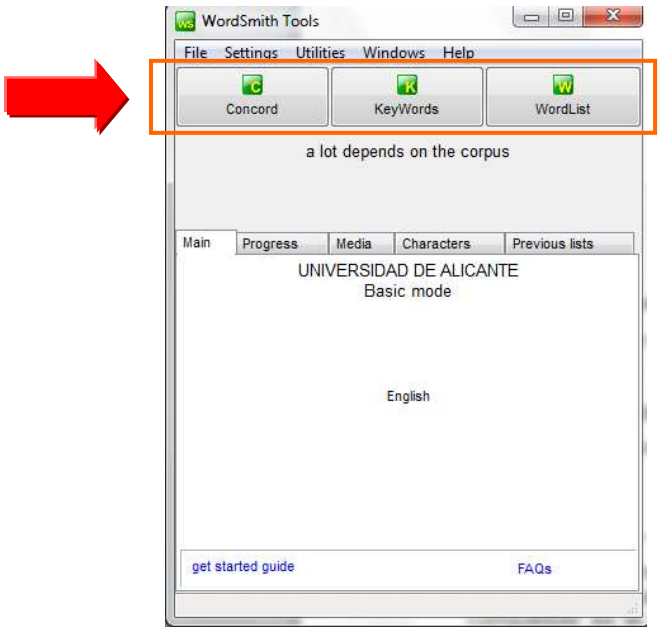
Análisis textual con el programa de concordancias *WordSmith Tools (WST)*

1. Para empezar…

Para abrir el programa, haga doble clic en el **icono** de *WordSmith Tools 5.0* (en adelante WST). Una vez que se ha abierto el programa, aparece la **pantalla principal**, en la que se aprecian de forma destacada los tres botones de sus herramientas específicas: C Concord, K KeyWords y W WordList.



WST está compuesto de: (a) **herramientas**; y (b) **utilidades**. Dentro de cada herramienta hay una serie de instrumentos de análisis y de funciones que permiten, entre otras acciones, elaborar listados de palabras monoléxicas, poliléxicas o polilexemáticas[1], de agrupamientos léxicos (*clusters*) —bien de todo el

[1] Éste es el término empleado habitualmente en la bibliografía sobre lingüística para referirse a una unidad léxica compuesta por dos o más palabras. Otro término compatible es el de n-grama, más común en el ámbito del Procesamiento del

conjunto de textos, o bien de una palabra base—, de palabras claves (*keywords*).

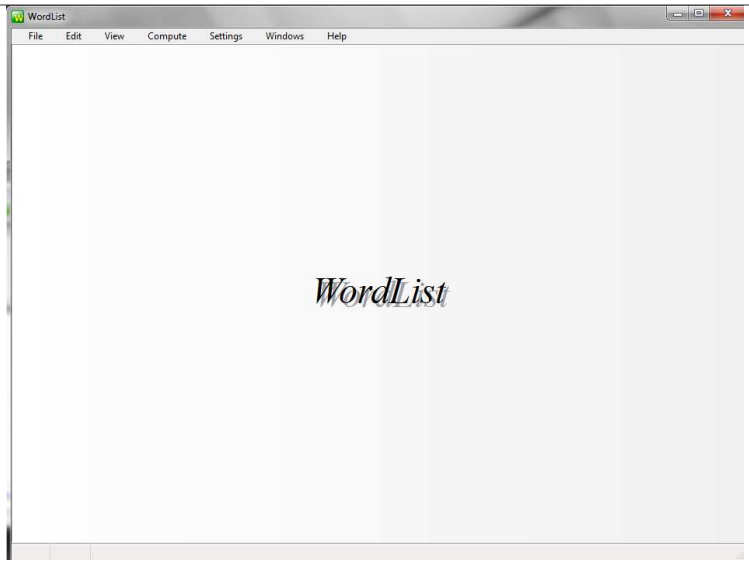
Las **herramientas** de las que se compone son: ***Wordlist*, *KeyWords*, *Concord*.**

Las **utilidades** de este programa en la versión 5.0 son:

- *Character profiler:* permite configurar los tipos de caracteres de un conjunto de textos;
- *Corpus Corruption Finder:* comprueba si alguno de los archivos del corpus está corrupto o está en otro idioma;
- *Data Converter*: permite convertir datos de versiones anteriores
- *File Utilities*: permite realizar diversas acciones sobre los ficheros (comparar dos, reducirlos de tamaño, encontrar ficheros duplicados y/o renombrarlos);
- *File Viewer:*
- *Languages Chooser*: permite seleccionar la lengua del texto o textos que se van a procesar;
- *Minimal Pairs*: encuentra palabras que difieren levemente en su grafía;
- *Text Converter*: permite editar los textos, renombrar los ficheros, cambiar sus atributos y moverlos a otra carpeta si contienen ciertas palabras o frases;
- *Viewer & Aligner*: permite examinar los ficheros en varios formatos. También se puede emplear para copiar un fichero y para alinear las frases de dos ficheros de distinta lengua;
- *Webgetter*: recupera textos directamente de Internet y los descarga con la ayuda de un motor de búsqueda;
- *WSConcGram*: permite encontrar combinaciones léxicas.

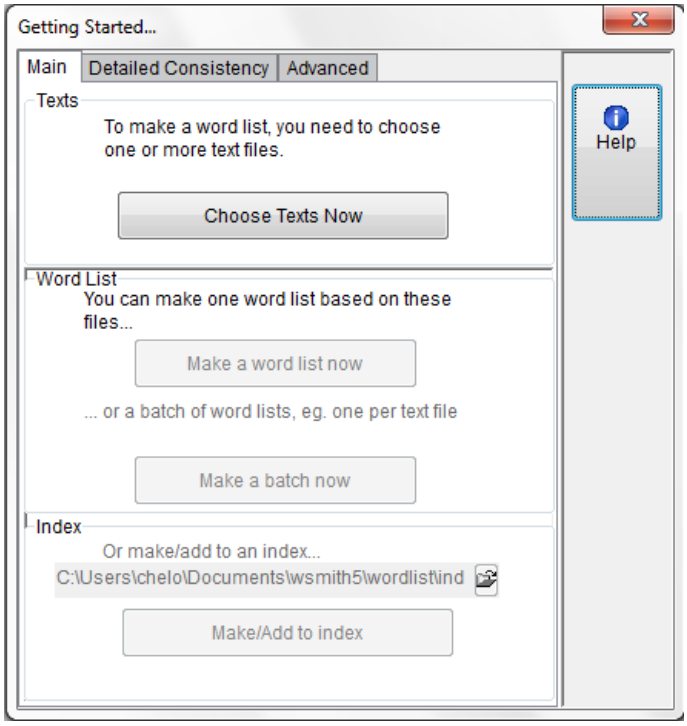
2. La extracción de listas de palabras: WordList

Lenguaje Natural. Concretamente, se utiliza bi-grama para conjuntos de dos palabras, tri-grama para tres, y así sucesivamente.



La herramienta *WordList* permite crear **un listado** de palabras a partir de todos los textos seleccionados (opción <*Make a word list now*>). Para crearla realice los siguientes pasos:

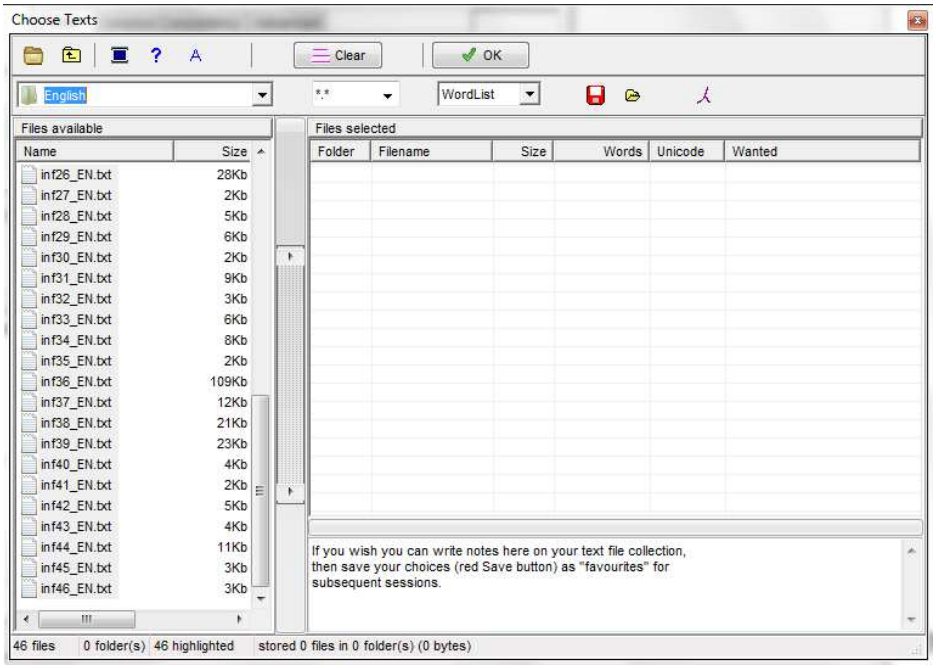
1. Dentro de WordList, seleccione **File >> New**, aparecerá la pantalla **Getting Started…**:



2. A continuación, pulsamos el botón **Choose Texts Now** con el fin de seleccionar los textos que formarán parte del corpus que pretendemos

analizar. Aparece ahora la pantalla **Choose Texts**, desde donde tendrá que navegar para acceder a la ubicación en donde se encuentran sus textos, tal y como se muestra en la imagen:

3. Una vez tenga la ubicación de sus textos, hay que seleccionarlos



También es posible generar un **grupo de listados** (opción <*Make a batch now*>), uno para cada texto seleccionado.

