AntConc (Windows, Macintosh OS X, and Linux)

Build 3.2.4

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
November 10, 2011

Introduction

AntConc is a freeware, multiplatform tool for carrying out corpus linguistics research and data-driven learning. It runs on any computer running Microsoft Windows (tested on Win 98/Me/2000/NT, XP, Vista, Win 7), Macintosh OS X (tested on 10.4.x, 10.5.x, 10.6.x), and Linux (tested on Ubuntu 10). It is developed in Perl using ActiveState's PerlApp compiler to generate executables for the different operating systems.

Installation

Windows

On Windows systems, simply double click the *AntConc* icon and this will launch the program. No installation is necessary.

Macintosh OS X

On Macintosh systems, first install and launch X11. X11 is a graphical toolkit that is available on the disks included with the computer or via the Apple website. Next, double click double click the *AntConc* icon and this will launch the program. No installation is necessary.

Linux

On Linux systems, change the permissions to allow AntConc to be run as an executable file. Next, double click double click the *AntConc* icon and this will launch the program. No installation is necessary.

Overview of Tools

AntConc contains seven tools that can be accessed either by clicking on their 'tabs' in the tool window, or using the function keys F1 to F7.

Concordance Tool:

This tool shows search results in a 'KWIC' (KeyWord In Context) format. This allows you to see how words and phrases are commonly used in a corpus of texts.

Concordance Plot Tool

This tool shows search results plotted as a 'barcode' format. This allows you to see the position where search results appear in target texts.

File View Tool

This tool shows the text of individual files. This allows you to investigate in more detail the results generated in other tools of AntConc.

Clusters (N-Grams):

This Clusters Tool shows clusters based on the search condition. In effect it summarizes the results generated in the Concordance Tool or Concordance Plot Tool. The N-Grams Tool, on the other hand, scans the entire corpus for 'N' (e.g. 1 word, 2 words, …) length clusters. This allows you to find common expressions in a corpus.

Collocates:
This tool shows the collocates of a search term. This allows you to investigate non-sequential patterns in language.

Word List:
This tool counts all the words in the corpus and presents them in an ordered list. This allows you to quickly find which words are the most frequent in a corpus.

Keyword List:
This tool shows the which words are unusually frequent (or infrequent) in the corpus in comparison with the words in a reference corpus. This allows you to identify characteristic words in the corpus, for example, as part of a genre or ESP study.

Concordance Tool

This tool shows search results in a 'KWIC' (KeyWord In Context) format. This allows you to see how words and phrases are commonly used in a corpus of texts.

To produce a set of concordance lines of text, you need to perform the following actions:

1) Select one or more files for processing from using the 'Open File(s)...' or 'Open Dir...' options 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 50 characters)
4) Click on the 'Start' button to start the concordance lines results generation. 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 three levels of sort are possible, with the second and third levels 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 **File View Tool** (see below).

Note that the total number of concordance lines generated (hits) is shown in the middle of the AntConc button 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.

Search terms can be specified as being "words"(default) or "word fragments" by choosing the "Word" search term option. Also, searches can be either case sensitive or case insensitive (default) by choosing the "Case" search term option. 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) or type "regular expressions" in a web search engine to find many sites on the subject.

Information about regular expressions can be found at: http://www.regular-expressions.info/quickstart.html

By clicking on the "Advanced Search" button, more complex searches become possible. The first advanced search option is to define a set of search terms, either by typing them one per line, or by loading in a list of search terms from a file. Note that each line will be treated as a separate search term. The feature allows the user to use a large set of search terms without having to re-type them each time. The second advanced search option is to define context words and a context window within which the search term(s) must appear. For example, to search for "student" where it appears at least three words to the left or right of the word "university," set the search term as "student," the context word as "university," and the context window as 'From' 3L 'To' 3R.

A number of menu preferences are available with this tool. (See below).

Concordance Plot Tool

This tool shows search results plotted as a 'barcode' format. This allows you to see the position where search results appear in target texts.

Generating concordance plots can be achieved using the same actions as when using the **Concordance Tool**. However, the **Concordance Plot Tool** 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

If you 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. See the **File View Tool** (see below).

File View Tool

This tool shows the text of individual files. This allows you to investigate in more detail the results generated in other tools of AntConc.

To produce a view of the original file, you need to perform the following actions:

1) Select a file to view in the file list frame to the left of the main window.
2) 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 Tool** and **Concordance Plot Tool**.
3) Use the "Hit Location" buttons to jump to the appropriate hit in the file.
4) Change the search term and click on the 'Start' button to view other hits in the file.
5) Clicking on the highlighted text will generate a set of KWIC lines using the highlighted text as the search term.

Below is a list of Shortcuts unique to the **File View Tool**.

