User Guide

User Thesaurus Plus

VERSION
1.3

support@dtsearch.co.uk

User Thesaurus Plus

Updated 6 Aug 2021 (1st Draft)

Contents

Introduction

User Thesaurus Plus is a multipurpose editing tool for professionals, IT Administrators, or software developers working with dtSearch User Thesaurus, Macro and Alphabet files.

Professionals -
Search with dtSearch Desktop or dtSearch Network and select from multiple user-defined thesaurus files for high specificity search expansion.

Software developers
Create multiple thesaurus and macro files for use with other applications that use the dtSearch Engine.

IT Administrators
Create multiple thesaurus and macro files for use with dtSearch Network or other applications that use the dtSearch Engine.

dtSearch Desktop and Network have built-in editors for the User Thesaurus and Macro files, however they can only create a single file, have no sorting, merging, automatic de-duplication, or text cleanup; the built-in alphabet file editor can only edit characters in the range 33 to 127, user Thesaurus Plus adds ability to edit all Unicode currency symbols.

User Thesaurus Plus is distributed on its own for use with dtSearch Desktop or dtSearch Network; it is also distributed as part of a Language Extension Pack (LEP500 series) for use by developers that need wider distribution licensing, and with the premium add-on to Site Search ONE WordPress Plugin.

Installation & Registration

Installation

- Download (**Save...**) the file UserThesaurusPlusSetup.msi to the machine where you want to install it.
- Double-click on UserThesaurusPlusSetup.msi and follow the instructions.
- The installer will create a short-cut on your Windows Programs menu named User Thesaurus Plus under the folder DTSUK.
- Install dtSearch Desktop or dtSearch Network on your machine and create a User Thesaurus file (it can be empty); if you have not done this, start dtSearch Desktop, click on the **Options|Preferences** menu, navigate to **Search Options**, click on **User Thesaurus** and click **OK** to close the dialog box.

Upgrading or Uninstalling

WARNING
Uninstalling or upgrading will remove all the sample files installed by User Thesaurus Plus. If you have modified any of the sample files you are advised to make copies and store them outside of the User Thesaurus Plus Data folder or rename them.

If you need to un-install, go to Windows **Settings|Control Panel|Add or Remove Programs** and click on User Thesaurus Plus (or UTP100 or LEP500 if it was installed as part of a Language Extension Pack or an earlier version).

Evaluation Version
When first installed the program will run in evaluation mode. The evaluation version will run for 30-days, it will allow you create and edit files and to select all the sample files in the Data folder. You can enter a serial number at any time. (This will restore functionality if the evaluation period has expired).

Registration
If you have purchased a license for UTP500 or LEP500 you will need to enter the serial number to continue beyond 30 days. From the **Help|About** menu press the **Add serial...** button.

Enter the serial number and press OK. (If you enter an invalid number the text boxes will turn red).

We recommend that you register your purchase online at www.dtsearch.co.uk from the Support menu choose Register.

Initial set-up for use with dtSearch Desktop or Network

User Thesaurus Plus does not require install of LEP500, it can be used with dtSearch Desktop/Network to change the synonym or macro files at any time.

1 - Locate the Thesaurus file path

- Open dtSearch Desktop and from the **Help|About dtSearch**... dialog, scroll to the bottom of the text and look for the path listed after **User Thesaurus**, select and copy the path.

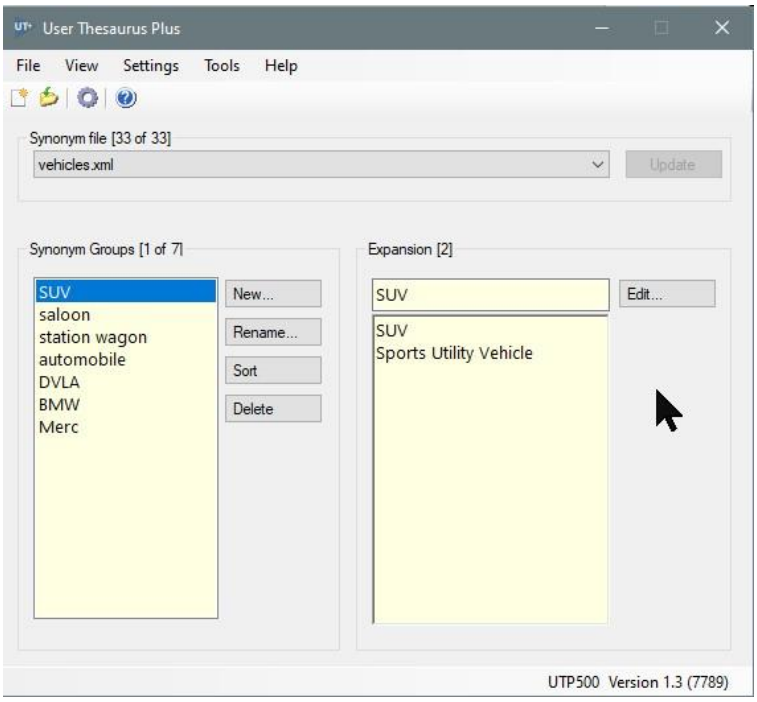
- Open User Thesaurus Plus.

2 - Import Existing Thesaurus and Macro Files

- If you have already edited the dtSearch thesaurus file you should import it into User Thesaurus Plus so that you can easily edit it in future. Click on the **File|Import file** and paste the path into the File name box. Press OK then rename the file.

- If you have already edited the dtSearch Macro file you should click on **File|Import file** and browse to the macros.xml file which you will find in the same location as above.

3 - Set the Target Path



Click on the **Settings|Target Path** menu then click on the **Set Target Path** dialog.

On the dialog that opens, paste the path into the **File name** box, then click **Open**, then on the **Set Target Path** dialog click **OK.**

You will be prompted with a dialog box asking if you want to update the Target file. See Update Target File.

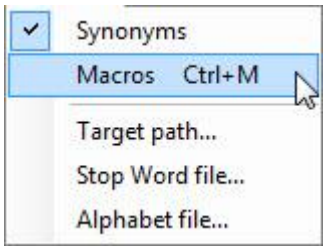
WARNING: Pressing yes will overwrite your target file with the file selected in the drop-down list, ensure you have imported your thesaurus file as above before doing this.

NOTE: You may need to change the user permissions so that you have read/write access to the file. On Windows 7 it is advisable to save the file in the User\AppData folder.

User Thesaurus Plus will make a copy of the thesaur.xml file from the Target path and save it in your My Documents folder under \User Thesaurus Plus\Data. You will be able to select this from the Synonym file drop-down list at any time.

4 - Select the Synonym or Macro file to be edited or used with dtSearch.

The **Settings** menu allows you to choose dtSearch Synonyms or Macro files; initially the default selection is dtSearch Synonym files.



You can toggle between Synonym and Macro files by pressing Ctrl+M

Select a file from the drop-down list, this can now be edited as required.

To use the selected file with dtSearch press the Update button, this will update the synonym or macro target file as appropriate.

dtSearch will use the synonym or macro target file that you have selected to expand your search queries; there is no need to update the index.

You can change the synonym or macro file at any time with dtSearch running or not; dtSearch will only use the selected file once the Update button is pressed.

You can either close User Thesaurus Plus or minimise it to the Windows task bar, your settings are automatically saved.

5 - If you want to use the stop word highlighting feature you will need to set the path to the Stop Word File.

6 - If you want to make currency symbols searchable in dtSearch set the path to the Alphabet file.

Notes on Thesaurus Construction

"General to specific" (i.e., a broader to narrower term) is usually a good way to create and name Synonym groups in the dtSearch User Thesaurus. You would use the Synonym Group Name for the general term, it is better to use the plural term, since when people are searching, they tend to think in terms of "where will I find information on clocks".

The relationship between the broader term and the narrower terms should be true independent of context, for example you could list mice under a group name of rodents, but if you listed it under pests, it would not be correct, because some mice – pet mice, laboratory mice – are not considered pests.

Example:

Group Name	Entries in the thesaurus
jackets	jackets anoraks blazers boleros dinner jacket donkey jacket flying jacket harrington jacket kagouls sports jacket tweed jacket

Although User Thesaurus Plus will de-duplicate terms it is better when preparing a long list to enter them in alphabetic order, so the likelihood of repeating items is reduced. It is important to make sure you include the general term in the synonym list, User Thesaurus Plus will automatically copy the Group Name you enter as the first item on the synonym list.

- If a user searches using a general term like jackets, they will be able to find information on all types of jacket, perhaps finding types of jacket that they were not aware of.

- If they search on a more specific term like blazers they will find information on all other types of jacket which may or may not be considered relevant; so the rule in search is that if you are searching using a very specific term and aren't interested in alternatives, then don't expand your search using the thesaurus options, but DO use **stemming** so that you will find documents containing 'blazer' as well as 'blazers'.

Because the thesaurus and stemming both conflate the search terms to enhance recall, a side effect is that precision will fall, therefore you may need to take extra steps to maintain precision. For example if your index contains documents on a wide range of topics, you may find that your search results have many non-relevant documents, for this reason it is preferable that where possible you should offer your users the possibility of searching in particular 'zones' – the index for each zone could contain just documents from a particular department or even just certain types of document from a department (e.g. research or project reports, marketing collateral).

When you are searching you can also narrow the domain of you search by using a proximity search, for example:

jackets not w/5 potato

will find documents containing the word jackets (or any of the synonyms you have set up in the user thesaurus) but not if the word potato is within a distance of 5 words away; be careful to not simply use - **jackets not potato** - because this could miss a document which happened to have the words potato and jacket(s) anywhere in the same document.

Sample Thesaurus and Macro files

Months

Days

Currencies
162 currency codes with synonyms of currency name, currency sign, including full-width signs and other variants.

Irregular verbs
Danish, Dutch, English, French, German, Italian, Norwegian, Spanish, Swedish.
Handles verb forms (irregular verbs, strong verbs, stem change verbs) that stemming may not, for example in English go, went, gone; run, ran; speak, spoke; drink, drank.

Irregular nouns
Danish, Dutch, English, Norwegian, Swedish
Handles plurals that stemming rules may not, for example in English woman - women, foot - feet, tooth - teeth, child - children.

Names
Genealogy – Cross-lingual – political

Examples of various methods of name searching, nick names, maiden/married names, diminutives, transliterations, political office, aliases.

Names Russian
Male and Female (2 separate files)
Examples of name searching - diminutives, transliterations, similar names.

Prenoms francaise masculin
Examples of Name searching - French male forenames and diminutives.

