August 2008
Volume XXXVII
Number 8
A Publication
of the
American
Translators
Association

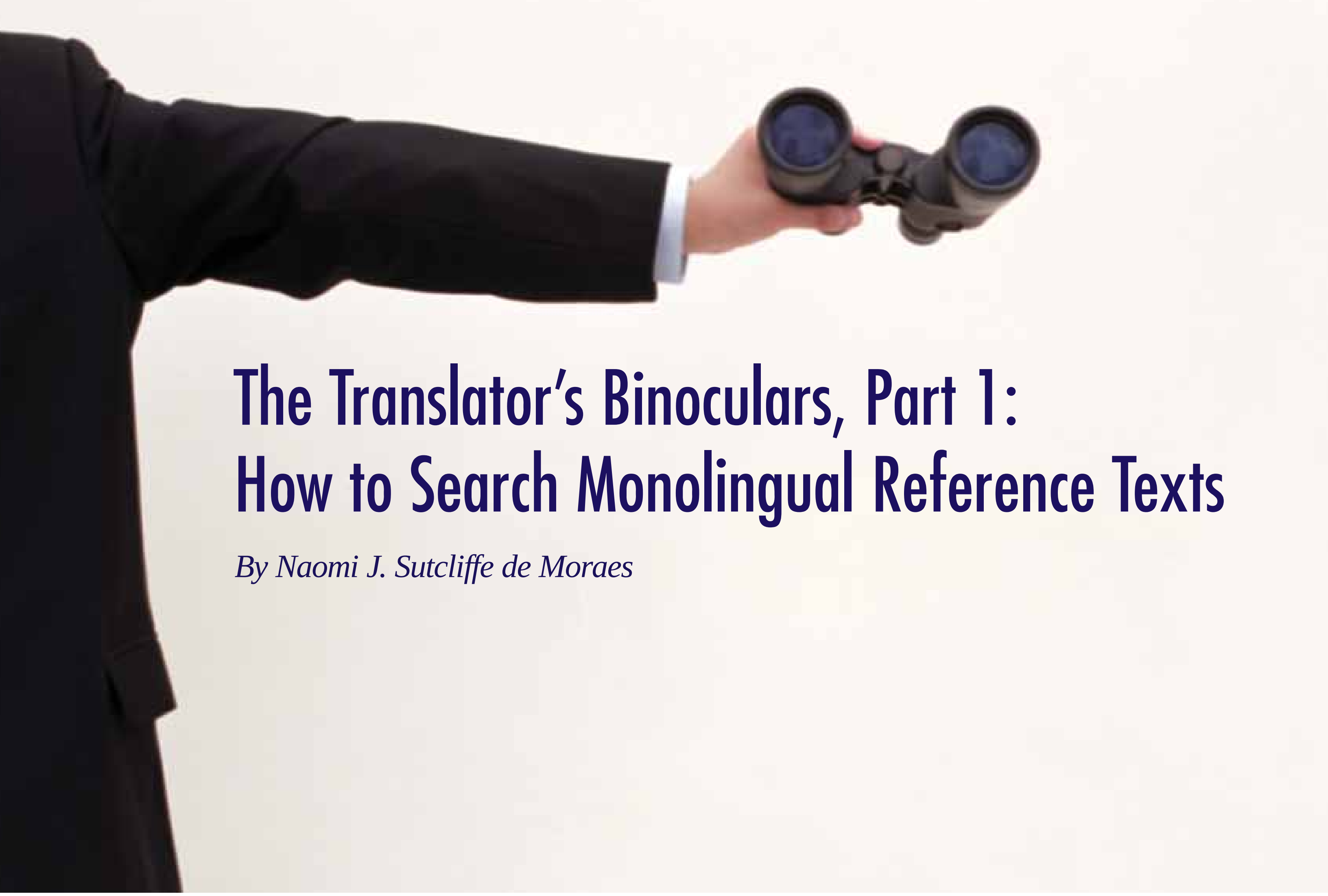
The ata CHRONICLE

In this issue:

Promoting Health Care Interpreting

PDF Files and Translation

Searching Monolingual Reference Texts



The Translator's Binoculars, Part 1: How to Search Monolingual Reference Texts

By Naomi J. Sutcliffe de Moraes

When most translators think of translation tools, they think of translation memory or terminology databases. But what about monolingual reference texts in the target language? There are three common situations where monolingual references will be provided or will prove useful:

1. When translating academic articles or reports containing a bibliography.

2. When translating marketing material or manuals for a client who provides you with similar material to consult in order to maintain consistent terminology.