1) CTRL-Click = Jumps to the nearest hit in the window

Clusters Tool

The Clusters Tool shows clusters based on the search condition. In effect it summarizes the results generated in the Concordance Tool or Concordance Plot Tool.

The clusters can be ordered either by frequency or the start or end of the word. They can also be ordered by the probability of the first word in the cluster preceding the remaining words. All list orderings can also be inverted. Also, a user can select the minimum and maximum length (number of words) in each cluster, and the minimum frequency of clusters displayed. It is also possible to select if the search term always appears on the left or right of the cluster. (Note: In the current version, if more than one word is specified as the search term, only the first word will appear on the right, if the "Search Term on Right" option is selected.)

To produce a cluster list, you need 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.
3) Clicking on the cluster will generate a set of KWIC lines using the text as the search term.

A number of menu preferences are available with this tool. (See below).

N-Grams Tool (part of the Clusters Tool)

The N-Grams Tool scans the entire corpus for 'N' (e.g. 1 word, 2 words, …) length clusters. This allows you to find common expressions in a corpus. For example, n-grams of size 2 for the sentence "this is a pen", are 'this is', 'is a' and 'a pen'.

As with the **Clusters Tool**, the n-grams can be ordered either by frequency or the start or end of the word. They can also be ordered by the probability of the first word in the cluster preceding the remaining words. All list orderings can also be inverted. Also, a user can select the minimum and maximum size (number of words) in each n-gram, and the minimum frequency of n-grams displayed.

To produce an N-gram list, you need to perform the following actions:

1) Click on the "N-Grams" option above the search entry box.
2) Choose the appropriate ordering options.
3) Press the 'Start' button. At any time, the generation of the n-grams list can be halted using the 'Stop' button.
4) Clicking on the lexical bundle will generate a set of KWIC lines using the text as the search term.

A number of menu preferences are available with this tool. (See below).

Collocates Tool

This tool shows the collocates of a search term. This allows you to investigate non-sequential patterns in language.

The collocates can be ordered either by frequency, frequency on the left or right of the search term, or the start or end of the word. They can also be ordered by the value of a statistical measure between the search term and the collocate. The value measures how 'related' the search term and the collocate are. Current

possible statistical measures are listed below. All list orderings can also be inverted. Also, a user can select the span of words to the left and right of the search term in which to find collocates, and the minimum frequency of collocates displayed. If only a one-word span is required, for example, to see which words appear directly on the right of the search term, check the "Same" box, to keep the minimum and maximum span size the same.

Statistical Measures:
(MI) Mutual Information: Using equations described in M. Stubbs, Collocations and Semantic Profiles, Functions of Language 2, 1 (1995)
(T-Score) T-Score: Using equations described in M. Stubbs, Collocations and Semantic Profiles, Functions of Language 2, 1 (1995)

To produce a collocate list, you need to perform the following actions:

1) Choose the appropriate ordering options.
2) Press the 'Start' button. At any time, the generation of the collocates list can be halted using the 'Stop' button.
3) Clicking on the collocates will generate a set of KWIC lines using the text as the search term.

A number of menu preferences are available with this tool. (See below).

Word List Tool

This tool counts all the words in the corpus and presents them in an ordered list. This allows you to quickly find which words are the most frequent in a corpus.

The words can be ordered either by frequency or the start or end of the word, and the list can be inverted. The word list can also be generated in case-insensitive mode, where words in upper and lower case are treated the same (default) or case-sensitive, where words in upper and lower case are treated separately.

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.
3) Clicking on the word will generate a set of KWIC lines using the text as the search term.

A number of menu preferences are available with this tool. (See below).

Keyword List

This tool shows the which words are unusually frequent (or infrequent) in the corpus in comparison with the words in a reference corpus. This allows you to identify characteristic words in the corpus, for example, as part of a genre or ESP study.

To produce a keyword list, you need to perform the following actions:
1) Select a set of target files.
2) Go to the 'Preferences' menu and chose the 'Keyword Preferences' option.
3) Choose a statistical measure to assess the 'keyness' of 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 one of the reference corpus options. Select "Use raw file(s)" when you will use raw text (.txt) files to serve as the reference corpus. Select "Use word list(s)" when you will use one of more word lists that are generated from a reference corpus. The "Use word list(s)" option allows you to generate keywords even when the original reference corpus is not available. The format for a word list is:

RANK	FREQUENCY		WORD (separated by any type of white space, including spaces and tabs).
e.g.:	1	12838	the
	2	11289	a
	3	8583	of
	...		

7) Load the reference corpus of text (.txt) files, in the same way that the target files are chosen.
8) 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.
9) Click 'OK' in the Keyword Preferences menu, and return to the main Keywords window.
10) Choose suitable options for displaying the list of generated Keywords (in a similar manner to the options for generating a Word List).
11) Press the 'Start' button. At any time, the generation of the keyword list can be halted using the 'Stop' button.
12) Clicking on the keyword will generate a set of KWIC lines using the text as the search term.

