Open research group

ESRC Centre for Corpus Approaches to Social Science

Lancaster University

https://cass.lancs.ac.uk/open-advanced-methods-research-group/

Finish update

CASS

ESRC Centre for Corpus Approaches to Social Science

ABOUT US BLOGS CASS PROJECTS CASS: BRIEFINGS

Open Advanced Methods Research Group

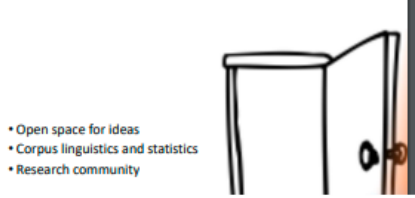
ed-methods-research-group/

Finish update

– 12:00pm–12:50pm (UK time): R Scripts and

n research group

for Corpus Approaches to Social Science
Lancaster University



- Open space for ideas
- Corpus linguistics and statistics
- Research community

Meeting1 Download

Topics

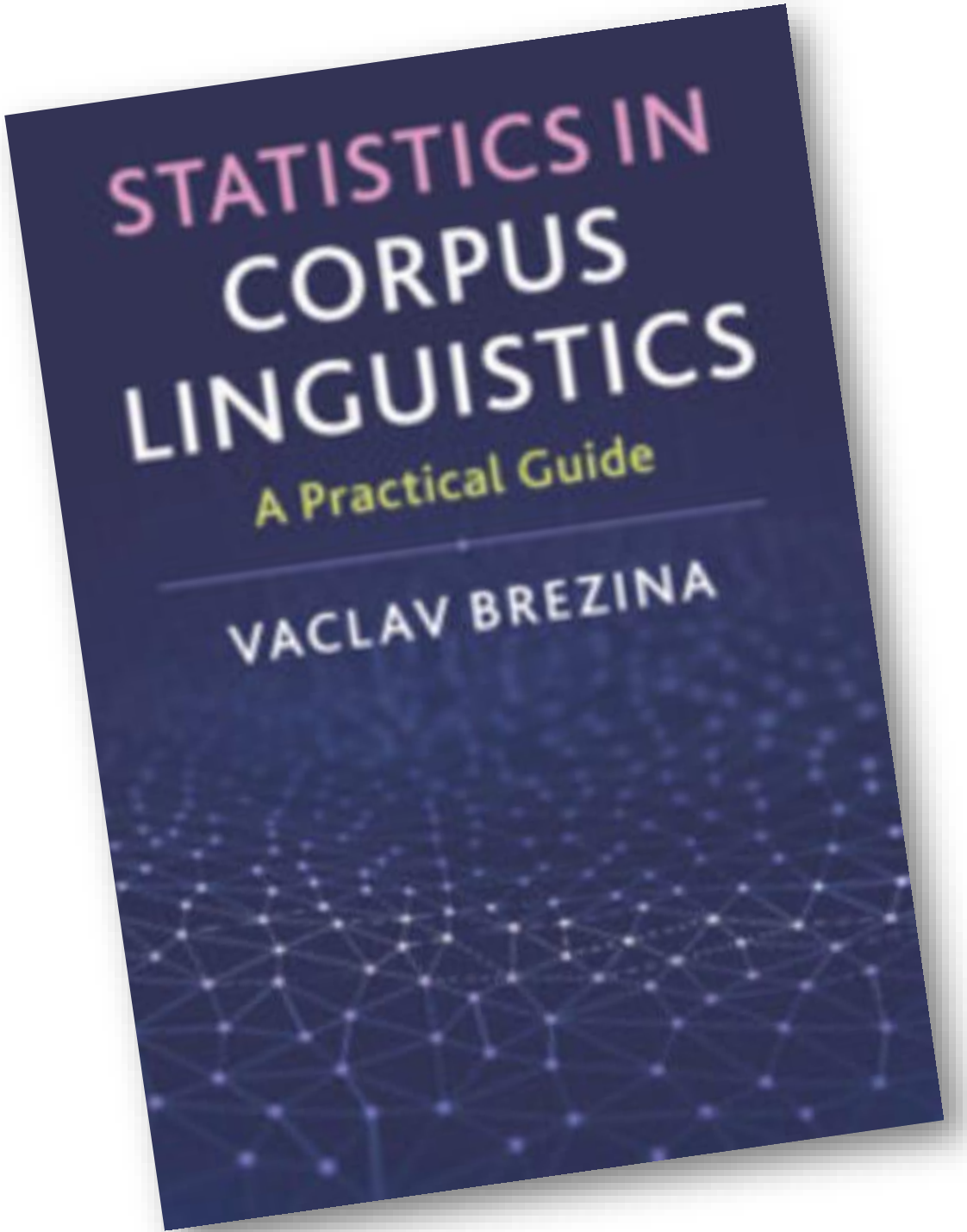
Wednesday 16 October 12.00pm - 12.50pm UK time, **Statistics and language analysis - #LancsBox KWIC**