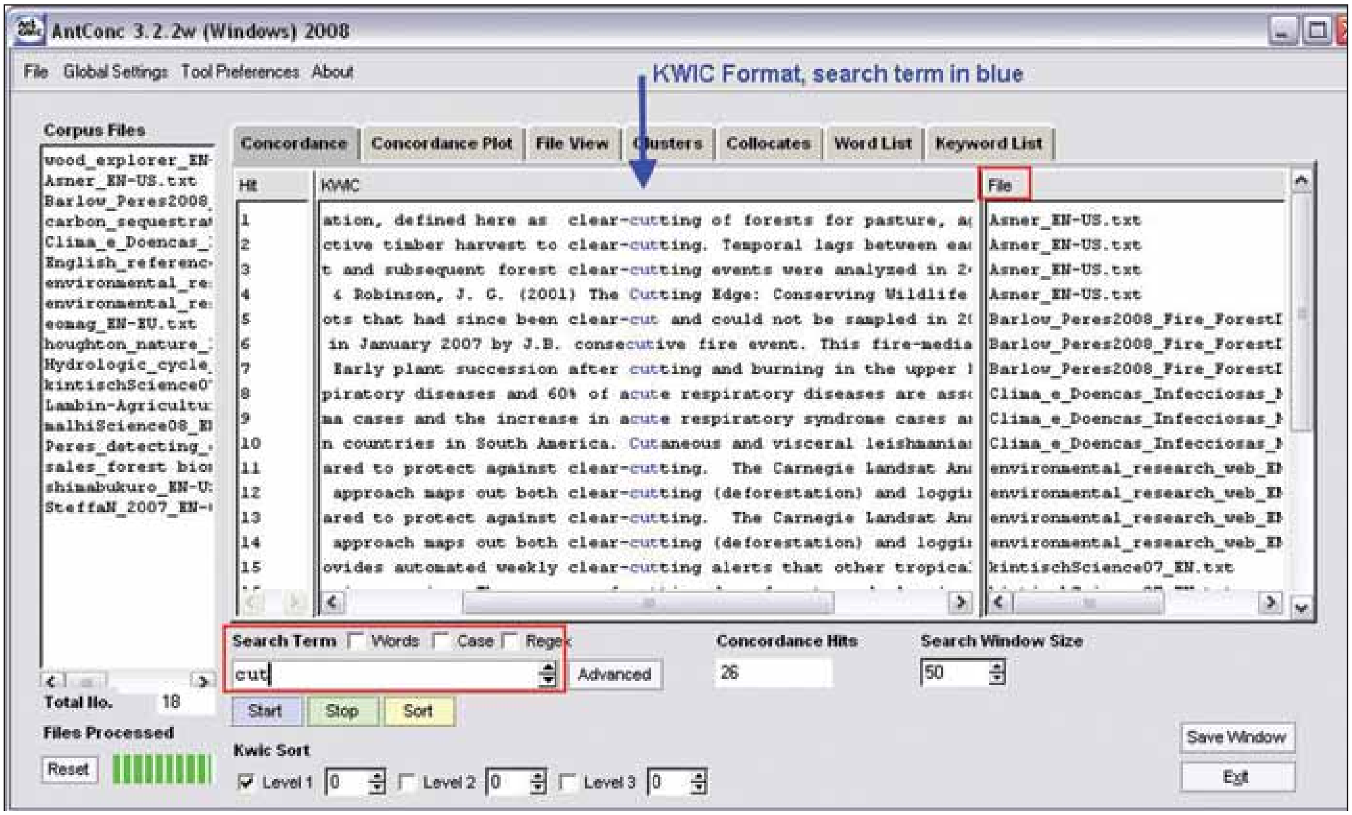
3. When translating legislation, where terminology must be consistent with prior laws.

> The primary advantage of using a tool like AntConc rather than just using Google to search the Internet at large is that you have much more control over which texts you are searching.

AntConc and What it Can Do

In linguistics, a corpus is a collection of texts of a specific type or on a specific subject that is stored electronically and used for lexical, grammatical, or other linguistic analyses. A corpus is often used to check for occurrences of commonly used words in related texts and to validate terminology choices. AntConc is what corpus linguists call a concordancer. AntConc imports your reference texts (your corpus) and can show you the contexts for all occurrences of a search term. Google also does this for texts on the Internet, but it uses algorithms to rank hits, whereas a normal concordancer like AntConc simply shows all occurrences of a term. Unlike Google, AntConc also shows collocations, that is, words that commonly appear together in the indexed texts, such as "dead serious" or "highly qualified."

In my article, "Techniques for



Corpus Files

```
wood_explorer_EN-
Asner_EN-US.txt
Barlow_Peres2008_
carbon_sequestra
Clima_e_Doencas_
English_referenc-
environmental_re:
environmental_re:
eomag_EN-EU.txt
houghton_nature_
Hydrologic_cycle_
kintischScience0
Lambin-Agricultu:
malhiScience08_EI
Peres_detecting_
sales_forest bio:
shimabukuro_EN-U:
SteffaN_2007_EN-
```

```
Hit KWIC                                                                    File
1   ation, defined here as  clear-cutting of forests for pasture, a;         Asner_EN-US.txt
2   ctive timber harvest to clear-cutting. Temporal lags between ea:         Asner_EN-US.txt
3   t and subsequent forest clear-cutting events were analyzed in 2         Asner_EN-US.txt
4    & Robinson, J. G. (2001) The Cutting Edge: Conserving Wildlife          Asner_EN-US.txt
5   ots that had since been clear-cut and could not be sampled in 2(        Barlow_Peres2008_Fire_ForestI
6    in January 2007 by J.B. consecutive fire event. This fire-media        Barlow_Peres2008_Fire_ForestI
7    Early plant succession after cutting and burning in the upper I         Barlow_Peres2008_Fire_ForestI
8   piratory diseases and 60% of acute respiratory diseases are ass         Clima_e_Doencas_Infecciosas_I
9   ma cases and the increase in acute respiratory syndrome cases a:         Clima_e_Doencas_Infecciosas_I
10  n countries in South America. Cutaneous and visceral leishmania:        Clima_e_Doencas_Infecciosas_I
11  ared to protect against clear-cutting.  The Carnegie Landsat An:        environmental_research_web_EI
12   approach maps out both clear-cutting (deforestation) and loggi:        environmental_research_web_EI
13  ared to protect against clear-cutting.  The Carnegie Landsat An:        environmental_research_web_EI
14   approach maps out both clear-cutting (deforestation) and loggi:        environmental_research_web_EI
15  ovides automated weekly clear-cutting alerts that other tropica:        kintischScience07_EN.txt
```

Search Term □ Words □ Case □ Regex: cut

Concordance Hits: 26

Search Window Size: 50

Total No. 18

Files Processed

Kwic Sort ☑ Level 1 0 □ Level 2 0 □ Level 3 0

Figure 1: Search on “cut” using AntConc

Teaching Medical Translation into English,” in the January 2004 issue of *The ATA Chronicle*, I mentioned the importance of background reading when beginning to translate a technical article. Using a concordancer lets you cover much more ground and saves time by pinpointing just the terms you need. For example, I recently translated a long article on satellite monitoring of deforestation in the Amazon. One of the first terms I came across (in Portuguese) was *corte raso*. I used what I call the “guess and check” method (see my article on IntelliWebSearch in the July issue) and guessed that *corte* was probably “cut.” I then searched in my reference texts for this particular project. The result is shown in Figure 1.