A number of menu preferences are available with this tool. (See below).

MENU OPTIONS

Menu options are divided into three groups, "File", "Global Settings" and "Tool Preferences". The options available in each group will be described below.

<FILE>
Options here relate to reading files into AntConc and writing files to the hard disk containing data of various types. There are also options to export all current settings to a file, and import user settings from a file. If a user settings file becomes corrupted for any reason, simply restart the program or use the "Restore Default Settings" option to return the program to its original state.

<GLOBAL SETTINGS>
Categories here will have an effect on multiple tools in AntConc:

<File Settings>
In the File Settings category, the user can choose to display the full path of a file or just the name. The user can also choose to show or hide any tags in the file. The tag boundaries can be specified.

<Tag Settings>
In the Tag Settings category, the user can choose to display or hide any tags that are contained in the corpus files. If tags are to be hidden, the opening and closing tag markers must be specified. The default is <>.

<Wildcard Settings>
In the Wildcard Settings category, users can edit the default wildcard characters so that they do not clash with a search entry. For example, the "or" wildcard default character (a 'pipe' character |) can be changed to a backslash / here.

<Token (Word) Definition>

In the Token (Word) Definition category, the user can choose which characters, numbers and so on will define a "word". For example, in some cases only letters will be considered words, but at other times, in might be desirable to include numbers, dashed and so on in the word definition. AntConc is fully Unicode compliant, meaning that it can handle data in any language, including all European languages and Asian languages. For this reason, the default options talk of letters in the broadest sense. Letters, for example, include all Japanese characters, if that language encoding is being used (see below). It is also possible for a user to define his or her own "word" definition.

For more information on the Unicode standards see:
http://www.cs.tut.fi/~jkorpela/unicode/guide.html
http://www.unicode.org/
http://www.unicode.org/Public/5.0.0/ucd/UCD.html
http://www.unicode.org/Public/UNIDATA/PropList.txt
http://www.unicode.org/charts/

<Color Settings>

In the Color Settings category, the user can edit the colors used to display results and other information.

<Font Settings>

In the Font Settings category, the user can edit the font types, sizes, and styles used to display results and other information.

<Language Encodings>

AntConc is fully Unicode compliant, meaning that it can handle data in any language, including all European languages and Asian languages. The language (encoding) of the data to be read by AntConc should be specified here. For example, if you are working with data saved in a Western language, it will usually be encoded in iso-8859-1 (default). On the other hand, Japanese texts are usually encoded in SHIFT-JIS. By specifying the correct encoding, data from all languages can be processed correctly within AntConc.

<TOOL PREFERENCES>

Each tool (with the exception of **Concordance Plot Tool** and **File View Tool**) has a preferences category, where settings can be fine-tuned. All tool preference categories allow the user to show or hide the different frames in which the results are displayed. For example, the use can choose to hide the frame showing file names in the **Concordance Tool** display window. Also, all tools have the option to treat all data as lowercase and use case when sorting. If results are displayed case sensitively, words including capital letters will appear higher up in the display.

<Concordance Preferences>

In addition to the above, the following settings can be made:
Instead of arranging results by words to the left or right of the search term, it is possible to arrange the results by LETTERS to the left or right of the first letter of the search term. This makes it possible to search for spelling differences. The search term can also be chosen to be hidden in the KWIC lines, allowing instructors to quiz students on possible words to fit the gap. Another option is to add tab spaces around the hit in the KWIC display. This makes it easier to see the hit and also eases later processing of the data in a spreadsheet software program.

<Clusters Preferences>
For this tool there are no additional settings that can be set other than those described above.

<Collocates Preferences>
In addition to the above the settings, the choice of statistical measure can be chosen here. Currently, two statistical measures can be used: Mutual Information (MI) and T-Score. See above.

<Word List Preferences>
In addition to the above the following settings can be made:
A 'lemma list' can be loaded from a file, which can then be used to generate a lemma list instead of a word list. When the lemma list function is used, the 'lemma word form(s)' column will show the words in the corpus associated with each lemma.
A lemma list can be created by specifying the 'lemma entry' follow by '->' followed by one or more 'words' that should be assigned to the lemma separated by one of more non-tokens. See the example below:
be->is, are
play->play, plays, playing, played

Note that in the example above, commas and spaces are assumed to be NOT defined as tokens. For this reason, if the lemma list available on the AntConc webpage is used, a 'dash' needs to be added to the token (word) definition for the lemma list to be processed correctly as the hyphenated words are used to the right of the lemma definition.

The wordlist can be generated using all words, or a specific set of words, or ignoring a certain set of words (a stop list). This is termed the "Wordlist Range". The range of words to be used (or ignored) can be entered directly by the user, or can be stored in files which are then read by AntConc by pressing the 'Open' button. A combination of words in a file and words directly entered by the user can also be used.

