Efficient Communication Forward and Backward

Richard Futrell (MIT), Kyle Mahowald (MIT), Steve Piantadosi (University of Rochester), and Edward Gibson (MIT)

Contact: *futrell@mit.edu*

Information-Theoretic Approaches to Language

Length and Informativity

General hypothesis: Word lengths reflect a pressure for *efficient communication.* Less informative words should be shorter and more reduced (Aylett & Turk, 2004). More informative words should be longer and more distinctive.

Informativity is Surprisal: -$\log_2 p(w)$

Example: The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point Frequently these messages have meaning that they refer to or are correlated according to some

(Shannon, 1948)

In an efficient code, you can minimize your expected code length by making uninformative words shorter and informative words longer (Shannon, 1948, 1951).

Average surprisal in context predicts word length

Efficiency and word length: Word lengths are determined by the *average* amount of information conveyed by a word *in context*. Information content (average surprisal in context) predicts word length better than frequency of usage (Zipf, 1936, 1949; Piantadosi, Tily, & Gibson, 2011, 2013).

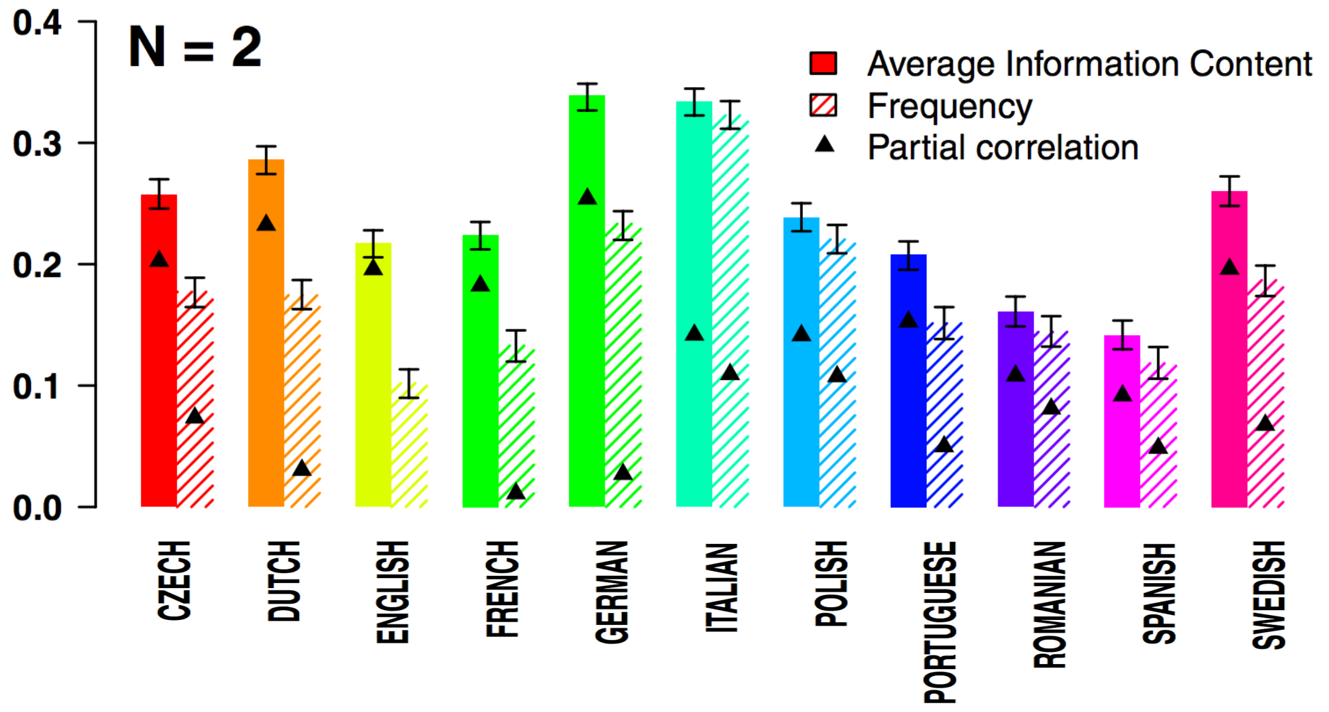


Figure 1. Relationship between frequency and length, and information content and length, in 11 languages, from Piantadosi, Tily, & Gibson (2011).

