On the relationship between form and information content

Cassandra L. Jacobs[1], Arya D. McCarthy[2], and Maryellen C. MacDonald[1]
1: University of Wisconsin, Madison; 2: Johns Hopkins University
jacobs.cassandra.l@gmail.com

Models of linguistic co-occurrence statistics, known broadly as *language models*, have shown considerable success in explaining human language performance (Smith & Levy, 2013; Seyfarth, 2014; Cohen Priva, 2015; 2017). Some have argued that *informativity*, a measure of the average probability of a word appearing in any particular context, predicts word forms that speakers select, with shorter words in more predictive contexts and longer words in less predictive ones, as measured by n-gram statistics (Piantadosi, Tily, & Gibson, 2011; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013). Recent advances in language modeling with neural networks have shown improvements over n-gram models in predicting linguistic statistical regularities, such as long-distance syntactic dependencies, which may inform these length-informativity relationships.

The present work addresses these issues by applying a powerful neural language model (Ng et al., 2019) to a large human cloze dataset (Luke & Christianson, 2016) with word-by-word predictions for 220 English sentences belonging to 55 paragraphs of news articles. Given a sentence like, "The watch doesn't have a...", participants might predict *battery* or *purpose*, even though the next word was actually *microphone*. The dataset contains 11,660 words and over 40,000 unique predictions. For comparison to the human participants, we use a pre-trained Transformer-based model whose learning objective is to predict the next word in a standard machine translation corpus using only prior context (Ng et al., 2019). As in Piantadosi et al. (2011), we quantify the *information content* of a word, defined as a word w's surprisal in context averaging across all contexts c:

$$\frac{-1 * \sum_{i=1}^{N} log(p(W = w|C = c_i))}{N}$$

For human participants, $p(W = w|C = c)$ is the proportion of participants who predict a given word from the context; for the models, $p(W = w|C = c)$ is defined as the probability the model assigns to the next word given the sentence it has processed so far.

We can make two general predictions. First, if participants are rationally expecting longer word forms to occur in less predictable contexts, then we expect to see longer words in less predictable contexts (or, conversely, less predictable contexts for longer words). Second, if speakers are aware of the relationship between word length and predictability, then participants should predict longer words in unpredictable contexts — even if they are incorrect ($\approx$ 80% of responses). The neural language model should similarly show that longer words occur in less predictive contexts. We show in Figure 1 the relationship between (scaled) information content, source of the informativity measure (human or model; Producer), whether a completion was correct or not (Match), and word length. A linear model (Table 1) predicting word length in characters confirmed that correct and incorrect completions differed in the effect of information content: for correct completions, the frequency-informativity relationship is robust but very weak for incorrect completions for both human participants and the model. An additional analysis showed that lexical frequency better predicted word length (R^2=0.13) than information content (R^2=0.04). Nevertheless, counter to the findings of Piantadosi et al. (2011) who found that information content alone best predicted word length, including both in a model resulted in the best fit to the data (R^2=0.14),

In sum, this work shows that while longer words do typically occur in less predictive contexts in natural texts, human participants do not necessarily predict longer words. These results raise important questions about the cognitive architectures that could support an information theoretic process model of language production, perhaps by accounting for the differences between online and offline demands on the language production system.

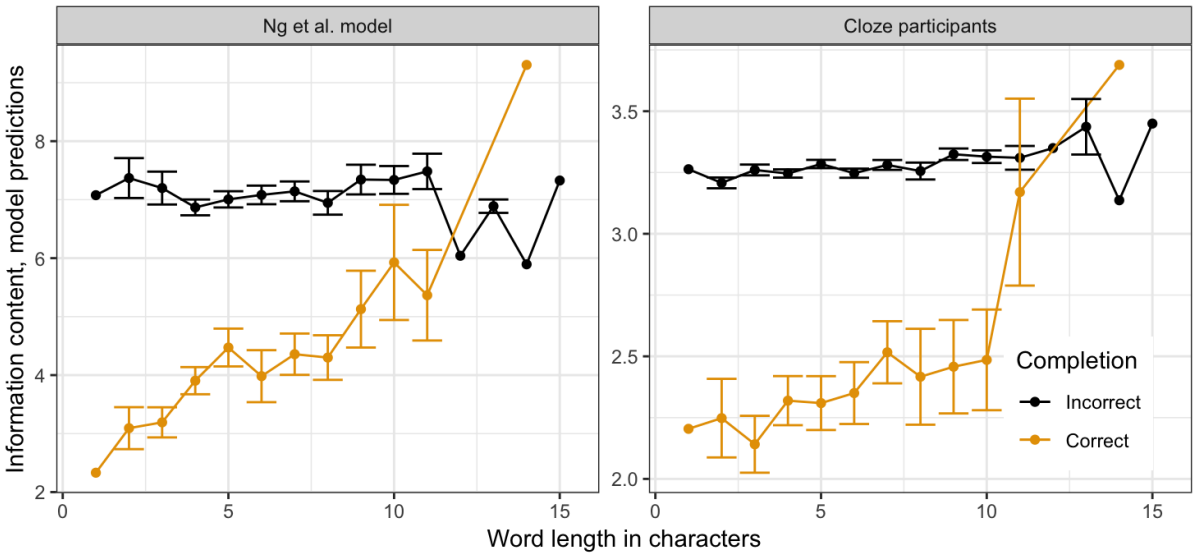

Figure 1: Relationship between word length and information content.

	β	SE	t	p
Intercept	6.33	0.05	118.74	< .001
Completion matches	-0.25	0.05	-4.83	< .001
Information content	-0.16	0.06	-2.54	< .05
Human vs. model (Producer)	0.02	0.05	0.41	n.s.
Information content x Match	0.60	0.06	9.17	< .001
Producer x Match	0.00	0.05	0.14	n.s.
Information content x Producer	0.09	0.06	1.39	< n.s.
Information content x Producer x Match	-0.29	0.06	-4.41	< .001

Table 1: Linear model of word length in characters.

References

[1] Cohen Priva, U. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, *6*, 243-278. **[2]** Cohen Priva, U. (2017). Informativity and the actuation of lenition. *Language*, *93*, 569-597. **[3]** Luke, S. G., Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22-60. **[4]** Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*, 313-318. **[5]** Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., & Edunov, S. (2019). Facebook FAIR's WMT19 News Translation Task Submission. In *Proceedings of the Fourth Conference on Machine Translation* (Volume 2: Shared Task Papers, Day 1), (pp. 314—319). **[6]** Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*, 3526-3529. **[7]** Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*, 140-155. **[8]** Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302-319.