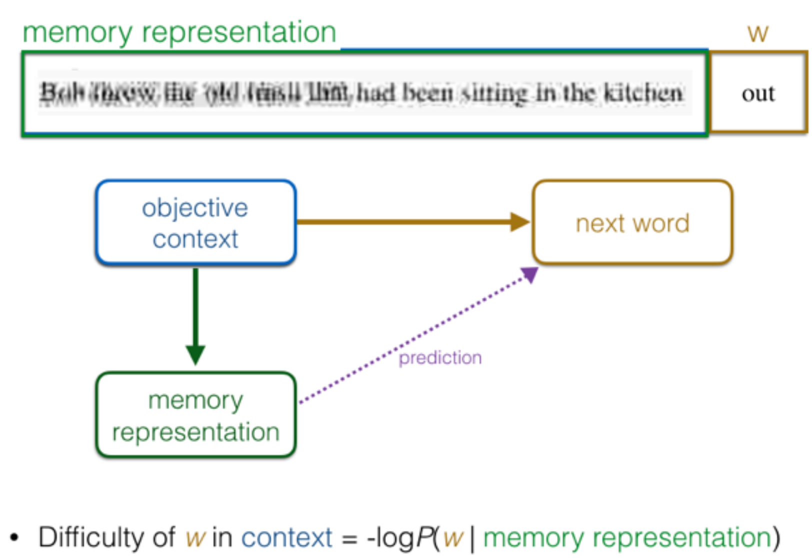
Richard Futrell
@rljfutrell

Follow ···

Lossy-context surprisal extends the reach of information-theoretic models of human language processing, and lets us make new predictions about how efficiency shapes language. Work with @languageMIT @roger_p_levy. Open-access paper: onlinelibrary.wiley.com/doi/full/10.11...



12:43 PM · Feb 27, 2020

4 32 65 16

Richard Futrell @rljfutrell · Feb 27, 2020

Quick summary: Lossy-Context Surprisal says that incremental processing difficulty for a word in context is given by -log P(word|memory). The memory is lossy, and this ends up explaining various effects in sentence processing. And now in more detail...

1 1

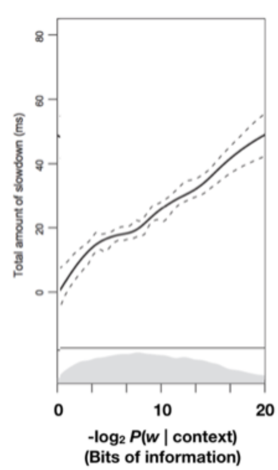
Richard Futrell @rljfutrell · Feb 27, 2020

The goal is to predict how much effort goes into processing each word in context during online language comprehension. Usually this effort is measured using reading times, based on various methodologies.

1 1

Richard Futrell @rljfutrell · Feb 27, 2020

One robust generalization is that words are hard to understand when they are unexpected in context. More precisely, word-by-word difficulty appears to scale with the negative logarithm of the probability of a word in context, as -log P(word | context).

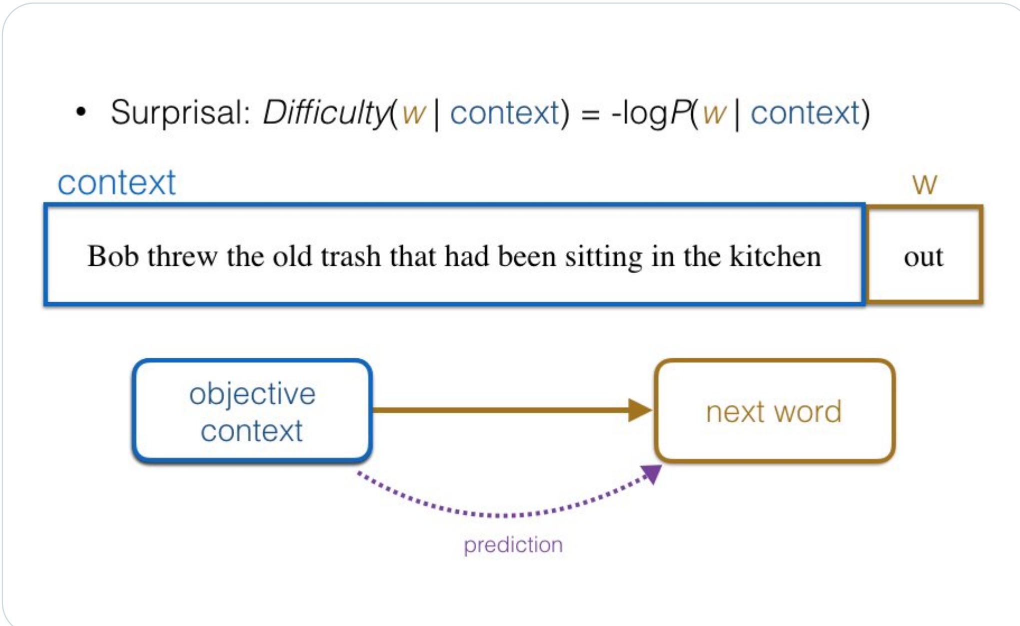


Smith & Levy (2013)

