

myCAT

Text Aligner User Manual

1. General Description

myCAT is a computer-assisted translation tool which includes several tools: a Text Aligner, a Quote Detector and a Self-Quote Detector. This document is the User Manual for the Text Aligner.

The Text Aligner is a search engine with a bi-text alignment function:

- It allows searching for terms or expressions in full-text mode within existing document pairs (bi-texts); and
- It shows how terms were translated in their complete context by automatically aligning the source and target versions and highlighting the relevant text part.

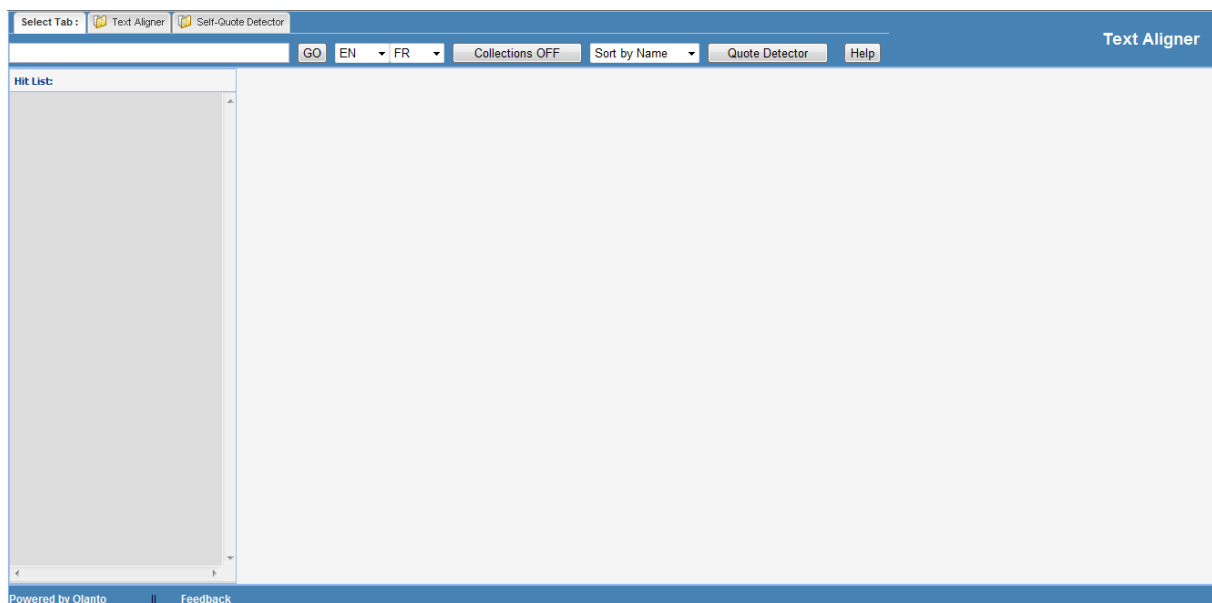
Thus the Text Aligner is a tool which facilitates the re-use of terminology found in the full text of previously-translated documents.

2. Accessing the Text Aligner

The Text Aligner is a web-based tool: it can be accessed like any web site by entering the relevant URL (*i.e.* Internet address) in the navigation bar of a browser.

Supported Browsers: All of myCAT's tools support Internet Explorer 8 or later and Mozilla Firefox 3.6 or later.

The home page of the Text Aligner looks like this :

