Using rank-frequency and type-token statistics to compare morphological typology in the Celtic languages

Andrew Wilson* and Rosie Harvey

Department of Linguistics and English Language, Lancaster University, Lancaster, UK

Provide full correspondence details here including e-mail for the *corresponding author

Department of Linguistics and English Language, Lancaster University, County South Building, Lancaster LA1 4YL, UK

a.wilson@lancaster.ac.uk

Provide short biographical notes on all contributors here if the journal requires them.

N/A

Using rank-frequency and type-token statistics to compare morphological typology in the Celtic languages

Tristram (2009) applied Greenberg's (1960) synthetism index to compare three of the Celtic languages: Irish, Welsh, and Breton. She did not analyse samples of the other three Celtic languages – Scottish Gaelic, Manx, and Cornish. This paper expands on her work by comparing all six Celtic languages, including two periods of Irish (Early Modern and Present Day). The analysis is based on a random sample of 210 parallel psalm texts (30 for each language). However, Greenberg's synthetism index is problematic because there are no operational standards for counting morphemes within words. We therefore apply a newer typological indicator (B7; Popescu, Mačutek & Altmann, 2009), which is based solely on lexical rank-frequency statistics. Following Kelih (2010), we also explore whether type-token counts alone can provide similar information. The B7 indicator shows that both varieties of Irish, together with Welsh and Cornish, tend more towards synthetism, whereas Manx tends more towards analytism. Breton and Scottish Gaelic do not show a clear tendency in either direction. Rankings using type-token statistics vary considerably and do not tell the same story.

Keywords: typology; synthetism; rank-frequency statistics; type-token statistics; Celtic.

Introduction

In this paper, we explore the application of a newer generation of typological indicators, based on lexical rank-frequency and type-token statistics, to examine the structure of the Celtic group of languages on the morphological dimension of synthetism versus analytism. In doing so, we expand on earlier work by Tristram (2009), who examined just three of these languages using Greenberg's original (1960) synthetism index. We provide a comparison of all six Celtic languages, including two different periods in the development of Irish. We also place Celtic typology in the context of a number of other world languages, using the data provided by Popescu, Mačutek & Altmann (2009).

The Celtic languages are traditionally divided, on genetic grounds, into two groups: Q-Celtic (or Goidelic) and P-Celtic (or Brythonic). The names Q-Celtic and P-Celtic refer to the reflexes of an hypothesized Proto-Indo-European labiovelar (*k^w), which evolved to become a velar plosive in Q-Celtic and a bilabial plosive in P-Celtic – e.g., in the word for "four" (*k^w*etuor*), which, for instance, in modern Welsh became ***p****edwar* and in modern Manx ***k****iare*. Of the three attested Q-Celtic languages, Irish is the oldest, with Scottish Gaelic and Manx being later independent developments from it. There are similarly three attested P-Celtic languages, namely Welsh, Cornish, and Breton. All six of the modern Celtic languages are currently spoken, although the present-day versions of Cornish and Manx stem largely from revival movements in the 20th and 21st centuries; Cornish was effectively a dead language by the end of the 19th century (Mills, 2010), whilst the native-speaker continuity of Manx since the middle of the 1970s has been the subject of some debate (see Ager, 2009).[1]

Different approaches to language typology can be considered (Popescu & Altmann, 2008a). Some, traditionally, focus merely on classifying languages, either by tracing their genetic descent or, empirically, in terms of specific shared characteristics. Others, however, attempt to move away from mere classification, towards explanation. This latter course is characteristic of contemporary linguistic synergetics, which attempts to account for the nature of language itself by means of laws that are generally applicable across all known languages. A synergetic approach to typology attempts to derive and account for classifications of languages (and other groupings of linguistic entities, such as text types) in terms of variations in those laws. These variations are

---

[1] Further information about the Celtic languages can be obtained from the standard overviews by Ball & Fife (1993), MacAulay (1992), and Russell (1995).

distinct from the rules and exceptions of many traditional linguistic theories because they must be derivable mathematically from the general statement of the relevant language law: they cannot be simply a list of ad hoc additions or exceptions (Köhler, 1987).

Notwithstanding the shift of emphasis from mere classification to explanation, a continuing focus of interest has been in the morphological typology of languages. Whilst a number of different typologies have been proposed, one of the more enduring contrasts has been the distinction between synthetic and analytic languages. This has its classic statement in the work of Sapir (1921). A "purely" analytic language is one that has only one meaningful unit (morpheme) per word form, whilst a "purely" synthetic language is one that has more than one morpheme per word form. Sapir (1921) also added the term "polysynthetic" to describe those languages that have a particularly high number of morphemes per word. However, it is clear that most languages cannot be classified straightforwardly as synthetic or analytic (or even polysynthetic) but, rather, occupy a place on a continuum, such that they are "more analytic than synthetic", or vice versa. For example, the simple English sentence *She was running* tends somewhat towards analytism (in contrast to one of its possible Latin equivalents, *currebat*, where the subject's gender is derived only from the context and all other information about tense and aspect is encoded in the verb inflection); but it nevertheless still involves the inflection (or, rather, suppletion) of the verb *to be* as *was* and the inflection of the verb *to run* by the addition of the *-ing* suffix.

Greenberg (1960) took Sapir's work a step further by describing a number of quantitative indices for morphological typology, including a simple synthetism index, M/W, which measures the proportion of morphemes (M) to running words (W) in a representative text sample. Undertaking a preliminary analysis on eight languages, he

noted that "even cursory inspection of the indices set forth here shows that, if we define an analytic language as one with a synthetic index of 1.00-1.99, synthetic as 2.00-2.99, and polysynthetic as 3.00+, the results would conform to the usual nonquantitative judgments" (p. 194).

Tristram (2009) applied Greenberg's synthetism index to a number of different periods, dialects, and text types of three out of the six Celtic languages - namely Irish, Welsh, and Breton. For Old Irish, she obtained a mean synthetism index of 3.57 (SD 0.03), for Classical Irish 2.14 (SD 0.22), for modern Irish 1.94 (SD 0.04), and for modern Breton 1.68 (SD 0.17).[2] The results for Welsh she took from Parina's earlier (2006) study: for Old Welsh the synthetism index was 1.28 and for modern Welsh it was 1.35.[3] Using Greenberg's (1960) rule of thumb, this means that Old Irish can be categorized as "polysynthetic", Classical Irish as "synthetic", and Welsh and Breton as "analytic". Modern Irish stands very much on the border of "synthetic" and "analytic" (1.94, close to Greenberg's cut-off of 1.99). However, as we suggested earlier, it is probably preferable to avoid simplistic classifications and to talk rather of stronger or weaker tendencies along the analytic-synthetic continuum. Unfortunately, Tristram did not analyse any texts from the remaining three Celtic languages - Scottish Gaelic, Manx, and Cornish. Though she actually cites time pressures as her grounds for doing so, the omission of Scottish Gaelic is, to a certain extent, excusable for theoretical reasons, since Classical Irish functioned as the literary language of Gaelic Scotland until

---

2 We do not include all of Tristram's statistics here, only those which have a bearing on the present experiment.

3 No standard deviations are given for the Welsh statistics from Parina (2006), as she only analysed one text sample per period. Furthermore, two different figures are provided for Welsh, depending on whether word-initial mutations are included in the morpheme counts. Generally, it seems that Tristram does not consider the mutations as morphemic. It is worth noting that Parina (2006) also analysed a text that is closer in date to our own sample – John 1.1-7 from the Welsh Bible of 1588. This had a synthetism index of 1.42 (ignoring mutations). However, this result makes no difference to Tristram's rankings of languages.

at least the 18th century (MacCoinnich, 2008). Nevertheless, it would be interesting to know how Scottish Gaelic has evolved since then, in comparison with both Classical Irish and Present-Day Irish. As regards the other two languages (Manx and Cornish), Tristram (2009, p. 256) gives the misleading impression that they are no longer spoken: "Von diesen Sprachen werden heute noch vier gesprochen: das Walisische, Bretonische, Irische und Schottisch Gälische". This is untrue (see, e.g., Ager, 2009; Mills, 2010); sufficient textual material – both historical and contemporary – is, in fact, available for both Manx and Cornish.

An important point about Greenberg's synthetism index is that it requires a fairly deep knowledge of the grammar and semantics of a language in order to apply it reliably. Indeed, it is questionable whether it can be applied reliably at all, unless very clear operational criteria are laid down for the definition of words and morphemes. For instance, in a further study, Tristram (2010) asked 32 specialists in Old Irish to code a selection of eight orthographic word forms with, first, a word count and, second, a morpheme count. Fourteen people responded, to which she also added her own analyses, giving a total of fifteen suggested analyses. The suggested word counts for these single orthographic forms ranged from one to three (with unanimous agreement on only two out of the eight word forms), and the morpheme counts ranged from two to ten (with no unanimous agreement on any of the eight word forms). These results suggest not only that Greenberg's index is difficult, or at least unreliable, to apply, even when used by experts in a given language, but also that any published analyses that use it should be approached with a degree of caution. Whilst some of the broader conclusions that are based upon it may not be entirely misleading, it should certainly not be seen as the decisive "gold standard" for the quantitative study of language typology.

But this is not the only problem with Greenberg's indices. For one thing, even if the language constructs on which they are based can be defined operationally to an acceptable degree of inter-rater reliability, these analyses will still involve a fairly substantial manual effort, given that the automatic morphemic analysis of unrestricted text is not a straightforward task for natural language processing. Simpler alternative indicators that can measure the same theoretical constructs reliably are therefore to be welcomed.

Drawing on data from twenty languages, which included known extremes of synthetism and analytism, Popescu and Altmann (2008b) demonstrated that a rank-frequency-based indicator, known as *B*, could be used to help gauge the tendency of a language towards synthetism or analytism. In a book-length treatment of aspects of word frequencies (Popescu, Mačutek & Altmann, 2009), they later added a number of other possible rank-frequency-based indicators to this one. In the 2009 book, it should be noted that they also renamed the earlier indicator *B* as *B7*. This name will be used in the remainder of the present paper.