1 1

Richard Futrell @rljfutrell · Feb 27, 2020

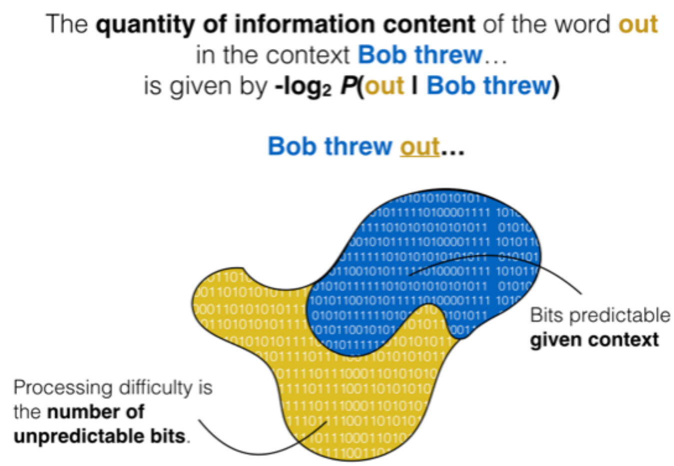
Surprisal Theory is a psycholinguistic theory based on this idea. It says that the comprehender uses context to form expectations about the next word, and things are hard when the next word is surprising given those expectations.



1 1

Richard Futrell @rljfutrell · Feb 27, 2020

You can think of Surprisal Theory in terms of information. Below, the blob represents all the bits of information in the word "out". Some of those bits (the blue ones) are predictable. The remaining (yellow) bits are not, and they determine the processing effort for the word.



1

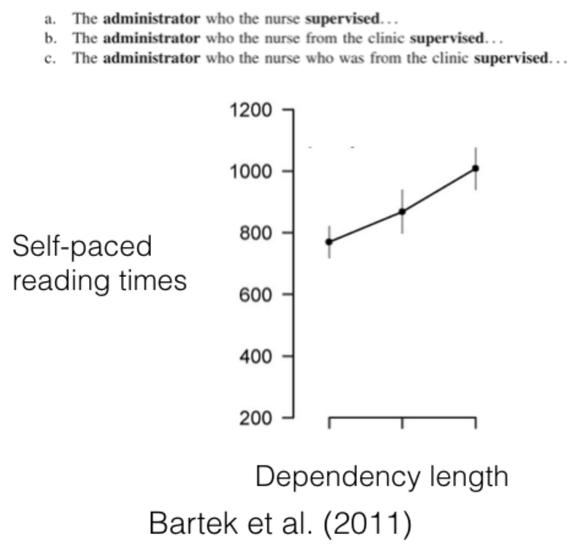
Richard Futrell @rljfutrell · Feb 27, 2020

Surprisal Theory can predict many empirical phenomena (including many garden path effects), and it has multiple converging theoretical justifications. But there is a class of sentence processing phenomena that it cannot handle: effects of memory.

1 1

Richard Futrell @rljfutrell · Feb 27, 2020 ···

Words are hard to understand when they require difficult memory retrieval operations. For example, when a word is distant from another word that it depends on, memory retrieval difficulty increases, and reading time slows down. This effect is called dependency locality.



