Uncertainty reduction as a measure of cognitive processing effort in sentence comprehension

Stefan Frank
University College London
contact: s.frank@ucl.ac.uk

It has been argued that the amount of information conveyed by each word in a sentence is a measure of the amount of cognitive effort needed to process the word. Two complementary formalizations of word information have been proposed: surprisal (Hale, 2001) and entropy reduction (Hale, 2006). These quantify, respectively, the extent to which a word’s occurrence was unexpected, and the word’s effect on the uncertainty about the rest of the sentence. The goal of this study was to investigate whether these two information measures indeed predict processing effort, as observed in word-reading times.

Both information measures can be computed from the probabilities of words given the sentence-so-far. Formally, if $w_{1...t}$ denotes the first $t$ words of the sentence and $w_{t+1}$ is the upcoming word, then the surprisal of that word equals

$$\text{surprisal}(w_{t+1}) = -\log P(w_{t+1}|w_{1...t}).$$

The entropy directly before processing $w_{t+1}$ is defined as

$$H(t) = -\sum_{w_{t+1}...t+n} P(w_{t+1}...t+n|w_{1...t}) \log P(w_{t+1}...t+n|w_{1...t}),$$

(1)

where $w_{t+1}...t+n$ denotes the rest of the sentence, that is, from word $w_{t+1}$ onwards. The reduction in entropy due to $w_{t+1}$ is simply

$$\Delta H(w_{t+1}) = H(t) - H(t + 1).$$

Although the effect of surprisal on reading time is well established (e.g., Demberg & Keller, 2008; Frank & Bod, 2011), the role of entropy reduction has not received much attention, possibly because the estimation of entropy is computationally very demanding. Hale (2006) showed how entropy reduction can be computed from probabilistic context-free grammars, but this is not computationally feasible for grammars of realistic size. Frank (2010) found that entropy-reduction values, as computed relatively efficiently by recurrent neural nets (RNNs), account for word-fixation times in eye-tracking data. However, these values were estimated only for the words’ syntactical categories rather than for the words themselves.

Here, we show for the first time that there is a general positive relation between the reading time on a word and the reduction in sentence-entropy due to processing that word. A RNN was trained on 702,412 sentences (comprising over 7.6 million word tokens; 7,754 word types) from the British National Corpus. At several point during training, the network was made to generate surprisal and entropy-reduction estimates for 5,043 word tokens of 361 sentences, selected from three novels on www.freeonlinenovels.com. In theory, entropy is defined, as in Eq. (1), over the probabilities of all possible continuations $w_{t+1}...t+n$ of the sentence-so-far. However, in order to make the estimation of entropy-reduction computationally feasible, the number of continuations was restricted in two ways. First, instead of looking at complete continuations, only the upcoming three words were taken into account (i.e., $w_{t+1}...t+3$). Second, only the 40 most probable length-$m$ continuations $w_{t+1}...t+m$ were retained for estimating the probabilities of the length-$m+1$ continuations $w_{t+1}...t+m+1$. 
Reading times on each word of the same set of 361 sentences were collected in a self-paced reading task involving 54 native speakers of English. Mixed-effect regression analyses (including several covariates such as word frequency and length) showed that both surprisal and entropy reduction are positively related to word-reading time, with either factor having a significant effect over and above the other. Moreover, as shown in Figure 1, these effects grew stronger as the network learned to capture the language statistics more accurately. At the end of RNN training, the unique effects of surprisal and entropy reduction were, respectively, $\chi^2 = 145 \ (p \approx 0)$ and $\chi^2 = 12.3 \ (p < .001)$.

These results show that the reduction in uncertainty about sentence continuations is indeed a cognitively relevant measure of the amount of information conveyed by a word, and provides further support for the theory that word-information content quantifies cognitive processing effort.

Figure 1: Effect sizes of surprisal (left) and entropy reduction (right) as a function of the network’s accuracy as a language model (quantified by average negative surprisal). Effect size is the $\chi^2$-statistic of either the surprisal or entropy-reduction predictor in the mixed-effect model.

References: