Chapter 3: Exercises

COLLOCATIONS

1) Which association measures would you use in the following research scenarios? Note that more than one answer is possible in each case – think of your rationale for the answer you choose.

   a) MI score, MU
   b) MI score, MU
   c) Log Dice, MI2

2) Use the online *Collocation Calculator* to calculate four association measures: MI, LL, Delta P and log Dice. N. B. uncorrected versions of the association measure values are displayed.

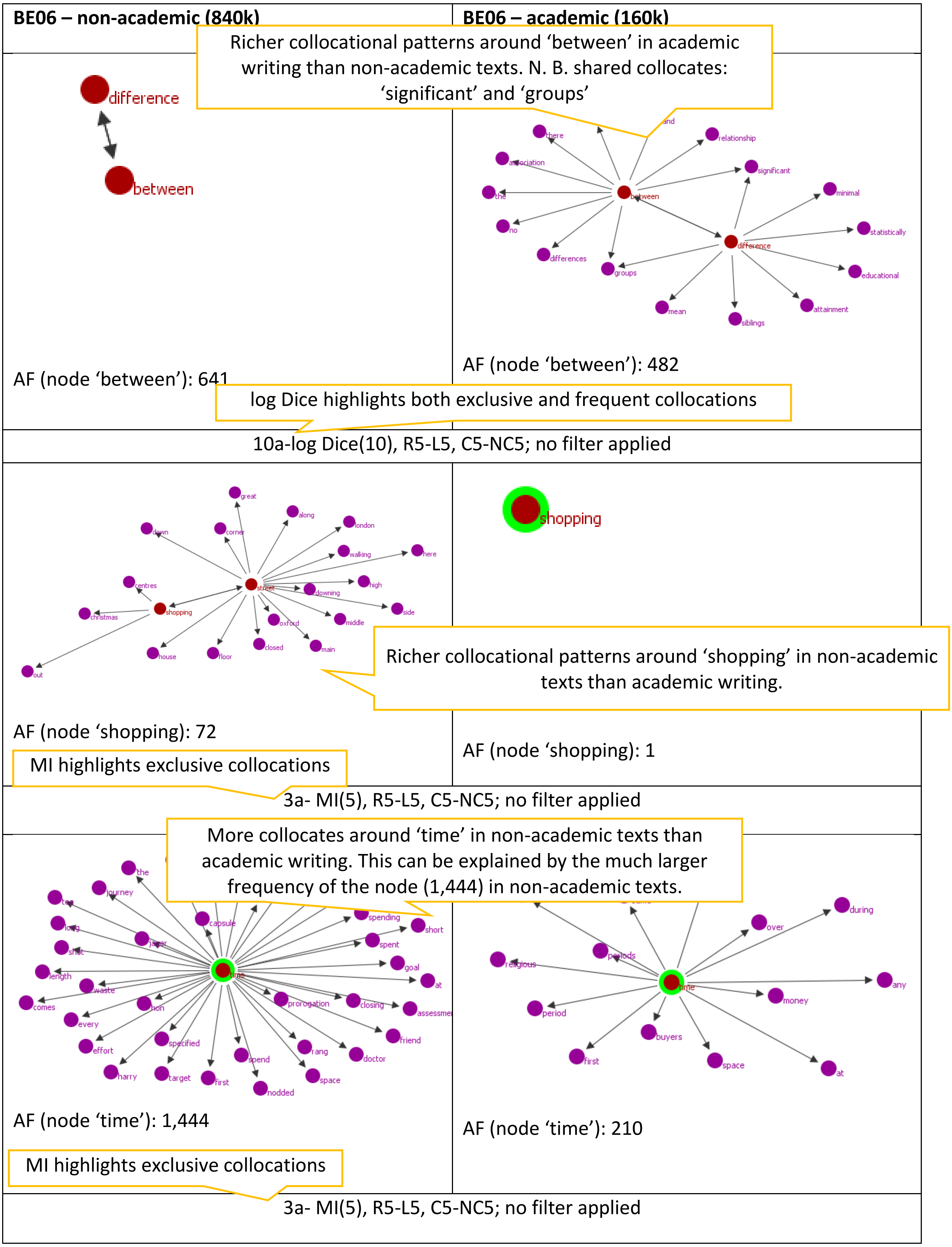
Table 3.15: Collocates of 'issue' in BE06

