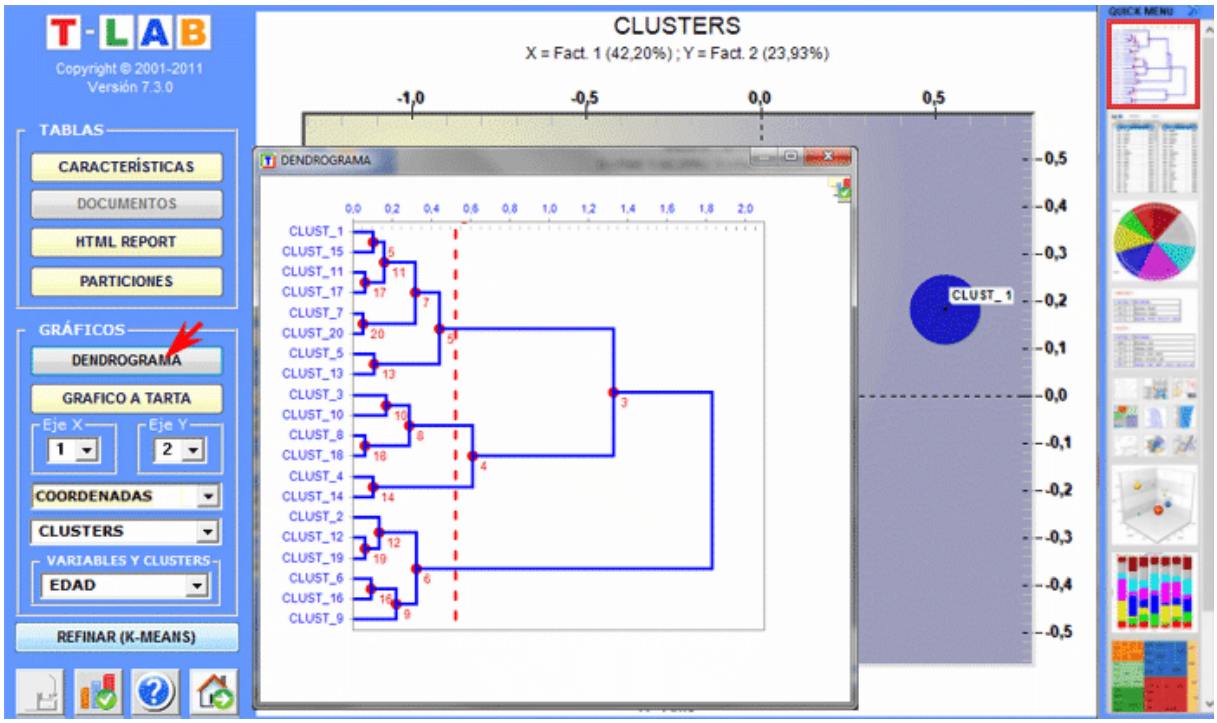


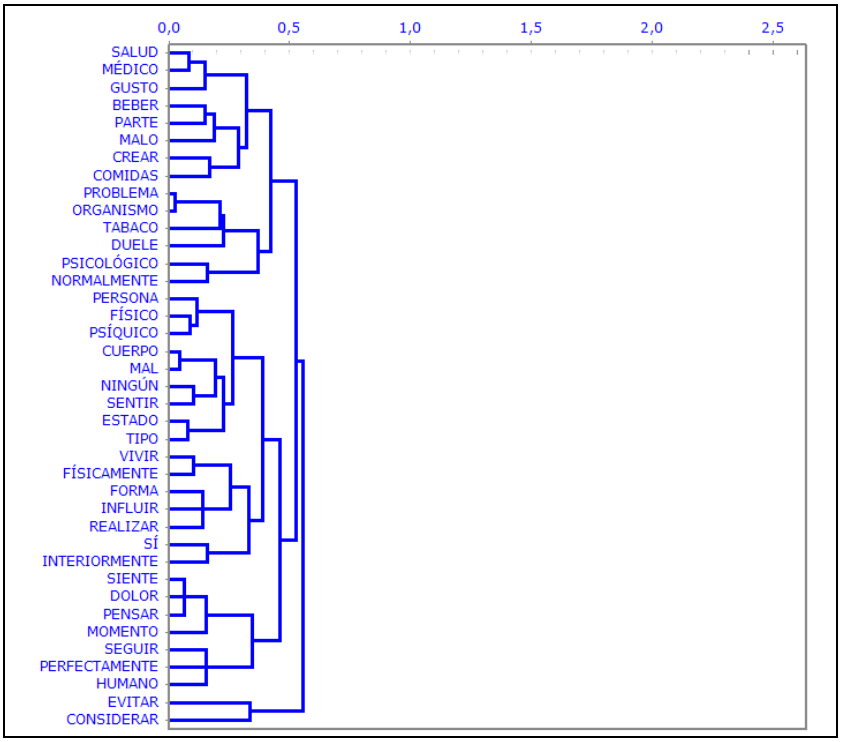
El **Análisis de Correspondencias** permite explorar varios tipos de relaciones (semejanzas y diferencias) entre grupos de unidades de contexto.





El **Cluster Analysis** permite encontrar grupos de unidades de texto que presentan dos características complementarias: máxima homogeneidad interna y máxima heterogeneidad entre cada clúster y todos los demás. Se puede implementar recurriendo a múltiples técnicas y requiere, previamente, un análisis de las correspondencias o un SVD.





C : INSTRUMENTOS PARA ANÁLISIS TEMÁTICOS

Estos instrumentos permiten individuar, examinar y trazar el mapa de los "temas" que emergen de los textos analizados.

Puesto que "**Tema**" es una palabra polisémica, cuando se usa software para análisis temático es útil hacer referencia a algunas definiciones operativas.

En los instrumentos **T-LAB**, "tema" es una etiqueta usada para indicar cuatro diferentes entidades:

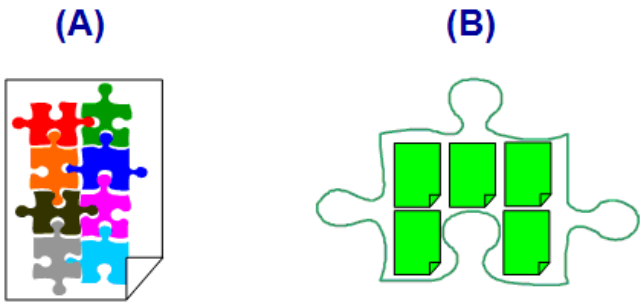
1- un **clúster temático de unidades de contexto** caracterizados por los mismos modelos de palabras clave (ver los instrumentos **Análisis Temático de Contextos Elementales** y **Clasificación Temática de Documentos**);

2- un **grupo temático de palabras-clave** clasificadas en términos de pertenencia a una misma categoría (véase la herramienta **Clasificación Basada en Diccionarios**);

3- un **componente de un modelo probabilista** que representa cada unidad de contexto (contextos elementales o documentos) generado de una mezcla de "temas" (ver los instrumentos **Modelización de los Temas Emergentes** y **Textos y Discursos como Sistemas Dinámicos**);

4- una **específica palabra clave** ("temática") usada para extraer un conjunto de contextos elementales. Esta palabra está asociada con un específico conjunto de palabras preseleccionadas por el usuario (ver el instrumento **Contextos Clave de Palabras Temáticas**).

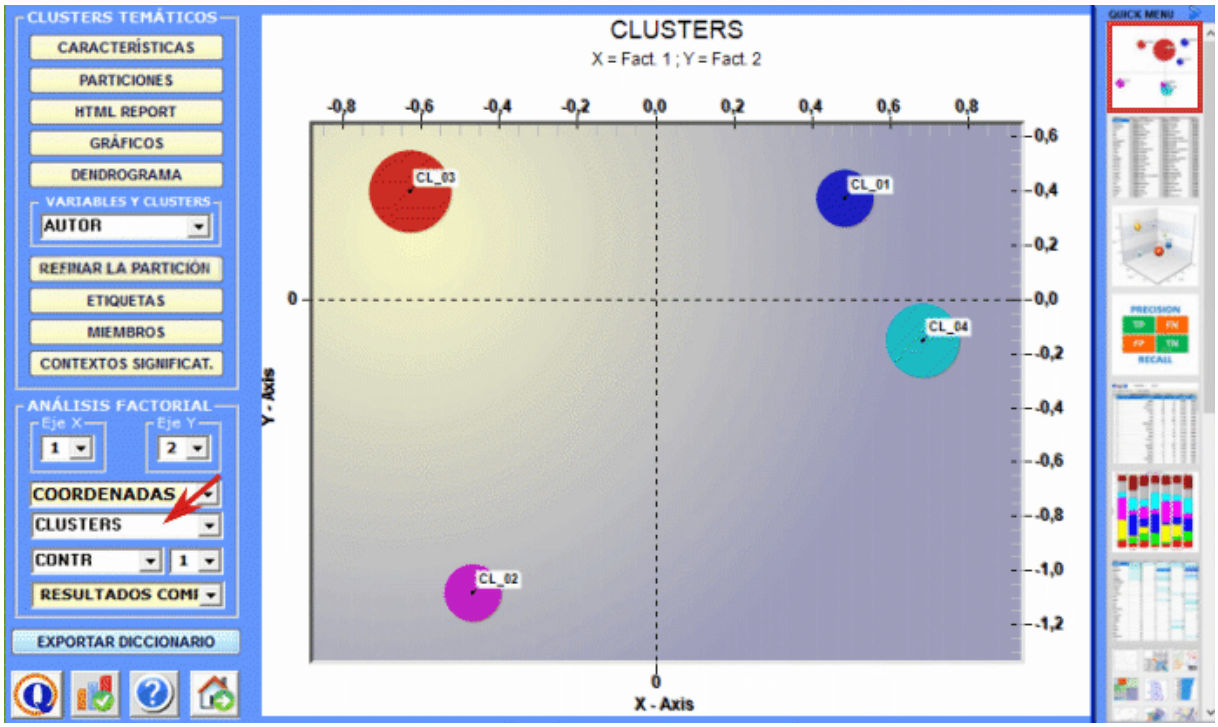
Por ejemplo, según el tipo de herramienta que estemos utilizando, un documento concreto puede ser analizado bien en términos de co-presencia de varios 'temas' en un único documento (véase 'A' abajo) o bien como parte de un conjunto de documentos que conciernen el mismo 'tema' (véase 'B' abajo). De hecho, en el caso 'A', cada tema puede corresponder a una palabra o frase mientras que, en el caso 'B', un tema puede representar una etiqueta asignada a un conjunto de documentos que presentan los mismos patrones de palabras-clave.

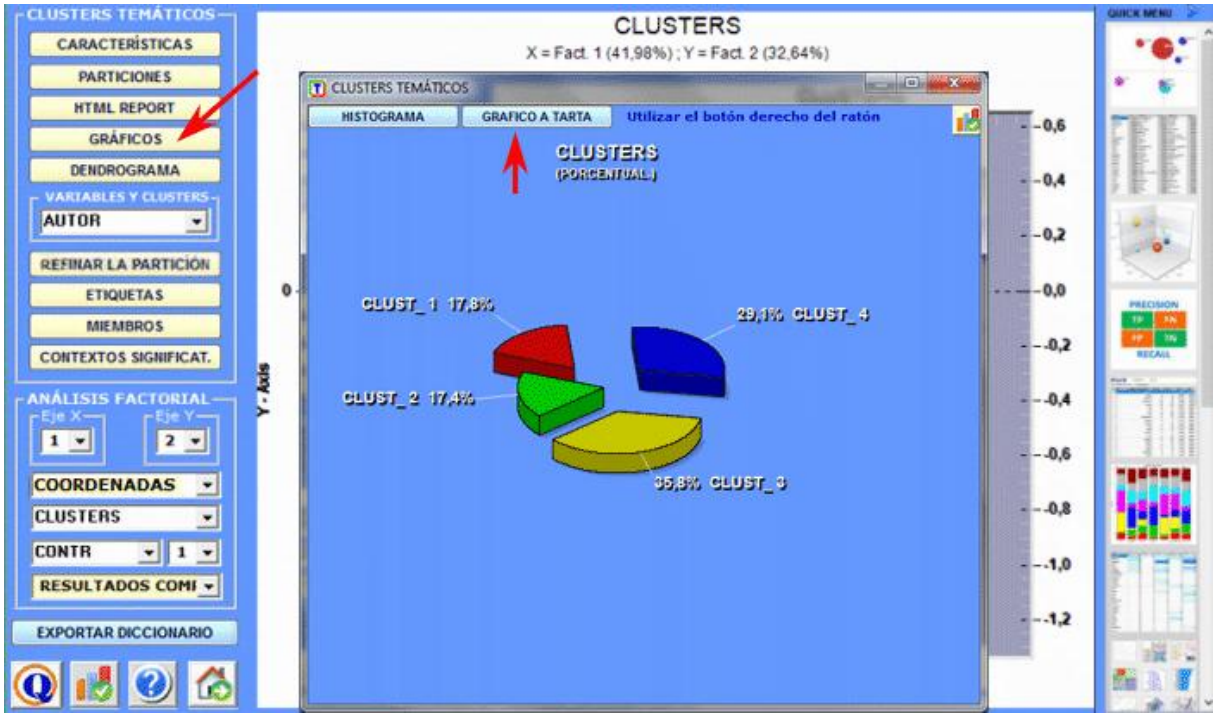


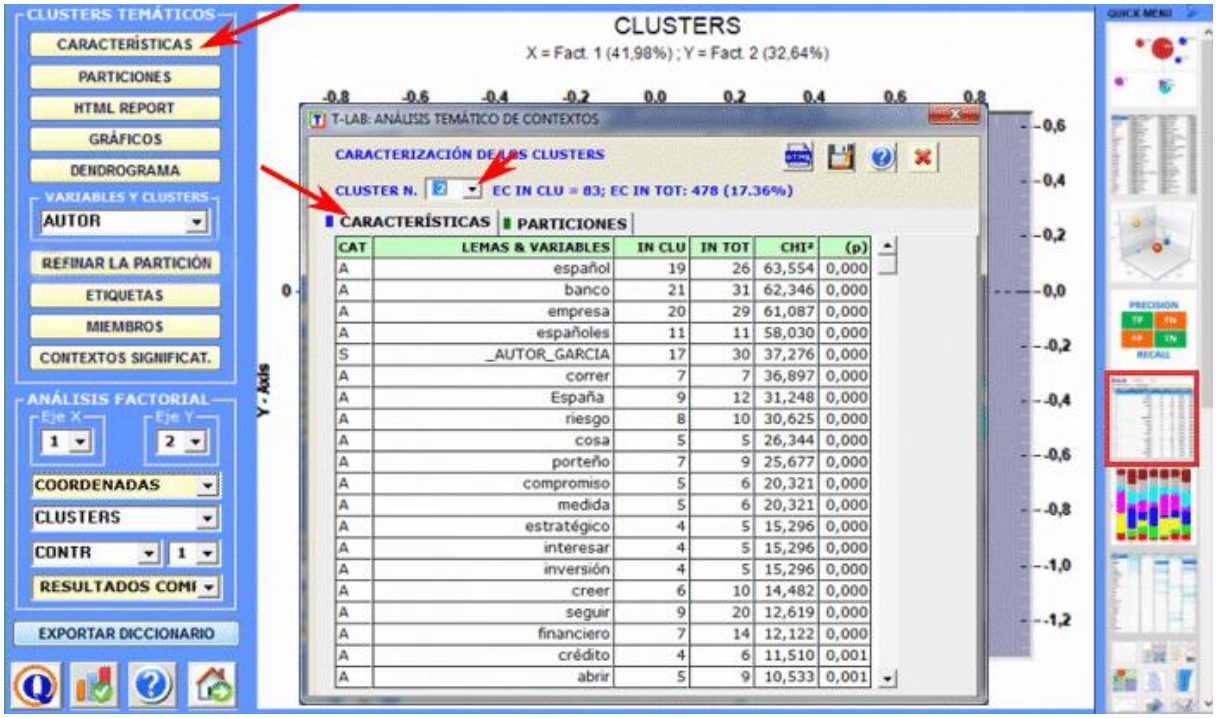
Más en detalle, **T-LAB** 'extrae' los temas utilizando las siguientes metodologías:

1 - tanto el **Análisis Temático de Contextos Elementales** como la **Clasificación Temática de Documentos** funcionan en la siguiente manera:

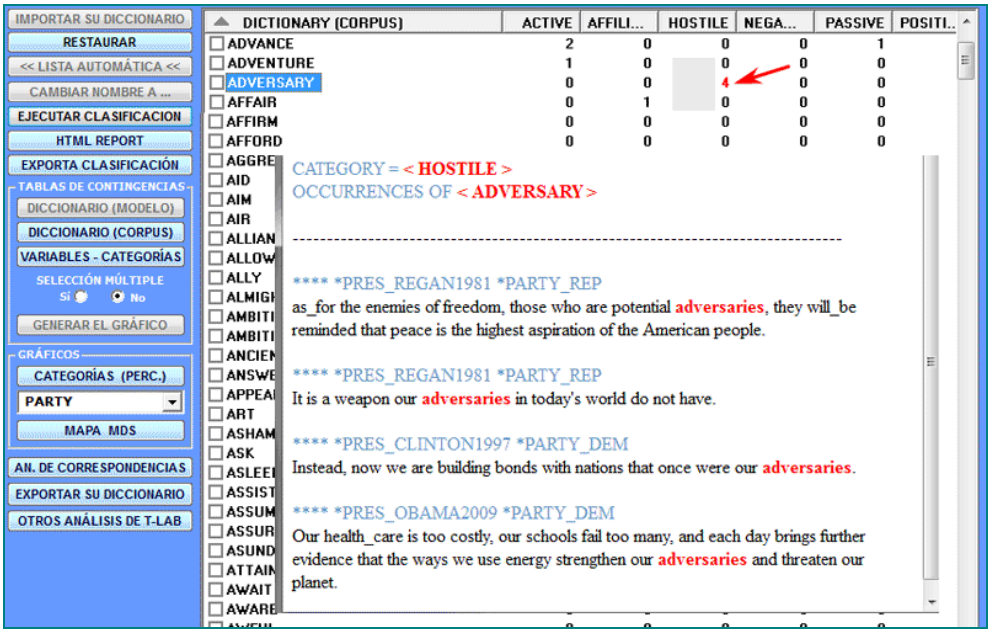
a- realizan un **análisis de co-ocurrencias** para obtener los clusters temáticos de unidades de contexto;
b- realizan un **análisis comparativo** para confrontar los perfiles de los distintos clusters;
c- generan varios tipos de gráficos y tablas (ver a continuación);
d- permiten archivar las **nuevas variables** obtenidas (clusters temáticos) y utilizarlas en análisis posteriores.

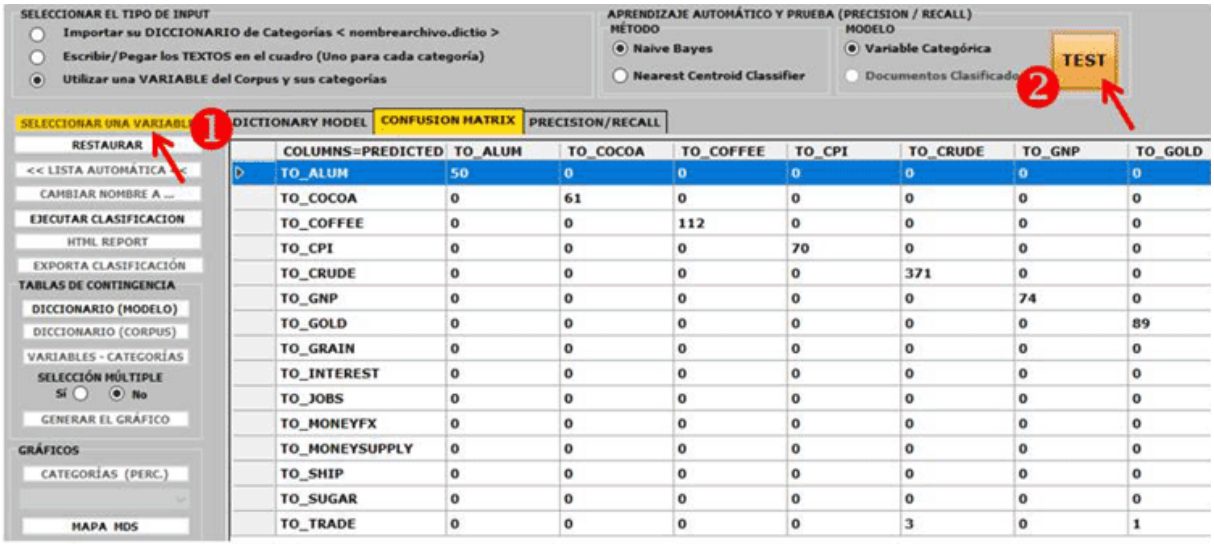



CLUSTERS TEMÁTICOS
CARACTERÍSTICAS
PARTICIONES
HTML REPORT
GRÁFICOS
DENDROGRAMA
VARIABLES Y CLUSTERS
AUTOR
REFINAR LA PARTICIÓN
ETIQUETAS
MIEMBROS
CONTEXTOS SIGNIFICAT.
ANÁLISIS FACTORIAL
Eje X
Eje Y
1
2
COORDENADAS
CLUSTERS
CONTR
1
RESULTADOS COMI
EXPORTAR DICCIONARIO
CLUSTERS
X = Fact. 1 (41,98%) ; Y = Fact. 2 (32,64%)
CLUSTERS TEMÁTICOS
HISTOGRAMA
GRAFICO A TARTA
Utilizar el botón derecho del ratón
CLUSTERS
(PORCENTUAL)
CLUST_1 17,8%
29,1% CLUST_4
CLUST_2 17,4%
35,8% CLUST_3
Y - Axis
QUICK MENU

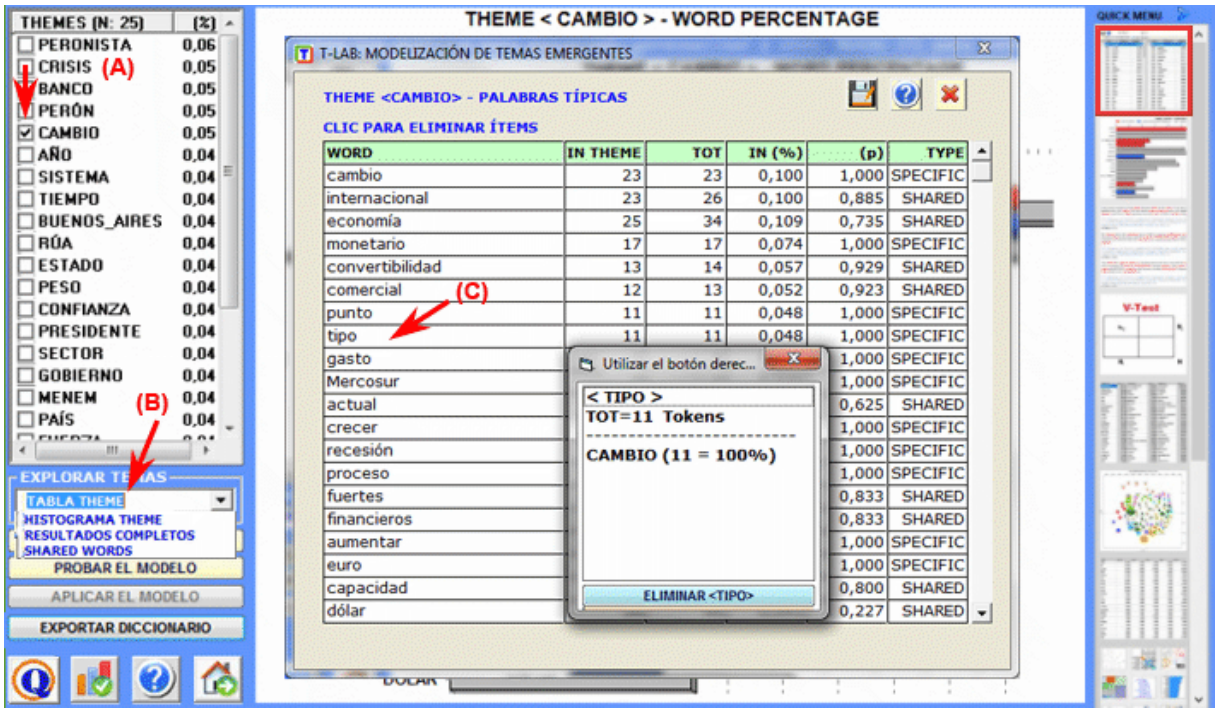

CLUSTERS TEMÁTICOS
CARACTERÍSTICAS
PARTICIONES
HTML REPORT
GRÁFICOS
DENDROGRAMA
VARIABLES Y CLUSTERS
AUTOR
REFINAR LA PARTICIÓN
ETIQUETAS
MIEMBROS
CONTEXTOS SIGNIFICAT.
ANÁLISIS FACTORIAL
EXPORTAR DICCIONARIO
CLUSTERS
X = Fact. 1 (41,98%) ; Y = Fact. 2 (32,64%)
T-LAB: ANÁLISIS TEMÁTICO DE CONTEXTOS
CARACTERIZACIÓN DE LOS CLUSTERS
CLUSTER N. 2
EC IN CLU = 83; EC IN TOT: 478 (17.36%)
CARACTERÍSTICAS
PARTICIONES
CAT | LEMAS & VARIABLES | IN CLU | IN TOT | CHI² | (p)
A | español | 19 | 26 | 63,554 | 0,000
A | banco | 21 | 31 | 62,346 | 0,000
A | empresa | 20 | 29 | 61,087 | 0,000
A | españoles | 11 | 11 | 58,030 | 0,000
S | _AUTOR_GARCIA | 17 | 30 | 37,276 | 0,000
A | correr | 7 | 7 | 36,897 | 0,000
A | España | 9 | 12 | 31,248 | 0,000
A | riesgo | 8 | 10 | 30,625 | 0,000
A | cosa | 5 | 5 | 26,344 | 0,000
A | porteño | 7 | 9 | 25,677 | 0,000
A | compromiso | 5 | 6 | 20,321 | 0,000
A | medida | 5 | 6 | 20,321 | 0,000
A | estratégico | 4 | 5 | 15,296 | 0,000
A | interesar | 4 | 5 | 15,296 | 0,000
A | inversión | 4 | 5 | 15,296 | 0,000
A | creer | 6 | 10 | 14,482 | 0,000
A | seguir | 9 | 20 | 12,619 | 0,000
A | financiero | 7 | 14 | 12,122 | 0,000
A | crédito | 4 | 6 | 11,510 | 0,001
A | abrir | 5 | 9 | 10,533 | 0,001
Y - Axis
QUICK MENU

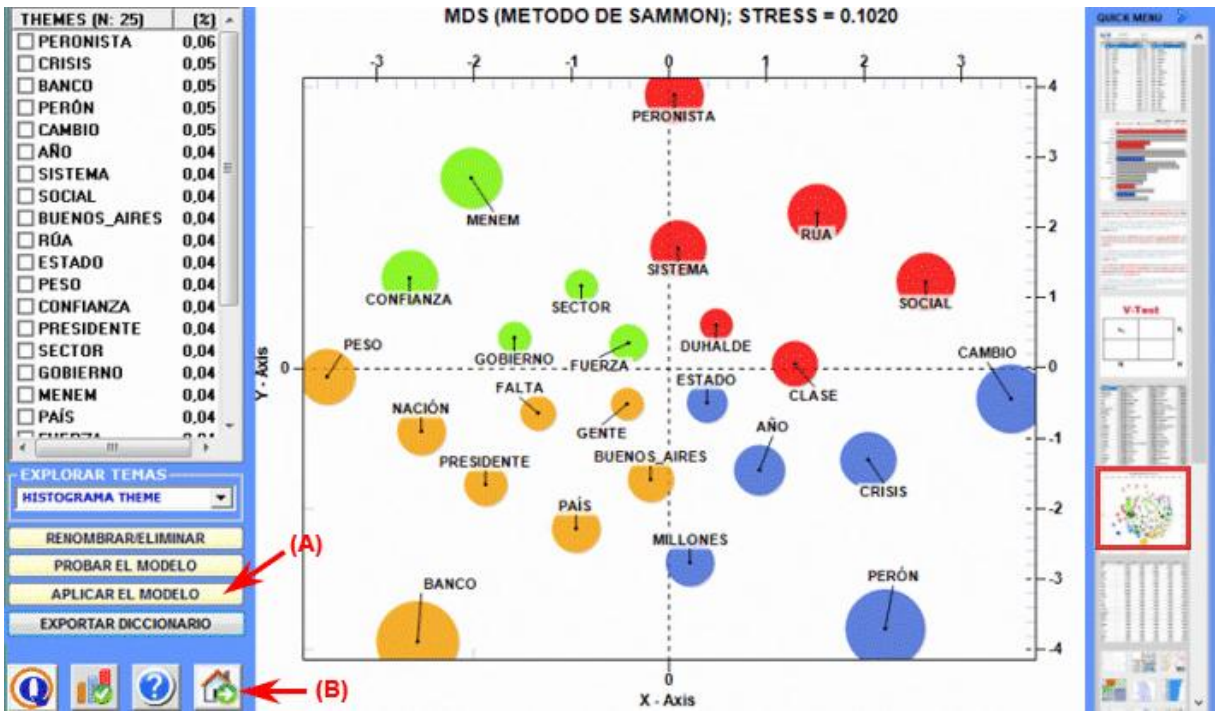
2 - a través de la herramienta **Clasificación Basada en Diccionarios**, podremos fácilmente construir/testar/aplicar modelos (p. ej.: Diccionarios de categorías) tanto para el análisis de contenido clásico como para el sentiment analysis. De hecho, esta herramienta nos permite implementar una clasificación automática de tipo top-down de las unidades lexicales (es decir, palabras y lemas) y también de las unidades de contexto (es decir, frases, párrafos y pequeños documentos).



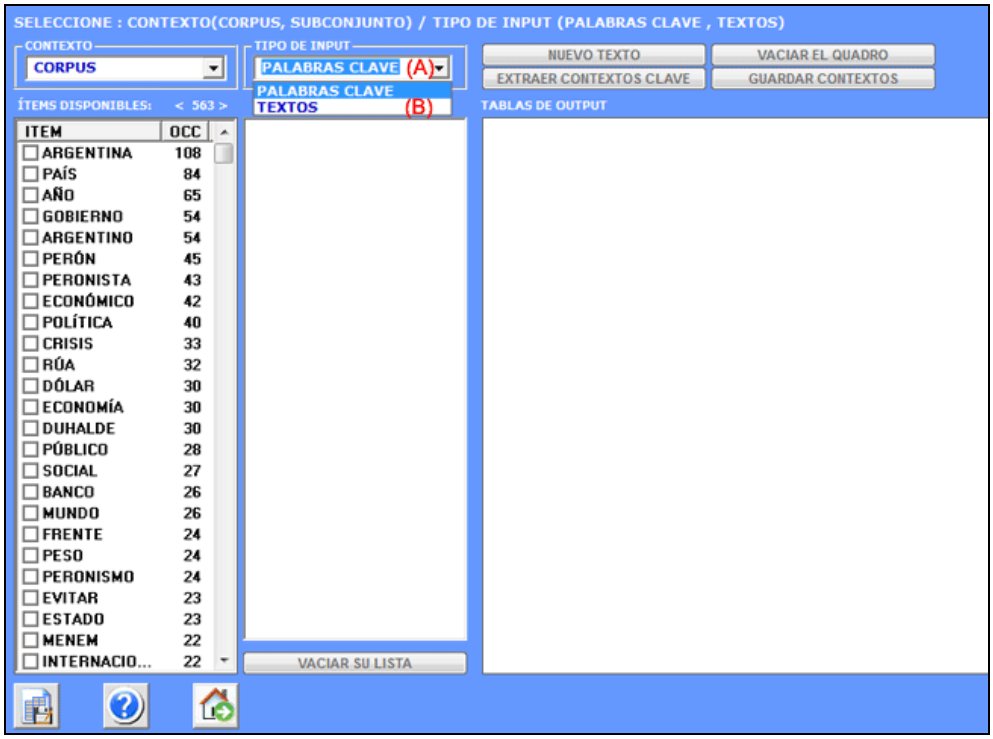


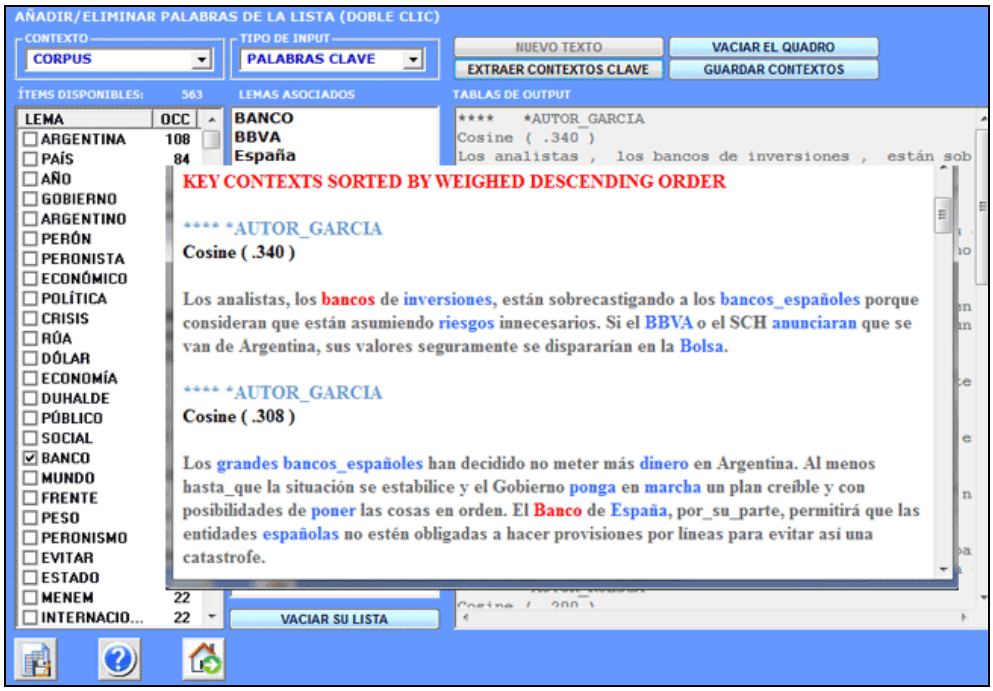
3 - mediante la herramienta **Modelización de los Temas Emergentes** (véase abajo), los componentes de la 'mixtura' temática pueden ser descritos a través de su vocabulario característico y pueden ser utilizados para construir tablas para el análisis cualitativo y/o para la clasificación automática de las unidades de contexto (es decir, contextos elementales o documentos).





4 - La herramienta **Contextos Clave de Palabras Temáticas** se puede utilizar para alcanzar dos objetivos: (a) extraer listados de unidades de contexto (es decir, contextos elementales) que permitan profundizar el valor temático de palabras-clave especificas; (b) extraer grupos de unidades de contexto que resulten ser similares a algún texto de 'ejemplo' escogido por el usuario.

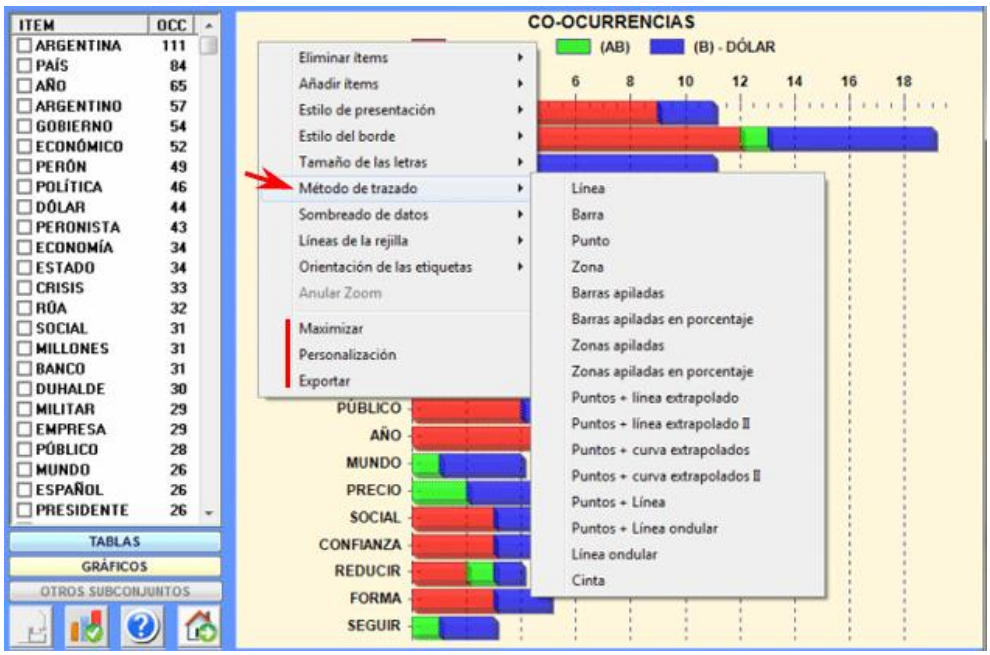


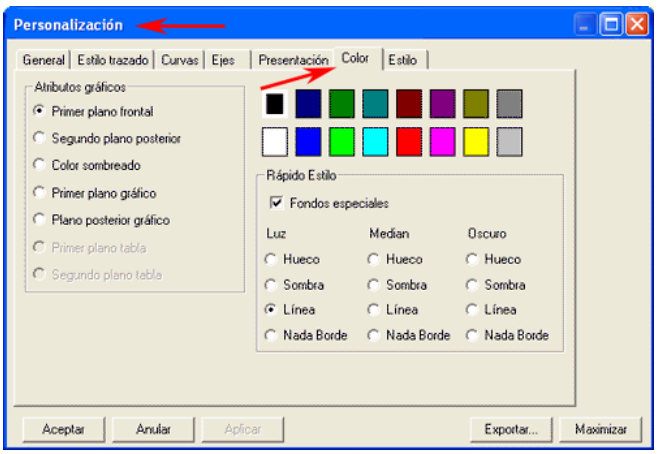


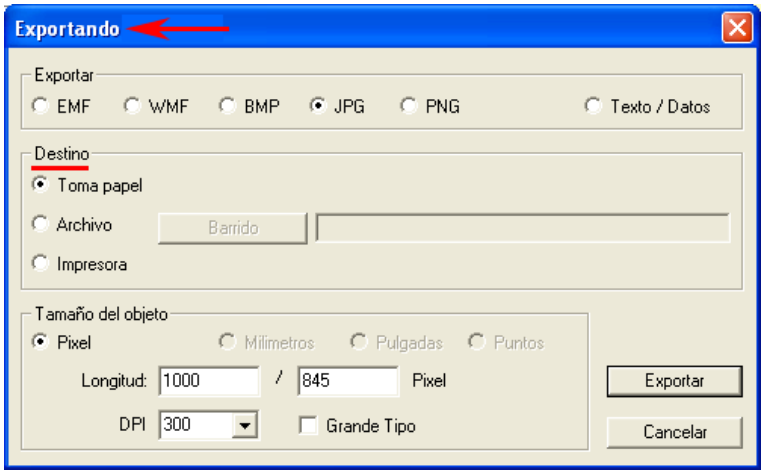
6 - LA INTERPRETACIÓN DE LOS OUTPUT consiste en la consulta de las tablas y de los gráficos producidos por **T-LAB**, en la eventual personalización de su formato y en el hacer inferencias sobre el significado de las relaciones en los mismos representados.

En el caso de las **tablas**, según los casos, **T-LAB** permite exportarlas en filas con las siguientes extensiones: **.DAT**, **.TXT**, **.CSV**, **.XLXS**, **.HTML**. Esto significa que, utilizando cualquier editor de textos y/o de cualquier aplicativo de la suite Microsoft Office, el usuario puede, fácilmente, importarlos y reelaborarlos.

Todos los **gráficos** y **tablas** pueden ser maximizados (hacer clic con el botón izquierdo y arrastre), personalizados y exportados en diferentes formatos (hacer clic con el botón derecho del ratón para ver los pop up menús).





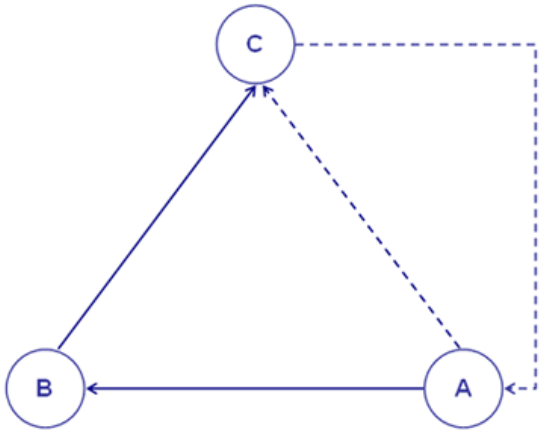


En un paper citado en **Bibliografía** y disponible en el sitio https://www.tlab.it (Lancia F.: 2007) se mencionan algunos criterios generales para la interpretación de los outputs **T-LAB**. En el mismo se propone la hipótesis que los output de las elaboraciones estadísticas (tablas y gráficos) son un tipo particular de textos, es decir son objetos multi-semióticos caracterizados por el hecho que las relaciones entre los signos y los símbolos están ordenadas por medidas que redireccionan a **códigos** específicos.

En otros términos, tanto en el caso de textos escritos en lenguaje natural como en los escritos en el lenguaje de la estadística, la posibilidad de hacer inferencias sobre las relaciones que organizan las **formas del contenido** está garantizada por el hecho de que las relaciones entre las **formas de la expresión** no son casuales (random); de hecho, en el primer caso (lenguaje natural) las unidades significantes se subsiguen ordenadas según un modo lineal (una tras otra en la cadena del discurso), mientras que en el segundo caso (tablas y gráficos) los principios de ordenación están constituidos por las medidas que determinan la organización de los **espacios semánticos** multidimensionales.

Si bien los espacios semánticos representados en los mapas **T-LAB** son muy variados y cada uno de esos requiere procedimientos de interpretación específicos, se puede suponer que - en general - la lógica del proceso inferencial es la siguiente:

A - sacar cualquier relación significativa entre las unidades "presentes" en el plano de la expresión (por ej. entre "datos" de tablas y/o entre "etiquetas" de gráficos);
B - explorar y comparar los componentes semánticos de las mismas unidades y los contextos a los que están mentalmente y culturalmente asociadas (plano del contenido);
C - construir algunas hipótesis o algunas "categorías" de análisis que, en el contexto definido por el corpus den cuenta de las relaciones entre formas de la expresión y formas del contenido.



Actualmente, las opciones de **T-LAB** presentan las siguientes **restricciones**:

- tamaño del corpus: máximo 90Mb (casi 55.000 páginas en formato ASCII);
- documentos primarios: máximo 30.000 (o 99.999 textos cortos Max. 2.000 caracteres cada uno; Eje. respuestas a preguntas abiertas, mensajes de Twitter etc.);
- variables categóricas: máximo 50, cada una con máximo 150 modalidades;
- modelización de los temas emergentes: máximo 5.000 unidades lexicales (*) por 5.000.000 ocurrencias;
- análisis temático de contextos elementales: máximo 300.000 filas (unidades de contexto) por 5.000 columnas (unidades lexicales);
- clasificación temática de documentos: máximo 99.999 filas (unidades de contexto, documentos) por 5.000 columnas (unidades lexicales);
- análisis de especificidades (unidades lexicales x categorías de una variable): máximo 10.000 filas por 150 columnas;
- análisis de correspondencias (unidades lexicales x categorías de una variable): máximo 10.000 filas por 150 columnas;
- análisis de correspondencias (unidades de contexto x unidades lexicales): máximo 10.000 filas por 5.000 columnas;
- análisis de correspondencias múltiples (contextos elementales x categorías de dos o más variables): máximo 150.000 filas por 250 columnas;
- descomposición de valores singulares (SVD): máximo 300.000 filas por 5.000 columnas;
- cluster analysis que utiliza los resultados de un precedente análisis de correspondencias (o SVD): máximo 10.000 filas (unidades lexicales o contextos elementales);
- asociaciones de palabras, comparaciones entre parejas de palabras-clave y análisis de co-palabras: máximo 5.000 unidades lexicales;
- análisis de secuencias: máximo 5.000 unidades lexicales (o categorías) por 3.000.000 ocurrencias.

(*) En **T-LAB**, las 'unidades léxicales' son palabras, palabras múltiples, lemas y categorías semánticas. Así que, cuando se aplica la lematización automática, 5.000 unidades léxicales corresponden a cerca de 12.000 palabras.

CONFIGURACIONES DE ANÁLISIS

Configuración Automática y Configuración Personalizada

La elección entre la configuración **automática** (A) o **personalizada** (B) está relacionada con el listado de Palabras Clave que se utilizan en todos los análisis implementadas por **T-LAB**. Dicha elección será reversible siempre y cuando el usuario no haya aportado modificaciones al diccionario del corpus.

A) CONFIGURACIÓN AUTOMÁTICA

Escoger a la **configuración automática** implica que el listado de palabras clave incluya hasta un máximo de **5000** unidades lexicales, todas ellas escogidas automáticamente por T-LAB y pertenecientes a las categorías gramaticales que son más densas de significado: nombres, verbos, adjetivos y adverbios.

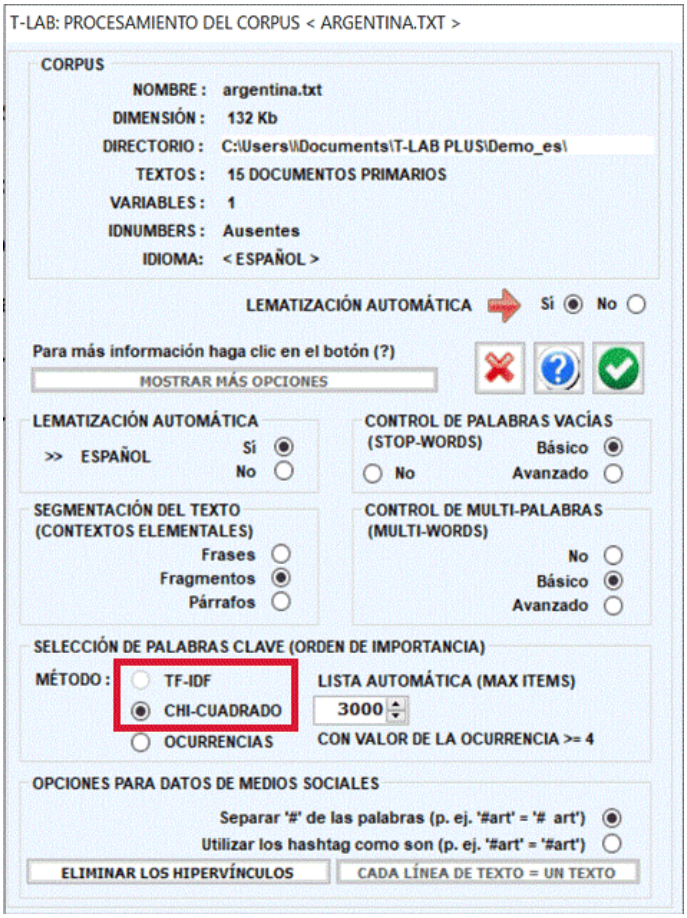
El criterio de selección varía en función del tipo de corpus analizado.

Si el corpus se compone de un único texto **T-LAB** selecciona simplemente las unidades lexicales con los valores más altos de **ocurrencia**.

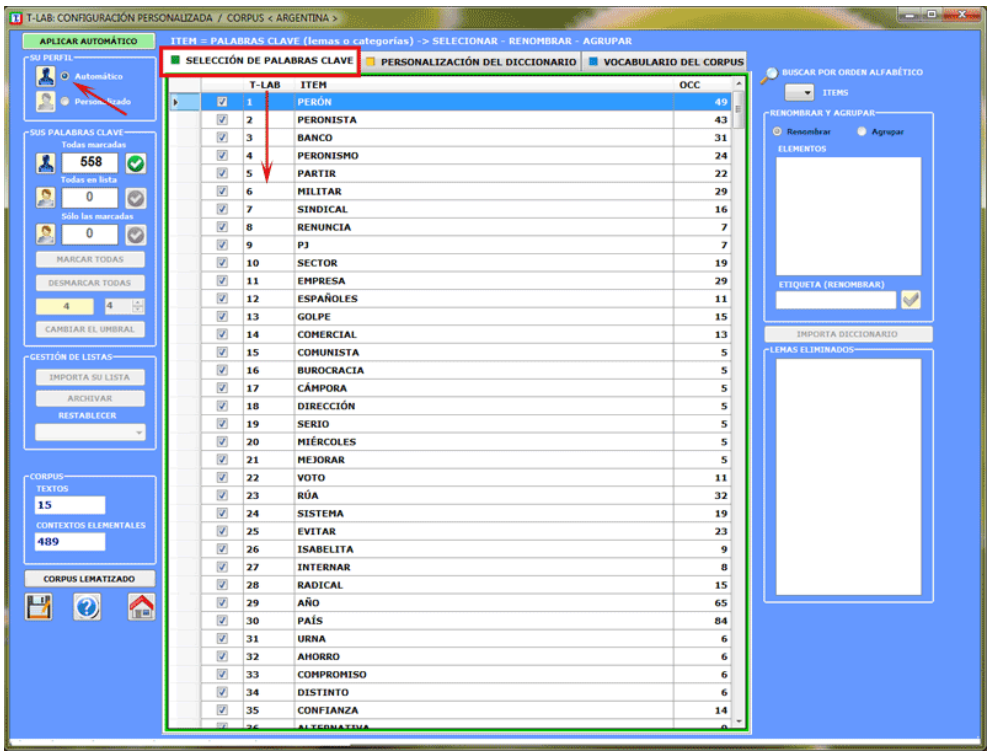
Si el corpus se compone de dos o más textos **T-LAB** utiliza el algoritmo ilustrado en la nota siguiente:

- selecciona las palabras con valores de frecuencia superiores al umbral mínimo;
- aplica el TF-IDF o el test del chi-cuadrado a todos los cruces de cada palabra seleccionada para todos los textos analizados (NOTA: En el caso del chi cuadrado, los textos deben ser máximo 500);
- selecciona las palabras con los valores más altos en el TF-IDF o en el test de chi cuadrado, o sea esas palabras que, en el texto, hacen la diferencia.

- En el caso de que el corpus esté compuesto por dos o más textos, el usuario puede escoger el criterio de selección a utilizar en la fase de importación (CHI cuadrado o TF-IDF);



- Al activarse la opción de impostaciones automáticas, en la tabla que contiene el listado de **Palabras Clave** aparecerá también la columna 'T-LAB'. En ésta se recoge el grado de relevancia de cada ítem en base al criterio escogido por el usuario (véase abajo).



B) **CONFIGURACIÓN PERSONALIZADA**

En el caso de escoger la **configuración personalizada**, el usuario podrá seleccionar las unidades lexicales (palabras, lemas o categorías) a incluir en los análisis **T-LAB** que implementará posteriormente. También podrá **cambiar de nombre** y **agrupar** dichas unidades lexicales.

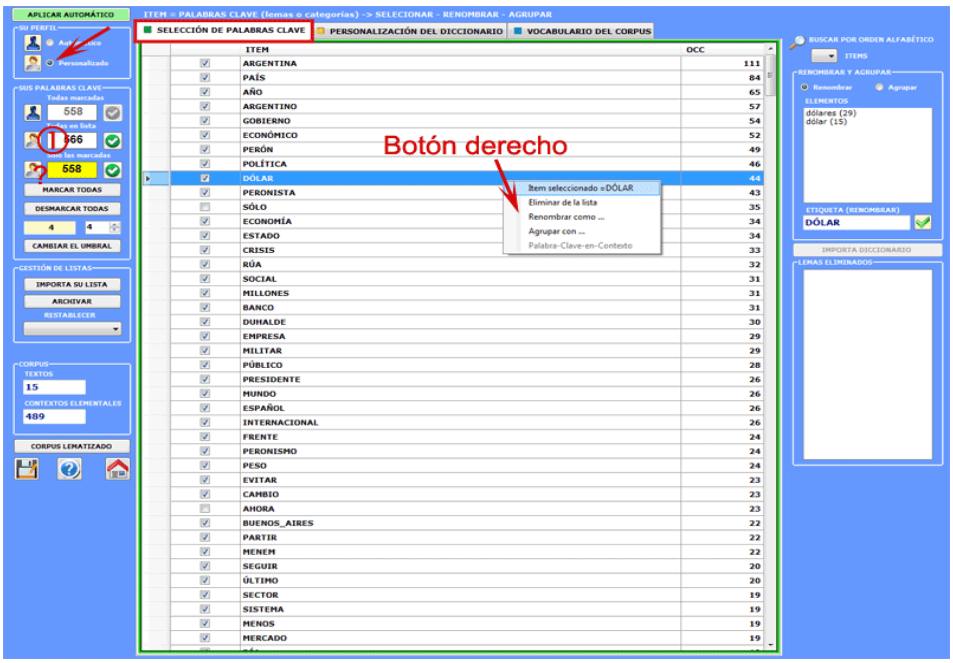
En la tabla se reproduce la lista (lista 1) de las unidades lexicales con valores de ocurrencia iguales o superiores al **umbral** prefijado. Algunas de estas, las indicadas con una "☑", forman parte de una sub-lista (lista 2) sugerida por **T-LAB** (véase **Configuración Automática**).
En función de los análisis que desea efectuar, el usuario decidirá si usar/modificar la lista (1) o la lista (2).

En ambos casos las operaciones posibles son las siguientes:

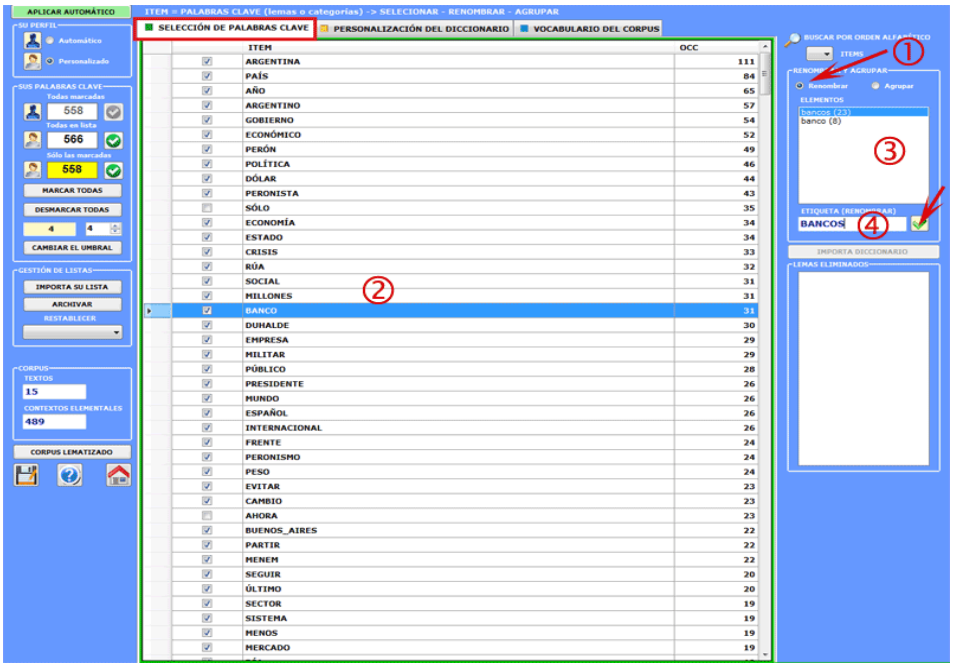
- **cambiar** el valor umbral;
- **seleccionar** qué lemas deben ser excluidos del análisis;
- **restablecer** el uso de uno o varios;
- **seleccionar/de-seleccionar** los ítems de la lista.

Un clic sobre el botón "(1)" o sobre el botón "(2)" permite la opción "personalizada" de análisis.

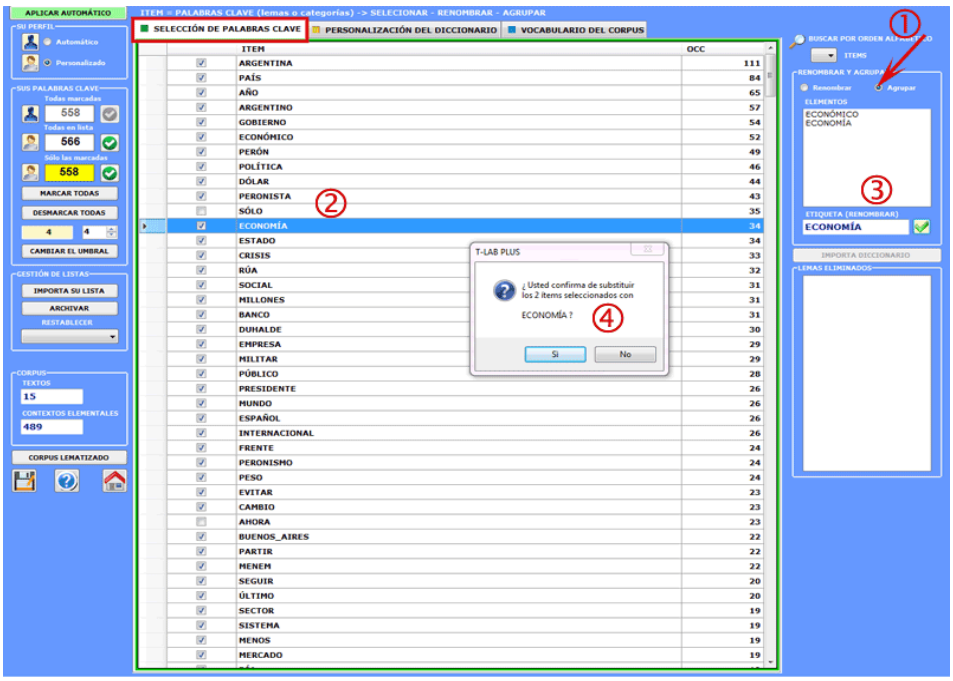
Para cada lema, es posible acceder a las opciones relacionadas con las diferentes acciones a ello asociadas. Para ello, es suficiente seleccionar cualquier ítem de la tabla y hacer clic con el botón derecho del ratón (véase abajo).



Renombrar un solo lema debe llevarse a cabo como sigue:
1 - asegúrese de que está activada la opción "RENOMBRAR";
2 - haga clic sobre ítem de la lista;
3 - elige una de las palabras o tecla una ETIQUETA en su elección;
4 - haga clic en "SUSTITUIR".

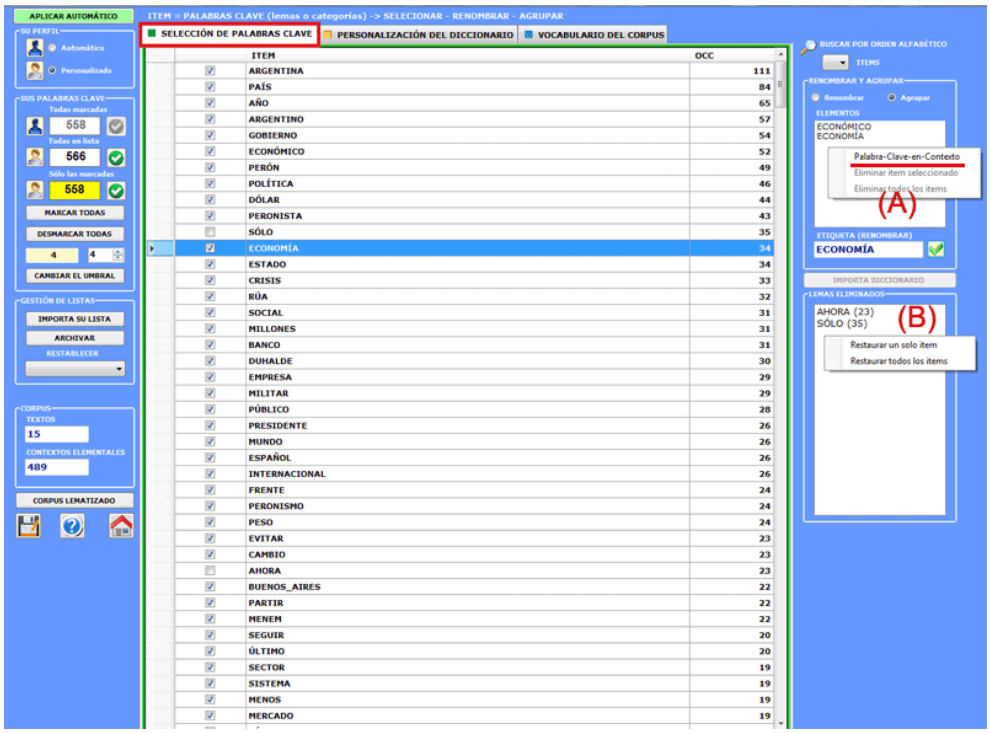


Agrupamientos de dos o más términos se hará como sigue:
1 - seleccionar "AGRUPAR";
2 - haga clic en dos o más ítems de la lista;
3 - elija en uno de los términos o escriba una ETIQUETA en su elección;
4 - haga clic en "SUSTITUIR".

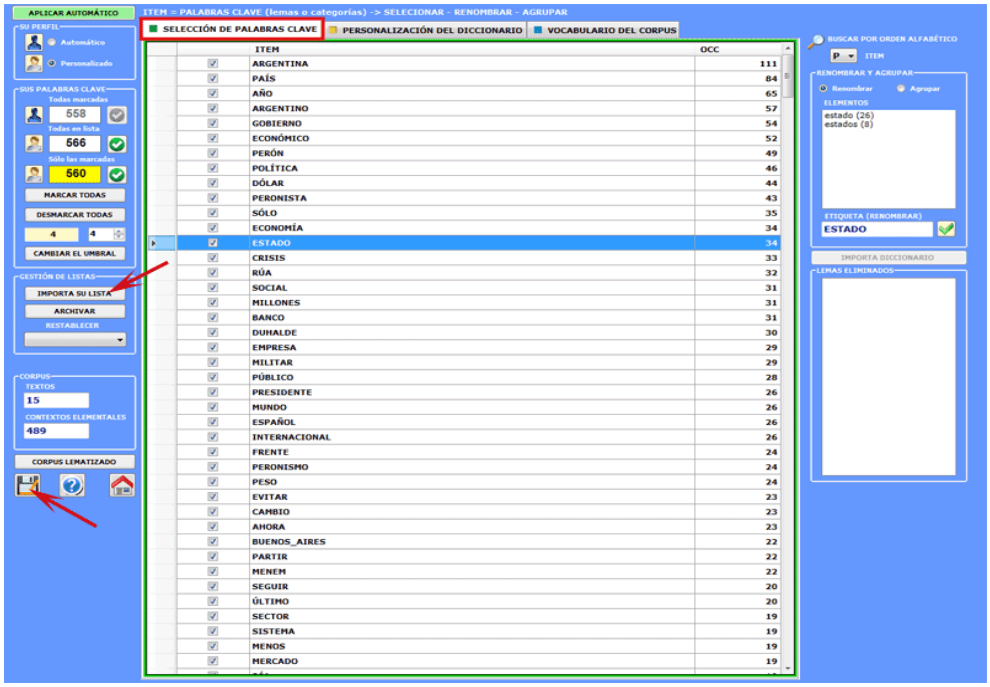


Es posible activar ulteriores opciones si se hace clic con el botón derecho del ratón en el cuadro que incluye los ítems a renombrar/agrupar (A) o en el cuadro que incluye los 'lemas eliminados' (B).

Más en concreto, y en relación al caso (A), si se escoge la opción 'Key-Word-in-Context' es posible acceder automáticamente a la herramienta "concordancias", que permite verificar los contextos de ocurrencia de los diferentes ítems (véase abajo).



Un botón específico (véase abajo) permite **importar listas personalizadas de palabras clave**. Cada lista (archivo llamado MyList.diz) puede incluir hasta un máximo de 10.000 palabras (min = 20). Cada registro de la lista debe ser una palabra sin signos de puntuación y sin espacios en blanco.
Un modelo de archivo MyList.diz se crea automáticamente por T-LAB al guardar cualquier lista de su palabras (véase el botón correspondiente en la parte inferior izquierda).



Las **configuraciones de cada análisis** (hasta un máximo de 10) pueden ser guardadas y ser restablecidas. Eso significa que lo mismo corpus - sin necesidad de una importación adicional - puede ser analizado con varios diccionarios y varias selecciones de palabras.

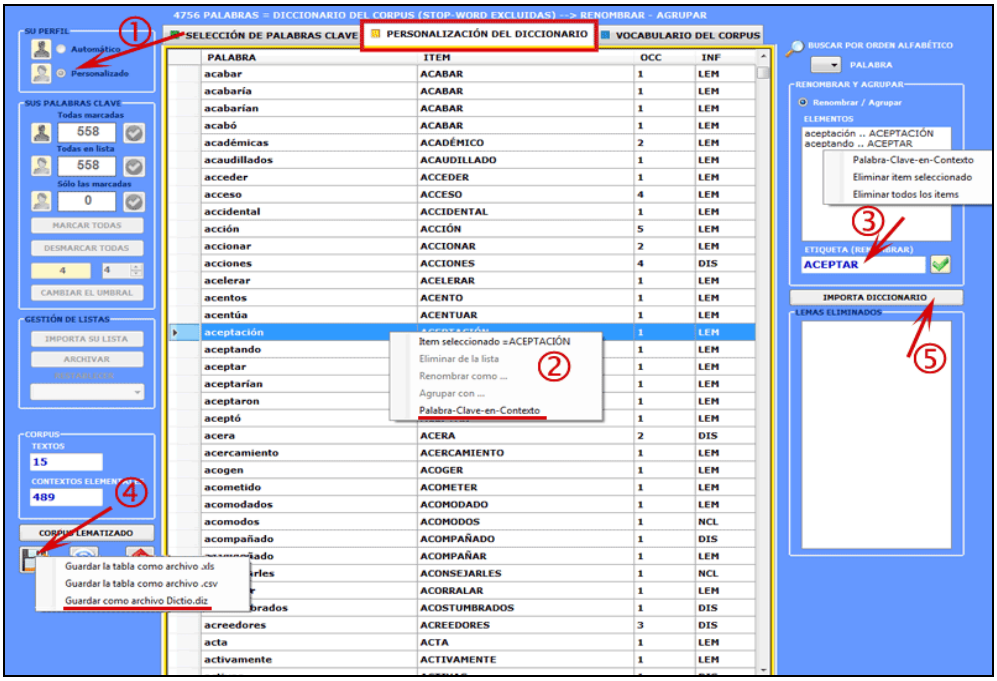
Además, **T-LAB** permite que las configuraciones personalizadas sean realizadas y modificadas durante varias sesiones, también después de operaciones como **Personalización del Diccionario**.

Personalización del Diccionario

La opción **Personalización del Diccionario** abre una ventana en la que el usuario puede realizar algunas operaciones en el diccionario del corpus. Para los propósitos de los análisis siguientes, **su uso es opcional**.

El usuario puede volver a denominar o agrupar los **lemas** disponibles (véase abajo la opción '3');; además, puede exportar el diccionario construido (véase abajo la opción '4') o importar un **diccionario personalizado** (véase abajo la opción '5').

El punto de partida es una tabla (el **vocabulario de T-LAB**) con todas las correspondencias palabra/lema, sus respectivas ocurrencias en el corpus y algunas etiquetas que se refieren a la lematización automática (columna "INF").



Antes de realizar cualquier tipo de operación, es posible verificar las **concordancias** relevantes para el usuario (Key-Word-in-Context) utilizando el botón derecho del ratón (véase arriba la opción '2'); además, después de hacer clic en la pestaña "selección de palabras clave", debe ser seleccionada la configuración personalizada (véase arriba la opción '1').

Las **operaciones posibles**, aún persiguiendo finalidades diversas (revisión de las lematizaciones y/o usos de plantillas para el análisis del contenido), todas se traducen en una reorganización de la base de datos **T-LAB**, y por tanto en tablas diferentes para el análisis de los datos. En particular, se modifica el vocabulario del corpus. De ello se desprende que todas las operaciones se deben realizar en las palabras (lemas o categorías) consideradas interesantes para los análisis sucesivos. **T-LAB**, de hecho, pone a disposición otra opción, **Configuración Personalizada** (véase 'Selección de Palabras Clave'), con la cual los usuarios pueden decidir qué lemas "conservar" y cuáles "descartar".

Las dos funciones (Personalización del Diccionario y Configuración Personalizada) se relacionan mucho entre ellas y el usuario puede moverse fácilmente de acá para allá, también para cambiar su elecciones.

En **Personalización del Diccionario** para cambiar las etiquetas (o 'lemas') asignadas a las palabras, están previstas dos modalidades de acción:

- mover las palabras seleccionadas (clic) al box de la derecha y, sucesivamente, volverlas a denominar usando la opción "substituye" (N.B.: En este caso, la nueva etiqueta puede ser definida usando uno de los lemas seleccionados - haga clic en un elemento presente en el box de "renombrar/agrupar" - o escribiendo en "etiqueta");
- importar un "diccionario personalizado" (N.B.: Esta opción es particularmente aconsejada a los usuarios expertos, puesto que ya disponen de sus listados para clasificar las palabras contenidas en uno o más corpus).

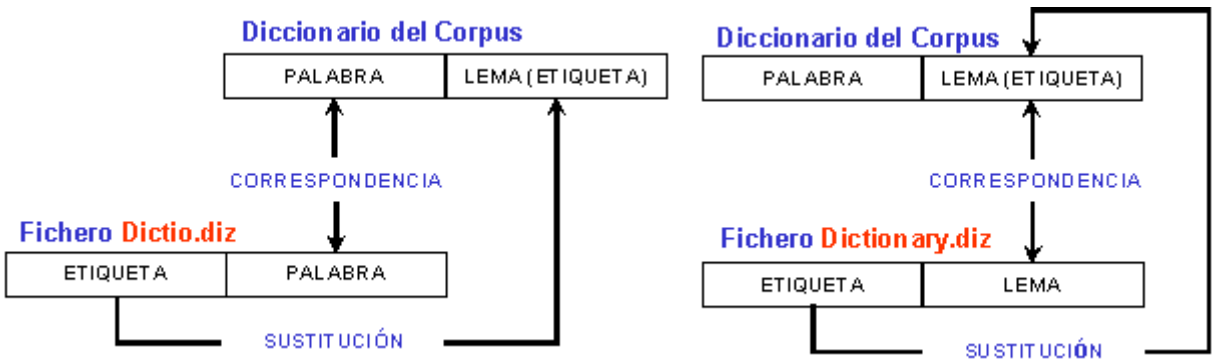
NOTA: Haciendo clic con el botón derecho del ratón en el cuadro "renombrar/agrupar" se activa un menú contextual que permite 3 operaciones: a) verificar las concordancias del ítem seleccionado (KeyWord-in-Context) b) eliminar del cuadro el ítem seleccionado c) eliminar del cuadro todos los ítems seleccionados.

Para poder importar un **diccionario personalizado** es necesario que el usuario haya predispuesto previamente un archivo llamado **Dictio.diz** o un archivo **Dizionary.diz**.
Estos pueden componerse de "n" líneas, cada una con un par de cadenas, separadas por medio del carácter ";".
La longitud máxima de una cadena (palabra, lema o categoría) es de 50 caracteres y en su interior no debe tener ni espacios en blanco ni apostrófos.
Para cada par, la primera cadena - la de izquierda - indica la etiqueta (lema o categoría) definida por el usuario, la segunda indica la palabra (caso **Dictio.diz**) o el lema (caso **Dictionary.diz**) correspondiente ya presente en el diccionario **T-LAB**.

He aquí algunos ejemplos:

(Fichero **Dictio.diz**)	(Fichero **Dictionary.diz**)
CARGAR;carga CARGAR;cargaba CARGAR;cargábamos CARGAR;cargaban ------ ARISTOCRÁTICO;aristocrática ARISTOCRÁTICO;aristocráticas ARISTOCRÁTICO;aristocrático ARISTOCRÁTICO;aristocráticos	 BIOTECH;biotech BIOTECH;biotecnología BIOTECH;biotecnológico --- MENTE_ABSTRACTA;clasifiar MENTE_ABSTRACTA;discernir MENTE_ABSTRACTA;análisis

Según el tipo de fichero que usted importa, los cambios serán como sigue:



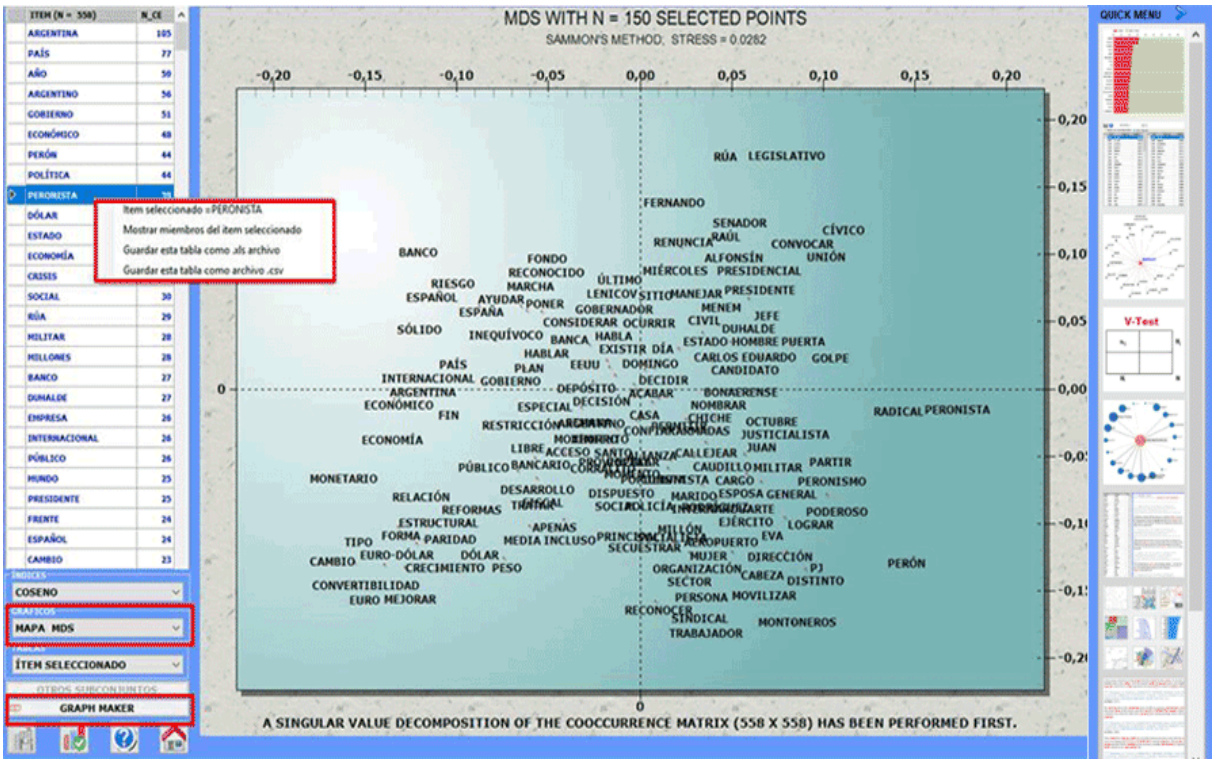
ATENCIÓN:

- Mediante la opción **Corpus Lematizado** es posible exportar una copia del corpus (archivo .txt) en la cual cada palabra será substituida por el lema correspondiente;
- Cuando se ha modificado el diccionario, los análisis siguientes (en el mismo corpus) están disponibles solamente como "configuración personalizada".

ANÁLISIS DE CO-OCURRENCIAS

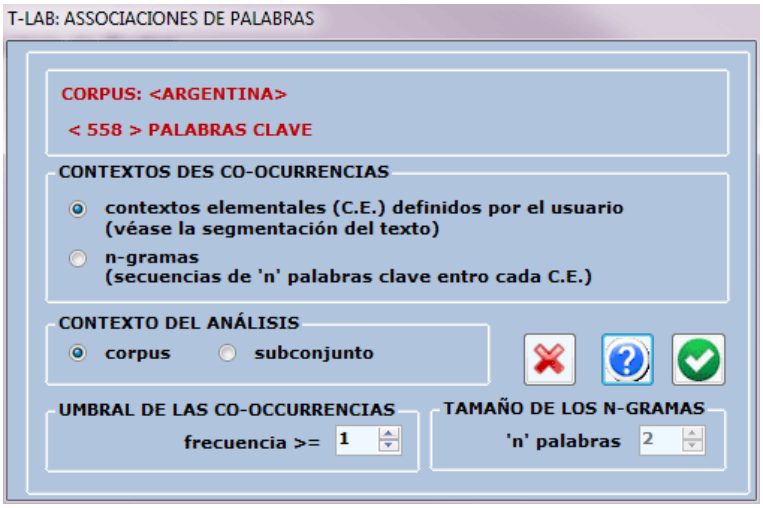
Asociaciones de Palabras

NOTA: Las imagenes contenidas en este apartado hacen a una versión anterior de T-LAB, ya que el interfaz de **T-LAB 10** cambia ligeramente. Además: a) hay una nueva opción que permite al usuario trazar una ' Map Overview' con las palabras más relevantes; b) una nueva herramienta (**'GRAPH MAKER'**) permite crear y exportar diferentes tipos de gráficos dinámicos en formato HTML; c) cliqueando con el **botón derecho** del ratón sobre las tablas que incluyen las palabras clave se accede al listado de las opciones avanzadas; d) una galería de imágenes de acceso rápido que funciona como un menú adicional permite cambiar entre varias salidas con un solo clic.
Algunas de estas nuevas características se destacan en la imagen de abajo.



Esta herramienta de **T-LAB** permite comprobar las relaciones de **co-ocurrencia** y de **semejanza** que, dentro del corpus o de sus sub-conjuntos, determinan el significado local de las **palabras clave** seleccionadas por el usuario.

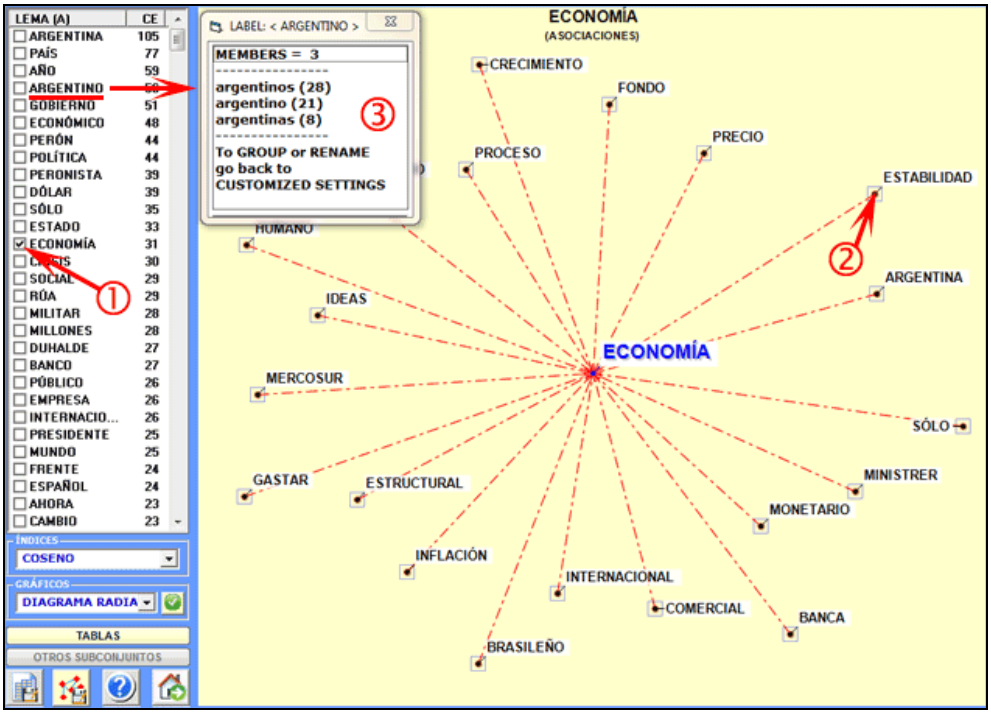
Dicha comprobación puede hacerse mediante las **opciones predeterminadas** (A) o mediante las **opciones personalizadas** por el usuario (B).



En el primer caso (**A**: opciones **predeterminadas**) las **co-ocurrencias** de las palabras se calculan dentro de los **contextos elementales** seleccionados durante la importación del corpus (ejemplo: frases, fragmentos, párrafos, etc.). En el segundo caso (**B**: opciones **personalizadas**), las co-ocurrencias pueden ser calculadas también dentro de las secuencias de palabras que presentan longitud variable (es decir, los **n-gramas**, véase sección del glosario correspondiente). En este último caso, también es posible decidir el umbral mínimo (es decir, la frecuencia) a partir del cual considerar las co-ocurrencias.

Una vez calculadas las co-ocurrencias entre todas las palabras escogidas por el usuario aparece la ventana de trabajo (véase abajo).

En la parte izquierda de la ventana se ubica una tabla. Dicha tabla contiene las palabras y los valores numéricos que indican la cantidad de contextos elementales o de n-gramas en los que aparecen dichas palabras.



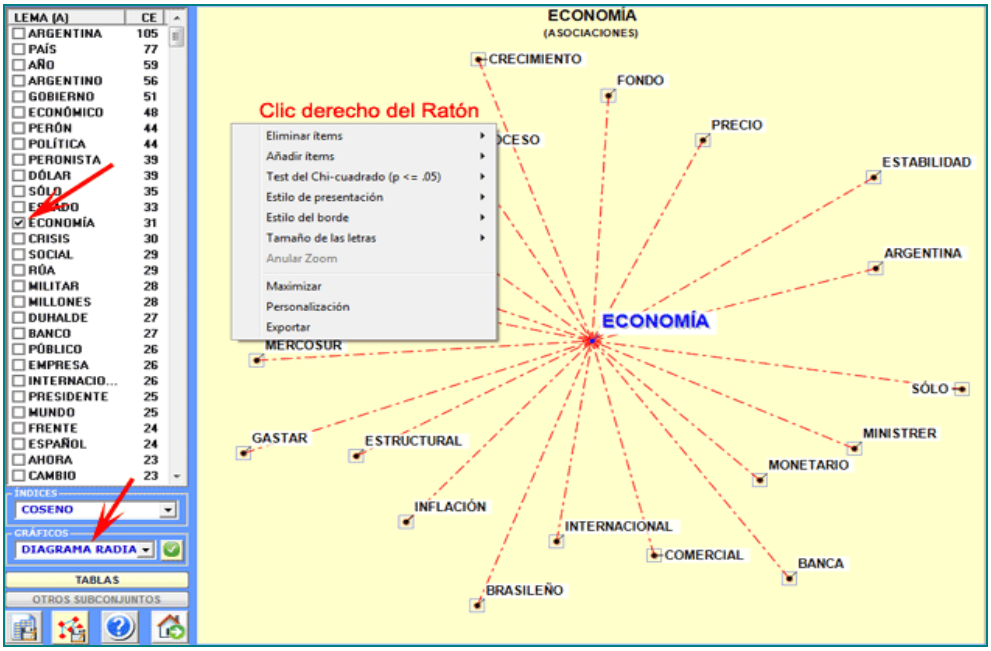
Simplemente con cliquear en los ítems de la tabla (véase arriba, opción 1) o en los puntos de los gráficos (opción 2), se hace posible comprobar las asociaciones de cada una de las palabras objetivo. Por otra parte, si se cliquea en las etiquetas incluidas en la tabla (opción 3), es posible verificar los ítems incluidos en cada lema.

En cada paso, la selección de las palabras asociadas se realiza bien calculando un **Índice de Asociación** (véase el glosario en la sección correspondiente) o bien utilizando un índice de semejanza de **segundo orden**. En el primer caso, hay seis índices a disposición (**Coseno, Dice, Jaccard, Equivalencia, Inclusión y Información Mutua**), cuyo calculo es de rápida ejecución. Sin embargo, en el caso de los índices de segundo orden, el análisis puede tardar algunos minutos, especialmente si el corpus tiene una extensión elevada. Además, es importante considerar que, en el caso de los índices de segundo orden, la fiabilidad de los resultados aumenta al aumentar las palabras incluidas en la lista.

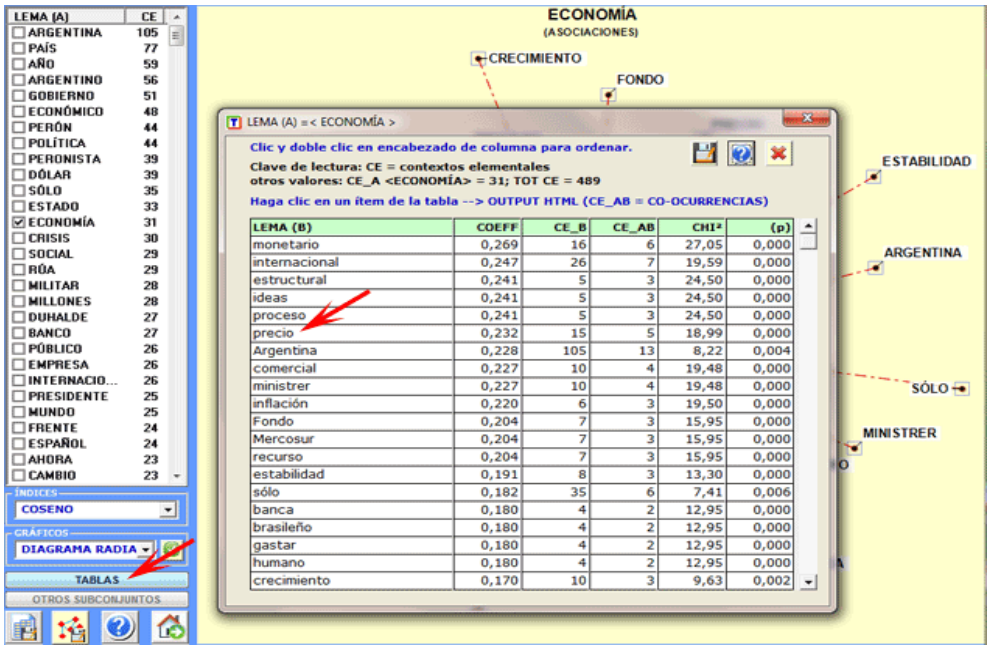
Para cada pregunta, **T-LAB** produce gráficos y tablas.
Tanto las tablas como los gráficos pueden ser guardados utilizando los apropiados botones.

En el **diagrama radial** (véase abajo), el lema seleccionado está en el centro. Los otros se distribuyen alrededor de él, cada uno a una distancia proporcional a su grado de asociación. Por tanto, las relaciones significativas son del tipo uno a uno, entre el lema central y cada uno de los otros.

NOTA: Cada clic en un punto produce un nuevo gráfico y, usando el botón derecho del ratón, es posible abrir una caja de diálogo que permite varias personalizaciones.



Las **tablas** contienen datos que permiten verificar las relaciones entre ocurrencias y co-ocurrencias de las palabras que presentan la asociación más fuerte con aquella seleccionada (máximo 50).

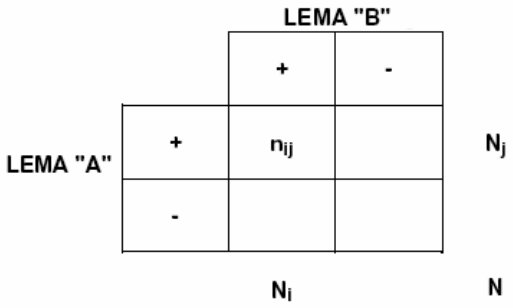


LEMA (B)	COEFF	CE_B	CE_AB	CHI²	(p)
monetario	0,269	16	6	27,05	0,000
internacional	0,247	26	7	19,59	0,000
estructural	0,241	5	3	24,50	0,000
ideas	0,241	5	3	24,50	0,000
proceso	0,241	5	3	24,50	0,000
precio	0,232	15	5	18,99	0,000
Argentina	0,228	105	13	8,22	0,004
comercial	0,227	10	4	19,48	0,000
ministrer	0,227	10	4	19,48	0,000
inflación	0,220	6	3	19,50	0,000
Fondo	0,204	7	3	15,95	0,000
Mercosur	0,204	7	3	15,95	0,000
recurso	0,204	7	3	15,95	0,000
estabilidad	0,191	8	3	13,30	0,000
sólo	0,182	35	6	7,41	0,006
banca	0,180	4	2	12,95	0,000
brasileño	0,180	4	2	12,95	0,000
gastar	0,180	4	2	12,95	0,000
humano	0,180	4	2	12,95	0,000
crecimiento	0,170	10	3	9,63	0,002

Las llaves de lectura son las siguientes:

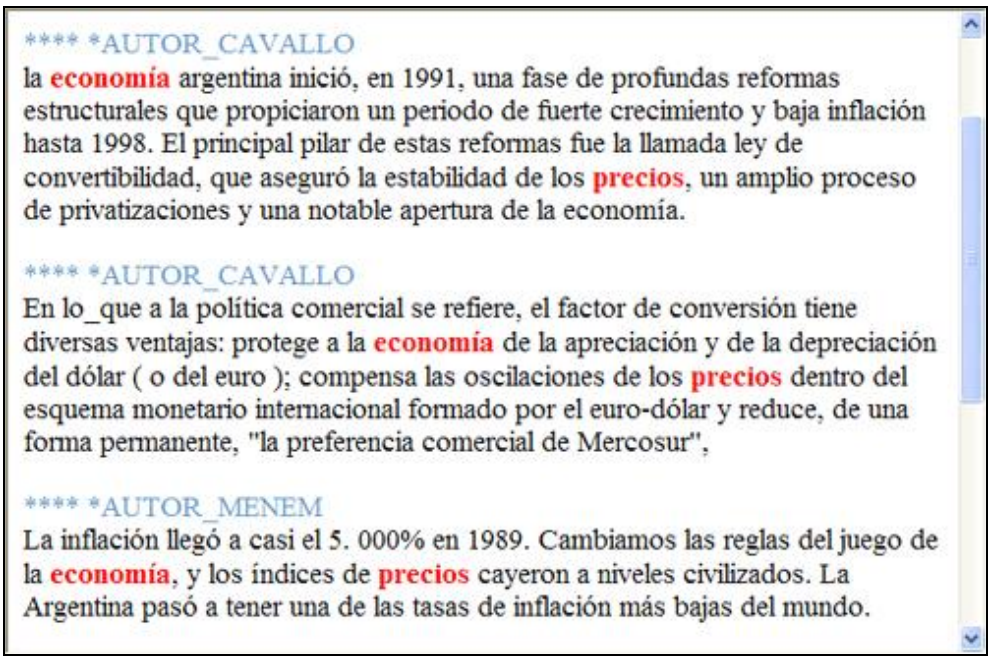
- **LEMA (A)** = lema seleccionado;
- **LEMA (B)** = lemas asociados al LEMA (A);
- **COEFF** = valor del índice de asociación seleccionado;
- **TOT CE** = total de los contextos elementales (CE) o de los n-gramas analizados;
- **CE_A** = total de los CE en los que está presente el lema seleccionado (A);
- **CE_B** = total de los CE en los que está presente cada lema asociado (B);
- **CE_AB** = total de los CE en los que los lema "A" e "B" están asociados (co-ocurrencias);
- **CHI2** = valor del chi cuadrado para verificar la significación de las co-ocurrencias;
- **(p)** = probabilidad asociada a cada valor del chi-cuadrado (def=1).

En el caso del **chi cuadrado**, para cada pareja de lemas ("A" y B") la estructura de la tabla analizada es la siguiente:



En la que : nij = CE_AB ; Nj = CE_A ; Ni = CE_B ; N = TOT CE.

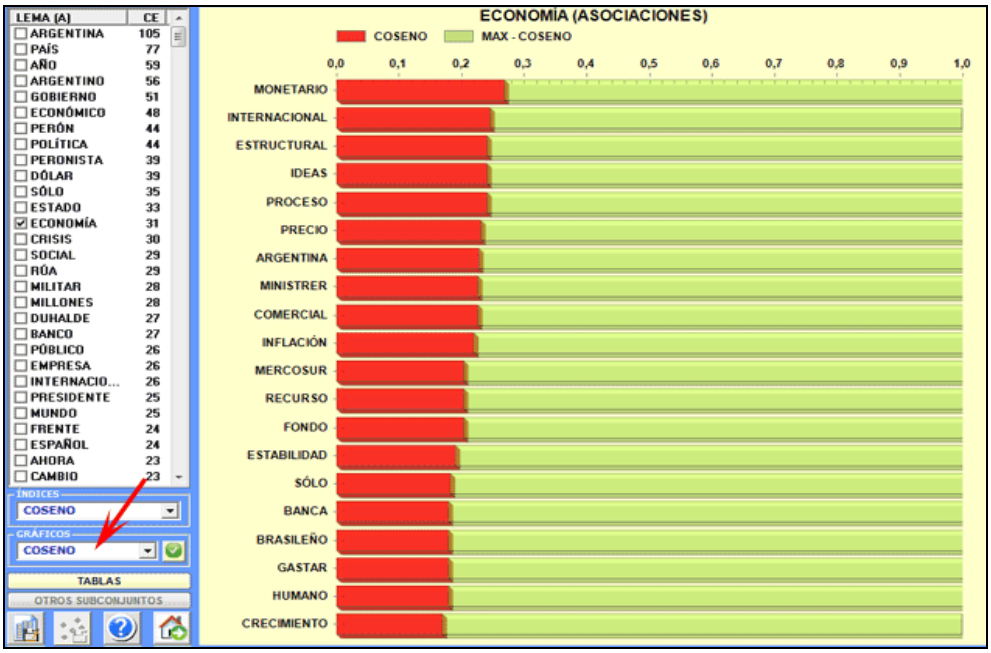
Un clic en cada etiqueta (ej. "precio") de la tabla permite de guardar un archivo con todos los **contextos elementales** donde empareja con la palabra seleccionada (ej. co-ocurrencias de "economía" y " precio").

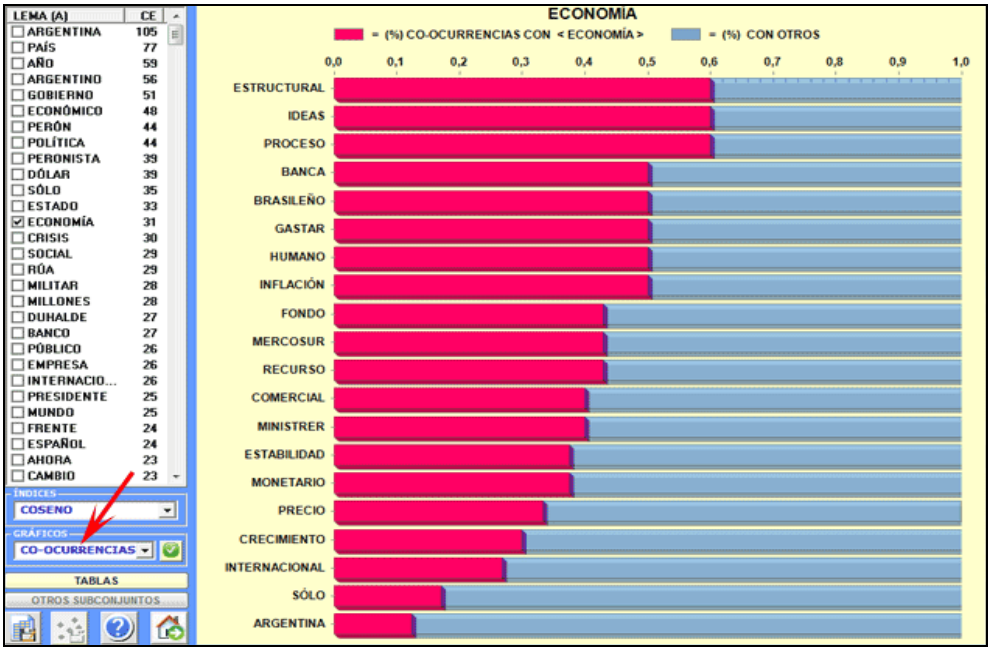

**** *AUTOR_CAVALLO
la **economía** argentina inició, en 1991, una fase de profundas reformas estructurales que propiciaron un periodo de fuerte crecimiento y baja inflación hasta 1998. El principal pilar de estas reformas fue la llamada ley de convertibilidad, que aseguró la estabilidad de los **precios**, un amplio proceso de privatizaciones y una notable apertura de la economía.

**** *AUTOR_CAVALLO
En lo_que a la política comercial se refiere, el factor de conversión tiene diversas ventajas: protege a la **economía** de la apreciación y de la depreciación del dólar (o del euro); compensa las oscilaciones de los **precios** dentro del esquema monetario internacional formado por el euro-dólar y reduce, de una forma permanente, "la preferencia comercial de Mercosur",

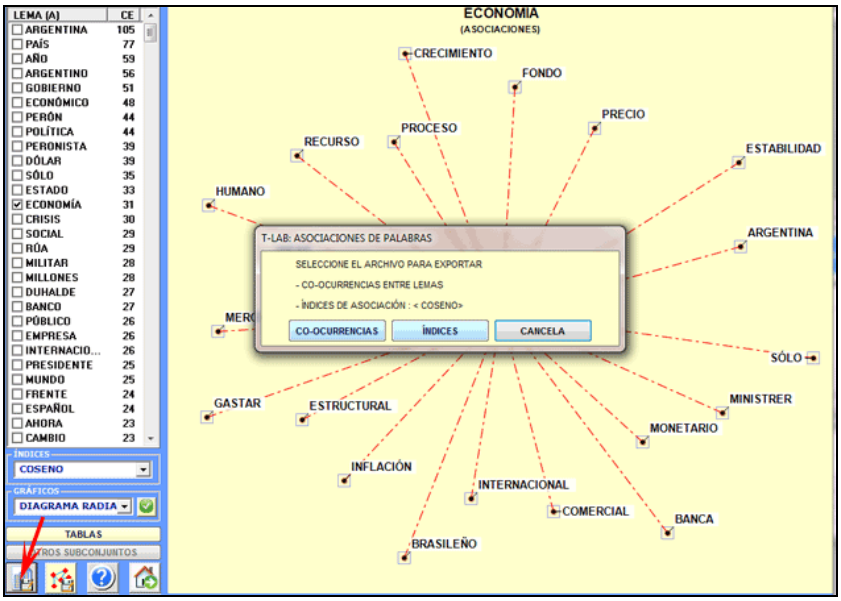
**** *AUTOR_MENEM
La inflación llegó a casi el 5. 000% en 1989. Cambiamos las reglas del juego de la **economía**, y los índices de **precios** cayeron a niveles civilizados. La Argentina pasó a tener una de las tasas de inflación más bajas del mundo.

Ulteriores gráficos (Histogramas) permiten apreciar los valores del coeficiente utilizado y los porcentajes de co-ocurrencias (véase abajo).



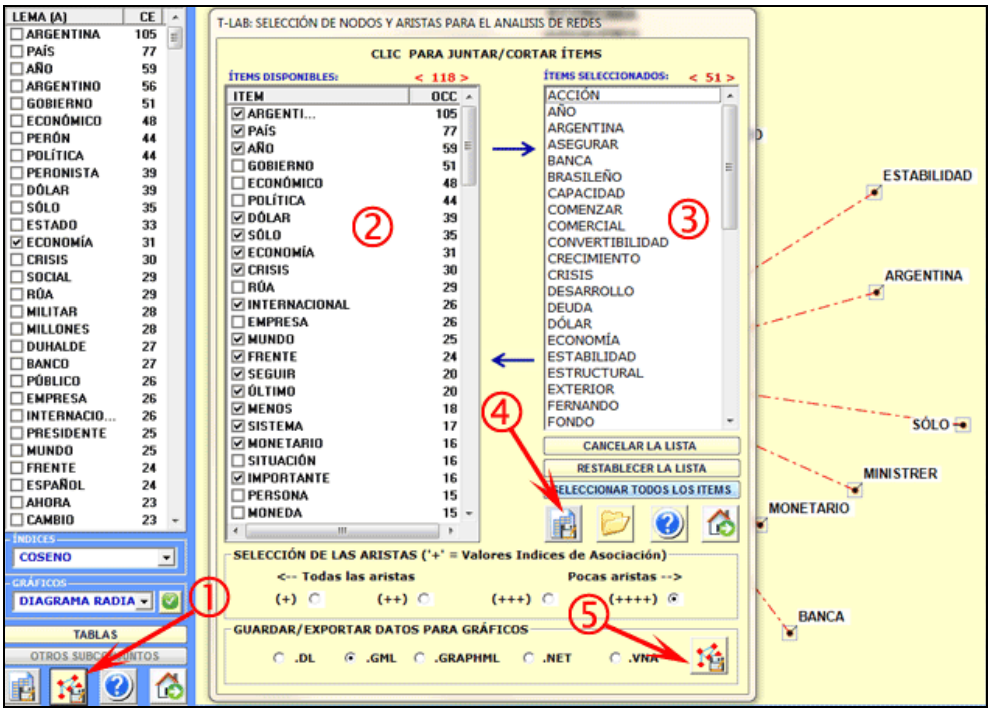


Haciendo clic en el botón de abajo a la izquierda, el usuario puede exportar diferentes tipologías de tablas (véase imagen siguiente).

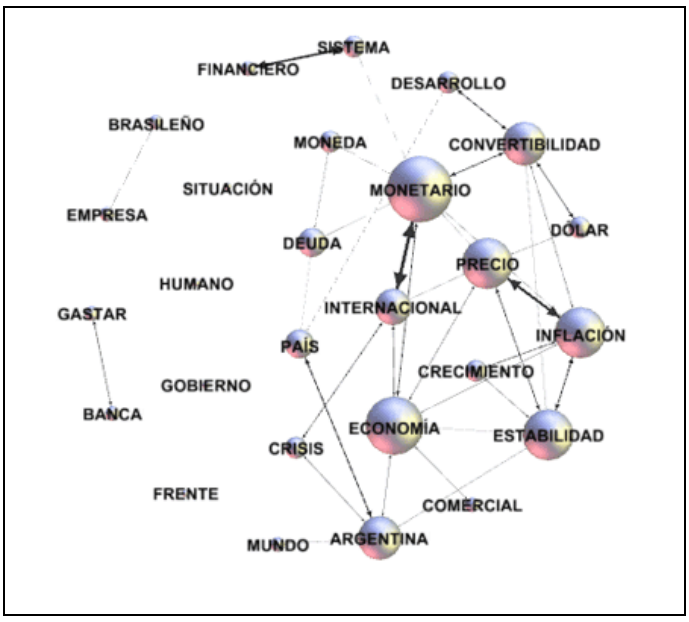


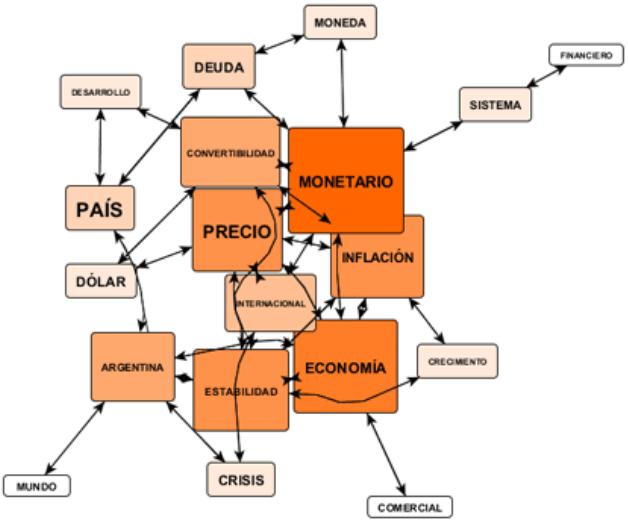
Otra ventana de **T-LAB** (véase imagen siguiente, paso 1) permite crear archivos gráficos que se pueden editar mediante los softwares para el network analysis, como Gephi, Pajek, Ucinet, yEd y otros. En este caso, los **nudos** de la red están formados por las palabras asociadas con la palabra objetivo. Las tres opciones disponibles son: seleccionar los ítems (es decir, los nudos) a insertar en los gráficos (véase abajo, pasos 2 y 3), exportar la matriz de adyacencia correspondiente (véase abajo, paso 4), exportar el tipo de archivo grafico seleccionado (véase abajo, paso 5).

NOTA: En **T-LAB 10** la ventana que se muestra a continuación ha sido sustituida por la herramienta **GRAPH MAKER.**



Por ejemplo, los archivos en formato .gml exportados por **T-LAB** permiten realizar gráficos como los siguientes.





N.B.: El primero de los gráficos se ha creado por medio de Gephi (https://gephi.org/), el segundo por medio de yEd (http://www.yworks.com/en/products_yed_download.html/), ambos softwares disponibles como descarga gratuita.

