[From: *Information theory*: papers read at a symposium on 'Information theory' held at the Royal Institution, London, September 12th to 16th, 1955.; edited by Colin Cherry (London: Butterworths, 1956).]

16

# NEGATIVE ENTROPY OF WELSH WORDS

D. A. BELL AND ALAN S. C. ROSS
*University of Birmingham*

AFTER SHANNON[1] had experimentally investigated the degree of redundancy of printed English, BELL[2] (one of the authors of the present note) showed that the extent to which letters could be omitted from printed English without losing the sense could be estimated reasonably closely from the extent of the known constraints of English spelling; since comparatively few of the possible combinations of letters of the alphabet are accepted English words, the omission (or falsification) of some of the letters of a word will usually leave one and only one recognizable possibility for the original word. Since entropy can be broadly defined as a measure of average variability or of uncertainty, any factor which reduces variability, as does the imposition of a system of spelling on a generator of letters of the alphabet, reduces entropy; and Shannon estimated that the entropy of printed English averaged 2.62 bits per letter, whereas the entropy of a random selection from the 26-letter alphabet would be 4.7 bits per letter. This reduction of entropy by 2.08 bits per letter is equivalent to a certain amount of information *(i.e.* the recipient's prior knowledge that the message is to be in English allows the transmission to be less specific than would be necessary if the message consisted of arbitrary groups of letters) and this quality of English language was named by BELL[2] 'internal information'. In the title of the present paper this property of word structure has been denoted by negative entropy in order to emphasize the equivalence of negative entropy to information. This equivalence has been demonstrated by SZILARD[3] and others in relation to the Maxwell-demon paradox and by RAYMOND[4] and by BELL[5] in relation to a 'reversible' communication system. The latter is one in which the rate of communication per unit bandwidth is extremely slow *(cf.* the very slow changes in a reversible heat engine) and the information communicated may be precisely equated to the increase of entropy of the system due to the power used to effect communication. Entropy is relevant to printed information since the record will be permanent only if it involves a potential energy greater than $\frac{1}{2}kT$ per bit—though in practice the stability provided is far in excess of this, because the mechanisms tending to destroy the record are often more powerful than thermal agitation.

The calculation of negative entropy for the English language involved two stages:

(1) The reduction of entropy inherent in English spelling was estimated in terms of the ratio of number of words to number of letter-combinations for various lengths of word, by means of samples from the *Concise Oxford Dictionary.*

(2) The frequencies of usage of various lengths of word were estimated from DEWEY's list[6].

From the first consideration one obtains a measure of the reduction of entropy for words of different length, which may be expressed in bits per letter. By combining the second factor with these results one finds the reduction of entropy, or internal information, for English language.

In the discussion on a previous paper ROSS[7] suggested (1) that English was the world's most unsuitable language for this kind of test in that its sound-symbol correspondence (in contradistinction to that of, for instance, Welsh and Finnish) was most markedly not one-one ('thought' has three phonemes but seven letters); (2) that the test should have been carried out upon the whole of the English vocabulary (as set out in the *New English Dictionary)* and not upon a sample of it. The Research Committee of the University of Birmingham was kind enough to make a grant to the authors to enable them to implement these recommendations *apropos* Welsh. The Welsh 'count' was made by two persons, Revd. Arthur Davies (Welsh Presbyterian Minister for Wolverhampton and Birmingham) and Mr. R. Aled Roberts (Ysgol Cymraeg, Llandudno) on the basis of *Y geiriadur newydd; the New Welsh Dictionary* (H. M. Evans and W. O. Thomas, Llandebie, 1953) and *Spurrell's Welsh-English Dictionary* (Ed: J. Bodvan Anwyl, 10th Edition, Carmarthen, 1925). Welsh lexicography is (like that of Spanish) still in its infancy; the method adopted *(faute de mieux)* by the two counters was that they first looked at each dictionary entry to see whether or not they recognized the word entered, and secondly added words of which the entry reminded them but which were not entered in the dictionaries.

We may note here that in the count (1) no inflexional forms were included, (2) plurals and 'collectives' were entered in the form having the less number of letters (and in the singular if the forms had equal numbers of letters). Thus there were entered *dyn* (man), *sêr* (stars), *anghydffuriwr* (non-conformist) and not *dynion* (men), *seren* (star), *anghydffurwyr* (non-conformists). For the purpose of this count, Welsh may be considered to have the following 32 phoneme-characters:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| A | Â | B | C | CH | D | DD | E | Ê | F | FF |
| G | NG | H | I | L | LL | M | N | O | Ô | P |
| R | RH | S | T | TH | W | Ŵ | U | Y | Ŷ | |

This list calls for several comments:

(1) The long vowels* have been reckoned as separate characters since they are both phonematically distinct from the short ones and also kept distinct in print *(tân* 'fire'/*tan* 'under').