Derivation of indicator B7

Perhaps the simplest thing that can be done with a machine-readable text is to produce a frequency list of the word forms within it. When presented in descending order of frequency, with each word-type assigned a serial number from 1 (the most frequent) to V (the least frequent), this then becomes a ranked frequency list. Ranked frequency lists are most often used to gain an insight into the content of a text, based on the assumption that the words that are used most frequently - at least in so far as they are autosemantic (i.e. 'open-class' or 'content') words - are also those which are most central to the theme(s) of the text. However, setting aside the actual words that are

used, the purely formal properties of a ranked frequency list can also provide typological information for the linguist.

To almost any rank-frequency list, it is possible to fit the Zipfian (also known as the Zeta or power) function:

$$f_r = \frac{a}{r^b}$$

where r = the rank of a given word, f_r = the frequency of this word, and a and b are parameters to be estimated. However, the theoretical Zipfian curve typically does not fit naturally occurring data exactly and usually crosses the observed frequencies at some point within the range of the hapax legomena (words occurring only once). Popescu and Altmann (2008a) have observed that, if the curve crosses the observed frequencies early, so that most of the hapax legomena lie above the curve, this indicates a tendency to synthetism; however, if the curve crosses the observed frequencies late, so that most of the hapax legomena lie below the curve, then this indicates a tendency to analytism. This is because the more analytic languages will tend to use the same word-form multiple times to communicate a concept (it does not change to signal different grammatical relations - e.g. subject vs. object), whilst the more synthetic languages will use a greater number of unique forms (because the same lexeme changes its form to signal grammatical information). In the former case, the theoretical curve underestimates the number of hapax legomena, hence the predicted frequencies lie above the observed frequencies for longer; in the latter case, the theoretical curve overestimates the number of hapax legomena, hence the predicted frequencies fall below the observed frequencies at an earlier stage.

Popescu and Altmann (2008b) go on to point out that, in the former case, the mean of the Zipfian function must be smaller than the empirical mean of the observed data. The empirical mean of the data is given by:

$$M_e = \bar{r} = \frac{1}{N}\sum_{i=1}^{V} r\, f_r$$

where N is the number of tokens in the text, V is the number of types, r are the ranks, and f_r is the frequency of the word at rank r. The mean of the Zipfian function (M_f) is given by the same equation, but substituting the predicted frequency for the observed frequency.

The indicator B7 is then given by:

$$B7 = \frac{M_e - M_f}{M_e}$$

and it has the properties that, if B7 is greater than zero, the language tends more towards synthetism and, if B7 is less than zero, the language tends towards analytism.

As regards the other indicators proposed by Popescu, Mačutek & Altmann (2009), named B8 and B9, neither of these has, so far, a stated range of "analytic" versus "synthetic" tendencies. Furthermore, they are rank-order identical with the indicator B7, and all three can be derived mathematically from each other. For these reasons, we do not consider the indicators B8 and B9 any further in our experiment.

Materials and methods

The data chosen for this study consists of a parallel translation corpus of thirty psalms in all six of the Celtic languages, i.e., Welsh, Scottish Gaelic, Cornish, Breton, Manx, and Irish. For Irish, we were also able to consider two different time periods: Early Modern Irish and Present-Day Irish. (Early Modern Irish corresponds to what Tristram [2009] calls "Classical Irish".) In total, then, our sample contains 210 individual texts.

The thirty psalms were selected at random, ignoring texts numbered higher than 111, as these were not readily available in two of the language varieties analysed (i.e. Breton and Early Modern Irish). The psalms analysed were numbers 1, 2, 3, 6, 12, 17,

20, 21, 27, 28, 29, 32, 41, 42, 43, 51, 54, 56, 79, 81, 84, 85, 90, 91, 95, 96, 97, 98, 99, and 101.

Using bibliographic data, the various translations can be dated roughly as follows:

- Welsh ca. 1620
- Early Modern Irish ca. 1640-1685
- Manx 1765
- Scottish Gaelic 1794
- Breton 1893 (revised in 2004-2011)
- Cornish 1997
- Present-Day Irish 2004

The use of parallel translation texts should provide a fair comparison of the six Celtic languages, in so far as it largely allows us to rule out variation owing to content, text type, text length, etc. For Irish, as has already been noted, it has also been possible to process texts from the Early Modern and contemporary eras, giving some insight into diachronic developments over the course of roughly 350 years. However, in other cases, it was simply not possible to avoid, or to control for, differences in date whilst retaining a parallel translation corpus covering all of the Celtic languages. In a couple of cases, this was due simply to issues of access and copyright, but there were other reasons too. For instance, in contrast to most of the other languages, no Cornish translation of the psalms had been produced prior to the translation used here, which dates from as late as 1997. A Breton version was also quite late in coming; the text used here originates in the late 19th century, but it has been subjected to some modernization more recently, in 2004-2011.

In order to calculate the typological indicator B7, it is necessary to produce a rank-frequency list of the words in each text in each language. (We do not mix the texts into aggregated corpora because this introduces heterogeneity and affects the properties of the Zipfian distribution - see Altmann, 1992.) The production of the rank-frequency lists was achieved using a bespoke word-counting program written in Ruby 1.8. However, prior to producing the lists, it was first necessary to decide on how the texts were to be tokenized.

Tokenization is not a trivial problem for the Celtic languages, since hyphenation (especially in Manx) and the use of apostrophes (especially in Irish and Scottish Gaelic) are both commonplace. More generally, each of the languages has its own conventions governing the use of punctuation marks, division into orthographic word forms, and so on. However, orthographic conventions are merely that: conventions. Furthermore, most of them substantially post-date the texts being analysed here and have a prescriptive, rather than a merely descriptive, role. If we are looking for systematic relationships and differences between languages, we need to use tokenization criteria that are minimally theory-bound and independent of the post-hoc decisions of the codifiers of individual languages. However, even phonetic transcriptions of spoken texts would not help us here, since their tokenization into words for counting is still an artificial task governed by rules derived from grammars, dictionaries, and the orthographic conventions just mentioned. (A detailed consideration of tokenization issues, exemplified on the Slavic languages, can be found in Antić, Kelih and Grzybek, 2006.)

For the present experiment, it was therefore decided to tokenize all of the lists using broadly the same cross-linguistic conventions that have been applied for over a decade in the systematic study of word-length typology (Best, 2009). Thus, a word was

defined as being a string of printed characters with white space or punctuation at either end. Internal hyphens were treated as part of the word, so that a form such as the Manx *cur-jee* counts as one word and not two. In our study, apostrophes were always treated as part of a word, even when they occurred at the beginning or end of a string (but excluding cases which were obviously quotation marks). For instance, Scottish Gaelic distinguishes orthographically between, e.g., *m'*, *'m*, and *m*; these three forms represent different underlying lexemes, but their disambiguation or lemmatization lies beyond the scope of this experiment (and also contradicts its aim of a rapid, minimally knowledge-based assessment of typology). Further special cases covered by Best's conventions - i.e., numerals, abbreviations, and acronyms - did not occur in these data.

The rank-frequency lists were then read into the R environment for statistical computing (Ihaka & Gentleman, 1996) and the Zipfian distribution fitted using the *nls* function for nonlinear regression. Using the parameter estimates from this stage, the typological indicator B7 was calculated for each text. A measure of goodness of fit was also calculated for the Zipfian distribution. This is the determination coefficient R^2, which is given by:

$$R^2 = 1 - \frac{\sum_{r=1}^{V}(f - f_{pred})^2}{\sum_{r=1}^{V}(f - \bar{f})^2}$$

where f is the observed frequency and f_{pred} is the predicted frequency. A good fit is indicated by $R^2 > 0.9$; a fit is still acceptable when $R^2 > 0.8$ (but in psychology one allows even smaller R^2).

Results

Appendix 1 shows the values of the indicator B7 for each of the 210 individual texts. It also shows the type (N) and token (V) counts for each text, as well as the parameter estimates (A and b) and goodness of fit (R^2) for the Zipfian function.

Figure 1 summarizes the B7 values for each language in the form of a boxplot, together with the mean values. The means are also presented in Table 1, along with their standard deviations and 95% t-confidence intervals. The confidence intervals are depicted visually in Figure 2.

INSERT FIGURE 1 HERE

INSERT TABLE 1 HERE

INSERT FIGURE 2 HERE

It will be seen from Table 1 and Figure 2 that Early Modern Irish has the highest mean B7 (i.e., the highest degree of synthetism), followed (in descending order) by Present-Day Irish, Welsh, Cornish, Breton, Scottish Gaelic, and finally Manx.

The indicator B7 makes a clear statement about where the theoretical dividing line between analytic and synthetic tendencies falls; the more synthetic languages will show values of B7 in the positive range, whereas the more analytic languages will fall in the negative range. Both varieties of Irish, together with Welsh and Cornish, can thus be said to tend more towards synthetism, whilst Manx tends more towards analytism. Breton and Scottish Gaelic fall very much on the dividing line between synthetism and analytism, just into the negative (analytic) range of the indicator. However, it should be remembered that these figures are means, calculated from thirty individual texts per language. To check the degree of a tendency in one direction or the other, whilst taking sampling error into account, we therefore also calculated 95% t-confidence intervals for the means. If a confidence interval includes zero, then, on the basis of the current evidence, we cannot reject the hypothesis that the mean value of B7 for that language could, in fact, be zero. However, if the upper and lower ends of the interval both fall on the same side of zero, we may infer that the value of B7 also falls on that side of zero.

This means that, for Breton and Scottish Gaelic, we cannot reject the hypothesis that the value of B7 is zero; for the other languages, we do reject this hypothesis.

Although we had reason to question the reliability of Greenberg's original (1960) synthetism index, it is nevertheless of interest to compare these results with Tristram's (2009) calculations. Both her study and ours suggest that Early Modern Irish (which she refers to as "Classical Irish") tends the most towards synthetism. However, that is where the similarities end. Tristram reported that Present-Day Irish tended towards analytism, whereas our results suggest it tends towards synthetism. Tristram's results also suggested that both Welsh and Breton are analytic, whereas our results suggest that Welsh tends towards synthetism, whilst Breton cannot be classified clearly as either synthetic or analytic.

