Spillover as Rational Processing Delay in Sentence Comprehension

Kohei Kajikawa and Ethan Gotlieb Wilcox

Spillover effects in sentence comprehension are rarely the primary focus and are typically treated as a byproduct of modular processing (e.g., in Bartek, Lewis, Vasishth, & Smith, 2011). Instead, we hypothesize that spillover can arise from rational decisions to delay processing a word under memory constraints. Adopting rational analysis (Anderson, 1990), we assume a comprehender seeks to minimize expected processing costs. Sentence comprehension is incremental at a macro level but not necessarily fully incremental at a micro level: comprehenders may defer integrating a word w_i until subsequent input arrives. Such delays can reduce the surprisal of w_i by allowing interpretation in a richer context, but they also incur storage costs, because w_i must be held in working memory until it is integrated. We develop a rational-delay framework in which spillover reflects a tradeoff between the benefit of such *backward* surprisal reduction and storage costs.

Proposal. The backward reduction in surprisal of w_i, or Δsurp_{w_i}, is the extent to which knowledge of future word(s), $\boldsymbol{w}_f$, reduces the information content of w_i, given past context $\boldsymbol{w}_{<i}$. It is equivalent to the contextual pmi between w_i and future words $\boldsymbol{w}_f$ (eq. 1). We quantify the *expected* backward reduction in two ways, which correspond to different assumptions about lexical identification: (i) $\mathbb{E}$pmi, the expectation over $\boldsymbol{w}_f$ given w_i and $\boldsymbol{w}_{<i}$ (eq. 2), and (ii) MI, the expectation over both $\boldsymbol{w}_f$ and w_i given $\boldsymbol{w}_{<i}$. This corresponds to a processing context where an expectation is taken before w_i is lexically identified (eq. 3). Here, we approximate the future context $\boldsymbol{w}_f$ with the immediately following word w_{i+1}. For the cost term, we leave the exact functional form unspecified but assume that the storage cost increases additively with the number of future time steps for which w_i must be retained. We treat processing decisions as approximately maximizing expected backward surprisal reduction minus a penalty for storage cost. Given that mutual information in language decays as a power law with distance (Lin & Tegmark, 2017) while storage cost grows with delay, this framework predicts that (a) higher benefit at w_i should be associated with greater spillover at w_{i+1}, (b) that spillover should be confined to a small number of subsequent words and (c) modulated by information locality of languages (Futrell, 2019). Here, we focus on testing prediction (a).

Methods. We analyze two English self-paced reading time (RT) datasets, Brown (Smith & Levy, 2013) and Natural Stories (NS; Futrell et al., 2021), which preclude parafoveal preview and word skipping. Word-by-word spillover is estimated with a generalized additive model (GAM) predicting RT from word length, frequency, and GPT-2 surprisal at w_i and at the three preceding positions. For each word, we define *relative* spillover as the contribution of the lag terms, obtained by summing their 10-fold cross-validated predictions. For NS we also derive a spillover estimate from a GAM using A-maze (Boyce & Levy, 2023), treated as a proxy for word-by-word processing load because the paradigm minimizes spillover. Figure 1 shows the distributions of spillover estimates; maze-based and lag-based methods of estimation are strongly correlated (Figure 2). We then predict spillover at w_{i+1} using $\mathbb{E}$pmi and MI at w_i, both computed from GPT-2 with GAMs.

Results & Discussion. Figure 3 shows a dissociation between the two benefit variants. Across both datasets, MI at w_i exhibits a significant, approximately monotonic *positive* relationship with the estimated relative spillover at w_{i+1}. By contrast, the effect of $\mathbb{E}$pmi is significant but non-monotonic: it increases as predicted at low values, but reverses at high values. This suggests that our current operationalization of the benefit term via $\mathbb{E}$pmi is entangled with other confounding properties: since $\mathbb{E}$pmi measures how strongly w_i constrains w_{i+1} once w_i is observed, higher $\mathbb{E}$pmi likely reflects stronger preactivation of w_{i+1} and thus shorter reading times at w_{i+1}. These findings provide qualified support for the view that spillover effects arise, in part, from rational processing delays that are sensitive to MI. Future work will refine the specification of the benefit and cost terms to derive more precise quantitative predictions and will examine typologically diverse languages to test the predicted limits on spillover range by information locality (e.g., SOV vs. SVO).

