

7 Words tool

The Words tool allows in-depth analysis of frequencies of types, lemmas and POS categories as well as comparison of corpora using the keywords technique.

It can be used, for example, to:

- Compute frequency and dispersion measures for types, lemmas and POS tags.
- Visualize frequency and dispersion in corpora.
- Compare corpora using the keyword technique.
- Visualize keywords.

7.1 Visual summary

The screenshot shows the #LancsBox v 3.0 interface. The top menu bar includes KWIC, Whelk, GraphColl, Words, and Text. The main window is divided into two panels.

Left Panel: A table with columns for Corpus, LOB, Frequency, Dispersion, and Type. The table is filtered to show the top 10 words. A callout box points to the table header with the text: "Right-click on the table header to activate filter." Below the table, another callout box points to the table content with the text: "Right-click inside the table to activate a Whelk pop-up."

Type	Frequency: 01 - Freq	Dispersion: 01 - CV
the	69970.000000	0.162650
of	36408.000000	0.291942
and	27503.000000	0.078241
to	22750.000000	0.081485
a	21197.000000	0.114962
in	11196.000000	0.208365
that	11003.000000	0.175456
is	10516.000000	0.565229
was	10034.000000	0.461677
it	9820.000000	0.225496

Right Panel: A network graph visualization showing relationships between corpora. The graph has a central node labeled "Brown" and several other nodes representing different corpora. A callout box points to the "Brown" node with the text: "Left-double-click on the corpus to see its internal structure." Another callout box points to the graph with the text: "Drag corpora together to produce keywords."

Left: Creating frequency lists, computing dispersion and keywords.

Right: Visualizing frequencies, dispersions and keywords.

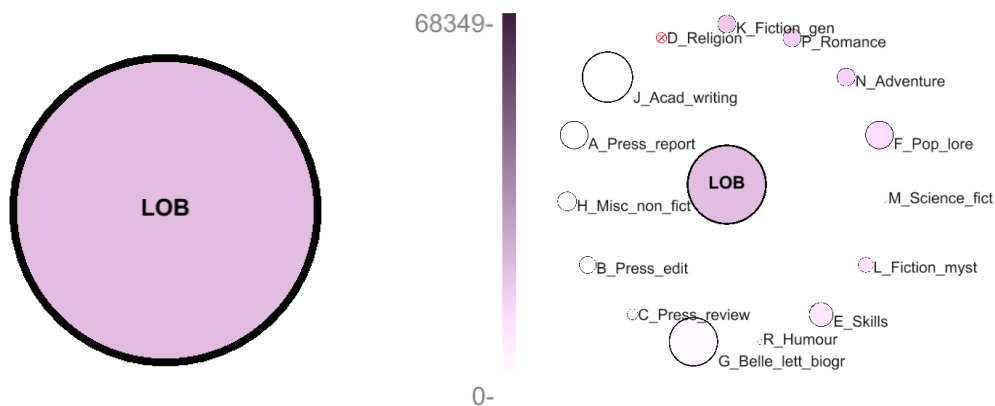
7.2 Producing frequency list

On start, Words produces a frequency list (table) based on the default corpus (see Section 2.5, point 2) and default settings. These settings can be changed and a different frequency list is produced.

1. The following are the settings for frequency lists:
 - i) Corpus: The corpus that is being used.
 - ii) Frequency: Absolute or relative frequency [default: absolute frequency].
 - iii) Dispersion: The dispersion statistic [default: coefficient of variation (CV)].
 - iv) Unit: The unit used in the frequency list (type, lemma or part of speech tag).
2. Changing any of these settings triggers re-computing of the frequency list.
3. Frequency lists can be searched using the search box (top left).
4. Frequency lists can be sorted by left-clicking on the header.
5. Frequency lists can be filtered by right-clicking on the header and applying a filter.
6. Two different frequency lists can be computed in the split-screen view, which is triggered by left-clicking on a bar with three triangles: ▲▲▲. This brings up the bottom panel.

7.3 Visualizing frequency and dispersion

The Words module displays corpora and corpus files (when a corpus is left-double-clicked). It visualises frequency and dispersion of words using intensity of colour and position of individual files displayed as circles; the size of the circle indicates the relative size of the corpus/file.




Display of frequency in the whole corpus on the scale of 0 - 68,349 (most frequent item). Display of frequency per file (when corpus is left-double-clicked).

1. To visualize frequency of an item in the table, left-click on the item in the frequency table. The shade of the colour of the corpus will change according to the frequency value of this item. The scale on the right offers a reference point for interpretation.
2. To visualize dispersion of an item in the table, left-double-click on the corpus (large circle). The corpus will expand to display individual files (small circles) of which the corpus consists. The size of each circles is proportional to the size of the corpus subpart. The shade of the colour of the

small circles will change according to the frequency value of the item in the frequency list. Crossed-out (⊗) circles indicate that the item does not occur in the given corpus file. In addition, the corpus files are ordered according to the relative frequency of the item with the file with the largest relative frequency of the item appearing at the 12-o'clock position () and the other files ordered clockwise according to decreasing relative frequency of the item ().

7.4 Producing keywords

The Words module computes a comparison of frequencies between two corpora/wordlists using a selected statistical measure. It identifies and visualizes positive keywords, negative keywords and lockwords.

1. Left-click on ▲▲▲ to bring up the bottom panel.
2. In the bottom panel, select a comparison (reference) corpus, while in the top panel keep your corpus of interest.
3. In the visualisation panel (right), drag the circles that represent the two corpora together .
4. The resulting table will display frequency and dispersion info about the two corpora as well as the keyword statistic; the graphics will identify top 10 positive keywords, top 10 negative keywords and top 10 lockwords.
5. In the settings, you can change the i) keyword statistic and ii) threshold.
Keyword statistic: This is a measure that compares two frequency lists [default: simple maths with constant $k = 100$].
Threshold: Threshold values for the identification of positive keywords, negative keywords and (by implication) lockwords.

► Did you know?

The statistical technique of keyword analysis was originally developed by Mike Scott (1997) and it was implemented in WordSmith Tools. It relied on corpus comparison using the chi-squared test or the log-likelihood test. As Kilgarriff pointed out, the chi-squared test and the log-likelihood test are not entirely appropriate for this type of comparison. Kilgarriff's solution implemented in Sketch Engine was to compare corpora using a 'simple maths' procedure, a simple ratio between relative frequencies of words in the two corpora we compare. In addition to 'simple maths', #LancsBox offers also other types of solutions for corpus comparison.

Scott, M. (1997). PC analysis of key words—and key key words. *System*, 25(2), 233-245.

Kilgarriff, A. (2009, July). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*. Liverpool, UK.