If you are translating material for a business client, even through an intermediary, ask for reference material if you think it would be useful.

In Figure 1, AntConc shows the search term in blue in the middle of the window in what is called the Key Word In Context (KWIC) format. The file in which the segment of text was found is listed in the right window. Note that I have indicated the dialect (U.S. or U.K. or foreign English) in the file name to help me separate out native collocations from non-native ones. In this particular job the client wanted British English, so when faced with different collocations in British and U.S. English, I followed the U.K. standard. This search on the word “cut” shows me that “clear-cut” is a very common term in my reference texts and is probably the translation for *corte raso*. This can then be confirmed through other research methods, including reading sections of the reference texts to find a definition of the English term.

Examples of Searches

Another example of how I used my reference corpus for the above project was when I was confronted with several terms used in a classification of forest cover types. I had to translate *Floresta Ombrófila Densa* and *Floresta Ombrófila Aberta*. I guessed that *densa* was probably “dense,” and the search results for “dense” are shown in Figure 2 on page 28.

The second hit in Figure 2 is “dense closed-canopy forest,” and the same sentence mentions "open canopy forests." Clicking on "dense" (in blue) opens the File View tab to reveal about 100 words before and after the search term. Note that AntConc allows you to sort on the form of the search word (e.g., dense, densely, denser), on the word immediately to the right (in red) or left, on the word two words to the right or left, and on various combinations of these. In Figure 2, I sorted only on the word directly to the right, which is why it is highlighted in red. Other searches I performed for this translation were:

1. Question: Can *sumidoro de carbono* be translated as carbon sink? Search on: *sink*
2. Question: Is *cobertura florestal* forest cover or forest coverage? Search on: *cover**
3. Is *regime hydrológico* hydrologic regime or hydrologic cycle, or hydrological cycle or water cycle? Search on: *Hydro** then on *cycle*

Note that, unlike search engines, AntConc lets you search using wildcards and regular expressions. Regular expressions provide a concise and flexible means for identifying strings of text of interest, such as particular characters, words, or patterns of characters. Wildcards are characters that substitute for other characters in regular expressions. Some of the default wildcard settings are:

* = one or more characters
(tree* = tree, trees, treed)

? = any one character
(wom?n = woman, women)

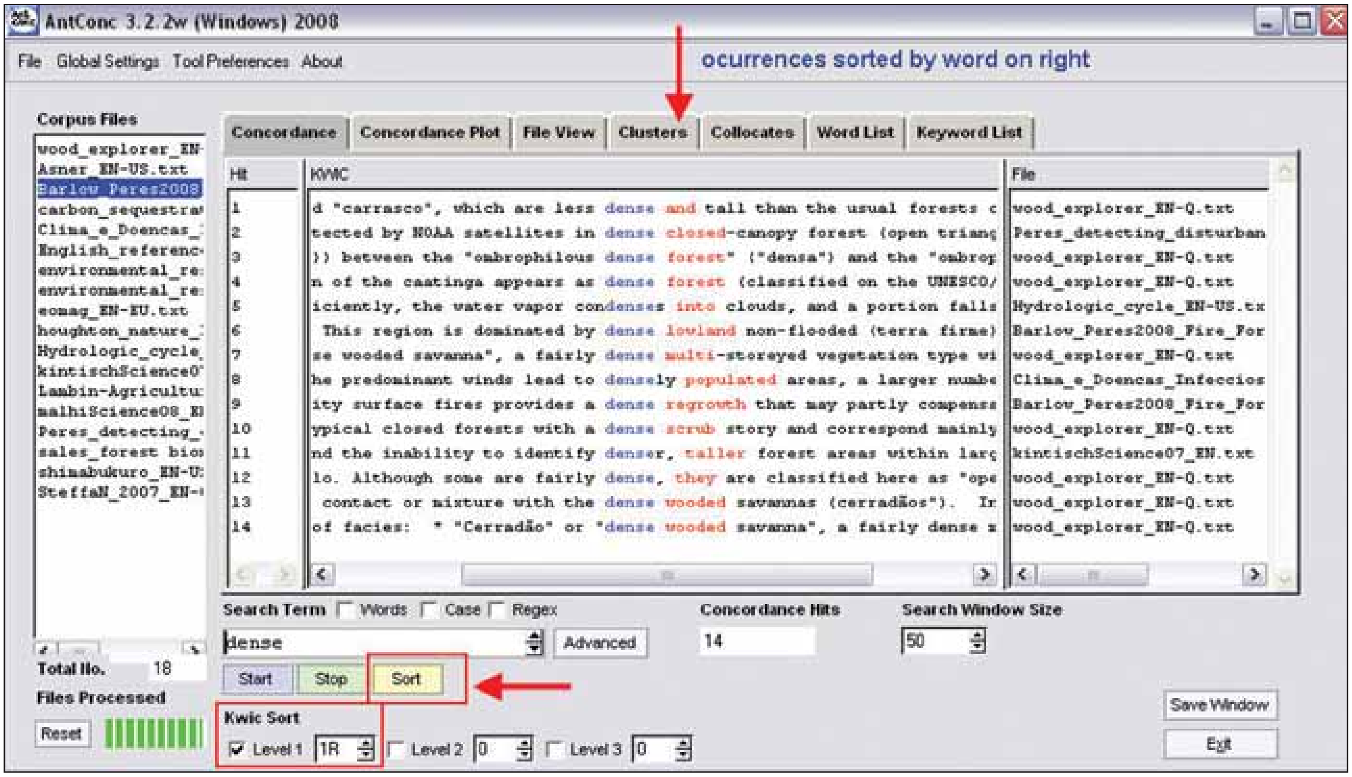
@ = zero or one word
(red @ brown = red brown, red and brown)