$$\Delta\mathsf{surp}_{w_i} := \mathsf{surp}(w_i \mid \boldsymbol{w}_{<i}) - \mathsf{surp}(w_i \mid \boldsymbol{w}_{<i}, \boldsymbol{w}_f)$$
$$= \mathsf{pmi}(w_i; \boldsymbol{w}_f \mid \boldsymbol{w}_{<i}) \qquad (1)$$

$$\mathbb{E}_{\boldsymbol{w}_f \mid \boldsymbol{w}_{<i}, w_i}\left[\Delta\mathsf{surp}_{w_i}\right] = \mathbb{E}_{\boldsymbol{w}_f \mid \boldsymbol{w}_{<i}, w_i}\left[\mathsf{pmi}(w_i; \boldsymbol{w}_f \mid \boldsymbol{w}_{<i})\right]$$
$$:= \mathbb{E}\mathsf{pmi} \qquad (2)$$

$$\mathbb{E}_{w_i \mid \boldsymbol{w}_{<i}}\left[\mathbb{E}_{\boldsymbol{w}_f \mid \boldsymbol{w}_{<i}, w_i}\left[\Delta\mathsf{surp}_{w_i}\right]\right] = I(W_i; \boldsymbol{W}_f \mid \boldsymbol{w}_{<i})$$
$$:= \mathsf{MI} \qquad (3)$$

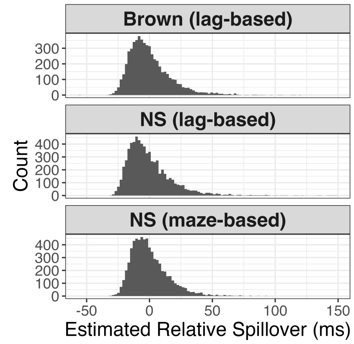

Figure 1: Distributions of estimated *relative* spillover.

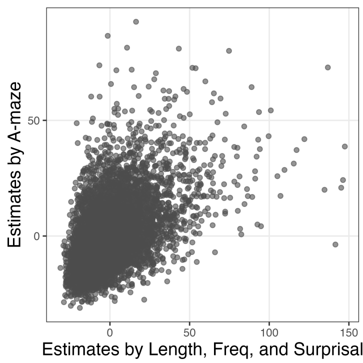

Figure 2: Relationship between two types of estimated spillovers in Natural Stories, either from word length, frequency, and surprisal (x-axis) or from A-maze (proxy for word-by-word processing, y-axis). The high correlation (r=.51) indicates that the lexical predictors are good-enough as approximation to word-by-word processing load.

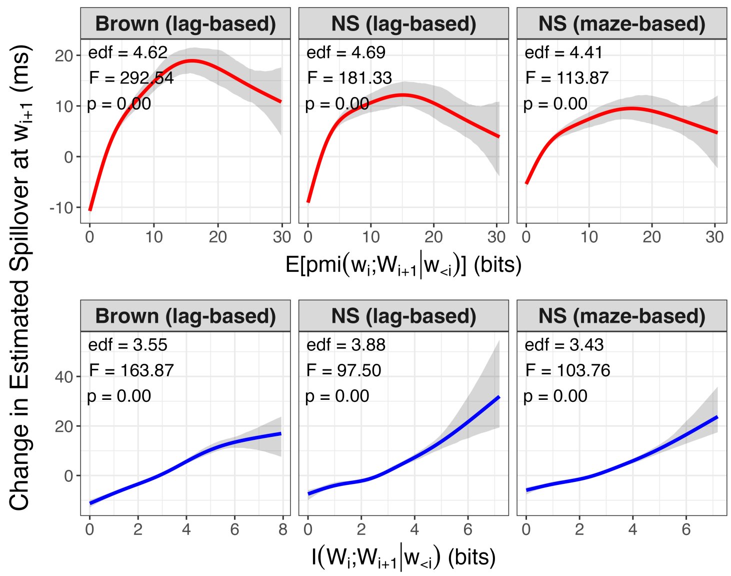

Figure 3: Relationship between $\mathbb{E}$pmi and MI at w_i and estimated relative spillover at w_{i+1}, based on lexical predictors and the maze data). The lines and ribbons show GAM-fitted smooths and their bootstrapped 95% confidence intervals. $\mathbb{E}$pmi and MI have a significant effect across datasets.

References

Anderson, J. R. (1990). *The adaptive character of thought*.

Bartek, B., Lewis, R. L., Vasishth, S., & Smith, M. R. (2011). In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1178–1198.

Boyce, V., & Levy, R. (2023). A-maze of Natural Stories: Comprehension and surprisal in the Maze task. *Glossa Psycholinguistics*, *2*(1).

Futrell, R. (2019). Information-theoretic locality properties of natural language. In *Proceedings of the First Workshop on Quantitative Syntax* (pp. 2–15). Paris, France: ACL.

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2021). The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, *55*, 63–77.

Lin, H. W., & Tegmark, M. (2017). Critical Behavior in Physics and Probabilistic Formal Languages. *Entropy*, *19*(7), 299.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.