<Keyword List Preferences>
In addition to the above the following settings can be made:
As described in the section on the **Keyword List Tool**, to generate a keyword list, the user needs to specify a reference corpus, and a statistical measure of 'keyness'. Although, the default options for the 'keyness' measure and threshold values are recommended, changes can be made in this menu. By choosing the "Show Negative Keywords" option, words that are unusually INFREQUENT in the target corpus compared with the reference corpus will be displayed. Also, here you can choose to use raw reference corpus file(s), or word list(s) that correspond to a reference corpus. In addition, you can swap the main and reference corpora.

SHORTCUTS

Here is a list of Shortcuts that apply to all tools using window panes for results.

CTRL-C = Copies the currently selected text
CTRL-A = Selects all text in the window pane
ALT-A = Selects all text in all window panes showing

Double click = Selects the current word
Triple click = Selects the current line in the window pane
SHIFT-click = Selects continuous lines across all window panes showing
CTRL-click = Selects discontinuous lines across all window panes showing

DELETE = This deletes any selected lines that span across all window panes
INSERT = This keeps any selected lines that span across all window panes, and deletes all others

For any 'spinbox' widgets (e.g. the search term entry box) the 'UP' and 'DOWN' arrow keys on the keyboard can be used to activate the up and down buttons.

NOTES

Saving Results
Results can be either saved to the clipboard, saved to a text file (..txt) or saved to a new window using keyboard commands, the appropriate option in the 'File Menu', or by clicking on the "Save Window" button in each tool, respectively. Also, it is possible to launch multiple clones of AntConc by double clicking on the .exe file.

Comments/Suggestions/Bug Fixes
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 a 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' (see Legal Matter below), but it is important for my funding to hear about any successes that people have with the software. Therefore, if you find the software useful, please send me an e-mail briefly describing how it is being used.

ACKNOWLEDGEMENTS

I would like to say thank you to the users of AntConc who have taken the trouble to e-mail me with feedback on the software and suggestions for improvements and/or changes.
The development of AntConc is supported by a Grant-in-aid for Scientific Research by the Japan Society for the Promotion of Education, Science, Sports and Culture, Japan (No. 16700573), and by a Waseda University Grant for Special Research Projects, Japan (No. 2004B-861).

LEGAL MATTER

AntConc3.1.302 can be used freely for individual use for non-profit research purposes, and freely distributed on the condition that this read me file is attached in an unaltered state. If the software is planned to be used in a group environment, you are required to inform me how the software is to be used, and I will then determine if you can have permission to use it. 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. File Opening: On Linux systems, with the "Open File(s)" option, there appears to be a maximum number of files that can be opened at a single time. If more than this number is selected, none will be read into AntConc. On my computer, the maximum number appears to be around 950. To get around this problem, I advise the user to either select files in two or three batches, or use the "Open Directory" option, for which there is no maximum limit. On WINDOWS, a different file open method is used, but this creates ghost dialog boxes if the main File Open Dialog box is moved. This is another strange bug that I cannot fix.

2. Scrollbar: When a large number of concordance lines are generated (or words or keywords), the scroll bar 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 scrollbars. This is an annoying bug in the scroll bar subroutine (not mine!) and I am waiting for someone to fix this.

3. In the **Word Clusters Tool**, if more than one word is specified as the search term, only the first word will appear on the right, if the "Search Term on Right" option is selected.

4. There is a strange but serious bug that causes the program to sometimes crash if the 'hide tags option' is activated, and then the File View tool is used before using any other tool. I do not understand the cause of the crash, but a solution is being investigated now. Also, the bug seems to only appear on certain machines at certain times. Possibly this is a problem related to using the program with files stored on the Desktop under a Japanese version of Windows. Therefore, I advise that users do not store files under non-Latin1 (Western) path names, for example, the 'Desktop' on Asian systems.

5. Related to point 4, there are many issues with language encodings on non-Western systems. For example, many users have found it difficult to view Chinese characters correctly until I have suggested the correct encoding for their systems. I advise that you load the corpus files and then try each font encoding until one shows characters correctly.

6. There is a report that when the "Word End" sort option is chosen in the Word List tool, the program crashes in some special cases.

7. I have heard reports that AntConc on Macintosh OS X sometimes does not correctly display rare font characters associated with some non-English languages. This appears to be a problem with X11 (the graphic engine used to display the AntConc windows) rather than AntConc itself. I recommend installing the latest version of X11 before using the software.

8. On some Windows systems, if corpus files are placed on the desktop and opened twice, on the second time, if the file name appears in the system help popup, AntConc may suddenly crash. This is *not* a bug in AntConc. It is a Microsoft Windows bug. The same effect can be replicated in a software program like Microsoft Notepad. The workaround is to not save files on the desktop, or just select them quickly before the popup appears.

