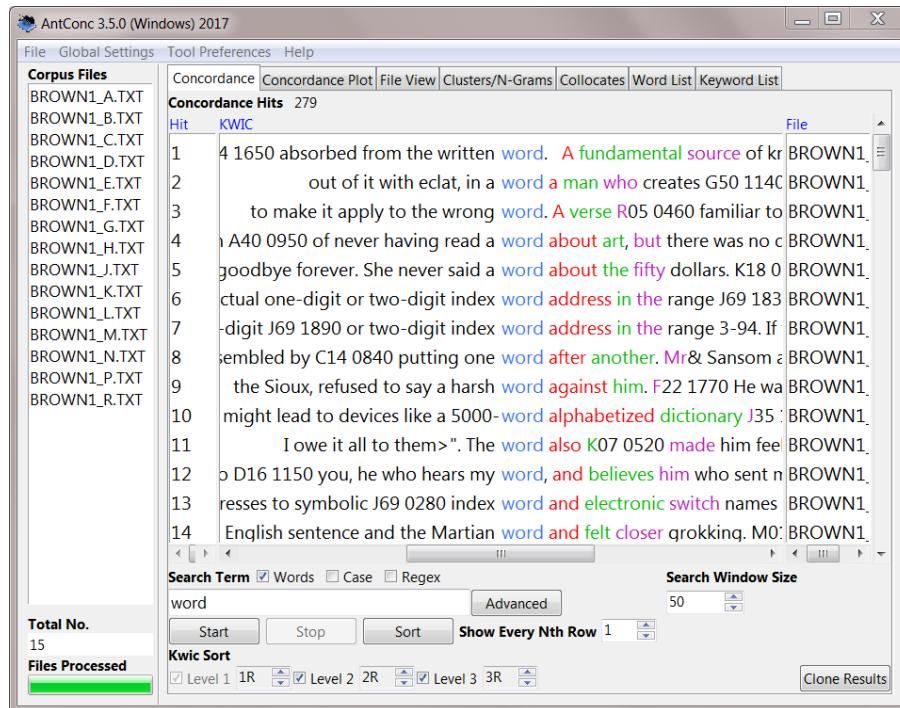# AntConc (Windows, Macintosh OS X, and Linux)
Build 3.5.1 (January 22, 2017)

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
Help file version: 001.

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, Linux Mint). It is developed in Perl using various compilers to generate executables for the different operating systems.

Getting Started (No installation necessary)

Windows
On Windows systems, simply double click the *AntConc* icon and this will launch the program.

Macintosh OS X
On Macintosh systems, simply double click the *AntConc* icon and this will launch the program.

Linux
On Linux systems, change the permissions to allow AntConc to be run as an executable file. Next, double click the *AntConc* executable and it will launch.

Caution: Do not place user settings files from older versions of AntConc in the same folder as new versions. This can cause unexpected problems and may even prevent AntConc from starting. It is recommended that you delete your earlier settings file and export it again from the AntConc file menu.
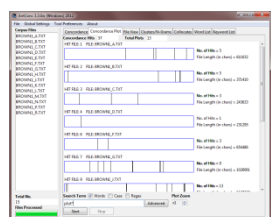
Overview of Tools

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.
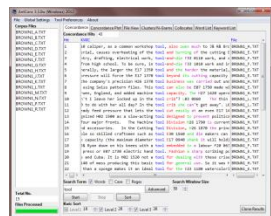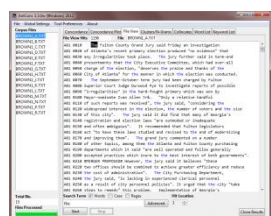

Concordance Tool:
This tool shows search results in a 'KWIC' (KeyWord In Context) format. This allows you to see how words and phrases are commonly used in a corpus of texts.
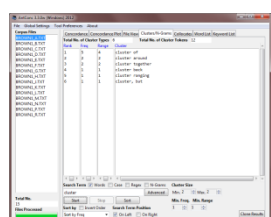

Concordance Plot Tool
This tool shows search results plotted as a 'barcode' format. This allows you to see the position where search results appear in target texts.