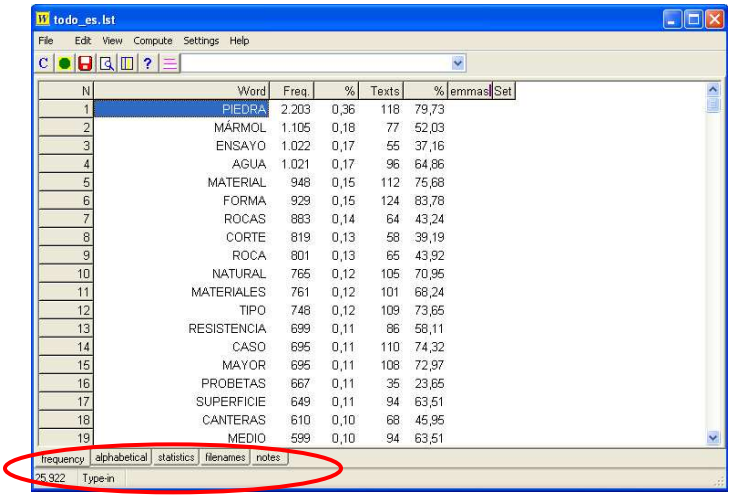
Los **resultados** se muestran en tres tipos de listados:

1) las palabras están ordenadas **alfabéticamente**;
2) la ordenación es por la **frecuencia** de las palabras; y,
3) en el tercer listado, aparecen las **estadísticas** relativas a los textos cargados[2] para la producción de las listas.

Cada uno de estos listados está contenido en una ventana diferente, a la que se accede seleccionando la pestaña correspondiente situada en la parte inferior de la

[2] **FORMATOS DE TEXTO**: La codificación o extensión de los archivos textuales con los que WST trabaja son: texto plano, es decir, .txt., html, SGML o XML. Por defecto el programa tiene activada la opción 'Plain text'. El formato de texto se configura desde la pantalla principal (Setting>Languages).

pantalla, como se podrá apreciar en la figura siguiente:



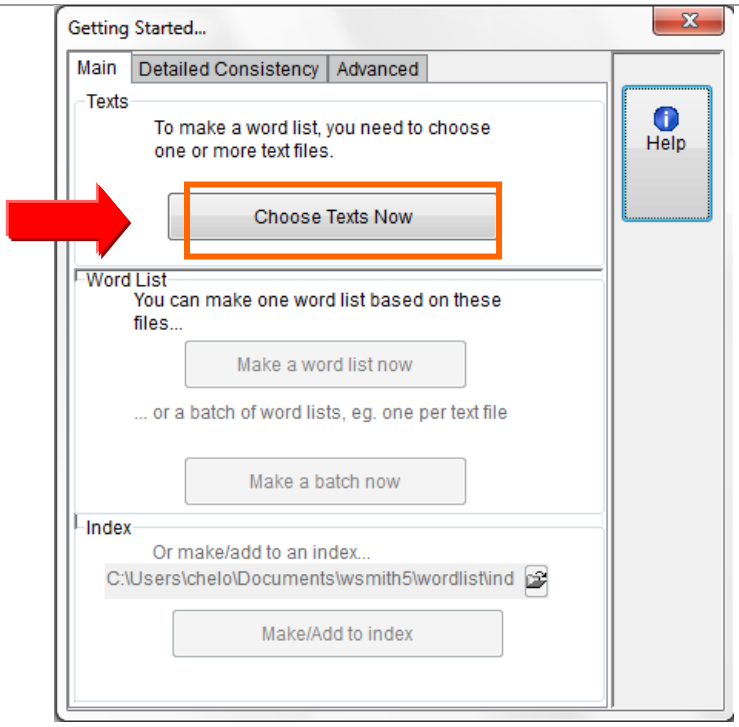
N	Word	Freq.	%	Texts	%	emmasl	Set
1	PIEDRA	2.203	0,36	118	79,73		
2	MÁRMOL	1.105	0,18	77	52,03		
3	ENSAYO	1.022	0,17	55	37,16		
4	AGUA	1.021	0,17	96	64,86		
5	MATERIAL	948	0,15	112	75,68		
6	FORMA	929	0,15	124	83,78		
7	ROCAS	883	0,14	64	43,24		
8	CORTE	819	0,13	58	39,19		
9	ROCA	801	0,13	65	43,92		
10	NATURAL	765	0,12	105	70,95		
11	MATERIALES	761	0,12	101	68,24		
12	TIPO	748	0,12	109	73,65		
13	RESISTENCIA	699	0,11	86	58,11		
14	CASO	695	0,11	110	74,32		
15	MAYOR	695	0,11	108	72,97		
16	PROBETAS	667	0,11	35	23,65		
17	SUPERFICIE	649	0,11	94	63,51		
18	CANTERAS	610	0,10	68	45,95		
19	MEDIO	599	0,10	94	63,51		

2.1 Cargar los textos

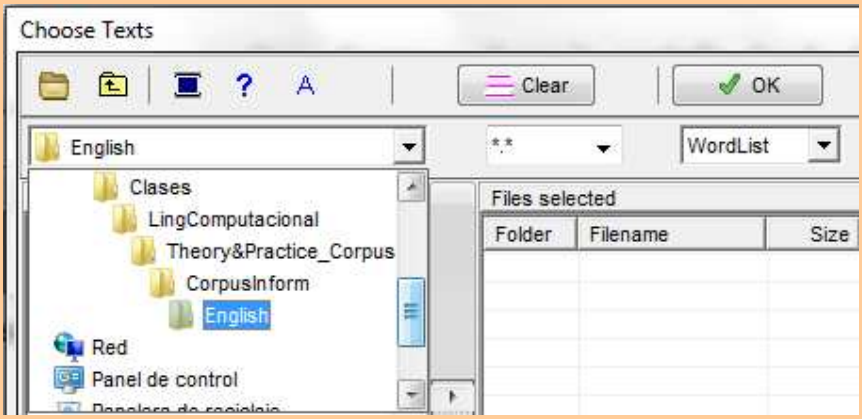
Para empezar a utilizar cualquiera de las herramientas de WST será necesario indicarle **cuál es el conjunto de textos** sobre los que se quiere trabajar.

Vamos a generar el corpus de estudio. Para ello, desde la pantalla principal de *WordList*:

- **1** Selecciona **File>New** y en la ventana *Getting Started* elige **Choose Texts Now**



➢ **2** Tal y como hacemos con el explorador de Windows, buscamos los archivos que se van a cargar. Para ello, **despliega** la flecha del navegador de archivos y accede a la unidad en donde se encuentren sus textos, tal y como se muestra en la imagen:



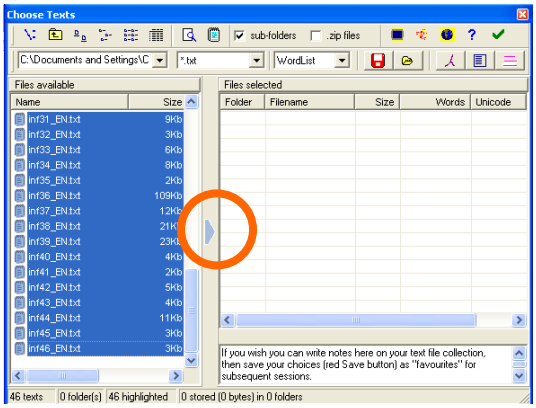
En el cuadro inferior (*Files available*) podrá ver todo lo que hay en esa unidad.

➢ **3** Busque la **carpeta** 'CorpusInform'

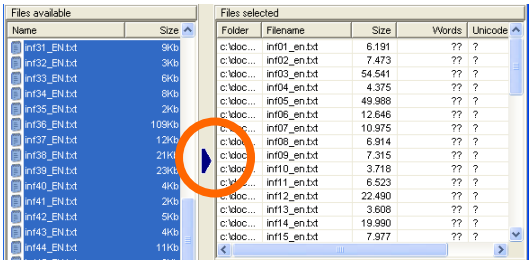
➢ **4** En primer lugar generaremos la lista con los textos en inglés (X:\CorpusInform\English). Cuando los veamos en la pantalla de la izquierda (*Files available*), los seleccionamos todos y hacemos clic en la flecha para pasarlos a la ventana de la derecha (*Files selected*), como se

aprecia en las siguientes figuras:

1



2

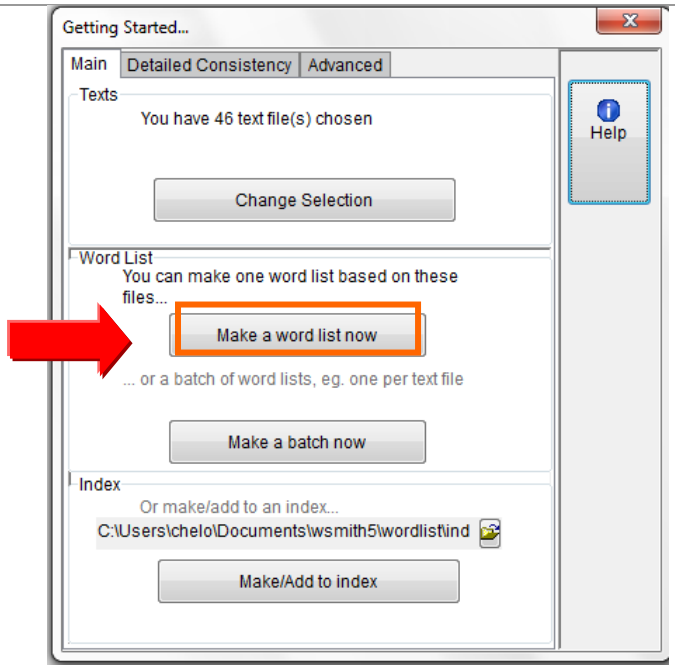


➢ **5** Cuando tenga los textos en la ventana de la derecha, podrá entonces podemos validar la operación. Para ello, dele el visto bueno haciendo **clic en el icono** ✓ situado en la parte superior derecha de la pantalla.