Colours
List of synonyms of a sample of colours for example azure - cornflower - blue.

Trade
This file is a CLIR (Cross Lingual Information Retrieval) example, it has the English word Invoice with equivalents in over 20 languages.

Fashion
This file is an example of using brand names as synonyms for items of clothing.

Geographic
Small sample from: http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes

Legal & Medical
These files are examples of using legal/medical terms (in Latin) with common English synonyms. The

medical file also includes an example of using synonyms that include trade names, chemical names, and 'street names' for drugs.

Macro Files

- UT_sample_macros_3 contains macro called IPV4 and IPv6. These will enable you to search for all IPv4 and IPv6 addresses in dtSearch Desktop/Network by using the search request @IPV4 or @IPV6 rather than having to enter a long Regular Expression.

- UT_sample_macros_4 contains a macro for BTC - Bitcoin searching.

IPV4 searching

To be able to use this macro the IP addresses need to have the full stop (period) character indexed. This can be done in dtSearch Desktop\Network from the **Options|Preferences|Letters and Words - Edit Alphabet** dialog, make the character 46 a "Letter" then use Save as... to save the modified alphabet file with a unique name -**`"fullstop-as-letter.abc"`** for example, then update the index.

A serious side effect of making the period (full stop) a searchable letter is that words at the end of a sentence will not be found in a normal search, so it is best to make this a separate index which is only used for searching or extracting IP addresses. Remember to select the **`default.abc`** alphabet file for any other indexes. A copy of the modified alphabet file is saved with the index, so there is no need to select the alphabet file each time an index is updated.

IPV6 searching

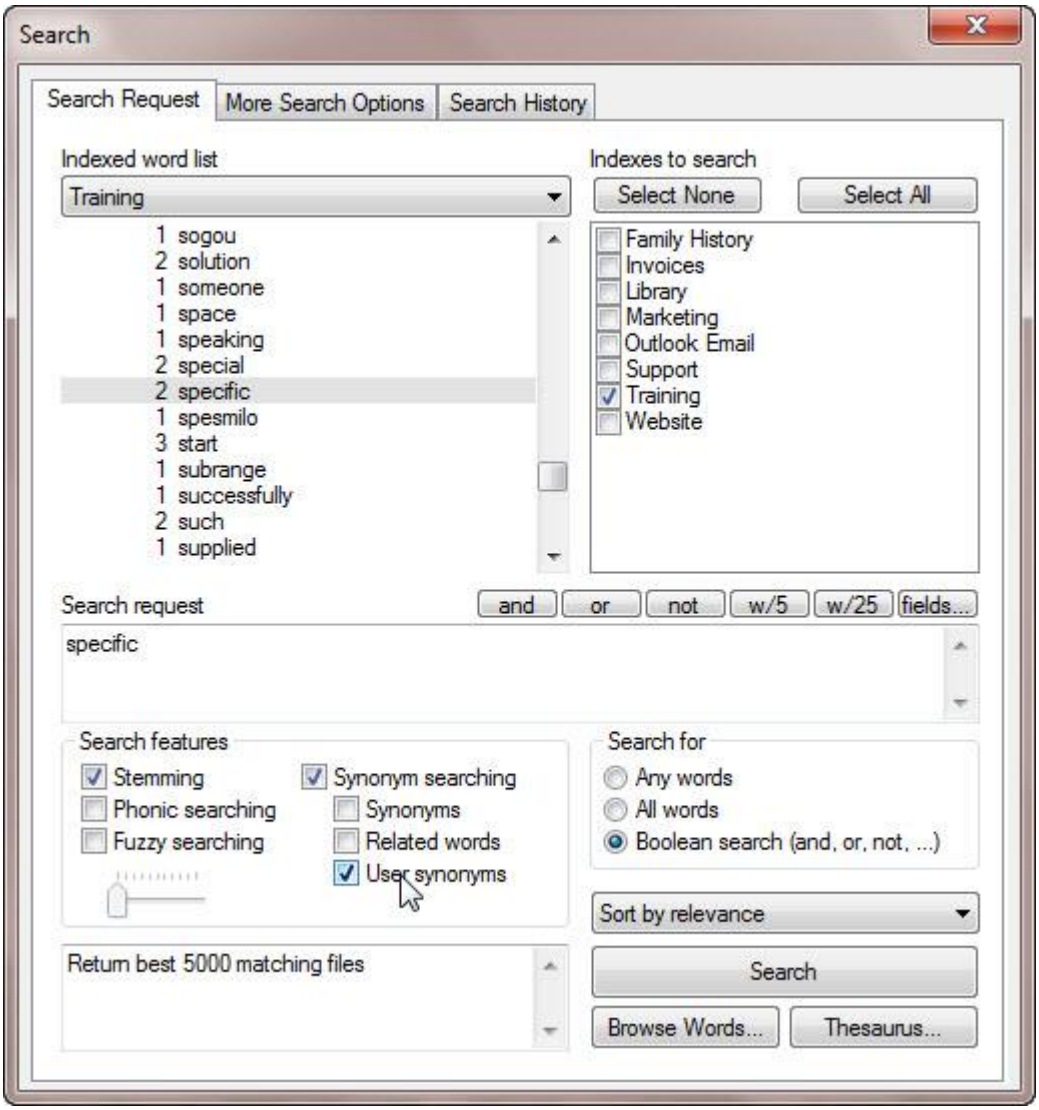
This is more complicated since, apart from editing the Alphabet file, it is necessary to edit the maximum word length that dtSearch stores in the index, and a Windows Registry edit is required. For detailed instructions see the Training article T205.

BitCoin

For detailed instructions see the Training article T205.

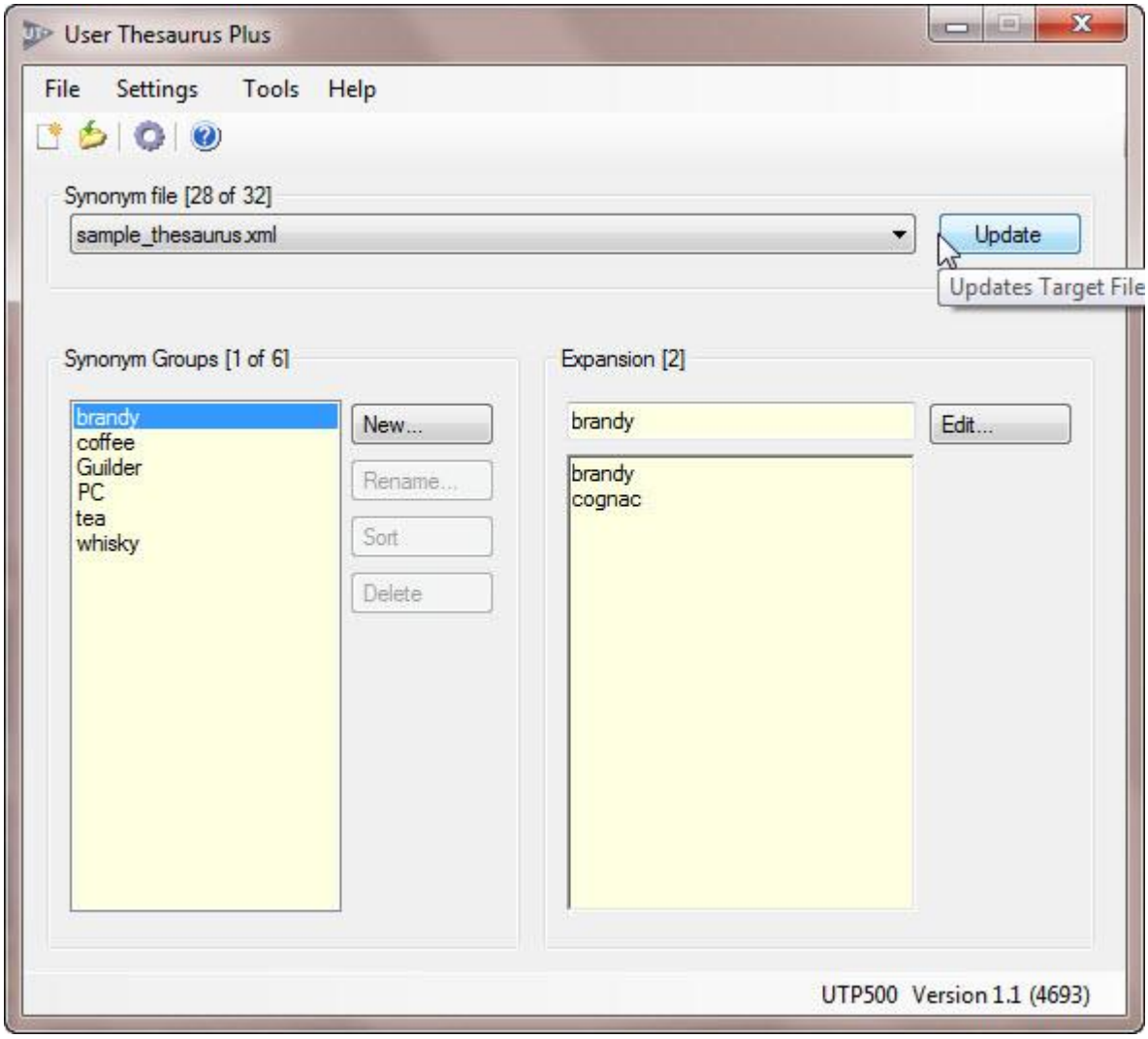
Using the Sample Files with dtSearch Desktop or Network

In the dtSearch Desktop/Network search dialog make sure **Synonym searching** and **User synonyms** are selected.



In User Thesaurus Plus select the sample file you want to work with from the drop-down list, now press the **Update** button.

You can check the user synonyms or macros that are in use in dtSearch Desktop from the **Options|Preferences** menu, choose **User thesaurus** or **Macros**

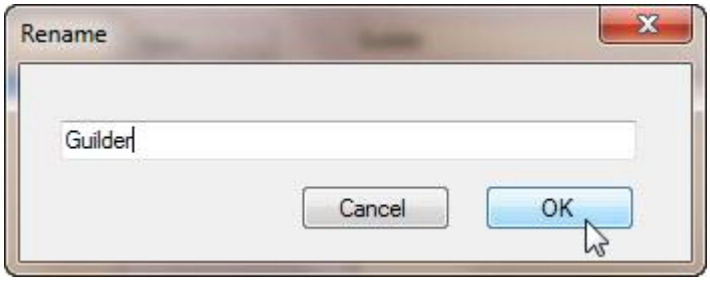


You search queries will now be automatically expanded.

