**Shared Lexis Tool User Manual**

Table of Contents

[What is the Shared Lexis Tool? 2](#_Toc533335026)

[Installing the tool 3](#_Toc533335027)

[Interface and features 4](#_Toc533335028)

[Starting up 4](#_Toc533335029)

[Basic operation 5](#_Toc533335030)

[Changing the text size and panel width 7](#_Toc533335031)

[Identifying words shared between two or more lists 8](#_Toc533335032)

[Additional controls 9](#_Toc533335033)

[Menus 10](#_Toc533335034)

[Temporary files created by the shared lexis tool that can be safely deleted 12](#_Toc533335035)

[Contextual search / “text grab” functions (available by request) 13](#_Toc533335036)

# What is the Shared Lexis Tool?

The Shared Lexis Tool provides a means of exploring all of the words that are statistically associated (viz., coassociated[[1]](#footnote-1)) with a word provided by the user, at a given distance, in a given corpus of text, using a given measure of coassociation, over a given date range. In contrast to the Coassociation Grapher, which requires all words of interest to be provided explicitly by the user, the Shared Lexis tool allows one to discover what terms are are highly statistically associated with a word of interest, and to explore how these differ across time/corpora/distance/measures.

One of the key tenets of the Concept Lab is that concepts are not identical to words. Rather, concepts leave ‘traces’ in the lexical record: evidence for a given conceptual structure may be distributed across many lexical items, or in the mutually reinforcing behaviour of a small number of words. The Shared Lexis Tool has been designed to help discover such patterns of lexical behaviour. When plotted in rank order, the terms co-associated with any given word exhibit a characteristic “decline curve” whereby a relatively small number of highly associated words slowly gives way to a mass of weakly associated ones. The small number of highly associated words—those above the “bend in the curve” (Fig. 1) —and the strengths of their associations constitute a ‘signature’ list that can be compared with those of other words (Fig. 2).

In some cases, comparison of multiple signature lists can reveal heretofore unknown properties of a word’s contexts of use, or those of an entire group of interrelated words. For example, the words “monarchy,” “aristocracy,” and “democracy” all appear very near to the very top of each others’ signature lists in Eighteenth Century Collections Online (to a strength and degree that far outstrips all other words for political concepts in the eighteenth century). This suggests that all three terms are conceptually bound up to an enormous degree, a fact which demands further qualitative investigation in the form of close reading. The Shared Lexis Tool also facilitates this kind of investigation with search functionality that allows texts which match particular criteria to be located within the corpus.

# Installing the tool

**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/sharedlexis/sharedlexis.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 **sl.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 **sharedlexis/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:



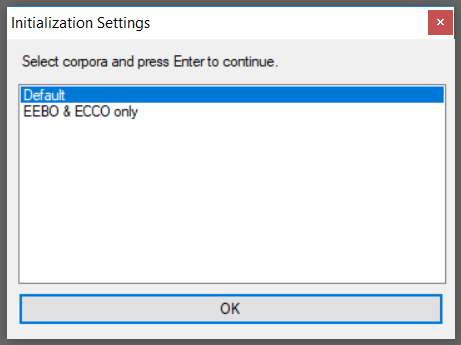
To get around this, you can try the following:  
  
a. Click the underlined “More info”. This should give you more information, as well as a “Run anyway” option. If so, click “Run anyway” to run the application.  
b. If this does not work, you may need to temporarily turn off SmartScreen:

* 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

## Starting up

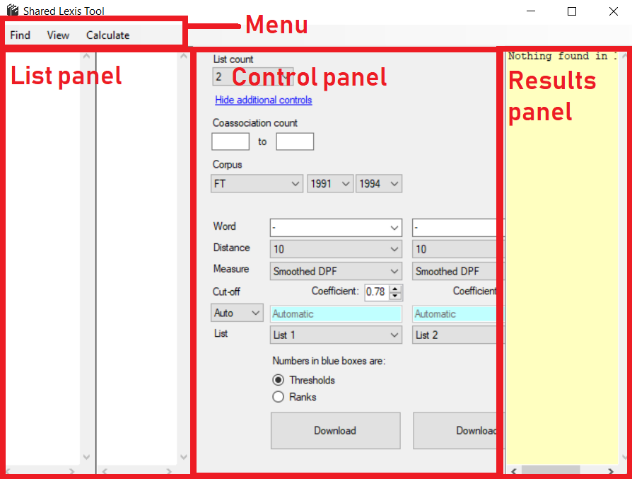
When you first start the tool you will see an Initialization Screen similar to this:



This shows various different sets of corpora that you can choose to load into the tool. To access the corpora most frequently used by the Concept Lab, you can leave it on Default and click OK.

