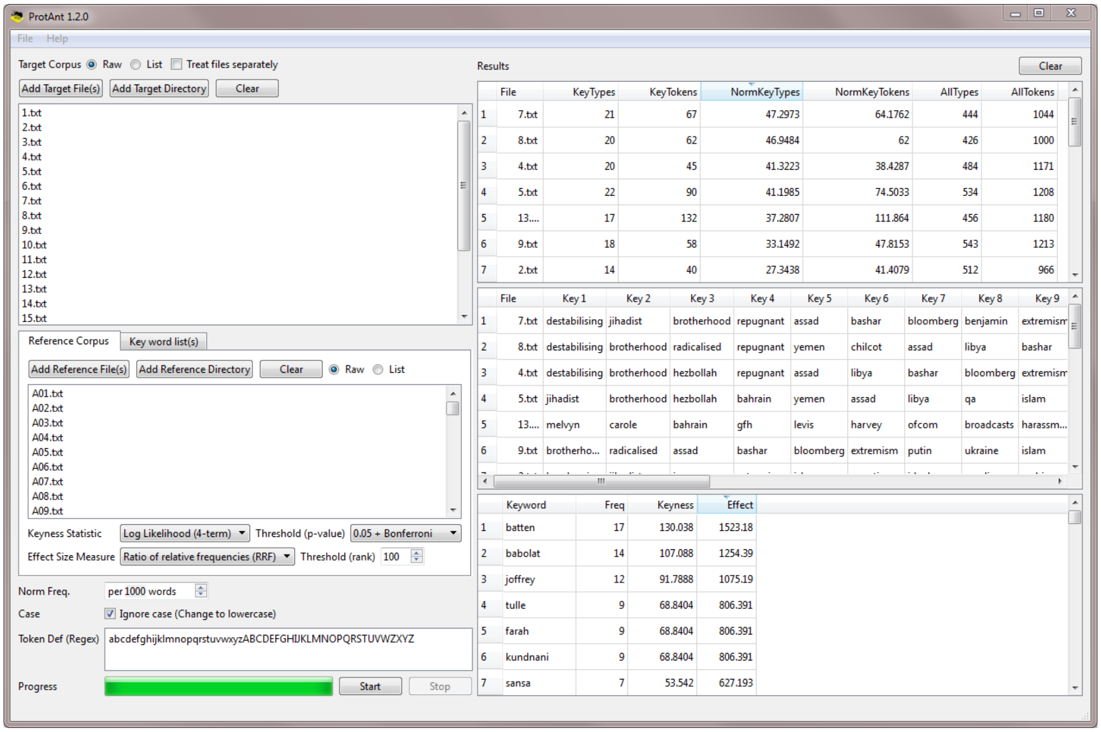
ProtAnt (Windows)

Build 1.2.1 (Released March 21, 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.



Target Corpus ● Raw ○ List □ Treat files separately

1.txt
2.txt
3.txt
4.txt
5.txt
6.txt
7.txt
8.txt
9.txt
10.txt
11.txt
12.txt
13.txt
14.txt
15.txt

Reference Corpus: ● Raw ○ List

A01.txt
A02.txt
A03.txt
A04.txt
A05.txt
A06.txt
A07.txt
A08.txt
A09.txt

Keyness Statistic: Log Likelihood (4-term)
Threshold (p-value): 0.05 + Bonferroni
Effect Size Measure: Ratio of relative frequencies (RRF)
Threshold (rank): 100
Norm Freq.: per 1000 words
Case: ☑ Ignore case (Change to lowercase)
Token Def (Regex): abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWZXYZ
Progress

	File	KeyTypes	KeyTokens	NormKeyTypes	NormKeyTokens	AllTypes	AllTokens
1	7.txt	21	67	47.2973	64.1762	444	1044
2	8.txt	20	62	46.9484	62	426	1000
3	4.txt	20	45	41.3223	38.4287	484	1171
4	5.txt	22	90	41.1985	74.5033	534	1208
5	13....	17	132	37.2807	111.864	456	1180
6	9.txt	18	58	33.1492	47.8153	543	1213
7	2.txt	14	40	27.3438	41.4079	512	966