As mentioned earlier, however, it is perhaps more meaningful to consider all of the Celtic languages together on a single continuum from analytic to synthetic, and then to compare the ranking of the languages, rather than merely their binary classifications. Examining Figures 1 and 2 suggests that there may be three groupings within the data, with Manx, Scottish Gaelic, and Breton having a "low" B7; Cornish, Welsh, and Present-Day Irish having an "intermediate" B7; and Early Modern Irish having a "high" B7. (We have placed these descriptors within quote marks, because they are only relative, within the present set of languages.) To test for statistically significant differences between the individual pairs of languages, we calculated a set of pairwise Welch-Satterthwaite confidence intervals for the differences in means (Welch, 1947). Owing to the relatively large number of pairwise comparisons (21), we applied a Bonferroni correction to keep the family-wise Type I error rate below 5%. For pairwise confidence intervals, the Bonferroni formula is:

$$100 - \frac{(\alpha/100)}{m}$$

where α is the desired family-wise Type I error rate (as a percentage) and *m* is the number of pairwise intervals being calculated. For our data, this computed as:

$$100 - \frac{(5/100)}{21} = 99.9976$$

meaning that we needed to calculate a set of 99.9976% intervals, rather than the usual 95% intervals. These are shown in Table 2 and, graphically, in Figure 3. The test for each pair of languages is then whether the null hypothesis of no difference (i.e. zero) falls within the interval or not; if it does not, then we reject the null hypothesis for that pair, otherwise we cannot reject it on the basis of these data.

From Table 2 and Figure 3, we see that Early Modern Irish, which has the highest mean B7, is significantly different from all the other languages. Below that, Welsh, Cornish, and Present-Day Irish are also all significantly different from Manx, which has the lowest mean B7. The only other two significant differences are between Scottish Gaelic and Present-Day Irish (borderline, with zero at the very end of the interval), and between Scottish Gaelic and Welsh. The tests thus broadly support the informal inference from Figures 1 and 2 – namely, that there exist three levels of "low", "middle", and "high" ranking languages on the B7 indicator. However, the positions of Scottish Gaelic and Breton are ambiguous. The mean of Scottish Gaelic differs significantly from two members of the "middle" group (Welsh and Present-Day Irish), but not from the third member, Cornish. It is also not significantly different from Manx, hence it appears to occupy a borderline position between the "low" and "middle" ranks. Breton is not significantly different from any other language, apart from Early Modern Irish, and can thus not be classified straightforwardly as having either a "low" or a "middle" mean value of B7.

There were four languages in common between this experiment and Tristram's: Welsh, Breton, Early Modern Irish, and Present-Day Irish. Tristram reported that Welsh

ranked as the most analytic of these four, followed in order by Breton, Present-Day Irish, and Early Modern Irish. Our results suggested that Breton is the most analytic of the four, followed by Welsh, Present-Day Irish, and Early Modern Irish. However, as the difference in means between Breton and Welsh was not statistically significant, we cannot claim that our ranking is any different from the ranking obtained by Tristram; we are seeing broadly the same pattern.

Perhaps the most interesting finding in our study actually relates to two of the languages that Tristram (2009) did not consider. Manx shows an unusually high degree of analytism when compared to the two diachronic varieties of Irish (both quite strongly synthetic), as does Scottish Gaelic, though to a marginally lesser extent. Again, further research is required to properly support and account for this finding, which is suggestive of an important diachronic divergence within the Q-Celtic branch. It is clear that the developments in Manx and Scottish Gaelic are not of very recent origin, since the data processed here for both languages date from the 18th century. Unfortunately, however, the written tradition in Manx began only in the 17th century (Sebba, 1998), making investigations of its earlier history difficult. Similarly, a distinctive written literature in Scottish Gaelic does not begin to emerge until around this time (MacCoinnich, 2008, p. 330).

In terms of how the Celtic languages compare with a number of other world languages, we can compare our B7 values with those obtained by Popescu, Mačutek & Altmann (2009, p. 115). As all of our data were parallel psalm texts, and Popescu, Mačutek & Altmann used different quantities of texts drawn from different text types, a strict statistical comparison would not be appropriate. However, some raw comparisons of the means may nevertheless help to contextualize our results and suggest hypotheses for future comparative studies. That said, then, our most synthetic language, Early

Modern Irish, shows a value of B7 that is fairly close to that obtained for German (0.0738). This is lower (i.e. more analytic) than the value for Latin (0.1612) and considerably lower than the most synthetic of Popescu, Mačutek & Altmann's languages, Hungarian (at 0.6309). Our most analytic language, Manx, does not have a very close counterpart in Popescu, Mačutek & Altmann's list; however, it falls in the span between Bulgarian (0.0055) and Indonesian (-0.0501). Breton and Scottish Gaelic also fall inside this span. According to Kelih (2010), Bulgarian (along with Macedonian) has the most limited morphological case-flexion system of all the Slavic languages. However, this still means that our most analytic Celtic language ranks as considerably less analytic than other Indo-European languages such as Italian (-0.0744) and English (-0.1617). It is also far less analytic than the most analytic language in Popescu, Mačutek & Altmann's study – Hawaiian – which had a mean B7 of -1.2484.[4] The three remaining Celtic languages in the middle of our rankings – Present-Day Irish, Welsh, and Cornish – all show B7 values that are similar to that obtained for Russian (0.0349).

Type-token statistics

A final footnote to the present experiment seems appropriate. Kelih (2010) examined a parallel translation corpus of the twelve Slavic languages - a set of translations of the novel *How the steel was tempered* by N.A. Ostrovskij. His aim was to examine how far the raw type-token statistics might function as an indicator of language typology,

---

[4] The figure quoted by Popescu, Mačutek & Altmann (2009) for Hawaiian is actually -12.484, but this is clearly a misprint; recalculation from their table of individual text statistics shows that the correct value is -1.2484.

without any need to consider fitting the Zipfian (or any other) function. This makes some sense, because an increase in the number of hapax legomena, which is the key underlying factor for all of the rank-frequency-based indicators in Popescu, Mačutek & Altmann (2009), necessarily leads to a change in the overall type-token relationship. Kelih (2010) found that his ranking of token counts matched very closely the traditional sub-classification of the Slavic languages into the South, West, and East Slavic branches. However, the clustering was less evident in the type counts and hardly visible at all when the type and token counts were taken together.

Perhaps the simplest way to compare type and token counts is to make a scatterplot, as Kelih did. However, with the present data, we have thirty distinct texts for each language, which leads to a far more complex plot (Figure 4) than Kelih had for his data, where only one data point per language was plotted. It is, in fact, difficult to establish any clear pattern of clustering using this graph, except to note a general linear increase of type counts in relation to token counts. This is clearly because we are combining texts with quite different lengths on a single plot.

INSERT FIGURE 4 HERE

One alternative would be to produce a set of thirty scatterplots, one for each parallel psalm. However, these would be quite difficult to analyse systematically in terms of two-dimensional clustering patterns. A more straightforward alternative is to calculate a single summary measure – the token-type ratio (TTR) – for each language on each psalm:

$$TTR = \frac{N}{V}$$

where, as before, N = the number of tokens and V = the number of types. The individual TTR values are shown in Appendix 2. However, as it is well known that the TTR grows larger with increasing N, independently of V, we refrained from

undertaking a numerical comparison of the means for each language.[5] Instead, we took each psalm in turn and ranked the TTR values for the different languages. We then counted how often each language occupied each rank position across the thirty psalms (Table 3).

INSERT TABLE 3 AROUND HERE

It can be seen from Table 3 that Welsh nearly always has the highest TTR; it ranks first for 28 out of the thirty psalms. However, the pattern amongst the remaining languages is less constant. There are a couple of other noticeable trends: Breton occurs 23 times within the top three ranks, whilst Present-Day Irish and Manx both occur frequently within the bottom three ranks (29 and 27 times, respectively). This might seem, at first glance, to suggest a split along the conventional P- versus Q-Celtic lines; but, apart from Welsh, it has to be noticed that no language occupies a single rank for more than around half of the total number of psalms (16 being the next highest cell frequency). Early Modern Irish, in particular, is quite evenly dispersed across all rank positions, apart from the first; and Cornish and Scottish Gaelic also show quite even dispersion across a range of three or four different rank positions. Overall, then, these figures do not seem to provide any consistent support for a ranking of the languages in terms of the type-token relationship.

We should return finally, however, to Kelih's (2010) observation that, in his study, the relationship between type-token statistics and typology was only clearly reflected in the token counts. To test for this with our data, we undertook a similar

---

[5] In our case, Spearman's rho for the relationship between TTR and N was 0.409 ($p < 0.001$) but, for the relationship between TTR and V, rho was just -0.0098 ($p = 0.8869$).

analysis to the previous one, but using the token counts in place of the TTR values. In other words, we took each psalm in turn, ranked the token counts for the different languages, and then counted how often each language occupied each rank position across the thirty psalms. The results are shown in Table 4.

INSERT TABLE 4 AROUND HERE

As with the TTR, the token-count patterns are not conclusive, with only one cell of Table 4 showing a frequency higher than 15 out of thirty. On the whole, however, Present-Day Irish and Cornish tend to have the smallest token counts (falling, respectively, 28 and 24 times out of thirty in the bottom two ranks). Scottish Gaelic and Welsh have the highest counts (falling 25 and 26 times, respectively, in the top two ranks). Breton falls mostly in the middle of the range, occupying the middle three ranks for 25 times out of thirty. Early Modern Irish shows a similar pattern to Breton, falling 24 times out of thirty in the middle three ranks. Taken together, these rankings do not match either the results from our analysis of the B7 indicator or the traditional division into the P- and Q-Celtic sub-groups. It would appear, then, that Kelih's (2010) success in clustering his Slavic parallel texts according to the traditional sub-classification of languages does not replicate consistently for our Celtic parallel texts.

Discussion

This study has examined the application of a newer quantitative typological indicator, named B7, to the Celtic languages. This indicator is distinctive from other earlier typological indicators (such as Greenberg's [1960] synthetism index) in that it requires no morphological analysis but relies purely on lexical rank-frequency statistics.