Wednesday 30 October 12.00pm - 12.50pm UK time, **Collocations - #LancsBox GraphColl**

Wednesday 13 November 12.00pm - 12.50pm UK time, **Group comparison – Text tool**

Wednesday 27 November 12.00pm - 12.50pm UK time, **Wordlists and keywords - Words**

Wednesday 11 December 12.00pm - 12.50pm UK time, **R scripts and #LancsBox Wizard**



Brezina (2018)

“

Identifying keywords is one of the crucial techniques in corpus linguistics (Scott 1997), yet it is also a procedure that is often misunderstood. Keywords are words that are considerably more frequent in one corpus than in another corpus; we can therefore say that keywords are words that are typical of the corpus of interest when compared with another corpus. However, it is important to remember that ‘keywords’ is a relative term depending on the differences in lexical frequencies in the two corpora in question. Keywords are important when identifying key concepts in discourses, typical vocabulary in a genre/language variety, lexical development over time, etc.

Think about and discuss

1. What is the difference between keywords and collocations?
2. Are keywords *always* the most important words in a text?
3. What are the keywords in this short paragraph?

Storm Bert caused devastating flooding in the UK this week, taking lives and destroying homes and businesses in what has become a frequent occurrence during autumns and winters.

Climate breakdown is making these extreme weather events more probable. Extreme rainfall is more common and more intense because of human-caused global heating across most of the world, and particularly in Europe. This is because warmer air can hold more water vapour, and flooding has become more frequent and severe as a result. But floods are also hitting communities with more intensity because of inadequate, underfunded flood defences.

Storm Bert caused devastating flooding in the UK this week, taking lives and destroying homes and businesses in what has become a frequent occurrence during autumns and winters.

Climate breakdown is making these extreme weather events more probable. Extreme rainfall is more common and more intense because of human-caused global heating across most of the world, and particularly in Europe. This is because warmer air can hold more water vapour, and flooding has become more frequent and severe as a result. But floods are also hitting communities with more intensity because of inadequate, underfunded flood defences.

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Keywords

Reference corpus: BNC2014 newspapers

Terms: **284,427**

1	Focus rel. freq. (...	Reference rel. fre...	Simple maths	Log likelihood	% difference	Log ratio
frequent	20,833.33	11.21	188.23	26.09	185,741.56	10.86
flooding	20,833.33	17.11	178.75	24.41	121,658.26	10.25
extreme	20,833.33	36.24	153.65	21.42	57,392.37	9.17
autumns	10,416.67	0.15	105.01	20.03	7,061,879.17	16.11
human-caused	10,416.67	0.25	104.91	19.12	4,237,087.50	15.37
underfunded	10,416.67	0.74	104.40	17.05	1,412,295.83	13.79
vapour	10,416.67	1.03	104.09	16.39	1,008,754.17	13.30
bert	10,416.67	2.26	102.84	14.85	460,463.86	12.17
occurrence	10,416.67	2.41	102.69	14.72	432,266.07	12.08
rainfall	10,416.67	4.03	101.09	13.70	258,265.09	11.34

