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	Tachwedd 24, 1999:	Stamp Cristnogaeth ar y Mileniwm
Text C2	Tachwedd 24, 1999:	Diffyd deddf addysg i Gymru yn araith fawr y Frenhines
Text C3	Tachwedd 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:	Glinfyrddau i Nyrsys
Text N4	Tachwedd 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:

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

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 = α 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 α = 0.6694103212

DF = 1 X^2= 1.298 P[X^2]= 0.2546 C= 0.0043

X[i]	F[i]	NP[i]
1	202	201.60
2	69	71.15
3	29	25.12
4	5	7.13

Text B2

Parameters: a = 0.6213749753 α = 0.7945991211

DF = 1 X^2= 0.049 P[X^2]= 0.8247 C= 0.0002

X[i]	F[i]	NP[i]
1	201	201.06
2	84	84.34
3	27	26.20
4	6	6.40

Text B3

Parameters: a = 0.598712857760379 α = 0.999952378637506

DF =1 X^2 = 0.6123 P(X^2) = 0.4339 C = 0.0043

X[i]	F[i]	NP[i]
1	77	78.0347
2	48	46.7163
3	15	13.9848
4	2	3.2642

Text B4

Parameters: a = 0.930837900126862 α = 0.687377351544798

DF =2 X^2 = 4.6516 P(X^2) = 0.0977 C = 0.0241

X[i]	F[i]	NP[i]
1	114	112.6353
2	45	48.6820

3	29	22.6575
4	5	7.0302
5	0	1.9949

Text C1

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

X[i]	F[i]	NP[i]
1	106	105.4171
2	53	49.5337
3	20	25.0011
4	8	8.4125
5	4	2.6356

Text C2

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

X[i]	F[i]	NP[i]
1	124	123.2119
2	81	85.1975
3	43	35.3750
4	7	9.7921
5	1	2.4234

Text C3

X^2= 0.000   d.f.= 1 $P[X^2]$= 0.9906 C= 0.0000
Parameters: a = 0.5989163857   α = 0.9506187404

X[i]	F[i]	NP[i]
1	183	182.93
2	100	100.09
3	30	29.97
4	7	7.01

Text C4

X^2= 1.361   d.f.= 1 $P[X^2]$= 0.2434 C= 0.0043
Parameters: a = 0.6098844120 α = 0.9594112977

X[i]	F[i]	NP[i]
1	179	178.69
2	99	101.11
3	35	30.83
4	5	7.37

Text N1

Parameters: a = 0.687604226884657   α = 0.85190747385472

DF =1 X^2 = 0.0613 $P(X^2)$ = 0.8045   C = 0.0003

X[i]	F[i]	NP[i]
1	125	125.0818
2	63	63.9099
3	23	21.9724
4	4	5.0361
5	2	0.9999

Text N2

Parameters: a = 1.29279036316005 α = 0.709022979483008
DF =2 X^2 = 0.0418 $P(X^2)$ = 0.9793   C = 0.0002

X[i]	F[i]	NP[i]
1	84	84.0099
2	43	43.5294
3	29	28.1372
4	12	12.1252
5	5	5.1983

Text N3

Parameters: a = 1.02546219539642 α = 0.786768275976181
DF =2 X^2 = 4.2498 $P(X^2)$ = 0.1194   C = 0.0190

X[i]	F[i]	NP[i]
1	112	110.9676
2	60	64.8130
3	41	33.2316
4	10	11.3593
5	1	3.6286

Text N4

Parameters: a = 1.14298103265859 α = 0.726448497589307
DF =0 X^2 = 0.3066 $P(X^2)$ = 0.0000   C = 0.0012

X[i]	F[i]	NP[i]
1	127	128.3188
2	65	67.2492
3	51	38.4323|
4	11	19.9997|

CONCLUSION

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

Ball MJ (ed) 1992 *The Celtic languages*. Routledge, London.

Beibl Cymraeg Newydd 1985 *Y Beibl Cymraeg Newydd. Y Testament Newydd. Y Salmau.* Y Gymdeithas Feiblaidd Frytanaidd a Thramor, Llundain.

Best, K-H 1999 Quantitative Linguistik: Entwicklung, Stand und Perspektive. *Göttinger Beiträge zur Sprachwissenschaft* 2: 7-23.

Best, K-H (ed) 2001 *Häufigkeitsverteilungen in Texten*. Peust & Gutschmidt, Göttingen.

Drechsler, J 2001 Häufigkeitsverteilungen von Wortlängen in gälischen Texten. In: Best, K-H (ed.), *Häufigkeitsverteilungen in Texten*. Peust & Gutschmidt, Göttingen, 115-123.

King G 1993 *Modern Welsh: A comprehensive grammar.* Routledge, London.

Macaulay D 1992 *The Celtic languages.* Cambridge University Press, Cambridge.

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

Williams SJ 1980 *A Welsh grammar*. University of Wales Press, Cardiff.

Wilson, A 2001 Word Length Distributions in Classical Latin Verse. *Prague Bulletin of Mathematical Linguistics* 75: 69-84.

Wimmer G, Altmann G 1996 The theory of word length: Some results and generalizations. *Glottometrika* 15: 112-133.

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