# GraphColl guide

**GraphColl is** a free multi-platform tool for the analysis of collocation networks. It has been developed at the ESRC Centre for Corpus approaches to Social Science, Lancaster University. GraphColl works on Windows, Linux and Mac. An important note: Please make sure that you have an updated version of Java on your computer before you run GraphColl.

**How to cite GraphColl?**

Brezina, V., McEnery, T. & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, *20*(2).

**To start using GraphColl:**

1. Download GraphColl zip file.
2. Extract the contents of the zip into a directory of your choice.
3. Run the GraphColl.jar file using Java.

The following is a practical guide describing the basic features of GraphColl. For demonstration, it uses the LOB corpus available from the following link: http://corpora.lancs.ac.uk/stats/data/LOB_genres.zip

**Task 1.** Upload and explore corpora.

a) Upload the LOB corpus to GraphColl.

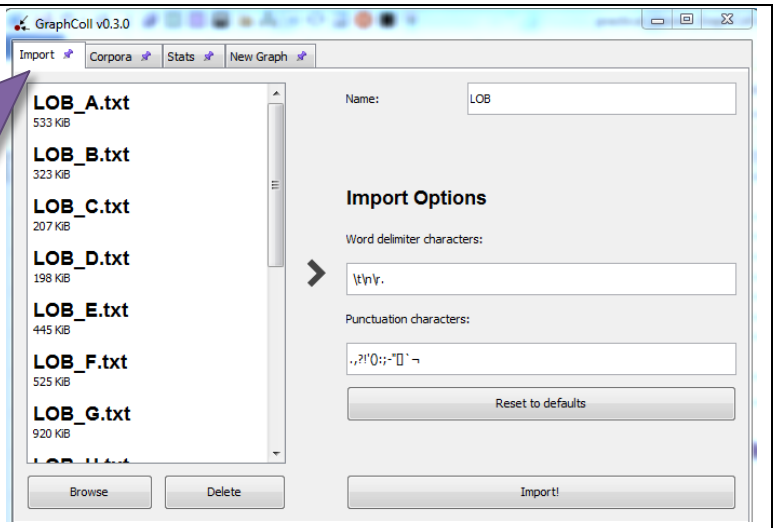
b) Upload the Academic writing (file J), newspapers (A, B and C) and fiction (K, L, M, N, P and R) subcorpora to GraphColl.

c) Check and note down the token counts for each of these under the 'Corpora' tab.

| | Tokens |
|---|---|
| **Academic writing** | |
| **Newspapers** | |
| **Fiction** | |
| **[...]** | |
| **LOB (whole corpus)** | |

***GraphColl* tip:**

1. Start GraphColl and go to the 'Import' tab.
2. Click on the 'Browse' button to select files.
3. Type in the corpus name and click on the 'Import!' button.




**Task 2.** Create graphs. Work with the whole LOB corpus.

a) Build the first-order collocation network around the word *time* using MI score and the default settings.

```
return (
    s   >= 3    &&  // MI >= 3
    o11 >= 5        // Over 5 within-window
);
```

b) How many collocates does the graph display? Are all of them useful?

c) Go back to 'New Graph', select MI as the statistic, change the default settings as indicated in the figure below (MI = 5 and above) and search for the node *time* again.

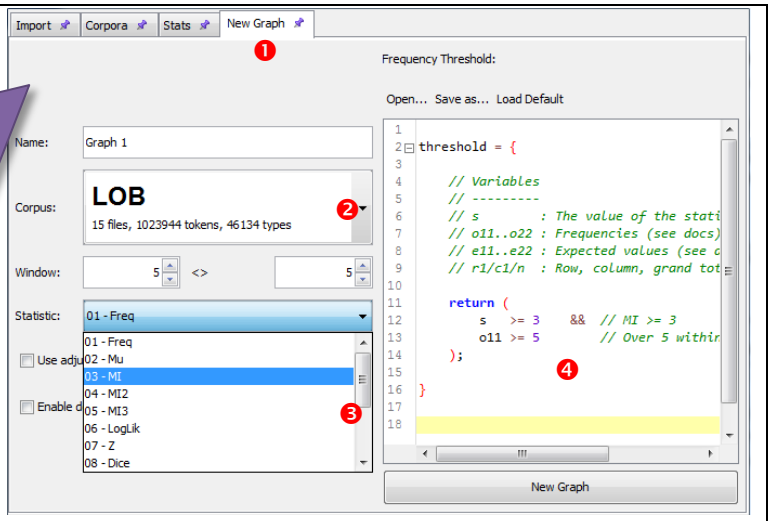
```
return (
    s   >= 5    &&  // MI >= 3
    o11 >= 5        // Over 5 within-window
);
```

How many results did you get this time?