1 1

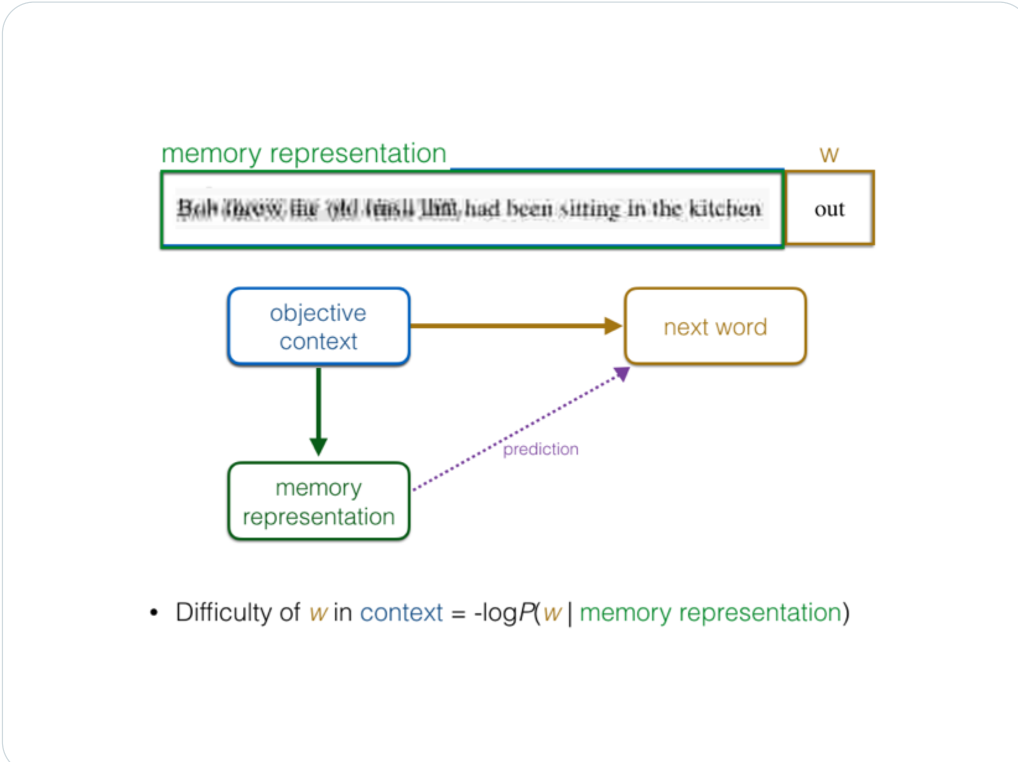
Richard Futrell @rljfutrell · Feb 27, 2020 ···

There are also many other memory effects in sentence processing. Our goal is to capture these memory effects within an information-theoretic, expectation-based framework like Surprisal Theory.

1 1

Richard Futrell @rljfutrell · Feb 27, 2020 ···

Lossy-Context Surprisal says the comprehender is predicting the next word given a *lossy memory representation* of the context. "Lossy" means that the memory representation does not contain complete information about the context.



1 2

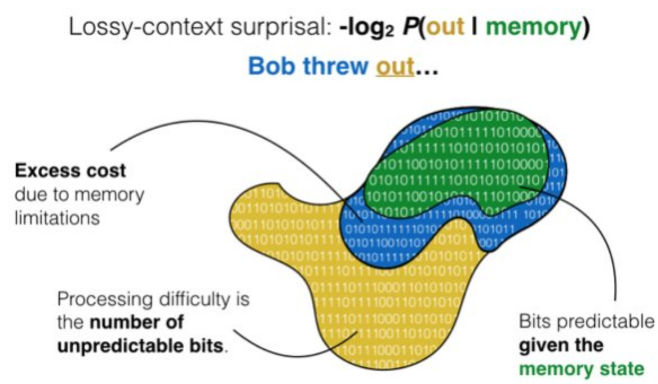
Richard Futrell @rljfutrell · Feb 27, 2020 ···

So the comprehender's expectations are different from what they would be if the comprehender knew the complete context. So the comprehender will experience extra surprisal at the next word. That extra surprisal constitutes the effects of memory on sentence processing.

1 1

Richard Futrell @rljfutrell · Feb 27, 2020

Here's the information-based picture. The green bits are predictable from the memory representation, and the blue ones would be predictable from the true context, but not from the memory state. Those blue bits convert into processing difficulty, on top of Surprisal Theory.



1

Richard Futrell @rljfutrell · Feb 27, 2020

Now, we haven't specified anything about what the memory representation looks like yet. But before doing so, we can use information-theoretic principles to make some general deductions about what *any* lossy-context surprisal theory must look like.

1 1

Richard Futrell @rljfutrell · Feb 27, 2020

When you are predicting the next word given your memory representation, you have to do noisy-channel inference to figure out what the real underlying context was. All the principles of noisy-channel processing apply.