(See the references on page 32 for an excellent article on regular expressions by Jonathan Lukens that was published in the March 2008 issue of *LTD News*, the newsletter of ATA's Language Technology Division.)

Where / How to Get Reference Texts

The primary advantage of using a tool like AntConc rather than just using Google to search the Internet at large—in addition to the ability to use wildcards and regular expressions when searching—is that you have much more control over which texts you are searching. By looking at the names of article authors, their academic institutions, and where the articles were pub-

Figure 2: Search on "dense" using AntConc



Corpus Files

wood_explorer_EN-
Asner_EN-US.txt
Barlow_Peres2008
carbon_sequestra
Clima_e_Doencas_
English_referenc
environmental_re
environmental_re
eomag_EN-EU.txt
houghton_nature_
Hydrologic_cycle_
kintischScience0
Lambin-Agricultu
malhiScience08_E
Peres_detecting_
sales_forest bio
shimabukuro_EN-U
SteffaN_2007_EN-

Hit	KWIC	File
1	d "carrasco", which are less dense and tall than the usual forests c	wood_explorer_EN-Q.txt
2	tected by NOAA satellites in dense closed-canopy forest (open triang	Peres_detecting_disturban
3	)) between the "ombrophilous dense forest" ("densa") and the "ombrop	wood_explorer_EN-Q.txt
4	n of the caatinga appears as dense forest (classified on the UNESCO/	wood_explorer_EN-Q.txt
5	iciently, the water vapor condenses into clouds, and a portion falls	Hydrologic_cycle_EN-US.tx
6	This region is dominated by dense lowland non-flooded (terra firme)	Barlow_Peres2008_Fire_For
7	se wooded savanna", a fairly dense multi-storeyed vegetation type wi	wood_explorer_EN-Q.txt
8	he predominant winds lead to densely populated areas, a larger numbe	Clima_e_Doencas_Infeccios
9	ity surface fires provides a dense regrowth that may partly compensa	Barlow_Peres2008_Fire_For
10	ypical closed forests with a dense scrub story and correspond mainly	wood_explorer_EN-Q.txt
11	nd the inability to identify denser, taller forest areas within larg	kintischScience07_EN.txt
12	lo. Although some are fairly dense, they are classified here as "ope	wood_explorer_EN-Q.txt
13	contact or mixture with the dense wooded savannas (cerradãos"). In	wood_explorer_EN-Q.txt
14	of facies: * "Cerradão" or "dense wooded savanna", a fairly dense s	wood_explorer_EN-Q.txt

Search Term □ Words □ Case □ Regex: dense

Concordance Hits: 14

Search Window Size: 50

Total No. 18

Files Processed

Kwic Sort: ☑ Level 1 1R □ Level 2 0 □ Level 3 0

Table 1: Corpus Tools

Tool	Price	Type of Files Indexed	Ease of Use	Languages	Operating Systems
AntConc	Free	Text and HTML/.xml files	Very intuitive. Must convert files into text format before indexing. Has best results screen.	All,Unicode compliant	Windows, Mac, and Linux
LogiTerm Pro	$535 (includes one year of support and updates)	MS Word, Word Perfect, HTML, Excel, PowerPoint, PDF files not created as images, and text formats (.txt, .rtf, etc.)	Part of a larger program. No need to convert most files before indexing. Results screen requires extra click to see each hit.	Latin alphabet only	Windows
TextSTAT	Free	Windows MS Word files (.doc and .rtf), OpenOffice files (.sxw), ASCII/ANSI .txt files, and HTML files	Fewer options than AntConc, but very simple and easy to use. Results do not show file name.	All, Unicode compliant	Windows, Mac, and Linux

lished, you can get a feel for how trustworthy the English (or whatever language you translate into) will be before adding them to your corpus.

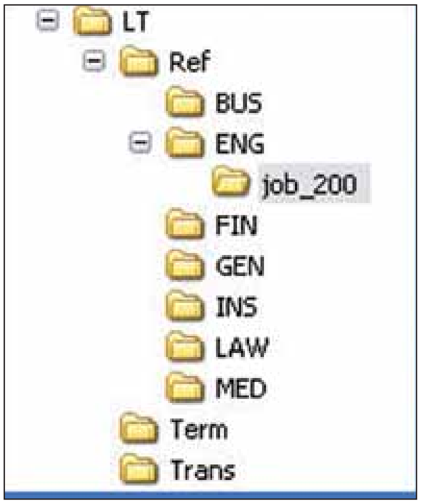
If you are translating an article for the authors, they probably have many of the references on their PCs, which they can send to you with no extra work. Some material can be downloaded from the Internet as you perform searches related to the translation, either in HTML or by cutting and pasting the text into a simple text (.txt) file. If the bibliography of the source document contains mostly English-language references, I usually copy the bibliography itself into a file and include it with the files I search. Finding a Wikipedia article on the general subject may also provide useful background material.