Una vez seleccionado el corpus, vuelve a aparecer la pantalla desde donde se realizan todos los tipos de listados (*Getting Started*). Las opciones son:

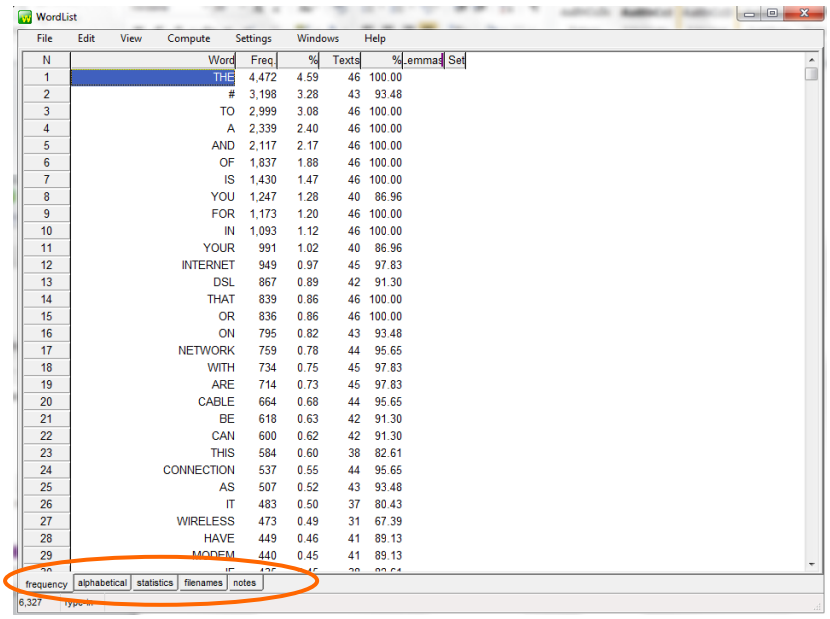
1) **Make a word list now**, para generar una única lista con todos los textos del corpus;
2) **Make a batch now**, para realizar una lista por cada uno de los textos seleccionados;
3) **Make/Add to index**, para elaborar un índice a partir del cual obtener diferentes tipos de listados poliléxicos o de colocaciones.

➢ **6** Inicia el proceso haciendo clic en **Make a word list now.**



Como resultado de esta operación surge una pantalla con varias pestañas en la parte inferior. Las más interesantes son:

a) **frequency**: contiene el listado de frecuencia;
b) **alphabetical**: con el listado alfabético;
c) **statistics**: proporciona un conjunto de datos numéricos (número total de palabras, de cada texto, de párrafos, etc.)



2.2 Cargar un listado de palabras gramaticales (*stopword list*)

Si la finalidad de uso de WST es confeccionar un glosario u observar palabras o grupos de palabras especializadas en contexto, antes de crear un listado de palabras ordenado alfabéticamente o por frecuencia, conviene que alimentemos previamente el programa con unas listas de exclusión que contienen palabras gramaticales, conocidos también por el nombre *stopword list*[3]. Estas listas pueden contener palabras de clase cerrada, es decir, unidades léxicas sin contenido específico, no válidas para un trabajo terminológico, que salen con una elevada frecuencia en los textos y que generan lo que se denomina «ruido». En definitiva, se trata de palabras como artículos definidos e indefinidos, numerales, adverbios, palabras de contenido muy general, etc.

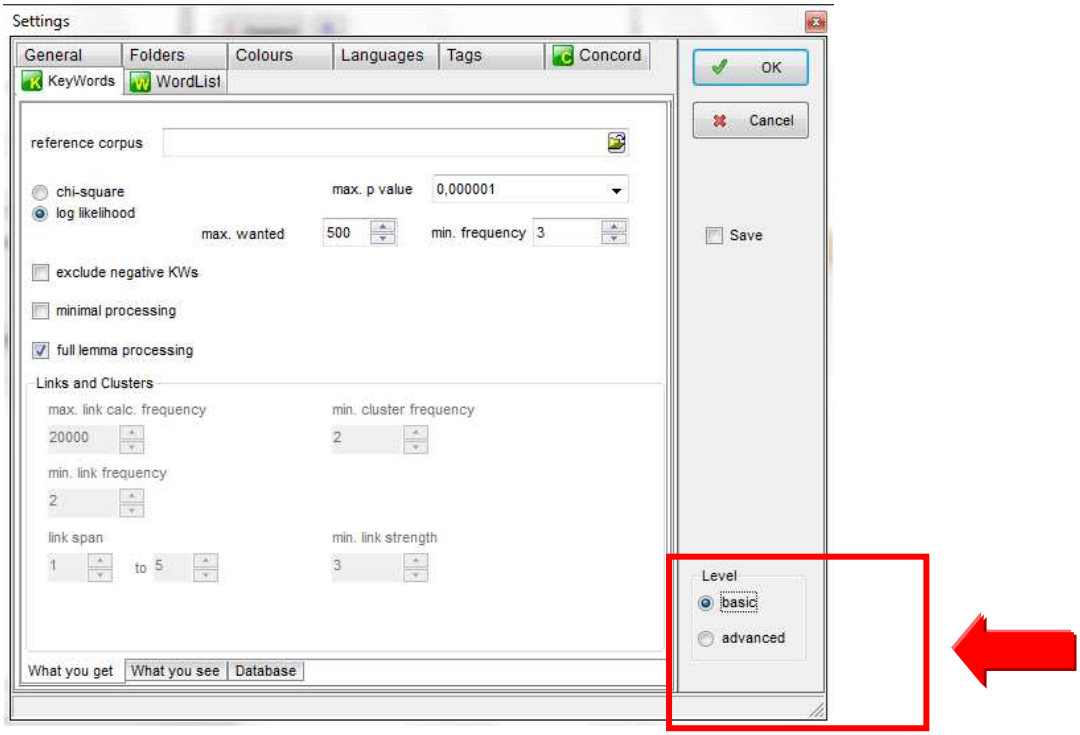
Para poder emplear estos listados, debemos preparar un archivo en texto plano (.txt) en el Bloc de Notas u otra aplicación similar, con todas aquellas palabras que queremos que WST no saque en el listado. Las palabras a excluir deberán estar separadas entre sí por comas o por saltos de párrafo (¶).

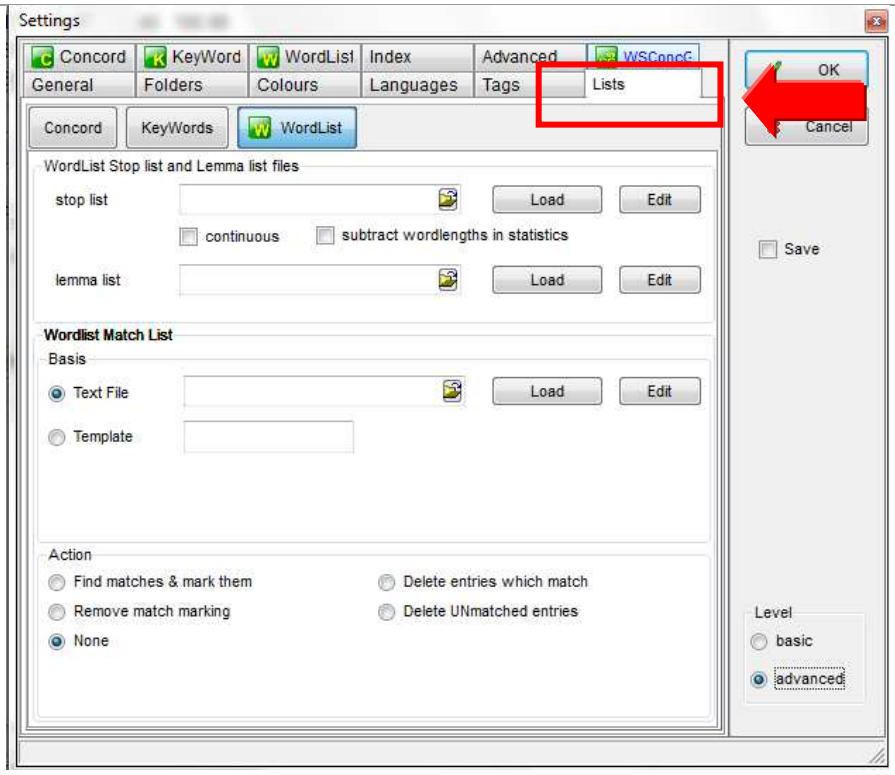
Para cargar una lista de exclusión:

➢ **1** Desde la pantalla principal de WST, vaya a **Setting>Adjust Setting**

[3] Pueden conseguirse listados ya confeccionados en varios idiomas en http://www.unine.ch/info/clef/

➢ 2 En la pantalla que aparece, **Settings**, puede que tenga que cambiar de nivel a **advanced** para que aparezca la pestaña **List**, que ha de seleccionar:





- **3** En el campo 'stop list' ha de indicar la ruta donde se encuentra el listado. Para ello, **haga clic en la carpeta amarilla** y se abrirá el navegador de Windows.
- **4** Busque en la unidad que corresponda el archivo `StopEspa.txt` para el corpus en español.
- **5** Haga clic en **Load** (observa que el botón cambia a *Clear)*
- **6** Haga clic en **OK** para salir de la patalla *Settings*
- **7** Vuelva a generar el listado de palabras como se indica en 2.1 y observe los resultados

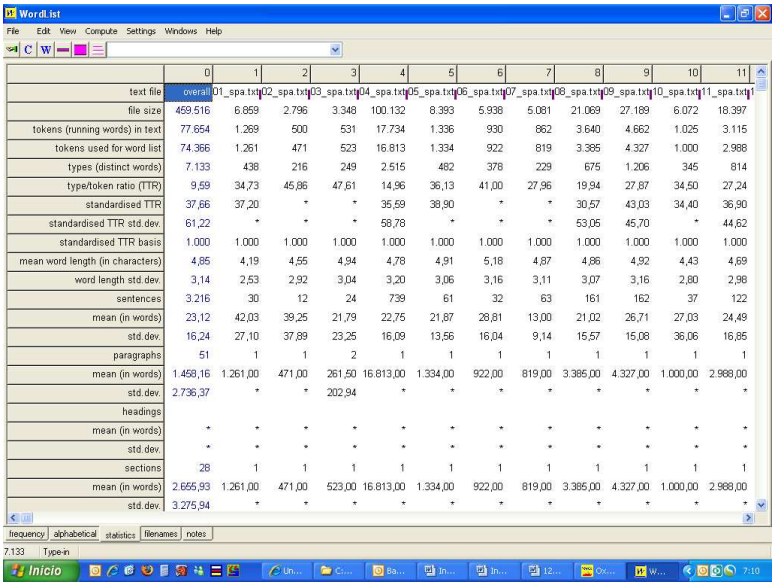
2.3 Guardar los listados

Si queremos guardar las listas hay que:

- **1** Ir a **File>Save** o presionar **Ctrt+F2**. Los datos se guardan por defecto con la extensión **.lst** (formato específico de WordList) para que así puedan volverse a recuperar sin tener que cargar de nuevo los textos. También puede guardarse con otros formatos (**Save as**)

Vuelva a repetir las operaciones anteriores pero ahora cargue los textos en inglés (carpeta CorpusInform\English). Utilice la lista de exclusión en inglés denominada STOPLIST_en.txt

2.4 Las estadísticas



	0	1	2	3	4	5	6	7	8	9	10	11
text file	overall	01_spa.txt	02_spa.txt	03_spa.txt	04_spa.txt	05_spa.txt	06_spa.txt	07_spa.txt	08_spa.txt	09_spa.txt	10_spa.txt	11_spa.txt
file size	459.516	6.859	2.796	3.348	100.132	8.393	5.938	5.081	21.069	27.189	6.072	18.397
tokens (running words) in text	77.654	1.269	500	531	17.734	1.336	930	862	3.640	4.662	1.025	3.115
tokens used for word list	74.366	1.261	471	523	16.813	1.334	922	819	3.385	4.327	1.000	2.988
types (distinct words)	7.133	438	216	249	2.515	482	378	229	675	1.206	345	814
type/token ratio (TTR)	9,59	34,73	45,86	47,61	14,96	36,13	41,00	27,96	19,94	27,87	34,50	27,24
standardised TTR	37,66	37,20	*	*	35,59	38,90	*	*	30,57	43,03	34,40	36,90
standardised TTR std.dev.	61,22	*	*	*	58,78	*	*	*	53,05	45,70	*	44,62
standardised TTR basis	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
mean word length (in characters)	4,85	4,19	4,55	4,94	4,78	4,91	5,18	4,87	4,86	4,92	4,43	4,69
word length std.dev.	3,14	2,53	2,92	3,04	3,20	3,06	3,16	3,11	3,07	3,16	2,80	2,98
sentences	3.216	30	12	24	739	61	32	63	161	162	37	122
mean (in words)	23,12	42,03	39,25	21,79	22,75	21,87	28,81	13,00	21,02	26,71	27,03	24,49
std.dev.	16,24	27,10	37,89	23,25	16,09	13,56	16,04	9,14	15,57	15,08	36,06	16,85
paragraphs	51	1	1	2	1	1	1	1	1	1	1	1
mean (in words)	1.458,16	1.261,00	471,00	261,50	16.813,00	1.334,00	922,00	819,00	3.385,00	4.327,00	1.000,00	2.988,00
std.dev.	2.736,37	*	*	202,94	*	*	*	*	*	*	*	*
headings												
mean (in words)	*	*	*	*	*	*	*	*	*	*	*	*
std.dev.	*	*	*	*	*	*	*	*	*	*	*	*
sections	28	1	1	1	1	1	1	1	1	1	1	1
mean (in words)	2.655,93	1.261,00	471,00	523,00	16.813,00	1.334,00	922,00	819,00	3.385,00	4.327,00	1.000,00	2.988,00
std.dev.	3.275,94	*	*	*	*	*	*	*	*	*	*	*

Algunos de los elementos de la ventana de estadísticas son:

- Las columnas 0, 1, 2, 3 muestran los datos del conjunto de archivos cargados (0) y de cada uno de los archivos individualmente (1, 2, 3, etc.).
- En la columna 0 se puede apreciar que hay un total de ítems o palabras (*tokens*) de 77.654, dato que corresponde al tamaño global del corpus en español.
- La fila *text file* indica la ruta y el nombre de cada archivo.
- En la fila *types* (tipo) se muestra el número de palabras diferentes.
- La ratio tipo/ítem (*type/token ratio*) se expresa en porcentaje y se ha obtenido dividiendo el total de tipos por el total de ítems. Cuanto mayor sea este valor más palabras diferentes contiene el texto. En contrapartida, un valor bajo indicará un número alto de repeticiones, aspecto que se podría traducir en que el texto es menos rico o variado desde el punto de vista del vocabulario. Puede apuntar, por tanto, al nivel de especialización que tiene un texto. Un valor bajo podría indicar que nos hallamos ante un texto con un

grado alto de especialización.

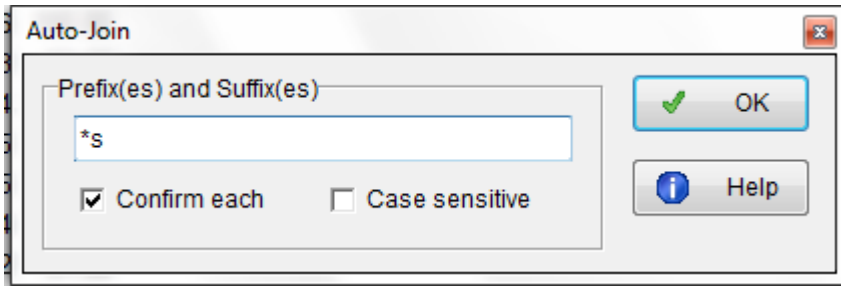
- La ratio tipo/ítem (RTI) estandarizada (*standardised TTR*) calcula la RTI en intervalos regulares. Se emplea para neutralizar la influencia del tamaño del texto en el cálculo de la RTI, ya que los textos de mayor tamaño presentan más repetición y por eso se obtienen valores más bajos que los textos menores. La RTI es sensible a la extensión de la muestra textual, no siendo por tanto del todo fiable para emplearlo en comparaciones entre textos de diferente tamaño. Un texto que es mayor da lugar a más repeticiones y de ahí que su valor pueda ser más bajo. La RTI estandarizada, por su parte, no permite que se tenga en cuenta la repetición de las palabras que aparecen en otra parte del texto, resultando en un valor medio más alto.
- Número de frases y párrafos. Dichas medidas dependen de las convenciones utilizadas para definir tales unidades y, por ello, es necesario asegurarse de que los textos cargados las respetan. Así, según hayamos configurado estos parámetros del texto, el programa identifica como frase la cadena de caracteres entre marcas de puntuación (!?.) y como párrafo el espacio de texto que termina con una línea en blanco, es decir, cuando encuentra dos marcas de párrafo consecutivas (¶), que se consigue al presionar dos veces seguidas la tecla <*intro*>, como es sabido. Por tanto, los textos deben seguir rigurosamente estas normas de delimitación de frase y de párrafo, pues, de otro modo, los datos presentados por el programa no serán correctos.
- Longitud de las palabras. WST puede llegar a contabilizar hasta aquéllas que contienen 50 letras.

