

## 5 Whelk tool

The Whelk tool provides information about how the search term is distributed across corpus files.

It can be used, for example, to:

- Find absolute and relative frequencies of the search term in corpus files.
- Filter the results according to different criteria.
- Sort files according to absolute and relative frequencies of the search term.

### 5.1 Visual summary of Whelk tab

The screenshot shows the Whelk tool interface. The top panel displays search results for the term 'love'. The bottom panel shows a table with the following data:

File	Tokens	Frequency	Relative frequency per 10k
P_Romance.txt	58197	75	12.887262
C_Press_review.txt	34289	39	11.37391
K_Fiction_gen.txt	58515	69	10.253783
L_Fiction_myst.txt	48259	15	3.1022284
F_Pop_lore.txt	88742	26	2.928415
N_Adventure.txt	58322	16	2.7433903
G_Belle_let_niogr.txt	155271	35	2.2543234
E_Skills.txt	76613	16	2.0884185
M_Science_ict.txt	12037	2	1.6615435
D_Religion.txt	34257	4	1.1675448
J_Acad_writing.txt	161289	10	0.6200051
A_Press_report.txt	88805	5	0.5630314
R_Humour.txt	18087	1	0.55288327
B_Press_end.txt	54397	0	0.0
H_Misc_non_ict.txt	60627	0	0.0

Top panel: Searching corpora

You can:

- Search, sort and filter.
- Use simple and advanced searching functionality.
- Use 'smart' searches.

Bottom panel: Displaying distribution

You can:

- View the distribution of the search term in individual files.
- Sort, filter and copy/paste.

### 5.2 Top panel: KWIC

The top panel in Whelk has the same powerful search, sort and filter functionalities as the KWIC tool (see Section 4). It is directly connected to the bottom panel: any update in the top panel is immediately reflected in the bottom panel.

### 5.3 Bottom panel: Frequency distribution

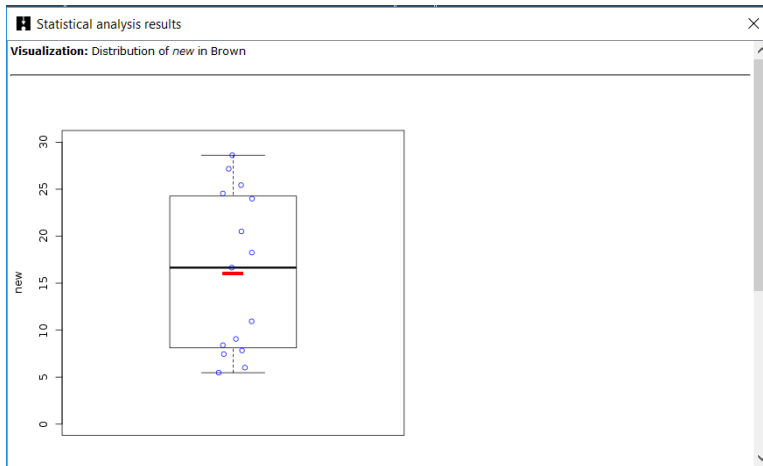
The bottom panel in Whelk provides detailed information about the distribution of the search term.

1. 'File' column lists the name of the individual files in the corpus.
2. 'Tokens' column provides the information about the size of each file in running words (tokens).
3. 'Frequency' column provides absolute frequencies of the search term i.e. refers to how many instances of the search term there are in each file.
4. 'Relative frequency per 10k' provides relative frequency normalised to the basis of 10,000 tokens; this value is comparable across files and corpora.

## 5.4 Statistical analysis

Whelk connects to Lancaster Stats Tools online to perform statistical analysis of the data.

When search results appear, these can be visualised using a boxplot by clicking on the statistical analysis button (📊). The tool automatically connects to Lancaster Stats Tools online (Brezina 2018) and displays the result:



### ► Did you know?

The Whelk tool (both the name and the functionality) is inspired by Kilgarriff's (1997: 138ff) notion of the 'whelks problem'. Imagine, says Kilgarriff, that you have a corpus which includes one text (a book) about whelks – small snail-like sea creatures (🐌). In this text, the word *whelks* will appear many times and hence will appear as a frequent word in the entire corpus, although its use is limited to one specific context. To overcome the problem and present more accurate information about word distribution, the Whelk tool shows the frequency distribution of search terms in individual corpus files.