REVISION HISTORY

3.2.4

This is a minor upgrade fixing two problems (one only applicable to OS X). No new features have been added.
Bug fixes:

1) Modified the OS X version so that lemma list files will load correctly regardless of the OS version. The bug was more general and possibly caused errors when using any single file open file dialog box, e.g., importing settings files, search list files, and word list files. I only heard one or two reports of problems on OS X, so perhaps the problem was introduced with upgrades to recent versions of OS X. It is difficult to tell.
2) Fixed a bug that caused the total number of lemma entries to not be displayed properly.

3.2.3

This is a major upgrade introducing several new features and addressing bugs that appeared in version 3.2.2.1.
New Features:

3) Massively simplified the engine used for sorting and displaying results. This will allow for improved performance later.

4) Introduced a Concordance preference setting allowing tab spaces to be added around the search hit in the KWIC concordance view. Introducing this means that the toggle shortcut key 'x' to show/hide the KWIC search term has had to be removed. This function can now be accessed via the (fixed) menu option. I also hope to reintroduce this toggle shortcut key later.
5) Introduced the option to use word list(s) of reference corpus files instead of directly processing the raw files. This allows fast generation of keywords and also allows users to generate keyword lists even when the reference corpus is not available (but a word list is). See the Keyword List section for further details.
6) Introduced feature to output the counts of types and tokens for each tool when saving results as a text file.
7) Changed some of the default settings in response to user feedback:
8) Concordance Tool sort level s are set as 1R, 2R, 3R.
   The Collocates Tool Stats measure is now displayed and calculated.

Bug fixes:

1) The splash screen that appeared when a user settings file was added has been removed. Now, the words "User Settings" appears next to the title of the software at the top of main window, whenever a user settings file is used.
9) The layout of the main window has been improved so that the main work area is maximized for all sizes of main window.
10) Unicode "Marks" can now be added/removed in the token definition settings
11) Occasionally, the last bar of the progress bar remained blank at the end of processing. This has now been fixed.
12) Under tag settings, the "Hide Header Tags" option will now work for headers than span across multiple lines.
13) The main window and all pop-up windows now appear centered in the display.
14) When viewing the global settings preferences, it was possible to accidentally activate the main window by clicking on it. This has been fixed.
15) The list of files in the left pane was often too narrow to view easily. This has been fixed by introducing a two-pane approach, with the left pane showing the files, and the right pane showing the results and controls. The width of both panes can be adjusted with the default size being the minimum size.
16) It was possible to adjust the main window to be smaller than the default size of 800px by 600px. This caused the positioning of some widgets to become misaligned. Now, 800px by 600px has been set as the minimum size of the main window.
17) Various problems experienced when cutting, copying, and pasting text to/from the Search boxes in each tool have been addressed by removing the "Spinbox" widget type and replacing this with a standard "Entry" widget. The history of previous searches can be now accessed using the "Up" (previous) and "Down" (next) arrow keys.
18) After moving the cursor over the corpus files list while using the File View Tool, it should change into a "Pointing Finger" cursor. However, this cursor shape was returned incorrectly back to a standard "Edit" cursor when one of the files was selected. This has now been fixed.
19) Changed foreground to white when AntConc reports that the KWIC analysis has finished or has no results. This should be easier to read.
20) Fixed a bug that caused the Concordance preference option "Hide search term in KWIC display" to only show the hit. (The opposite of the stated function!). This is now fixed.
21) Fixed some typos in this Read Me document.
22) Fixed a major bug that caused a warning to always appear when using the T-Score statistical measure with the Collocates Tool. Note that the warning could be ignored as the results would always be correct.

3.2.2.1

This is a minor upgrade addressing bugs that appeared in version 3.2.2.

Bug fixes:

1) The legacy character encodings for Japanese, Chinese, Taiwanese, and Korean were inadvertently left out of the compiled executable for version 3.2.2 meaning that files encoded in these languages would not open correctly in AntConc. These have now been added.
2) In the Mac OS X version, the compiled version did not include several important components which are essential for the X11 graphical toolkit to run correctly. These have been added.

3.2.2

This is a minor upgrade addressing several minor bugs that appeared in version 3.2.1. It is also the first version of AntConc to be complied with Perl 5.10.

Bug fixes:

3) Fixed a bug that caused back references to not work correctly when the 'Regex' search option was selected. This now works correctly. However, note that the first back reference should be 2 and not 1 (due to an implicit back reference 1 that holds the entire search result.
4) Fixed error message when the sort by Stat option was selected.

3.2.1

This is a minor upgrade addressing several bugs that appeared in version 3.2.0, as well as introducing a few new features requested by users.