Las modalidades de cálculo de los diferentes índices de 'asociación' (o proximidad) están indicadas en la sección correspondiente del Manual/Aiuda (véase glosario). Tal y como se podrá observar, todos estos índices se obtienen mediante una normalización de los valores de co-ocurrencia vinculados a las parejas de palabras. Consecuentemente, en los cálculos de **primer orden**, dos palabras que no co-ocurren obtendrían un índice de asociación igual a '0'. Sin embargo, los índices de **segundo orden** evidencian fenómenos de **semejanza** relacionados con el uso (y, por ende, con el significado) de las palabras que no dependen directamente de sus co-ocurrencias. De hecho, en este caso, dos palabras que no co-ocurren pueden llegar a tener un índice de asociación muy elevado.

Utilizando algunos conceptos de la Lingüística Estructural, podemos afirmar que, mientras los índices de 'primer orden' permiten destacar fenómenos asociados al eje sintagmático (combinación y proximidad 'in praesentia', es decir, palabras que dentro de una frase concreta están 'una al lado de la otra'), los índices de 'segundo orden' destacan fenómenos vinculados al eje paradigmático (asociación y semejanza 'in absentia', es decir relaciones de casi-sinonimia entre dos o más términos usados por el mismo autor).

Para comprender la manera con la que **T-LAB** calcula los índices de 'segundo orden' es útil recordar que los índices de 'primer orden' pueden usarse para construir matrices de proximidad como la siguiente (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

Matriz 'A': semejanza de primer orden.

En esta matriz simétrica (A), el valor 0.373 (en amarillo) coincide con el índice de 'primer orden' más alto, e indica la asociación entre las palabras 'w_03' y 'w_10'. Más en concreto, se trata de un índice de equivalencia obtenido dividendo el cuadrado de sus co-ocurrencias entre el producto de sus ocurrencias (360^2/627*553).

A partir de la matriz recién descrita (A), **T-LAB** construye una segunda matriz (B). Para ello, se calculan los cosenos de las comparaciones entre todas las columnas que incluyen los índices de 'primer orden' (véase matriz A). Observando la tabla 'B', es posible constatar cómo el valor de 'semejanza' más elevado es aquel que caracteriza la relación entre las palabras 'w_06' e 'w_08'. Esto quiere decir que los vectores correspondientes (véanse las dos columnas en verde de la matriz 'A') son muy parecidos entre sí (coseno = 0.905), pese a que la asociación de 'primer orden' entre las dos palabras en cuestión es bastante baja (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

Matriz 'B': semejanza de segundo orden.

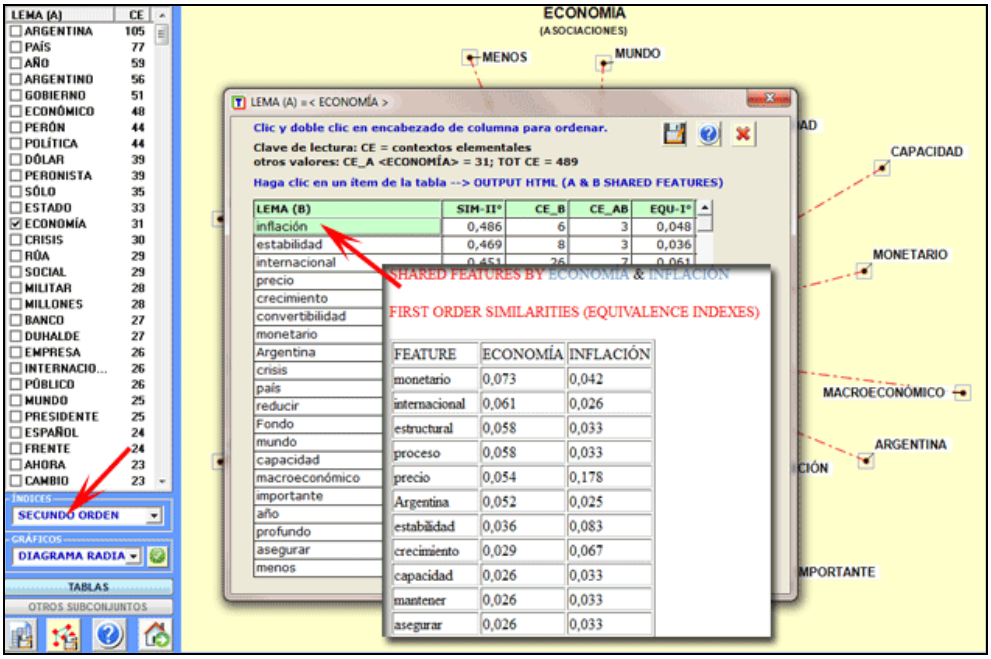
Dicho de otro modo, un índice de 'primer orden' se obtiene a partir de una fórmula que incluye valores de co-ocurrencia y ocurrencia, mientras que un índice de 'segundo orden' se obtiene multiplicando dos vectores normalizados.

Más allá de las modalidades de cálculo, cabe recordar que en los dos casos ('A' e 'B') subyacen dos fenómenos distintos. En el primer caso ('A') nos centramos en las co-ocurrencias mientras que, en el segundo caso ('B'), y independientemente de las co-ocurrencias, nos centramos en la semejanza entre 'perfiles' cuyos datos hacen referencia al uso de palabras por parte de los autores de los textos analizados.

A modo de ejemplo, se considera el análisis de primer orden de **Pinocho**, donde el término 'hada' está principalmente asociado (véanse co-ocurrencias) con 'buena' y 'pelo turquesa'. Sin embargo, en el análisis de segundo orden el término que resulta ser más parecido a 'hada' es 'mamá'. Todo ello, pese a que las co-ocurrencias entre los términos 'hada' y 'mamá' son, dentro del cuento de Collodi, prácticamente irrelevantes (sólo 3).

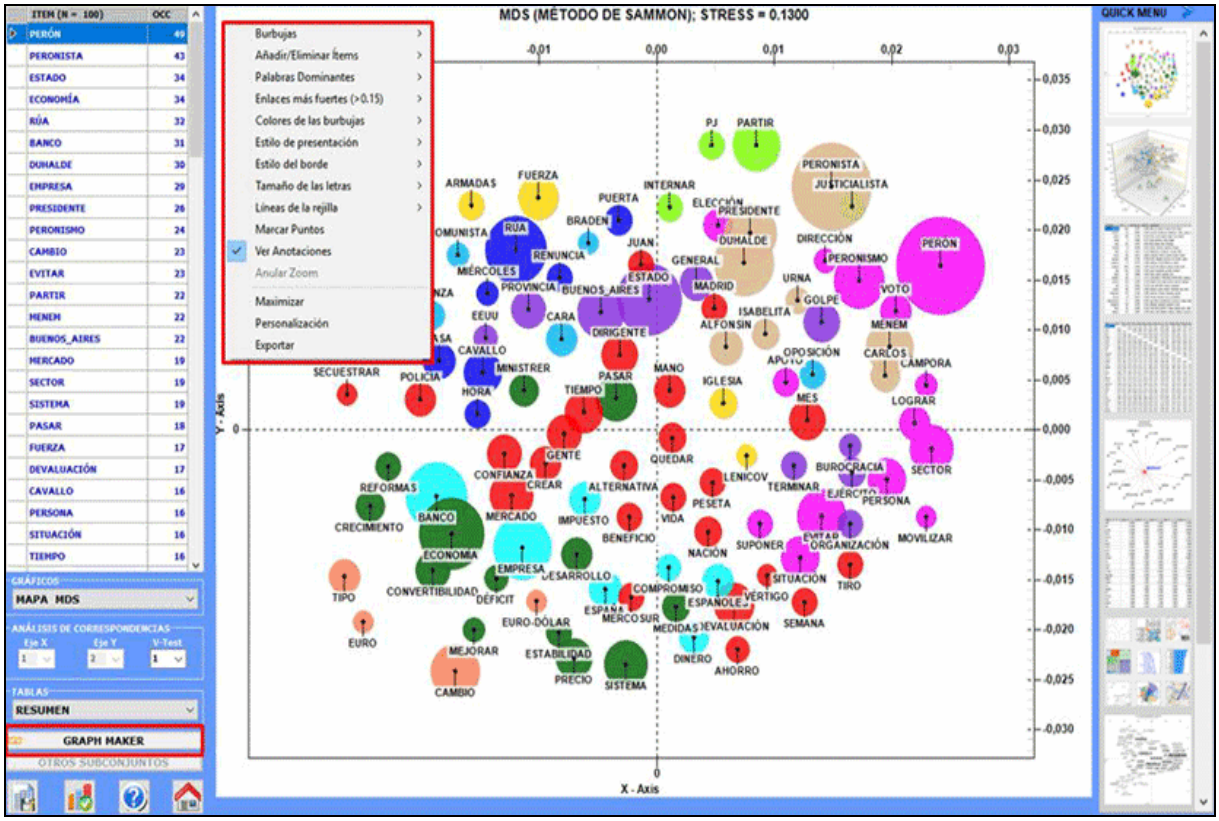
Las tablas visualizadas por **T-LAB** permiten verificar tanto las semejanzas de segundo orden (véase abajo columna SIM-II°), como los índices de primer orden (EQU-I°, es decir, índices de equivalencia).

Además, cliqueando en cada ítem de esta tabla, es posible abrir unos archivos HTML que permiten verificar qué características (‘features’) determinan las semejanzas de segundo orden entre cada pareja de palabras. Por ejemplo, en la siguiente tabla se observa como la semejanza de segundo orden entre ‘economia’ y ‘inflación’ está determinada principalmente por características compartidas como ‘monetario’, ‘internacional’, ‘estructural’, etc.



Análisis de Co-Palabras y Mapas Conceptuales

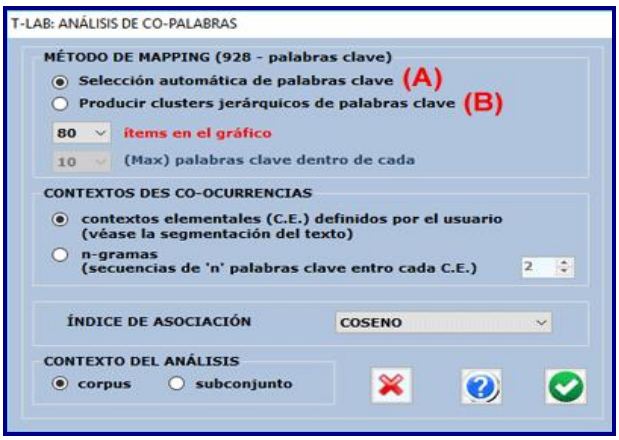
NOTA: Las imagenes contenidas en este apartado hacen referencia a una versión anterior de T-LAB, ya que el interfaz de **T-LAB 10** cambia ligeramente. Además: a) si se escoge la opción 'selección automática de palabras clave', los diferentes clústeres de elementos vienen representados en los mapas MDS con colores diferentes; b) se ha agregado la técnica de visualización llamada t-SNE (t-Distributed Stochastic Neighbor Embedding); c) una nueva herramienta (‘**GRAPH MAKER**’) permite crear y exportar diferentes tipos de gráficos dinámicos en formato HTML; d) el uso del **botón derecho del ratón** sobre las tablas que incluyen las palabras clave permite acceder a las las opciones avanzadas; e) una galería de imágenes de acceso rápido que funciona como un menú adicional permite cambiar entre varias salidas con un solo clic. Algunas de estas nuevas características se destacan en la imagen de abajo.



Esta herramienta **T-LAB** permite analizar dos tipos de relaciones concernientes las **co-ocurrencias** de palabras:

A - entre las **palabras-clave** seleccionadas (lemas o categorías), si sus cantidad no excede los 500 elementos (mínimo 10).

B - entre (y dentro) **clusters** (es decir **Núcleos Temáticos**), si la cantidad de **palabras-clave** seleccionadas excede los 100 elementos (máximo 3.000);



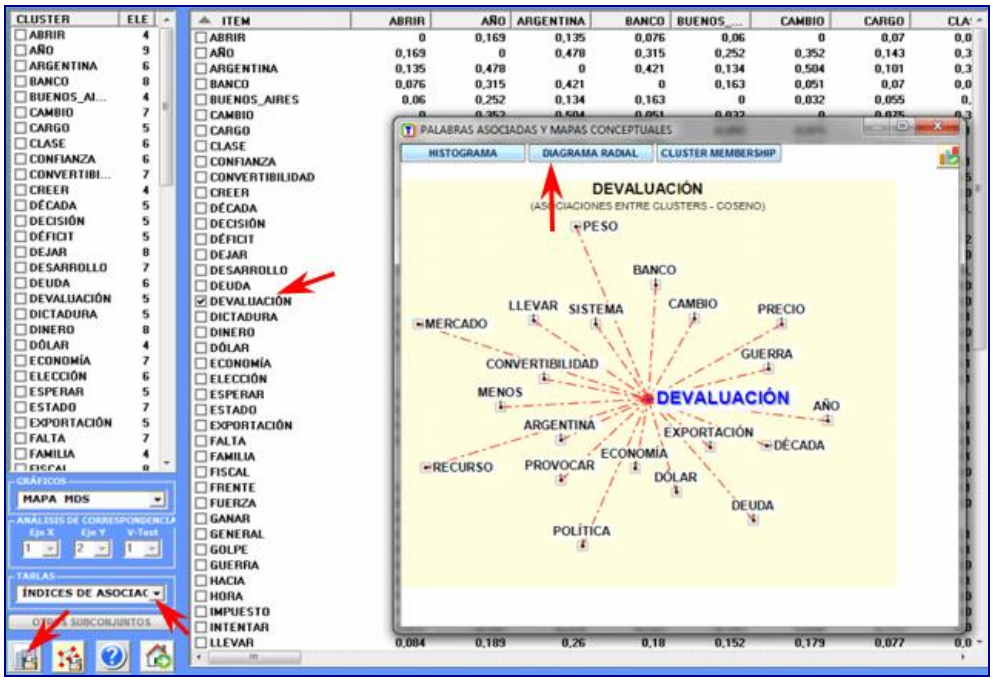
El usuario puede seleccionar el **índice de asociación** a ser utilizado y, sólo en el caso de la opción B, puede seleccionar tanto la cantidad máxima de clusters a obtener (de 50 a 100) como la cantidad máxima de palabras clave por cluster.

El proceso de cálculo empleado prevé los pasos siguientes:

1- construcción de una matriz de co-ocurrencias (palabra x palabra);
2- cálculo de los índices de asociación seleccionados (Coseno, Dice, Jaccard, Equivalencia, Inclusión, Información Mutua);
3- clustering jerárquico de la matriz de la desemejanza;
4- construcción de una segunda matriz de co-ocurrencias (cluster x cluster);
5- representación gráfica de las relaciones a través del modelo del Multidimensional Scaling y del Análisis de Correspondencias.

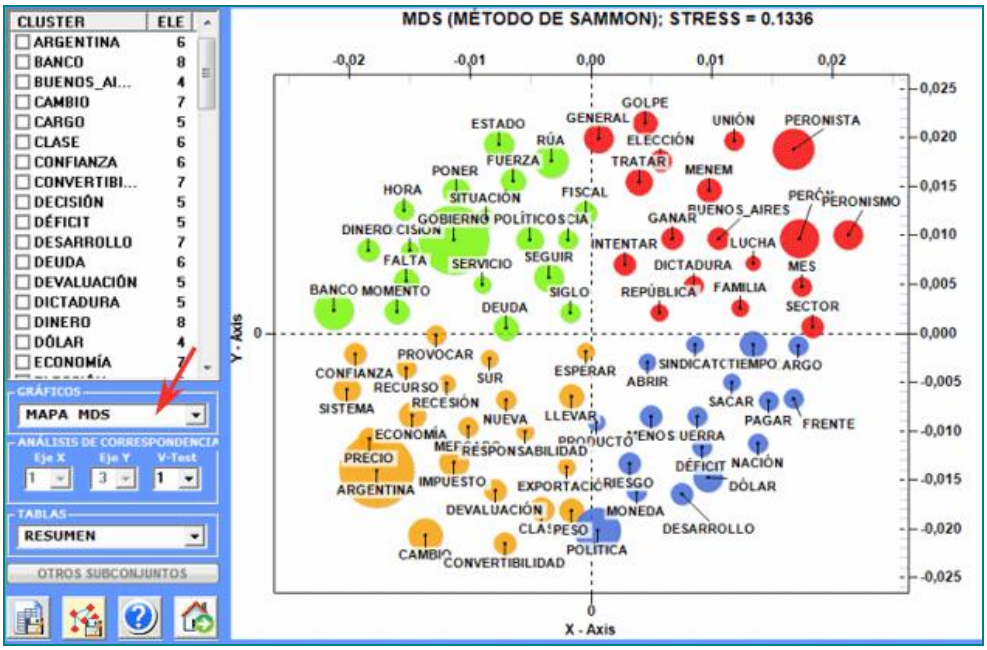
N.B:
- en el caso 'A' (véase arriba), el usuario puede revisar y personalizar la selección de las palabras-clave (véase imagen siguiente) y **T-LAB** no implementaría las fases 3 y 4;

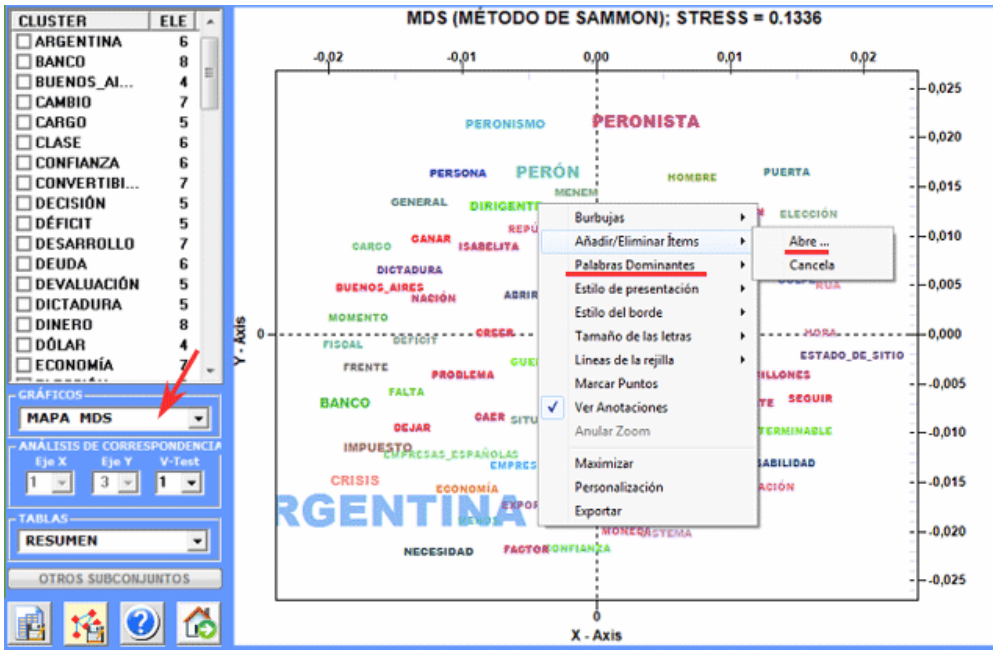
- la calidad de resultados depende de una cuidadosa **selección de palabras clave**;
- puesto que las **multi-palabras** (multiwords) no catalogadas por **T-LAB** son casos específicos de co-ocurrencia y que la opción "B" trata esos como unos pequeños racimos (ej. "Twin" + "Towers"), se aconseja de resolver estos casos durante la fase de **importación**. En todo caso, sin la repetición de la importación del corpus, es posible realizar cambios por medio de la función **Personalización del Diccionario** (ej. asignando la etiqueta "Twin_Towers" a los dos diversos items "Twin" + "Towers");
- haciendo clic sobre los botones apropiados todas las tablas de datos pueden ser comprobadas (véase abajo).



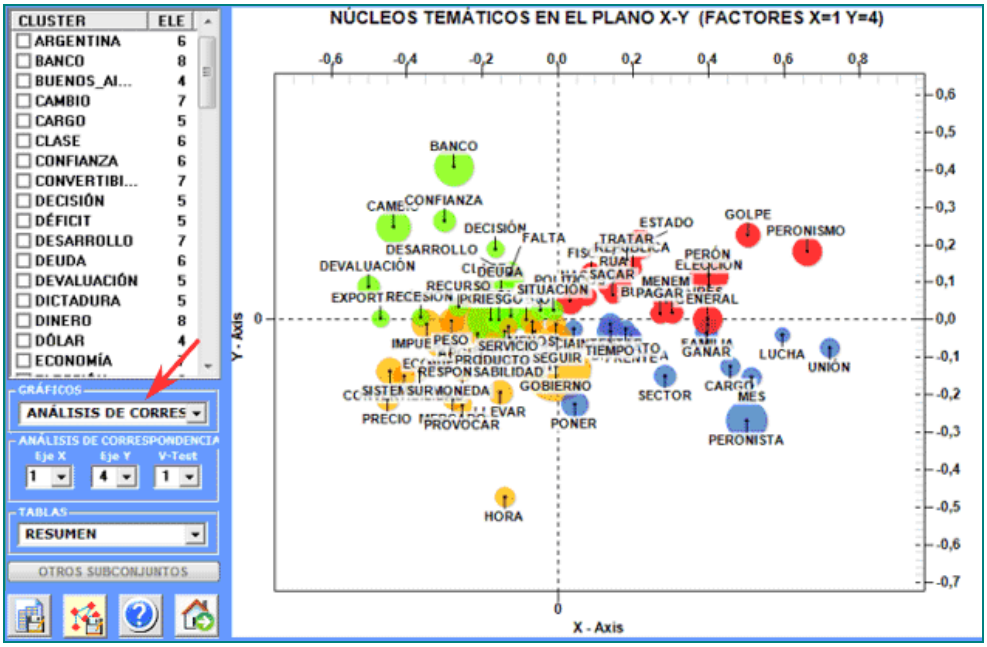
Al término del análisis automático, están disponibles cuatro tipos de gráficos que pueden ser personalizados utilizando el botón derecho del ratón.

1 - **Mapa MDS**

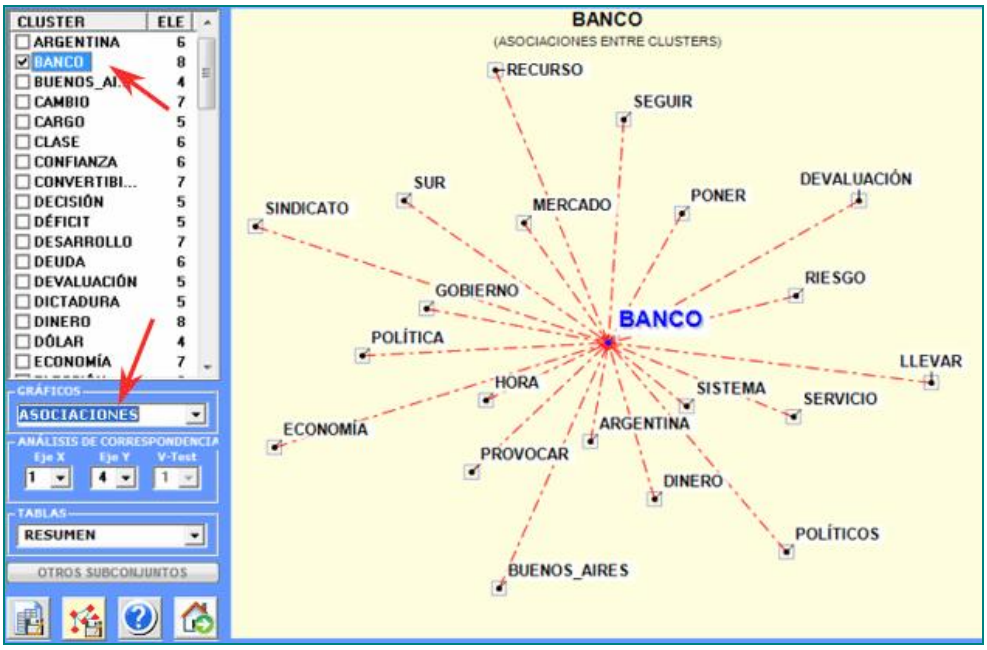

MDS (MÉTODO DE SAMMON); STRESS = 0.1336
CLUSTER
ELE
ARGENTINA 6
BANCO 8
BUENOS_AI... 4
CAMBIO 7
CARGO 5
CLASE 6
CONFIANZA 6
CONVERTIBI... 7
DECISIÓN 5
DÉFICIT 5
DESARROLLO 7
DEUDA 6
DEVALUACIÓN 5
DICTADURA 5
DINERO 8
DÓLAR 4
ECONOMÍA 7
GRÁFICOS
MAPA MDS
ANÁLISIS DE CORRESPONDENCIA
Eje X
Eje Y
V-Test
TABLAS
RESUMEN
OTROS SUBCONJUNTOS
X - Axis
Y - Axis


MDS (MÉTODO DE SAMMON); STRESS = 0.1336
Burbujas
Añadir/Eliminar Ítems
Palabras Dominantes
Estilo de presentación
Estilo del borde
Tamaño de las letras
Líneas de la rejilla
Marcar Puntos
Ver Anotaciones
Anular Zoom
Maximizar
Personalización
Exportar
Abre ...
Cancela
X - Axis
Y - Axis

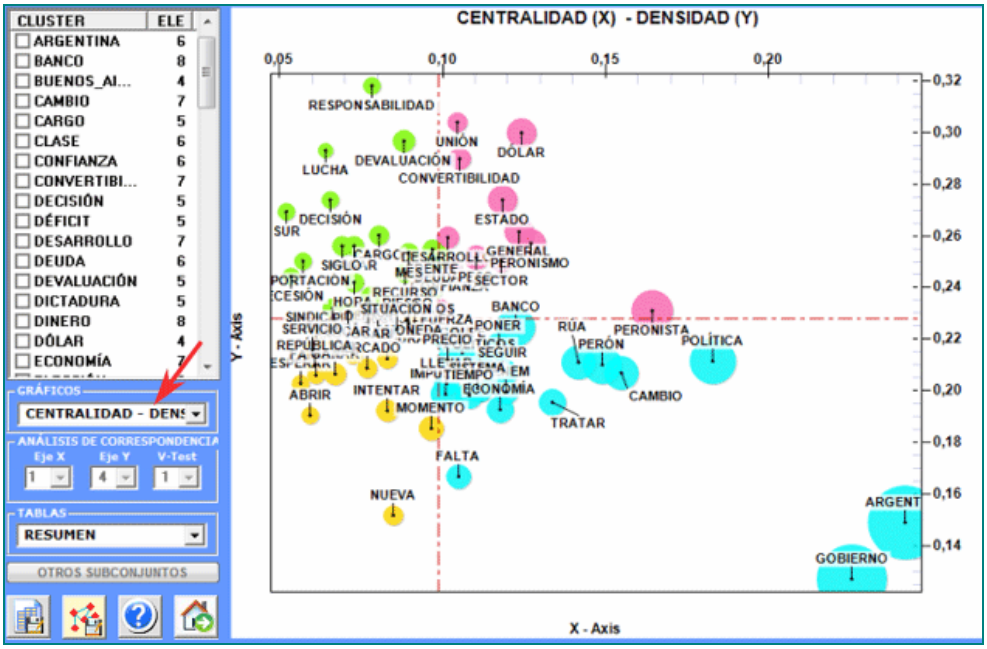
2 - Análisis Factorial de Correspondencias



3 - Diagrama de Asociaciones



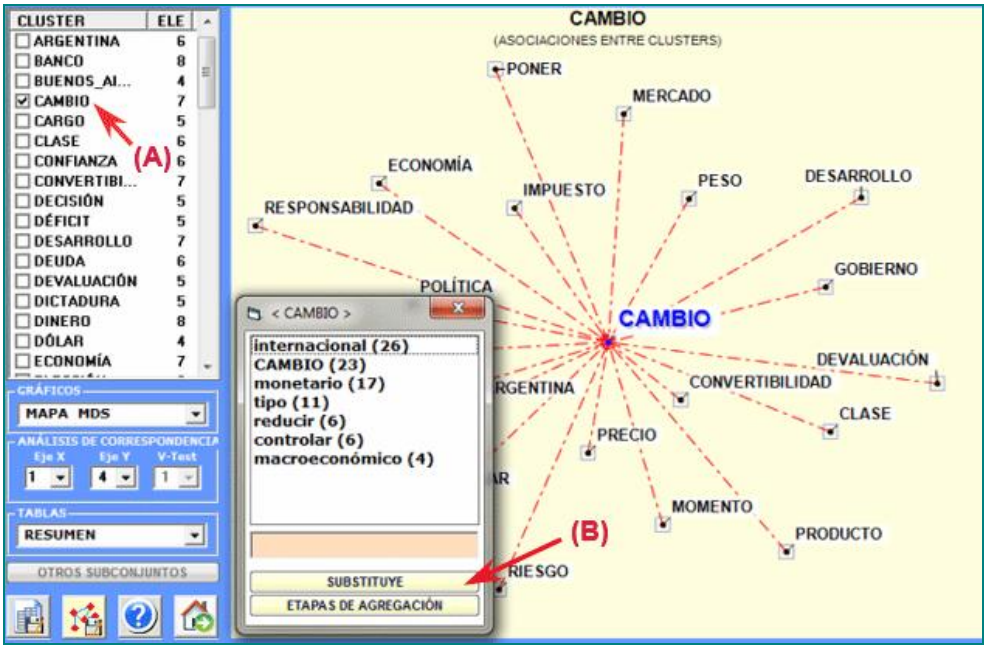
4 - Mapa con las medidas de **Centralidad** y **Densidad** (solo después de un cluster análisis)

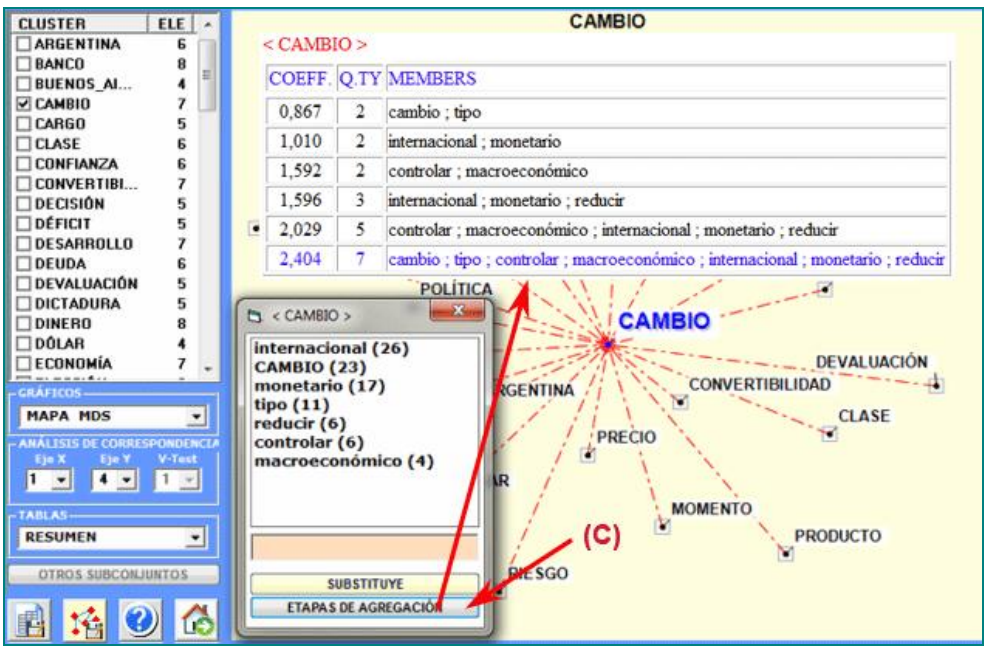


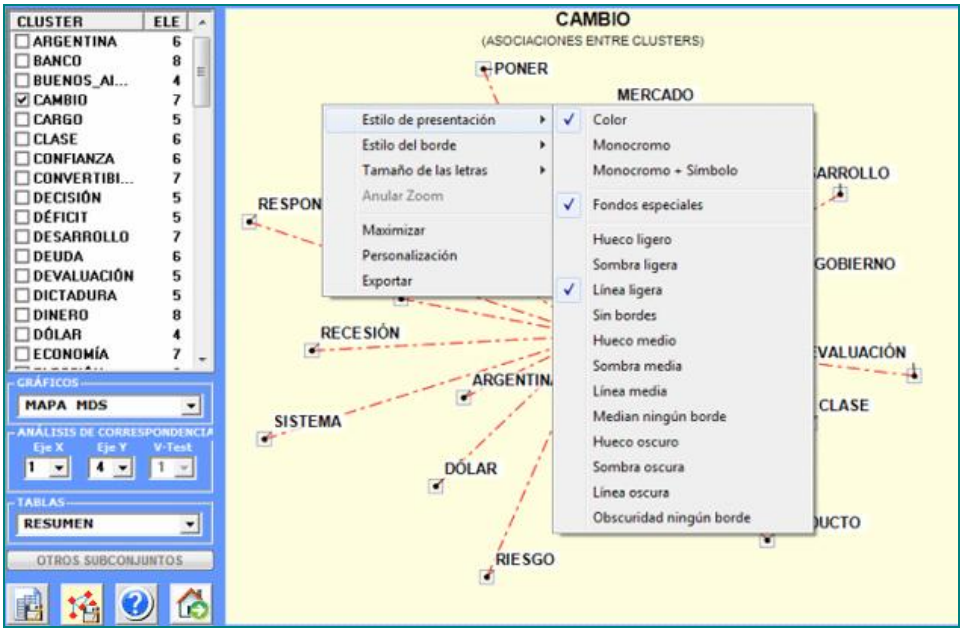
En detalle, los resultados obtenidos por el **Análisis de Correspondencias** se pueden visualizar usando las coordenadas de los primeros diez ejes (véase "A" abajo). Puesto que **T-LAB** nos permite verificar los valores test de cada factor (véase "B" abajo), esto tipo de output se puede utilizar para una cuidadosa interpretación de las relaciones entre clusters y/o entre palabras clave.

Los gráficos pueden ser explorados y modificados de las maneras siguientes:

ACCIÓN	RESULTADO
clic en un ítem de la tabla o en un punto del gráfico	diagrama de las asociaciones correspondientes
clic en una etiqueta de la columna "CLUSTER" (véase "A" abajo)	lista con los elementos del cluster
clic en el botón "Substituye" (véase "B" abajo)	nueva etiqueta atribuida al cluster
clic en el botón "Etapas de agregación" (véase "C" abajo)	agregaciones en el cluster
botón derecho del ratón	caja de diálogo para personalizar los gráficos

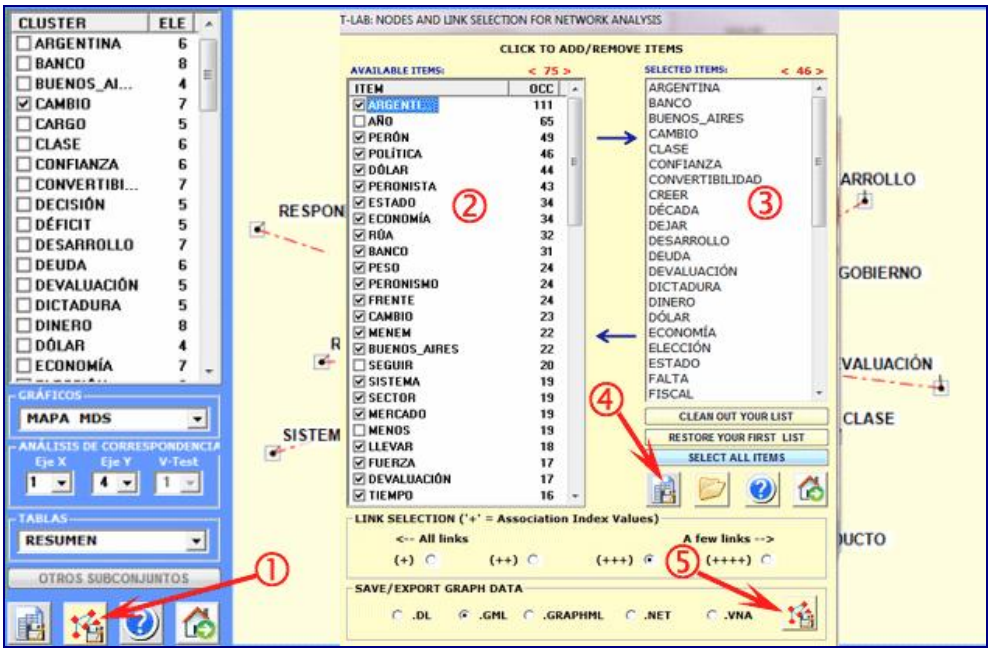




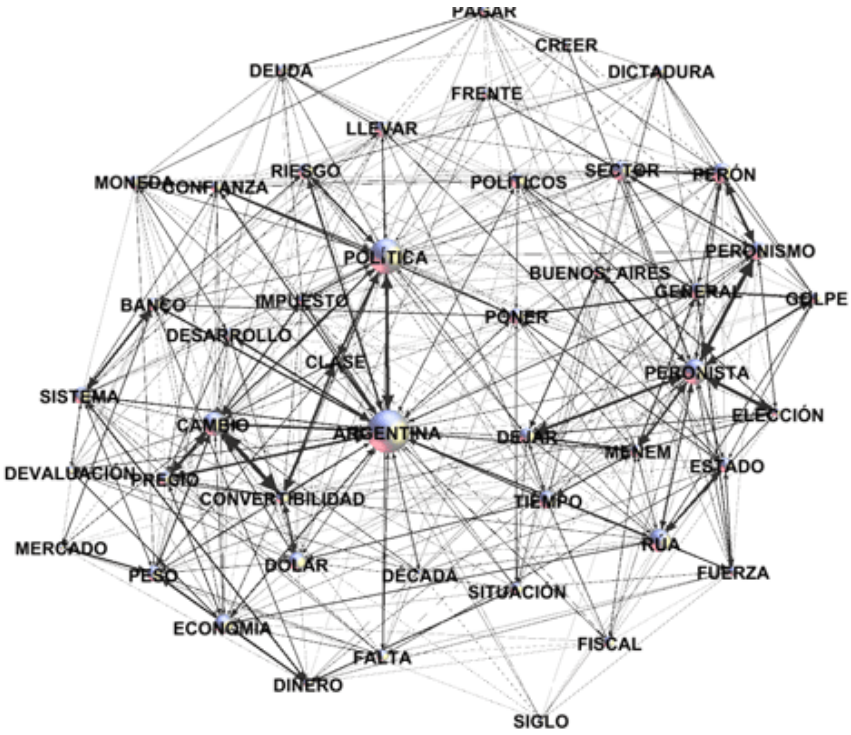


Otra ventana **T-LAB** (véase ventana siguiente, paso 1) permite crear archivos gráficos que pueden ser editados mediante los softwares para el **network analysis**, como Gephi, Pajek, Ucinet, yEd entre otros. En este caso, las opciones disponibles son las siguientes: seleccionar los ítems (es decir, los nudos) a insertar en los gráficos (véase abajo, pasos 2 y 3), exportar la matriz de adyacencia correspondiente (véase abajo, paso 4), exportar el tipo de archivo seleccionado (véase abajo, paso 5).

NOTA: La ventana que se presenta a continuación ha sido sustituida por la herramienta **GRAPH MAKER**.

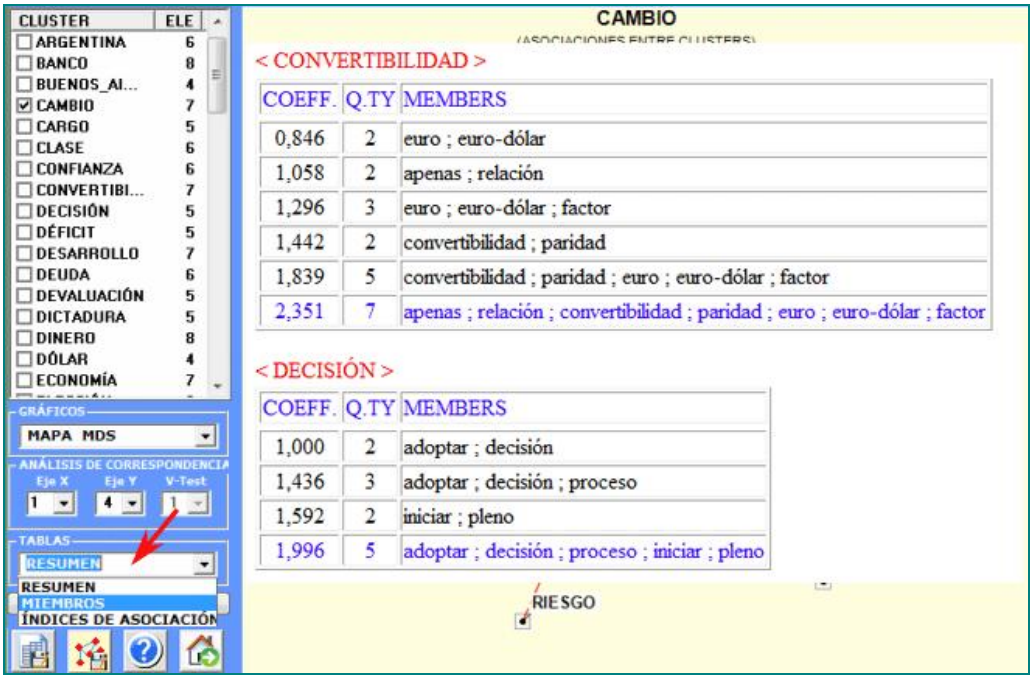


Por ejemplo, un archivo en formato .gml exportado por **T-LAB** permite realizar un grafico como el siguiente.



Hay tres tipologías de tablas exportables a través de esta herramienta de **T-LAB**:

1 - la tabla "Miembros" se refiere a la agregación jerárquica de palabras dentro de cada cluster;



2 - la tabla "Resumen" (véase abajo) incluye las medidas siguientes:

- ECQ = cantidad de contextos elementales en los cuales dos o más palabras del cluster son co-ocurrentes;
- Centrality (Centralidad) = media de índices de asociación referentes a las relaciones entre clusters;
- Density (Densidad) = media de índices de asociación de palabras dentro de cada cluster.

CLUSTER	ECQ	CENTRALITY	DENSITY	MEMBERS
ARGENTINA	85	0,242	0,149	ARGENTINA; CAPACIDAD; CRISIS; MUNDO; PAÍS; SOCIEDAD
BANCO	59	0,122	0,225	BANCO; BBVA; EMPRESA; ESPAÑOL; ESPAÑOLES; ESTRATÉGICO; INVERSIÓN; PERDIDO
BUENOS_AIRES	13	0,083	0,233	BUENOS_AIRES; DIFÍCIL; ENTENDER; PROVINCIA
CAMBIO	55	0,155	0,207	CAMBIO; CONTROLAR; INTERNACIONAL; MACROECONÓMICO; MONETARIO; REDUCIR; TIPO
CARGO	14	0,081	0,260	CARGO; CONVERTIR; CREAR; LÍDER; NOMBRAR
CLASE	27	0,112	0,201	CLASE; FORMA; INCLUSO; MEDIA; PARTE; PESADO
CONFIANZA	23	0,102	0,248	CONFIANZA; CREDIBILIDAD; DEMOCRÁTICO; INSTITUCIÓN; INTERIOR; SÓLIDO
CONVERTIBILIDAD	26	0,105	0,290	APENAS; CONVERTIBILIDAD; EURO; EURO-DÓLAR; FACTOR; PARIDAD; RELACIÓN
DECISIÓN	11	0,066	0,274	ADOPTAR; DECISIÓN; INICIAR; PLENO; PROCESO
DÉFICIT	12	0,078	0,237	BUROCRACIA; DÉFICIT; POBLACIÓN; SINDICAL; SUMA
DESARROLLO	18	0,102	0,259	DEPENDER; DESARROLLO; GRADO; MANTENER; PUNTO; RESULTAR; SUBIR

3 - la tabla "Índices de asociación" (véase abajo) incluye medidas de las relaciones entre (between) y dentro (within) los clusters.

Between

< ARGENTINA >

CLUSTER	INDEX
politica	0,591
precio	0,511
cambio	0,504
GOBIERNO	0,478
riesgo	0,431
Desarrollo	0,426
banco	0,421
clase	0,391
economia	0,370
momento	0,355
fuerza	0,355
tiempo	0,350
sistema	0,342
confianza	0,342
menos	0,329
seguir	0,320
Menem	0,313
general	0,310

Within

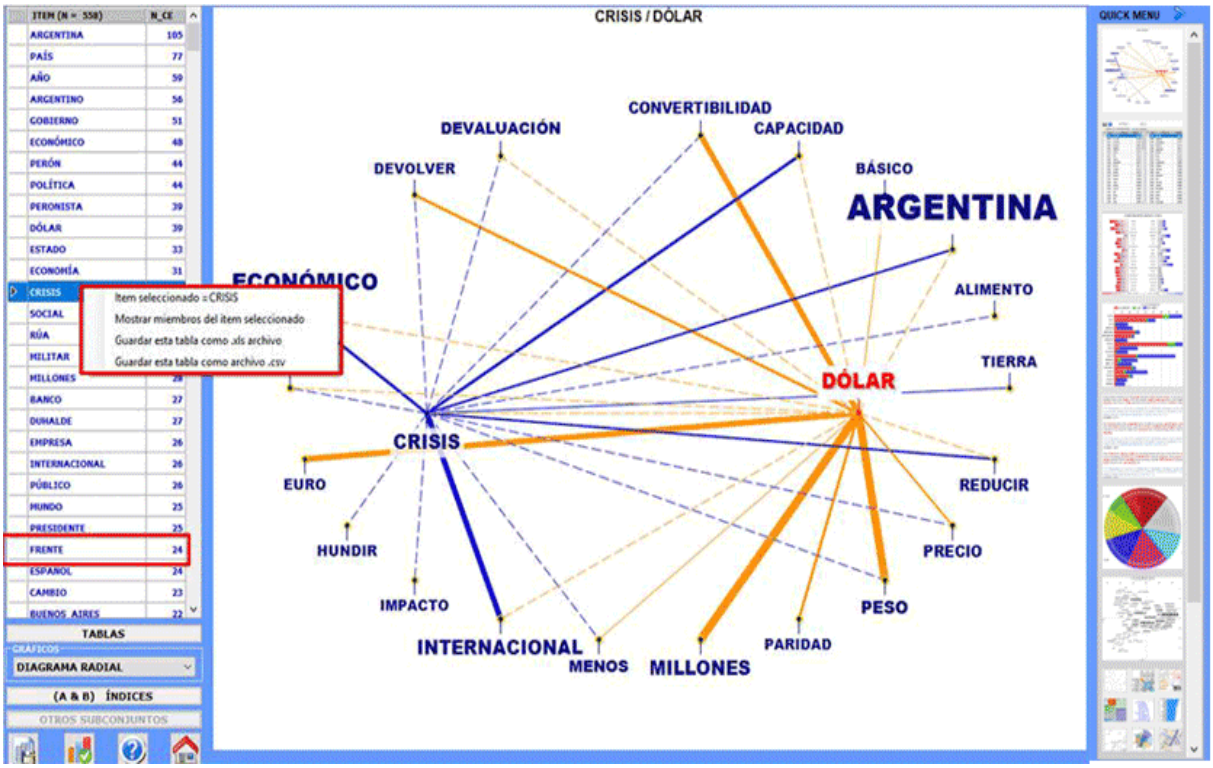
< ARGENTINA >

LEMMA_A	LEMMA_B	INDEX
Argentina	pais	0,300
capacidad	crisis	0,245
Argentina	capacidad	0,218
Argentina	crisis	0,214
Argentina	mundo	0,195
mundo	sociedad	0,173
Argentina	sociedad	0,169
crisis	pais	0,146
capacidad	sociedad	0,129
crisis	sociedad	0,105
pais	sociedad	0,099
capacidad	mundo	0,089
mundo	pais	0,068
capacidad	pais	0,051
crisis	mundo	0,037

N.B.:
- cuando un cluster análisis no se ha hecho, la tabla "Miembros" no está disponible, la tabla "Resumen" se simplifica y la tabla "Índices de asociación" se refiere solamente a las co-ocurrencias de palabras.
- al final del análisis, el diccionario de Núcleos Temáticos (es decir la lista de las etiquetas asignadas a cada cluster de las palabras) se puede exportar y, después de una cuidadosa revisión, puede ser importado por medio de la función **Personalización del Diccionario**. De esta manera el usuario podrá realizar algunos análisis de segundo orden (es decir análisis concernientes "temas" o "conceptos").

Comparaciones entre Parejas de Palabras-Clave

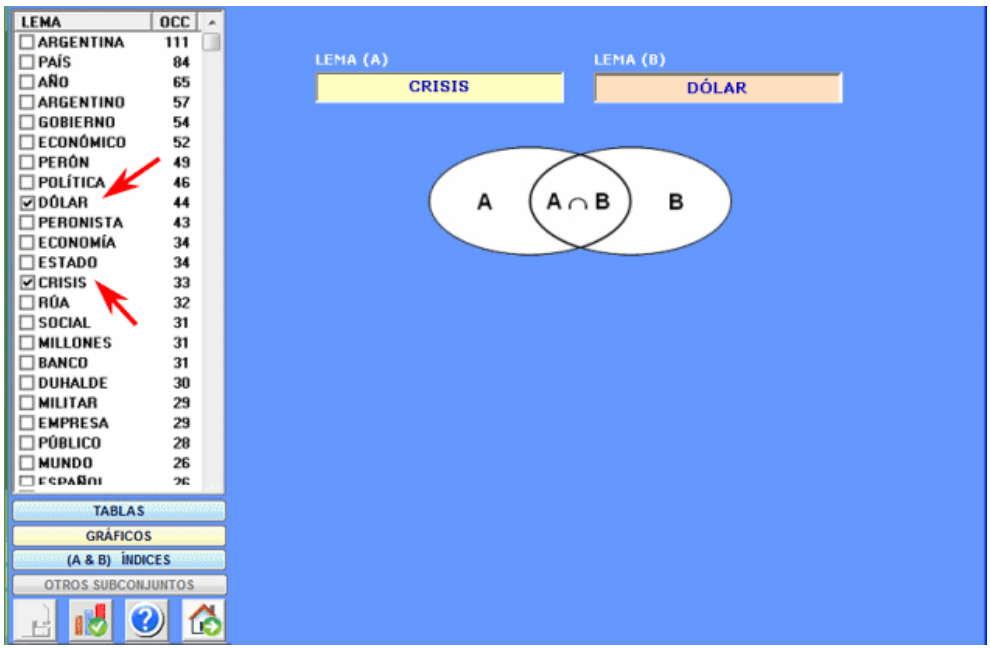
NOTA: Las imagenes contenidas en este apartado hacen referencia a una versión anterior de T-LAB, ya que el interfaz de **T-LAB 10** cambia ligeramente. Además, el uso del **botón derecho** del ratón sobre las tablas que incluyen las palabras clave permite acceder a otras opciones. También está disponible un nuevo **diagrama radial**. Este permite comprobar de forma rápida las diferencias que existen entre asociaciones de palabras. Algunas de estas nuevas características se destacan en la imagen de abajo.



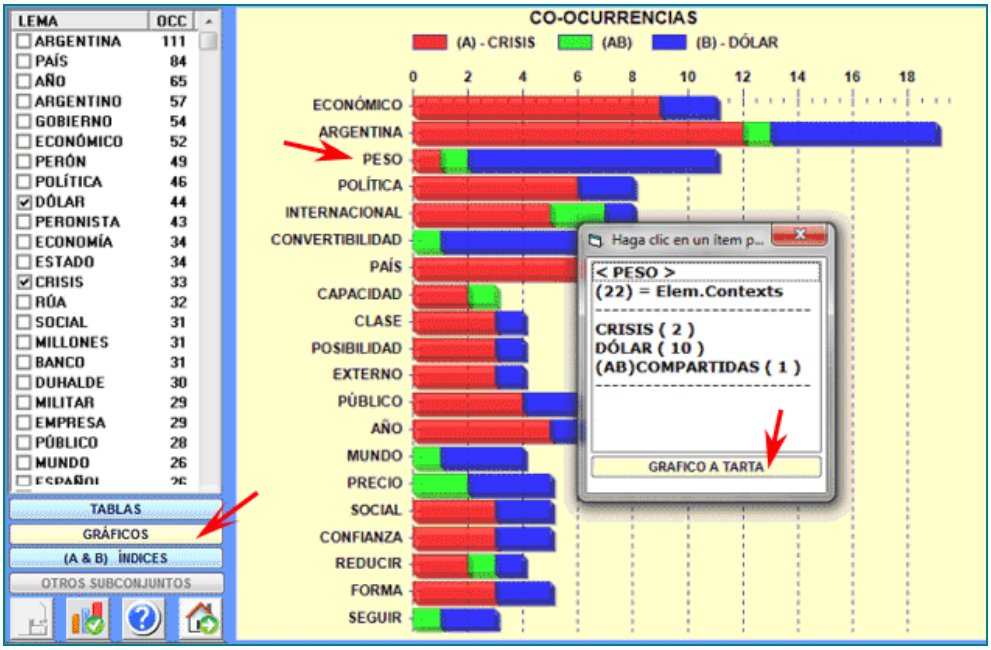
Esta herramienta de **T-LAB** nos permite comparar los conjuntos de **contextos elementales** (es decir contextos de co-ocurrencia) en los cuales los miembros de una pareja de **palabras-clave** están presentes.

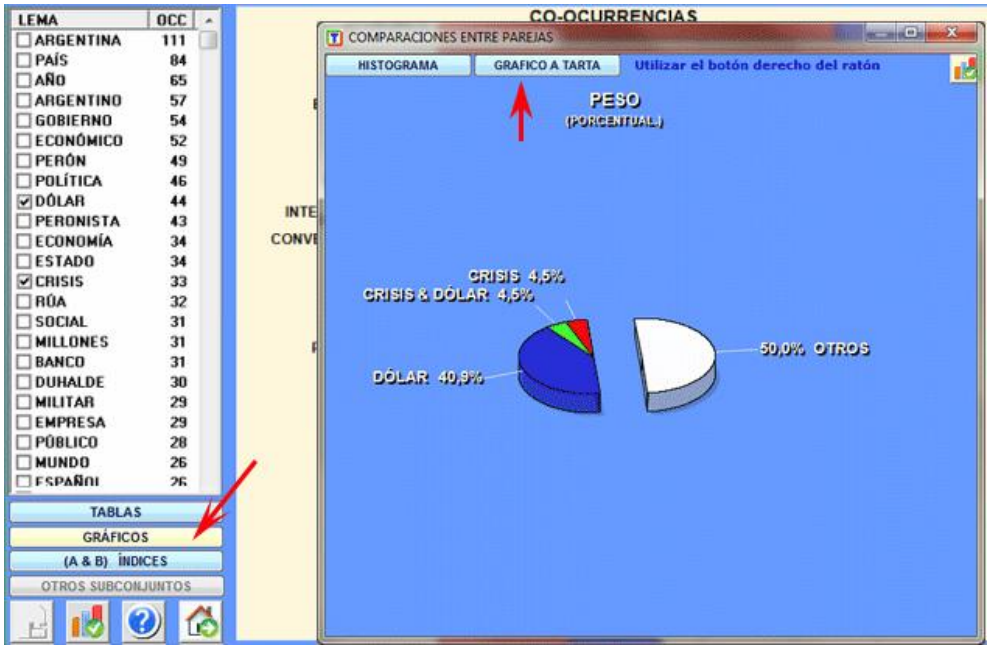
A la izquierda se muestra la tabla con los **lemas** seleccionados y los correspondientes valores de **ocurrencia** en el **corpus** o en un sus **subconjunto**.

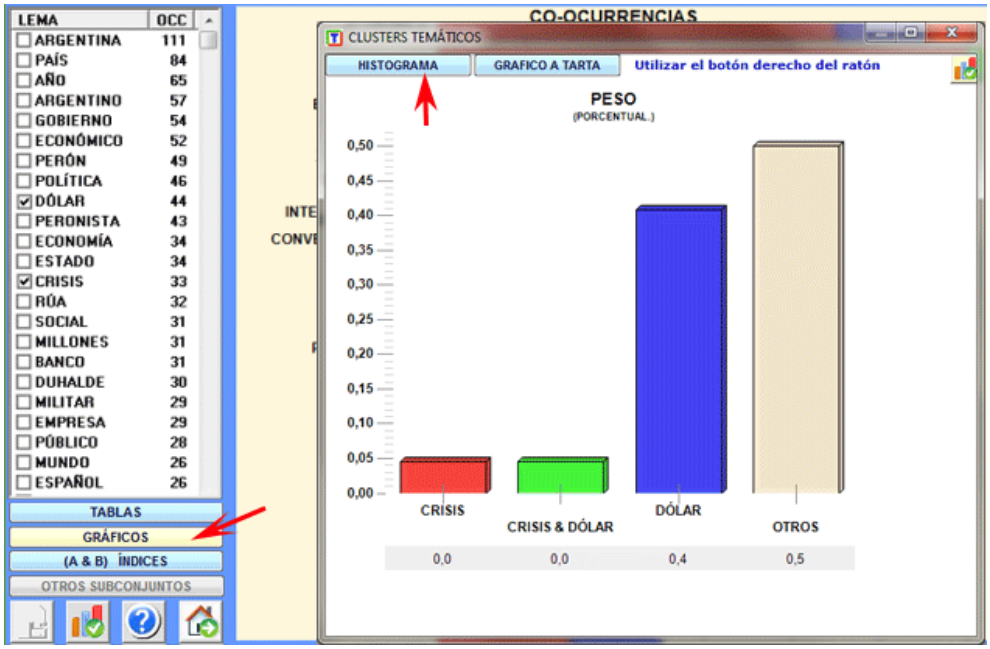
El usuario, con un simple clic, es invitado a seleccionar - uno después el otro - dos de éstos (una "pareja").



Un histograma (véase abajo) nos permite apreciar el número de contextos elementales en los cuales cada lema co-ocurre con el término "A" (color rojo), con el término "B" (color azul) y con ambos "A" y "B" (color verde). Con un simple clic sobre cada etiqueta de la carta es también posible comprobar sus valores correspondientes de co-ocurrencia y obtener un gráfico a tarta y un histograma.







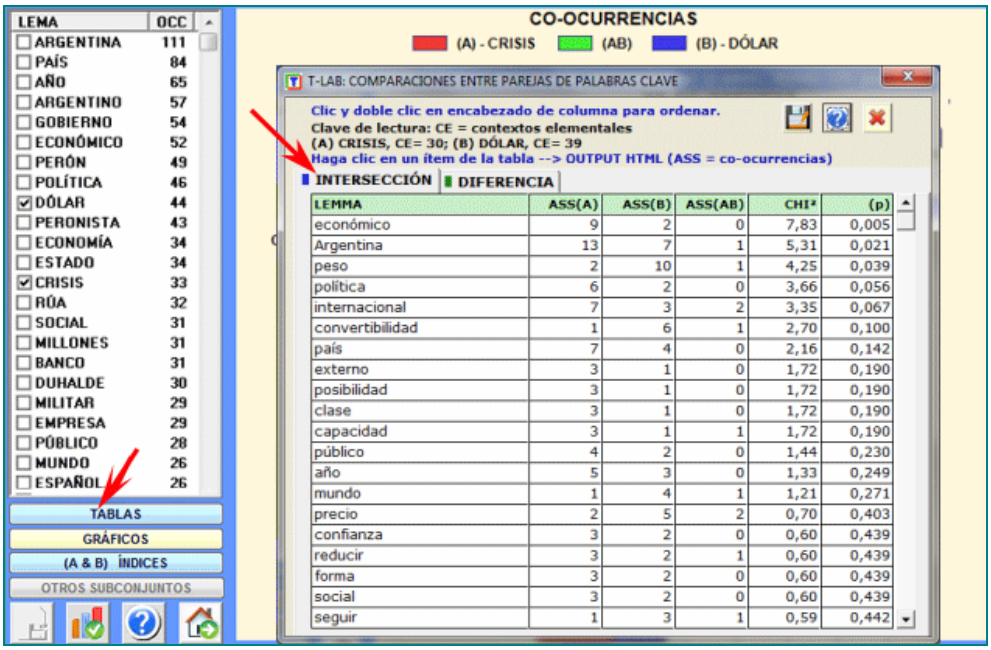
Las comparaciones propuestas por **T-LAB** conciernen las co-ocurrencias entre los elementos de la "pareja" y cada una de las palabras contenidas en la tabla (véase abajo).

Sean:

A = conjunto de los **contextos elementales** (TOT. C.E. = 30) en los que se halla la primera palabra de la pareja (ej. "crisis");
B = conjunto de los contextos elementales (TOT. C.E. = 39) en los que se halla la segunda palabra de la pareja (ej. "dólar").

El primer tipo di comparación concierne las **asociaciones compartidas** (véase el botón **intersección**) y considera todas las palabras que se hallan tanto en "A" como en "B".

En la tabla cada fila muestra los valores que corresponden a las comparaciones de cada lema.



Las llaves de lectura son las siguientes:

- **ASS (A)** = número de contextos elementales en los que cada lema está asociado (co-ocurrencias) con (A);
- **ASS (B)** = número de contextos elementales en los que cada lema está asociado con (B);
- **ASS (AB)** = número de contextos elementales en los que cada lema está asociado con (A) y con (B);
- **CHI2** = valores del CHI cuadrado;
- **(p)** = probabilidad asociada a cada valor del chi-cuadrado (def=1).

En este caso, para cada palabra clave (ej. "económico") **T-LAB** construye una tabla como la siguiente y le aplica el test del **CHI cuadrado**:

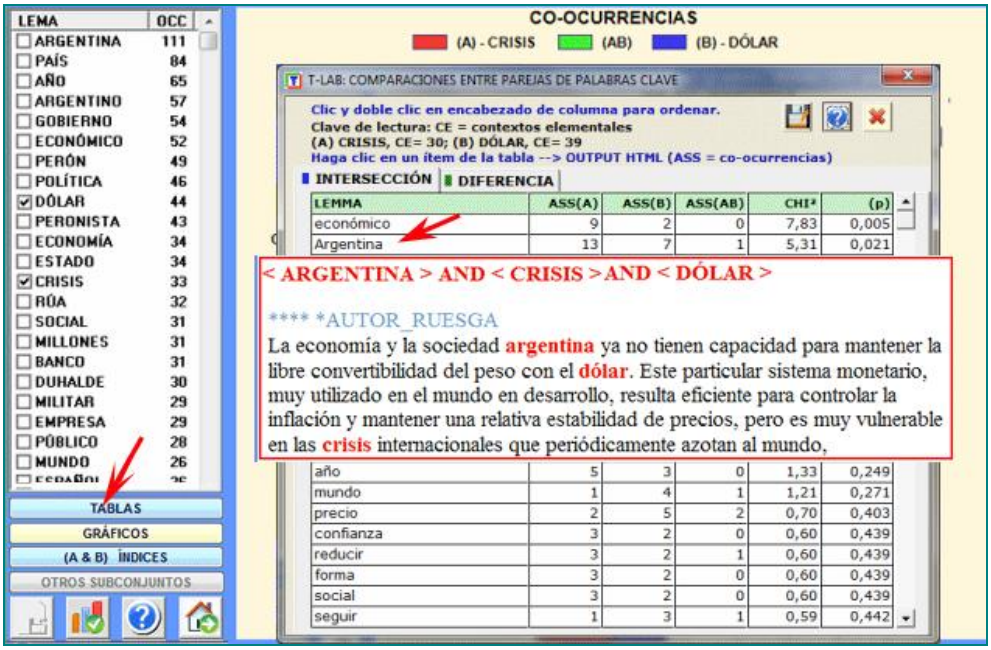
	ASSOC.	NON ASSOC.	TOT.
A	9	21	30
B	2	37	39
	11	58	69

En la línea (A) se indica el número de contextos elementales en los que "económico" está presente (9) o ausente (21) en el conjunto de contextos (30) propios de la primera palabra de la pareja ("crisis").

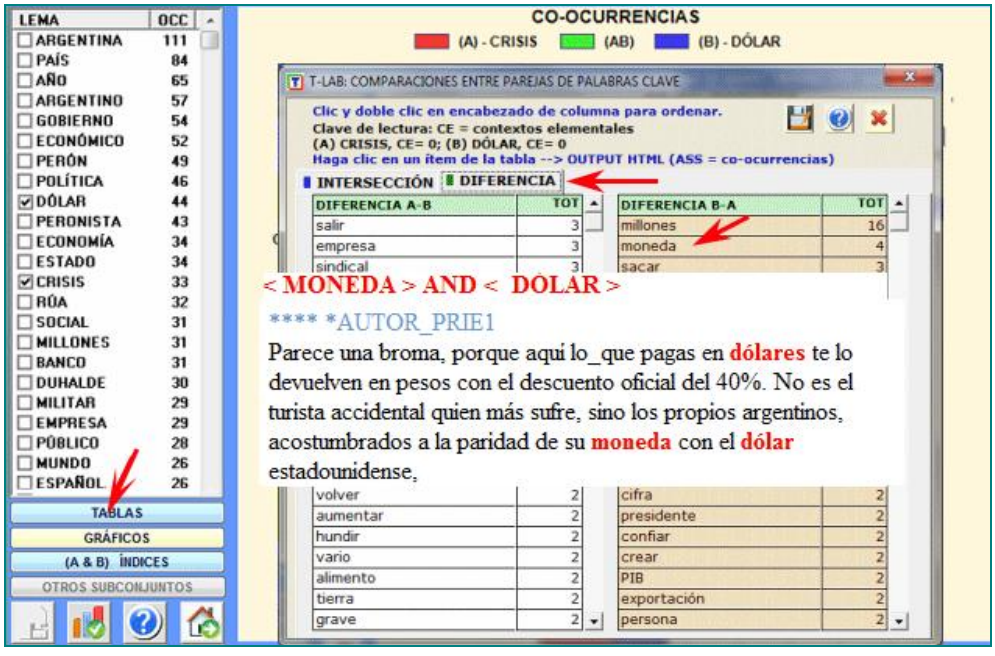
En la línea (B) se indica el número de contextos elementales en los que "económico" está presente (2) o ausente (37) en el conjunto de contextos (39) propios de la segunda palabra de la pareja ("dólar").

NOTA En este caso, el valor del CHI cuadrado es igual a 7.828.

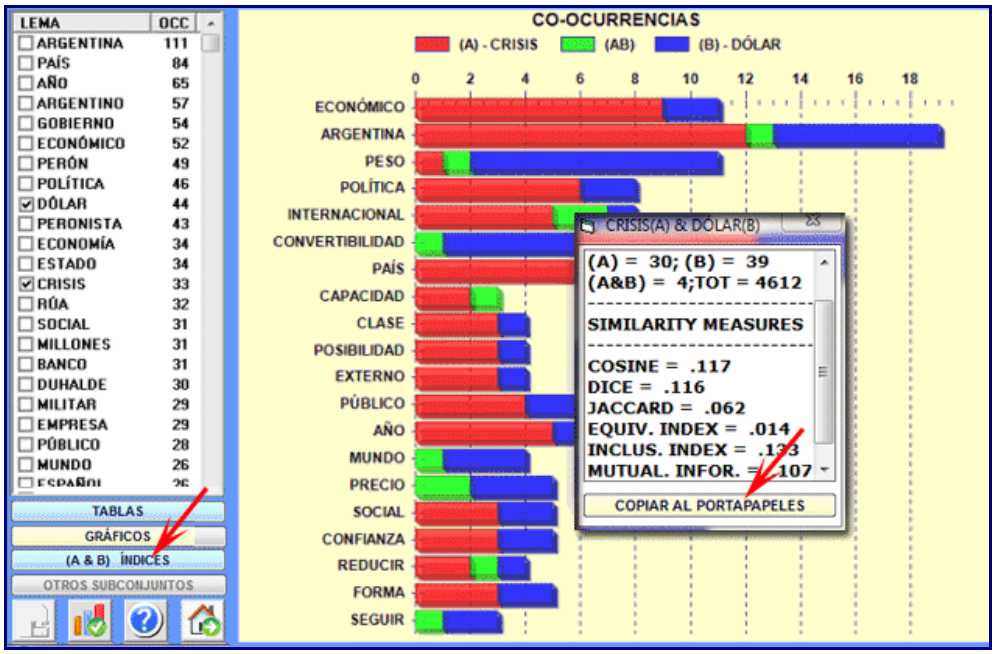
Por otra parte un doble clic sobre cada ítem de la tabla nos permite ahorrar un archivo HTML con el número de contextos elementales en la columna correspondiente.



El segundo tipo de comparación concierne las diferencias entre A y B (A - B y B - A).
En este caso T-LAB propone dos tablas con las palabras clave que, de manera exclusiva, están asociadas al primero "o" al segundo término de la pareja. En ambas tablas, la columna "TOT" indica la cantidad de contextos elementales en los que cada lema está asociado sólo con uno de los dos términos de la pareja.



Finalmente, cliqueando en el botón correspondiente (véase imagen siguiente), se pueden verificar y exportar todos los índices de semejanza que caracterizan la pareja de palabras en examen.



Análisis de Secuencias y Análisis de Redes

Esta herramienta de **T-LAB** tiene en cuenta las posiciones de las diferentes unidades lexicales dentro de las frases, permitiendonos así analizar y representar cualquier texto como si fuera una **red** de relaciones.

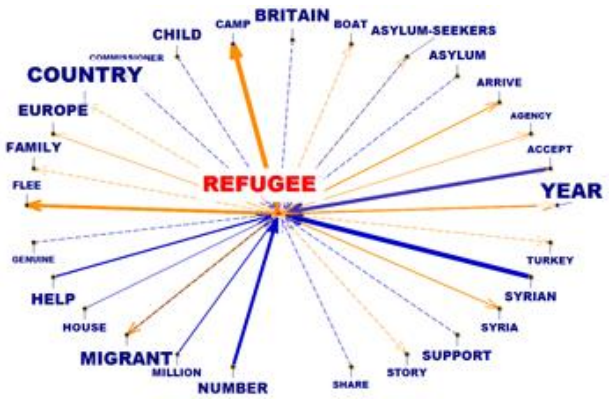
Las opciones a dispición del usuario permiten implementar analisis de Co-Word, analisis tematicos y desambiguaciones.

De hecho, una vez construidas las dos matrices que incluyen todas las parejas de predecesores y sucesores, **T-LAB** calcula las **probabilidades de transición** (cadenas de markov) y proporciona diferentes output relacionados con las palabras objeto de estudio.
.

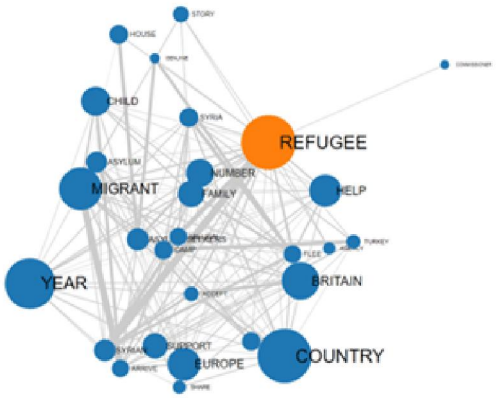
Además, es posible realizar un **analisis de clústeres**. Como consecuencia, se podrán explorar las relaciones semanticas que existen entre palabras tanto dentro de la red entera como dentro de de los "clústeres temáticos" (N.B.: en este caso, el algoritmo usado para la clusterización coincide con el 'Louvain method' desarrollado por Blondel V.D., Guillame J.-L , Lambiotte R., Lefebre E., 2008).

Por tanto, una vez implementado este tipo de analisis, el usuario podrá explorar las relaciones que existen entre nodos de la red (esto es, las palabras clave) a diferentes niveles: a) dentro de las relaciones del tipo uno-a-uno; b) dentro de un "ego network"; c) dentro de las comunidades a las que pertenecen; d) dentro de la red formada por el texto analizado.

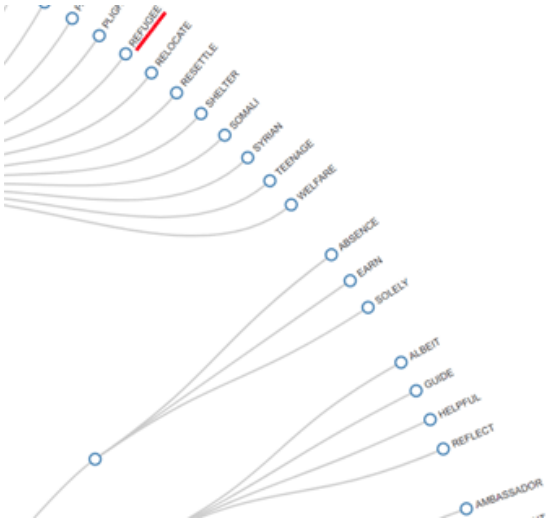
RELACIONES DEL TIPO UNO-AD-UNO



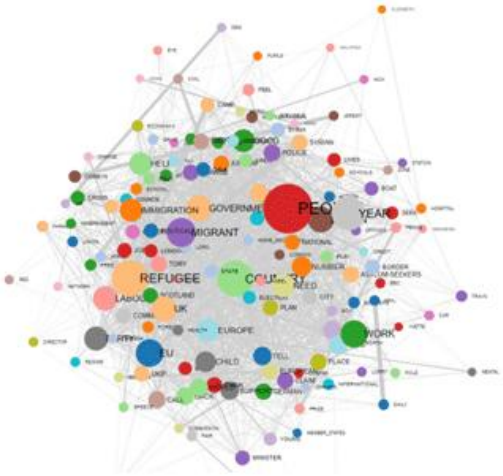
EGO-NETWORK



COMUNIDADES



RED ENTERA



Las informaciones necesarias para utilizar las diferente opciones de analisis están organizadas en tres secciones

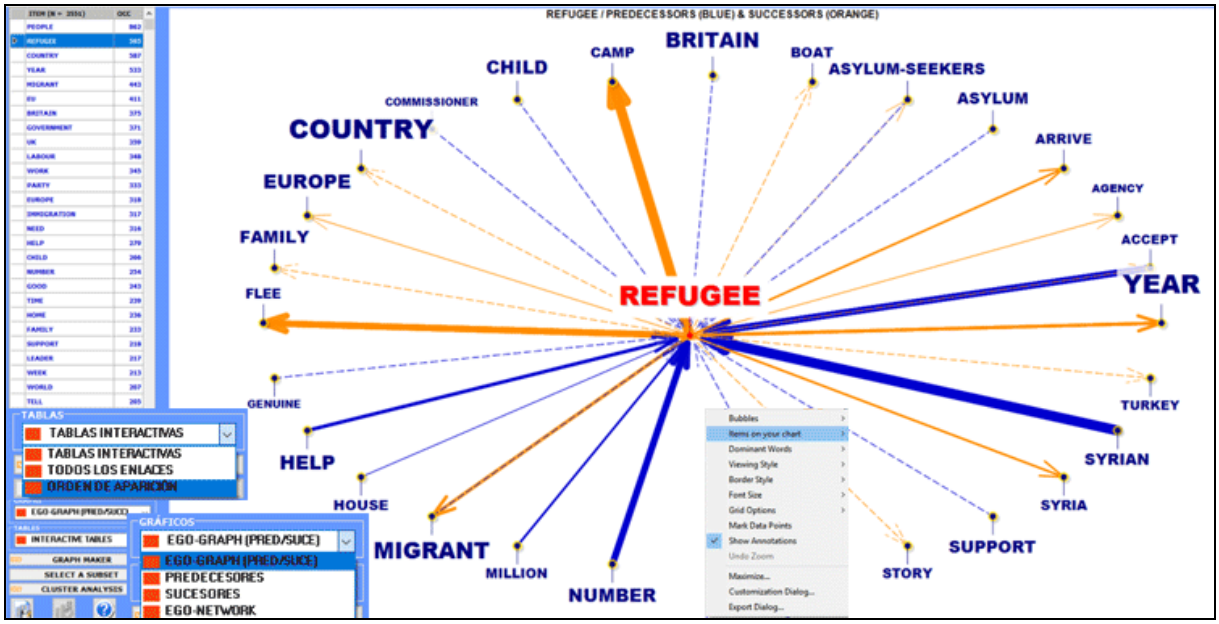
A - Explorar las conexiones del tipo uno-a-uno y las "ego networks";
B - Explorar las "comunidades" (clústeres temáticos) y la red entera;
C - Algunos detalles técnicos.

N.B.: Por razones de carácter editorial, esta página incluye ejemplos de analisis basados en un corpus escrito en lengua inglesa.

A - EXPLORAR LAS CONEXIONES DEL TIPO UNO-A-UNO Y LAS "EGO NETWORKS"

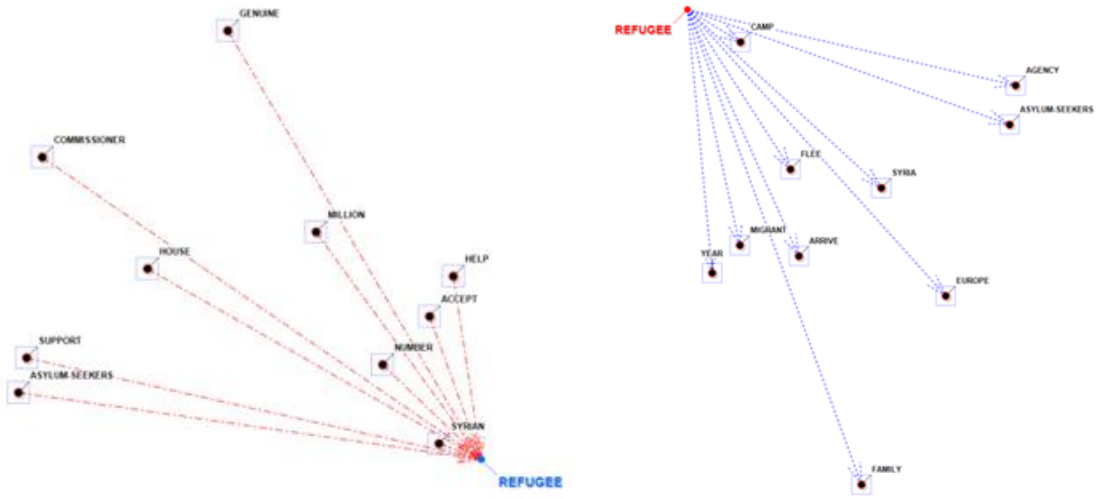
Una vez acabado el análisis automático, se dispondrá de diferentes tablas y gráficos que permitirán explorar las relaciones y los datos asociados a las palabras clave seleccionadas (N.B.: Para obtenerlos basta con hacer clic sobre uno de los ítems incluidos en las tablas o en cualquiera de los puntos que componen los gráficos).

Usando el botón derecho del ratón, será posible personalizar cualquier tipo de gráfico y exportarlo a diferentes formatos.

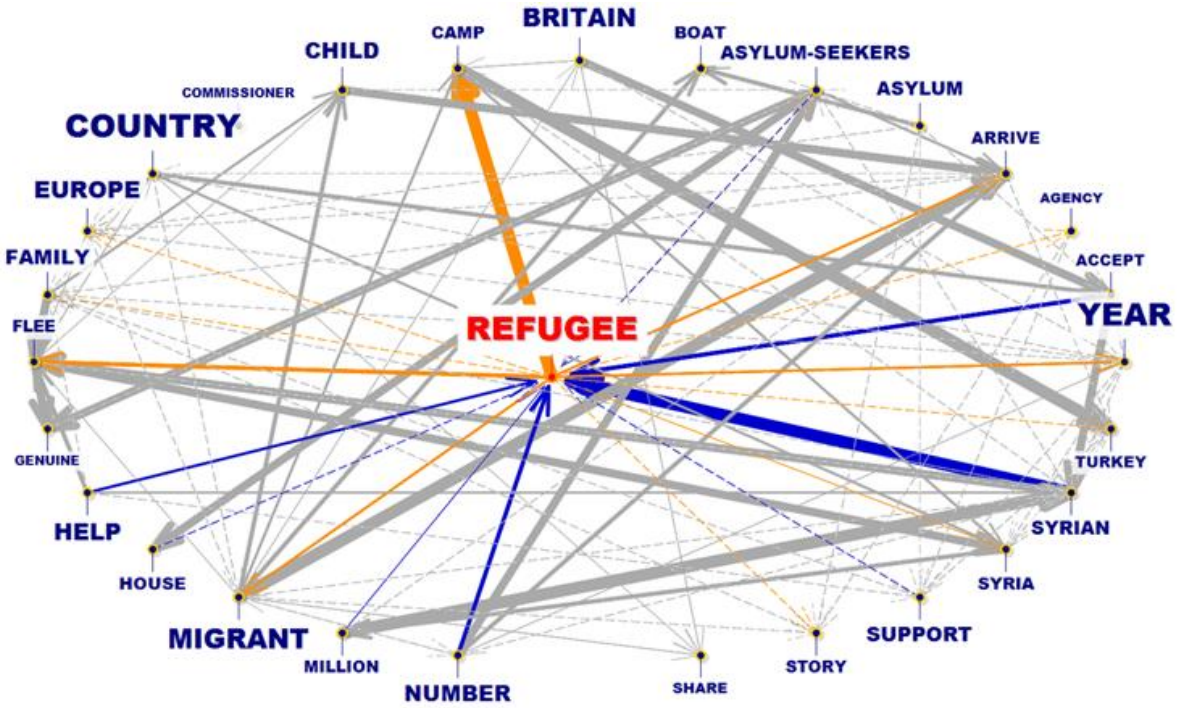


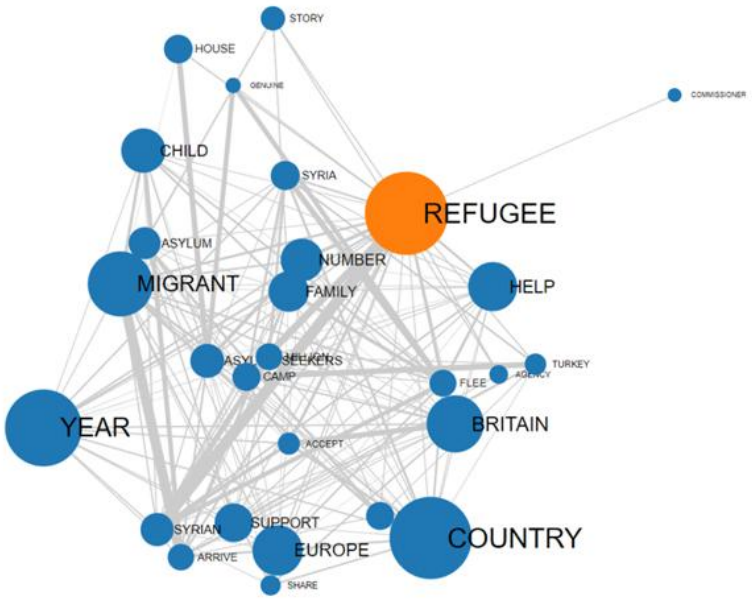
En dos de los gráficos los elementos más cercanos a los seleccionados son aquellos que mayor probabilidad tienen de estar delante (predecesores) o detrás (sucesores) de los mismos.

PREDECESORES **SUCESORES**



En los demás casos, la cercanía entre palabras-clave viene representada graficamente mediante el grosor de las flechas que las conectan.



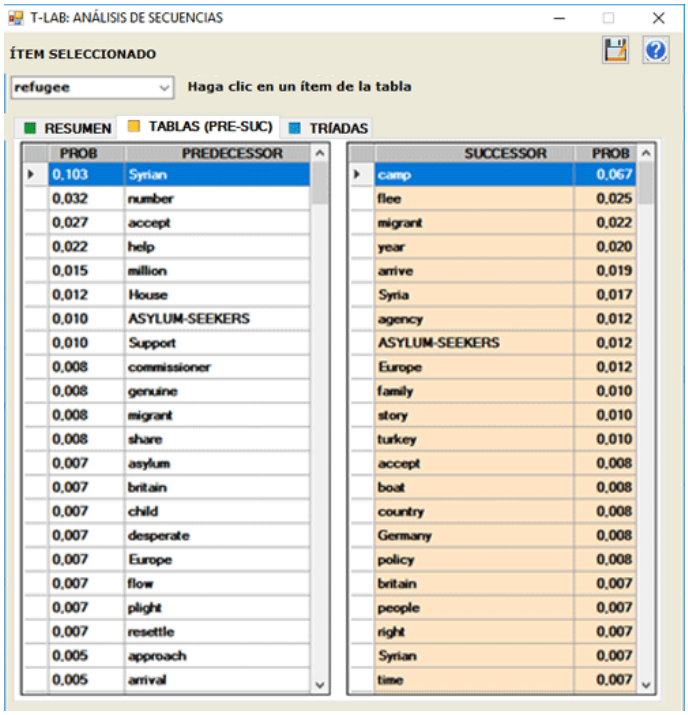


Es posible comprobar todos los datos utilizando las diferentes tipologías de **tablas**.

Más en detalle:

Las **TABLAS INTERACTIVAS** muestran los listados de predecesores y sucesores vinculados a las palabras clave seleccionadas.

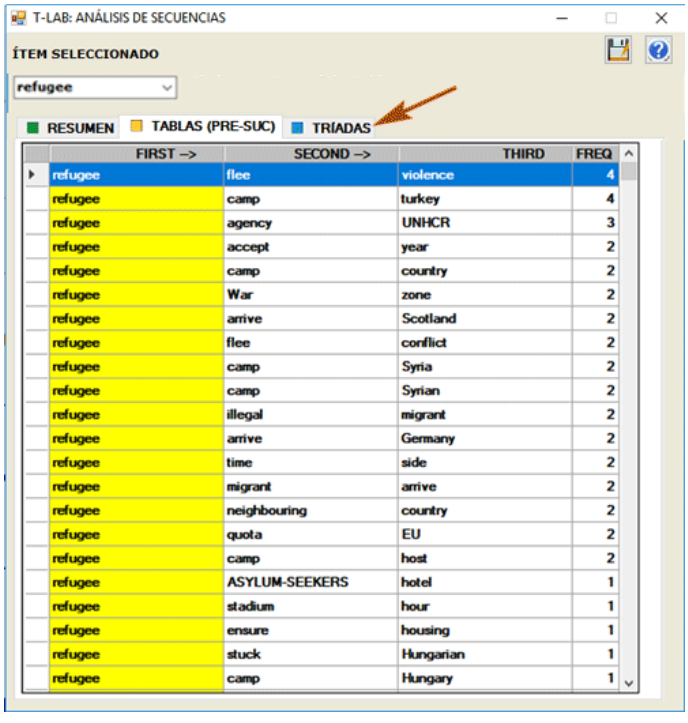
La lista está en una orden descendente según los valores de probabilidad ("PROB"). Por ejemplo, en la tabla siguiente, la probabilidad de que "camp" siga "refugee" es igual a 0.067, que es 6.7%.



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

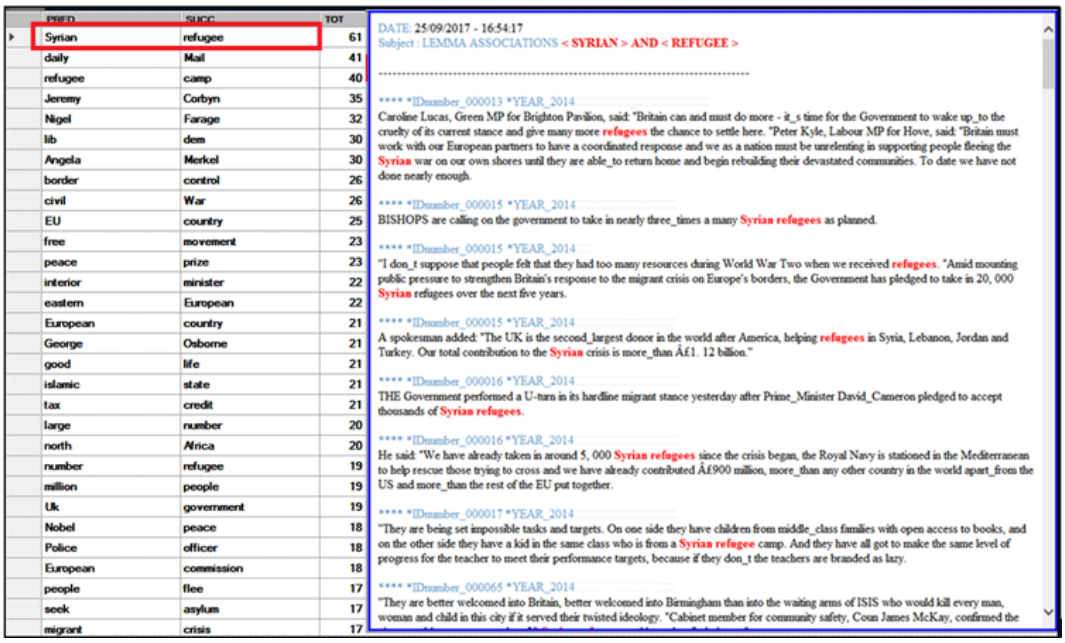
La opción **TRÍADAS** nos permite visualizar algunas tablas con secuencias de tres elementos en las cuales la palabra seleccionada está en la primera, en la segunda o en la tercera posición. Para cada tríada **T-LAB** muestra los correspondientes valores de ocurrencia.
(N.B.: Dentro de las tríadas las palabras vacías no son incluidas).



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

La tabla **TODOS LOS ENLACES** (véase abajo) es particularmente util para desambiguar los significados de las palabras, y contiene todas las parejas de predecesores y sucesores junto con las ocurrencias a ellas correspondientes.
Cliqueando en una de la líneas de esta tabla, será posible visualizar, en el lado derecho de la misma y en formato HTLM, todos los segmentos de texto (esto es, los contextos elementales) en los cuales aparecen conjuntamente dos elementos de una misma pareja (esto es, las co-ocurrencias).

.



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.

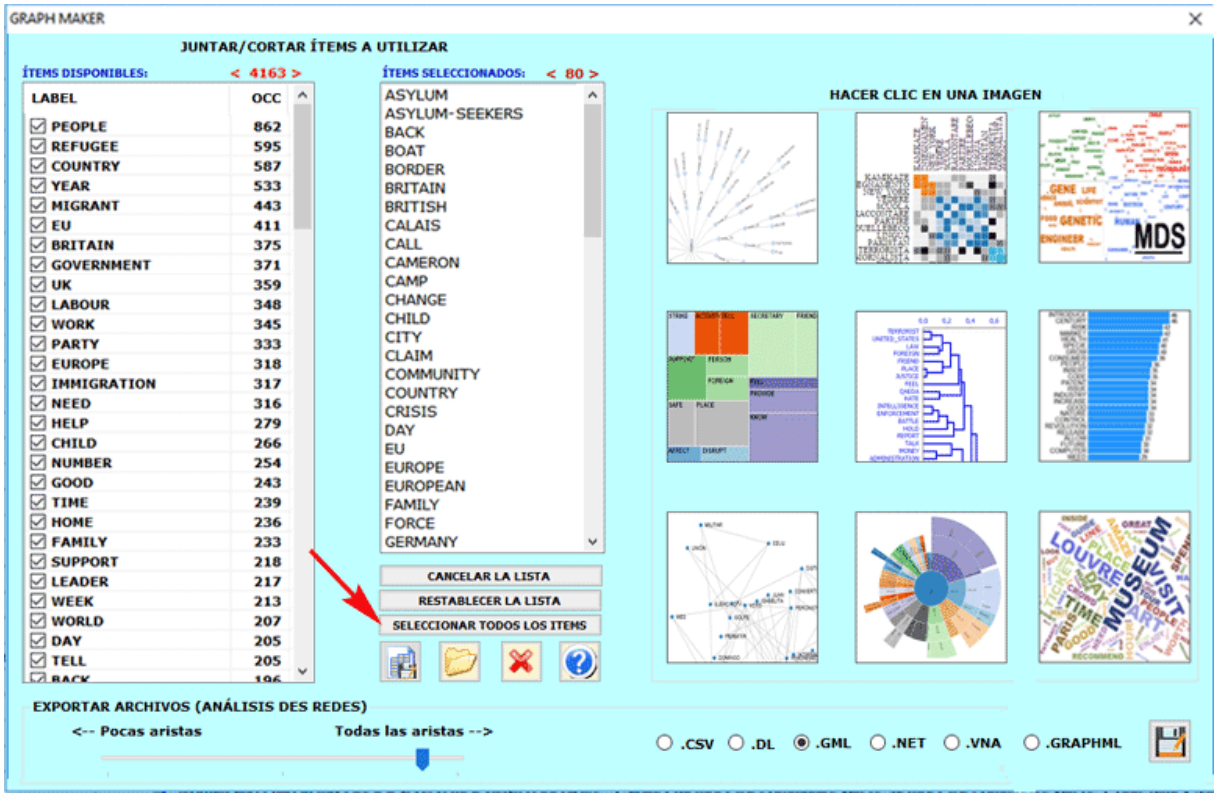
**** *IDnumber_000016 *YEAR_2014
He said: "We have already taken in around 5, 000 Syrian refugees since the crisis began, the Royal Navy is stationed in the Mediterranean to help rescue those trying to cross and we have already contributed Â£900 million, more_than any other country in the world apart_from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

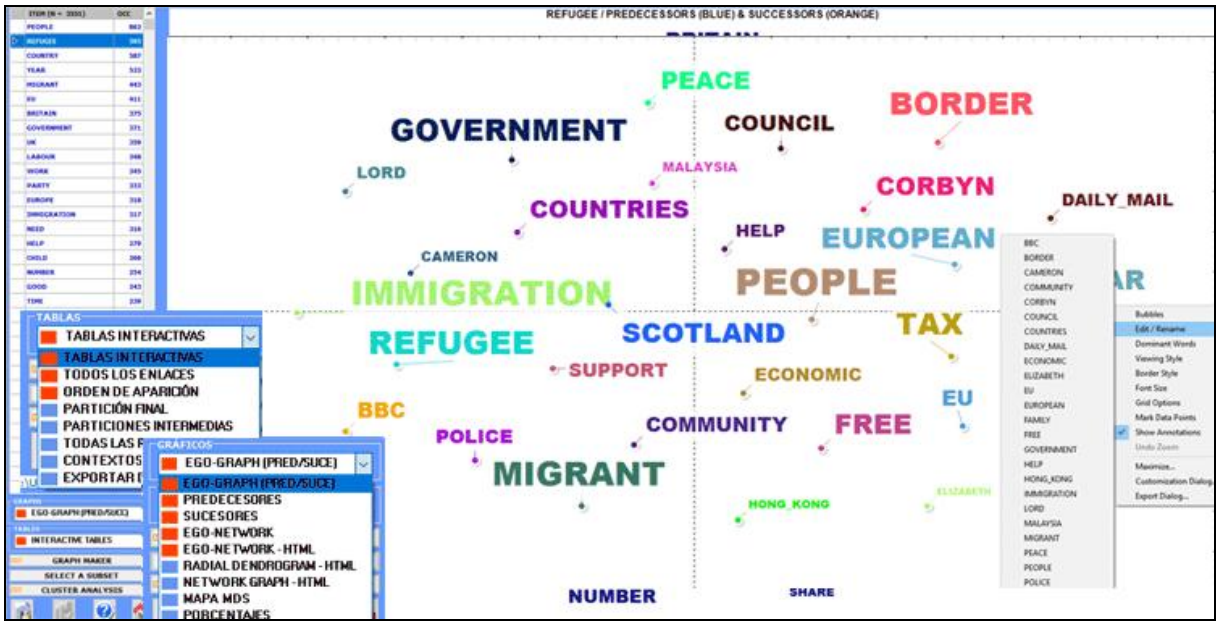
La tabla **RANGO DE APARICIÓN** incluye la frecuencia y el orden medio de aparición (o evocación) de cada palabra dentro de un segmento de texto. Sólo es posible ver esta tabla cuando el corpus está compuesto por textos cortos, como por ejemplo respuestas a preguntas abiertas.

Cliqueando en la opción **GRAPH MAKER** el usuario podrá crear, en todo momento y a partir de los listados personalizados de palabras clave, diferentes tipos de graficos (véase abajo). Los usuarios avanzados que estén interesados en exportar archivos a formatos diferentes (p.e. .dl .gml .vna .graphml) junto con los datos relativos a todos los enlaces, pueden hacer click en el botón 'SELECCIONAR TODOS LOS ITEMS.



B - EXPLORAR LAS COMUNIDADES (CLÚSTERES TEMÁTICOS) Y LA RED ENTERA

Una vez realizado un análisis de clústeres, se vuelven disponibles **nuevos gráficos y tablas**. Todos ellos están indicados por pequeños rectángulos azules (véase abajo).

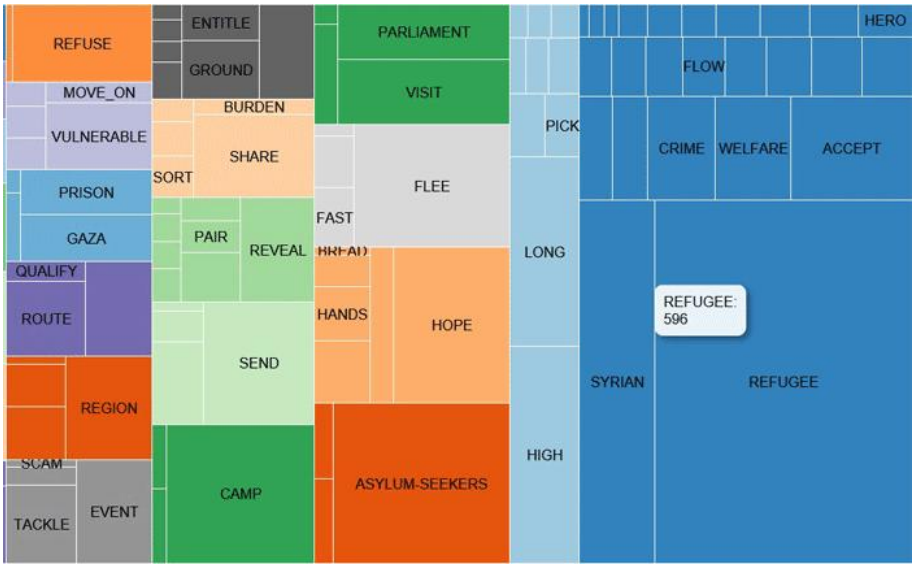


La primera tabla resume las mejores características (palabras clave) de la PARTICIÓN FINAL obtenida a partir del algoritmo de clusterización.

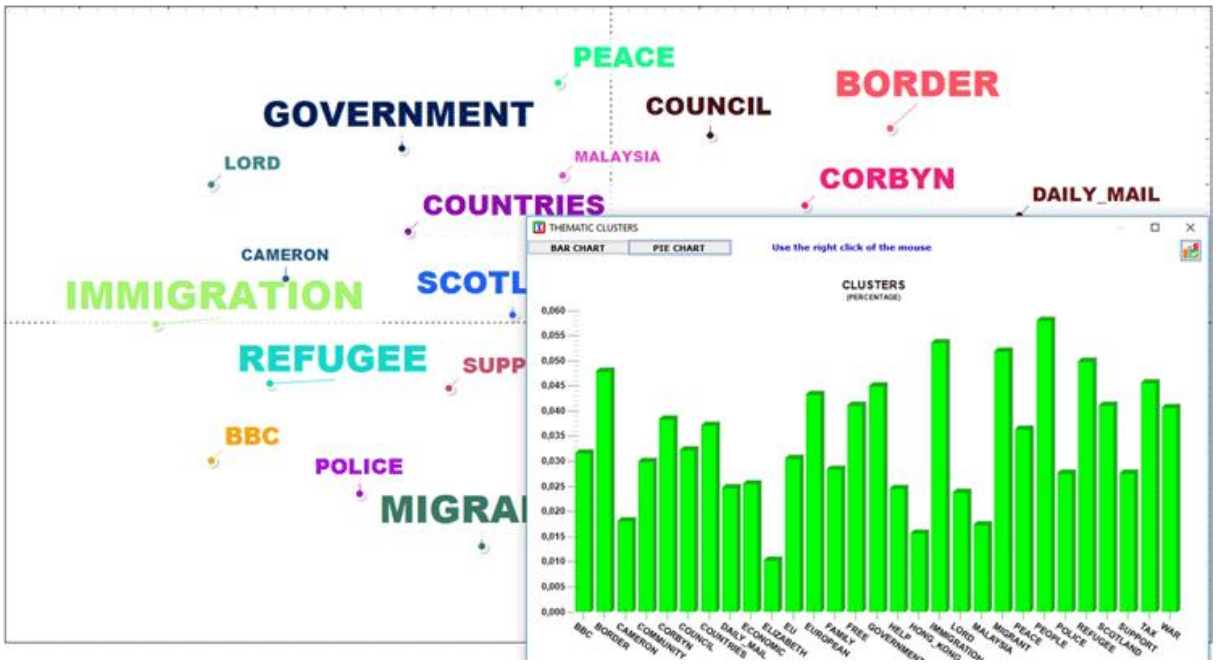
En dicha tabla se encuentran ordenadas en base a su valor **TF-IDF** (véase abajo) las características de cada clúster

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

Haciendo clic en cualquier palabra de la tabla anterior (así como de la tabla **TODAS LAS PARTICIONES**), un TreeMap nos permite verificar las comunidades a las que pertenece (ver abajo).

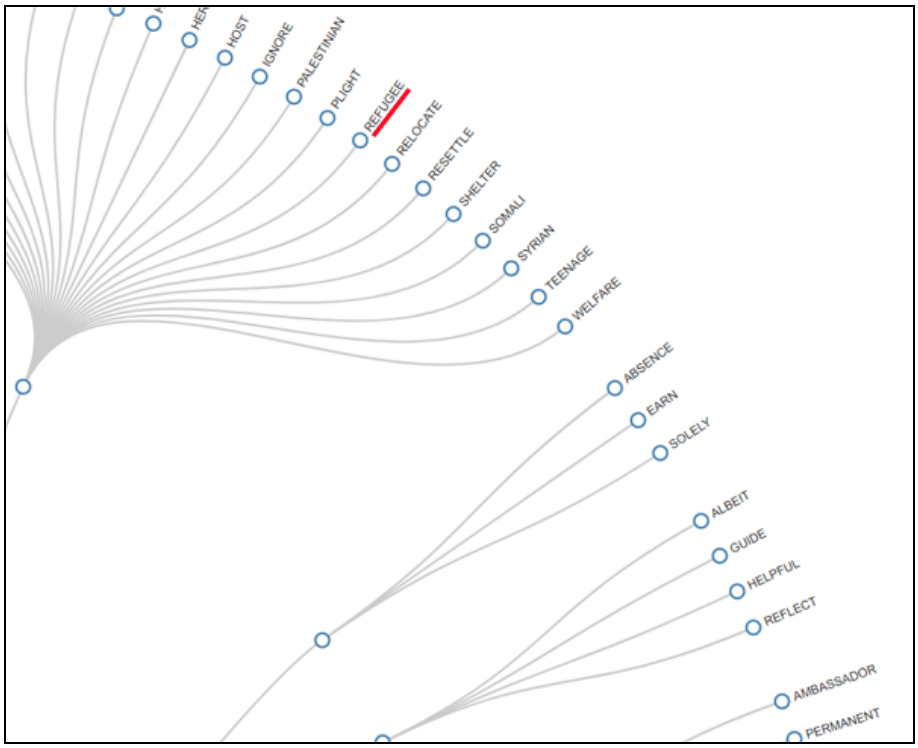


El **MAPA MDS** y el gráfico **PORCENTAJES** (véase abajo) permiten comprobar el 'peso' de cada clúster, así como las relaciones entre diferentes clústeres dentro de la partición final encontrada (véase abajo).