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
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 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.
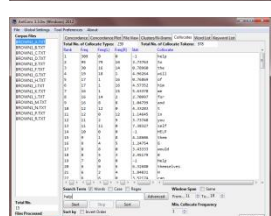
Collocates:
This tool shows the collocates of a search term. This allows you to investigate non-sequential patterns in language.
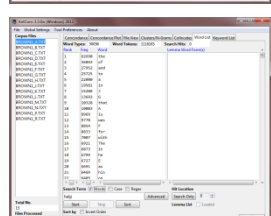
Word List:
This tool counts all the words in the corpus and presents them in an ordered list. This allows you to quickly find which words are the most frequent in a corpus.
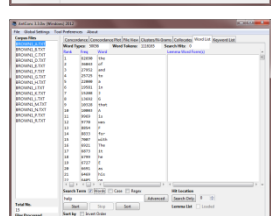
Keyword List:
This tool shows the which words are unusually frequent (or infrequent) in the corpus in comparison with the words in a reference corpus. This allows you to identify characteristic words in the corpus, for example, as part of a genre or ESP study.

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.

The following steps produce a set of concordance lines from a corpus and demonstrate the main features of this tool.

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 search box.
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) Choose a 'thinning' value via the "Show Every Nth Row" option to limit the number of rows presented.
5) 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.
6) Use the Kwic Sort options to rearrange the concordance lines at three different levels. 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.
7) Click on the 'Sort' button to start the sorting process.
8) 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 you to view the search term hit as it appears in the original file via the File View Tool (see below).
9) Click on the "Clone Results" button to create a copy of the results so that different sets of results can be compared.
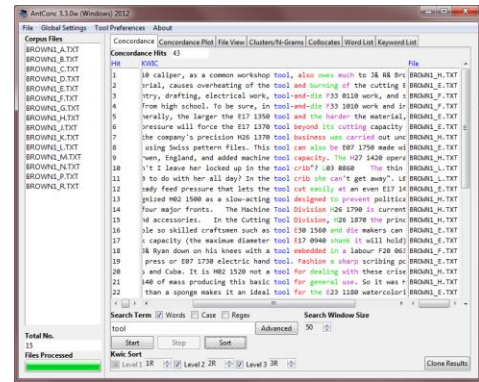
The total number of concordance lines generated (Concordance Hits) is shown at the top of the tool window. This number will flash with the word "FINISHED" when processing has been completed, and will flash with the word "NO HITS", if not hits are generated for a particular search term.
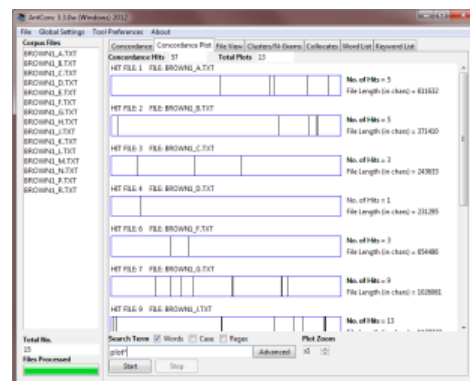
Search terms can be specified as being "words" (default) or "character strings" by activating or deactivating the "Word" search term option. Also, searches can be either "case insensitive" (default) or "case sensitive" by activating or deactivating the "Case" search term option. Searches can also be made using full regular expressions by activating 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 (e.g., http://www.regular-expressions.info/quickstart.html). AntConc supports Perl regular expressions.

By clicking on the "Advanced Search" button, more complex searches become possible. The first advanced search option allows you to import a set of search terms, either by typing them one per line, or by loading in a list of search terms from a file. Here, each line will be treated as a separate search term. This feature allows you to use a large set of search terms without having to re-type them each time. The second advanced search option allows you 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 set 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 concordance search results plotted in a 'barcode' format, with the length of the text normalized to the width of the bar and each hit shown as a vertical line within the bar. This allows you to see the position where search results appear in target texts. The tool also allows you 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 Tool is in determining where specific content words appear in a technical paper, or where an actor or story character appears during the course of a play or novel.

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 "Plot Zoom" buttons. A "Show Every Nth Row" thinning option is also available.

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 you to view the search term hit as it appears in the original file via the File View Tool (see below).