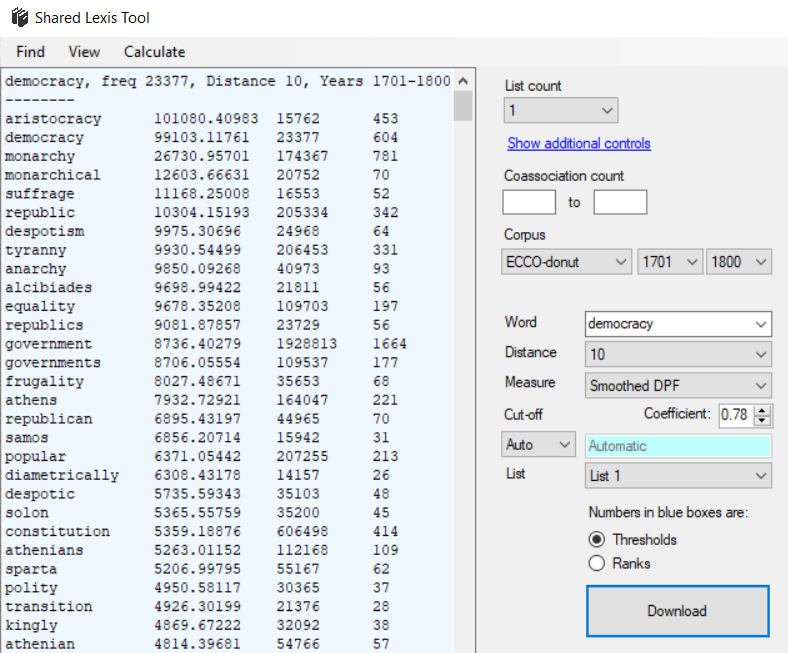
Note that some of these sets contain fewer corpora than are available in the Default set. This is so if you’re only interested in a few corpora (say, EEBO and ECCO only) and you don’t want to wait for all corpora to be loaded up, you can select a set that contains only the corpora you are interested in.

After you click OK and the corpora are loaded, you will see the main interface. The “list panels” are to the left, the “control panel” in the center, the “results panel” on the right and the “menu” up top.



## Basic operation

The basic operation of the tool is to permit the user to select a word **(“focal token”)**, corpus, timeframe, and measure of association and to display lists of terms **(“bound tokens”)** that are strongly associated with it. For example, in the figure below, we can see that the user has selected “ECCO-donut” as the corpus, 1701-1800 as the date range, “democracy” as the focal token, “smoothed DPF” as the measure, and “10” as the distance. After the user selected these parameters and clicked the “Download” button, the leftmost list panel was populated with a list of the words that are most closely associated with “democracy” within ECCO from 1701-1800 at this distance. These parameters are explained in more detail in the following sections.



Some further features of the above image require explication:

***List panel (far left).***This shows the words that are most associated with “democracy” at a distance of 10 in ECCO from 1701-1800, according to the default association measure of *smoothed DPF* (explained later). They are sorted by their degree of association and displayed in the format  
*word association frequency number-of-coassociations*

For example, in the image we can see that “aristocracy” is the word most associated with “democracy” on these parameter settings, with a smoothed DPF of 101,080; it appears 15,762 times in the corpus as a whole within the specified date slice (which in this case was 1701-1800), and 453 of those times it appeared at a distance of approximately 10 words away from “aristocracy” (specifically, 8 to 12 words to the left or right).

***List count.***More than one list pane can be open at a time. You can see which words appear on two or more lists in the yellow *results pane* on the far right, discussed later.

***“Show/Hide additional controls”.***Clicking this link toggles the visibility of additional controls on the control panel, discussed later.

***Coassociation count.*** These boxes allow you to set a minimum and maximum number of coassociations that a bound token must have with the focal token in order to appear in one of the lists.