Collocate	C_1	O_{11}	MI value	LL value (long)	Delta P values	log-Dice value
the	58,591	101	3.396	362.695	0.557; 0.002	5.816
this	4,815	38	5.591	229.6	0.226; 0.008	7.966
important	322	7	7.053	54.994	0.042; 0.022	8.883
address	88	6	8.702	61.046	0.037; 0.068	9.608
bbc	98	5	8.283	47.859	0.030; 0.051	9.289
HUPO-PSI	1	1	12.576	17.440	0.006; 1.000	7.634

3) -

COLLOCATION NETWORKS

4) Compare the following pairs of collocation networks based on a) BE06 – non-academic subcorpus, an 840,000 word sample of written British English ranging from newspapers and general prose to fiction, and b) the academic subcorpus of BE06, which consists of over 160,000 words of academic English. Note that the BE06-non-academic is more than five times larger than its academic English counterpart. Pay attention to the frequencies of the initial node and the CPN parameters, especially the cut-off points and their effect on the collocates that are shown in the graphs.



BE06 – non-academic (840k)

BE06 – academic (160k)

Richer collocational patterns around ‘between’ in academic writing than non-academic texts. N. B. shared collocates: ‘significant’ and ‘groups’

AF (node ‘between’): 641

AF (node ‘between’): 482

log Dice highlights both exclusive and frequent collocations

10a-log Dice(10), R5-L5, C5-NC5; no filter applied

Richer collocational patterns around ‘shopping’ in non-academic texts than academic writing.

AF (node ‘shopping): 72

AF (node ‘shopping): 1

MI highlights exclusive collocations

3a- MI(5), R5-L5, C5-NC5; no filter applied

More collocates around ‘time’ in non-academic texts than academic writing. This can be explained by the much larger frequency of the node (1,444) in non-academic texts.