Importing User Thesaurus or Macro files

You may have colleagues that have created User Thesaurus or Macro files on other machines that you may want to share. From the **File** menu choose **Import file**... to import these into your User Thesaurus Plus Data folder. User Thesaurus Plus will automatically identify the type.

A dialog will be displayed so that you can rename the thesaur.xml or macros.xml file.

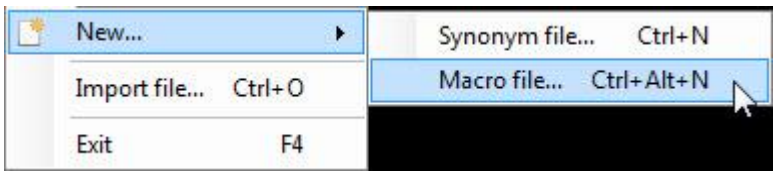


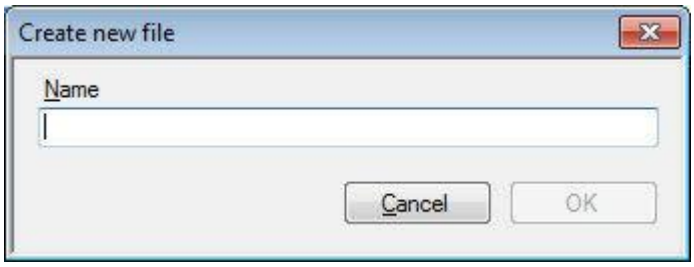
You **must** rename the file, If you attempt to save a file with a name that already exists in the Data folder it will display a message that the file has not been copied. (Do not add the .xml extension to the name).

IMPORTANT: If at any time you edit the user thesaurus in dtSearch Desktop you will need to import it back into the User Thesaurus Plus Data folder with a new name; since this may not be desirable it is better to avoid editing files in dtSearch Desktop.

Creating a New Synonym or Macro File

From the File menu select New > Synonym file... or New > Macro file... as appropriate.



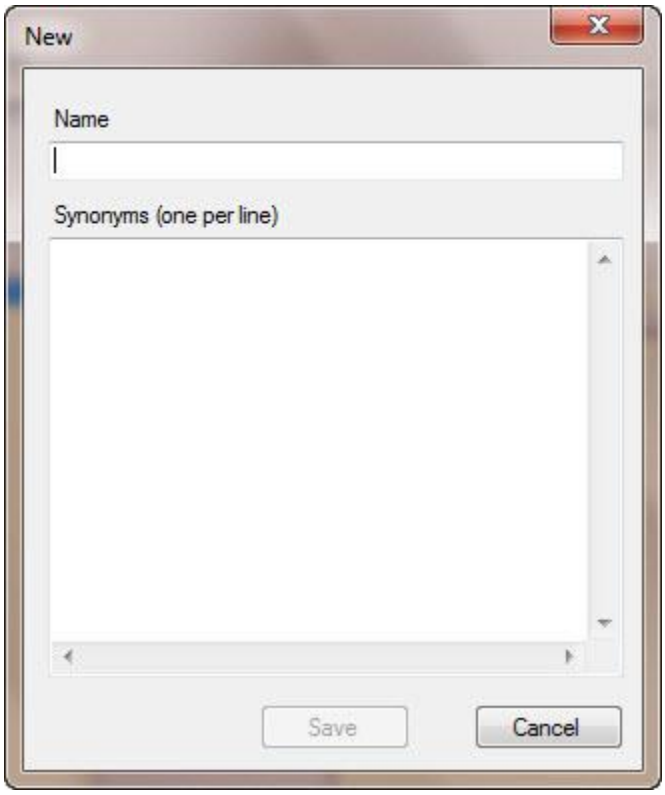


Enter a suitable name and press OK; the file will be saved in your Data folder and will appear selected in the drop-down list ready for you to enter new synonym groups or macros. If you attempt to save a file with a name that already exists, it will not copy the file but will display the existing file.

Creating, Renaming and Sorting Synonym Groups or Macros

Creating a new synonym or macro group

Click the New button.



Name
Enter a suitable group name, for dtSearch synonym and macro files the name of the group is for reference only, it will not be used in a search. The name can be up to 32 characters in length and can be more than one word, duplicates are not allowed.

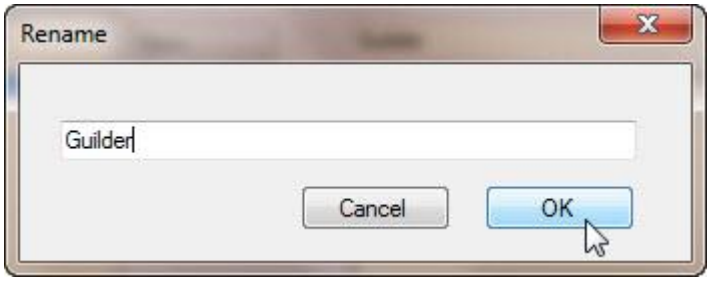
Once a name is entered the Save button will be enabled, you can save the new empty group or proceed to enter synonyms; for synonyms the group name will be copied automatically into the list of synonyms when you press the Save button.

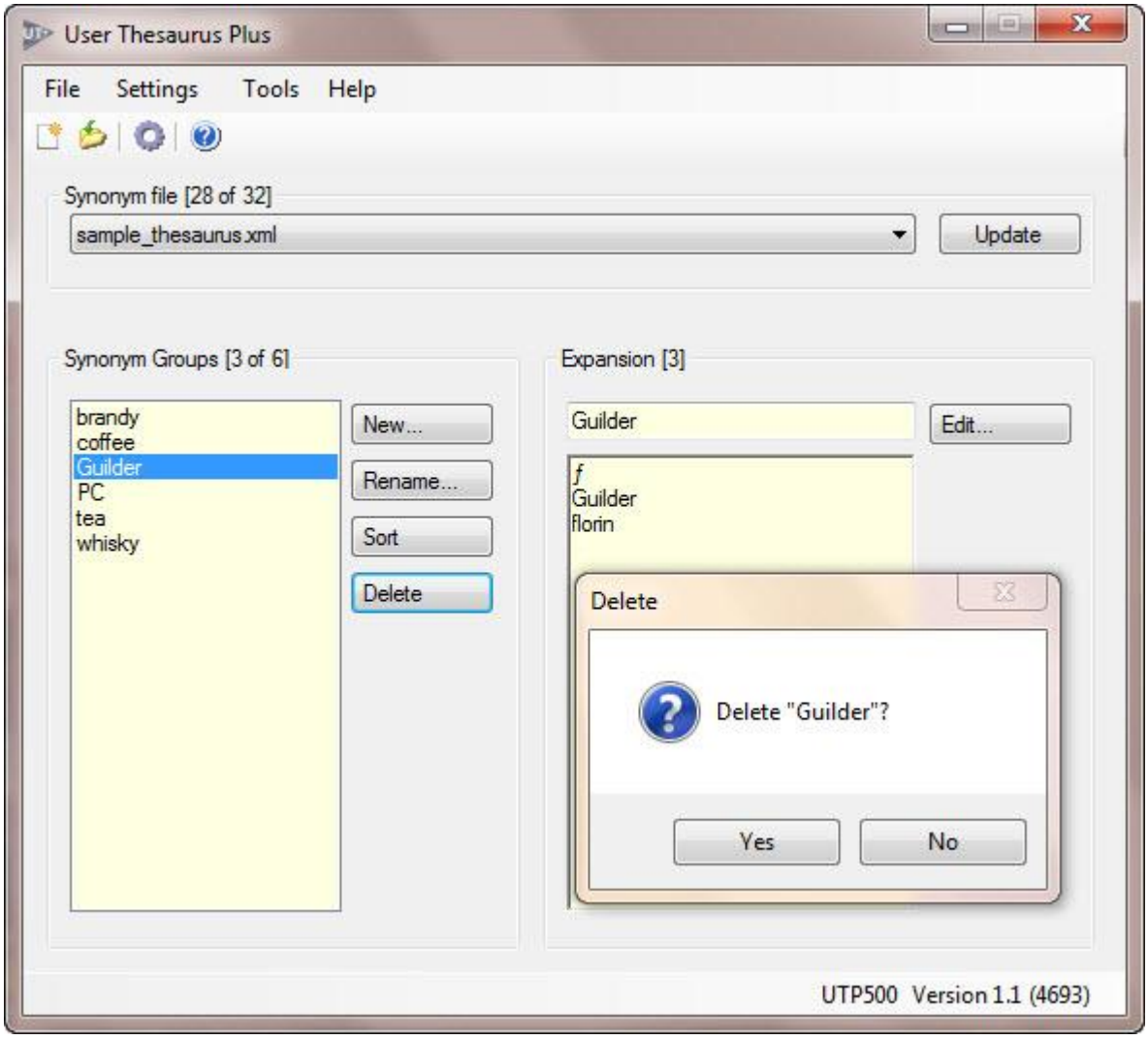
Synonyms
Press the tab key to move to the Synonyms text box, now enter one word or phrase per line. When you are finished press the tab key then click the **Save** button or press the Enter key.

For dtSearch synonym groups, when entering more than one word on a line you do not need to enter quotation marks, these will be automatically added to the saved XML file but are not displayed in the list of synonyms.

You can right-click in the text box to paste text. Duplicated entries, blank lines, any spaces before or after words, or punctuation accidentally entered in any of the lines of text will be removed automatically when you click Save.