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

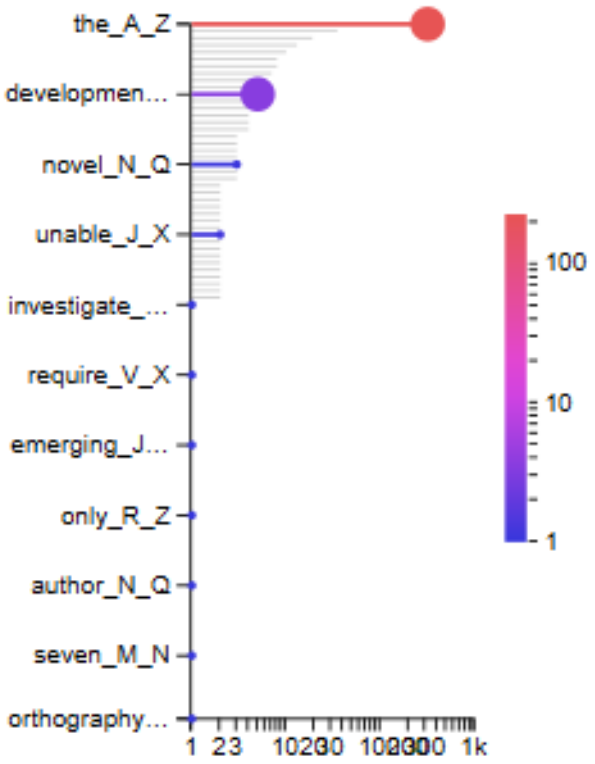
#LancsBox X 2.0.0

c*

Example cor... 1.0 CLAWS7 | whole corpus | ... | lexeme

Terms: **1,253**

Term	Freq...	Rel. ...	ARF ...	Range	Ran...	CV (...	Juill..D
right_N_S	3	68...	1.12	1	50....	1.00	0
three_M_N	3	68...	2.44	1	50....	1.00	0
range_N_N	3	68...	2.27	2	10...	0.12	0.88
now_R_T	3	68...	2.01	2	10...	0.12	0.88
creation_N_A	3	68...	2.05	1	50....	1.00	0
success_N_X	3	68...	2.25	2	10...	0.12	0.88



Closed keywords table.