2.4 Elementos del listado alfabético y por frecuencia

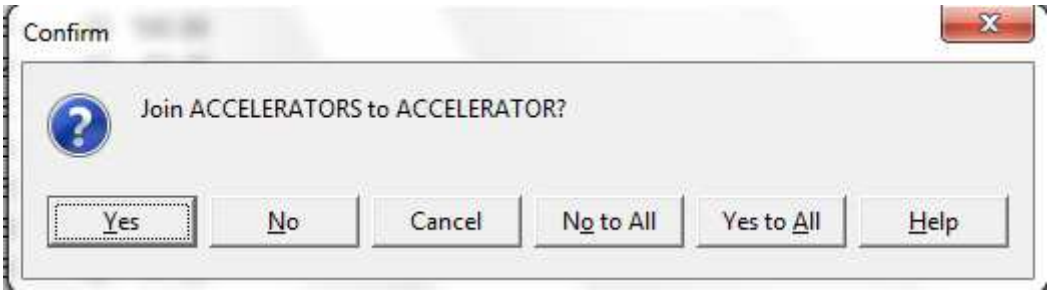
Tanto el listado ordenado alfabéticamente, como el que indica el índice de frecuencia de aparición de las palabras contienen los siguientes elementos:

- Columna "Word", que relaciona las palabras contenidas en el corpus;
- Columna "Freq.", que nos indica el número de veces que aparece la palabra a su derecha;
- Columna "%" o porcentaje de aparición de la palabra calculado a partir del total de palabras del corpus;
- Columna "Texts", número de textos en los que aparece la palabra;

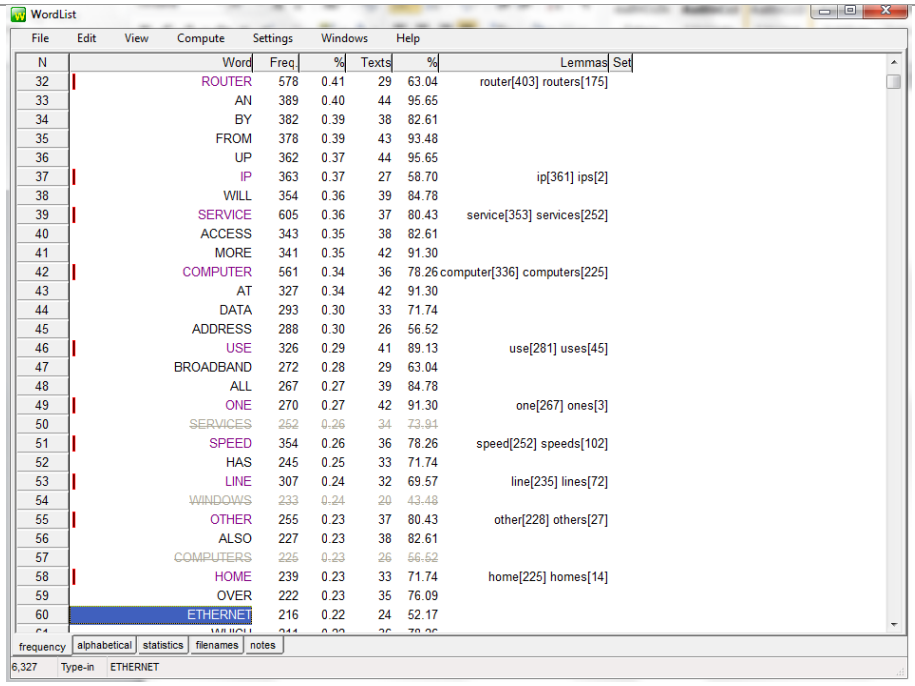
- Columna "%" o porcentaje de aparición de la palabra calculado a partir del total de textos;
- **Columna "Lemmas":** los lemas son unidades léxicas que incorporan formas derivadas. Por ejemplo, en la imagen siguiente podemos apreciar que la columna "Lemmas" suma al término *granite* en singular las veces en que aparece también en plural. Lo mismo ocurre con *specimen*, *type*, *building*, etc. Este proceso se puede hacer de forma manual o de forma automática. El procedimiento manual se hace arrastrando la palabra derivada y soltando cuando nos hallemos en el lema correspondiente. El proceso automático se lleva cabo con el comando <*Auto-Join*> ( ), al que se accede a través de **Edit >> Joining >> Auto-Join**. Con esta opción aparece la pantalla Auto-Join. Si queremos lematizar los plurales en inglés ha de escribir en la caja de Prefijos/Sufijos: *s, tal y como se muestra en la siguiente imagen:



Tras ello, el programa le pedirá confirmación de todas las palabras que coinciden con los criterios. En este caso, puede indicarle **Yes to All:**



- Así, obtenemos una indicación más fiel con respecto a la frecuencia y una lista reducida en su tamaño.



3. La extracción de concordancias: Concord

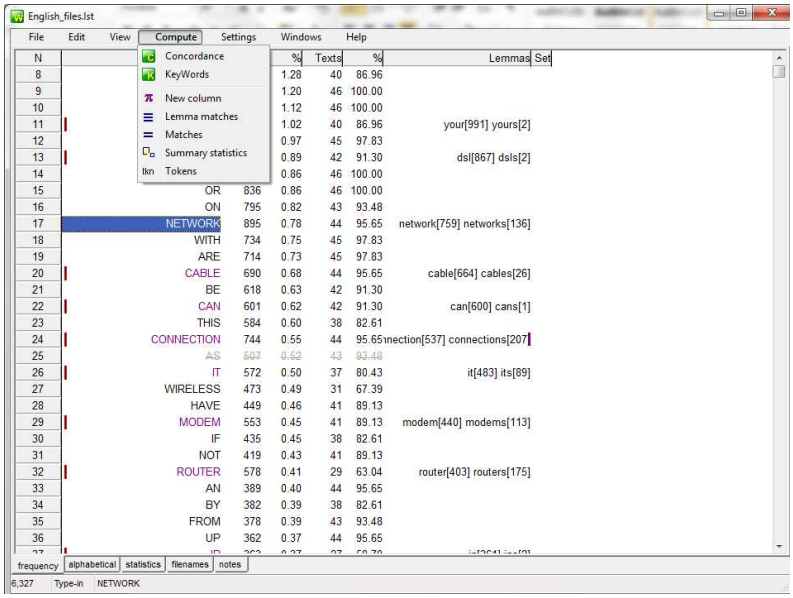
Concord es la aplicación de concordancias de *WordSmith*. Esta herramienta produce concordancias o listados de aparición de una palabra específica —llamada *palabra de búsqueda, palabra base* y también *palabra clave* que puede estar formada por una unidad, varias o parte de ésta— acompañada del texto que la rodea (co-texto).

El tipo de concordancia más común es *Key Word In Context* (KWIC) o palabra clave en contexto. Una lista KWIC agrupa las apariciones de la palabra de búsqueda, que aparece destacada en el centro, lo cual permite analizar y detectar con rapidez sus colocadores o palabras que aparecen en su entorno. Esta opción hace posible el análisis de patrones lingüísticos que salen con una determinada frecuencia en el corpus, aspecto que refleja el comportamiento real en contexto de una palabra, ya se trate de un corpus general o de uno especializado. Las concordancias son instrumentos consolidados ya como indispensables en el estudio de las colocaciones y patrones léxicos; por ello, resulta una pieza clave en la investigación de un corpus.

3.1 Visualizar una concordancia

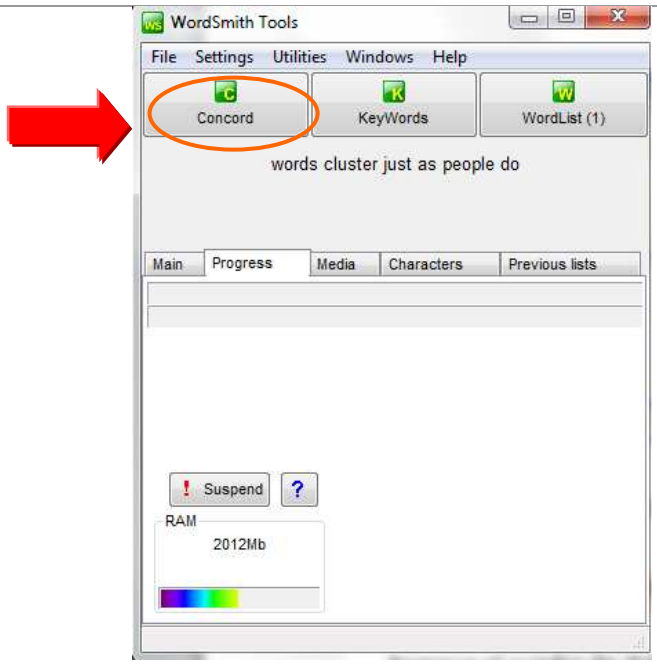
Una manera de acceder a las concordancias de una palabra es seleccionándola en una de las listas de palabras y solicitando desde aquí sus concordancias. Para ello, desde la pantalla de WordList, pestalla *frequency*:

- **1** Abra la lista de palabras (archivo con extensión lst) que generó para el corpus en inglés.
- **2** Seleccione como palabra clave "network".
- **3** Desde el **Menú** seleccione **Compute>Concordance** (o pinchar en la **C** de la barra de herramientas).