	File	Key 1	Key 2	Key 3	Key 4	Key 5	Key 6	Key 7	Key 8	Key 9
1	7.txt	destabilising	jihadist	brotherhood	repugnant	assad	bashar	bloomberg	benjamin	extremism
2	8.txt	destabilising	brotherhood	radicalised	repugnant	yemen	chilcot	assad	libya	bashar
3	4.txt	destabilising	brotherhood	hezbollah	repugnant	assad	libya	bashar	bloomberg	extremism
4	5.txt	jihadist	brotherhood	hezbollah	bahrain	yemen	assad	libya	qa	islam
5	13....	melvyn	carole	bahrain	gfh	levis	harvey	ofcom	broadcasts	harassm...
6	9.txt	brotherho...	radicalised	assad	bashar	bloomberg	extremism	putin	ukraine	islam

	Keyword	Freq	Keyness	Effect
1	batten	17	130.038	1523.18
2	babolat	14	107.088	1254.39
3	joffrey	12	91.7888	1075.19
4	tulle	9	68.8404	806.391
5	farah	9	68.8404	806.391
6	kundnani	9	68.8404	806.391
7	sansa	7	53.542	627.193

Introduction

ProtAnt is a freeware prototypical text detection tool developed in collaboration with Paul Baker of Lancaster University, UK. *ProtAnt* takes a corpus of texts (***UTF-8 encoded***) and compares them either individually or as a whole against a reference corpus (***UTF-8 encoded***) or list of 'key' words (***UTF-8 encoded***) to find characteristic features in the target files. Then, *ProtAnt* looks at each individual target file and counts how many of these characteristic features are in them. The target files are then ranked in terms of their prototypicality by the number of characteristic features they contain. *ProtAnt* runs on any computer running Microsoft Windows (tested on Win 98/Me/2000/NT, XP, Vista, Win 7, Win 8) and Macintosh OS X computers (tested up to OS X 10.9 Mavericks). *ProtAnt* is developed in Python and Qt using the *PyInstaller* compiler to generate executables for the different operating systems.

Getting Started (No installation necessary)

Windows

On Windows systems, simply double click the *ProtAnt* icon to launch the program.

Macintosh OS X

On Macintosh systems, simply double click the *ProtAnt* zip file. The zip file will unzip the *ProtAnt* application. Then, you can drag the *ProtAnt* application to your application folder, your desktop, or anywhere else you like. Throw away the zip file when you are finished.

Finding prototypical texts using a reference corpus

Step 1: Select the type of target corpus files you are going to use (raw files or word lists)

Step 2: Select if you want keywords to be generated by comparing all target file together with the reference files (unchecked option) or each target file separately with the reference files (checked option)

Step 3: Select the target corpus files that you want to analyze. You can do this in three ways:

1) Click on the File->Open Target File(s) menu option or the "Add Target File(s)" button below the "Target Corpus" label and select the files you want to analyze;
2) Click on the File->Open Target Dir menu option or the "Add Target Directory" button below the "Target Corpus" label and select a directory of files you want to analyze;
3) Drag and drop files directly onto the *ProtAnt* application.

Step 4: Select the reference corpus files that you want to analyze. You can do this in three ways:

1) Click on the File->Open Reference File(s) menu option or the "Add Reference File(s)" button below the "Reference Corpus" label and select the files you want to add;
2) Click on the File->Open Reference Dir menu option or the "Add Corpus Directory" button below the "Reference Corpus" label and select a directory of files you want to add;
3) Drag and drop files directly onto the *ProtAnt* application.