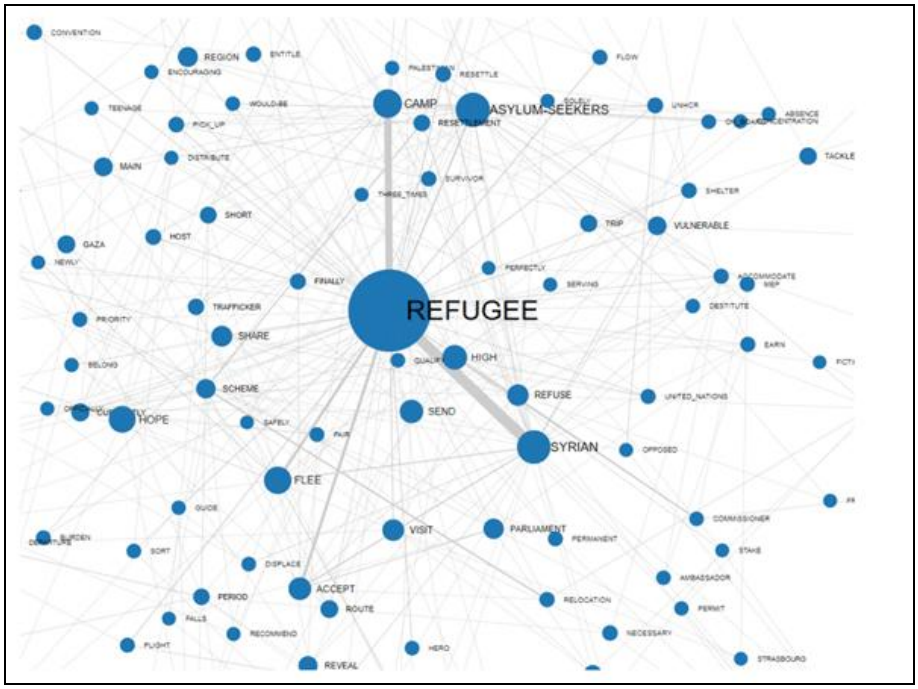


En función del número de palabras clave, será posible explorar las relaciones entre ellas, utilizando dos gráficos en formato HTLM. Todo ello, tanto dentro de la entera red como dentro de los clústeres a los que pertenecen las palabras clave (véase abajo).

DENDROGRAMA RADIAL



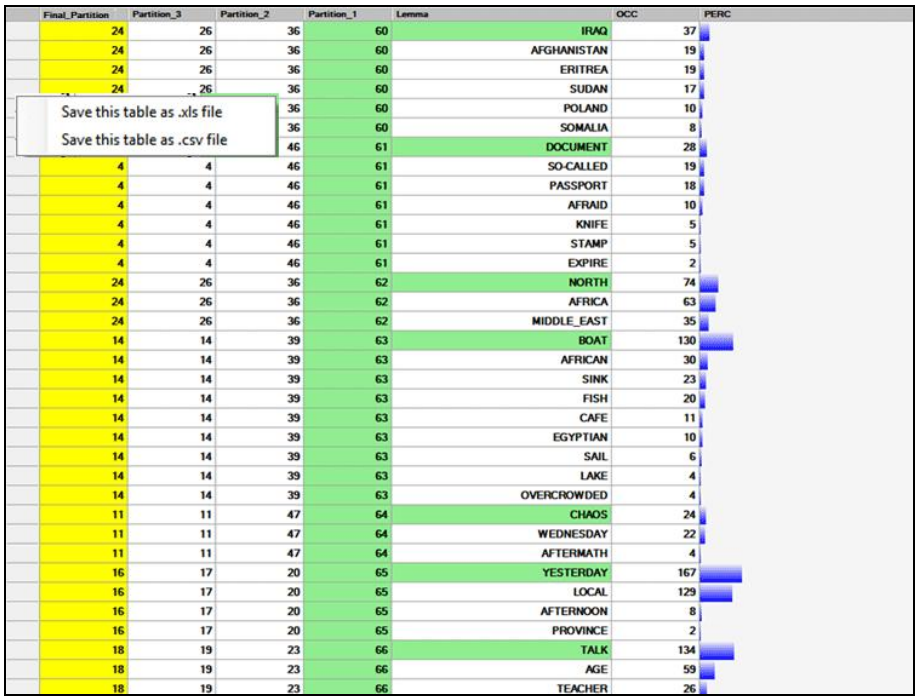
NETWORK (FORCE-DIRECTED GRAPH)



Tres nuevas tablas proporcionarán ulterior información obtenida a partir de los análisis de clústeres.

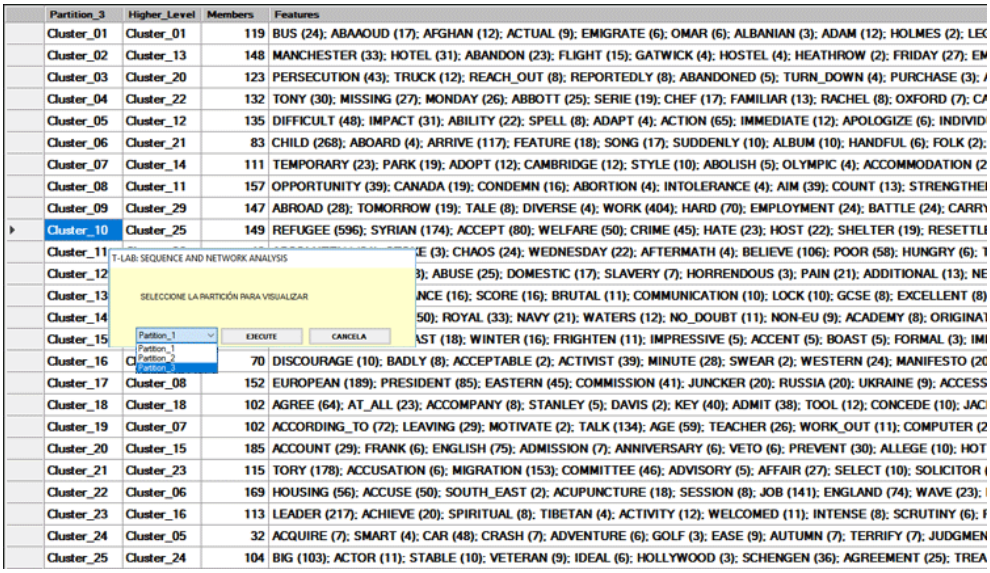
En concreto:

La tabla TODAS LAS PARTICIONES permitirá comprobar como las palabras claves estén agrupadas a partir de cada una de las particiones del analisis de clústeres (véase la tabla abajo. Los números incluidos en las columnas de las particiones hacen referencia a los diferentes clústeres).
N.B.: Por defecto, esta tabla viene ordenada en base a la primera partición, que presenta el numero más alto de clústeres. Cada movimiento de un clúster pequeño hacia otro viene puesto de releve marcando en color verde la primera palabra que lo compone.



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	

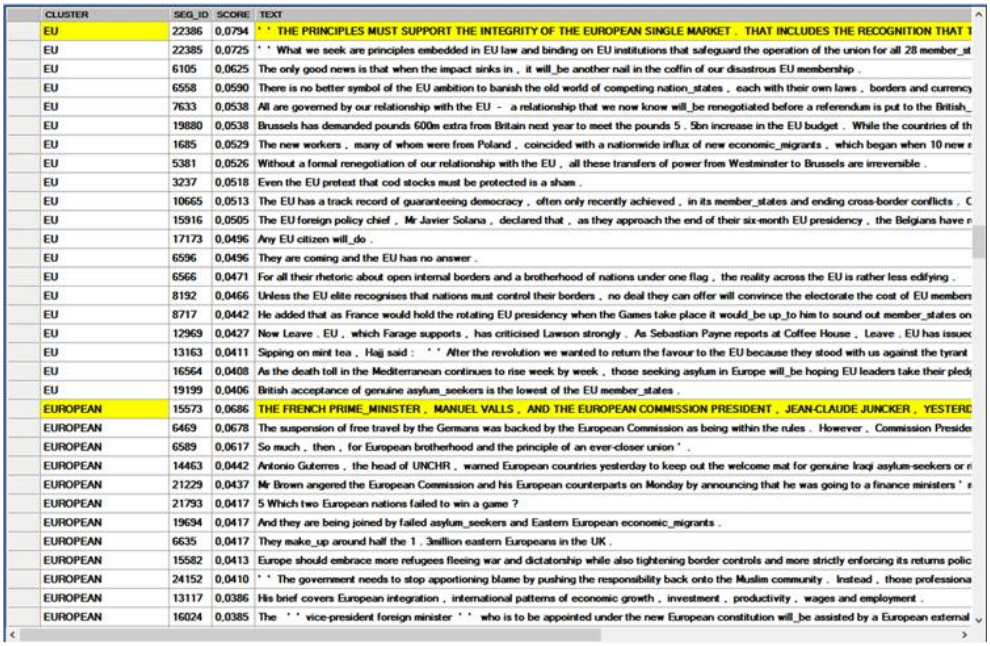
La tabla **PARTICIONES INTERMEDIAS** permitirá explorar de qué manera hayan sido agrupadas las palabras claves dentro de cada una de las particiones seleccionadas. Paso a paso, éstas vendrán ordenadas de forma descendente en base a los valores de sus co-ocurrencias (véase abajo).



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); CA
Cluster_05	Cluster_12	135	DIFFICULT (48); IMPACT (31); ABILITY (22); SPELL (8); ADAPT (4); ACTION (65); IMMEDIATE (12); APOLOGIZE (6); INDIVID
Cluster_06	Cluster_21	83	CHILD (268); ABOARD (4); ARRIVE (117); FEATURE (18); SONG (17); SUDDENLY (10); ALBUM (10); HANDFUL (6); FOLK (2);
Cluster_07	Cluster_14	111	TEMPORARY (23); PARK (19); ADOPT (12); CAMBRIDGE (12); STYLE (10); ABOLISH (5); OLYMPIC (4); ACCOMMODATION (2
Cluster_08	Cluster_11	157	OPPORTUNITY (39); CANADA (19); CONDEMN (16); ABORTION (4); INTOLERANCE (4); AIM (39); COUNT (13); STRENGTHE
Cluster_09	Cluster_29	147	ABROAD (28); TOMORROW (19); TALE (8); DIVERSE (4); WORK (404); HARD (70); EMPLOYMENT (24); BATTLE (24); CARR
Cluster_10	Cluster_25	149	REFUGEE (596); SYRIAN (174); ACCEPT (80); WELFARE (50); CRIME (45); HATE (23); HOST (22); SHELTER (19); RESETTLE
Cluster_11			E (3); CHAOS (24); WEDNESDAY (22); AFTERMATH (4); BELIEVE (106); POOR (58); HUNGRY (6);
Cluster_12			3); ABUSE (25); DOMESTIC (17); SLAVERY (7); HORRENDOUS (3); PAIN (21); ADDITIONAL (13); NE
Cluster_13			NCE (16); SCORE (16); BRUTAL (11); COMMUNICATION (10); LOCK (10); GCSE (8); EXCELLENT (8)
Cluster_14			50); ROYAL (33); NAVY (21); WATERS (12); NO_DOUBT (11); NON-EU (9); ACADEMY (8); ORIGINA
Cluster_15			AST (18); WINTER (16); FRIGHTEN (11); IMPRESSIVE (5); ACCENT (5); BOAST (5); FORMAL (3); IM
Cluster_16	Cl	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

La tabla CONTEXTOS TÍPICOS permite explorar los segmentos de texto que mayor puntuación de asociación presentan en relación con los clústeres de la partición final. En esta tabla se utiliza el índice de coseno para medir la semejanza entre el vector de las características de cada clúster y el vector que contiene los segmentos de texto.

N.B. Viene marcado en color amarillo el segmento de texto más significativo de cada clúster.



CLUSTER	SEG_ID	SCORE	TEXT
EU	22386	0,0794	' ' THE PRINCIPLES MUST SUPPORT THE INTEGRITY OF THE EUROPEAN SINGLE MARKET . THAT INCLUDES THE RECOGNITION THAT T
EU	22385	0,0725	' ' What we seek are principles embedded in EU law and binding on EU institutions that safeguard the operation of the union for all 28 member_st
EU	6105	0,0625	The only good news is that when the impact sinks in , it will_be another nail in the coffin of our disastrous EU membership .
EU	6558	0,0590	There is no better symbol of the EU ambition to banish the old world of competing nation_states , each with their own laws , borders and currency
EU	7633	0,0538	All are governed by our relationship with the EU - a relationship that we now know will_be renegotiated before a referendum is put to the British_
EU	19880	0,0538	Brussels has demanded pounds 600m extra from Britain next year to meet the pounds 5 . 5bn increase in the EU budget . While the countries of th
EU	1685	0,0529	The new workers , many of whom were from Poland , coincided with a nationwide influx of new economic_migrants , which began when 10 new m
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 m
EU	17173	0,0496	Any EU citizen will_do .
EU	6596	0,0496	They are coming and the EU has no answer .
EU	6566	0,0471	For all their rhetoric about open internal borders and a brotherhood of nations under one flag , the reality across the EU is rather less edifying .
EU	8192	0,0466	Unless the EU elite recognises that nations must control their borders , no deal they can offer will convince the electorate the cost of EU members
EU	8717	0,0442	He added that as France would hold the rotating EU presidency when the Games take place it would_be up_to him to sound out member_states on
EU	12969	0,0427	Now Leave . EU , which Farage supports , has criticised Lawson strongly . As Sebastian Payne reports at Coffee House , Leave . EU has 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 pledg
EU	19199	0,0406	British acceptance of genuine asylum_seekers is the lowest of the EU member_states .
EUROPEAN	15573	0,0686	THE FRENCH PRIME_MINISTER , MANUEL VALLS , AND THE EUROPEAN COMMISSION PRESIDENT , JEAN-CLAUDE JUNCKER , YESTERD
EUROPEAN	6469	0,0678	The suspension of free travel by the Germans was backed by the European Commission as being within the rules . However , Commission Preside
EUROPEAN	6589	0,0617	So much , then , for European brotherhood and the principle of an ever-closer union ' .
EUROPEAN	14463	0,0442	Antonio Guterres , the head of UNCHR , warned European countries yesterday to keep out the welcome mat for genuine Iraqi asylum-seekers or r
EUROPEAN	21229	0,0437	Mr Brown angered the European Commission and his European counterparts on Monday by announcing that he was going to a finance ministers ' m
EUROPEAN	21793	0,0417	5 Which two European nations failed to win a game ?
EUROPEAN	19694	0,0417	And they are being joined by failed asylum_seekers and Eastern European economic_migrants .
EUROPEAN	6635	0,0417	They make_up around half the 1 . 3million eastern Europeans in the UK .
EUROPEAN	15582	0,0413	Europe should embrace more refugees fleeing war and dictatorship while also tightening border controls and more strictly enforcing its returns polic
EUROPEAN	24152	0,0410	' ' The government needs to stop apportioning blame by pushing the responsibility back onto the Muslim community . Instead , those professiona
EUROPEAN	13117	0,0386	His brief covers European integration , international patterns of economic growth , investment , productivity , wages and employment .
EUROPEAN	16024	0,0385	The ' ' vice-president foreign minister ' ' who is to be appointed under the new European constitution will_be assisted by a European external

Así como ocurre para otros tipos de análisis temático**, T-LAB** permite exportar el diccionario de la partición final. De este modo, su uso estará disponible para ulteriores análisis.

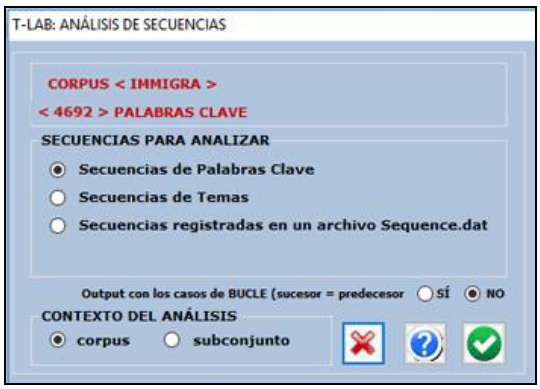
C - ALGUNOS DETALLES TÉCNICOS

Esta herramienta de **T-LAB** puede ser implementada a partir de las siguientes tipologías de secuencias:

1- **Secuencias de palabras-clave**, cuyos elementos son unidades lexicales (es decir, palabras o lemas) presentes en el corpus o un subconjunto del corpus mismo. En este caso, el número máximo de 'nudos' (es decir, los 'tipos' de unidades lexicales) es 5.000;
N.B.: Quando se aplica la lematización automática, 5.000 unidades léxicales corresponden a cerca de 12.000 palabras.

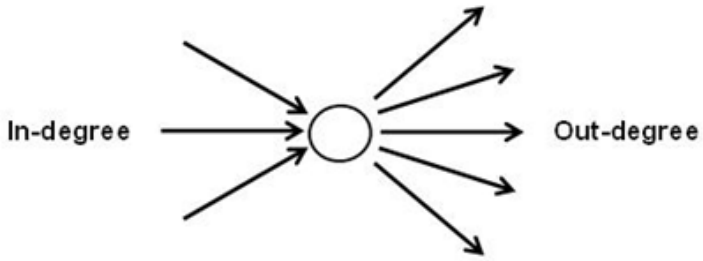
2- **Secuencias de Temas**, cuyos elementos son las unidades de contexto (es decir, contextos elementales) clasificadas por una de las herramienta de **T-LAB** para el análisis temático.
N.B.: En este caso, ya que la secuencia de los contextos elementales (frases o párrafos) caracteriza la 'cadena' entera del corpus (predecesores y sucesores), T-LAB implementa una forma concreta de Análisis del Discurso, cuyos nudos (es decir los 'temas') varían de un mínimo de 5 a un máximo de 50.

3 - **Secuencias registradas en un archivo Sequence.dat** predispuesto por el usuario (véanse las explicaciones pertinentes al final de esta sección). En este caso, el número máximo de records es 50.000 y el numero de 'tipos' (es decir, los nudos) no debe superar los 5.000.

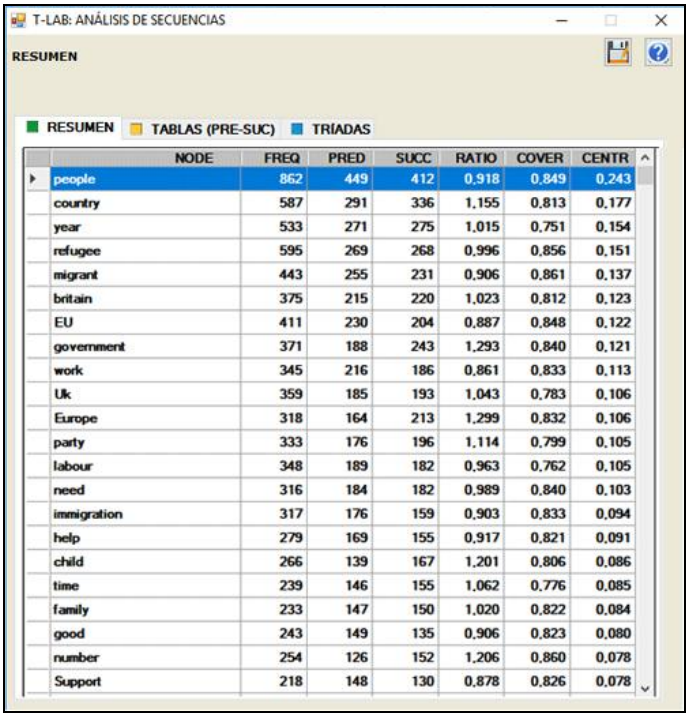


Las informaciones que siguen vienen proporcionadas para que el usuario comprenda mejor los datos incluidos en la tabla RESUMEN.

Según la teoría de gráficos, los predecesores y los sucesores de cada nodo (en este caso, unidad lexical) pueden ser representados por medio de flechas (arcos) entrantes (in-degree = los tipos de predecesores) y salientes (out-degree = los tipos de sucesores).



Por ejemplo, en la tabla siguiente "people" tiene 412 tipos de sucesores y 449 tipos de predecesores.
Y el centrality degree es igual a 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

Según el cociente (sucesores/predecesores), es posible verificar la variedad semántica engendrada por cada nodo:

- si el cociente es mayor de 1, el nodo es definido "fuente";
- si el cociente es igual a 1, el nodo es definido "relais"
- si el cociente es más bajo de 1, el nodo es definido "pozo".

En la misma tabla, para cada unidad lexical, la columna "cover" (cobertura) indica el porcentaje de sus ocurrencias precedidas o seguidas por las unidades lexicales incluidas en la lista del usuario.

Cuando las unidades analizadas "cubren" la totalidad de los presentes dentro del corpus, el valor de "cover" es igual a 1; si no, es un valor inferior. Por otra parte: cuando el valor de "cover" es igual a 1, también las adiciones de los valores de probabilidad (de predecesores y de sucesores) son iguales a 1; si no, son valores inferiores. En ambos casos, el porcentaje "residual" es determinado por el hecho de que hay predecesores y sucesores no incluidos en el análisis.

Por ejemplo, la secuencia representada en la imagen siguiente es constituida por 39 acontecimientos: de éstos, solamente 16 (las hipotéticas unidades en análisis) "se cubren" (las cajas grises). Es porque algunos de ellos (véase aquéllos correspondientes a las ocurrencias de la unidad "A") tienen predecesores y sucesores no incluidos en el análisis (cajas blancas).

Diferentemente, cuando el usuario analiza Secuencias de Temas o un archivo externo todos los acontecimientos "se cubren".

N.B.: Para analizar un fichero externo es necesario preparar el fichero 'Sequence.dat' correspondiente. Sucesivamente, y una vez abierto un proyecto ya existente, el usuario debe escoger la opción "Secuencias registradas en un archivo Sequence.dat".

El método del cálculo, los gráficos y las tablas son análogos a ésos ya descritos (véase arriba).

El archivo de Sequence.dat, que puede contener cada clase de etiquetas (e.g. los nombres en una conversación, las categorías obtenidas por análisis del contenido, las clases de acontecimientos,etc.), se debe componer por "N" líneas (mínimo 50 máximo 10.000), cada una con una etiqueta de máximo 50 caracteres, sin signos de puntuación o espacios en blanco.

Los tipos de etiquetas deben ser máximo 5.000.

He aquí algunos ejemplos de Sequence.dat en el formato correcto:

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

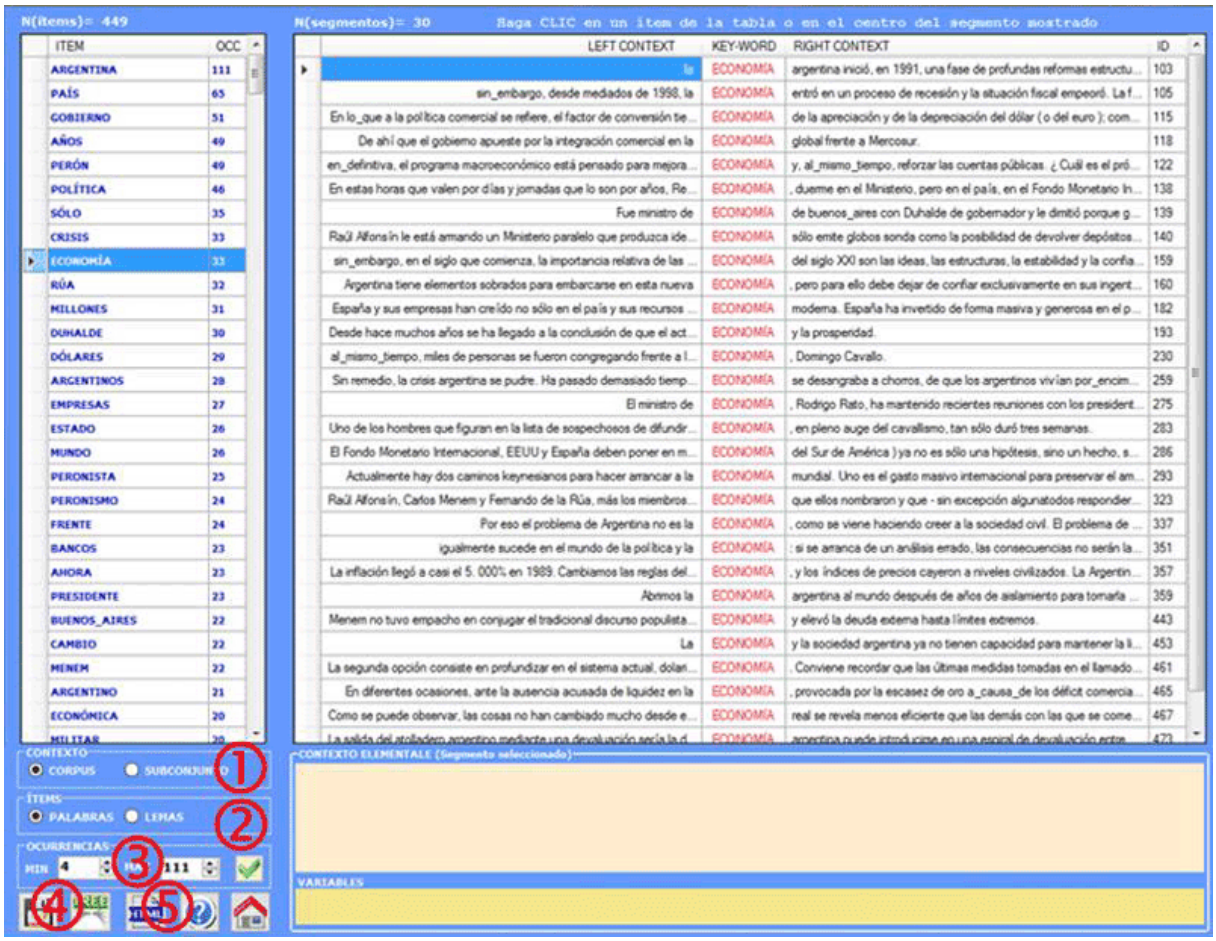
Tanto en el caso de secuencias de unidades lexicales (corpus analizado) como en el de secuencias incluidas en un archivo externo (Sequence.dat), T-LAB produce algunas tablas en la carpeta MY-OUTPUT.

Concordancias

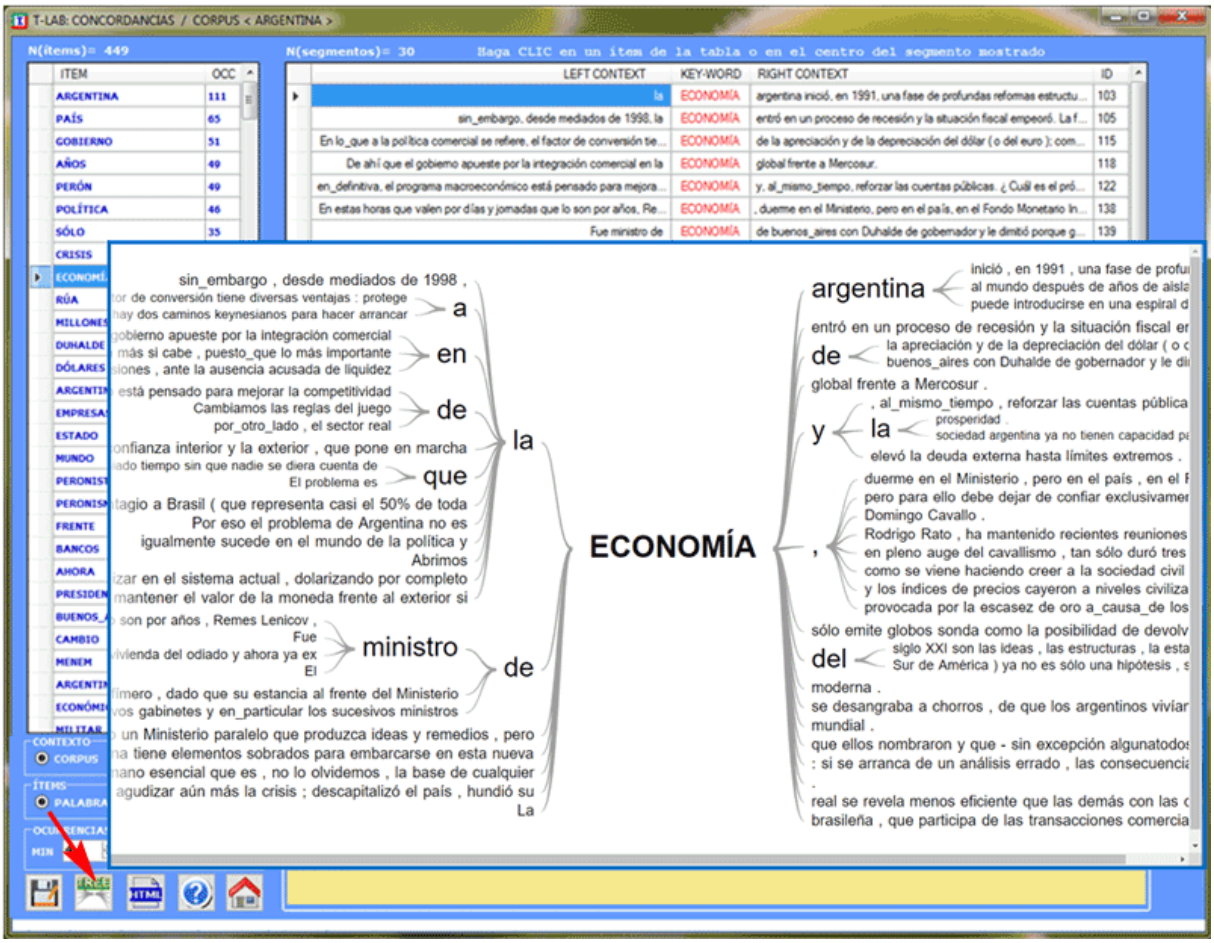
Esta herramienta de **T-LAB** nos permite comprobar los contextos de ocurrencia de cada unidad lexical.

Las búsquedas del tipo KWIC (Key-Word in Context) pueden ser realizadas tanto por **palabras** que por **lemas** (véase abajo la opción '2'). Además, dichas búsquedas peuden ser implementadas bien en todo el **corpus** o bien sólo en un **subconjunto** del mismo (véase abajo la opción '1').

Además es posible definir la gama de las ocurrencias a visualizar (véase abajo la opción '3').



Para cada unidad lexical del corpus, con un simple clic en la columna correspondiente, es posible comprobar cuáles son sus contextos de ocurrencia (los **contextos elementales**); además, es posible crear un 'Word Tree' dinámico (véase abajo la '4') o guardar un archivo HTML con todos los contextos seleccionados (véase abajo la la opción '5').



DATE: 20/06/2015 - 09:32:26
CONCORDANCES **ECONOMÍA**

**** *AUTOR_CAVALLO
la **ECONOMÍA** argentina inició, en 1991, una fase de profundas reformas estructurales que propiciaron un periodo de fuerte crecimiento y baja inflación hasta 1998. El principal pilar de estas reformas fue la llamada ley de convertibilidad, que aseguró la estabilidad de los precios, un amplio proceso de privatizaciones y una notable apertura de la economía.

**** *AUTOR_CAVALLO
sin_embargo, desde mediados de 1998, la **ECONOMÍA** entró en un proceso de recesión y la situación fiscal empeoró. La fuerza del dólar provocó importantes presiones deflacionistas que, al_contrario de lo que sucedió en los primeros años de la convertibilidad, no se vieron compensadas por un crecimiento de la productividad.

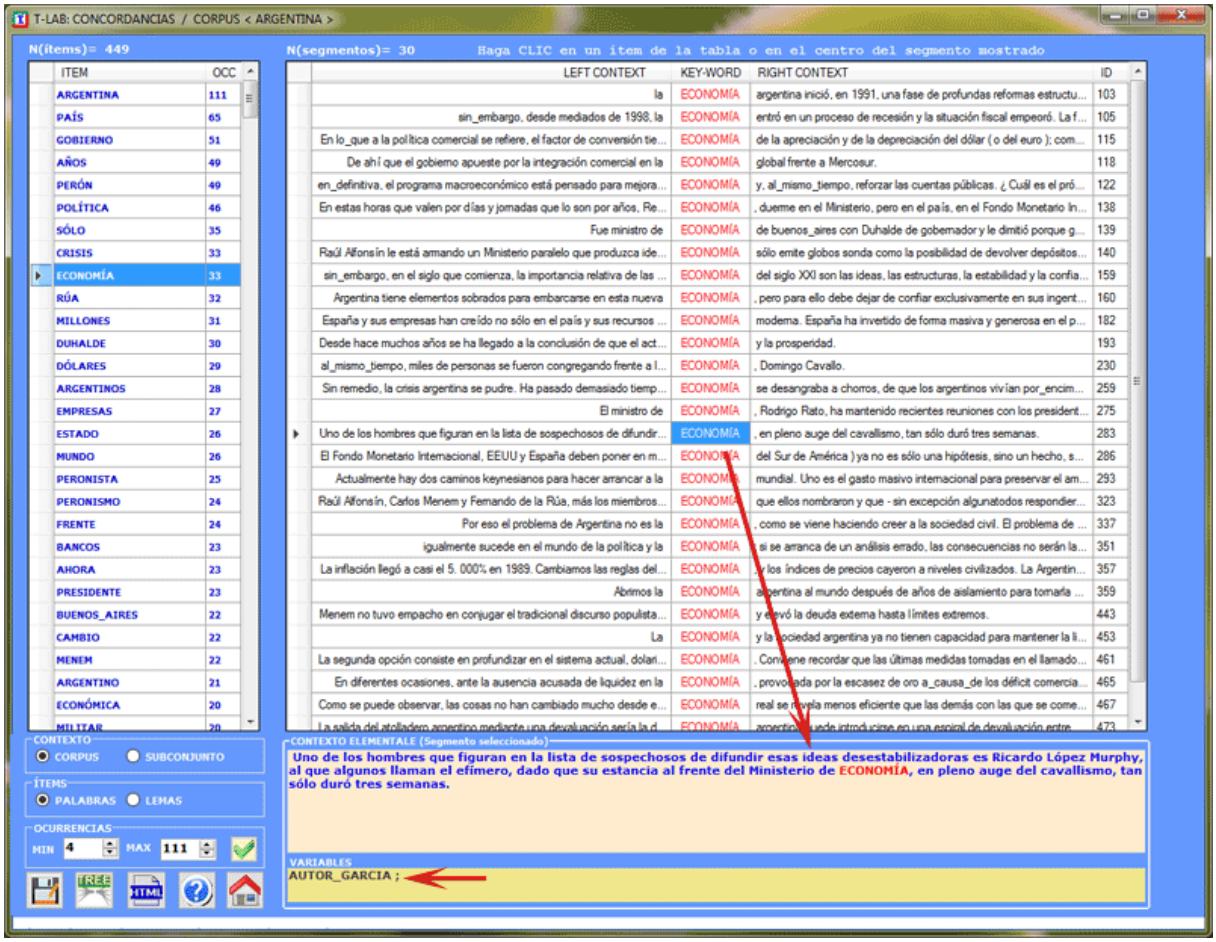
**** *AUTOR_CAVALLO
En lo_que a la política comercial se refiere, el factor de conversión tiene diversas ventajas: protege a la **ECONOMÍA** de la apreciación y de la depreciación del dólar (o del euro); compensa las oscilaciones de los precios dentro del esquema monetario internacional formado por el euro-dólar y reduce, de una forma permanente, "la preferencia comercial de Mercosur".

**** *AUTOR_CAVALLO
De ahí que el gobierno apueste por la integración comercial en la **ECONOMÍA** global frente a Mercosur.

**** *AUTOR_CAVALLO
en_definitiva, el programa macroeconómico está pensado para mejorar la competitividad de la **ECONOMÍA** y, al_mismo_tiempo, reforzar las cuentas públicas. ¿ Cuál es el próximo punto de la agenda? En el nuevo trimestre, el Gobierno se concentrará en las reformas estructurales que deben asegurar a Argentina un crecimiento económico sostenido.

**** *AUTOR_PRIE3
En estas horas que valen por días y jornadas que lo son por años, Remes Lenicov, ministro de **ECONOMÍA**, duerme en el Ministerio, pero en el país, en el Fondo Monetario Internacional, en las sedes de las multinacionales, se sigue esperando un papel con las nuevas reglas del juego.

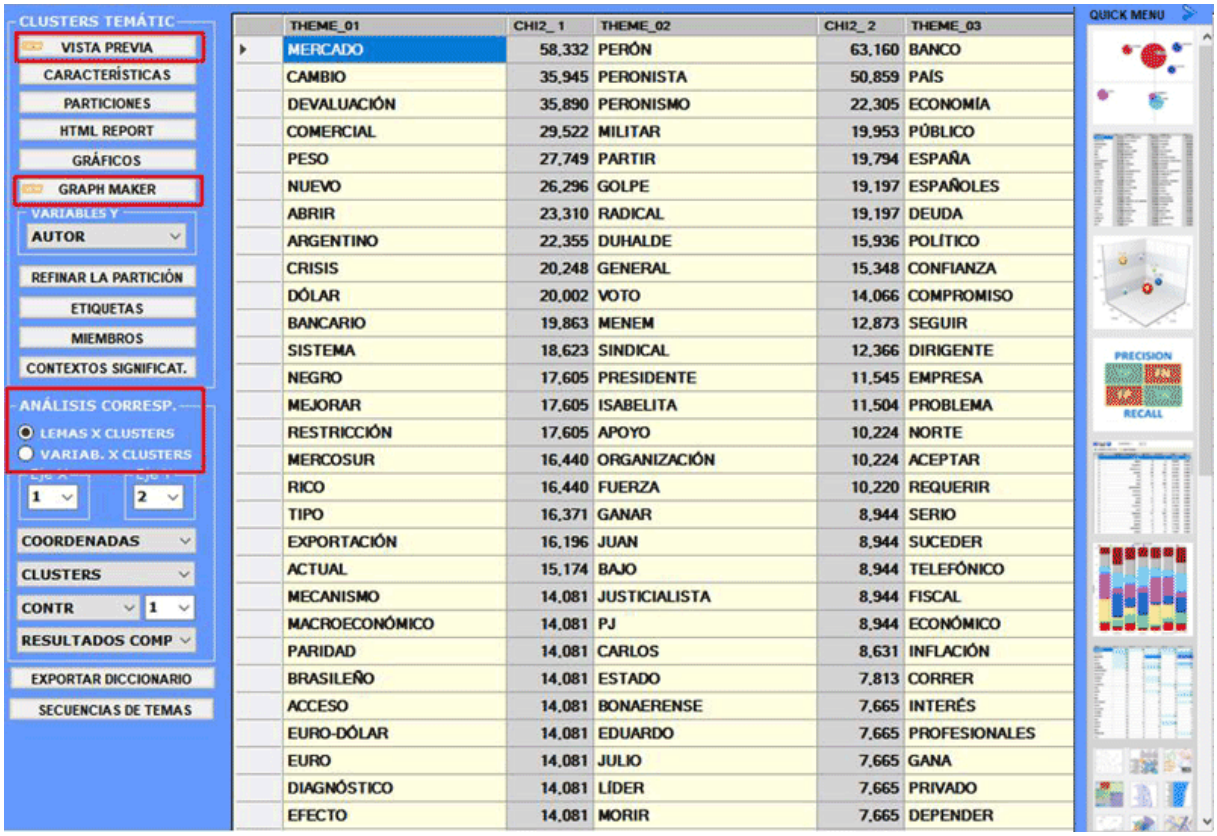
Por otra parte, chascando el centro de un segmento es posible visualizar su contenido y comprobar los categorías usadas en sus líneas de codificación (véase abajo).



ANÁLISIS TEMÁTICOS

Análisis Temático de Contextos Elementales

NOTA: Las imagenes contenidas en este apartado hacen referencia a una versión anterior de T-LAB, ya que el interfaz de **T-LAB 10** cambia ligeramente. Además: a) se ha incluido un nuevo botón (**TREE MAP PREVIEW**) que permite crear gráficos dinámicos en formato HTML; b) el botón Dendrograma ha sido sustituido por la herramienta **GRAPH MAKER**.; c) es posible implementar ulteriores análisis de las correspondencias entre los clústeres temáticos y cada una de las variables disponibles; d) una galería de imágenes de acceso rápido que funciona como un menú adicional permite cambiar entre varias salidas con un solo clic. Algunas de estas nuevas características se destacan en la imagen de abajo.



THEME_01	CHI2_1	THEME_02	CHI2_2	THEME_03
MERCADO	58,332	PERÓN	63,160	BANCO
CAMBIO	35,945	PERONISTA	50,859	PAÍS
DEVALUACIÓN	35,890	PERONISMO	22,305	ECONOMÍA
COMERCIAL	29,522	MILITAR	19,953	PÚBLICO
PESO	27,749	PARTIR	19,794	ESPAÑA
NUEVO	26,296	GOLPE	19,197	ESPAÑOLES
ABRIR	23,310	RADICAL	19,197	DEUDA
ARGENTINO	22,355	DUHALDE	15,936	POLÍTICO
CRISIS	20,248	GENERAL	15,348	CONFIANZA
DÓLAR	20,002	VOTO	14,066	COMPROMISO
BANCARIO	19,863	MENEM	12,873	SEGUIR
SISTEMA	18,623	SINDICAL	12,366	DIRIGENTE
NEGRO	17,605	PRESIDENTE	11,545	EMPRESA
MEJORAR	17,605	ISABELITA	11,504	PROBLEMA
RESTRICCIÓN	17,605	APOYO	10,224	NORTE
MERCOSUR	16,440	ORGANIZACIÓN	10,224	ACEPTAR
RICO	16,440	FUERZA	10,220	REQUERIR
TIPO	16,371	GANAR	8,944	SERIO
EXPORTACIÓN	16,196	JUAN	8,944	SUCEDER
ACTUAL	15,174	BAJO	8,944	TELEFÓNICO
MECANISMO	14,081	JUSTICIALISTA	8,944	FISCAL
MACROECONÓMICO	14,081	PJ	8,944	ECONÓMICO
PARIDAD	14,081	CARLOS	8,631	INFLACIÓN
BRASILEÑO	14,081	ESTADO	7,813	CORRER
ACCESO	14,081	BONAERENSE	7,665	INTERÉS
EURO-DÓLAR	14,081	EDUARDO	7,665	PROFESIONALES
EURO	14,081	JULIO	7,665	GANA
DIAGNÓSTICO	14,081	LÍDER	7,665	PRIVADO
EFECTO	14,081	MORIR	7,665	DEPENDER

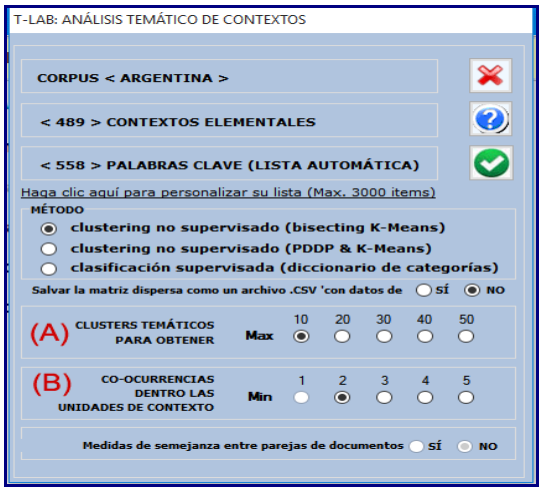
Esta herramienta de **T-LAB** nos permite obtener una **representación de los contenidos del corpus** mediante pocos y significativos **clusters temáticos** (de 3 a 50), de modo que cada uno de ellos:

a) resulta constituido de un conjunto de **contextos elementales** (ej. frases, párrafos, fragmentos de texto, respuestas a preguntas abiertas) caracterizados por los mismos patrones (patterns) de palabras clave;

b) puede ser descrito por las **unidades lexicales** (palabras, lemas o categorías) y por las **variables** (si presentes) que más caracterizan los contextos elementales de los cuales se compone.

Por muchos motivos, los resultados del análisis se pueden interpretar como mapas de **isotopías** (iso = igual; topos = lugar), es decir como mapas de temas "genéricos" o "específicos" (Rastier, 2002: 204) caracterizados por la co-ocurrencia de componentes semánticos.

El proceso de análisis puede ser implementado bien a través de un método de **clustering no supervisado** (en el caso concreto, un algoritmo bisecting k-means) o bien a través de una **clasificación supervisada** (es decir, el enfoque top-down). Si se elige el segundo procedimiento, (es decir, clasificación supervisada), se requiere la importación de un diccionario de las categorías, resultado de un anterior análisis **T-LAB** o de una elaboración del usuario.



Una caja de diálogo (véase arriba) permite que el usuario fije algunos parámetros del análisis

En particular:

- el parámetro (A) permite que el usuario fije el número máximo de clusters que se incluirán en los outputs de **T-LAB**;
- el parámetro (B) permite que el usuario excluya del análisis cualquier unidad del contexto que no contenga un número mínimo de palabras clave incluidas en la lista que él está utilizando.

NOTA:
- Ambos los parámetros antedichos producen cambios significativos en los resultados del análisis solamente cuando el número de las unidades del contexto es muy grande y/o cuando los textos analizados son cortos;
- A la hora de seleccionar la opción 'clasificación supervisada', ya que el numero de clústeres que hay que obtener coincide con el numero de categorías del diccionario, el parámetro 'A' no aparece como disponible

En el caso de **clustering no supervisado** (opción por defecto), el proceso de análisis se compone de las siguientes etapas:

a - construcción de una tabla unidades de contexto x unidades lexicales (hasta 300.000 filas x 3.000 columnas), con valores de tipo presencia-ausencia;
b - cálculo de pesos **TF-IDF** y normalización de los vectores (norma euclídea);
c - clusterización de las unidades de contexto (medida de semejanza: coeficiente del coseno; método de clusterización: bisecting K-means; referencias: Steinbach, Karypis, & Kumar,

2000; Savaresi, Booley, 2001);
d - salvaguardia de las particiones obtenidas y, para cada una de ellas:
e - construcción de una tabla de contingencia unidades lexicales x clusters (n x k);
f - test del chi cuadrado aplicado a todos los cruces unidades lexicales x clusters.
g - análisis de las correspondencias de la tabla de contingencia (referencias: Benzécri, 1984; Greenacre, 1984; Lebart, Salem, 1994).

NOTA : A partir de la versión de **T-LAB 2016** , el agrupamiento no supervisado de las unidades de contexto (véase más arriba elel paso 'c' más) puede ser realizado de dos maneras (1) bien usando el algoritmo bisecting k-means o (2) bien usando una versión no centrada del PDDP (es decir, Principal Direction Divisive Partitioning) propuesto por D. Booley (1998) para seleccionar los centroides de las diferentes bisecciones K-means.
Las principales diferencias entre los métodos anteriores se basa en cómo se computan los dos centroides de cada una de las bisecciones; de hecho, en el caso (1) que son el resultado de un procedimiento iterativo, mientras que en el caso (2) se calculan a través de SVD (Singular Value Decomposition), es decir a través de un algoritmo "one-shot '(ver Savaresi, SM, y Boley, DL, 2004).

Así, este procedimiento realiza un **análisis de las co-ocurrencias** (pasos a-b-c) y, a continuación, un **análisis comparativo** (e-f-g). En particular, el análisis comparativo utiliza como columnas de las tablas de contingencia las modalidades (niveles o categorías) de la "nueva variable" derivada del análisis de las co-ocurrencias (modalidades de la nueva variable = clusters temáticos).

En el caso de **clasificación supervisada**, las fases del análisis comparativo son las mismas (véase arriba e-f-g), mientras que el análisis de las co-ocurrencias se ejecuta como sigue:

a) Normalización de los seed vectors (es decir, los perfiles de las co-ocurrencias) correspondientes a las 'k' categorías del diccionario importado;

b) Calculo de los índices de coseno y de las distancias euclidianas entre cada 'i' unidad de contexto y cada 'k' ‘semilla’ de vectores;

c) Asignación de cada 'i' unidad de contexto a la 'k' clase o categoría que tiene la maximiza la semejanza con la semilla correspondiente (en este caso, la máxima semejanza del coseno y la mínima distancia euclidiana deben coincidir. De no ser así, **T-LAB** considera la 'i' unidad de contexto como no clasificada).

NOTA: Cuando el usuario decide repetir/aplicar los resultados de un análisis anterior (es decir, un **Análisis Temático de los Contextos Elementales** o una **Modelización de los Temas Emergentes**), **T-LAB** sólo ejecuta un análisis comparativo (pasos e-f-g).

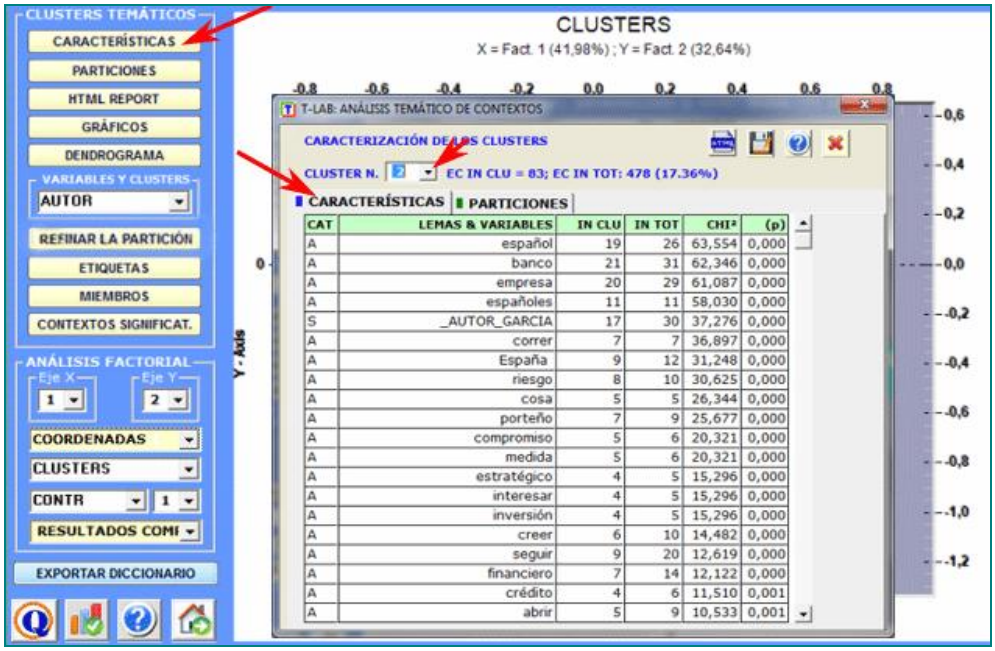
Al término del análisis, el usuario puede efectuar rápidamente las siguientes operaciones:

1 - explorar las características de los clusters;
2 - explorar las relaciones entre clusters;
3 - explorar las relaciones entre clusters y variables;
4 - explorar las diversas particiones de los clusters;
5 - refinar los resultados de la partición elegida y, si es necesario, repetir unos pasos antedichos

(1,2,3);
6 - asignar etiquetas a los clusters;
7 - verificar qué contextos elementales pertenecen a qué clusters;
8 - verificar el peso de cada uno de los contextos elementales dentro del cluster al que pertenece;
9- obtener una clasificación temática de los documento (proporcionada solamente cuando el corpus se compone por lo menos de 2 documentos primarios y éstos no son textos breves como las respuestas a preguntas abiertas);.
10- archivar la partición seleccionada para explorarla con otras herramientas **T-LAB**;
11- exportar un diccionario de las categorías;
12 – verificar la calidad de la partición elegida y la coherencia semántica entre los diferentes temas;
13 - además, cuando el corpus se articula como un discurso o como una conversación, es decir cuando las unidades de contexto se suceden con un orden temporal preciso, se pueden explorar las secuencias de temas de forma dinámica (véase abajo, en la parte final de la sección).

En detalle:

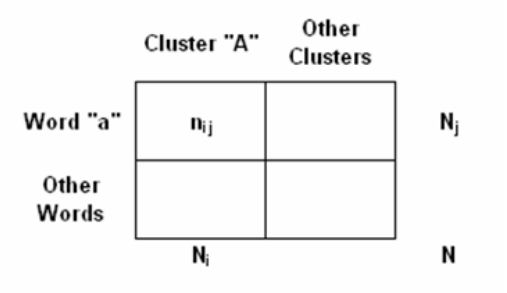
1 - Explorar las características de los clusters



CAT	LEMAS & VARIABLES	IN CLU	IN TOT	CHI²	(p)
A	español	19	26	63,554	0,000
A	banco	21	31	62,346	0,000
A	empresa	20	29	61,087	0,000
A	españoles	11	11	58,030	0,000
S	_AUTOR_GARCIA	17	30	37,276	0,000
A	correr	7	7	36,897	0,000
A	España	9	12	31,248	0,000
A	riesgo	8	10	30,625	0,000
A	cosa	5	5	26,344	0,000
A	porteño	7	9	25,677	0,000
A	compromiso	5	6	20,321	0,000
A	medida	5	6	20,321	0,000
A	estratégico	4	5	15,296	0,000
A	interesar	4	5	15,296	0,000
A	inversión	4	5	15,296	0,000
A	creer	6	10	14,482	0,000
A	seguir	9	20	12,619	0,000
A	financiero	7	14	12,122	0,000
A	crédito	4	6	11,510	0,001
A	abrir	5	9	10,533	0,001

Haciendo clic en el botón **características**, para cada cluster se muestran los valores siguientes: Chi-cuadrado y sumatoria de contextos elementales en los que cada característica (lemma o variable) se encuentra presente, bien sea en el interior del cluster seleccionado ("IN CLUST") o en el interior del conjunto analizado ("IN TOT"). Además, la columna "CAT" indica si la característica ha sido seleccionada por el usuario ("A"), con la función **Configuración del Análisis**, o si ha sido sugerida por **T-LAB** como descripción "suplementaria" ("S").

En el caso del **chi cuadrado** la estructura de la tabla analizada es la siguiente:



Donde:
n_{ij} se refiere a las ocurrencias de la palabra (a) dentro del cluster seleccionado (A)
N_j se refiere a todas las ocurrencias de la palabra (a) dentro del corpus (o del subconjunto) analizado
N_i se refiere a todas las ocurrencias de palabras dentro del cluster seleccionado (A)
N se refiere a todas las ocurrencias de la tabla de la contingencia palabras x clusters.

Un **informe HTML** (ver a continuación) permite verificar en detalle las características de los clusters. En éste, además de la lista de palabras típicas, se muestran los contextos elementales que más caracterizan el cluster seleccionado, ordenados de manera descendente según el respectivo peso (score).

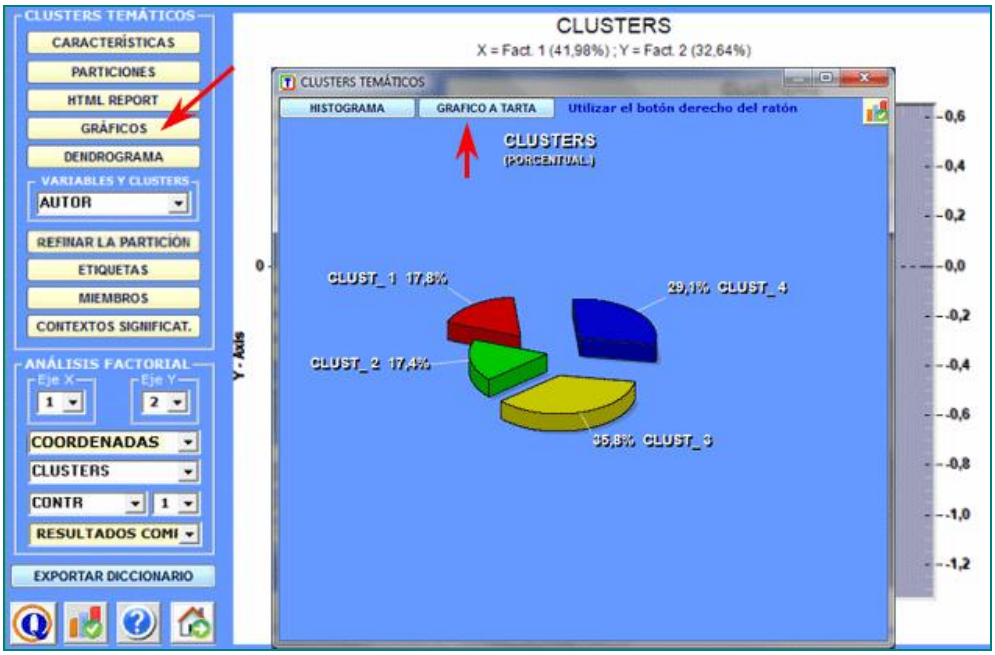
LEMMA	CHI SQUARE	WORD	OCC
español	63.554	español	4
español	63.554	española	3
español	63.554	españolas	12
banco	62.346	banco	5
banco	62.346	bancos	16
empresa	61.087	empresa	1
empresa	61.087	empresas	19
españoles	58.03	españoles	11

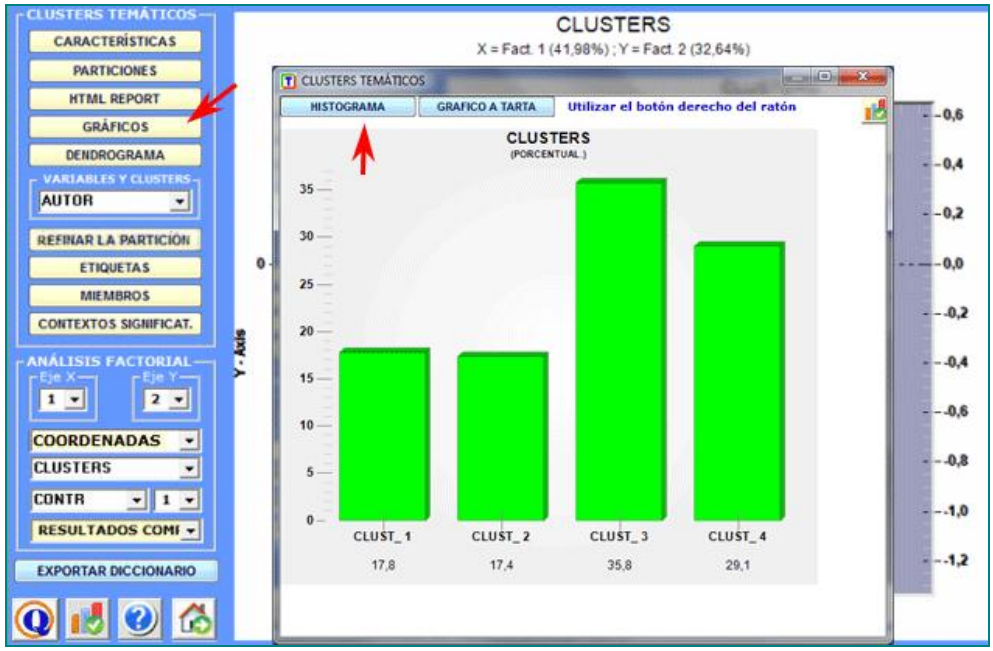
"Las **medidas adoptadas** por el **Gobierno argentino** son una confiscación de hecho de toda la banca", comentaba esta **semana** en privado un **financiero español**, buen conocedor de la **realidad argentina**. Los **bancos españoles**, que **suponen** un porcentaje **importante** del **sistema**, **sufren** a diario los ataques de los desesperados.

**** *AUTOR_DEARIST
SCORE (24.620)

España y sus **empresas** han **creído** no **sólo** en el **país** y sus **recursos** sino, sobre_todo, en su ciudadanía, en el elemento humano esencial que es, no lo olvidemos, la base de cualquier **economía** moderna. **España** ha invertido de **forma masiva** y generosa en el **país**.

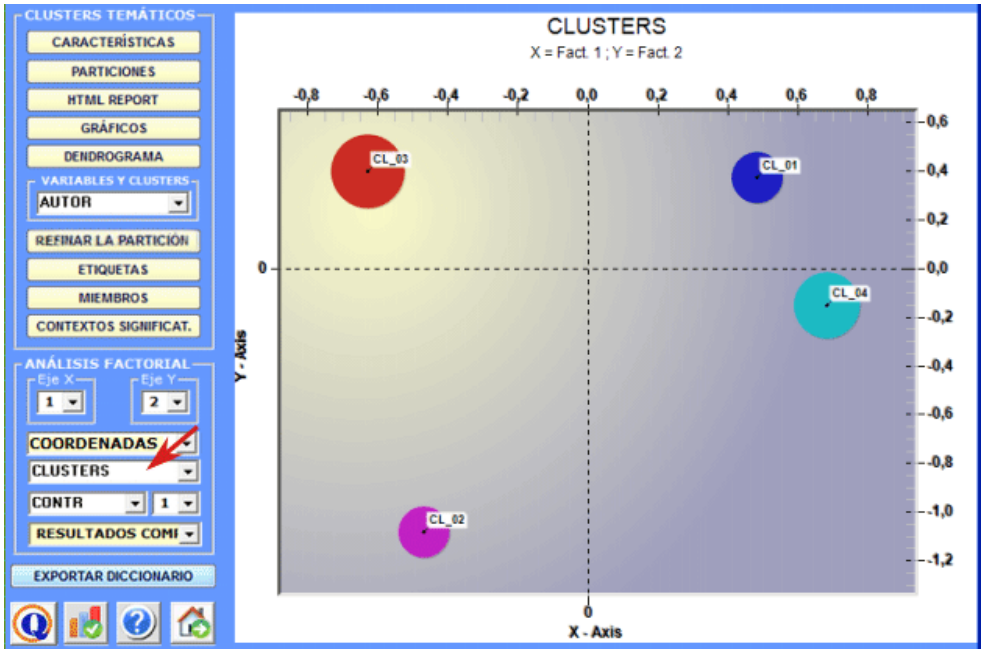
Gráficos a tarta y histogramas (véase abajo) permiten verificar el porcentaje de unidades de contexto que pertenece a cada cluster.



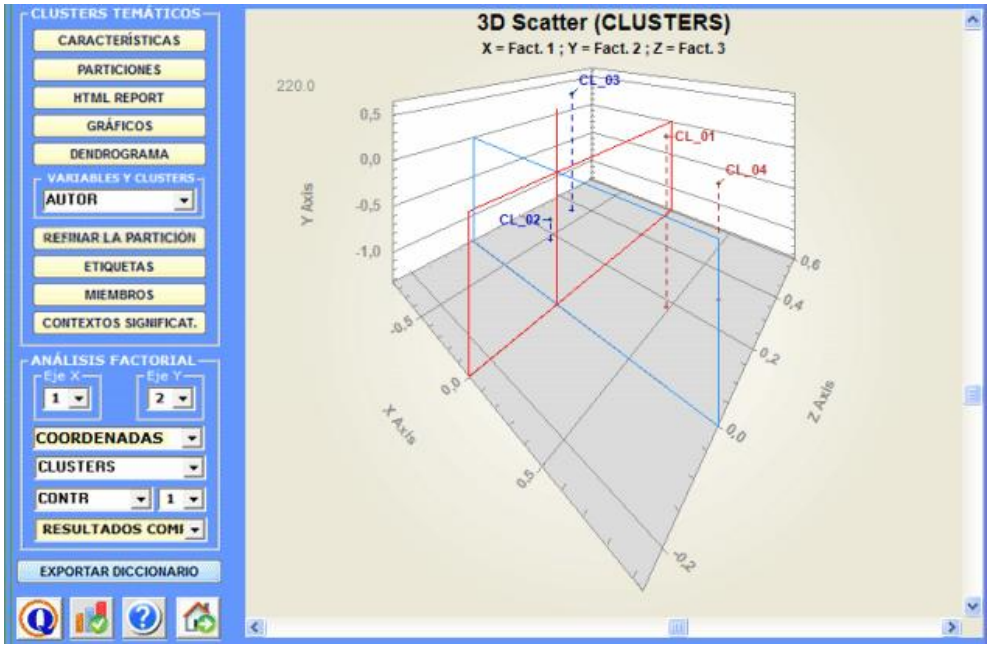


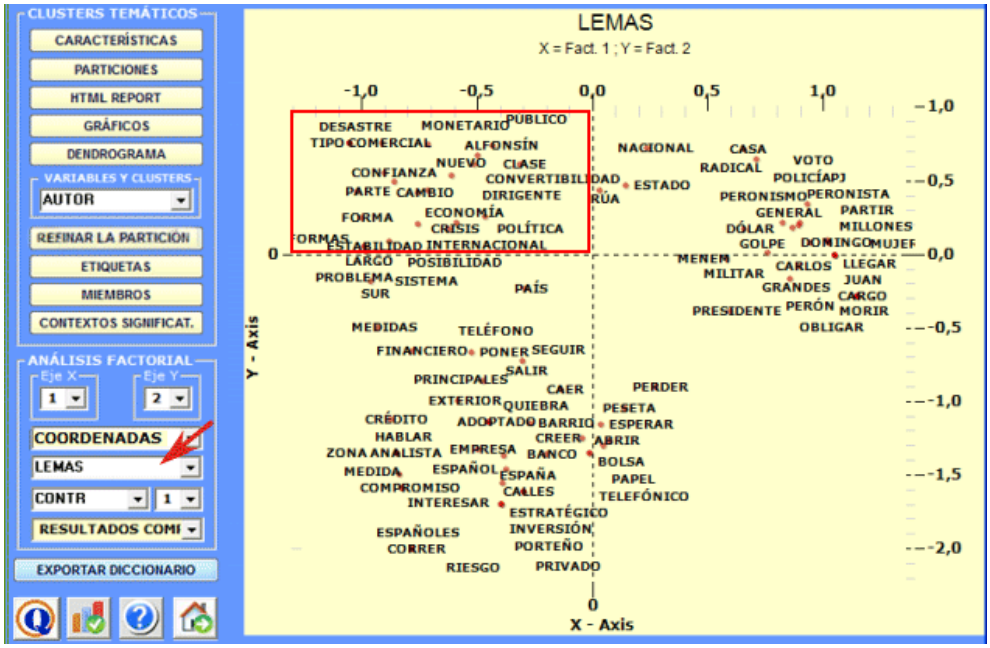
2 - Explorar las relaciones entre clusters

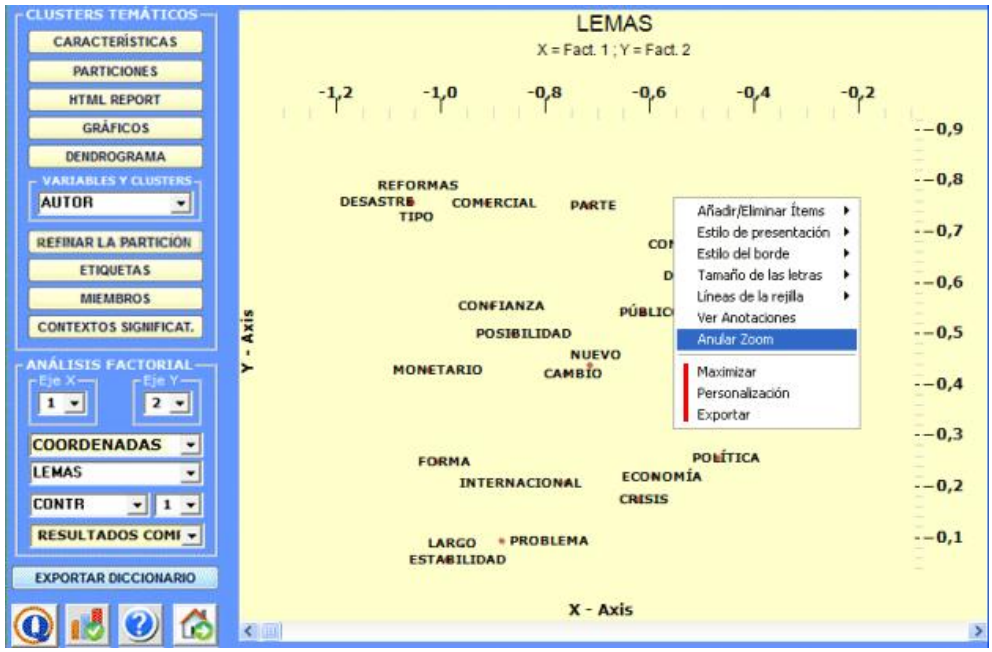
Algunos gráficos, obtenidos por medio de **Análisis de Correspondencias**, permiten explorar las relaciones entre clusters en espacios bidimensionales.
De forma más específica:
- para explorar las distintas combinaciones de los ejes factoriales es suficiente seleccionarlos en los boxes apropiados ("Eje X", "Eje Y");
- para cada una de las combinaciones (X-Y), es posible visualizar los distintos tipos de elementos (clusters, lemas y variables).



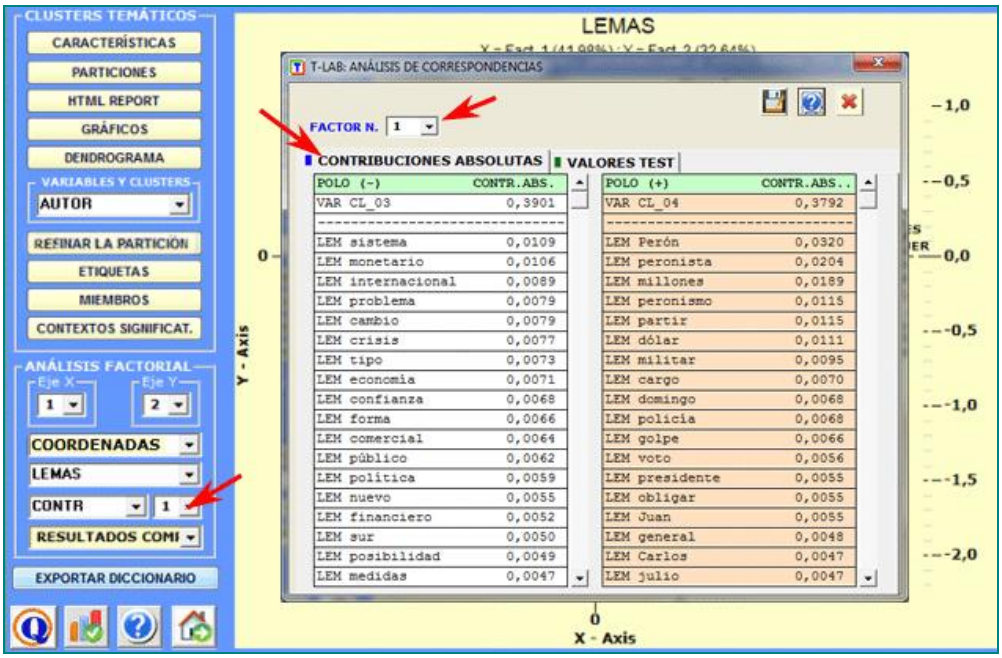
Todos los gráficos pueden ser personalizados usando el apropiado cuadro de diálogo (hacer clic en el botón derecho del ratón). Además cuando los clústers temáticos son más de tres, sus relaciones pueden ser exploradas en **gráficos 3D** (ver abajo).



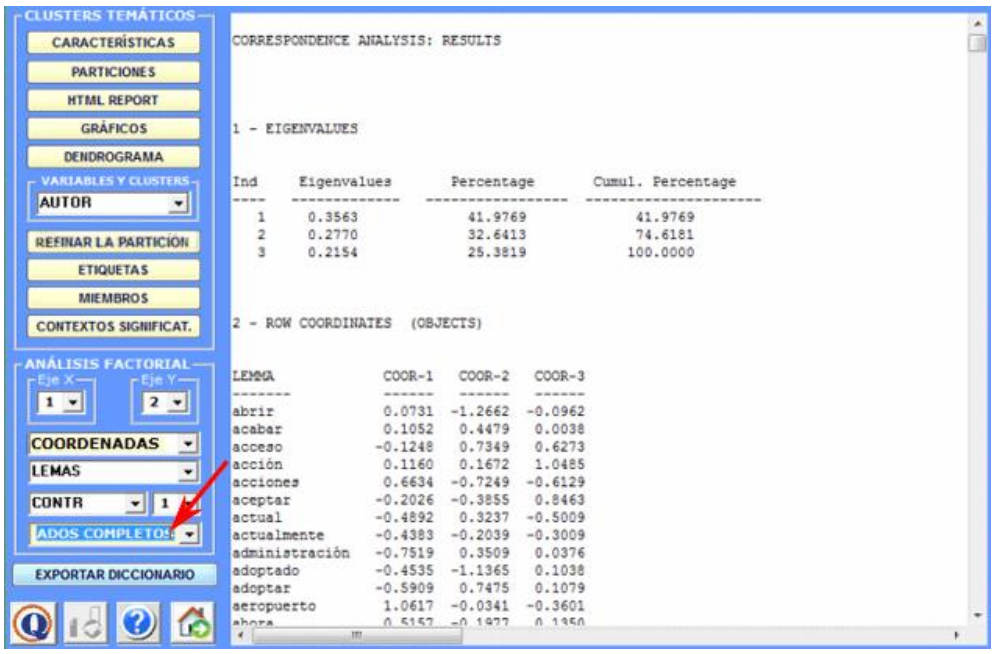




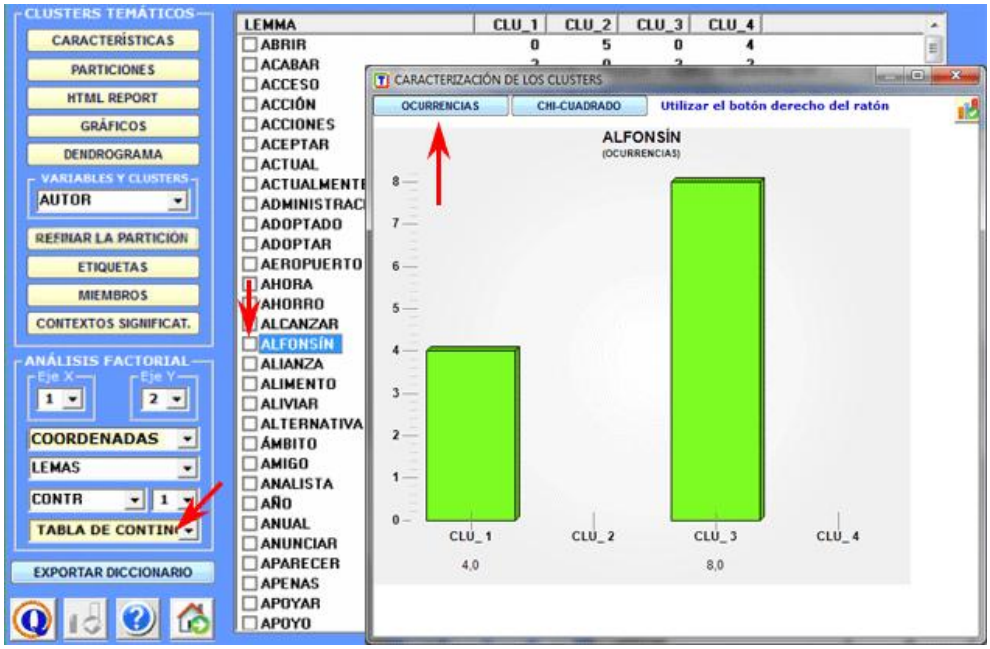
Las características de cada polo factorial pueden ser exploradas haciendo clic en los botones marcados en rojo (véase abajo).

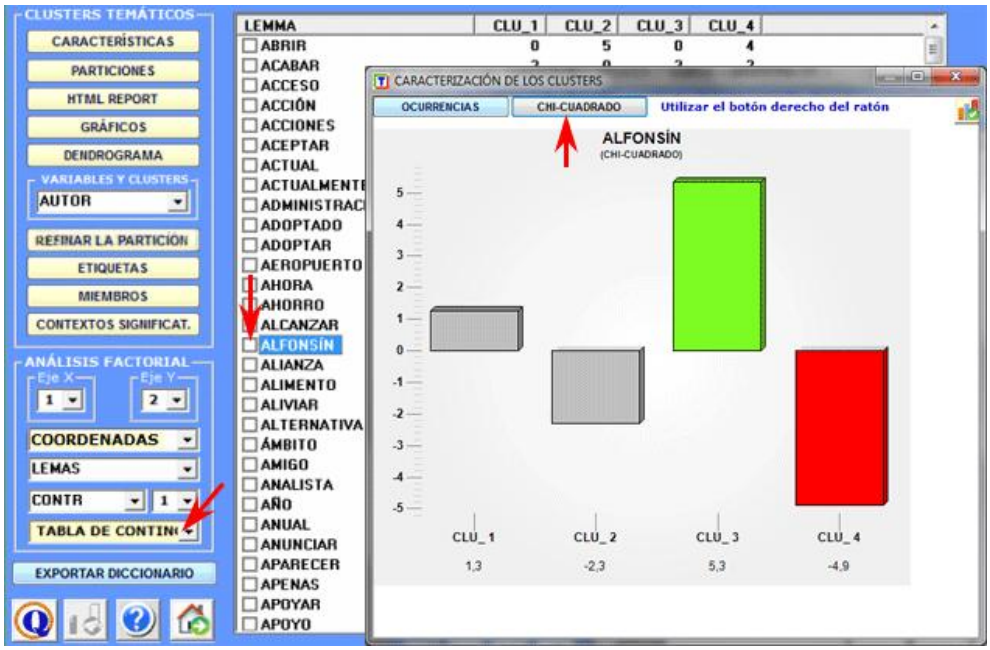


Un clic en el botón correspondiente permite que usted visione y guarde el archivo que contiene los **resultados completos** del análisis: valores propios, coordenadas, aportes absolutos y relativos, valores test.



Una opción específica (véase más abajo) nos permite visualizar/exportar la **tabla de contingencia** y crear gráficos que muestran la distribución de cada palabra dentro de los clusters.



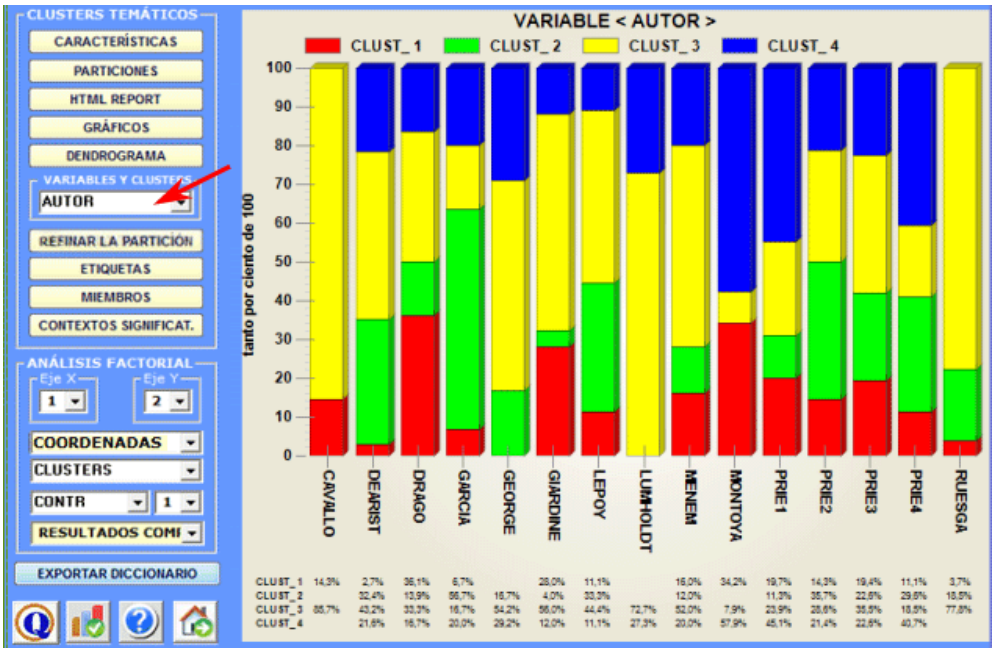


Además, haciendo clic en específicas células de la tabla, es posible crear un archivo HTML que incluye todos los contextos elementales en que la palabra en la fila está presente en el cluster correspondiente.

NOTA: Esta tabla incluye tanto las palabras clave activas ("A") como aquellas suplementarias ('S').

3 - Explorar las relaciones entre clusters y variables

Algunos **histogramas** permiten verificar las relaciones entre los clusters y las variables.



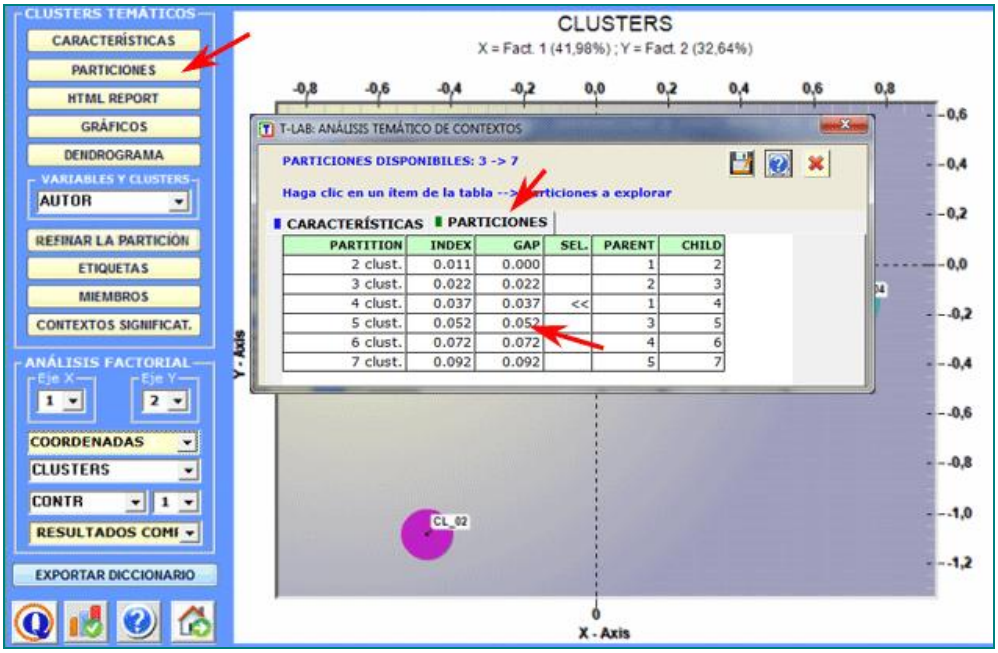
Además es posible explorar ulteriores relaciones entre clusters y variables con las opciones disponibles en la sección **Análisis factorial**" (ver más arriba).

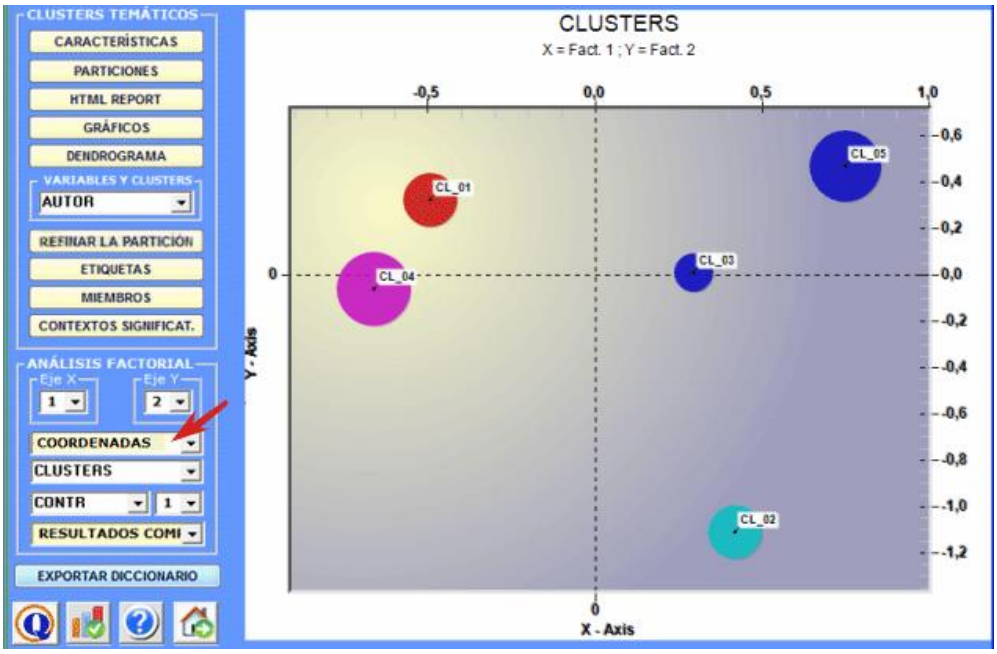
4 - Explorar las diversas particiones de los clusters

Posto que el algoritmo usado produce una clusterización jerárquica, el usuario puede explorar fácilmente diferentes soluciones del análisis: particiones de 3 a 50 clusters.

Para cada una de las particiones obtenidas existe una tabla (ver a continuación) con los siguientes valores:
- "Index", que corresponde a la relación entre la varianza intercluster y la varianza total;
- "Gap", que indica la diferencia entre el valor del índice y el de la partición inmediatamente anterior;
- Número del cluster "hijo" (child) obtenido por medio de la bi-sección del "progenitor" (parent) correspondiente.

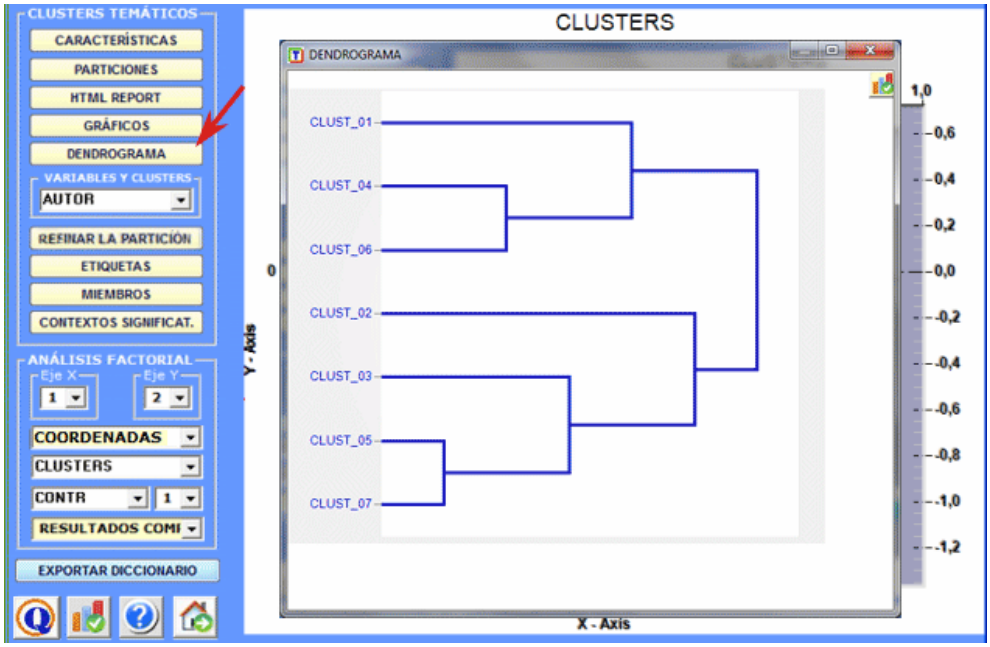
La opción **particiones** permite explorar las características de las soluciones disponibles (clic en los ítems de la tabla).


CLUSTERS TEMÁTICOS
CARACTERÍSTICAS
PARTICIONES
HTML REPORT
GRÁFICOS
DENDROGRAMA
VARIABLES Y CLUSTERS
AUTOR
REFINAR LA PARTICIÓN
ETIQUETAS
MIEMBROS
CONTEXTOS SIGNIFICAT.
ANÁLISIS FACTORIAL
Eje X
Eje Y
COORDENADAS
CLUSTERS
CONTR
RESULTADOS COMI
EXPORTAR DICCIONARIO
CLUSTERS
X = Fact. 1 (41,98%) ; Y = Fact. 2 (32,64%)
T-LAB: ANÁLISIS TEMÁTICO DE CONTEXTOS
PARTICIONES DISPONIBLES: 3 -> 7
CARACTERÍSTICAS
PARTICIONES
PARTITION INDEX GAP SEL. PARENT CHILD
2 clust. 0.011 0.000 1 2
3 clust. 0.022 0.022 2 3
4 clust. 0.037 0.037 << 1 4
5 clust. 0.052 0.052 3 5
6 clust. 0.072 0.072 4 6
7 clust. 0.092 0.092 5 7
CL_02
Y - Axis
X - Axis

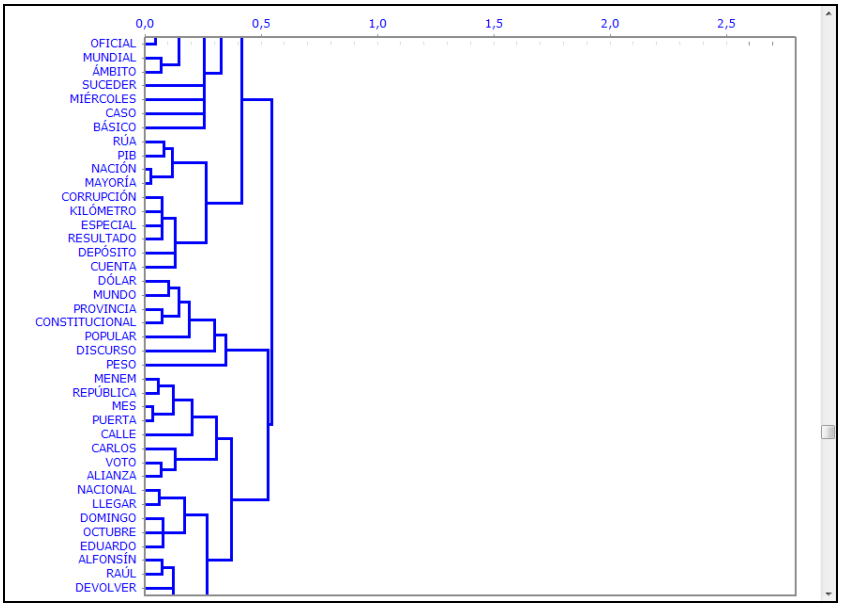

CLUSTERS TEMÁTICOS
CARACTERÍSTICAS
PARTICIONES
HTML REPORT
GRÁFICOS
DENDROGRAMA
VARIABLES Y CLUSTERS
AUTOR
REFINAR LA PARTICIÓN
ETIQUETAS
MIEMBROS
CONTEXTOS SIGNIFICAT.
ANÁLISIS FACTORIAL
Eje X
Eje Y
COORDENADAS
CLUSTERS
CONTR
RESULTADOS COMI
EXPORTAR DICCIONARIO
CLUSTERS
X = Fact. 1 ; Y = Fact. 2
CL_01
CL_02
CL_03
CL_04
CL_05
Y - Axis
X - Axis

Además, la opción dendrograma (véase abajo) permite dos posibilidades:

A) verificar el árbol de las diferentes bisecciones de los clústeres;



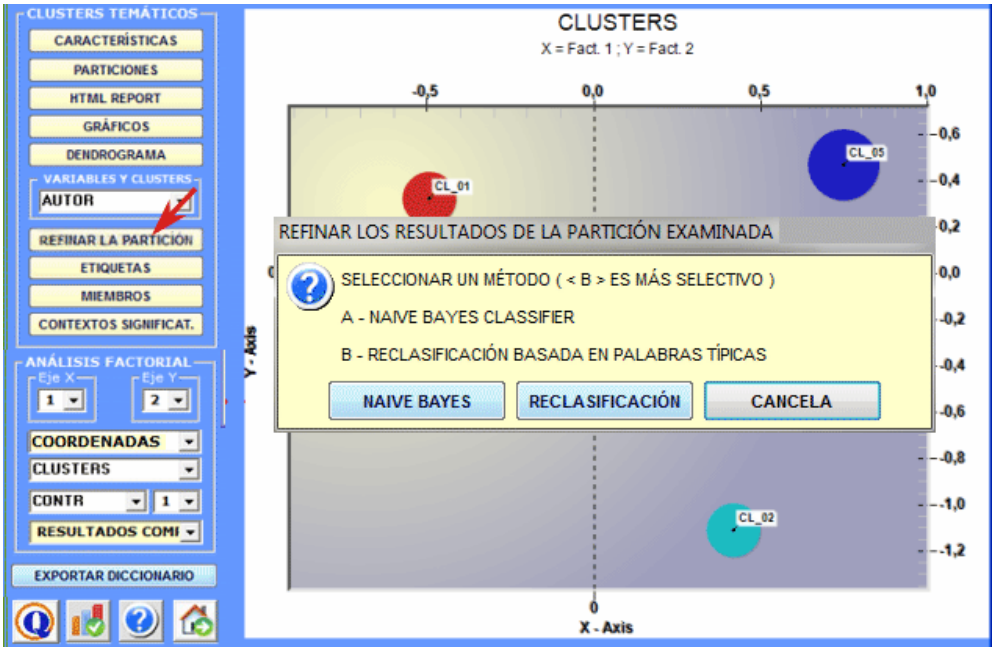
B) verificar el árbol de las palabras características de cada clúster.



5 - Refinar los resultados de la partición elegida

Después de haber explorado diversas soluciones, el usuario puede refinar los resultados de la partición elegida y, si es necesario, repetir unos pasos antedichos (1,2,3).

Para alcanzar este objetivo, hay dos métodos disponibles (véase imagen siguiente).



Cuando se elige el método '**A**' (es decir, el **Naïve Bayes Classifier**), esta opción de **T-LAB** permite que el usuario suprima del análisis todas las unidades del contexto cuya pertenencia a un cluster no satisface los criterios siguientes:
a) por cada unidad de contexto, el cluster asignado mediante el método del bisecting K-Means (unsupervised clustering) y aquel asignado mediante el clasificador Naive Bayes (supervised clustering) deben ser los mismos;
b) el valor máximo de la probabilidad a posteriori que corresponde a la pertenencia de la i-unidad de contexto al k-cluster debe ser, en términos porcentuales, por lo menos 50% más grande que sus valores restantes (es decir las probabilidades a posteriori en otros clusters).

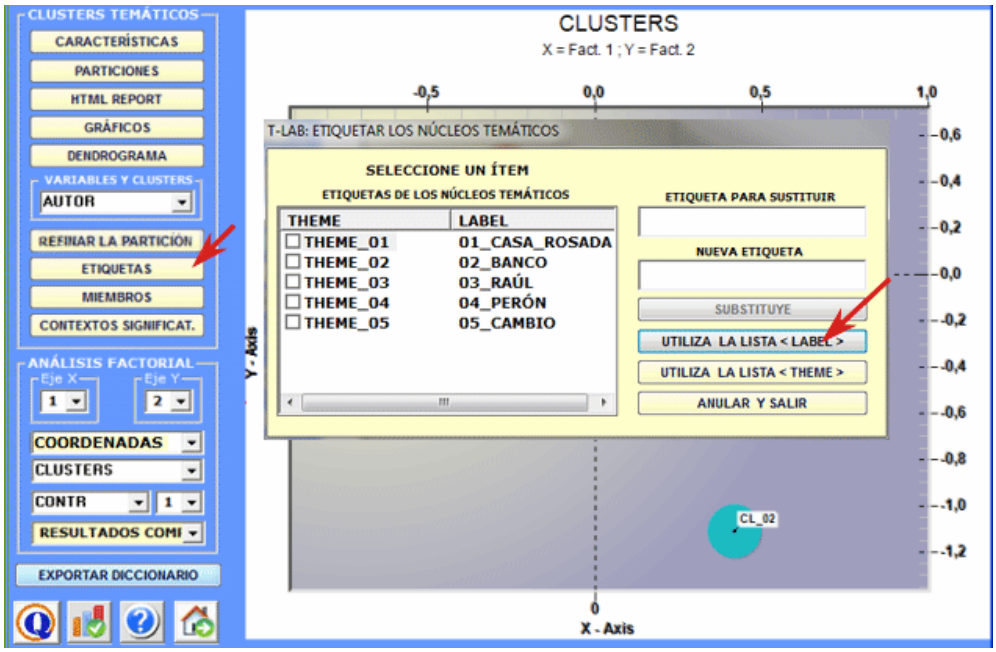
Por otro lado, en el caso del método '**B**' (es decir, **Re-clasificación basada en las Palabras Típicas**), **T-LAB** considera las características de los clústeres, eso es, las palabras que presentan valores significativos de Chi-Cuadrado, como ítems de un diccionario de las categorías y ejecuta las tres fases de la 'clasificación supervisada' descritas al encomienzo de esta sección. Consecuentemente, si el usuario está interesado en volver a aplicar los diccionarios y en comparar los resultados, se recomienda vivamente utilizar este método.

Todos los resultados de este cómputo están en una tabla exportada por **T-LAB** (véase abajo), la cuál contiene los valores de probabilidad expresados en términos porcentuales.

Context_ID	OLD	NEW	MATC	CL1	CL2	CL3	CL4	CL5	CONTEXT
00001000001	1	1	YES	1	0	0	0	0	Los argentinos ya no tienen libre acceso a su dine
00001000002	1	1	YES	1	0	0	0	0	Hace millones de años , cuando en el siglo pasad
00001000003	3	3	YES	0	0	1	0	0	que regresaban al continente exhaustos de combu
00001000004	3	3	YES	0	0	1	0	0	Nada ni nadie ; soledad total y vientos obsesivos
00001000005	5	5	NO	0	0	0,27	0	0,73	Tras horas manejando el auto con un horizonte inr
00001000006	5	5	YES	0	0	0	0	1	Había sentido que tras aquel horizonte sólo había
00001000007	4	4	YES	0	0	0	1	0	en Buenos Aires , los argentinos , ensopados de
00001000008	4	4	YES	0	0	0	1	0	En el aeropuerto internacional de Ezeiza ya no te
00001000009	4	4	YES	0	0	0	1	0	Parece una broma , porque aquí lo_que pagas en
00001000010	4	4	YES	0	0	0	1	0	saqueados por la devaluación y encerrados en un
00001000011	1	1	YES	1	0	0	0	0	Ni conozco otra situación parecida , excepto en e
00001000012	5	5	YES	0	0	0,18	0	0,82	El país de los alimentos no está hambreado como
00001000013	2	2	YES	0	1	0	0	0	Un kilo de pan en buenos_aires CF se pone en 25
00001000014	1	1	YES	1	0	0	0	0	Claro que el problema es cobrar tu salario . Es una
00001000015	4	4	YES	0	0	0	1	0	Asidos todos a las ubres del Estado o de las gobe
00001000016	4	4	YES	0	0	0	1	0	TELEFÓNICA SABRÁ La otra cara de la moneda
00001000017	5	5	YES	0	0	0	0	1	por lo incontable de su fortuna , hecha en la políti
00001000018	4	2	NO	0	0,67	0	0,33	0	César Alierta por Telefónica , Martín Villa por End
00001000019	1	1	YES	1	0	0	0	0	Argentina está quebrada , pero habiendo tenido a
00001000020	2	2	YES	0	1	0	0	0	penthouse (un tríplex es cosa de arribistas social
00001000021	2	2	YES	0	0,98	0	0,02	0	Por supuesto que según qué huéspedes se habla
00001000022	4	5	NO	0,06	0,07	0,19	0,24	0,45	Para ser verdaderamente rico en el mundo hay que
00001000023	4	4	YES	0	0	0	1	0	En uno de esos salones Guido di Tella , de una d
00001000024	1	1	YES	1	0	0	0	0	Yo y otros periodistas extranjeros presentes estuv
00001000025	4	4	YES	0	0	0	1	0	Y su esposa Hilda ' ' Chiche ' ' Duhalde gastab
00001000026	4	4	YES	0	0	0	1	0	Ahora La Chiche quiere resucitar a Evita , firmó c
00001000027	1	1	YES	1	0	0	0	0	El dimisionario de la Rúa (en realidad dio una esp
00001000028	3	3	YES	0	0	1	0	0	La medida yuguló la hiperinflación causante de la
00001000029	1	1	YES	1	0	0	0	0	Mingo Cavallo , un monetarista neoliberal formado
00001000030	1	1	YES	1	0	0	0	0	derogar la ley federal que , lógicamente , hacía in
00001000031	4	4	YES	0	0	0	1	0	UN TIRO EN EL by_pass El suicidio sucedió hace
00001000032	3	3	YES	0	0	1	0	0	Lo rechazó todo en Estados Unidos para regresar
00001000033	4	4	YES	0,11	0	0	0,89	0	A la caída de la dictadura militar daba charlas tele
00001000034	4	4	YES	0	0	0	1	0	Creó una fundación con su nombre , desahuciada

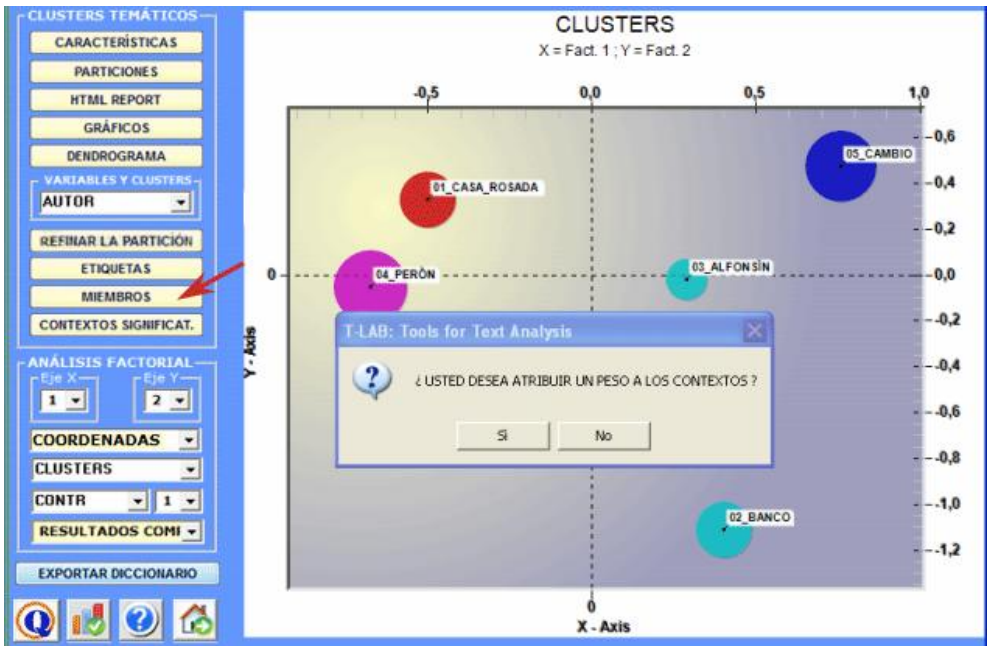
6 - Asignar etiquetas a los clusters

Una función de **T-LAB** permite atribuir etiquetas a los clusters.
(Nota: Al primer uso, algunas de las etiquetas son asignadas automáticamente por el programa).



Las etiquetas atribuidas a los distintos clusters pueden ser visualizadas en los distintos gráficos disponibles (ver a continuación).

7 - Verificar qué contextos elementales pertenecen a qué clusters;
8 - Verificar el peso de cada uno de los contextos elementales dentro del cluster al que pertenece;
9 - Obtener una clasificación temática de los documentos.



De hecho el botón **Miembros** permite exportar tres tipos de tablas en formato MS Excel:

a - " Cluster_Partitions.xls " (véase abajo) con todas las correspondencias de unidad de contexto x cluster en el interior de las distintas particiones;

(IDNUMBE	PART-2	PART-3	PART-4	PART-5	PART-6	PART-7	PART-8	PART-9	PART-10
1	1	1	1	1	1	1	1	1	1
2	1	1	4	4	6	6	6	9	9
3	1	1	4	4	6	6	6	9	9
4	2	3	3	3	3	3	3	3	3
5	2	2	2	2	2	2	2	2	2
6	2	2	2	2	2	2	2	2	2
7	2	3	3	5	5	7	7	7	10
8	2	2	2	2	2	2	2	2	2
9	2	2	2	2	2	2	2	2	2
10	1	1	4	4	6	6	6	9	9
11	1	1	4	4	6	6	6	9	9
12	1	1	1	1	1	1	1	1	1
13	2	2	2	2	2	2	8	8	8
14	2	3	3	5	5	5	5	5	5
15	2	3	3	5	5	7	7	7	10
16	1	1	1	1	1	1	1	1	1
17	2	3	3	3	3	3	3	3	3
18	1	1	1	1	1	1	1	1	1
19	2	3	3	5	5	7	7	7	7
20	2	3	3	5	5	7	7	7	10
21	2	3	3	5	5	5	5	5	5
22	2	3	3	5	5	7	7	7	10
23	2	3	3	5	5	7	7	7	10
24	2	3	3	5	5	7	7	7	10
25	2	3	3	3	3	3	3	3	3
26	1	1	1	1	1	1	1	1	1
27	2	3	3	5	5	5	5	5	5
28	2	2	2	2	2	2	2	2	2
29	2	3	3	5	5	5	5	5	5
30	2	3	3	5	5	5	5	5	5
31	2	3	3	5	5	5	5	5	5

b - Themes-Contexts.xls (véase abajo) con las correspondencias de unidad de contexto x cluster en el interior de la partición seleccionada.

(IDNUMBER)	THEME	SCORI	CONTEXT
00001000001	01_CASA_ROSADA	0,57	Los argentinos ya no tienen libre acceso a su dinero . El Gobierno determina
00001000002	01_CASA_ROSADA	1,47	Hace millones de años , cuando en el siglo pasado vivía en este país , me en
00001000003	03_ALFONSÍN	2,04	que regresaban al continente exhaustos de combustible tras atacar con sus c
00001000004	03_ALFONSÍN	2,98	Nada ni nadie ; soledad total y vientos obsesivos . Ni las ovejas eran visibles
00001000006	05_CAMBIO	0,37	Había sentido que tras aquel horizonte sólo había un vacío vertical dispuesto a
00001000007	04_PERÓN	0,1	en Buenos Aires , los argentinos , ensopados de sudor por el verano y la hur
00001000008	04_PERÓN	0,72	En el aeropuerto internacional de Ezeiza ya no te espera la policía militar con
00001000009	04_PERÓN	0,79	Parece una broma , porque aquí lo_que pagas en dólares te lo devuelven en p
00001000010	04_PERÓN	0,2	saqueados por la devaluación y encerrados en un ' ' corralazo ' ' (más_q
00001000011	01_CASA_ROSADA	0,57	Ni conozco otra situación parecida , excepto en economías de guerra , en do
00001000012	05_CAMBIO	0,08	El país de los alimentos no está hambreado como sugieren los asaltos a supe
00001000013	02_BANCO	4,16	Un kilo de pan en buenos_aires CF se pone en 250 pesetas ; uno de papas e
00001000014	01_CASA_ROSADA	2,75	Claro que el problema es cobrar tu salario . Es una nación devorada por el clie
00001000015	04_PERÓN	6,31	Asidos todos a las ubres del Estado o de las gobernaciones provinciales , los
00001000016	04_PERÓN	2,68	TELEFÓNICA SABRÁ La otra cara de la moneda revela la corrupción institucio
00001000017	05_CAMBIO	0,61	por lo incontable de su fortuna , hecha en la política de San Luis , o el mismo
00001000019	01_CASA_ROSADA	0,6	Argentina está quebrada , pero habiendo tenido acceso como periodista a ma
00001000020	02_BANCO	0,6	penthouse (un triplex es cosa de arribistas sociales) en el metro cuadrado m
00001000021	02_BANCO	1,44	Por supuesto que según qué huéspedes se habla mejor inglés que español y s
00001000023	04_PERÓN	2,67	En uno de esos salones Guido di Tella , de una de las familias patricias del pa
00001000024	01_CASA_ROSADA	1,47	Yo y otros periodistas extranjeros presentes estuvimos en un tris de levantarn
00001000025	04_PERÓN	7,49	Y su esposa Hilda ' ' Chiche ' ' Duhalde gastaba 250 millones de dólares a
00001000026	04_PERÓN	4,92	Ahora La Chiche quiere resucitar a Evita , firmó como aquélla el acta presiden
00001000027	01_CASA_ROSADA	11,99	El dimisionario de la Rúa (en realidad dio una espantada) , representante d
00001000028	03_ALFONSÍN	8,67	La medida yuguló la hiperinflación causante de la derrota a Raúl Alfonsín , y ,
00001000029	01_CASA_ROSADA	3,07	Mingo Cavallo , un monetarista neoliberal formado en Estados Unidos y no se
00001000030	01_CASA_ROSADA	2,48	derogar la ley federal que , lógicamente , hacía intocables los depósitos banc
00001000031	04_PERÓN	1,31	UN TIRO EN EL by_pass El suicidio sucedió hace unos meses y debió haberr
00001000032	03_ALFONSÍN	4,54	Lo rechazó todo en Estados Unidos para regresar a Argentina y hacer país .
00001000033	04_PERÓN	0,72	A la caída de la dictadura militar daba charlas televisadas sobre civismo , aust
00001000034	04_PERÓN	0	Creó una fundación con su nombre , desahuciada por los impagos federales y
00001000035	02_BANCO	0,29	Viudo y sin hijos , tenía buena salud , era respetado y hasta tenía una novia
00001000036	04_PERÓN	3,03	Julio César Strassera , quien fuera Fiscal General de la República en la demo

En particular, el valor de importancia (score) asignado a cada contexto elemental (j) que pertenece al racimo (k) viene de la fórmula siguiente

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

Donde:

Score$_j$ = valor de la importancia asignado al contexto elemental (j);
ΣX_{ij} = suma de los valores del Chi-cuadrado asignados a las palabras clave (i) encontradas en el contexto elemental (j) y que son típicas del racimo (k);
n_j = total de palabras clave (palabras distintas), típicas del cluster (k), encontradas en el contexto elemental (j);
N = total de las palabras clave (palabras distintas) típicas del cluster (k).

c - " Ec_Document_Classification.xls " (proporcionado solamente cuando el cluster se compone por lo menos de 2 documentos primarios y éstos no son textos breves como las respuestas a preguntas abiertas) enumera las pertenencias mezcladas de cada documento (véase abajo).

DOC_ID	VAR_01	BEST_CL	CLUST_1	CLUST_2	CLUST_3	CLUST_4	CLUST_5
1	AU_PRIE1	4	0,164	0,049	0,326	0,455	0,005
2	AU_PRIE2	1	0,443	0,273	0,196	0,066	0,022
3	AU_CAVA	5	0,019	0	0	0	0,981
4	AU_PRIE3	3	0,218	0,173	0,513	0,025	0,072
5	AU_DEAR	2	0,002	0,675	0,164	0,046	0,113
6	AU_PRIE4	2	0,064	0,494	0,202	0,184	0,056
7	AU_DRAG	1	0,685	0,095	0,047	0,055	0,118
8	AU_GARC	2	0,016	0,905	0,031	0,04	0,008
9	AU_GEOR	3	0	0,099	0,455	0,064	0,382
10	AU_GIARD	1	0,459	0,018	0,342	0,039	0,142
11	AU_LUMH	5	0	0	0,07	0,026	0,905
12	AU_MENE	5	0,248	0,092	0,021	0,005	0,633
13	AU_MONT	4	0,434	0	0	0,562	0,004
14	AU_RUES	5	0	0,171	0	0	0,829
15	AU_LEPO	2	0,118	0,761	0,082	0,033	0,005

En este caso los valores derivan de la fórmula antedicha (véase "b") sumando los scores de los contextos elementales que pertenecen a cada documento y aplicando un cálculo de porcentaje.

10 - Archivar la partición seleccionada para explorarla con otras herramientas T-LAB

A la salida de la función de **Análisis temático de los Contextos elementales**, algunos mensajes recuerdan que es posible explorar la partición seleccionada con otras herramientas **T-LAB**.

Seleccionando la opción **Guardar**, la variable < **CONT_CLUST** > (cluster de contextos elementales) queda disponible sólo en algunos tipos de análisis (por ej. Secuencias de Temas, Asociaciones de Palabras, Comparación entre Parejas, Análisis de Co-Palabras) y hasta que:

- la misma función (**Análisis temático de los Contextos elementales**) se selecciona para un nuevo uso;
- el usuario modifica su lista de palabras clave.

11 - Exportar un diccionario de las categorías

Cuando se selecciona esta opción, **T-LAB** genera dos archivos:
a) un archivo diccionario, con extensión .dictio, que puede ser importado directamente a través de una de las herramientas disponibles para el análisis temático. En dicho diccionario, cada clúster corresponde a una categoría descrita mediante sus palabras características, es decir, mediante todas las palabras de este clúster que presentan un **Chi-Cuadrado** significativo;
b) un archivo **MyList.diz** listo para la importación mediante la función 'Ajustes personalizados'. Este archivo contiene el listado, ordenado alfabéticamente, de todas las palabras que presentan un valor significativo de Chi cuadrado, es decir, de todas aquellas palabras que determinan las diferencias entre clústeres temáticos. Así pues, su uso permite repetir determinados análisis siguiendo una perspectiva aún más selectiva y discriminante.

12 – Verificar la calidad de la partición elegida y la coherencia semántica entre los diferentes temas

Al hacer clic sobre el icono 'Índices de Calidad' (véase arriba), **T-LAB** genera un archivo HTML que contiene diferentes medidas.
Las primeras de ellas se refieren a la calidad de las particiones en 'k' clústeres. Esto es, por ejemplo, el cociente entre varianza externa e interna.
Otro conjunto de medidas se refiere a la 'coherencia semántica' de cada clúster, es decir, las semejanzas entre las 10 primeras palabras características de cada tema.
Más en concreto:
- Las primeras 10 palabras son aquellas caracterizadas por un valor de probabilidad más alto
- las medidas de semejanza están calculadas con base en el coeficiente del coseno;
-Al igual que para la herramienta '**Asociación de Palabras**', el coeficiente del coseno se calcula verificando las co-ocurrencias de las palabras contenidas en los segmentos de texto definidos como contextos elementales.

13 – Explorar secuencias de temas

Al contrario de la herramienta 'secuencias de temas', incluida en un submenú **T-LAB** de análisis de las co-ocurrencias, esta opción ha sido generada específicamente para integrar el análisis temático de los contextos elementales. Más en concreto, su uso adquiere sentido sólo cuando el corpus entero se considera como un discurso y/o cuando sus diferentes secciones

(por ejemplo: capítulos de libro, partes de una entrevista, intervenciones de diferentes participantes en una conversación o en un debate, etc.) se alternan siguiendo un preciso orden temporal..

En este caso, las relaciones analizadas son aquellas que se instauran entre contextos elementales (hasta un máximo de 100.000) a lo largo de la cadena lineal del corpus. Cada uno de ellos - tanto si son 'predecesores' como si son 'sucesores' - viene tratado como una unidad de análisis que pertenece a un clúster temático (o no clasificado).

Todos los resultados proporcionados permiten al usuario explorar las relaciones secuenciales entre 'temas', bien de forma 'estática' o bien de forma 'dinámica'. Más en concreto, el usuario puede verificar cuándo las personas abarcan temas específicos (véanse, por ejemplo, en las imágenes a continuación, los puntos presentes en la diagonal de las matrices) y cuándo pasan de un tema central a otro. Todo ello, contemplando la dinámica temporal de las secuencias a través de gráficos animados.

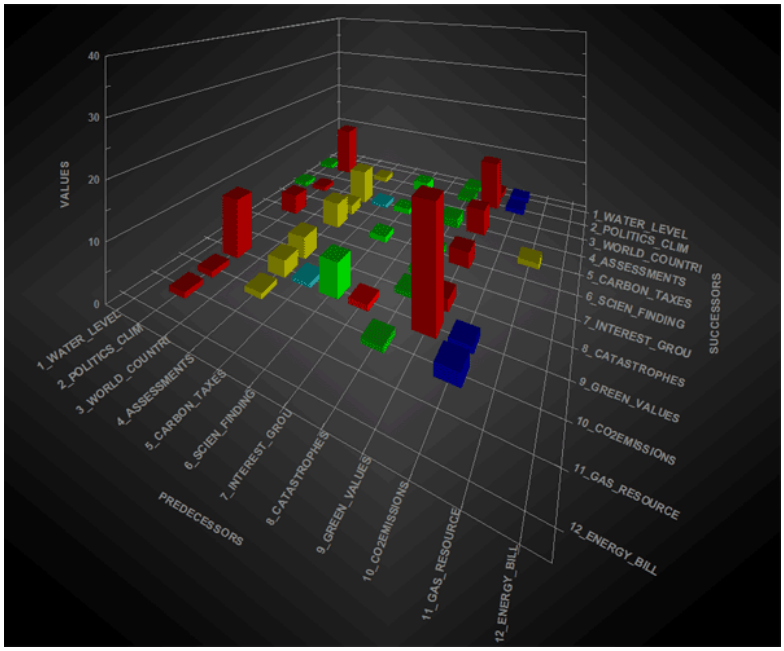
A continuación se proporciona, paso a paso, una breve descripción de las diferentes opciones disponibles.

(N.B.: Todos los resultados contenidos en el ejemplo se han obtenido a partir de un análisis temático del libro 'The Politics of Climate Change " de Antony Giddens publicado en el sitio web de T-LAB).

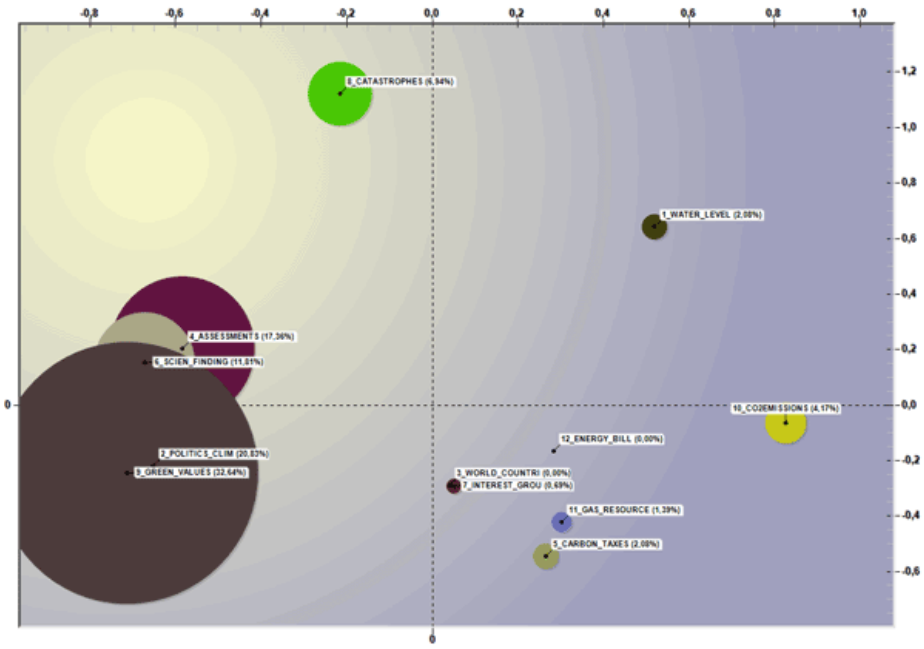
Una vez habilitado el botón ‘Secuencias de Temas', cliqueando el mismo se vuelve visible y activo el siguiente 'player'.

La opción '1' (véase arriba) hace referencia a la tipología de gráfico escogida para la visualización de las secuencias, tanto dentro del corpus entero como dentro de una parte del mismo (véase arriba la opción '2').

La opción 'matriz' devuelve un grafico en 3D que sintetiza las relaciones entre predecesores y sucesores mediante barras de colores ubicadas en los respectivos cruces. Respecto a los gráficos animados en 3D, cabe destacar que el aumento de longitud de las barras implica el aumento del número de ocurrencias dentro de las secuencias correspondientes (véanse relaciones binarias entre 'predecesores' y 'sucesores' en el grafico siguiente).



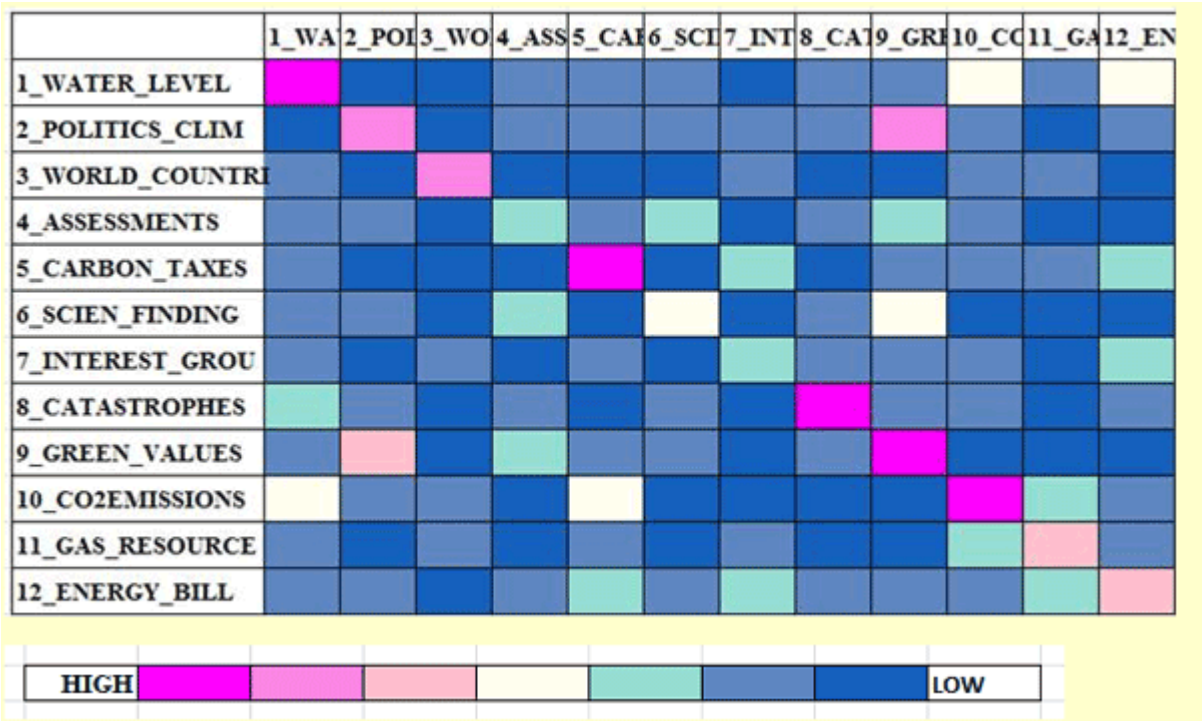
La opción 'espacio' genera un grafico en 2D en el que las dimensiones (es decir, los porcentajes) y las relaciones entre grupos temáticos están representadas en un plano compuesto por dos ejes factoriales escogidos por el usuario. En este caso, a la hora de visualizar gráficos animados, las dimensiones de las 'burbujas' - que vienen constantemente ajustadas a un total del 100% - indican cómo los porcentajes relativos a cada clúster varían en el tiempo. Al mismo tiempo, el movimiento de las flechas indica el orden según el cual se van alternando los temas.



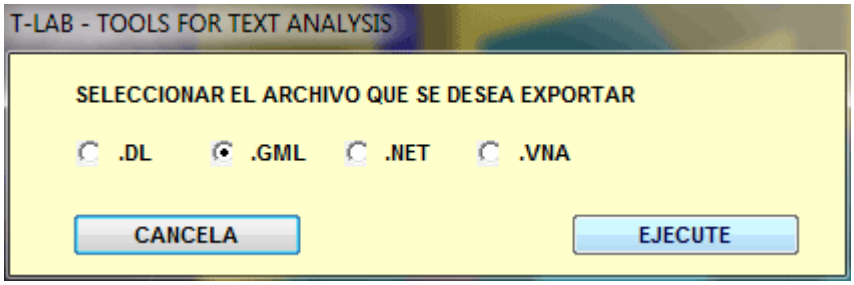
En cada una de las situaciones recién descritas es posible, tras haber parado el video (véase botón 'pausa'), visualizar dos resultados ulteriores:

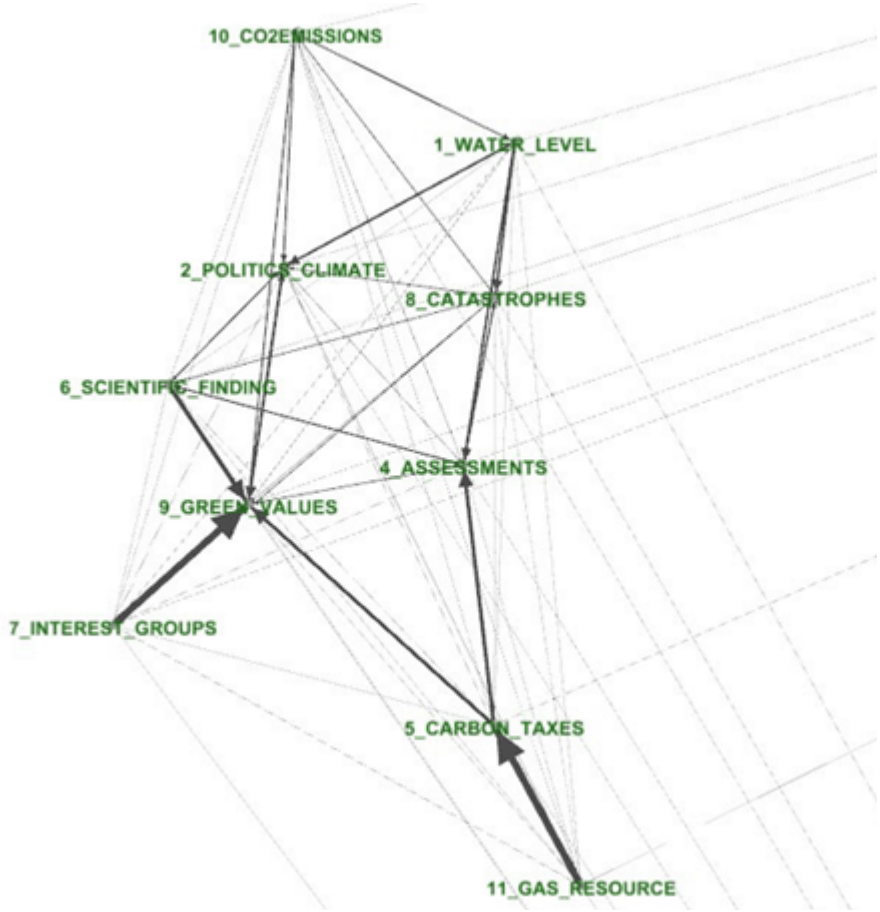
A - tablas html que resumen las relaciones entre predecesores y sucesores (véase abajo);

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



B - archivos gráficos que pueden ser importados por programas para el análisis de redes.



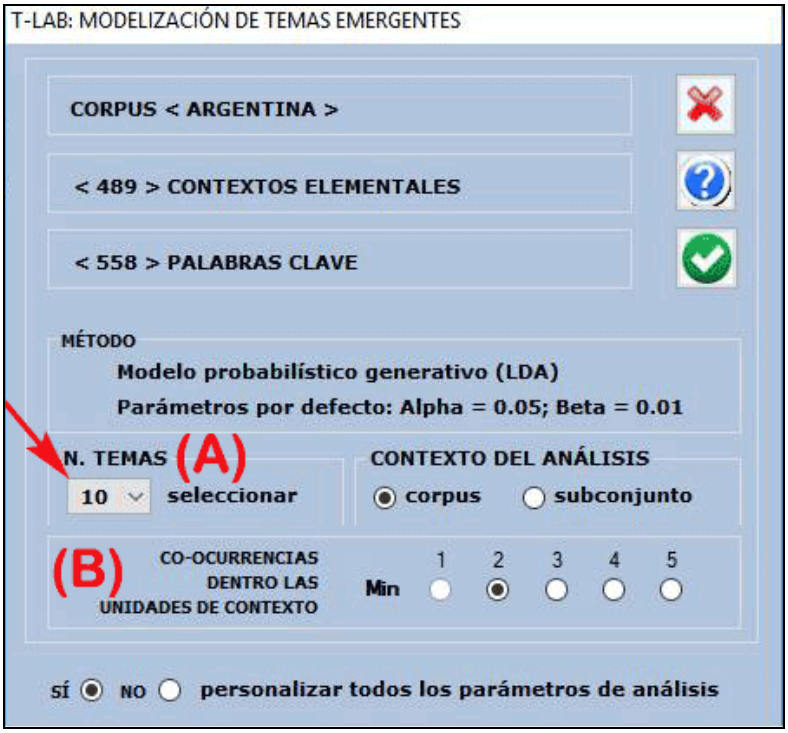


N.B.: El grafico anterior, que hace referencia al tercer capitulo del libro de Giddens, ha sido creado tramite el programa Gephi (véase https://gephi.org/).

Modelización de Temas Emergentes

Este instrumento **T-LAB** permite **individualizar, analizar y modelizar los principales temas que emergen de los textos** y, consecuentemente, utilizarlos en ulteriores análisis, tanto de tipo cualitativo como de tipo cuantitativo.

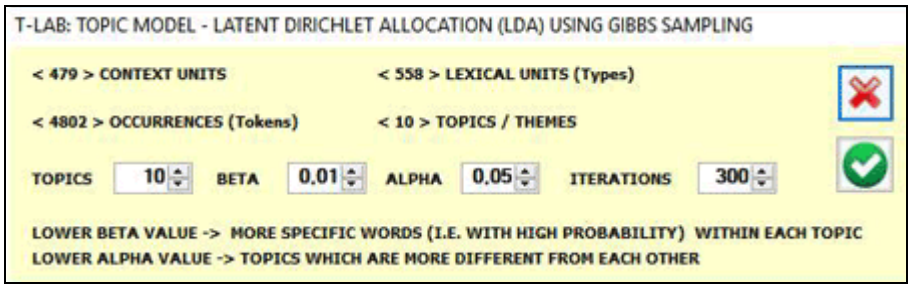
Los temas emergentes - que están descritos a través de sus vocabulario característico, es decir a través de un conjunto de palabras clave que se presentan en coocurrencia en las unidades de contextos examinados - pueden ser utilizados para **clasificar** estas unidades (tanto documentos como contextos elementales) y **obtener nuevas variables** utilizables en nuevas análisis **T-LAB.**



Un cuadro de diálogo **T-LAB** (véase arriba) permite que el usuario fije dos parámetros de análisis.
En particular:

- el parámetro (A) permite establecer el número de temas que se obtendrán. (Tenga en cuenta que cuanto mayor sea este número, más coherentes serán las relaciones de co-ocurrencia dentro de cada tema, y si es necesario, algunos temas - por ejemplo, los que son redundantes o difíciles de interpretar - pueden ser eliminados en un segundo momento a través de una funcionalidad específica del instrumento en examen);

- el parámetro (B) permite excluir del análisis cualquier unidad de contexto que no contenga un número mínimo de palabras clave incluidas en la lista utilizada.
Solo cuando usted elija personalizar todos los parámetros de análisis (véase la opción 'Sí' arriba), se mostrará la ventana siguiente y habrá más opciones disponibles. (Tenga en cuenta que en la siguiente imagen el número de unidades de contexto está determinado por el parámetro "B" mencionado anteriormente).



El **proceso automático de análisis** sigue los siguientes pasos:

a – construcción de una matriz documentos por palabras, donde los documentos son siempre contextos elementales que corresponden a las unidades de contexto (es decir, fragmentos, frases, párrafos) en los que se ha subdividido el corpus;

b – análisis de datos a través un modelo probabilístico que usa la Latent Dirichlet Allocation y el Gibbs Sampling (para más información se pueden consultar las siguientes Web de Wikipedia:
http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation;
http://en.wikipedia.org/wiki/Gibbs_sampling);
c – descripción de cada tema a través de los valores de probabilidades asociados a sus palabras características, tanto “específicas” como “compartidas” por uno o más temas.

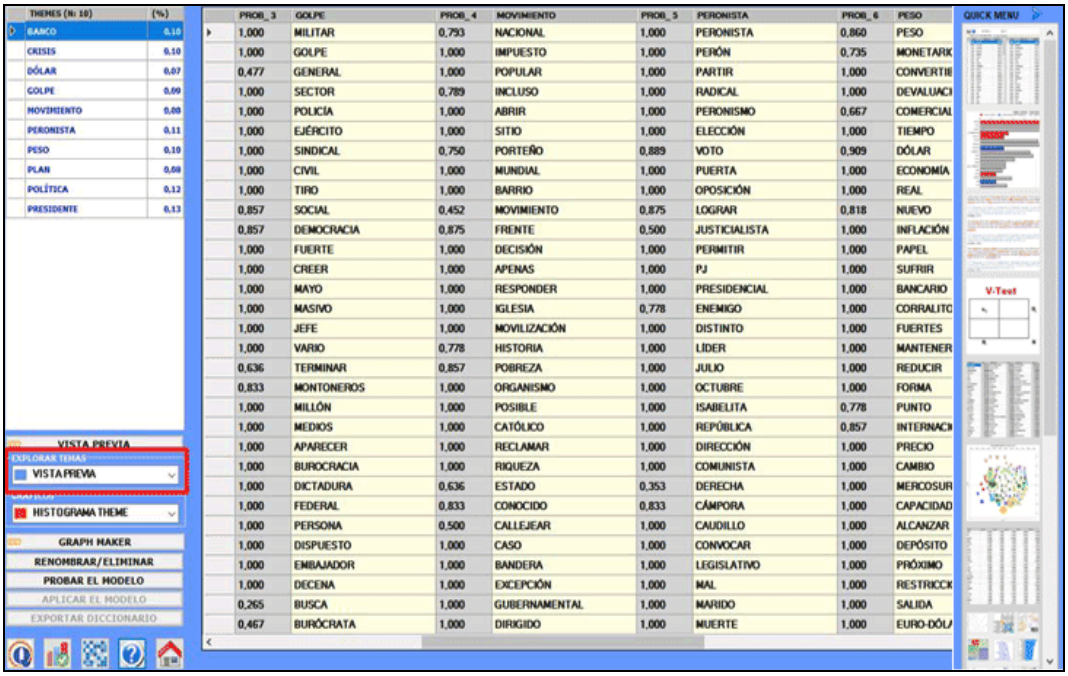
Al final del proceso de análisis, el usuario puede fácilmente efectuar las siguientes operaciones:

1 – explorar las características de cada tema;
2 – explorar las relaciones entre los diversos temas;
3 – renombrar o eliminar temas específicos;
4 - verificar la coherencia semántica entre los diferentes temas;
5 – probar el modelo y asignar los temas a las unidades del contexto, tanto documentos como contextos elementales;
6 – aplicar el modelo y crear una nueva variable temática, cuyos valores son los temas elegidos**;**
7– exportar un diccionario de las categorías, que se puede utilizar en un análisis posterior.

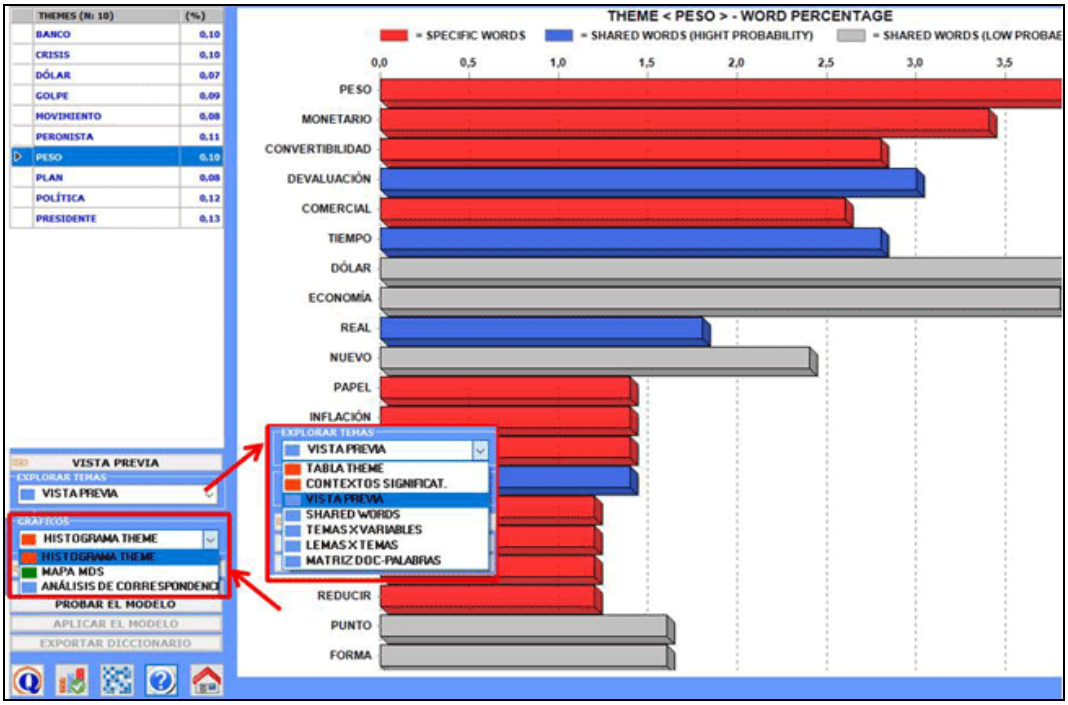
En el detalle:

1 – Explorar las características de cada tema

El primer resultado que se puede consultar y guardar consiste en una tabla con una 'Vista previa' de todos los temas. Y, cuando se desee, se puede volver a acceder fácilmente utilizando el botón correspondiente (ver a a continuación).

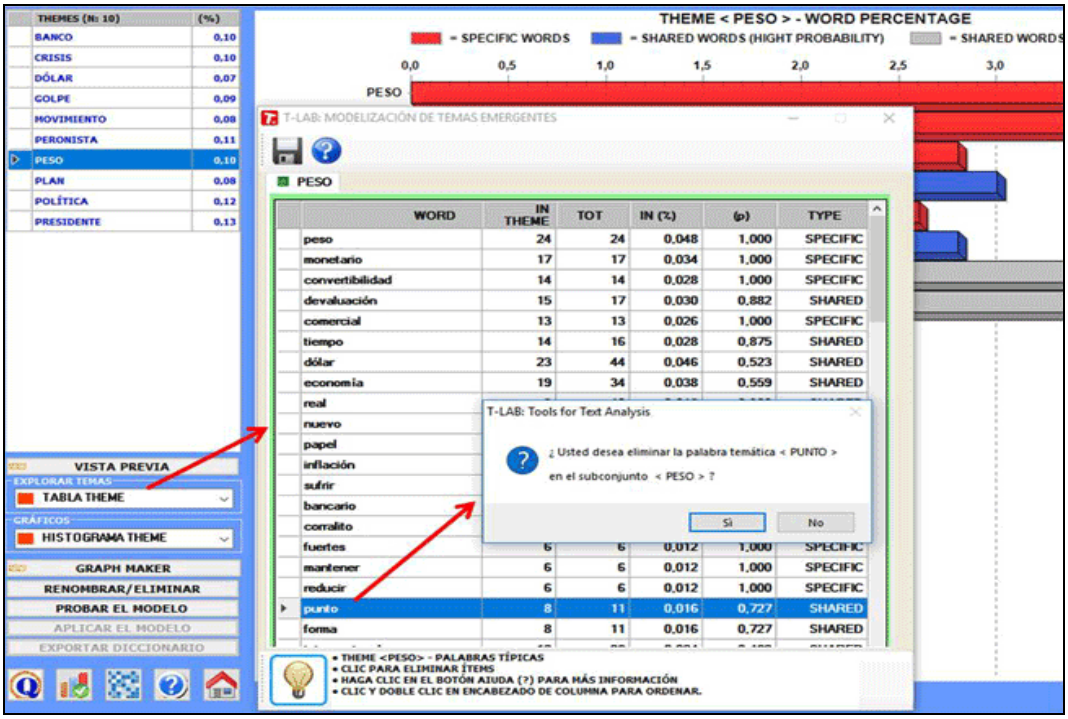


Además se puede acceder a otros tipos de resultados eligiendo una de las opciones resaltadas en la imagen siguiente.



NOTA: En este gráfico “hight probability” indica una probabilidad >=.75.

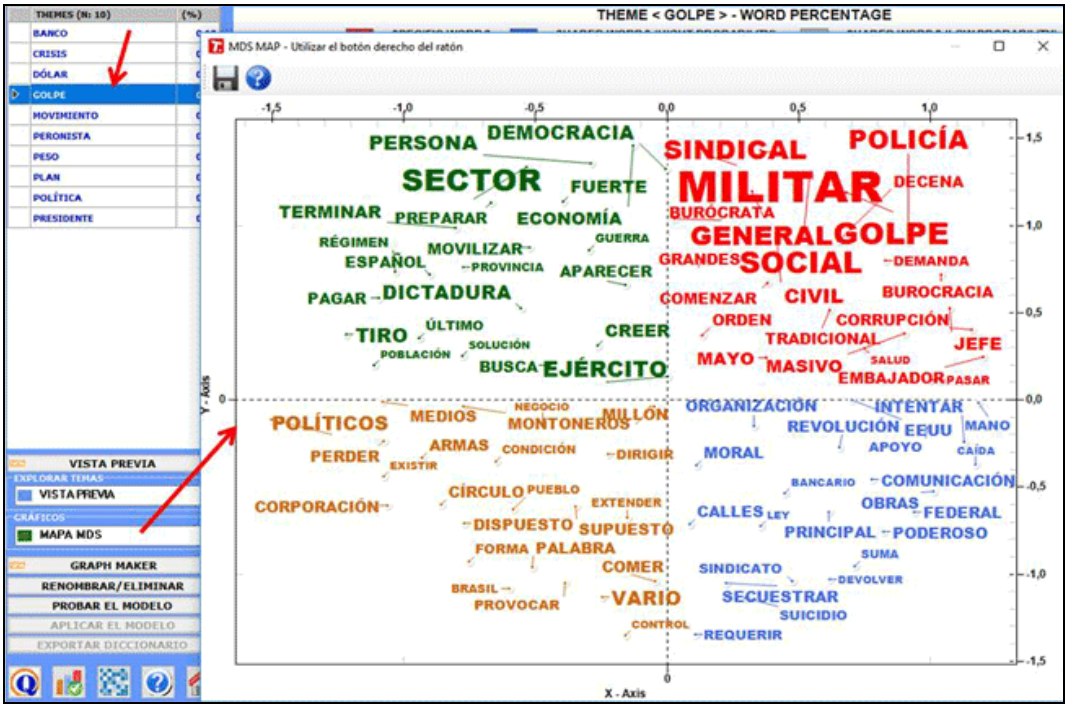
Cuando se selecciona un tema, al hacer clic en la opción "Tabla Theme", se pueden verificar sus características; además, al hacer clic en cualquier palabra de la tabla que se muestra, parece una opción adicional que permite "eliminar" el elemento seleccionado (ver imagen a continuación).



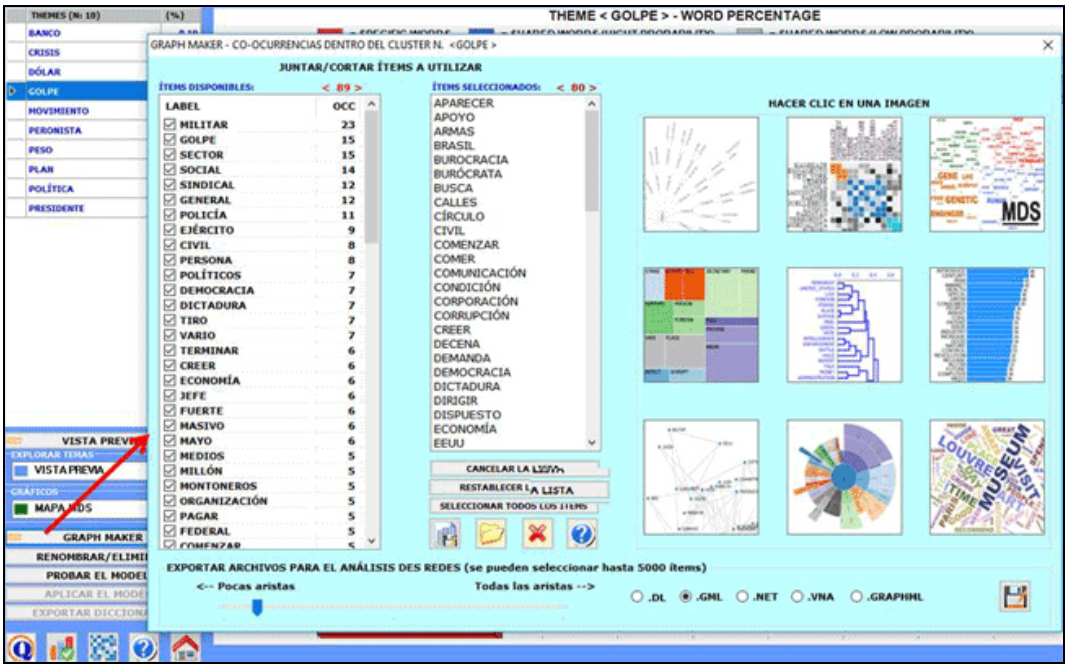
Las claves de lectura de la tabla anterior son las siguientes:

IN THEME = ocurrencias (tokens) de cada palabra dentro del tema seleccionado;
TOT = ocurrencias (tokens) de cada palabra dentro del corpus o del subconjunto analizado;
IN (%) = peso porcentaje de cada palabra dentro del tema seleccionado;
(p) = valor de probabilidad asociado a cada relación palabra x tema;
TYPE = marcado como "specific" cuando la palabra (con p = 1) pertenece solo al tema seleccionado y como "shared" en todos los otros casos (es decir cuando la palabra es presente, en maneras diferentes, en mas de un tema).

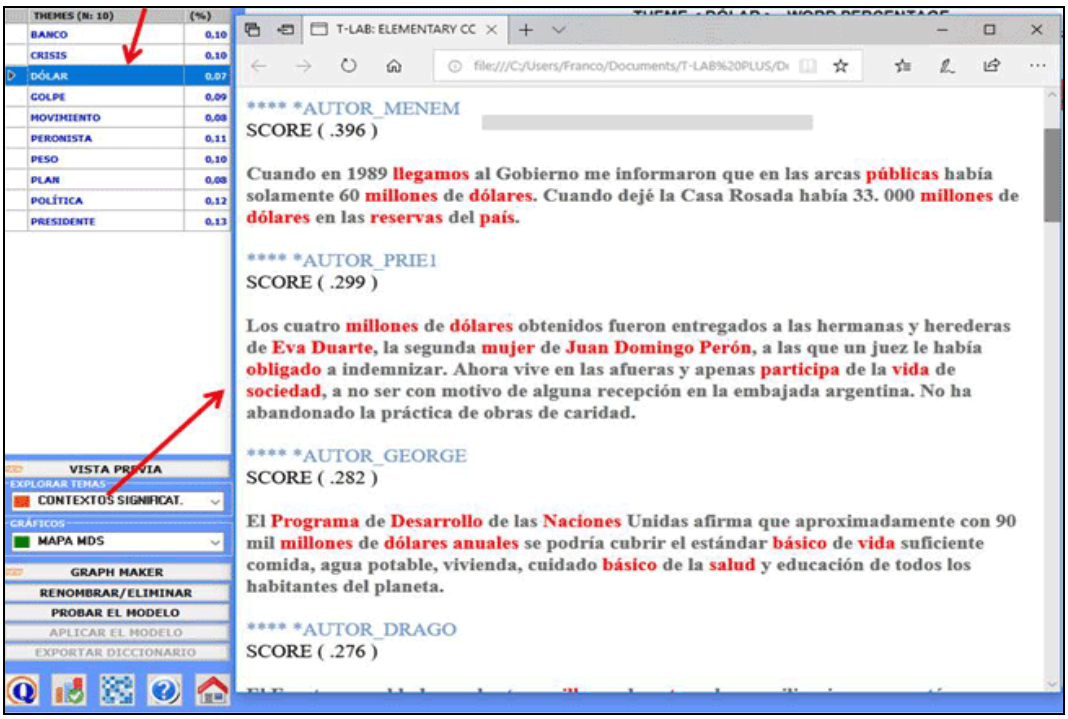
Cuando se selecciona un tema, al hacer clic en la opción "Mapa MDS" se pueden explorar fácilmente las relaciones semánticas entre las palabras que son más características (ver la imagen siguiente).



Además, utilizando la herramienta 'Graph Maker', algunas opciones gráficas adicionales están disponibles (ver las imágenes siguientes).



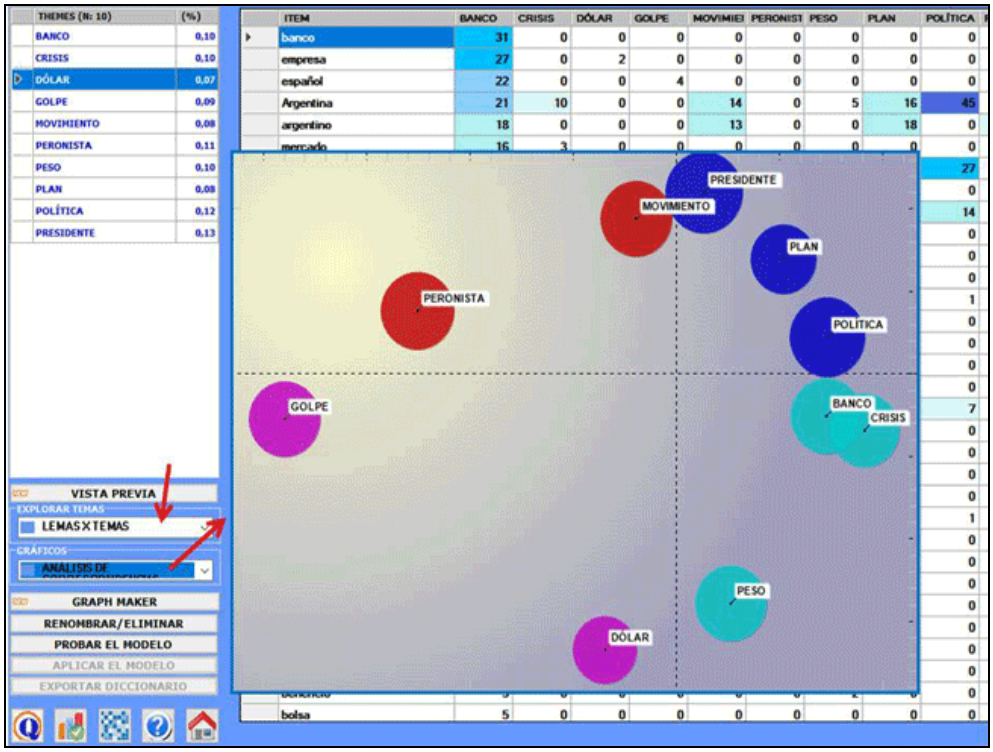
Cuando se selecciona un tema, al hacer clic en la opción "contextos significativos", se crea un archivo HTML donde se muestran los 20 segmentos de texto principales, que se corresponden más con las características del tema (ver la imagen siguiente).



2 – Explorar las relaciones entre los diversos temas

Usando la herramienta **Análisis de Correspondencia,** se pueden crear y explorar dos tipos de tablas de contingencia:

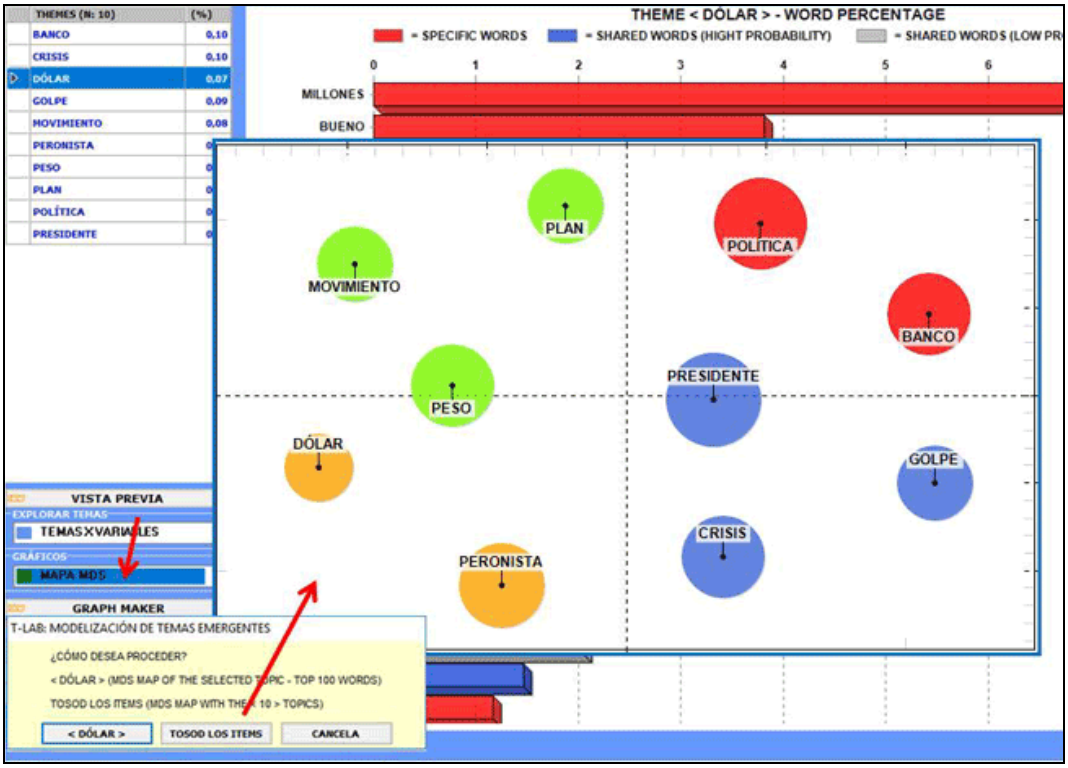
2.1) una tabla palabras por tema (ver abajo)



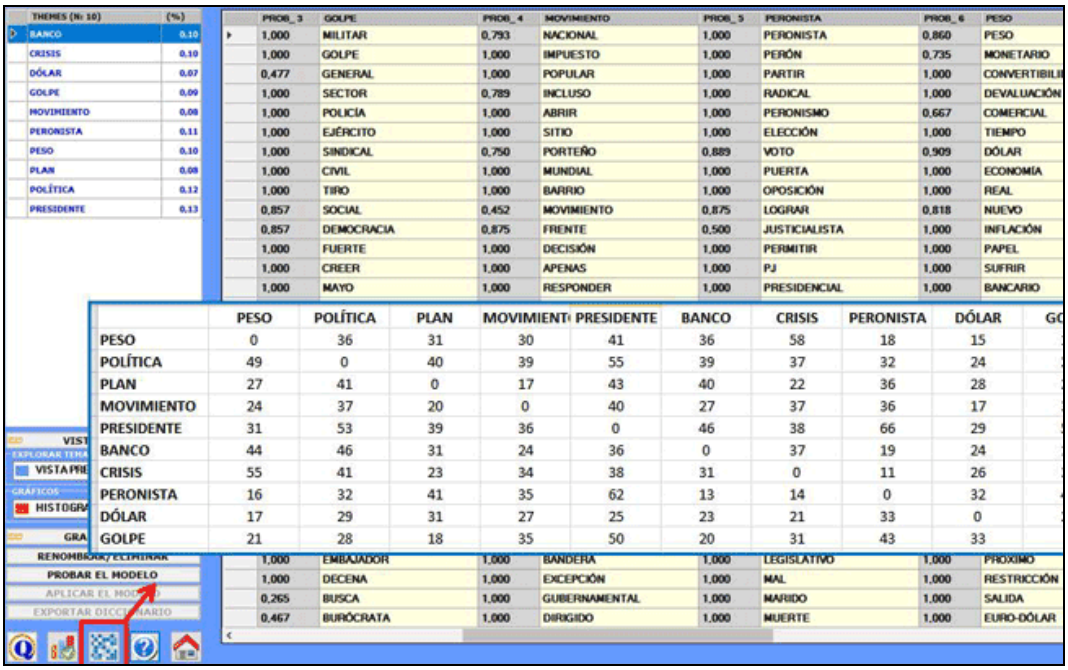
2.2) una tabla que cruza los temas con las modalidades de la variable seleccionada

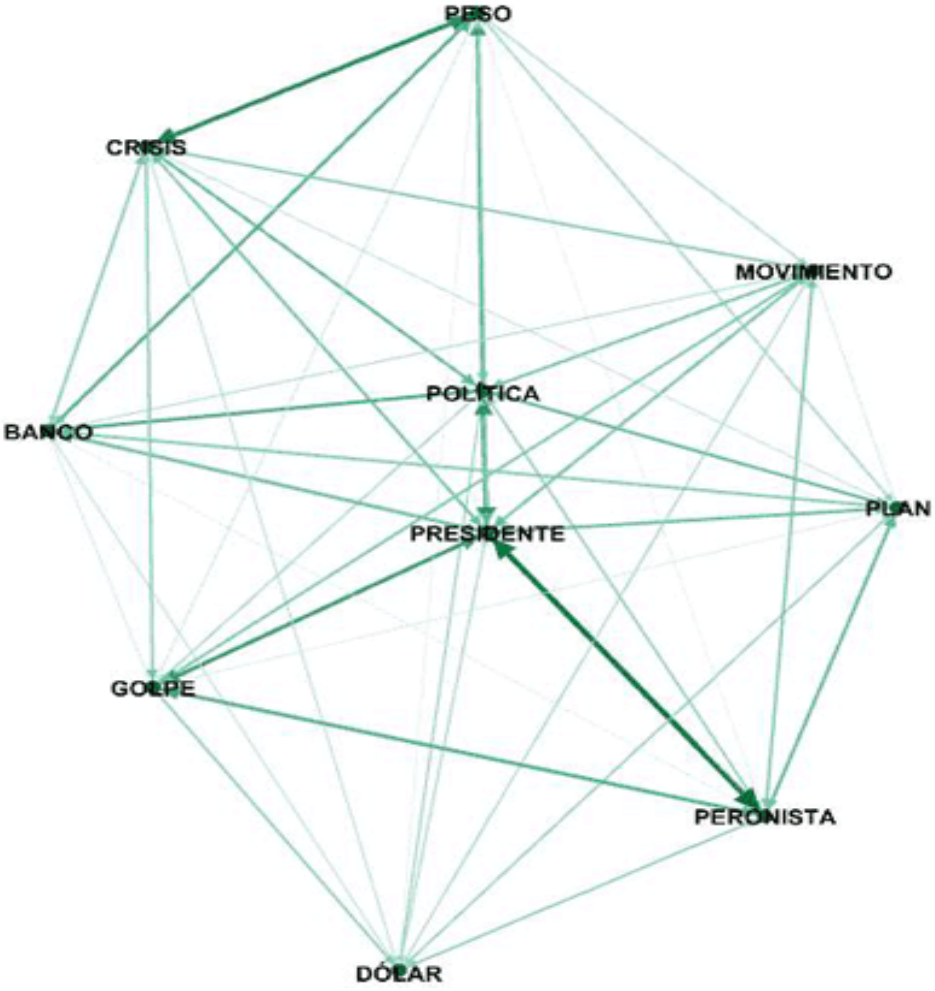
También hay otras dos opciones gráficas disponibles que nos permiten mapear las relaciones entre los diversos temas:

2.3) un mapa MDS



2.4) un gráfico de red obtenido al exportar / importar la tabla de adyacencia creada por **T-LAB** (ver a continuación)



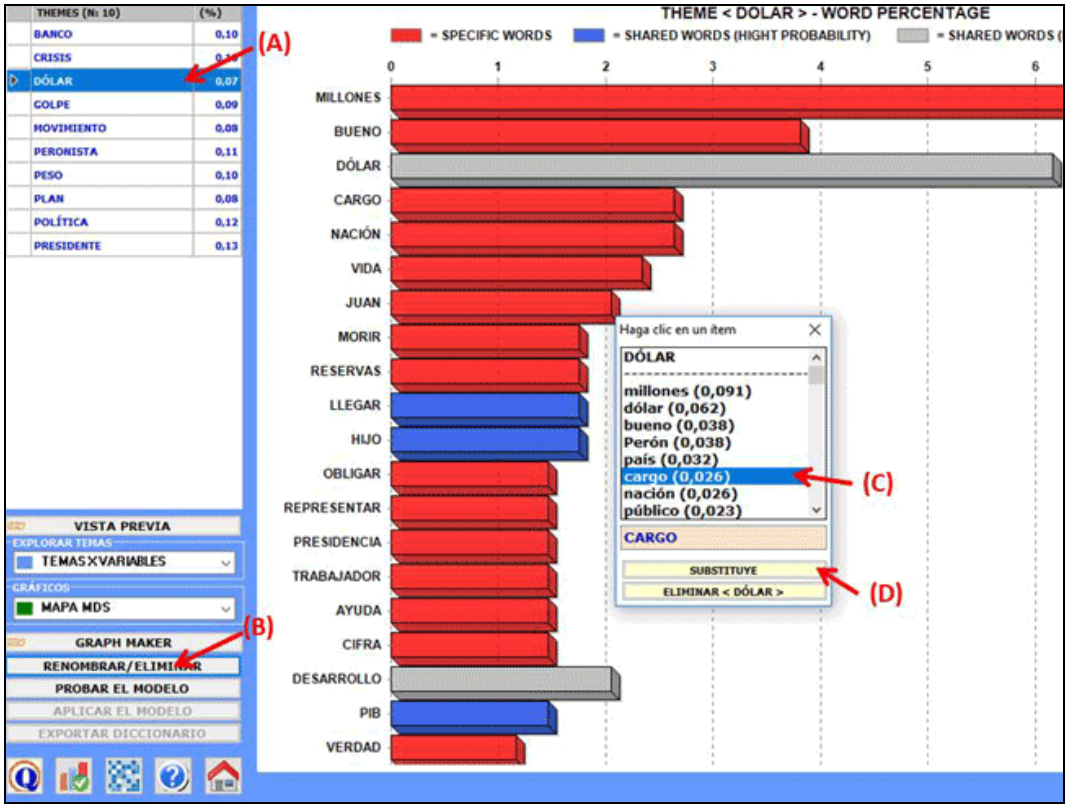


NOTA: El gráfico anterior se ha creado por medio del programa Gephi (**https://gephi.org/**), después haber importado una tabla creada por **T-LAB**.

3 – Renombrar o eliminar temas específicos

Para renombrar o eliminar temas específicos es suficiente seleccionar los ítems correspondientes (ver "A" en el cuadro siguiente) y pinchar sobre el botón "renombrar/eliminar" (ver "B" en el cuadro siguiente).

Cuando aparece el cuadro con las varias opciones, el usuario puede, según su objetivos, cambiar la etiqueta del tema (tanto eligiendo entre las palabras disponibles como tecleando una nueva palabra; ver "C" en el cuadro siguiente) o eliminar el tema seleccionado pinchando sobre el botón correspondiente (ver "D" en el cuadro siguiente).



4 – Verificar la coherencia semántica entre los diferentes temas

Al hacer clic sobre el icono 'Índices de Calidad' (véase arriba), **T-LAB** calcula las semejanzas entre las primeras 10 palabras características de cada tema (top 10).

Más en concreto:

- Las primeras 10 palabras son aquellas caracterizadas por un valor de probabilidad más alto
- las medidas de semejanza están calculadas con base en el coeficiente del coseno;
-Al igual que para la herramienta 'Asociación de Palabras', el coeficiente del coseno se calcula verificando las co-ocurrencias de las palabras contenidas en los segmentos de texto definidos como contextos elementales.

Come resultado**, T-LAB** genera un archivo HTML en el cual los 'k' temas están recogidos en un listado y van asociados a sus respectivos índices de 'coherencia semántica'.

NOTA: Las medidas de semejanza varían en función de los cambios en las palabras seleccionadas. Por ello, se recomienda repetir el procedimiento cada vez que alguna de las diez palabras asociadas a un tema haya sido eliminada por el usuario.

5 –**Probar el modelo y asignar los temas a las unidades del contexto**

Al final del análisis de los datos (ver los puntos "a" y "b" del proceso de análisis) cada unidad de contexto (por ejemplo un documento o un contexto elemental) resulta constituido como una mixtura de temas. De otra manera, el proceso de clasificación utilizado para probar/aplicar el modelo asocia cada unidad de contexto al tema que mas lo caracteriza. Como resultado, en esta fase, cada tema se pone de hecho como un clúster de unidad de contexto.

Por esa razón, cuando se selecciona la opción "Probar el Modelo", **T-LAB** produce dos archivos XLS (ver abajo) que permiten a el usuario de verificar la pertenencia de cada unidad de contexto a un tema específico.

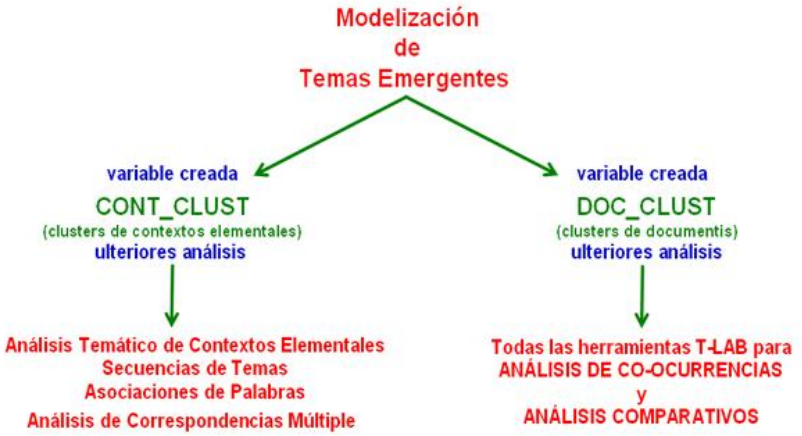
ID_DOC	BEST	BANCO	CRISIS	DÓLAR	GOLPE	MOVIMIENT	PERONISTA	PESO	PLAN	POLÍTICA	PRESIDENTE
1	3	1,379	0,275	2,015	1,075	0,425	1,241	1,323	1,172	1,289	1,445
2	10	0,46	0,12	0,399	0,356	0,157	0,48	0,277	0,377	0,685	1,563
3	7	0,289	1,028	0,196	0,092	0,145	0,036	2,014	0,029	0,507	0,144
4	10	0,536	0,371	0,082	0,396	0,533	0,245	0,493	0,485	0,611	0,981
5	9	2,275	1,02	0,877	0,104	0,494	0,049	0,453	0,711	3,034	0,401
6	3	0,462	0,165	0,789	0,302	0,558	0,19	0,196	0,446	0,359	0,362
7	5	0,113	0,542	0,677	0,382	1,31	0,854	0,399	0,838	1,299	1,245
8	1	2,864	0,693	0,73	0,373	0,071	0,15	0,433	1,061	0,26	0,734
9	3	0,413	0,564	1,006	0,177	0,283	0	0,325	0,414	0,842	0
10	9	0,225	0,151	0,155	0,429	0,612	0,15	0,226	0,267	1,252	1,083
11	7	0,213	0,548	0,091	0	0,108	0,021	0,894	0,181	0,667	0,015
12	9	0,322	0,942	0,674	0,068	0,437	0,161	0,503	0,49	1,156	0,505
13	6	0,239	0,88	1,141	3,219	1,032	7,459	0,253	1,206	0,922	1,229
14	2	0,951	2,06	0,032	0,113	0,032	0,017	1,555	0,216	0,859	0,19
15	8	0,206	0,127	0,205	0,271	0,425	0,03	0,038	0,468	0,109	0,036

NOTA: En la tabla anterior cada documento tiene un valor de probabilidad asociado con cada tema.

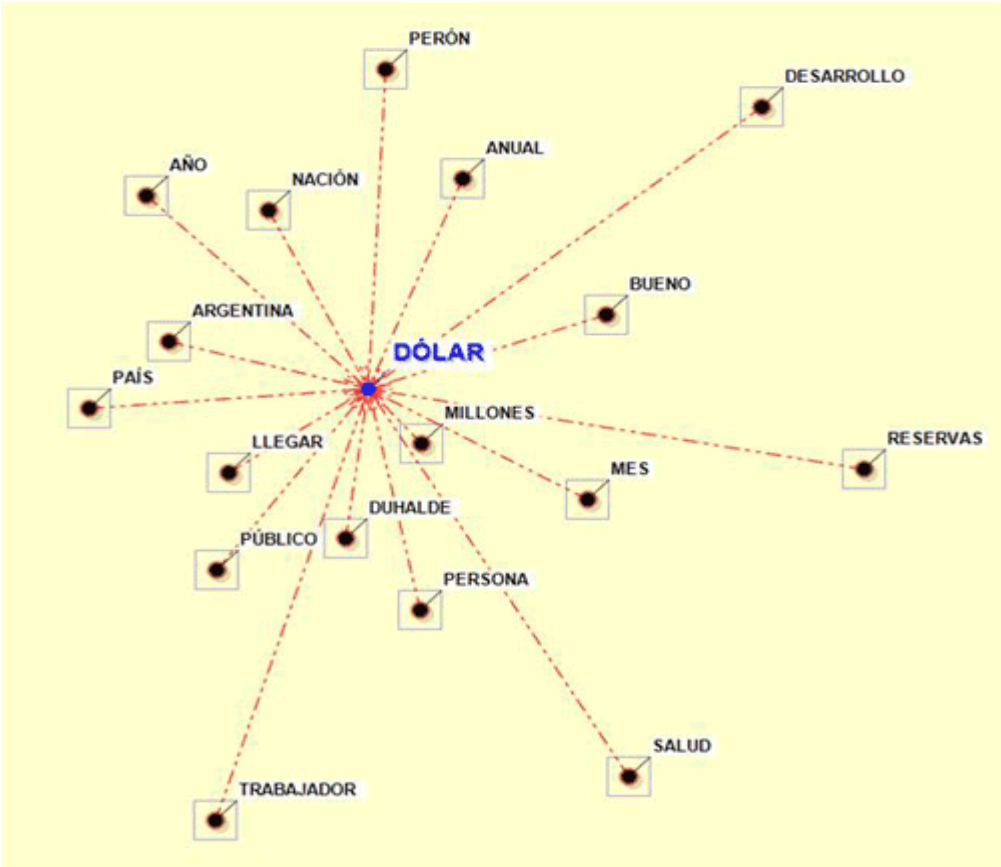
IdDoc	IdSeg	Topic	Score	Segm
12	363	DÓLAR	0,396	Cuando en 1989 llegamos al Gobierno me informaron que en las arcas públicas había solamente 60 millones d
5	194	POLÍTICA	0,361	Estoy convencido de que la crisis en Argentina más_que económica es política y de confianza del pueblo hacia
10	337	POLÍTICA	0,352	Por eso el problema de Argentina no es la economía , como se viene haciendo creer a la sociedad civil . El pro
5	157	POLÍTICA	0,347	la profunda crisis económica y política que atraviesa Argentina ha desatado algunas polémicas y reavivado alg
14	453	PESO	0,323	La economía y la sociedad argentina ya no tienen capacidad para mantener la libre convertibilidad del peso co
5	162	POLÍTICA	0,32	La dictadura militar y otros factores de inestabilidad política han quebrado la confianza del pueblo argentino e
5	172	POLÍTICA	0,315	Otro problema al que los analistas recurren para explicar la crisis es el de la clase política argentina . El verdad
8	270	BANCO	0,314	Los grandes bancos españoles han decidido no meter más dinero en Argentina . Al menos hasta_que la situaci
7	254	POLÍTICA	0,301	Porque el mayor problema a la hora de buscar soluciones duraderas en la Argentina es el desprestigio de la cla
13	428	PERONISTA	0,3	Los sectores más radicales del peronismo , Montoneros , las Fuerzas Armadas Peronistas y las Fuerzas Armad
1	63	DÓLAR	0,299	Los cuatro millones de dólares obtenidos fueron entregados a las hermanas y herederas de Eva Duarte , la seg
13	381	PERONISTA	0,296	Ninguno de los barones del partido se atreve hoy día a desembarazarse de la mítica tutela de Perón y Evita . T
5	189	BANCO	0,294	sin_embargo , la acusación injusta de que las empresas españolas obtienen ingentes beneficios de Argentina
13	426	PERONISTA	0,294	Primero , tantearía a_través_de un testaferro político , Héctor Cámpora , el grado de aceptación electoral del
3	113	PESO	0,293	Para anticipar las ventajas de la convertibilidad ampliada se creó un mecanismo de compensación comercial ,
5	181	BANCO	0,285	Esto nos lleva , necesariamente , a hablar del compromiso de España , de su sociedad y de sus empresas con e
9	294	DÓLAR	0,282	El Programa de Desarrollo de las Naciones Unidas afirma que aproximadamente con 90 mil millones de dólare
13	388	PLAN	0,277	Estados Unidos , preocupado por las buenas relaciones de los militares nacionalistas argentinos con Mussolin
7	246	DÓLAR	0,276	El Frente , respaldado por los tres millones de votos y las movilizaciones espontáneas reclama el establecimie
3	115	PESO	0,273	En lo_que compensa las oscilaciones de los precios dentro del esquema monetario internacional formado po
1	25	DÓLAR	0,267	Y su esposa Hilda ' ' Chiche ' ' Duhalde gastaba 250 millones de dólares anuales en las manzaneras , mujere
8	274	BANCO	0,261	Los analistas , los bancos de inversiones , están sobrecastigando a los bancos españoles porque consideran qu
13	393	PERONISTA	0,258	Perón salió de la prisión fortalecido , blandiendo como eslogan electoral el simbólico ' ' Braden o Perón ' '
8	265	BANCO	0,257	' ' Las medidas adoptadas por el Gobierno argentino son una confiscación de hecho de toda la banca ' ' , cor
13	413	PERONISTA	0,256	GOBIERNO RADICAL La autollamada Revolución Libertadora que dio el golpe de Estado contra Perón , terminó
5	160	POLÍTICA	0,252	Argentina tiene elementos sobrados para embarcarse en esta nueva economía , pero para ello debe dejar de
8	267	DÓLAR	0,252	Tanto el BBVA como el SCH dan ya por perdida su inversión en Argentina . El banco que preside Emilio Botín h
10	330	POLÍTICA	0,252	' ' Esta política económica es la única posible . Cualquier otra nos llevará al desastre ' ' , dicen los que gobi
13	416	GOLPE	0,252	El clima de inestabilidad social hizo que grandes empresarios industriales y poderosos ganaderos golpearan la
13	415	PERONISTA	0,25	y dirigiendo una ola de sabotajes . La ofensiva peronista no logró sin_embargo solidificar su frente interno . La
5	187	POLÍTICA	0,247	Las empresas españolas van_a seguir allí , no sólo porque la inversión exterior en sectores estratégicos es con
13	431	PERONISTA	0,247	Tras la llegada de Perón a Argentina y la matanza de Ezeiza que tuvo lugar el mismo día de su retorno , Cámpo

6– Aplicar el modelo

Después haber aplicado y guardado el modelo, por tanto que los temas son archivados por **T-LAB** como modalidades de dos nuevas variables que se refieren a clúster de contextos elementales (**CONT_CLUST**) y/o a clúster de documentos (**DOC_CLUST**), las relaciones entre los mismos temas y/o sus características pueden ser más explorados con diferentes instrumentos de análisis (ver el cuadro siguiente)



Por ejemplo, usando la herramienta **Asociaciones de palabras** y seleccionando el subconjunto (es decir, el tema) "Dólar", se puede crear el siguiente gráfico.



7 - Exportar un diccionario de las categorías modelo

Cuando se selecciona esta opción, **T-LAB** genera un archivo diccionario con extensión .dictio listo para ser importado a través de una de las herramientas disponibles para el análisis temático. En dicho diccionario, cada categoría viene descrita a través de sus palabras características.

Clasificación Temática de Documentos

Esta función sólo está habilitada cuando el corpus en análisis incluye un número de documentos primarios comprendido entre un mínimo de 20 hasta un máximo de 99.999.

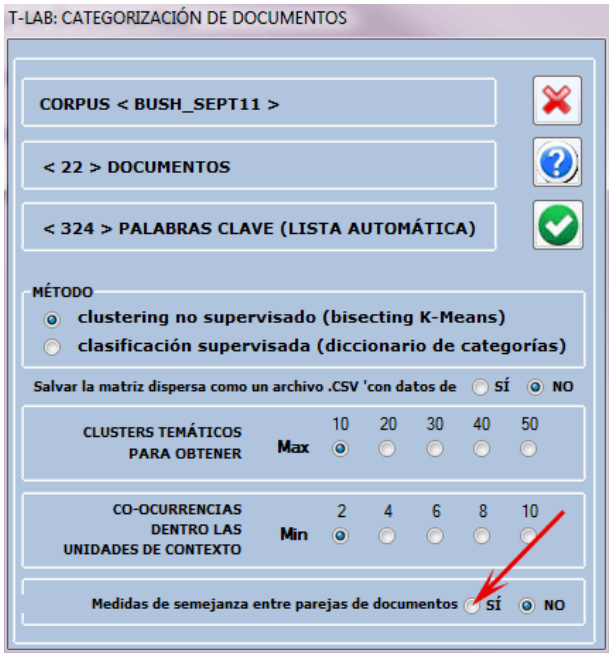
El proceso de análisis puede ser ejecutado o con un método de clustering 'no supervisado' (en el caso concreto, un algoritmo de bisecting K-Means) o con una clasificación supervisada (es decir, un enfoque top-down). Cuando se elige la segunda vía, es decir, la clasificación supervisada, se requiere la importación de un diccionario de las categorías, bien creado por un anterior análisis **T-LAB**, o bien construido por el usuario.

Su uso permite construir clusters de documentos y explorar sus características por medio de operaciones/opciones similares a las descritas en la sección del manual dedicado al **Análisis Temático de Contextos Elementales.**

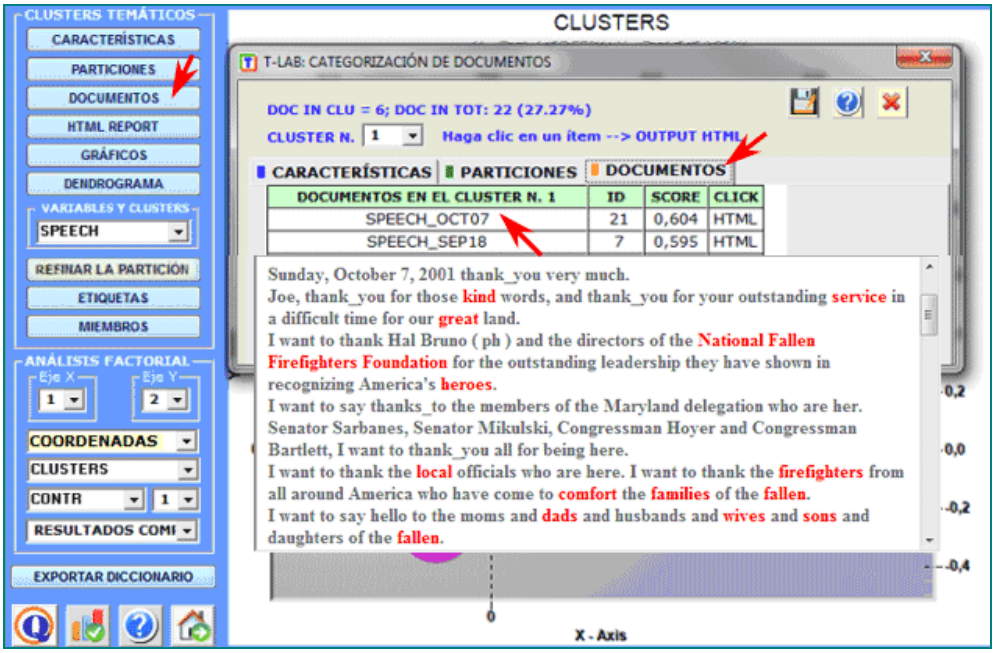
Su especificidad consiste en el hecho de que la tabla analizada se compone de tantas líneas como contenga el documento del corpus, cada una de las cuales se representa como un vector de valores que indican la ocurrencia de la palabra presente en el mismo.

Además, cuando el número de documentos analizados no es superior a 3000, es posible obtener medidas de semejanza (índice de coseno) entre cada uno de ellos y todos los demás (véase abajo).

N.B.: En este caso el nivel mínimo de aceptación del índice de semejanza está fijado en 0.05.



Consecuentemente, los resultados específicos de esta función son los siguientes:



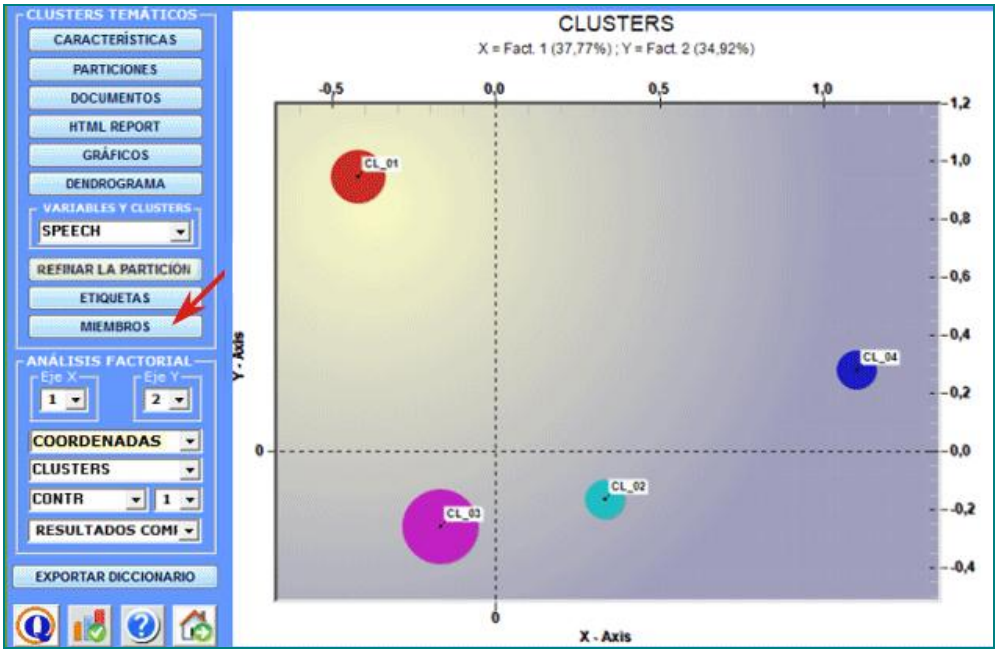
Los documentos que pertenecen a cada cluster son ordenados por el valor decreciente de importancia y se pueden examinar en formato HTML.

En este caso el valor de importancia (score) asignado a cada documento (i) en el cluster (k) es obtenido aplicando la fórmula siguiente:

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

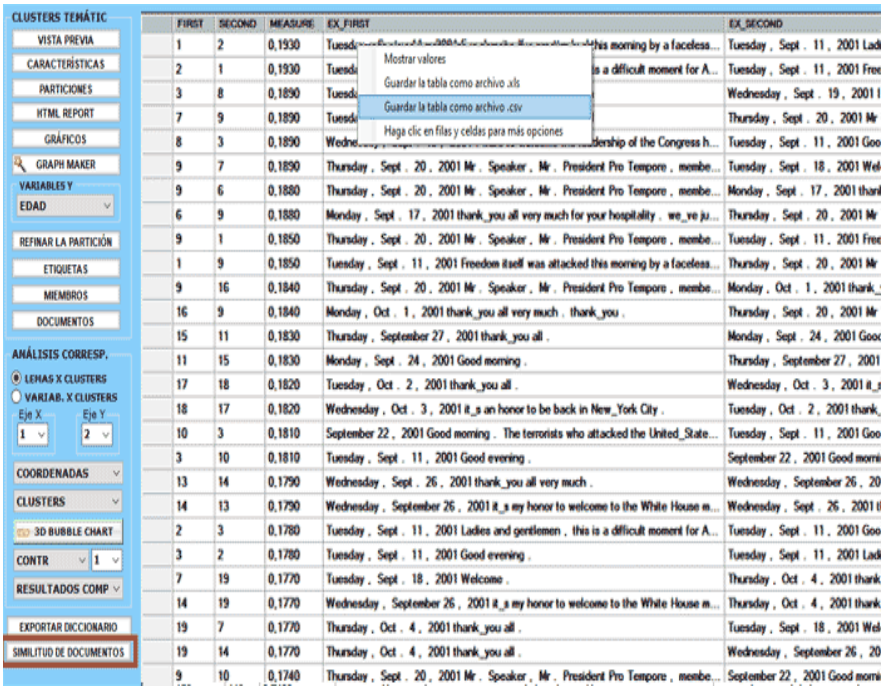
Donde:
i – se refiere al documento **i**;
k – se refiere al cluster **k**;
cos - es el símbolo del coseno;
$\mathbf{d_i}$ - es el vector normalizado de $\mathbf{TF_{j,\,i}\ IDF_j}$, donde **j** se refiere a una palabra del documento **i**
$\mathbf{c_k}$- es el vector normalizado de $\mathbf{TF_{j,\,k}\ IDF_j}$, donde **j** se refiere a palabra del cluster **k.**

Usando los valores (scores) obtenidos por la fórmula antedicha, que son transformados en porcentajes, **T-LAB** hace disponible el archivo " Document_Membership_Degree.xls " (véase abajo) que contiene los clusters a los cuales pertenecen los diferentes documentos, tanto sea por el bisecting K-Means (donde cada documento pertenece exclusivamente a un cluster) como por el TF-IDF (donde cada documento es caracterizado da una pertenencia mezclada a varios clusters).

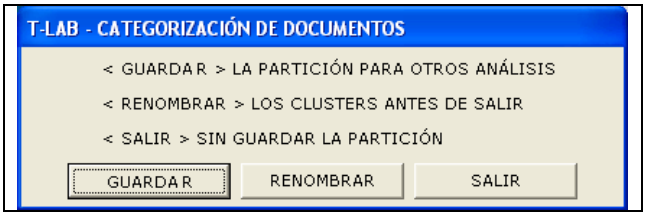


DOC_ID	VAR_01	CLUST_K-	BEST_TF-	MATCHIN(	CLUST-1	CLUST-2	CLUST-3	CLUST-4
1	BU_SEP11a	4	4	1	0,323	0,148	0,177	0,352
2	BU_SEP11b	1	1	1	0,478	0,107	0,159	0,256
3	BU_SEP11c	4	4	1	0,246	0,223	0,165	0,366
4	BU_SEP14	4	4	1	0,27	0,164	0,079	0,487
5	BU_SEP15	4	4	1	0,251	0,188	0,157	0,404
6	BU_SEP17	2	2	1	0,145	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

Cliqueando el botón 'Similitud de Documentos', tras haberlo habilitado, se puede verificar en qué medida cada documento es similar a cada uno de los demás. En este caso, la medida de semejanza es el coeficiente de coseno, y su valor varía en función del número de palabras utilizadas para la clasificación temática. La imagen siguiente presenta las diferentes opciones disponibles para este tipo de verificación.



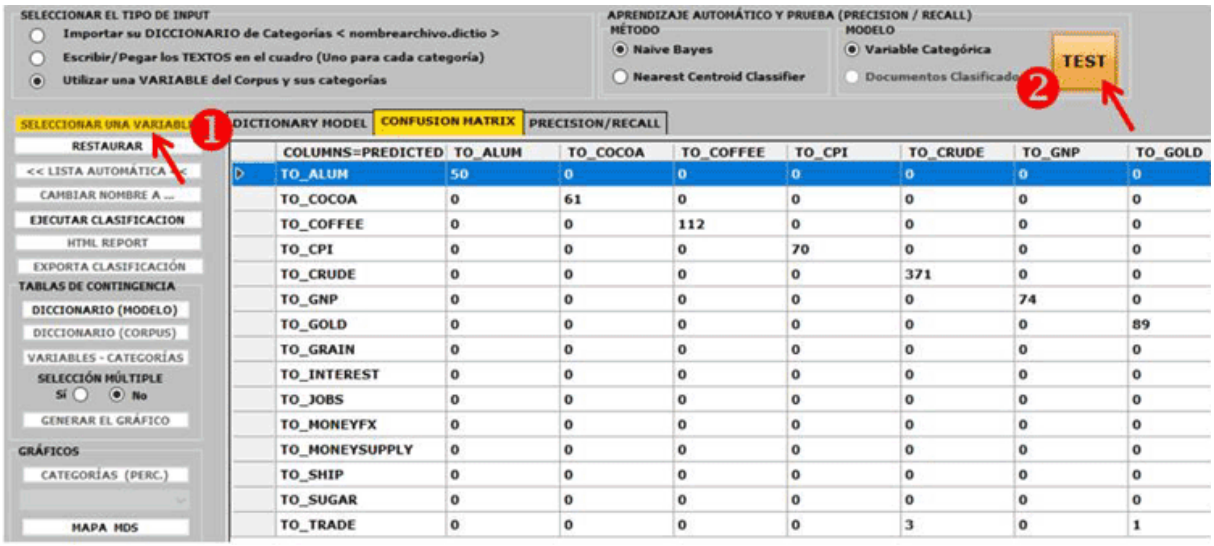
A la salida de esta función, algunos mensajes recuerdan que es posible explorar el cluster obtenido con otras herramientas **T-LAB**.



Seleccionando la opción "GUARDAR", será posible utilizar la variable < **DOC_CLUST** > (cluster de documentos) en todos los sucesivos análisis del mismo corpus realizados con otras herramientas **T-LAB.**

Clasificación basada en Diccionarios

NOTA: Las imagenes contenidas en este apartado hacen referencia a una versión anterior de T-LAB, ya que el interfaz de **T-LAB 10** cambia ligeramente. En particular, a partir de la versión 2021, una nueva característica permite probar fácilmente cualquier modelo en datos etiquetados (por ejemplo, datos que incluyen temas obtenidos de un análisis cualitativo anterior) y obtener resultados como matrices de confusión y métricas de precisión / recall (ver imagen a continuación).



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

Esta herramienta de **T-LAB** permite implementar una **clasificación automática** tanto de las **unidades lexicales** (es decir, palabras y lemas, incluidas los multiworlds) como de las **unidades de contexto** (frases, párrafos o pequeños documentos) presentes en un corpus. Todo esto aplicando un conjunto de categorías predefinidas o elegidas por el usuario

Según el tipo de categorías elegidas, que pueden ser importadas a través de un diccionario o generadas por **T-LAB**, dicha clasificación puede considerarse como una variedad de **análisis del contenido** o como una tipología de **sentiment analysis**.

Ya que el proceso de análisis permite la creación de variables nuevas y de ulteriores diccionarios que se pueden importar y exportar en otros proyectos de análisis, dicho instrumento se puede también utilizar para explorar el mismo corpus según perspectivas diferentes. Además, esta herramienta permitiría analizar dos o más conjuntos de textos aplicando los mismos modelos.

Entre los **posibles usos** de la herramienta destacan:

- Codificación automática de las respuestas a preguntas abiertas;
- Análisis top-down de los discursos políticos;
- Sentiment Analysis de los comentarios sobre productos específicos;

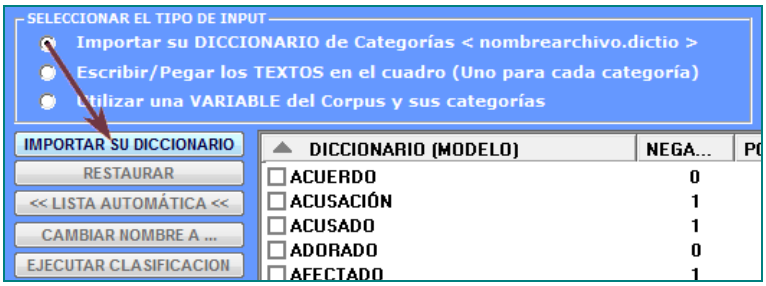
- Verificación del proceso psicoterapéutico;
- Validación de metodologías para el análisis cualitativo.

A continuación se proporciona una breve descripción de las cuatro fases principales del proceso de análisis. Éstas, sin embargo, tienen que ser consideradas como independientes las unas de las otras. De hecho, el investigador también tiene la opción de utilizar esta herramienta sólo para personalizar sus diccionarios o para explorar su conjunto de datos.

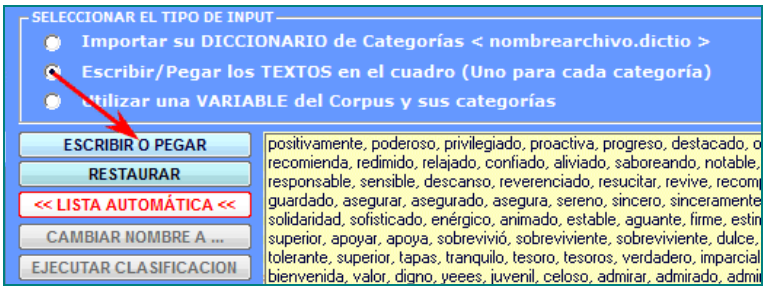
A) - FASE DE PRE-PROCESSING

Existen, para la fase de pre-processing, tres posibles puntos de partida, con **tipologías distintas de input** asociadas a ellos:

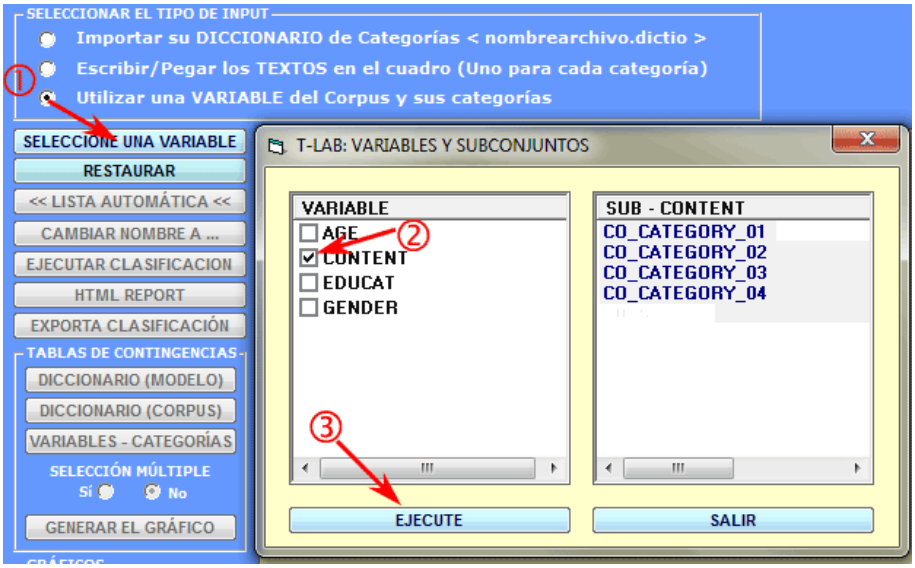
1 - Un **diccionario** pre-configurado de las categorías en el formato apropiado, y que ya se encuentra disponible (véanse las informaciones que a ello conciernen en la sección 'E' de este documento). En este caso hay que seleccionar la opción '**Importar su Diccionario**' (véase abajo);



2 - Un diccionario de las categorías que hay que generar a partir de **ejemplos de textos** o a partir de **listas de palabras** proporcionadas por el usuario. En este caso es suficiente teclear o copiar/pegar los textos en la casilla apropiada (un ejemplo por cada categoría, en secuencia y con un máximo 100.000 caracteres por cada uno);

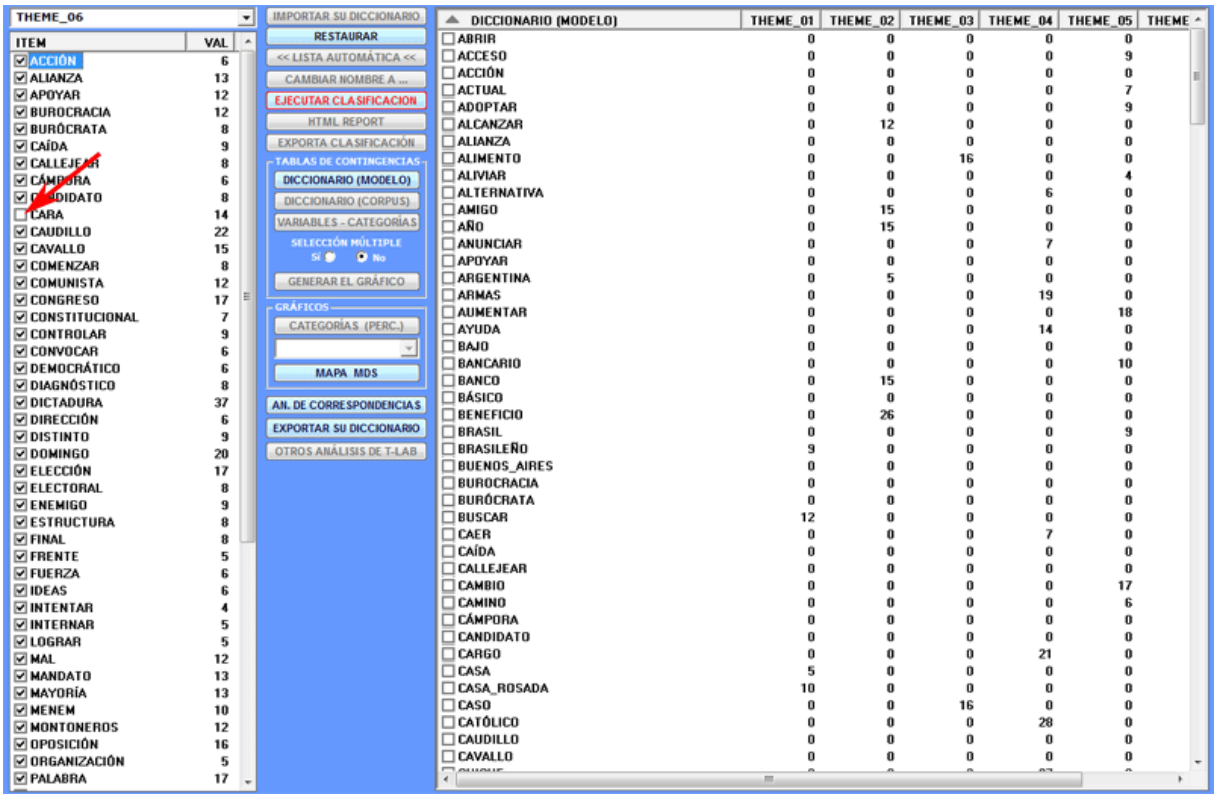


3 - Un diccionario que hay que generar a partir de las categorías de una **variable** obtenida en un análisis anterior de contenido. En este caso es suficiente hacer clic en la opción '**Seleccione una variable**' y realizar las elecciones apropiadas (véase abajo).

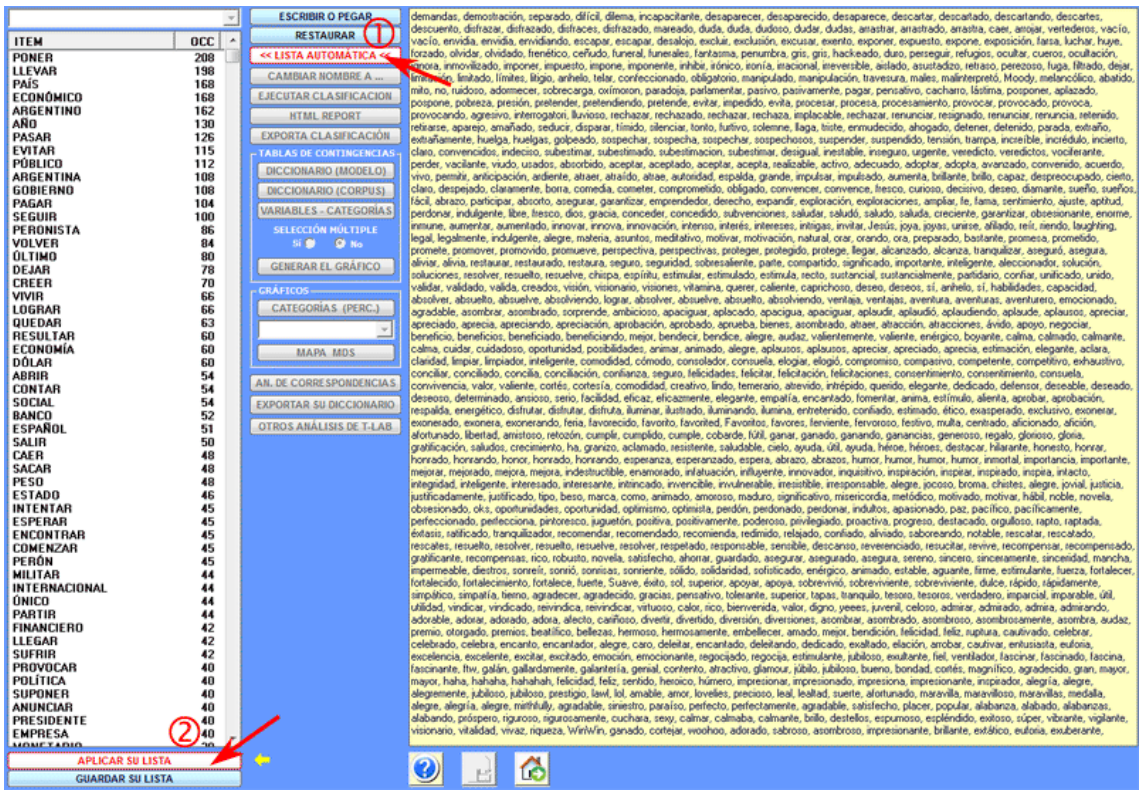


En base al punto de partida en el que se encuentre el usuario, y antes de habilitar la función 'Ejecutar Clasificación', **T-LAB** funciona de la siguiente manera:

1 - Se transforma el diccionario importado en una tabla de contingencia que el usuario puede utilizar de distintas maneras (véase la sección 'C' de este documento). Además, seleccionando cada categoría, es posible eliminar uno o más de los elementos correspondientes (véase imagen de abajo).

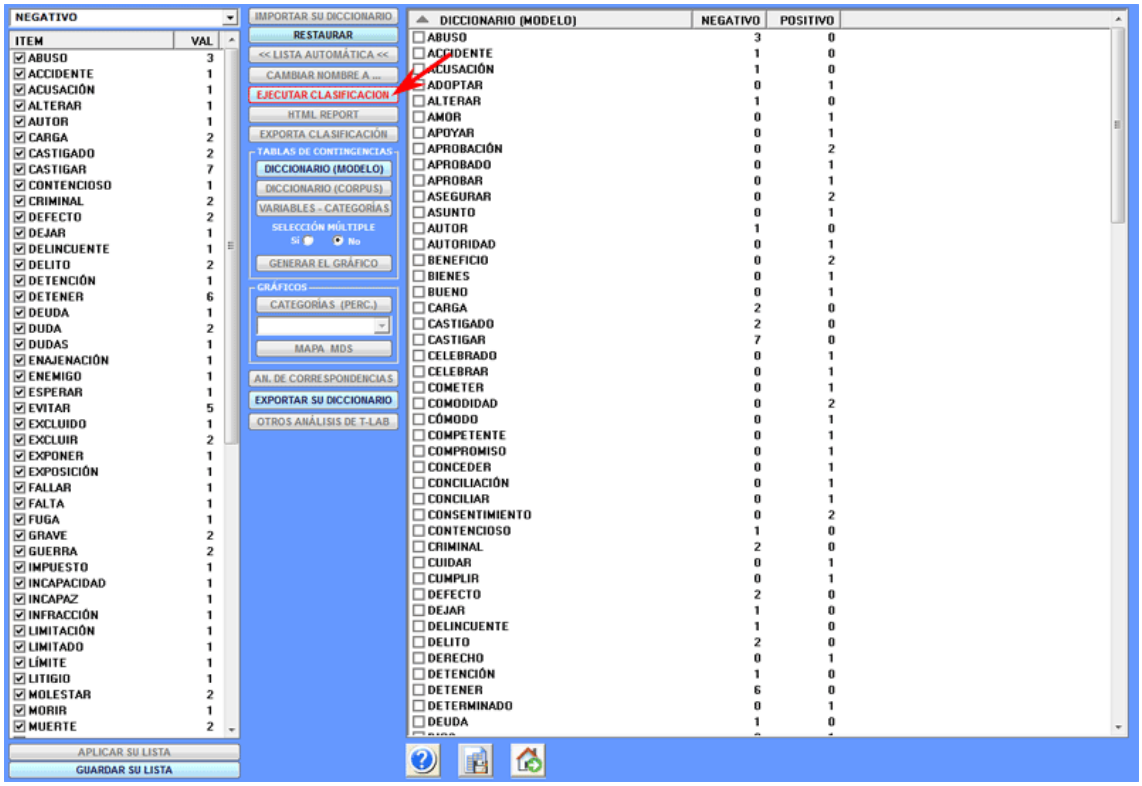


2 - Una vez que se hayan introducido los textos de ejemplo en la casilla correspondiente, y después de haber seleccionado la opción '**Lista automática**' (véase abajo), **T-LAB** ejecuta una tipología de lematización específica que sólo utiliza el diccionario del corpus seleccionado (véase el listado de palabras en la zona de izquierda de la imagen siguiente) y luego transforma cada texto en un listado cuyos elementos pueden ser incluidos o excluidos en la selección. Sucesivamente, para convalidar cada lista de palabras (es decir, cada categoría del diccionario), se necesita seleccionar la opción '**Aplicar su lista**' (véase abajo). Es necesario repetir cada una de las operaciones recién mencionadas para cada categoría presente en el diccionario. Después de haberlo hecho, el usuario está en disposición de ejecutar las operaciones descritas en la sección 'C' de este documento.

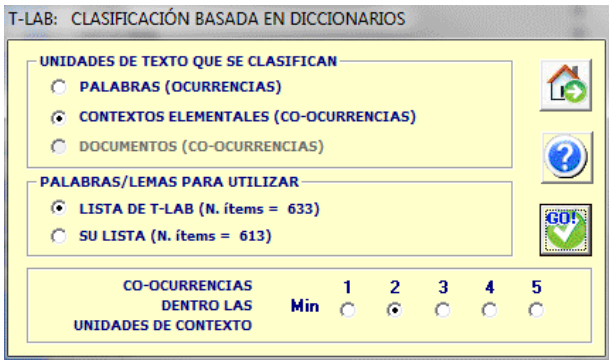


3 - Cuando se selecciona una variable proporcionada por un anterior análisis del contenido, **T-LAB** le asocia una tabla de contingencia palabras por categorías. De esta manera el usuario puede ejecutar todo tipo de operaciones de exploración de los datos (véase la sección 'C' del presente documento).

B) - PROCESO DE CLASIFICACIÓN



Después de haber seleccionado la opción '**Ejecutar Clasificación**' (véase arriba), el usuario puede elegir, según el tipo de corpus que esté analizando, entre las siguientes opciones.



En este punto, si el usuario decide **clasificar las palabras**, no hay ulteriores opciones disponibles. De hecho, en este caso, las ocurrencias de cada palabra (es decir, los word tokens) simplemente se enumeran como ocurrencias de la categoría correspondiente. Pongamos el caso de que en nuestro diccionario exista la categoría 'religión', y que ésta incluya las palabras 'fe' y 'oración'. A la hora de analizar un documento que contenga ambas palabras, **T-LAB** se limitaría a juntar sus ocurrencias. Por ejemplo, 2 ocurrencias de la palabra 'fe' y 3 ocurrencias de la palabra 'oración', se convertirían en 5 ocurrencias de la categoría 'religión'.

Por otro lado, si el usuario decide **clasificar las unidades de contexto** (es decir 'contextos elementales', como frases y párrafos, o 'documentos'), **T-LAB**considera tanto las categorías del diccionario como las unidades de contexto a clasificar en términos de perfiles de co-ocurrencia (es decir, term vectors), y calcula sus medidas de semejanza. Para ello, se pueden

filtrar los perfiles de co-ocurrencia bien a través de una 'lista de T-LAB' (es decir una lista que incluya todas aquellas palabras-clave que tengan valores de ocurrencia mayores o iguales al umbral mínimo de 4), bien mediante una lista personalizada (es decir, un listado de palabras-clave elegidas por el usuario). Dichas listas, sin embargo, pueden a veces resultar iguales. Además, en estos casos, **T-LAB**permite excluir del análisis las unidades de contexto que no incluyan un número mínimo de palabras-clave (véase arriba el parámetro 'co-ocurrencias dentro de las unidades de contexto').

Cuando, como en el caso recién descrito, los 'objetos' a clasificar son las unidades de contexto, **T-LAB** procede de la siguiente forma:

a) Normaliza los vectores correspondientes a las 'k' categorías (perfiles columna) del diccionario utilizado;
b) Normaliza los vectores correspondientes a las unidades de contexto que hay que analizar;
c) Calcula medidas de semejanza (coseno) y diferencia (distancia euclidiana) entre cada uno de los 'i' vectores, correspondientes a todas de las unidades de contexto, y cada uno de los 'k' vectores, correspondientes a todas las categorías del diccionario utilizado;
d) Asigna cada unidad de contexto ('i') a la clase o categoría ('k') con la que mantiene la relación de semejanza más alta. (Nota: En todos los casos, para cada pareja 'unidad de contexto' / 'categoría', el valor máximo del coseno y el valor mínimo de la distancia euclidiana deben coincidir. De no ser así, **T-LAB**considera la unidad de contexto 'i' como 'no clasificada').

En otras palabras, en el caso recién descrito, **T-LAB** utiliza algo parecido a un método K-means , donde los 'k' centroides se definen a priori y no vienen actualizados durante el proceso de análisis.

Debido a que, en este caso, la clasificación es de tipo top-down, la calidad de los resultados obtenidos depende, básicamente, de dos factores:
1 - La 'pertinencia' del diccionario utilizado (véase relación entre léxico del corpus y diccionario de las categorías),
2 - La capacidad 'discriminante' de cada una de las categorías (véase relación entre las categorías del diccionario).
De hecho, cuando estos dos factores alcanzan el nivel óptimo, ambos parámetros de 'precision' y 'recall' (véase http://en.wikipedia.org/wiki/Precision_and_recall) toman valores comprendidos entre 80% y 95%.

Cabe recordar que, de momento, **T-LAB** no tiene en consideración las fórmulas de negación. Consecuentemente, si a la hora de implementar una sentiment analysis, una frase como 'No odies tu enemigo' podría ser clasificada con tonalidad 'negativa'. Los usuarios expertos pueden gestionar este problema mientras se importa el corpus (véase el uso de listas para stop-words y multi-words). Por ejemplo, la expresión 'no odies' se puede transformar en 'no_odies' y, si se considera oportuno, se puede incluir en la categoría 'positivo'.

