Crosslinguistic Word Orders Enable an Efficient Tradeoff of Memory and Surprisal

Michael Hahn (Stanford), Judith Degen (Stanford), Richard Futrell (UCI)
mhahn2@stanford.edu

Online memory limitations are well-established as a factor impacting sentence processing and have been argued to account for crosslinguistic word order regularities. Building off expectation-based models of language processing (Hale, 2001; Levy, 2008; Futrell & Levy, 2017), we provide an information-theoretic formalization of these memory limitations. We introduce the idea of a **memory-surprisal tradeoff**: comprehenders can achieve lower average surprisal per word at the cost of storing more information about past context. We show that the shape of the tradeoff is determined in part by word order. In particular, languages will enable more efficient tradeoffs when they exhibit **information locality**: when predictive information about a word is concentrated in the word's recent past (Futrell & Levy, 2017). We show evidence from corpora of 46 real languages showing that languages allow for more efficient memory-surprisal tradeoffs than random baseline word order grammars. **Theoretical results.** 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: 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 surprisal. We derive the precise form of the memory-surprisal tradeoff in Theorem 1 (Appendix). We demonstrate that languages with stronger information locality lead to more favorable tradeoffs, enabling listeners to incur lower surprisal at the same level of memory load. **Experiment 1:** We analytically calculated the comprehender's memory-surprisal tradeoff for the artificial languages from Fedzechkina, Chu and Jaeger (2017). In that work, learners were given input consistent with two possible word order grammars: one favoring short dependencies and the other favoring long dependencies. Learners consistently had a bias to infer the grammar favoring short dependencies. We find that the grammar favored by learners enables more efficient memory-surprisal tradeoffs than the one dispreferred by learners.

Experiment 2: We investigated whether word orders as found in natural language grammars optimize the comprehender's memory-surprisal tradeoff by comparing corpora of real languages against hypothetical reorderings of those languages under random baseline grammars. *Data* We used treebanks of 46 languages, mostly from Universal Dependencies. *Baseline Orders* For each language, we constructed counterfactual word order rules by sampling, for each syntactic relation (*subject*, *object*, …) used in the treebank annotation, its position relative to the head and other of siblings (similar to Gildea & Temperley, 2010). For each language and each such set of rules, we reordered the treebank according to these counterfactual word order rules. *Estimating Memory* For each language and its counterfactually ordered versions, we estimated the listener's surprisal-memory tradeoff (Theorem 1) using an LSTM recurrent neural language model. We encoded the input by concatenating POS and morphological information provided in the treebanks. Experiments with full word forms on high-resource languages led to similar results. *Results* Tradeoff curves are shown in Figure 2. In 43 out of 46 languages, the observed orderings led to more favorable tradeoffs than 50% of the counterfactual orderings (Exceptions: Armenian, North Sami, Polish). 33 of the languages produce tradeoffs more favorable than 90% of the counterfactual orderings.

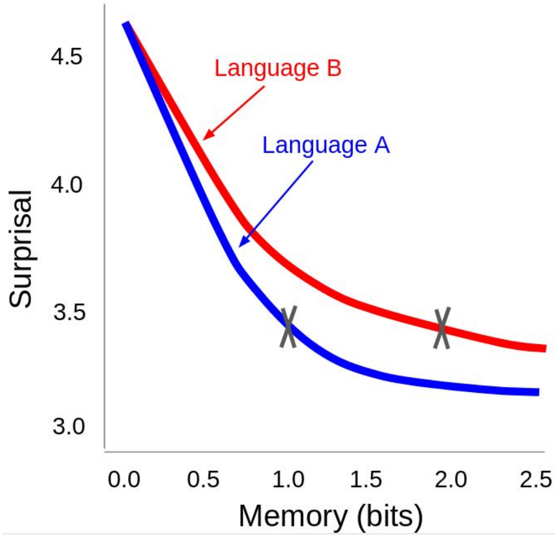

Figure 1: Conceptual tradeoff between memory and surprisal for two languages. In Language A (blue), a comprehender storing 1 bit can achieve average surprisal 3.5, while the same level of surprisal requires 2 bits of memory for a comprehender in Language B (red).

Theorem 1. *For each positive integer t, define $I_t := I[X_t, X_0|X_{1\ldots t-1}]$, i.e., the mutual information between words at distance t, controlling for redundancy with the intervening words. Let T be a positive integer, and consider a comprehender using at most $\sum_{t=1}^{T} t I_t$ bits of memory on average. Then this comprehender will incur average surprisal at least $H[X_t|X_{<t}] + \sum_{t>T} I_t$.*

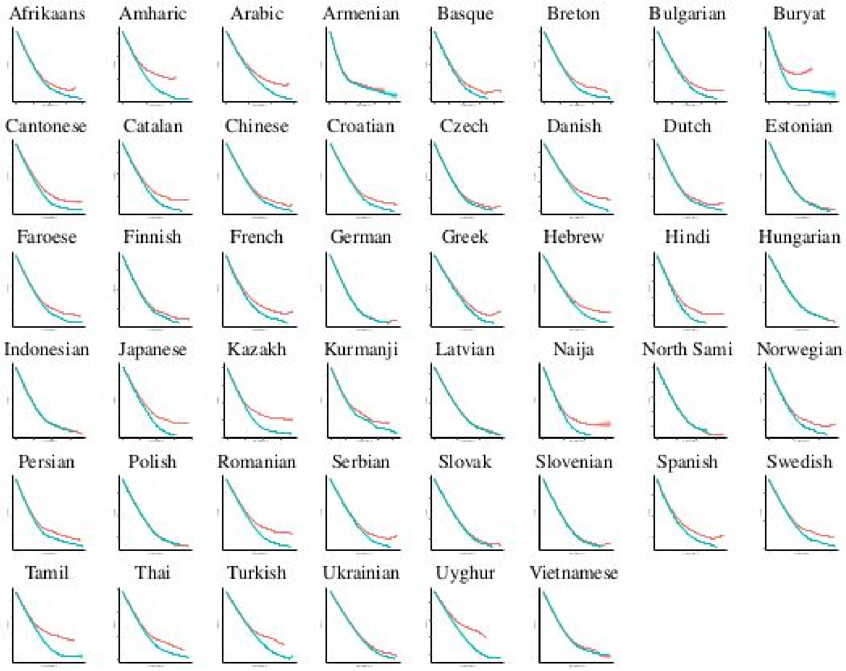

Figure 2: Estimated tradeoffs between memory and surprisal in 46 languages. Blue curves represent actual orderings; red curves represent the average over counterfactual orderings. Axes are as in Figure 1.