In so far as comparative data have been available (which was the case for Early Modern and Present-Day Irish, Welsh, and Breton), the indicator B7 provided quite

similar results to Greenberg's synthetism index, as computed by Tristram (2009), when considered as rankings on a continuum; however, there was a greater degree of discrepancy when the binary classification of languages into analytic versus synthetic was considered. Although this broad similarity in rankings is pleasing, we do not consider the comparison as a gold-standard test of the indicator B7, for the reasons discussed in the Introduction.

The indicator B7 suggested not only that Irish had evolved diachronically from a greater to a lesser degree of synthetism but also that the overall synthetic versus analytic tendencies within Celtic were not straightforwardly linked to the ancestral Q- versus P-Celtic classification. This picture was not visible in Tristram's (2009) study, since she had not computed synthetism indices for Scottish Gaelic, Manx, or Cornish. In the present study, Manx (a Q-Celtic language) was the most analytic of all, whereas two of the other Q-Celtic languages (i.e. Early Modern and Present-Day Irish) both tended quite strongly towards synthetism; in contrast, Welsh (a P-Celtic language) proved to be more synthetic than both Manx and Scottish Gaelic (the other Q-Celtic language). Cornish was also more synthetic than Manx, but the difference with Scottish Gaelic was not statistically significant. Since the diachronic tendency in most Indo-European languages has been to move away from synthetism towards analytism, it seems unlikely that disparities in date lie behind these results: indeed, the Cornish texts are some of the most recently composed in this study, whereas the Manx texts only post-date the Early Modern Irish data by about a century. It thus seems, from these figures, that Manx especially (but also, to a slightly lesser extent, Scottish Gaelic) evolved earlier and more decisively towards analytism than has the present-day standard written variety of Irish on the island of Ireland. This view can be supported by comments in the non-quantitative literature. For instance, Broderick (1999, p. 77) has noted that, after the

fifteenth century, “Manx became more progressive in its development from a synthetic to an analytic type”. Similarly, when writing about the verbal system in a spoken dialect of Irish that is considered to be particularly close to Scottish Gaelic (Rathlin Island, Co. Antrim), Holmer (1942, p. 129) noted that “[t]he analytic conjugation, which is typical of Scottish Gaelic, is properly a simplification of the original synthetic conjugation, and the former is gaining ground also in Northern Irish, especially among the younger generation”. In other words, by the time Holmer was writing, Scottish Gaelic had already shifted in the direction of analytism, but some spoken dialects of Irish were only just starting to move in that direction through the influence of younger speakers. We cannot firmly identify any similar trends within the P-Celtic group, since none of the pairwise contrasts between Welsh, Cornish, and Breton were statistically significant.

Comparison with the B7 values obtained for a range of other languages by Popescu, Mačutek & Altmann (2009) showed that our most synthetic language, Early Modern Irish, demonstrates a similar degree of synthetism to modern German, which, in turn, is less synthetic (according to this indicator) than Latin. However, the maximal degree of shift towards analytism within Celtic (with Manx being the only language that has a mean B7 significantly on the negative side of zero) is considerably less than has occurred in some other modern Indo-European languages, such as English and Italian. It would be interesting to analyse some more recent Manx data using B7, to see whether there has been any further shift towards analytism since the eighteenth century.

Following on from Kelih’s (2010) consideration of the Slavic languages, the study also looked at whether simple type-token statistics could be used for typological classification, in place of calculating an indicator such as B7, which requires the prior fitting of the Zipfian function. However, the results from this analysis showed few

conclusive tendencies. Neither the summary Token-Type Ratio (TTR), nor the token count alone (which succeeded in Kelih's study), produced rankings where most of the languages fell in the same rank position on more than fifteen out of the thirty parallel psalm texts in our sample. Comparing broader tendencies (i.e. taking two or three contiguous rank positions at a time) showed some degree of patterning (especially for the token counts alone) but did not match with either the B7 results or the traditional classification of the languages into P- and Q-Celtic.

In summary, the present research, despite drawing only on psalm texts, and with severe limitations on controls for date, suggests that the typological indicator B7 may be of considerable value in investigating typological variation across the world's languages. Certainly, the story that it tells in relation to these Celtic texts ties in very well with the accounts of Celtic language change that can be found in more traditional qualitative scholarship. Further work, drawing on other text types and other dates in the history of Celtic, will surely tell interesting stories, not only about language evolution but also about text typology within individual languages. However, the future value of using raw type-token statistics for typological purposes seems unclear to us.

Data sources

Welsh

http://justus.anglican.org/resources/bcp/

Breton

http://bibl.monsite-orange.fr/

Manx

http://mannin.info/MHF/

Cornish

Courtesy of Keith Syed, Cornish Bible Project

Scottish Gaelic

Digital Archive of Scottish Gaelic, Text No. 152 (http://www.dasg.ac.uk)

Present-Day Irish

https://www.ireland.anglican.org/prayer-worship/book-of-common-prayer/2004-texts

Early Modern Irish

http://macmate.macace.net/~macfhionn@macace.net/index.html/Psailm.html

References

Ager, S. (2009). *A study of language death and revival with a particular focus on Manx Gaelic.* MA dissertation, Bangor University.

Altmann, G. (1992). Das Problem der Datenhomogenität. In B. Rieger (Ed.), *Glottometrika 9* (pp. 287–298). Bochum: Brockmeyer.

Antić, G., Kelih, E. & Grzybek, P. (2006). Zero-syllable words in determining word length. In P. Grzybek (Ed.), *Contributions to the science of text and language: Word length studies and related issues* (pp. 117-156). Dordrecht: Springer.

Ball, M.J. & Fife, J. (Ed.) (1993). *The Celtic languages.* London: Routledge.

Best, K.-H. (2009*). Quantitative Linguistik: Einführung und "Führer" zur Sprachstatistik des Deutschen sowie Principles for Word-Length Count.* Retrieved July 25, 2011, from http://wwwuser.gwdg.de/~kbest/einfueh.htm

Broderick, G. (1999). *Language death in the Isle of Man.* Berlin: Walter de Gruyter.

Greenberg, J.H. (1960). A quantitative approach to the morphological typology of languages. *International Journal of American Linguistics, 26,* 178-194.

Holmer, N.M. (1942). *The Irish language in Rathlin Island, Co. Antrim.* Dublin: Royal Irish Academy.

Ihaka, R. & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5,* 299-314.

Kelih, E. (2010). The type-token relationship in Slavic parallel texts. *Glottometrics, 20,* 1-11.

Köhler, R. (1987). Systems theoretical linguistics. *Theoretical Linguistics, 14,* 241-257.

MacAulay, D. (Ed.) (1992). *The Celtic languages.* Cambridge: Cambridge University Press.

MacCoinnich, A. (2008). Where and how was Gaelic written in late medieval and early modern Scotland? Orthographic practices and cultural identities. *Scottish Gaelic Studies, 24,* 309-356.

Mills, J. (2010). Genocide and ethnocide: The suppression of the Cornish language. In J. Partridge (Ed.), *Interfaces in language* (pp. 189-206). Newcastle upon Tyne: Cambridge Scholars Publishing.

Parina, E. (2006). О морфологическом типе валлийского языка. In А.Б. Кошелев (Ed.), Аналитизм в языках различных типов: сорок лет спустя. К 100-летию со дня рождения В. Н. Ярцевой (pp. 236-246). Moscow: Institute of Linguistics of the Russian Academy of Sciences.

Popescu, I.-I., Altmann, G. (2008a). Hapax legomena and language typology. *Journal of Quantitative Linguistics, 15*(4), 370-378.

Popescu, I.-I., Altmann, G. (2008b). Zipf's mean and language typology. *Glottometrics, 16,* 31-37.

Popescu, I.-I., Mačutek, J. & Altmann, G. (2009). *Aspects of word frequencies.* Lüdenscheid: RAM-Verlag.

Russell, P. (1995). *An introduction to the Celtic languages.* London: Longman.

Sapir, E. (1921). *Language: An introduction to the study of speech.* New York: Harcourt Brace and Company.

Sebba, M. (1998). *Orthography as practice and ideology: The case of Manx.* Centre for Language in Social Life Working Paper 102. Lancaster University.

Tristram, H.L.C. (2009). Wie weit sind die inselkeltischen Sprachen (und das Englische) analytisiert? In U. Hinrichs, N. Reiter & S. Tornow (Ed.), *Eurolinguistik: Entwicklung und Perspektiven* (pp. 255-280). Wiesbaden: Harrassowitz.

Tristram, H.L.C. (2010). Probleme bei der Quantifizierung morphologischer Komplexität im Altirischen. In: K. Stüber, T. Zehnder & D. Bachmann (Eds.), *Akten des 5. Deutschsprachigen Keltologensymposiums* (pp. 407−426). Vienna: Praesens Verlag.

Welch, B.L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika, 34*, 28-35.

Appendix 1

Results for the individual texts. Additional abbreviations: N = token count; V = type count; A and b = parameters of the Zipfian function; SG = Scottish Gaelic; EMI = Early Modern Irish; PDI = Present-Day Irish.

