An evaluation of Shannon's "Basic English" redundancy conjecture.

The decades leading to WWII saw the flourishing of the International Language movement. In 1933 E. Sapir advocated for the development of "a highly efficient and maximally simple international language" that, eschewing emotional and psychological factors, would provide "the individual with a fit symbol of solidarity with the international world." C. K. Ogden's "Basic English" (BE) enjoyed broad support as an international auxiliary language built for that purpose. Instead of being a whole new creation, BE was a simplified form of Standard English (SE), with a vocabulary limited to 850 words. The lexicon of SE was reduced to periphrases in BE based on philosophical analyses and a theory of definitions. The consequences of a simplified version of English for a mathematical theory of information did not go unnoticed. C. Shannon (1948) observed that "the Basic English vocabulary is limited to 850 words and the redundancy is very high." adding that "this is reflected in the expansion that occurs when a passage is translated into Basic English." Shannon (1950) explained "redundancy" as "the measure of the extent to which it is possible to compress (a language) if the best possible code is used." He calculated the redundancy of English to be around 50%, far below that of BE. To my knowledge, no attempt has been made to test Shannon's "Basic English Conjecture" empirically. In this project, I propose to create a parallel corpus of SE texts with their BE translations (including the Bible, stories by Poe, Hawthorne, and others, plays by Shakespeare and Shaw, political speeches, etc.). The two corpora will be compared to find out exactly how much more redundant BE is than SE.

Shannon (1948) defined redundancy as one minus the relative entropy of the language, where relative entropy is the ratio of the actual entropy of a source (given actual probabilities for the N symbols in the language) to its maximum entropy (assuming equiprobability over the N symbols). For a language with a vocabulary of size N, then, the redundancy R can be calculated with the following formula:

$$R = 1 - \frac{-\sum_{i=1}^{n} p_i \log p_i}{\log N}$$

Experiment 1 will determine the values of R for the sub-corpora in the parallel SE-BE corpus. Given that the vocabulary size of a corpus of SE will be much larger than the 850 words in its translation into BE, Shannon's conjecture is expected to be verified. But the comparison is not completely accurate. A single word in SE is rendered by a periphrasis in BE, which makes BE texts much longer than their translations. Nevertheless, a BE text communicates the same content as its SE translation, only in a more "rational" form of expression (according to Ogden's principles). The comparison, then should be with a model of BE that takes into account the probabilities of word sequences of length *f*, for different values of *f*. **Experiment 2** will run the different n-gram models over the BE corpus to calculate the redundancy R_f at different values of *f*. The prediction is that, as *f* increases, the redundancy of BE will converge with that of SE. The results of Experiment 2 will lead to a discussion of what the values of *f* mean for the claims made Ogden that auxiliary languages like BE offer a "simple", "regular", and "economical" way of constructing messages, which satisfy the "minimum of demands on the learning capacity of the humblest individual" (Ogden 1931: 27).

Bibliography

Falk, J. S. (1995). Words without Grammar: Linguists and the International Auxiliary Language Movement in the United States. *Language & Communication*, Vol. 15, No. 3, pp. 241-259.

McElvenny, J. (2017). *Language and Meaning in the Age of Modernism: C.K. Ogden and His Contemporaries.* Edinburgh University Press.

Ogden, C. K. (1931). *Debabelization (With a Survey of Contemporary Opinion on the Problem of a Universal Language).* London: Kegan Paul, Trench, Trubner & Co., Ltd

Ogden, C. K. (1933). Is an artificial auxiliary language necessary? Simplified English as an alternative. *Actes du deuxième Congrès international de linguistes: Genève, 25-29 août 1931*. Paris: Adrien-Maisonneuve. pp. 84-85.

Ogden, C. K. (1934). *The System of Basic English*. New York: Harcourt, Brace and Company.

Sapir, E. (1933). The case for constructed international language. *Actes du deuxième Congrès international de linguistes: Genève, 25-29 août 1931*. Paris: Adrien-Maisonneuve. 86-88.

Shannon, C. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656,

Shannon, C. (1950). The Redundancy of English. In: Claus Pias (ed.), *Cybernetics. The Macy Conferences 1946-1953, vol. 1: Transactions*, Berlin/Zurich 2003. 248-272