(2) The groups 'ch', 'dd', 'ff', 'ng', 'll', 'rh', 'th' have each been reckoned as one character, for they are clearly phonemes.

(3) Welsh has no long consonants, so 'nn', as of *tynnu* 'to pull', and 'rr', as of *gyrru* 'to drive', have been reckoned as single characters, for they are mere orthographies.

(4) The groups 'mh', 'ngh', 'nh' are difficult and require a little discussion. The 'grammar' of Welsh (unlike that of most European languages) is, in large

* Those marked with a circumflex accent (as Â) in the list.

part, effected by 'mutation'* as in *cadair* 'chair', *ei gadair* 'his chair', *ei chadair* 'her chair'; *gwraig* 'wife', *y wraig* 'the wife', *fy ngwraig* 'my wife'. The last example shows 'nasal' mutation; the nasal mutations of 'p', 't', 'c' are, phonetically, the unvoiced nasals corresponding to the voiced nasals 'm', 'n', 'ng'†, respectively; these unvoiced nasals are written 'mh', 'nh', 'ngh', respectively *(pen* 'head', *fy mhen* 'my head'; *tad* 'father', *fy nhad* 'my father'; *cadair* 'chair', *fy nghadair* 'my chair'). In Welsh, the 'privative'‡ prefix may be considered, from the point of view of the present study, to be of two types, *viz.* (1) *a-* + nasal + non-nasal (as *anfad* 'nefarious': *mad* 'good') and (2) *a-* + nasal mutation (as *amharod* 'unprepared': *parod* 'ready'). *Anhapus* 'unhappy' ( : *hapus* 'happy') is thus of type 1 whereas *anhebyg* 'dissimilar' ( : *tebyg* 'similar') is of type 2. There might thus, at first sight, well seem reason to regard 'nh' of *anhebyg* as one phoneme and 'nh' of *anhapus* as two phonemes. To a Welsh speaker, non-cognizant of Phonematology, these two 'nh's' are however identical (and, it may be added, almost invariably regarded as two units, not one). In our counting, we therefore had to take the decision whether to regard 'nh' invariably as one phoneme or invariably as two and we decided upon the latter alternative. Having so decided the case of 'nh', we could, it seemed to us, hardly come to the opposite conclusion in that of 'mh' and 'ngh'. All three—'mh', 'ngh' and 'nh'—have thus been reckoned by us as two characters each.

It is now possible to compare, for each word-length, the number of words with the number of possible written combinations of the 32 characters (the number of possible written combinations is, of course, far larger than the number of phonetically possible combinations). The results are recorded in the following table, and the only comment required is that in the single-lettered words difficulty is caused by homonyms: the single letter 'A'

*Table*

| *Word-length, characters* | *No. of words* | *No. of combinations of 32 characters* | *Reduction of entropy (bits per character)* |
|---|---|---|---|
| 1 | 4 | 32 | 3.0 |
| 2 | 87 | $1.02.10^3$ | 1.78 |
| 3 | 531 | $3.28.10^4$ | 1.98 |
| 4 | 1056 | $1.05.10^6$ | 2.49 |
| 5 | 1576 | $3.36.10^7$ | 2.89 |
| 6 | 1876 | $1.07.10^9$ | 3.19 |
| 7 | 1626 | $3.44.10^{10}$ | 3.48 |
| 8 | 1211 | $1.10.10^{12}$ | 3.72 |
| 9 | 810 | $3.51.10^{13}$ | 3.93 |
| 10 | 449 | $1.13.10^{15}$ | 4.12 |
| 11 | 267 | $3.40.10^{16}$ | 4.27 |
| 12 | 89 | $1.15.10^{18}$ | 4.46 |
| 13 | 30 | $3.68.10^{19}$ | 4.62 |
| 14 | 6 | — | — |
| 15 | 1 | — | — |

* That is, change of the initial consonant.
† As of English *ram, ran, rang.*
‡ That is, the prefix corresponding (both etymologically and semantically) to English 'un'- of *unkind.*

appears as five separate dictionary entries. Since these five are indistinguishable in print they have for the present purpose been counted as a single word and this together with the letters 'I', 'O', and 'Y' provides a total of four single-letter words.