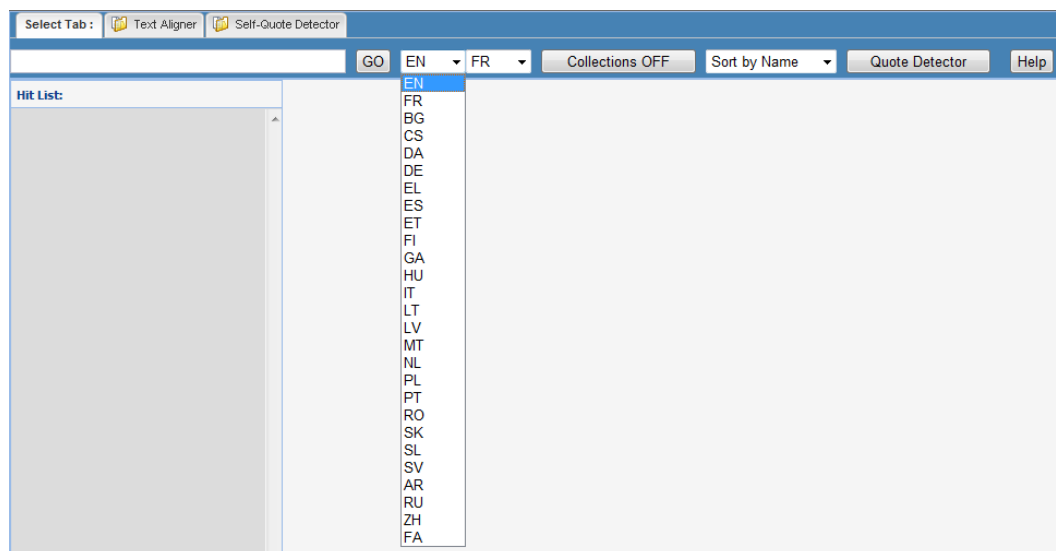


3. Using the Text Aligner

The Text Aligner is extremely simple to use: The user defines the source and target languages, possibly a specific collection of documents and enters the term(s) to be looked for.

3.1. Defining the Language Pair

First define your source and target languages from the scroll-down menus at the top of the screen:



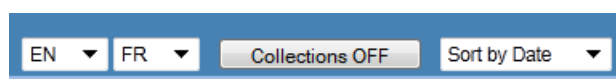
The languages are indicated in the ISO-639-1 code: EN: English ES: Spanish FR: French, etc.

The complete list of ISO-639-1 codes can be found here:

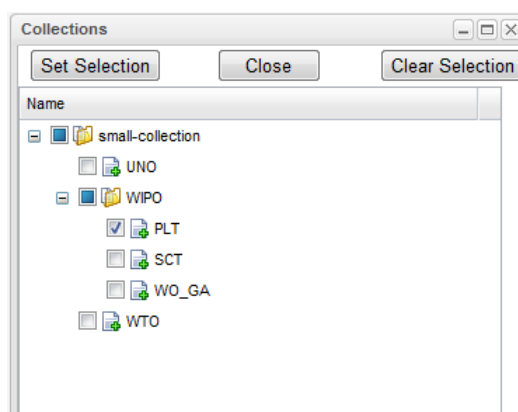
http://en.wikipedia.org/wiki/List_of_ISO_639-1_codes

3.2. Choosing Collections

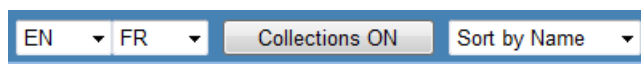
The terms you are looking for may be very specific to a given type of documents, for example meeting reports of a particular type. You can limit the search to one or several collections of reference documents by clicking on the button called "Collections OFF" at the top of the screen:



This button opens the Collection box; you can then click on one or several collections or sub-collections, for example "WIPO/PLT":



Validate your choice by clicking on the “Set Selection” button. The Collections’ button now indicates “Collections ON”:



Some collections may bear a + sign on their left side: it means that a number of sub-collections are available. You can expand the list of sub-collections by clicking on the + sign.

