



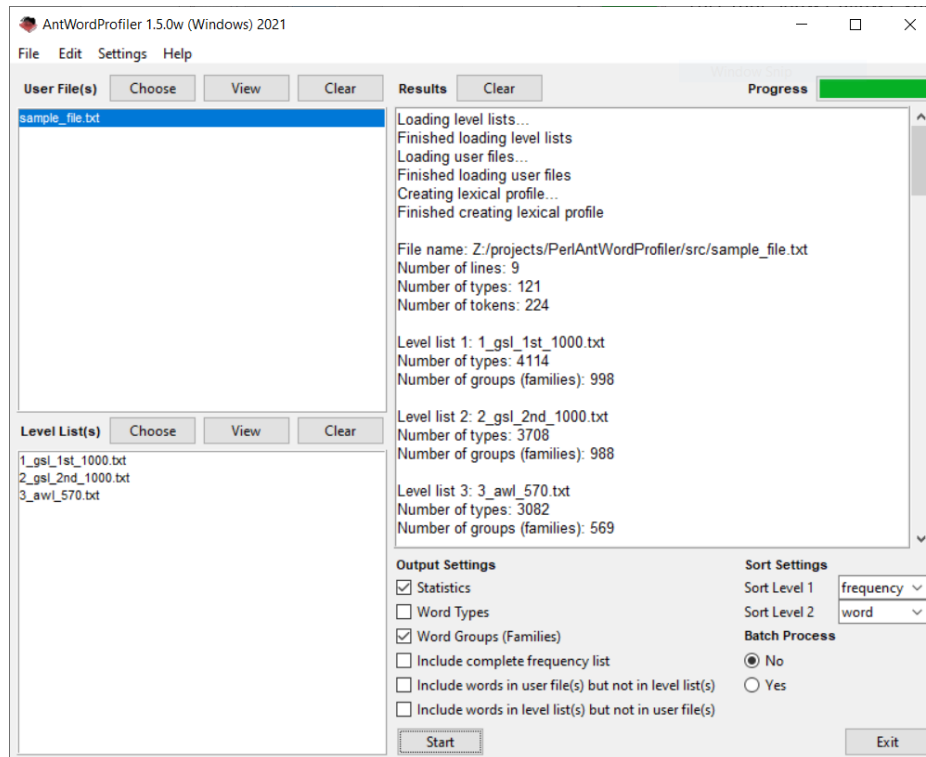
# AntWordProfiler (Windows, Macintosh OS X, and Linux)

Build 1.5.1

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

March 4, 2021



## Introduction

*AntWordProfiler* is a freeware, multiplatform tool for carrying out corpus linguistics research on vocabulary profiling. It runs on any computer running Microsoft Windows, Macintosh OS X, and Linux computers. 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

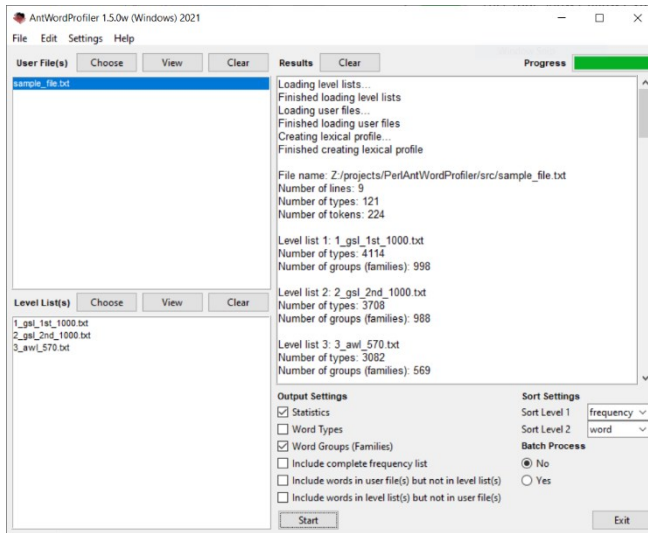
On Macintosh systems, simply double click the *AntWordProfiler* icon and this will launch the program.

### 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.

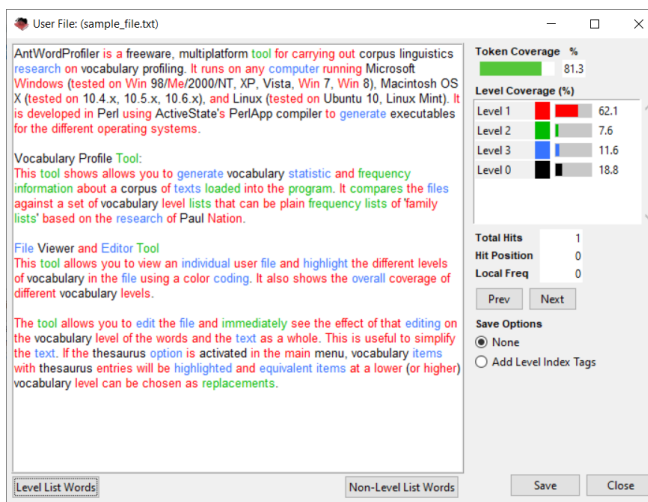
## Overview of Tools

*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.

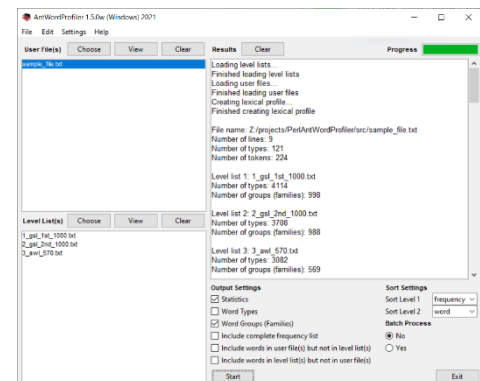
The tool allows you to edit the file and immediately see the effect of that editing on the vocabulary level and the text as a whole. This is useful to simplify the text. If the thesaurus option is activated in the main menu, vocabulary items with thesaurus entries will be highlighted and equivalent items 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
  - 3awl\_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
...
```

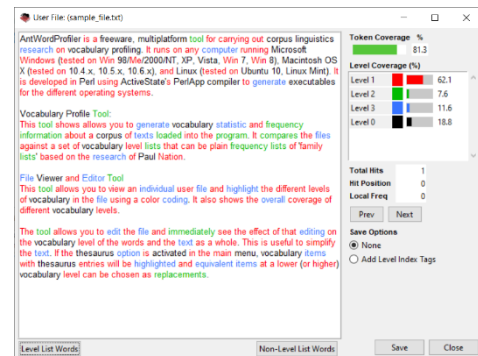
- 3) Import one or more target user files using the Choose button or the file menu option.
- 4) Select one or more of the output data options
  - a) Statistics: this will produce a set of statistics about the target files.
  - b) Word Types: this will produce a list of word types in each target file 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, it 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 CTRL-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 1.5.0*, which was released in 2021, you would cite/reference it as follows:

Anthony, L. (2021). *AntWordProfiler* (Version 1.5.0) [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.

Early development of *AntWordProfiler* was 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 was also helped by generous support from Compass Media, Seoul.

## **KNOWN ISSUES**

None at present.

Copyright: Laurence Anthony