Surprisal in context predicts word duration

Efficiency and word duration: Controlling for length, words that are more predictable in a certain context are spoken faster (Bell et al., 2009). Their vowels are also more reduced (Aylett & Turk, 2004). Also, speakers are more likely to use abbreviated forms (Mahowald et al., 2012).

What Context Matters?

Forward Surprisal

If informativity is predictability in context, that raises the question: what context?

Most information theoretic studies examine the effects of predictability *given previous words*. Similarly, most *N*-gram language models use preceding context, even when this is not required by the engineering application (Jurafsky & Martin, 2008).

We call surprisal estimated as the probability given previous words ***forward surprisal.*** For example, we can quantify the surprisal of a word in this context:

fundamental problem of _?_

Backward Surprisal

Here we study ***backward surprisal****:* the predictability of words given *following* context. We quantify the surprisal of a word in this kind of context:

__?__ problem of communication

Bell et al. (2009) find that it is backward surprisal—*not* forward surprisal!—that predicts word duration for most words. Also, Popel & Marecek (2010) find that backward *N*-gram models have lower perplexity than forward models.

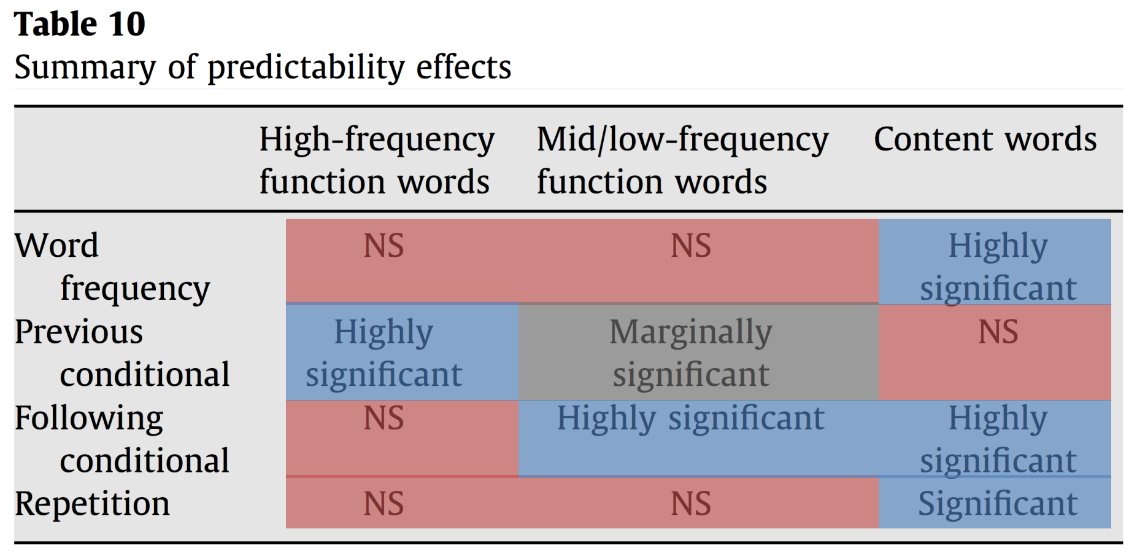


Table 10
Summary of predictability effects

	High-frequency function words	Mid/low-frequency function words	Content words
Word frequency	NS	NS	Highly significant
Previous conditional	Highly significant	Marginally significant	NS
Following conditional	NS	Highly significant	Highly significant
Repetition	NS	NS	Significant

Figure 2. Summary of findings from Bell et al. (2009).

Crucially, the surprisal of a word for the hearer is based on forward surprisal, while only the speaker has access to following words.