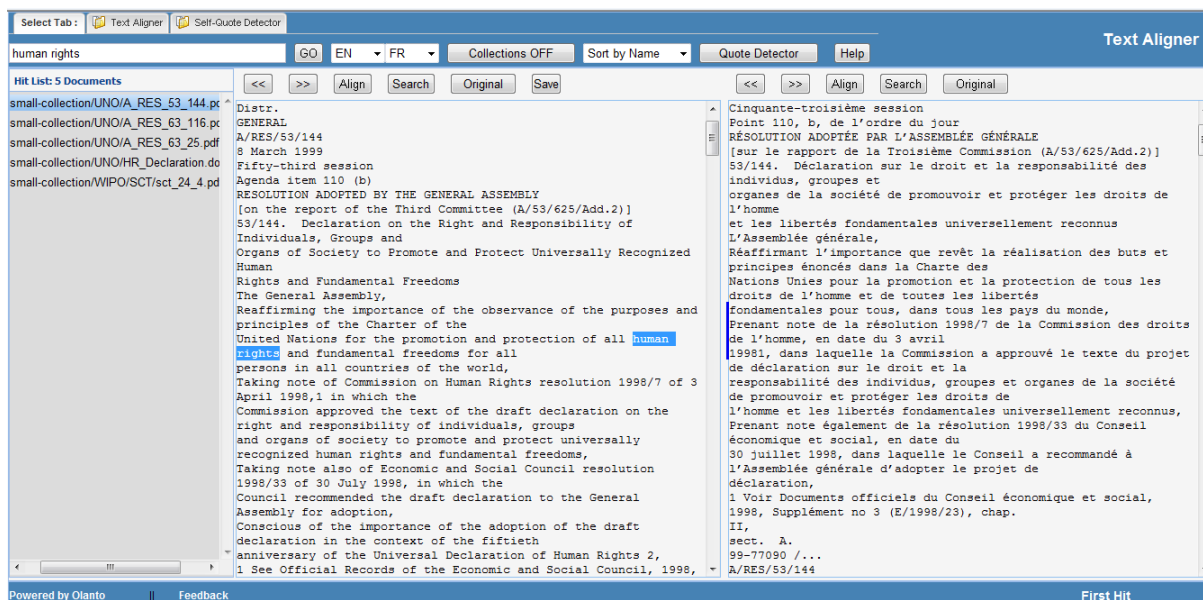
If you click on “Close” the Collection box will be closed but your selection will not be saved. If you click on “Clear Selection” all the ticked boxes will be cleared so you can make other choices.

It is possible to set a priority order in the chosen collections: if you would like to see the results of a particular collection appearing before the others, just click on that collection first. Example: If you want the “Formation” sub-collection results to be displayed first, and then the results of the “Marketing” sub-collection, simply click those two collections in that order.

- Note about personal preferences: After you chose a given language pair and (possibly) specific collections, those preferences are automatically saved in a cookie (*i.e.* a small file) on your computer, so the next time you connect to the Text Aligner your preferences will be remembered. If you change those preferences later on, the new preferences will automatically replace the old ones.

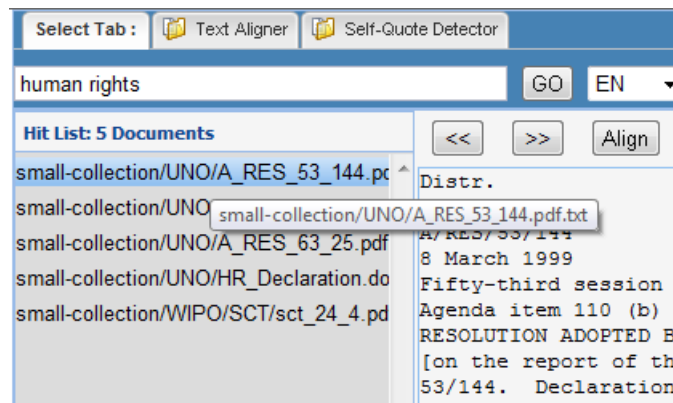
3.3. Searching for Terms or Expressions

After you selected the relevant language pair and collections, all you have to do is type your term query in the text field and execute the search by clicking on the “GO” button (or hitting the ENTER key on your keyboard) :



In the example above we searched for the expression “human rights”. The result list displayed in the Hit List frame includes all the documents which contain that term. If more than 200 results are found, the other results are discarded and a warning message is displayed in the bottom bar. You can navigate in the result list with the scroll bar located at the right of the frame.

The documents are listed by their path and file name. You can view the entire path (*i.e.* the collection and sub-collections in which they are located) by positioning the mouse on the file name; the complete path will be shown in a grey label, as illustrated below:



Clicking on any file name will display the source and target versions of the corresponding documents, and will align them on the first occurrence of the term in the text. That occurrence is highlighted in the source frame, and an indicative vertical blue bar shows where the corresponding term should be located in the target frame.