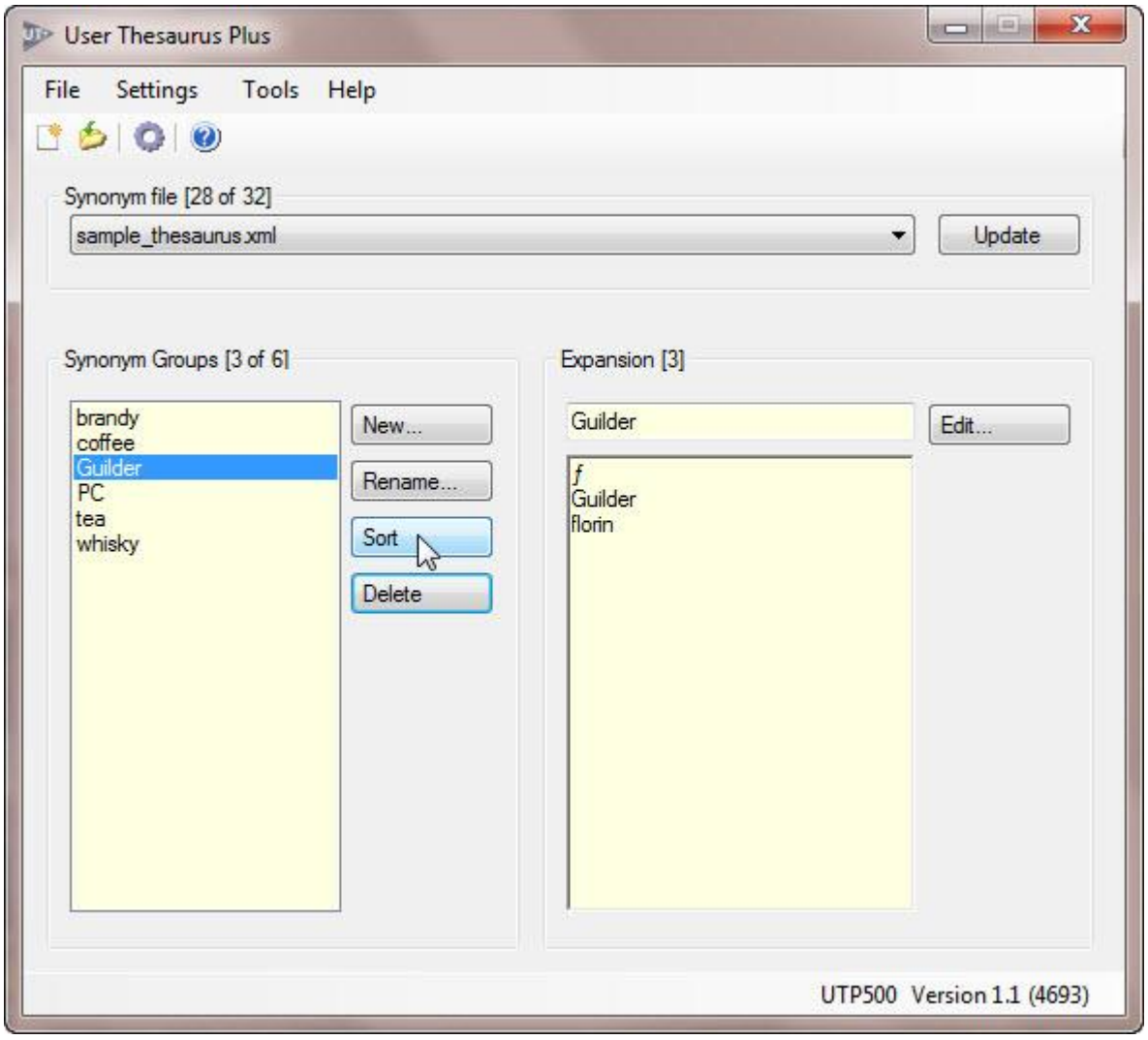
Rename or Delete a group





If you have made errors in the list of Synonym Group or Macro Names you can Rename or Delete any of them by clicking on the item in the list, then pressing the Rename or Delete button.

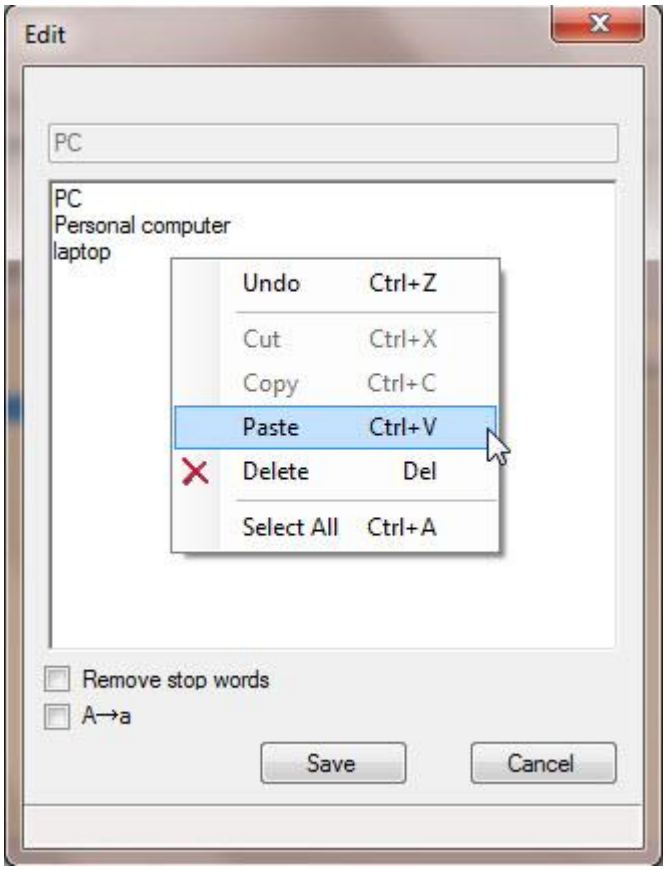
Sort a group



By default, the Sort button will sort items in the Synonyms Group numerically then alphabetically according to English alphabet rules. The rules for sorting are controlled by XSL Transform files in your **`MyDocuments\User Thesaurus Plus`** folder.

If the language is not specified in the XSL Transform file, sorting will be according to the language of the operating system, if you are sorting items in some other language you will need to edit the XSL Transform files.

Editing Synonyms or Macros



Enter each word or phrase on a separate line.

You can right-click in the text box to Copy or Paste text.

Note
Punctuation marks and characters that are by default not searchable in dtSearch are automatically blocked from being entered. Macros will allow characters needed for a dtSearch search query

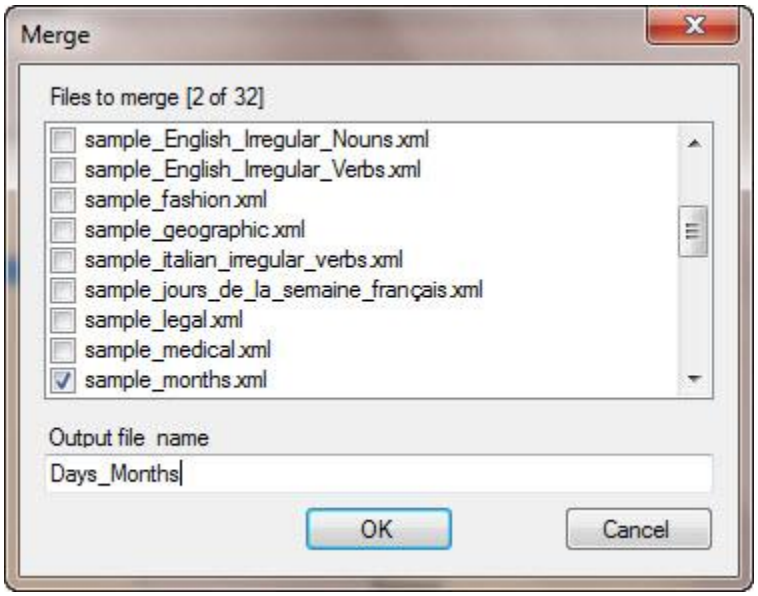
Duplicated entries, blank lines, any white space before or after words, or punctuation characters entered in any of the lines of text accidentally will be removed automatically when you click Save.

Remove stop words
You can remove any single stop words in the list by checking the 'Remove stop words' checkbox, they will be removed when you press the Save button. Any phrase containing a stop word will NOT be removed.

A->a
You can convert all the text to lower case by checking the A->a checkbox; the text will be converted when you press the Save button.

Merging Files

Select the type of file (synonyms, macros) from the **Settings** menu, select the **Tools|Merge..** menu.

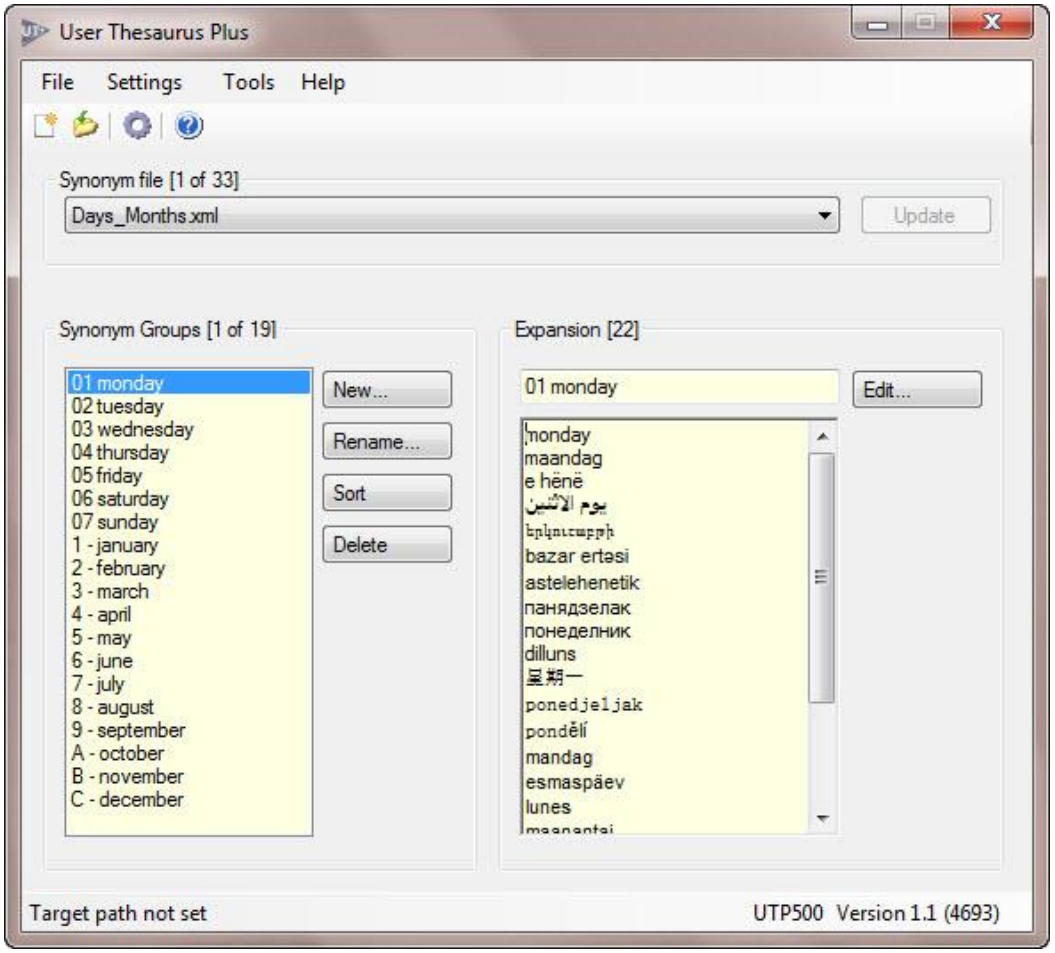


File to merge
Select any two files from the list.

