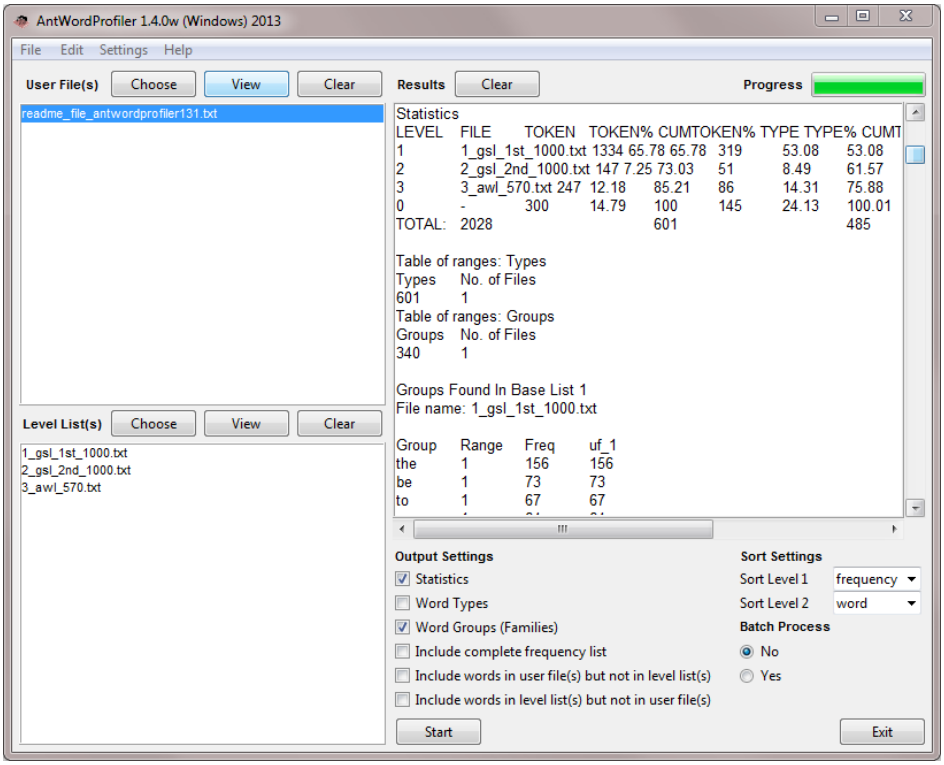
AntWordProfiler (Windows, Macintosh OS X, and Linux)

Build 1.4.1.0

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 (October 5, 2014).



Introduction

AntWordProfiler is a freeware, multiplatform tool for carrying out corpus linguistics research on vocabulary profiling. It runs on any computer running Microsoft Windows (tested on Win 7), Macintosh OS X computers (tested up to OS X 10.9 Mavericks), 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 *AntWordProfiler* icon and this will launch the program.

Macintosh OS X

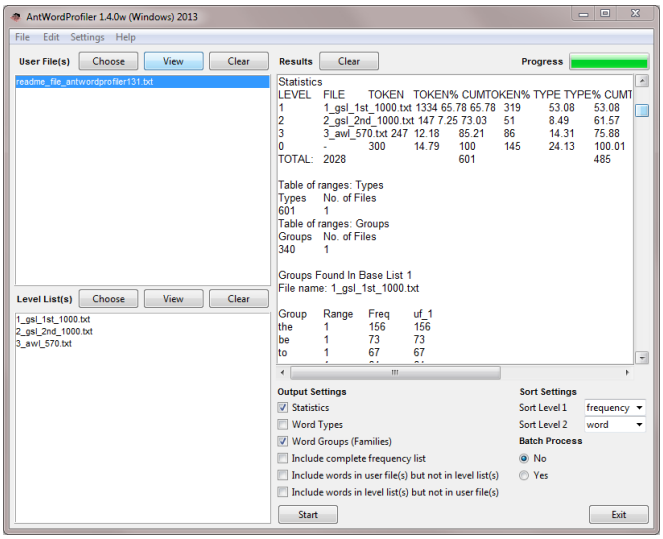
On Macintosh systems, simply double click the *AntWordProfiler* icon and this will launch the program.

