Crosslinguistic Word Orders Enable an Efficient Tradeoff of Memory and Surprisal (Abstract)

Michael Hahn
Stanford University
mhahn2@stanford.edu

Richard Futrell
University of California, Irvine
rfutrell@uci.edu

Memory limitations are well-established as a factor in human online sentence processing (Gibson, 1998; Lewis and Vasishth, 2005), and have been argued to account for crosslinguistic word order regularities. For example, the Performance–Grammar Correspondence Hypothesis of Hawkins (1994) holds that forms which are easier to produce and comprehend end up becoming part of the grammars of languages. We build on expectation-based models of language processing (Levy, 2008) and on the theory of lossy compression (Cover and Thomas, 2006) to develop a highly general information-theoretic notion of memory efficiency in language processing, in terms of a trade-off of surprisal and memory usage. We derive a method for estimating a lower bound on the memory efficiency of languages from corpora, and apply our method to corpora from 54 languages to test the idea that word order is structured to reduce processing effort under memory limitations. We find that word orders tend to support efficient tradeoffs between memory and surprisal, suggesting that word order rules are structured to enable efficient online processing.

Background Surprisal theory (Levy, 2008) posits that the processing effort on a word w_t in context $w_1 \dots w_{t-1}$ is proportional to the **surprisal** of the word in context:

$$S = -\log \mathrm{P}(w_t | w_1 \dots w_{t-1}). \quad (1)$$

Experimental work has confirmed that surprisal is a reliable and linear predictor of processing effort as reflected in reading times (Smith and Levy, 2013).

However, surprisal theory as presented above cannot in principle account for effects of memory limitations on online processing, because Equation 1 represents surprisal as experienced by an idealized listener who accurately remembers the entire history of previous words $w_{1\dots t-1}$. More realistically, human listeners deploy memory resources that maintain imperfect representations of the preceding context (Lewis and Vasishth, 2005; Futrell and Levy, 2017). If m_t is a listener's memory state after hearing $w_1 \dots w_{t-1}$, then the true surprisal experienced by the listener will be:

$$S_M := -\log_2 \mathrm{P}(w_t | m_t), \quad (2)$$

which must be larger than Eq. 1 on average (Cover and Thomas, 2006).

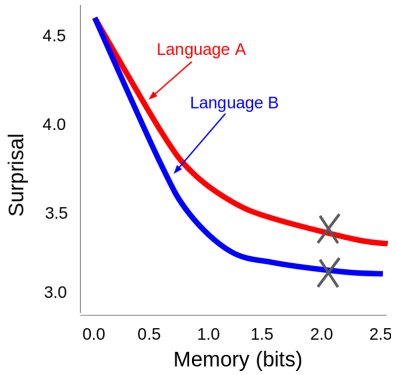

Figure 1: Conceptual tradeoff between memory and surprisal for two languages. In Language A (blue), a listener storing 1 bit can achieve average surprisal 3.5, while the same level of surprisal requires 2 bits of memory for a listener in Language B (red).

Memory–surprisal tradeoff. These considerations imply a *tradeoff between memory and surprisal*: A listener maintaining higher-precision memory representations m_t will, on average, incur lower surprisal, at the cost of higher memory load. The idea of the memory-surprisal tradeoff is visualized in Fig. 1: for each desired level of average surprisal, there is a minimum number of bits of information which must be stored about context. The shape of the trade-off is determined by the language, and in particular its word order: some languages enable more efficient trade-offs than others by forcing a listener to store more bits in memory to achieve the same level of average

Proceedings of the Society for Computation in Linguistics (SCiL) 2020, pages 468-469.
New Orleans, Louisiana, January 2-5, 2020