Search terms can be specified as being "words" (default) or "character strings", and searches can be "case insensitive" (default), "case sensitive," or "Regex" based. Advanced searches are also available. For details see the Concordance Tool explanation.

File View Tool

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

The following steps produce a view of the original file and demonstrate the main features of this tool.
1) Select a file to view in the "Corpus Files" list on 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) Click on the highlighted text to generate a set of KWIC lines using the highlighted text as the search term.
6) Click on the "Clone Results" button to create a copy of the results so that different sets of results can be compared.
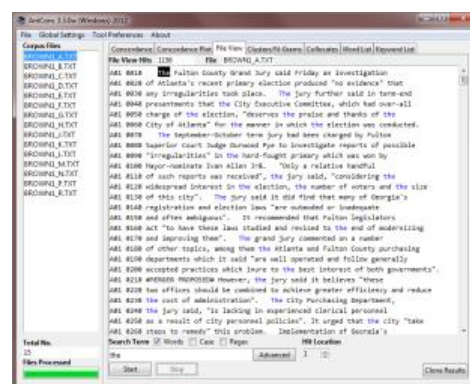
Search terms can be specified as being "words" (default) or "character strings", and searches can be "case insensitive" (default), "case sensitive," or "Regex" based. Advanced searches are also available. For details see the Concordance Tool explanation.

The following shortcut is unique to the File View Tool:
CTRL-Click = Jumps to the nearest hit in the window
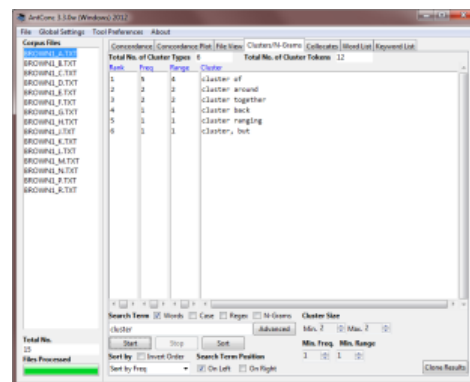
Clusters/N-Grams Tool

The Clusters Tool
This allows you to search for a word or pattern and group (cluster) the results together with the words immediately to the left or right of the search term. In effect it summarizes the results generated in the Concordance Tool or Concordance Plot Tool.

The clusters can be ordered by frequency, the start or end of the word, the range of the cluster (number of files in which the cluster appears), or the probability of the first word in the cluster preceding the remaining words. All list orderings can also be inverted by activating the "Invert Order" option. Also, you 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 (default) 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.)

The following steps produce a set of cluster results and demonstrate the main features of this tool.
1)	Choose the appropriate ordering options (see above for details).
2)	Press the 'Start' button. At any time, the generation of the clusters list can be halted using the 'Stop' button.
3)	Click on the cluster to generate a set of KWIC lines using the text as the search term.
4)	Click on the "Clone Results" button to create a copy of the results so that different sets of results can be compared.
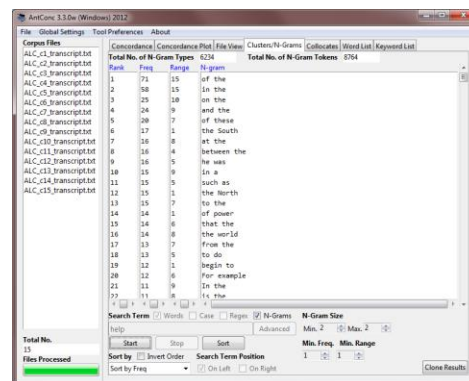
The N-Grams Tool
This allows you to scan the entire corpus for 'N' word clusters (e.g. 1 word, 2 words,…). This allows you to find common expressions in a corpus. For example, n-grams of size 2 for the sentence "this is a pen" are 'this is', 'is a' and 'a pen'.

All ordering options available in the Clusters Tool are also available in the N-grams tool. You can also select the minimum and maximum size (number of words) in each n-gram, and the minimum frequency and range of n-grams displayed.

The following steps produce a set of N-gram results and demonstrate the main features of this tool.
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)	Click on the n-gram to generate a set of KWIC lines using the text as the search term.
5)	Click on the "Clone Results" button to create a copy of the results so that different sets of results can be compared.

