

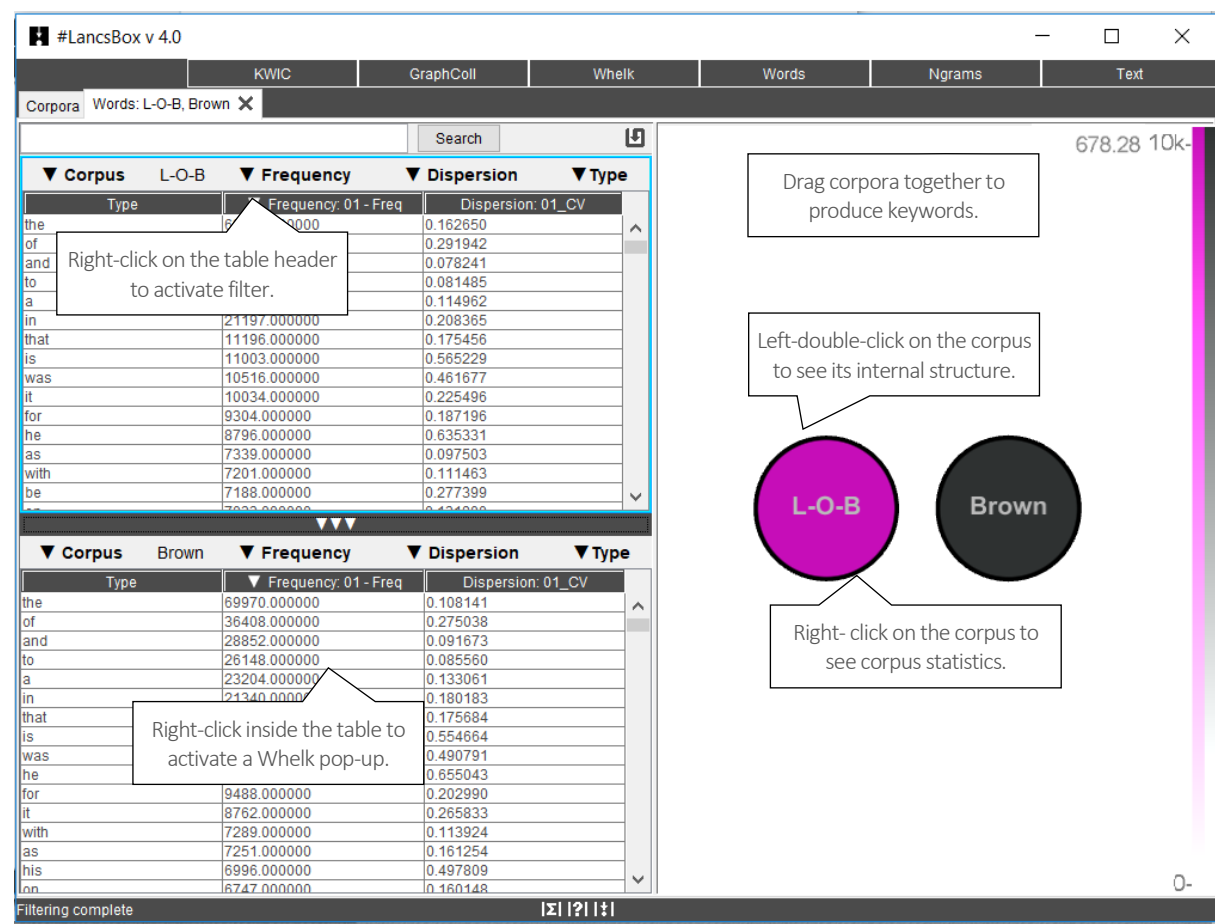
7 Words tool

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



Left: Creating frequency lists, computing dispersion and keywords.

Right: Visualizing frequencies, dispersions and keywords.