Psalm	Language	N	V	Empirical mean	Zipfian mean	B7	A	b	R^2
1	Breton	129	73	23.6047	23.809	-0.0087	11.7245	0.6245	0.9291
2	Breton	186	113	37.8763	38.4367	-0.0148	11.4586	0.5565	0.9005
3	Breton	135	77	25.3333	24.1596	0.0463	12.7007	0.6578	0.9652
6	Breton	149	89	29.8792	28.1659	0.0573	12.5441	0.6394	0.9691
12	Breton	160	95	31.2625	28.4174	0.091	14.6615	0.6857	0.9656
17	Breton	271	161	52.5793	49.6329	0.056	17.5116	0.6299	0.9353
20	Breton	154	93	31.3831	31.3266	0.0018	10.9134	0.5765	0.9433
21	Breton	222	126	39.7568	37.491	0.057	17.8786	0.6722	0.9739
27	Breton	295	158	47.6475	46.8462	0.0168	21.6387	0.6631	0.966
28	Breton	192	107	35.2917	37.2983	-0.0569	11.3096	0.5344	0.9381
29	Breton	175	74	19.7486	20.701	-0.0482	22.8481	0.7635	0.8747
32	Breton	237	131	41.0717	39.3855	0.0411	18.2117	0.6613	0.9816
41	Breton	237	137	43.8354	41.5029	0.0532	17.5788	0.6525	0.9421
42	Breton	259	122	35.1969	38.2522	-0.0868	19.5305	0.6289	0.9399
43	Breton	112	70	24.4554	23.4862	0.0396	9.4863	0.5991	0.9563
51	Breton	329	180	55.6657	56.1374	-0.0085	19.7448	0.6153	0.9476
54	Breton	131	83	28.9008	27.3838	0.0525	10.4197	0.6041	0.9203
56	Breton	246	116	32.4106	36.0017	-0.1108	19.9676	0.6407	0.9109

79	Breton	245	146	48.3918	42.2194	0.1276	18.5136	0.6873	0.9689
81	Breton	267	153	48.9326	50.2782	-0.0275	15.2755	0.5758	0.9112
84	Breton	222	113	34.027	36.727	-0.0793	15.8549	0.6002	0.9285
85	Breton	218	111	33.5229	35.1608	-0.0489	16.71	0.6249	0.9438
90	Breton	293	161	50.1399	50.1517	-0.0002	18.6771	0.621	0.9514
91	Breton	256	143	44.6484	42.7898	0.0416	19.0799	0.6606	0.9792
95	Breton	168	103	35.744	37.3533	-0.045	8.9565	0.4943	0.868
96	Breton	223	96	26.5695	29.4724	-0.1093	20.5597	0.662	0.9281
97	Breton	218	107	30.0413	29.8341	0.0069	23.0625	0.7358	0.9407
98	Breton	163	83	25.6135	27.6119	-0.078	13.1357	0.5959	0.9295
99	Breton	169	89	27.8047	29.6589	-0.0667	12.9462	0.5898	0.9328
101	Breton	159	91	29.4025	28.0341	0.0465	14.2312	0.6626	0.9418
1	Cornish	116	67	22.2931	21.6833	0.0274	11.0888	0.6389	0.9711
2	Cornish	178	116	41.0337	37.6351	0.0828	11.8399	0.6005	0.924
3	Cornish	124	70	23.0726	21.9689	0.0478	12.3175	0.6652	0.9725
6	Cornish	149	88	28.698	25.5419	0.11	15.2838	0.7173	0.9407
12	Cornish	139	91	32.3165	28.0932	0.1307	11.8606	0.6607	0.9394
17	Cornish	283	160	51.3852	50.0227	0.0265	17.7254	0.6182	0.9605
20	Cornish	143	82	26.9441	28.1485	-0.0447	10.5564	0.5648	0.9086
21	Cornish	210	117	36.9524	33.2225	0.1009	18.9781	0.7151	0.9767
27	Cornish	284	151	44.8099	44.3078	0.0112	22.1933	0.6738	0.9473
28	Cornish	173	104	35.4682	37.3333	-0.0526	9.4867	0.5049	0.8971
29	Cornish	166	71	19.1566	19.2205	-0.0033	23.2573	0.7951	0.8869
32	Cornish	217	131	44.2212	44.3349	-0.0026	12.2774	0.5546	0.9239
41	Cornish	217	128	42.1613	41.5036	0.0156	14.0111	0.5961	0.9602
42	Cornish	245	126	38.7796	38.6326	0.0038	18.3553	0.6468	0.9725

43	Cornish	107	71	25.6916	24.0604	0.0635	8.6647	0.5878	0.9383
51	Cornish	274	158	50.9051	48.2274	0.0526	18.3323	0.6391	0.9431
54	Cornish	100	69	26.13	24.4472	0.0644	7.2608	0.5421	0.8886
56	Cornish	213	107	32.3944	33.9801	-0.0489	16.5119	0.6247	0.9459
79	Cornish	227	134	43.3524	42.2683	0.025	15.5716	0.6187	0.9228
81	Cornish	244	145	47.1967	44.6118	0.0548	16.815	0.6365	0.9572
84	Cornish	200	117	38.14	35.7851	0.0617	15.5483	0.653	0.9738
85	Cornish	186	100	31.1505	29.6653	0.0477	17.1023	0.6893	0.9799
90	Cornish	273	155	49.5934	49.3554	0.0048	16.7433	0.6036	0.9437
91	Cornish	240	145	48.2083	46.0808	0.0441	15.0508	0.6083	0.9691
95	Cornish	167	109	39.4611	39.9648	-0.0128	8.1535	0.4797	0.9203
96	Cornish	203	91	25.7438	26.0932	-0.0136	21.5342	0.725	0.9823
97	Cornish	178	101	31.4157	27.9735	0.1096	18.7429	0.7454	0.9653
98	Cornish	143	80	25.9091	24.2658	0.0634	14.0359	0.6861	0.9662
99	Cornish	159	74	21.7358	24.0197	-0.1051	14.7507	0.6282	0.9371
101	Cornish	147	86	28.4898	25.4314	0.1073	14.3295	0.703	0.9442
1	EMI	120	78	27.4917	25.451	0.0742	10.1514	0.6193	0.9213
2	EMI	185	127	43.8731	39.0871	0.1091	14.2126	0.643	0.9647
3	EMI	116	90	31.9286	29.7209	0.0691	10.3711	0.5981	0.9619
6	EMI	163	95	34.4028	33.9978	0.0118	8.3477	0.5126	0.9008
12	EMI	144	93	33.6357	31.7986	0.0546	9.3176	0.5615	0.9522
17	EMI	264	182	63.6117	53.5673	0.1579	17.7982	0.6623	0.9522
20	EMI	150	88	30.7111	27.4659	0.1057	11.7503	0.653	0.94
21	EMI	214	135	44.8157	35.6894	0.2036	20.3396	0.7607	0.9653
27	EMI	298	167	51.6782	46.6207	0.0979	23.5221	0.7074	0.943
28	EMI	183	109	36.0865	34.0214	0.0572	14.2824	0.6393	0.9746

29	EMI	158	75	21.0188	20.8614	0.0075	20.5082	0.767	0.9397
32	EMI	219	140	46.7424	41.8068	0.1056	16.81	0.6634	0.9751
41	EMI	237	147	49.8584	45.5977	0.0855	15.3294	0.6289	0.9621
42	EMI	269	140	44.94	47.1879	-0.05	13.9402	0.5557	0.9244
43	EMI	128	73	25.8462	25.7246	0.0047	8.4681	0.5444	0.9574
51	EMI	308	166	51.4812	45.9588	0.1073	24.0275	0.714	0.9243
54	EMI	107	76	30.402	29.3711	0.0339	5.2787	0.4353	0.9106
56	EMI	230	115	36.3785	33.8679	0.069	18.0838	0.6863	0.9769
79	EMI	283	154	51.9212	44.1909	0.1489	18.6139	0.6905	0.9001
81	EMI	253	154	51.172	44.1684	0.1369	18.9872	0.6909	0.9752
84	EMI	227	126	40.9437	34.3537	0.161	19.631	0.7421	0.9709
85	EMI	197	123	42.0105	34.1684	0.1867	17.1955	0.7292	0.9552
90	EMI	278	177	59.2384	51.5155	0.1304	19.5207	0.6724	0.9264
91	EMI	271	149	48.812	45.8052	0.0616	16.9061	0.6359	0.9543
95	EMI	191	116	41.2543	36.7379	0.1095	12.2866	0.6227	0.9296
96	EMI	213	111	32.977	31.9019	0.0326	20.4953	0.7088	0.9659
97	EMI	192	122	40.145	34.9525	0.1293	17.5312	0.7052	0.9421
98	EMI	137	93	30.0127	27.0464	0.0988	15.7475	0.7114	0.9574
99	EMI	187	95	31.1779	30.4035	0.0248	13.0273	0.6247	0.9599
101	EMI	152	102	36.7582	34.5933	0.0589	9.8466	0.5646	0.9204
1	Manx	120	83	28.9248	28.6715	0.0088	9.416	0.5575	0.9146
2	Manx	197	141	49.0541	47.7106	0.0274	11.8863	0.5516	0.9348
3	Manx	140	82	28.5267	28.7141	-0.0066	9.0552	0.5439	0.8673
6	Manx	144	100	33.6325	33.1638	0.0139	11.7001	0.5877	0.9501
12	Manx	140	111	38.3352	38.7999	-0.0121	10.1871	0.5298	0.9077
17	Manx	273	182	56.6163	58.1746	-0.0275	18.5796	0.5936	0.9248

20	Manx	135	94	30.2061	30.0159	0.0063	13.4771	0.6275	0.9451
21	Manx	217	136	42.0729	43.6714	-0.038	16.234	0.602	0.9273
27	Manx	289	167	47.8537	48.0645	-0.0044	25.8592	0.684	0.9737
28	Manx	185	119	38.296	41.7763	-0.0909	12.1116	0.5222	0.9266
29	Manx	160	91	26.2162	25.6813	0.0204	20.5093	0.7382	0.9567
32	Manx	229	141	46.214	47.0238	-0.0175	13.7863	0.5654	0.9441
41	Manx	233	136	41.4008	43.0355	-0.0395	17.8125	0.6152	0.9564
42	Manx	250	154	46.8904	48.9881	-0.0447	18.2587	0.6048	0.9254
43	Manx	117	76	24.7941	24.5805	0.0086	12.1361	0.6298	0.9544
51	Manx	293	168	51.4121	52.231	-0.0159	19.6208	0.6208	0.9646
54	Manx	102	76	28.3636	25.6501	0.0957	8.4941	0.5872	0.9455
56	Manx	214	135	42.747	43.7167	-0.0227	15.7466	0.5947	0.9636
79	Manx	241	160	49.2199	48.3024	0.0186	20.2323	0.6477	0.9457
81	Manx	250	163	50.8783	53.4481	-0.0505	16.8501	0.5752	0.9185
84	Manx	213	139	45.5277	46.8552	-0.0292	13.1654	0.5559	0.8966
85	Manx	190	118	37.9766	38.5278	-0.0145	14.3675	0.5937	0.9305
90	Manx	281	172	50.2399	54.9083	-0.0929	20.3717	0.597	0.9232
91	Manx	250	154	48.2708	50.5495	-0.0472	15.9428	0.5765	0.9057
95	Manx	173	118	38.6049	38.3406	0.0068	13.8312	0.5983	0.9426
96	Manx	217	112	32.1351	36.6628	-0.1409	18.4786	0.594	0.9002
97	Manx	200	135	45.4163	45.3672	0.0011	12.5914	0.5601	0.9033
98	Manx	158	97	30.4611	31.3586	-0.0295	14.0262	0.6139	0.9609
99	Manx	163	103	34.5287	34.8352	-0.0089	11.4491	0.5668	0.9444
101	Manx	153	97	35.0667	35.6832	-0.0176	7.8728	0.4812	0.8716
1	PDI	121	75	25.9083	23.7424	0.0836	11.0852	0.6514	0.9663
2	PDI	173	124	44.1351	39.7628	0.0991	12.1993	0.6078	0.9537

