Information Equilibration in English and Japanese Morphemes

Alexander Kilpatrick[1]

[1]*Nagoya University of Commerce and Business (Japan)*

In Information Theory [1], improbable events express more information than probable ones. Information is measured using the Surprisal equation ($S = -\log^2 P$, where P is the probability of an event). Informativity—the mean Surprisal within a word—has been shown to be an excellent cross-linguistic predictor of word length whereby longer words tend to express information more sparsely [2]. The present study extends the literature on this phenomenon [e.g., 3,4] by exploring how morphemes contribute to informativity and how surprisal is expressed at the beginning of words in two unrelated languages: American English (hereafter: English) and Standard Japanese (hereafter: Japanese). It tests the following hypotheses: **H1**) reaffirm the established inverse relationship between informativity and word length, **H2**) explore how morphemes contribute to Informativity in English, and **H3**) examine Surprisal at the beginning of words and morphemes to test if length influences information expression across the entire word or only as they grow longer.

All data and code relating to this project can be found here: https://tinyurl.com/4f8tfuwj. Bigram Surprisal is calculated on diphone transitional probability and Informativity is calculated on the average Surprisal within words and morphemes. Only words and morphemes with more than one phoneme are included. Surprisal at the first position is the Surprisal of the second phoneme given the first phoneme. The English dataset comes from SUBLEX-US [5] a corpus of around 50 million instances of around 50,000 unique words taken from subtitle data of spoken American English. This was cross-referenced with the Carnegie Mellon University Pronouncing Dictionary [6] to obtain a phonemic transcription. Surprisal was calculated on the combined dataset. Unmatched samples were discarded. Morpheme counts were obtained by cross-referencing an additional database [7]. The Japanese dataset comes from the Corpus of Spontaneous Japanese [8] which lists morphemes as samples. Surprisal was calculated on romanization due to its relatively good phonemic match. Regression models were constructed in R [9].

H1: Two simple linear regression models were constructed to test the influence of length on Informativity. Both the English ($t(1,44558) = -42.45$, $p < .001$, $R^2 = 0.039$) and Japanese ($t(1,12775) = -15.62$, $p < .001$, $R^2 = 0.019$) models revealed significant negative correlations. These are illustrated in Figures 1 and 2. **H2**: A multiple linear regression model constructed to test the influence of length and morpheme count on English words ($F(2,37235) = 860.3$, $p < 0.001$, $R^2 = 0.044$) revealed significant effects for both length ($\beta = -0.094$, $p < .001$) and morpheme count ($\beta = 0.08$, $p < .001$), showing that for every additional phoneme, average Surprisal decreases by approximately 0.094, but for every additional morpheme, average Surprisal increases by approximately 0.08, provided the opposing metric remains stable. This calculation was not conducted for Japanese because samples in that dataset are morphemes, not words. **H3**: Two simple linear regression models were constructed to test whether information equilibration can be observed at the beginning of words by testing the influence of length on Surprisal of the second phoneme given the first. Both the English ($t(1, 44558) = -38.84$, $p < .001$, $R^2 = 0.033$) and Japanese ($t(1, 12775) = -13.25$, $p < .001$, $R^2 = 0.013$) models revealed significant negative correlations showing that length influences information expression at the very beginning of words/morphemes. The relationship between length and Surprisal at the first position is illustrated in Figures 3 and 4.

Longer words and morphemes express information more sparsely while shorter words and morphemes express information more densely (**H1**). This appears to be specifically tied to morphemes because—at least in English—when the number of phonemes remains constant, additional morphemes increase information density (**H2**), although further research should be conducted to examine the influence of compound words. One might consider that the relationship between Informativity and length is the result of phonotactic constraints—such as no coda /h/ in English and only coda nasals in Japanese—which increase the predictability of the latter parts of words and morphemes by decreasing possible diphone combinations; however, information equilibration was exhibited at the very beginning of words (**H3**) showing that the influence of word length on Informativity occurs across the word, not just as words get longer and phonotactic possibilities decrease. Research is currently underway to explore these effects in other languages.

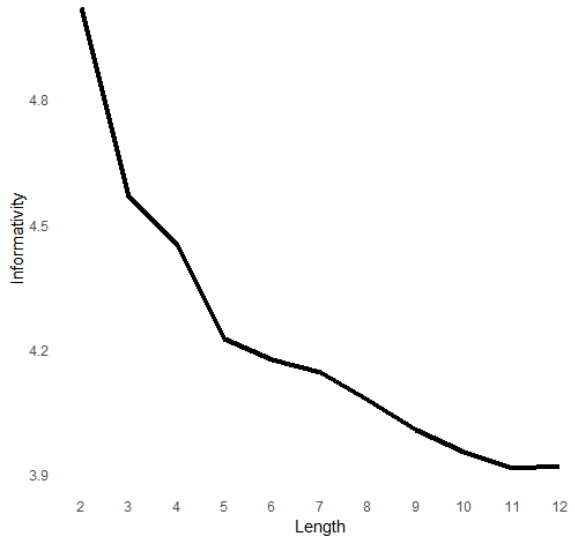

Fig. 1. Average Informativity of English words according to length. Only lengths with more than 500 samples are included.