1 1

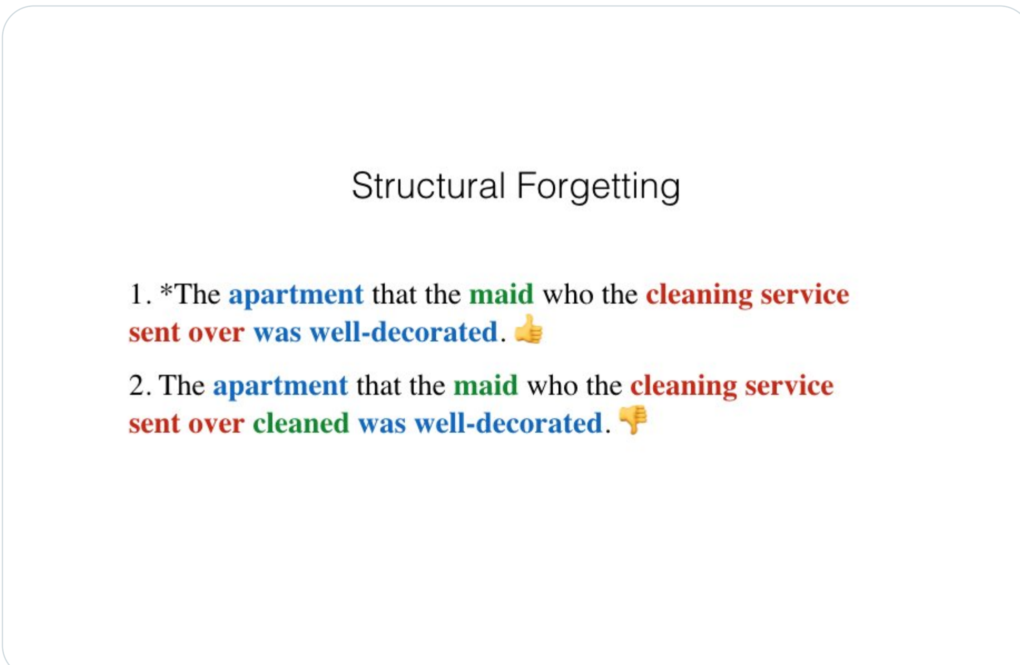
Richard Futrell @rljfutrell · Feb 27, 2020

For example: Noisy-channel inference is based in part on prior expectations. So the comprehender's expectations under Lossy-Context Surprisal will be biased towards continuations that are probable a priori, without regard to context.

1 1

Richard Futrell @rljfutrell · Feb 27, 2020

This turns out to explain structural forgetting, a puzzling phenomenon in sentence processing that involves both expectations and memory. In English, sentence (1) below sounds as acceptable as sentence (2), even though (1) is ungrammatical—it needs the verb "cleaned".



1 2

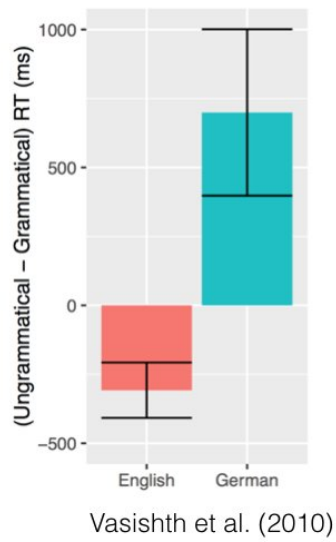
Richard Futrell @rljfutrell · Feb 27, 2020

The usual explanation is that processing the word "cleaned" in (2) is so difficult, due to memory effects, that people prefer the ungrammatical (1). But what makes structural forgetting a big puzzle is what happens in other languages.

1 1

Richard Futrell @rljfutrell · Feb 27, 2020

In English, RT is faster for the ungrammatical sentence. In German and Dutch, it's faster for the grammatical sentence. It seems that the statistics of these languages somehow interact with the structure of memory to produce different behaviors.