Step 5: Choose the keyness statistic, keyness threshold value (p-value), effect size measure, and effect size threshold (the cutoff rank) you would like to use to generate keywords for the target corpus. Note that a keyness threshold of 0 means all (positive) keywords will considered, and an effect size threshold of -1 means that all effect values will be shown. A positive keyword means that it is relatively more frequent in the target corpus than the reference corpus. Negative keywords (relatively more frequent in the reference corpus than the target corpus) are not considered.

Step 6: Choose the normalization function for frequencies displayed in the results window.

Step 7: Choose to ignore case in the target corpus (change all words to lowercase)

Step 8: Decide a suitable token definition based on regular expression (regex) syntax.

Step 9: Click "Start" to begin the analysis.

Finding prototypical texts using 'key' word lists

Step 1: Select the type of target corpus files you are going to use (raw files or word lists)

Step 2: Select the target corpus files that you want to analyze. You can do this in three ways (see Step 3 above):

Step 3: Select the 'key' word list files that you want to analyze. You can do this in three ways:

1) Click on the File->Open Keywords File(s) menu option or the "Add Keywords File(s)" button below the "Key word list(s)" label and select the files you want to add;
2) Click on the File->Open Keywords Dir menu option or the "Add Keywords Directory" button below the "Key word list(s)" label and select a directory of files you want to add;
3) Drag and drop files directly onto the *ProtAnt* application.

Step 4: Choose the normalization function for frequencies displayed in the results window.

Step 5: Choose to ignore case in the target corpus (change all words to lowercase)

Step 6: Decide a suitable token definition based on regular expression (regex) syntax.

Step 7: Click "Start" to begin the analysis.

Note 1: If you click on the File->Close Target Files menu option, the File->Close Reference Files menu option, or the File->Close Keywords Files menu option, the files will removed from the relevant list.

Note 2: If you click on the "Clear" button below the "Target Corpus" label "Key word list(s)" label, or the "Clear" button below the "Reference Corpus" label, or the "Clear" button below the "Key word list(s)" label, the files will removed from the relevant list.

Note 3: The results of the prototypical text detection are shown in the top right window. The keywords appearing in each (ranked) corpus file are shown in the middle right window. The complete list of keywords is shown in the bottom right window. All columns can be sorted in either ascending or descending order by clicking on the column headers.

Note 4: The prototypical text detection analysis can be stopped at any time by clicking the "Stop" button.

Note 5: If you are using target corpus or reference corpus word lists, they should be formatted as RANK, FREQ, TYPE separated by tabs. Any line beginning with # will be ignored.

Note 6: If you are using 'key' word files, they should be formatted with each 'key' word on a new line.

Additional Features

The output display can be selected, copied, and pasted as is standard on the operating system:

Windows:	CTRL-A ⇨ Select All	CTRL-C ⇨ Copy	CTRL-V ⇨ Paste
Macintosh:	CMD-A ⇨ Select All	CMD -C ⇨ Copy	CMD -V ⇨ Paste

NOTES

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 have any suggestions for improving the program, please let me know and I will try to address the issues in a future version.

This software is available as 'freeware' according to the license below. 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 *ProtAnt*

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

Anthony, L. and Baker, P. (YEAR OF RELEASE). *ProtAnt* (Version VERSION NUMBER) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

For example if you download *ProtAnt 1.2.0*, which was released in 2016, you would cite/reference it as follows: Anthony, L. and Baker, P. (2016). *ProtAnt* (Version 1.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http:// www.laurenceanthony.net/

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/

LICENSE for ProtAnt

ProtAnt 1.0 and any minor updates issued by AntLab Solutions (collectively 'the Software')

TERMS GOVERNING THE USE OF THE SOFTWARE

The Software is protected by copyright and must not be used, displayed, modified, adapted, distributed, transmitted, transferred or published or otherwise reproduced in any form by any means other than strictly in accordance with the terms set out below. By installing the Software, you agree to be bound by the terms of the license. This ProtAnt License (“License”) is made between AntLab Solutions, Tokyo, Japan as licensor, and you, as licensee, as of the date of your use of the Software. The Software is in use on a computer when it is loaded into the RAM or installed into the permanent memory of that computer, e.g., a hard disk or other storage device.

