

# AntGram (Windows)

Build 1.2.0 (Released August 8, 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.

AntGram Edit Help								-		×
orpus Files		Main Results								
BIO.GO.01.1_F_NS.txt	∧ Sta	atistics						Gram Types 13622 Gra	m Tokens	6644(
BIO.G0.02.1_M_NS.txt										•
BIO.G0.02.2_M_NS.txt		gram	freq	doc_freq	ttr_s1	norm_ent_	51			~
IO.G0.02.3_M_NS.txt	1	the * of	1900	67	0.35	0.89				
IO.G0.02.4_M_NS.txt										
IO.G0.02.5_M_NS.txt	2	to * the	363	62	0.52	0.92				
O.G0.02.6_M_NS.txt										
IO.G0.03.1_F_NS.txt	3	the * and	357	61	0.76	0.98				
O.G0.03.2_F_NS.txt										
IO.G0.03.3_F_NS.txt	4	a * of	296	64	0.46	0.92				
IO.G0.04.1_F_NS.txt										
IO.G0.04.2_F_NS.txt IO.G0.04.3_F_NS.txt	5	of * and	242	57	0.77	0.98				
0.G0.05.1_F_NS.txt	6	and * of	104	54	0.75	0.07				
0.G0.06.1 F NS.txt	0	and ^ of	194	54	0.75	0.97				
0.G0.07.1 F NS.txt	7	and * the	186	59	0.69	0.94				
0.G0.07.2_F_NS.txt		and the	100	55	0.05	0.34				
IO.G0.07.3_F_NS.txt	8	of * in	180	46	0.64	0.93				
O.G0.08.1_F_NS.txt	-									
O.G0.09.1_M_NS.txt	9	in * to	178	50	0.15	0.59				
O.G0.10.1_F_NS.txt										
O.G0.10.2_F_NS.txt	10	0 the * that	159	48	0.47	0.84				~
O.G0.11.1_M_NS.txt			_							-
O.G0.11.2_M_NS.txt				_						
O.G0.11.3_M_NS.txt		Start Ap	ply Filter	s						
O.G0.12.1_F_NS.txt		I-Gram Min. Length	3 🔻	May Long	H 2	T Open S	ata 🔳 💌 🖂 🖉	nly inner slots 🗌 No linebreak	crossing	
O.G0.12.2_F_NS.txt		-Grant Min. Lengur	• ·	wax. ceny	ui s	Open si			uossing	
IO.G0.13.1_F_NS.txt	т	oken Definition (Reg	ex): [\p	(Letter}N	Jumber)	]+				
O.G0.14.1_F_NS.txt		Result Filter		So	rt Filter		Case	Number Processing		
O.G0.15.1_M_NNS.txt								-		
IO.G0.16.1_M_NS.txt		Results Displayed	100	∼ fr	eq	-	insensitive	replace numbers by #		
IO.G0.17.1_M_NS.txt	~	Minimum freg	1	•			Sensitive	replace each digit by #		
otal Files 67				•			0	<ul> <li>Ieave as is</li> </ul>		
		Minimum docfreq	1	•				<ul> <li>reave as is</li> </ul>		
les Processed	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.

## Kacintosh 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.

# Å 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).

Corpus Files		Main Results							
BIO.G0.01.1_F_NS.txt	^ St	atistics						Gram Types 13622 Gram	Tokens 6
BIO.G0.02.1_M_NS.txt		gram	freq	doc frea	ttr s1	norm ent	e1		
BIO.G0.02.2_M_NS.txt BIO.G0.02.3_M_NS.txt		-							
BIO.G0.02.4 M NS.bt	1	the * of	1900	67	0.35	0.89			
BIO.GO.02.5 M NS.bt	2	to * the	363	62	0.52	0.92			
BIO.G0.02.6_M_NS.txt	4	to - the	303	02	0.52	0.92			
BIO.G0.03.1_F_NS.txt		the * and	357	61	0.76	0.98			
BIO.G0.03.2_F_NS.txt									
BIO.G0.03.3_F_NS.txt	4	a * of	296	64	0.46	0.92			
BIO.G0.04.1_F_NS.txt									
BIO.G0.04.2_F_NS.txt BIO.G0.04.3 F_NS.txt	5	of " and	242	57	0.77	0.98			
BIO.GO.04.3_F_NS.txt BIO.GO.05.1 F_NS.txt	6	and * of	194	54	0.75	0.97			
BIO.G0.06.1 F NS.txt	e	and " of	194	54	0.75	0.97			
BIO.GO.07.1_F_NS.txt		and * the	186	59	0.69	0.94			
BIO.GO.07.2 F NS.bd		und unc	100			0.24			
BIO.G0.07.3_F_NS.txt	8	of * in	180	46	0.64	0.93			
BIO.G0.08.1_F_NS.txt									
BIO.G0.09.1_M_NS.txt	s	in * to	178	50	0.15	0.59			
BIO.G0.10.1_F_NS.txt									
BIO.G0.10.2_F_NS.txt	1	0 the * that	159	48	0.47	0.84			
BIO.G0.11.1_M_NS.txt									
BIO.G0.11.2_M_NS.bd		Start Ap	ply Filter						
BIO.G0.11.3_M_NS.txt BIO.G0.12.1 F NS.txt		зан мр	ply rikel	>					
BIO.G0.12.2 F NS.txt		N-Gram Min. Length	3 🔻	Max. Lengt	th 3	<ul> <li>Open SI</li> </ul>	lots 1 🔻 🗹 C	Inly inner slots 🗌 No linebreak cro	ossing
BIO.GO.13.1 F_NS.txt		Token Definition (Reg	w): Din	ll etteri) p/b	lumber	04			
BIO.GO.14.1 F NS.bd									
BIO.G0.15.1_M_NNS.txt		Result Filter		So	rt Filter		Case	Number Processing	
BIO.G0.16.1_M_NS.txt		Results Displayed	100	✓ fn	ea		insensitive	replace numbers by #	
BIO.G0.17.1_M_NS.txt	~			•			O sensitive	O replace each digit by #	
510.00.10.1 1.1 MICAA		Minimum freq	1	×.			O sensitive	<ul> <li>replace each digit by #</li> </ul>	