***Corpora.*** Allows you to select the corpus that you want your analysis to be based on.

***Year dropdowns.*** Allows you to select a range of years (e.g., 1701 to 1710); the resulting analysis will be based only on documents with publication years in that range. The range is inclusive: the first year and last year are included in the analysis. To conduct an analysis over documents in only a single year, make sure that the start and end years are set to the same year (e.g., 1759 to 1759). These dropdowns may not be visible on corpora that have not been subdivided by year.

***Word.*** The focal token; the word that the analysis will be with reference to.

***Distance.*** The number of words away from the focal token that the coassociation window is centred around; all analyses in the shared lexis tool use a window size of 5. Because the selected distance in the example image above was 10, the association scores appearing in the list were computed on the basis of all words appearing 10 words away from every instance of the focal token (“democracy”) in the corpus, using a window size of 5 words (so, 8, 9, 10, 11, or 12 words away). Because the selected corpus was “ECCO-donut”, these 5-word windows were calculated to the left and to the right[[2]](#footnote-2) of every instance of “democracy”. Counts for all corpora have been computed using this “donut” measure unless the corpus is described as “directional” (in which case counts are computed by looking to the right of the focal token, rather than to the left and to the right).

***Measure.*** The default measure, “Smoothed DPF,” uses a measure we have referred to variously as the “smoothed distributional probability factor (dpf)”, and “simple smoothing” in Recchia & Nulty (2017)[[3]](#footnote-3). This is a simple variant of pointwise mutual information (PMI) which drops the logarithm and introduces a smoothing exponent in the denominator because we are primarily interested in the rank order of tokens as they are co-associated with a focal token *x*. Since the logarithm does not affect this rank ordering it can be dropped without loss of generality. Doing so highlights the fact that this is essentially a small modification to the measure sometimes referred to simply as ‘observed over expected’: the number of times two words are observed in conjunction, divided by the number of times one would expect to see them together by chance. The measure is computed as follows:

|  |  |
| --- | --- |
|  | (2) |

-- the probability of observing *x* and *y* together -- is estimated by dividing the total number of times *x* and *y* are observed “together” by the total number of tokens N in the corpus, e.g. Because the code treats it as a co-occurrence (i.e., *cooc(x, y)* is incremented) if *x* appears at any one of ten locations (-*d-2,* -*d-1,* -*d,* -*d+1,* -*d+2,* +*d-2,* +*d-1,* +*d,* +*d+1,* or +*d+2*)relative to any instance of the focal token *y,* this meansthat if we were to assume that tokens were uniformly (randomly) distributed, the number of times we would expect to see *x* co-occur with *y* would be ; this is the quantity that is used to estimate   
  
When we calculated the value of α that minimized the inverse correlation between rank and frequency described in the previous section, we found the optimal value for our corpus to be 0.78, very close to the value of 0.75 that has been cited in the literature (without any theoretical justification) as one that works well on practical tasks. Higher values corresponded to a negative correlation between rank and frequency, while lower values corresponded to a positive correlation. In other words, with an appropriate value of alpha, smoothed DPF is sensitive to data sparsity as well as frequency, and retains the simplicity and transparency of PMI.

***Coefficient.***The value of alpha to be used in the Smoothed DPF formula previously described.

***Cut-off.***Defines the criteria that determine the length of each list, i.e., the number of words that appear on it. If set to “Auto,” a power curve will be fit to the ‘decline plot’[[4]](#footnote-4) and the list terminated at the “bend in the curve,” e.g. the point at which the derivative of the curve is equal to -1.

If set to “Manual”, the user can manually specify a numeric cut-off. The user may choose to specify either:

* a minimum *threshold value* (e.g., the minimum smoothed DPF) that a word must achieve in order to be included on the list. If the “Thresholds” radio button is selected, the number in the blue box will be interpreted in this way. [[5]](#footnote-5)
* The maximum *rank* to include. If the “Ranks” radio button is selected, the number in the blue box will be interpreted in this way. For example, selecting “Ranks” and putting 20 in the blue box would result in only the top 20 bound tokens (i.e., the 20 bound tokens that score highest on the specified association measure) being displayed in each list panel.

***List.*** Which list pane to place the results in when the “Download” button is clicked (List 1, List 2, etc.)

***“Download” button.*** Click this button to download and/or update a list after parameters have been selected or changed.