New Features:

1) Better display of long lists of fonts and font sizes in the global options/font menu. Now the lists appear as an easy to navigate list with attached scrollbar.
2) When results windows are saved, the cloned windows now display summary results information.
3) Word range lists can now be used as lemma range lists.
4) New feature allowing tagged data to be searched while remaining hidden. See the tag settings preferences. Pressing CONTROL and the START button (or the ENTER button if the search entry box has the focus) temporarily disables the new feature, allowing the user to switch easily between a 'non-tagged' or 'tagged' display.
5) New options in the tag settings preferences that allow embedded and non-embedded tags to be shown, ignored, or hidden. This enables data of the form of the BROWN and BNC corpora to be processed easily.
6) Improved the updating of the progress bar display, which may also improve the speed of processing in some cases.
7) Improved the images used for icons within the program.

Bug fixes:

1) Fixed bug that caused user defined token definitions containing special regular expression characters from not working properly
2) Fixed bug that caused "Treat all data as lowercase" option to ignore the wordlist range and lemma lists
3) Fixed bug that caused the lemma list "Load" button to not ignore the currently opened file if a new file dialog was opened and then "Cancel" was pressed
4) Fixed bug that caused file searches to not work if the search entry box was blank.

3.2.0

This is a major upgrade with a completely redesigned interface, several new features, and several bug fixes. The new interface follows the basic design used in previous versions, although users should find it 'cleaner' and more intuitive. In particular, all global and tool menu settings have been combined into two groups, where all the related settings can be accessed and adjusted within the same window. This should dramatically improve the usability. All tools now have access to the search engine (including the **Word List Tool** and **Keyword List Tool**) and there is also a new advanced search window that can be used to perform list (file) searches, and searches within a particular context. Due to the nature of the changes, this version will not be compatible with the settings files for previous versions. Another huge change is that this version will run on Macintosh OS X systems.

New Features:

1) Completely redesigned interface
2) Added search and advanced search features to all tools (including the W**ord List Tool** and **Keyword List Tool**).
3) Created new list (file) search available in all tools.
4) Created new context search option in all tools except the **Word List Tool** and **Keyword List Tool** where it has no meaning.
5) Busy cursors are used to indicate when very long sorting operations are being carried out (e.g. when sorting large N-gram list results).
6) Case options affecting whether or not data is converted to lower case are now more intuitive. For example, the 'Case' option in the main window now only affects the operation of the search itself and has no impact on the data under observation. Data can be treated as lowercase (for example in Word list tool) by choosing the 'Treat all data as lowercase' under the appropriate category in the 'Tool Preferences' menu.
7) The number of corpus files (and reference corpus files) being analyzed is now displayed.
8) Correct some mistakes in this readme file
9) My name has been removed from the top of the main window! However, please remember that my name is Laurence with a U if you are ever citing me in your research papers!!
10) Now works with Macintosh OSX

Bug fixes:

1) The program no longer crashes when the 'All Values' option is chosen as the threshold value in Keywords.
2) Negative keywords are now highlighted correctly when the 'Show Negative Keywords' option is selected.
3) The KWIC lines are now aligned correctly even when the hit appears near the very start or end of a file.
4) Collocates frequency values are now correctly calculated even when the span extends further left than the start of the file
5) The action of the 'one word only' wildcard is now more intuitive.
6) Some operations (e.g. creating a word list) now do not crash after restoring the default settings and then performing an operation.

Bug fixes (since beta1 version):

1) The program now (correctly) only shows files that generate hits in the Concordance Plot tool.
2) The sort function in the Keywords Tool now works correctly. In previous versions, even when the 'Frequency' option was selected, the sort would be based on Keyness. Also, in some cases inverted sorting did not work.

Bug fixes (since beta2 version):

1) The program now (correctly) hides the various Concordance Tool panes depending on the chosen Display Options. In the earlier beta versions, the options were ignored.
2) The default file type to use when opening directories now works correctly. In previous beta versions, after hitting the apply button, the default file type reverted back to the .txt type.
3) Fixed a bug that prevented the n-grams option in the Clusters Tool from working when the search term entry box was empty.

Bug fixes (since beta3 version):

1) Fixed bug that caused the program to not be able to open files with non-English names correctly if the full-pathname option was selected. There are potentially many problems with non-English filenames, so I recommend that users use English filenames for their corpus files, and also save them under a pathname which only contains English characters.
2) Fixed bug that caused the 'OR' wildcard to not work correctly if a character other that '|' was user defined.

New Features (since beta4 version):

1) Made some small changes so that the program could be more easily ported to Macintosh and Linux platforms.

3.1.303

This is a very minor upgrade with just one change:

Bug fixes:

1) Corrected problem that caused the No. of Hits to not be indicated correctly in the Concordance Plot Tool display when more than one corpus file was being used.