El otro modo es:

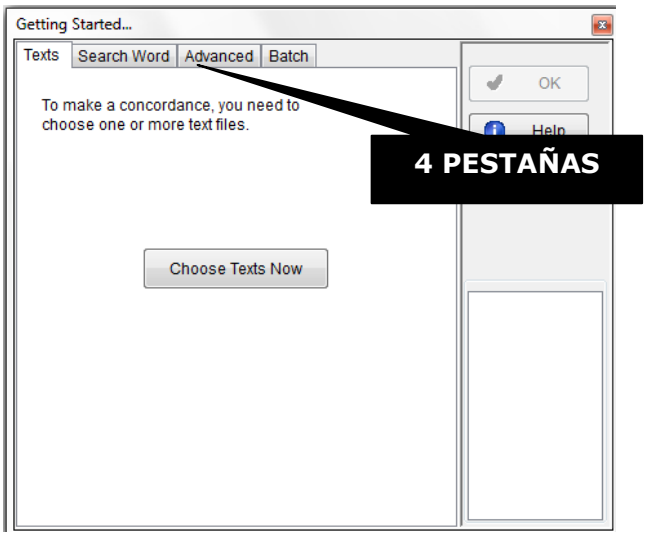
- **1** Seleccione la herramienta *Concord* desde el menú principal de WST:



La solicitud de una KWIC desde *Concord* se realiza del siguiente modo:

➢ **2** Vaya a **File**>**New** del menú de la herramienta *Concord*.

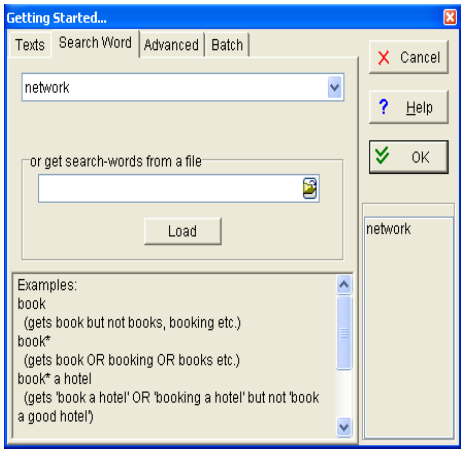
Aparece el cuadro de diálogo *Getting Started…* que se divide en cuatro pestañas, tal y como se muestra en la siguiente figura:



La primera pestaña (**Text**) nos sirve para seleccionar los textos con los que vamos a trabajar. Si pincha el botón **Choose Texts Now** se abrirá la misma pantalla que le ofreció *WordList* para seleccionar los textos.

➢ 3 Seleccione los textos en el idioma inglés tal y como se indica en 2.1.

La segunda pestaña (**Search Word**) sirve para especificar la palabra de la cual se quieren obtener las concordancias.



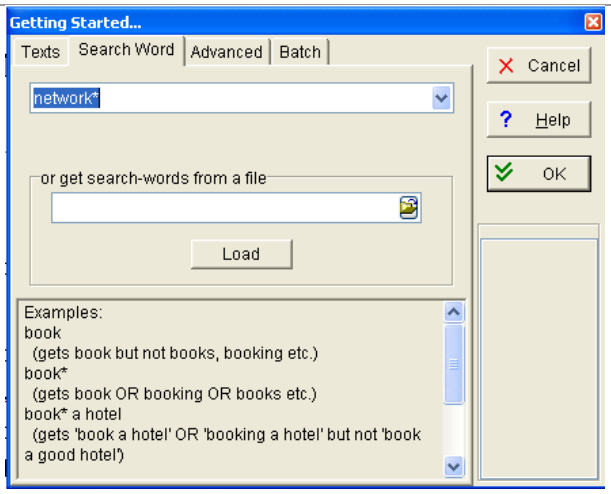
Al escribir la palabra de búsqueda podemos utilizar una serie de comodines[4] que nos permitirán recuperar contextos de más de una opción. Por ejemplo, puede ser interesante recoger tanto las formas singulares como las plurales del sustantivo 'wire' y 'wires', sus compuestos 'wireless', 'unwire', etc. También se puede preparar un fichero de texto con diferentes palabras de búsqueda (**or get search-words from a file**).

Para obtener la concordancia,

➢ 4 escriba, por ejemplo, **network*** y haga clic en **OK**

[4] **COMODINES**

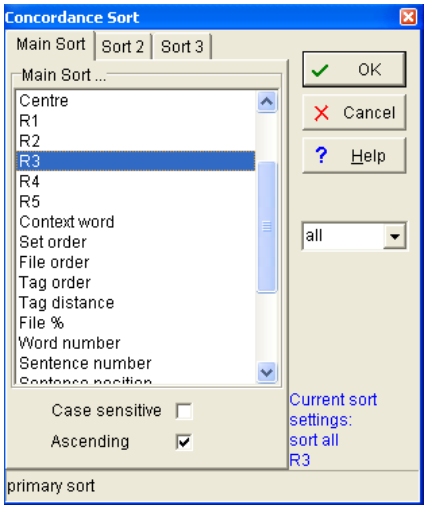
*****: sustituye un número indeterminado de letras tanto delante (*wire) como detrás (wire*) de la palabra buscada;
?: sustituye cualquier carácter o también signos de puntuación ('wire?'-> 'wireless' y 'wire,');
^: sustituye cualquier carácter del alfabeto. De este modo, si buscamos 'wire^' se puede recuperar 'wireless' o 'wires', pero no 'wire,' ;
==: al poner dos == delante y detrás de la palabra, se distinguen mayúsculas y minúsculas (*case sensitive*). Si buscamos '==Wire==' recuperará únicamente 'Wire', pero no 'wire' o 'WIRE'.



3.2 Ordenar los resultados

Una vez que obtenemos la lista de concordancias, los resultados se pueden reordenar en función de diferentes parámetros a fin de detectar patrones de forma visual. Podemos indicar a WST que ordene teniendo en cuenta *n* número de palabras a la derecha o a la izquierda del núcleo o palabra base de la concordancia. Para ello,

- **1** Vaya a **Edit** y seleccione **Resort** o pulse **F6**. Aparecerá la ventana *Concordance Sort:*

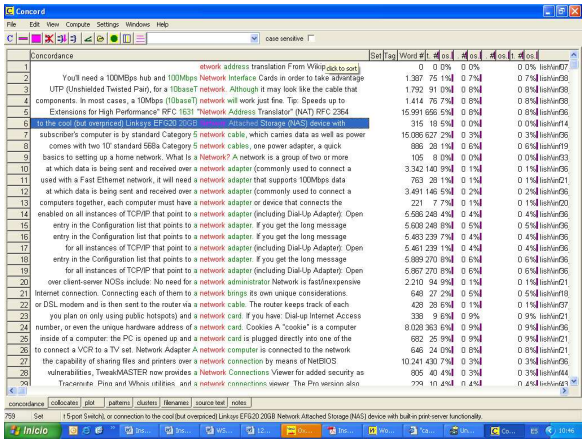


- **2** Configure el orden de la concordancia de la siguiente manera:
  - Para la pestaña **Main sort** selecciona **Centre**
  - Para la pestaña **Sort 2** selecciona **L1** (que significa primera palabra a

la izquierda)

- Para la pestaña **Sort 3** seleccione **R1** (que significa primera palabra a la derecha).

Como se observa en este listado, lo primero que aparece ordenado alfabéticamente son las palabras que ocurren inmediatamente antes de la palabra base.



El objeto de configurar el orden en que WST debe mostrar los datos es hacer posible la detección de patrones léxicos característicos. Distinguir de manera visual dichos patrones no resulta una tarea fácil si no ordenamos los datos de algún modo. Sin embargo, al indicar a WST cómo queremos que reordene las líneas de concordancias la búsqueda de patrones léxicos se simplifica enormemente.

En nuestra búsqueda de *network*,* al indicarle al programa que destaque la primera palabra a la izquierda (Sort2: L1) a partir de la palabra central (Main sort: centre), y la segunda palabra a partir de la palabra central (Sort3:R1), nos es posible observar que frecuentemente se repiten expresiones del tipo ***network*** *adapter*, ***network*** *card, area* ***network***, etc. que responden al patrón *network*+sustantivo o adjetivo/sustantivo+*network,* etc.

3.3 Tipos de búsquedas

Las búsquedas en *Concord* pueden ser simples o complejas.

Las **búsquedas simples** se realizan en la pestaña *Search word* y, como hemos visto, se realizan a partir de una palabra o palabras clave y añadiendo o no

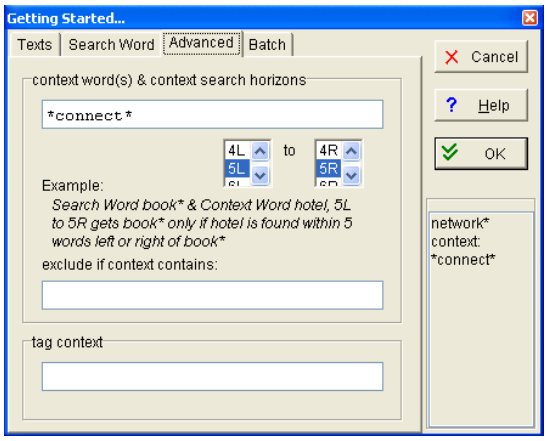
comodines.

Las **búsquedas avanzadas** se llevan a cabo desde la pestaña *Advanced*. Aquí podemos buscar una palabra de dos modos:

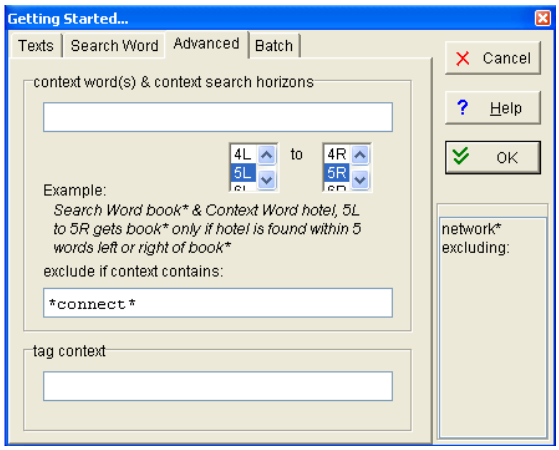
a. Buscando una palabra con otra palabra en contexto (*context word(s) & context search horizons*).
b. Buscando palabras con comodines, eliminando posibles opciones (*exclude if context contains*).

