Open research group

ESRC Centre for Corpus Approaches to Social Science

Lancaster University

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

Topics

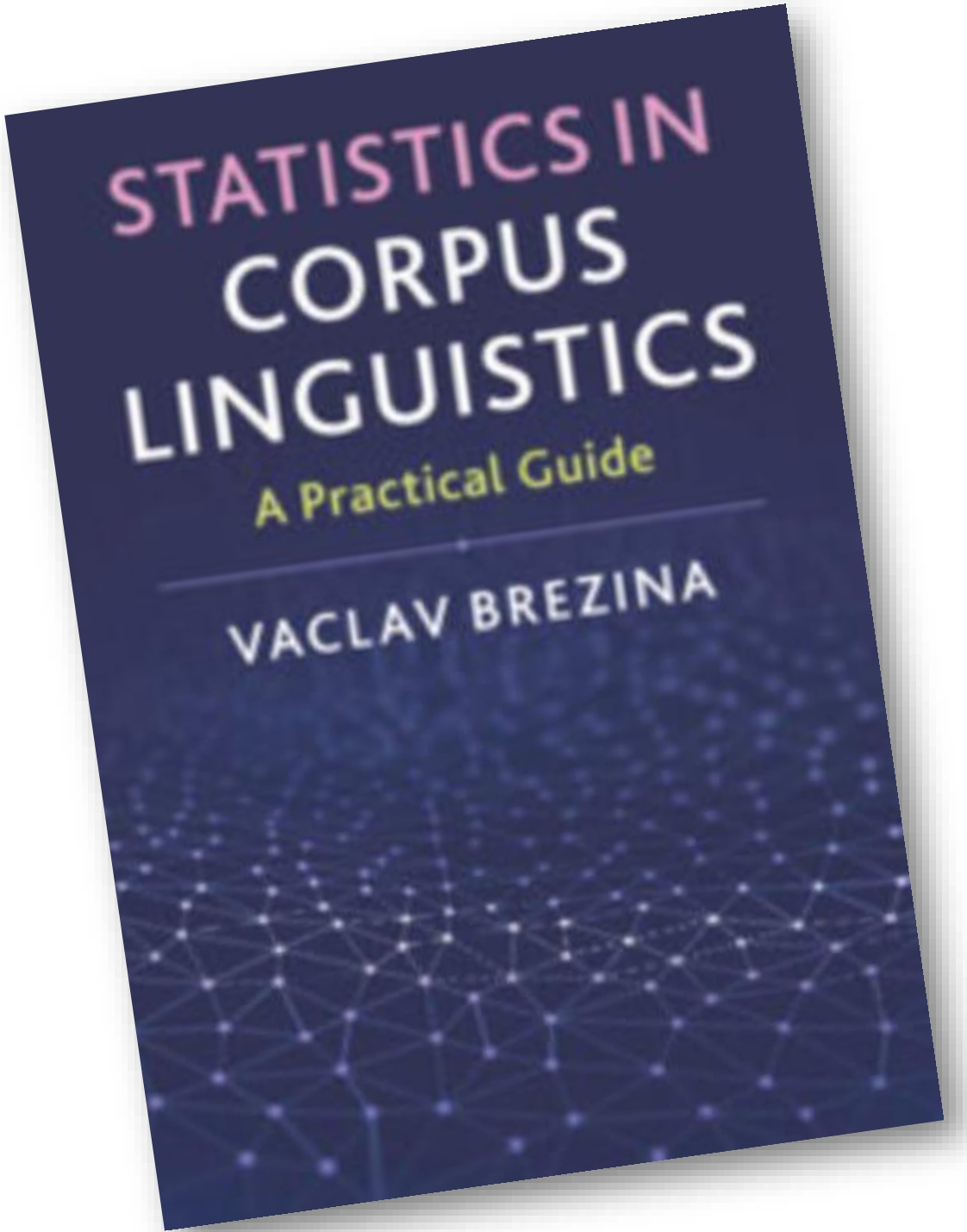
Wednesday 16 October 12.00pm - 12.50pm UK time, **Statistics and language analysis - #LancsBox KWIC**

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)

#LancsBox X 2.0.0

Exampl... 1.0 CLAWS7 | whole corpus | ... | word (lo...