You can then copy the corresponding part of the target text and paste it into your current translation work. Copies can be performed through the usual CTRL+C keys, or through the browser's Edit/Copy menu.

3.4. Using the Navigation Buttons

The source and target frames have an identical button bar. You can navigate from one term occurrence to the next one by clicking on the >> button, and to the previous occurrence by clicking on the << button.

When the highlighted term or expression is the first one in the reference document, the message "First Hit" appears in the bottom bar.

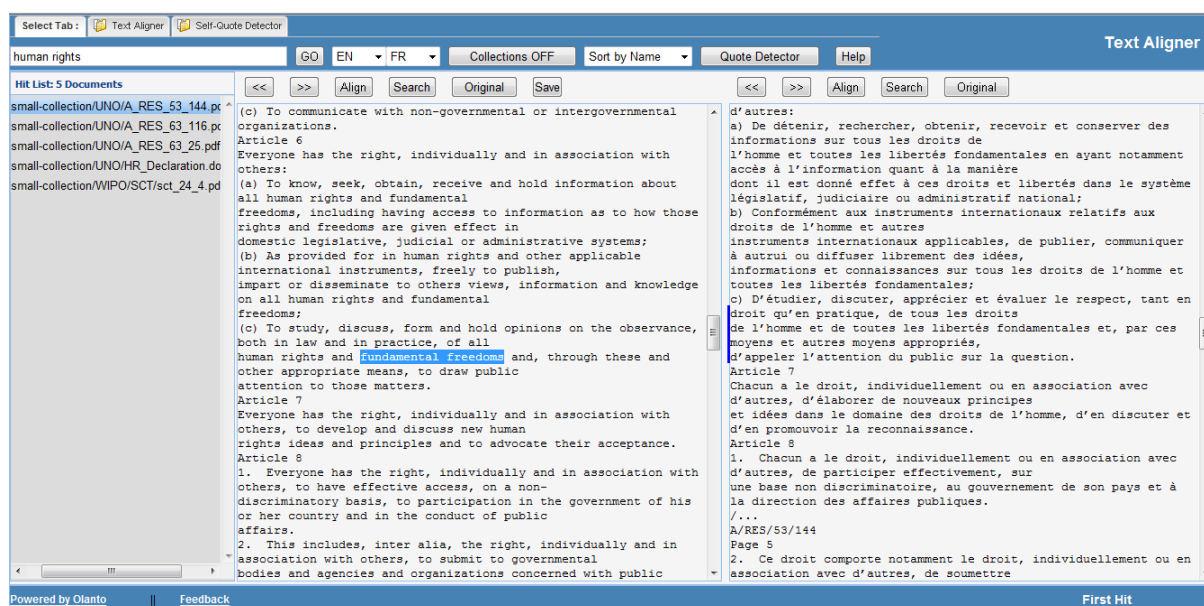
If you use the navigation buttons, you will navigate from a hit to the next one until you reach the final hit in the reference document. In that case the message "Last Hit" will appear in the bottom bar.

It is possible to search for terms which include a mix of letters and figures.

3.5. Using the « Align » Button

One of the core features of The Text Aligner is that the complete reference document is always provided both in the source and in the target languages. Thus you can always use the scroll bar on the right side of the frame to read more of any reference document. If you navigate far up or down in one version (source or target) of the document, you may want to display the corresponding part of the other version. In that case, simply double-click on any term you are interested in and then click on the "Align" button. The other version of the document will be automatically aligned.

Please remember to double-click a word in the text in which you are interested, otherwise the other version will not be aligned (because the system will not know which part of the text it should focus on).



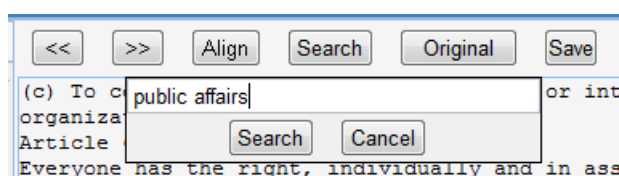
In the example above, although the initial search term was “human rights”, we highlighted in a source text the expression “Fundamental Freedoms” and clicked on the “Align” button. The corresponding part of the target version was then displayed and correctly aligned.