Output file name
Enter a suitable output file name, without the .xml file extension.

Click the **OK** button.

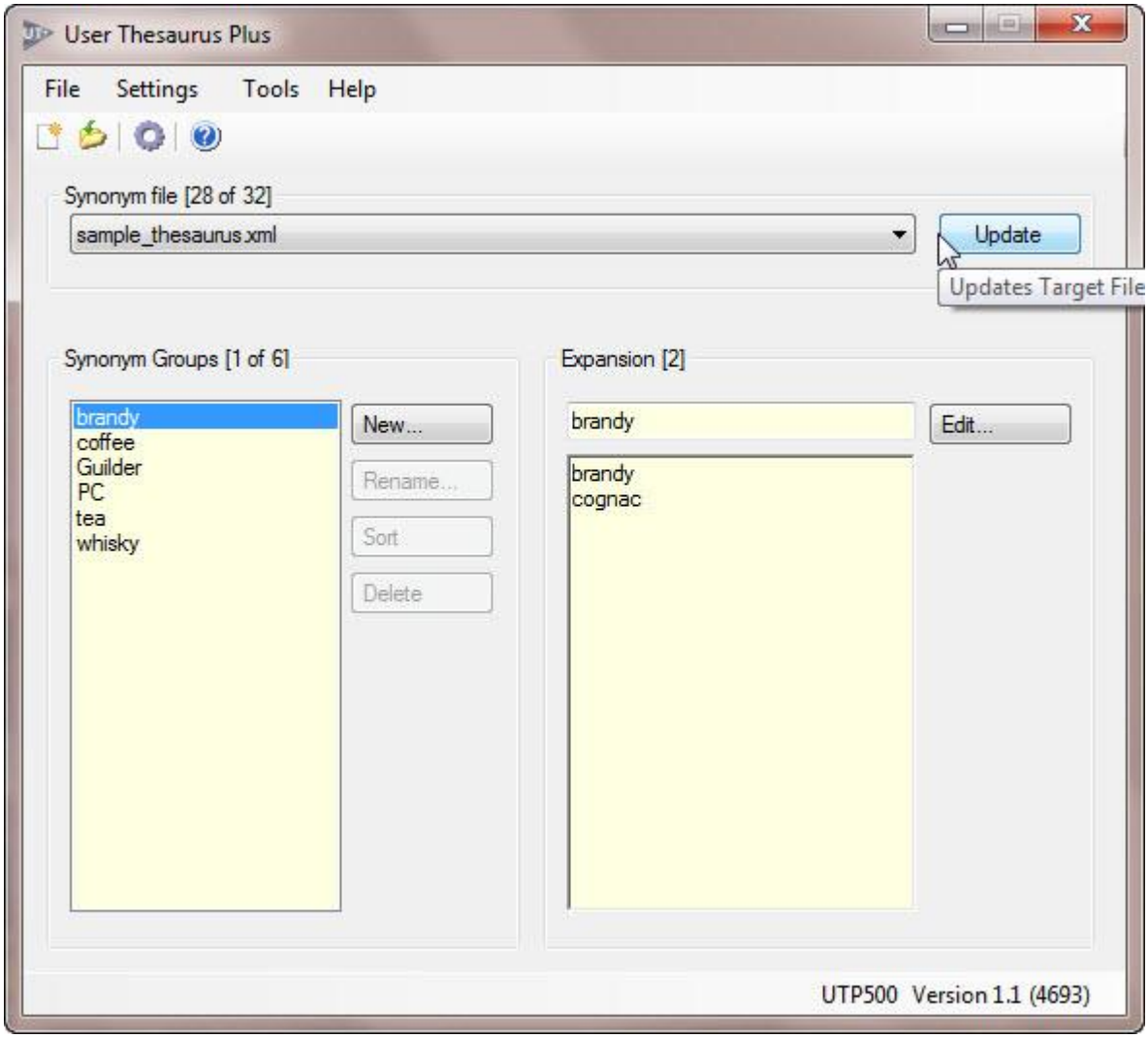
You may want to try merging the Days and Months sample files as shown, this demonstrates a technique of prefixing group names with numbers and letters to force the order in which items are sorted.



After a merge the new file is automatically selected in the drop-down list.

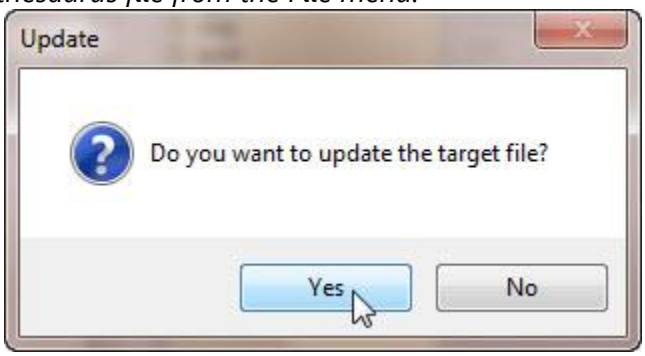
Update Target File

The target file can be updated using the **Update** button, the button is enabled each time a file is edited or sorted, or when a group has been renamed or deleted.



Each time a target file is set from the Set Target Path dialog, you will be prompted with a dialog box asking if you want to update the Target file.

WARNING: Pressing yes will overwrite your target file with the file selected in the drop-down list. If want to keep a copy of your original target thesaurus file, make a backup and keep it safe and/or import your thesaurus file from the File menu.



Noise words

Noise words, also known as Stop words, are words that are not of value in a search; common words such as **the, a, and** are typically chosen as stop words in English language search applications. To make it easy to recognise if a stop word has been entered User Thesaurus Plus will highlight them.

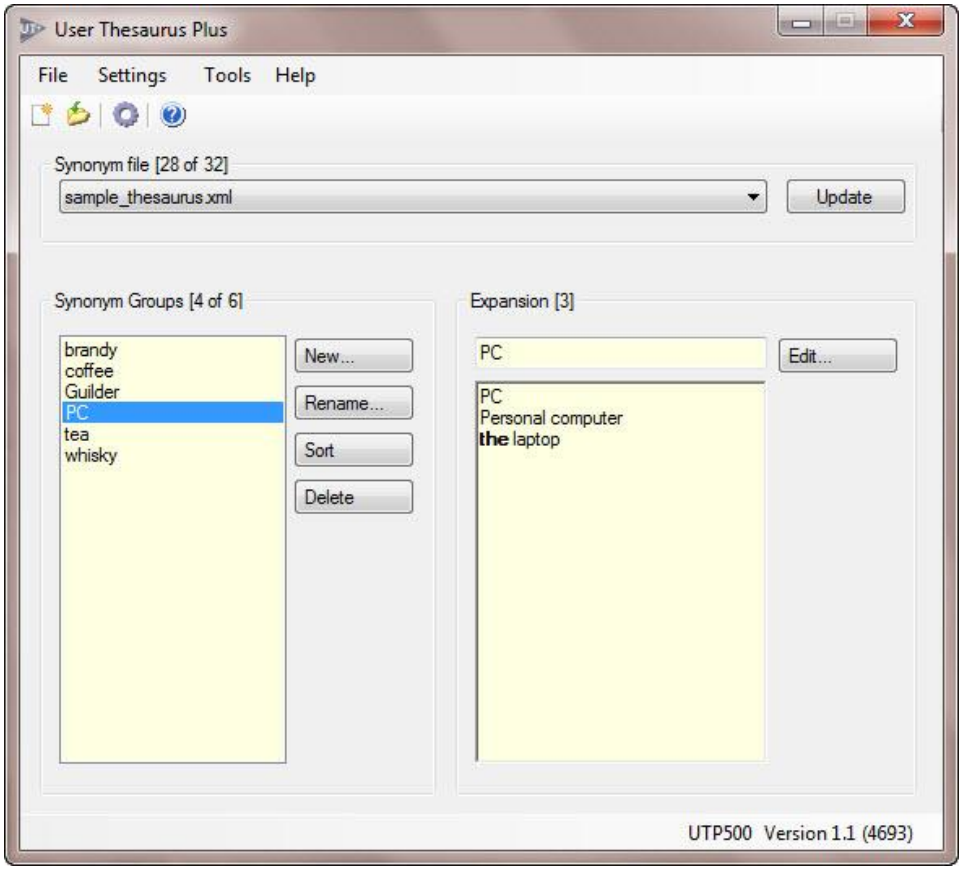
Set path to stop word file

To use the noise word highlighting feature in User Thesaurus Plus (recommended), you will need to set the path to the stop-word file from the **Settings|Stop word file...** menu, then click on the Stop Word dialog.

You can find the path from within dtSearch Desktop from the **Options|Preferences|Letters and Words** menu; you can change the file name or path by pressing the **Edit...** button, then the **Save As...** button. The default file is called noise.dat,

Noise word highlighting

User Thesaurus Plus will highlight noise words in the expansion list as shown below



Edit

If the noise words are not required as part of a Boolean search, you can manually remove them by clicking the Edit button.

If you select the **Remove Stop Words** option, any noise words that are each listed on a single line will be removed from the list when you press the Save button.