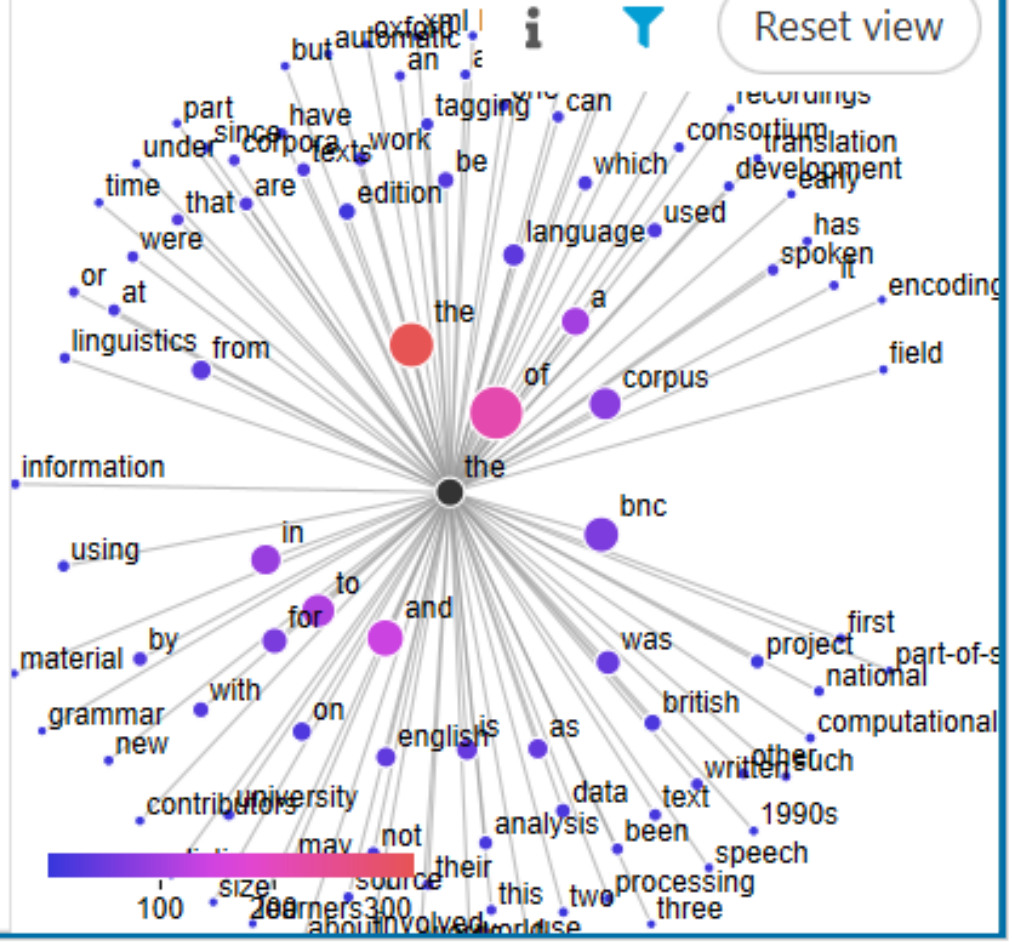
the ×

the Hits: 322 (73,921.03) Texts:

"Log Dic...

Colloc...	Distrib...	Frec	Freq. (...	Lc
of		220	219	13.7
bnc		92	66	12.9
the		152	322	12.9
and		102	140	12.8
corpus		78	78	12.6
to		83	113	12.6

Reset view



Maximized tool.

GRAPH

#LancsBox X 2.0.0

data

Example corpus 1.0 CLAWS7 | whole corpus | 4K

[hw="p.*" pos="n.*"] Hits: 91 (20,890.73) Texts: 2/2

File	Left	Node	Right
text1.x	science,containing relevant computer science	publications	with sentient metadata such as
text1.x	NLP Scholar,a combination of	papers	of the ACL Anthology and
text1.x	which attempt to trace a	path	from data to theory. Wallis
text1.x	what they called the 3A	perspective:	Annotation,Abstraction and Analysis. Ar
text1.x	may include e.g.,rule-learning for	parsers.	Analysis consists of statistically probing,
text1.x	lexical corpora today are part-of-speech-tagged(	POS-tagged).	However even corpus linguists who
text1.x	with other interests and differing	perspectives	than the originators'can exploit