Linux

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

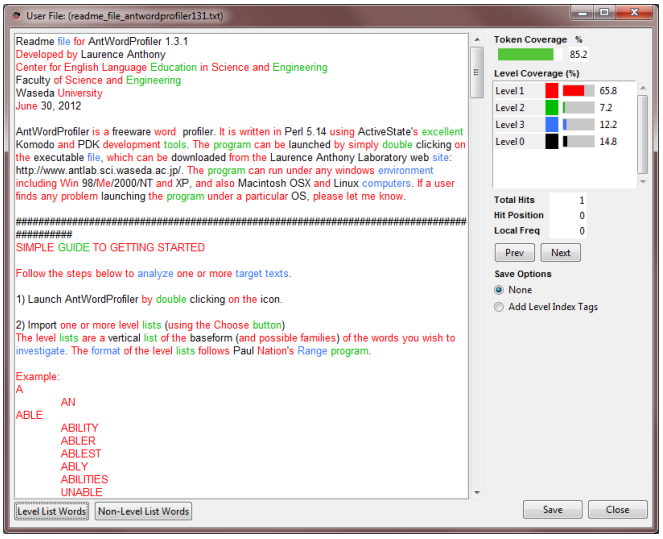
Overview of Tools

AntWordProfiler contains two tools. The main tool is a general vocabulary profiling tool. It appears in the main window of the program.



Vocabulary Profile Tool:

This tool shows allows you to generate vocabulary statistic and frequency information about a corpus of texts loaded into the program. It compares the files against a set of vocabulary level lists that can be plain frequency lists of 'family lists' based on the research of Paul Nation.

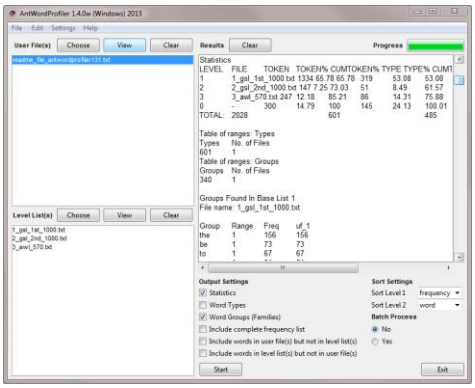


File Viewer and Editor Tool

This tool allows you to view an individual user file and highlight the different levels of vocabulary in the file using a color coding. It also shows the overall coverage of different vocabulary levels.

The tool allows you to edit the file and immediately see the effect of that editing on the vocabulary level of the words and the text as a whole. This is useful to simplify the text. If the thesaurus option is activated in the main menu, vocabulary items with thesaurus entries will be highlighted and equivalent items at a lower (or higher) vocabulary level can be chosen as replacements.

Vocabulary Profile Tool



This tool shows allows you to generate vocabulary statistic and frequency information about a corpus of texts loaded into the program.

To generate a profile, you need to perform the following actions:

1) Launch *AntWordProfiler* by double clicking on the icon.
2) Check which level lists to use. Three baseword lists are included with the program by default. These are:
   1_gsl_1st_1000.txt
   2_gsl_2nd_1000.txt
   3_awl_570.txt

The level lists are a vertical list of the base form (and possible families) of the words you wish to investigate. The format of the level lists follows Paul Nation's *Range* program. The lists can be in any case (upper, lower, or mixed). *AntWordProfiler* will convert all list members to lower case before processing them.

Example:

A
    AN
ABLE
    ABILITY
    ABLER
    ABLEST
    ABLY
    ABILITIES
    UNABLE
    INABILITY
...

NOTE: The baseword will also be included as a family member. In other words, the list above will be treated as equivalent to the list below.

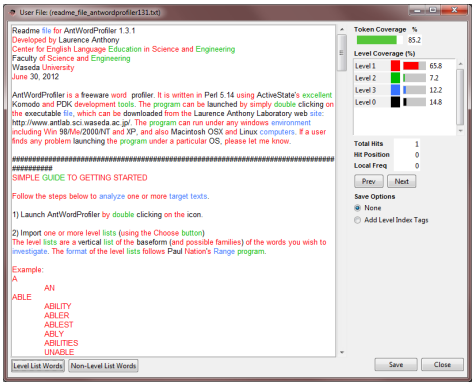
A
    A
    AN
ABLE
    ABLE
    ABILITY
    ABLER
    ABLEST
    ABLY
    ABILITIES
    UNABLE
    INABILITY
...