Macros

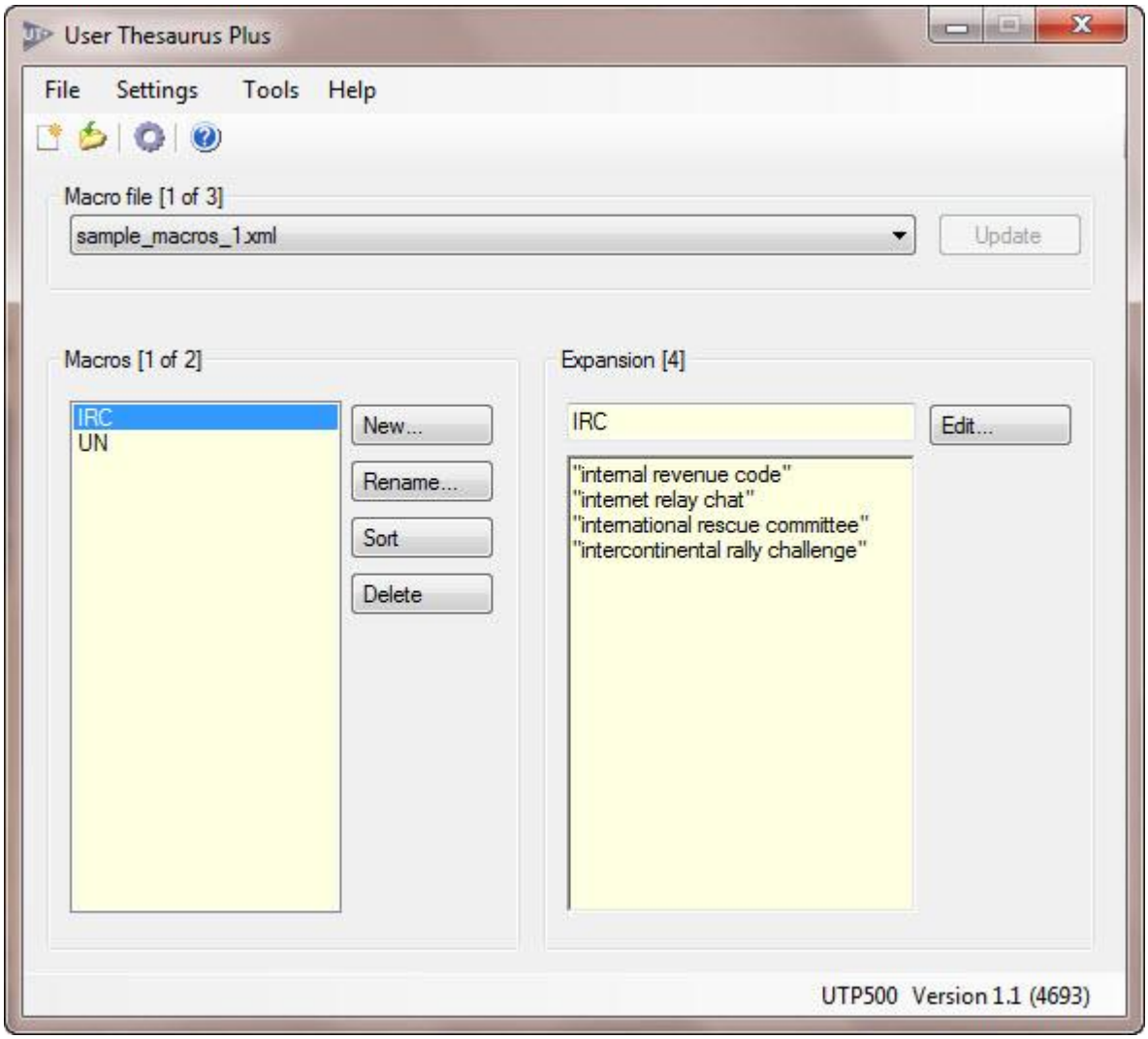
User Thesaurus Plus can be used to import and edit Macro files for use with dtSearch. These files are used when you perform a search, for example if you enter a macro such as **@IRC** in the Search dialog, dtSearch will replace that term with `internal revenue code` in the actual search query.

Macros are useful for frequently used search terms to save typing a long query. There is no need to have more than one word or phrase, unlike a synonym file where it is always necessary to have more than one word or phrase in the set of synonyms (also referred to as a synset, synonym set or synonym ring).

Note: Phrases in a Macro should be quoted to ensure that they will be searched as a phrase in an 'All Words' search.

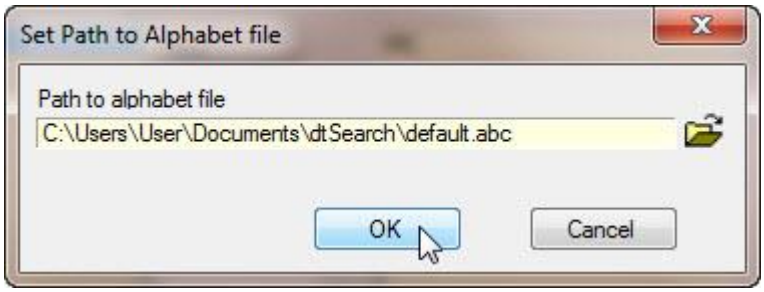
To import a macro file, from the **File** menu choose **Import file...**, User Thesaurus Plus will automatically recognise that it is a macro file and will change the user interface accordingly.

In version 1.2 from the View menu, select **Macro** or **Synonym** to change the list of files in the drop-down list.



Alphabet File Editor

The Alphabet file editor allows you to make Unicode currency symbols searchable in dtSearch. Before you can edit a dtSearch alphabet file in User Thesaurus Plus you must set the path to the alphabet file from the **Settings|Alphabet file...** menu.



You can find the path from within dtSearch Desktop from the **Options|Preferences|Letters and Words** menu; you can change the file name or path by pressing the **Edit...** button, then the **Save As...** button.

The default file is called `default.abc`, if you want to create alternative files, rename the default file, and then edit it in dtSearch Desktop; next set the path to the new alphabet file in User Thesaurus Plus and edit it to make the desired currency symbols searchable.
Note:

The symbol **$** (Dollar sign, 36 decimal, 0024 Hex) can be made searchable in dtSearch Desktop from the **Options|Preferences|Indexing Options|Letters and Words** menu, press the Edit button alongside the Alphabet file text box, select character code 36 then choose Letter instead of Space, now Save and Close the dialog box.

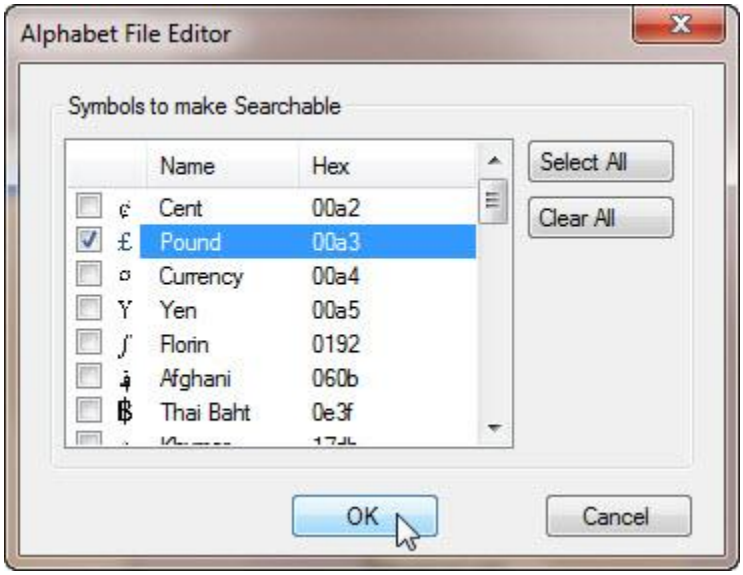
You can edit the alphabet file at any time from either dtSearch Desktop or the Alphabet File Editor in User Thesaurus Plus.

The dollar sign and a few other currency signs have alternative Unicode code points, for example if you make the **$** character searchable as above, dtSearch will not by default index the $ symbol at code points FF04 (Full width Dollar sign) and FF69 (small Dollar sign).

To make these symbols searchable, first make the dollar sign searchable as above, then in User Thesaurus Plus **Settings** menu set the Alphabet file path to **`default.abc`,** then open the Alphabet File Editor from the Tools menu and press the Save button.

The Alphabet File Editor will automatically make the alternative code points searchable for any of the selected symbols marked with an asterisk in the table below.

To edit the dtSearch alphabet file in User Thesaurus Plus select **Tools|Alphabet Editor...**



Select the currency symbols that you want to make searchable, then press the OK button. You will need to update the dtSearch index(es) before the settings take effect.

A copy of the current alphabet file will be saved with your index when you create or update an index.

Currency symbols that can be made searchable (34)

The items marked with an asterisk also make full-width character code points searchable.

Decimal code	Symbol	Name	Hex code
162	¢	CENT SIGN	00A2 *
163	£	POUND SIGN	00A3 *
164	¤	CURRENCY SIGN	00A4
165	¥	YEN SIGN (also YUAN)	00A5 *
402	ƒ	FLORIN/GUILDER SIGN	0192
1547	؋	AFGHANI SIGN	060B
3647	฿	THAI BAHT SIGN	0E3F
6107	៛	KHYMER RIEL SIGN	17DB
8353	₡	COLON SIGN	20A1
8354	₢	CRUZEIRO SIGN	20A2
8355	₣	FRENCH FRANC SIGN (present in WGL4)	20A3
8356	₤	LIRA SIGN	20A4
8357	₥	MILL SIGN	20A5
8358	₦	NAIRA SIGN	20A6
8359	₧	PESETA SIGN	20A7

8360	₨	RUPEE SIGN	20A8
8361	₩	WON SIGN	20A9*
8362	₪	NEW SHEKEL SIGN	20AA
8363	₫	DONG SIGN	20AB
8364	€	EURO SIGN	20AC
8365	₭	KIP SIGN	20AD
8366	₮	TUGRIK SIGN	20AE
8367	₯	DRACHMA SIGN	20AF
8368	₰	GERMAN PENNY SYMBOL	20B0
8369	₱	PESO SIGN	20B1
8370	₲	GUARANI SIGN	20B2
8371	₳	AUSTRAL SIGN	20B3
8372	₴	HRYVNIA SIGN	20B4
8373	₵	CEDI SIGN	20B5
8376	₸	TENGE SIGN	20B8
8377	₹	INDIAN RUPEE SIGN	20B9
8378	₺	TURKISH LIRA	20BA
8379		NORDIC MARK SIGN	20BB
8380		MANAT SIGN	20BC
8381		RUBLE SIGN	20BD
8382		LARI SIGN	20BE
8388		BITCOIN SIGN	20BF
20803	元	CJK UNIFIED YUAN	5143
65020	﷼	RIAL SIGN (IRAN)	FDFC

Search Days and Months

To include dates in dtSearch Desktop/Network from the **Menu option: Options > Preferences > Indexing Options** select **"Automatically recognize dates, email addresses and credit card numbers in text**

Date recognition in dtSearch looks for anything that appears to be a date, using English-language months (including common abbreviations) and numerical formats.

Examples of date formats that are recognized include:

```
January 11, 2010
11 Jan 10
2010/01/11
1/11/10
1-11-10
The eleventh of January, two thousand ten
```

To search for a date, put **`date()`** around the date expression or range.