1. License Material

These terms govern your use of the Software but not including subsequent versions (e.g. ProtAnt 2.0').

2. License Grant

AntLab Solutions grants to you a personal non-exclusive non-transferable license ('the License') to use the Software in the following specific contexts.

a) Non-Commercial (Freeware) Use:

You may use the software for non-profit purposes on more than one computer or on a network so long as you are the sole user of the Software. (A “network” is any combination of two or more computers that are electronically linked and capable of sharing the use of a single software program.) You are not permitted to sell, lease, distribute, transfer, sublicense, or otherwise dispose of the Software, in whole or in part, for any form of actual or potential commercial gain or consideration.

b) Commercial Evaluation (Trial) Use:

You may evaluate (trial) the software for commercial purposes for a period of no more than fourteen (14) days from the date of download on more than one computer or on a network so long as you are the sole user of the Software.

c) Commercial Use

When you pay the commercial license fee established by AntLab Solutions, you may use the software for non-profit or commercial purposes on more than one computer or on a network so long as you are the sole user of the Software. (A “network” is any combination of two or more computers that are electronically linked and capable of sharing the use of a single software program.) You will obtain a separate license for each additional user of the Software (whether or not such users are connected on a network). You are not permitted to sell, lease, distribute, transfer, sublicense, or otherwise dispose of the Software, in whole or in part, for any form of actual or potential commercial gain or consideration.

3. Termination

You may terminate this License at any time by uninstalling the Software and deleting it. The License will also terminate if you breach any of the terms of the License.

4. Proprietary Rights

The Software is licensed, not sold, to you. AntLab Solutions reserves all rights not expressly granted to you. Ownership of the Software and its associated proprietary rights, including but not limited to patent and patent applications, are retained by AntLab Solutions. The Software is protected by the copyright laws of Japan and by international treaties. Therefore, you must comply with such laws and treaties in your use of the Software. You agree not to remove any of AntLab Solutions' copyright, trademarks, and other proprietary notices from the Software.

5. Distribution

Except as may be expressly allowed in Section 2, or as otherwise agreed to in a written agreement signed by both you and AntLab Solutions, you will not distribute the Software, either in whole or in part, in any form or medium.

6. Transfer and Use Restrictions
You may not sell, license, sub-license, lend, lease, rent, share, assign, transmit, telecommunicate, export, distribute or otherwise transfer the Software to others, except as expressly permitted in this License Agreement or in another agreement with AntLab Solutions. You may not modify, reverse engineer, decompile, decrypt, extract, or otherwise disassemble the Software.

7. Warranties
ANTLAB SOLUTIONS MAKES NO WARRANTIES WHATSOEVER REGARDING THE SOFTWARE AND IN PARTICULAR, DOES NOT WARRANT THAT THE SOFTWARE WILL FUNCTION IN ACCORDANCE WITH THE ACCOMPANYING DOCUMENTATION IN EVERY COMBINATION OF HARDWARE PLATFORM OR SOFTWARE ENVIRONMENT OR CONFIGURATION, OR BE COMPATIBLE WITH EVERY COMPUTER SYSTEM. IF THE SOFTWARE IS DEFECTIVE FOR ANY REASON, YOU WILL ASSUME THE ENTIRE COST OF ALL NECESSARY REPAIRS OR REPLACEMENTS.