1 · 1

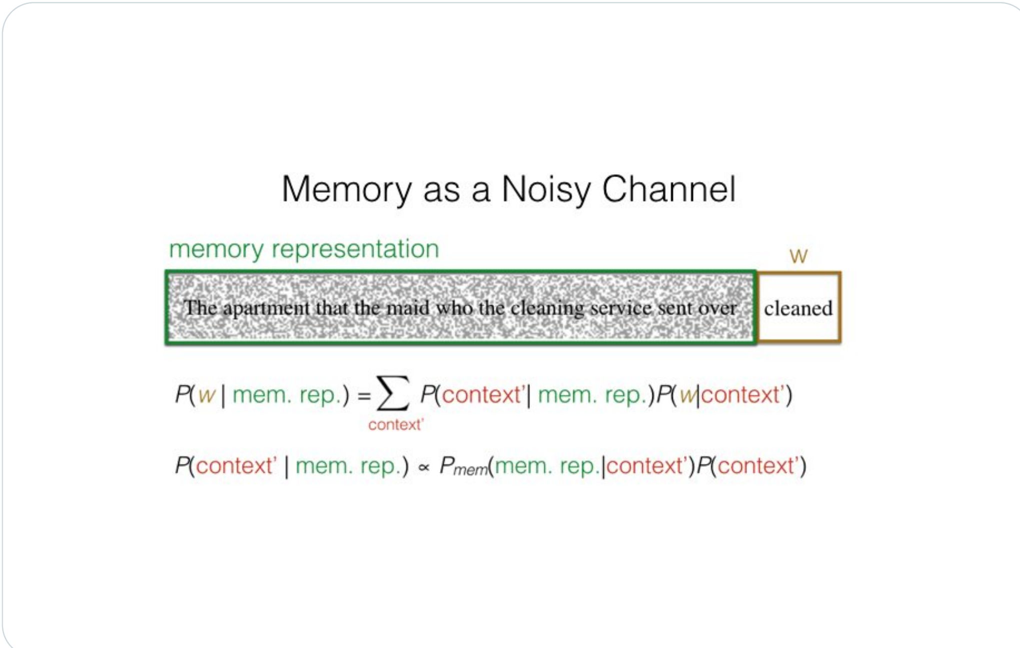
Richard Futrell @rljfutrell · Feb 27, 2020

(Thanks to @shravanvasishth for sharing data!)

1

Richard Futrell @rljfutrell · Feb 27, 2020

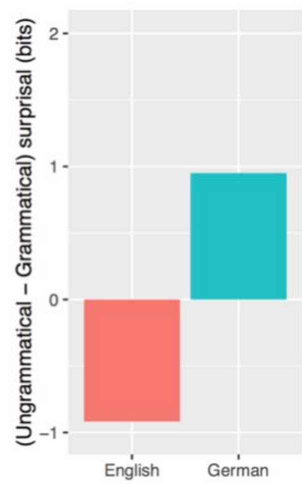
Lossy-context surprisal explains this by looking at the probability of verb completions given a noisy memory of the context, as below. The grammatical thing to do is to complete the sentence with three verbs.



2 1

Richard Futrell @rljfutrell · Feb 27, 2020

Based on toy grammars of English vs. German, and modeling noise in memory using random deletions, we can reproduce the language-dependent structural forgetting effect using lossy-context surprisal values:



1 1

Richard Futrell @rljfutrell · Feb 27, 2020

What's going on? In English, nested verb-final constructions are rare, so a two-verb completion is much more a priori probable than a three-verb completion. So given noisy memory, people gravitate towards the two-verb completion.

1 1

Richard Futrell @rljfutrell · Feb 27, 2020

That is, even if the two-verb completion has probability zero (i.e., is ungrammatical) in the true context, comprehenders still end up assigning it high probability due to their lossy memory. In this way, the model has a competence-performance distinction.

1 1

Richard Futrell @rljfutrell · Feb 27, 2020

In German/Dutch, on the other hand, nested verb-final constructions are more common, so the three-verb completion is relatively more probable a priori. So, people are less drawn toward the two-verb completion in these languages. This follows from noisy-channel principles.

1 1

Richard Futrell @rljfutrell · Feb 27, 2020

Previously, @StefanLFrank and colleagues showed that neural network language models also reproduce the language-dependent structural forgetting effect. We think this is because these models have lossy memory representations.

1 1

Richard Futrell @rljfutrell · Feb 27, 2020

Next, we show how you can derive the existence of dependency locality effects in Lossy-Context Surprisal. The derivation requires an assumption that memory representations degrade over time. I won't go as deep into this one, except to say...