- 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.
  - $\circ$  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)

am							- 🗆 ×	🦔 AntGram								- 0	×
Help								File Edit Help									
8	м	in Results						Corpus Files	N	Main Resu	its the * of	f: slot 1	8				
	Stat	stics					Gram Types 13622 Gram Tokens 6644	BIO.GO.01.1_F_NS.txt	Sta	tistics						Item Types 662 Item To	Akens 190
_M_NS.bt M_NS.bt		gram	freq	doc_freq	ttr_s1	norm_ent_s1	^	BIO.G0.02.1_M_NS.txt BIO.G0.02.2_M_NS.txt		slot	gram	freq	doc_freq				^
lot lot	1	the " of	1900	67	0.35	0.89		BIO.G0.02.3_M_NS.txt			number	95	29				
								810.G0.02.4_M_N5.txt 810.G0.02.5_M_N5.txt									_
	2	to * the	363	62	0.52	0.92		BIO.GO.02.6_M_NS.bt	2	1	effects	60	21				
	3	the " and	357	61	0.76	0.98		BIO.G0.03.1_F_NS.txt BIO.G0.03.2_F_NS.txt	3	1	evolution	56	11				
	4	a * of	296	64	0.46	0.92		80.G0.03.2_F_NS.bt	١.		amount	35	18				
								BIO.G0.04.1_F_NS.txt	11		amoura						
	5	of * and	242	57	0.77	0.98		810.G0.04.2_F_NS.btt 810.G0.04.3 F_NS.btt	5	1	presence	32	22				
	6	and * of	194	54	0.75	0.97		BIO.GO.05.1_F_NS.txt	6	1	use	30	18				
	7	and " the	186	50	0.69	0.04		810.G0.06.1_F_NS.txt 810.G0.07.1 F NS.txt	١.			27	11				
	111		100		0.00			BIO.GO.07.2_F_NS.txt	1	·	origin	21	11				
	8	of " in	180	46	0.64	0.93		810.G0.07.3_F_NS.txt 810.G0.08.1 F NS.txt	8	1	frequency	22	9				
	9	in * to	178	50	0.15	0.59		BIO.GO.09.1_M_NS.bt	9	1	spread	21	10				
	10	the " that	159	10	0.47	0.94		80.G0.10.1_F_NS.txt 80.G0.10.2_F_NS.txt				19					
		une uner	1.5.8	~	0.47	0.04	~	BIO.GO.11.2_F_VS.bit		1	effect	19	12				~
		Start Ap	pły Filter					BIO.G0.11.2_M_NS.txt		Start	Apply	-					
								810.G0.11.3_M_NS.bit 810.G0.12.1_F_NS.bit									
							Only inner slots No linebreak crossing	BIO.GO.12.2_F_NS.txt							ots 📘 🕈 🗹 0	inly inner slots 🗌 No linebreak crossin	19
		en Definition (Rege	ext: [NP					BIO.GO.13.1_F_NS.txt BIO.GO.14.1_F_NS.txt			nition (Regex):	[\p[Le					
	R	esult Filter		Sc	ort Filter	Case	Number Processing	BIO.G0.15.1_M_NNS.bd	11	Result Filt	ler		Sort Filte	er	Case	Number Processing	
	F	esults Displayed	100	~ fi	req	<ul> <li>insensitive</li> </ul>	<ul> <li>replace numbers by #</li> </ul>	80.G0.16.1_M_NS.txt 80.G0.17.1_M_NS.txt		Results D	isplayed 100		freq	•	insensitive	replace numbers by #	
`	<b>1</b> ,	tinimum freq	1	\$		<ul> <li>sensitive</li> </ul>	replace each digit by #	ALS / A + A + A + A + A + A + A + A + A + A		Minimum	freq 1	\$			Sensitive	replace each digit by #	
		finimum docfreq	1	\$			leave as is	Total Files 67		Minimum	docfreg 1	\$				leave as is	
100	6							Files Processed 100%									

### 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:

tep 2. Save the results via one of the File men

- File->Save Display Results As...
  - $\circ$   $\;$  Save the filtered results (i.e. those shown on the screen)
- File->Save All Results As...
  - o 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 nonprofit 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**

1.2.0 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

Copyright: Laurence Anthony 2019