Changing the text size and panel width  
  
You can change the text size in either of the following ways:

* Hover the mouse over the list panel or results panel, hold down the Ctrl key, and scroll the mouse wheel up or down.
* Alternatively, you can click in the list panel or results panel, and press Ctrl+= (the equals sign while holding down Ctrl) to increase the font size, or Ctrl+- (the minus sign while holding down Ctrl) to decrease the font size.

You can change the panel width in either of the following ways:

* Hover the mouse over the list panel or results panel, hold down the Shift key, and scroll the mouse wheel up or down.
* Alternatively, you can click in the list panel or results panel, and press Shift+= (the equals sign while holding down Shift) to increase the panel width, or Shift+- (the minus sign while holding down Shift) to decrease the panel width.

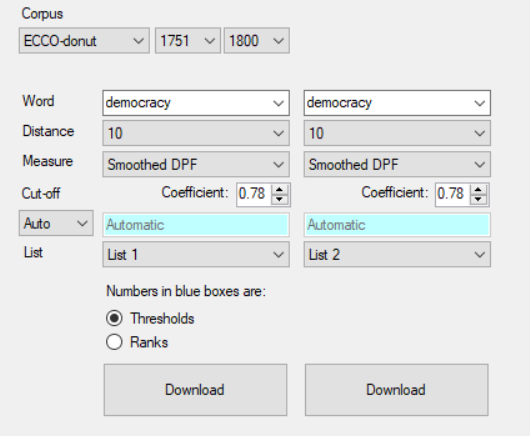
### Identifying words shared between two or more lists

The “shared lexis” tool is so named because one of its primary functions is to identify words that are shared among one or more lists. Here are two methods for doing so.

*Method 1: Using the “List” dropdown (underneath the pale blue threshold box)*

* Change the “List Count” (first dropdown on the control pane) to 2.
* Select a word, distance, and timespan – for this example, select “democracy”, distance 10, 1701-1750 – and click “Download” to download the list of words that are most associated with “democracy” at distance 10 from 1701-1750. The list length is determined by the ‘cut-off’ criteria described earlier.
* Change the word, distance, date range, and/or corpus as you please. For this example, change the date range to “1751-1800”.
* Change the “List” to “List 2” and click “Download”. The list of words most associated with “democracy” at distance 10 from 1751-1800 appears in List 2.

*Method 2: Using additional controls*

* Change the “List Count” (first dropdown on the control pane) to 2.
* Change the date range from 1701 to 1750.
* If the blue link underneath the List Count dropdown reads “Show additional controls”, then click it to display a second set of controls for the word, distance, etc. Note that we now have a leftmost and rightmost set of identical controls:  
  
* Ensure that the “List” dropdown reads “List 1” in the leftmost set, and “List 2” in the rightmost set.
* Select a word and distance in the leftmost set of controls – for this example, select “democracy”, distance 10 – and click the left “Download” button to download the list of words that are most associated with “democracy” at distance 10 from 1701-1750 into list 1. The list length is determined by the ‘cut-off’ criteria described earlier.
* Change the word, distance, date range, and/or corpus as you please. For this example, change the date range to “1751-1800”.
* Select a word and distance in the rightmost set of controls – for this example, select “democracy”, distance 10 – and click the right “Download” to download the list of words that are most associated with “democracy” at distance 10 from 1750-1800 into list 2.

After completing the steps listed in Method 1 or Method 2, we should now see a list of the words that are most associated with “democracy” at distance 10 from 1701-1750 in List 1, and a list of the words that are most associated with “democracy” at distance 10 from 1751-1800 in List 2.

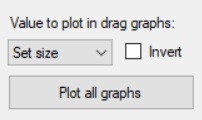
We can now look in the yellow results pane to see: the sum and mean of the associations in each list; the number of words that appear on exactly 2 lists (i.e., both lists) and the number that appear on exactly 1 list (i.e., one or the other list, but not both); the percentage of words in each list that are “shared by all” (i.e., that appear on both lists); and the percentage of the words that appear on *any* list (i.e., at least one list) that are “shared by all.”

Note that we can change the list count to 3, 4, or more to see what is shared among 3 or more lists. If the list count we have selected is 3 or more, we will have to use the “List” dropdown underneath the pale blue threshold box (Method 1) in order to download data into list 3, 4, etc.