3	PDI	113	80	29.6552	26.9783	0.0903	8.7171	0.5846	0.9501
6	PDI	137	104	36.5153	35.6416	0.0239	10.1116	0.5531	0.9468
12	PDI	132	98	36.0347	33.7996	0.062	8.9931	0.5498	0.9521
17	PDI	256	165	54.678	47.3834	0.1334	19.5426	0.6863	0.9641
20	PDI	128	94	32.1067	26.8584	0.1635	14.6181	0.7254	0.9533
21	PDI	201	128	41.1542	38.5484	0.0633	16.9515	0.6612	0.8927
27	PDI	280	158	47.2718	48.4723	-0.0254	20.5195	0.6349	0.9162
28	PDI	177	111	37.3224	35.1309	0.0587	13.4509	0.6257	0.9632
29	PDI	157	78	23.0506	22.7778	0.0118	17.6599	0.7219	0.9425
32	PDI	200	142	50.0091	46.9796	0.0606	12.3454	0.5727	0.9332
41	PDI	218	147	49.9536	46.7582	0.064	14.4916	0.6069	0.9461
42	PDI	235	134	41.0743	42.5773	-0.0366	17.9662	0.6123	0.9774
43	PDI	103	71	22.8672	24.3938	-0.0668	10.4703	0.5734	0.8592
51	PDI	277	158	46.2208	48.0826	-0.0403	21.7393	0.6416	0.9185
54	PDI	96	77	29.8598	29.4647	0.0132	5.7234	0.4467	0.8794
56	PDI	195	125	40.2304	39.4378	0.0197	15.993	0.6221	0.9738
79	PDI	229	178	59.424	54.4753	0.0833	17.6451	0.6316	0.9268
81	PDI	225	161	55.2292	52.1394	0.0559	14.1644	0.5869	0.9579
84	PDI	195	131	42.5595	40.4282	0.0501	16.1574	0.639	0.9703
85	PDI	170	118	38.3198	35.2929	0.079	16.3646	0.6717	0.9213
90	PDI	274	169	55.9964	54.0244	0.0352	15.9256	0.5965	0.9318
91	PDI	243	151	46.4797	44.5739	0.041	20.3194	0.6689	0.942
95	PDI	160	116	39.1937	38.8601	0.0085	11.9851	0.5701	0.9309
96	PDI	195	106	31.3286	34.567	-0.1034	15.8832	0.6006	0.8739
97	PDI	184	121	40.7708	37.8354	0.072	14.102	0.6318	0.9037
98	PDI	130	83	27.8029	26.0659	0.0625	12.306	0.6514	0.9465

99	PDI	150	94	27.9947	30.3782	-0.0851	15.1899	0.6162	0.9237
101	PDI	142	97	34.1118	32.9433	0.0343	10.1405	0.5659	0.9159
1	SG	133	87	29.6014	29.1797	0.0142	10.6451	0.5848	0.9539
2	SG	222	136	47.7129	45.5859	0.0446	11.7351	0.5622	0.9242
3	SG	131	92	33.0567	33.6458	-0.0178	7.8505	0.4904	0.8742
6	SG	166	107	35.5337	35.6194	-0.0024	12.0152	0.5804	0.9156
12	SG	179	115	40.0279	39.6622	0.0091	10.3842	0.5418	0.8829
17	SG	331	202	65.5616	62.4483	0.0475	19.1459	0.618	0.92
20	SG	165	97	31.5629	32.8242	-0.04	11.6063	0.5696	0.8805
21	SG	247	132	41.3882	40.6993	0.0166	17.3123	0.6394	0.9645
27	SG	335	183	57.7757	56.0704	0.0295	19.864	0.6294	0.9421
28	SG	223	126	41.1081	42.9998	-0.046	12.692	0.5481	0.914
29	SG	185	93	28.8857	29.942	-0.0366	14.2945	0.6204	0.9139
32	SG	243	146	48.5436	49.3443	-0.0165	12.9448	0.5513	0.8963
41	SG	262	145	45.0736	43.9741	0.0244	18.8046	0.6487	0.9414
42	SG	292	161	48.637	49.2926	-0.0135	20.4672	0.6357	0.9457
43	SG	136	81	26.6667	27.0127	-0.013	11.2262	0.5951	0.9412
51	SG	313	177	54.5906	53.6156	0.0179	20.8221	0.6403	0.9672
54	SG	110	79	31.4667	30.1031	0.0433	5.5551	0.4504	0.881
56	SG	249	136	43.132	46.7051	-0.0828	13.4581	0.5387	0.9154
79	SG	291	167	53.3808	52.7373	0.0121	17.2506	0.6077	0.8514
81	SG	304	163	53.7582	53.5838	0.0032	14.8334	0.5729	0.9239
84	SG	235	139	44.15	47.0593	-0.0659	14.3189	0.5517	0.9442
85	SG	214	116	38.8731	37.5297	0.0346	13.4624	0.6031	0.9461
90	SG	346	193	60.608	63.8534	-0.0535	17.1988	0.5608	0.9391
91	SG	277	164	53.1053	54.1176	-0.0191	15.355	0.5691	0.9143

95	SG	205	118	39.9792	37.4329	0.0637	13.5585	0.6202	0.9616
96	SG	259	126	37.8538	40.7864	-0.0775	17.0563	0.5984	0.9392
97	SG	221	131	43.8356	43.1113	0.0165	13.3919	0.5812	0.9268
98	SG	180	96	31.4061	31.3962	0.0003	12.4896	0.6036	0.9447
99	SG	174	96	30.3807	31.8112	-0.0471	13.0049	0.5909	0.9289
101	SG	150	113	40.3636	40.0049	0.0089	9.3643	0.5158	0.9545
1	Welsh	143	73	24.5372	24.2163	0.0131	10.4558	0.6076	0.9259
2	Welsh	209	117	42.052	37.8067	0.101	11.5527	0.6038	0.9494
3	Welsh	141	71	25.5575	26.2295	-0.0263	7.2747	0.493	0.9155
6	Welsh	178	84	27.9781	26.0409	0.0692	12.7139	0.6625	0.926
12	Welsh	179	89	32.2652	28.4983	0.1167	10.5232	0.6284	0.9494
17	Welsh	333	165	56.3203	52.8944	0.0608	14.7388	0.595	0.8993
20	Welsh	167	79	27.2969	26.0932	0.0441	10.4028	0.6064	0.9337
21	Welsh	237	120	39.4229	38.3144	0.0281	14.1349	0.6135	0.9413
27	Welsh	321	150	45.4286	46.1904	-0.0168	19.3857	0.6341	0.9535
28	Welsh	222	113	39.6271	39.3573	0.0068	10.021	0.5326	0.9144
29	Welsh	175	72	20.2102	20.3132	-0.0051	19.9926	0.7589	0.946
32	Welsh	241	133	46.86	42.5762	0.0914	12.8767	0.6059	0.9082
41	Welsh	258	131	43.2569	42.9338	0.0075	13.5891	0.5851	0.9347
42	Welsh	303	130	41.8468	42.1961	-0.0083	15.0189	0.5944	0.9656
43	Welsh	141	71	26.6893	25.951	0.0277	6.7763	0.5054	0.9098
51	Welsh	320	162	51.6209	51.1832	0.0085	17.1309	0.6085	0.9085
54	Welsh	105	65	24.1771	23.4394	0.0305	7.0092	0.5265	0.9084
56	Welsh	250	109	35.3744	34.9436	0.0122	14.2662	0.615	0.9426
79	Welsh	281	145	49.2576	43.8355	0.1101	16.1072	0.6514	0.9498
81	Welsh	273	145	50.52	49.1284	0.0275	11.8244	0.5492	0.9253

84	Welsh	260	123	41.8103	39.6058	0.0527	13.1339	0.6045	0.9159
85	Welsh	197	111	39.2824	31.4756	0.1987	14.9019	0.7197	0.9261
90	Welsh	352	156	48.1642	43.6084	0.0946	22.792	0.7101	0.9503
91	Welsh	285	139	44.0082	41.962	0.0465	17.975	0.6547	0.9674
95	Welsh	192	107	38.5125	36.2953	0.0576	10.0044	0.5619	0.9259
96	Welsh	253	100	30.6103	33.1476	-0.0829	14.2133	0.5882	0.9074
97	Welsh	219	100	30.2391	28.3064	0.0639	18.8872	0.7285	0.9576
98	Welsh	165	82	28.9385	27.9046	0.0357	9.5572	0.5738	0.9594
99	Welsh	176	86	28.34	29.7445	-0.0496	10.5008	0.5541	0.9086
101	Welsh	176	90	31.9085	32.3559	-0.014	8.4681	0.5104	0.8996

Appendix 2

Token-Type Ratio (TTR) for each psalm in each language. Abbreviated labels: ScotG = Scottish Gaelic, PDI = Present-Day Irish, EMI = Early Modern Irish.

