LETTER AND WORD FREQUENCIES

REX GOOC H
Letchworth Garden City, Herts, England
rexgooch@ntlworld.com

Counts of letters and words have been made from various bodies of text over the years. Perhaps the earliest well-known corpus was that assembled by W.N. Francis and H. Kucera of Brown University (hence known as the Brown corpus). It contains one million words of American English from 1961 in 15 categories of 500 texts. If it has a use today, it is because matching studies were done of British English (LOB corpus) and Indian English (Kolhapur Corpus) among others. In Word Ways, I have preferred to use the Collins COBUILD corpus, known as the Bank of English®: it runs to hundreds of millions of words of English text from British, US, Australian, and Canadian sources (including textbooks, novels, newspapers, guides, magazines, and websites). The corpus has been automatically word-class tagged, and a 200-million-word corpus has been parsed. It is at http: //www.cobuild.collins.co.uk/. Recently, I have used the British National Corpus, which is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written (available from http: //www.natcorp.ox.ac.uk/, the Oxford University website). There is a Baby version, and the original Brown corpus, tagged, is also available.

In the February issue of Word Ways (page 41), Anil remarked that the order of popularity of letters varied greatly depending on whether the letters are counted from running text, or from a list of the most popular words. I believe he was talking about column 6 below, to which I have added actual frequencies. He was mainly comparing the order with that in column 4, the Linotype machine. The Linotype order is very close to that of the BNC running text in column 3. There is a significant difference in order between letters and running text. However, in all cases E is dominant, followed by the same cluster of letters — differences in frequency are small in this cluster, so letters easily change places, and the differences in frequency are not truly significant. I thought that perhaps any difference between running text and words might be due to a difference between the most popular words (often repeated in running text), and others. I therefore decided to count the letters in the most popular 999 words in the BNC, then in the next 85,801: the results in columns 1 and 2 do in my view show significant differences, eg in the more popular words (which are shorter), E is more frequent, and H less so. Similarly, T seems to be more common in running text. I leave the reader to browse the results further. Column 5 is from the same source as Anil’s column 6, but has no plural words or words with common suffixes: the lack of terminal -S and -ING are just apparent.

Finally, I include the top 100 words from the BNC, and from the American Heritage Word Frequency Book (Carroll, Davies and Richman, 1971). It draws upon 5 million running words used in US schools. The numbers are the number of occurrences per 10,000 words for the BNC.

I trust the tables will also serve as a reference for Word Ways contributors.

© Rex Gooch 2006
### Parts per thousand for individual letters in various selections of text

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>E 134</td>
<td>E 108</td>
<td>E 125</td>
<td>E 114</td>
<td>E 117</td>
</tr>
<tr>
<td>2</td>
<td>A 73</td>
<td>A 85</td>
<td>T 92</td>
<td>T 85</td>
<td>I 86</td>
</tr>
<tr>
<td>3</td>
<td>T 73</td>
<td>P 81</td>
<td>A 79</td>
<td>T 74</td>
<td>R 74</td>
</tr>
<tr>
<td>4</td>
<td>R 73</td>
<td>S 76</td>
<td>O 76</td>
<td>O 75</td>
<td>A 79</td>
</tr>
<tr>
<td>5</td>
<td>O 71</td>
<td>R 73</td>
<td>I 73</td>
<td>I 74</td>
<td>R 74</td>
</tr>
<tr>
<td>6</td>
<td>N 70</td>
<td>N 72</td>
<td>N 70</td>
<td>N 71</td>
<td>N 74</td>
</tr>
</tbody>
</table>