In both the Clusters Tool and N-Grams Tool, search terms can be specified as being "words" (default) or "character strings", and searches can be "case insensitive" (default), "case sensitive," or "Regex" based. Advanced searches are also available for the Clusters Tool. For details see the Concordance Tool explanation. A number of menu preferences are available with this tool. (See below).

Collocates Tool

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

The collocates can be ordered either by total 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, you 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.
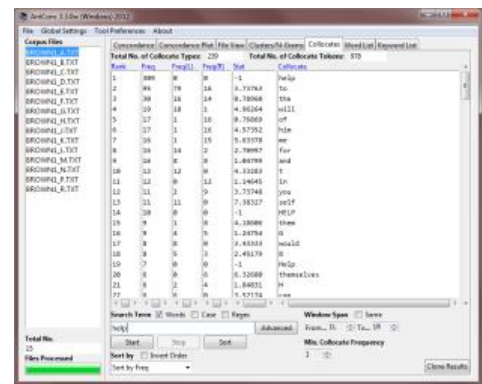
Statistical Measures:
Mutual Information (MI): Using equations described in M. Stubbs, Collocations and Semantic Profiles, Functions of Language 2, 1 (1995)
Log Likelihood: Using equations described in P. C. Gomez, Statistical Methods in Language and Linguistic Research. Milton Keynes, UK: Equinox Pres. (2013).
T-Score: Using equations described in M. Stubbs, Collocations and Semantic Profiles, Functions of Language 2, 1 (1995)

The following steps produce a set of collocate results and demonstrate the main features of this tool.
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)   Click on one of the collocates to generate a set of KWIC lines using the text as the search term.
4)   Click on the "Clone Results" button to create a copy of the results so that different sets of results can be compared.

Search terms can be specified as being "words" (default) or "character strings", and searches can be "case insensitive" (default), "case sensitive," or "Regex" based. Advanced searches are also available for the Collocates Tool. For details see the Concordance Tool explanation. A number of menu preferences are available with this tool. (See below).
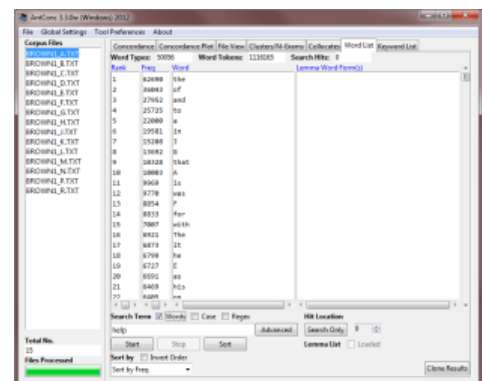
Word List Tool

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 ordering 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.

The following steps produce a word list and demonstrate the main features of this tool.
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)   Click on the word to generate a set of KWIC lines using the text as the search term.
4)   Click on the "Clone Results" button to create a copy of the results so that different sets of results can be compared.

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.

The following steps produce a keyword list and demonstrate the main features of this tool.
1) Select a set of target files.
2) Go to the 'Preferences' menu and chose the 'Keyword Preferences' option.
3) Choose the keyword generation method (a statistical measure) to calculate the 'keyness' of the target file words. The default setting of Log Likelihood is recommended.
4) Choose a significance value (p value) for keyness statistic.
5) Choose an effect size measure to rank the keywords.
6) Choose a threshold for the number of keywords to be displayed.
7) Choose whether or not to view 'Negative Keywords' (target file words with an unusually low frequency compared with the frequency in the reference corpus)
8) 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 as follows with the values in the rows separated by tabs:

RANK	FREQUENCY	WORD.
1	12838	the
2	11289	a
3	8583	of
...		

Note that blank lines and lines beginning with # will be ignored. Also, *AntConc* will check that the file(s) are correctly formatted and report any errors.
9) Load the reference corpus of text (.txt) files, in the same way that the target files are chosen.
10) 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.
11) Click 'Apply' in the Keyword Preferences menu and return to the main Keywords window.
12) Choose suitable options for displaying the list of generated Keywords (in a similar manner to the options for generating a Word List).
13) Press the 'Start' button. At any time, the generation of the keyword list can be halted using the 'Stop' button.
14) Click on the keyword to generate a set of KWIC lines using the text as the search term.
15) Click on the "Clone Results" button to create a copy of the results so that different sets of results can be compared.