3.1.302

This is a very minor upgrade with the following changes:

Bug fixes:

1) Corrected problem which caused the program to not launch when the path of the default temporary folder on the system contained non-English characters.
2) (Linux only): Corrected problem that caused the Open Dir menu option to not work correctly.
3) (Linux only): Corrected problem that caused font selections to not work correctly.

New Features: Improved speed and memory handling when calculating collocates. Over 10 times faster than in previous versions (including version 3.1.3).

3.1.3

This is a minor upgrade containing an important bug fix that prevented files with non-ASCII filenames being used. There are also some major performance improvements. For example, n-grams will now be processed over 10 times faster on small corpora and many more times faster on larger corpora. A list of all important changes is below:

New Features:

The history feature for search term entries has been changed. I have heard two reports of the 3.1.2 version not starting on computers. Hopefully, this change will allow the program to start correctly on all machines.

The performance of tools such as Collocates, Clusters and N-grams, has been significantly improved. (Over 10 times faster on small corpora and many more times faster on larger corpora.)

The Open Dir option now open files in all sub-directories too.

The program will automatically look for a user defined settings file named "antconc_settings.ant" in the directory where the program is saved. If this file is found, this settings file will be used instead of the default

settings. Also a splash window will be displayed when a user settings file is found. If no file is found, the default settings will be used. In this case, no splash file will be shown. This feature allows users to save their setting preferences and use them again without having to load the preferences each time.

Bug Fixes:

1) Files with non-ASCII file names were incorrectly processed preventing them being used. Introduced in version 3.1.2. Fixed.
2) In the N-grams tool, non-word units (e.g. spaces) at the beginnings of lines were treated as words. This caused some n-grams of n-1 size to also appear in the results list. Fixed
3) When "ALL" was selected as the file type option in Open Dir, sub-directories were also displayed even though these could not be opened or processed. Fixed
4) In some tools, non-ASCII filenames were sometimes displayed incorrectly, even after the correct encoding was chosen. Fixed.
5) If 'Exit' was chosen from the File menu options, the program exited without a warning. Fixed.
6) The "Add" button for the 'Add Word' option was accidentally deleted from the Word List Preferences menu. This meant that words could only be added by hitting the return button. Fixed.

Changes:

1) The default file type for Open Dir has been changed to .txt
2) The 'Directory' displays in both the main window and keywords list preferences window have been deleted. As directories and sub-directories can now be used, this feature has become redundant and perhaps confusing.

3.1.2

This is a semi-major upgrade containing a new Lemma List tool, and numerous bug fixes and interface improvements. A list of all major changes is below:

1) Binding to plot canvas lines to allow jumping to hit in file (same as in concordance view).
2) Binding to file list allowing the user to simply click on the file name when using the File View tool to view the file
3) History feature for all entries (use up and down arrows on keyboard)
4) Much faster concordancer processing (especially sorting... up to 10 times faster)
5) Ability to clear content of tools (Clear Tool, Clear All Tools, Clear All Tools and Files)
6) Ability to save list of all loaded files in settings file
7) Redesigned "one or more words" and "any one word" wildcards to act more sensibly. Now, the wildcards incorporate any non-tokens between possible words. Therefore, use "is#dog" and "is@@dog" etc. to search for the hit "is a dog" in the sentence "This is a dog."
8) New Lemma tool allowing all lemmas of a word list word to be displayed. Note: A lemma list is required.
9) Added bindings to Wordlist Range list to allow words to be deleted (<Delete> button) or kept (<Insert> button) as with other tools.
10) Swap button to switch the main and reference corpora when doing keyword analyses. (Accessible from the Keyword preferences Menu).
11) Fixed serious bug that caused the program to freeze or crash when an incorrectly formed regular expression was used as a search term.
12) Fixed serious bug that caused some word, keyword and font encodings to not be loaded correctly from the user settings files
13) Fixed bug that caused the 'Reset' buttons to not work correctly.
14) Fixed bug that caused 'user defined token definitions' to not be loaded correctly.

3.1.1

This is a minor upgrade although it contains a new T-Score stats feature, and has a new editing feature, allowing you to select and then delete, or keep certain results lines. A list of all major changes is below:

1) Added T-Score to stats measure in the Collocates Tool.
2) Changes the Collocates Tool menu to allow one a several statistical measure to be chosen.
3) Adding a feature so that any selected results lines can be either deleted or kept (after deleting all others).
4) The tag settings options in the File Settings menu have now been moved to a separate menu.
5) Labels in the software (in particular those related to language encodings) have been slightly altered to make them clearer.
6) Improved the warnings given when the Collocates Tool is not in sync with the Word List Tool, during the calculation of collocate significance.
7) Changed the internal workings to allow LINUX and WINDOWS ports of AntConc to be easily created from the same source code. This should allow future versions to be released at the same time for both operating systems.
8) Fixed a bug that caused the Transition probabilities calculation to not initialize correctly. This meant that although the first calculation was correct, all later calculations would be give false measurements.