For example, to find any of the expressions above near the word "apple", search for:

date(jan 11 2010) w/10 apple

Searching for Money

dtSearch Desktop/Network does not index currency signs by default, nevertheless a search for $120 will find documents containing $120 because it will ignore the $ sign; unfortunately, this means it may also find documents containing £120 or 120 € or 120 Yen or even 120 chickens! If you are searching in a small document database this may be acceptable but for many this lack of precision will be frustrating.

If you use the Alphabet File Editor to make currency signs searchable your searches will have much higher precision, but you need to be aware that you will need to adapt your search queries accordingly. For example, if you make the Dollar sign a searchable letter a search for $2000000 will only find documents containing $2000000, you will not find a document that contains $ 2000000 (i.e. with a space between the sign and the digits). To ensure that you will find documents with or without a space between the currency sign and the amount, you will need to change your search query to:

($ w/1 2000000) or $2000000

By also using the currencies sample thesaurus file and other techniques it is possible to improve both the precision and the recall when searching on monetary amounts, for example it is possible to find documents containing amounts in various currencies when they are expressed in these forms:

$ 2 million (including if the $ sign is the full-width or small character variants)
$ 2 m
$2000000
$ 2000000
2000000 Dollars
2000000 USD
USD 2000000
$ 2,000,000
2,000,000 dollars
two million dollars

See Training article T201 “Search for money” for more information.

Search on Names

See Training article T209 “Searching for people”

Installed files and folders

UserThesaurusPlus.exe will normally be installed in the Program Files(x86) folder under DTSUK\UTP100. The installer will also create a set of folders in the Documents folder as shown below.

Folder Name	Contents	Description
User Thesaurus Plus	UTP100_thes.xlt UTP100_macros.xlt	Used for sorting the contents of the XML files
User Thesaurus Plus\Data	UT_sample_colours.xml UT_sample_currencies.xml UT_sample_danish_irregular_nouns.xml UT_sample_danish_irregular_verbs.xml UT_sample_days.xml UT_sample_dutch_irregular_nouns.xml UT_sample_dutch_irregular_verbs.xml UT_sample_English_irregular_nouns.xml UT_sample_English_irregular_verbs.xml UT_sample_fashion.xml UT_sample_geographic.xml UT_sample_italian_irregular_verbs.xml UT_sample_jours_de_la_semaine_francaise.xml UT_sample_legal.xml UT_sample_medical.xml UT_sample_months.xml UT_sample_names_cross_lingual.xml UT_sample_names_genealogy.xml UT_sample_names_political.xml UT_sample_names_russian_female.xml UT_sample_names_russian_male.xml UT_sample_norwegian_irregular_nouns.xml UT_sample_norwegian_irregular_verbs.xml UT_sample_prenoms_francaise_masculin.xml UT_sample_spanish_irregular_verbs.xml UT_sample_swedish_irregular_nouns.xml UT_sample_swedish_irregular_verbs.xml UT_sample_thesaurus.xml UT_sample_trade.xml UT_sample_unregelmäßige_deutsche_Verben UT_sample_verbes_irréguliers_français.xml UT_sample_Wochentage_deutsch.xml (index.xml)	Thesaurus files that will be displayed in the drop-down list in User Thesaurus Plus. A file generated during merging, do not remove.

User Thesaurus Plus\Test	UTT_combine.txt UTT_sample_English_Irregular_Verbs.txt UTT_sample_fashion.txt UTT_sample_legal.txt UTT_sample_macros.txt UTT_sample_medical.txt UTT_sample_names_cross_lingual.txt UTT_sample_names_genealogy.txt UTT_sample_names_political.txt UTT_sample_names_trade.txt UTT_sample_thesaurus.txt	Test files for checking that each file is working correctly.
User Thesaurus Plus\Data\Macros	UT_sample_macros_1.xml UT_sample_macros_2.xml UT_sample_macros_3.xml UT_sample_macros_4.xml	Macro files that will be displayed in the drop-down list in User Thesaurus Plus.
User Thesaurus Plus\Data\temp	empty	

dtSearch User Thesaurus File Format

The file generated by dtSearch Desktop/Network is always called thesaur.xml and is stored in the Users private directory. The file encoding is UTF-8.

Whenever you select a synonym file from the drop-down list in User Thesaurus Plus and then press the Update button it generates a thesaur.xml file and copies it to the target path you have set; normally this will be in the dtSearch Users private directory as above.

All files imported into User Thesaurus Plus are prefixed with UT, example UT_sample_thesaurus.xml and saved in your My Documents folder under User Thesaurus Plus\Data.

Example: sample_thesaur.xml

```
<?xml version="1.0" encoding="UTF-8" ?>
<!-- Generated by User Thesaurus Plus : 2012-11-30 01:00:00 -->
<dtSearchUserThesaurus>
<Item>
<Name>Personal computer</Name>
<Synonyms>"Personal computer" PC laptop</Synonyms>
</Item>
<Item>
<Name>how much</Name>
<Synonyms>"how much is" "what's the price of" "what's the cost of" "how much
does"</Synonyms>
</Item>
<Item>
<Name>Guilder</Name>
<Synonyms>ƒ Guilder florin</Synonyms>
</Item>
<Item>
<Name>sing</Name>
<Synonyms>sing sang sung</Synonyms>
</Item>
</dtSearchUserThesaurus>
```

Search Queries are only expanded with the synonyms from the user thesaurus when the User Thesaurus checkbox on the Search dialog is selected. Because dtSearch Desktop/Network is supplied with a single thesaurus file this can make searches unnecessarily broad, harming the precision.

The User Thesaurus Plus utility makes it possible to have multiple smaller files that can be designed for very specific search tasks, thus enabling better recall while maintaining a high precision.

The is no hard limit to the size of each XML file, but performance will fall off as the size increases, a recommended maximum is 10,000 items per file.

Files generated in User Thesaurus Plus when using **File|New > synonym file...** contain a time stamp to distinguish them from files generated by dtSearch.

Example:

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- Generated by User Thesaurus Plus : 2012-11-30 01:00:00 --
><dtSearchUserThesaurus />
```

Macro file format

The file generated by dtSearch Desktop/Network is always called macros.xml and is stored in the Users private directory. The file encoding is UTF-8.

When you select a macro file from the drop-down list in User Thesaurus Plus and then press the Update button it generates a macros.xml file and copies it to the target path you have set; normally this will be in the dtSearch Users private directory as above.

All macro files imported into User Thesaurus Plus are prefixed with UT, Example UT_sample_macros.xml and saved in your My Documents folder under User Thesaurus Plus\Data\Macros.

Example: Macros which contain more than one line are saved within CDATA tags as below:

```
UT_sample_macros_1.xml file:
<?xml version="1.0" encoding="utf-8"?>
<!-- Generated by User Thesaurus Plus : 2012-11-30 01:00:00 -->
<dtSearchMacros>
  <Item>
    <Name>IRC</Name>
    <Expansion><![CDATA["internal revenue code"
"internet relay chat"
"international rescue committee"
"intercontinental rally challenge"]]></Expansion>
  </Item>
  <Item>
    <Name>UN</Name>
    <Expansion>"United Nations"</Expansion>
  </Item>
</dtSearchMacros>
```

Macros which contain only one line are saved as below:

```
UT_sample_macros_2.xml file:
<?xml version="1.0" encoding="utf-8"?>
<!-- Generated by User Thesaurus Plus : 2012-11-30 01:00:00 -->
<dtSearchMacros>
  <Item>
    <Name>IRC</Name>
    <Expansion>"internal revenue code"</Expansion>
  </Item>
  <Item>
    <Name>UN</Name>
    <Expansion>"United Nations"</Expansion>
  </Item>
</dtSearchMacros>
```

Files generated in User Thesaurus Plus when using **File|New > Macro file...** contain a time stamp to distinguish them from files generated by dtSearch.

Example:

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- Generated by User Thesaurus Plus : 2012-11-30 01:00:00 -->
<dtSearchMacros />
```

Testing dtSearch Desktop User Thesaurus & Macros

Test files are supplied so that you can check that each file is working correctly.

To carry out a test in dtSearch Desktop first create a new index, then in the update index dialog choose **Add Folder**... browse to your My Documents\User Thesaurus Plus folder and open the Test folder.

In the Filename filters box add ***.txt,** now click the **Start button.**

When the indexing is complete, open the Search dialog and select just the index you have created. Check the **Synonym searching** and **User Synonyms** boxes and unselect all other search expansion checkboxes (i.e., Stemming, Phonic, Fuzzy, Synonyms, Related words).

Enter any of the words from the indexed word list; you should find that whatever word you search with dtSearch will open **UTT_combine.txt** and the test file that matches the word in your search request.

Now search again using a word or phrase that appears in the synonym file that you have selected in User Thesaurus Plus, all words on the same line should be highlighted*.

Test again with the **User Synonyms** checkbox unselected, you should find that dtSearch will find just the word you searched on and no other words will be highlighted.

Note: if you search on a single word that is part of a phrase it will not be highlighted unless you enter the complete phrase.

Combine.bat

The Test folder contains a file `combine.bat`, this is a batch file that combines all the .txt files in the folder into a single file UTT_combine.txt. If you change the text files in the Data folder and need to create a new UTT_combine.txt, first delete UTT_combine.txt then double-click on combine.bat to create a new UTT_combine.txt file.

Months and Days Sample files

Months
This file is designed to demonstrate a typical cross-lingual information retrieval (CLIR) application. The sample file has group names for all the months in English arranged in chronological order, obviously you are unlikely to want to sort the month names alphabetically, so the months have been prefixed with numbers and letters 1-9, A-C (so that if the sort button was clicked accidentally or deliberately, they get sorted numerically to match the chronological order. When you use this file with dtSearch, you will be able to search for a month in any of over 20 languages* and will find documents referring to that month in any of those languages.

A problem to be aware of is that if you are searching for month 11, November in English, documents that are returned containing **listopad** will be correct for Polish, Czech, and archaic Slovene (the Cyrillic equivalent is лістапада in Belarusian, листопаді in Ukrainian) but in Croatian be aware that **listopad** is the month of October. A similar problem exists with month 12 - December, the word prosinec (Czech) or prosinac (Croatian) refers to December, but prosinec may be found referring to the month of January in some archaic Slovene documents.

Finally note the similar words sierpień (Polish) and srpen (Czech) is month 8 – August, but srpanj is July in Croatian and veliki srpan ("large sickle") is August and mali srpan ("small sickle") is July in archaic Slovene.

