

## 6 GraphColl

The GraphColl tool identifies collocations and displays them in a table and as a collocation graph or network.

It can be used, for example, to:

- Find the collocates of a word or phrase.
- Find colligations (co-occurrence of grammatical categories).
- Visualise collocations and colligations.
- Identify shared collocates of words or phrases.
- Summarise discourse in terms of its 'aboutness'.

### 6.1 Visual summary of GraphColl tab

The screenshot shows the GraphColl interface with the search term 'spent'. On the left, a table displays collocates with their status, position, and frequency. On the right, a network graph visualizes the relationships between 'spent' and its collocates. Callout boxes provide instructions: 'Save results.' (top right), 'Change collocation settings.' (top left), 'Display collocates in a table.' (middle left), 'Display collocation graphs and networks.' (middle right), and 'Pull up the bottom panel.' (bottom right).

Status	Position	Colloc	
●	R	time	
○	R	their	
○	R	on	5.07229d
○	M	much	6.15385f
○	L	money	8.82401h

### 6.2 Producing a collocation graph

GraphColl produces collocations graphs on the fly. After selecting the appropriate settings you can start searching for the node and its collocates.

1. Select the appropriate settings for the collocation search:
  - i) Span: how many words to the left (L) and to the right (R) of the node (search term) are being considered when searching for collocates [default: 5L, 5R].

- ii) Statistics: the association measure used to compute the strength of collocation [default: frequency – no association measure is preferred because the choice depends on the research question].
  - iii) Threshold: The minimum frequency and statistics cut-off values for an item (word, lemma, POS) to be considered a collocate.
  - iv) Corpus: The corpus that is being searched.
  - v) Unit: The unit (type, lemma, part of speech [POS] tag) used for collocates.
2. Type the search term into the search box (top left) and left-click 'Search'.
  3. This will produce a collocation table (left) and a collocation graph (right).

### 6.3 Reading collocation table

A collocation table is a traditional way of displaying collocates. In GraphColl, the table shows the following pieces of information for each collocate: i) status, ii) position, iii) stat, iv) collocation frequency and v) frequency of the collocate anywhere in the corpus. By default, the table is sorted according to the selected collocation statistic (largest-smallest).

1. The following is a visual description of the collocation table.

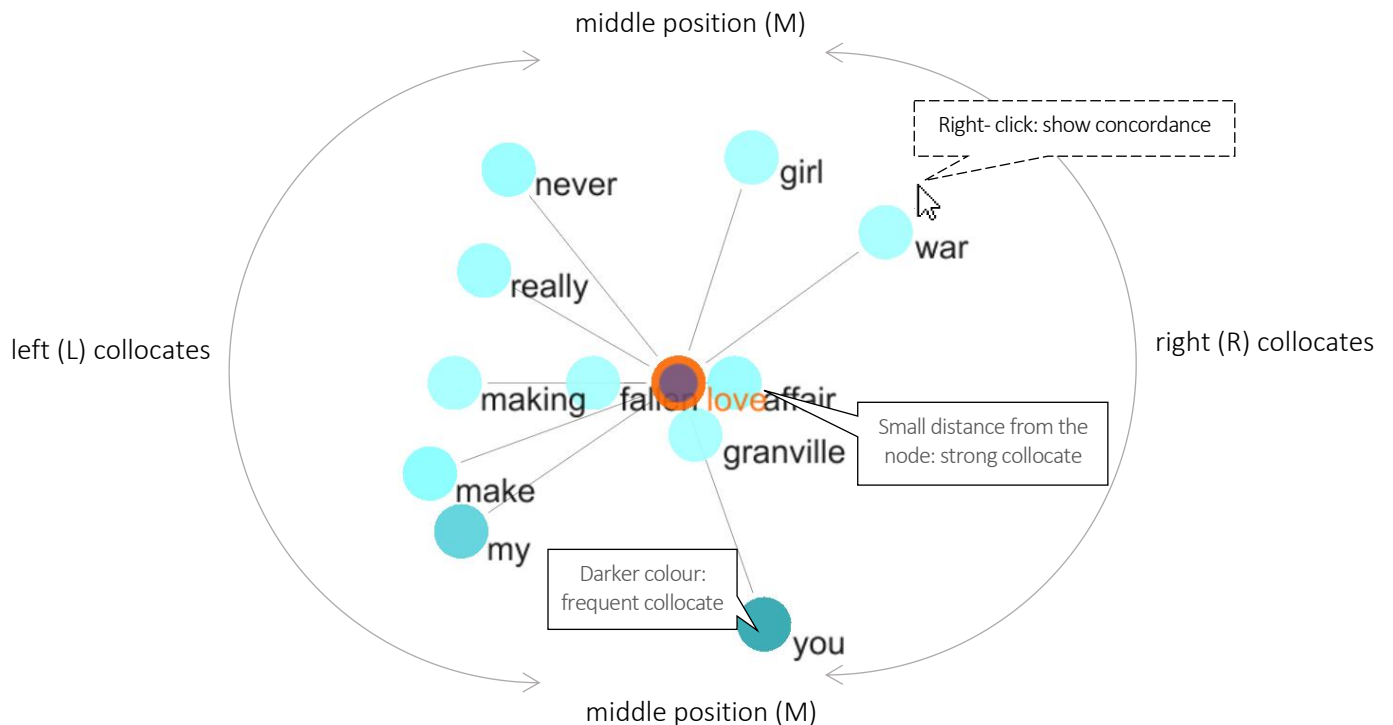
Status	Position	Collocate	▼ Stat	Freq (coll.)	Freq (corpus)
○	R	granville	9.208810483...	5	28
○	R	affair	9.179667251...	7	40
○	L	fallen	8.623849712...	5	42
●		love	6.768239667...	10	304
○	L	making	5.999358660...	5	259
○	L	really	5.989182871...	6	313
○	R	girl	5.754072064...	5	307
○	R	you	5.392047146...	46	3630
○	R	war	5.386810358...	5	396
○	L	my	5.256041267...	21	1821
○	L	make	5.223919105...	9	798
○	L	never	5.075197...	7	688