- **BNC**
  - THE 651 the
  - OF 309 of
  - AND 282 and
  - TO 269 a
  - A 226 to
  - IN 198 in
  - THAT 117 is
  - IT 114 you
  - IS 105 that
  - WAS 97 it
  - I 95 he
  - FOR 89 for
  - ON 76 was
  - YOU 73 on
  - HE 71 are
  - BE 69 as
  - WITH 68 with
  - AS 54 his
  - BY 53 they
  - AT 50 at
  - HAVE 49 be
  - ARE 49 this
  - THIS 48 from
  - NOT 48 I
  - BUT 47 have
  - HAD 46 or
  - HIS 45 by
  - THEY 45 one
  - FROM 43 had
  - SHE 40 not
  - WHICH 39 but
  - OR 39 what
  - WE 37 all
  - AN 36 were
  - THERE 34 when
  - HER 34 we
  - WERE 33 there
  - ONE 30 can
  - DO 29 an
  - BEEN 28 your
  - ALL 27 which
  - THEIR 27 their
  - HAS 27 said
  - WOULD 26 if
  - WILL 26 do
  - WHAT 26 will
  - IF 24 each
  - CAN 24 about
  - WHEN 22 how
  - SO 22 up

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NO 22</td>
<td>out</td>
<td>SAID 22</td>
<td>them</td>
<td>WHO 21</td>
</tr>
<tr>
<td></td>
<td>MORE 21</td>
<td>she</td>
<td>ABOUT 20</td>
<td>many</td>
<td>UP 19</td>
</tr>
<tr>
<td></td>
<td>THEM 18</td>
<td>so</td>
<td>SOME 18</td>
<td>these</td>
<td>COULD 17</td>
</tr>
<tr>
<td></td>
<td>HIM 17</td>
<td>other</td>
<td>INTO 17</td>
<td>into</td>
<td>ITS 17</td>
</tr>
<tr>
<td></td>
<td>THEN 16</td>
<td>more</td>
<td>TWO 16</td>
<td>her</td>
<td>OUT 16</td>
</tr>
<tr>
<td></td>
<td>TIME 16</td>
<td>like</td>
<td>LIKE 16</td>
<td>him</td>
<td>ONLY 16</td>
</tr>
<tr>
<td></td>
<td>MY 16</td>
<td>time</td>
<td>DID 15</td>
<td>could</td>
<td>OTHER 14</td>
</tr>
<tr>
<td></td>
<td>ME 14</td>
<td>make</td>
<td>YOUR 14</td>
<td>than</td>
<td>NOW 14</td>
</tr>
<tr>
<td></td>
<td>OVER 13</td>
<td>been</td>
<td>JUST 13</td>
<td>its</td>
<td>MAY 13</td>
</tr>
<tr>
<td></td>
<td>THESE 13</td>
<td>now</td>
<td>NEW 13</td>
<td>people</td>
<td>ALSO 13</td>
</tr>
<tr>
<td></td>
<td>PEOPLE 13</td>
<td>made</td>
<td>ANY 13</td>
<td>over</td>
<td>KNOW 12</td>
</tr>
<tr>
<td></td>
<td>VERIFY 12</td>
<td>down</td>
<td>SEE 12</td>
<td>only</td>
<td>FIRST 12</td>
</tr>
<tr>
<td></td>
<td>USE 12</td>
<td>after</td>
<td>AFTER 12</td>
<td>use</td>
<td>SHOULD 11</td>
</tr>
<tr>
<td></td>
<td>MAY 11</td>
<td>over</td>
<td>THAN 11</td>
<td>water</td>
<td>WHERE 11</td>
</tr>
<tr>
<td></td>
<td>BACK 10</td>
<td>little</td>
<td>HOW 10</td>
<td>very</td>
<td>GET 10</td>
</tr>
<tr>
<td></td>
<td>WATER 10</td>
<td>words</td>
<td>WAY 10</td>
<td>called</td>
<td>DOWN 10</td>
</tr>
<tr>
<td></td>
<td>OUR 9</td>
<td>where</td>
<td>MADE 9</td>
<td>most</td>
<td>GOT 9</td>
</tr>
</tbody>
</table>

1. top 999 words in BNC (5.5 letters per word)
2. next 85801 words in BNC (7.7 letters per word)
3. running text, BNC
4. Linotype
5. 18584 Common Base words
6. 45406 Common Words