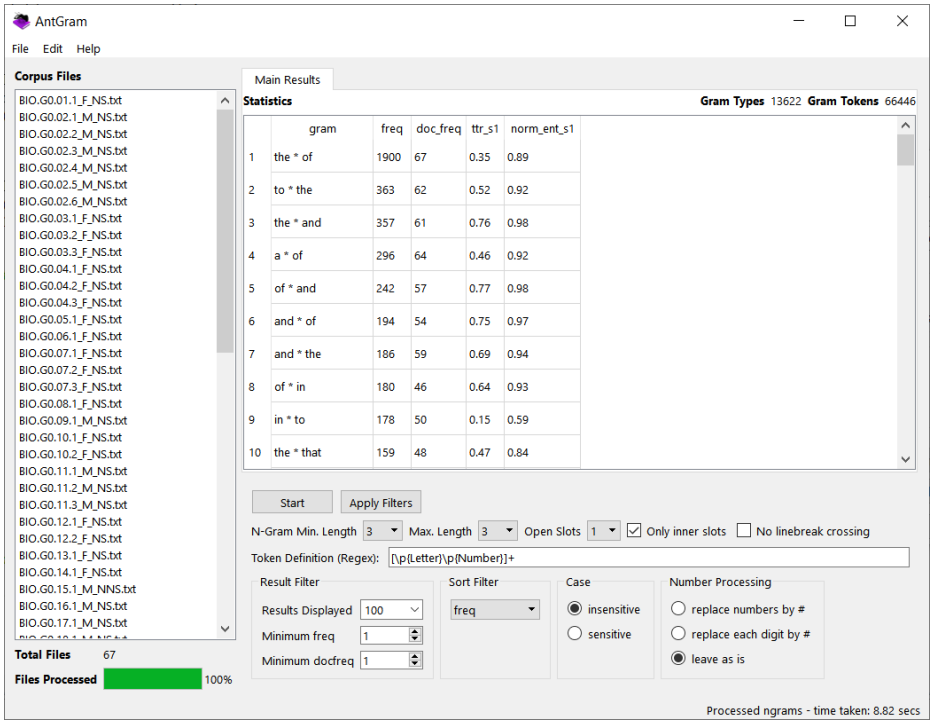
AntGram (Windows)

Build 1.2.2 (Released October 21, 2019)

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: 100.



Introduction

AntGram is a freeware n-gram and phrase frame (p-frame) generation and profile tool that produces results based on a corpus of texts (UTF-8 encoded). *AntGram* runs on any computer running Microsoft Windows (tested on Win 10), Macintosh OS X (tested on OS X 10.9 Mavericks), and Linux (tested on Linux Mint 17) computers. It 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 *AntGram* icon to launch the program.

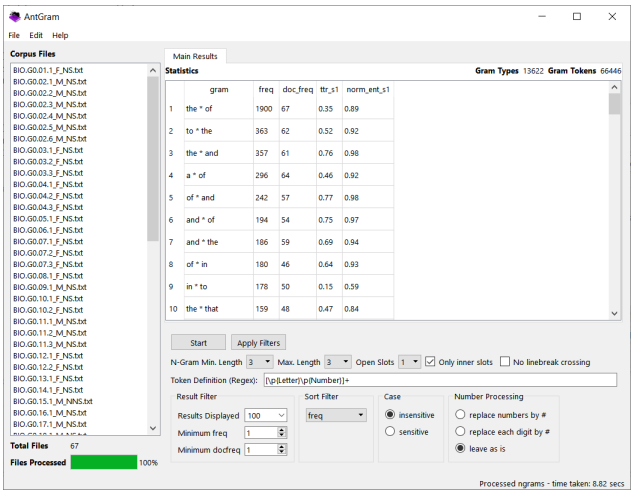
Macintosh OS X

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

Linux

On Linux systems, set the permissions to run the executable, then double click the *AntGram* icon to launch the program.

