**Coassociation Grapher User Manual**

## What is the Coassociation Grapher?

Given a particular word of interest (a “focal token”) in a particular corpus of text, the Coassociation Grapher allows you to view the relative probability of observing other terms (“bound tokens”) before or after the focal token (anywhere from 999 words prior to the focal token to 999 words after it). For example, in the following graph, we see clearly that the focal token “cat” is far more likely to appear immediately after the word “black” then it is to appear at a randomly selected point in the corpus—over 25 times more likely, in fact.



Such graphs can yield information about the overall patterns with which two words appear relative to each other in a body of texts. Furthermore, because this tool allows for particular timeslices to be selected from a diachronic corpus, it allows for comparisons of the ways words conjoin with each other (or fail to) across time.

## Installation

 **Please note that this tool will only run on Microsoft Windows.**

1. Download the tool from its current location (currently <https://concept-lab.lib.cam.ac.uk/grapher/ca-grapher.zip> ). If this link is broken, the current location of the tool should be described at <https://concept-lab.lib.cam.ac.uk/> .
2. Unzip (extract) the tool to a folder. To do this, find the file after it has downloaded, right-click it and select “Extract All…”
3. To run the tool, double-click on **ca\_grapher.exe**. If you have downloaded the tool only, this will be located in the **Debug** folder. If you have downloaded the version of the tool that includes the source code, this will be located in **grapher/bin/Debug**.

**Troubleshooting**

Because this application has not been signed by a publisher and is rarely downloaded, Windows may prevent it from starting. If so, you may receive the following message:



To get around this, you can try the following:

a. Click the underlined “More info”. This should give you more information, as well as a “Run anyway” option. If so, click “Run anyway” to run the application.
b. If this does not work, you may need to temporarily turn off SmartScreen:

* Open the Start Menu by clicking in the bottom left hand corner, or by pressing the Windows Key.
* Type “Windows Defender Security Center” and click on it to open it.
* Click *App & browser control*.
* Under the sub-heading Check apps and files, select the Off button.

## Interface and features: initial screen

Upon opening the Coassociation Grapher, you are presented with the following interface:



The input parameters governed by the different controls are as follows.

**Focal token:** Word that the graph is centered around.

**Bound tokens:** One or more bound tokens can be specified on separate lines or separated with commas. The conditional probability or relative conditional probability of occurrence of the focal token, conditioned on these bound tokens, will be plotted on separate lines.

**Exclusions:** It is possible to exclude certain documents from the subset of the corpus from which the counts are derived.

To specify that counts should exclude documents that contain particular words, type those words into the Exclusions field and prefix them with a hyphen. For example, “-sublime -beauty” will exclude documents containing either the word sublime or the word beauty.

To specify that counts should exclude cases in which a particular word falls between the focal and bound tokens, prefix those words with a tilde. For example, if the focal token is “sublime” and the bound token is “beauty”, then specifying “~beautiful ~beauty” in the Exclusions field would count up all cases in which “beauty” precedes or follows “sublime”, as long as neither the word “beautiful” nor the word “beauty” falls between them.

**Max lines**

Maximum number of bound tokens that the tool will attempt to put on a single graph before creating multiple graphs.

**Proximity**

Maximum distance (in both directions away from the focal token) to display on the graph.

**Corpus**

Corpora for which coassociation graphs can be constructed are as follows:

*ECCO:* Eighteenth Century Collections Online, 1701-1800
*EEBO:* Early English Books Online, 1600-1700
*msbooks*: The British Library Labs 19th Century Books corpus, 1800-1899[[1]](#footnote-1)
*NANT:* The North American News Text Corpus
*NYT:* The New York Times Annotated Corpus, 1987-2007
*Times:* The Times Online Digital Archive, 1785-2010 *Aberdeen, Edinburgh, Glasgow:* Subsets of ECCO consisting of books having Aberdeen, Edinburgh, or Glasgow (respectively) in their “publisher’s imprint” metadata
*aberdeen-filtered, edinburgh-filtered, glasgow-filtered:* Subsets of ECCO consisting of books having Aberdeen, Edinburgh, or Glasgow (respectively) in their “publisher’s imprint” metadata, and further restricted to books which also contain in their imprint a publisher name from the set of notable Scottish Enlightenment publishers identified by Dr John Regan
*adamsmith:*An Inquiry into the Nature and Causes of the Wealth of Nations (1776)
Lectures on Justice, Police, Revenue and Arms (1763)
The Theory of Moral Sentiments (1759) *hume:*A Treatise of Human Nature [1739]
Enquiries Concerning the Human Understanding and Concerning the Principles of Morals [1751][[2]](#footnote-2)
The Natural History of Religion (1889 ed.) [1757].
Essays Moral, Political, Literary (LF ed.) (Essays and Treatises on Several Subjects) [1777]
The History of England [1778], all volumes (excluding front material: the preface & the 'LETTER FROM ADAM SMITH, LL.D. TO WILLIAM STRAHAN, ESQ.')