Conversely you can highlight a term and click on the “Align” button in the target frame; it is the source frame which will then be aligned.

3.6. Using the « Search » Button

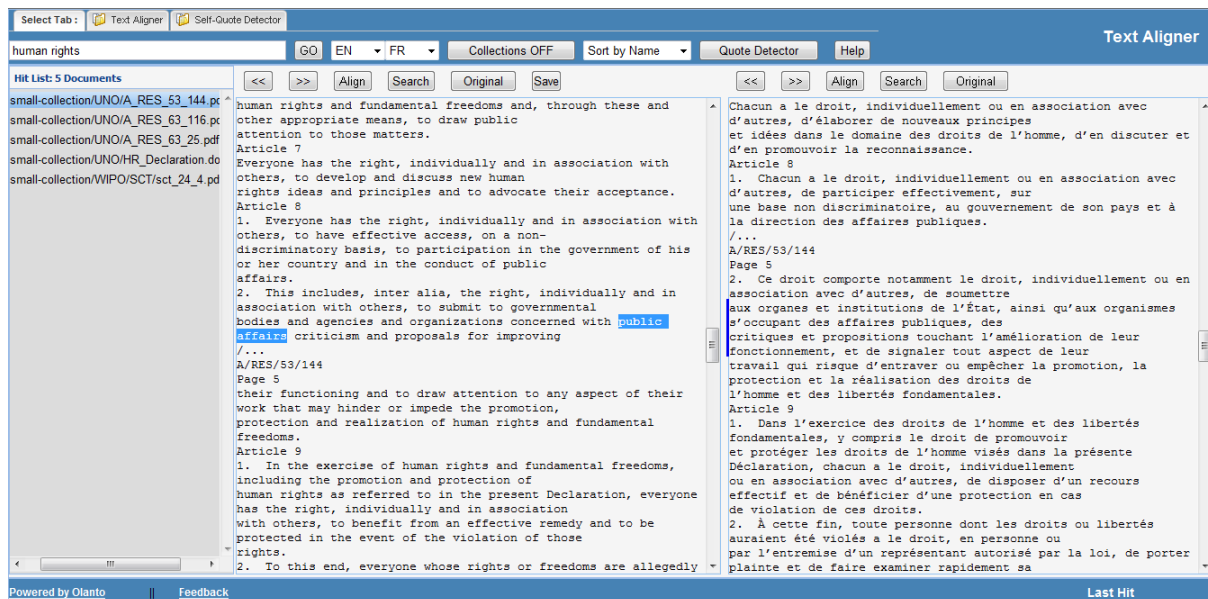
Some reference documents may include a lot of useful terminology for your current translation work. If such is the case, you may not want to perform again a complete search under The Text Aligner, but rather limit the search to the document at hand.

To do so, click on the “Search” button. A text field will appear, in which you can type your new query:



In this example we performed a new search on the expression “public affairs”. To execute that search, click on the “Search” button. (Clicking on the “Cancel” button will simply close the box without any further action.)

The result of this additional search within a hit document is the following:



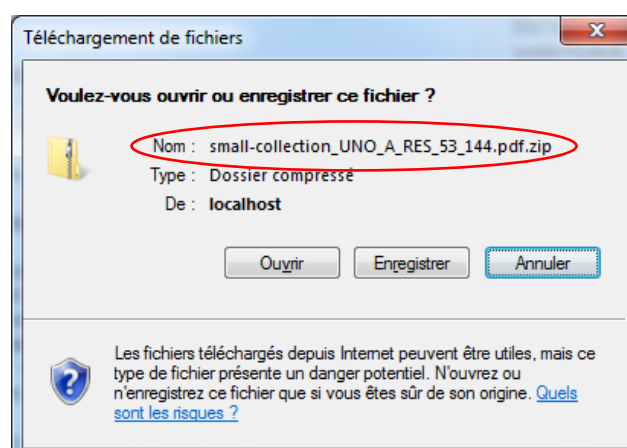
Although the initial search (targeting the whole corpus) was on “human rights”, the new search (targeting only the current document named small-collection/UNO/A_RES_53_144.pdf.) retrieved the expression “public affairs” and aligned the target version accordingly.

