Read me file for AntConc 2.6.0 (Windows)

###

Laurence Anthony, Ph.D.
Center for English Language Education in Science and Engineering
School of Science and Engineering
Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
Sept. 13th, 2004
###

AntConc started out as a relatively simple concordance program, but has been slowly progressing to become a rather useful text analysis tool. It is written in Perl 5.6 usingActiveState's excellent Komodo 2.5 development environment. The program can be launched by simply double clicking on the .exe executable file, which can be downloaded from the Laurence Anthony laboratory website. The program can run under any windows environment including Win 98/Me/2000/NT and probably will run with no problems on Win XP.An earlier version of the software for Linux is also available. (If anyone wants the latest version of AntConc ported to Linux please let me know). If a user finds any problem launching the program under a particular OS, please let me know.

AntConc contains the following tools that will be explained separately.

Concordancer

The concordancer generates concordance lines (or KWIC: key word in context) lines from one or more target texts chosen by the user.

To produce a set of concordance lines of text, a user needs to perform the following actions:

1) Select either a single file or set of files stored in a single directory using the 'Open File' option in the 'File Menu'. The list of selected files is shown in the left frame of the main window.
2) Enter a search term on which to build concordance lines in the entry box on the left of the button bar.
3) Choose the number of text characters to be outputted on either side of the search term, using the increase and decrease buttons on the right of the button bar under the "Search Window Size" title. (default value is 30 characters)
4) Click on the 'Start' button to start the concordance lines results generation (Note: The concordance generation can be halted at any time by clicking on the 'Stop' button)
5) Select a target word on which to rearrange the concordance lines, using the buttons to the right of the button bar. 0 is the search word, 1L, 2L... are words to the left of the target word, 1R, 2R .. are words to the right of the target word. Note that two levels of sort are possible, with the second level not-activated when the software is first launched.
6) Click on the 'Sort' button to start the sorting process.
7) Move the cursor over the highlighted search term in one of the concordance lines. The cursor will change to a small hand icon. Clicking on the highlighted search term, will allow the user to view the search term hit as it appears in the original file via. the View Files Tool (see below).

Search words can be specified as being "Whole Words" by clicking on the search term option (default), and searches can be either case sensitive, or case insensitive (default). Searches can also be made using full regular expressions by choosing the "Regex" option. For details on how to use regular expressions, consult one of the many texts on the subject. E.g., Mastering Regular Expressions (O'Reilly & Associates Press).

The default colors used to highlight the search term, and the Level 1 and Level 2 sort words, can be changed using

options in the Settings Menu. The fonts used to display search terms and results can also be changed using options in the Settings Menu.

The user can choose to show or not show the full pathnames of the files chosen via the 'File Preferences' option in the 'Preferences' Menu. Similarly, the user can choose to show or not show the file names corresponding to hits in the concordance lines window via the 'File Preferences' option in the 'Preferences' Menu. Also, the user can choose to show or not show tags that exist in a file via the 'File Preferences' option in the 'Preferences' Menu.

Note that the total number of concordance lines generated (hits) is shown in the top right of the AntConc window. This number will flash with the word "FINISHED" when processing has been completed, and will flash with the word "NO HITS", if not hits are generated for a particular search term. In this case, the concordance lines view will not be updated, and the previous set of concordance lines will remain visible.

Concordance line results can be either saved to the clipboard, or saved to a text file (..txt) using the appropriate option in the 'File Menu'.

Concordance Search Term Plot

Generating Concordance Search Term Plots can be achieved using the same actions as when using the Concordancer. However, the Concordance Search Term Plot offers an alternative view of concordance lines. Here, all the hits for each file are plotted in the form of a 'barcode' indicating the position in the file where the hit occurred. The plot provides an easy way to see which files include the target search term, and can also be used to identify where the search term hits cluster together. An example of the use of the plot is in determining where specific content words appear in a technical paper, or when a character appears during the course of a novel or play.

The number of hits and length of each text is shown to the right of the barcode plot, and the plot itself can be enlarged or reduced in size using the zoom buttons

View Files

At any time a target file can be viewed in its original form using the View Files feature.