Psalm	Language	TTR
1	Breton	1.7671
2	Breton	1.646
3	Breton	1.7532
6	Breton	1.6742
12	Breton	1.6842
17	Breton	1.6832
20	Breton	1.6559
21	Breton	1.7619
27	Breton	1.8671
28	Breton	1.7944
29	Breton	2.3649
32	Breton	1.8092
41	Breton	1.7299
42	Breton	2.123
43	Breton	1.6
51	Breton	1.8278
54	Breton	1.5783
56	Breton	2.1207
79	Breton	1.6781
81	Breton	1.7451

84	Breton	1.9646
85	Breton	1.964
90	Breton	1.8199
91	Breton	1.7902
95	Breton	1.6311
96	Breton	2.3229
97	Breton	2.0374
98	Breton	1.9639
99	Breton	1.8989
101	Breton	1.7473
1	Cornish	1.7313
2	Cornish	1.5345
3	Cornish	1.7714
6	Cornish	1.6932
12	Cornish	1.5275
17	Cornish	1.7688
20	Cornish	1.7439
21	Cornish	1.7949
27	Cornish	1.8808
28	Cornish	1.6635
29	Cornish	2.338
32	Cornish	1.6565
41	Cornish	1.6953
42	Cornish	1.9444

43	Cornish	1.507
51	Cornish	1.7342
54	Cornish	1.4493
56	Cornish	1.9907
79	Cornish	1.694
81	Cornish	1.6828
84	Cornish	1.7094
85	Cornish	1.86
90	Cornish	1.7613
91	Cornish	1.6552
95	Cornish	1.5321
96	Cornish	2.2308
97	Cornish	1.7624
98	Cornish	1.7875
99	Cornish	2.1486
101	Cornish	1.7093
1	EMI	1.5385
2	EMI	1.4567
3	EMI	1.2889
6	EMI	1.7158
12	EMI	1.5484
17	EMI	1.4505
20	EMI	1.7045
21	EMI	1.5852

27	EMI	1.7844
28	EMI	1.6789
29	EMI	2.1067
32	EMI	1.5643
41	EMI	1.6122
42	EMI	1.9214
43	EMI	1.7534
51	EMI	1.8554
54	EMI	1.4079
56	EMI	2
79	EMI	1.8377
81	EMI	1.6429
84	EMI	1.8016
85	EMI	1.6016
90	EMI	1.5706
91	EMI	1.8188
95	EMI	1.6466
96	EMI	1.9189
97	EMI	1.5738
98	EMI	1.4731
99	EMI	1.9684
101	EMI	1.4902
1	Manx	1.4458
2	Manx	1.3972

3	Manx	1.7073
6	Manx	1.44
12	Manx	1.2613
17	Manx	1.5
20	Manx	1.4362
21	Manx	1.5956
27	Manx	1.7305
28	Manx	1.5546
29	Manx	1.7582
32	Manx	1.6241
41	Manx	1.7132
42	Manx	1.6234
43	Manx	1.5395
51	Manx	1.744
54	Manx	1.3421
56	Manx	1.5852
79	Manx	1.5063
81	Manx	1.5337
84	Manx	1.5324
85	Manx	1.6102
90	Manx	1.6337
91	Manx	1.6234
95	Manx	1.4661
96	Manx	1.9375

97	Manx	1.4815
98	Manx	1.6289
99	Manx	1.5825
101	Manx	1.5773
1	PDI	1.6133
2	PDI	1.3952
3	PDI	1.4125
6	PDI	1.3173
12	PDI	1.3469
17	PDI	1.5515
20	PDI	1.3617
21	PDI	1.5703
27	PDI	1.7722
28	PDI	1.5946
29	PDI	2.0128
32	PDI	1.4085
41	PDI	1.483
42	PDI	1.7537
43	PDI	1.4507
51	PDI	1.7532
54	PDI	1.2468
56	PDI	1.56
79	PDI	1.2865
81	PDI	1.3975

84	PDI	1.4885
85	PDI	1.4407
90	PDI	1.6213
91	PDI	1.6093
95	PDI	1.3793
96	PDI	1.8396
97	PDI	1.5207
98	PDI	1.5663
99	PDI	1.5957
101	PDI	1.4639
1	ScotG	1.5287
2	ScotG	1.6324
3	ScotG	1.4239
6	ScotG	1.5514
12	ScotG	1.5565
17	ScotG	1.6386
20	ScotG	1.701
21	ScotG	1.8712
27	ScotG	1.8306
28	ScotG	1.7698
29	ScotG	1.9892
32	ScotG	1.6644
41	ScotG	1.8069
42	ScotG	1.8137

43	ScotG	1.679
51	ScotG	1.7684
54	ScotG	1.3924
56	ScotG	1.8309
79	ScotG	1.7425
81	ScotG	1.865
84	ScotG	1.6906
85	ScotG	1.8448
90	ScotG	1.7927
91	ScotG	1.689
95	ScotG	1.7373
96	ScotG	2.0556
97	ScotG	1.687
98	ScotG	1.875
99	ScotG	1.8125
101	ScotG	1.3274
1	Welsh	1.9589
2	Welsh	1.7863
3	Welsh	1.9859
6	Welsh	2.119
12	Welsh	2.0112
17	Welsh	2.0182
20	Welsh	2.1139
21	Welsh	1.975

27	Welsh	2.14
28	Welsh	1.9646
29	Welsh	2.4306
32	Welsh	1.812
41	Welsh	1.9695
42	Welsh	2.3308
43	Welsh	1.9859
51	Welsh	1.9753
54	Welsh	1.6154
56	Welsh	2.2936
79	Welsh	1.9379
81	Welsh	1.8828
84	Welsh	2.1138
85	Welsh	1.7748
90	Welsh	2.2564
91	Welsh	2.0504
95	Welsh	1.7944
96	Welsh	2.53
97	Welsh	2.19
98	Welsh	2.0122
99	Welsh	2.0465
101	Welsh	1.9556

Table 1. Mean values, standard deviations, and 95% CIs for indicator B7.

Language	Mean B7	SD	95% Lower CI	95% Upper CI
Early Modern Irish	0.0861	0.0580	0.0645	0.1078
Present-Day Irish	0.0370	0.0607	0.0144	0.0597
Welsh	0.0367	0.0559	0.0159	0.0576
Cornish	0.0324	0.0550	0.0119	0.0529
Breton	-0.0018	0.0613	-0.0247	0.0211
Scottish Gaelic	-0.0048	0.0424	-0.0191	0.0094
Manx	-0.0181	0.0381	-0.0339	-0.0023

Table 2. 99.7619% Welch-Satterthwaite confidence intervals for differences in mean B7 between pairs of languages. Abbreviated labels: ScotG = Scottish Gaelic, PDI = Present-Day Irish, EMI = Early Modern Irish.

Comparison	Difference in means	99.7619% Lower CI	99.7619% Upper CI
EMI-Breton	0.0879	0.039	0.1369
PDI-Manx	0.0552	0.012	0.0983
Welsh-Manx	0.0548	0.014	0.0957
EMI-Cornish	0.0537	0.0074	0.1001
Welsh-ScotG	0.0416	0.0021	0.0811
PDI-Breton	0.0389	-0.0112	0.0889
Welsh-Breton	0.0385	-0.0096	0.0867
Cornish-Breton	0.0342	-0.0136	0.082
ScotG-Manx	0.0133	-0.0198	0.0464
PDI-Cornish	0.0046	-0.0429	0.0522
Welsh-Cornish	0.0043	-0.0412	0.0498
Welsh-PDI	-0.0003	-0.0482	0.0476
ScotG-Breton	-0.003	-0.0453	0.0392
Manx-Breton	-0.0163	-0.0598	0.0272
ScotG-Cornish	-0.0372	-0.0763	0.0018
ScotG-PDI	-0.0419	-0.0838	0
PDI-EMI	-0.0491	-0.0978	-0.0004
Welsh-EMI	-0.0494	-0.0962	-0.0026
Manx-Cornish	-0.0505	-0.0909	-0.0101
ScotG-EMI	-0.091	-0.1316	-0.0504
Manx-EMI	-0.1042	-0.1461	-0.0624

Table 3. Token-type ratios ranked for each psalm: numbers of psalms for which each language occupies each rank position. (Rank 1 = highest TTR, Rank 7 = lowest TTR.) Abbreviated labels: ScotG = Scottish Gaelic, PDI = Present-Day Irish, EMI = Early Modern Irish.

Rank	Breton	Cornish	EMI	Manx	PDI	ScotG	Welsh
1	1	1	0	0	0	0	28
2	15	5	5	0	0	4	1
3	7	9	5	0	0	9	0
4	5	8	5	3	1	7	1
5	2	5	6	7	3	7	0
6	0	1	5	12	10	2	0
7	0	1	4	8	16	1	0

Table 4. Token counts ranked for each psalm: numbers of psalms for which each language occupies each rank position. (Rank 1 = highest token count, Rank 7 = lowest token count.) Abbreviated labels: ScotG = Scottish Gaelic, PDI = Present-Day Irish, EMI = Early Modern Irish.

Rank	Breton	Cornish	EMI	Manx	PDI	ScotG	Welsh
1	2	0	1	0	1	14	12
2	3	0	1	1	0	11	14
3	14	1	8	2	0	3	2
4	8	1	8	10	1	1	1
5	3	4	8	13	0	1	1
6	0	15	4	4	7	0	0
7	0	9	0	0	21	0	0

Figure 1. Boxplot of B7 values. Means are shown by the "+" symbol. Abbreviated labels: ScotG = Scottish Gaelic, PDI = Present-Day Irish, EMI = Early Modern Irish.