So if words shorten due to backward surprisal, then word lengths are not only optimized for the hearer's surprisal in the moment. **This would rule out the subset of audience design theories where the speaker is trying to minimize immediate surprisal for the hearer.**

Length Correlates with Backward Surprisal

Crosslinguistic Corpus Study

Using the paradigm of Piantadosi et al. (2011), we correlate word length with average surprisal as measured in the Google Web N-Gram corpora. We include more diverse languages than the previous study, and we measure the effect of average backward surprisal.

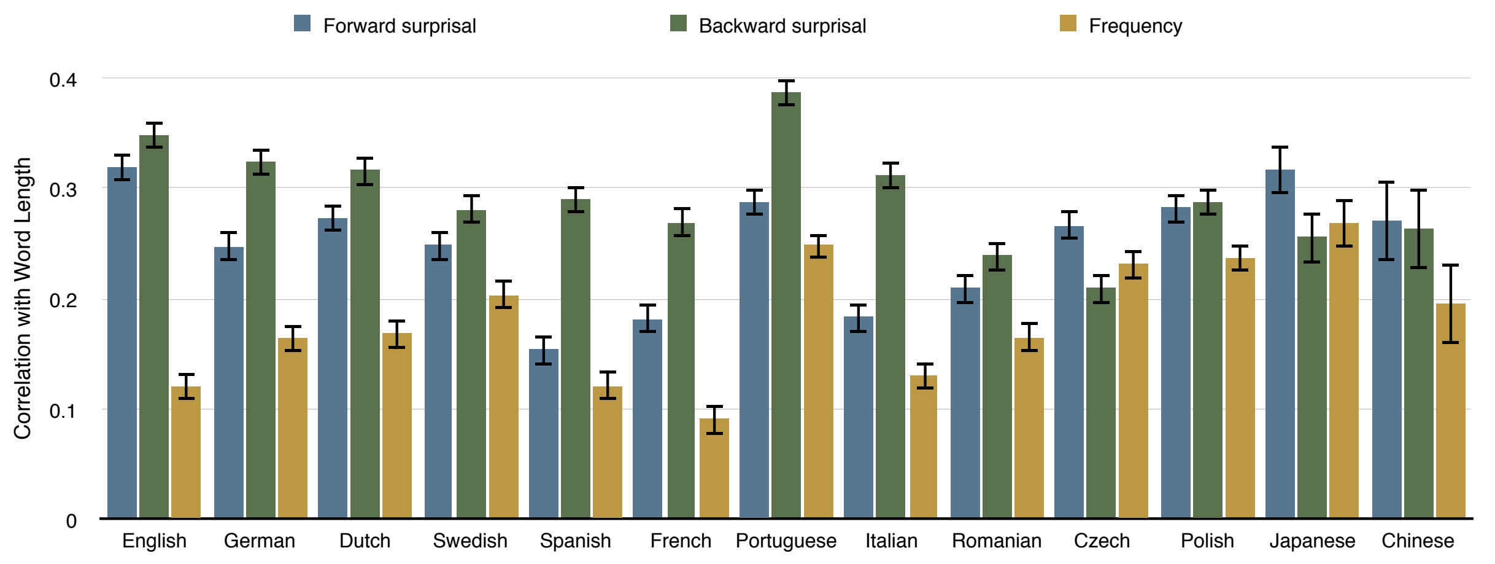


Figure 3. Spearman correlations of surprisal with word length in 13 languages in Google Web *N*-grams, *N*=2. Length in Japanese is measured in morae; length in Chinese is measured in pinyin characters.

As Figure 3 shows, in many languages, average backward surprisal is a significantly better predictor of word length than average forward surprisal. But the two are highly correlated. Figure 4 shows length correlations of average forward surprisal and average backward surprisal, when controlling for frequency and for the other kind of surprisal.

