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) You need to identify technical terms connected with the word *process* in a corpus of research articles on organic chemistry *e*.g. *petrochemical process*. Note that technical terms are exclusive and relatively rare combinations of words with a specific meaning.

   b) You want to study the associations that the word *enemy* (node) has in the newspaper discourse. You are interested to see content words around the node rather than frequent grammatical words.

   c) You want to write a dictionary of collocations for learners of English that would include a broad range of fixed expressions such as *find out, take responsibility, dire consequences* etc. The collocations you include need to be both recognisable as specific meaningful units and they need to occur as frequent combinations.

2) Look at the information about the co-occurrence of the word *issue* in a L3 – R3 collocation window in *BE06,* a one-million corpus of written English. Use the online *Collocation Calculator* to calculate four association measures: MI, LL, Delta P and log Dice.

- Number of tokens in the whole corpus (N): 1,001,514
- Frequency of the node in the whole corpus (R_1): 164
- Collocation window size: 6 (3L, 3R)

Table 1: Collocates of 'issue' in BE06

Collocate	C_1	O_{11}	MI value	LL value	Delta P values	log-Dice value
the	58,591	101				
this	4,815	38				
important	322	7				
address	88	6				
bbc	98	5				
HUPO-PSI	1	1				

3) Discuss how the association measures from exercise 2 rank the six collocates. Which association measure would you choose?

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.

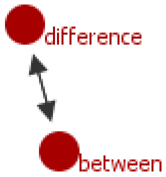
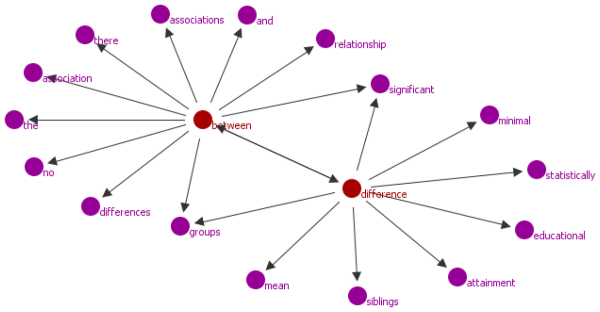
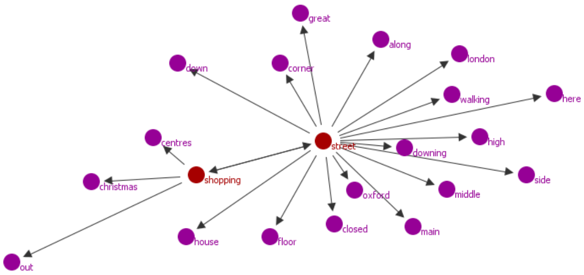
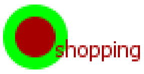
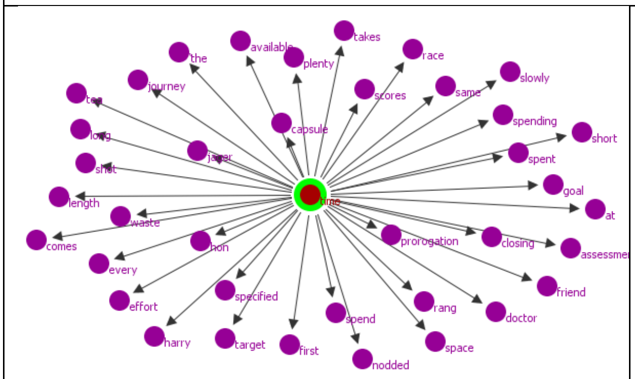
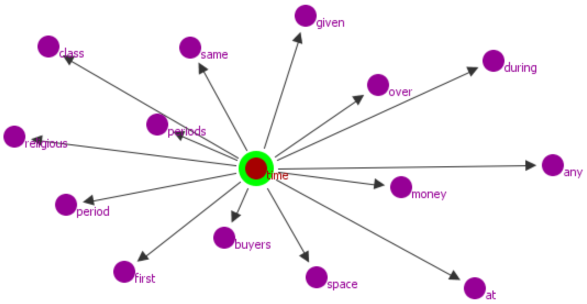
BE06 – non-academic (840k)	**BE06 – academic (160k)**
 AF (node 'between'): 641	 associations and there relationship association significant the between minimal no statistically difference differences groups educational attainment mean siblings AF (node 'between'): 482
10a-log Dice(10), R5-L5, C5-NC5; no filter applied	
 great along down corner london walking here centres street high downing shopping christmas side oxford middle closed main house floor out AF (node 'shopping): 72	 AF (node 'shopping): 1
3a- MI(5), R5-L5, C5-NC5; no filter applied	
 takes available the plenty race journey slowly scores same tea spending capsule long short jager spent shot goal length time at waste prorogation closing comes hon assessmer every friend specified effort rang spend doctor harry target first space nodded AF (node 'time'): 1,444	 given class same during over periods religious any time money period buyers first space at AF (node 'time'): 210
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*

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.
   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.
   c) We are interested in keywords typical of spoken language. Our corpus of interest is the spoken part of the British National Corpus.

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).

Table 1: Keywords

Word	C (tokens: 1,007,532)	R (tokens: 1,017,879)	SMP (Simple Maths Parameter)	Decision (+/-/0)
BBC	106	3		
before	970	854		
London	471	119		
nation	51	195		
she	4,162	4,494		
slowly	83	94		
today	270	278		
tomorrow	47	48		
Washington	27	222		
which	2,680	2,056		

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

   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

   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

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

After the rating, answer the following questions:

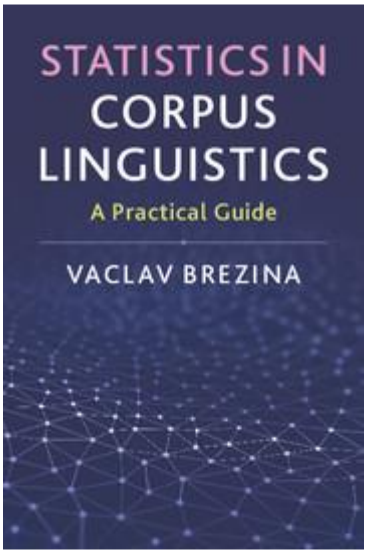
- How confident are you about the ratings you have provided?
- Would you consider politeness a robust judgement variable?
- How important do you think it is to have another rater for this judgement variable?

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: ________, Value:_____________________

- 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.



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

Do you use language corpora in your research or study, but find that you struggle with statistics? This practical introduction will equip you to understand the key principles of statistical thinking and apply these concepts to your own research, without the need for prior statistical knowledge. The book gives step-by-step guidance through the process of statistical analysis and provides multiple examples of how statistical techniques can be used to analyse and visualise linguistic data. It also includes a useful selection of discussion questions and exercises which you can use to check your understanding.

The book comes with a Companion website, which provides additional materials (answers to exercises, datasets, advanced materials, teaching slides etc.) and Lancaster Stats Tools online, a free click-and-analyse statistical tool for easy calculation of the statistical measures discussed in the book.