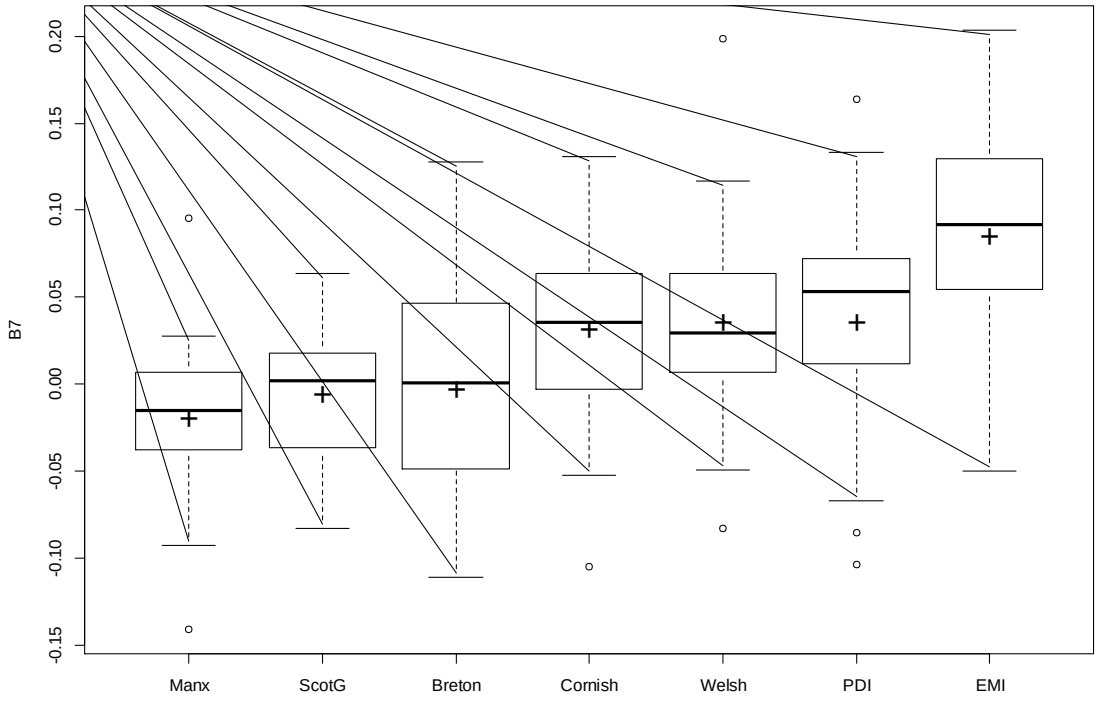

Figure 2. 95% t-confidence intervals for B7. Abbreviated labels: ScotG = Scottish Gaelic, PDI = Present-Day Irish, EMI = Early Modern Irish.

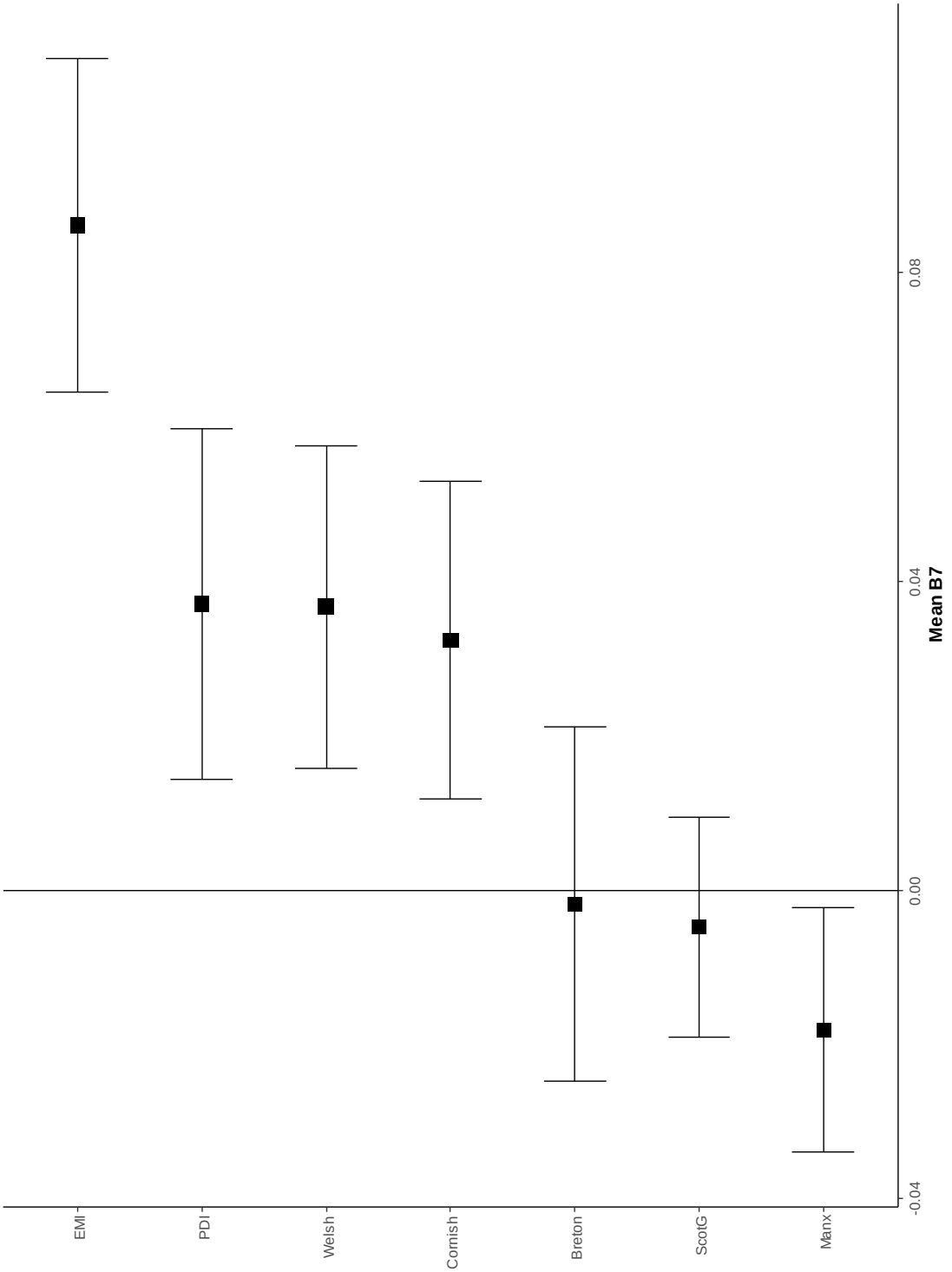

Figure 3. 99.7619% Welch-Satterthwaite confidence intervals for differences in mean B7 between pairs of languages. Abbreviated labels: ScotG = Scottish Gaelic, PDI = Present-Day Irish, EMI = Early Modern Irish.

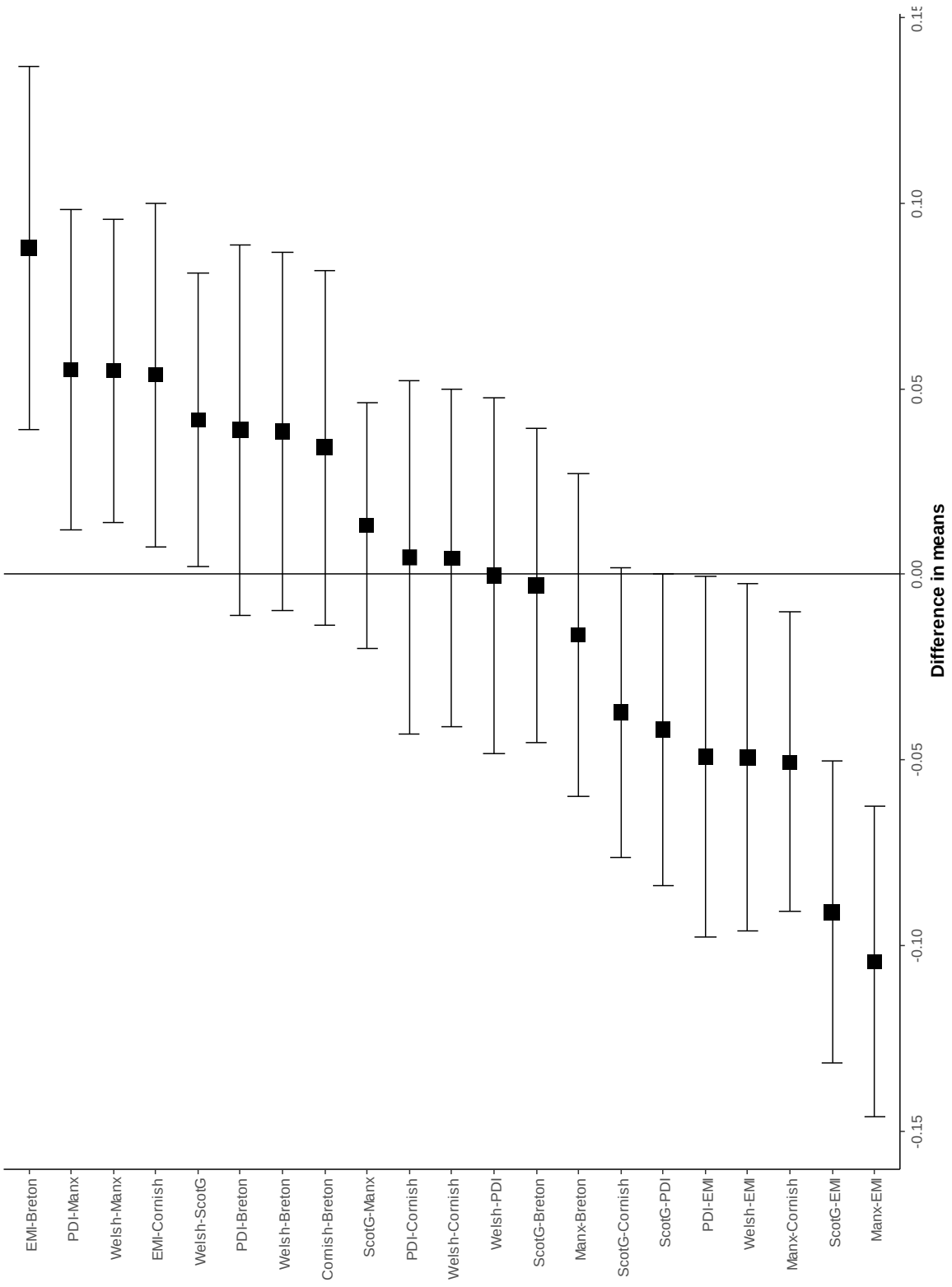

Figure 4. Plot of N (number of tokens) against V (number of types) for all 210 texts. (Labels: 1 = Breton, 2 = Cornish, 3 = Early Modern Irish, 4 = Manx, 5 = Present-Day Irish, 6 = Scottish Gaelic, 7 = Welsh.)