2. The meaning of the individual columns is:
  - i) Status: shows whether the collocate has been expanded; ○ indicates a non-expanded collocate, while ● indicates expanded collocate (node) in a collocation network.
  - ii) Position: shows textual position of the collocate, which can be either left (L) of the node, right (R) of the node or middle (M), i.e. with equal frequency L and R.
  - iii) Collocate: shows the collocate in question.
  - iv) Stat: displays the value of the selected association measure.
  - v) Freq (coll): displays the frequency of the collocation (combination of node + collocate).
  - vi) Freq (corpus): displays the frequency of the collocate anywhere in the corpus.

## 6.4 Reading collocation graph

The graph displays three dimensions: i) strength of collocation, ii) collocation frequency and iii) position of collocates. To find out more about a collocate, right-click on it to obtain concordance lines (KWIC), in which the collocates co-occurs with the node.

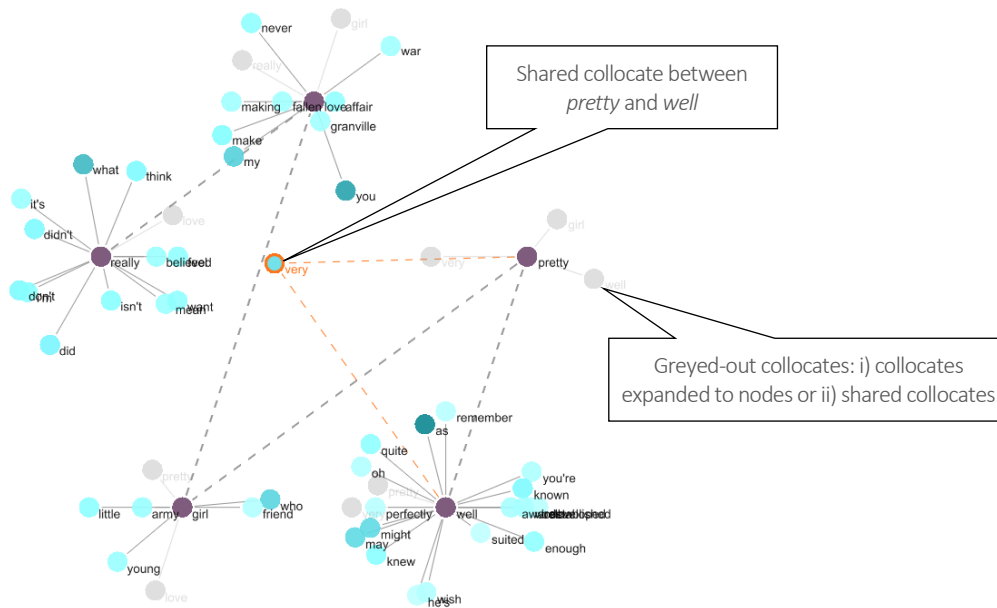
1. **Strength:** The strength of collocation as measured by the association measure is indicated by the distance (length of line) between the node and the collocates. The closer the collocate is to the node, the stronger the association between the node and the collocate ('magnet effect').
2. **Frequency:** Collocation frequency is indicated by the intensity of the colour of the collocate. The darker the shade of colour, the more frequent the collocation is.
3. **Position:** The position of collocates around the node in the graph reflects the exact position of the collocates in text: some collocates appear (predominantly) to the left of the node, others to the right; others still appear sometimes left and sometimes right (middle position in the graph). For the ease of display (if multiple collocates appear in a similar position and hence overlap), the tool allows 'spreading out' collocates evenly around the node. This is done by clicking on the 'Spread out' button (top right). When this is done, the collocates are dispersed evenly around the node with a 'L' or 'R' index displayed above the collocate circle indicating their original position to the left and to the right respectively.



