Word-Length Distribution in Modern Welsh Prose Texts

Andrew Wilson  
*Lancaster University*

**INTRODUCTION**

The Celtic language family is made up of two distinct sub-groups: P-Celtic (or Brythonic) and Q-Celtic (or Goidelic). The P-Celtic group consists of Welsh, Cornish and Breton, whilst the Q-Celtic group consists of Irish, Scottish Gaelic and Manx.

Although very little Celtic data has yet been examined within the Göttingen project on word-length distributions, one set of Q-Celtic data has already been processed – a set of 31 Scottish Gaelic e-mails, for which the best-fit distribution was the 1-displaced hyperpoisson distribution (Drechsler 2001). This study will add data from a P-Celtic language – Welsh – in order to obtain a preliminary impression of whether Celtic is likely to show the same distribution for all its member languages, or whether there are likely to be differences, perhaps along the Q-Celtic versus P-Celtic dimension.

**DATA**

The data for this study consisted of twelve Welsh prose texts. All were written within the past twenty years and most of them within the past one to two years. They were selected from two main genres: Bible texts and news reports. Within the category of Bible texts, two psalms and two short epistles from the *Beibl Cymraeg Newydd* (1985) were processed; these translations date from the late 1970s. Within the category of news reports, four texts from the Welsh weekly newspaper *Y Cymro* were processed (two general news items and two sports items) as well as four texts from the Welsh-language version of the University of Wales Bangor’s in-house newsletter, *Newyddlen*.

The texts used were as follows:

*Beibl Cymraeg Newydd:*

- Text B1  2 John
- Text B2  3 John
- Text B3  Psalm 20
- Text B4  Psalm 21

*Y Cymro:*

- Text C1  Tachwed 24, 1999: Stamp Cristnogaeth ar y Mileniwm
- Text C2  Tachwed 24, 1999: Diffyd deddf addysg i Gymru yn araith fawr y Frenhines
- Text C3  Tachwed 24, 1999: Inter yn brawf i Llanelli (sport)
- Text C4  Rhafgyr 1, 1999: Cadw safonau y clybiau (sport)

*Newyddlen:*

- Text N1  Hydref 1999: Cyfnod Newydd o Hanes ar Safle’r George
- Text N2  Mehefin 2000: Gradd i Paxman y Dyfodol
- Text N3  Mehefin 2000: Glinfyrrdau i Nyrsys
- Text N4  Tachwed 2000: Argyfwng – Pa Argyfwng?
METHOD

For each text analysed, the number of words falling into each word-length class was counted.

Using the standard rules of pronunciation for spoken Welsh (Rowland 1857, Williams 1980), the word lengths were determined in accordance with the usual principles of the Göttingen project, i.e., in terms of the number of spoken syllables per orthographic word.

In line with the general guidelines of the project, abbreviations were treated as instances of the full word for which the abbreviation stands – so, for example, Dr is counted as a two-syllable word (doctor). Acronyms, in contrast, are treated as single words and the number of syllables counted as pronounced; thus, for instance, PCB is counted as one word with three syllables.

Like abbreviations, numerals are treated as instances of the fully spelled out forms. Thus, 111 is treated as three words of one, two and one syllables respectively (cant undeg un). It should be noted here that a special feature of the Welsh language is that it has two counting systems: a decimal system and a vigesimal system (King 1993: 111-114). This means that a number such as 21 can be spoken as either dauddeg un (decimal) or un ar hugain (vigesimal). For the purposes of the present work, the decimal system was used throughout.

Clitics (e.g. the i in o’i) are pronounced as an integral part of the preceding word and were treated as such here.

The word-length frequency statistics for each text were then run through the Altmann Fitter software at Göttingen to determine which probability distribution was the most appropriate model.

STATISTICS

The Altmann Fitter compares the empirical frequencies obtained in the data analysis with the theoretical frequencies generated by the various probability distributions (Wimmer & Altmann 1996; 1999). The degree of difference between the two sets of frequencies is measured by the chi-squared test and also by the discrepancy coefficient C; the latter is given by $X^2/N$ and is used especially where d.f. = 0. A probability distribution is considered an appropriate model for the data if the difference between the empirical and theoretical frequencies is not significant, i.e., if P($X^2$) > 0.05 and/or C < 0.02. The best distribution is that which shows the highest P and/or lowest C.

RESULTS

The best results were achieved by fitting the 1-displaced Singh-Poisson distribution, which is given by:

$$P_x = \begin{cases} 1 - \alpha + \alpha e^{-\alpha}, & x = 1 \\ \alpha \frac{x^{-1} e^{-\alpha}}{(x - 1)!}, & x = 2, 3, 4, \ldots \end{cases}$$

Although two other distributions – the positive Singh-Poisson distribution and the 1-displaced hyperpoisson distribution – could also be fitted, these showed poorer goodness of fit.

The individual results for the 1-displaced Singh-Poisson distribution are shown in the tables below, where:
X = number of syllables in the word  
E[X] = frequency of words with X syllables in the text  
Np[X] = expected frequency calculated by the relevant probability formula  
\(X^2\) = chi-square value  
d.f. = number of degrees of freedom  
P[\(X^2\)] = probability of chi-square value  
C = the coefficient \(X^2/N\)  
Parameters = \(\alpha\) and a in the above equation.

In the case of text N4, it was necessary to merge two length classes in order to obtain a satisfactory fit.