***GraphColl* tip:**

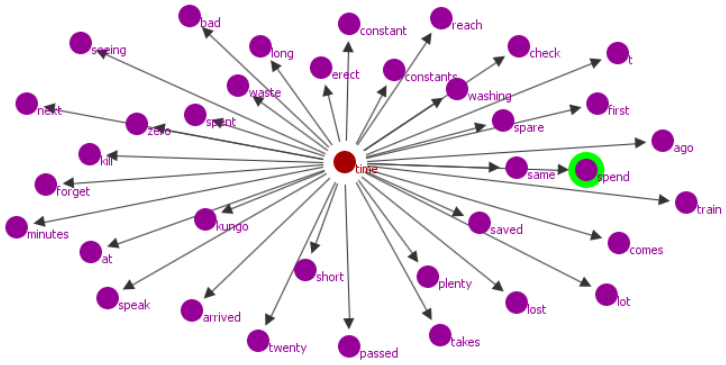
1. Go to the 'New Graph' tab.
2. From the drop-down menu select the required corpus.
3. From the drop-down menu select the required statistic.
4. Review and/or change the threshod settings.

s .... minimum value of the statistics
o11 .... minimum collocation frequency

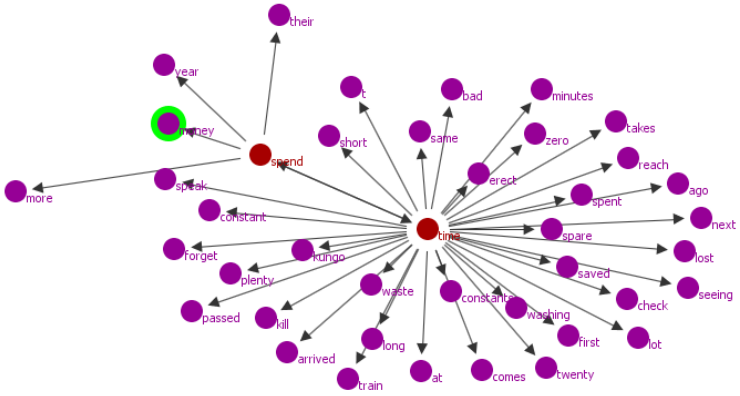



**Task 3.** Build collocation networks and explore graphs.

a) Go to the graph you have created in Task 2 c). It should be similar to the graph displayed in the figure below:




b) Find the collocate *spend* in the graph and double click on it. You should get a collocation network similar to the one displayed below:



c) Find the second-order collocate *money* in the graph and double click on it. Comment on the connection between *time* and *money* that you can see in the resulting graph that shows collocates around the node *money*.

***GraphColl* tip:**

Double clicking on a collocate creates a new branch of the collocation network treating the original collocate as a new node.


**Task 4.** Interpret graphs.

To help interpret graphs, GraphColl offers a concordance function which displays nodes in context (KWIC).

a) In the graph built in Task 3, create concordances for *time*, *money* and *spend* to help you interpret the collocation network.

b) One of the strong collocates of *time* is *kungo.* What are the contexts in which this collocate occurs? Does this collocate occur across different contexts?

c) Interpret the following collocates of *time: erect*, *washing*, *zero* and *t*

**Task 5.** Compare graphs.

Compare the use of the word *time* in the academic writing, fiction and news subcorpora of the LOB corpus. Use the MI score with the following threshold settings:

```
return (
    s    >= 5     &&  // MI >= 3
    o11  >= 5         // Over 5 within-window
);
```

**Task 6.** Explore statistics.

Use different statistical measures to build collocational networks of *time* and other words of interest. Note how the collocation networks differ with different settings:

a) Statistics

b) Collocation window

c) Thresholds

**Reporting statistics**

For the sake of replicability of results, all major parameters that can affect collocate identification should be reported. For this purpose, Brezina et al. (2015) introduce Collocation parameters notation (CPN) that captures all important parameters for collocate identification. CPN has the following format:

| Statistic ID | Statistic name | Statistic cut-off value | L and R span | Minimum collocate freq. (C) | Minimum collocation freq. (NC) | Filter |
|---|---|---|---|---|---|---|
| 4b | MI2 | 3 | L5-R5 | 5 | 1 | function words removed |
| 4b-MI2(3), L5-R5, C5-NC1; function words removed | | | | | | |

Example