## 6.5 Extending graph to a collocation network

A collocation network is an extended collocation graph that shows i) shared collocates and ii) cross-associations between several nodes.

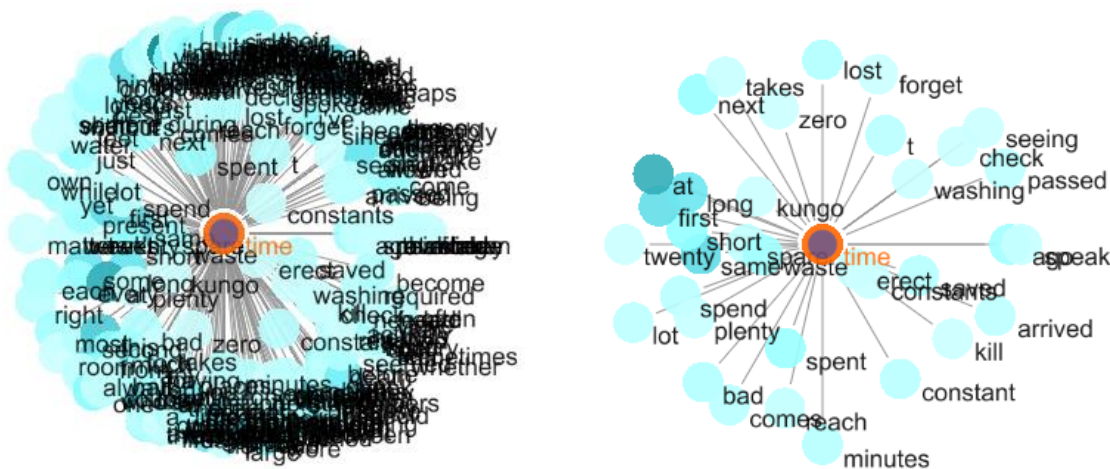
1. To expand a simple collocation graph (see above) into a collocation network, either search for more nodes or left-double-click on a collocate in either the table or the graph.
2. A collocation network displays nodes with unique collocates (outer rim of the graph) and shared collocates (middle of the graph). The links between nodes and shared collocates are indicated by a dash-dot line (— · — · —).



## 6.6 Problems with graphs: overpopulated graphs

If a collocation graph or network includes too many nodes and collocates, it becomes hard to interpret. We call this type of graph/network an overpopulated graph/network. The solution is either to change the graph's settings making the threshold values more restrictive (see Section 6.2) or filtering some of the results based on a clearly specified criterion (e.g. function words).

The following figure shows an overpopulated graph on the left and a graph that is more easily interpretable on the right. Note the difference in settings recorded in CPN (see Section 6.7)



Collocation graph of 'time' 3a-MI(3), L5-R5, C5-NC5      Collocation graph of 'time' 3a-MI(5), L5-R5, C5-NC5

## 6.7 Reporting collocates: CPN

It is important to realise that there is no one definite sets of collocates: different statistical procedures and threshold values highlight different sets of collocates. We therefore need to report the statistical choices involved in the identification of collocations using standard notation called Collocation Parameters Notation (CPN). When saving the results, GraphColl saves the settings in the form of CPN.

Brezina et al. (2015) propose CPN as a specific notation to be used for accurate description of collocation procedure and replication of the results. The following parameters are reported.

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						

### ► Did you know?

The name GraphColl is an acronym for *graphical collocations* tool. GraphColl was the first module in #LancsBox (v.1.0) with the other tools being added at a later stage. Graphical display of collocations and collocation networks is inspired by the work of Phillips (1985), who demonstrated the concept of lexical networks (Phillip's term for 'collocation networks') with small specialised corpora. GraphColl takes this notion further, offering different statistical choices and producing collocation networks on the fly with both small and large corpora.

Phillips, M. (1985). *Aspects of text structure: An investigation of the lexical organisation of text*. Amsterdam: North-Holland.