3.7. Using the « Original » Button






The “Original” button allows displaying the document in its original format (for example MS-Word or PDF) so the user can see and possibly import the real page layout. Simply click that button to display, in a new window or tab, the document in its original format.

3.8. Using the « Save » Button

The “Save” button allows saving in a zip file all the different language versions of the displayed document (in their original format). The zip file has the same name as the target files, with a .zip extension:



In this example the document named A_RES_53_144.pdf existed in 5 different languages and its original format was PDF, as illustrated below:

Nom	Taille compressée
 small-collection_UNO_A_RES_53_144_AR.pdf	65 Ko
 small-collection_UNO_A_RES_53_144_EN.pdf	27 Ko
 small-collection_UNO_A_RES_53_144_ES.pdf	20 Ko
 small-collection_UNO_A_RES_53_144_FR.pdf	20 Ko
 small-collection_UNO_A_RES_53_144_ZH.pdf	182 Ko

This function is useful for example if you intend to prepare a set of reference documents for a translation to be outsourced.

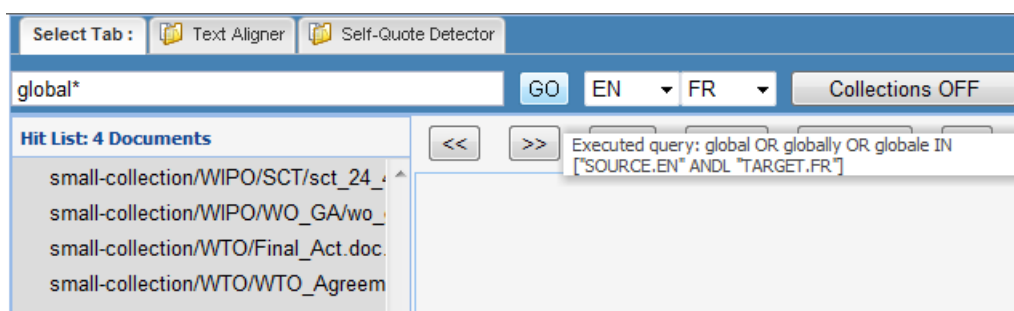
3.9. Using the Wildcard (Joker) Character

It is possible to extend a search by using a “wildcard” character, which is the * star character, and which can be replaced in the search by any possible sequence of characters.

- The wildcard can be added anywhere in a word:
 - A search on **global*** could retrieve words such as **global**, **globalized**, **globalization**, etc.
 - A search on ***conditioning** could retrieve **conditioning**, **preconditioning**, **reconditioning**, etc.
 - A search on **labo*r** could retrieve both **labor** and **labour**.
- It is possible to use several wildcards in a word:
 - A search on ***condition*** could retrieve **condition**, **conditioned**, **preconditioning**, etc.
 - A search on ***con*ition*** could retrieve **condition**, **contrition**, **conditioned**, **preconditioning**, etc.

There is an important constraint on the use of the wildcard character: it may only be used for a search on a single word, not on a combination of words (such as in an expression) because that would create too many possibilities (combinatory explosion).

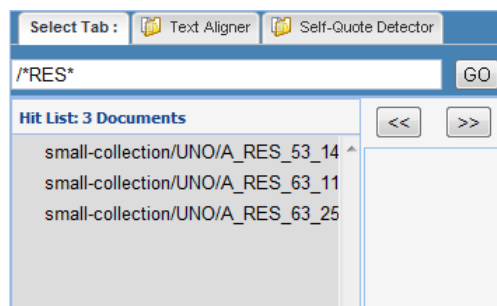
It is possible to check out the complete (*i.e.* extended) request by hovering the mouse over the GO button, as illustrated in the simple example below:



In that example we looked for the root word “global*”, which means we were interested in the word “global” but also in any longer word such as “globalized”, “globally”, etc. By placing the mouse on the GO button we could view the complete list of possible words which was built by the Text Aligner for this operation.