If you are translating material for a business client, even through an intermediary, ask for reference material if you think it would be useful. Look at the client's site in the target language to see what is available. You can actually download an entire website automatically using a tool like the free HTTrack, which is sort of like a vacuum cleaner that picks out the kinds of files you want (e.g., .txt, HTML, and PDF, but not images like .jpg or .gif) and places them on your hard drive with the same file structure (see references on page 32). Note that HTTrack should not be used to download files for translation because it changes the HTML code.

This kind of tool can also be used with *source language* files. Some projects contain many small files in no particular order. Sometimes the meaning of a word is not clear in the first context in which you encounter it, but a search on the word in all the project files will shed light on the meaning. This is very difficult to do without some kind of tool. I also found corpus tools useful when writing my Ph.D. dissertation. Each chapter was in a different file, and sometimes I would struggle to remember where I had written something I needed to reference. I simply added all the files to my corpus and performed a search on them.

AntConc lets you search using wildcards and regular expressions.

Figure 3: My file structure for organizing reference files



How AntConc Works

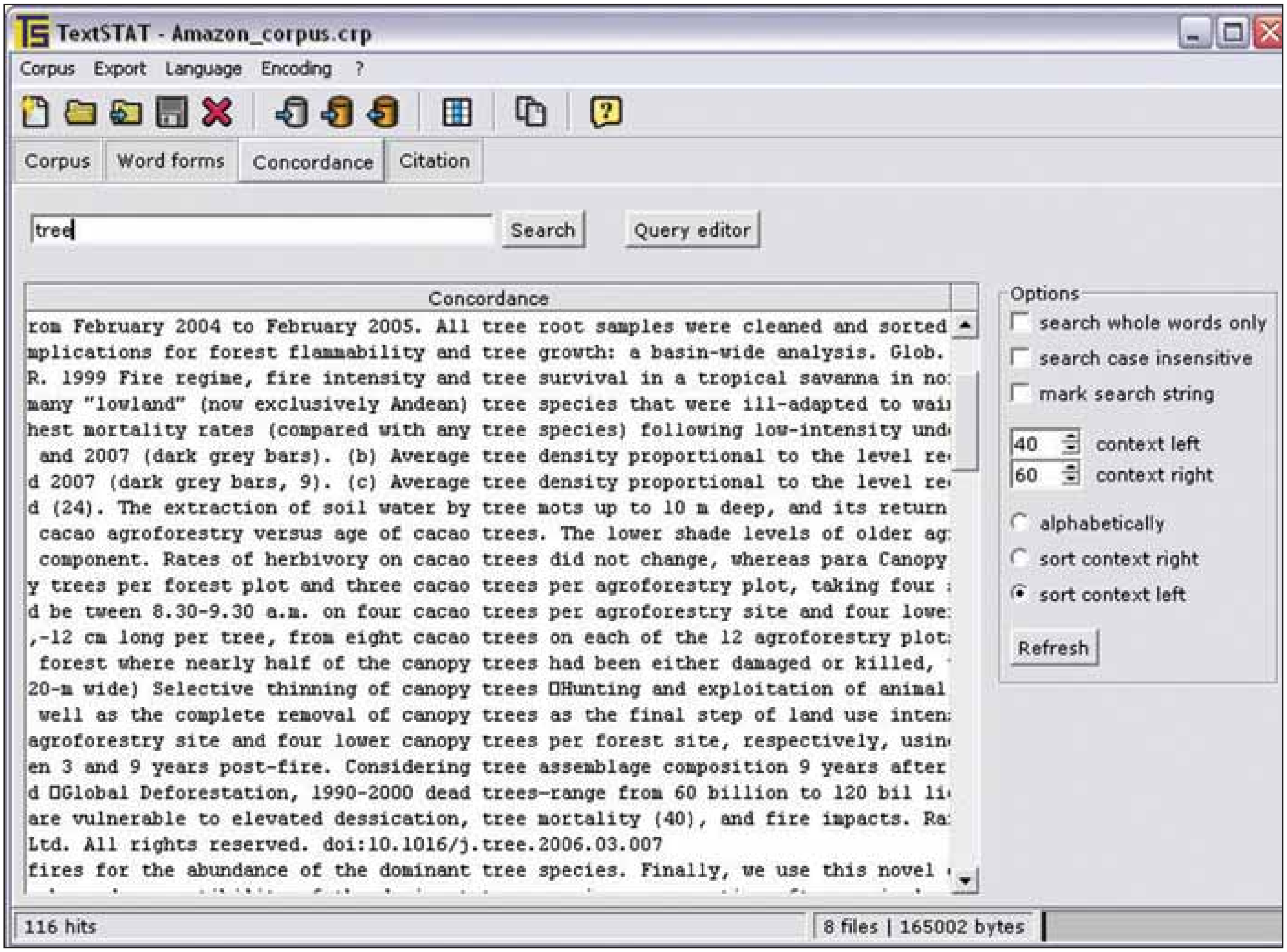
AntConc is very simple. Convert your reference texts into .txt or HTML format and place them in a folder. In the AntConc File menu, choose Add Files and select the files you want to search. You can also choose entire directories. Next, fill in the search term and AntConc shows the results.

Instead of placing reference files in the project file directory, I keep all terminology, prior aligned translations, and reference files on my D drive in the file structure shown in Figure 3 on page 29. I use subfolders for reference texts pertaining to business, engineering, financial, insurance, and other areas. Sometimes I create a folder for the job, as shown in Figure 3, and then move the files into the main folder for that category at a later time. You will often find that a reference file you downloaded for an earlier project will be useful again. If you convert the reference files into .txt files they will not occupy much space. I do not use AntConc to search my terminology files, but I certainly could if I converted them into text or HTML format.