AF (node ‘time’): 1,444

AF (node ‘time’): 210

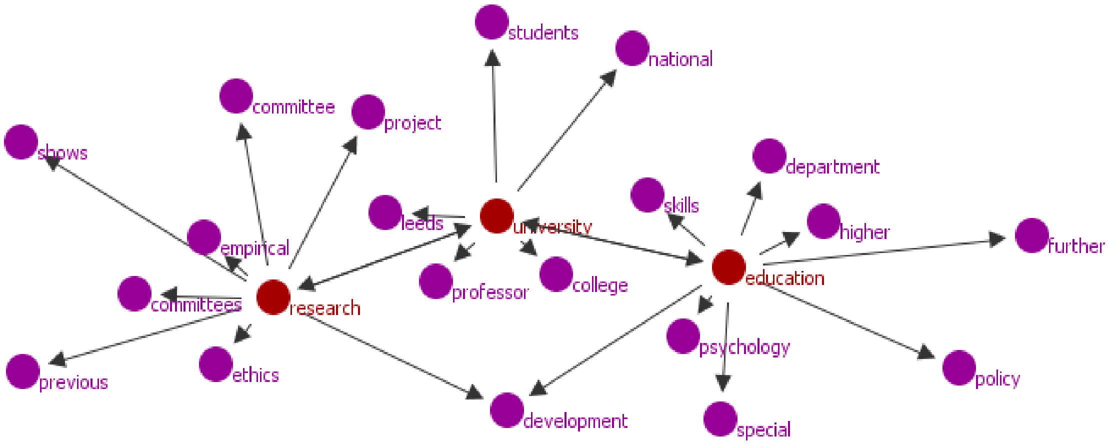
MI highlights exclusive collocations

3a- MI(5), R5-L5, C5-NC5; no filter applied

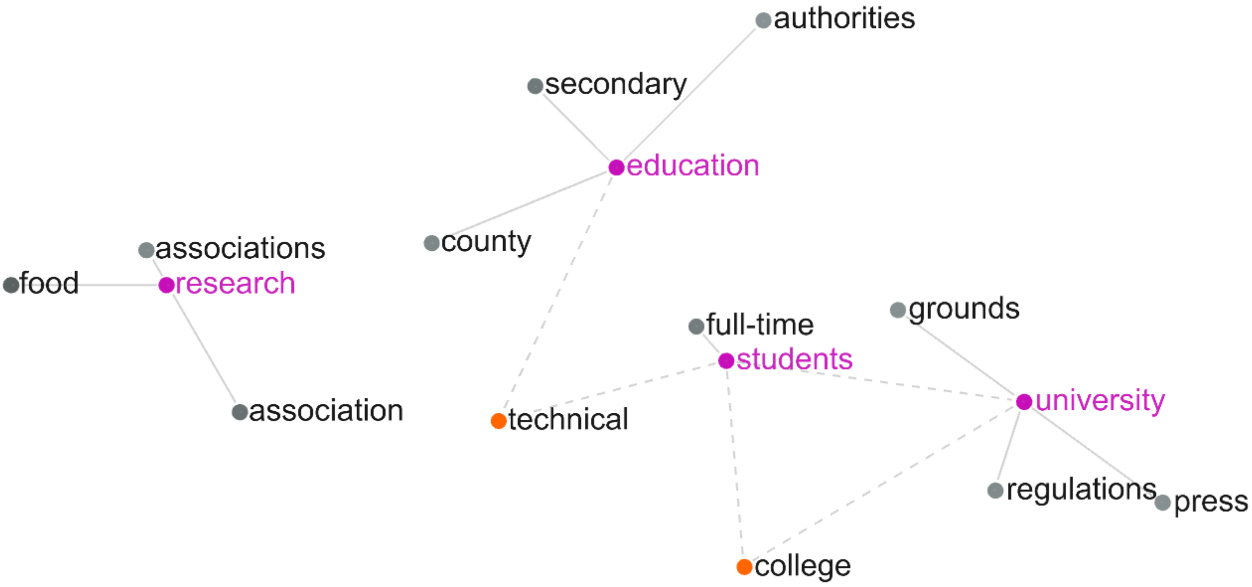
Figure 3.8. Selected collocation networks

5) Use #LancsBox, which is downloadable from http://corpora.lancs.ac.uk/lancsbox, to build collocation networks based on the LOB corpus (available from the Companion website). LOB is a one-million-word corpus representing written British English of the 1960s.
Nodes to search for:

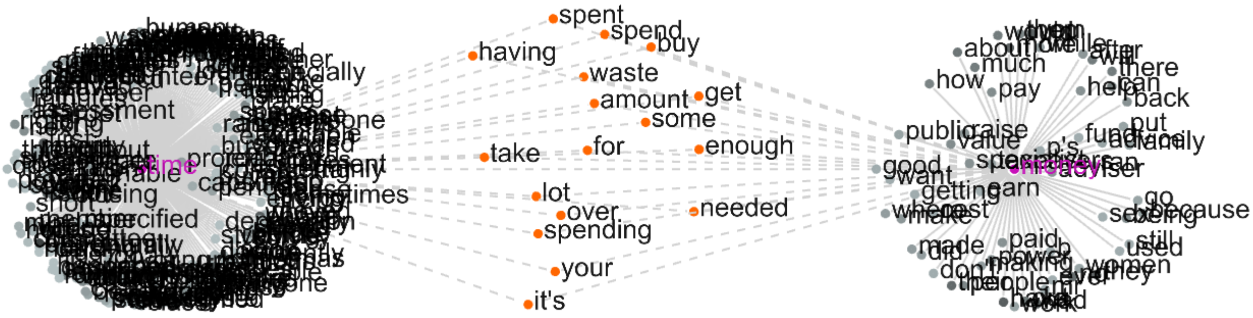
- *university*
- *time*



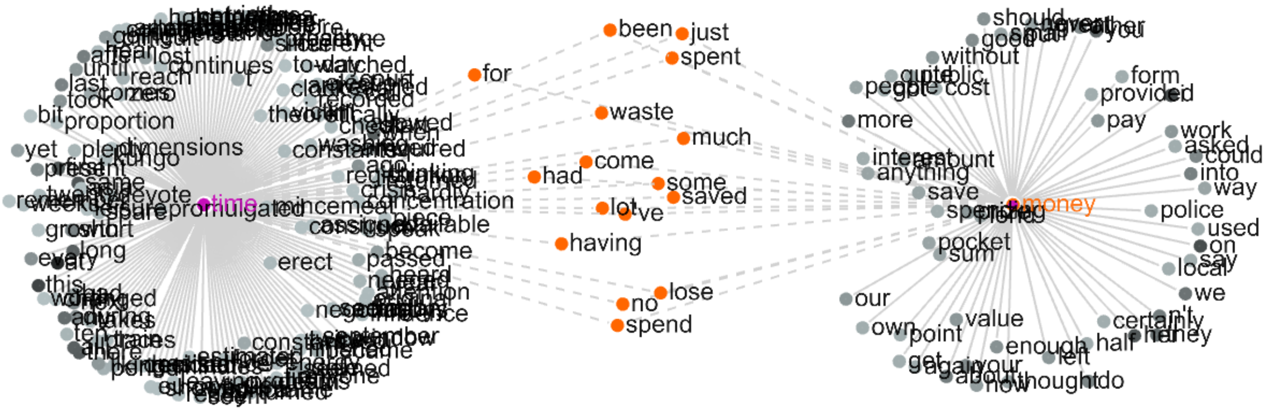
Collocation network of 'university' based on BE06 [3b-MI(3), L5-R5, C8-NC8]



Collocation network of 'university' based on LOB [3b-MI(3), L5-R5, C8-NC8]



Collocation network of 'time' based on BE06 [3a-MI(5`), R5-L5, C4-NC4; no filter applied]



Collocation network of 'time' based on LOB [3a-MI(5`), R5-L5, C4-NC4; no filter applied]

Compare the collocation networks of *time* and *university* based on LOB with the collocation networks built using BE06, which represents British English around 2006, shown in section 3.3. Is there any difference/indication of language development?

KEYWORDS

6) Review the following situations and decide upon an appropriate type of the reference corpus (e.g. general language corpus, specialised corpus representing...) Justify your answer.

a) In a literary stylistic study, we compiled a corpus of all works by a certain author; we want to identify keywords typical of this author of interest. Reference corpus: works of the same genre by other authors from the same period to control for as many relevant variables (genre, time period) as possible.

b) We are interested in keywords typical of the genre of academic writing. We have compiled a corpus of research articles and books in multiple disciplines representing all major academic fields. Reference corpus: a general corpus such as the BNC or BNC2014 representing a variety of genres/registers from which the academic component has been excluded. This will make the investigation more focused with the differences more visible than if we use just a general reference corpus including the academic component.

c) We are interested in keywords typical of spoken language. Our corpus of interest is the spoken part of the British National Corpus. Reference corpus: a balanced written corpus representing a variety of registers/genres (e.g. written BNC) to highlight words that are typical of speech.

7) Calculate the SMP statistic for the words below. Decide which of the words belongs to i) positive keywords (+), ii) negative keywords (-) and iii) lockwords (0).
The following calculations of SMP are made with the constant k = 100;

Table 1: Keywords

Word	C (tokens: 1,007,532)	C Relative freq.	R (tokens: 1,017,879)	R Relative freq.	SMP (Simple Maths Parameter)	Decision (+/-/0)
BBC	106	105.21	3	2.95	1.99	+
before	970	962.75	854	839.00	1.13	+
London	471	467.48	119	116.91	2.62	+
nation	51	50.62	195	191.57	0.52	-
she	4,162	4130.89	4,494	4415.06	0.94	0?
slowly	83	82.38	94	92.35	0.95	0?
today	270	267.98	278	273.12	0.99	0
tomorrow	47	46.65	48	47.16	1.00	0
Washington	27	26.80	222	218.10	0.40	-
which	2,680	2659.97	2,056	2019.89	1.30	+

INTER-RATER AGREEMENT

8) The following ratings we obtained in three situations involving a judgement variable. Calculate the inter-rater agreement in each situation.

A) Situation 1: In a discourse analysis study, a judgement variable with three possible values (1, 2 and 3) was coded by three independent raters. The variable of interest was a nominal variable capturing a discourse category.

Rater A: 2, 1, 1, 2, 1, 1, 3, 3, 2, 2, 3, 1
Rater B: 2, 1, 2, 2, 1, 1, 2, 2, 2, 2, 3, 1

Rater C: 2, 1, 1, 2, 1, 2, 2, 2, 2, 2, 3, 1

The appropriate agreement statistic is either Gwet's AC_1 and Fleiss' κ; $AC_1 = 0.68$ ($p = 0.0004$); Fleiss' kappa = 0.63 ($p = 0.0024$). AC_1 is just above the standard threshold (0.67) for acceptable agreement. N. B. Fleiss' kappa is below this threshold but AC_1 is more robust.