Also, a simple list of vocabulary items can also be used. Here, *AntWordProfiler* will consider each word to be a single family with no other members.

A
AN
ABLE
ABILITY
ABLER
ABLEST
ABLY
ABILITIES
UNABLE
INABILITY
...

3) Import one or more target user files using the Choose button or the file menu option.
4) Select one or more of the output data options
   a) Statistics: this will produce a set of statistics about the target files.
   b) Word Types: this will produce a list of word types in each target file that are also in the level lists
   c) Word Groups (Families): this will produce a list of base word groups (families) in each target file that are also in the level lists
   d) Include a complete frequency list of all words in the target file together with some valuable statistics
   e) Include words not in list(s): this will include words not in the level list(s) as part of the results
   d) Include words not in file(s): this will include words not in the target file(s) as part of the results
5) Decide on a sort setting:
   a) Sort by word spelling
   b) Sort by word range (document frequency)
   c) Sort by word frequency
6) Decide how to process the files
   a) All files together (Batch Process = No)
   b)) all files separately (Batch Process = Yes)
7) Press START to generate the results:
8) Select a save option from the FILE menu to save the results
   a) Paul Nation's Range program format
   b)) tabbed-spaced text format

File Viewer and Editor Tool



This tool allows you to view the vocabulary profiler of individual user files and edit (e.g., simplify) them.

Follow the steps below to view (and simplify) a target file.

1) Select a target file from the list of User File(s)
2) Click on the 'View' button
3) Click the "Level List Words' button to color-highlight the words included in the level lists.
4) Click the "Non-Level List Words' button to color-highlight words not included in the level lists.
5) Use the 'Prev' and 'Next' buttons to navigate to words that you want to edit or simplify.
6) If the <Thesaurus> option is activated in Menu->Settings->Thesaurus Settings, click the 'Edit' button to view synonyms of selected words at different levels. Words that have a thesaurus entry will be underlined. Use the 'Replace' or 'Replace All' buttons to edit the target text.
7) Choose one of the 'Save Options' to decide how edited target texts will be saved.
   The 'None' will leave the text unchanged. The 'Add Level Index Tags' will tag all words with an underbar followed by a level number.
8) Click the 'Save' button to save the edited target text.

Notes: When using the File View tool, the following features are available:
Click [Return] to choose the selected thesaurus entry in the thesaurus viewer.
Hover over the Level labels in the 'Level Coverage' display to see the full pathname.

MENU OPTIONS

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

<FILE>
Options here relate to reading files into *AntWordProfiler* 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.

<Edit>
Options here relate to cutting, copying, pasting, and deleting the text that you have selected. Note that a few of the shortcuts related to editing are non-standard. I hope to address this problem in the future.

<SETTINGS>
Categories here will have an effect on multiple tools in *AntWordProfiler*:

<Global Settings>

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

<File Settings> In the File Settings category, you can choose to display the full path of a file or just the name. You can also choose whether or not to automatically load the default level lists at startup.

<Tag Settings> In the Tag Settings category, you can choose to display or hide information enclosed in angle (<>) tags.

<Token Definition Settings> In the Token Definition Settings 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 in the word definition. *AntWordProfiler* is fully Unicode compliant, meaning that it can handle data in any language, including all European languages and Asian languages. The default option is that a word must contain Unicode-defined 'letters' and be followed optionally by Unicode-defined numbers (but not preceded by them). Here, 'letters' are used in the broadest sense, for example, they include all letters in the French, German and even Japanese language. Similarly, 'numbers' are equally broad and cover numbers used in all the world's languages. It is possible for you to define your own "token" definition by specifying a suitable regular expression. A simple example is [a-zA-Z0-9]+ to capture 'words' containing a string of letters of the alphabetic and numbers.

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

<Thesaurus Settings>

Here, you can choose to activate the internal thesaurus to help you edit (e.g., simplify) target texts. User thesauruses can also be loaded, and a reference list of word families can be loaded to guide the simplification.