Other Corpus Tools

Table 1 on page 29 provides a summary of some corpus tools and their features. I prefer AntConc because of its results screen. TextSTAT displays the word with context, but you must click once to see the file in which it was found for each occurrence. LogiTerm Pro displays the names of the files in which occurrences were found, but you must click once to see

Figure 4: TextSTAT results screen



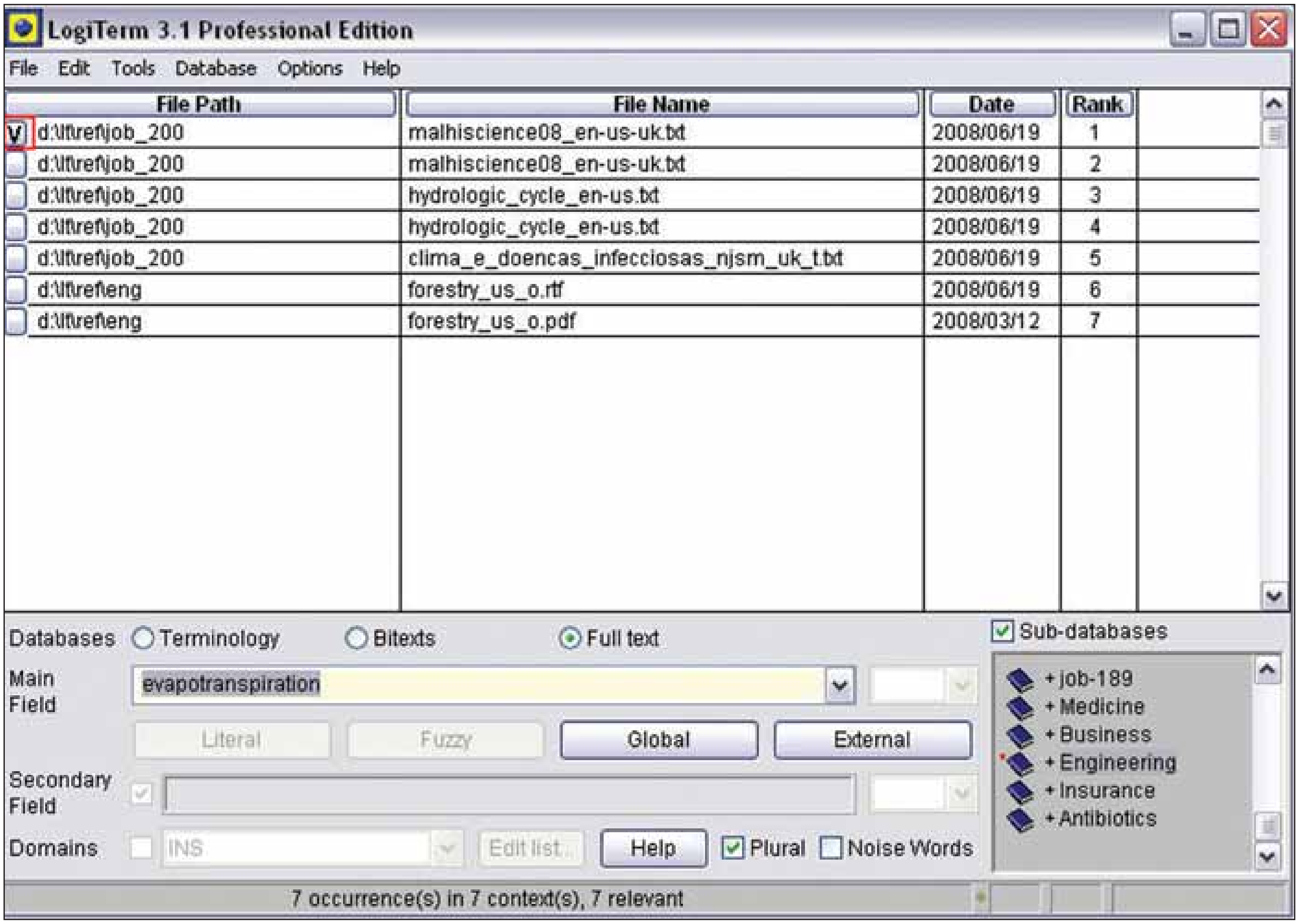
tree

```
rom February 2004 to February 2005. All tree root samples were cleaned and sorted
mplications for forest flammability and tree growth: a basin-wide analysis. Glob.
R. 1999 Fire regime, fire intensity and tree survival in a tropical savanna in no:
many "lowland" (now exclusively Andean) tree species that were ill-adapted to wai
hest mortality rates (compared with any tree species) following low-intensity und
 and 2007 (dark grey bars). (b) Average tree density proportional to the level re
d 2007 (dark grey bars, 9). (c) Average tree density proportional to the level re
d (24). The extraction of soil water by tree mots up to 10 m deep, and its return
 cacao agroforestry versus age of cacao trees. The lower shade levels of older ag:
 component. Rates of herbivory on cacao trees did not change, whereas para Canopy
y trees per forest plot and three cacao trees per agroforestry plot, taking four :
d be tween 8.30-9.30 a.m. on four cacao trees per agroforestry site and four lowe:
,-12 cm long per tree, from eight cacao trees on each of the 12 agroforestry plot:
 forest where nearly half of the canopy trees had been either damaged or killed,
20-m wide) Selective thinning of canopy trees □Hunting and exploitation of animal
 well as the complete removal of canopy trees as the final step of land use inten:
agroforestry site and four lower canopy trees per forest site, respectively, usin
en 3 and 9 years post-fire. Considering tree assemblage composition 9 years after
d □Global Deforestation, 1990-2000 dead trees-range from 60 billion to 120 bil li
are vulnerable to elevated dessication, tree mortality (40), and fire impacts. Ra:
Ltd. All rights reserved. doi:10.1016/j.tree.2006.03.007
fires for the abundance of the dominant tree species. Finally, we use this novel
```