WORDS

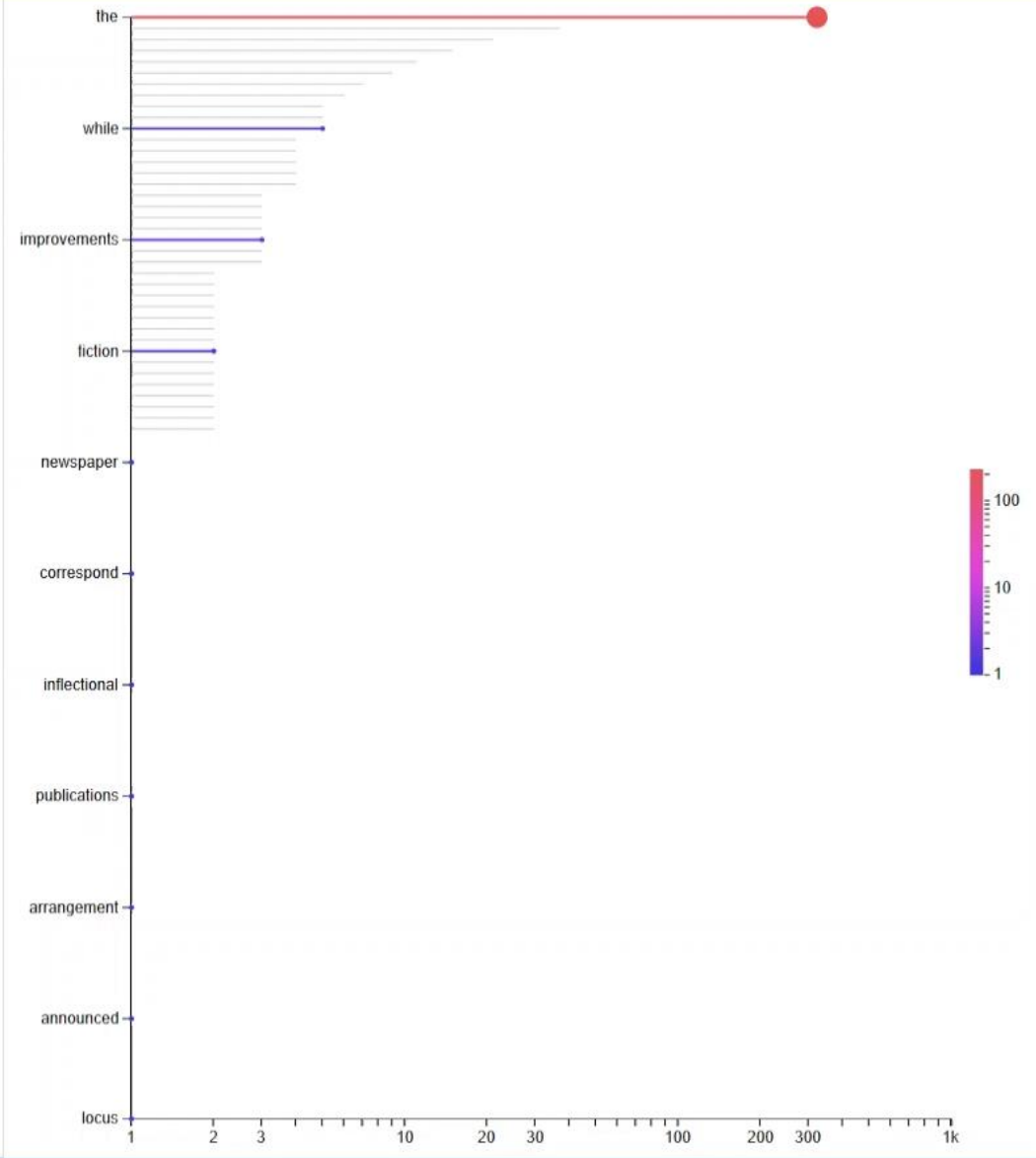
#LancsBox X 2.0.0

corpus

Example corpus 1.0 CLAWS7 | whole corpus | 4K | word (lowercase)

Terms: **1,265**

Term	Frequency	Rel. frequency	ARF (averag...	Range	Range %	CV (coeffici...	Juilland's D	DP (deviatio...
their	17	3,902.66	6.47	1	50.00	1.00	0	0.28
work	17	3,902.66	11.63	2	100.00	0.17	0.83	0.07
been	16	3,673.09	8.36	2	100.00	0.08	0.92	0.03
at	16	3,673.09	10.19	2	100.00	0.08	0.92	0.03
text	16	3,673.09	9.64	2	100.00	0.26	0.74	0.09
such	15	3,443.53	10.11	2	100.00	0.12	0.88	0.05
not	15	3,443.53	6.89	2	100.00	0.69	0.31	0.21
it	15	3,443.53	8.87	2	100.00	0.69	0.31	0.21
this	15	3,443.53	9.38	2	100.00	0.43	0.57	0.15
an	14	3,213.96	9.19	2	100.00	0.01	0.99	0.006
words	14	3,213.96	8.50	2	100.00	0.40	0.60	0.14
linguistics	14	3,213.96	7.15	2	100.00	0.81	0.19	0.51
can	14	3,213.96	7.75	2	100.00	0.18	0.82	0.07
have	13	2,984.39	7.59	2	100.00	0.23	0.77	0.10
also	13	2,984.39	8.65	2	100.00	0.13	0.87	0.05
one	13	2,984.39	7.22	2	100.00	0.36	0.64	0.13
tagging	13	2,984.39	3.67	2	100.00	0.36	0.64	0.13
may	12	2,754.82	7.25	2	100.00	0.32	0.68	0.11
has	11	2,525.25	6.34	2	100.00	0.19	0.81	0.08



Maximized tool.