References:

1 *http://en.wikipedia.org/wiki/Slovene_months*
2 *https://digital.lib.washington.edu/ojs/index.php/ssj/article/viewFile/4179/3518*

Days
Days of the week in over 20 languages**. In a similar fashion to the Months file, the days of the week have been prefixed with numbers so that they will sorted correctly, in addition they have been prefixed with a zero so that they will appear in order if you merge the file with the Months sample file. The order is in accordance with international standard ISO 8601 with Monday as the first day of the week, however many countries such as the USA still have their calendars refer to Sunday as the first day of the week. You can Rename the list to suit the order and language you prefer.

* Languages: Arabic, Belarusian, Bosnian, Bulgarian, Croatian, Czech, Chinese, Danish, Dutch, English, Estonian, Finnish, French, German, Greece, Hungarian, Italian, Japanese, Latvian, Lithuanian, Norwegian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Turkish, Ukrainian, Uzbek, Welsh. (Note: for Chinese and Japanese the month name is a 'numbered month' in traditional Chinese; for Arabic there are variants for names as used in Algeria, Egypt, Iran, Iraq, Jordan, Lebanon, Palestine, Sudan, Syria).

** Languages: Afrikaans, Albanian, Arabic, Armenian, Azerbaijani, Basque, Belarusian, Bengali, Bulgarian, Catalan, Chinese, Croatian, Czech, Danish, Dutch, English, Estonian, Filipino, Finish, French, Galacian, Georgian, German, Greek.

Currencies

The currencies sample file provides 162 currency codes (ISO 4217 current and some not current) and some additional commercially used codes, with synonyms of currency name and currency sign, so that a search for 100 EUR will find 100 euro or 100 €.

The currencies sample also includes full width currency sign variants for the Dollar, Cent, Pound, Yuan (Yen) and Won.

It should be noted that by default dtSearch does not index currency symbols, to be able to index and search currency symbols it is necessary to edit the alphabet file. The editor built into dtSearch Desktop controls the processing of characters in the range from 33 to 127 only, the dollar sign ($, character code 36) can be made searchable by selecting the Character type 'Letter' instead of the default 'Space'; characters above 127 are processed according to the Unicode specification and dtSearch does not treat other symbols as searchable characters, however it is possible to manually edit the alphabet file (default.abc) to make additional Unicode characters searchable, an easier and faster method is to use the Alphabet file Editor in User Thesaurus Plus.

Note, the WordNet thesaurus built into dtSearch Desktop has no entry for euro, but does have entries for dollar, pound, drachma, dirham and others but contains no symbols or ISO codes.

According to the European Union's Publication Office, in English, Irish, Latvian and Maltese texts, the ISO 4217 code is followed by a fixed space and the amount: *a sum of EUR 30*; In Bulgarian, Czech, Danish, Dutch, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Lithuanian, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish and Swedish the order is reversed; the amount is followed by a fixed space and the ISO 4217 code: *une somme de 30 EUR*

The above should be taken with a pinch of salt, since individuals within each country do not necessarily write in that format, to ensure a search for EUR 100 also finds 100 EUR you should search using **EUR w/1 100** (i.e., "EUR within one word of 100").

Reference
http://en.wikipedia.org/wiki/ISO_4217

Irregular verbs

This thesaurus file enables dtSearch to handle verb forms (irregular verbs, strong verbs, stem change verbs) that stemming may not, for example in English go, went, gone; run, ran; speak, spoke; drink, drank.

Irregular verbs listed in the sample files include those which may be listed as strong verbs and stem change verbs in various grammatical texts.

Not all conjugations for each verb may be listed, if a verb follows regular verb patterns in most of the forms the dtSearch stemmer may handle those correctly without the need to have them included in the thesaurus, listing all verb conjugations may slow searches.

Language	**Number of verbs**
Danish	22
Dutch	224
English	233
French	106
German	183
Italian	18
Norwegian	19
Spanish	31
Swedish	112

Irregular Nouns

This thesaurus file will enable dtSearch to handle plurals of words that stemming rules may not, for example in English woman - women, foot - feet, goose - geese, child - children.

The normal plural of English nouns ends in -(e)s, for example dog, dogs; bush, bushes. Nouns which end in a consonant plus -y change the -y to -ies, for example pony, ponies; lady, ladies. Finally, some nouns that end in -f(e) change the -f to -ve in the plural, for example calf, calves; knife, knives. When searching for such nouns in dtSearch always keep English stemming turned on to ensure that a search will always find documents containing either the singular or plural form of the noun.

English
There is a small group of nouns for which the plural form involves a vowel change rather than a change in the word ending, dtSearch stemming alone will not find these, some common words with irregular plurals of this kind are: man, men; foot, feet; tooth, teeth.

Danish, Dutch, Norwegian and Swedish
Just as with English and other Germanic languages* the Danish, Norwegian and Swedish plural indefinite forms of the words man (mande, mann, man), foot (fod, fot, fot) and tooth (tand, tann, tand) are irregular. (Note that Dutch plural forms of these words use a regular -en suffix to form the plural: man - maanen; voet - voeten, tand - tanden).

dtSearch Danish, Dutch, Norwegian and Swedish stemming rules will handle most regular plurals and some irregular forms. The thesaurus sample noun files will improve the recall of irregular nouns in these languages.

Language	Number of nouns in sample file
English	80
Danish	30
Dutch	39
Norwegian	37
Swedish	31

* en.wikipedia.org/wiki/Germanic_languages

License Agreement for the
User Thesaurus Plus ("Licensed Software") for dtSearch.

By installing this software, you (the Licensee) agree to the terms of this License Agreement, if you do not agree then do not proceed with the installation. "Licensor" is ElectronArt Design Ltd trading as dtSearch UK. Tel: 0845 299 7307 Fax: +44 (0)207 900 6021 https://www.dtsearch.co.uk

License

1. In the absence of a signed written agreement with dtSearch UK, or if purchased as part of the Language Extension Pack LEP500 series, this Software is licensed for use by a single individual. The single individual may use the software on up to two computers with dtSearch Desktop, so long as the dtSearch Desktop application will be used on both computers exclusively by that individual. This product is an end-user application only and may not be distributed with any other application that uses the dtSearch developer API. If wider distribution is needed, please upgrade to a Language Extension Pack license.

If purchased with a Language Extension Pack LEP500 series, the license supplied with the Language Extension Pack will apply, including any distribution with an application incorporating the dtSearch Text Retrieval Engine under a license issued by dtSearch Corp.

2.1 ` The Licensed Software is protected by copyright laws and other international treaties, as well as other intellectual property laws and treaties. The Licensed Software is licensed, not sold, according to the terms of this License Agreement.

2.2 Licensor grants, and Licensee hereby accepts, an irrevocable, non-exclusive, and world-wide license to use the Licensed Software.

2.3 Except as expressly provided herein, Licensee may not distribute, copy, reproduce, sub-license, sell or otherwise transfer the Licensed Software. Licensee may make backup copies of the Licensed Software strictly for Licensee's archival purposes.

2.4 Licensee may not modify, adapt, translate or create derivative works based on the Licensed Software, with the exception of those Re-distributable files identified as being Modifiable, and only to the extent specified in the documentation accompanying the Licensed Software. Any copies that Licensee is permitted to make pursuant to this License Agreement must contain the same copyright and other proprietary notices that appear on or in the Licensed Software.

2.5 Licensee may not assign or otherwise transfer its rights under this agreement without written permission from the Licensor, which permission shall not be unreasonably withheld, except in the case of a sale of substantially all assets of Licensee's company or the merger or acquisition of Licensee's company into a new company, in which case Licensee may transfer its rights under this agreement to the acquiring party.

3. Maintenance and Support

3.1 Licensor agrees to make available to Licensee any maintenance releases, new and enhanced versions, or upgrades of the Licensed Software for a period of one year.

3.2 Licensor agrees to provide technical support via email to a single point of contact designated by the Licensee for a period of one year from the date of this license.

4 Limitations on Warranty and Liability.

4.1 The Licensed Software is provided AS IS. To the extent permitted by applicable law, ANY AND ALL OTHER REPRESENTATIONS AND WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NON-INFRINGEMENT, except as stated in section 4.4, below, ARE EXPRESSLY EXCLUDED AND DISCLAIMED. LICENSOR AND ITS SUPPLIERS DO NOT AND CANNOT WARRANT THE PERFORMANCE OR RESULTS LICENSEE OR LICENSEE'S END-USERS MAY OBTAIN BY USING THE SOFTWARE.

4.2 IN NO EVENT SHALL LICENSOR BE LIABLE FOR INCIDENTAL OR CONSEQUENTIAL DAMAGES, including lost profits, lost savings, lost opportunities or other incidental or consequential damages arising out of the use of or inability to use the Licensed Software, even if Licensor has been advised of the possibility of such damages.

4.3 UNDER NO CIRCUMSTANCE MAY LICENSOR'S LIABILITY TO LICENSEE, UNDER ANY AND ALL PROVISIONS OF THIS AGREEMENT, EXCEED THE LICENSE FEE.

4.4 To the extent that Licensor remains legally liable to Licensee, such liability shall expire one year from the date of this Agreement.

4.5 Subject to the limitations on liability contained in sections 4.1, 4.2, 4.3, and 4.4 Licensor represents and warrants: that it owns the Licensed Software or has requisite authority to enter into this transaction; that the Licensed Software, when properly used as contemplated herein, will not infringe or misappropriate any copyright, trademark, or the trade secrets of any third persons; and that to the best of Licensor's knowledge, the Licensed Software does not infringe any patents of third persons.

General Terms.

5.1 The failure to immediately enforce any provisions, rights or remedies under this contract shall not constitute a waiver by the party failing to enforce such provision, even if the party failing to enforce such provisions, rights or remedies is aware of the other party's contractual breach.

5.2 This agreement shall be interpreted under English Law.

5.3 Either party may terminate this agreement upon written notice if the other party materially violates any provision of this agreement and fails to remedy such violation within twenty-eight (28) days of receipt of a second written notice thereof, if such violation has not already been remedied following twenty-eight (28) days from receipt of a first written notice thereof. Evidence of delivery of such written notice, by recorded delivery for instance, is required.

5.4 This agreement contains the entire agreement between the parties, superseding all previous agreements. This agreement may not be amended other than by a written agreement.

5.5 If any portion of this agreement is found to be invalid, the remainder shall continue in force.

Get Help

If you need assistance with setting up User Thesaurus Plus, please contact **support@dtsearch.co.uk** or use the **Contact Us** form on the About menu or in the footer section of the dtSearch UK website. Quote your serial number or date of purchase if known.

Feedback

Please give us feedback on this user guide so that we can provide content that is useful and helpful. Thank you!

All suggestions for improvement of the product and this manual are very welcome!