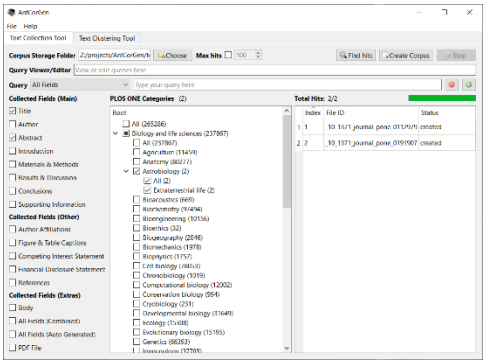
# AntCorGen (Windows, MacOS, Linux)

## Build 1.3.2

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
April 6, 2026




## Introduction

*AntCorGen* is a freeware corpus generation tool. AntCorGen lets you search for documents in the PLOS ONE research database via search queries and/or subject category browsing and decide which sections (e.g. title, abstract, introduction) of these documents should be stored. *AntCorGen* then accesses the database, downloads the sections, and saves each one as a text file in an appropriate folder. *AntCorGen* can also analyze the different parts of speech (e.g. adjectives, verbs) of words in the files and cluster similar sentences into sub-groups. These sub-groups will show similar patterns of language use.

*AntCorGen* runs on any computer running Microsoft Windows (tested on Win 7), 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 - Installer

Double click the *AntCorGen.exe* file and follow the instructions to install the application into your Programs folder. You can delete the .exe file when you are finished. You can start the application via the Start Menu.

### Windows - Portable

Unzip the *AntCorGen.zip* file into a folder of your choice. In the *AntCorGen* folder, double click the *AntCorGen.exe* file to launch the program.

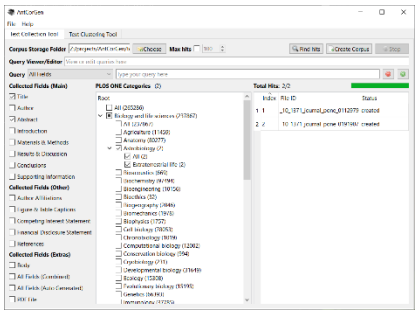
### Macintosh OS X

Double click the *AntCorGen.dmg* file to create a AntCorGen disk image on your desktop. Open the disk image and drag and drop the AntCorGen app onto the Applications folder (or into another location if you desire). You can then launch the app by double clicking on the icon in the Applications folder or the Launchpad.

### Linux

Decompress the *AntCorGen.tar.gz* file into a folder of your choice. In the *AntCorGen* folder, double click the *AntCorGen.sh* file to launch the software. On the command line, type ./AntCorGen.sh to launch the software.

## Text Collection - Quick Guide



**Step 1:** Select a corpus storage folder into which the corpus files will be saved using the "Choose" button.

**Step 2:** Choose documents to be included in the corpus collection.
**Option A:** Search for relevant documents using the "Query Viewer/Editor" and/or "Query" settings:

1) The "Query Viewer/Editor" will show a complete query in the Solr search query language used by PLOS ONE. More information about the query language can be found at the following links:
    - [Tutorial] http://www.solrtutorial.com/solr-query-syntax.html
    - [Examples] http://api.plos.org/solr/examples/
    - [Main Solr site] http://lucene.apache.org/solr/
2) The "Query" tool will allow you build a query using field names, queries items, and AND/OR/NOT operations. To add/delete parts to the query, using the -/+ buttons. All changes made in the "Query" tool will be reflected in the "Query Viewer/Editor".

**Option B:** Browse for relevant documents using the "PLOS ONE Categories" browser tree:

1) Click on category branches in the browser tree to expand the branches and show sub-categories. The number of documents in each category is shown in parentheses.
2) Select categories to be included in the collection. The total number of documents within the selected categories is shown in the browser tree header.

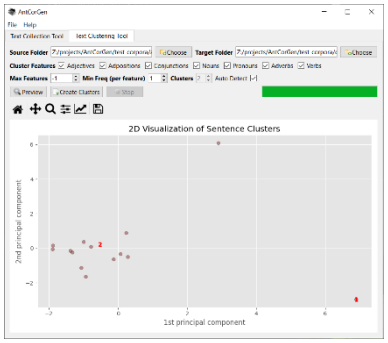
**Step 3:** Decide whether or not to set a maximum number of corpus files to collect using the "Max hits" checkbox option and spinbox values widgets.

**Step 4:** Click the "Find Hits" button to show an estimate of the total number of documents ("Total Hits") that will be collected. The result is shown in the status window.

**Step 5:** Click the "Create Corpus" button to collect the documents and store them in the corpus storage folder. The total number of documents (hits) will be updated to show how many have been collected. The id and status of the collection for each document is shown in the status window.

**Step 6:** Click the "Stop" button to stop the collection process at any time.

## Text Clustering - Quick Guide




**Step 1:** Select a "source folder" of corpus files that you want to cluster using the "Choose" button.

**Step 2:** Select a "target folder" into which the clustered files will be saved using the "Choose" button.

**Step 3:** Choose features that you want to include as part of the clustering algorithm.

**Step 4:** Choose parameters (max number of features, min frequency of features, number of clusters) that you want to use in the clustering algorithm. To use all the possible features, set the "Max features" option to -1 (the default). If you are not sure how many clusters to pick, use the "Auto Detect" option.

**Step 5:** Click the "Preview" button to show a scatter-plot visualization of the clusters. If the clusters are not separated in the visualization, adjust the features and parameters as necessary. The scatter-plot can be resized, zoomed, label-adjusted, and saved using the icons above the plot image.
**Step 6:** Click the "Create Clusters" button to cluster the document sentences and store them in the target folder.

**Step 7:** Click the "Stop" button to stop the clustering process at any time.

## 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 *AntCorGen*

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

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

For example if you download *AntCorGen 0.0.1* which was released in 2017, you would cite/reference it as follows: Anthony, L. (2017). *AntCorGen* (Version 1.0.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

## KNOWN ISSUES

None at present.