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 use 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.
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 Error! No text of specified style in document.1). Electronic version of this table is available from the Companion website.

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

For each word, calculate:

a) the Range.
b) the Standard deviation.
c) the Coefficient of variation.
d) Juillard’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):

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