Recognizing formulaic sequences: The subjective frequency of n-grams

Cyrus Shaoul[1], Harald Baayen[1,2], & Chris Westbury[2]
[1]*University of Tübingen* & [2]*University of Alberta*

The subjective frequency of individual words has been extensively studied and linked to performance on a variety of linguistic tasks. In the present study, we first replicated and then extended this work by asking subjects to judge the subjective frequency of combinations of n words (n-grams) to study what properties of the n-gram influence their responses. We investigated the capacity of people to gauge both the absolute and the relative frequencies of n-grams.

179 pairs of n-grams (matched within each pair on the geometric mean of their constituent word frequencies) were chosen from the Google Web1T data set (Brants & Franz, 2006): 60 pairs of 2-grams, 43 pairs of 3-grams, 36 pairs of 4-grams and 38 pairs of 5-grams. They covered a broad range of frequencies, from 1139 occurrences per million for the phrase *to the* to 0.00006 per million for the phrase *to know and keep the*. As these examples show, the n-grams were not restricted to being clausal phrases.

In Experiment 1, subjective frequency ratings were obtained from 1048 subjects who rated 31 n-grams each on a 7-point Likert scale, so that each phrase was rated about 1300 times. The ratings showed a strong correlation with corpus frequency, in particular for n-grams with the highest subjective frequency. We found effects of both whole n-gram and component n-gram frequencies. These effects varied with the length of the phrase.

In Experiment 2, the paired n-grams were used in a forced-choice relative frequency decision task in which 33 participants decided which of the paired n-grams was more frequent. Accuracy on this task was reliably above chance. The trial-level accuracy was best predicted by a model that included the ratio of the corpus frequencies of the two whole n-grams, for values of n < 4, or the ratio of frequencies of some of their shorter component n-grams, for n >= 4.

These results support models of reading that posit traces in long-term memory for (especially short) n-grams as well as words, models that take advantage of the probabilistic information in each n-gram.

References

Brants, T. & Franz, A. (2006). *Web 1T 5-gram version 1.* Philadelphia, PA: Linguistic Data Consortium.