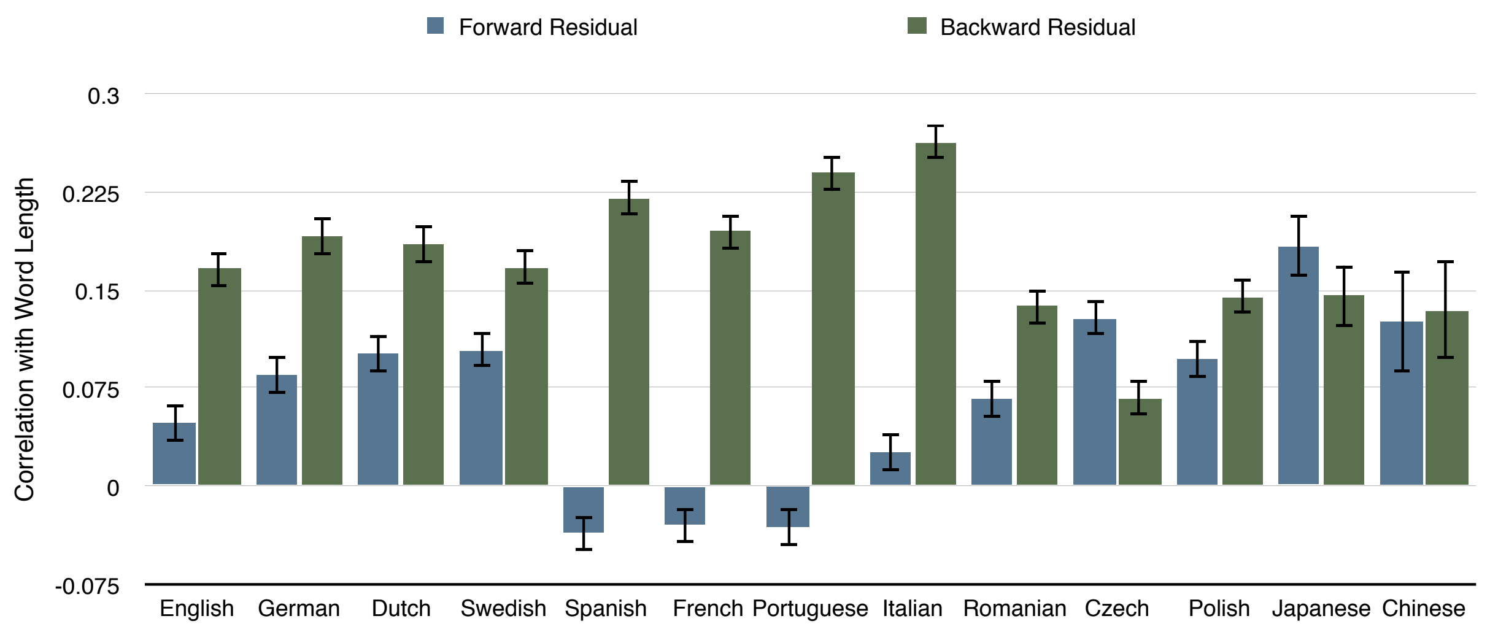


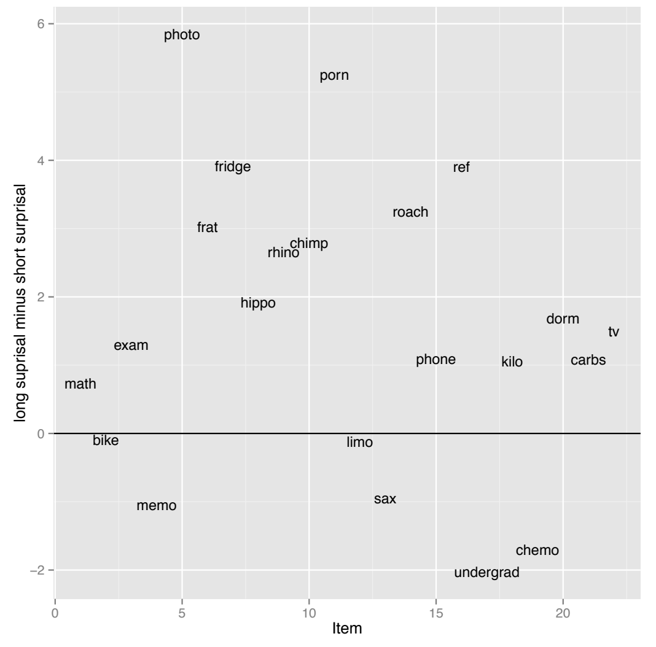
Figure 4. Spearman correlations of **residualized** surprisal with word length in 13 languages in Google Web *N*-grams, *N*=2. Length in Japanese is measured in morae; length in Chinese is measured in pinyin characters.