☐ search whole words only
☐ search case insensitive
☐ mark search string
40 context left
60 context right
○ alphabetically
○ sort context right
● sort context left

116 hits | 8 files | 165002 bytes

Figure 5: LogiTerm Pro search of reference texts


LogiTerm 3.1 Professional Edition
File Edit Tools Database Options Help
Literal
Fuzzy
Global
External
Edit list...
Help

File Path	File Name	Date	Rank
☑ d:\lft\ref\job_200	malhiscience08_en-us-uk.txt	2008/06/19	1
□ d:\lft\ref\job_200	malhiscience08_en-us-uk.txt	2008/06/19	2
□ d:\lft\ref\job_200	hydrologic_cycle_en-us.txt	2008/06/19	3
□ d:\lft\ref\job_200	hydrologic_cycle_en-us.txt	2008/06/19	4
□ d:\lft\ref\job_200	clima_e_doencas_infecciosas_njsm_uk_t.txt	2008/06/19	5
□ d:\lft\ref\eng	forestry_us_o.rtf	2008/06/19	6
□ d:\lft\ref\eng	forestry_us_o.pdf	2008/03/12	7

Databases ○ Terminology ○ Bitexts ● Full text

Main Field: evapotranspiration

Secondary Field: ☑

Domains □ INS

☑ Plural □ Noise Words

☑ Sub-databases
- +job-189
- +Medicine
- +Business
- +Engineering
- +Insurance
- +Antibiotics

7 occurrence(s) in 7 context(s), 7 relevant

each occurrence. I like to see the context and the file name at the same time, since 10 hits in the same file is not as significant as 10 hits in 10 different files. These tools all permit the use of wildcards for searching, though the exact options differ from program to program.

TextSTAT

If you cannot run AntConc for some reason or do not want to convert your .doc/.rtf files into text, TextSTAT is a good alternative. It is simpler and also free. Figure 4 on page 30 shows a search on the word "tree" in the same corpus used in Figures 1 and 2. The context is shown in the KWIC format. Unlike AntConc, you can only sort on the first word to the right or left. You can, however, use wildcards and regular expressions. An added advantage is that TextSTAT contains a Web spider that can download files from a website and add them to your corpus automatically. The interface is available in English, Dutch, and German.

LogiTerm Pro

For detailed information on LogiTerm, see my two-part review in the November 2007 and January 2008 issues of *The ATA Chronicle*, especially the second part where reference texts are discussed. The program can index almost any file type, including PDFs not created as images. This is a great advantage because almost no work is required. Just drag the reference texts into a folder, tell LogiTerm Pro where to look, and then update the index.

The LogiTerm Pro "Full text" search results window is shown in Figure 5. The search was performed for "evapotranspiration," and seven hits were found in four different files, including one PDF file.

Note that LogiTerm Pro tells

you the file name and the folder where it can be found. In this case, it found a file on forestry in my general engineering reference texts that I had not even remembered I had, and it turned out to be extremely useful for this project. LogiTerm Pro allows you to use the * wildcard to represent zero or any number of characters and the ? wildcard to represent a single character in a word. It also lets you use quotes to find a specific string of text (do not forget to check the "Noise Words" checkbox).

The drawback of using LogiTerm Pro to search reference texts is that the results are not in KWIC format—you must click on the little "button" at the

References

Online

AntConc
www.antlab.sci.waseda.ac.jp/software.html

HTTrack Website Copier (Windows and Linux)
www.httrack.com

LogiTerm
www.terminotix.com

Lukens, Jonathan. "Turbocharging Your QA: Regular Expressions for Translators." ***LTD News*** **(March 2008), 8.**
www.ata-divisions.org/LTD/documents/newsletter/2008-1_LTDnewsletter.pdf