SHORTCUTS

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

[ALT .] (alt period) = makes the text larger
[ALT ,] (alt comma) = makes the text smaller

In the target file VIEW window , the following shortcuts will work:
[ALT n] moves the cursor to the 'n'ext color-highlighted word
[ALT p] moves the cursor to the 'p'revious color-highlighted word
[ALT l] activates the "Highlight Level List Words" button
[ALT h] activates the "Highlight Non-Level List Words" button
[ALT u] makes the non-highlighted words appear underlined
[ALT b] makes the highlighted words appear in black
[ALT g] makes the highlighted words appear in grey
[ALT w] makes the highlighted words appear in white

Other standard shortcut keys, such as CRTL-X (cut), CTRL-C (copy), CTRL-V (paste), CTRL-Z (undo), CTRL-Y (redo), should work as expected.

NOTES

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.

CITING/REFERENCING ANTWORDPROFILER

Use the following method to cite/reference *AntWordProfiler* according to the APA style guide:

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

For example if you download *AntWordProfiler 1.4.0*, which was released in 2014, you would cite/reference it as follows:
Anthony, L. (2012). *AntWordProfiler* (Version 1.4.0) [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 thank Paul Nation of Victoria University of Wellington for being a constant source of excellent suggestions and feedback during the development of this program. I would also like to thank him for giving me permission to include his baseword lists in the software. Without these, the value of the software would be severely diminished.

In addition, I would like to thank Chris Sheppard of Waseda University for encouraging me to develop this software at the outset, and his suggestions and comments during the early testing stages. The development of *AntWordProfiler* has been supported by a Grant-in-aid for Scientific Research by the Japan Society for the Promotion of Education, Science, Sports and Culture, Japan (No. 18700658). Development of version 3.3 has also been helped by generous support from Compass Media, Seoul.

LEGAL MATTER

KNOWN ISSUES

None at present.

REVISION HISTORY

1.4.1

This is a minor upgrade addressing a refresh issue that was noted on Macintosh OS X 10.9 (Mavericks) operating systems.

Bug fixes:

1. Due to changes introduced to the graphics rendering on Macintosh OS X 10.9 (Mavericks), AntWordProfiler did not refresh correctly after certain actions making the screen look to freeze. The earlier workaround was click outside of the AntWordProfiler window and back into the window to force a refresh. This new version of AntWordProfiler fixes the problem at the code level and so no refresh issues should be seen in the future. The workaround has involved creating a standard build of the Perl programming language based on Citrus Perl (5.16.0) and the latest Tcl graphics engine. The introduction of the latest versions of these tools may introduce other very small changes in results due to the use of the latest Unicode standard, but these are unlikely to affect normal users.

1.4.0

This is a major update that adds a few new features.

New Features

1) Sorting of results can now be made a two independent levels chosen by the user from the choices of (frequency, range, or word).
2) Users can now opt to use their own token definition instead of the default one.
3) The help menu option now launches this PDF file instead of a basic text file.
4) Warnings are now issued before closing the file viewer window (this might become annoying!)

Changes

1) Some of the widgets in the interface have been renamed to make their actions easier to understand.

2) The ordering of global settings categories is now alphabetical.
3) Highlight colors have now been adjusted to be more pleasant on the eye.
4) The regular expression associated with the default token definition is now shown.
5) The thesaurus option is now turned OFF by default.
6) This readme file is now formatted as a PDF file.

Bug fixes

1) The Stop button has been removed from the main window because it never 'stopped' the program properly before. To force close the program just click on the button at the top of the window.
2) The progress bar now in the main window now updates properly during a long operation.

1.3.1

This is a minor update addressing one small bug and adding one or two new features

New Features

1) Added an option in the thesaurus settings menu to allow users to select to automatically update the file view after editing it with the thesaurus tool.
2) Massively improved (>10 times) the speed of the file viewer tool when highlighting texts.

Bug fixes

1) Corrected a bug that caused the local frequency count in the file viewer to report values for characters as well as words.
2) Tidied the interface a little so large frequency values do not become hidden in the file viewer tool
3) Fixed a bug that caused the program to hang if the Profiler Tool start button was clicked when no files or lists were imported.

1.3.0

This is a major update adding new features and correcting various bugs in the 1.2x versions.