These data reiterate Bell et al.'s (2009) finding that backward surprisal affects reduction; in this case we find that it affects the length of words as stored in the lexicon. This means that word lengths are not optimized for minimizing the hearer's immediate surprisal, but rather for creating a code that is efficient for the speaker or for a hearer who cares more about global coherence than momentary surprisal.

Check out https://github.com/piantado/ngrampy for the code to replicate these results!

Word Shortening Does Too

For words that have short and long variants, such as *chimp/ chimpanzee*, the shorter variant has lower average surprisal—measured with both forward and backward surprisal. This indicates that speakers prefer shorter forms when a word is predictable given following material.



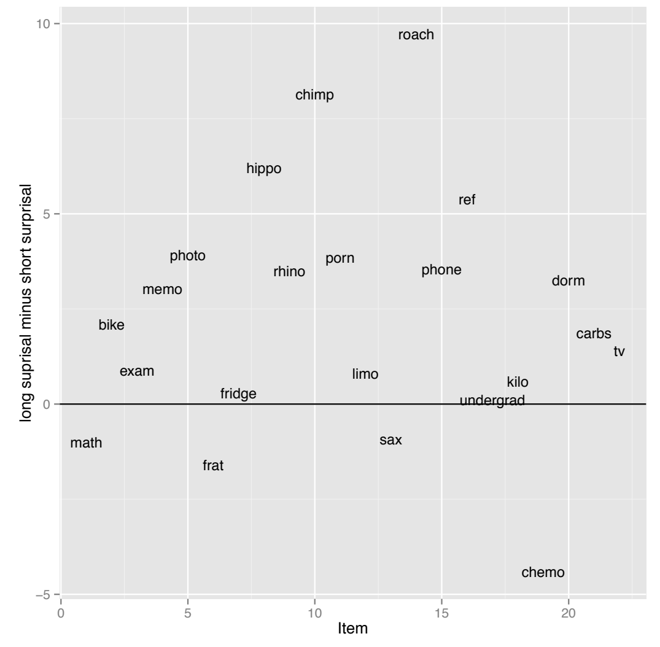


Figure 5. Words with long and short variants by difference in average **forward** surprisal between variants.

Figure 6. Words with long and short variants by difference in average **backward** surprisal between variants.

Conclusions

The relevance of preceding context for surprisal seems intuitively obvious: it is the order in which linguistic units are sent and received. However, when viewed in the context of the whole processes of production and comprehension, it is less clear that forward context is what matters.

In producing an utterance, a speaker has knowledge of what she is going to talk about, so the decision to produce each word is not conditioned solely on the words previous produced. Similarly, a comprehender is trying to understand a whole utterance; if one part of the utterances doesn't make sense when he initially hears it, that is fine if following parts of the utterance make it make sense in retrospect.

We hope the demonstrated effect of backward surprisal motivates a move away from models that use preceding context only.

References

M. Aylett & A. Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47: 31–56.

A. Bell, J. Brenier, M. Gregory, C. Girand, & D. Jurafsky. 2009. Predictability Effects on Durations of Content and Function Words in Conversational English. *Journal of Memory and Language,* 60(1): 92-111.

D. Jurafsky & J. H. Martin. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall.

K. Mahowald, E. Fedorenko, S.T. Piantadosi, & E. Gibson. 2012. Info/information theory: speakers actively choose shorter words in predictable contexts. *Cognition*, 126: 313-318.

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

S. T. Piantadosi, H. Tily, & E. Gibson. Information content versus word length in natural language: A reply to Ferrer-i-Cancho and Moscoso del Prado Martin. [arXiv:1209.1751]. ArXiv e-prints.

M. Popel & D. Mareček. 2010. Perplexity of n-gram and dependency language models. *Text, Speech and Dialogue*, 173-180.

C. E. Shannon, 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3): 379–423.

C. E. Shannon, 1951. Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1): 50-64.