TextSTAT
www.niederlandistik.fu-berlin.de/ textstat/software-en.html

Print

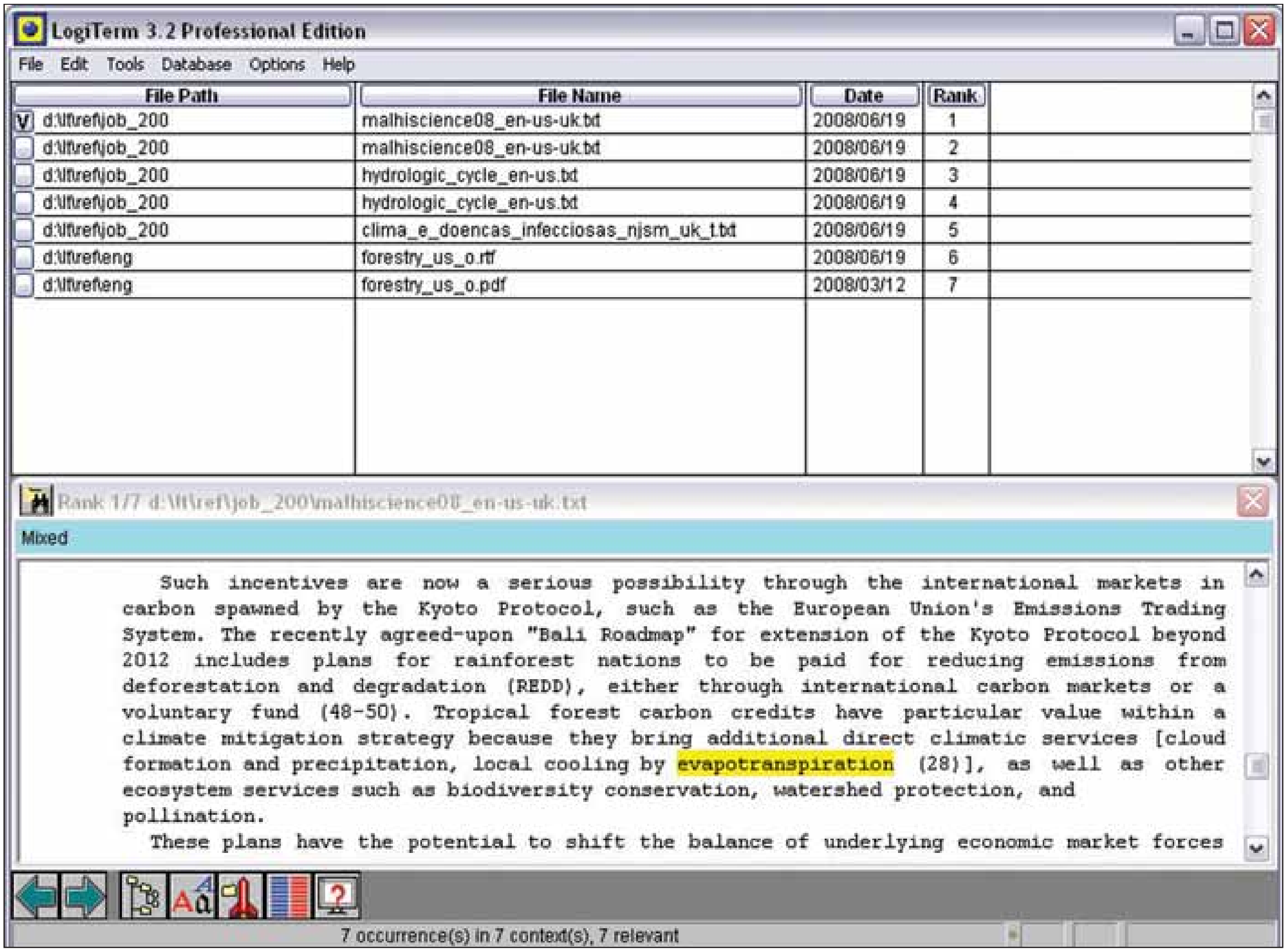
Sutcliffe de Moraes, Naomi. "Techniques for Teaching Medical Translation into English." *The ATA Chronicle* (January 2004), 30.

Sutcliffe de Moraes, Naomi. "LogiTerm: Your Personal Search Engine, Part I." *The ATA Chronicle* (November/December 2007), 32.

Sutcliffe de Moraes, Naomi. "LogiTerm Part II." *The ATA Chronicle* (January 2008), 22.

Sutcliffe de Moraes, Naomi. "IntelliWebSearch: A Configurable Search Tool for Translators." *The ATA Chronicle* (July 2008), 26.

Figure 6: LogiTerm Pro context for reference text search



	File Path	File Name	Date	Rank
V	d:\lt\ref\job_200	malhiscience08_en-us-uk.txt	2008/06/19	1
□	d:\lt\ref\job_200	malhiscience08_en-us-uk.txt	2008/06/19	2
□	d:\lt\ref\job_200	hydrologic_cycle_en-us.txt	2008/06/19	3
□	d:\lt\ref\job_200	hydrologic_cycle_en-us.txt	2008/06/19	4
□	d:\lt\ref\job_200	clima_e_doencas_infecciosas_njsm_uk_t.txt	2008/06/19	5
□	d:\lt\ref\eng	forestry_us_o.rtf	2008/06/19	6
□	d:\lt\ref\eng	forestry_us_o.pdf	2008/03/12	7

Rank 1/7 d:\lt\ref\job_200\malhiscience08_en-us-uk.txt

Mixed

Such incentives are now a serious possibility through the international markets in carbon spawned by the Kyoto Protocol, such as the European Union's Emissions Trading System. The recently agreed-upon "Bali Roadmap" for extension of the Kyoto Protocol beyond 2012 includes plans for rainforest nations to be paid for reducing emissions from deforestation and degradation (REDD), either through international carbon markets or a voluntary fund (48-50). Tropical forest carbon credits have particular value within a climate mitigation strategy because they bring additional direct climatic services [cloud formation and precipitation, local cooling by evapotranspiration (28)], as well as other ecosystem services such as biodiversity conservation, watershed protection, and pollination.

These plans have the potential to shift the balance of underlying economic market forces

7 occurrence(s) in 7 context(s), 7 relevant

beginning of each line (circled in red at the top of Figure 5 on page 31) to see the context, as shown in Figure 6. I do not understand why, since when searching terminology files and bitext files the term or translation is shown without the need to click on an extra box. It must be due to some internal limitation in how information is stored.

In the context window (Figure 6), LogiTerm Pro highlights the word to make spotting it easier. If you want to see the original file in its original format, click on the red rocket ship at the bottom of the screen to open the file. Note that after you click on a button to see the context, LogiTerm Pro marks the button with a V to indicate that you have already viewed the context.

I certainly would not recommend buying LogiTerm Pro *just* to search reference texts, since that is not its principal function and other tools can do this for free. However, if you already own it because of all the other functions it performs and are not using it for this purpose, take a look at the manual to familiarize yourself with this function. I collect reference files all the time and just drag them over to my LogiTerm reference folders (shown in Figure 3 on page 29) for later indexing and searching.

Next month, the second part of this review will present some desktop search tools that can be used to search reference material and find files on your computer's hard drives.