New Features

1) Revised the names and cleaned contents of the internal baseword lists (approved by Paul Nation)
2) Added an option to view a complete frequency list of the target files in the profile
3) Added an option to view words not in the target files in the profile
4) Added new sort options (word, range, freq)
5) Added a batch process option (main window only)
6) Revised this readme file
7) Added a new thesaurus viewer (activated via a menu option)
8) Revised the profile output to include cumulative frequencies
9) Improved the User File Viewer tool to display all level lists in a scrolled pane.
10) Added new editing functions (Previous, Next) to the User File Viewer tool
11) Added a new edit function to simplify the target text via a level-marked thesaurus.

Bug fixes

1) Corrected an unintuitive feature that duplicated the listing of internal baseword lists after importing a settings file.
2) Corrected a bug that caused the windows to not be centered when first displayed.
3) Corrected a bug that prevented more than eight baseword lists to function correctly.

1.2.1

This is a minor update to correct a few bugs and minor issues in the 1.200 version.

New Features

1) Added a new icon to the software.

Bug fixes

1) Fixed an issue that caused the global settings options to not appear on Macintosh OS X.

Minor issues

1) The icon for AntWordProfiler has been update to a graphic that matches other *AntLab* tools. I hope you like it!
2) The height of widgets has now been fixed to make them fluid and thus allow them to take their full height on each operating system.

1.200

This version includes many changes and so deserves its 1.2 status. A list of new features and changes are below.

New Features

1) Redesigned the software engine so that the program now works smoothly with Unicode (UTF-8) encoded files. This means the program will work with any language (including Asian languages) provided that the target files, and baseword lists are saved in the UTF-8 encoding.
2) Redesigned the settings files to be easier to edit and work with (for creating user generated settings files). The settings file is no longer an XML file. Rather, it is a special file with a very simple format.
3) The main window title now shows if the program is using a user-settings file or not.
4) Any errors do not show appear on the main window anymore. This makes debugging a little more difficult, but improves the user experience.

1.104

Bug fixes

1) Fixed code to correctly update the level list word database when analyzing new texts. Previous versions would produce strange results if the User File view was used without generating a main set of results first.

1.103

New Features

1) Recoded to allow Macintosh OSX version to be compiled easily.

Bug fixes

None

1.102

Bug fixes

1) Deleted two file menu options that were redundant (Choose Main Word List, Choose User Word List)
2) Added widget names to interface that were included in the settings file but not reflected.
3) Repositioned widgets to allow non-English names to be used successfully.
4) Revised the internal level lists to exactly match those in Paul Nation's original Range program (not Range BNC).The previous versions of AntWordProfiler used the first three Range BNC lists. Note that either set of lists can be imported and used as is. Note also that the placement of some entries in the Range lists is questionable (see Bug fix 1 in version 1.101).

1.101

Bug fixes

1) Corrected the placement of "M" in the Nation Baseword List 1. Originally this had its own entry, but now it is correctly placed as a family member of "BE"
2) Corrected the trimming of non-token characters from the beginning of file lines. These caused the not-in-list results to include a stray entry count.
3) Updated and redefined the settings of the compiler for AntWordProfiler. This may resolve the problems reported by some users in Asia who cannot launch the program.
4) Added XML header to user settings file

1.100

New Features

1) Introduced a database back engine, allowing much larger files to be processed.
2) Hugely improved the speed of processing. Now, 7 million words can be processed in around 1 minute.
3) Introduced a tag ignore feature allowing tagged files to be processed without change.
4) Greatly simplified the program (and interface) by adopting Paul Nation's Range format for level list files.
5) Importing and exporting of the settings is now possible via an XML file.
6) Simplified the save format options. The Excel save format may be reintroduced later.
7) Changed the name of the program to more accurately represent what it does.
8) Re-wrote this readme file to reflect the new changes.

Bug fixes

1) Several mistakes in the language of the interface have been corrected.
2) Errors reports that occur when pressing Cancel in some dialog boxes have been fixed.

1.01

Bug fixes

1) Corrected how to read in main word lists and level lists so that blank lines and lines starting with non-token characters are ignored. In version 1.00, blank lines would generate very strange results.
2) Added character encodings for Chinese, Taiwanese, and Korean (Japanese encodings were already included in version 1.00)

1.00 This is the first version of the program.