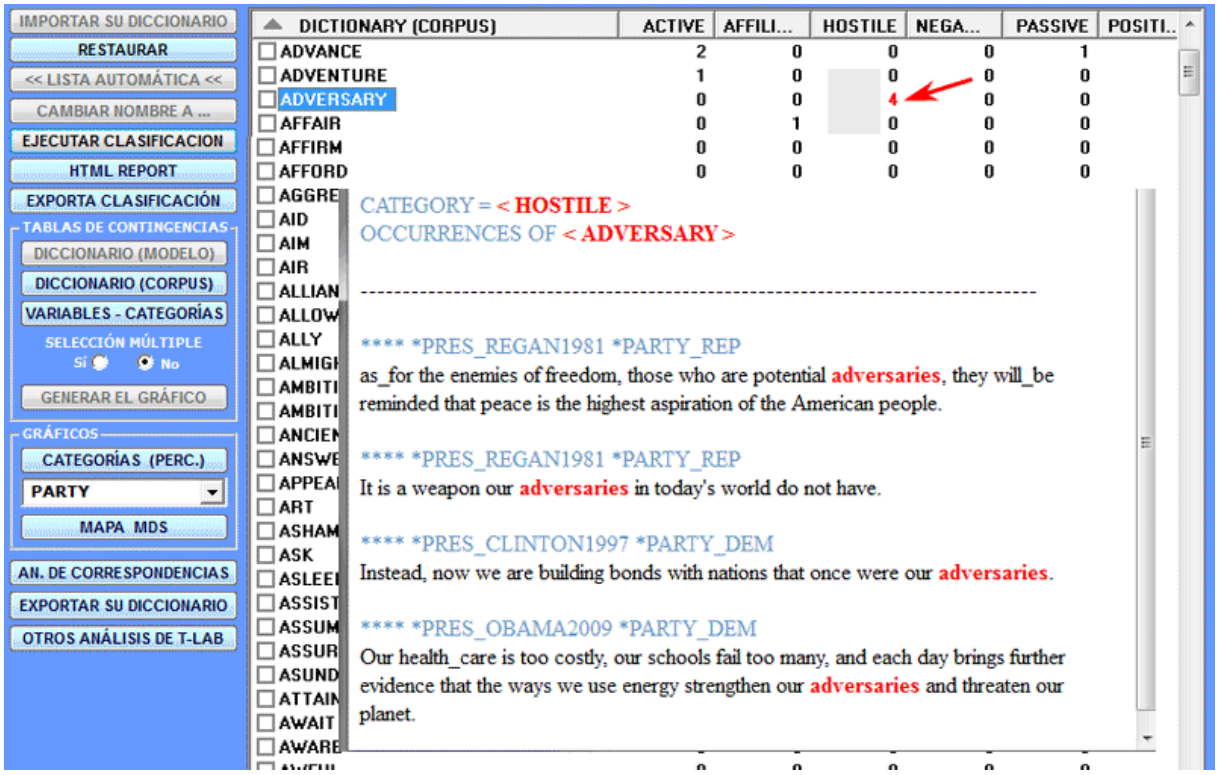
C) - EXPLORACIÓN DE LOS DATOS

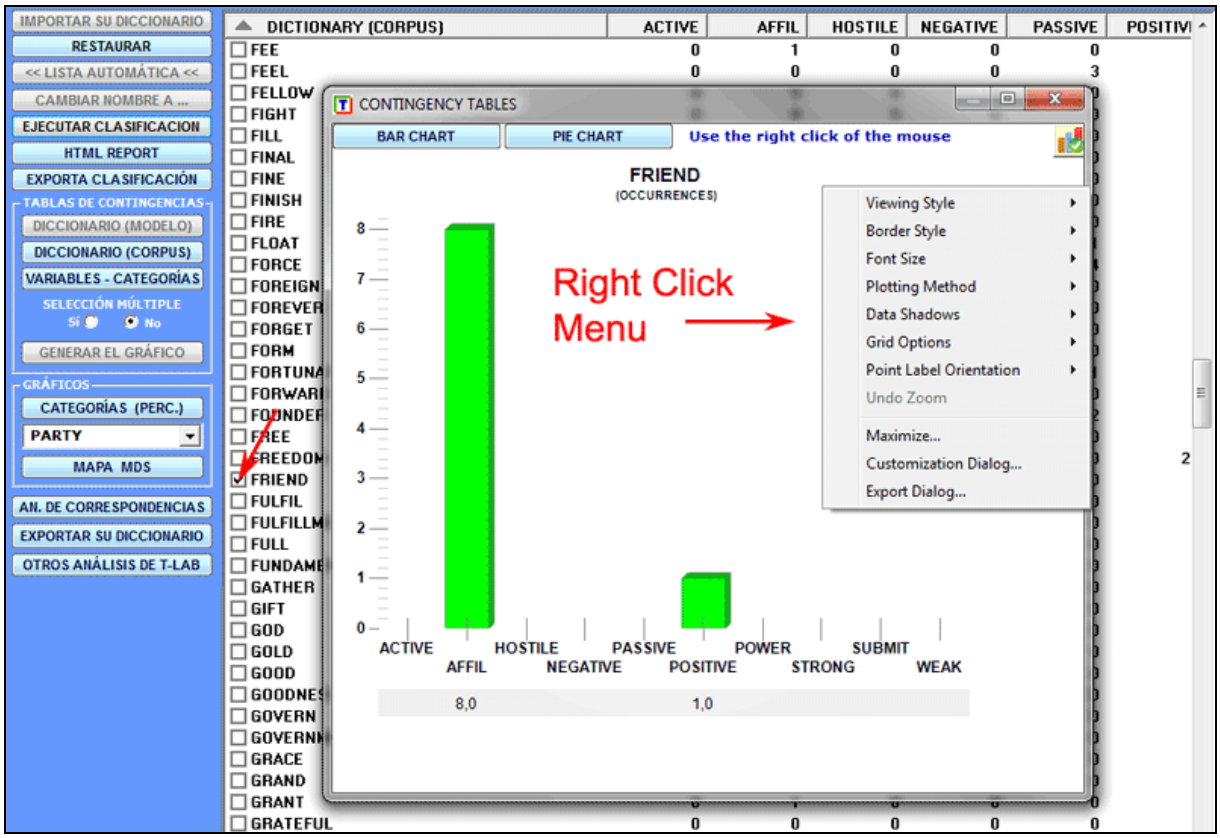
En el uso de esta herramienta, toda actividad de exploración hace referencia a **tablas de contingencias** que, según los casos, pueden incluir tanto los datos de input (por ejemplo, un diccionario de categorías) como los de output (por ejemplo, los resultados del proceso de clasificación).

Más en concreto, concerniente a los resultados del análisis, y dependiendo del tipo de unidad textual clasificada - (a) 'palabras', (b) 'contextos elementales' o (c) 'documentos' - las celdas de

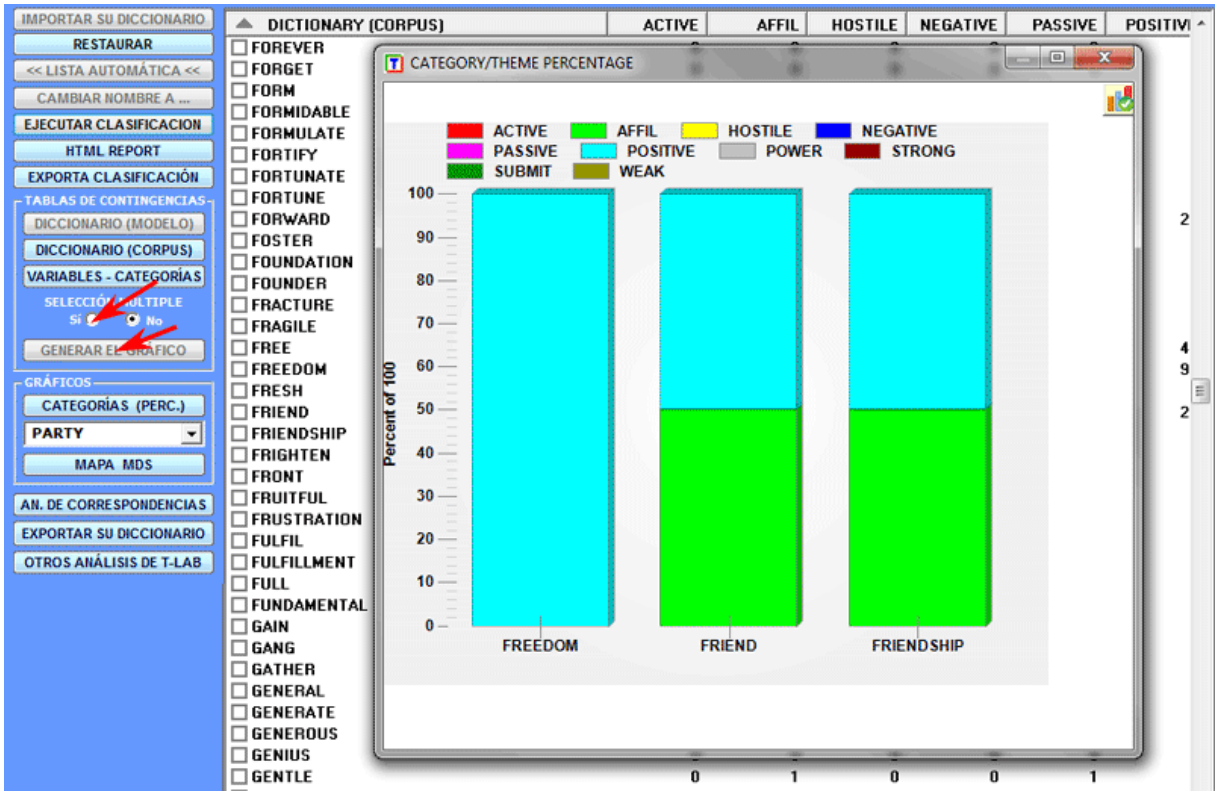
las tablas visualizadas pueden contener los siguientes valores:
a) El total de las ocurrencias de cada palabra que, dentro del corpus analizado o de un subconjunto del corpus, ha sido clasificada como perteneciente a una categoría predefinida (es decir, a la 'j' columna de la respectiva tabla de contingencia). Cabe destacar que, en este tipo de clasificación, las palabras que pertenecen simultáneamente a dos o más categorías tienen los mismos valores repetidos en las columnas correspondientes;
b) El total de los contextos elementales asociados a una categoría determinada (es decir, la 'j' columna) donde está presente la palabra en la línea 'i' correspondiente;
c) Total de las ocurrencias de cada palabra (véanse líneas de la relativa tabla de contingencia) dentro de los documentos asociados a cada categoría (véanse columnas de la tabla de contingencia) .

Haciendo clic en los check-box correspondientes a los diferentes ítems puestos en las líneas de la tabla, es posible obtener gráficos que se pueden personalizar de distintas maneras. Además, en el caso de la clasificación de tipo 'b' (véase arriba), si se hace clic en los valores contenidos en las celdas, es posible visualizar los contextos de ocurrencia de cada palabra.
A continuación, se presentan los output de un análisis en el que se han aplicado algunas categorías de un diccionario 'clásico' en el análisis del contenido (Harvard IV-4) a los discursos inaugurales de los presidentes de EEUU.

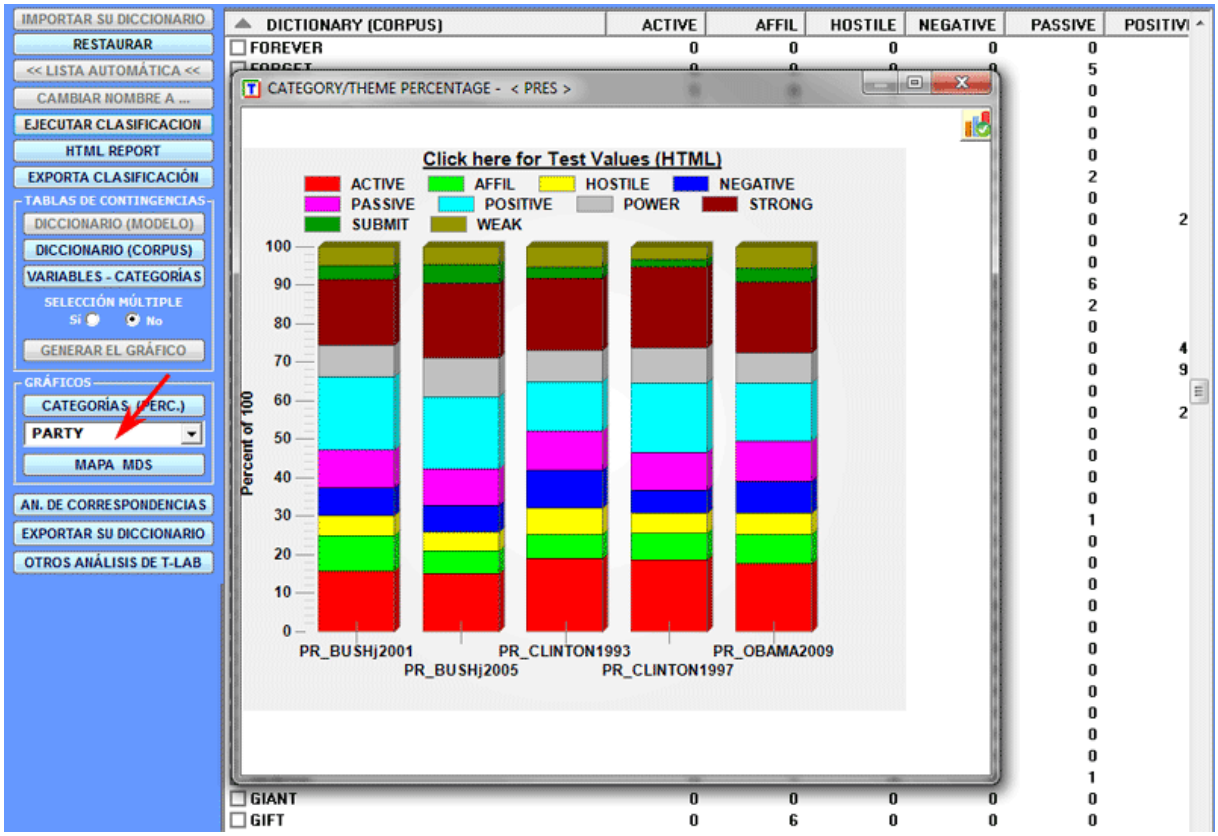




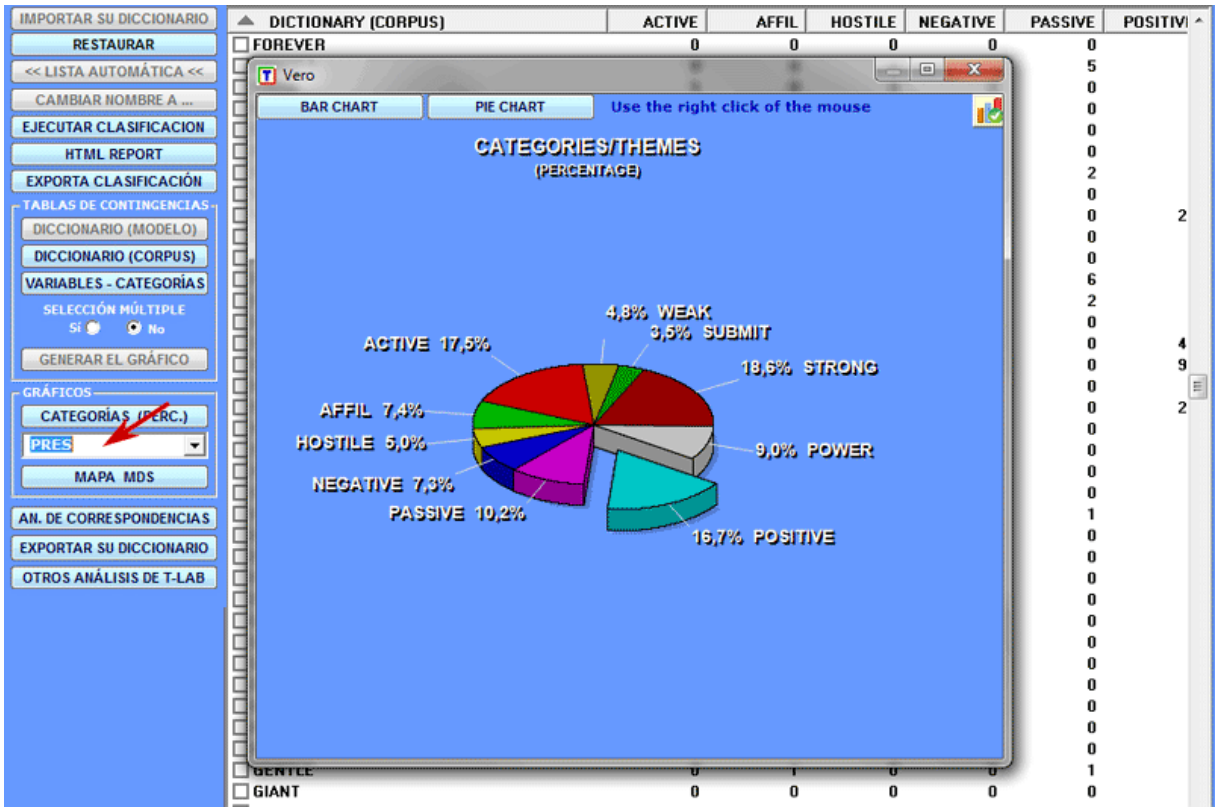
Para realizar gráficos con diferentes series de datos, a las cuales corresponderán diferentes líneas de las tablas de contingencia, es suficiente escoger la opción '**Selección Múltiple**' (opción 'SÍ'), seleccionar los elementos deseados, hasta un máximo de 20, y hacer clic en el botón '**Generar Gráfico**' (véase abajo).



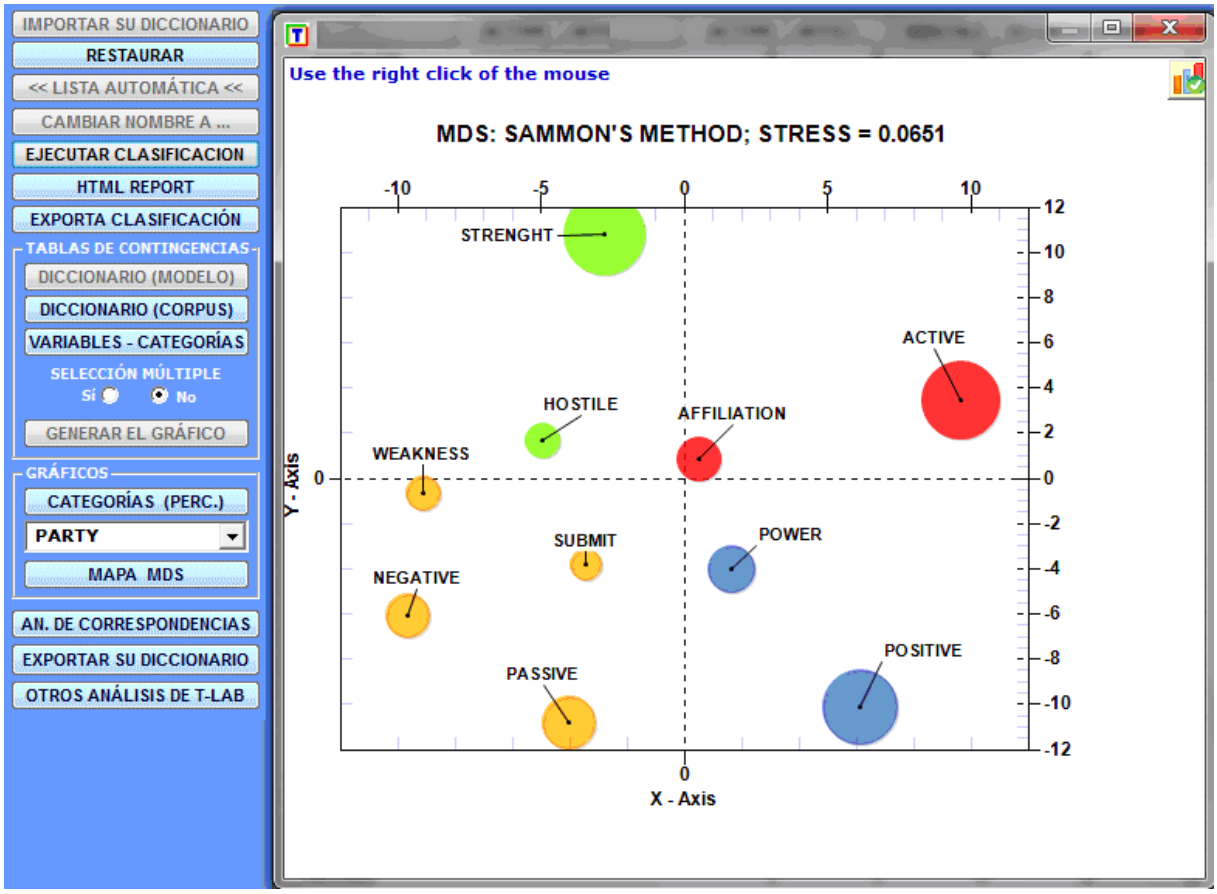
Las dos opciones recién mencionadas, también están disponibles para las tablas que incluyen los valores de las variables.

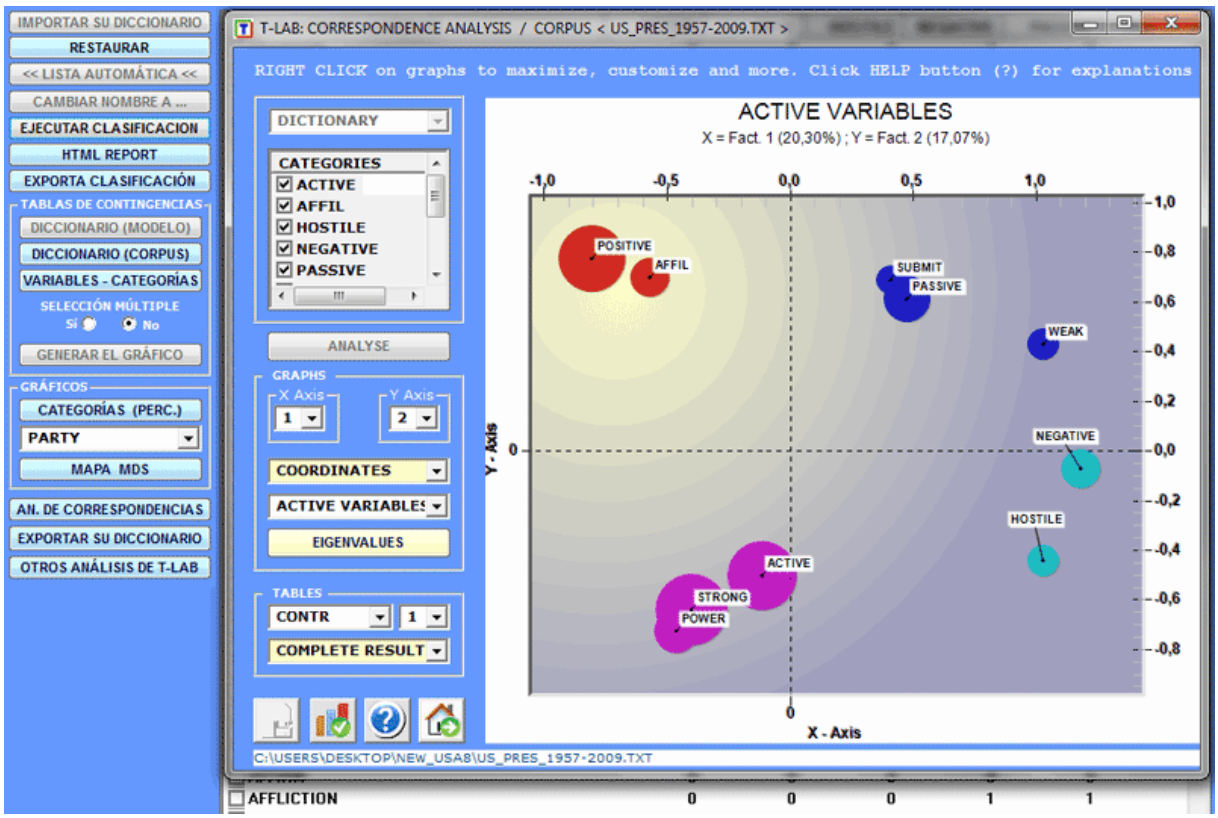


Existen distintas maneras de verificar los porcentajes de las diferentes categorías (véase abajo).



Para explorar la estructura entera de los datos incluidos en las tablas de contingencia se puede utilizar tanto la opción '**MDS**' como la opción '**Análisis de Correspondencias**' (véase abajo).





Sólo en el caso en que las unidades de contexto hayan sido clasificadas, es posible visualizar y exportar otros output con sus datos correspondientes. Además, en este caso, también se pueden guardar los resultados de los análisis en una variable nueva, y así seguir la exploración con otras herramientas del menú **T-LAB**.

Más en concreto, haciendo clic en el botón '**HTML Report**', es posible visualizar algunos de los resultados del proceso de clasificación en el que se asigna una puntuación de semejanza (coseno) a todos los 'contextos elementales' o 'documentos' que pertenecen a las diferentes categorías (Nota: Las imágenes que siguen se refieren a un corpus de documentos que contienen breves descripciones de empresas).

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 ...

Datos parecidos pueden ser exportados en archivos XLS (véase abajo) que contienen todas las informaciones inherentes a los contextos elementales ('Context_Classification.xls') o los documentos ('Document_Classification.xls') clasificados correctamente;

(1) - Context_Classification.xls

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

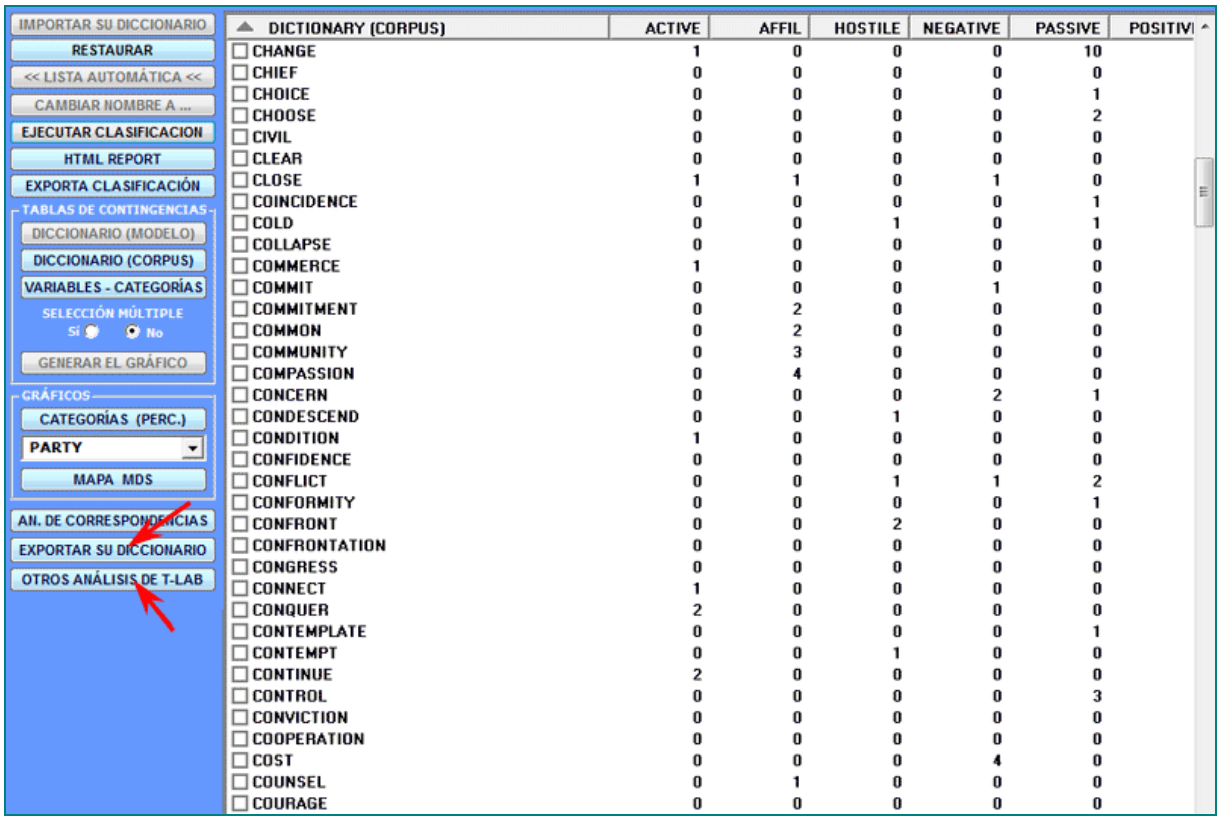
(2) - Document_Classification.xls

	'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) - FASES POSTERIORES DEL PROCESO DE ANÁLISIS

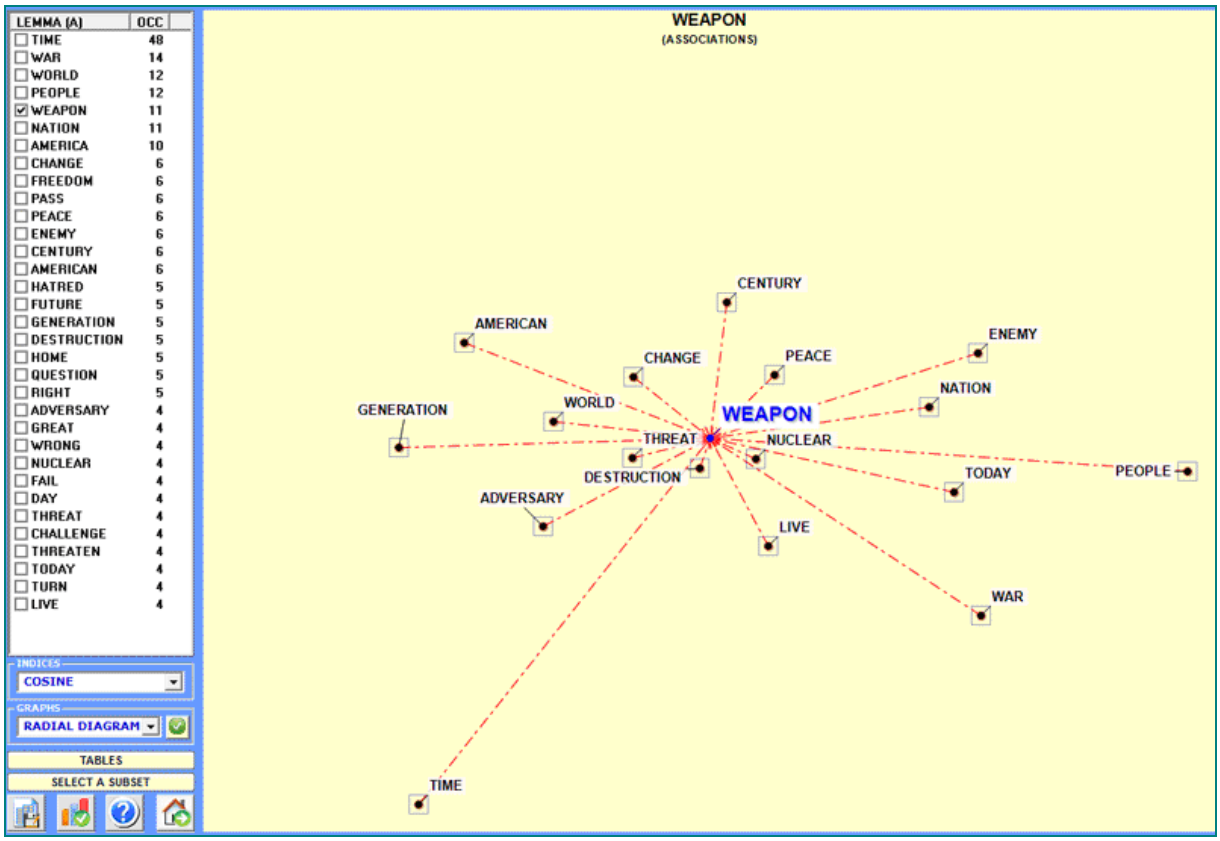
Una vez que el proceso de clasificación haya producido sus output, existen dos opciones disponibles:

-- '**Exportar su diccionario**', que genera un diccionario listo para ser importado y utilizado en otras herramientas de **T-LAB** para los análisis temáticos;
- '**Otros análisis de T-LAB**', que, en función de la estructura del corpus analizado, del tipo de clasificación implementado y del número de categorías aplicadas, produce una nueva variable que puede ser utilizada por otros instrumentos de **T-LAB** (véase abajo).



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	

A continuación se muestra un ejemplo construido a través del análisis de un 'subconjunto' de contextos clasificados por la herramienta '**Asociaciones de Palabras**' (véase el menú principal **T-LAB**).



E) - FORMATO INPUT/OUTPUT DE LOS DICCIONARIOS T-LAB

Se presentan aquí todas las informaciones acerca de los formatos de diccionarios que pueden ser importados por esta herramienta de **T-LAB:**

- Todos los diccionarios deben ser archivos de texto (ASCII/ANSI) con extensión 'dictio.' (ej. Mycategories.dictio);
- Todos los diccionarios creados por herramientas **T-LAB** para los análisis temáticos, incluidos los creados por la herramienta 'Clasificación basada en Diccionarios', están listos para la importación, sin necesidad de posteriores modificaciones por parte del usuario;
- Otros diccionarios, tanto estándar como personalizados, deben de ser creados siguiendo las presentes indicaciones:

1 - Cada diccionario se compone de 'n' líneas y no puede superar las 100.000 record ;
2 - Cada línea del diccionario incluye dos o tres 'cadenas' separadas por el signo de punto y coma (ejemplo: económico; crédito);
3 - Para cada línea, la primera cadena debe ser una 'categoría', la segunda una 'palabra' (o lema) y la tercera - si la hay - debe ser un numero real positivo (es decir, un numero entero), comprendido entre '1' y '999', y que representa el 'peso' de cada palabra dentro de la categoría correspondiente;
4 - El tamaño máximo de una cadena (palabra, lema o categoría) es de 50 caracteres y no debe contener ni espacios vacios ni apóstrofos;
5 - Cuando el diccionario incluye multi-words (ej. Gobierno Federal), los espacios deben ser sustituidos por el carácter '_' (ej. Gobierno_Federal);
6 - En cada diccionario, el número de categorías utilizadas puede variar de un mínimo de 2 a un máximo de 50. Cuando el numero de categorías es superior a 50, se aconseja utilizar un diccionario de diferente formato e importarlo a través de la herramienta **Personalización del**

diccionario (véanse 'Herramientas de Léxico' en el menú **T-LAB**). Cabe recordar que, en este caso, cada palabra debe tener una correspondencia unívoca con una sola categoría.

A continuación se presentan dos extractos de archivos .dictio, con dos y tres cadenas por línea respectivamente:

a) Caso con dos cadenas (es decir 'parejas' de categorías y palabras)

```
...
negativo;catastrófico
negativo;nocivo
...
positivo;fantástico
positivo;satisfecho
...
```

b) caso con tres cadenas (es decir, categorías, palabras y números)

```
...
negativo;catastrófico;10
negativo;nocivo;8
...
positivo;fantástico;9
positivo;satisfecho;7
```

Textos y Discursos como Sistemas Dinámicos

NOTA: Esta sección solo está disponible en inglés.

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.

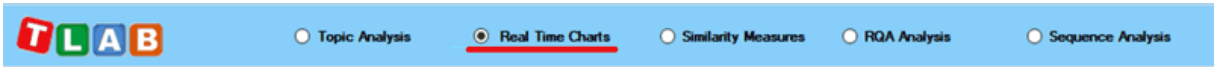
(**) 'Topic' and 'Theme' will be hereafter treated as synonymous terms.

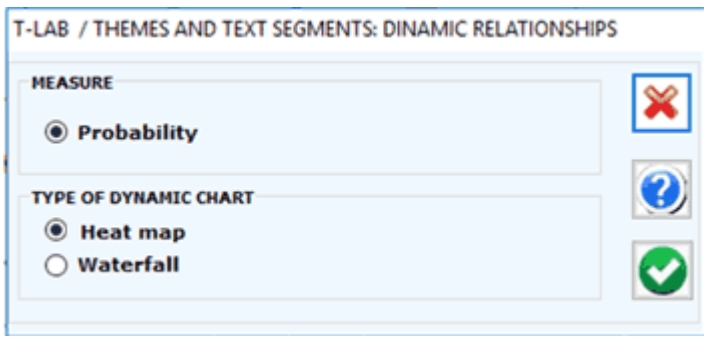
Please note that the main goal of the above automatic steps is to extract 'k' latent dimensions (where 'k' varies from 20 to 30) which determine the content structure of the analysed text and which – like a mixture model - can be used for exploring both text dynamics and similarities between text segments.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

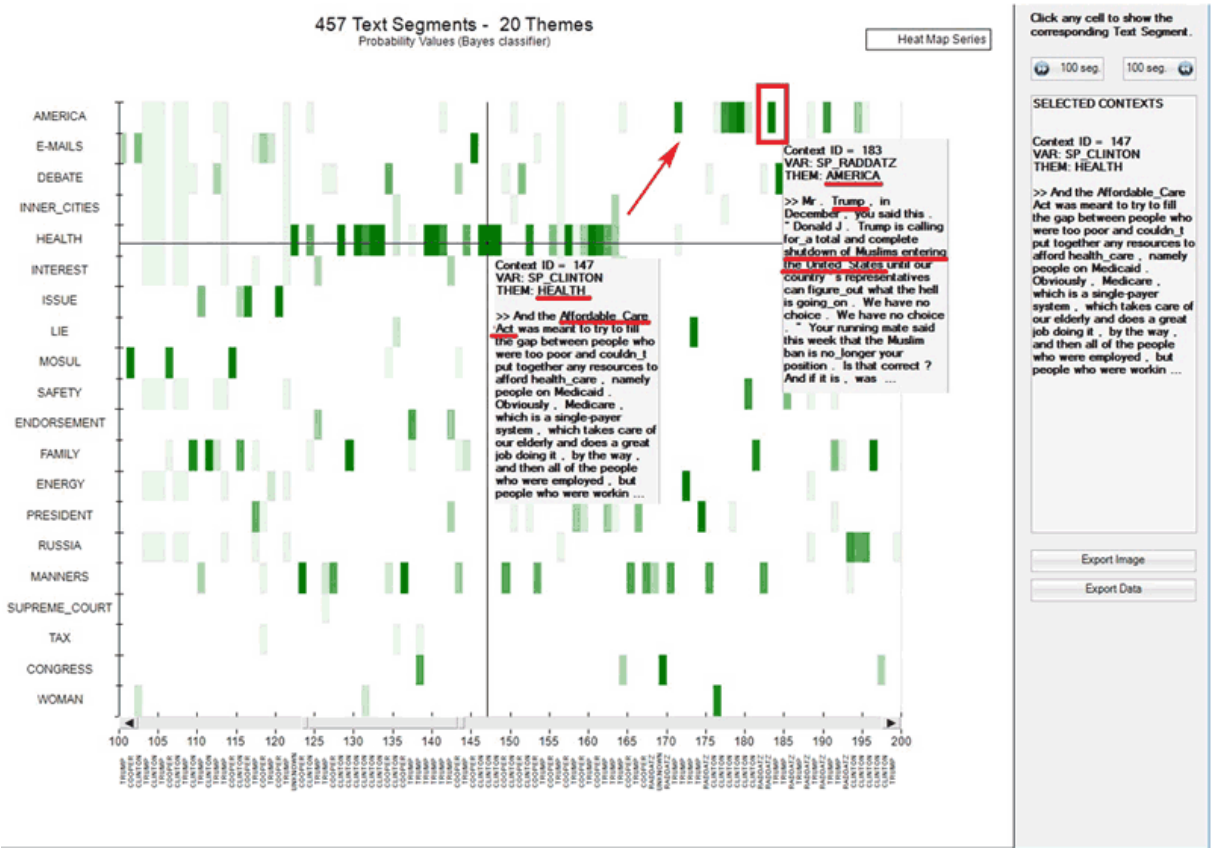
Two complementary charts allows us to easily appreciate various types of events, including the **strong recurrences** of some themes or the **shifts** from a theme to another (see the below pictures, obtained by analysing a presidential debate between Hillary Clinton and Donald Trump which took place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

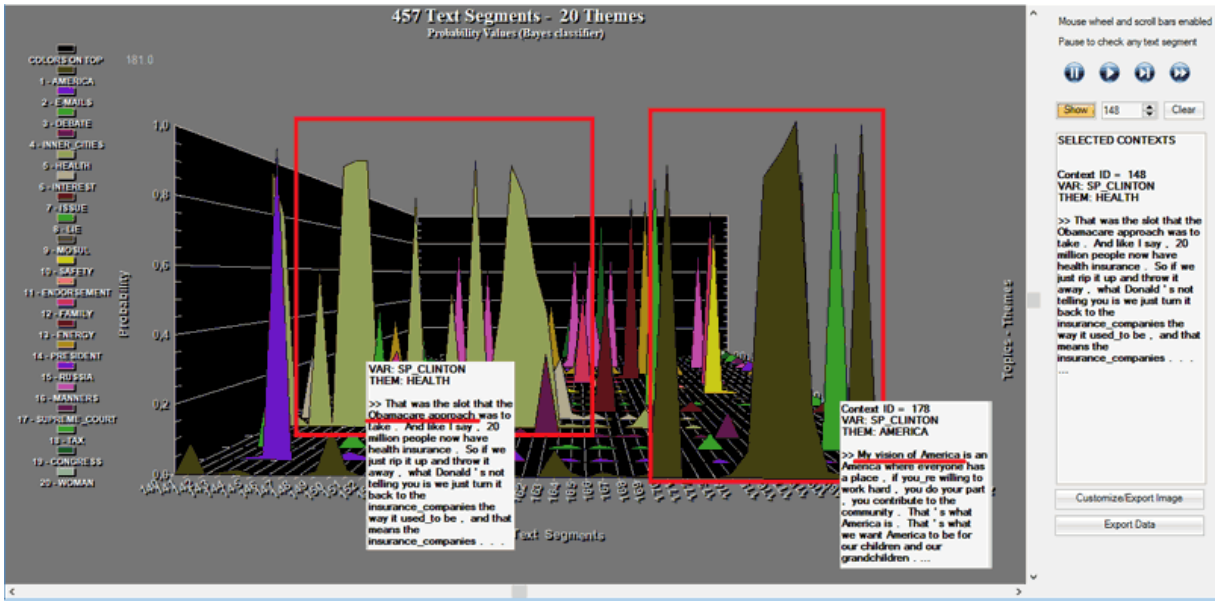
As the information summarized by these types of charts mainly refers to formal aspects of text contents, the same charts may be regarded as some sort of musical scores where the sequence of themes and their 'intensity' (i.e. probability) vary in time.

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

A.1 - Heat map



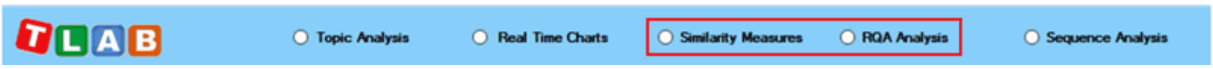
A.2 - Waterfall



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.

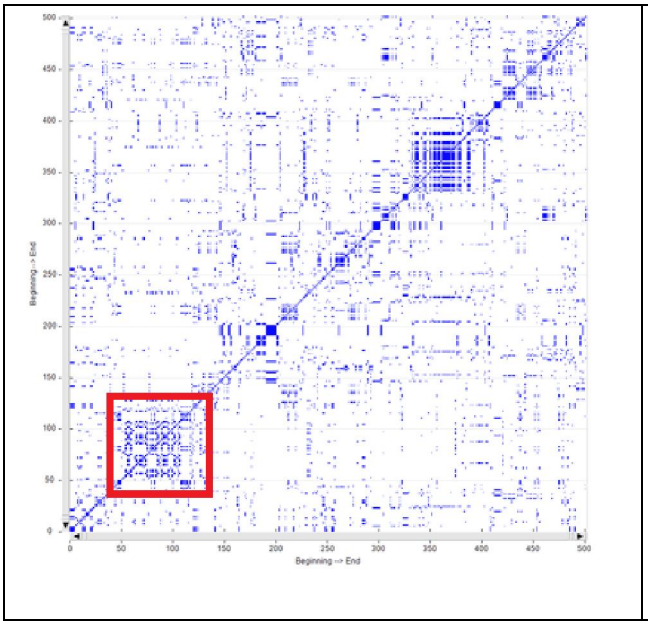


Fig. 1 - Time series

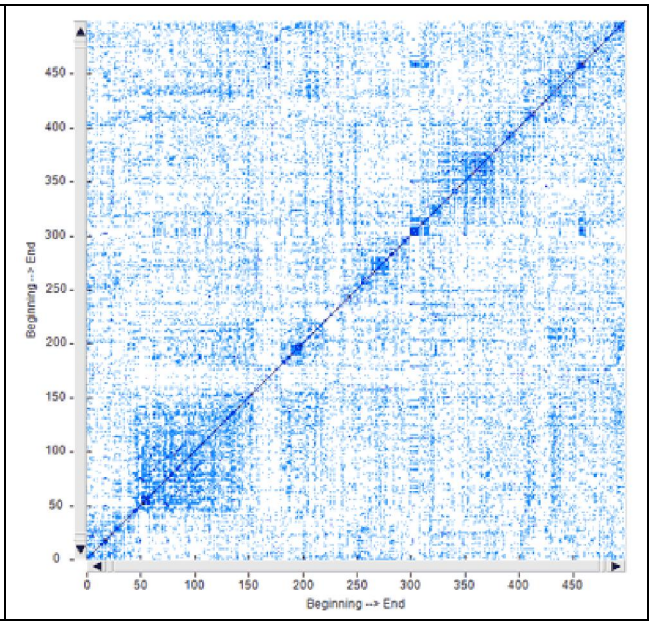
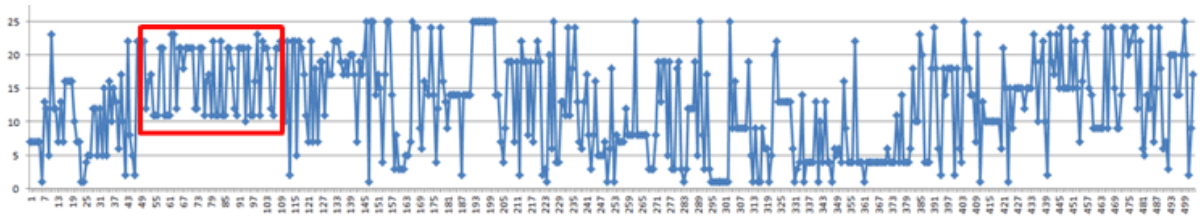
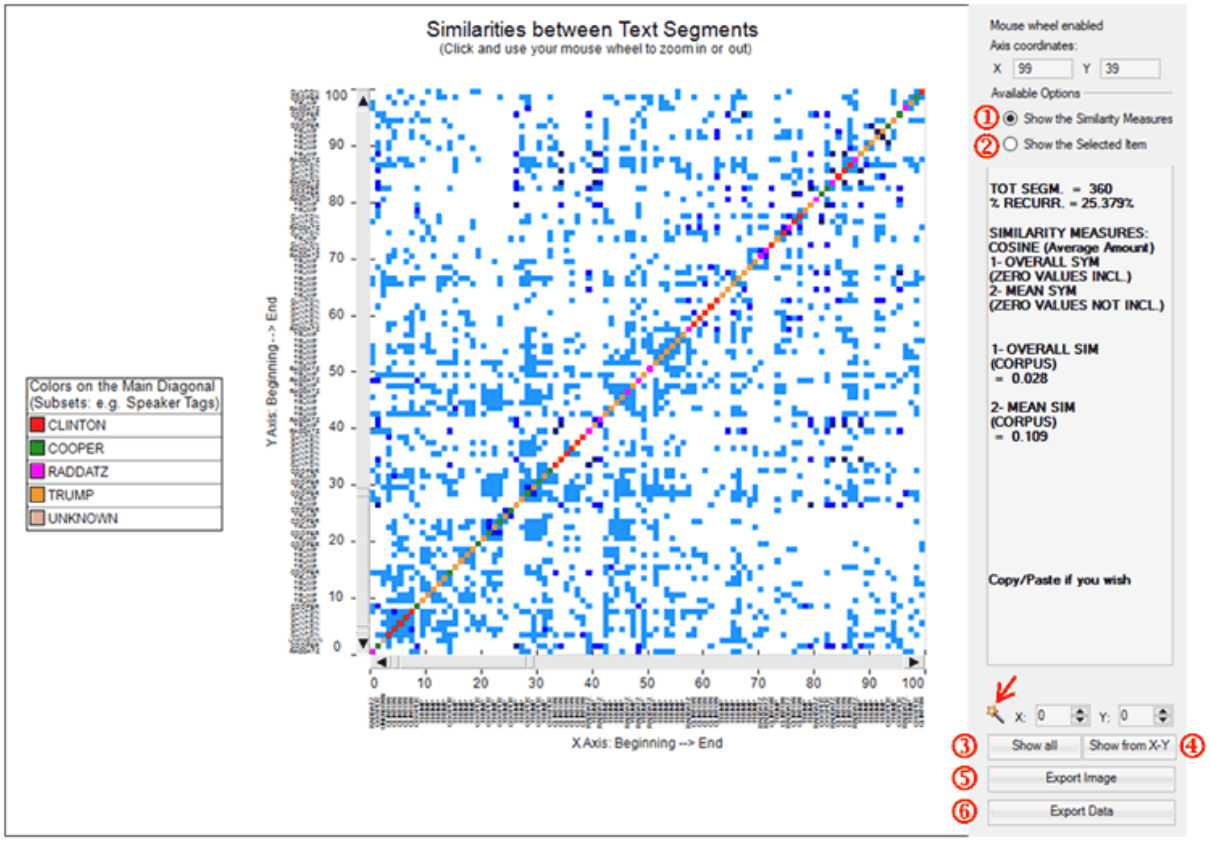


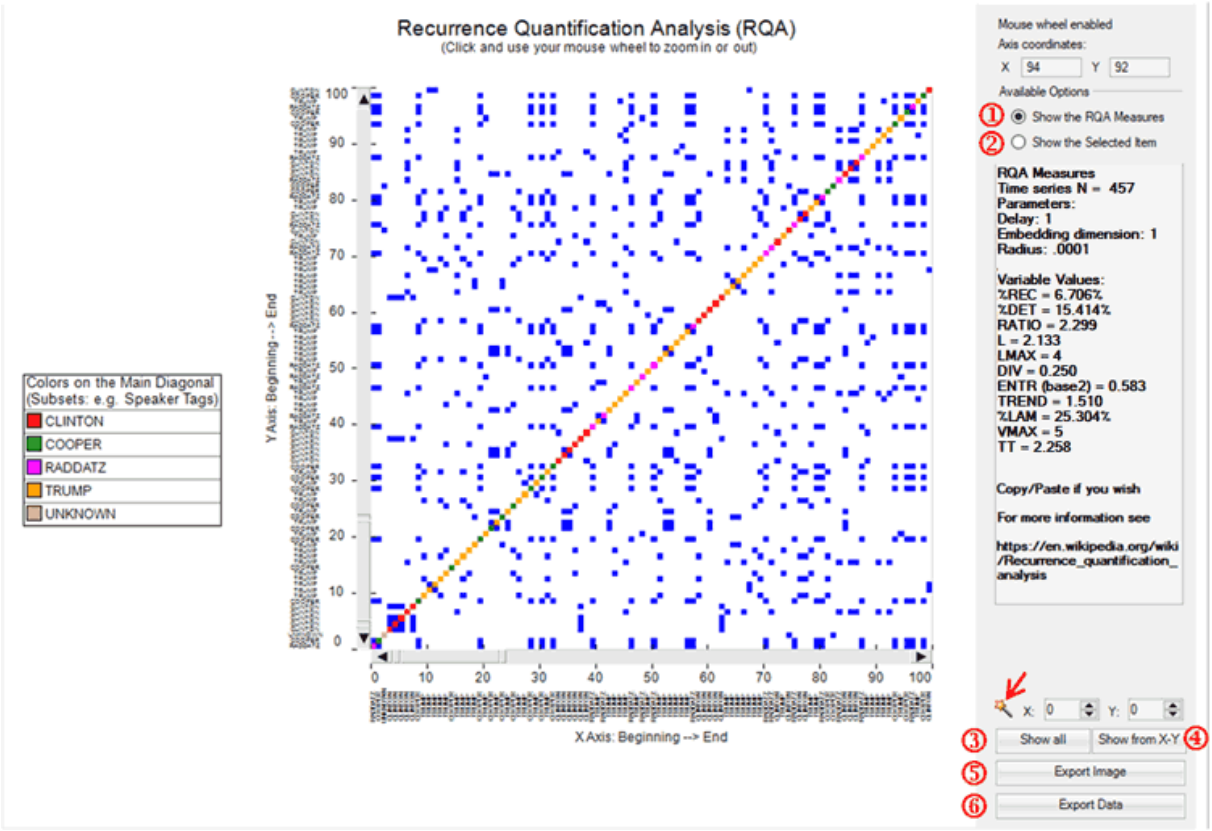
Fig2 - Similarities

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).



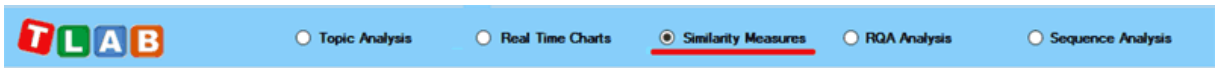


In particular:
-options ‘1’ and ‘2’ allow us to visualize the general measures (‘1’) or the transcript of the selected segment (‘2’);
-options ‘3’ and ‘4’ allow us to visualize the complete recurrence plot (‘3’) or a subsection of it (‘4’);
-options ‘5’ and ‘6’ allow us to export the image in different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

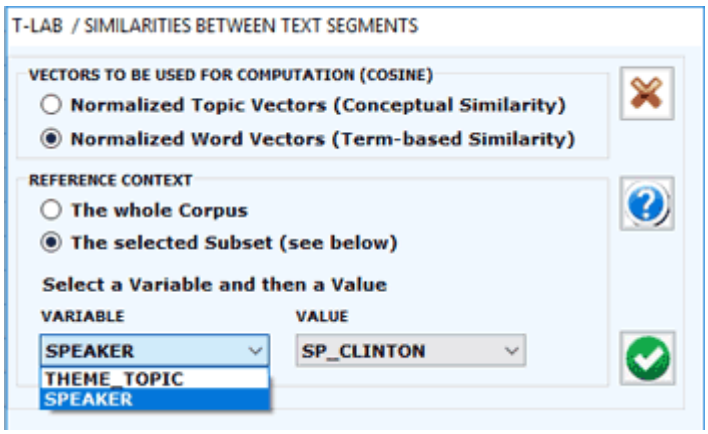
FIRST	SECOND	Cosine
SP_CLINTON	SP_CLINTON	0.0961
SP_CLINTON	SP_COOPER	0.1099
SP_CLINTON	SP_RADDATZ	0.1025
SP_CLINTON	SP_TRUMP	0.0847
SP_CLINTON	SP_UNKNOWN	0.1087
SP_COOPER	SP_CLINTON	0.1099
SP_COOPER	SP_COOPER	0.3106
SP_COOPER	SP_RADDATZ	0.2359
SP_COOPER	SP_TRUMP	0.1432
SP_COOPER	SP_UNKNOWN	0.1446
SP_RADDATZ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

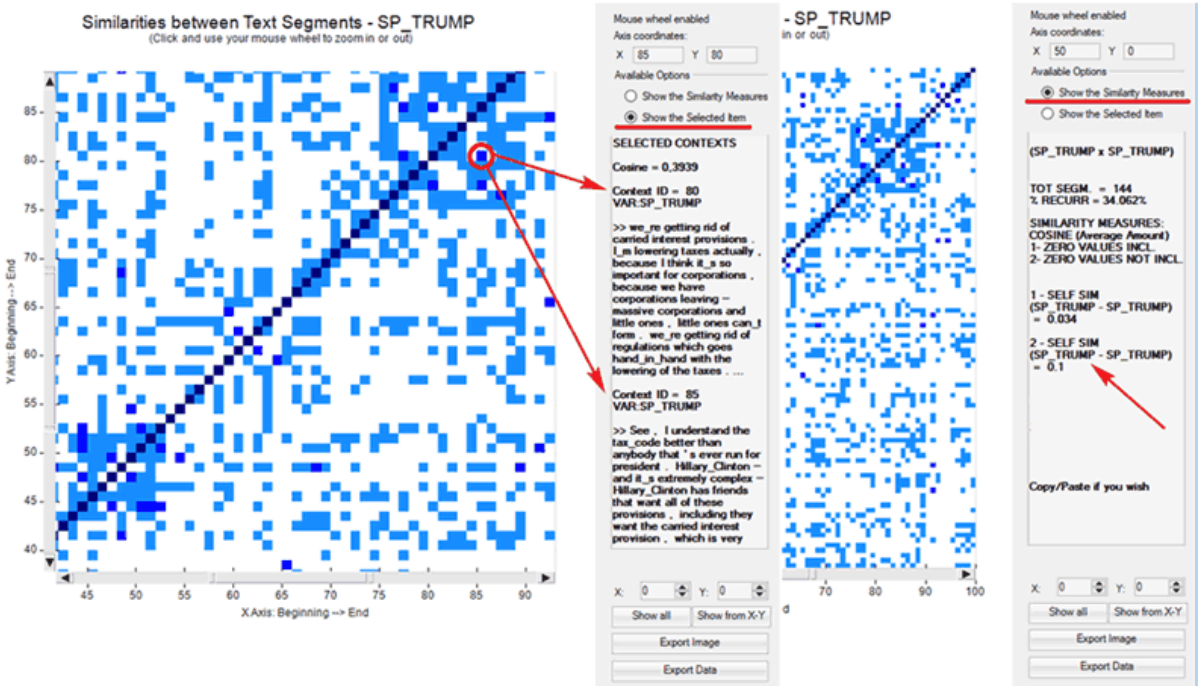


When choosing 'Similarity Measures', several options are made available (see picture below) which allow the user to select both the vectors to be used for the similarity computation and the reference context to be analysed (i.e. either the entire corpus or a subset of it).

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.

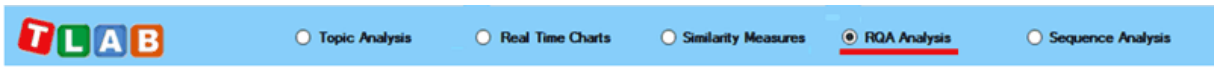


According to the design of the user interface, in this case - like in the RQA analysis (see section 'D' below) - the user can choose between visualizing the global measures or the transcripts of recurrent segments (see picture below). Moreover, when a corpus subset is selected, two further measures are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

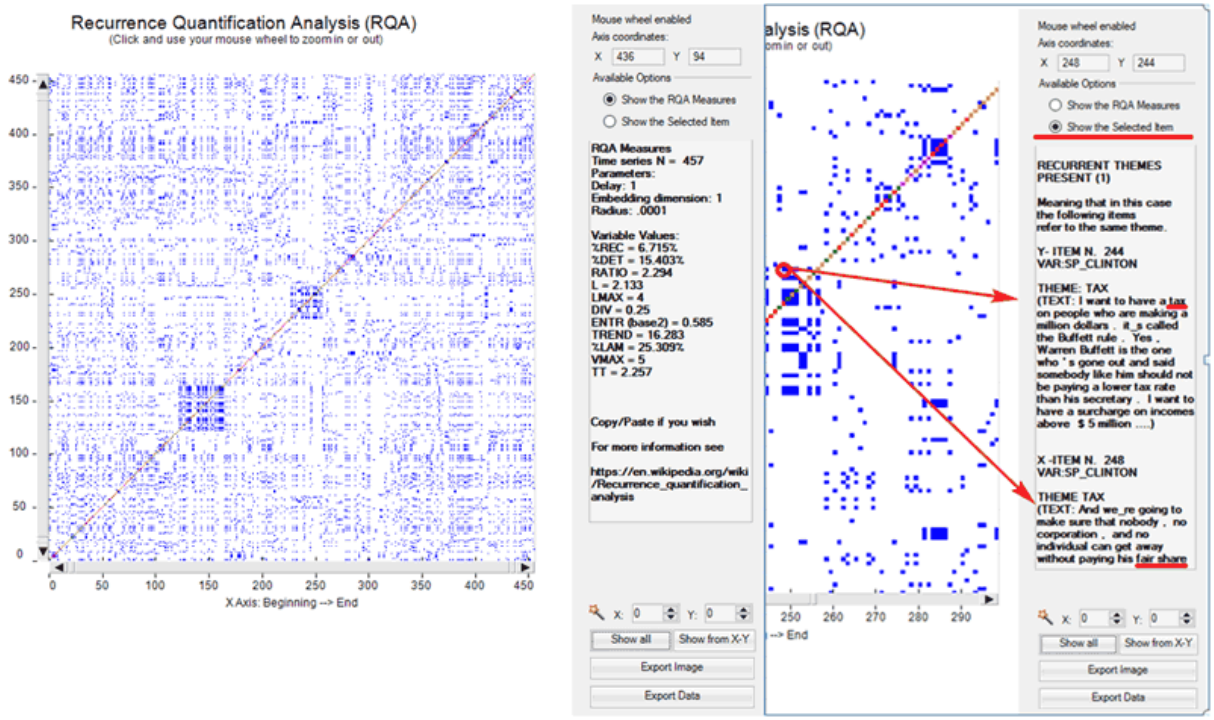
In this **T-LAB** tool, both in the case of the RQA Analysis and in the case of the Sequence Analysis (i.e. Markovian Analysis), a time series is represented by a categorical vector where each element is an integer which corresponds to the topic assigned to the 'i' text segment. However only in the case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

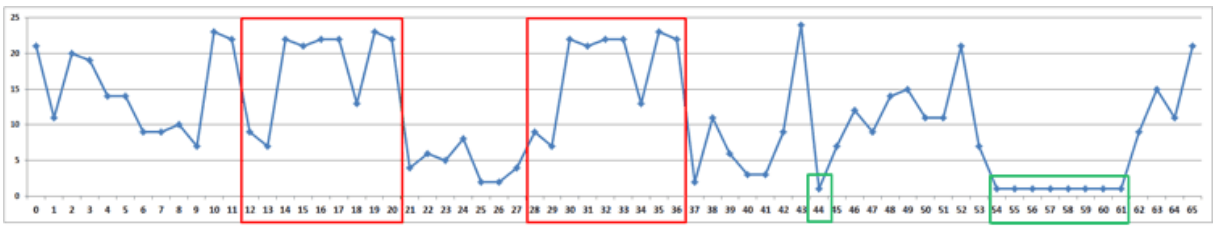
(**) For more information about the RQA measures see section 'E' below.



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.

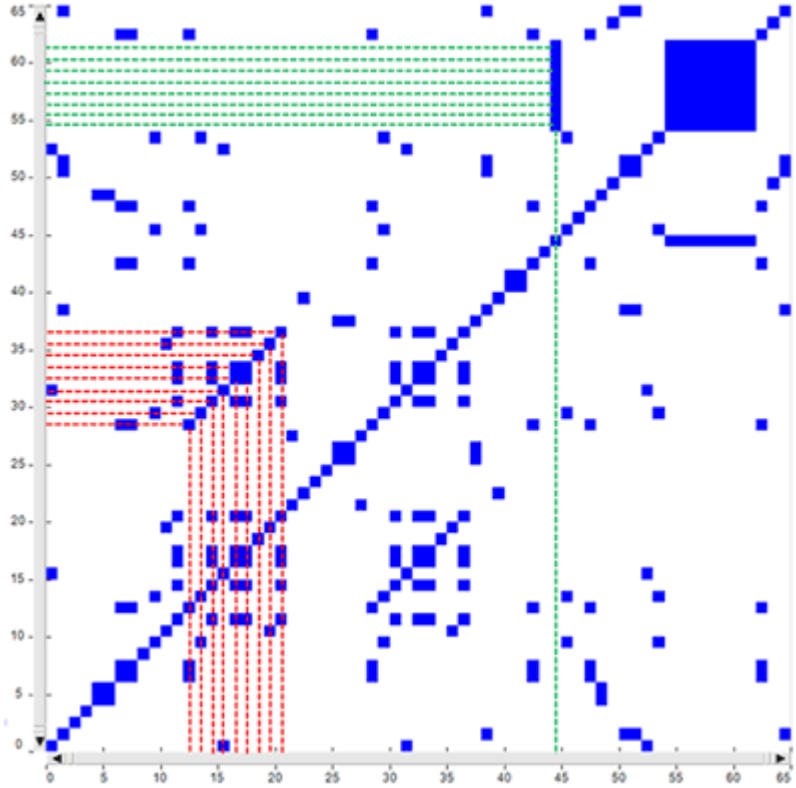
In particular, the distribution of diagonal lines allows for the investigation of **determinism** (i.e. the predictability of the system) and the distribution of vertical lines allows for the investigation of **intermittency** (i.e. the sequences which are interspersed by erratic breaks).



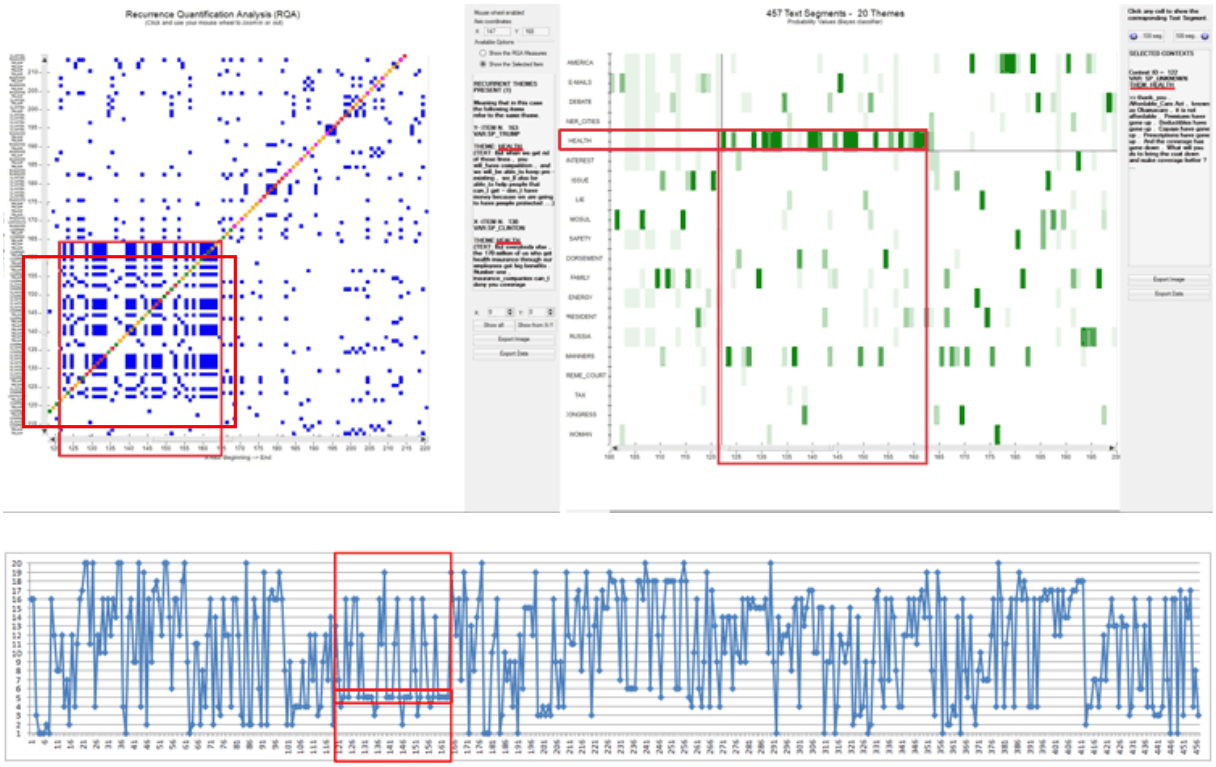
As an example, just consider the above fictitious time series. In it the same sequence of nine points/themes is repeated two times in different time spans (see the above red rectangles), respectively from t-12 to t-20 and from t-28 to t-36, where each 't' stands for a different text segment. In the same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

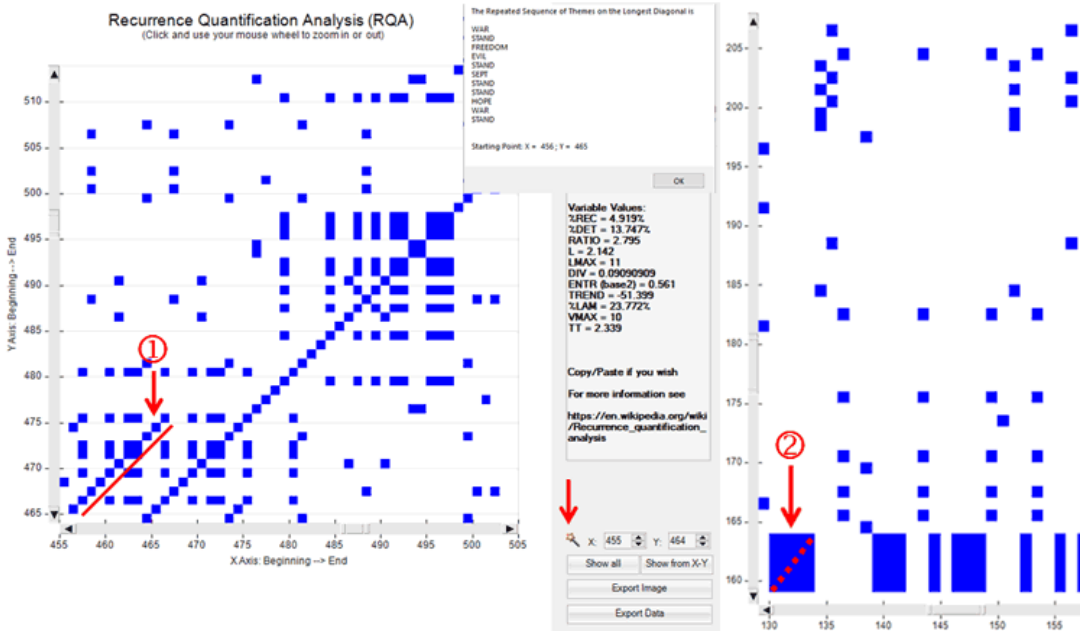


Regarding the **rectangular block** structures – which actually include both vertical and diagonal lines - they can be seen as referring to recurrences of the same topics in sub sections of the time series, i.e. to groups of overall similar feature vectors. In fact each dot in the graph represents a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

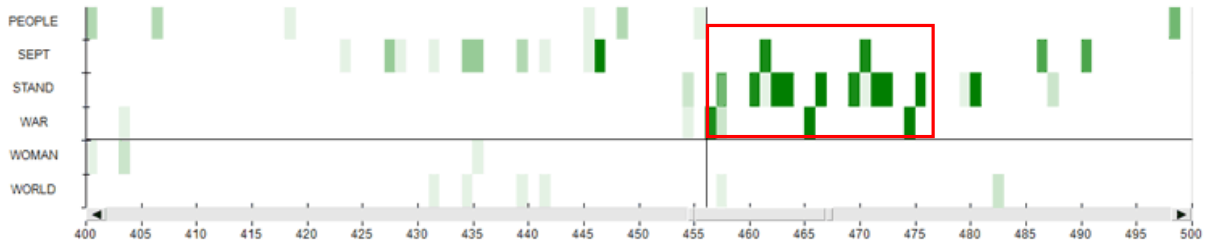


As stated above, in the RQA outputs the **longest diagonals** parallel to the main diagonal allow us to detect interesting repetitions of the same thematic sequence. However their shapes are not so evident as the rectangular block structures, also because sometimes they can be hidden inside one of them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

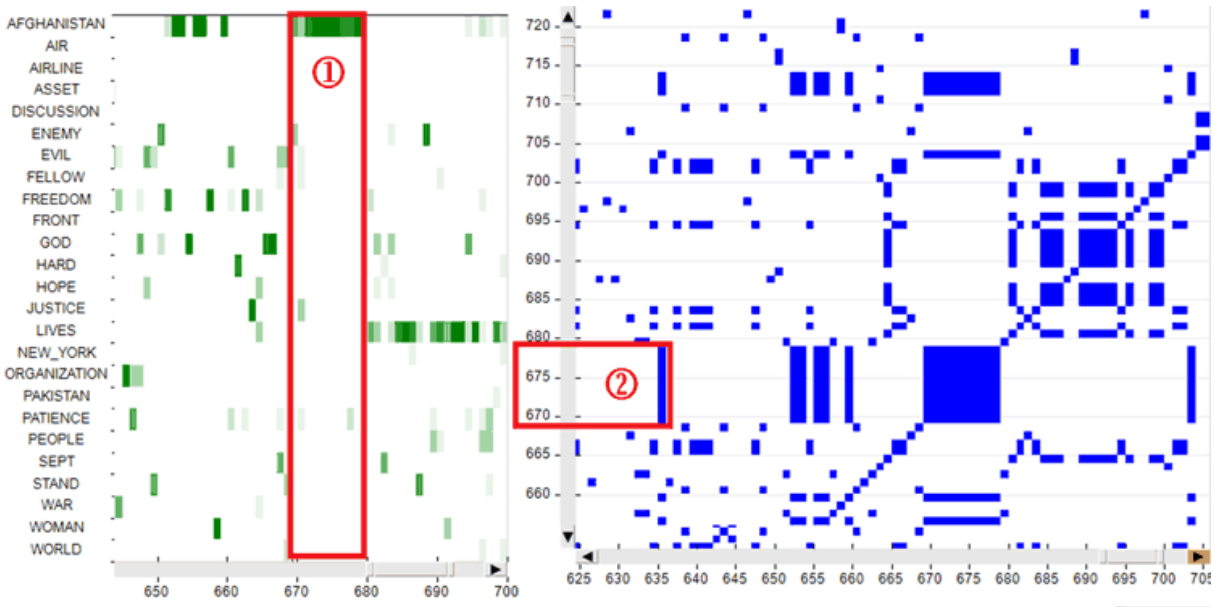
N.B.: Soon after the longest diagonal is detected **T-LAB** allows the user to export a file with the most frequent **repeated sequences**, each one of them including at least three concatenated themes. Such a file can be considered a sort of summary of the main themes - and of the corresponding variations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

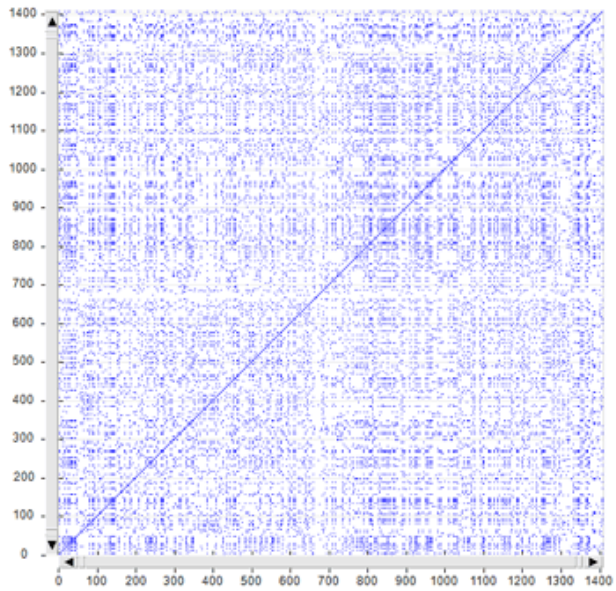
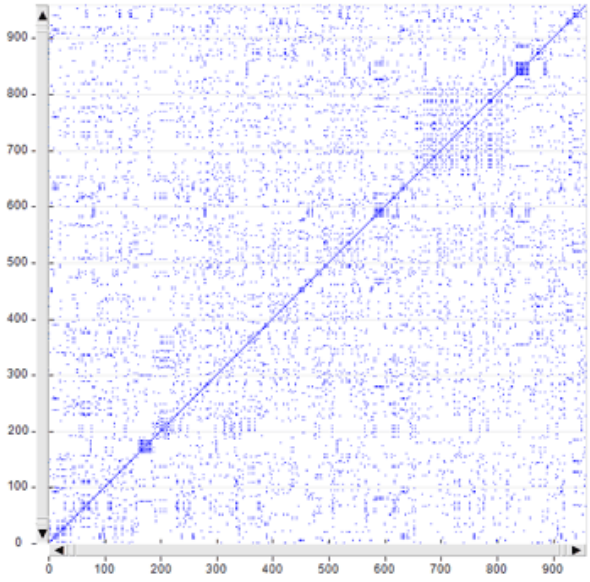
Measure	**Definition**
%REC - Recurrence Rate	The percentage of recurrence points in a Recurrence Plot which fall within a specified radius.
%DET - Determinism	The percentage of recurrence points which form diagonal line structures, main diagonal not included (N.B.: In RQA the main diagonal is also called 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% **%DET = 16.474%** RATIO = 2.009 L = 2.093 LMAX = 6 DIV = 0.167 ENTR (base2) = 0.460 **TREND = 4.705** %LAM = 30.717% VMAX = 7 TT = 2.263	**%REC = 3.525%** **%DET = 9.676%** RATIO = 2.745 L = 2.089 LMAX = 5 DIV = 0.2 ENTR (base2) = 0.435 **TREND = -5.599** %LAM = 23.194% VMAX = 6 TT = 2.267

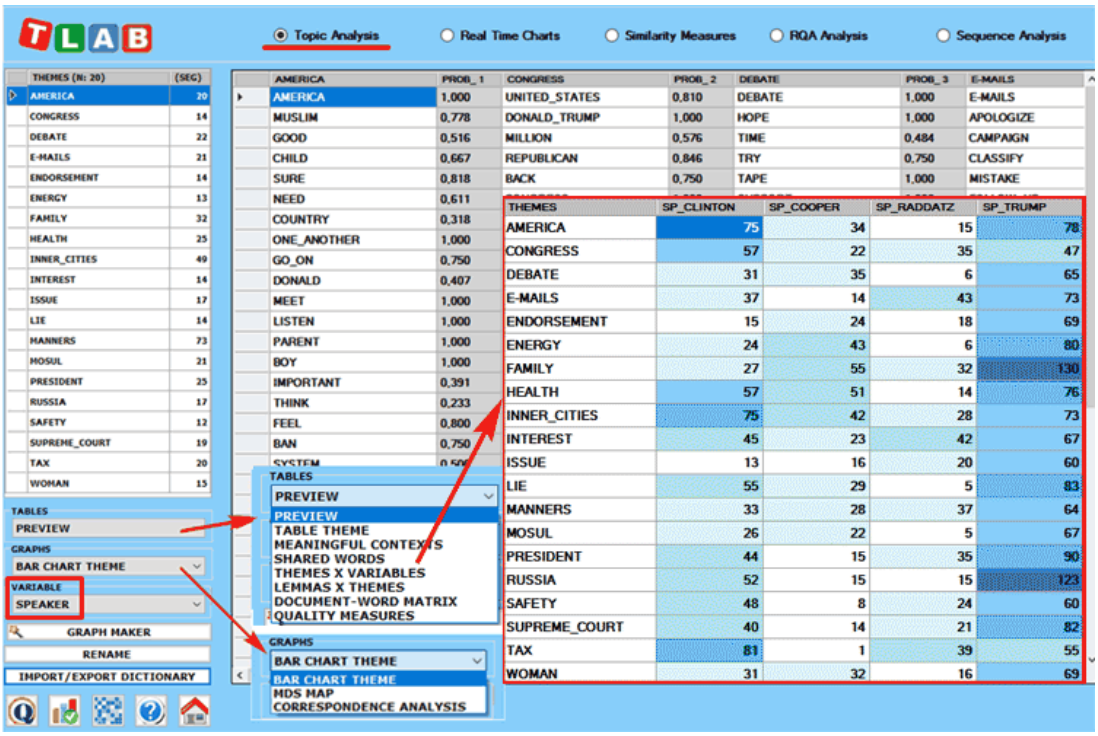
Here are the two corresponding recurrence plots.

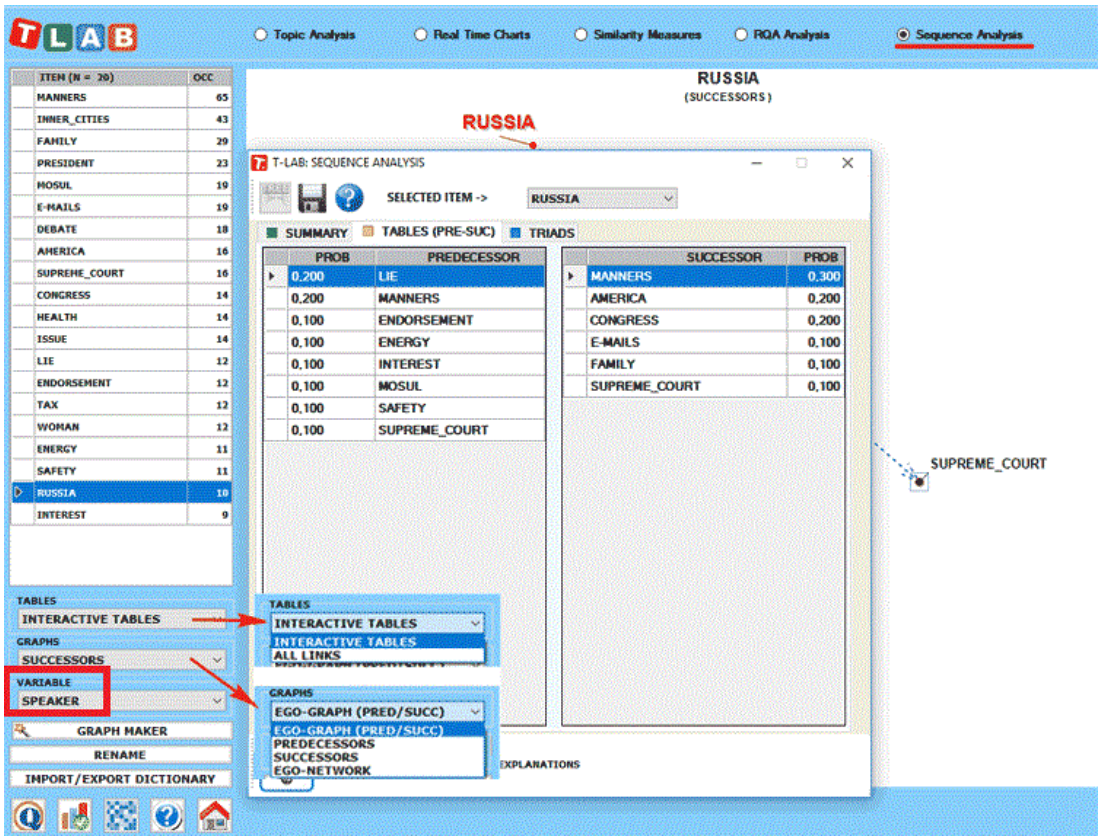
On the Origin of Species (a)	The adventures of Pinocchio (b)
	

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

F) Topic Analysis and Sequence Analysis

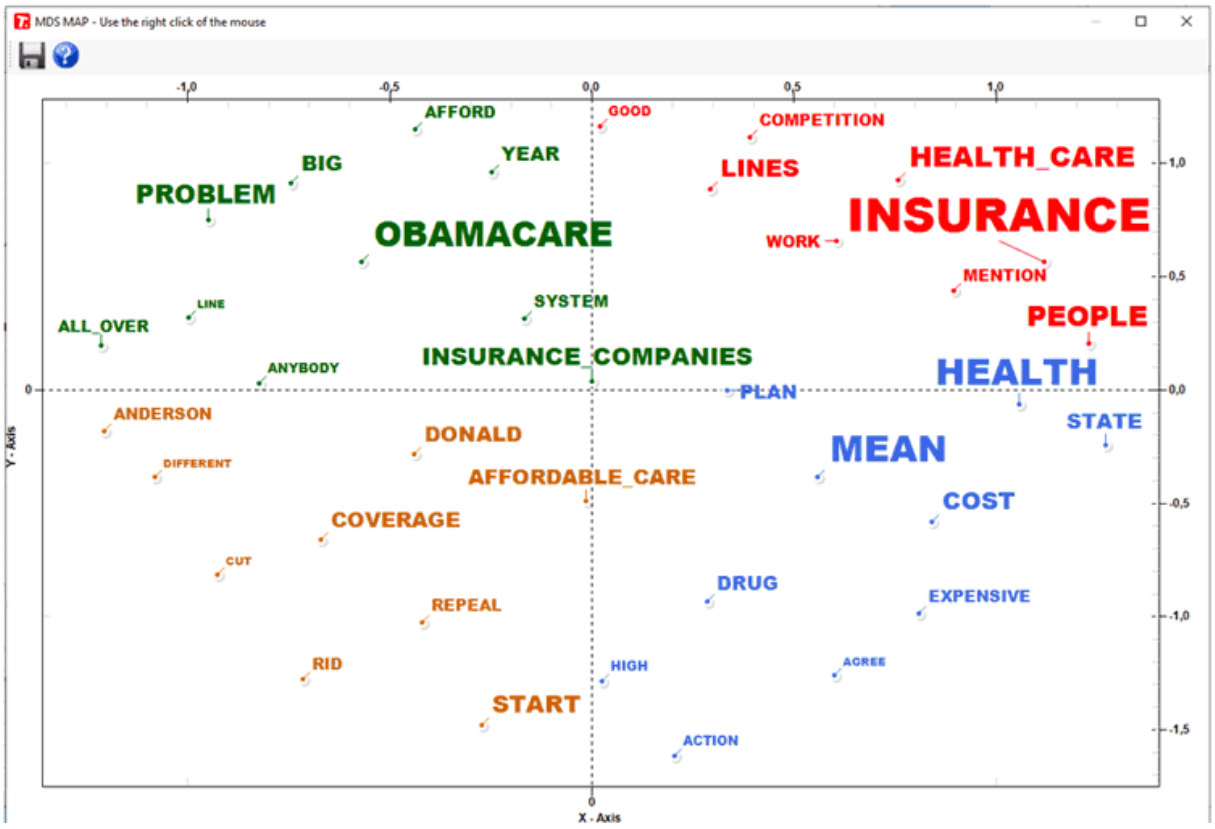
The below pictures summarize the main options of two tools already present in the **T-LAB** menu, which are integrated with the new ones and which are explained in the corresponding sections of this manual/help, i.e. 'Modeling of Emerging Themes' and 'Sequence and Network Analysis'.





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

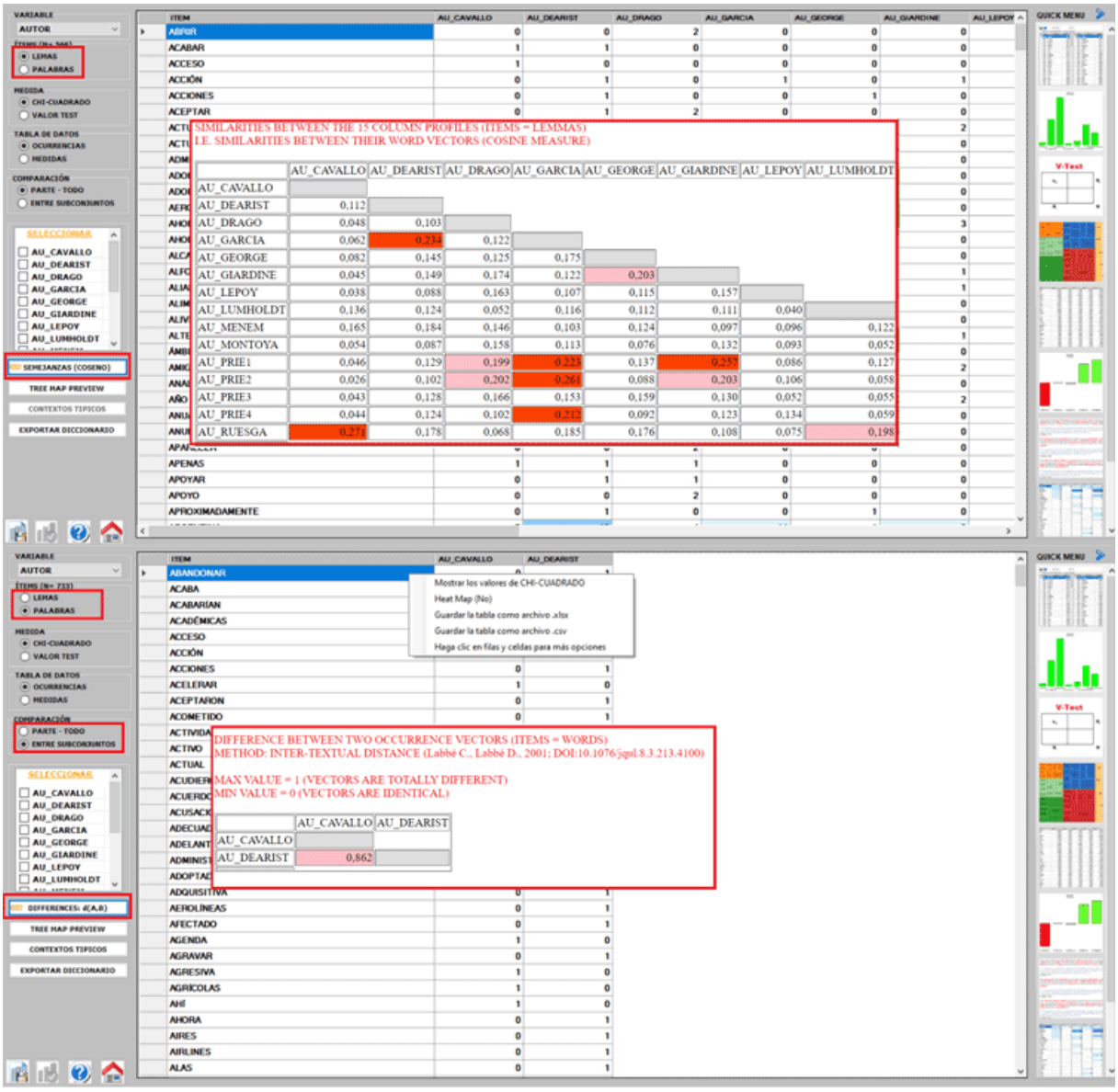
Topic Analysis
Real Time Charts
Similarity Measures
RQA Analysis
Sequence Analysis
GRAPH MAKER - CO-OCCURRENCES WITHIN THE CLUSTER N. <HEALTH>
ADD/REMOVE ITEMS TO BE USED
CLICK A PICTURE TO DISPLAY THE GRAPH
EXPORT DATA FILES FOR NETWORK ANALYSIS (up to 5.000 items can be selected at a time)
GRAPH MAKER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


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

ANÁLISIS COMPARATIVOS

Especificidades

NOTA: Las imagenes contenidas en este apartado hacen referencia a una versión anterior de T-LAB, ya que el interfaz de **T-LAB 10** cambia ligeramente. En particular, una galería de imágenes de acceso rápido que funciona como un menú adicional permite cambiar entre varias salidas con un solo clic.Además, el usuario puede evaluar fácilmente **similitudes** (es decir, coseno) y **diferencias** (es decir, Inter -Distancia textual) entre subconjuntos de corpus (de 2 a 150), y también para detectar documentos duplicados y casi duplicados (ver imágenes a continuación).



Esta herramienta de **T-LAB** permite verificar cuáles son las **unidades lexicales** (es decir, palabras, lemas o categorías) **típicas** o **exclusivas** de un texto o de un subconjunto del corpus definido por una variable categorial. Además, permite localizar las **unidades de contexto características** de los diferentes subconjuntos en análisis (por ejemplo las frases 'típicas' que mejor diferencian a los discursos de los diferentes líderes políticos).

Más en detalle:

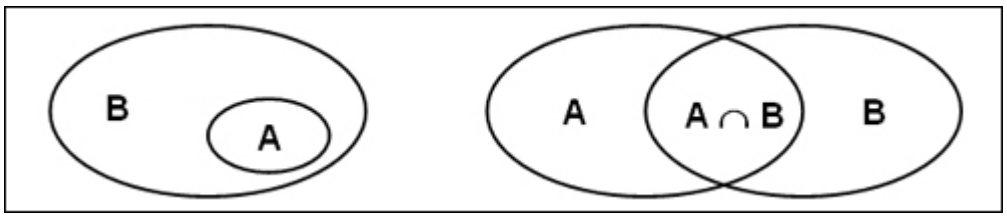
Las **unidades lexicales típicas**, definidas por las proporciones de las respectivas ocurrencias (es decir, por su sobre/sub-utilización, exceso/carencia de uso), se eligen en base al cálculo del **Chi-cuadrado** o del **valor-test**.

Las **unidades de contexto características** se obtienen calculando y sumando los valores TF-IDF normalizados asignados a las palabras que componen cada frase o párrafo.

El análisis de especificidades permite realizar dos tipos de **comparaciones** concernientes filas y columnas de las tablas de contingencia:

1- entre una **parte** (ej. el subconjunto "A") y el **todo** (ej. todo el corpus analizado, "B");

2- entre parejas de **subconjuntos** del corpus ("A" y "B").



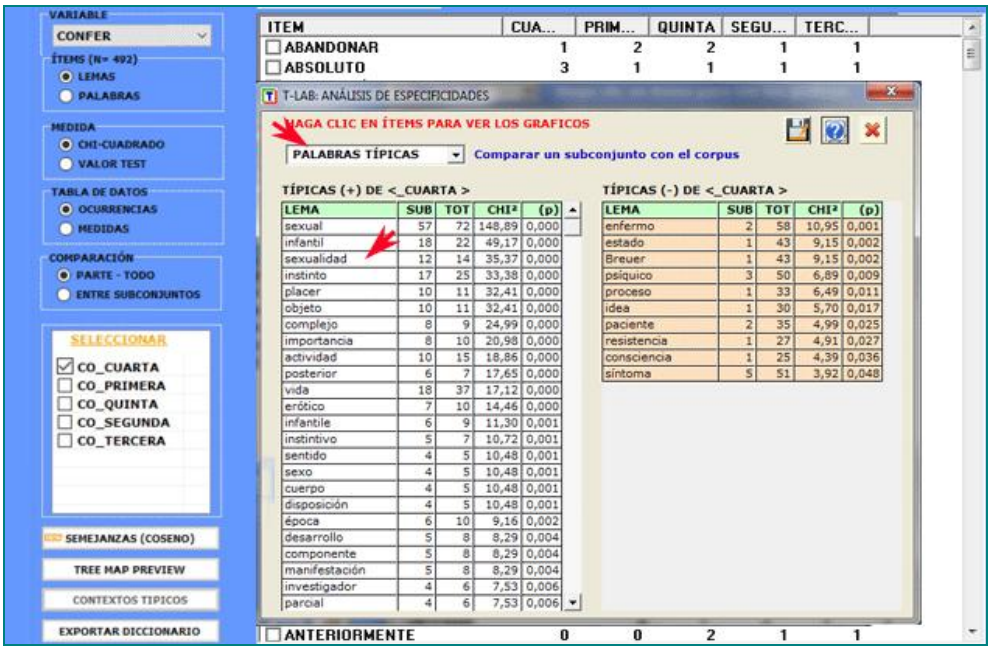
En ambos casos pueden ser analizadas tanto las Especificidades relativas a las **intersecciones**, como las relativas a las **diferencias**.

Las modalidades del cálculo se muestran en la entrada correspondiente del **glosario**.

Las unidades lexicales consideradas pueden ser todas (configuración automática) o solamente las seleccionadas por el usuario (configuración personalizada).

En sucesión, los cuatro tipos de comparaciones posibles:

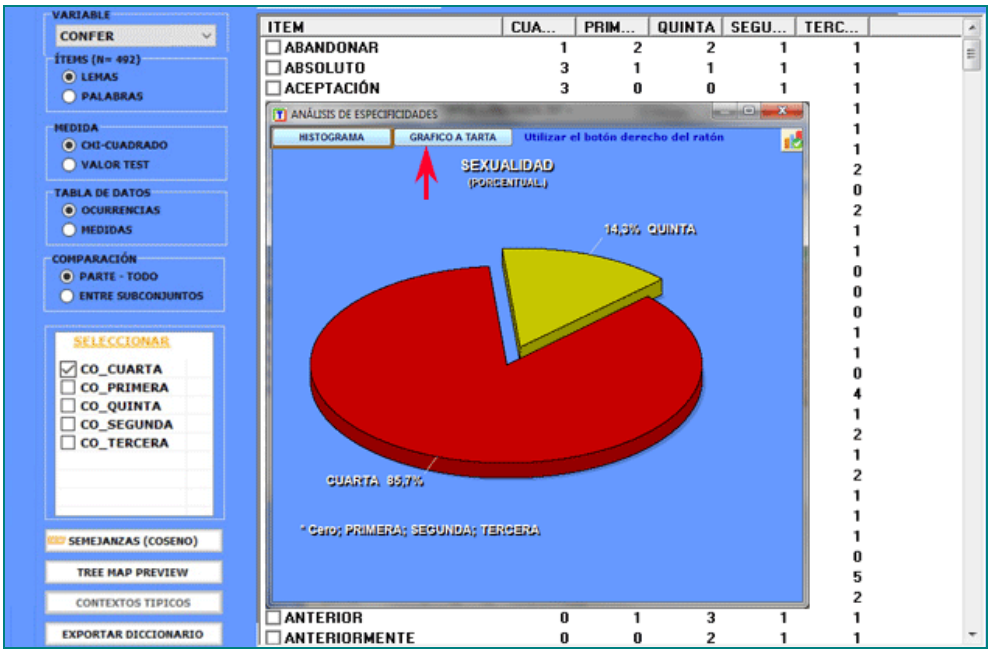
1.1 - **parte/todo**: unidades lexicales "típicas"

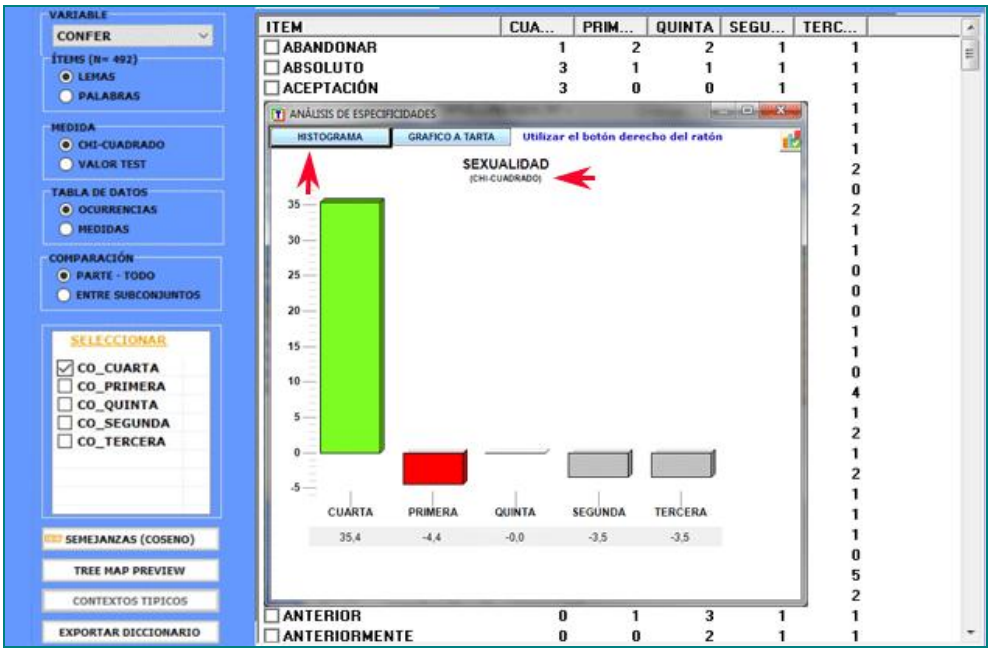


Las llaves de lectura son las siguientes:

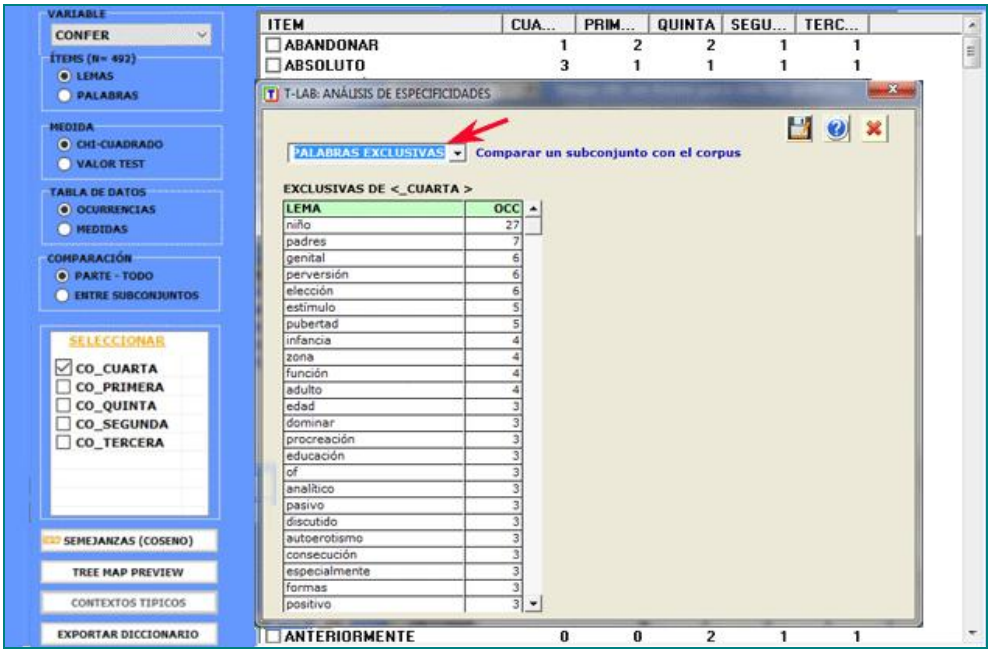
- LEMA = unidades lexicales típicas (en exceso o en defecto);
- CHI2 = valor del chi cuadrado (o VTEST = Valor Test);
- SUB = ocurrencias de cada LEMMA en el Subconjunto;
- TOT = ocurrencias de cada LEMMA en el Corpus o en los dos Subconjuntos (véase 2.1);
- CHI2 = valor del chi cuadrado (o VTEST = Valor Test);
- (p) = probabilidad asociada a cada valor del chi-cuadrado (def=1).