### Additional controls

Clicking the *“Show/Hide additional controls”* link toggles the visibility of additional controls on the control panel. Some of these controls are duplicate dropdowns, buttons, etc. that simply provide an alternative means of downloading data into multiple lists; this functionality is described in the previous section “Identifying words shared between two or more lists,” Method 2.

The other additional controls allow you to view various graphs that visualise the data in different ways, and appear underneath the “Download” button:



***Value to plot in drag graphs.*** When the “plot all graphs” button is clicked, various graphs are generated, the first of which plots how the focal token’s “drag” changes as distance increases. There are different ways to quantify “drag”. In this dropdown, one can choose from the following options:

* *Set size.* This is the total number of words exceeding the specified threshold (or above the bend in the curve if the threshold is set to “Automatic”).
* *Median.* This is the median coassociation (e.g., the median smoothed DPF) of all words exceeding the specified threshold (or above the bend in the curve if the threshold is set to “Automatic”).

***“Invert” checkbox.***

The construct of “drag” can be conceived of either as something which increases as the number of words exceeding a particular threshold increases (e.g., if a focal token’s drag is high, more words are “dragged in” to its conceptual orbit), or the reverse (e.g., the fewer words that are in a focal token’s conceptual orbit, the more relative influence it has on each (at least, this is one hypothesis)). Selecting this checkbox inverts the “drag” measure (i.e., plots the reciprocal of set size or median, depending on what was selected in the adjoining dropdown), corresponding to this latter interpretation.

***“Plot all graphs” button.*** Plots a selection of graphs illustrating changes in the properties of the coassociation lists associated with the current focal token under different parameter settings. As usual, the coassociation lists are composed of all words exceeding the specified threshold (or above the bend in the curve if the threshold is set to “Automatic”). The graphs generated are as follows:  
  
*Drag.* Plots how the focal token’s “drag” changes as distance increases. See the previous two subsections (“’Invert’ checkbox”, “Value to plot in drag graphs”) for more detail about this visualisation.

*Shared Lexis – Bar Chart.* This graph illustrates the properties of what is *shared* between coassociation lists of *adjacent* distances (one list more proximal, one more distal). For example, the first bar contains information about the relationship between distance 5 (more proximal) and distance 10 (more distal). The second bar contains information about the relationship between distance 10 (more proximal) and distance 20 (more distal); the third bar contains information about the relationship between distance 20 (more proximal) and distance 30 (more distal); and so on. Specifically:

* The height of the bar corresponds to the total number of items on the more distal list.
* The percentage indicates the percentage of words that were “retained”: the percentage of words on the more distal list that were also on the more proximal list. The hue of the bars vary from bright red when this is 100% to a dark blue at 0%.
* The number preceded by the + sign indicates the number of words that were “introduced”: the number of words on the more distal list that were *not* also on the more proximal list.

*Shared Lexis – Grid.* The rows of this chart are labeled with distances (5, 10, 20, 30, 40, 50…) as are the columns. Let a distance labelling a given column be denoted by *c* and a distance labelling a given row be denoted by *r.* The cell of this grid at location *(r, c)* contains the percentage of lexis on the list at distance *c* that is shared with the list at distance *r.*

*Decline plots.* For each distance, these display the curve that results when the coassociation measure (e.g., smoothed DPF) is plotted on the y-axis, against the rank (whether a term is #1 on the list, or #2, or #3…) on the x-axis.   
  
*“Save” and “Save all” buttons*. These save plots as image files in the folder “Shared lexis tool graphs”.

### Menus

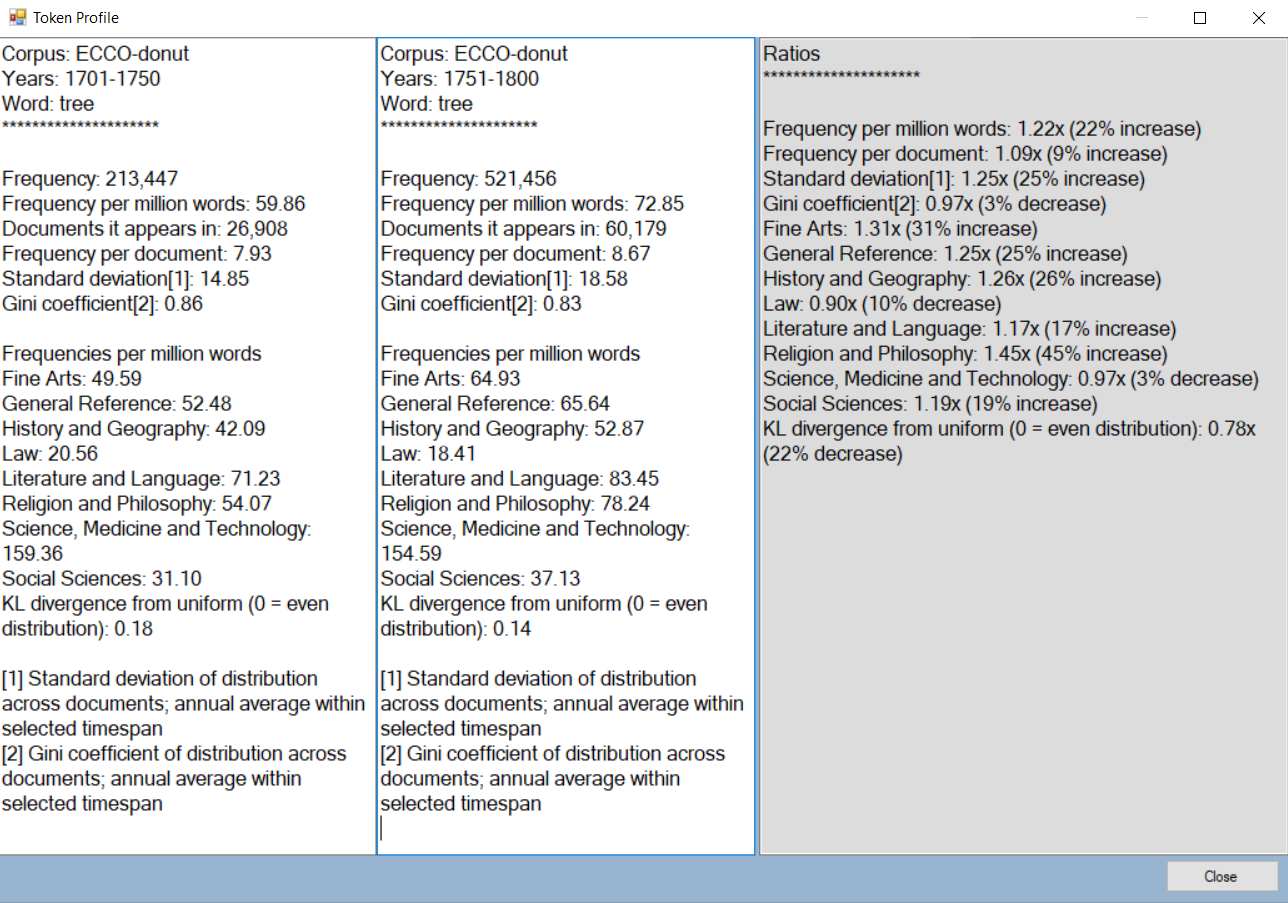
This describes the functionality of the main menu at the top of the screen.  
  
**Find menu.** Selecting “Find in lists…” (Ctrl+F) allows one to search for a particular word or part of a word in all open lists. To continue searching for other examples of the term, you can repeatedly press F3 or select “Find again”.

**View menu.**   
View -> Profile.(Only applies to the ECCO corpus). By clicking on a list and going to View->Profile… (Ctrl+P), you can view various statistics about the focal token as it appears within the date span corresponding to the selected list. These include the word’s frequency, frequency per million words, number of documents it appears in, frrequency per document, standard deviation, and Gini coefficient.

You can also view the term’s frequency (per million words) in each ECCO topic subset: Fine Arts, General Reference, History and Geography, Law, Literature and Language, Religion and Philosophy, Science Medicine and Technology, and Social Sciences.

Standard deviation, Gini coefficient, and “KL divergence from uniform” can be viewed as measures of “inequality” with respect to how the term is distributed across documents (in the case of standard deviation and Gini coefficient) and topic areas (in the case of KL divergence). Higher numbers correpond to less equitable distributions across documents (or topics).