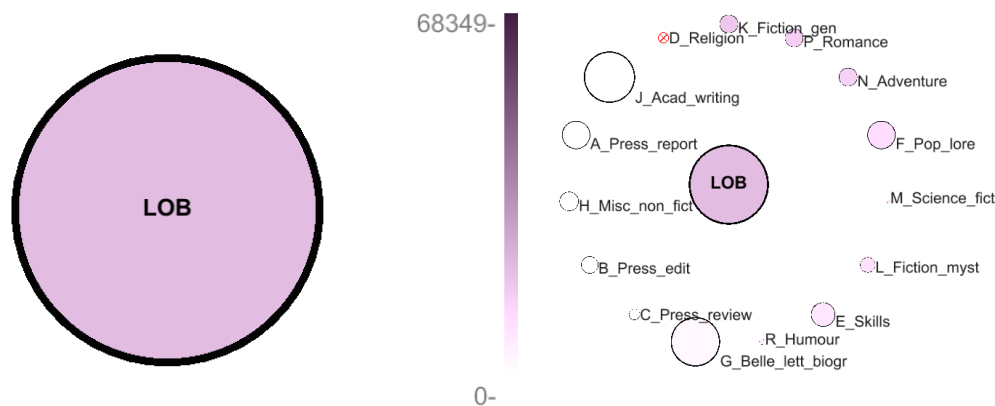
7.2 Producing frequency list

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 . Alternatively, press the space bar.
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.

7.5 Producing corpus statistics

The Words module computes essential corpus statistics: i) Complexity stats and ii) Lexical stats

1. Right-click on corpus .
2. In the pop-up table toggle between Complexity stats and Lexical stats.

Mean sentence length and Standard deviation (SD)

L-O-B

▼ Complexity Stats		▼ Lexical Stats		
File	Sentence Length (mean)	Sentence Length (SD)	Word Length (mean)	Word Length (SD)
A_Press_report.bt	19.159855	11.671002	4.745014	2.592452
B_Press_edit.bt	20.061825	12.509228	4.734839	2.6490588
C_Press_review.bt	22.179173	14.621512	4.77955	2.7150402
D_Religion.bt	19.105968	13.838464	4.5256734	2.5267594
E_Skills.bt	20.938234	13.569921	4.603331	2.51522
F_Pop_lore.bt	21.013971	12.89571	4.6807714	2.5748186
G_Belle_left_biogr.bt	24.429043	15.205565	4.714493	2.6827366
H_Misc_non_fict.bt	25.527159	20.760244	4.882379	2.7997973
J_Acad_writing.bt	26.358719	16.505852	4.851614	2.8534663
K_Fiction_gen.bt	14.338397	12.206561	4.3068104	2.27138
L_Fiction_myst.bt	12.934602	9.881333	4.30815	2.259926
M_Science_fict.bt	12.371017	11.275793	4.5213094	2.4187284
N_Adventure.bt	11.963488	9.186616	4.262817	2.1768787
P_Romance.bt	12.555987	9.679649	4.236387	2.1641104
R_Humour.bt	17.87253	14.513976	4.5027366	2.506564

Type-token ratio (TTR), Standardised type-token ratio (STTR), Moving average type-token ratio (MATTR)

L-O-B

▼ Complexity Stats		▼ Lexical Stats			
File	Types	Tokens	TTR	STTR	MATTR
A_Press_report.bt	12079	88805	0.13601711	0.7342071	0.7342669
B_Press_edit.bt	7909	54367	0.14547427	0.73095614	0.7306529
C_Press_review.bt	7703	34289	0.22464931	0.74618065	0.74707484
D_Religion.bt	5399	34257	0.15760283	0.69137025	0.6896752
E_Skills.bt	10808	76613	0.14107266	0.72006595	0.7209448
F_Pop_lore.bt	12274	88742	0.13831106	0.72313124	0.72300106
G_Belle_left_biogr.bt	17485	155271	0.112609565	0.7196904	0.7203814
H_Misc_non_fict.bt	6717	60627	0.11079222	0.6818785	0.6824127
J_Acad_writing.bt	15743	161289	0.097607404	0.685145	0.685025
K_Fiction_gen.bt	7841	58515	0.13399982	0.7243858	0.72329557
L_Fiction_myst.bt	6632	48259	0.13742514	0.7332717	0.7323574
M_Science_fict.bt	3187	12037	0.26476696	0.7563636	0.7587221
N_Adventure.bt	7638	58322	0.13096258	0.73029095	0.7307091
P_Romance.bt	6525	58197	0.11211918	0.7355844	0.73544407
R_Humour.bt	4452	18087	0.24614364	0.7351933	0.73470604

► 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.