Haciendo clic en los ítems de las tablas, es posible crear varios tipos de gráficos**.**

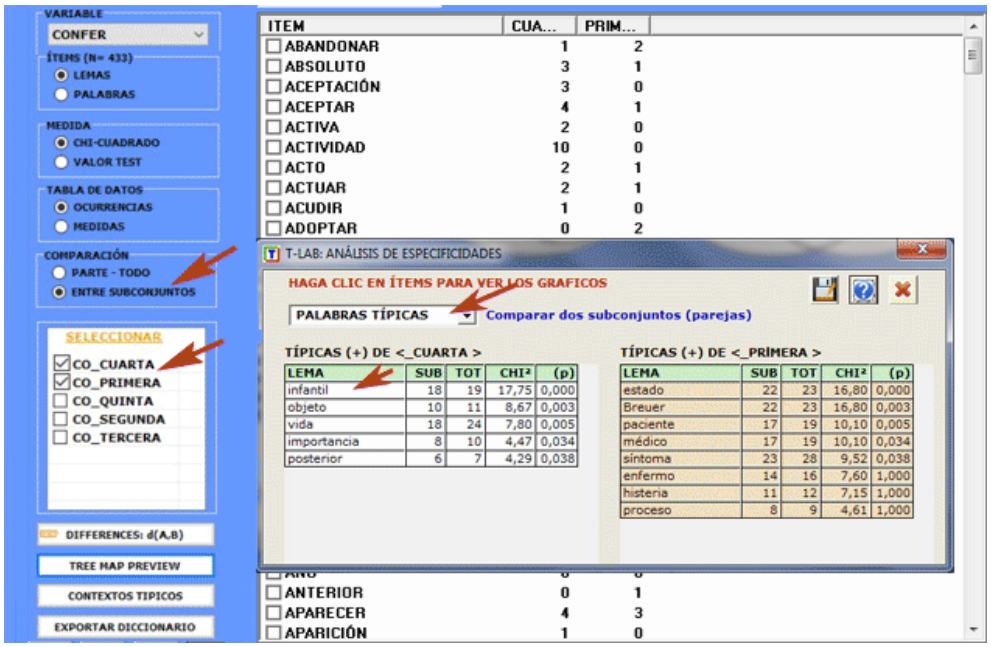

VARIABLE
CONFER
ÍTEMS (N= 492)
LEMAS
PALABRAS
MEDIDA
CHI-CUADRADO
VALOR TEST
TABLA DE DATOS
OCURRENCIAS
MEDIDAS
COMPARACIÓN
PARTE - TODO
ENTRE SUBCONJUNTOS
SELECCIONAR
CO_CUARTA
CO_PRIMERA
CO_QUINTA
CO_SEGUNDA
CO_TERCERA
SEMEJANZAS (COSENO)
TREE MAP PREVIEW
CONTEXTOS TIPICOS
EXPORTAR DICCIONARIO
ITEM
CUA...
PRIM...
QUINTA
SEGU...
TERC...
ABANDONAR 1 2 2 1 1
ABSOLUTO 3 1 1 1 1
ACEPTACIÓN 3 0 0 1 1
ANÁLISIS DE ESPECIFICIDADES
HISTOGRAMA
GRAFICO A TARTA
Utilizar el botón derecho del ratón
SEXUALIDAD
(PORCENTUAL)
14,3% QUINTA
CUARTA 85,7%
* Cero: PRIMERA; SEGUNDA; TERCERA
ANTERIOR 0 1 3 1 1
ANTERIORMENTE 0 0 2 1 1

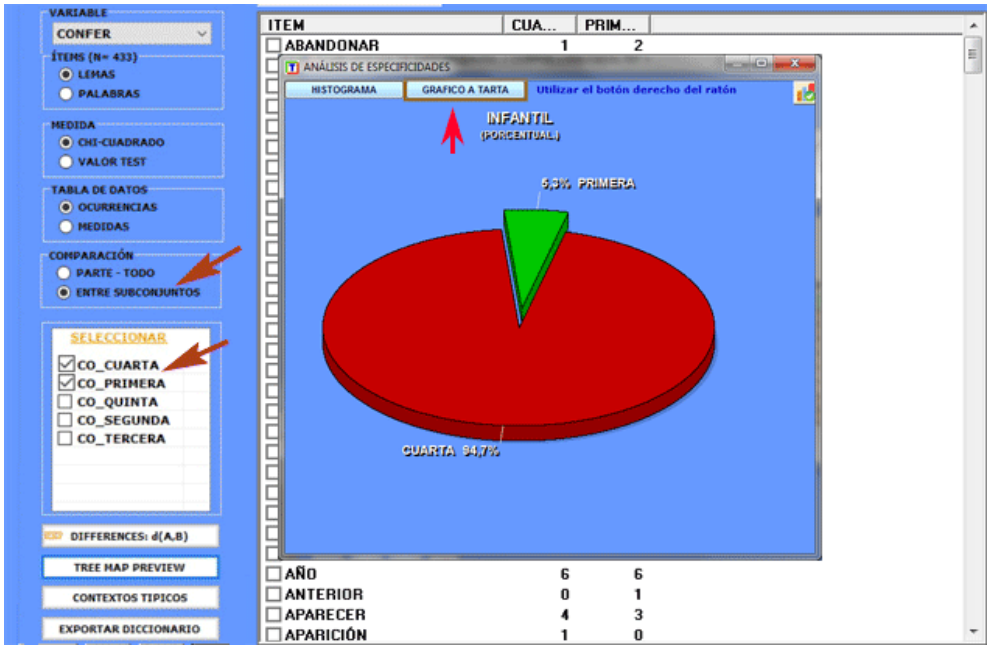

VARIABLE
CONFER
ÍTEMS (N= 492)
LEMAS
PALABRAS
MEDIDA
CHI-CUADRADO
VALOR TEST
TABLA DE DATOS
OCURRENCIAS
MEDIDAS
COMPARACIÓN
PARTE - TODO
ENTRE SUBCONJUNTOS
SELECCIONAR
CO_CUARTA
CO_PRIMERA
CO_QUINTA
CO_SEGUNDA
CO_TERCERA
SEMEJANZAS (COSENO)
TREE MAP PREVIEW
CONTEXTOS TIPICOS
EXPORTAR DICCIONARIO
ITEM
CUA...
PRIM...
QUINTA
SEGU...
TERC...
ABANDONAR 1 2 2 1 1
ABSOLUTO 3 1 1 1 1
ACEPTACIÓN 3 0 0 1 1
ANÁLISIS DE ESPECIFICIDADES
HISTOGRAMA
GRAFICO A TARTA
Utilizar el botón derecho del ratón
SEXUALIDAD
(CHI-CUADRADO)
35
30
25
20
15
10
5
0
-5
CUARTA
PRIMERA
QUINTA
SEGUNDA
TERCERA
35,4
-4,4
-0,0
-3,5
-3,5
ANTERIOR 0 1 3 1 1
ANTERIORMENTE 0 0 2 1 1

1.2 - **parte/todo**: unidades lexicales "exclusivas"

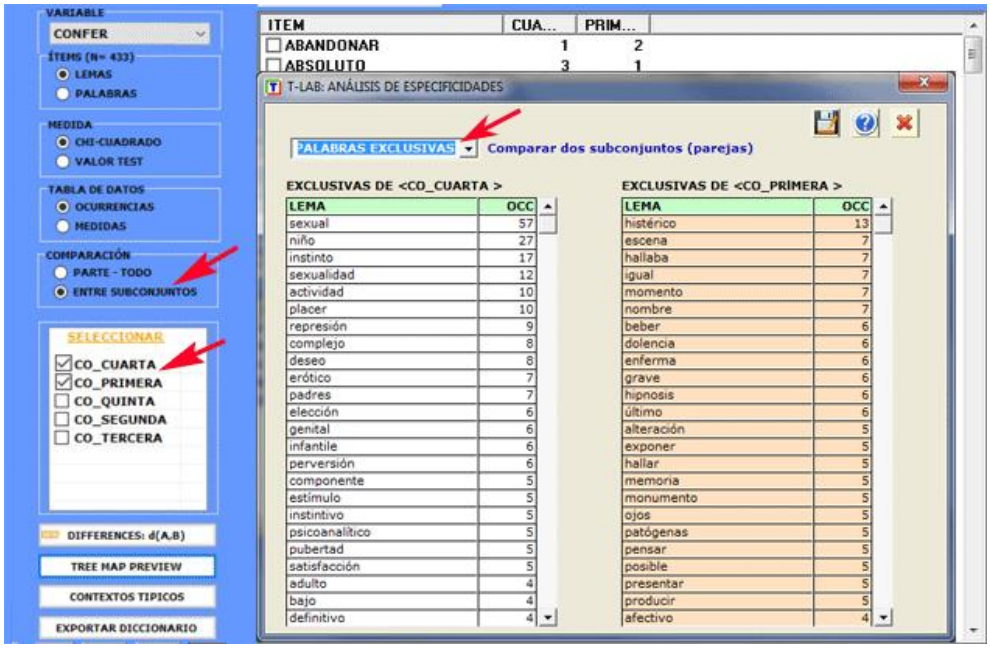


2.1- **subconjunto/subconjunto**: unidades lexicales "típicas"

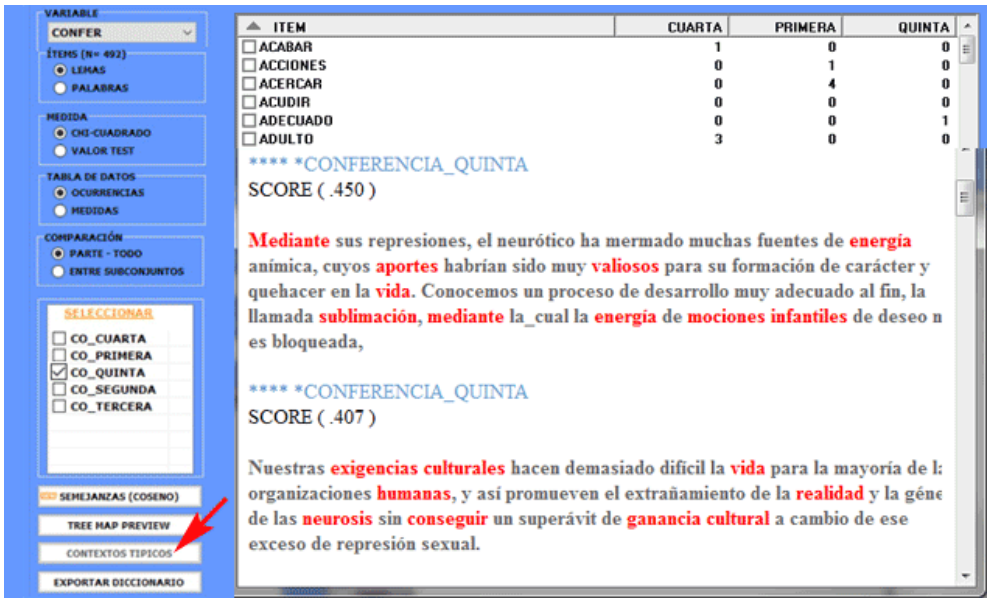




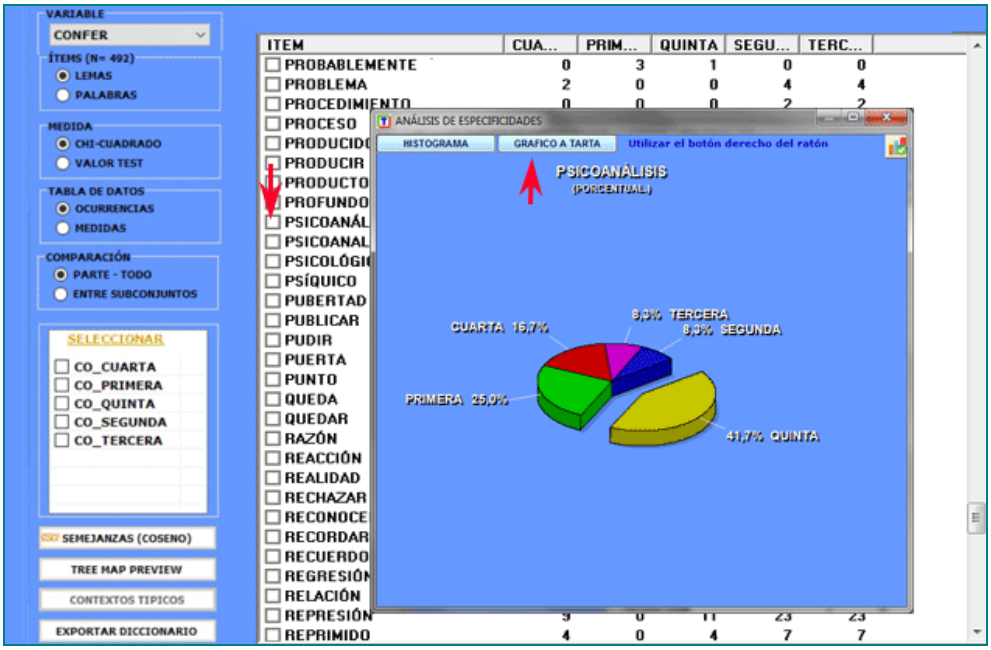
2.2 - **subconjunto/subconjunto**: unidades lexicales "exclusivas"

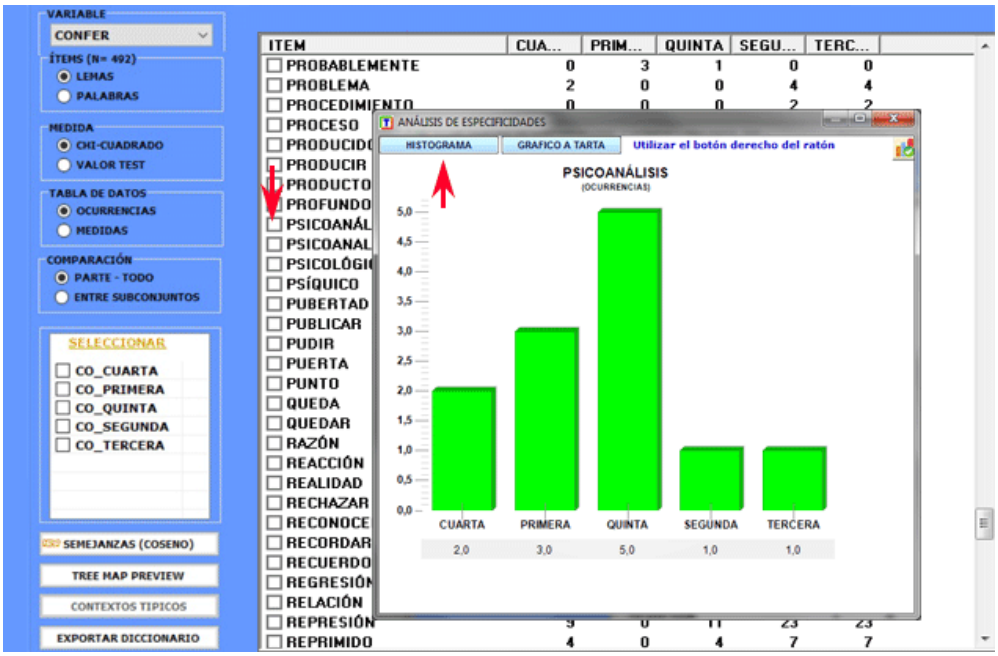


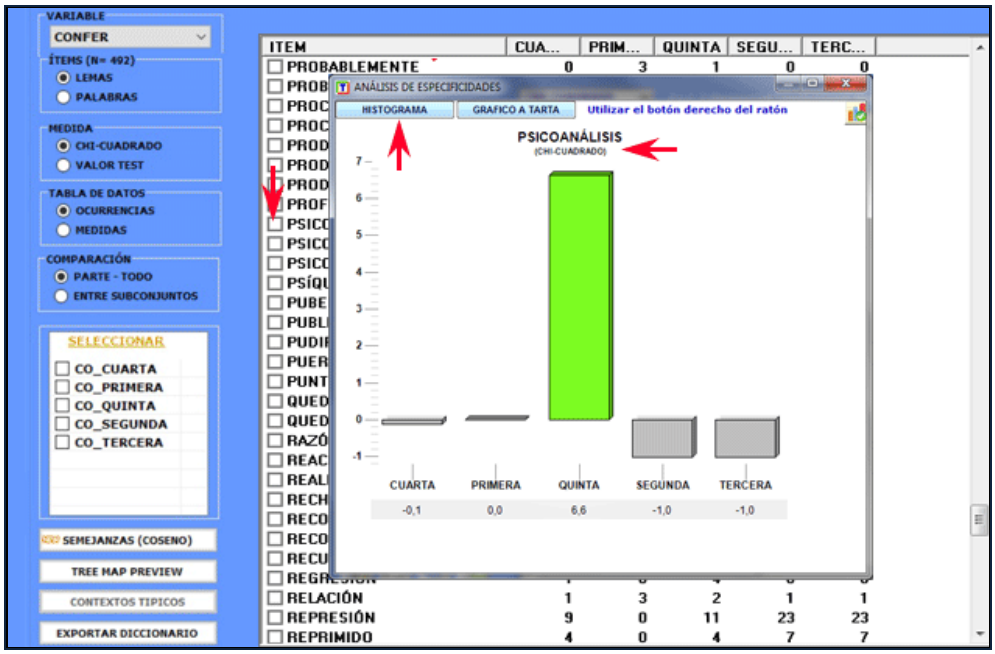
En cada subconjunto analizado es posible verificar cuáles son los contextos elementales (es decir, frases o párrafos) que mejor lo distinguen de los demás. En este caso, la 'especificidad' llega del cálculo de los valores TF-IDF normalizados. Más en concreto, el score asignada a cada contexto elemental (véase imagen siguiente) es el resultado de la suma de los valores TF-IDF de las palabras que lo componen.

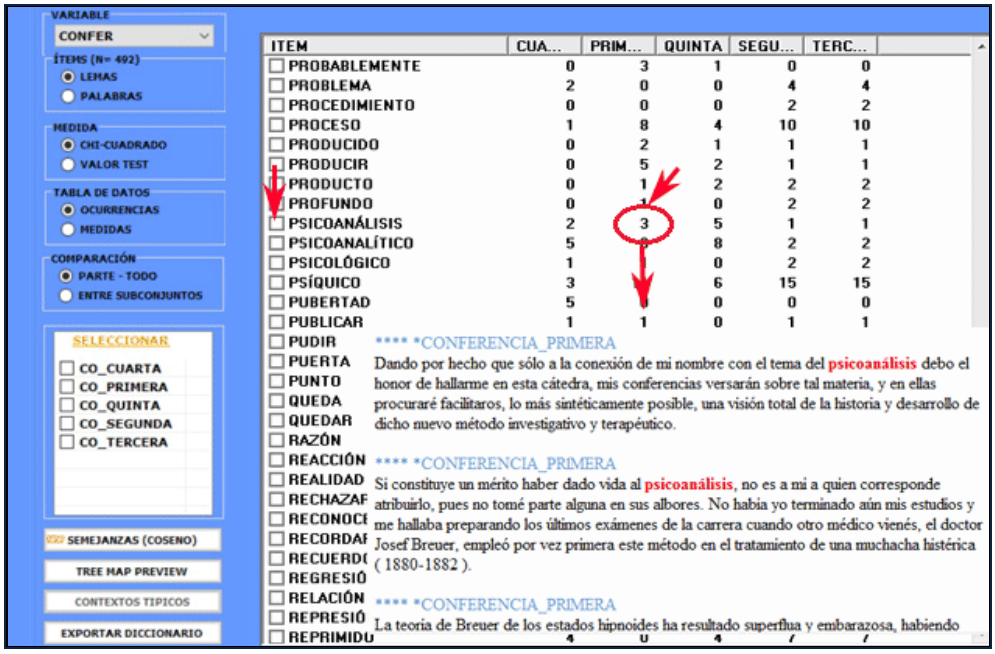


Todas las tablas de contingencia pueden ser fácilmente exploradas y nos permiten crear varios tipos de gráficos. Además, haciendo clic en específicas células de la tabla (véase abajo), es posible crear un archivo HTML que incluye todos los contextos elementales en que la palabra en la fila está presente en el subconjunto correspondiente.

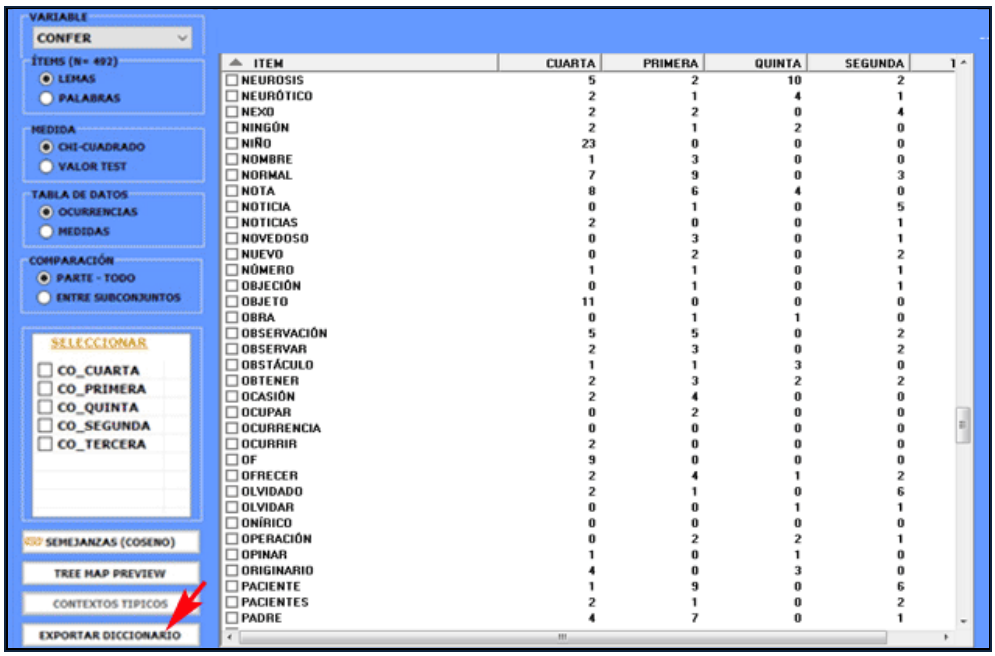



VARIABLE
CONFER
ÍTEMS (N= 492)
LEMAS
PALABRAS
MEDIDA
CHI-CUADRADO
VALOR TEST
TABLA DE DATOS
OCURRENCIAS
MEDIDAS
COMPARACIÓN
PARTE - TODO
ENTRE SUBCONJUNTOS
SELECCIONAR
CO_CUARTA
CO_PRIMERA
CO_QUINTA
CO_SEGUNDA
CO_TERCERA
SEMEJANZAS (COSENO)
TREE MAP PREVIEW
CONTEXTOS TIPICOS
EXPORTAR DICCIONARIO
ITEM
CUA...
PRIM...
QUINTA
SEGU...
TERC...
PROBABLEMENTE 0 3 1 0 0
PROBLEMA 2 0 0 4 4
PROCEDIMIENTO 0 0 0 2 2
PROCESO
PRODUCIDO
PRODUCIR
PRODUCTO
PROFUNDO
PSICOANÁL
PSICOANAL
PSICOLÓGIC
PSÍQUICO
PUBERTAD
PUBLICAR
PUDIR
PUERTA
PUNTO
QUEDA
QUEDAR
RAZÓN
REACCIÓN
REALIDAD
RECHAZAR
RECONOCE
RECORDAR
RECUERDO
REGRESIÓN
RELACIÓN
REPRESIÓN 9 0 11 23 23
REPRIMIDO 4 0 4 7 7
ANÁLISIS DE ESPECIFICIDADES
HISTOGRAMA
GRAFICO A TARTA
Utilizar el botón derecho del ratón
PSICOANÁLISIS
(OCURRENCIAS)
5,0
4,5
4,0
3,5
3,0
2,5
2,0
1,5
1,0
0,5
0,0
CUARTA
PRIMERA
QUINTA
SEGUNDA
TERCERA
2,0
3,0
5,0
1,0
1,0


VARIABLE
CONFER
ÍTEMS (N= 492)
LEMAS
PALABRAS
MEDIDA
CHI-CUADRADO
VALOR TEST
TABLA DE DATOS
OCURRENCIAS
MEDIDAS
COMPARACIÓN
PARTE - TODO
ENTRE SUBCONJUNTOS
SELECCIONAR
CO_CUARTA
CO_PRIMERA
CO_QUINTA
CO_SEGUNDA
CO_TERCERA
SEMEJANZAS (COSENO)
TREE MAP PREVIEW
CONTEXTOS TIPICOS
EXPORTAR DICCIONARIO
ITEM
CUA...
PRIM...
QUINTA
SEGU...
TERC...
PROBABLEMENTE 0 3 1 0 0
RELACIÓN 1 3 2 1 1
REPRESIÓN 9 0 11 23 23
REPRIMIDO 4 0 4 7 7
ANÁLISIS DE ESPECIFICIDADES
HISTOGRAMA
GRAFICO A TARTA
Utilizar el botón derecho del ratón
PSICOANÁLISIS
(CHI-CUADRADO)
7
6
5
4
3
2
1
0
-1
CUARTA
PRIMERA
QUINTA
SEGUNDA
TERCERA
-0,1
0,0
6,6
-1,0
-1,0

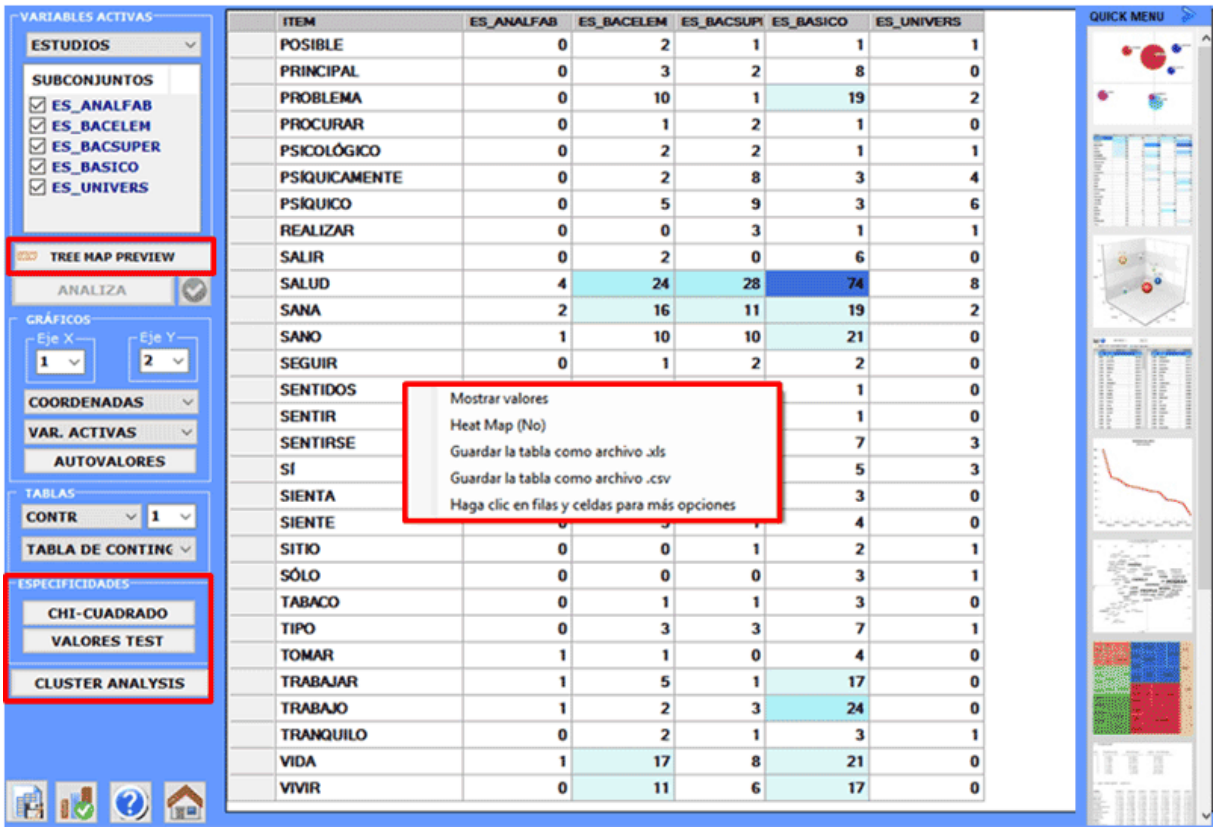


Finalmente, haciendo clic en la opción correspondiente (véase abajo), se genera un archivo **diccionario** con la extensión .dictio, que está listo para ser importado por cualquier herramienta de **T-LAB** para el **análisis temático**. Tal diccionario incluye todas las palabras típicas de la variable categórica seleccionada.



Análisis de Correspondencias

NOTA: Las imagenes contenidas en este apartado hacen a una versión anterior de T-LAB, ya que el interfaz de **T-LAB 10** cambia ligeramente. Además: a) el uso del **botón derecho** del ratón sobre las tablas que incluyen las palabras clave permite acceder a otras opciones; b) una nueva herramienta (**GRAPH MAKER**) permite crear y exportar diferentes tipos de gráficos dinámicos en formato HTML; c) dos nuevos botones nos permiten verificar las especificidades de cada valor de variable utilizando la prueba de chi-cuadrado o el valor test d) se incluye un botón que permite implementar un **análisis de clúster** y que utiliza las coordenadas de los objetos (unidades lexicales o de contexto) relativas a los primeros ejes factoriales (hasta un máximo de 10); e) se pueden visualizar las tablas de contingencia en modalidad 'head-map'; f) las palabras pueden ser representadas en los gráficos utilizando un tamaño de letras proporcional a la cantidad de ocurrencias que las caracterizan; g) una galería de imágenes de acceso rápido que funciona como un menú adicional permite cambiar entre varias salidas con un solo clic.
Algunas de estas nuevas características se destacan en la imagen de abajo.



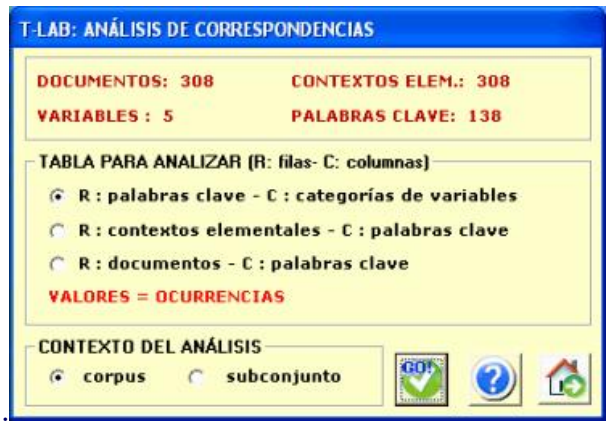
ITEM	ES_ANALFAB	ES_BACELEM	ES_BACSUP	ES_BASICO	ES_UNIVERS
POSIBLE	0	2	1	1	1
PRINCIPAL	0	3	2	8	0
PROBLEMA	0	10	1	19	2
PROCURAR	0	1	2	1	0
PSICOLÓGICO	0	2	2	1	1
PSÍQUICAMENTE	0	2	8	3	4
PSÍQUICO	0	5	9	3	6
REALIZAR	0	0	3	1	1
SALIR	0	2	0	6	0
SALUD	4	24	28	74	8
SANA	2	16	11	19	2
SANO	1	10	10	21	0
SEGUIR	0	1	2	2	0
SENTIDOS				1	0
SENTIR				1	0
SENTIRSE				7	3
SÍ				5	3
SIENTA				3	0
SIENTE				4	0
SITIO	0	0	1	2	1
SÓLO	0	0	0	3	1
TABACO	0	1	1	3	0
TIPO	0	3	3	7	1
TOMAR	1	1	0	4	0
TRABAJAR	1	5	1	17	0
TRABAJO	1	2	3	24	0
TRANQUILO	0	2	1	3	1
VIDA	1	17	8	21	0
VIVIR	0	11	6	17	0

Esta herramienta de **T-LAB** tiene como finalidad la de destacar las **semejanzas** y **diferencias** entre **unidades del contexto**.

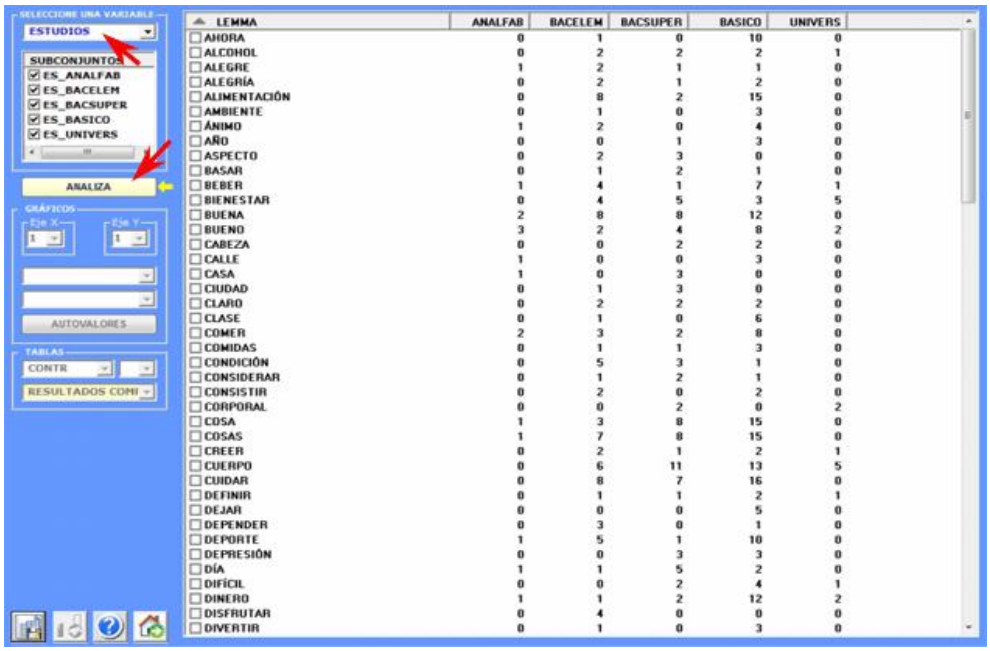
En particular, en **T-LAB**, el Análisis de Correspondencias permite analizar tre tipos de tablas:

(**A**) tablas palabras por categorías de variables con los valores de **ocurrencias**;
(**B**) tablas contextos elementales por palabras con los valores de **co-ocurrencias**;
(**C**) tablas documentos por palabras con los valores de **ocurrencias**.

.

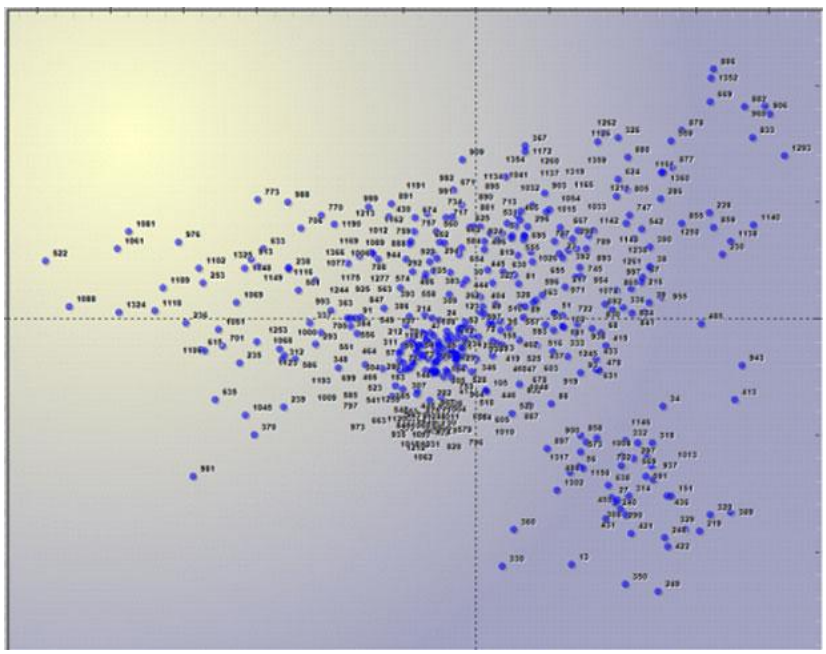


Para analizar las tablas (A) lemas (or palabras) por variables, el corpus se debe componer de un mínimo de tres textos o debe ser codificado con algunas variables (no menos de tres categorías).
Las variables son enumeradas en un box apropiado y pueden ser usadas de una en una.
Después de cada selección, en secuencia, se muestra la tabla de contingencia y hay que hacer clic en el botón **analiza** (véase abajo).

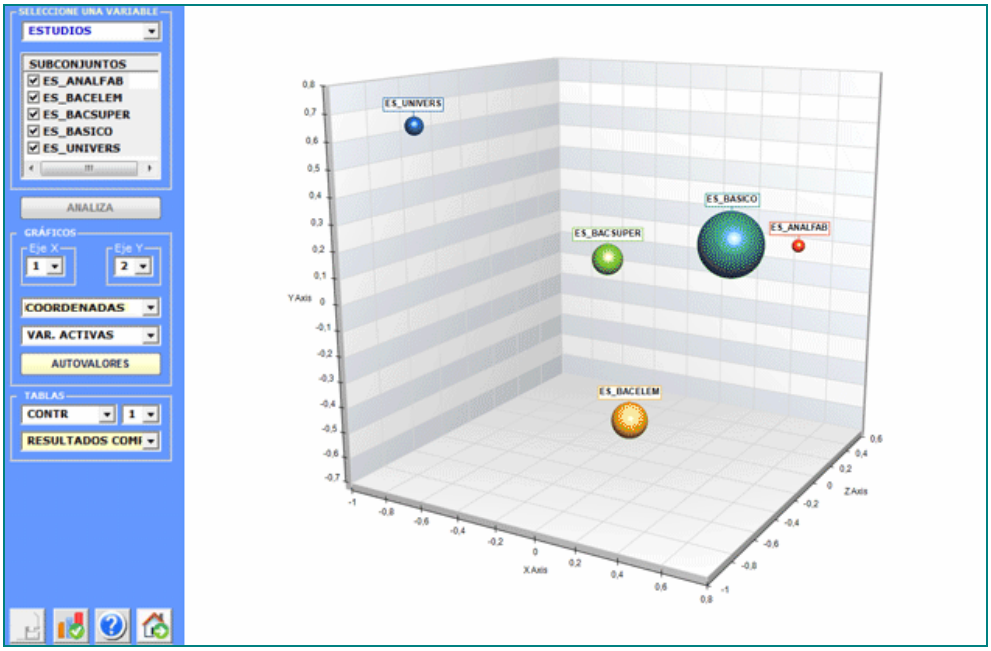


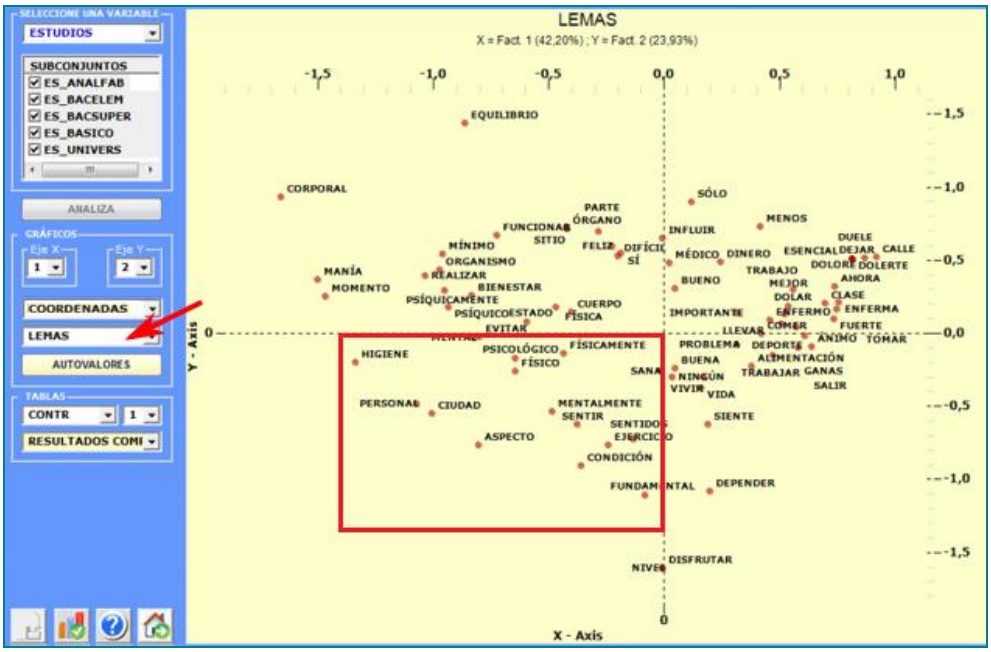
Como resultado del análisis se obtienen tablas, a partir de las cuales se pueden producir los gráficos que - en planos cartesianos - muestran las relaciones entre los subconjuntos del corpus y entre las unidades lexicales (palabras o lemas).
En particular, según los casos, los tipos de gráficos disponibles muestran las relaciones entre **variables activas**, entre **variables ilustrativas**, entre lemas o entre lemas y variables.

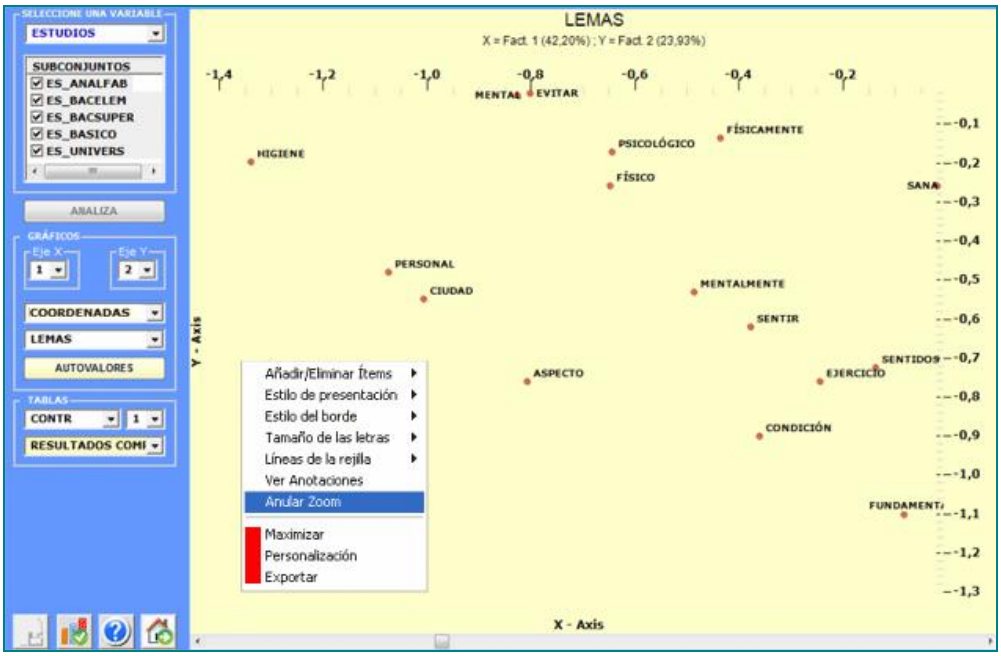
Además, cuando la tabla a analizar es parte de la tipología "documentos x palabras", es posible visualizar los puntos (máximo 3000) correspondientes a cada documento.



Todos los gráficos pueden ser maximizados y personalizados usando la caja de diálogo apropiada (botón derecho del ratón). Por otra parte, cuando las categorías variables son 3 o más, sus relaciones se pueden explorar en **3D** (véase abajo).

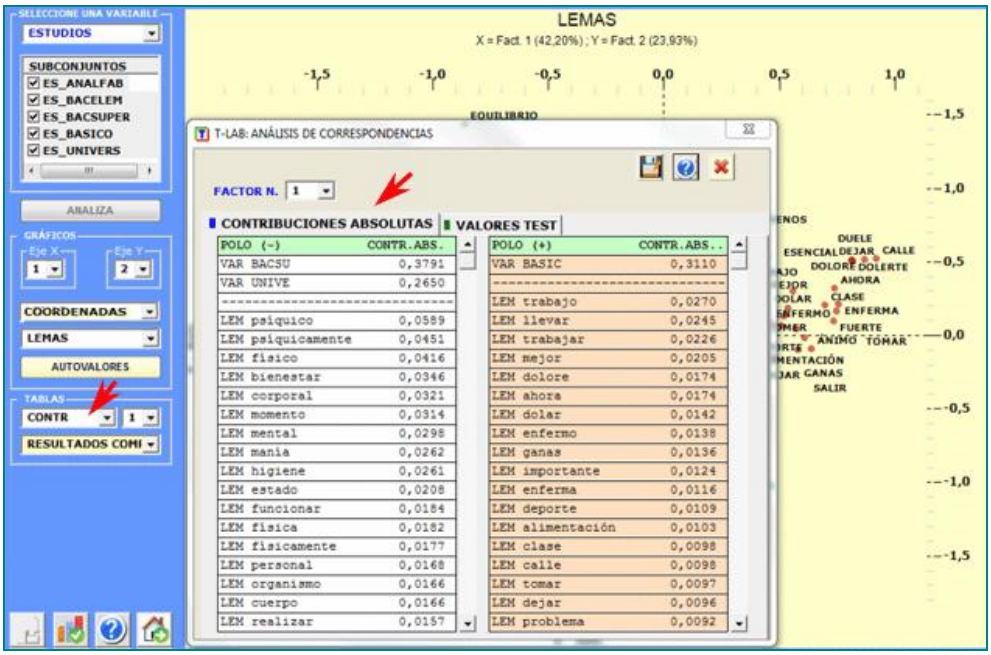




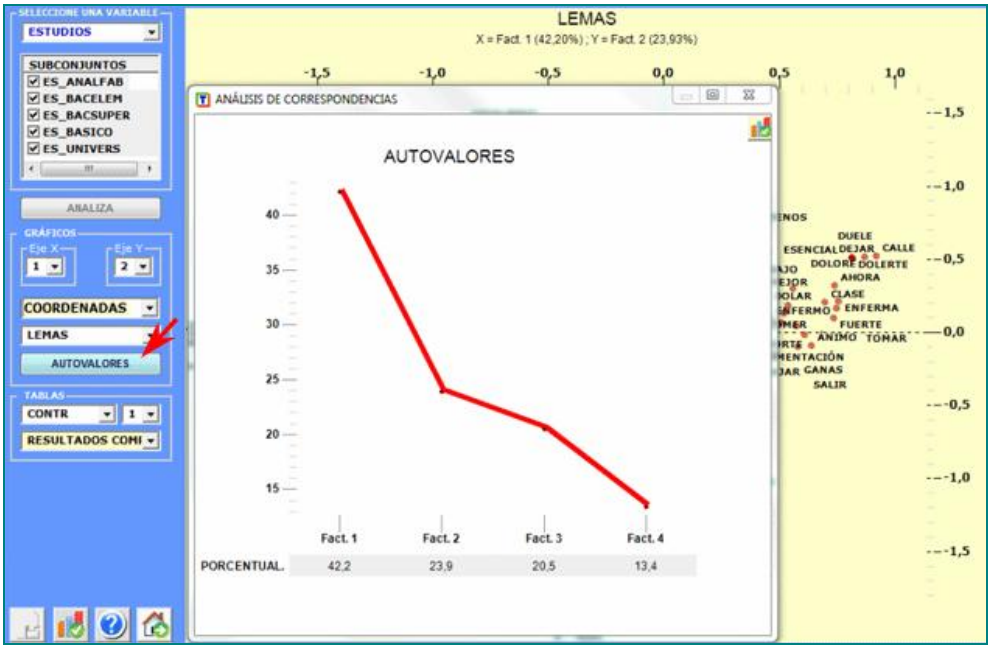


Para explorar las varias combinaciones de los ejes factoriales, es suficiente seleccionarlos en los boxes apropiados ("Eje X", "Eje Y").

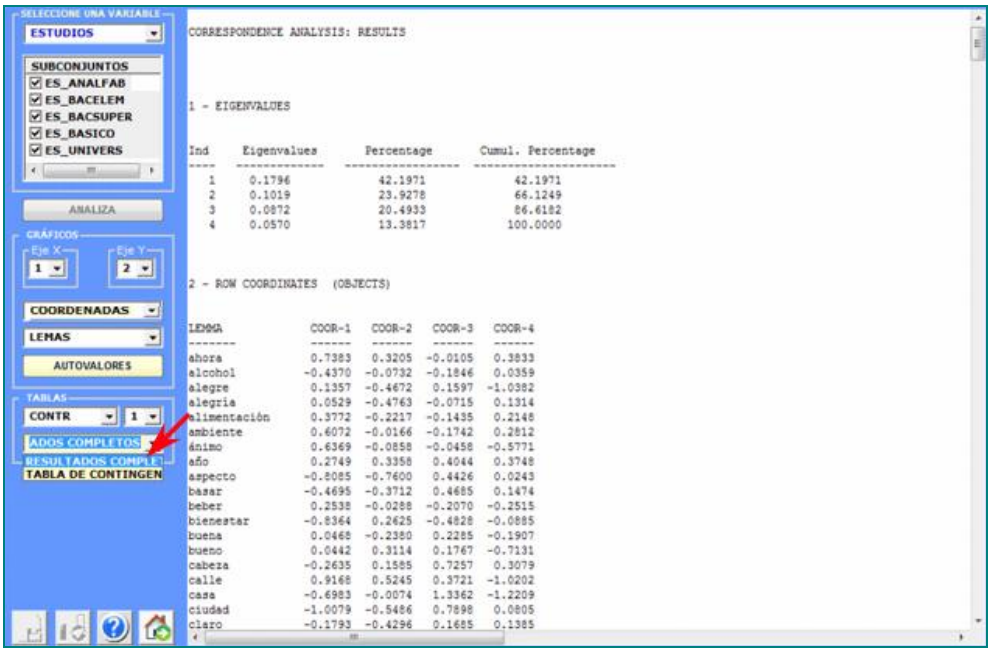
En **T-LAB** las características de cada polo factorial (las oposiciones mostradas en los ejes de los gráficos) se marcan usando dos medidas: las **Contribuciones Absolutas**, cuyo umbral es 1/N (N = filas de la tabla analizada) y los **Valores Test** ("Valeur Test"), cuyo umbral es +/- 1.96.



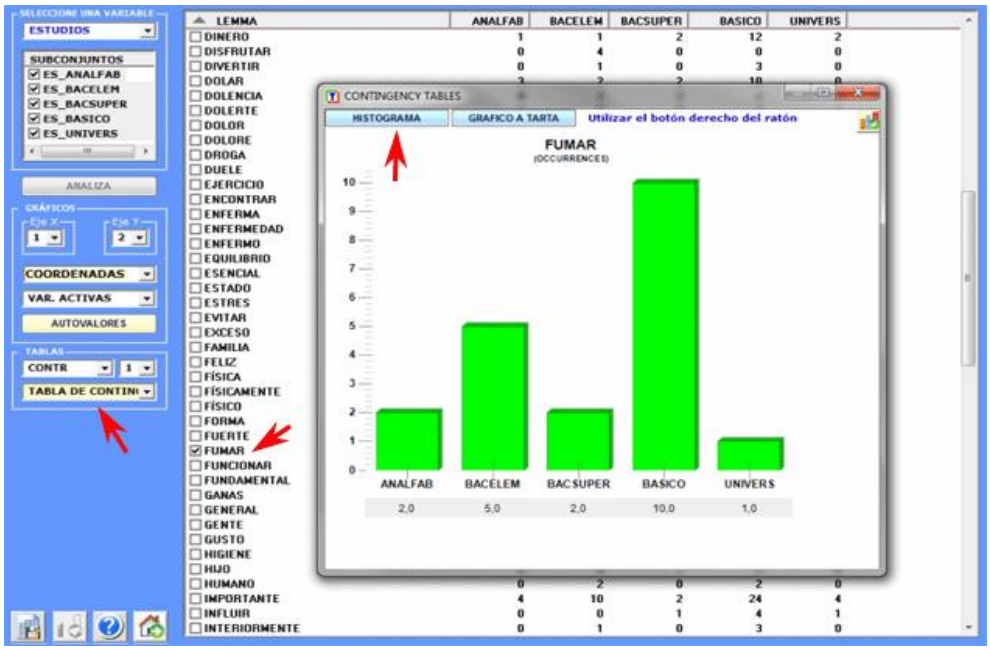
Usando el gráfico **autovalores** es posible apreciar la importancia relativa de cada factor, es decir el porcentaje de variancia que explican.

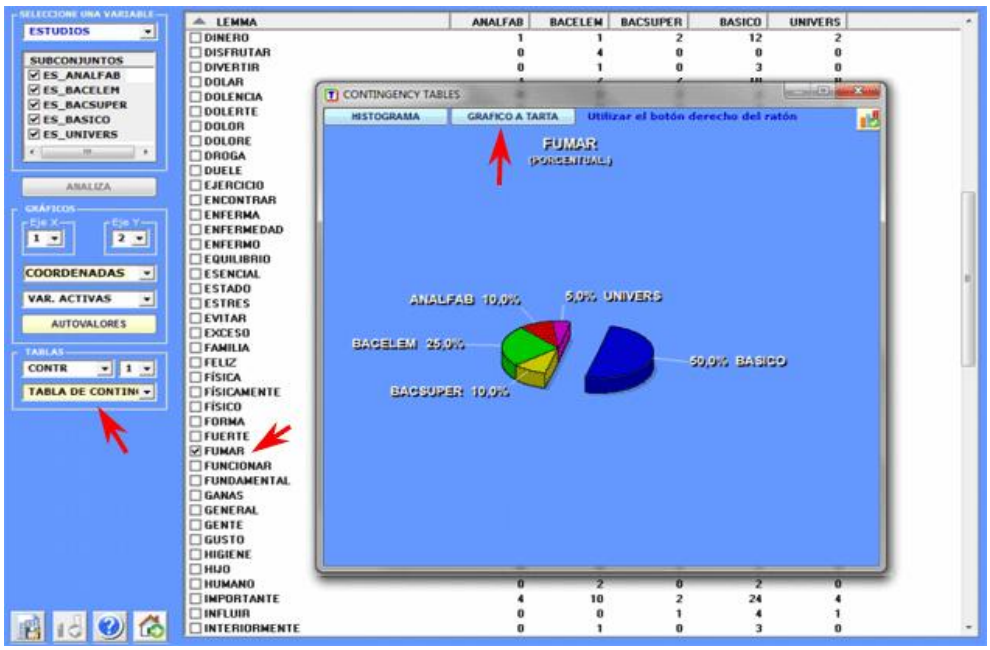


Finalmente, un clic en el botón **Resultados Completos** permite que usted visione y guarde el archivo que contiene todos los resultados del análisis: valores propios, coordenadas, contribuciones absolutas y relativas, valores test.



Todas las tablas de contingencia pueden ser fácilmente exploradas y nos permiten crear varios tipos de gráficos.
Además, haciendo clic en específicas células de la tabla (véase abajo), es posible crear un archivo HTML que incluye todos los contextos elementales en que la palabra en la fila está presente en el subconjunto correspondiente.


CONTINGENCY TABLES
HISTOGRAMA
GRAFICO A TARTA
Utilizar el botón derecho del ratón
FUMAR
(OCCURRENCES)
ANALFAB
BACELEM
BACSUPER
BASICO
UNIVERS
2,0
5,0
2,0
10,0
1,0


CONTINGENCY TABLES
HISTOGRAMA
GRAFICO A TARTA
Utilizar el botón derecho del ratón
FUMAR
(PORCENTUALE)
ANALFAB 10,0%
5,0% UNIVERS
BACELEM 25,0%
50,0% BASICO
BACSUPER 10,0%

Además, sucesivamente es posible efectuar una **Cluster Analysis**.

En los análisis de tablas (**B**) y (**C**), esas están constituidas por tantas líneas como las unidades de contextos (max 10.000) y tantas columnas como palabras clave seleccionadas (max 1.500).

El algoritmo de cálculo y los output son análogos a los del análisis unidades lexicales por variables, sólo que - en este caso - para limitar el tiempo de elaboración, **T-LAB** se limita a extraer los 10 primeros factores: un número más que suficiente para resumir la variabilidad de los datos.

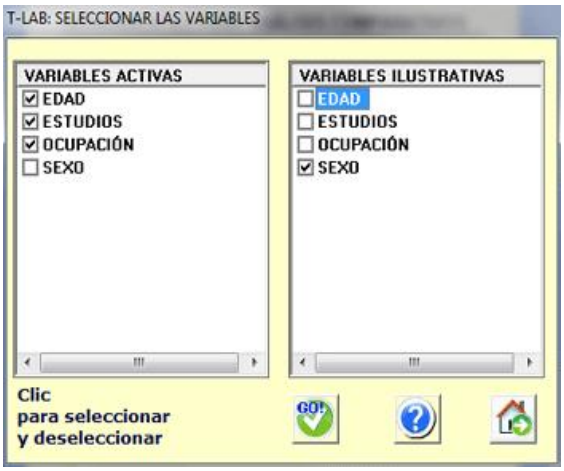
Análisis de Correspondencias Múltiples

NOTA: Las imagenes contenidas en este apartado hacen referencia a una versión anterior de T-LAB, ya que el interfaz de **T-LAB 10** cambia ligeramente. Además, se incluye un botón que permite implementar un **análisis de clúster** que utiliza las coordenadas de los objetos relativas a los primeros ejes factoriales (hasta un máximo de 10).

El Análisis de Correspondencias Múltiples, que se puede considerar una extensión del Análisis de Correspondencia simple (véase arriba), permite analizar las relaciones entre dos o más variables categóricas.

En **T-LAB**, las limitaciones de este tipo de análisis son las siguientes:
- 150.000 contextos elementales como filas;
- 250 categorías de variables como columnas;
- 3,000 palabras clave, como columnas suplementarias (véase Lebart L., Salem A., 1994).

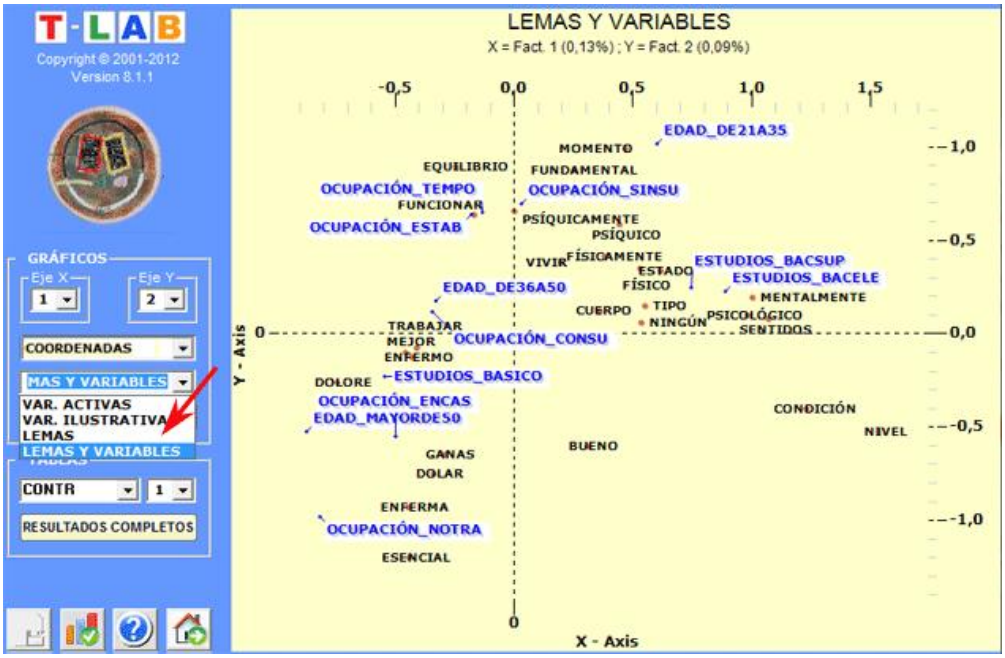
El Análisis de Correspondencias Múltiples, disponible solamente si el corpus incluye por lo menos dos variables, requiere que el usuario seleccione sus opciones en la ventana siguiente:

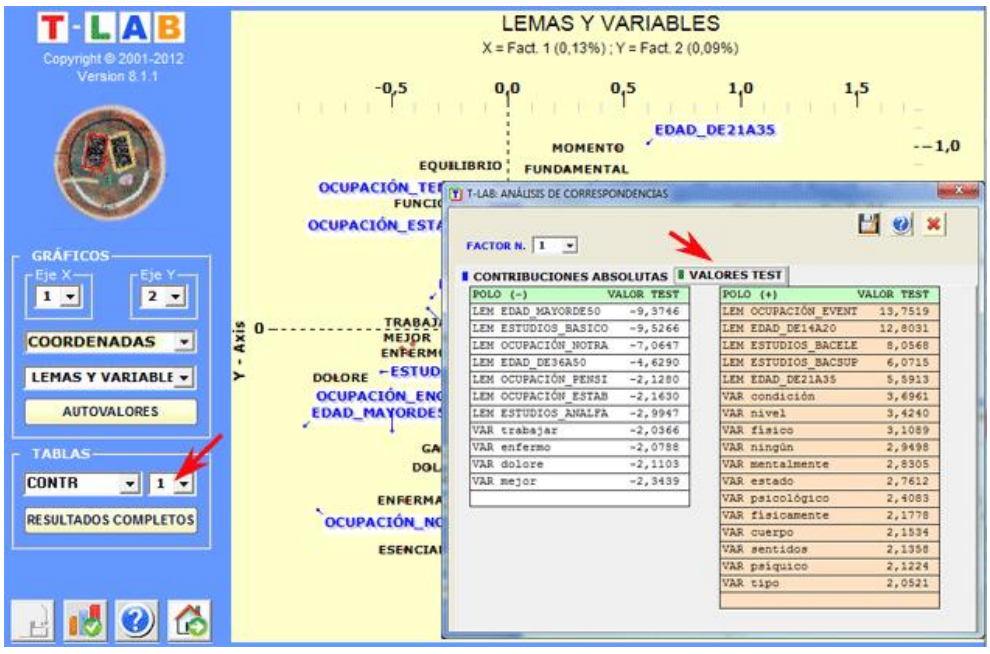


Al final del análisis:

- los outputs de **T-LAB** son iguales a los del Análisis de Correspondencia (véase abajo) con además la tabla de Burt (Burt_Table.xls) que incluye todas las variables cruzadas;

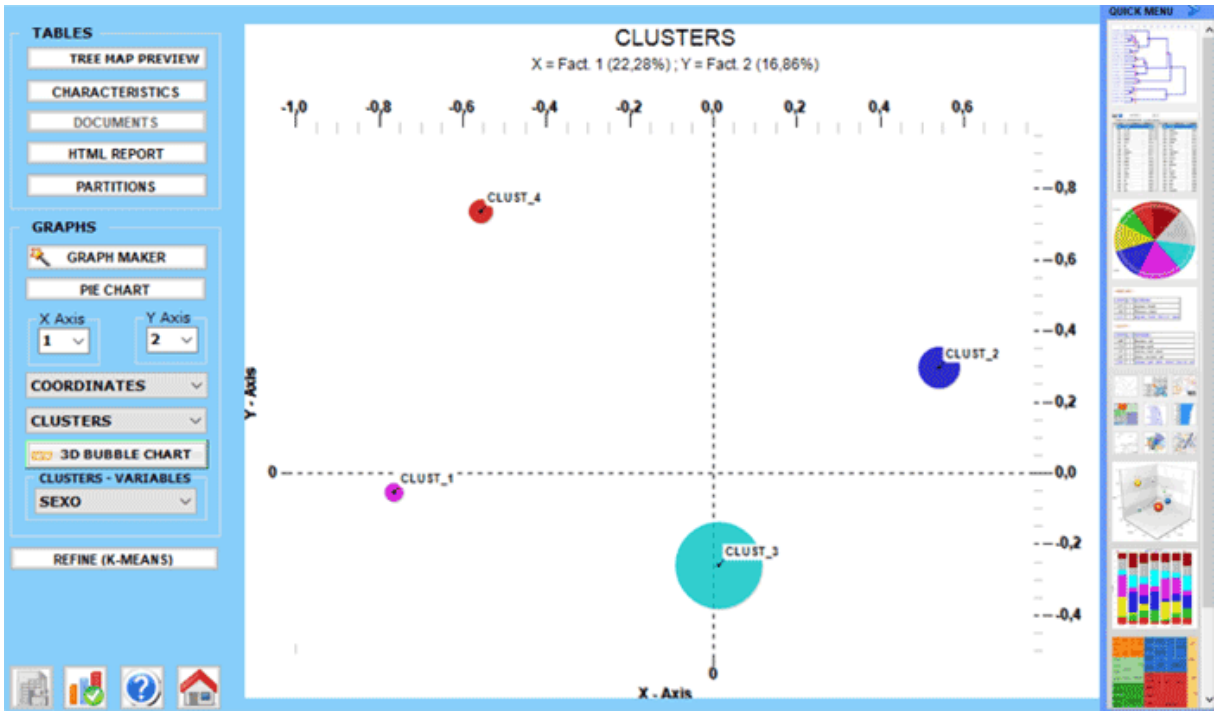
- solamente cuando los contextos elementales corresponden a los documentos primarios (por ej. respuestas a preguntas abiertas) es posible hacer un **cluster analysis**.


T-LAB
Copyright © 2001-2012
Version 8.1.1
GRÁFICOS
Eje X
Eje Y
1
2
COORDENADAS
MAS Y VARIABLES
VAR. ACTIVAS
VAR. ILUSTRATIVA
LEMAS
LEMAS Y VARIABLES
TABLAS
CONTR
1
RESULTADOS COMPLETOS
LEMAS Y VARIABLES
X = Fact. 1 (0,13%) ; Y = Fact. 2 (0,09%)
-0,5
0,0
0,5
1,0
1,5
EDAD_DE21A35
MOMENTO
FUNDAMENTAL
EQUILIBRIO
OCUPACIÓN_TEMPO
FUNCIONAR
OCUPACIÓN_SINSU
OCUPACIÓN_ESTAB
PSÍQUICAMENTE
PSÍQUICO
VIVIR
FÍSICAMENTE
ESTADO
FÍSICO
ESTUDIOS_BACSUP
ESTUDIOS_BACELE
MENTALMENTE
EDAD_DE36A50
CUERPO
TIPO
NINGÚN
PSICOLÓGICO
SENTIDOS
TRABAJAR
MEJOR
OCUPACIÓN_CONSU
ENFERMO
ESTUDIOS_BASICO
DOLORE
OCUPACIÓN_ENCAS
EDAD_MAYORDE50
CONDICIÓN
NIVEL
BUENO
GANAS
DOLAR
ENFERMA
OCUPACIÓN_NOTRA
ESENCIAL
Y - Axis
X - Axis
1,0
0,5
0,0
-0,5
-1,0

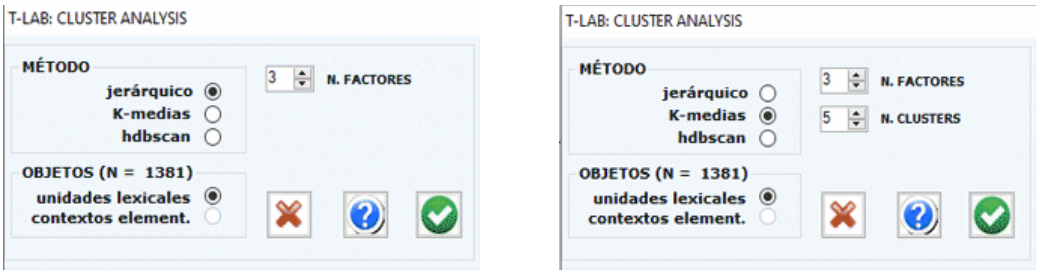

T-LAB
Copyright © 2001-2012
Version 8.1.1
GRÁFICOS
Eje X
Eje Y
1
2
COORDENADAS
LEMAS Y VARIABLE
AUTOVALORES
TABLAS
CONTR
1
RESULTADOS COMPLETOS
LEMAS Y VARIABLES
X = Fact. 1 (0,13%) ; Y = Fact. 2 (0,09%)
EDAD_DE21A35
MOMENTO
FUNDAMENTAL
EQUILIBRIO
T-LAB: ANÁLISIS DE CORRESPONDENCIAS
FACTOR N. 1
CONTRIBUCIONES ABSOLUTAS
VALORES TEST
POLO (-) VALOR TEST
LEM EDAD_MAYORDE50 -9,3746
LEM ESTUDIOS_BASICO -9,5266
LEM OCUPACIÓN_NOTRA -7,0647
LEM EDAD_DE36A50 -4,6290
LEM OCUPACIÓN_PENSI -2,1280
LEM OCUPACIÓN_ESTAB -2,1630
LEM ESTUDIOS_ANALFA -2,9947
VAR trabajar -2,0366
VAR enfermo -2,0788
VAR dolore -2,1103
VAR mejor -2,3439
POLO (+) VALOR TEST
LEM OCUPACIÓN_EVENT 13,7519
LEM EDAD_DE14A20 12,8031
LEM ESTUDIOS_BACELE 8,0568
LEM ESTUDIOS_BACSUP 6,0715
LEM EDAD_DE21A35 5,5913
VAR condición 3,6961
VAR nivel 3,4240
VAR físico 3,1089
VAR ningún 2,9498
VAR mentalmente 2,8305
VAR estado 2,7612
VAR psicológico 2,4083
VAR físicamente 2,1778
VAR cuerpo 2,1534
VAR sentidos 2,1358
VAR psíquico 2,1224
VAR tipo 2,0521
Y - Axis
X - Axis
1,0
0,0
-1,0

Cluster Analysis

NOTA: Las imagenes contenidas en este apartado hacen referencia a una versión anterior de T-LAB, ya que el interfaz de **T-LAB 10** cambia ligeramente. Además: a) una nueva herramienta (**GRAPH MAKER**) permite crear y exportar diferentes tipos de gráficos dinámicos en formato HTML; b) el botón DENDROGRAMA ha sido sustituido por la herramienta GRAPH MAKER; c) una galería de imágenes de acceso rápido que funciona como un menú adicional permite cambiar entre varias salidas con un solo clic.



La opción **Cluster Análisis** pone en marcha un cómputo que utiliza los resultados de un precedente **análisis de correspondencias**; más concretamente, el cómputo utiliza las coordenadas de los objetos (unidades lexicales o unidades de contexto) de los primeros ejes factoriales (hasta un máximo de 7).



Según los casos, el usuario puede elegir entre tres técnicas de cluster analysis:

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

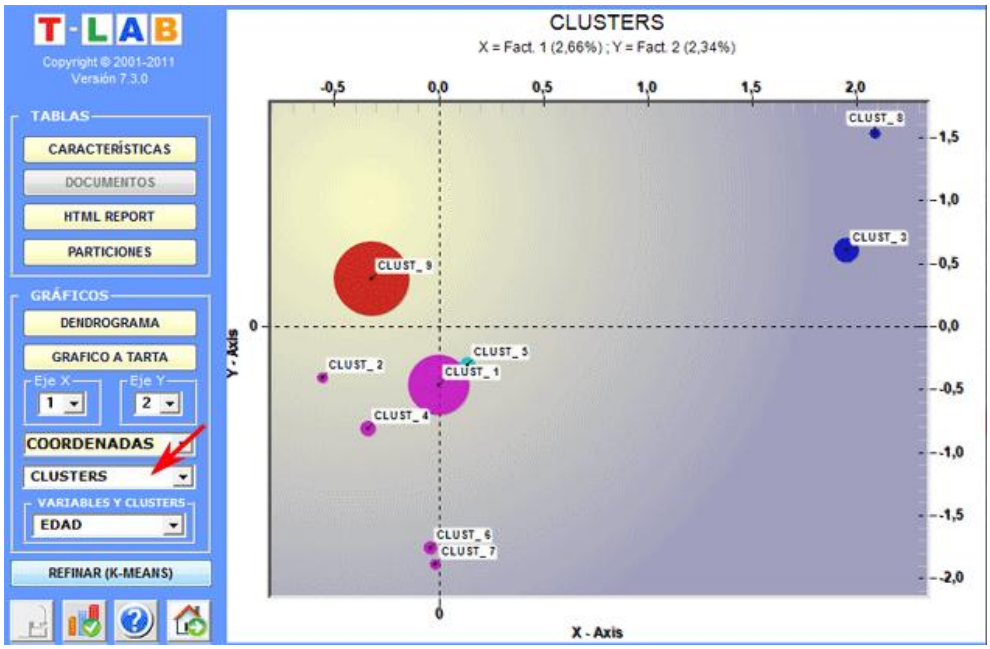
Las dos primeras (a, b) permiten explorar (tablas y gráficos) soluciones de 3 a 20 cluster; mientras que la tercera (c), que requiere un parámetro adicional (es decir, el número mínimo de palabras dentro de un clúster), permite al usuario explorar solo una solución.

NOTA: Cuando se utilisa el método jerárquico **T-LAB** hace disponible una opción (véase el botón 'Refinar') que permite al usuario combinar los métodos de Ward y K-means..

Una breve descripción de las tres técnicas está contenida en el **glosario** de este manual.

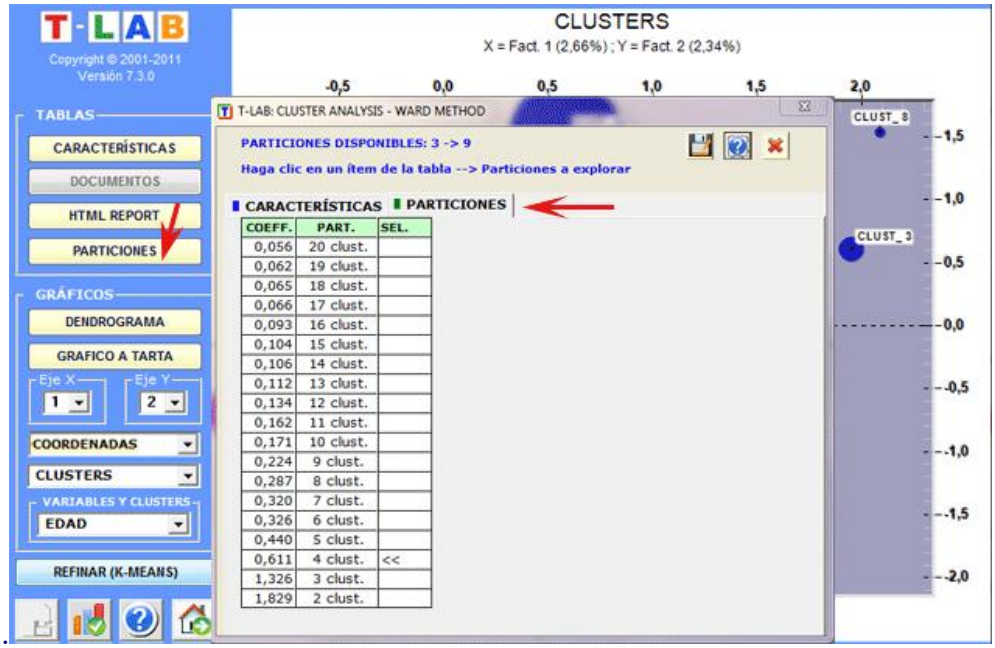
Al término de proceso, **T-LAB** muestra gráficos y tablas.

Los gráficos representan los clusters (racimos) en el mismo espacio detectado por el análisis de correspondencias (véase abajo).

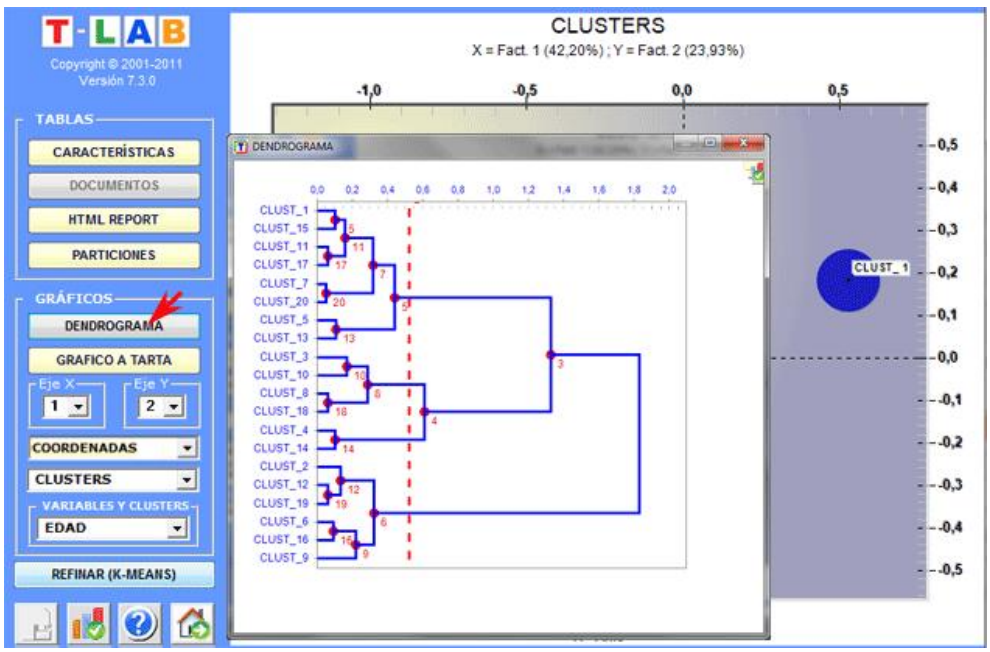


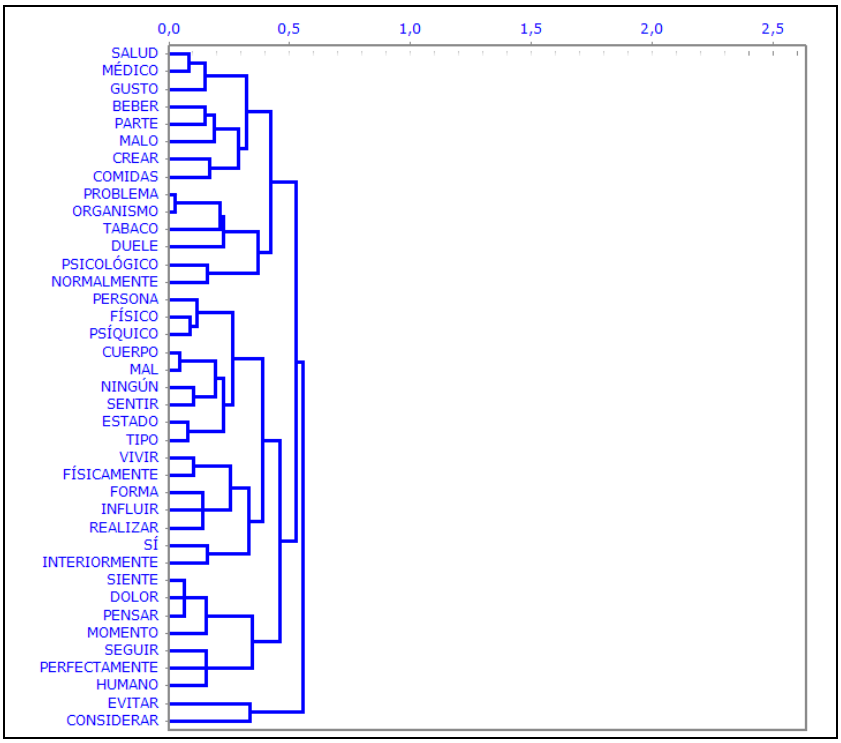
NOTA: Para explorar las varias combinaciones de los ejes factoriales, es suficiente seleccionarlos en los boxes apropiados ("Eje X", "Eje Y").

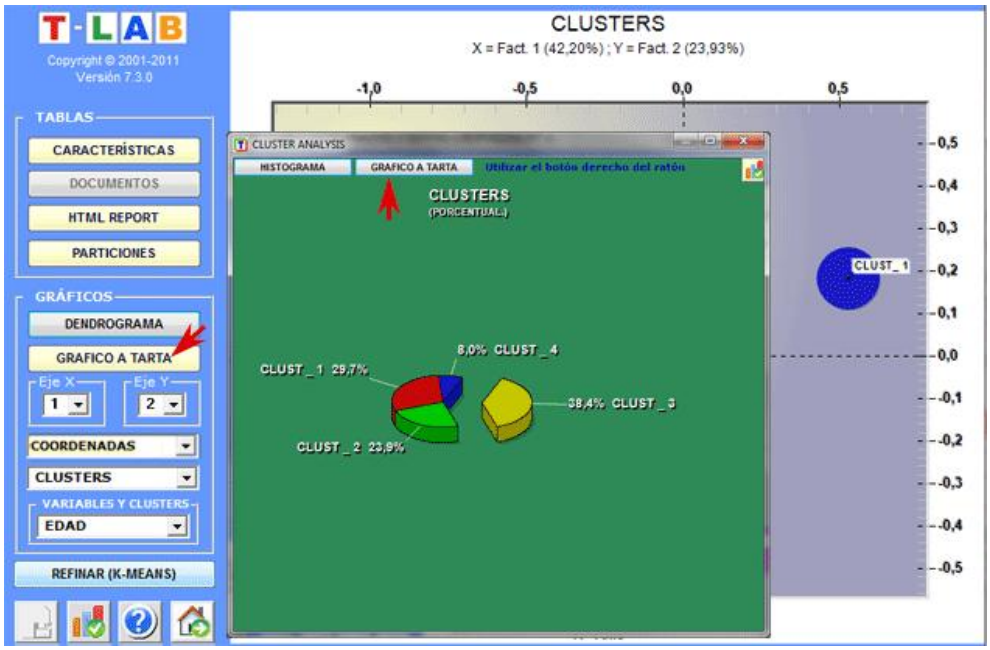
En el caso de clusterización jerárquica el usuario puede explorar fácilmente las particiones diferentes.

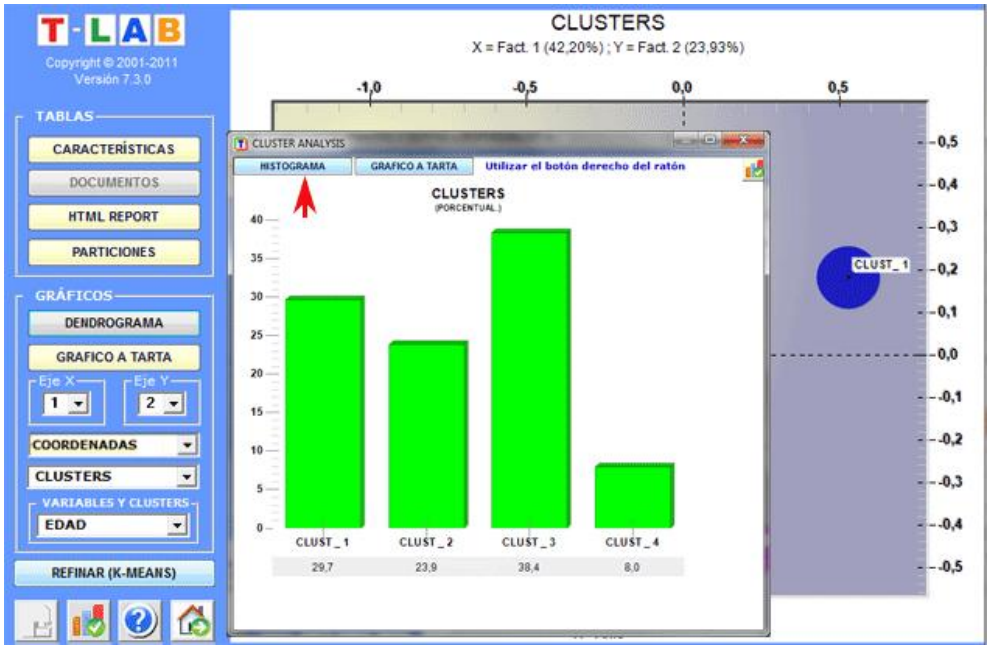


Dendrogramas, gráficos a tarta y histogramas permiten verificar las características de cada partición.

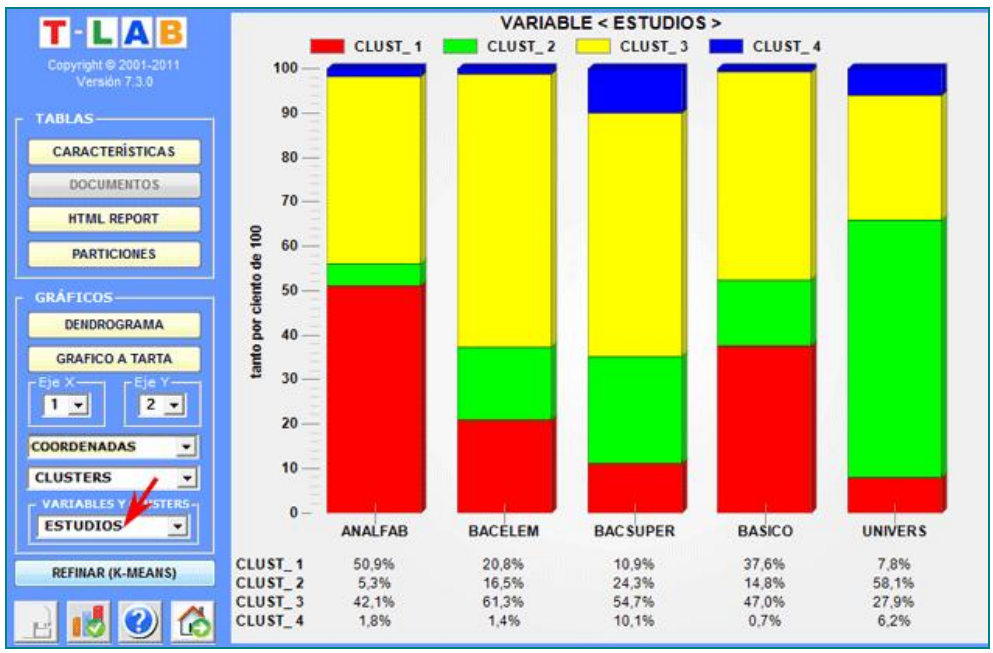



0,0
0,5
1,0
1,5
2,0
2,5
SALUD
MÉDICO
GUSTO
BEBER
PARTE
MALO
CREAR
COMIDAS
PROBLEMA
ORGANISMO
TABACO
DUELE
PSICOLÓGICO
NORMALMENTE
PERSONA
FÍSICO
PSÍQUICO
CUERPO
MAL
NINGÚN
SENTIR
ESTADO
TIPO
VIVIR
FÍSICAMENTE
FORMA
INFLUIR
REALIZAR
SÍ
INTERIORMENTE
SIENTE
DOLOR
PENSAR
MOMENTO
SEGUIR
PERFECTAMENTE
HUMANO
EVITAR
CONSIDERAR


T-LAB
Copyright © 2001-2011
Versión 7.3.0
TABLAS
CARACTERÍSTICAS
DOCUMENTOS
HTML REPORT
PARTICIONES
GRÁFICOS
DENDROGRAMA
GRAFICO A TARTA
Eje X
1
Eje Y
2
COORDENADAS
CLUSTERS
VARIABLES Y CLUSTERS
EDAD
REFINAR (K-MEANS)
CLUSTERS
X = Fact. 1 (42,20%) ; Y = Fact. 2 (23,93%)
-1,0
-0,5
0,0
0,5
CLUSTER ANALYSIS
HISTOGRAMA
GRAFICO A TARTA
Utilizar el botón derecho del ratón
CLUSTERS
(PORCENTUAL)
CLUST_1 29,7%
8,0% CLUST_4
38,4% CLUST_3
CLUST_2 23,9%
CLUST_1
0,5
0,4
0,3
0,2
0,1
0,0
-0,1
-0,2
-0,3
-0,4
-0,5

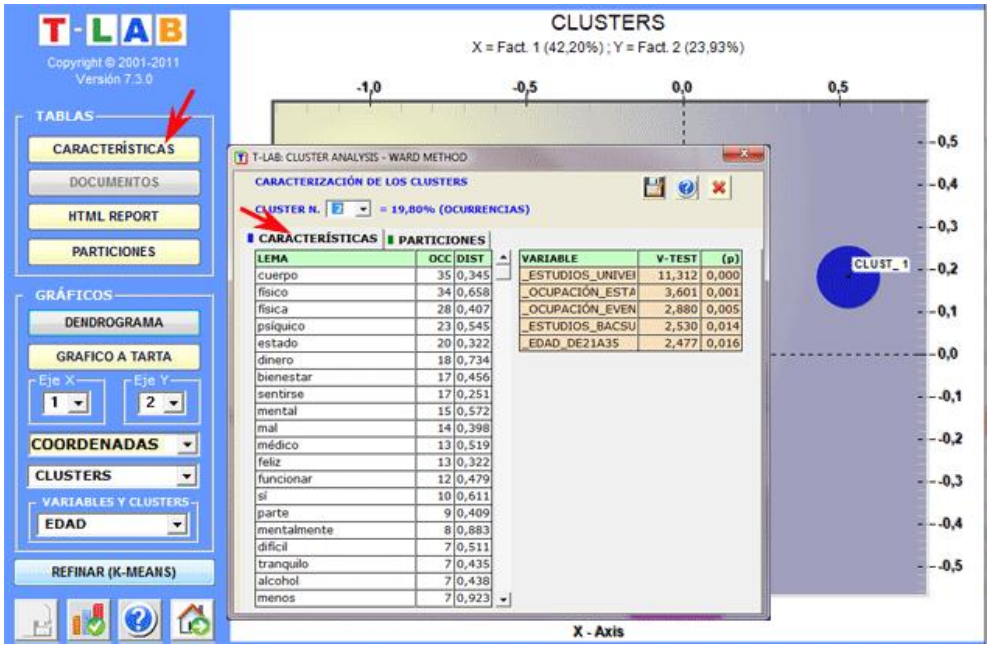


Algunos histogramas nos permiten verificar las relaciones entre los clusters y las variables.

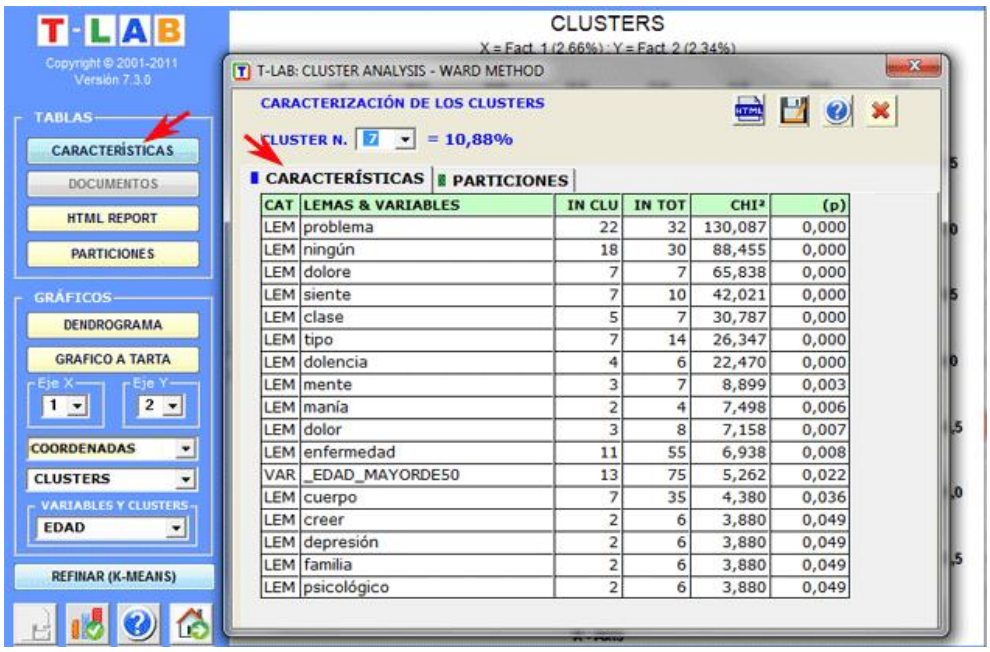


Las tablas son de dos tipos:

(**A**) si los objetos arracimados son unidades lexicales, por cada una de ellas (y por cada cluster) se exhiben las ocurrencias respectivas ('OCC') y las distancias ('DIST') respectivas de los centroides; también, por cada variable que se asocie significativamente al cluster examinado, se exhibe el valor respectivo del **Valor Test**.



(**B**) si los objetos arracimados son contextos elementales, las características de cada cluster (las unidades lexicales y las variables) se describen por medio del mismo método usado en **Análisis Temático de Contextos Elementales** (véase abajo).



En el caso de análisis realizados con métodos jerárquicos o K-means **T-LAB** permite visualizar y exportar un archivo (ver botón "Output HTML") en el que se muestran las características de los cluster y algunas medidas relativas a la calidad de la partición en estudio.

BETWEEN-CLUSTER VARIANCE (S2b) : 1,5247

WITHIN-CLUSTER VARIANCE (S2w) : 0,8747

CLUSTERS 1	0,2606
CLUSTERS 2	0,0726
CLUSTERS 3	0,1701
CLUSTERS 4	0,3713

S2b / (S2b + S2w) : 0,6355

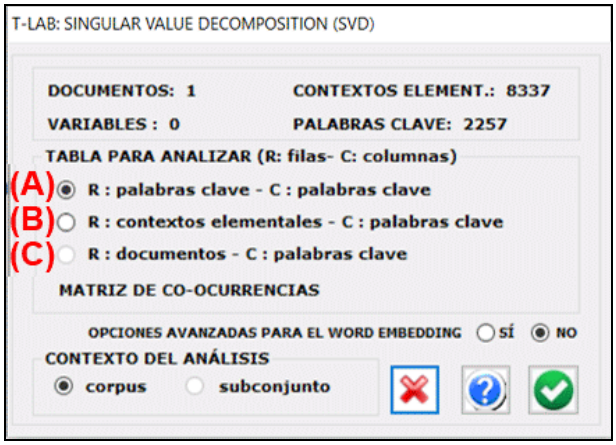
CENTROID COORDINATES

CLUSTERS 1	0,1262	-0,4623	0,3170
CLUSTERS 2	2,4026	1,0441	-0,1707
CLUSTERS 3	-0,8683	1,1000	0,3094
CLUSTERS 4	-0,3475	-0,1648	-0,8953

Descomposición de Valores Singulares (SVD)

La **Descomposición de Valores Singulares** (**SVD** - Singular Value Decomposition) es una técnica de reducción de dimensiones que, en minería de textos, puede utilizarse para descubrir las **dimensiones latentes** (o componentes) que determinan **similitudes semánticas** entre las palabras (es decir, unidades léxicas) o entre los documentos (es decir, unidades de contexto).

T-LAB nos permite realizar un SVD de **tres tipos de tablas de datos**. En el primer caso (ver 'A' a continuación), la tabla de datos es una matriz de co-ocurrencias con - en filas y en columnas - las palabras clave seleccionadas. En el segundo caso (ver 'B' a continuación), la tabla de datos contextos elementales X palabras clave contendrá valores de presencia / ausencia (es decir, '1' y '0'). En el tercer caso (ver 'C' a continuación) la tabla de datos documentos X palabras clave contendrá valores de ocurrencia.
NOTA: Tenga en cuenta que, al analizar la matriz de co-ocurrencias cuyas filas y columnas son términos clave (ver 'A' a continuación), **T-LAB** proporciona vectores densos de alta calidad (es decir, word embeddings).



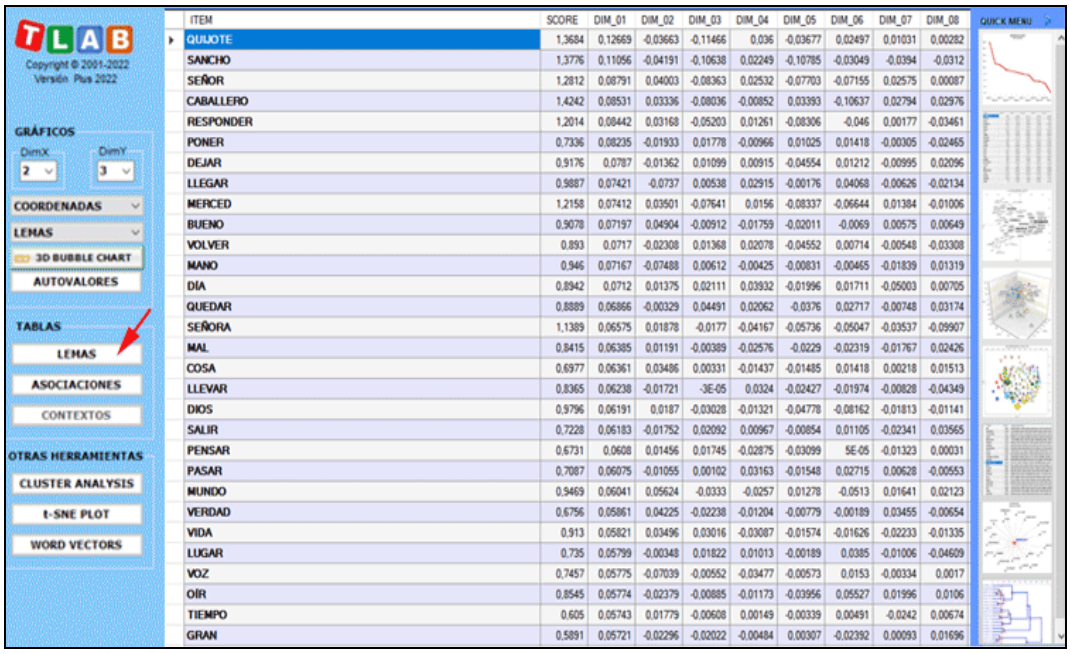
El procedimiento de análisis consta de los siguientes pasos:
1 - construcción de la tabla de datos a analizar (hasta 300,000 filas x 5,000 columnas);
2 - normalización TF-IDF y escalado de vectores de fila a longitud de unidad (norma euclidiana);
3 - extracción de las primeras 20 'dimensiones latentes' a través del algoritmo Lanczos.
NOTA:
- En el caso de las matrices de co-ocurrencia (ver "A" arriba), la normalización de los datos se obtiene usando la medida del coseno;
- Cuando se seleccionan las opciones avanzadas para el word embedding, T-LAB calcula los valores de PPMI (Positive Pointwise Mutual Information) y hace posible utilizar las primeras 50 dimensiones de la SVD.

Los resultados del análisis se muestran en **tablas** y **gráficos**.

En detalle:

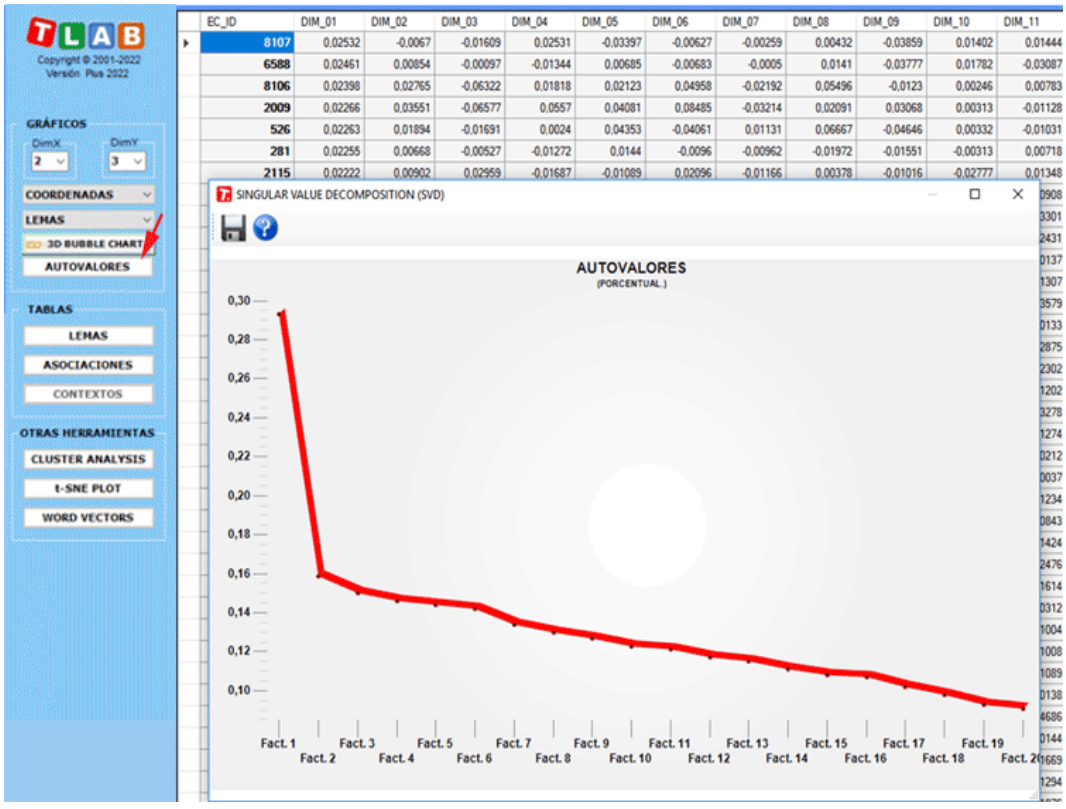
Dos tablas, cuyas filas pueden ser unidades léxicas o unidades de contexto, tienen tantas columnas como las dimensiones extraídas.

En el caso de la tabla LEMAS (es decir, unidades léxicales), se muestra una columna más, en la que se informan los puntajes de importancia (ver a continuación).



ITEM	SCORE	DIM_01	DIM_02	DIM_03	DIM_04	DIM_05	DIM_06	DIM_07	DIM_08
QUIJOTE	1,3684	0,12669	-0,03663	-0,11466	0,036	-0,03677	0,02497	0,01031	0,00282
SANCHO	1,3776	0,11056	-0,04191	-0,10638	0,02249	-0,10785	-0,03049	-0,0394	-0,0312
SEÑOR	1,2812	0,08791	0,04003	-0,08363	0,02532	-0,07703	-0,07155	0,02575	0,00087
CABALLERO	1,4242	0,08531	0,03336	-0,08036	-0,00852	0,03393	-0,10637	0,02794	0,02976
RESPONDER	1,2014	0,08442	0,03168	-0,05203	0,01261	-0,08306	-0,046	0,00177	-0,03461
PONER	0,7336	0,08235	-0,01933	0,01778	-0,00966	0,01025	0,01418	-0,00305	-0,02465
DEJAR	0,9176	0,0787	-0,01362	0,01099	0,00915	-0,04554	0,01212	-0,00995	0,02096
LLEGAR	0,9887	0,07421	-0,0737	0,00538	0,02915	-0,00176	0,04068	-0,00626	-0,02134
MERCED	1,2158	0,07412	0,03501	-0,07641	0,0156	-0,08337	-0,06644	0,01384	-0,01006
BUENO	0,9078	0,07197	0,04904	-0,00912	-0,01759	-0,02011	-0,0069	0,00575	0,00649
VOLVER	0,893	0,0717	-0,02308	0,01368	0,02078	-0,04552	0,00714	-0,00548	-0,03308
MANO	0,946	0,07167	-0,07488	0,00612	-0,00425	-0,00831	-0,00465	-0,01839	0,01319
DÍA	0,8942	0,0712	0,01375	0,02111	0,03932	-0,01996	0,01711	-0,05003	0,00705
QUEDAR	0,8889	0,06866	-0,00329	0,04491	0,02062	-0,0376	0,02717	-0,00748	0,03174
SEÑORA	1,1389	0,06575	0,01878	-0,0177	-0,04167	-0,05736	-0,05047	-0,03537	-0,09907
MAL	0,8415	0,06385	0,01191	-0,00389	-0,02576	-0,0229	-0,02319	-0,01767	0,02426
COSA	0,6977	0,06361	0,03486	0,00331	-0,01437	-0,01485	0,01418	0,00218	0,01513
LLEVAR	0,8365	0,06238	-0,01721	-3E-05	0,0324	-0,02427	-0,01974	-0,00828	-0,04349
DIOS	0,9796	0,06191	0,0187	-0,03028	-0,01321	-0,04778	-0,08162	-0,01813	-0,01141
SALIR	0,7228	0,06183	-0,01752	0,02092	0,00967	-0,00854	0,01105	-0,02341	0,03565
PENSAR	0,6731	0,0608	0,01456	0,01745	-0,02875	-0,03099	5E-05	-0,01323	0,00031
PASAR	0,7087	0,06075	-0,01055	0,00102	0,03163	-0,01548	0,02715	0,00628	-0,00553
MUNDO	0,9469	0,06041	0,05624	-0,0333	-0,0257	0,01278	-0,0513	0,01641	0,02123
VERDAD	0,6756	0,05861	0,04225	-0,02238	-0,01204	-0,00779	-0,00189	0,03455	-0,00654
VIDA	0,913	0,05821	0,03496	0,03016	-0,03087	-0,01574	-0,01626	-0,02233	-0,01335
LUGAR	0,735	0,05799	-0,00348	0,01822	0,01013	-0,00189	0,0385	-0,01006	-0,04609
VOZ	0,7457	0,05775	-0,07039	-0,00552	-0,03477	-0,00573	0,0153	-0,00334	0,0017
OÍR	0,8545	0,05774	-0,02379	-0,00885	-0,01173	-0,03956	0,05527	0,01996	0,0106
TIEMPO	0,605	0,05743	0,01779	-0,00608	0,00149	-0,00339	0,00491	-0,0242	0,00674
GRAN	0,5891	0,05721	-0,02296	-0,02022	-0,00484	0,00307	-0,02392	0,00093	0,01696

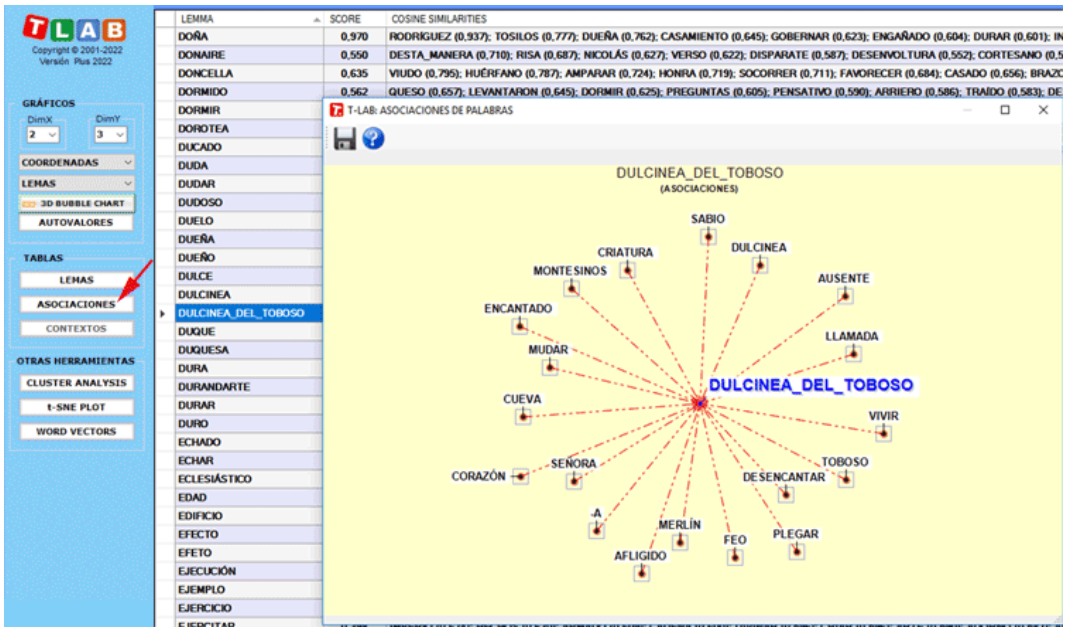
NOTA: La puntuación (es decir el 'score') de **importancia** de cada lema se calcula sumando los valores absolutos de sus primeras 20 coordenadas (es decir, los vectores propios), cada uno multiplicado por lo valor propio correspondiente.