The relation between word-length and negative-entropy in bits/letter is also displayed graphically as the full-line curve *in Figure 1* where the previously published[2] results for English are indicated by circles. The general form is similar but the significant differences are: (1) the occurrence of the lowest negative-entropy being for 2-letter words in Welsh, in contrast to 3-letter

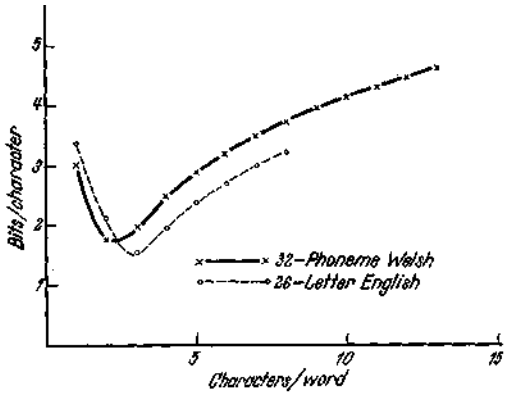


*Figure 1.*

words in English, and (2) the generally greater magnitude of negative-entropy in Welsh. The shift in position of the minimum must be attributed to some linguistic characteristic, but the general difference in level merely expresses the fact that the Welsh dictionaries contain far fewer words than the English dictionaries: the full count recorded here amounts to only 9,620 words, compared with an estimate of approximately 28,000 from the English samples. A large part of this difference corresponds to a comparative absence of long words in Welsh: words up to 8 letters inclusive account for 7,470 Welsh words, or 78 per cent of those recorded, but according to the previous sample they accounted for only 19,400 or 69 per cent of the English words. The increase in the proportion of long words is probably a late development resulting from the incorporation in a language of various groups of esoteric words. For example, GOOD[8] suggests an increase over the author's figures for English words by roughly 1.5 times for 3- and 4-letter words but nearer 2.5 times for 7-letter words.

The previous work in English was extended to the internal information of the language *i.e.* the type of assortment of words commonly used. Dewey's table of frequencies of occurrence of English words was used as a guide to the weighting factor to be applied to the negative-entropy of various word-lengths in order to find the contribution from word structure to the internal information of English language. It has not been possible to take this further step in Welsh since there is no analysis of word-frequencies comparable with Dewey's table for English.

## REFERENCES

[1] SHANNON, C. E. 'Prediction and Entropy of Printed English', *Bell Syst. Tech. J.*, 30 (1951) 50
[2] BELL, D. A. 'The "Internal Information" of English Words', *Communication Theory*, ed. W.Jackson, London; Butterworths, 1953
[3] SZILARD, L. *Zeits.fürPhys.*, 53 (1929) 840
[4] RAYMOND, R. C. *American Scientist*, 38 (1950) 273
[5] BELL, D. A. *Ibid.*, 40 (1952) 682
[6] DEWEY, G. *Relativ Frequency of English Speech Sounds*, Harvard, 1923
[7] See reference 2, p. 390
[8] GOOD, I. J. Discussion on reference 2

## DISCUSSION

COLIN CHERRY: In his introductory remarks, Dr. Bell has referred to certain thermodynamic concepts and measures; in particular he mentions $\frac{1}{2}kT$. I suggest that such truly thermodynamic topics have nothing whatever to do with the theme of this paper—the entropy of the Welsh language.

On the other hand, concerning the authors' investigation of the selective entropy of Welsh, involving such laborious studies of its syntactic constraints, I can only applaud their industry.

R. A. FAIRTHORNE: I suggest that the increase of physical entropy due to an act of decision, which could be reasonably taken as a lower bound, should not be identified with increases of entropy arising from the consequences of such a decision (such as an explosion!).

T. J. MCDERMOTT: I should like to ask Professor Ross a purely technical question. He appears to have treated the Welsh vowel diphthongs, such as the 'oe', and 'ai' in *coed* and *gwraig*, as two and not one phoneme. Does he think that this is valid ?

A. S. C. ROSS in reply: The decision whether to count a diphthong as one or two phonemes is always difficult. The latter choice is obviously the more desirable and should be made where there is no objection to it (as here). It is only where there is an objection to the latter choice that the former must, perforce, be adopted—as in English, for instance, where the word *ride* is almost universally considered to have three phonemes, not four.

D. A. BELL in reply: The relevance to information of thermodynamical concepts is justified in the introductory part of the paper.