3.1

AntConc 3.1 is a major, major upgrade. I was very tempted to name this AntConc 4.0, but in the end chose to keep it in the 3s. To list the improvements and changes will take up many pages, so only the major differences are listed here in no particular order. Of course, with so many changes, there will inevitably be new bugs that have crept into the program. However, I hope you will find that this release is an improvement over previous releases.

1) Implementation of new **Collocates Tool**.
2) Total reworking of the programming underneath all tools for performance enhancements.
3) Adding of "sort" features to many tools, enabling quick and efficient re-ordering of results information.
4) Sensible treatment of word case in all tools. Now searches are truly case sensitive or insensitive, and results can be displayed according the
1) case setting.
5) All settings can now be imported or exported to a file. This allows users to easily upload their preferred settings, avoiding the need to constantly make changes to the defaults. This is a HUGE improvement (I think!).
6) The "Open Dir" option has been restored. File types that will be read into AntConc via the "Open Dir" option can now be specified in the "File Settings" menu.
7) AntConc is now FULLY Unicode compliant. It should work with any language in the world. I am very interested to hear how AntConc works in different languages. Please let me know. Token definitions can be made using all Unicode character classes, or a user defined token definition can be made. However, I would advise against using the user defined token definition, as it is so easy to overlook possible characters that might need to be processed.
8) As AntConc is now fully Unicode compliant, all possible encodings of characters are listed under the Encodings menu option. For English and all other Western Language, the default option, iso-8859-1 should work fine. Note that many Windows systems actually save files in the cp-1252 encoding which resembles iso-8859-1 but is a little different, just to confuse people! Users with Japanese texts will probably find shifts to be the best option.
9) Some users have said they could not open the font selection windows. This problem has been fixed.
10) Many new ordering options have been added, for example, from word ends.
11) The overall design of the interface has been improved, allowing for new options to be easily added without cluttering the display. Also most menu options now use simple checkboxes instead of the confusing "yes", "no" radio boxes used in previous versions.

12) Some of the tool names have been changed. In particular "lexical bundles" are now called "N-Grams", which is a more familiar term.
13) The Word Stems feature has been removed. This only worked with English texts, and the results were of questionable value. If a user would like to see the feature in a future version, please let me know.
14) The size of the program has increased by around 1MB. This is so that all the fonts and encodings for true internationalization of AntConc can be included.
15) This read me file has been largely rewritten. I hope there are fewer spelling mistakes this time!
16) Many, many small bugs have been fixed, mainly related to the way the interface responded to user actions.

3.0.1

A few bugs in 3.0 have been corrected. Also, the View Files tool has been redesigned to work much, much faster. To do this, the program now does no processing at the end of lines. Therefore, ambiguous line endings will stay ambiguous!

1) Improved performance of View Files tool
2) Corrected bug which caused the first file in Keywords to appear even after the list of files is deleted from the system.
3) Added feature to allow files to be added to the system from different folders. Now, if a file is opened into the system it will be added to the current list without deleting the previous list files. The same applied to files added in the Keywords List preference menu.

3.0

There have been so many changes that it is almost impossible to list them. Here are some of the more obvious differences between AntConc3.0 and previous versions.

1) Wildcard implementation
2) Word stem implementation
3) Save current window feature added to all tools (except 'Concordance Search Term Plot')
4) Three levels of sort implemented
5) Complete redesign and implementation of the 'View Files' tool, making it generate results much faster. (But it is still too slow).
6) Hyperlinks added to the results of all tools.
7) Rearrangement of the menu options
8) Redesign of the file list window, allowing one or more files to be closed.
9) Redesign of the main results windows, placing information in different window 'panes'. Each pane can be resized, or hidden.
10) Redesign of the data selection methods. The selection methods now comply with most other software products.
11) Complete redesign of many sub-routines enabling quicker processing.
12) Many, many bugs found and corrected. Please tell me if you find a bug because I WILL correct it.

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 an 800 x 600 or higher resolution monitor.
2) Corrected several spelling errors in this read me 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 read me 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

1) First, progress indicators were added to 'pages' of AntConc.
2) Second, a new file view feature was added to view target files in their original state.
3) Third, a keyword generation feature has been added using log-likelihood and chi-squared methods.
4) Finally several bugs were found, in particular, bugs centered on 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

1) 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.
2) Numerous bugs centered on the way the software entered a 'Busy' mode were corrected.
3) The main core of the software was also updated resulting in a quicker, 'cleaner' processing of the data.
4) Performance improvements should be noticed as a result.

2.2.3

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

2.2.2

1) 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

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

1) 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

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

2.0

1) Added new sort feature, for rearranging concordance lines.
2) 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

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

1.0
First version