Para extraer los contextos de una palabra en los que también encuentra otra palabra…

- **1** Ve a la pestaña **Advanced** y en el cuadro **Context Word(s) & Context Search Horizons** indicamos qué palabra se tiene que encontrar y entre qué posiciones.
- **2** Buscamos contextos de "network" que contengan la palabra *"connect"* con sus variantes (*connect*) entre las posiciones 5L (la 5ª posición a la izquierda de "network") y 5R (la 5ª posición a la derecha de "network"). Así, los contextos de "network" que no contengan la palabra "*connect*" entre estas posiciones no serán tenidas en cuenta.



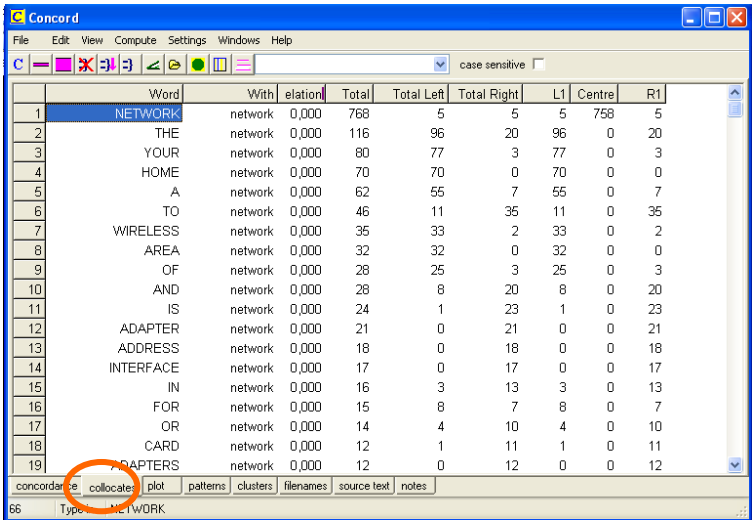
- **3** Ahora los contextos que no nos interesan son los que contienen "connect" o sus variantes. Por lo tanto, la expresión en contexto que indicaremos en el cuadro **exclude if context contains** es *connect*. De esta manera, WST no extraerá ningún contexto de "network" en el que también se encuentre *connect*.



3.4 La extracción de colocaciones: Collocates, Patterns y Clusters

Las colocaciones son, en el marco del programa WST, los patrones sintácticos y/o semánticos que se repiten a lo largo de una lista de concordancias. Hay tres formas básicas de obtener información sobre los colocados que rodean la palabra base (en nuestro ejemplo, "network") de un listado de concordancias:

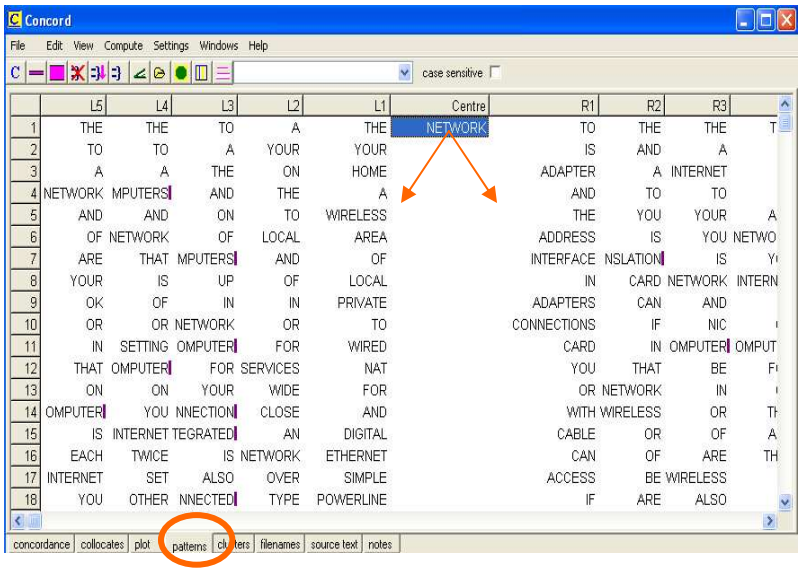
- **1 Pestaña inferior *Collocates***. Esta pantalla nos muestra qué palabras acompañan a la base de las concordancias en cada una de las posiciones de su contexto y con qué frecuencia:



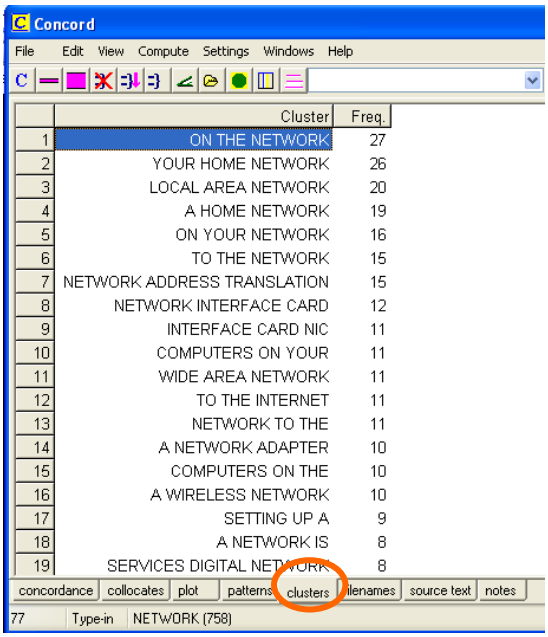
	Word	With	elation	Total	Total Left	Total Right	L1	Centre	R1
1	NETWORK	network	0,000	768	5	5	5	758	5
2	THE	network	0,000	116	96	20	96	0	20
3	YOUR	network	0,000	80	77	3	77	0	3
4	HOME	network	0,000	70	70	0	70	0	0
5	A	network	0,000	62	55	7	55	0	7
6	TO	network	0,000	46	11	35	11	0	35
7	WIRELESS	network	0,000	35	33	2	33	0	2
8	AREA	network	0,000	32	32	0	32	0	0
9	OF	network	0,000	28	25	3	25	0	3
10	AND	network	0,000	28	8	20	8	0	20
11	IS	network	0,000	24	1	23	1	0	23
12	ADAPTER	network	0,000	21	0	21	0	0	21
13	ADDRESS	network	0,000	18	0	18	0	0	18
14	INTERFACE	network	0,000	17	0	17	0	0	17
15	IN	network	0,000	16	3	13	3	0	13
16	FOR	network	0,000	15	8	7	8	0	7
17	OR	network	0,000	14	4	10	4	0	10
18	CARD	network	0,000	12	1	11	1	0	11
19	ADAPTERS	network	0,000	12	0	12	0	0	12

- **2 Pestaña inferior *Patterns***. En esta ventana obtenemos un listado resumen de los colocados agrupados en las posiciones en que aparecen

más frecuentemente.



➢ **3 Pestaña inferior *Clusters*** o agrupaciones de palabras que se repiten un mínimo determinado de veces a lo largo de un listado de concordancias. Se obtiene un listado de las agrupaciones con una frecuencia mínima de aparición que son candidatas a términos sintagmáticos o a expresiones fraseológicas.

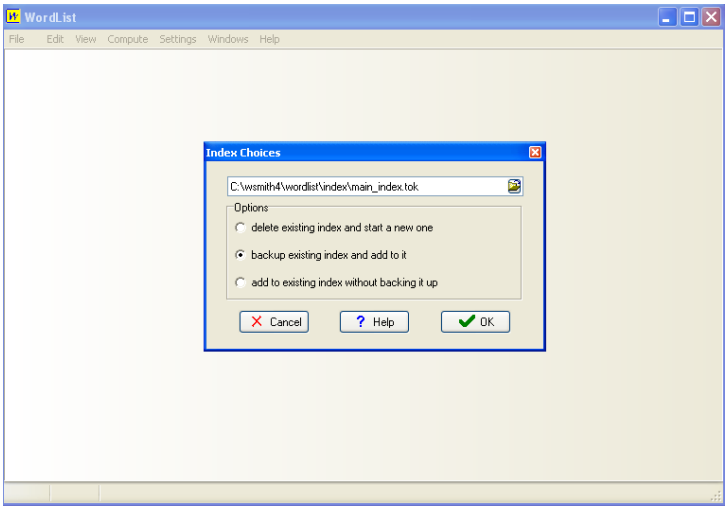


4. La creación de listados poliléxicos

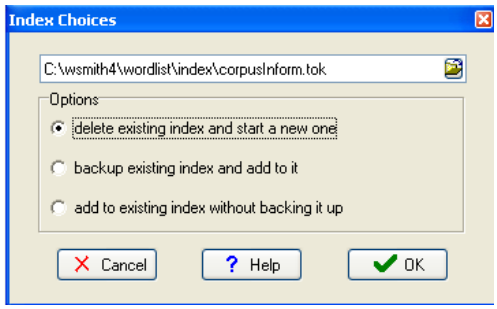
Con *WST* es también posible generar listados poliléxicos, a saber: de dos palabras, de tres, de cuatro, hasta un total de ocho.