3.10. Performing a Search on File Names

So far the term search operations were performed in full text, *i.e.* within the content of the reference documents. However it is also possible to perform a search on file names or their path, for example if you are looking for a document whose name or path you already know, at least partially. In that case your search should begin by the following symbols: `/*` to indicate that you are targeting file names, not file contents. Then you can add more information about the name or path of the files, as illustrated below:



In that example we were looking for any file whose name includes the letters “RES” in any position. You can see that 3 documents matched that request.

Here are some tips to use the Search on File Names feature efficiently:

- If you want to see all the documents in the reference corpus simply type `/*`, or click on the GO button when the search field is empty. However please note that the maximum number of documents which can be retrieved from that feature is limited to 5000.
- You can select any particular collection or sub-collection by using the “Filter OFF” button. Then you can perform a search with `/*` which will retrieve all the documents within that specific collection. (Of course you can perform a more precise search on file names within that collection.)
- Don’t forget to always start your search on file names by `/*` and to add a wildcard at the end of your search expression.
- If you click on a document in the Hit List but the source and target windows remain empty (white), it might be because the document contained only pictures, so the txt conversion is empty. (This can be the case for example with glossaries which were scanned into a pdf file which is a single picture). In that case please click on the “Original” button: you might be able to access the document in its original format.

3.11. Using the Exact Search Function

Normally the Text Aligner only looks for words which are indexed; thus a search for an expression such as “information for users” actually only looks for “information” and “users”, because the preposition “for” is not indexed. The Text Aligner might also retrieve expressions such as “information from users” or “information to users”, since “from” and “to” are not indexed either.

However it is possible to force the Text Aligner to search only for the exact string of characters which were entered; to do so, simply use the quotation marks “ ” before and after the string. In the example above, a search for “information for users” will only retrieve that specific expression, because it will check out in the retrieved documents that the preposition “for” is actually present between the other two words.

This exact search function is not used by default because it is costly in terms of searching time, *i.e.* it takes some more time to get a reply from the Text Aligner when you use the quotation marks.

3.12. Boolean Operators and Other Search Tips

Three Boolean operators (i.e. words which allow for various types of search operations) are available under the Text Aligner:

- term1 term2 termN: Search for a single term or a complete expression or phrase. When no Boolean operator and no quotation marks are used, all the terms are searched for and highlighted. Thus the AND Boolean operator is implicit.
- term1 OR term2: Only one of the terms is included in the hit document. This operator is useful to find a term which can be spelled out in various ways (such as “labour” and “labor”, “organization” and “organisation”, etc.)
- term1 NEAR term2 : Both terms are included in the hit document and they appear at a maximum distance of 5 words. Not-indexed words (see below) are excluded from that count. Only the first term is highlighted.

Below are a few additional search tips:

- Search is case insensitive (term = Term = TERM) when searching in full texts, but it is case sensitive when searching on file names (i.e. when using the /* feature).
- Stop Words (not-indexed words)
 - Articles, prepositions and other very short and common words in English and French are not indexed, in order to streamline the index size and improve the reply time. Thus words such as le, la, les, du, des, the, of, etc... will not be found. The list of stop words can be customized for each language.

3.13. Using the “Help” Button

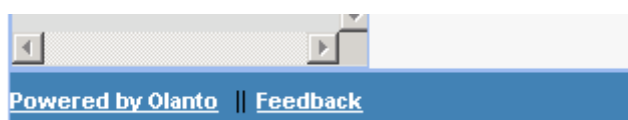
This User Manual can be accessed at anytime from the Text Aligner interface by clicking on the Help button at the top right of the screen.

3.14. Using the “Quote Detector” Button

A quote detection tool called the “Quote Detector” can be accessed from the Text Aligner interface by clicking on the “Quote Detector” button at the top right of the screen. A separate User Manual for the Quote Detector is available from the Help button in the Quote Detector interface.

3.15. Links in the Bottom Bar

Two links are provided at the bottom left corner of the screen, as illustrated below :



- **Powered by Olanto** : This link points to the web site of the Olanto Foundation, which is the publisher of myCAT (see <http://olanto.org> ; all the software published by Olanto is distributed under the AGPL open source license).

- **Feedback** : This links allows you to send an email in order to provide your feedback about myCAT. The default email address is info@olanto.org; it should be changed for the email of your Translation Support Unit.