**Vertical axis**

If this is set to “Conditional probability”, what is plotted is the probability of observing the focal token, given that the bound token has been observed at the given distance. In this case the baseline will be the overall probability of observing the focal token in the corpus.

If this is set to “Relative probability”, you get the conditional probability, but normalized by the overall probability of observing the focal token in the corpus; the baseline in this case is therefore always 1. The lines in this case answer the question “how much more frequently do you observe the focal token at this distance away from the bound token *than you would expect to* given the focal token’s overall frequency in the corpus.*”*

**Graph type**

“Original” displays the graph as previously described under “Bound tokens”. “Two-baseline” can be used if there is only one word in the “bound token” list. In addition to the standard graph, it creates another version of the graph in which the bound and focal tokens are flipped, and overlays it on top of the first.

**Start year**

The “start year” of the temporal slice of the corpus to use; this option will only appear for diachronic corpora. Documents with publication dates earlier than this year will not be included in the counts that are used to derive the probabilities displayed in the graphs.

**End year**

The “end year” of the temporal slice of the corpus to use; this option will only appear for diachronic corpora. Documents with publication dates later than this year will not be included in the counts that are used to derive the probabilities displayed in the graphs.

**Only use documents that are *X* words long**

If checked, excludes documents shorter than *X* words long from the counts that are used to derive the probabilities displayed in the graphs. *X* is 200 by default but can be set to any positive integer less than one million.

**Delete contents of download folders when grapher is closed**

To avoid repeatedly downloading large files if you are repeatedly making graphs that contain some of the same words, the grapher caches files it downloads from the internet into a folder on your computer called “Downloaded index files”. However, these take up space on your computer. So, when this is checked, the contents of this folder will be auto-deleted when the grapher is closed.

**Load graph from backup**

For every graph created by the grapher, the numbers required to generate the graph are automatically saved in migrated-current\grapher\backup. To redraw a previously generated graph without waiting, you can click this text and load a graph from migrated-current\grapher\backup. Note that you may want to clear out the contents of migrated-current\grapher\backup periodically.

**“Go” button**

Upon clicking this button, the necessary files to compute the probabilities are downloaded from the server if they are not already cached on disk, and the coassociation graph is displayed.

## Interface and features: graph screen

In the following example, we walk through the interpretation of the results of an example graph with *monarchy* as the focal token and *tyranny, kings, king, kingly* and *cat* as bound tokens.

**Explanation of the graph**



In the legend, “freq” is frequency of the token in the specified timeslice. “Peak” is the number of occurrences (e.g., the number of of co-associations of the focal and bound token) that the highest point on the bound token’s line corresponds to. “Percentile” gives an indication of how frequent each word is. For example, of all words in the lexicon[[3]](#footnote-3), *monarchy* is in the top 18.5% in terms of its overall frequency.

Notice that although there are frequent “spikes” for the word *kingly*—for example, we are more than 120 times more likely to see the word *monarchy* two words before *kingly* than we are to see it at a randomly selected location in the corpus—the import of this is mitigated by the fact that *kingly* is a relatively rare word. We can see from the legend that is is in the bottom 39% of all words in the overall ECCO lexicon in terms of its frequency, and that the data that underlies the highest “kingly” peak in the graph only consists of 31 cases of the word “kingly” appearing two words after “monarchy”. Therefore we may not wish to read too much into the specific distances at which “monarchy” and “kingly” coassociate. Somewhat more meaningful may be the increased frequency of “tyranny” shortly after the word “monarchy” (as opposed to before it). We can deselect the **Variable height (size to fit)** option button on the right side of the screen by simply selecting the **Fixed height** button, and changing the number in the associated text box. This allows us to specify the height of the vertical axis to “zoom in” vertically to get more fine-grained detail on the graph. When doing so, we can get a clearer sense of what is happening. For example, it has now become apparent that *monarchy* is actually *less* likely to appear in the presence of the word *cat* than it would be to appear in a randomly selected location within the corpus.