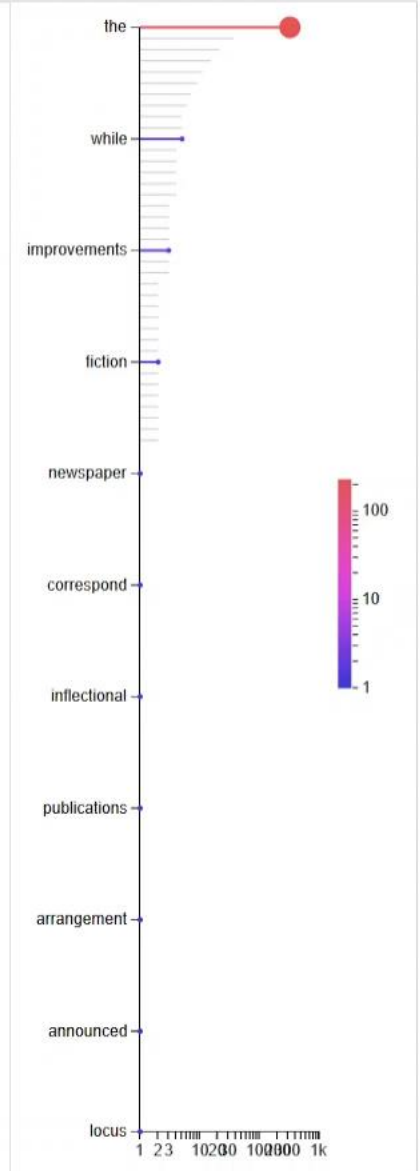
Example c... 1.0 CLAWS7 | whole corpus | ... | word (lowerca...

Reset view

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

Terms: 1,265

Term	Freq...	Rel. f...	ARF (...	Range	Rang...	CV (c...	Juilla...	DP (d...
the	322	73,...	225...	2	100...	0.11	0.89	0.04
of	219	50,...	145...	2	100...	0.13	0.87	0.06
and	140	32,...	96.12	2	100...	0.08	0.92	0.03
to	113	25,...	72.62	2	100...	0.26	0.74	0.09
a	102	23,...	66.51	2	100...	0.12	0.88	0.05
in	91	20,...	59.43	2	100...	0.14	0.86	0.06
cor...	78	17,...	42.26	2	100...	0.35	0.65	0.17
bnc	66	15,...	34.47	1	50.00	1.00	0	0.28
for	64	14,...	40.05	2	100...	0.26	0.74	0.09
was	46	10,...	23.85	2	100...	0.17	0.83	0.06
is	46	10,...	24.13	2	100...	0.05	0.95	0.02
as	43	9,8...	28.45	2	100...	0.06	0.94	0.02
eng...	41	9,4...	16.63	2	100...	0.50	0.50	0.26
lan...	37	8,4...	19.84	2	100...	0.37	0.63	0.18
from	35	8,0...	21.80	2	100...	0.31	0.69	0.11
be	29	6,6...	16.95	2	100...	0.68	0.32	0.21
with	27	6,1...	15.66	2	100...	0.12	0.88	0.05
that	27	6,1...	15.96	2	100...	0.12	0.88	0.05
by	26	5,9...	19.35	2	100...	0.03	0.97	0.01



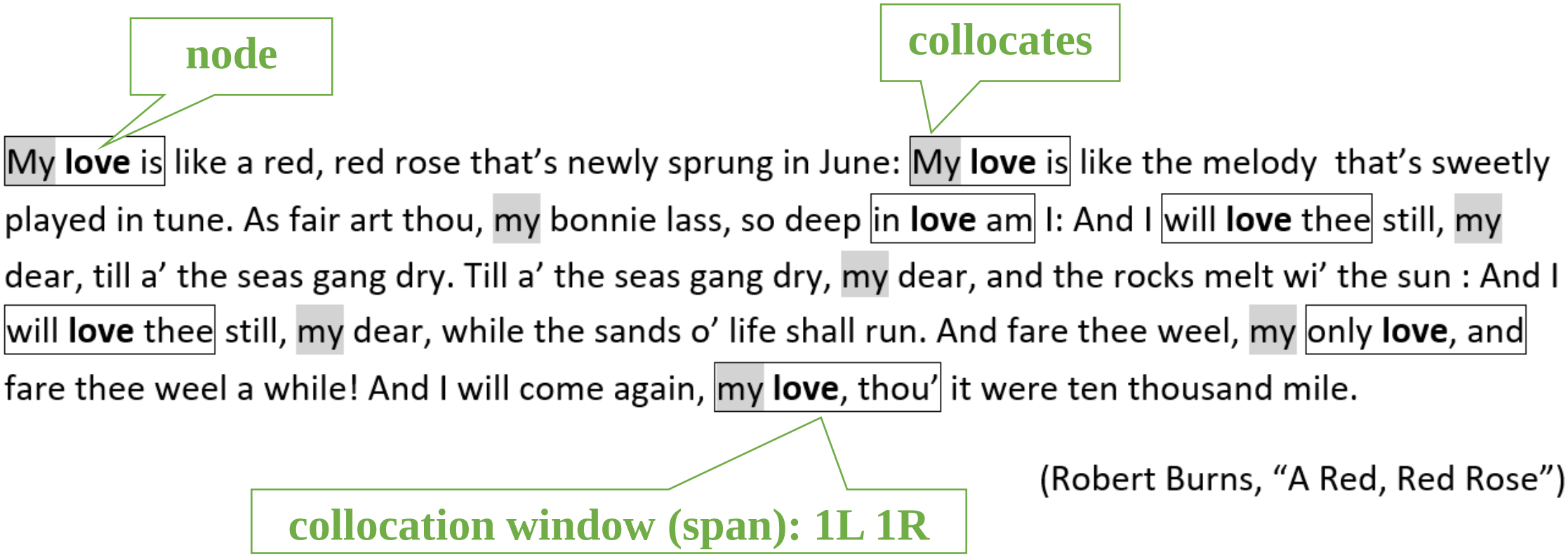
Created GraphColl tool.