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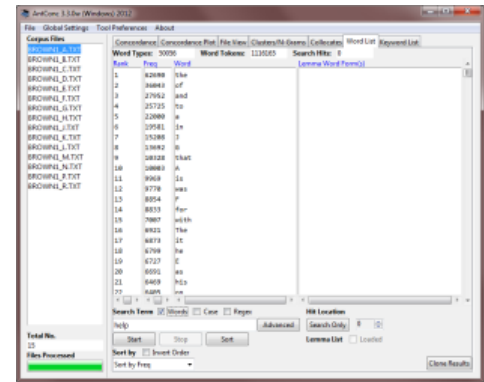
MENU OPTIONS

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*:

> **<Character Encoding>** *AntConc* is fully Unicode compliant, meaning that it can handle data in any language, including all European languages and Asian languages. The character 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. On the other hand, Japanese texts are usually encoded in shiftjis. By specifying the correct encoding, data from all languages can be processed correctly within *AntConc*. The default is Auto-Detect, meaning it will try to detect the encoding that you are using – defaulting to Unicode UTF-8, if it does not know. UTF-8 is an international standard designed to display all characters of the languages of the world in a single encoding. I recommend you use this encoding if you create any corpus.

> **<Colors>** In the Color Settings category, you can edit the colors used to display results and other information.

> **<Files>** In the File Settings category, you can choose to display the full path of a file or just the name.

> **<Fonts>** In the Font Settings category, you can edit the font types, sizes, and styles used to display file names, results, and the search term.

> **<Tags>** In the Tag Settings category, you can choose to display or hide any tags that are contained in the corpus files. You can also choose to search with tags but hide them in the results display. Embedded tags (e.g. book_NN1), non-embedded tags (e.g. <noun>book</noun>, and header tags can be shown or hidden by activating or deactivating the option.

> **<Token (Word) Definition>** In the Token (Word) Definition category, you 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, dashes and so on. *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 option refers to 'letters' in the broadest sense. Letters, for example, include all English letters (a to z, A to Z) but also all Japanese 'letter' characters. It is also possible to define your own "token" definition, or append characters to the standard classes.

> 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/

<Wildcard Settings> In the Wildcard Settings category, you 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. There are special wildcards to deal with whitespace issues.

<TOOL PREFERENCES>
Each tool (with the exception of the File View Tool) has a preferences category, where settings can be fine-tuned. All tool preference categories (with the exception of the Concordance Plot Tool and the File View Tool) allow you to show or hide the different frames in which the results are displayed. For example, you can choose to hide the frame showing file names in the Concordance Tool display window.

<Concordance> In addition to the above, the following settings can be made:
● Using the "Treat case in sort" option causes capitalized words to appear before lower-case words.
● Using the "Sort by characters instead of words" option, it is possible to arrange the results by CHARACTERS to the left or right of the first letter of the search term. This makes it possible to search for spelling differences.
● Using the "Hide search term in KWIC display" option, search term can be hidden in the KWIC lines, allowing instructors to quiz students on possible words to fit the gap.
● Using the "Put delimiter around hits in KWIC display" option (the default), the chosen delimiter character is added 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.
● Use the "Delimiter" option to select the delimiter character.
● Use the "Line break replacement" option to select a character to replace line breaks with.

<Concordance Plot> In addition to the above, the following settings can be made:
● Using the "Plot Length Options" allows the plots to be normalized to a fixed length or plotted at a relative length to the number of characters in the files..
● When the normalized length option is selected in the "Plot Length Options", the normalized plot maximum length determines the length of the plot.
● When the relative length option is selected in the "Plot Length Options", the relative plot scale factor determines the scaling of the relative plots.
● If the "Plot Hit Labels" option is selected, hits within each plot will be given a hit number label.
● If the "Plot Show Blank Files" option is selected, files with blanks will be shown. Otherwise, they will be hidden.

<Clusters/N-Grams> In addition to the above, the following settings can be made:
● Using the "Treat all data as lowercase" option (default) causes all words to be transformed to lower-case words. This is useful to get accurate counts of words in certain cases.
● Using the "Treat case in sort" option causes capitalized words to appear before lower-case words.

