Chapter 2: Exercises

1) Look at sentences a) – d) below and count the number of *tokens, types, lemmas* and *lexemes* in each.

   a) The City is braced for far worse figures to come in the coming months, unless the Government recovery package produces a startling turn round in optimism. [source: BNC, CEN]

   b) Of 354 fifth- and sixth-formers who left Sharon's school in the summer of 1981 forty had found real jobs by 18 November, four of these having entered military service. [source: BNC, GUR]

   c) Erm erm erm but, yeah and people er have great areas of that taken. [source: BNC, KC3]

   d) Homonyms are headwords to different entries that are spelt in the same way, e.g. bow (the weapon), bow (the action), bow (the verb expressing the action). [source: BNC, EAT]

2) Use the online *Word Calculator* to compare your results from Exercise 1 to the automatically generated token, type and lemma counts. Did you get the same results? If not can you explain the differences?

3) Use the online *Word Calculator* to compare different texts from the Internet. Calculate the lexical density using the three measures discussed in this chapter: simple TTR, STTR and MATTR. Compare the findings and think about which of the measures would be most appropriate to us with the text? What are your reasons for selecting the measure?

4) Calculate the relative frequencies of the following items. In each case, choose an appropriate basis for normalization.

   a) word: *muggle*
   absolute frequency: 2
   corpus size: 100,000

   b) word: *intriguingly*
   absolute frequency: 3,035
   corpus size: 11,191,860,036

   c) word: *worse*
   absolute frequency: 50
   corpus size: 1,007,299

5) Look at the frequency list below. It shows ten words from the BNC, together with their ranks. Use Zipf's law to predict the absolute frequency of the items presented in the table.

rank	word	absolute frequency
1.	the	6,041,234
2.	of	
3.	and	
4.	to	
5.	a	
10.	was	
50.	so	
100.	way	
1,000.	limited	
10,000.	conveniently	

6) Compare your results from question 5 with the actual frequencies provided in the Answers section at the end of this book. How well did Zipf's law predict the frequencies?
7) Look at the absolute frequencies of five selected words in the broadly-defined genre parts of the BNC (Table 1). Electronic version of this table is available from the Companion website.

Table 1 BNC: Distribution of five selected words

BNC part	Total no. of tokens	*some* (AF)	*smile* (AF)	*theory* (AF)	*chance* (AF)
Fiction and verse	16,143,913	24,616	5,498	347	2,645
Newspapers	9,412,174	10,520	304	266	2,589
Non-academic prose and biography	24,178,674	43,161	385	3,977	2,191
Academic prose	15,778,028	30,297	58	6,588	923
Other written material	22,390,782	37,867	488	1,268	3,323
Speech	10,409,858	20,589	112	363	1,138
Whole corpus	98,313,429	167,050	6,848	12,809	12,809

For each word, calculate:

a) the Range.
b) the Standard deviation.
c) the Coefficient of variation.
d) Juilland's D.
e) DP

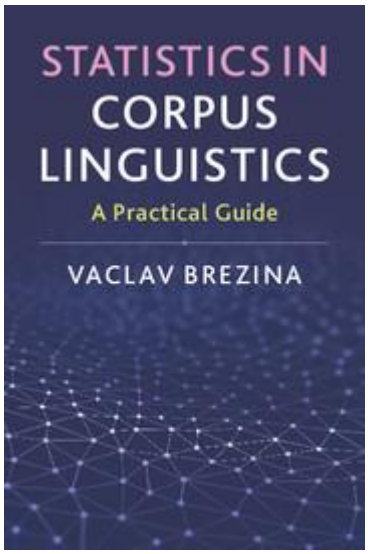
8) Use the *Dispersion calculator* to check your results from question 7.

9) Calculate the ARF of the following words in the *BE06* corpus (985,628 tokens):
a) *frigid*: AF: 2, corpus positions: 840,797 – 848,280

b) *chemistry*: AF = 7, corpus positions: 160,129 – 589,607 – 594,834 – 596,351 – 611,214 – 948,612 – 950,458

c) *porn*: AF = 14, corpus positions: 16,602 – 16,792 – 28,191 – 49,606 – 161,929 – 170,396 – 268,155 – 497,891 – 497,916 – 498,146 – 498,205 – 498,216 – 498,246 – 498,361

10) Use the *ARF Calculator* to compare AF and ARF for different words in texts of your choice from the Internet.



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.