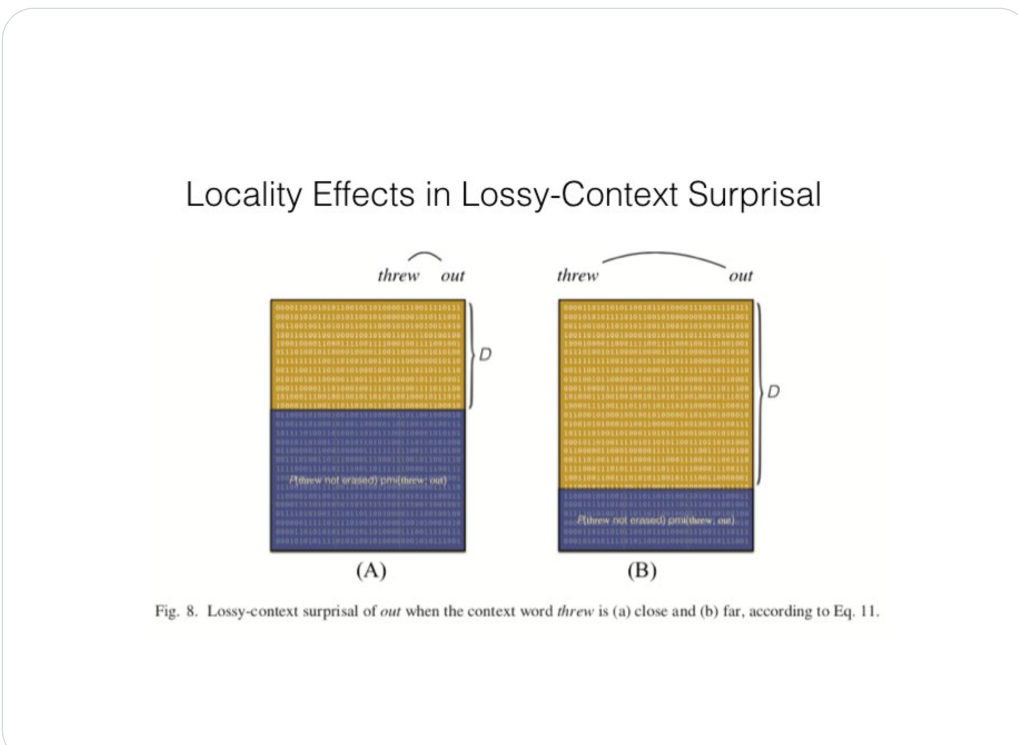


Fig. 8. Lossy-context surprisal of *out* when the context word *threw* is (a) close and (b) far, according to Eq. 11.

1 1

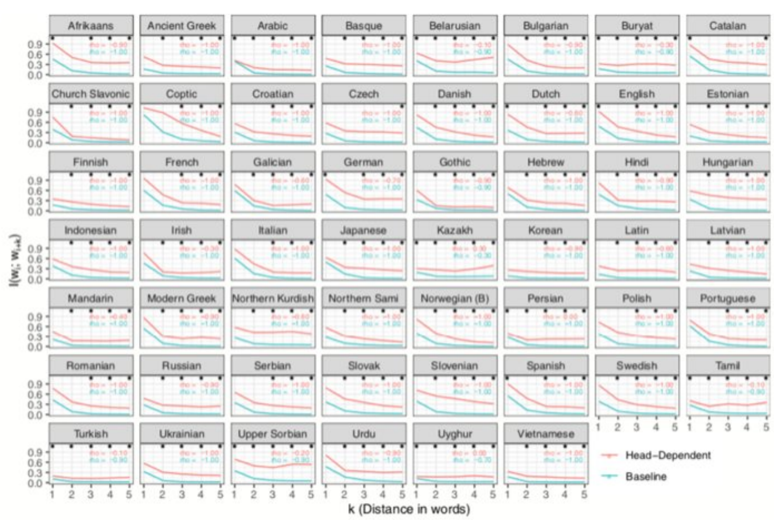
Richard Futrell @rljfutrell · Feb 27, 2020

We end up predicting a new, generalized form of dependency locality effect, which we call information locality. We predict extra processing difficulty whenever any words that *predict each other* are separated from each other—dependency locality is a special case of this.

1 2

Richard Futrell @rljfutrell · Feb 27, 2020

If people have preference for information locality in production, and/or if languages are shaped by a pressure for processing efficiency, then words that predict each other should be close to each other generally. We find this is the case in 54 Universal Dependencies corpora:



1 1 1

Richard Futrell @rljfutrell · Feb 27, 2020

So, to wrap up. Lossy-Context Surprisal extends the reach of information-theoretic models in linguistics. It is a resource-rational model, in the sense that it models rational behavior under resource constraints in the form of lossy memory.

1 3

Richard Futrell @rljfutrell · Feb 27, 2020

There are still memory effects in sentence processing that we do not explain, for example similarity-based interference. It remains to be seen whether or not these effects can be captured by lossy-context surprisal.

1 2

Richard Futrell @rljfutrell · Feb 27, 2020

Thanks for reading, and we hope our paper gives you ideas to test!

1 3