B) Situation 2: In an applied linguistic study, texts from second language speakers were used. Based on the texts, the proficiency of the second language speakers was coded using hierarchically ordered categories (ordinal variable) ranging from 1 (lowest proficiency) to 6 (highest proficiency). A random sample of 20 per cent of the texts was double coded to assess the robustness of the coding.

Rater A: 4, 4, 4, 3, 4, 4, 3, 3, 4, 4, 3, 3, 2, 4, 4, 4, 3, 4, 4, 4
Rater B: 4, 4, 4, 3, 4, 3, 3, 3, 3, 4, 4, 5, 2, 5, 5, 4, 4, 4, 4, 5

The appropriate agreement statistic is Gwet's AC_2; $AC_2 = 0.82$ ($p = 0$). AC_2 is above the standard threshold for very good agreement (0.8).

C) Situation 3: Two transcribers were given the same recording to transcribe. It contains a spoken interaction between six different speakers. Because speaker attribution in a dialogue between multiple speakers is notoriously difficult, the reliability of the speaker codes (1 to 6) at the beginning of each turn was checked by an inter-rater agreement measure.

Transcriber A: 1, 4, 5, 4, 3, 4, 2, 4, 1, 2, 6, 1, 4, 2, 1, 6, 1, 6, 4, 1
Transcriber B: 1, 4, 5, 4, 3, 4, 2, 4, 1, 2, 6, 2, 4, 6, 2, 4, 2, 4, 6, 2

The appropriate agreement statistic is either Gwet's AC_1 and Cohen's κ ; $AC_1 = 0.53$ ($p = 0.0009$); Cohen's kappa = 0.5 ($p = 0.0015$). Both AC_1 and Coken's κ are way below the standard threshold for acceptable agreement (0.67).

9) Look at the examples below taken from the Trinity Lancaster Corpus. They show how speakers of English as a foreign language express *disagreement*. Decide how polite (or impolite) these speakers are when they express disagreement. Use the following rating on a 5-point Likert scale:

1	2	3	4	5
←				→
very polite	polite	neutral	impolite	very impolite

Table 2. Examples for rating

Example	Rating 1	Rating 2
A) I completely disagree with this because er I I repeat as I said …	2	3
B) I agree with this point but don't you think maybe the ti= fact that times are changing is a good thing?	1	1
C) but I personally would disagree that that money would necessarily be spent on that	1	2
D) erm no no it's not so	3	4
E) well I 'm not totally convinced but er you know I live in a really traditional family	2	2
F) mm I can understand your opinion erm but I was still wondering…	2	1
G) I can't agree with you	3	2
H) er er I I think erm I I think they I I think they are wrong	3	1
I) I think they're completely wrong	3	2
J) no way	3	4
K) I think he's stupid	4	5
L) I I I can understand what you 're saying but I'm not I don't agree with that	2	1

After the rating, answer the following questions:

- How confident are you about the ratings you have provided? Moderately – pragmatic interpretation largely depends on the context – cultural, social, situational.
- Would you consider politeness a robust judgement variable? No, politeness offers a lot of scope for disagreement.
- How important do you think it is to have another rater for this judgement variable? Very important.

10) Compare your coding in exercise 9 with the coding of the same dataset by a different rater (e.g. ask a friend to help you with this exercise). Using the Agreement calculator, calculate the appropriate agreement measure.

Measure calculated: __ Gwet's AC$_2$______, Value:_____0.7 _______________

- If available, keep adding more raters and calculating the inter-rater agreement.

11) Imagine you need to produce a research report based on the dataset discussed in exercises 9 and 10. Report the results of the inter-rater agreement measure from exercise 10. Refer back to the 'Reporting statistics' box.

Two raters coded 12 concordance lines from the Trinity Lancaster Corpus independently to identify how polite or impolite the disagreement statements were. The coding was done one a 5-point Likert scale and was assumed to produce an ordinal judgement variable. Gwet's AC_2 measure showed agreement between the raters ($AC_2 = 0.7$, $p < 0.001$). A review of the differences between raters found no systematic pattern of disagreement. Given the nature of the judgement variable the amount of agreement was deemed sufficient.