To produce a view of the original file, a user needs to perform the following actions:

1) Select a file to view in the file list frame to the left of the main window.

2) Hit the start button

3) If a search term has been specified, the search term hits will be highlighted throughout the text. Search options are the same as for the Concordance function.

4) Use the "Previous Hit" and "Next Hit" buttons to jump to the appropriate hit in the file.

5) Left clicking anywhere in the view file window will immediately cause the nearest hit (to the left of the cursor) to be highlighted and shown.

Word Clusters

The Word Clusters feature is used to generate a ordered list of clusters that appear around a search term in the target files listed in the left frame of the main window.

The clusters can be ordered either by frequency or alphabetically, and the list can be inverted, and the frequency values shown or not shown using the appropriate options next to the start button. Also, a user can select the minimum and maximum length (number of words) in each cluster, the size of the cluster search window (i.e. the window of words on either side of the search term that will be used to generate clusters), and the minimum frequency of clusters displayed.

Note that the resulting clusters found by the system will be displayed in lower case. This enables the system to combine the results of clusters containing the same words but using different case. However, the clusters search function itself can be operated in either the 'case sensitive' or 'case insensitive' mode. Also note that the search term is treated as one 'word' in each cluster, regardless of its length.

To produce a cluster list , a user needs to perform the following actions:

1) Choose the appropriate ordering options.

2) Press the 'Start' button.

At any time, the generation of the clusters list can be halted using the 'Stop' button.

Lexical Bundles (part of Word Clusters)

The Lexical Bundles feature is used to generate a ordered list of lexical bundles that appear in the target files listed in the left frame of the main window. Lexical bundles are word n-grams, and therefore, large files will create huge numbers of n-grams. For example, lexical bundles of size 2 for the sentence "this is a pen", are 'this is', 'is a' and 'a pen'.

The bundles can be ordered either by frequency or alphabetically, and the list can be inverted, and the frequency values shown or not shown using the appropriate options next to the start button. Also, a user can select the minimum and maximum length (number of words) in each bundle, and the minimum frequency of bundles displayed.

Note that the current in the current version of AntConc, a 'word' in the bundles is defined as a series of characters from the set "a to z A to Z 0 to 9". Note, also, that the resulting lexical bundles found by the system will be displayed in lower case. This enables the system to combine the results of bundles containing the same words but using different case.

To produce a bundles list , a user needs to perform the following actions:
1) Click on the "Bundles" option above the search entry box.

2) Choose the appropriate ordering options.

2) Press the 'Start' button.

At any time, the generation of the bundles list can be halted using the 'Stop' button.

Word List

The Word List feature is used to generate a list of ordered words that appear in the target files listed in the left frame of the main window.

The words can be ordered either by frequency or alphabetically, and the list can be inverted, and the frequency values shown or not shown using the appropriate options next to the start button.

To produce a word list, a user needs to perform the following actions:

1) Choose the appropriate ordering options.

2) Press the 'Start' button.

At any time, the generation of the word list can be halted using the 'Stop' button.

A number of preferences are available when generating word lists, via the "Wordlist Preferences" option in the "Preferences" menu.

a) The wordlist can be generated using all words, or a specific set of words, or ignoring a certain set of words (a stoplist). This is termed the "Wordlist Range".

b) The range of words to be used (or ignored) can be entered directly by the user, or can be stored in a text file (..txt) when is then read by the AntConc program by pressing the 'Choose File' program. A combination of words in a file, a words directly entered by the user can also be used.

Keyword List

In addition to generating word lists, AntConc can compare the words that appear in the target files with the words that appear in a 'reference corpus' to generate a list of "Keywords", that are unusually frequent (or infrequent) in the target files.

To produce a keyword list , a user needs to perform the following actions:

1) Select a set of target files (as used to generate Concordance lines, Concordance Search Term Plots, or the Wordlist).

2) Go to the 'Preferences' menu and chose the 'Keyword Preferences' option.