Aside from the aforementioned “Variable height” and “fixed height” controls, the other controls in the right sidebar are as follows:

**Smoothing**

Setting this to “1” displays exact values. Setting this to a higher number *N* displays a so-called “moving window” average of the *N* data points to the right and left of every location on the horizontal axis. This has the effect of “smoothing” out the display and making general trends more apparent, at the expense of precision.

**Zoom**

Controls horizontal zoom (allowing the horizontal endpoints of the graph to be constricted), allowing one to see the distances close-up to the focal token in more detail.

**“Label” checkboxes: Baseline value(s), smoothing, highest peak, frequencies, corpus, percentile**

If checked, includes the corresponding label on the graph.

***Save****:* Saves an image file of the current graph in the *grapher* folder.
***Save as***: Saves an image file of the current graph, allowing you to specify a location.

***Save All****:* Saves an image file of all graphs that have been generated when you clicked the Go button, in the *grapher* folder. This is only relevant if you included more bound tokens than could be included on a single graph.

***Next, Back****:* Allows you to click through all graphs that were generated when you clicked the Go button. These buttons will generally be disabled; they are only relevant if you have included more bound tokens than could be included on a single graph.

NOTE: The percentiles in the grapher legend are based on cached word frequencies in the overall corpus, even if we are technically looking at a subcorpus due to the fact that we are restricting ourselves to documents of at least a certain length, or only counting tokens at least *n* words away from the beginning/end of the document.

## Caveats

There is currently an inconsistency in how multiple-word tokens are handled for indexed corpora and multi-word corpora *in the special case where one multi-word bound token that we care about* contains *another, shorter bound token that we also care about*. Suppose you are looking at binding between the focal token “democrat” and bound tokens “president” and “vice president”. Should the counts for “president” include cases where “president” appears as part of “vice president”, or not? There’s a reasonable case for ‘yes’, and a reasonable [better?] case for ‘no’ [namely: you can always get the results for ‘yes’ by plotting separately]. For all corpora, the code currently opts for “no”.

## Temporary files created by the grapher that can be safely deleted

The grapher creates backups of the data points of the graphs that are created within it. These can be found in the folder *grapher/backup*. Further files downloaded by the grapher may be stored in *grapher/temp*.

Finally, to avoid repeatedly downloading large files if you are repeatedly making graphs that contain some of the same words, the grapher caches files it downloads from the internet into the folder “*grapher*/*Downloaded index files*”.

However, these take up space on your computer, so the grapher provides a checkbox entitled "*Delete contents of download folders when grapher is closed*". When this is checked, the contents of these folders will be auto-deleted when the grapher is closed.

1. See [http://blogs.bl.uk/digital-scholarship/2013/12/a-million-first-steps.html#](http://blogs.bl.uk/digital-scholarship/2013/12/a-million-first-steps.html), <https://www.youtube.com/watch?time_continue=7&v=OGzsDQ_Nhyk> [↑](#footnote-ref-1)
2. Note: "Enquiry concerning the Principles of Morals" was first published in 1751, and "Enquiry concerning Human Understanding" in 1748.
Both were published together in "Enquiries Concerning the Human Understanding and Concerning the Principles of Morals" in 1777; it is this posthumous edition that appears in our tool. Nonetheless the date associated with it in the tools is set to "1751" so that it can be distinguished from the other 1777 collection "Essays Moral, Political, Literary (LF ed.)". Sources: <http://oll.libertyfund.org/titles/hume-a-treatise-of-human-nature>, <http://oll.libertyfund.org/titles/hume-enquiries-concerning-the-human-understanding-and-concerning-the-principles-of-morals>, <http://oll.libertyfund.org/titles/hume-the-natural-history-of-religion>, <http://oll.libertyfund.org/titles/hume-essays-moral-political-literary-lf-ed>, <http://oll.libertyfund.org/titles/hume-the-history-of-england-6-vols>. [↑](#footnote-ref-2)
3. Note: that’s all words in the *lexicon –­* that is, the set of words whose frequency threshold is high enough to be included in the tools *–* not all words in the corpus. [↑](#footnote-ref-3)