Para ello,

- **1** Desde la herramienta ***WordList*** cree un listado con la opción ***Make/Add to index.*** Aparecera la ventana *Index Choices:*



- **2** Pulse la carpeta amarilla para guardar el índice con otro nombre (corpusInform) y saber dónde se va a guardar (C:\wsmith4\wordlist\index\corpusInform.tok)



- **3** Seleccione la opción **delete existing index and start a new one** y haga clic en OK
- **4** Ahora abra el índice que acabas de crer: **File>Open** (busca en la carpeta

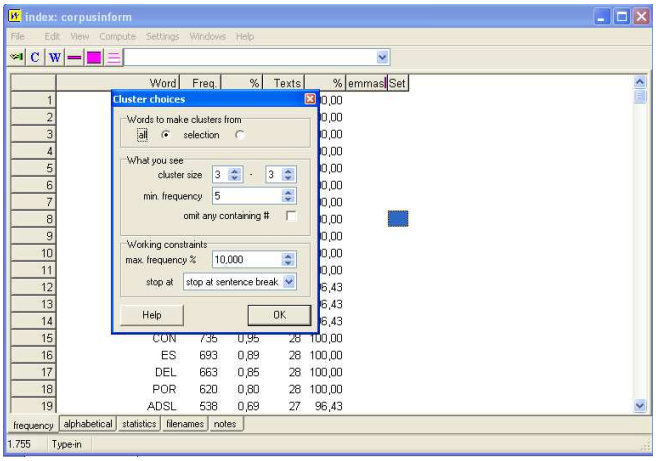
donde guardó el índice).

Una vez que el índice está elaborado y abierto, es posible generar bien un listado de la totalidad del corpus de agrupaciones léxicas (*clusters*) o bien pedirle al programa que calcule la información mutua.

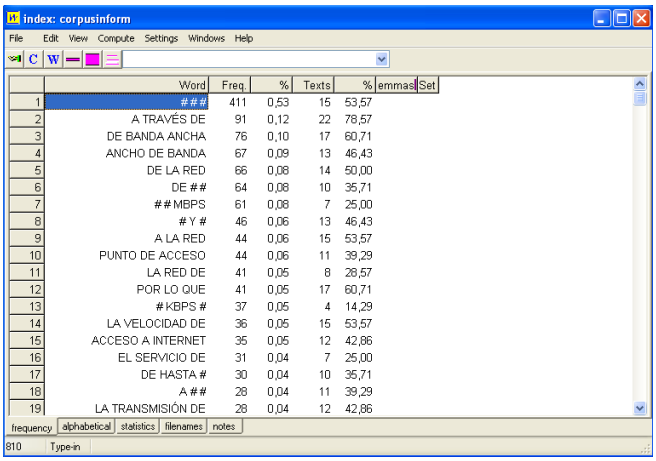
4.1 Generar el listado de agrupaciones léxicas

Para generar el listado de agrupaciones léxicas:

- **1** Vaya a **Compute>Clusters**. Aparecerá la ventana **Cluster choices** en donde podemos configurar las opciones de las que dispone WST:



- **2** Tras configurar las opciones que desee, haga clic en **OK** y obtendrá el listado de agrupaciones:



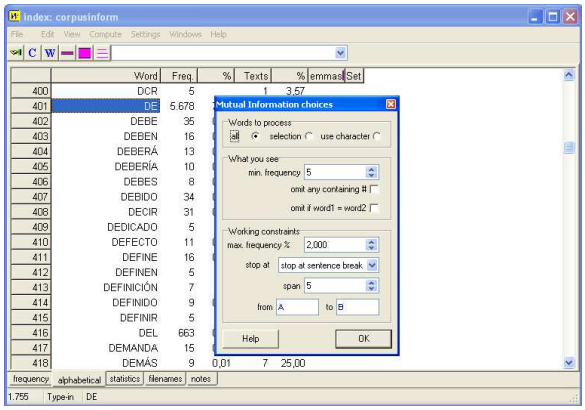
	Word	Freq.	%	Texts	%	emmas	Set
1	###	411	0,53	15	53,57		
2	A TRAVÉS DE	91	0,12	22	78,57		
3	DE BANDA ANCHA	76	0,10	17	60,71		
4	ANCHO DE BANDA	67	0,09	13	46,43		
5	DE LA RED	66	0,08	14	50,00		
6	DE ##	64	0,08	10	35,71		
7	## MBPS	61	0,08	7	25,00		
8	# Y #	46	0,06	13	46,43		
9	A LA RED	44	0,06	15	53,57		
10	PUNTO DE ACCESO	44	0,06	11	39,29		
11	LA RED DE	41	0,05	8	28,57		
12	POR LO QUE	41	0,05	17	60,71		
13	# KBPS #	37	0,05	4	14,29		
14	LA VELOCIDAD DE	36	0,05	15	53,57		
15	ACCESO A INTERNET	35	0,05	12	42,86		
16	EL SERVICIO DE	31	0,04	7	25,00		
17	DE HASTA #	30	0,04	10	35,71		
18	A ##	28	0,04	11	39,29		
19	LA TRANSMISIÓN DE	28	0,04	12	42,86		

4.2 Generar el listado de información mutua (IM)

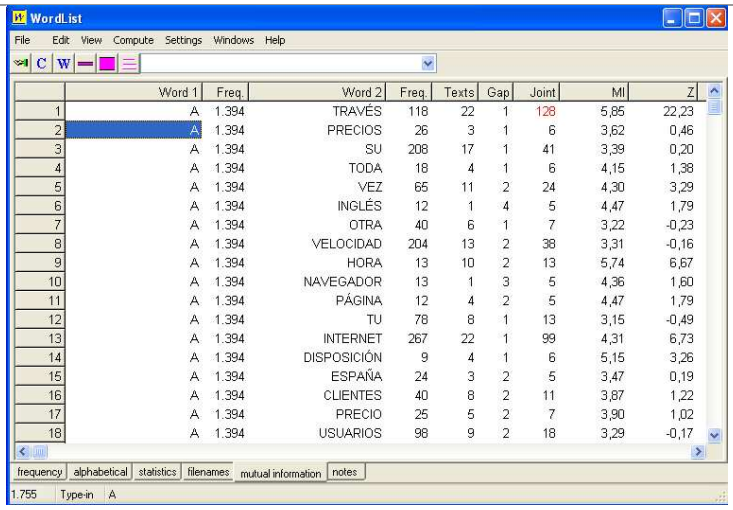
En el listado de información mutua, además de los índices de frecuencia, de la proximidad de las palabras que pone en relación, las veces que aparecen juntas, entre otros datos, se muestra una variedad de relaciones colocacionales; concretamente, MI, *Z Score*, MI3 y *Log Likelihood*.

Para generar el listado de IM:

- **1** Vaya a **Compute>Mutual Information**. Aparecerá la ventana **Mutual Information choices** en donde podemos configurar las opciones de las que dispone WST.



- **2** Tras configurar las opciones que desee, haga clic en **OK** y obtendrá el listado de agrupaciones:



	Word 1	Freq.	Word 2	Freq.	Texts	Gap	Joint	MI	Z
1	A	1.394	TRAVÉS	118	22	1	128	5,85	22,23
2	A	1.394	PRECIOS	26	3	1	6	3,62	0,46
3	A	1.394	SU	208	17	1	41	3,39	0,20
4	A	1.394	TODA	18	4	1	6	4,15	1,38
5	A	1.394	VEZ	65	11	2	24	4,30	3,29
6	A	1.394	INGLÉS	12	1	4	5	4,47	1,79
7	A	1.394	OTRA	40	6	1	7	3,22	-0,23
8	A	1.394	VELOCIDAD	204	13	2	38	3,31	-0,16
9	A	1.394	HORA	13	10	2	13	5,74	6,67
10	A	1.394	NAVEGADOR	13	1	3	5	4,36	1,60
11	A	1.394	PÁGINA	12	4	2	5	4,47	1,79
12	A	1.394	TU	78	8	1	13	3,15	-0,49
13	A	1.394	INTERNET	267	22	1	99	4,31	6,73
14	A	1.394	DISPOSICIÓN	9	4	1	6	5,15	3,26
15	A	1.394	ESPAÑA	24	3	2	5	3,47	0,19
16	A	1.394	CLIENTES	40	8	2	11	3,87	1,22
17	A	1.394	PRECIO	25	5	2	7	3,90	1,02
18	A	1.394	USUARIOS	98	9	2	18	3,29	-0,17

- 3 Este listado puede guardarlo como Excel (**File>Save as**) para eliminar manualmente el ruido o datos no válidos.

FIN DE ESTA PRÁCTICA

5. Bibliografía

Scott, M. (2003): *WordSmith Tools version 4.0*, Oxford: Oxford University Press.

Vargas Sierra, C. (2006): «El proceso terminográfico multilingüe con *WordSmith Tools*», *CONFLUENCIAS - Revista de Tradução Científica e Técnica*, n.4, pp. 84-107. [Disponible en: http://www.ua.es/personal/chelo.vargas/Documentos/n4_vargas-sierra.pdf].

Vargas Sierra, C (2005): *Aproximación terminográfica al lenguaje de la piedra natural. Propuesta de sistematización para la elaboración de un diccionario traductológico*. Universidad de Alicante. Tesis doctoral inédita.