3) Choose a statistical measure to assess the 'keyness' or the target file words. The default setting of Log Likelihood is recommended.

4) Choose a threshold for the number of keywords to be displayed.

5) Choose whether or not to view 'Negative Keywords' (target file words with an unusually low frequency compared with the frequency in the reference corpus)

6) Choose a reference corpus of text (.txt) files, in the same manner that the target files are chosen. Files can be chosen using either the 'Choose Dir' or 'Choose Files' option.

7) The reference corpus directory will be shown (if appropriate), and the list of reference corpus files will appear at the bottom of the Keyword Preferences option menu.

8) Click 'OK' in the Keyword Preferences menu, and return to the main Keywords window.

9) Choose suitable options for displaying the list of generated Keywords (in a similar manner to the options for generating a Word List).

10) Press the 'Start' button.

At any time, the generation of the keyword list can be halted using the 'Stop' button.

Other Comments

All new editions and bug fixes are listed in the revision history below. However, if a user finds a bug in the program, or has any suggestions for improving the program, please let me know and I will try to address the issues in an future version. Indeed, the revisions that have been made are largely due to the comments of users around the world, for which I am very grateful.

This software is available as 'Freeware', but it is important for my funding to hear about any successes that people have had with the software. Therefore, if you find the software useful, please send me an e-mail briefly describing how it is being used.

Legal Matter

AntConc can be used freely for individual use for non-profit research purposes, and freely distributed on the condition that this readme file is attached in an unaltered state. If the software is planned to be used in a group environment, please let me know how you plan to use the software, and I will then give you permission for it to be used. The software comes on an 'as is' basis, and the author will accept no liability for any damage that may result from using the software.

Known Issues

1. When the file or directory selection dialog boxes are accessed, if they are dragged across the top of the AncConc main window, they leave 'ghost' traces. These traces are removed when the dialog boxes are closed. This appears to be a problem with the perl modules I use to access the Windows API. I have yet to find a way to solve this problem. Any help or suggestions would be grateful.

2. When a large number of concordance lines are generated (or words or keywords), the scrollbar becomes sensitive to where on the bar the user clicks and drags to view lower down entries. Sometimes this results in a user not being able to view the last lines unless the cursor is repositioned on the scrollbar. The is an annoying bug in the scrollbar subroutine (not mine!) and I am waiting for someone to fix this.

Revision History
2.6.0
1) Corrected a bug (introduced in version 2.5.3) which caused cluster lists to be not be displayed alphabetically.

2) Corrected a bug (introduced in version 2.5.3) that prevented the 2nd sort color to be properly updated from the settings menu. Thanks to a user for pointing this out to me.

3) Corrected a bug (introduced in version 2.5.3) that prevented files on a Japanese system to be open correctly. This bug is related to the new implementation of the Perl programming language. Please point out any bugs due to the new version of Perl, as they are very difficult for one person to discover.

4) Added an option to either show or ignore tags embedded in files (which is useful when processing HTML or XML files).

5) Added a 'lexical bundles' option to the Word Clusters tool, that generates "word n-grams".

6) Deleted the "Open Directory" option, as it prevented files other than .txt from being uploaded.

7) Added menu setting options to allow the font for searches and results to be changed.

8) Added menu setting options to show Japanese fonts (in shift_jis encoding) to be displayed properly.

9) Reduced the overall size of the program by approx. 1 MB.

10) A number of small bugs were corrected.

2.5.4
1) Corrected a bug (introduced in version 2.5.3) which caused word lists and keyword lists to be not be displayed alphabetically.

2.5.3
1) Adjusted the positioning and size of application widgets so that the application will be displayed properly on a 800 x 600 or higher resolution monitor.

2) Corrected several spelling errors in this readme file.

3) Adjusted the parameter settings for the Word Clusters tool.

4) Re-labeled several buttons so that they display properly on low resolution monitors.

2.5.2
1) Enabled a new 'Clusters' tool, for generating word clusters centered on a target search term.