Think about and discuss

1. What associations come to your mind when you see the word *love*?
2. Why do you think the word has these associations for you?
3. How can collocations help with establishing links between words?

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

Collocations



Collocations (cont.)

- Is *my* really a genuine collocate of *love* in the poem?
- In other words, is *my* really strongly associated with *love*?
- Observed frequency (3) compared with:
  1) **No baseline:** We compare the observed frequencies of all individual words co-occurring with the node and produce a rank-ordered list.
  2) **Random co-occurrence baseline ('shake the box' model):** We compare the observed frequencies with frequencies expected by chance alone and evaluate the strength of collocation using a mathematical equation which puts emphasis on a particular aspect of the collocational relationship.
  3) **Word competition baseline:** We use a different type of baseline from random co-occurrence; this baseline is incorporated in the equation, which again highlights a particular aspect of the collocational relationship.

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

‘Shake the box’ model

fare art And like red, sweetly in **love** **love**, And gang wi’ played like dear, life shall rocks sprung the Till deep my my And still, weel, again, ten the the while! is till And As I: a’ only come were sands sun: dry, and gang it a’ the still, My thee will in my bonnie My red is a run. my **love** thee thou, melt the seas and th`ou’ I the I lass, I melody thee a my am rose **love** dear, that’s **love** newly **love** fare **love**, will o’ so dry. fair thee will that’s in while June: my seas tune. mile. thousand weel dear,

$$\text{expected frequency of collocation} = \frac{\text{node frequency} \times \text{collocate frequency}}{\text{no. of tokens in text or corpus}}$$