Finally, by copying and pasting the contents of one profile window into another, one can see the relative increase or decrease in each measure as one moves from the profile on the left to the profile on the right:



**Note:** Be aware that the reported “increase” or “decrease” in each measure always divides the numbers on the profile on the right by the numbers on the profile on the left, irrespective of their contents. In the example above, the left profile is 1701-1750 and the right is 1751-1800, so the 22% increase in frequency per million words indicates that the word’s relative frequency increased in the second half of the century. However, if the 1751-1800 profile had been on the left and the 1701-1750 profile on the right, it would have indicated a *decrease.* It is therefore always best to compare profiles with the oldest one on the left, for ease of interpretation.

View -> Stickiness Graph.View->Stickiness Graph brings up a graph that answers the question: Out of all the lists visible in the shared lexis token right now, what is the overall number of tokens that are shared by *N* lists? If four lists were visible, this would illustrate the overall number of tokens that (1) appear only on a single list; (2) are shared by exactly two lists; (3) are shared by exactly 3 lists; and (4) appear on all four lists.

View -> Sort Lists By…By default, lists in the list panel are sorted by coassociation value (e.g., smoothed DPF), but can also be sorted by frequency or total number of coassociations.

View -> Sort Output By…By default, words in the yellow results pane are sorted by their summed coassociation value (e.g. smoothed DPF) across all lists, but can also be sorted by their value on one particular list.

View -> Shortcuts Lists keyboard and mouse shortcuts for various functions.

View -> Options… Specifies options for “contextual search” functionality described earlier.

* Note that here, “window size for contextual search” truly refers to *window size around the bound token,* not distance. If the “Window size for contextual search” option there is set to anything other than 5, the number of matches in the contextual search window will not match the number of coassociations used to calculate DPF values.
* Changing “number of words to show before/after each coassociation” alters the overall amount of context text displayed when the contextual search functionality is deployed.
* “Special request” words refer to words that are not in the lexicon proper, such as “disinterestedness” and other focal tokens added by request of Pete, Ewan, Jamie Latham and Philip Connell.
* Feel free to click “Clear” if the “Size of files cached on hard drive” is quite large and you would like to free up space on your computer. The only downside is future searches may be slower as it will have to re-download any files that you may already have downloaded in the past.

**Calculate menu.** Allows the user to treat lists as if they are word vectors and perform elementwise addition, subtraction, multiplication and division. A word not appearing on a particular list is treated as if it has a value of 0. Common practice for these calculations is to apply them to lists that are as inclusive as possible (i.e., where you have manually set the threshold to .00001 or something low enough that it allows in all words in the lexicon that appear at least once).

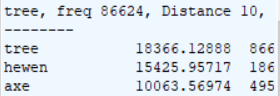
## Temporary files created by the shared lexis tool that can be safely deleted

To avoid repeatedly downloading large files if you are repeatedly making graphs that contain some of the same words, the shared lexis tool caches files it downloads from the internet on your system. However, these take up space on your computer (and in the Dropbox, if you are running the tool from there), so the tool provides a “Clear” button that allows you to delete these files from your system to free up space. This can be found in the window that appears when you select “Options…” from under the “View” menu.

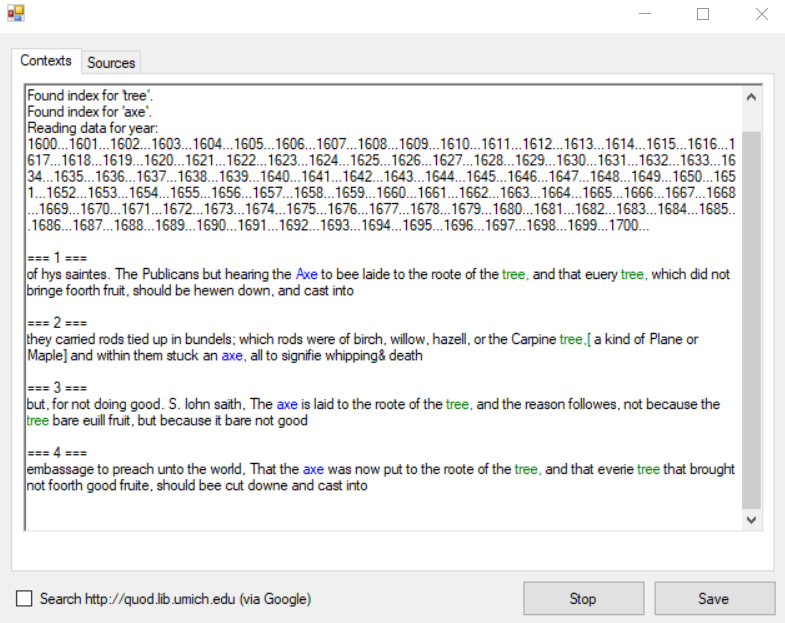
## Contextual search / “text grab” functions (available by request)