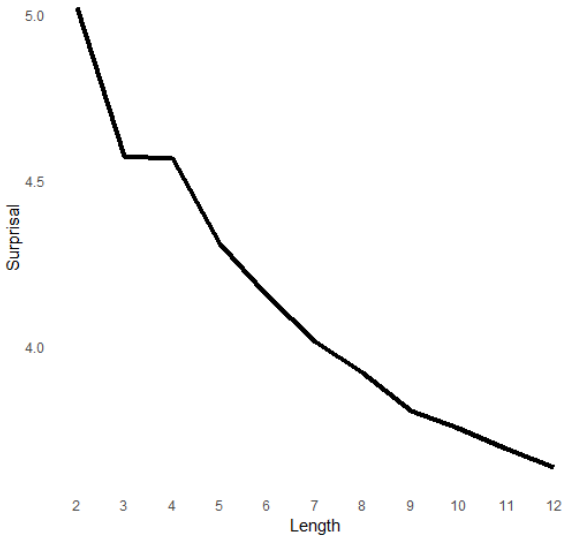

Fig. 3. Average Surprisal at the first position according to length in English. Only lengths with more than 500 samples are included.

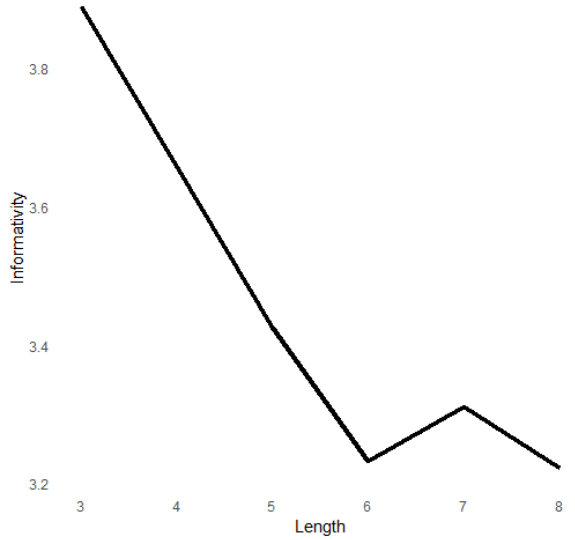

Fig. 2. Average Informativity of Japanese morphemes according to length. Only lengths with more than 500 samples are included.

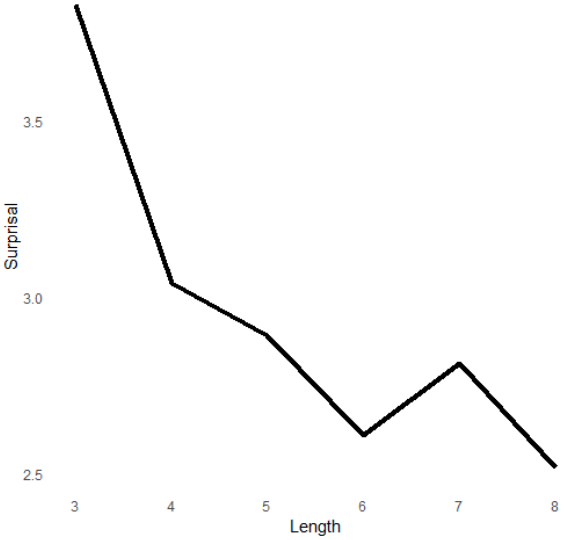

Fig. 4. Average Surprisal at the first position according to length in Japanese. Only lengths with more than 500 samples are included.

References

[1] Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

[2] Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526-3529.

[3] Hall, K. C., Hume, E., Jaeger, T. F., & Wedel, A. (2018). The role of predictability in shaping phonological patterns. *Linguistics Vanguard*, *4*(s2), 20170027.

[4] Cohen Priva, U. & Gleason, E. (2016). Simpler structure for more informative words: A longitudinal study. In *8th annual conference of the cognitive science society*, 1895–1900.

[5] Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, *41*(4), 977-990.

[6] Weide, R. (1998). *The Carnegie Mellon pronouncing dictionary*. Release 0.6, www.cs.cmu.edu.

[7] Sánchez-Gutiérrez, C. H., Mailhot, H., Deacon, S. H., & Wilson, M. A. (2018). MorphoLex: A derivational morphological database for 70,000 English words. *Behavior research methods*, *50*, 1568-1580.

[8] Maekawa, K. (2003). Corpus of spontaneous Japanese: Its design and evaluation. Proceedings *ISCA and IEEE workshop on spontaneous speech processing and recognition*, pp. 7–12.

[9] R Core Team (2020). R*: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/