<Collocates Preferences> In addition to the above, the following settings can be made:
● Use the "Selected Collocate Measure" to choose the statistical measure for measuring collocate strength. Currently, two statistical measures can be used: Mutual Information (MI) and T-Score. See the tool explanation above for references to the statistics.
● Using the "Treat all data as lowercase" option causes all words to be transformed to lower-case words. This is useful to get accurate counts of words in certain cases.
● Using the "Treat case in sort" option causes capitalized words to appear before lower-case words.

<Word List Preferences> In addition to the above the following settings can be made:
● Using the "Treat all data as lowercase" option causes all words to be transformed to lower-case words. This is useful to get accurate counts of words in certain cases.

- Using the "Treat case in sort" option causes capitalized words to appear before lower-case words.
- Use the lemma list options to select a lemma list. 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 commas or tabs (with or without additional space around the words). See the example below:

      be->is, are   be->is\<tab\>are
      play -> play, plays, playing, played   play->play \<tab\> plays \<tab\> playing \<tab\> played

- Using the "Use word list" option under the "Target Corpus" heading, a pre-built word list contained in a single file can be loaded after selecting the file by clicking on "Load Word List".
- Using the "Word List Range" option, a wordlist can be generated using all words ("Use all words"), or a specific set of words ("Use specific words below"), or ignoring a certain set of words ("Use a stoplist below"). The range of words to be used (or ignored) can be entered directly, 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 can also be used.

\<Keyword List Preferences\> In addition to the above the following settings can be made:
- Using the "Treat all data as lowercase" option causes all words to be transformed to lower-case words. This is useful to get accurate counts of words in certain cases.
- Using the "Treat case in sort" option causes capitalized words to appear before lower-case words.
- Use the "Show Negative Keywords" option to view words that are unusually INFREQUENT in the target corpus compared with the reference corpus.
- Use the "Keyness Values" options to choose the statistical measure for measuring keyword strength, the statistical threshold used to decide a cutoff point, the effect size measure used to rank the keywords, and the effect size threshold to decide the cutoff of the effect size. Currently, two statistical measures can be used: Chi-Squared and Log-Likelihood (the default). The default option for the 'keyness' measure is recommended, however, you may want to experiment with different effect size measure depending on the corpus you are using. The selection of statistical measures and effect size measures are inspired by the work of Andrew Hardie of Lancaster University.
- Use the "Use raw file(s)" option to use raw reference corpus file(s) as the reference corpus.
- Use the "Use word list(s)" option to use word list(s) that correspond to a reference corpus. The word list(s) should be formatted as described in the tool explanation.
- Click the "Add Directory" or "Add Files" buttons to select the reference corpus files.
- Click the "Swap with Target Files" button to swap the main and reference corpora. Note that this will only make sense when raw corpus files are being used.

\<HELP\>
The help menu provides access to this detailed readme file. It also shows some general information about AntConc, including the current version number and date of release.

SHORTCUTS

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

CTRL-C (COMMAND-C on Macintosh OS X) = Copies the currently selected text
CTRL-A (COMMAND-A on Macintosh OS X) = Selects all text in the window pane
ALT-A (OPTION-A or COMMAND+OPTION-A on Macintosh OS X) = 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 (BACKSPACE on Macintosh OS X) = This deletes any selected lines that span across all window panes
INSERT (SHIFT + BACKSPACE on Macintosh OS X) = 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 saved to the clipboard, saved to a text file (.txt), saved to a postscript file (.ps) -for the Concordance Plot tool, 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 you find 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.

CITING/REFERENCING ANTCONC
Use the following method to cite and reference AntConc according to the APA style guide:

Anthony, L. (YEAR OF RELEASE). AntConc (Version VERSION NUMBER) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

For example if you download AntConc 3.3.2, which was released in 2012, you would cite/reference it as follows:
Anthony, L. (2012). AntConc (Version 3.3.2) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Note that the APA instructions are not entirely clear about citing software, and it is debatable whether or not the "Available from ..." statement is needed. See here for more details:
http://owl.english.purdue.edu/owl/resource/560/10/

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.

KNOWN ISSUES

None at present.