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

Association measures

$$\log_2 \frac{O_{11}^{\,3}}{E_{11}}$$

$$\frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$$

$$O_{12} + \log_2 \frac{O_{12}}{E_{12}}$$

$$\text{Min window} - \text{M outside window}$$

$$\log_2 \frac{O_{11}^{\,2}}{E_{11}}$$

$$2 + \left(O_{11} \times \log \frac{O_{11}}{E_{11}} \right)$$

$$\text{Pooled SD}$$

$$\frac{2 \times O_{11}}{R_1 + C_1}$$

$$\log_2 \frac{O_{11} \times R_2}{O_{21} \times R_{1cor}}$$

$$\frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$$

Log-likelihood

Log ratio

Cohen's d

$$\frac{O_{11}}{R_1} - \frac{O_{21}}{R_2}$$

$$14 + \log_2 \frac{2 \times O_{11}}{R_{1cor} + C_1}$$

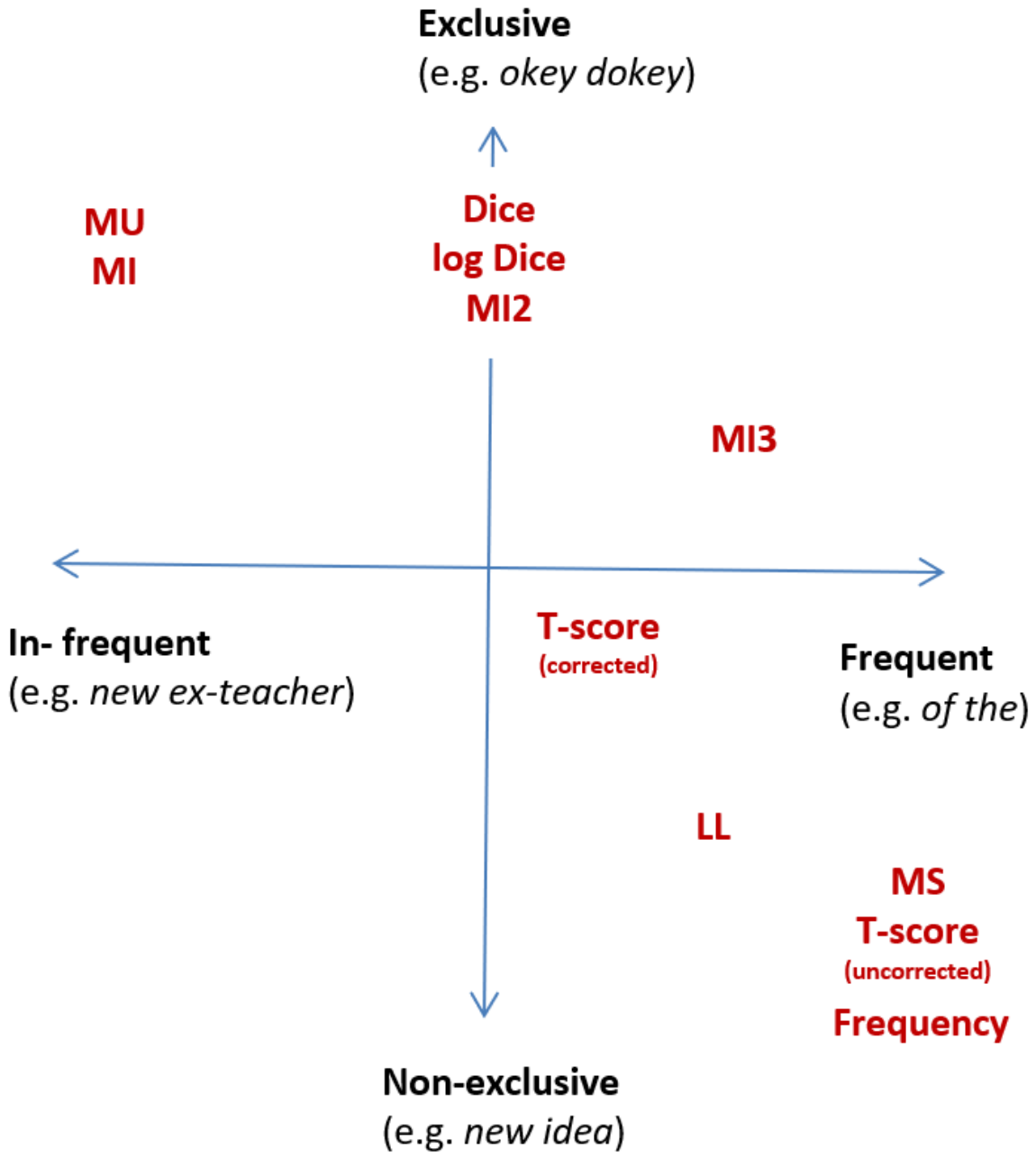
Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Association measures (cont.)

ID	Statistic	Equation	ID	Statistic	Equation
1	Freq. of co-occurrence	O_{11}	8	T-score	$\frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$
2	MU	$\frac{O_{11}}{E_{11}}$	9	DICE	$\frac{2 \times O_{11}}{R_1 + C_1}$
3	MI (Mutual information)	$\log_2 \frac{O_{11}}{E_{11}}$	10	LOG DICE	$14 + \log_2 \frac{2 \times O_{11}}{R_1 + C_1}$
4	MI2	$\log_2 \frac{O_{11}^{\ 2}}{E_{11}}$	11	LOG RATIO	$\log_2 \frac{O_{11} \times R_2}{O_{21} \times R_1}$
5	MI3	$\log_2 \frac{O_{11}^{\ 3}}{E_{11}}$	12	MS (Minimum sensitivity)	$min\left(\frac{O_{11}}{C_1}, \frac{O_{11}}{R_1}\right)$
6	LL (Log likelihood)	$2 \times \begin{pmatrix} O_{11} \times log \frac{O_{11}}{E_{11}} + O_{21} \times log \frac{O_{21}}{E_{21}} + \\ O_{12} \times log \frac{O_{12}}{E_{12}} + O_{21} \times log \frac{O_{22}}{E_{22}} \end{pmatrix}$	13	DELTA P	$\frac{O_{11}}{R_1} - \frac{O_{21}}{R_2}; \frac{O_{11}}{C_1} - \frac{O_{12}}{C_2}$
7	Z-score	$\frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$	14	Cohen's d	$\frac{Mean_{in\ window} - Mean_{outside\ window}}{pooled\ SD}$

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

Association measures (cont.)



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