***The functions described in the remainder of this document are available only in the version of the tool available to Cambridge University users, who have agreed to the conditions specified in the Cambridge University Library’s TEXT/DATA MINING ADDENDUM that grants permission to use ECCO metadata for text and data mining purposes, and who have been given explicit permission from Professor Peter de Bolla. They are not available in the open version of the tool. To obtain access, please send an e-mail to Professor Peter de Bolla to request permission to use the “text grab” version of the tool.   
  
Once he has given permission, Gabriel () can give you access.***

Holding down Ctrl while clicking on a bound token in a list allows one to search the corpus for focal/bound token associations at the particular distance, corpus, and date range that was used to generate that list (these should appear at the top of the list). For example, Ctrl+clicking “axe” in the following list…



…on EEBO results in the following output:

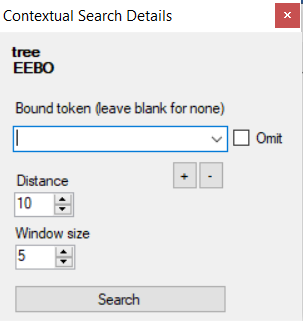


Clicking the Sources tab, if available, displays metadata for the source documents. Clicking “Save” allows the contents of the Contexts and Sources windows alike to be saved into a single file (Contexts at the top, Sources at the bottom).

**Important note:** These searches are governed by the contextual search options set in the View -> Options menu. If the “Window size for contextual search” option there is set to anything other than 5, the number of matches in the contextual search window will not match the number of coassociations used to calculate DPF values.

Holding down Ctrl while clicking on a list’s focal token – that is, while clicking on the very first header line of the list, above the hyphens that separate it from the rest of the list – allows one to search the corpus for all instances of the focal token at the particular distance, corpus, and date range that was used to generate that list.

Holding down Ctrl+Shift while clicking on list’s focal token – that is, while clicking on the very first header line of the list, above the hyphens that separate it from the rest of the list – brings up an advanced search interface that allows one to search for coassociations of the focal token and one or more bound tokens. Additional bound tokens can be added with the “+” button. Checking the “Omit” checkbox next to a bound token specifies that cases in which that token appears within the given distance and window size should be omitted from the results. Numbers specified in the “distance” and “window size” boxes override the default distance and window size settings.



1. The standard procedures for deciding whether two tokens x and y ‘co-occur’ are testing (a) whether y appears in a “window” of text that extends some specific number of words to the left and/or right of x, or (b) whether x and y appear in the same document. The most common way to implement (a) is for the window of text to begin at x and extend outward in both directions. As such, a classic “window” of 100 words will include all words that appear just 1, 2, 3… words away, which is not desirable if we wish to exclude words that primarily only appear with x in adjectival phrases, idiomatic expressions, and other relations that have more to do with syntax than conceptual relatedness. Because we are interested in long-range relationships and the ways in which associations vary with distance, we use the term ‘co-association’ rather than ‘co-occurrence,’ as the latter frequently implies that the two tokens appear in the same phrase or collocation. [↑](#footnote-ref-1)
2. This leaves a “hole” of in the middle (co-occurrences at distances of 1,2,3,4,5,6 or 7 were not counted), hence the “donut” nomenclature. [↑](#footnote-ref-2)
3. Recchia, G. & Nulty, P. (2017). Improving a fundamental measure of lexical association. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.),*Proceedings of the 39th Annual Conference of the Cognitive Science Society*(pp. 2963-2968)*.*Austin, TX: Cognitive Science Society. [↑](#footnote-ref-3)
4. That is, a power function fit to the rank-by-association plot with rank on the y-axis and the association measure on the x-axis. These plots can be viewed by clicking the “Plot all graphs” button and selecting any of the “Decline” plots. [↑](#footnote-ref-4)
5. Note that when using the default threshold of .00001, all items that occur with the focal token at least once will exceed the threshold; this is a way to display the focal token’s dpf with all other words. Words that do not co-occur with the focal token at all will not be included on the list. [↑](#footnote-ref-5)