2) Re-ordered Open File and Open Dir options in the File Menu. I think file navigation combined with 'select all files' <CTRL A> is easier than directory navigation,

3) Fixed a couple of small, insignificant bugs.

4) Fixed some spelling errors in this readme file.

2.5.1
1) Enabled a new feature, whereby clicking on the search term hit in any concordance line, allows the user to view the hit in the original file via the View Files tool.

2) Fixed a small bug in the View Files tool, which caused the searches to ignore the 'Case' setting.

3) Improved the performance when generating View Files by caching already processed files

2.5.0
A fairly major upgrade since 2.4.1
Here is a list of changes that have been made

1) Bug fix. When viewing files, and locating the next or previous hit, if the target file was changed and the hit number did not exist in the new file, the program would crash. This problem has been fixed.

2) Extension: In the view file window, hits would only appear if they occurred on a single line in the original file. This would result in different numbers of hits depending on if the search was made in the concordance window or the view files window. The view file processing has been completely revised enabling view file searching to correspond exactly with that used in the concordance window. Unfortunately, this has resulted in a small loss in performance when generating the highlighting in the view file. Also, clicking in the View File window now allows the user to immediately jump to the nearest hit.

3) The ability to show or not show full path names to files has been added as a system preference

4) The ability to show or not show file names in the concordance window has been added as a system preference.

5) The ability to set a wordlist 'range' has been added as a system preference.

6) Highlighting in the View Files tool has been changed to make the hits easier to see.

7) Pop-up windows that showed how many concordance hits were generated, and that reported when no hits were found have been removed. Instead, the status of the concordance hit processing is now shown in the top right of the main window.

8) Many small bugs relating to how the various tool displays are updated after preference changes are made have been corrected.

9) Processing that blocks user events (such as mouse clicks etc.) have been reduced.

10) The internal workings of the program have been re-written so that problems and future additions can be easily handled.

11) The general layout of the README file has be re-designed.

2.4.1
New since AntConc2.4.1 is the ability to choose whether or not to view 'Negative Keywords'. These are words in the target file that have an unusually 'low' frequency. In previous versions of AntConc, Negative Keywords were not distinguished from Keywords. However, now they are treated separately, and if the user chooses to display them, they appear after the Keywords, with a highlight color.

2.4.0
A major upgrade since 2.3.0
First, progress indicators were added to 'pages' of AntConc. Second, a new file view feature was added to view target files in their original state. Third, a keyword generation feature has been added using log-likelihood and chi-squared methods. Finally several bugs were found, in particular, bugs centered around the wordlist generation feature. This feature of the software should work much quicker now. Also, the user can interrupt the processing of files in any 'page' of the software.

2.3.0
A major upgrade since 2.2.3
First, the ability to view concordance search results as a barcode plot graph and a feature to produce wordlist according to different criteria were added. Numerous bugs centered around the way the software entered a 'Busy' mode were corrected. The main core of the software was also updated resulting in a quicker, 'cleaner' processing of the data. Performance improvements should be noticed as a result.

2.2.3

Updated file and directory selection dialog boxes to run smoothly in a Windows environment. Also, changed the default colors for sort highlighting, and search window frame size. A number of small bugs were also corrected

2.2.2
Corrected critical fault with compiler than caused program to expire when evaluation version of ActiveState Perl Development Kit expired. Sorry folks!! I didn't realize this would happen!!

2.2.1
Corrected bug which prevented new concordance lines being generated if the search term was left the same and then new files were selected. Port to Linux also completed.

2.2
Designed new subroutines for selecting directories and files to solve rendering of dialog windows problems. This also enables an easier port to Linux.

2.1
Added a second level of sort. Added ability to restrict searches to full-words only, case sensitive. Added ability to search using full Perl implemented regular expressions. Added ability to save results either to a file or the clipboard

2.0
Added new sort feature, for rearranging concordance lines. Tidied up the interface. Made the system more robust for novice users. (Now bad input will not cause the system to crash so easily).

1.1
Added binding to allow return key to launch concordance search. Also, recompiled software so that no console is required.

1.0
First version

Laurence Anthony