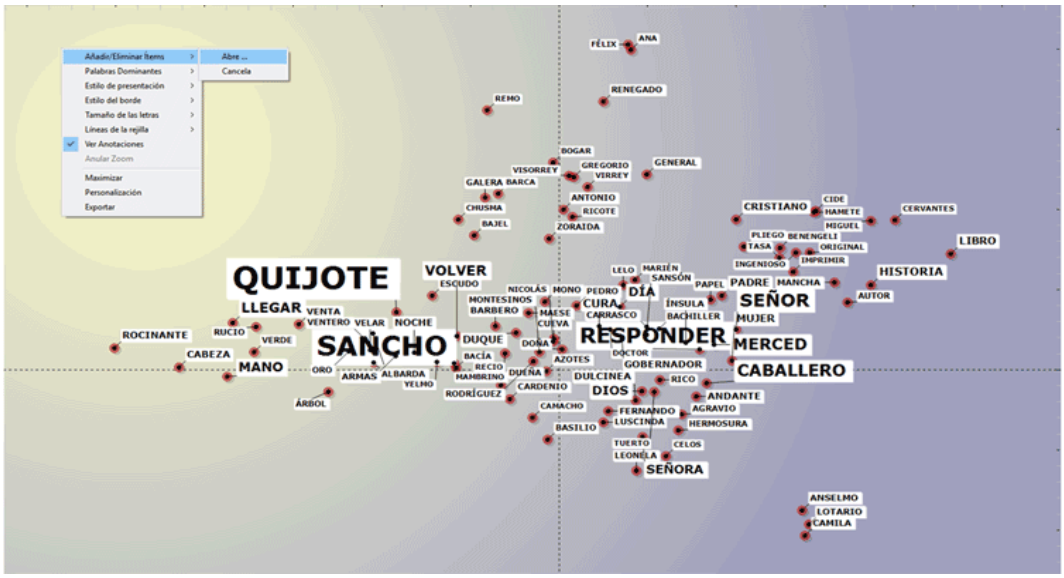
EC_ID	DIM_01	DIM_02	DIM_03	DIM_04	DIM_05	DIM_06	DIM_07	DIM_08	DIM_09	DIM_10	DIM_11
8107	0,02532	-0,0067	-0,01609	0,02531	-0,03397	-0,00627	-0,00259	0,00432	-0,03859	0,01402	0,01444
6588	0,02461	0,00854	-0,00097	-0,01344	0,00685	-0,00683	-0,0005	0,0141	-0,03777	0,01782	-0,03087
8106	0,02398	0,02765	-0,06322	0,01818	0,02123	0,04958	-0,02192	0,05496	-0,0123	0,00246	0,00783
2009	0,02266	0,03551	-0,06577	0,0557	0,04081	0,08485	-0,03214	0,02091	0,03068	0,00313	-0,01128
526	0,02263	0,01894	-0,01691	0,0024	0,04353	-0,04061	0,01131	0,06667	-0,04646	0,00332	-0,01031
281	0,02255	0,00668	-0,00527	-0,01272	0,0144	-0,0096	-0,00962	-0,01972	-0,01551	-0,00313	0,00718
2115	0,02222	0,00902	0,02959	-0,01687	-0,01089	0,02096	-0,01166	0,00378	-0,01016	-0,02777	0,01348

Cualquier tabla se puede **ordenar** en orden ascendente o descendente haciendo clic en cualquier encabezado de columna.
Para **exportar** cualquier tabla, solo use el botón derecho del mouse cuando se muestren los datos. Tenga en cuenta que, la primera vez que se exporta una tabla de este tipo, los valores propios también se exportan. De esta forma, el usuario puede evaluar el peso relativo de cada dimensión, es decir el porcentaje de varianza explicado por cada una de ellas.

Al hacer clic en el botón **Asociaciones** (véase a continuación), se muestra una tabla adicional con las medidas de similitud (es decir, el coseno) de cada palabra clave. Además, cuando se hace clic en cualquier fila de dicha tabla, se muestra un gráfico con los datos correspondientes.



Los **gráficos** principales muestran las relaciones entre las palabras clave (es decir, lemmas) en las dimensiones seleccionadas (véase a continuación).



Por defecto, el cuadro anterior incluye los 100 lemas más importantes. Sin embargo, el usuario puede personalizar tanto el número de lemas como las características del gráfico.

PREPARACION DEL CORPUS

Preparación del Corpus

En el caso de un **único texto** (o corpus considerado como único texto), **T-LAB** no necesita nada más: es suficiente seleccionar la opción 'Importar un único archivo ..' (vease la correspondiente voz del manual).

Cuando, en cambio, el corpus está compuesto por **más textos** y/o cuando se utilizan codificaciones que remiten al uso de alguna **variable**, en la fase de preparación se hace necesario utilizar el modulo **Corpus Builder** que permite transformar los textos a analizar en un corpus codificado y listo para la importación. Todo ello, de forma rápida y automática.

NOTA:

- en todo caso, aconsejamos una revisión ortográfica del material a analizar. Por otra parte, si algunas siglas importantes se espacian por medio de la puntuación (por ejemplo "N.U.") se recomienda su transformación en una sola secuencia (por ejemplo "NU" o " N_U"), porque, en la fase de la **normalización**, **T-LAB** interpreta los signos de puntuación como separadores;
- al término de la fase de preparación se recomienda crear una nueva carpeta de trabajo en cuyo interior sólo se encuentre el archivo corpus a importar.

Criterios Estructurales

Hay dos **criterios estructurales** que se tienen que respetar: el **tamaño** del **corpus** y su subdivisión en **partes**.

En cuanto al tamaño, todos los instrumentos **T-LAB** han sido probados con un corpus de 90Mb, equivalente a 55.000 páginas en formato texto.

Los límites para el **tamaño mínimo** requieren diversos criterios de la evaluación; esto es así porque, bajo un cierto umbral, el tamaño del corpus puede perjudicar la fiabilidad de muchos análisis estadísticos. Basta seguir estas simples instrucciones: utilice corpus con al menos 5.000 ocurrencias (aproximadamente 30 KB); si no, en el caso de preguntas abiertas, un mínimo de 50 respuestas. De hecho, en este último caso, cada respuesta constituye una unidad de contexto diferente.

Para ser procesado, una corpus se puede componer de: un único texto sin otras particiones, un único texto subdividido según los criterios establecidos por el usuario (por ejemplo, un libro dividido en capítulos), varios textos (por ejemplo, varias entrevistas o respuestas a preguntas abiertas) clasificados mediante el uso de etiquetas, que remiten a otras tantas **variables** o **IDnumber**.

En todos estos casos, el corpus se subdivide en partes que se deben definir con los **criterios formales** exactos.

Criterios Formales

En el caso de un **corpus** constituido por un solo texto, y cuando el usuario no recurra a las **variables**, **no se requiere ninguna otra operación**: se puede pasar directamente a la fase de **importación**.

Cuando, en cambio, el corpus está compuesto por **más de un texto** y/o cuando se utilizan **variables**, la preparación del corpus se debe realizar a través del modulo **Corpus Builder** que, de forma automática, respeta los siguientes criterios:

Cada texto o subconjunto del mismo (las "partes" individuadas por las variables) tienen que ir precedidas por **una línea de codificación**.

Cada línea de codificación tiene este formato:

- **comienza** con una cadena de **cuatro asteríscos** (****) seguida por un espacio en blanco. **T-LAB** lee esta cadena como: "aquí comienza un texto o una unidad de contexto definida por el usuario".
- **continua** con la adición de cadenas compuestas por **asteriscos aislados** y de etiquetas que definen **casos** (**IDnumber**), **variables** y las respectivas **modalidades**.
- **termina** con "vuelta a empezar".

Aquí hay algunos ejemplos.

La línea siguiente introduce un texto (o un subconjunto del corpus) codificado con tres variables - EDAD, SEXO y OCU (ocupación) - y sus modalidades (ADUL, FEM, PROF).

**** *EDAD_ADUL *SEXO_FEM *OCU_PROF

La línea siguiente introduce un texto (o un subconjunto del corpus) codificado con las mismas variables y la etiqueta **IDnumber**.

**** *IDnumber_0001 * EDAD _ADUL * SEXO _FEM * OCU_PROF

La línea siguiente introduce un texto (o un subconjunto del corpus) codificado con dos variables: AÑO, PERI (periódicos):

**** * AÑO _98 * PERI_PAÍS

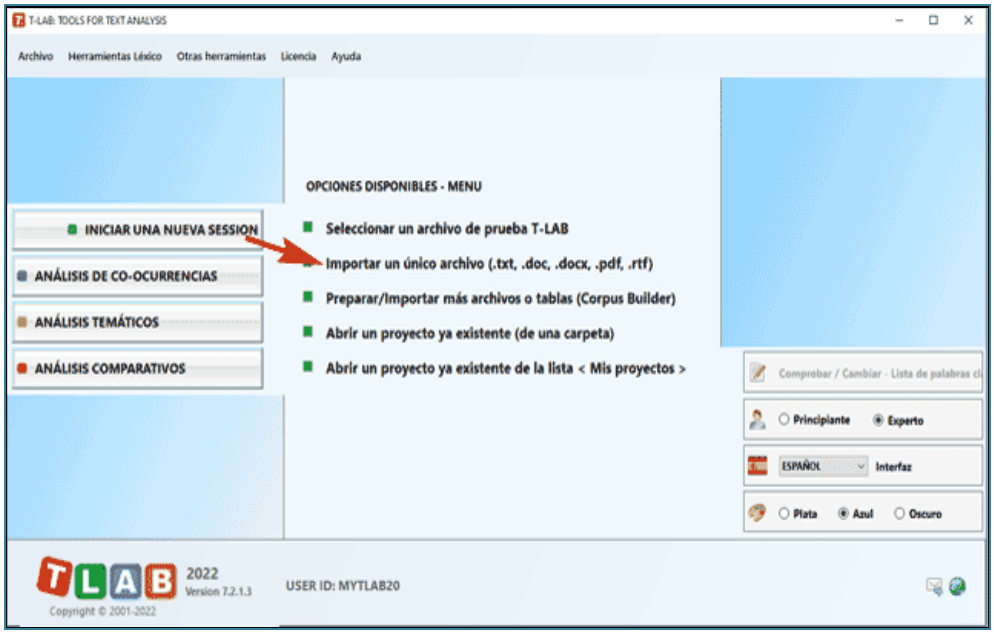
En cada línea de codificación, las reglas de T-LAB que se deben respetar son las siguientes:

- cada etiqueta (IDnumber, variables y modalidades) no puede ser distanciada por los espacios en blanco.
- cada etiqueta – tanto en el caso de las variables como en el de las modalidades - no puede superar 25 caracteres (min. 2).
- cada etiqueta de variables se debe ligar a la modalidad respectiva con un guión bajo ("_").
- entre dos variables, es decir antes del asterisco siguiente, se debe inserir un espacio en blanco.
- cada variable y respectivas modalidades se debe asignar para cada subconjunto del corpus.
- las variables utilizables son máximo 50, cada una con un máximo de 150 modalidades;
- el número máximo de IDnumber está fijado en 99.999 para textos cortos (Max. 2.000 caracteres cada uno. Eje. respuestas a preguntas abiertas), y en 30.000 para los demás casos.

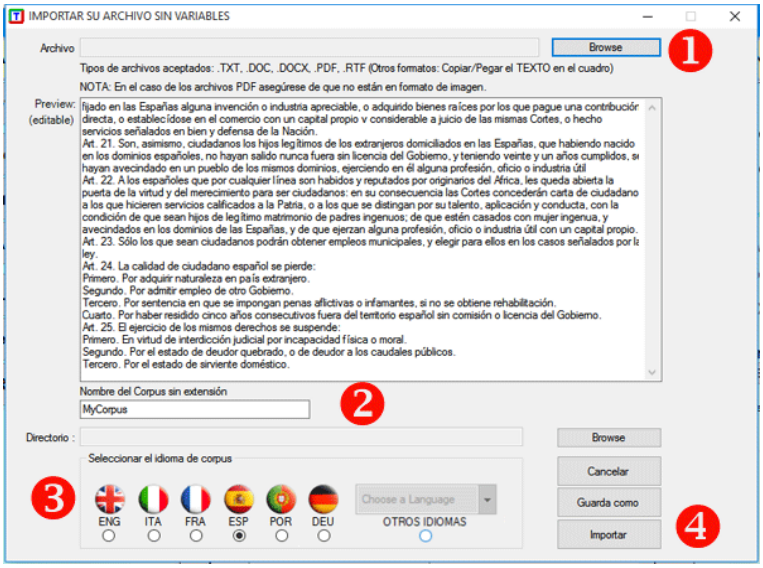
ARCHIVO

Importar un único archivo ...

En el caso de un único texto (o corpus considerado como único texto), **T-LAB** no necesita nada más: es suficiente seleccionar la opción 'Importar un único archivo (vease abajo)

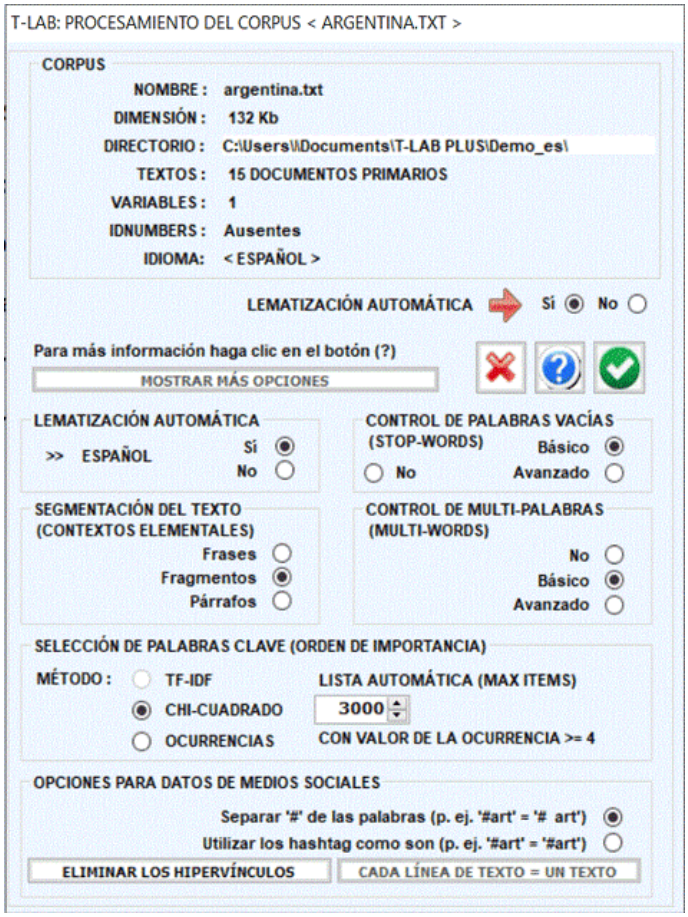


Entonces se requieren cuatro pasos (ver la imagen siguiente) : (1) seleccionar cualquier archivo; (2) elegir el nombre del proyecto; (3) seleccionar el idioma de su texto; (4) hacer clic en "Importar" .



Sucesivamente aparece una ventana (véase abajo) en la cual el usuario puede elegir algunos tratamientos.

NOTA:
- Porque las diferentes opciones determinan el tipo y la cantidad de unidades de análisis (es decir las unidades de contexto y las unidades lexicales), diversas opciones determinan diversos resultados del análisis (véase abajo las opciones avanzadas). Por esta razón, todos los outputs de **T-LAB** (es decir gráficos y tablas) utilizados en el manual del usuario y en la ayuda en red son solo indicativos;
- Todas las etapas de pre-procesamiento se realizan al importar cualquier tipo de corpus.



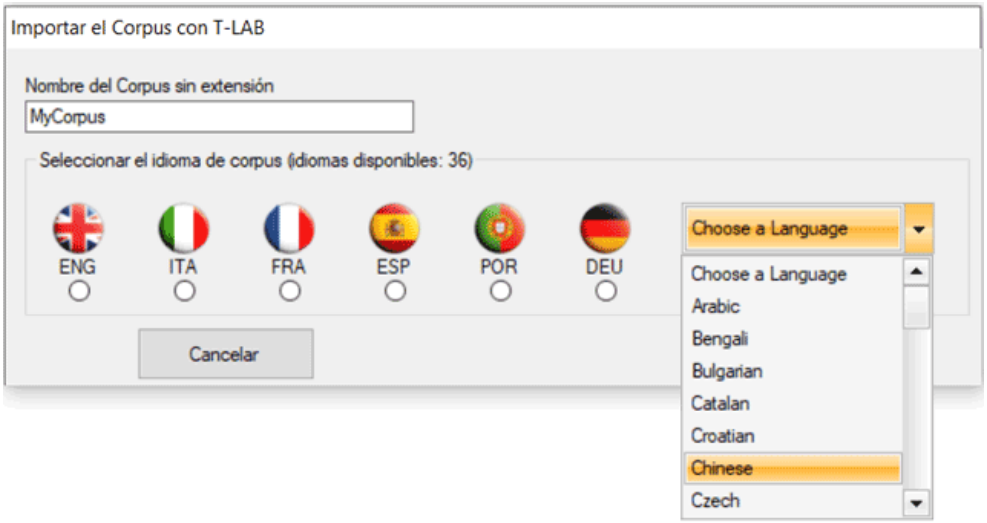
1 - LEMATIZACIÓN AUTOMÁTICA O STEMMING

A continuación, se presenta el listado completo de los 30 idiomas para los cuales **T-LAB** prevé la posibilidad de implementar procesos de lematización automática y de stemming.

LEMATIZACIÓN: alemán, catalán, croata, eslovaco, español, francés, inglés, italiano, latín, polaco, portugués, rumano, ruso, serbo, sueco y ucraniano.

STEMMING: árabe, bengalí, búlgaro, checo, danés, finlandés, griego, hindi, húngaro, indonesio, marathi, noruego, persa y turco.

En cualquier caso, sin lematización automática y / o mediante diccionarios personalizados, el usuario puede analizar textos **en todos los idiomas**. Lo importante es que las palabras estén separadas por espacios y/o signos de puntuación.



El resultado del proceso del lematización se puede verificar por medio de la función **Vocabulario** y se puede modificar por medio de la función **Personalización del Diccionario**.

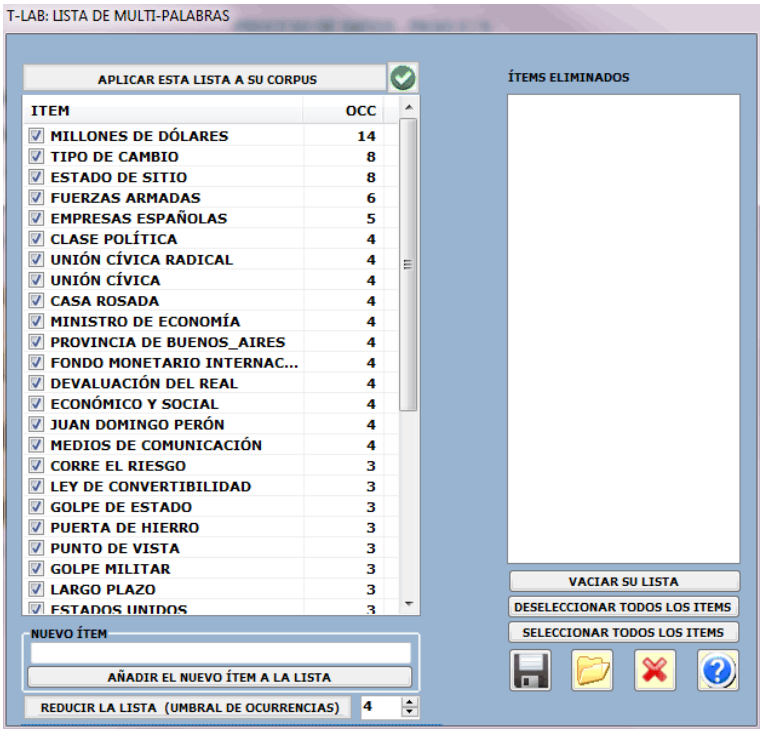
2 - SEGMENTACIÓN DE TEXTOS (CONTEXTOS ELEMENTALES)

Según la elección del usuario, los **contextos elementales** para el cómputo de **co-ocurrencias** pueden ser: frases, fragmentos de longitud comparable, párrafos o textos breves (por ejemplo, respuestas a las preguntas abiertas).

El fichero corpus_segments.dat contiene el resultado de la segmentación del corpus.

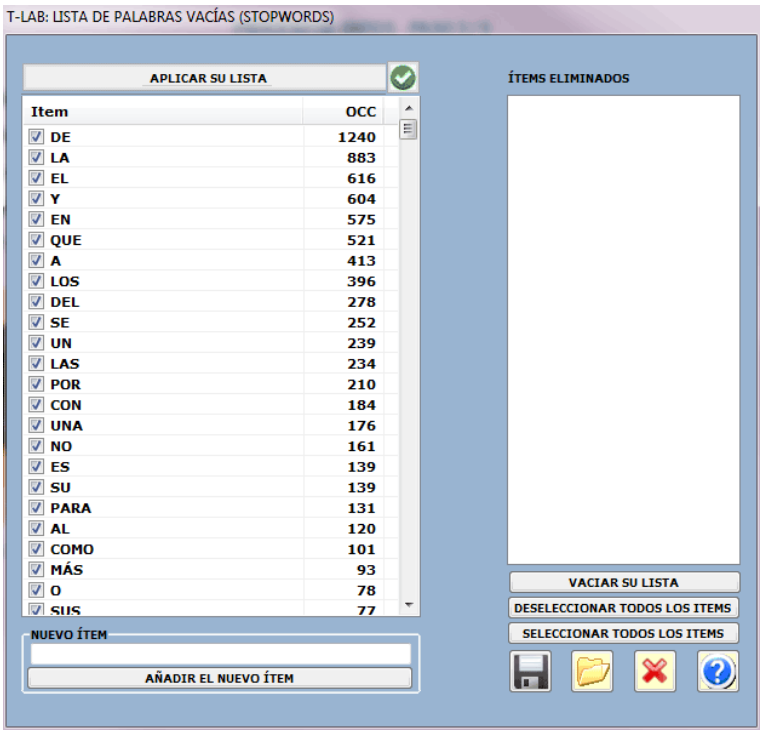
3 - CONTROL DE MULTI-PALABRAS

La opción "**Básico**" activa el uso automático de la lista **multi-palabras** de T-LAB.
Diferentemente la opción "**Avanzado**", disponible solamente con la lematización automática, permite las operaciones siguientes:
- verificar y modificar la lista de multi-palabras no incluidas en base de datos de T-LAB;
- importar y utilizar **listas personalizadas de multi-palabras** (archivos Multiwords.txt).



4 - CONTROL DE PALABRAS VACÍAS

La opción "**Básico**" activa el uso automático de la lista palabras vacías de T-LAB.
Diferentemente la opción "**Avanzado**" permite las operaciones siguientes:
- verificar y modificar la lista de palabras vacías presentes en el corpus;
- importar y utilizar **listas personalizadas de palabras vacías** (archivos StopWords.txt).



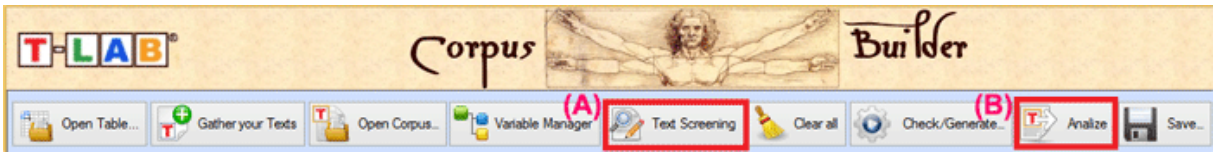
5 - SELECCIÓN DE PALABRAS CLAVE

Las opciones disponibles permiten que elijamos el método de la selección (**TF-IDF** o **Chi-cuadrado**) y la cantidad máxima de **unidades lexicales** que se incluirán en una lista usada por **T-LAB** para analizar los textos con la **configuración automática**.

NOTA: Al término de la fase de importación, utilizando la **configuración personalizada**, el usuario puede revisar la selección de palabras clave y crear varias **listas** para ser aplicadas.

Preparar un Corpus (Corpus Builder)

NOTA: Las imagenes contenidas en este apartado hacen referencia a una versión anterior de T-LAB, ya que el interfaz de **T-LAB 10** cambia ligeramente. Además**,** esta herramienta incluye dos ulteriores botones: a) uno para activar la opción **Text Screening**, en el caso de que el corpus no supere los 20 MB, y b) otro que permite la **importación** inmediata de los materiales textuales (véase imagen siguiente).



Esta herramienta ha sido diseñada para facilitar tanto la preparación de diferentes materiales textuales, como su transformación en un único archivo **corpus** listo para la importación en **T-LAB**.

De forma más concreta, esta herramienta permite ejecutar rápidamente las siguientes operaciones:

.

1. **Importar** automáticamente distintas tipologías de archivos;
2. **Editar** y modificar los textos importados;
3. Gestionar el uso de **variables categóricas**;
4. **Guardar** el resultado de un trabajo en un archivo que pueda ser directamente importado por **T-LAB**;
5. **Verificar** y **modificar** cualquier archivo corpus cuyo formato sea compatible con **T-LAB**.

Si bien la manera de importar los archivos (véase arriba '1') varia en base a los formatos que estos tengan, todas las demás operaciones siguen el mismo procedimiento.

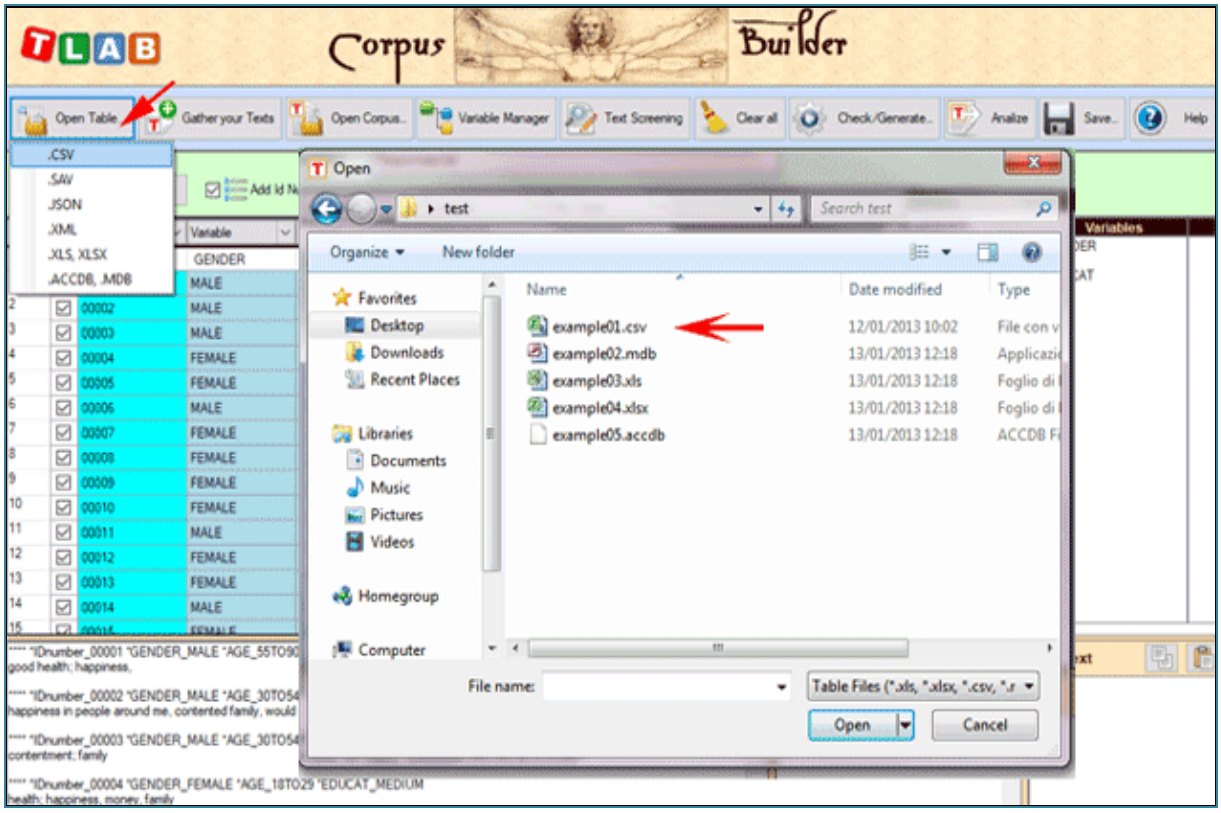
A continuación, se propone una breve descripción de las maneras de importar las diferentes tipologías de archivos.

A - **Importar un archivo en formato de tabla** (CSV, .SAV, .JSON, .XML, .XLS, XLSX, .MDB, .ACCDB).

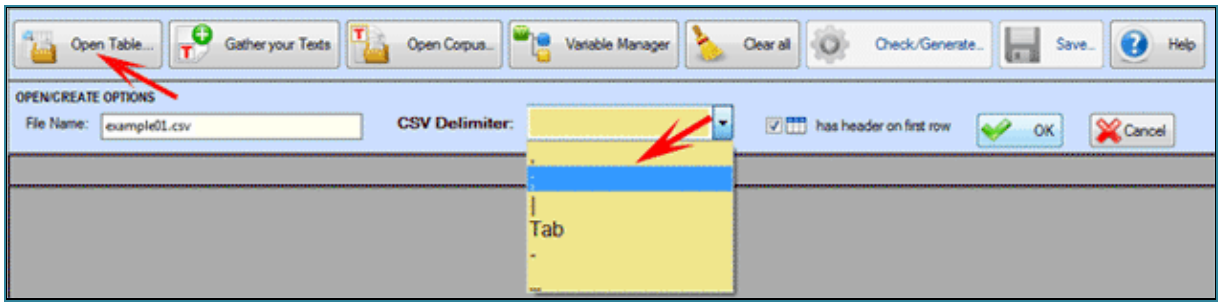
Para importar un **único archivo** que incluya hasta 30.000 entradas se puede utilizar tanto la opción 'Open Table' como el método drag and drop (NB: cuando ningún texto supera los 2.000 caracteres, se extiende a 99.999 el número máximo de entradas que se pueden importar).

Dicho archivo puede estar compuesto por diferentes columnas. Éstas pueden contener diferentes tipologías de datos:
-Variables categoriales (una por cada columna, hasta un máximo de 50)
-Textos a analizar (sólo una columna)
-IDnumbers. Es decir, identificativos de unidades de contexto o sujetos/casos.

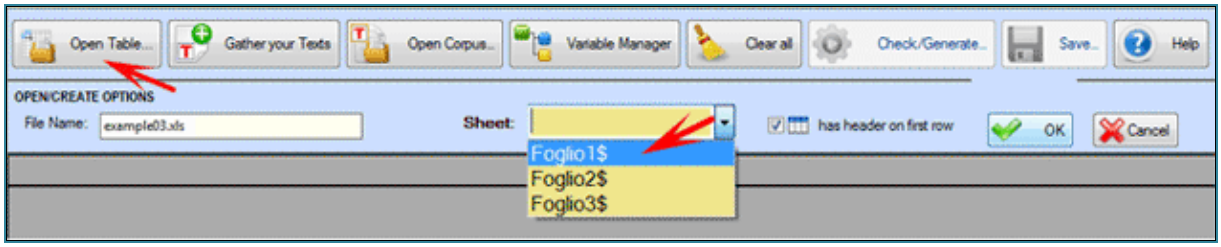
N.B.: La presencia de variables categoriales y IDnumbers es opcional. Sin embargo, la presencia de por lo menos una columna de texto es obligatoria.



A la hora de importar un archivo en formato .CSV, es necesario seleccionar de forma apropiada el delimitador a usar (véase abajo).



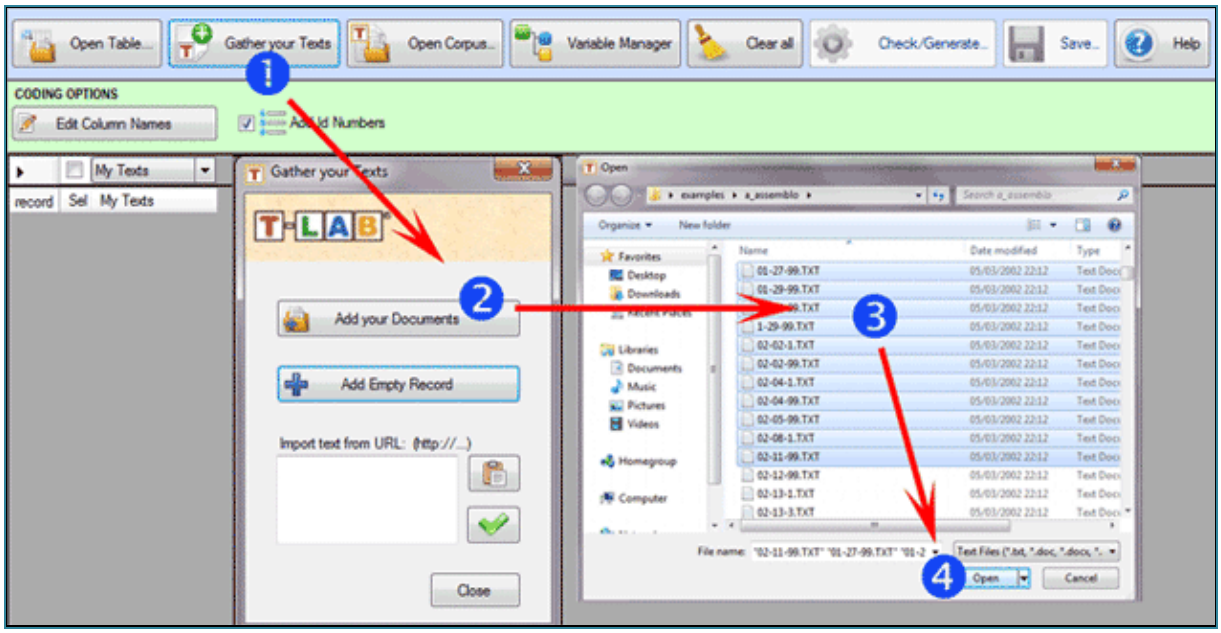
Cuando se importan archivos Excel o Access, sólo se puede seleccionar una tabla (véase abajo).



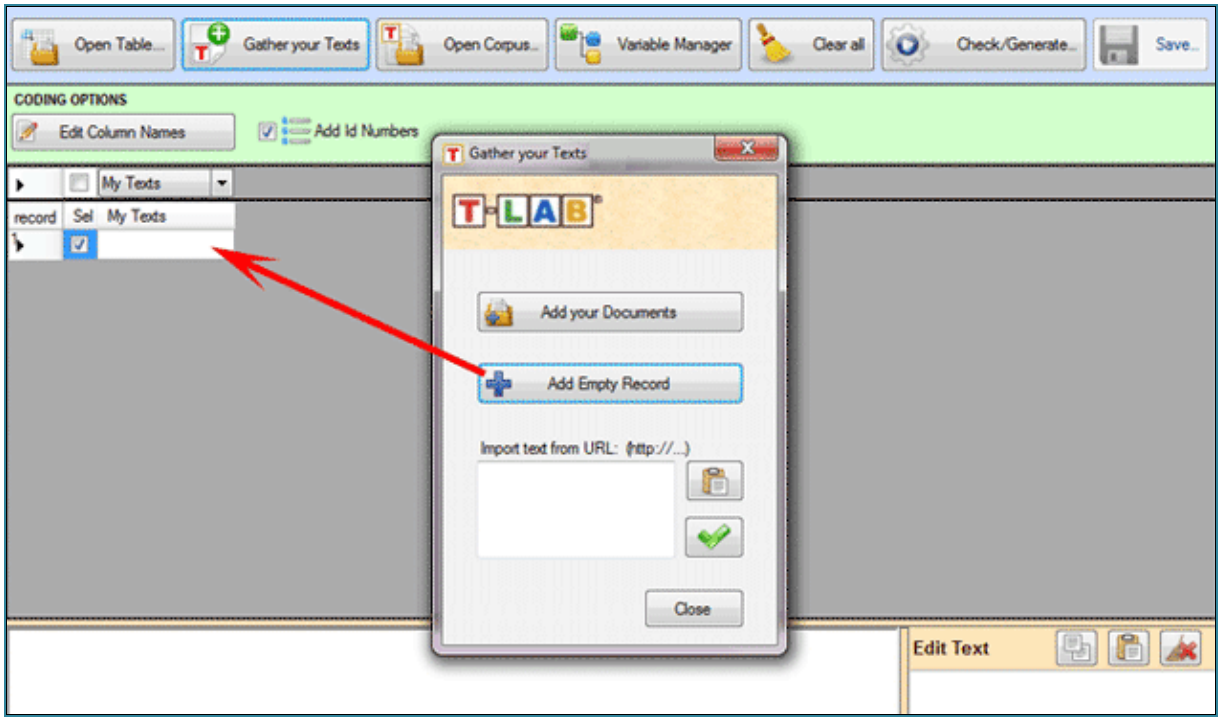
B - **Importación de textos y documentos** (.TXT, .DOC, .DOCX, .PDF, .RTF, .HTML).

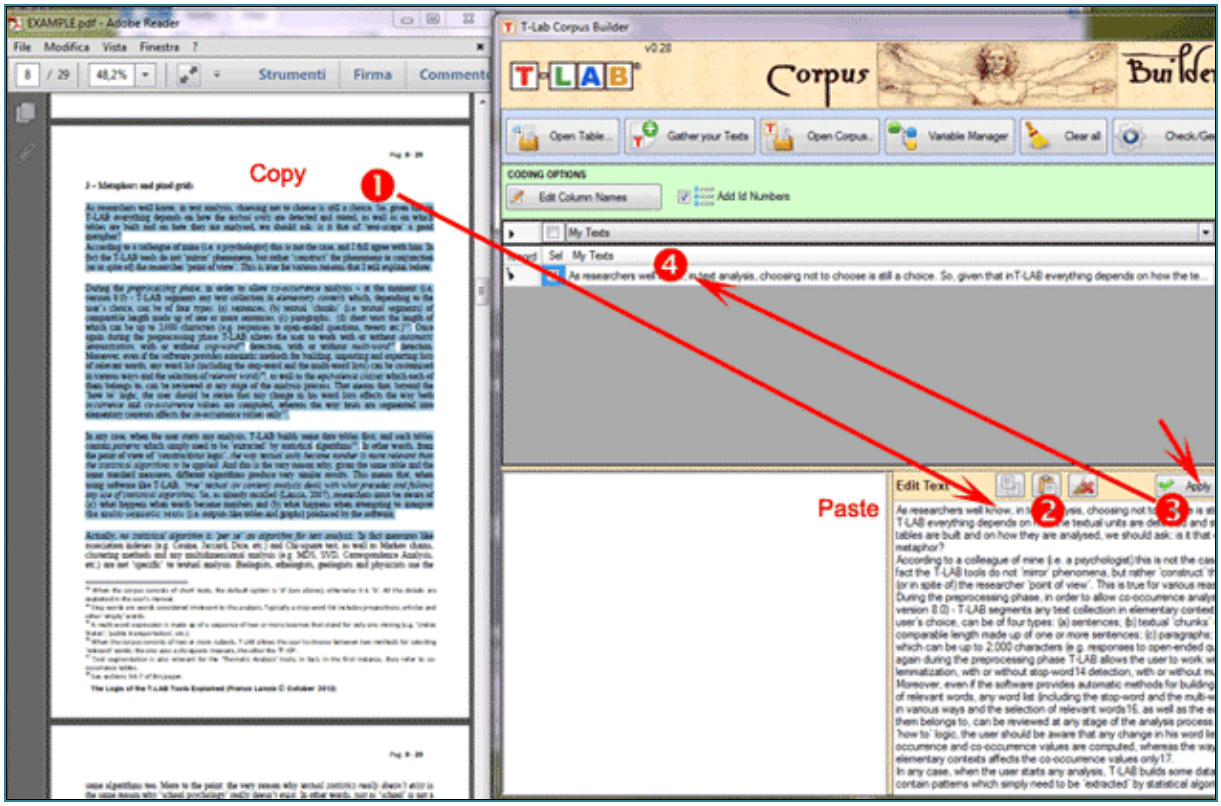
La opción 'Gather your Texts' (véase abajo) permite importar hasta un máximo de 30.000 documentos, bien de forma individual, bien mediante selección múltiple. Todo ello, mediante **3 posibles procedimientos**.

El **primer procedimiento** ('Add your Documents') prevé la importación automática de archivos .TXT, .DOC, .DOCX, .PDF, .RTF.



El **segundo procedimiento** ('Add EmptyRecord') permite importar records de forma individual. En cada uno de ellos es posible copiar/pegar cualquier tipología de texto (véase abajo).





El **tercer procedimiento** ('Import Text from URL') permite descargar archivos HTML directamente desde Internet. A la vez, permite editar el contenido de estos archivos y, consecuentemente, importarlos **a T-LAB**.

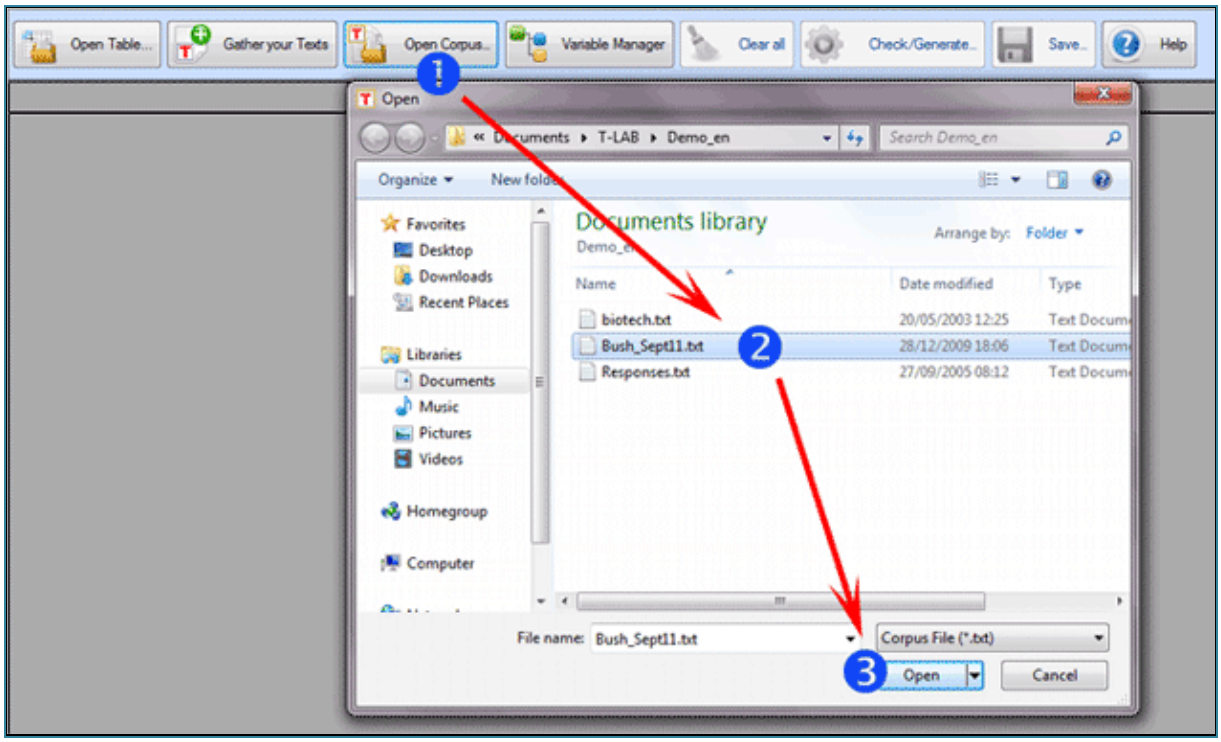
C - **Importación de un corpus** ya codificado según los criterios de compatibilidad con **T-LAB.**

El uso de la opción 'Open Corpus' está especialmente pensado para los siguientes casos:

1 – Si el usuario quiere modificar la estructura de un archivo corpus ya codificado (eje. Añadir otros textos mediante los métodos presentados en la anterior sección 'B', modificar los nombres de las variables, y/o de las modalidades, etc.);
2 – Si el usuario quiere verificar/corregir los eventuales errores que pueda presentar una codificación del corpus realizada manualmente (es decir, sin utilizar la herramienta Corpus Builder);
3 – Si el usuario quiere importar un archivo corpus que tenga una codificación "bruta" (véase imagen siguiente). Es decir un archivos cuyas partes (documentos o entradas/records) están precedidas exclusivamente de una línea de texto compuesta por 4 asteriscos seguidos por un espacio ('**** ').

```
**** ¶
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
```

Para implementar cada una de las opciones recién descritas es suficiente seleccionar un único archivo mediante la opción 'Open Corpus'. También se puede arrastrar el archivo utilizando el método drag and drop.

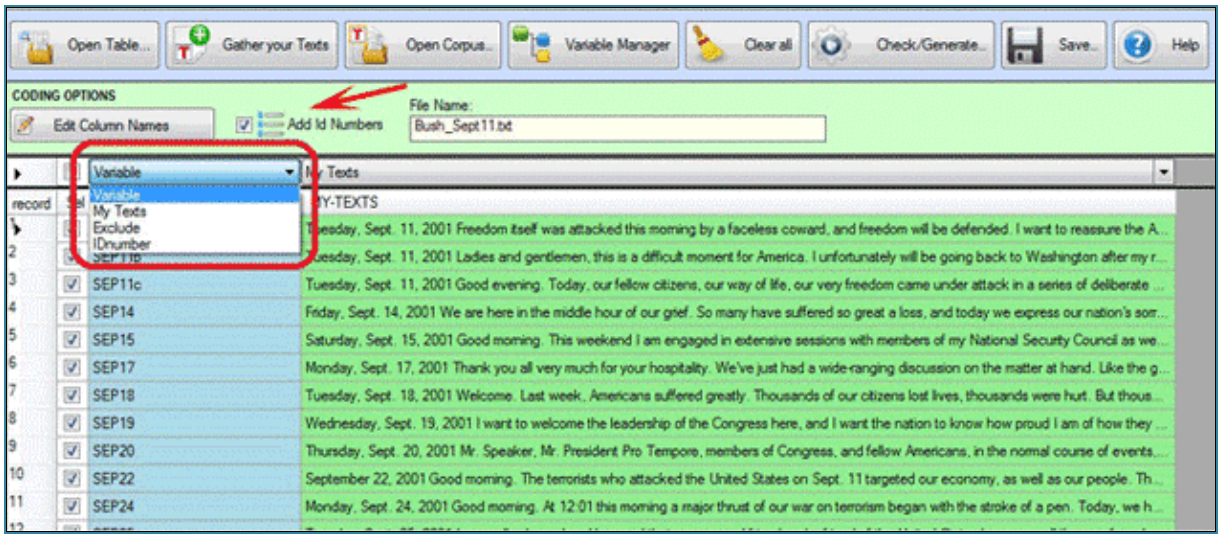


Operaciones posteriores a la importación del archivo

Una vez finalizada la fase de importación de los archivos mediante Corpus Builder, es posible escoger la opción 'Check /Generate' y – sucesivamente – guardar el corpus a importar en **T-LAB**. Esto, tanto en el caso en que no se quiera utilizar variables como en el caso en que ya se hayan efectuado las operaciones de codificación.

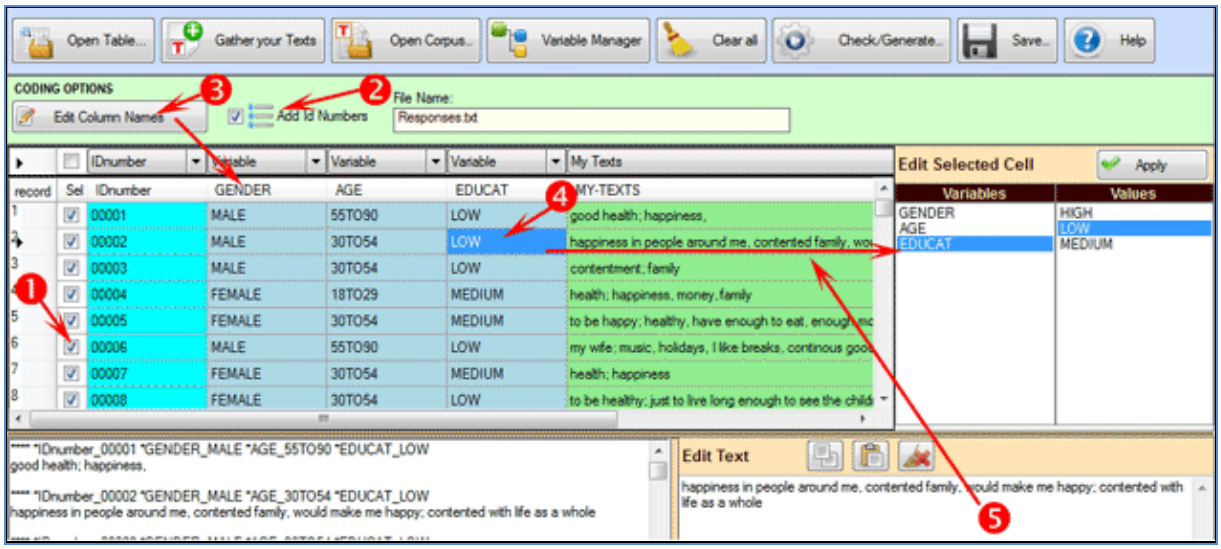
Si el corpus contiene codificaciones, es importante acordarse de que, en cada uno de los 3 métodos de importación descritos anteriormente (‘A’, ‘B’, ‘C’), los datos se visualizan en columnas diferentes. Éstas pueden tener distintas etiquetas:

- ‘Variable’, es decir variables categóricas cuyo uso es necesario cuando se quieren analizar las características de diferentes subconjuntos del corpus y las relaciones entre dichos subconjuntos;
- ‘IDnumber’, es decir, identificadores de casos/entradas y cuyo uso es opcional;
- ‘My Texts’, es decir textos a analizar y cuyo uso, obligatorio, está vinculado a una única columna;
- ‘Exclude’, se usa para indicar a Corpus Builder que no se deben utilizar los datos contenidos en la columna correspondiente.



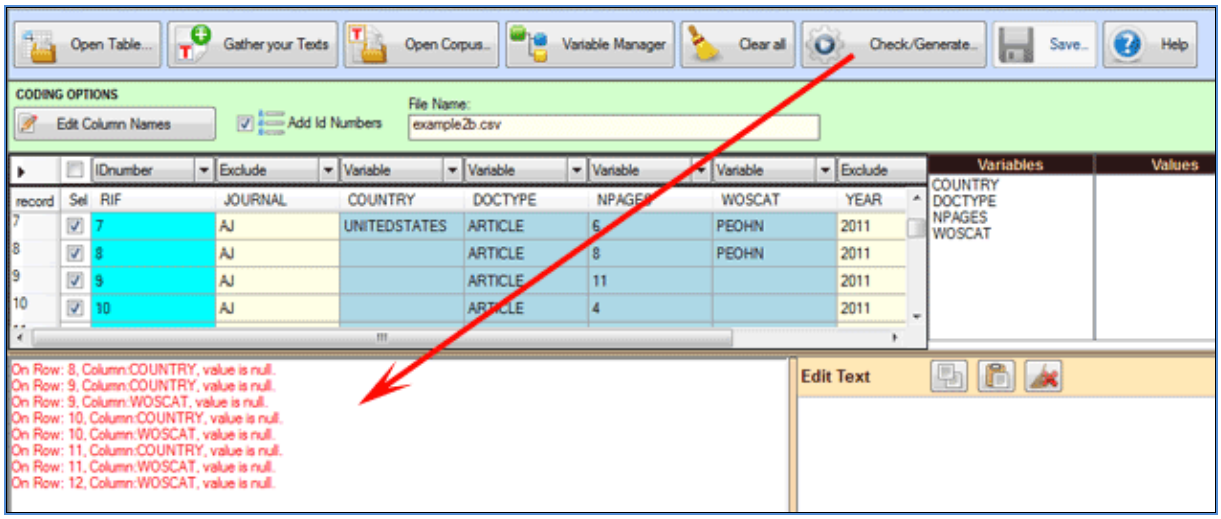
Para todos los casos, valen las siguientes indicaciones:

- Para cada entrada, existe tanto la posibilidad de seleccionar como de deseleccionar (véase abajo '1 ');
- Los IDnumber pueden ser añadidos de forma automática (véase abajo '2');
- Los nombres de las variables pueden ser editados y modificados (véase abajo '3');
- Cada valor de la variable puede ser editado y modificado (véase abajo '4');
- Cada campo ‘My Texts’ puede ser editado y modificado (véase abajo ‘5’).



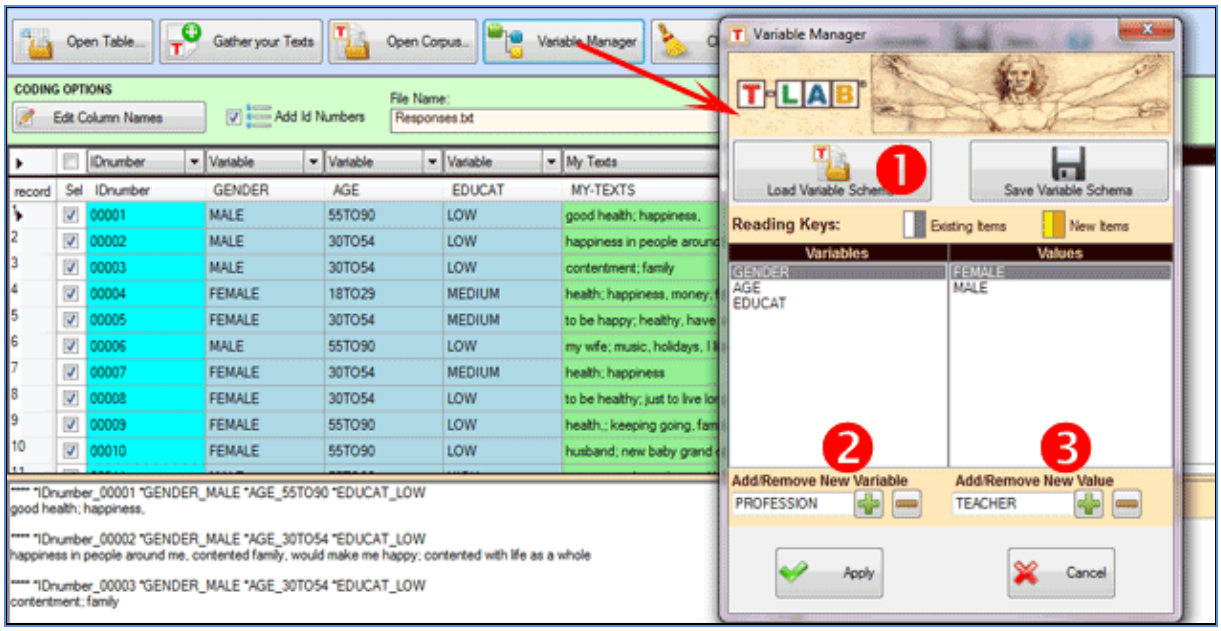
Otros aspectos a tener en cuenta son:

- El número máximo de columnas con variables categóricas es 50. Además, cada una las variables debe tener un numero de valores comprendido entre un mínimo de 2 y un máximo de 150;
- En el caso de utilizar los IDnumber, sus valores deben ser progresivos y empezar por el 1 (eje. 1, 2, 3, etc.);
- Tanto en el caso de las variables como en el de las modalidades, cada etiqueta debe tener una extensión no superior a los 25 caracteres alfanuméricos (min. 2) no tener espacios;
- En el modulo Corpus Builder los errores se visualizan en el cuadro abajo a la izquierda (véase abajo).



Uso de la herramienta Variable Manager

La herramienta 'Variable Manager' permite construir, editar, modificar y guardar cualquier esquema de codificación, incluso aquellos provenientes de corpus diferentes.
Cada esquema incluye el elenco de las variables y de sus valores correspondientes (véase abajo).

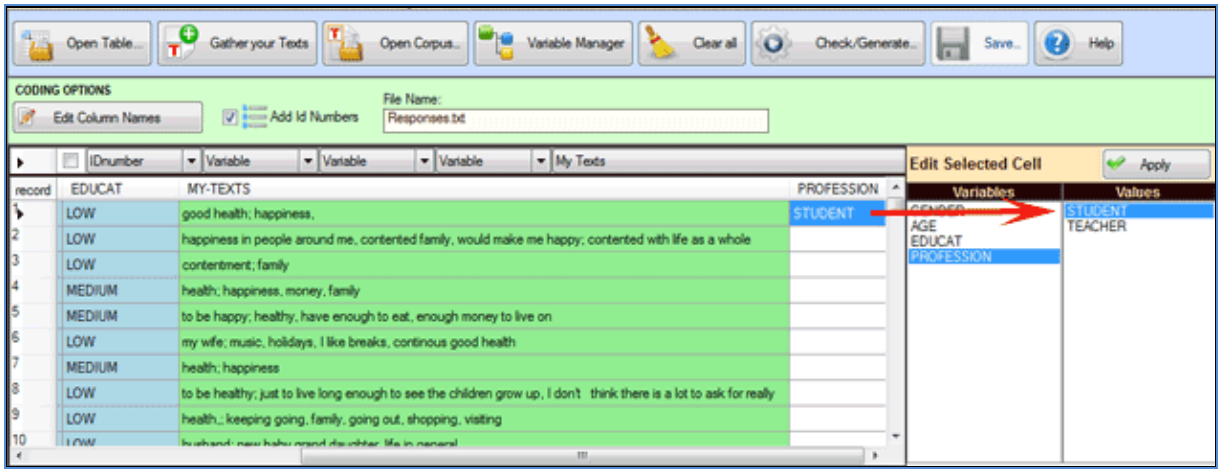


Para añadir las variables de otro corpus o de un esquema guardado anteriormente, es necesario seleccionar la opción '1' (véase arriba). Sin embargo, para añadir manualmente las variables y sus valores, hay que utilizar las opciones '2' y '3' consecutivamente. (véase arriba).

Para añadir los valores de las variables a las entradas es necesario proceder manualmente en una única sesión de trabajo (véase abajo). Esto es así, porque al guardar un esquema no se incluyen las modificaciones aportadas a cada entrada. De esta forma, en el caso en que el usuario tenga que codificar manualmente un corpus que incluya un número considerable de entradas y/o necesite más de una sesión de trabajo, se aconseja proceder de la siguiente manera:

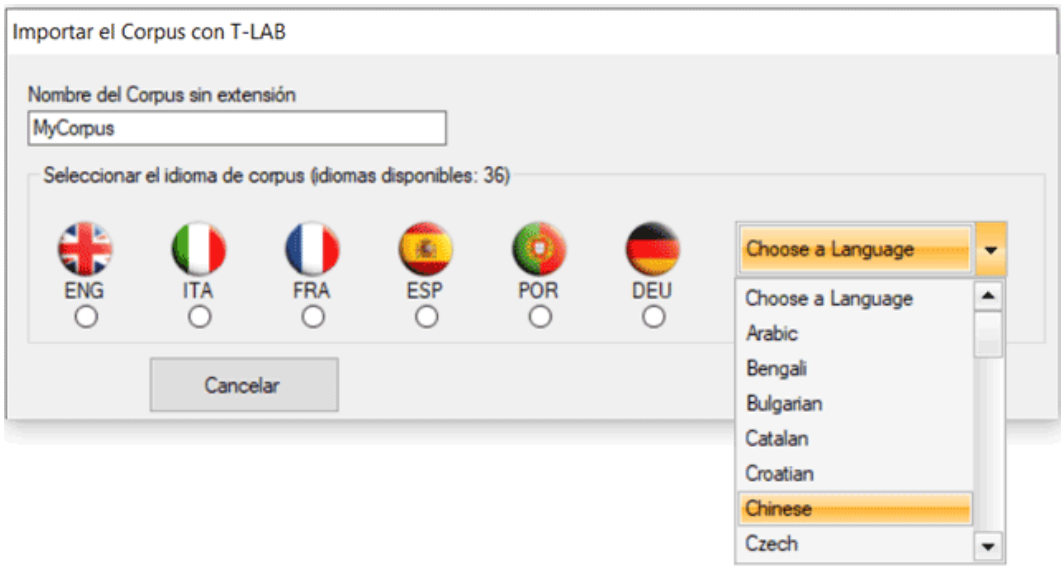
1 – Importar todos los archivos/records que se considera se puedan codificar en una única sesión de trabajo;
2 – guardar el trabajo como corpus (véase opción 'Save' del menú Corpus Builder).

Después, en la sesión siguiente, reimportar el corpus anteriormente guardado (véase arriba, punto '2'), y añadir otros records/archivos y continuar.



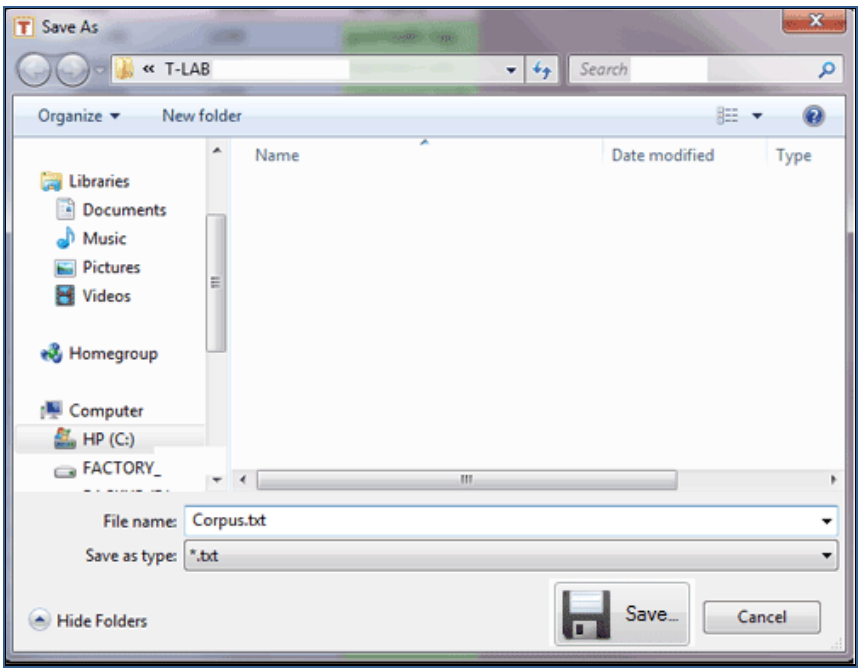
Una vez el usuario haya terminado las operaciones que considere oportunas, éstas pueden ser chequeadas mediante la opción 'Check/Generate'. Si todo está bien hecho, se puede exportar (A) o guardar (B) un corpus listo para la importación a **T-LAB.**

En el primer caso (A - véase abajo), Corpus Builder crea una nueva carpeta en el directorio ".. \ Mis documentos \ T-LAB PLUS \" y empieza automáticamente la importación del corpus.
NB: En este caso, la nueva carpeta tiene el mismo nombre del corpus.



En el segundo caso (B - véase abajo) el usuario puede guardar su corpus en cualquier carpeta desee. Sucesivamente, tiene que utilizar la opción de T-LAB "Importar un corpus".

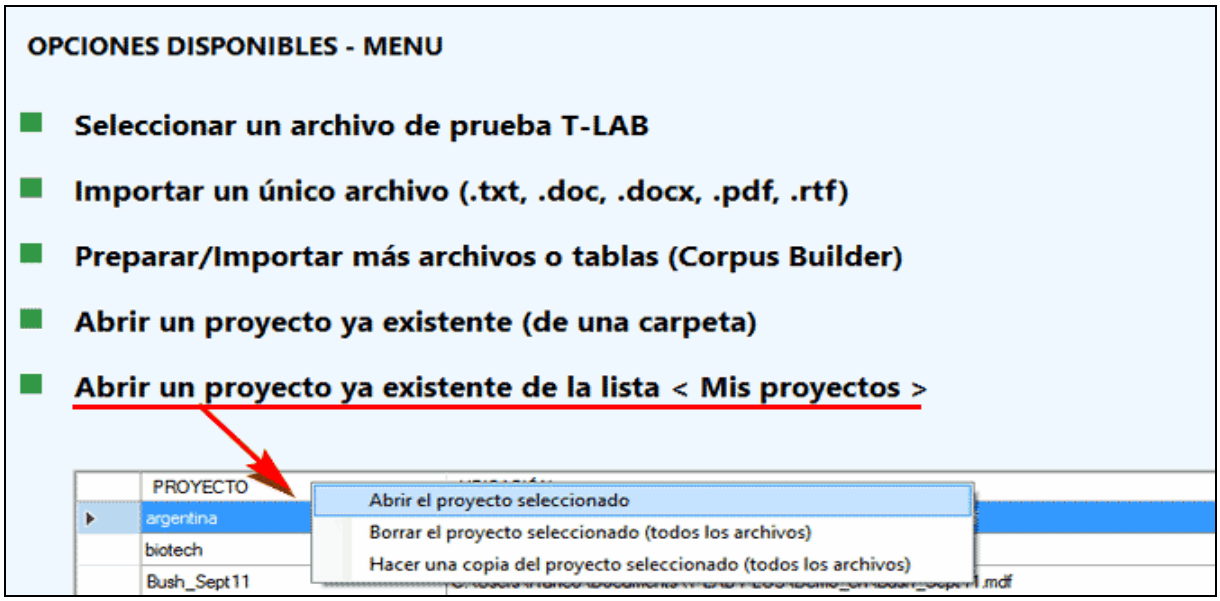
NB: En este caso, se recomienda crear – todas las veces – una carpeta de trabajo que contenga sólo el archivo a importar.



Abrir un Projecto ya existente

Gracias a esta opción el usuario puede volver a trabajar con un proyecto ya empezado. Esto es posible escogiendo el archivo bien de en una carpeta ya existente o bien de un listado sugerido por **T-LAB**.

Además, al seleccionar un ítem dentro del listado sugerido por **T-LAB**, el usuario puede eliminar los archivos a ello relacionados. También puede realizar una copia de seguridad de los mismos en otra carpeta. Ambas operaciones se implementan utilizando el botón derecho del ratón.



HERRAMIENTA LEXICO

Text Screening / Desambiguación de Palabras

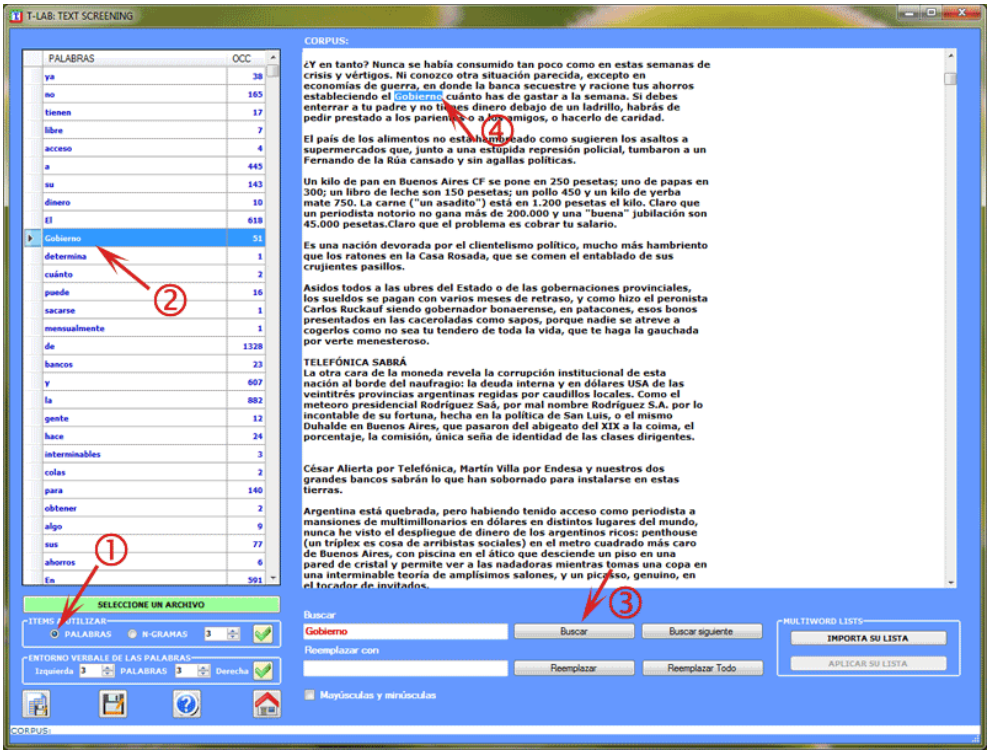
Esta herramienta de **T-LAB** permite editar cualquier archivo corpus (hasta 30 Mb de tamaño) e implementar un conjunto de operaciones que sirven tanto a la **exploración** de sus contenidos como a la **desambiguación** de unidades lexicales concretas.

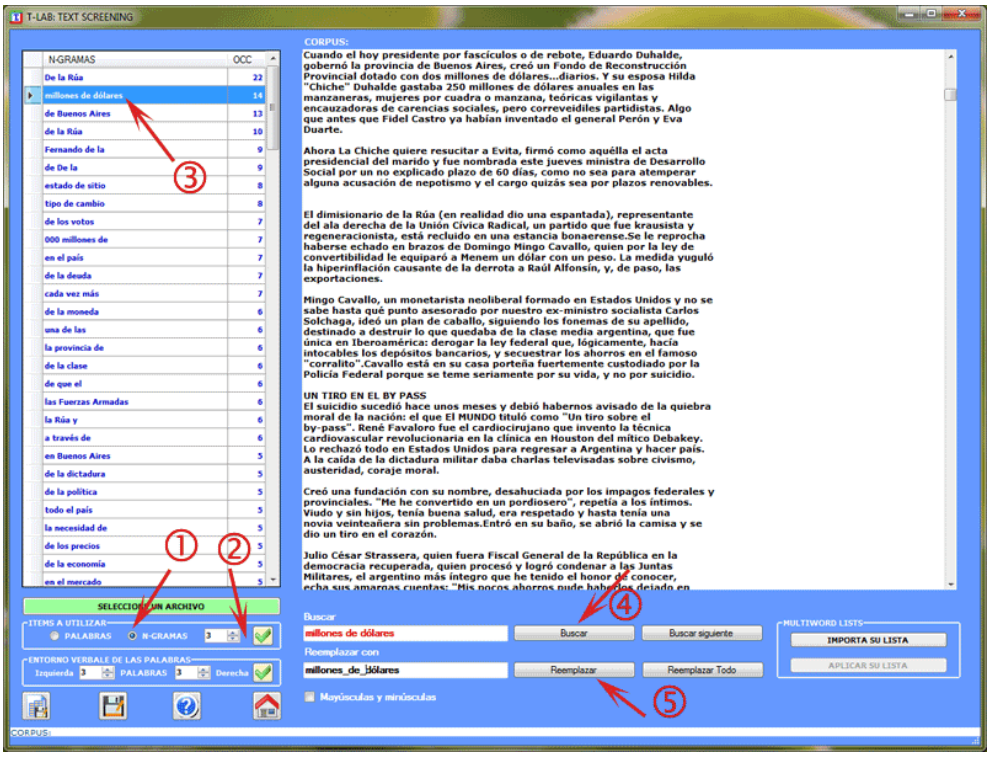
Más en concreto, esta herramienta produce de forma muy rápida un conjunto de **listados** y permite implementar operaciones relacionadas a las modalidades **buscar**/**reemplazar**.

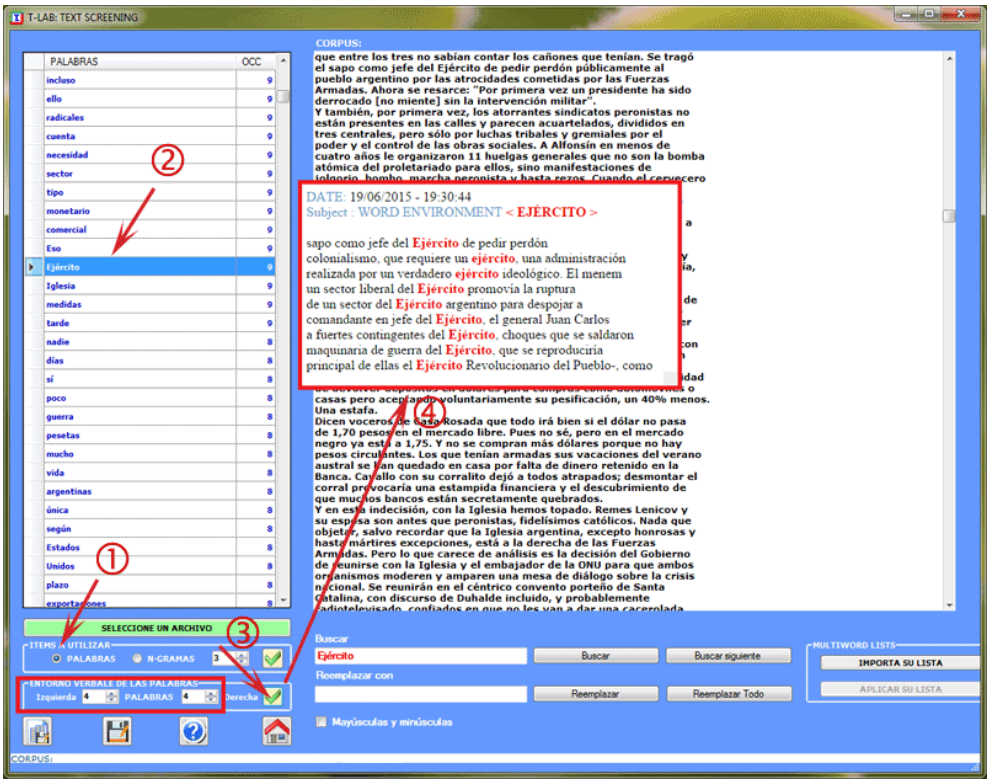
Las diferentes tipologías de listados que se pueden obtener mediante esta herramienta son:

a- **palabras** con sus respectivas ocurrencias (véase abajo la imagen 1);
b- **n-gramas** de palabras con sus ocurrencias (véase abajo la imagen 2);
c- **entornos verbales** de las palabras seleccionadas (véase abajo la imagen 2).

Las imágenes a continuación muestran las posibles operaciones en los tres casos (a-b-c)

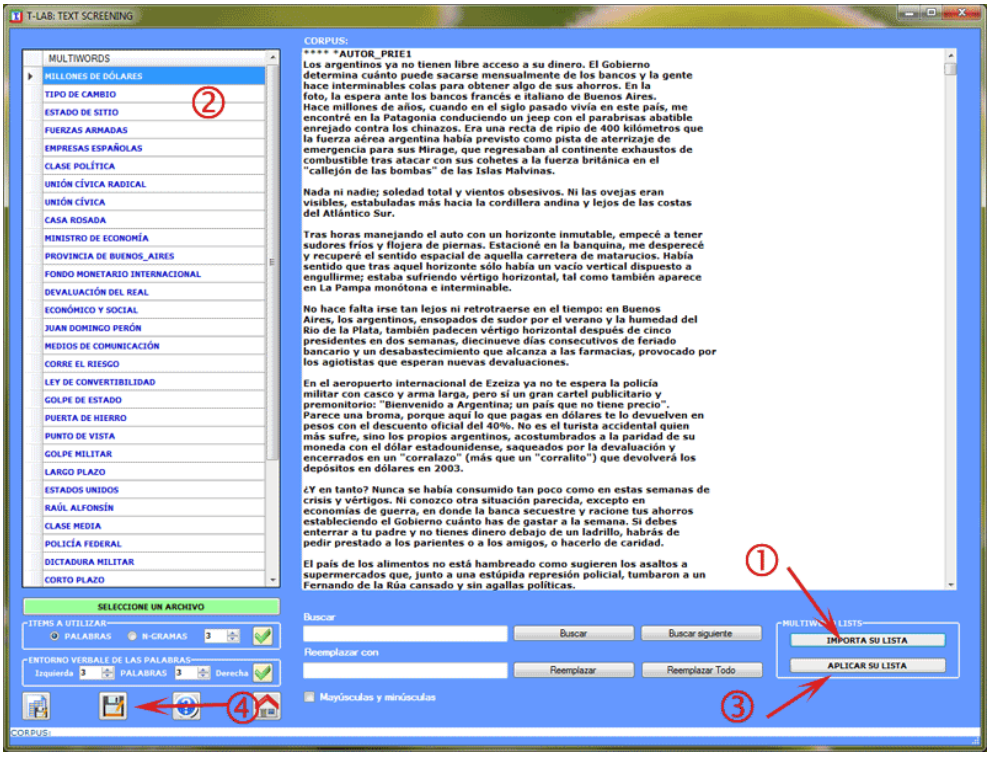






NOTA: Cliqueando el botón de abajo a la izquierda es posible exportar los listados de tipo 'a' y 'b' en formato Excel. Por otra parte, los listados de tipo 'c' se exportan automáticamente en formato .html.

Además, es posible importar listados personalizados de **Multiword**. Eventualmente, las multiwords pueden ser aplicada al corpus visualizado (véase imágen siguiente).



Una vez terminadas las operaciones, si el usuario ha modificado el texto y desea guardarlo, **T-LAB** permite crear un nuevo archivo (corpus_dis.txt). Una vez que a dicho archivo se le haya puesto el nombre más cómodo para el usuario, será posible su sucesiva importación análisis.

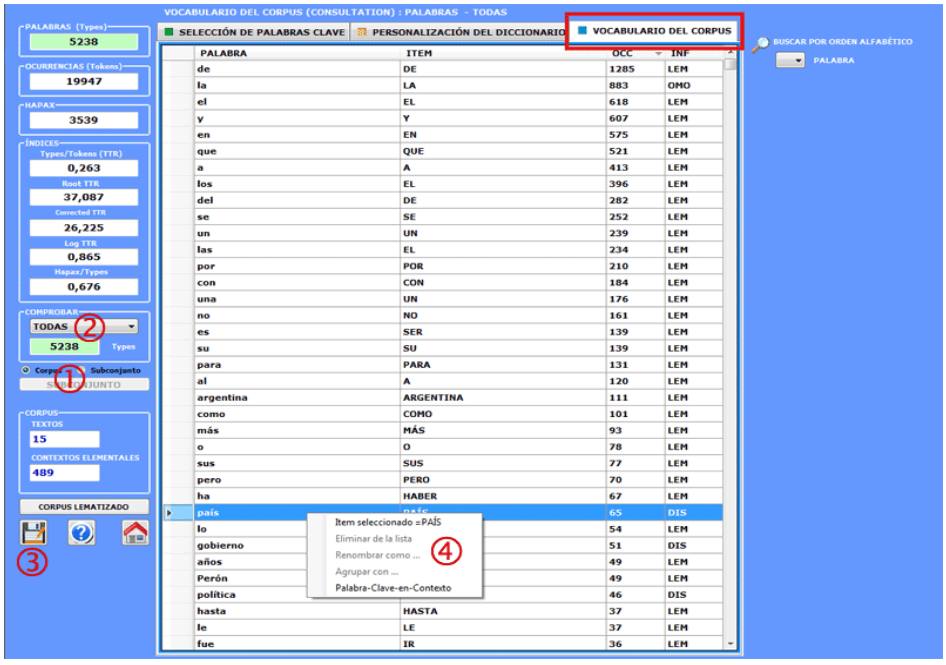
Vocabulario del Corpus

Esta herramienta de **T-LAB** nos permite comprobar el **Vocabulario** del corpus y de sus subconjuntos (véase abajo la opción '1').
Por otra parte se proporcionan algunas medidas de **riqueza léxica**.

La tabla Vocabulario es una lista que incluye todas las palabras distintas (es decir "word types"), la cantidad de sus ocurrencias (es decir "word tokens"), los lemas correspondientes y algunas categorías usadas por **T-LAB** (véase Glosario/Lematización).

El usuario puede seleccionar (véase abajo la opción '2') las unidades léxicas que pertenecen a cada categoría, consultar la tabla correspondiente y exportarla como archivo .xls (véase abajo la opción '3').

Además, usando el botón derecho del ratón, es posible verificar las **concordancias** (Key-Word-in-Context) de cada palabra (véase abajo la opción '4')



Las medidas de riqueza léxica son cinco:

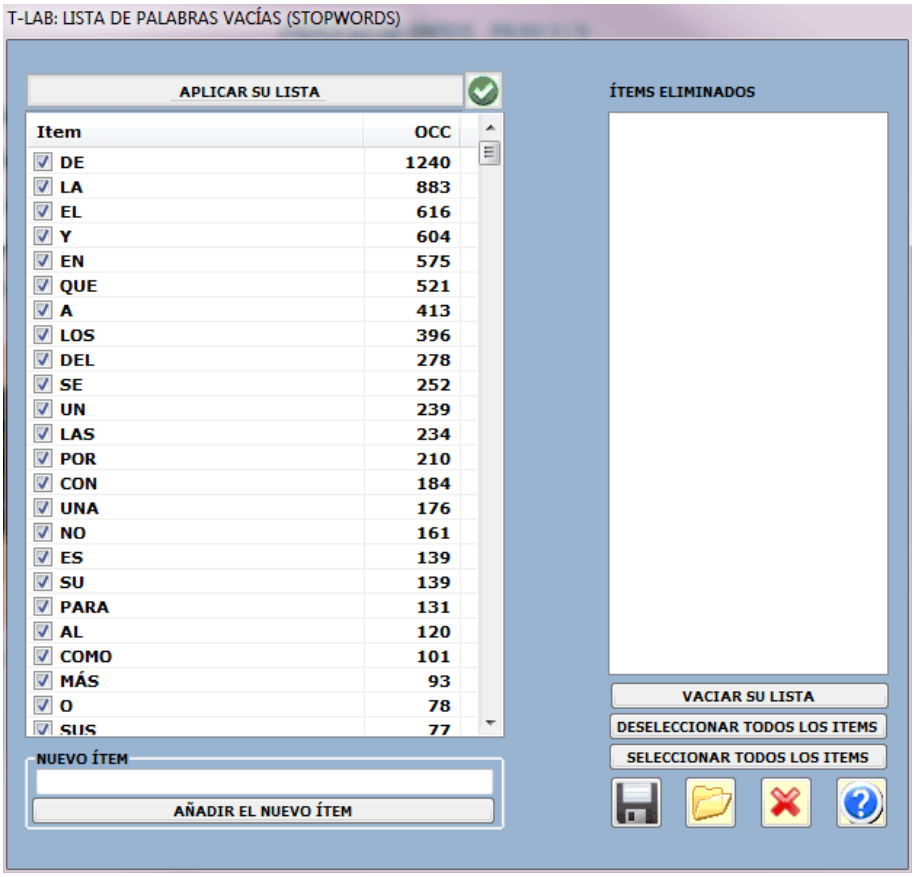
Type/Token ratio (TTR);
Root TTR (Guiraud, 1960), obtenida dividiendo el número de "types" por la raíz cuadrada del número de "tokens";

Corrected TTR (Carroll, 1964), obtenida dividiendo el número de "types" por la raíz cuadrada de dos veces el número de "tokens";
Log TTR (Herdan, 1960), obtenida dividiendo el logaritmo del número de "types" por el logaritmo del número de "tokens";
Hapax/Types ratio.

NOTA:
- Hapax (es decir Hapax Legomena) son las palabras que, en un corpus, ocurren solamente una vez; - cuando se analiza un subconjunto del corpus, todas las medidas de riqueza léxica no incluyen las palabras vacias (e.j. los artículos y las preposiciones).

Palabras Vacías

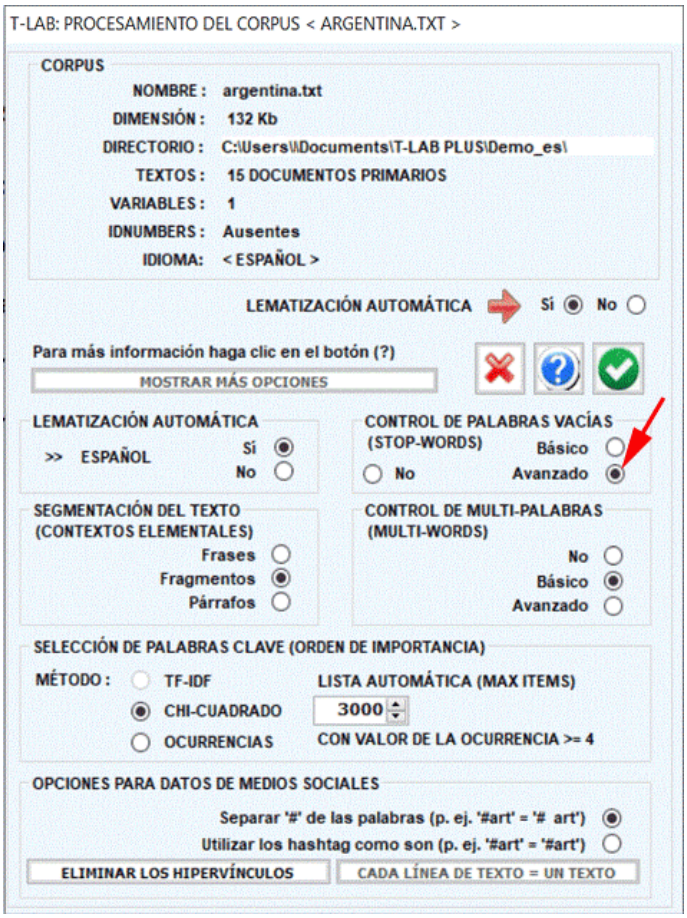
Esta opción permite crear/modificar listas de **Stop-Words** (Palabras Vacías) en la ventana siguiente:



En los archivos StopWord.txt preparados por el usuario deben ser respetadas las reglas siguientes:

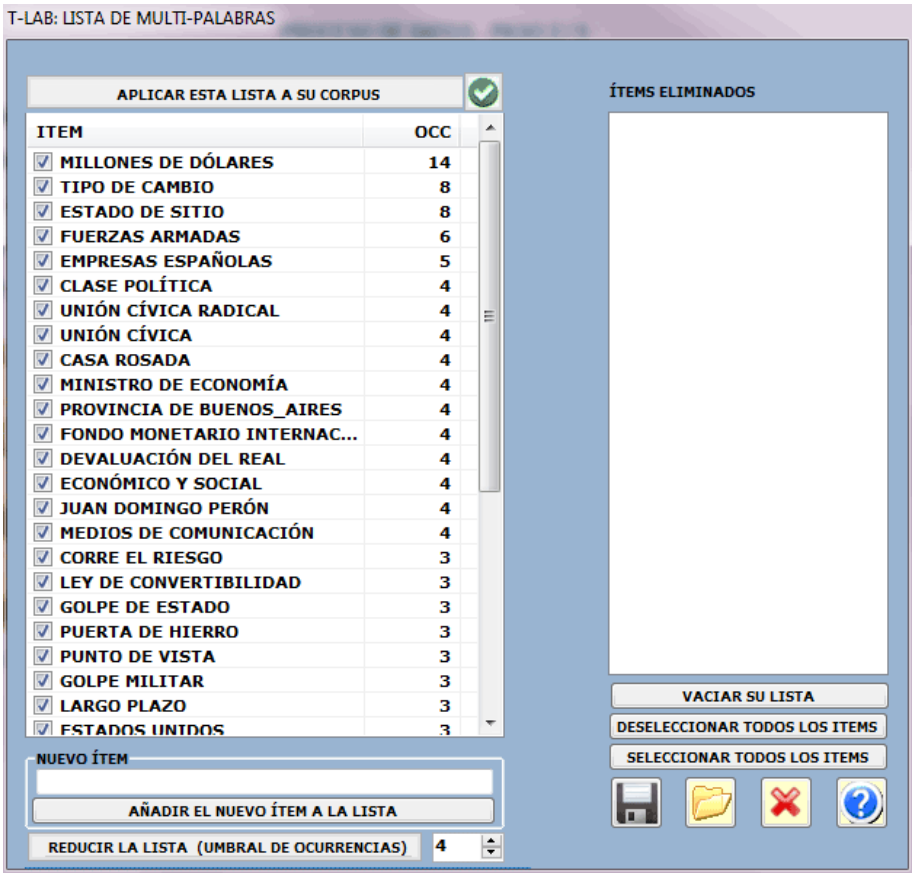
- la longitud máxima de una palabra es 50 caracteres
- no tiene que haber ni espacios en blanco ni signos de puntuación.

En todos casos, durante la importación de un **nuevo corpus** el usuario que desea verificar/utilizar listas de StopWords, tiene que seleccionar la opción "Avanzado" en la ventana siguiente:



Multi-palabras

Esta opción permite crear/modificar las listas de **Multi-Palabras** (Multi-Words) en la ventana siguiente.



Cada lista (archivo Multiwords.txt) tiene que ser compuesta de N líneas (máximo 5000), cada una con un conjunto de dos o más palabras (longitud máxima: 50 caracteres, sin signos de puntuación).

He aquí algunas líneas de Multiwords.txt en el formato correcto:

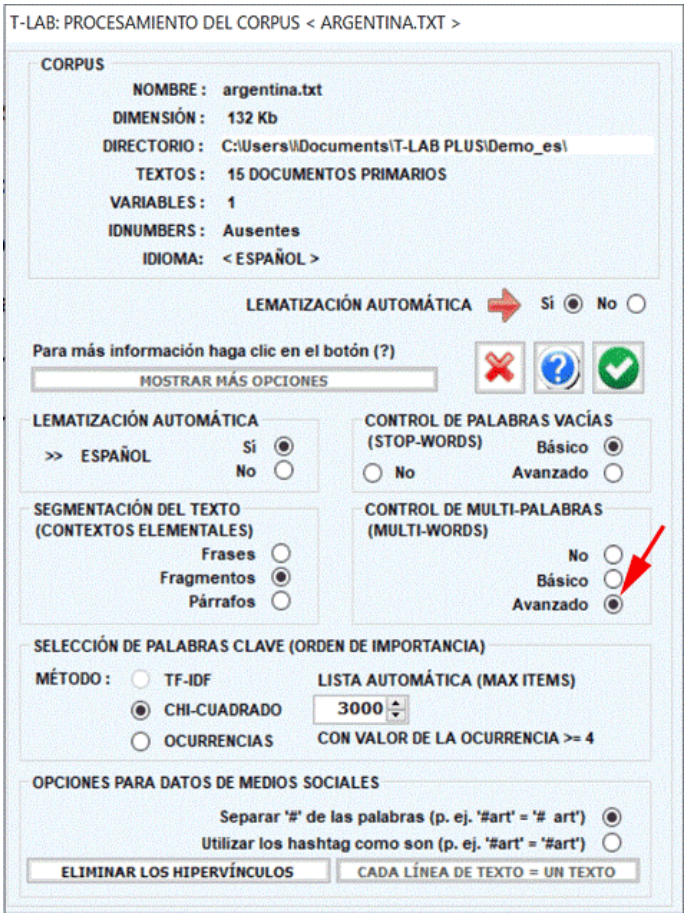
transporte público
sistema de información
banco de órganos

etc etc

Chascando en el botón "**Aplicar esta lista...** ", el usuario puede producir una rápida transformación de las multi-palabras presentes en un corpus en cadenas que pueden ser reconocidas y clasificadas por **T-LAB** (por ej. "sistema de información" es transformado en "sistema_de_ información").

Después del funcionamiento, esta opción genera un nuevo archivo (**New_Corpus.txt**) que, correctamente retitulado, puede ser analizado por **T-LAB**.

Para verificar/utilizar listas de Multi-Palabras durante la **importación de un nuevo corpus** el usuario tiene que seleccionar la opción "Avanzado" en la ventana siguiente:

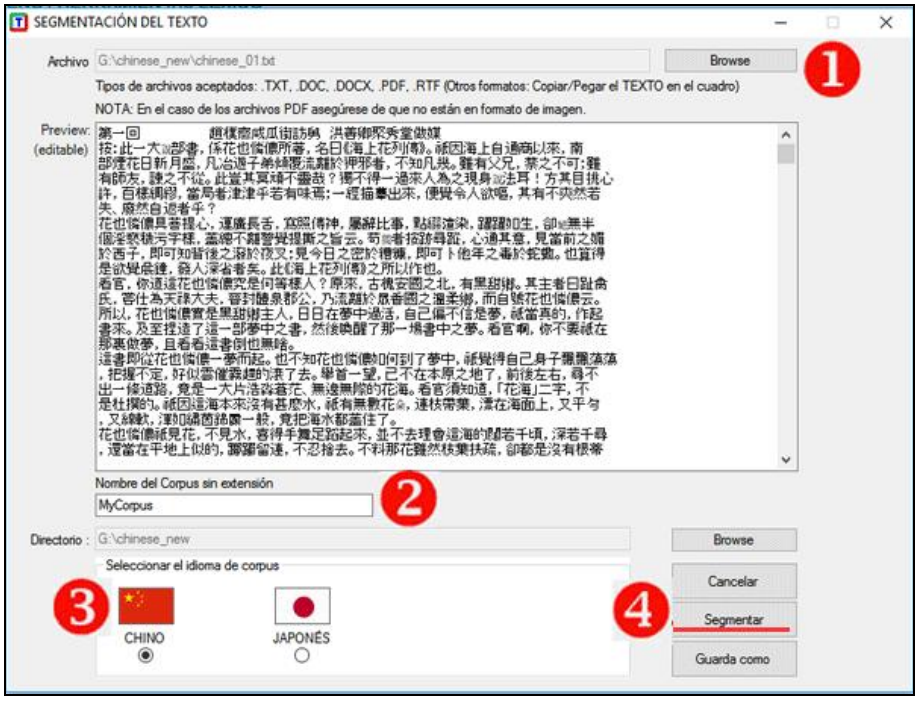


Segmentación de Palabras

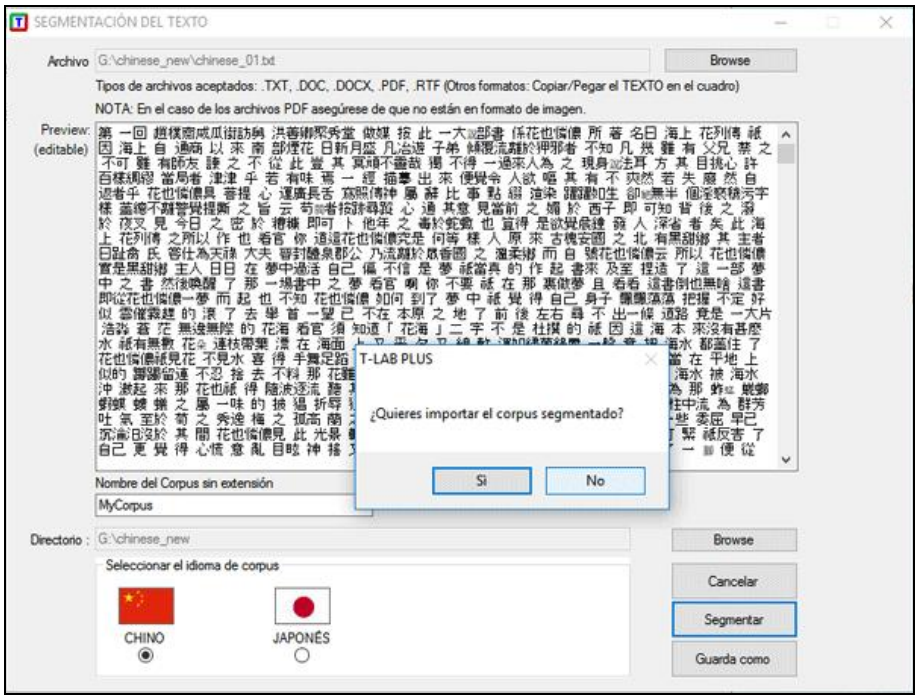
Esta herramienta de **T-LAB** puede ser utilizada previamente a la importación de cualquier texto escrito tanto tanto en chino como en japonés (*). Todo ello, siempre y cuando el texto en cuestión no presente delimitadores entre palabras (esto es, espacios y/o signos de puntuación).
(*) El texto a procesar puede estar compuesto bien por un unico documento o bien por un conjunto de documentos que incluyen variables categoriales.

Su uso es muy sencillo y se articula en 4 pasos (véase imagen abajo):

(1) seleccionar un archivo;
(2) escoger el nombre del proyecto;
(3) seleccionar la lengua del texto;
(4) seleccionar la opción 'Segmentar'.
Al final del proceso, el software habrá añadido espacios entre palabras.



En el caso de que, sucesivamente, se quiera proceder con la importación, será suficiente contestar 'Sí' a la pregunta "quieres importar el corpus segmentado?" (véase imagen abajo).



NOTA: En el caso de querer preparar un corpus compuesto por diferentes textos que incluyan líneas de codificación (esto es, variables categoriales), se aconseja proceder de la siguiente forma:

1- 'Fusionar' los textos no segmentados (*) mediante la herramienta Corpus Builder y, sucesivamente, guardar el archivo 'corpus';

2 - Importar el corpus recién creado mediante la herramienta Segmentación de Palabras.Luego, proceder según las indicaciones expuestas anteriormente.

(*) Lo cual implica que, a la hora de preparar el corpus, no es necesario segmentar previamente cada uno de los archivos.