8. Disclaimer
ANTLAB SOLUTIONS DOES NOT WARRANT THAT THE SOFTWARE OR SERVICE IS FREE FROM BUGS, DEFECTS, ERRORS OR OMISSIONS. THE SOFTWARE OR SERVICE IS PROVIDED ON AN “AS IS” BASIS AND ANTLAB SOLUTIONS MAKES NO OTHER WARRANTIES OR CONDITIONS, EXPRESS OR IMPLIED, WITH RESPECT TO THE SOFTWARE INCLUDING WITHOUT LIMITATION THE IMPLIED WARRANTIES OR CONDITIONS OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

9. Limitation of Liability
ANTLAB SOLUTIONS WILL HAVE NO LIABILITY OR OBLIGATION FOR ANY DAMAGES OR REMEDIES, INCLUDING, WITHOUT LIMITATION, THE COST OF SUBSTITUTE GOODS, LOST DATA, LOST PROFITS, LOST REVENUES OR ANY OTHER DIRECT, INDIRECT, INCIDENTAL, SPECIAL, GENERAL, PUNITIVE OR CONSEQUENTIAL DAMAGES, ARISING OUT OF THIS LICENSE OR THE USE OR INABILITY TO USE THE SOFTWARE OR SERVICE. IN NO EVENT WILL ANTLAB SOLUTIONS'S TOTAL AGGREGATE LIABILITY (WHETHER IN CONTRACT (INCLUDING FUNDAMENTAL BREACH), WARRANTY, TORT (INCLUDING NEGLIGENCE), PRODUCT LIABILITY, INTELLECTUAL PROPERTY INFRINGEMENT OR OTHER LEGAL THEORY) WITH REGARD TO THE SOFTWARE AND/OR THIS LICENSE EXCEED THE LICENSE FEE PAID BY YOU TO ANTLAB SOLUTIONS. FURTHER, ANTLAB SOLUTIONS WILL NOT BE LIABLE FOR ANY DELAY OR FAILURE TO PERFORM ITS OBLIGATIONS UNDER THIS LICENSE AS A RESULT OF ANY CAUSES OR CONDITIONS BEYOND ANTLAB SOLUTIONS' REASONABLE CONTROL

10. Jurisdiction
These terms will be governed by Japanese law and the Japanese courts shall have jurisdiction.

KNOWN ISSUES

None at present.

REVISION HISTORY

1.2.1
This is minor update
New features
Bug fixes:

1. The code has been adapted to ensure that logger files on Macintosh OSX are created in the correct location relative to the app.

1.2.0

This is minor update
New features

1. Either raw files or word lists can now be used as the target corpus and/or reference corpus.
2. Many more effect size measures can now be chosen.
3. The left and right panes in the main window and the individual results panes can now be dragged and hidden to maximize screen space.

1.1.0
This is minor update
New features

4. Either raw files or word lists can now be used as the target corpus and/or reference corpus.
5. Many more effect size measures can now be chosen.
6. The left and right panes in the main window and the individual results panes can now be dragged and hidden to maximize screen space.

Bug fixes:

2. In Version 1.0.2, effect sizes cut offs of types were based on alphabetically ordering instead of numerical ordering. This resulted in spurious rankings. This is now fixed.
3. In earlier versions, the last keyword in the list of keywords was not displayed in the file keyword middle panel. This is now corrected.

This is minor update
Bug fixes:

1. When files were dragged and dropped on the target corpus or reference corpus boxes, they were not loaded correctly. Drag and drop is now working.
2. When the Log Ratio option was chosen, in some cases no results were generated. This is now fixed.
3. In previous versions of ProtAnt, the keyness statistic (Log Likelihood) was based on a 2-term version of the calculation that is commonly used in corpus linguistics. This is an estimate of the more accurate 4-term calculation which is now the default. For backwards compatibility, the 2-term version is left as an option.
4. Although not strictly a bug, rather than allowing a ranking of keywords by keyness values (effectively a ranking by p-values) as in previous versions, ProtAnt now encourages rankings of keywords by an effect size measure. Two effect size measures are provided.

1.0.1
This is minor update
Bug fix: When a general Unicode character class token definition was supplied (e.g. \p{L} for letters), non-English texts were not being processed correctly. This is now fixed.

1.0.0
This is the first version of the program