# 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.

#LancsBox v 4.5 -				Top panel: Searching corpora
KWIC	GraphColl W	helk Words	Ngrams	Tet
orpora Whe X				
	Search			
	ces 304 (3.02) Texts 13/15	▼ Corpus L-O-B	▼ Context 7 ▼ Display Text	You can:
A Press rep	LoR and Juliof was the	Node	Right	
A_Press_rep	and Julief' was the irresponsibility of young love pushed into tragedy by Shakespeare." Othello" is a cuttivated, brave man who comes to love too late, and does not know what			Search, sort and filter.
A Press rep-	not to know what to do with love." Zeffirelii does not mention the colour of			Search, Sort and Inter.
A_Press_rep	Logue writes fierce, noisy poems about war, love, and Logue. Son of a Southampton civil			• Use simple and advanced searching functionality
A_Press_rep		up in flames one day. In love, he wrote:-		Use simple and advanced searching functionality.
C_Press_revi		him, the scandal of his love affair with		
C_Press_revi C_Press_revi		to a point that looks like love, had fanne st and most reasonable love, who does		Use 'smart' searches.
C_Press_revi		tudies of the renewal of love that the sc		
0 C_Press_revi		I sympathy—" you need someone to love you while you are looking for someone		
1 C_Press_revi	while you are	looking for someone to love". Miss Dora	Bryan plays the mother as	
2 C_Press_revi		ood picture," Let's Make Love," he was sw		
3 C_Press_revi 4 C_Press_revi		He wrote "My September Love," the big Da		
4 C_Press_revi 5 C_Press_revi		ult: Mr. Hudson and lady love Lollo find t himself that makes him love them. He r		
5 C_Press_real	1, 15 111		hay be able to express	
		***		
File Romance.txt	Tokens 58197	75	Ty Relative frequency ; 12.887262	left to the second s
Press review.bd	34289	39	11.37391	
Fiction_gen.txt	58515	60	10.253781	
_Fiction_myst.txt	48259	15	3.1082284	<b>Bottom panel:</b> Displaying distribution
_Pop_lore.txt	88742	26	2.9298415	
L_Adventure.txt	58322 155271	16	2.7433903 2.2541234	
Belle_lett_biogr.txt Skills.txt	76613	35	2.2541234	
Science_fict.txt	12037	2	1.6615435	You can:
_Religion.txt	34257	4	1.1676446	, eu eu
_Acad_writing.txt	161289	10	0.6200051	
_Press_report.bt	88805	5	0.5630314	View the distribution of the search term in
Lengen Leng Lengen Lengen	18087 54367	1	0.55288327	
i_Press_edit.txt i Misc non fict.txt	60627	0	0.0	individual files.
	In a second	14	la:a	
				<ul> <li>Sort, filter and copy/paste.</li> </ul>
		<b>X  ?  </b> ‡		

### 5.1 Visual summary of Whelk tab

#### 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

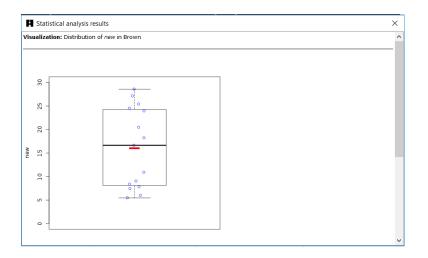
The bottom panel in Whelk provides detailed information about the distribution of the search term.

- 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 (<sup>‡</sup>). 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.