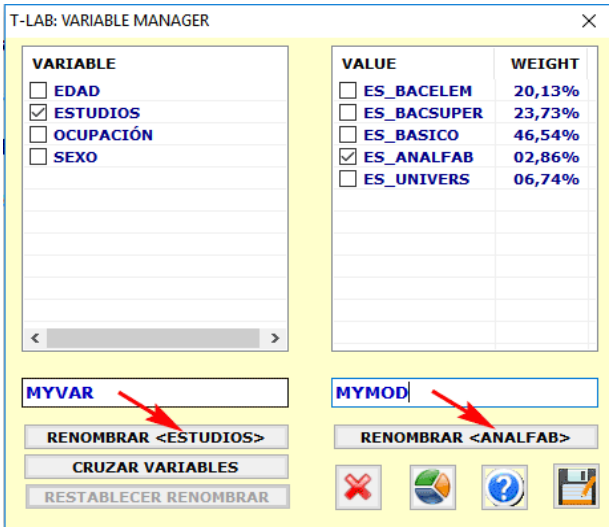
OTRAS HERRAMIENTAS

Variable Manager

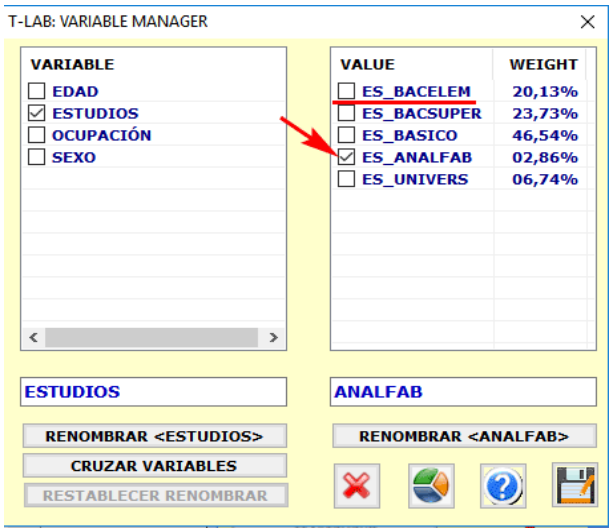
Esta opción, activa solamente cuando el corpus incluye particiones (variables y categorías), permite cinco tipos de operaciones:

a) **comprobar** las categorías de cada variable;

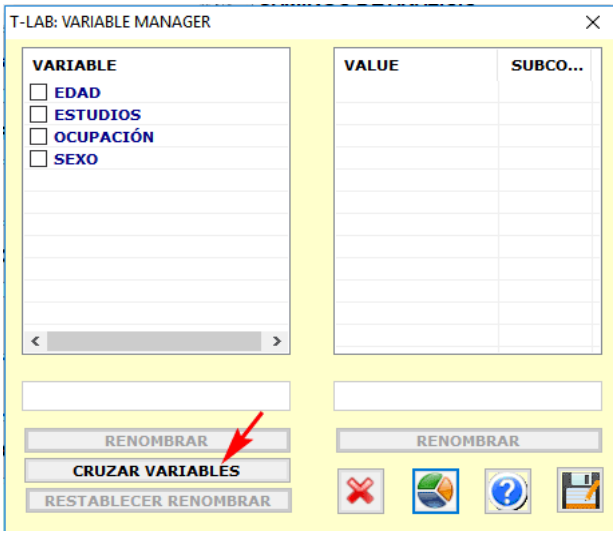
b) **renombrar** variables y categorías;

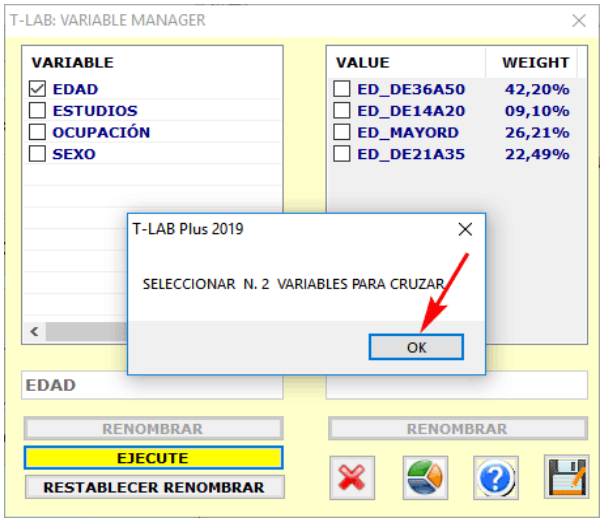


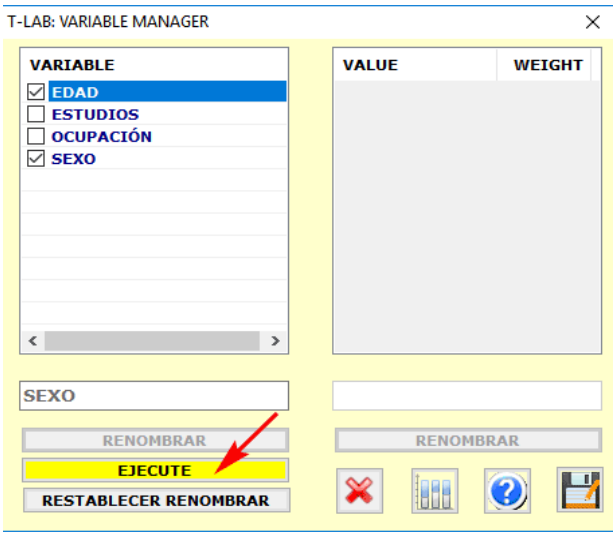
c) **agrupar** dos o más categorías asignándoles la misma etiqueta;

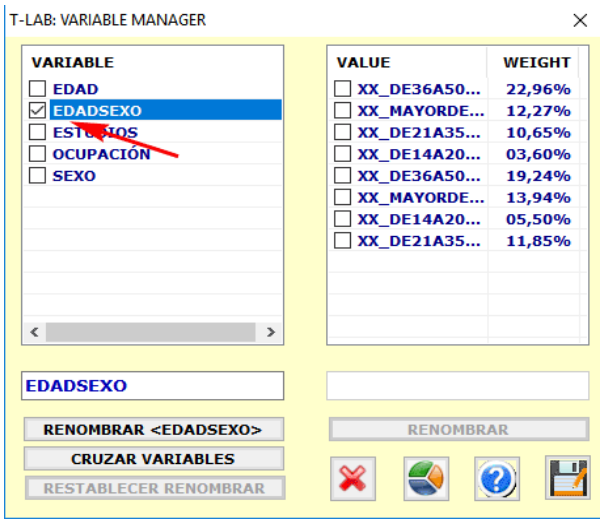


d) crear una **variable cruzada** disponible para otros análisis;

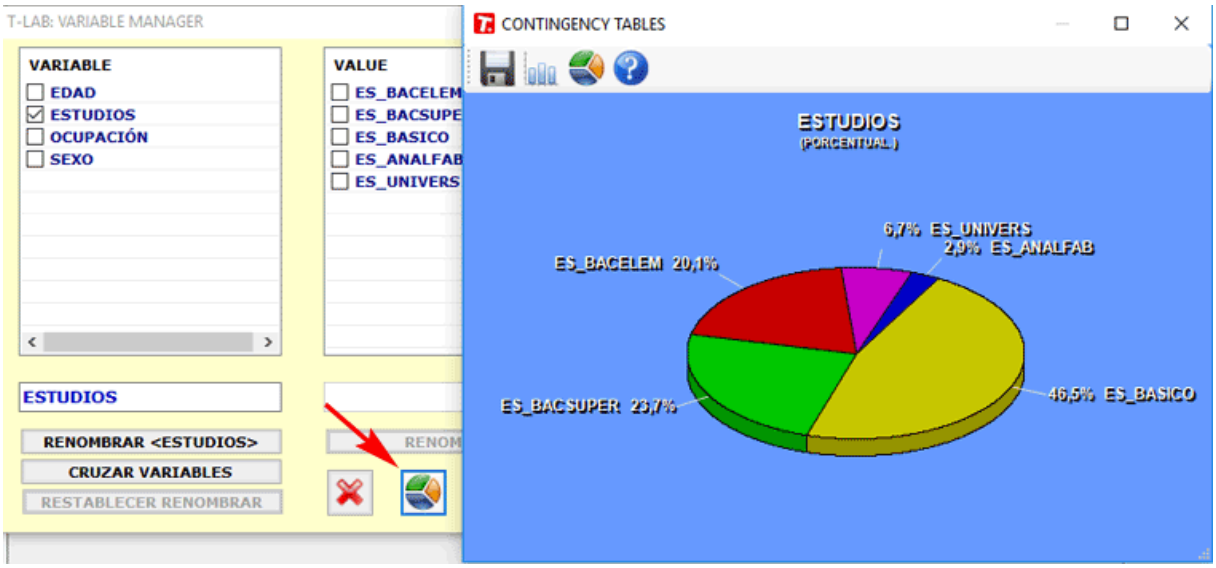


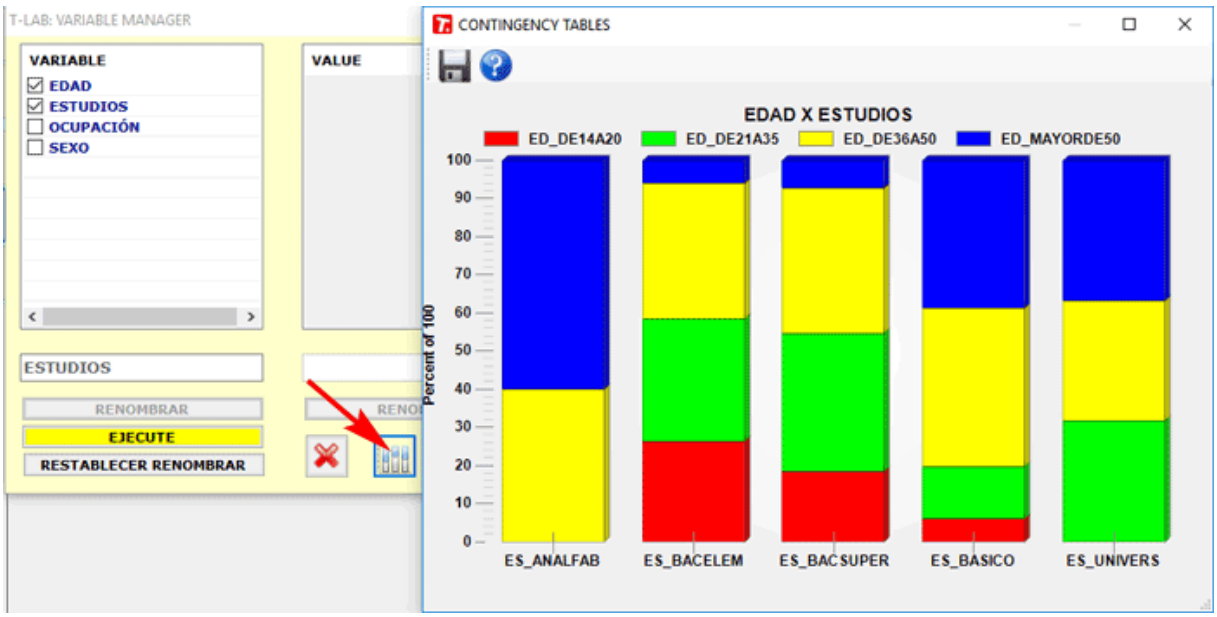






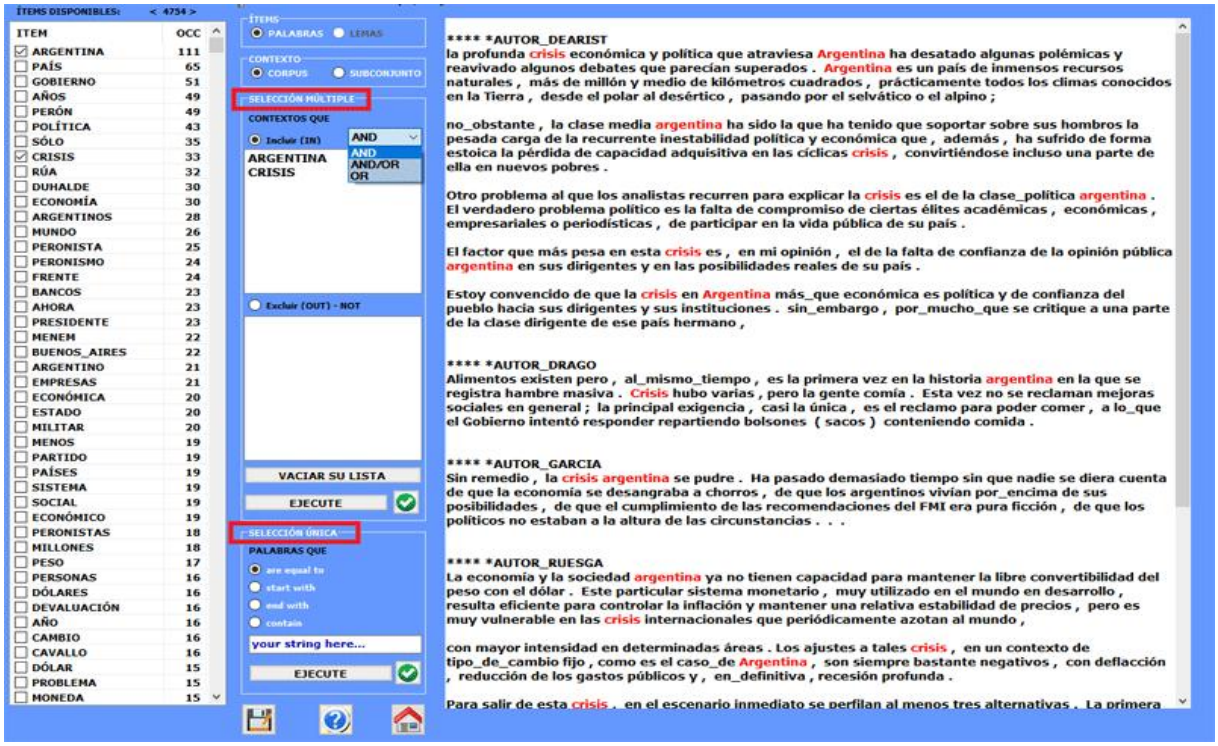
e) crear algunos **gráficos**.


T-LAB: VARIABLE MANAGER
VARIABLE
EDAD
ESTUDIOS
OCUPACIÓN
SEXO
VALUE
ES_BACELEM
ES_BACSUPE
ES_BASICO
ES_ANALFAB
ES_UNIVERS
ESTUDIOS
RENOMBRAR <ESTUDIOS>
CRUZAR VARIABLES
RESTABLECER RENOMBRAR
CONTINGENCY TABLES
ESTUDIOS
(PORCENTUAL)
6,7% ES_UNIVERS
2,9% ES_ANALFAB
ES_BACELEM 20,1%
ES_BACSUPER 23,7%
46,5% ES_BASICO


T-LAB: VARIABLE MANAGER
VARIABLE
EDAD
ESTUDIOS
OCUPACIÓN
SEXO
VALUE
ESTUDIOS
RENOMBRAR
EJECUTE
RESTABLECER RENOMBRAR
CONTINGENCY TABLES
EDAD X ESTUDIOS
ED_DE14A20
ED_DE21A35
ED_DE36A50
ED_MAYORDE50
Percent of 100
100
90
80
70
60
50
40
30
20
10
0
ES_ANALFAB
ES_BACELEM
ES_BACSUPER
ES_BASICO
ES_UNIVERS

Búsqueda avanzada en el Corpus

Esta herramienta de **T-LAB** permite extraer y exportar todos los segmentos del texto (esto es, frases o párrafos) que se corresponden con query de palabras sencillas o múltiples. Todo ello, tanto para el corpus entero como para los subconjuntos que lo componen.



Su uso es extremadamente intuitivo: basta con seleccionar las opciones deseadas dentro de los cuadros correspondientes.

En el caso de la selección "múltiple", las palabras pueden ser seleccionadas/añadidas cliqueando los ítems correspondientes. Dichos ítems están incluidos en la tabla a la izquierda.
En el caso de la selección 'individual', hay que digitar el texto a buscar dentro del cuadro correspondiente.
Una vez escogida la opción "ejecute", en el cuadro a la derecha se mostrarán los resultados de la búsqueda. Además, será posible guardarlos en un archivo .rtf.
Dicho archivo, que incluye todas las codificaciones T-LAB, también puede ser importado y analizado como si fuera un sub-corpus.

N.B.: El uso de esta herramienta se permite su uso sólo si se trabaja con un corpus ya importado y si ha sido seleccionada una lista de palabras clave (ver Configuración de Análisis).

Clasificación de nuevos documentos

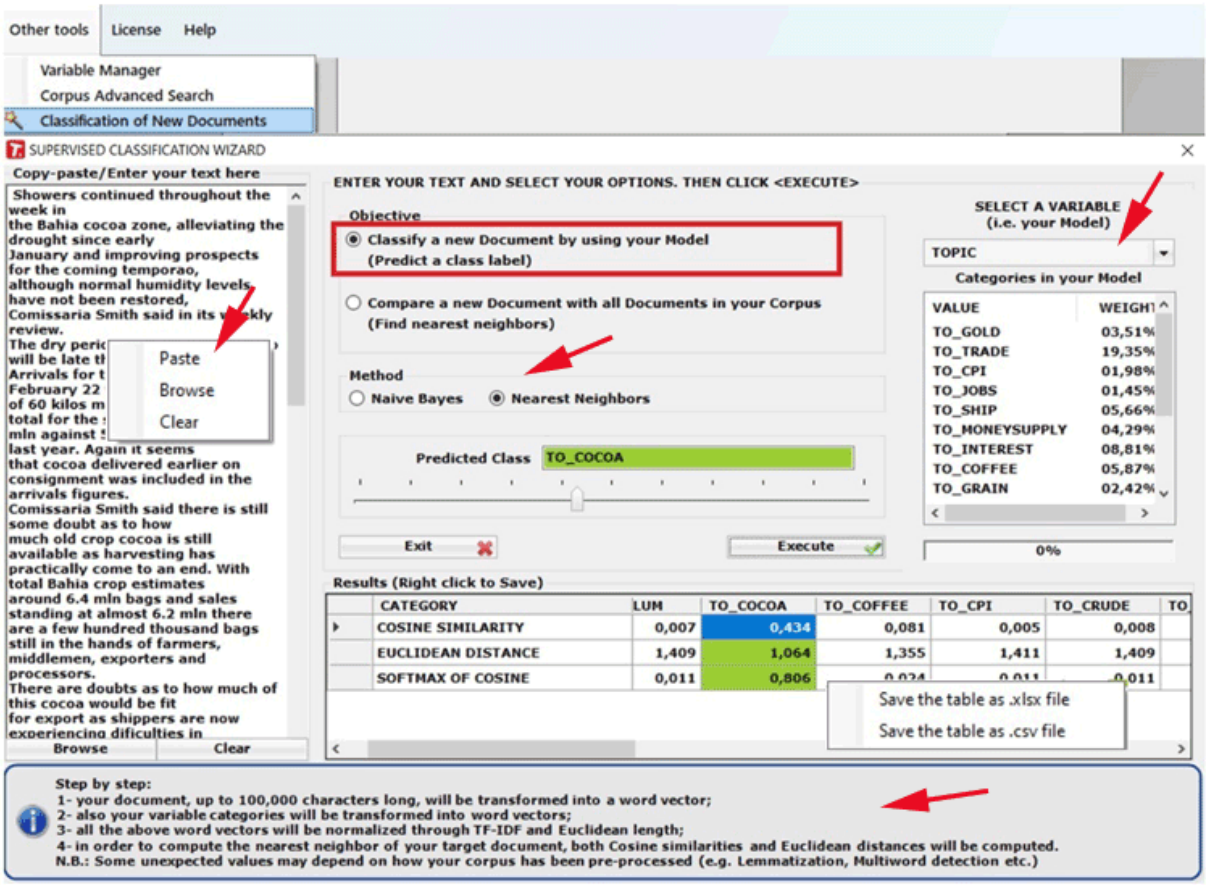
NOTA: Esta sección solo está disponible en inglés.

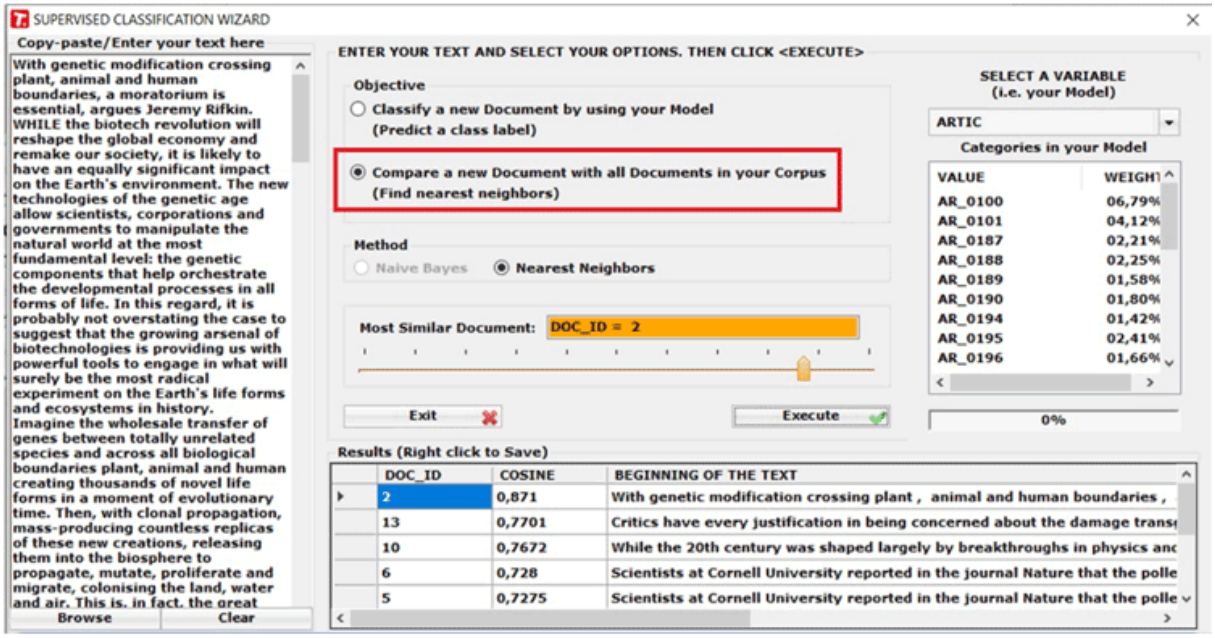
This tool, which is very easy to use, allows one to easily classify new documents according to a pre-existing model (i.e. any categorical variable) and also to compare any new document with all documents included in a corpus already analysed.

To this purpose, the following steps are required:

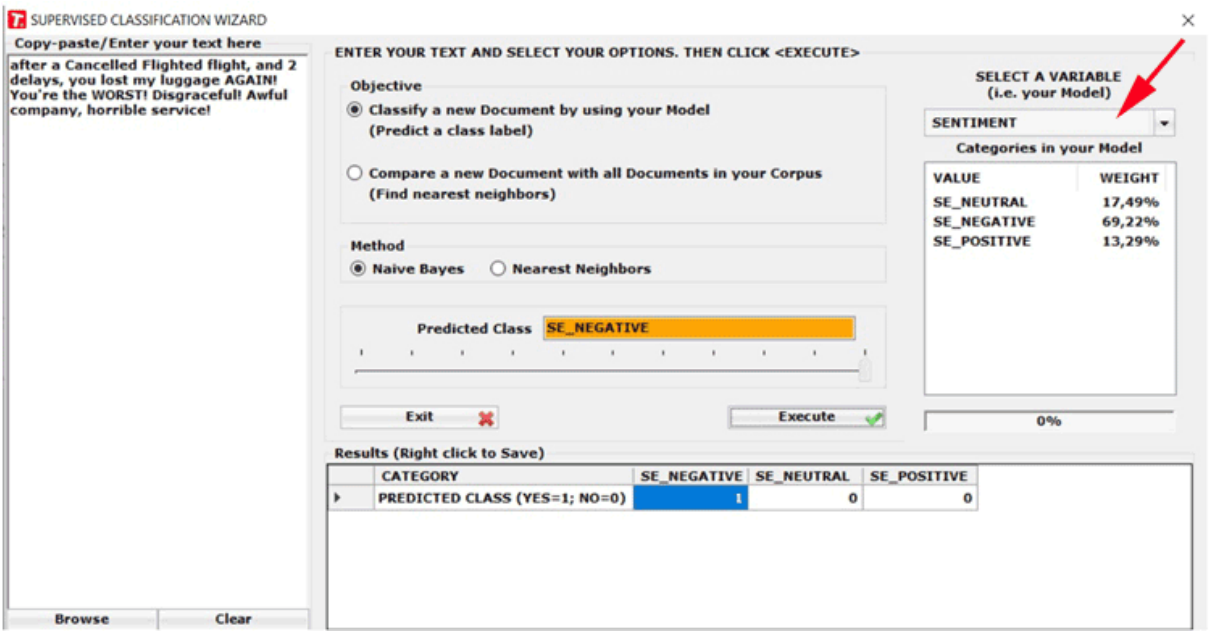
- enter a new document in the appropriate box;
- select a categorical variable to be used as a 'model';
- choose the desired 'objective' and a 'method';
- click 'execute'.

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).



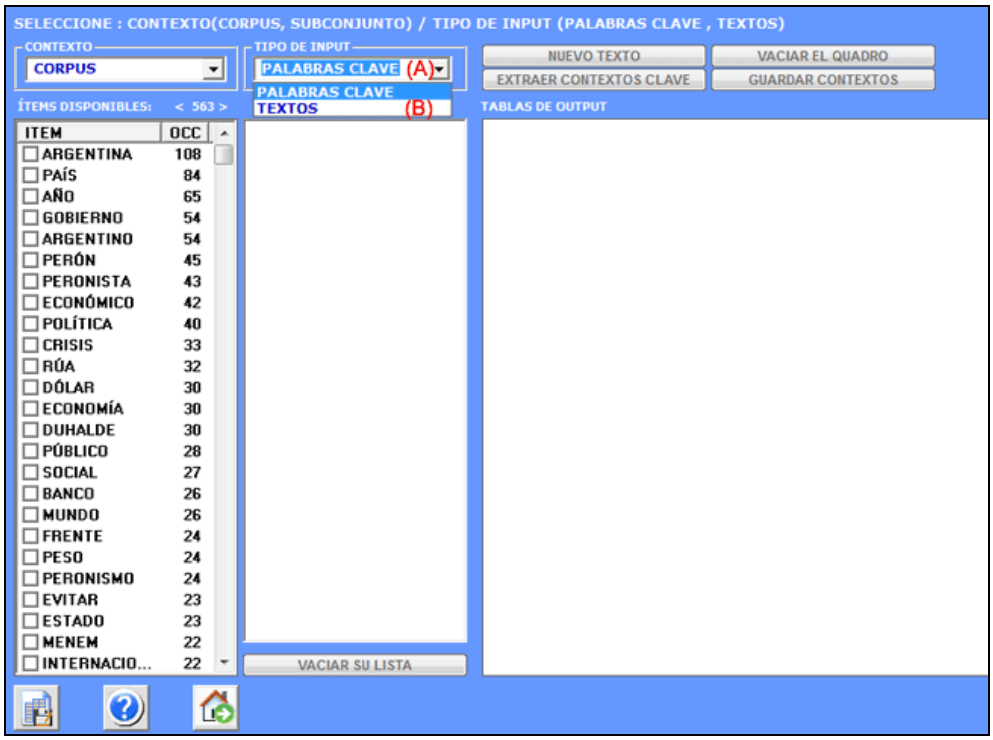
N.B.: When the user wishes to classify a dataset of new documents by using a supervised method, the dataset must be imported by T-LAB and then analysed by using a previously generated dictionary. To this purpose, the 'Thematic Document Classification' can be used, both for generating a dictionary of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

Contextos Clave de Palabras Temáticas

Esta herramienta de **T-LAB** puede ser utilizada para alcanzar dos objetivos diferentes:

a) extraer conjuntos de unidades de contexto que permiten profundizar el valor temático de **palabras-clave especificas**;

b) extraer las unidades de contexto que son más parecidas a los **textos de muestra propuestos por el usuario**.



Los procedimientos a seguir, paso a paso, son:

Caso (A)

Éste funciona de la siguiente manera:

1- el usuario elige (doble clic) una palabra temática "X" (véase abajo);

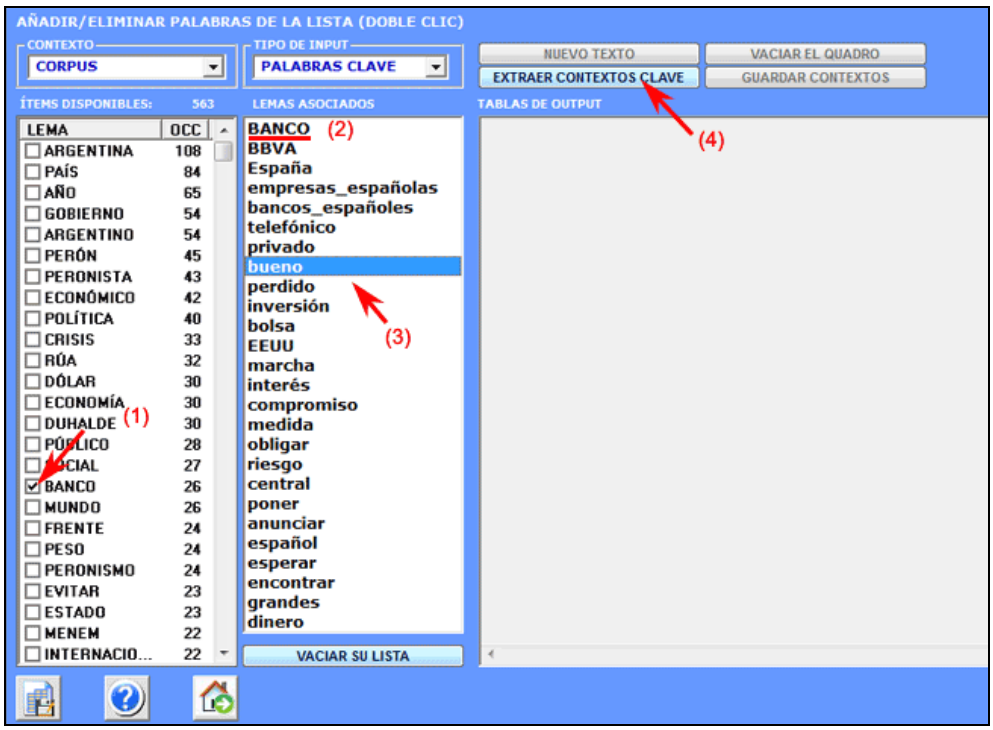
2- **T-LAB** propone una lista de palabras (máximo 50) los cuales valores de co-ocurrencia con "X" son más significativos;

3- el usuario puede quitar ítems irrelevantes de la lista proporcionada;

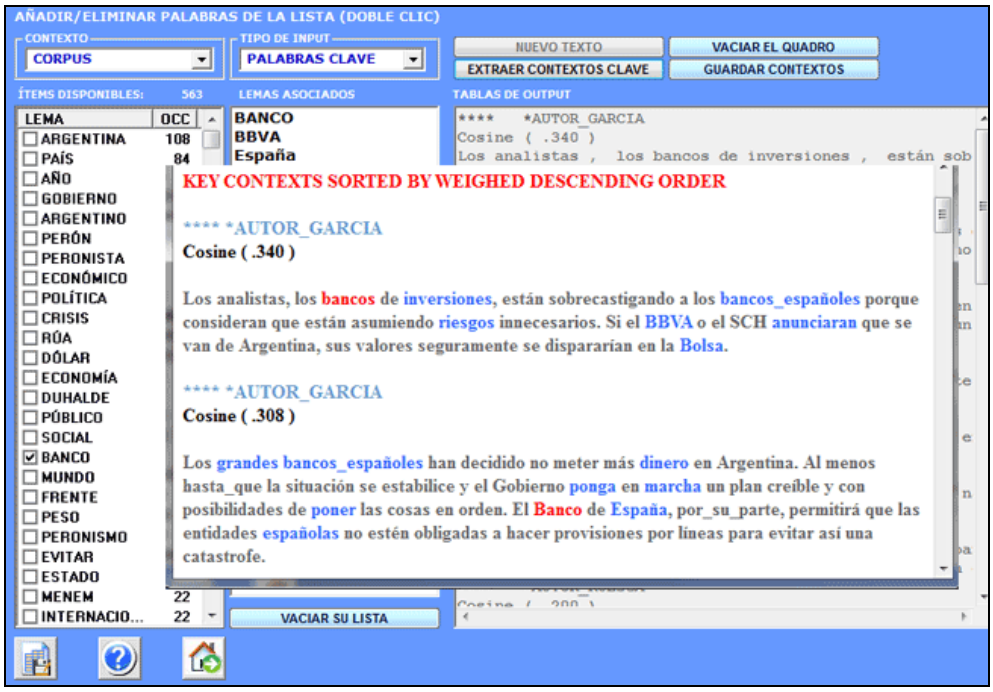
T-LAB asume que la lista del usuario es un vector de búsqueda (query vector) y computa sus **índices de asociación** (es decir los coeficientes del coseno) con todos los contextos elementales del corpus o del **subconjunto** seleccionado;

4- el output es un archivo **HTML** que contiene una lista de los contextos clave más significativos de "X" enumerados por orden descendente de sus índices de asociación (véase abajo);

Nota: Desemejante de la función **Concordancias**, que permite la extracción de todos los contextos elementales en los cuales las palabras clave seleccionadas son presentes (ocurrencias), y desemejante de la función **Asociaciones de Palabras**, que permite la extracción de todos los contextos elementales en los cuales las palabras clave seleccionadas se emparejan (co-ocurrencias), esta herramienta nos permite extraer los contextos elementales en los cuales cada palabra clave seleccionada se asocia a un conjunto de otras palabras (co-ocurrencias múltiples) que definen su campo temático.



Los resultados, tanto en formato HTML como en TXT, contienen un listado de los contextos clave más significativos de 'X' y están ordenados, de forma decreciente, en base a sus índices de asociación.



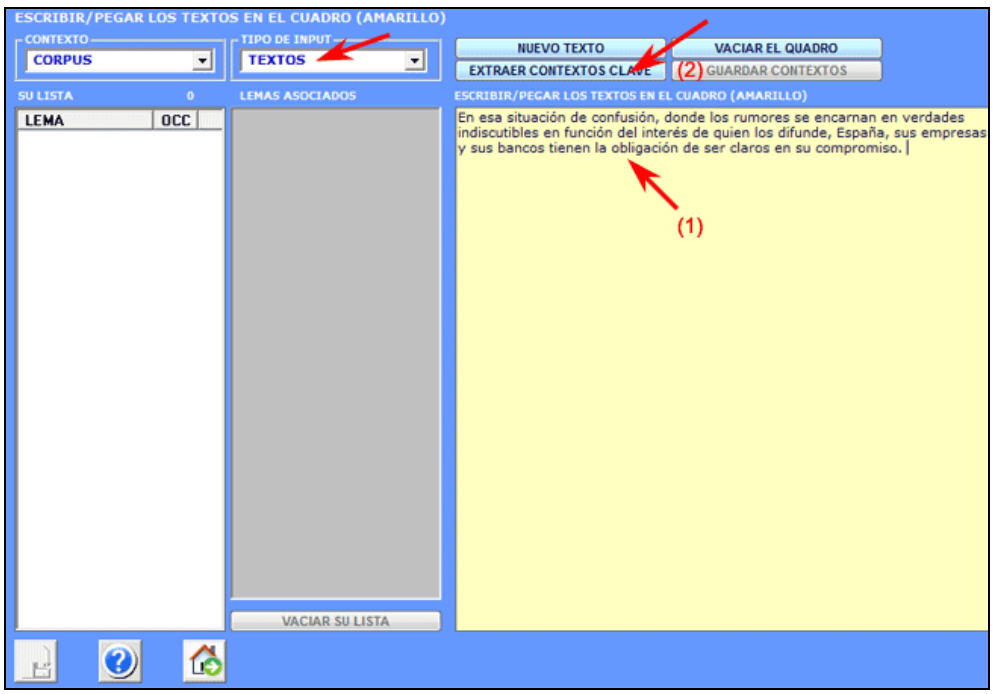
Los pasos 1-4 pueden ser reiterados para "n" palabras temáticas.

Caso (B)

Éste funciona de la siguiente manera:

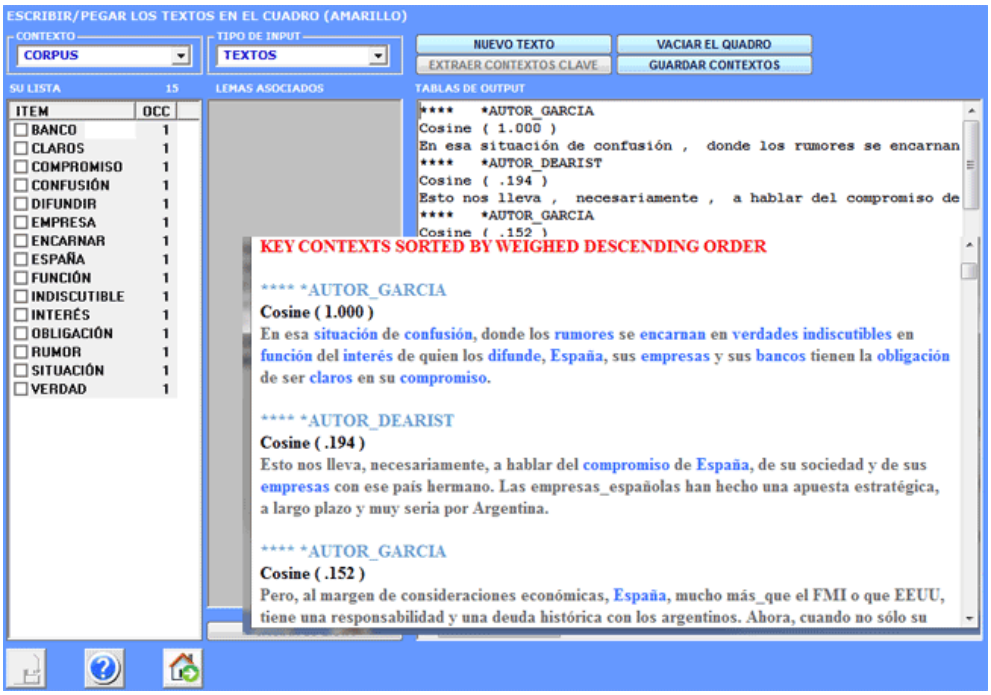
1 - El usuario copia/pega un 'modelo' de texto (máx. 5000 caracteres) en la casilla correspondiente;

2 - Después de haber seleccionado la opción 'extraer contextos clave', **T-LAB** transforma el texto introducido en un vector (query vector) y calcula los índices reativos de asociación (es decir, los coeficientes coseno) junto con todos los contextos elementales del corpus o del subconjunto seleccionado.



Los resultados, bien en formato HTML o bien en formato TXT, contienen un listado de los contextos clave que más semejanza tienen con el texto de input.

Nota: En este caso, la medida de semejanza no incluye a las palabras múltiples cuyas cadenas, con o sin el carácter de guion bajo ('_'), no correspondan al texto analizado.



Los pasos 1-2 pueden ser reiterados para "n" modelos de texto.

Exportar Tablas Personalizadas

NOTA: Las imagenes contenidas en este apartado hacen referencia a una versión anterior de T-LAB, ya que el interfaz de **T-LAB 10** cambia ligeramente. Sin embargo, las herramientas a disposición del usuario siguen siendo las mismas.

Esta opción permite crear, explorar y exportar tres tipos de tablas:

a) las con los valores de **ocurrencias** de las unidades lexicales dentro de los subconjuntos del corpus definidos por medio de alguna variable (matrices rectangulares);

b) las con los valores de las **co-ocurrencias** de las unidades lexicales (matrices cuadradas) dentro del corpus o dentro de los subconjuntos;

c) aquellas con las **ocurrencias** de las diferentes unidades lexicales presentes en todos los documentos (matrices dispersas con los índices de los diferentes elementos).

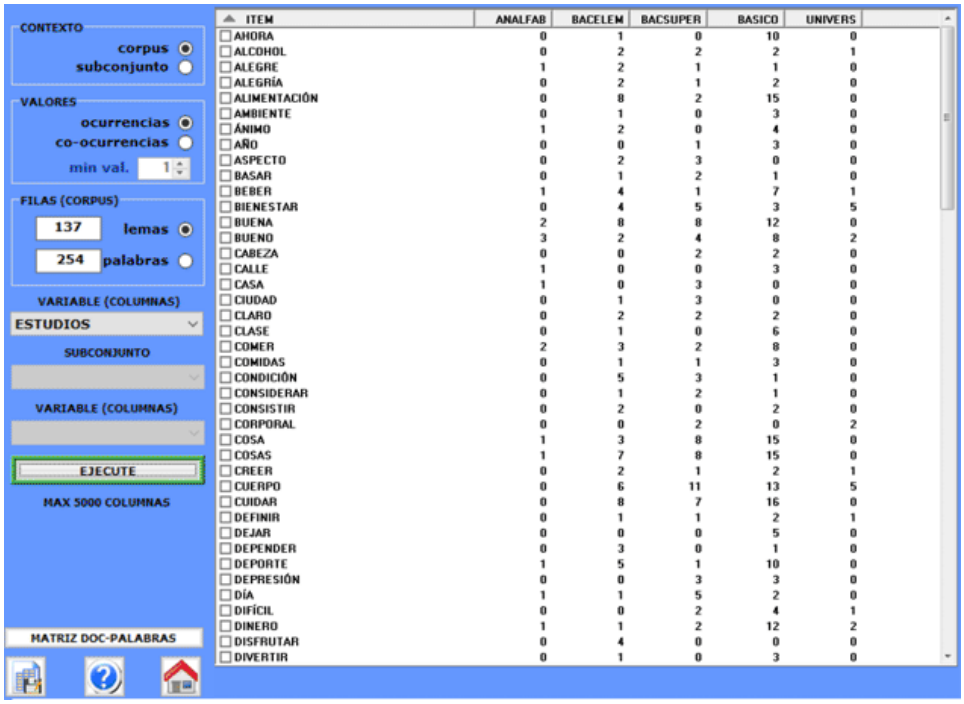
Los tamaños máximos de estas tablas son, respectivamente: a) 10.000 filas por 150 columnas, b) 5.000 filas por 5.000 columnas; c) 30.000 documentos por 10.000 unidades lexicales.

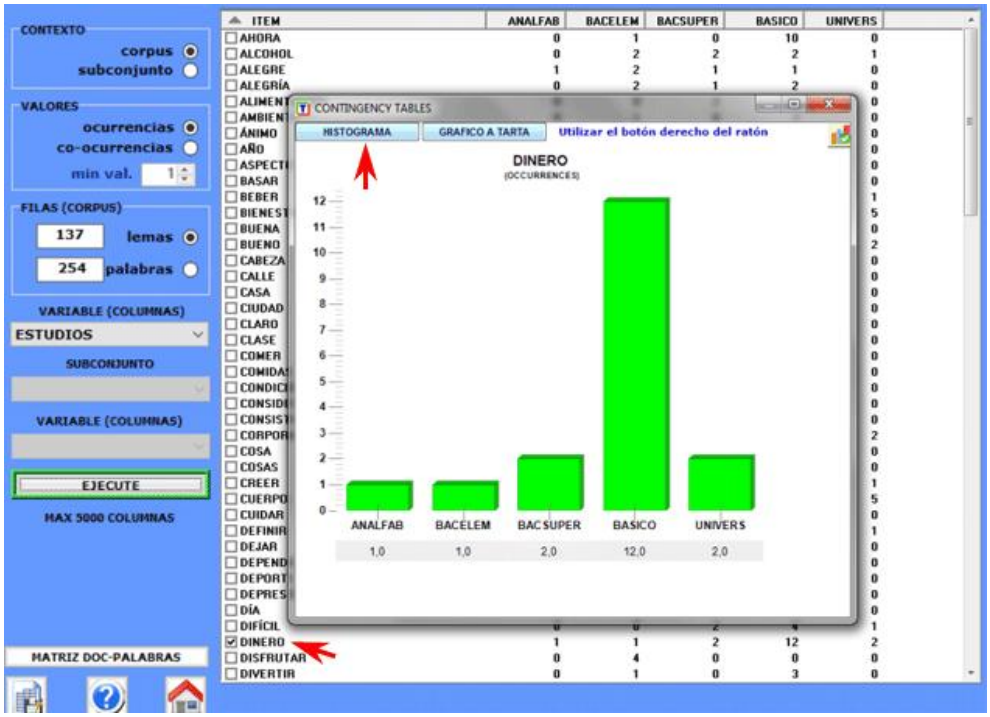
El uso de esta función es muy intuitivo.

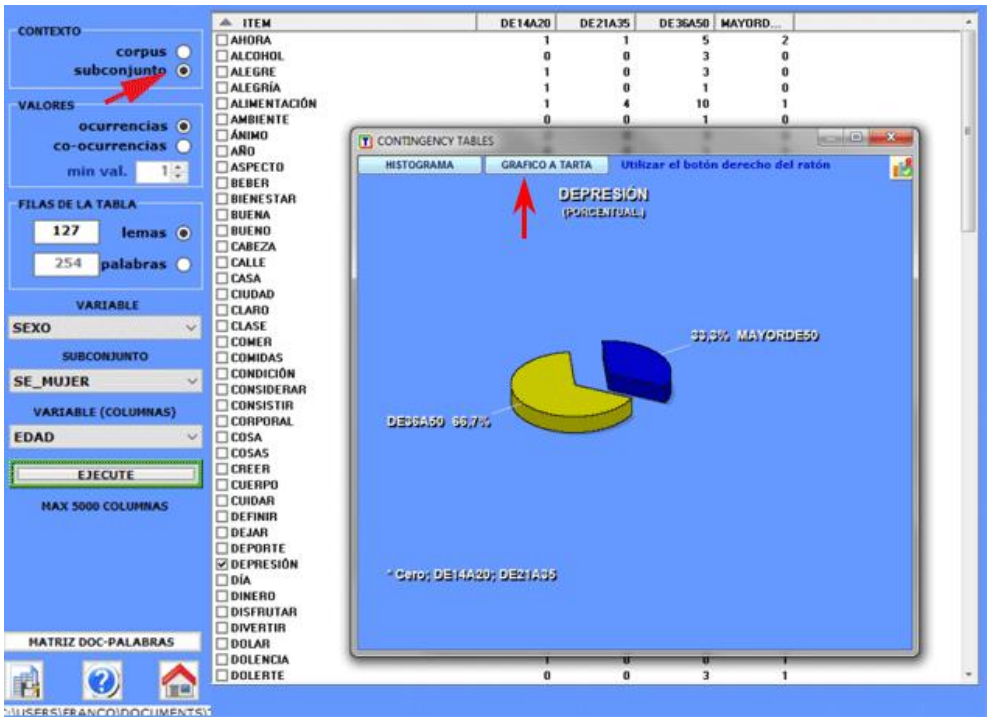
En los casos más simples, el usuario tiene que seleccionar una variable, cuya modalidades constituirán las columnas de la tabla.

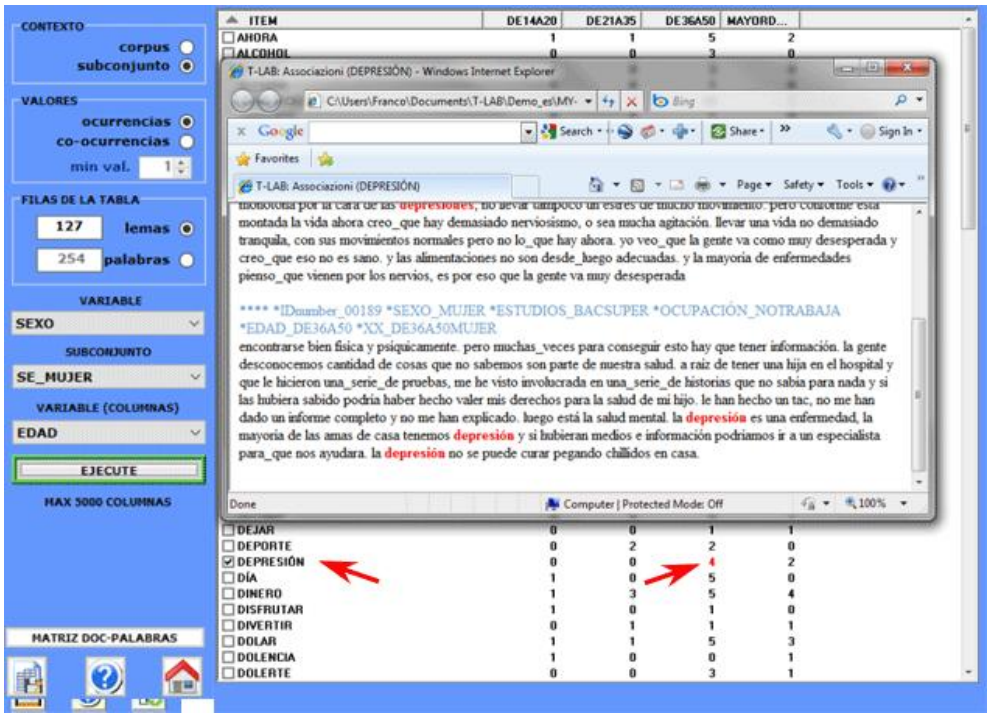
En los casos más complejos, se exige seleccionar una variable y un subconjunto.

Todas las tablas nos permiten crear varios tipos de gráficos.
Además, haciendo clic en específicas células de una tabla, es posible crear un archivo HTML que incluye todos los contextos elementales en que la palabra en la fila está presente en el subconjunto correspondiente (véase abajo).


CONTEXTO
corpus
subconjunto
VALORES
ocurrencias
co-ocurrencias
min val. 1
FILAS (CORPUS)
137 lemas
254 palabras
VARIABLE (COLUMNAS)
ESTUDIOS
SUBCONJUNTO
VARIABLE (COLUMNAS)
EJECUTE
MAX 5000 COLUMNAS
MATRIZ DOC-PALABRAS
ITEM ANALFAB BACELEM BACSUPER BASICO UNIVERS
AHORA 0 1 0 10 0
ALCOHOL 0 2 2 2 1
ALEGRE 1 2 1 1 0
ALEGRÍA 0 2 1 2 0
ALIMENTACIÓN 0 8 2 15 0
AMBIENTE 0 1 0 3 0
ÁNIMO 1 2 0 4 0
AÑO 0 0 1 3 0
ASPECTO 0 2 3 0 0
BASAR 0 1 2 1 0
BEBER 1 4 1 7 1
BIENESTAR 0 4 5 3 5
BUENA 2 8 8 12 0
BUENO 3 2 4 8 2
CABEZA 0 0 2 2 0
CALLE 1 0 0 3 0
CASA 1 0 3 0 0
CIUDAD 0 1 3 0 0
CLARO 0 2 2 2 0
CLASE 0 1 0 6 0
COMER 2 3 2 8 0
COMIDAS 0 1 1 3 0
CONDICIÓN 0 5 3 1 0
CONSIDERAR 0 1 2 1 0
CONSISTIR 0 2 0 2 0
CORPORAL 0 0 2 0 2
COSA 1 3 8 15 0
COSAS 1 7 8 15 0
CREER 0 2 1 2 1
CUERPO 0 6 11 13 5
CUIDAR 0 8 7 16 0
DEFINIR 0 1 1 2 1
DEJAR 0 0 0 5 0
DEPENDER 0 3 0 1 0
DEPORTE 1 5 1 10 0
DEPRESIÓN 0 0 3 3 0
DÍA 1 1 5 2 0
DIFÍCIL 0 0 2 4 1
DINERO 1 1 2 12 2
DISFRUTAR 0 4 0 0 0
DIVERTIR 0 1 0 3 0


CONTINGENCY TABLES
HISTOGRAMA
GRAFICO A TARTA
Utilizar el botón derecho del ratón
DINERO
(OCCURRENCES)
ANALFAB BACELEM BACSUPER BASICO UNIVERS
1,0 1,0 2,0 12,0 2,0


CONTEXTO
corpus
subconjunto
VALORES
ocurrencias
co-ocurrencias
min val. 1
FILAS DE LA TABLA
127 lemas
254 palabras
VARIABLE
SEXO
SUBCONJUNTO
SE_MUJER
VARIABLE (COLUMNAS)
EDAD
EJECUTE
MAX 5000 COLUMNAS
MATRIZ DOC-PALABRAS
ITEM
DE14A20
DE21A35
DE36A50
MAYORD...
CONTINGENCY TABLES
HISTOGRAMA
GRAFICO A TARTA
Utilizar el botón derecho del ratón
DEPRESIÓN
(PORCENTUAL)
33,3% MAYORDE50
DE36A50 66,7%
* Cero: DE14A20; DE21A35


CONTEXTO
corpus
subconjunto
VALORES
ocurrencias
co-ocurrencias
min val. 1
FILAS DE LA TABLA
127 lemas
254 palabras
VARIABLE
SEXO
SUBCONJUNTO
SE_MUJER
VARIABLE (COLUMNAS)
EDAD
EJECUTE
MAX 5000 COLUMNAS
MATRIZ DOC-PALABRAS
T-LAB: Associazioni (DEPRESIÓN) - Windows Internet Explorer
*ID_number_00189 *SEXO_MUJER *ESTUDIOS_BACSUPER *OCUPACIÓN_NOTRABAJA
*EDAD_DE36A50 *XX_DE36A50MUJER
encontrarse bien física y psiquicamente. pero muchas_veces para conseguir esto hay que tener información. la gente desconocemos cantidad de cosas que no sabemos son parte de nuestra salud. a raiz de tener una hija en el hospital y que le hicieron una_serie_de pruebas, me he visto involucrada en una_serie_de historias que no sabia para nada y si las hubiera sabido podria haber hecho valer mis derechos para la salud de mi hijo. le han hecho un tac, no me han dado un informe completo y no me han explicado. luego está la salud mental. la depresión es una enfermedad, la mayoria de las amas de casa tenemos depresión y si hubieran medios e información podriamos ir a un especialista para_que nos ayudara. la depresión no se puede curar pegando chillidos en casa.
ITEM
DE14A20
DE21A35
DE36A50
MAYORD...
DEPRESIÓN

Para exportar matrices dispersas de la tipología de documentos por palabras es suficiente cliquear en el botón correspondiente ('Matriz Documentos-Palabras').

En este caso, hay dos tipologías de output:

El primero (Sparse_Matrix.csv) tiene el siguiente formato:

Doc_Index; Word_Index; Word_Occ

00001; 1; 12

00001; 2; 5

…..

El segundo (Word_Indexes.csv) tiene el siguiente formato:

Word_Index; Word_Label

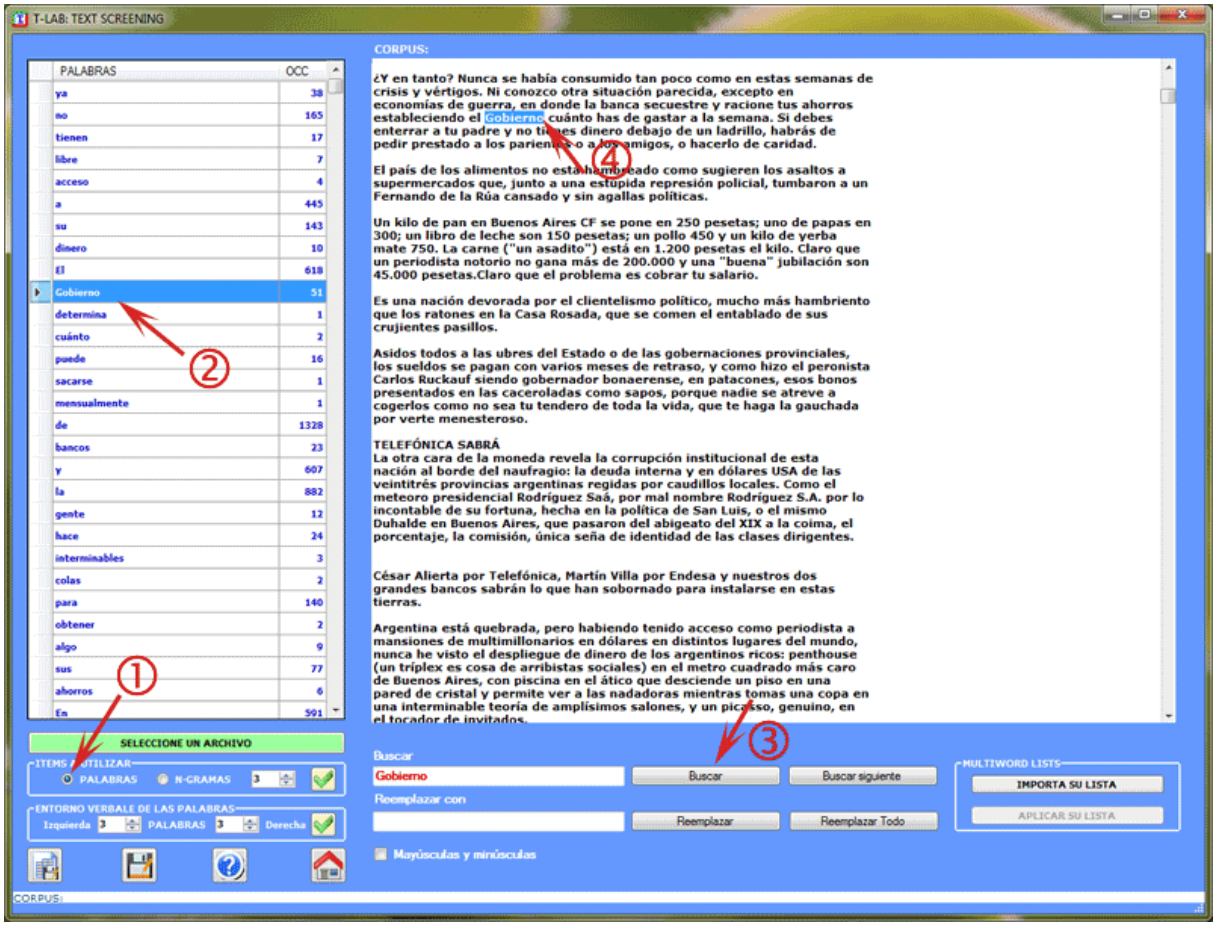
1; abrir

2; acabar

….

Editor

En **T-LAB 10** las herramientas para la edición del archivo en formato texto están incluidas en la herramienta **Text Screening** (véase abajo).



Importar-Exportar una Lista de Identificadores

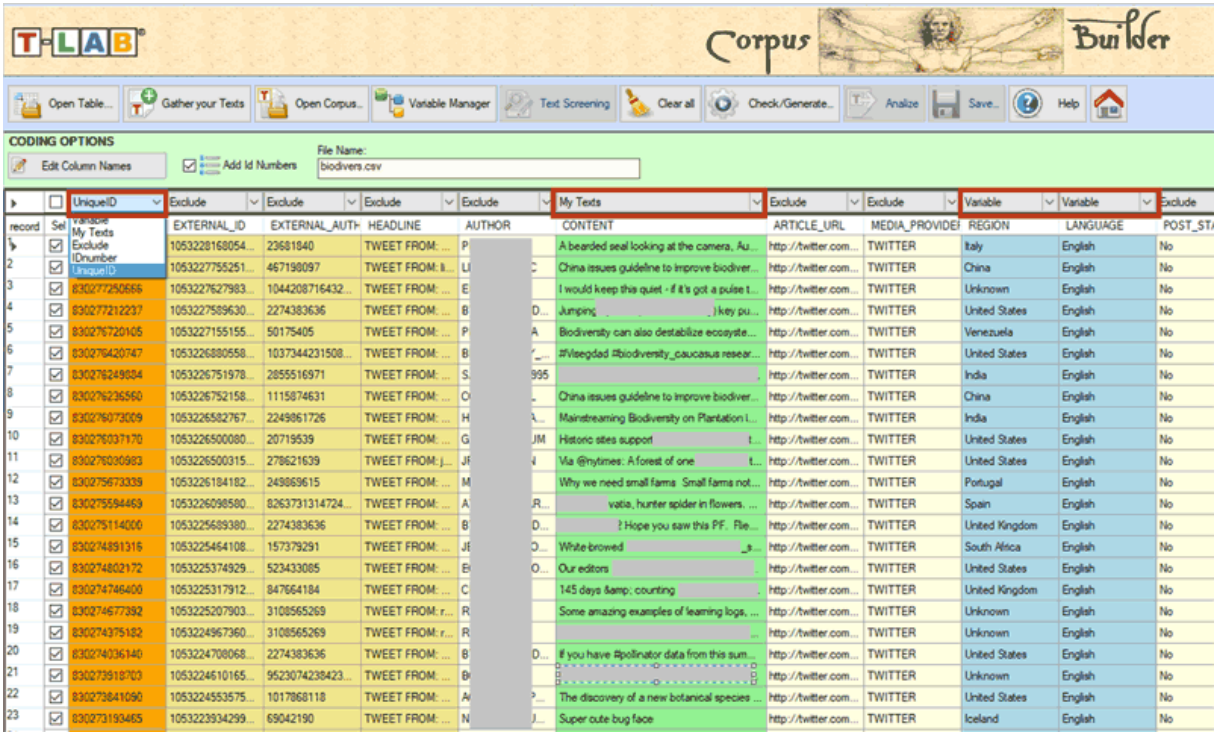
En **T-LAB**, un identificador único ('Unique Identifier') es una variable categórica con un valor distinto para cada documento (o caso).

Una lista de identificadores únicos puede consistir en cualquier tipo de cadenas alfanuméricas (por ejemplo, números de identificación, nombres propios, nombres geográficos, nombres de libros, etc.) de hasta 50 caracteres y sin espacios en blanco.

Como los identificadores únicos son singulares, es imposible realizar cualquier análisis de datos en ellos. En cambio, se utilizan para identificar resultados en las salidas de software.

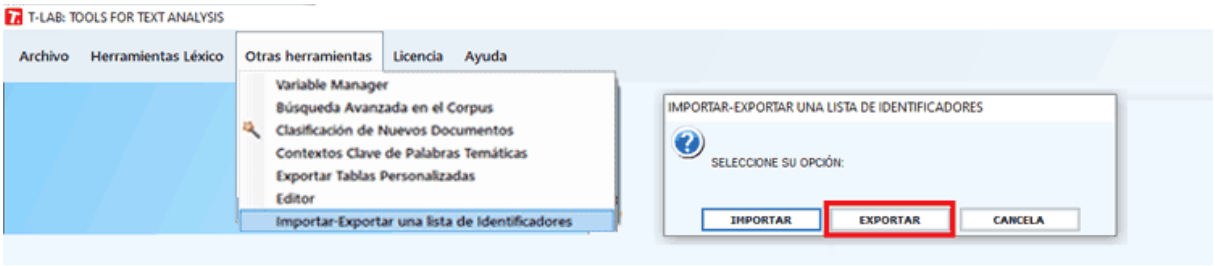
En **T-LAB**, a través de las opciones de importación / exportación, cualquier lista de identificadores únicos se puede modificar en cualquier momento.

Al importar datos en formato tabular, los identificadores únicos deben estar en la primera columna, como en el siguiente ejemplo con mensajes de Twitter.



En los otros casos (es decir, colecciones de documentos que no están en formato tabular) el procedimiento recomendado es el siguiente:

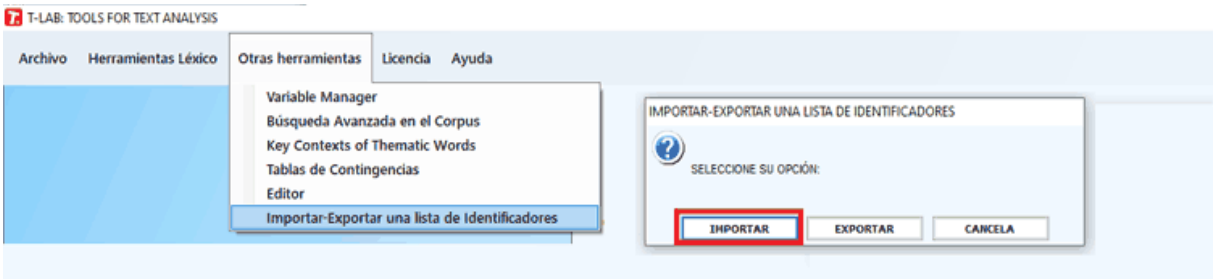
1- Importar un corpus;
2- Exportar la lista de identificadores creada automáticamente por **T-LAB**.



3- Editar y modificuar el archivo CSV creado por **T-LAB** (es decir, modificar los valores de "MyIdentifier" según sus necesidades. Véase la imagen a continuación).

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

4- Importar el archivo CSV que incluye los identificadores únicos revisados.



GLOSARIO

Análisis de Correspondencias

Técnica estadística de análisis factorial aplicada al estudio de **tablas de datos** cuyas celdas contienen valores de frecuencia (números reales positivos) o valores de tipo presencia-ausencia ("1" o "0").

Como todos los métodos de análisis factorial, el análisis de correspondencias permite la extracción de nuevas variables - los **factores** - que resumen de una manera organizada la información significativa contenida en los innumerables datos de las tablas; además, esta técnica de análisis permite la creación de gráficos que muestran - en uno o más espacios - los puntos que identifican los **objetos** en filas y/o en columnas, que - en nuestro caso - son las entidades lingüísticas (palabras, lemas, segmentos de textos y textos) con sus respectivas características de proveniencia.

En términos geométricos, cada factor organiza una dimensión espacial que puede ser representada como una línea o como un eje - en cuyo centro (o baricentro) está el valor "0 ", y que se desarrolla de una manera bipolar hacia los extremos negativos (-) y positivos (+), de modo que los objetos situados en polos opuestos sean los más diferentes, casi como la "izquierda" y la "derecha" en el eje de la política.

En **T-LAB** los resultados del análisis se resumen a través de gráficos bidimensionales (tipo planos cartesianos) que permiten evaluar las relaciones de proximidad/distancia - o sea de semejanza/diferencia - entre los objetos considerados.

Además, en **T-LAB** se proporcionan las medidas - en concreto **Contribuciones Absolutas** y **Valores Test** - que facilitan la interpretación de los **polos factoriales** que organizan las diferencias/semejanzas entre los objetos considerados.

Cadenas de Markov

Una cadena markoviana está constituida por una **sucesión** (o secuencia) de eventos, generalmente indicados como **estados**, caracterizada por dos propiedades:

- el conjunto de los eventos y de sus posibles resultados es finito;
- el resultado de cada evento depende sólo (o al máximo) del evento inmediatamente anterior.

Con la consecuencia de que a cada transición de un evento a otro le corresponde un valor de probabilidad.

En el ámbito científico, el modelo de las cadenas markovianas se utiliza para analizar las sucesiones de eventos económicos, biológicos, físicos, etc. En el ámbito de los estudios lingüísticos sus aplicaciones tienen como objeto las posibles combinaciones de las varias unidades de análisis en el eje de las relaciones sintagmáticas (una unidad tras otra).

En **T-LAB** el análisis de las cadenas markovianas concierne dos tipos de **secuencias**:

- las relativas a las relaciones entre unidades lexicales (palabras, lemas o categorías) presentes en el corpus en análisis;
- las presentes en archivos externos predispuestos por el usuario.

En ambos casos, en primer lugar se crean tablas cuadradas en las que se representan las ocurrencias de las transiciones, o sea cantidades que indican el número de veces en las que una unidad de análisis precede (o sigue) a la otra. Sucesivamente, las ocurrencias de las transiciones se transforman en valores de probabilidad (ver imágenes siguientes).

	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

Para más información véase el **Análisis de Secuencias**

Chi-cuadrado

Es un **test estadístico** para comprobar si los valores de frecuencia obtenidos por un examen, y registrados en una tabla cualquiera de doble entrada, son significativamente diferentes a los teóricos.

T-LAB aplica este tipo de test a tablas (2 x 2); por lo tanto el valor de umbral es 3.84 (df = 1; p. 0.05) o 6.64 (df = 1; p. 0.01).

Por ejemplo, para verificar la significación de las ocurrencias de una palabra ("x") dentro de una unidad de contexto ("A") el test se aplica a una tabla como sigue:

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

La fórmula de chi cuadrado, en su versión simplificada, es la siguiente :

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

Donde "O" y "E" representan respectivamente las frecuencias observadas y las teóricas.

Para cada célula, las ocurrencias teóricas (E) se calculan como sigue: (Ni x Nj)/Nij.

En consecuencia en el ejemplo considerado el valor del CHI cuadrado es igual a 19.38.

Puesto que es mayor que el valor crítico, la hipótesis nula (ausencia de la diferencia significativa) puede ser rechazada.

Cluster Analysis

Conjunto de técnicas estadísticas cuyo objetivo es individuar **grupos de objetos** que tengan dos características complementarias:

A - Máxima homogeneidad interna (dentro de cada cluster);

B - Máxima heterogeneidad externa (entre cluster y cluster).

En el lenguaje de la estadística, las características “A” y “B” corresponden respectivamente a la varianza interna (within cluster variance) y a la externa (between cluster variance).

En general, hay dos tipos de técnicas de Cluster Análisis:

- **Hierarchical methods** (métodos jerárquicos), cuyos algoritmos reconstruyen la jerarquía completa de los objetos analizados (el llamado "árbol"), tanto en orden ascendente como en orden descendente;

- **Partitioning methods** (métodos divisorios), cuyos algoritmos preveen que el usuario haya definido previamente el número de grupos en los cuales se dividen los objetos analizados.

En **T-LAB** se utilizan algoritmos de ambos tipos.

En particular:

- la función **Análisis de Co-Palabras y Mapas Conceptuales** utiliza un método jerárquico;
- la función **Cluster Analysis** permite utilizar tres métodos distintos: dos jerárquico y uno en particiones;
- la funciones **Análisis Temático de Contextos Elementales** y **Clasificación Temática de Documentos** utilizan un algoritmo del tipo bisecting K-means.

Algunas de las publicaciones citadas en la **Bibliografía** permiten profundizar tanto los aspectos generales de los varios métodos (Bolasco S., 1999; Lebart L., A. Morineau, M. Piron, 1995), como los aspectos específicos relativos a el Hdbscan (Campello R. J. G. B., Moulavi D., Zimek A. & Sander J. , 2015) y el método bisecting K-means (Steinbach, M., G. Karypis, V. Kumar, 2000; Savaresi S.M., D.L. Boley, 2001).

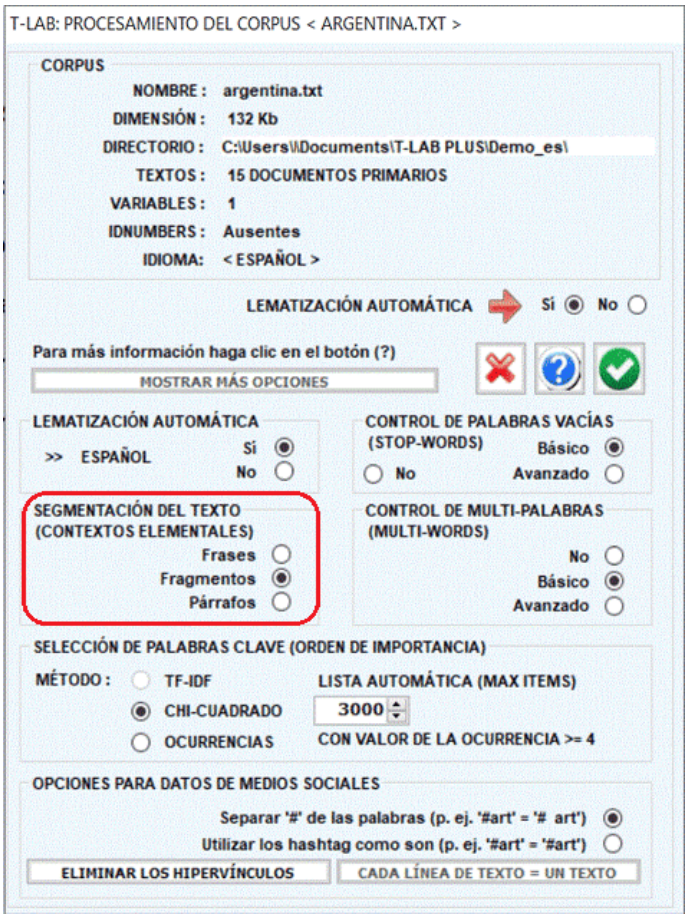
Codificación

Antes de la importación del corpus, el usuario puede introducir filas de codificación al inicio de cada **unidad de contexto** que desea clasificar por medio de una o más **variables**.

Normalmente, las unidades de contexto **clasificadas** corresponden a los **documentos primarios.**

Contextos Elementales

Durante la fase de la importación, **T-LAB** lleva a cabo una **segmentación** del corpus en **contextos elementales**, para facilitar las exploraciones del usuario y, sobre todo, para efectuar los análisis que requieren el cómputo de las **co-ocurrencias**.



Según la elección del usuario, los contextos elementales pueden ser:

1 - **Frases**

Contextos elementales que terminan con signos de puntuación (.?!) y que no superan longitud máxima de 1.000 caracteres.

2 - **Fragmentos**

Contextos elementales de longitud comparable y compuestos de uno o más enunciados.

En este caso, las reglas de segmentación usadas por T-LAB son las siguientes:

- considerar como contexto elemental cada secuencia de palabras interrumpida por el punto y a parte y cuyas dimensiones sean inferiores a la longitud de 400 caracteres;

- en el caso en el que, dentro de la longitud máxima, no haya ningún punto y a parte, buscar, en el orden, otros signos de puntuación (? ! ; : ,). Si no se encontraran, segmentar en base a un criterio estadístico, pero sin truncar las unidades lexicales.

3 - **Párrafos**

Contextos elementales que terminan con signos de puntuación (.?!) y retorno del carro (longitud máxima: 2.000 caracteres).

4 - **Textos Breves**

Esta opción se permite solamente cuando la longitud máxima de textos no supera los 2.000 caracteres (por ejemplo, las respuestas a preguntas abiertas).

NOTA:

- El fichero **corpus_segments.dat** contiene el resultado de la segmentación del corpus;
- En T-LAB, la opción **Concordancias** permite verificar los contextos elementales en los que está presente cada **palabra** (o **lema**).

Corpus y Subconjuntos

El **corpus** es una colección de uno o más textos seleccionados para el análisis.

Cada **subconjunto** del corpus se define por medio de una **modalidad** y de una **variable**.

T-LAB permite explorar y analizar las relaciones entre las unidades de análisis de todo el **corpus** o de sus **subconjuntos**.

Algunos ejemplos de **corpus**:

- un solo texto o documento que trate cualquier tema;
- un conjunto de artículos tomados de la prensa, referentes al mismo tema;
- una o varias entrevistas realizadas en el mismo proyecto de investigación;
- un conjunto de respuestas a una pregunta abierta de un cuestionario;
- una lista de direcciones sacada de internet;
- uno o varios libros del mismo autor que afronten temas similares;
- un conjunto de respuestas a una pregunta abierta de un cuestionario;
- transcripciones de focus groups.

NOTA: Algunos subconjuntos del corpus son los "**clusters temáticos**" de documentos o de contextos elementales obtenidos usando las herramientas correspondientes de **T-LAB**.

Algunos ejemplo de **subconjuntos**:

- unos o más capítulos de un libro
- unos o más artículos periodísticos publicados en el mismo año;
- unas o más entrevistas con la misma categoría de gente;
- un subconjunto de respuestas a una pregunta abierta.

En el caso de un corpus compuesto por varios textos, para hacer un **conjunto correctamente analizable**, se requiere que todas sus piezas tengan dos características que las hagan comparables:

a) una cierta homogeneidad temática y/o de contexto en el cual se han producido, para obtener datos comparables;

b) relaciones equilibradas entre sus dimensiones, tanto en términos de frecuencias como en términos de kilobytes, para no incurrir en anomalías estadísticas.

En la lógica de **T-LAB**, el corpus es una **base de datos** organizada en **registros** y **campos**. Más exactamente, los registros se componen de las entidades registradas (textos, segmentos de texto, palabras) y los campos se componen de las variables usadas para clasificar las diversas entidades (los autores del texto, los contextos de referencia, los tipos de temas, etc.).

Véase **Preparación del corpus**.

Desambiguación

Operación que intenta resolver casos de **ambigüedad** semántica, concretamente los atribuibles a los **homógrafos**, es decir, palabras con la misma **forma gráfica** pero con diversos significados.

NOTA: La herramienta **Text Screening**, propia de **T-LAB 10**, permite implementar funciones específicas para la desambiguación. Además, en la fase de importación, **T-LAB** reconoce y distingue entre ellos 3 tipos de objetos lingüísticos:

- nombres pro pios (de persona o lugar);
- **multipalabras** (palabras compuestas y modismos);
- los tiempos compuestos.

En los tres casos, **T-LAB** utiliza las listas de su base de datos, construidas y probadas para limitar los casos más frecuentes de ambigüedad (criterio de **eficacia**) y para moderar el tiempo de procesamiento (criterio de **eficiencia**).

Diccionario

Los diccionarios de **T-LAB** son tablas o archivos que contienen esquemas de clasificación para unidades léxicales (es decir, palabras).

Los esquemas de clasificación, y por lo tanto los diccionarios, pueden ser de dos tipos: (a) basados en **características lingüísticas** o (b) basados en **categorías temáticas**.

Ambos se pueden exportar y personalizar.

En el caso de "a" (es decir, cambiar el nombre o agrupar los elementos de la lista de palabras clave), el usuario puede consultar la herramienta **Personalización del Diccionario**.

En el caso de "b" (es decir, exportar / utilizar un diccionario para una clasificación supervisada), el usuario puede consultar cualquier herramienta **T-LAB** para el análisis temático (por ejemplo, **Clasificación basada en Diccionarios**, **Clasificación Temática de Documentos**, etc.).

Documentos Primarios

Los documentos primarios son textos (o partes del corpus) que corresponden a las unidades de contexto precedidas por una fila de **codificación**.

Según los casos, estos pueden ser: libros o capítulos de libros, artículos de periódicos, transcripciones de entrevistas, respuestas a preguntas abiertas etc.

Especificidad

En **T-LAB**, **Análisis de las Especificidades** es el nombre de una herramienta que permite verificar cuáles son las unidades lexicales (es decir, palabras, lemas o categorías) y los contextos elementales (es decir, frases o párrafos) que son típicos (o 'característicos') de un texto o de un subconjunto del corpus definido por una variable categorial.

Las unidades lexicales 'típicas', definidas por las proporciones de las respectivas ocurrencias (es decir, por su sobre/sub-utilización, exceso/carencia de uso), se eligen en base al cálculo del **Chi-cuadrado** o del **Valor-Test**.

Los contextos elementales 'típicos' se obtienen calculando y sumando los valores **TF-IDF normalizados** asignados a las palabras que componen cada frase o párrafo.

Graph Maker

La herramienta **Graph Maker** permite al usuario crear y exportar diferentes tipos de gráficos dinámicos en formato HTML.

Éstos pueden ser utilizados para alcanzar dos objetivos:

a) explorar las **relaciones de co-ocurrencia** entre palabras clave;

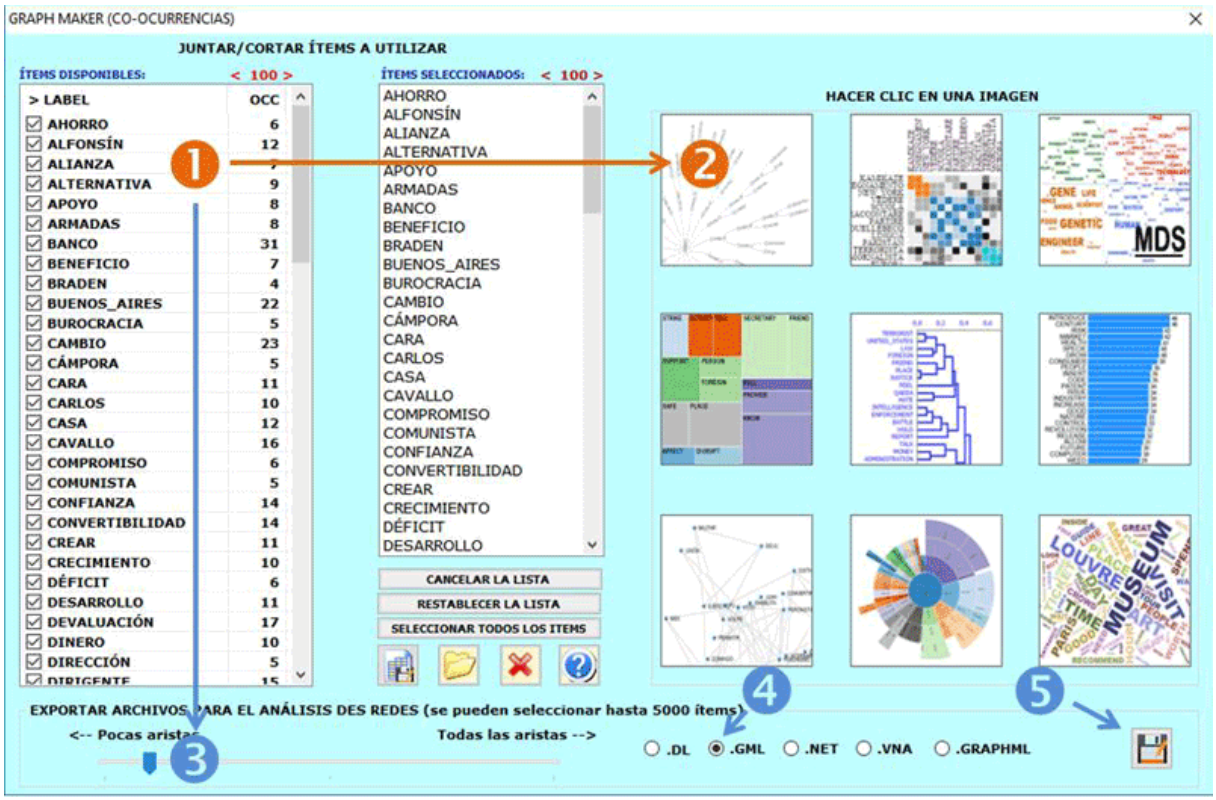
b) implementar algún tipo de **análisis de redes**.

En el caso (a) hay es necesario ejecutar dos operaciones (véase imagen siguiente):

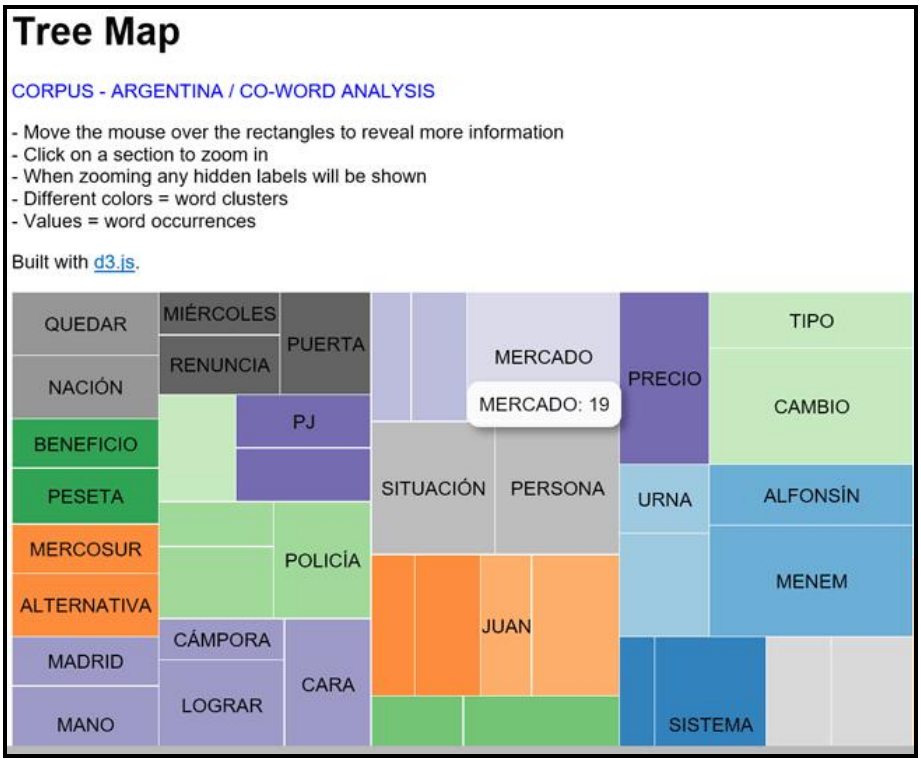
1-Seleccionar los ítems (esto es, las palabras clave) a usar

2-Cliquear sobre una imagen cualquiera para visualizar el grafico correspondiente.

En el caso (b), después de haber seleccionado las palabras clave (véase debajo "punto 1"), el usuario puede filtrar los enlaces a utilizar (véase debajo "punto 3"), escoger el formato del output (véase debajo "punto 4") y, finalmente, darle al botón "guardar" (véase debajo "punto 5").



NOTA: Cada output en formato HTLM incluye algunas instrucciones muy sencillas para facilitar su exploración (véase imagen siguiente).



Homógrafos

Dos o más palabras son **homógrafas** cuando tienen la misma forma gráfica (se escriben de la misma manera) pero tienen diversos significados.

En italiano, inglés y español hay millares de homógrafos. .

T-LAB utiliza un procedimiento de **desambiguación** que reduce su incidencia; en particular, la normalización de las **multipalabras** y la identificación de los verbos compuestos.

Así, por ejemplo - en la lengua italiana - la normalización de "il punto di vista" (modificada en "il_punto_di_vista") nos permite distinguir las ocurrencias específicas de "punto" y "vista" (dos típicos homógrafos). En la lengua inglesa, la normalización de la locución ("at_present") permite que distingamos "present" tanto como regalo como tiempo. Análogamente, en la lengua española, la normalización de "a_cada_rato" permite distinguir las diferentes acepciones de "rato".

IDnumber

IDnumber es una etiqueta que se puede introducir en las filas de codificación como elemento identificador de los sujetos (e.j. en el caso de respuestas a preguntas abiertas) o de las unidades de contexto en las que está subdividido el corpus a importar (ver **Preparación del corpus**).

En **T-LAB** la etiqueta "IDnumber", cada vez que se utiliza, tiene que estar seguida por un guión bajo ("_") y por un número progresivo de máx 5 cifras (ver ejemplo siguiente).

**** *IDnumber_00001 *EDAD_ADUL *SEXO_FEM *OCU_PROF

Sigue el texto de una respuesta o de un documento.
……

Cada corpus puede incluir numeraciones progresivas (IDnumber) de máx 30.000 sujetos o unidades de contexto.

NOTA:
El primer valor del Idnumber debe ser "1" (ej. IDnumber_00001).
En el caso en el que los textos recogidos por el usuario sean en formato MS Excel, en el paquete de instalación **T-LAB** se dispone de una macro que automáticamente los transforma en un corpus codificado y listo para su importación.

Índices de Asociación

En **T-LAB** los índices de asociación (o de similitud) se utilizan para analizar las **co-ocurrencias** de las **unidades lexicales** (**LU**, lexical units) en el interior de los **contextos elementales** (**EC**, elementary contexts), es decir datos binarios del tipo presencia/ausencia.

Por ejemplo, dadas dos **LU** y diez **EC**, se puede crear el siguiente ejemplo

	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

Los mismos datos se pueden representar del siguiente modo:

	LU_2		
LU_1	*Present*	*Absent*	*Total*
Present	*3*	*4*	*7*
Absent	*2*	*1*	*3*
Total	*5*	*5*	*10*

Generalizando y usando las letras del alfabeto:

	LU_2		
LU_1	*Present*	*Absent*	*Total*
Present	*a*	*b*	*a + b*
Absent	*c*	*d*	*c + d*
Total	*a + c*	*b + d*	*n*

Las fórmulas correspondientes a los seis índices de asociación usados por **T-LAB** son las siguientes:

Jaccard

$$\frac{a}{a+b+c}$$

Dice

$$\frac{2a}{2a+b+c}$$

Coseno

$$\frac{a}{\sqrt{(a+b)} \times \sqrt{(a+c)}}$$

Equivalencia

$$\frac{a^2}{(a+b) \times (a+c)}$$

Inclusión

$$\frac{a}{\text{Min}((a+b),(a+c))}$$

Información Mutua

$$\text{Log} \frac{a/N}{(a+b) \times (a+c)}$$

Suponiendo que se han obtenido los coeficientes de las relaciones entre diez LU, podemos crear una tabla como la siguiente:

	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 hecho, **T-LAB** crea y analiza tablas análogas de dimensiones N x N (en la que N puede corresponder a varios centenares de columnas), tanto con **Multidimensional Scaling** como con **Cluster Analysis**.

Las mismas tablas se utilizan también para calcular los índices de **semejanza de segundo orden** asociados a las parejas de palabras claves (véase herramienta de **Asociaciones de Palabras**).

Isotopía

Isotopía (iso = igual; topos = lugar) se refiere a un concepto de significado como "efecto del contexto", es decir, como algo que no pertenece a las palabras consideradas aisladamente, sino como resultado de sus relaciones en el interior de los textos o de los discursos.

La función de las isotopías es la de facilitar la interpretación de los discursos o de los textos; de hecho, cada una de ellas detecta un contexto de referencia común a varias palabras, que no derive de sus significados específicos. Esto en la lógica de que el conjunto es algo más que la adición de sus elementos.

La detección de una isotopía, por lo tanto, no es la mera observación de un "dato", sino el resultado de un proceso de interpretación (F. Rastier 1987).

La noción de isotopía fue propuesta inicialmente por el semiólogo A.J. Greimas (1966) para definir la repetición, dentro de las unidades sintagmáticas (frases e/o textos), de varias palabras con tratos semánticos comunes.

En la lógica de **T-LAB**, la identificación de las isotopías deriva del análisis de las ocurrencias y de las co-ocurrencias.

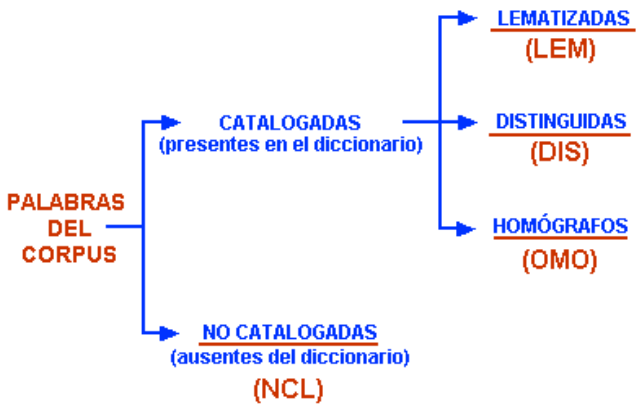
Lematización

En los diccionarios lingüísticos que consultamos, cada entrada corresponde a un lema que - generalmente - define un conjunto de palabras con la misma raíz lexical (el lexema) y que pertenece a la misma categoría gramatical (verbo, adjetivo, etc.).

En general, la **lematización** exige que las formas del verbo se pongan en infinitivo, los sustantivos en singular, etcétera.

Por ejemplo, las **formas flexivas** "hablan" y " hablando ", que resultan de la combinación de una **raíz** única con dos diversos sufijos (< - an > y < - ando >), se remiten al mismo lema "habl-ar". Sin embargo, hay algunos casos en los que la lematización no observa la regla de la raíz común; especialmente en los verbos irregulares.

Durante la fase de importación del corpus, **T-LAB** consiente hacer un tipo específico de lematización automática que sigue la lógica del árbol siguiente :



Obviamente, el diccionario de referencia es el que ha sido realizado en **T-LAB**.

Las abreviaturas de las cuatro-categorías se utilizan en muchas tablas, siempre en la columna "INF".

NOTA:
- la categoría "DIS " ("distinguir") significa que **T-LAB** no aplica la lematización estándar, para no anular las diferencias de significado entre las diversas palabras;
- a veces, para diferenciar homógrafos, **T-LAB** añade el carácter '_' (underscore) a su lema.

Lexia y Lexicalización

Según Pottier (ver **Bibliografía**), la **lexía** es una expresión constituida por una o más palabras que se comportan como una unidad lexical con significado autónomo.

Los tipos fundamentales son tres: *simple*, correspondiente a la palabra en el sentido común del término (ej. "*caballo*", "*comía*"); *compuesta*, constituida por dos o más palabras integradas en una única forma (ej. "*biotecnologías*", "*videoregistrador*"); *compleja*, constituida por una secuencia en vía de lexicalización (es. "*a mi juicio*", "*complejo industrial*").

La **lexicalización** es el proceso lingüístico a través del cual un sintagma o un grupo de palabras se convierten en una sola unidad lexical.

En **T-LAB** la función **Multi-palabras** permite crear una lista de las lexías complejas presentes en el corpus y de proceder a su transformación en cadenas unitarias (lexicalización).

MDS

Conjunto de técnicas estadísticas que permiten analizar matrices de semejanza para proporcionar una representación visual de las relaciones entre los datos dentro uno espacio de dimensiones reducidas.

En **T-LAB** un tipo de **MDS** (método de Sammon) se utiliza para representar las relaciones entre las unidades lexicales o entre los núcleos temáticos (véase **Análisis de Co-Palabras** y **Modelización de Temas Emergentes**).

Las tablas de input se constituyen de matrices cuadradas que contienen los valores de proximidad (desemejanzas) derivados a través del cálculo de un índice de asociación.

Los resultados obtenidos, como los del análisis de correspondencias, nos permiten interpretar las relaciones entre los "objetos" y las dimensiones que organizan el espacio en el cual se representan.

El grado de correspondencia entre las distancias, entre los puntos obtenidos por el mapa de MDS y aquellos de la matriz input, es medido (inverso) por una función de stress. Cuanto menor es el valor de lo stress (e.g. < 0,10), tanto mayor es la calidad del ajuste obtenido.

La fórmula de stress es la siguiente:

$$S = \sum_{i \neq j} \frac{(d_{ij}^{*} - d_{ij})^2}{d_{ij}^{*}}$$

Donde d_{ij}^{*} representa las distancias entre los puntos (ij) de la matriz input y d_{ij} representa las distancias entre los mimos puntos en el mapa MDS.

Multiwords (Multipalabras)

Un conjunto de dos o más formas gráficas que remiten a un **significado unitario**.

La categoría de las multipalabras, cuyos límites dependen del modelo analítico empleado, incluye subconjuntos como **nombres compuestos** (por ej. "transporte público" o "base imponible"), y las **locuciones** usadas como modismos (por ej. "en la medida en que", "con respecto a" ,o "en honor de la verdad").

La lista de las multipalabras realizada en **T-LAB** obviamente non es exhaustiva.
Ha sido construida y verificada con dos objetivos:

a) imitar los casos más frecuentes de ambigüedad (criterio de **eficacia**);

b) moderar el tiempo de ejecución del proceso de **normalización** (criterio de **eficiencia**).

En **T-LAB** es también posible utilizar una **lista personalizada de Multiwords**.

N-gramas

En **T-LAB** un **n-grama** es una secuencia de dos (bi-grama) o más palabras claves presentes dentro del mismo contexto elemental.

Su uso está vinculado exclusivamente al cálculo de las **co-ocurrencias**. Cabe destacar que, dentro del mismo contexto elemental, la contigüidad de las palabras consideradas no tiene en cuenta ni las 'palabras vacías' (es decir, stop-word) ni la puntuación.

Consideremos, a modo de ejemplo, el siguiente contexto elemental.

"La Nación española es libre e independiente, y no es ni puede ser patrimonio de ninguna familia ni persona"

En el supuesto de que los cinco ítems en rojo estén incluidos en nuestro listado de palabras claves, las segmentación en bi-gramas produciría los siguientes contextos de co-ocurrencia:

Nación & español
español & libre
libre & independiente
independiente & patrimonio
etc. etc.

Por otro lado, en el caso de tri-gramas, el resultado sería el siguiente:

Nación & español & libre
español & libre & independiente
libre & independiente & patrimonio
independiente & patrimonio & familia
etc. etc.

Es importante remarcar que, en el caso de los contextos elementales, las co-ocurrencias están basadas en la presencia de las palabras en el mismo 'lugar' (eje. frase, párrafo, etc.). Por otro lado, en el caso de n-gramas, las co-ocurrencias se fundamentan en una relación de contigüidad.

En **T-LAB** es posible implementar el análisis de co-ocurrencias basadas en n-gramas a través de la herramienta **Asociaciones de palabras**. Además, se puede implementar el análisis markoviano de bi-gramas utilizando la herramienta **Análisis de las Secuencias**.

Naïve Bayes Clasificador

La fórmula del clasificador Naive Bayes (NB) usada en **T-LAB** es la siguiente:

$$v_{\mathbf{NB}} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j)$$

En la que:

argmax - se refiere al máximo valor de la probabilidad a posteriori;
$v_j \in V$ - se refiere al j-cluster (v_j) de la partición (V);
$P(v_j)$ - se refiere a la probabilidad a priori de cada j-cluster;
$\prod_i P(a_i \mid v_j)$ - es el producto de las probabilidades de cada (a_i) palabra dendro cada (v_j) cluster.

Normalización

En **T-LAB**, la normalización del corpus tiene una meta doble:

a) permitir una detección correcta de las palabras como **formas gráficas**;

b) resolver previamente algunos casos de ambigüedad.

Esto significa que **T-LAB**, en primer lugar, realiza una serie de transformaciones del archivo que se está analizando: eliminación de los espacios vacios en exceso, adición del espacio después de signos de puntuación, reducción de las mayúsculas, etc.

En segundo lugar, **T-LAB** marca una serie de cadenas reconocidas como **nombres propios** (de persona y lugar); por tanto, convierte las secuencias de formas gráficas reconocidas como **multipalabras** en cadenas unitarias, para utilizarlas como tales durante el proceso de análisis ("en otras palabras" y "en tal caso" se transforman respectivamente en "en_otras_palabras" y "en_tal_caso").

Los parámetros de estas operaciones no pueden ser modificados por el usuario.

En la fase de normalización, para obtener un reconocimiento correcto de las formas gráficas, **T-LAB** utiliza la siguiente lista de **separadores**:

, ; : . ¡! ¿? ' " ( ) < > + / = [ ] { }

Núcleos Temáticos

T-LAB utiliza la locución **núcleos temáticos** en algunas rutinas que producen mapas de las **palabras clave**.

Esta indica los pequeños clusters de palabras, **co-ocurrentes** en los contextos elementales del corpus, que - en los mapas - se representan con etiquetas que pueden ser definidas y cambiadas por el usuario.

Ocurrencias y Co-ocurrencias

En análisis de textos, estos dos conceptos son de importancia fundamental.

Las **ocurrencias**, en efecto, son las cantidades que resultan del cómputo de cuántas veces (frecuencias) cada unidad lexical (**LU**, lexical units) se repite dentro del **corpus** o dentro las unidades de contexto (**CU**, context units) que lo constituyen.

Su distribución se puede representar en tablas de contingencia como la siguiente

	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

Las **co-ocurrencias** son las cantidades que resultan del cómputo del número de veces que dos o más unidades lexicales están presentes contemporáneamente en los mismos contextos elementales (**EC**, elementary context).

Su distribución se puede representar en tablas como la siguiente

(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

Con una simple transformación, las tablas del tipo "A" (rectangular) pueden transformarse en tablas del tipo "B" (cuadradas y simétricas) en las que para cada pareja de unidad lexical está indicada la cantidad de sus co-ocurrencias, es decir el total de contextos elementales en los que están contemporáneamente presentes.

(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	...	

En gran medida - en **T-LAB** - el análisis de textos se realiza mediante el estudio de las relaciones entre ocurrencias y entre co-ocurrencias, tanto con **índices de asociación** específicos, o con el uso de técnicas estadísticas multidimensionales como el **cluster análisis** y el **análisis de correspondencias**.

Palabras clave

En el interior de la lógica **T-LAB** son Palabras Clave todas las **unidades lexicales** (palabras, lemas, lexías, categorías) que, cada vez, se incluyen en las tablas a analizar.

Operativamente, la selección de las palabras clave se puede efectuar según dos modalidades: **automática** y **personalizada**.
NOTA: Solamente la segunda modalidad permite modificar las listas de unidades lexicales y de utilizar **diccionarios personalizados**.

Palabras y Lemas

Los software para el análisis de textos, en primer lugar, identifican las llamadas **formas gráficas**, es decir las cadenas de letras separadas por los espacios en blanco. Después, de acuerdo con sus algoritmos específicos o con las categorías usadas por los especialistas, el software reconoce **lemas**, **lexemas**, **palabras clave**, etc.

Las tablas **T-LAB**, para todas las unidades lexicales presentes en la base de datos del corpus, reproducen dos informaciones:

- la primera, denominada "**palabra**", contiene la trascripción de las unidades lexicales (palabras individuales, **lexias** o multi-palabras) como "cadenas" reconocidas por el software;

- la segunda, denominada "**lema**", contiene las etiquetas con las que están reagrupadas y clasificadas las unidades lexicales.

Según los casos, un lema puede ser:

- el resultado del proceso de lematización automática;
- una voz de un "diccionario personalizado";
- una categoría que indica un grupo di sinónimos;
- una categoría de análisis del contenido;
- etc.

Perfil

En **T-LAB** el perfil de una **unidad de análisis** (unidad lexical o unidad de contexto) corresponde al vector (fila o columna) de la tabla datos que contiene sus valores de **ocurrencia** o de **co-ocurrencia**.

NOTA:
En **Análisis de Correspondencias** se denominan **activos** los perfiles que intervienen en la construcción de los ejes factoriales; mientras que los que no intervienen en la determinación de los mismos se denominan **ilustrativos**.

Polos de Factores

En el **Análisis de Correspondencias** cada factor organiza una dimensión espacial que puede ser representada como una línea o como un eje - en cuyo centro (o baricentro) está el valor "0", y que se desarrolla de una manera bipolar hacia los extremos negativos (-) y positivos (+), de modo que los objetos situados en polos opuestos sean los más diferentes, casi como la "izquierda" y la "derecha" en el eje de la política.

Es útil recordar qué el matemático J.P Benzecri, uno de los contribuidores más importantes a este tipo de análisis, escribió sobre este tema:

"Entender un eje factorial significa descubrir lo que hay de análogo por un lado entre cuanto está a la derecha del origen (baricentro), por otro entre cuanto está a la izquierda de él, y después expresar de una manera concisa y precisa la oposición entre los dos extremos". (1984, p. 302, véase **Bibliografía**).

NOTA : Cuando los gráficos factoriales son bidimensionales (o tri-dimensionales) las oposiciones son más de dos: además de la
izquierda y de la derecha, hay hacia arriba y hacia abajo. Con todo los criterios de interpretación son iguales.

Stop Word List

En la práctica del análisis de textos, muchas palabras se definen "vacías" porque solas no tienen ningún contenido específico y/o significativo.

No existe un criterio estándar para construir una lista de estas palabras (**Stop Word List**).

En **T-LAB** la lista se toma de las categorías siguientes:

- adjetivos indefinidos;
- artículos;
- adverbios;
- exclamaciones;
- interjecciones;
- preposiciones;
- pronombres (demostrativos, indefinidos y relativos);
- verbos auxiliares (ser, haber)
- verbos modales (deber, poder, saber, soler, querer)

El usuario puede también utilizar **listas personalizadas de Stop-Words**.

Tablas de datos

Las tablas de datos (o **matrices**) se componen de filas, de columnas y de los valores registrados en las celdas respectivas

.
Estas tablas nos permiten sintetizar - de una manera ordenada - tanto las observaciones que hay que someter a análisis estadísticos (input), como los resultados obtenidos por su aplicación (output).

Por varios motivos, los estadísticos afirman que el éxito de un análisis se debe a la construcción de una "buena tabla".

En **T-LAB**, según los tipos de análisis, las tablas pueden ser de tres tipos, correspondientes a otras tantas maneras de construir cruces entre filas y columnas:

- lemas (o palabras) en fila y textos (o **variables**) en columna;
- textos (o fragmentos de textos) en fila y lemas (o palabras) en columna;
- lemas (o palabras) tanto en fila como en columna.

Los tres tipos, de diversas maneras, sintetizan los fenómenos de **ocurrencia y co-ocurrencia**.

TF-IDF

Esta medida, propuesta por G. Salton (1989), permite comprobar el peso de un termino (unidad lexical) en un documento (unidad de contexto).

Su fórmula es la siguiente:

w *i,j* = ***tf*** *i,j* x ***idf*** *i* (*Term Frequency* x *Inverse Document Frequency*)

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

En la que:

tf *i,j* = número de ocurrencias del termino ***i*** en el documento ***j***
df *i* = número de documentos que contienen ***i***
N = número de documentos en una colección (corpus)

El valor ***tf*** *i,j* (Frecuencia del Termino) puede ser normalizado en la manera siguiente:

tf *i,j* = ***tf*** *i,j* / Max (***f*** *i,j*)

en la que Max (***f*** *i,j*) es la frecuencia máxima de un cualquier termino ***i*** en el documento ***j*** .

Umbral de frecuencia

Durante el procesamiento preliminar de datos **T-LAB** calcula un umbral mínimo de frecuencia para seleccionar las palabras (o los lemas) que serán introducidas en los análisis del menú **Configuratión Automática** y, en particular, para construir la lista de **palabras clave**.

En cualquier caso, para garantizar la fiabilidad de algunos cálculos estadísticos, el umbral mínimo **T-LAB** no es inferior a 4.

*El algoritmo para este cómputo se documenta en uno de los libros de la **Bibliografía** (Bolasco, 1999) y prevé los pasos siguientes:*

- *detección de la gama de las frecuencia bajas, que, a partir de la frecuencia mínima "1", es definida por el primer "salto" en los valores crecientes de las frecuencias;*
- *elección del valor de umbral que, según las dimensiones del corpus, corresponde al valor mínimo en el primer o en el segundo decile (10% o 20%) de la gama.*

Unitad de Análisis

Las unidades de análisis de **T-LAB** son de dos tipos: **unidades lexicales** y **unidades de contexto**.

A - las **unidades lexicales** son **palabras**, simples o "múltiple", archivadas y clasificadas en base a algún criterio. En particular, en la base de datos **T-LAB**, cada unidad lexical constituye un registro clasificado con dos campos: **palabra** y **lema**. En el primer campo ("palabra") se enumeran las palabras así como aparecen en el corpus, mientras que en el segundo ("lema"), se enumeran las etiquetas atribuidas a grupos de unidades lexicales clasificadas según criterios lingüísticos (ej. **lematización**) o a través de **diccionarios** y **plantillas semánticas** definidas por el usuario.

B - Las **unidades de contexto** son porciones de texto en las que se puede dividir el corpus. En particular, en la lógica **T-LAB**, las unidades de contexto pueden ser de tres tipos:

B.1 **documentos primarios** correspondientes a la subdivisión "natural" del corpus (ej. entrevistas, artículos, respuestas a preguntas abiertas, etc.), o sea a los **contextos iniciales** definidos por el usuario;

B.2 **contextos elementales,** correspondientes a unidades sintagmáticas de una o más frases y definidas de modo automático (o semi-automático) por **T-LAB**. Por tanto, en la base de datos **T-LAB** cada documento primario está constituido por uno o más contextos elementales;

B.3 **subconjuntos del corpus** que corresponden a grupos de documentos primarios atribuibles a la misma "categoría" (es. entrevistas de "hombres" o de "mujeres", artículos de un determinado año o de un determinado periódico, y así sucesivamente).

Unidad de Contexto

Véase **unidad de análisis**.

Unitad Lexical

Véase **unidad de análisis**.

Valor Test

Es un índice estadístico que utiliza T-LAB para medir y caracterizar dos tipos de relaciones:

a) Aquellas que una unidad lexical cualquiera mantiene con una categoría cualquiera de una variable y cuyos valores de ocurrencia están incluidos en una tabla de contingencia;
b) Aquellas que conciernen cualquier línea o columna de una tabla de contingencia con factores extraídos mediante un análisis de las correspondencias de la misma tabla.

Según el tipo de relación analizado, las formulas del valor test, recogidas en uno de los volúmenes presentes en la bibliografía (Lebart L. Morineau A. Piron M., 1995, pp 181-184), son las siguientes:

a)

$$t_k(j) = \frac{n_{jk} - n_k \cdot \frac{n_j}{n}}{\sqrt{n_k \cdot \frac{n - n_k}{n - 1} \cdot \frac{n_j}{n} \cdot \left(1 - \frac{n_j}{n}\right)}}$$

En la que 'n_{jk}' indica las ocurrencias dentro de una celda, mientras que 'n_j' y 'n_k' representan, respectivamente, los marginales línea y columna;

b)

$$t\alpha(j) = \sqrt{\mathrm{nj}\,\frac{\mathrm{n}\,-1}{\mathrm{n}\,-\mathrm{nj}}}\,\varphi\alpha\mathrm{j}$$

En la que "n_j" y "$\varphi\alpha_j$" indican respectivamente las ocurrencias del j-ésimo objeto y su coordinada en el α-ésimo eje factorial.

El valor test tiene dos importantes propiedades: un valor umbral (1.96), correspondiente a la significatividad estadística de uso más común (p=0.05), y un signo (- / +).
Esto viene a decir que, ordenando los valores de forma ascendente o descendente, es posible comprobar, en poco tiempo, la relevancia de cada elemento analizado.

T-LAB permite una consulta de la tabla de **valores test** de una manera interactiva.

Variables y Modalidades

En **T-LAB**, las variables son las etiquetas usadas para identificar y clasificar diferentes partes del **corpus**: nombres con características que identifican tipos de sujetos, de textos y de contextos.

Cada variable tiene dos o más **modalidades**, cada una de las cuales, de manera inequívoca, corresponde a un valor de codificación: por ejemplo, la variable "sexo" tiene dos categorías (masculino y femenino).

En **T-LAB**, cada texto se puede identificar con un **máximo de 50 variables**.

Obviamente, para cada una de ellas, se deben indicar las categoría respectiva (max 150), según indicado en las instrucciones contenidas en **Preparación del corpus**.

Para conseguir más información, ver los ejemplos en el directorio demo.

BIBLIOGRAFIA BASICA

Alameda J.R. y Cuetos F. (1995):*Diccionario de frecuencias de las unidades lingüísticas del castellano, Universidad de Oviedo*

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*

Salton G. (1989):*Automatic text processing: the transformation, analysis, and retrieval of Information by Computer, Addison-Wesley, Reading, Massachussets*

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

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)