**Text B1**

Parameters: \(a = 0.7060646748\) \(\alpha = 0.6694103212\)  
DF = 1 \(X^2= 1.298\) \(P[\(X^2\)] = 0.2546\) \(C = 0.0043\)

<table>
<thead>
<tr>
<th>X[i]</th>
<th>F[i]</th>
<th>NP[i]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>202</td>
<td>201.60</td>
</tr>
<tr>
<td>2</td>
<td>69</td>
<td>71.15</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>25.12</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>7.13</td>
</tr>
</tbody>
</table>

**Text B2**

Parameters: \(a = 0.6213749753\) \(\alpha = 0.7945991211\)  
DF = 1 \(X^2= 0.049\) \(P[\(X^2\)] = 0.8247\) \(C = 0.0002\)

<table>
<thead>
<tr>
<th>X[i]</th>
<th>F[i]</th>
<th>NP[i]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>201</td>
<td>201.06</td>
</tr>
<tr>
<td>2</td>
<td>84</td>
<td>84.34</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>26.20</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>6.40</td>
</tr>
</tbody>
</table>

**Text B3**

Parameters: \(a = 0.598712857760379\) \(\alpha = 0.999952378637506\)  
DF =1 \(X^2= 0.6123\) \(P[\(X^2\)] = 0.4339\) \(C = 0.0043\)

<table>
<thead>
<tr>
<th>X[i]</th>
<th>F[i]</th>
<th>NP[i]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>77</td>
<td>78.0347</td>
</tr>
<tr>
<td>2</td>
<td>48</td>
<td>46.7163</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>13.9848</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>3.2642</td>
</tr>
</tbody>
</table>

**Text B4**

Parameters: \(a = 0.930837900126862\) \(\alpha = 0.687377351544798\)  
DF =2 \(X^2= 4.6516\) \(P[\(X^2\)] = 0.0977\) \(C = 0.0241\)

<table>
<thead>
<tr>
<th>X[i]</th>
<th>F[i]</th>
<th>NP[i]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>114</td>
<td>112.6353</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>48.6820</td>
</tr>
</tbody>
</table>

Text C1

Parameters: \( a = 1.0094578515139 \) \( \alpha = 0.704986474032476 \)
DF =2 \( X^2 = 1.9727 \) \( P(X^2) = 0.3729 \) \( C = 0.0103 \)

<table>
<thead>
<tr>
<th>( X[i] )</th>
<th>( F[i] )</th>
<th>( NP[i] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>106</td>
<td>105.4171</td>
</tr>
<tr>
<td>2</td>
<td>53</td>
<td>49.5337</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>25.0011</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>8.4125</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>2.6356</td>
</tr>
</tbody>
</table>

Text C2

Parameters: \( a = 0.830424053348335 \) \( \alpha = 0.919465873021633 \)
DF =2 \( X^2 = 3.4876 \) \( P(X^2) = 0.1749 \) \( C = 0.0136 \)

<table>
<thead>
<tr>
<th>( X[i] )</th>
<th>( F[i] )</th>
<th>( NP[i] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>124</td>
<td>123.2119</td>
</tr>
<tr>
<td>2</td>
<td>81</td>
<td>85.1975</td>
</tr>
<tr>
<td>3</td>
<td>43</td>
<td>35.3750</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>9.7921</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2.4234</td>
</tr>
</tbody>
</table>

Text C3

\( X^2 = 0.000 \) \ d.f. = 1 \( P(X^2) = 0.9906 \) \( C = 0.0000 \)
Parameters: \( a = 0.5989163857 \) \( \alpha = 0.9506187404 \)

<table>
<thead>
<tr>
<th>( X[i] )</th>
<th>( F[i] )</th>
<th>( NP[i] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>183</td>
<td>182.93</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>100.09</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>29.97</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>7.01</td>
</tr>
</tbody>
</table>

Text C4

\( X^2 = 1.361 \) \ d.f. = 1 \( P(X^2) = 0.2434 \) \( C = 0.0043 \)
Parameters: \( a = 0.6098844120 \) \( \alpha = 0.9594112977 \)

<table>
<thead>
<tr>
<th>( X[i] )</th>
<th>( F[i] )</th>
<th>( NP[i] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>179</td>
<td>178.69</td>
</tr>
<tr>
<td>2</td>
<td>99</td>
<td>101.11</td>
</tr>
<tr>
<td>3</td>
<td>35</td>
<td>30.83</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>7.37</td>
</tr>
</tbody>
</table>

Text N1

Parameters: \( a = 0.687604226884657 \) \( \alpha = 0.85190747385472 \)
These results suggest that the 1-displaced Singh-Poisson distribution is the best-fit probability distribution for word lengths in modern Welsh prose texts.

As this distribution could be fitted to all the text-types treated in the study, it seems that genre and domain of discourse are unlikely to affect the distribution of word lengths in Welsh prose. However, further studies are required to confirm this hypothesis. Word-length distributions in Welsh verse also deserve
attention, since, in some languages (such as Latin – Wilson 2001) the prose/verse distinction can be significant in determining the distribution of word lengths.

Comparing this set of Welsh data with Drechsler’s (2001) Scottish Gaelic data, it seems that the split between P- and Q-Celtic may also have led to different patterns of word-length distribution in the two branches. However, in order to obtain a fuller and more representative picture of word-length distributions in Celtic, further studies also need to be carried out on Breton, Irish, Cornish and Manx, as well as on other text-types of Welsh and Scottish Gaelic.

REFERENCES


Rowland T 1857 A grammar of the Welsh language, based on the most approved systems. Hughes & Butler, London.


Wimmer G, Altmann G 1999 Thesaurus of univariate discrete probability distributions. STAMM Verlag, Essen.