Creating an n-gram list (or p-frame list)



Step 1: Select the corpus you want to use. You can do this in four ways:

a) Click on the File->Open File(s) menu option and select the corpus you want to use;

b) Click on the File->Open Dir menu option and select a directory of corpus files you want to use;

d) Drag and drop corpus files directly onto the *AntGram* application.

Step 2: Select the parameters you want to use:

- *Min. Length:* The minimum length of n-grams, e.g. 2 => a b).
- *Max. Length*: he maximum length of n-grams, e.g. 4 => a b c d).
- *Open Slots*: The number of possible open slots in each n-gram, e.g. 2 => a # c #).
- *Only inner slots*. An option to restrict open slots to only appear inside the n-gram, e.g. YES => a # c d).
- *No linebreak crossing*: An option to restrict grams to only those that appear on the same line of a file.

Step 3: Create a definition of each token of the n-gram (if required):

a) Use the default setting (Unicode *letters* and *numbers*) unless you have specific reason to change it. Change the default setting by specifying an appropriate Perl-based regular expression (PCRE).

Step 4: Choose how to filter the results displayed on the screen (Result Filter).

- *Results displayed*: The number of results shown on the screen.
  - All results are generated by the program but only those set by the filters will be displayed. This feature allows huge numbers of n-gram/p-frames to be generated without using up all available memory when showing them.
- *Min. Freq*: n-grams with a frequency value less than this value will not be displayed.
  - The frequency value of an n-gram corresponds to the number of times it appears in the corpus
- *Min. DocFreq*: n-grams with a document frequency (range) value less than this value will not be displayed.
  - The document frequency value of an n-gram corresponds to the number of corpus files in which the n-gram appears.

Step 5: Choose how to sort the results displayed on the screen (Sort Filter).

- *alphabetical*: The sort order will alphabetical (according to the Unicode specification)
- *freq*: The sort order will be from the most frequent to the least frequent n-gram.
  - Ties are sorted alphabetically.
- *doc_freq*: The sort order will be from the n-gram with the highest docfreq value to that with lowest docfreq value.
  - Ties are sorted alphabetically.
- *ttr*: The sort order will be from the highest type-token ratio (TTR) value to the least.
  - Ties are sorted alphabetically
- *norm_ent*: The sort order will be from the highest normed entropy value to the least.
  - Ties are sorted alphabetically

Step 6: Choose how to treat letter case when processing the corpus files (Case)

- *insensitive*: (e.g. The 100 CATS => the 100 cats)
- *sensitive*:(e.g. The 100 CATS => The 100 CATS)

Step 7: Choose how to process numbers (Number Processing)

- *replace numbers by #*: (e.g. the 100 cats => the # cats)
- *replace each digit by #*: (e.g. the 100 cats => the ### cats)
- *leave as is*: (e.g. the 100 cats => the 100 cats)

Step 8: Click the "Start" button and wait for the results to be generated

Creating a p-frame slot profile

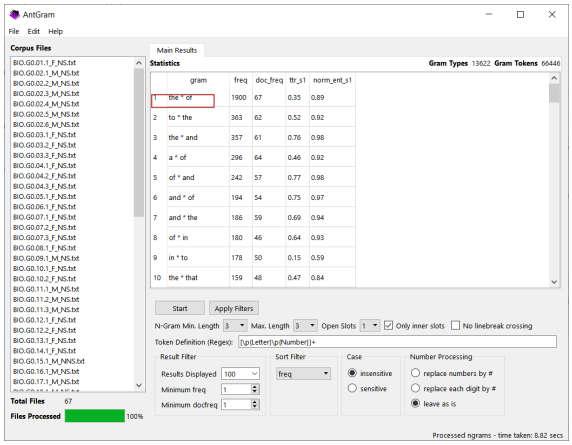
Step 1: Create an p-frame list:

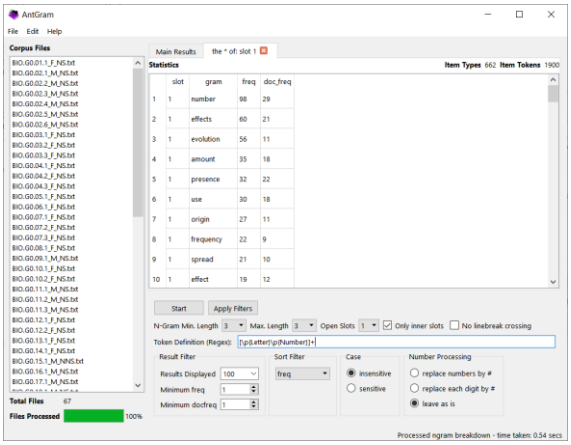
a) Follow the steps above and create a p-frame list (making sure the Open Slots option is set to 1 or more)

Step 2: Create a p-frame slot profile

a) Double click on one of the p-frame entries.

b) Wait for *AntGram* to create one or more new tabs showing the items that fill the slot(s)





Saving results

Step 1: Select the tab of the results that you want to save

Step 2: Save the results via one of the File menu options:

- File->Save Display Results As…
  - Save the filtered results (i.e. those shown on the screen)
- File->Save All Results As…
  - Save all the generated results (including those hidden by the filters)

Additional Features

Selected files can be closed via the File->Close Selected Files menu option. All files can be closed via the File->Close All Files menu option.

Results 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' (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 *AntGram*

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

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

For example if you download *AntGram 1.0*, which was released in 2018, you would cite/reference it as follows:
Anthony, L. (2018). *AntGram* (Version 1.0) [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/

KNOWN ISSUES

None at present.

LICENSE for *AntGram*

AntGram 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 *AntGram* 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. *AntGram* 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.

REVISION HISTORY

1.2.2 This is a minor update with several bug fixes:

- Bug fixes
    - In previous versions, typing a value for the number of shown results (rather than selecting one of the options) would only be reflected after the return key was pressed. Now, the setting is reflected as the user types the value so no return key press is needed.
    - A bug in version 1.2.0 caused the program to show ngram results with open slots at the end of the ngram even when the "only inner slots" option was selected. This only happened when a variable range of ngram sizes was chosen. This has now been fixed.

1.2.1 This is a minor update with one bug fix:

- Bug fixes
    - A bug in version 1.2.0 caused the program to crash when used with files with certain Unicode characters in them. This is now fixed.

1.2.0 This is a minor update with various features added and several bug fixes included:

- New features
    - The interface has now been revamped to improve usability:
        - The order of columns in the results window has now been changed so that grams, frequency, document frequency (range), type-token ratio (ttr), and normed_entropy (norm_ent) appear in that order. The ttr and norm_ent values will only show when an open slots exist in the ngrams.
        - The sort filter options can now be selected for a combobox
        - Open slots are now shown with a "*" character
        - Numbers are now be replaced with a "#" character
        - The default token definition now only includes Unicode *letter* and *number* characters
    - A new option ("No linebreak crossing") has been added to prevent the Ngram detection crossing line breaks.
    - A new internal function has been added to ensure that Ngram never cross document boundaries.
    - Processing should now be generally faster than in the original 1.0 version regardless of the target corpus size. It should also be faster than the 1.1.0 version on larger datasets as it uses an external database to store all information. Unfortunately, there is a performance loss compared with version 1.0.0 on smaller datasets. I am attempting to improve the speed for smaller datasets now.
    - The results window now always shows the total gram types/gram tokens for the dataset at the top right of the window.
- Bug fixes
    - A bug in version 1.1.0 that prevented results from being saved correctly has now been fixed.
    - Various bugs that caused the program to crash when particular combinations of settings parameters where chosen has now been fixed.

1.1.0 This is a minor update with various features added

- New features
    - Processing should now be generally faster than in the original 1.0 version
    - Only “true” open slot entries are shown (i.e., those that have two or more variable items that can fit in the empty slot
    - Several typos and unclear sentences in this help page have been corrected.
- Bug fixes
    - When the ‘Results Displayed’ filter was set to “all” the program would crash. This has now been fixed.

1.0 This is the first version of the program