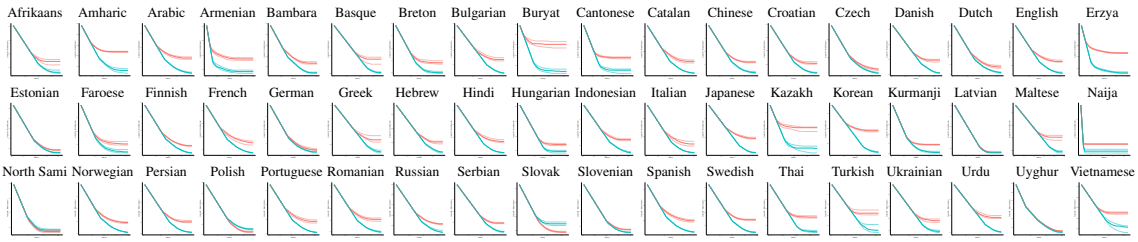

Figure 2: Tradeoffs between memory (x axis) and surprisal (y axis) in 54 languages, for real orderings (blue) and counterfactual baseline grammars (red). We provide 95% confidence bands for different model runs on the real languages, and for the median across different baseline grammars.

surprisal.

Theoretical Results In Theorem 1 below, we derive a bound on the memory-surprisal tradeoff curve which can be easily estimated from corpora. Let I_t be the conditional mutual information between words that are t steps apart, conditioned on the intervening words:

$$\mathrm{I}_t := \mathrm{I}[w_t, w_0 | w_{1\ldots t-1}].$$

This quantity measures how much predictive information the word t steps in the past contains about the current word.

Theorem 1. *Let T be a positive integer, and consider a listener using at most $\sum_{t=1}^{T} t\,\mathrm{I}_t$ bits of memory on average. Then this listener will incur average surprisal at least $\mathrm{H}[w_t|w_{<t}] + \sum_{t>T} \mathrm{I}_t$.*

The theorem allows us to estimate the extra surprisal associated with each amount of memory capacity for a language. The quantities I_t can be estimated as the difference between the cross-entropy of language models that have access to the last $t-1$ or t words. Given such estimates of I_t, we estimate tradeoff curves as in Figure 1 by tracing out $T = 1, 2, \ldots$.

Experimental Results We tested whether word orders as found in natural language grammars provide efficient memory-surprisal tradeoffs. To this end, we compared corpora of real languages against hypothetical reorderings of those languages under random baseline grammars. We used treebanks of 54 languages from Universal Dependencies 2.3 (Nivre et al., 2018).

For each language, we constructed counterfactual word order rules by adapting the methodology of Gildea and Temperley (2010) to Universal Dependencies: For each syntactic relation (subject, object, ...) used in the treebank annotation, we randomly sampled its position relative to the head and other of siblings. For each language and each such set of rules, we reordered the treebank according to these counterfactual word order rules.

For each language and its counterfactually ordered versions, we estimated the memory-surprisal tradeoff (Theorem 1) using an LSTM recurrent neural language model, considering all integers $T = 1, \ldots, 20$. Hyperparameters were tuned, for each language, to minimize average cross-entropy on counterfactual versions, introducing a conservative bias against our hypothesis.

Tradeoff curves are shown in Figure 2. In 50 out of 54 languages, the observed orderings led to more favorable tradeoffs than 50% of the counterfactual orderings ($p < 0.0001$; Exceptions: Latvian, North Sami, Polish, and Slovak).

Taken together, our results suggest that, across languages, word order in part reflects pressures towards efficient online processing under memory limitations.

References

Thomas M. Cover and J.A. Thomas. 2006. *Elements of Information Theory*. John Wiley & Sons, Hoboken, NJ.

R Futrell and R Levy. 2017. Noisy-context surprisal as a human sentence processing cost model. In *EACL*.

E. Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

D Gildea and D Temperley. 2010. Do Grammars Minimize Dependency Length? *Cognitive Science*.

JA Hawkins. 1994. *A performance theory of order and constituency*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

R Lewis and S Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*.

Joakim Nivre et al. 2018. Universal dependencies 2.3.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.