Entropy- and Distance-Based Predictors From GPT-2 Attention Patterns Predict Reading Times Over and Above GPT-2 Surprisal

Byung-Doh Oh William Schuler

oh.531@osu.edu
github.com/byungdoh/attn_dist

The Ohio State University

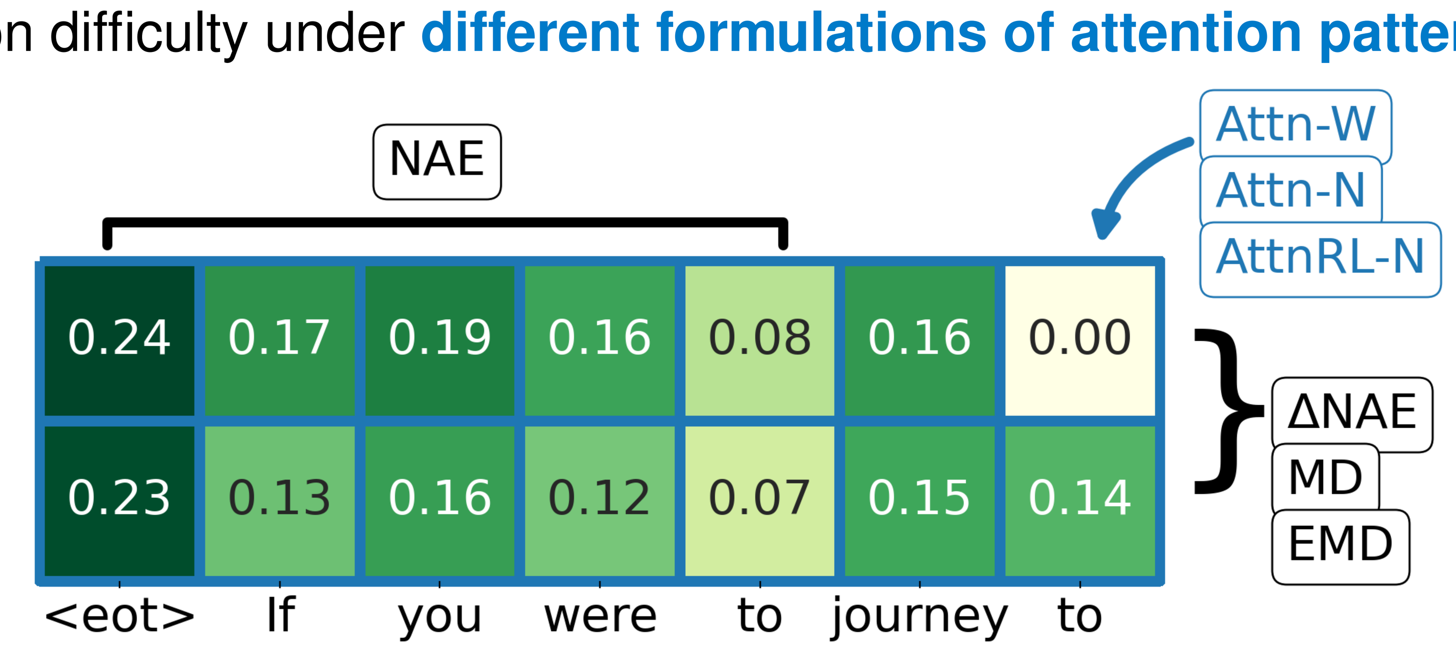
PAPER

Introduction

- There are empirical shortcomings of LM surprisal as expectation-based predictors of comprehension difficulty, such as underprediction of garden-path effects [9]
- As such, there are recent efforts to identify memory-based effects from LM representations
- For example, a connection has been made between Transformer self-attention weights and cue-based retrieval [7], but their entropy was not predictive over surprisal [8]
- Self-attention weights proper do not accurately reflect the importance of each token in context [2, 4]

Entropy- and Distance-Based Predictors

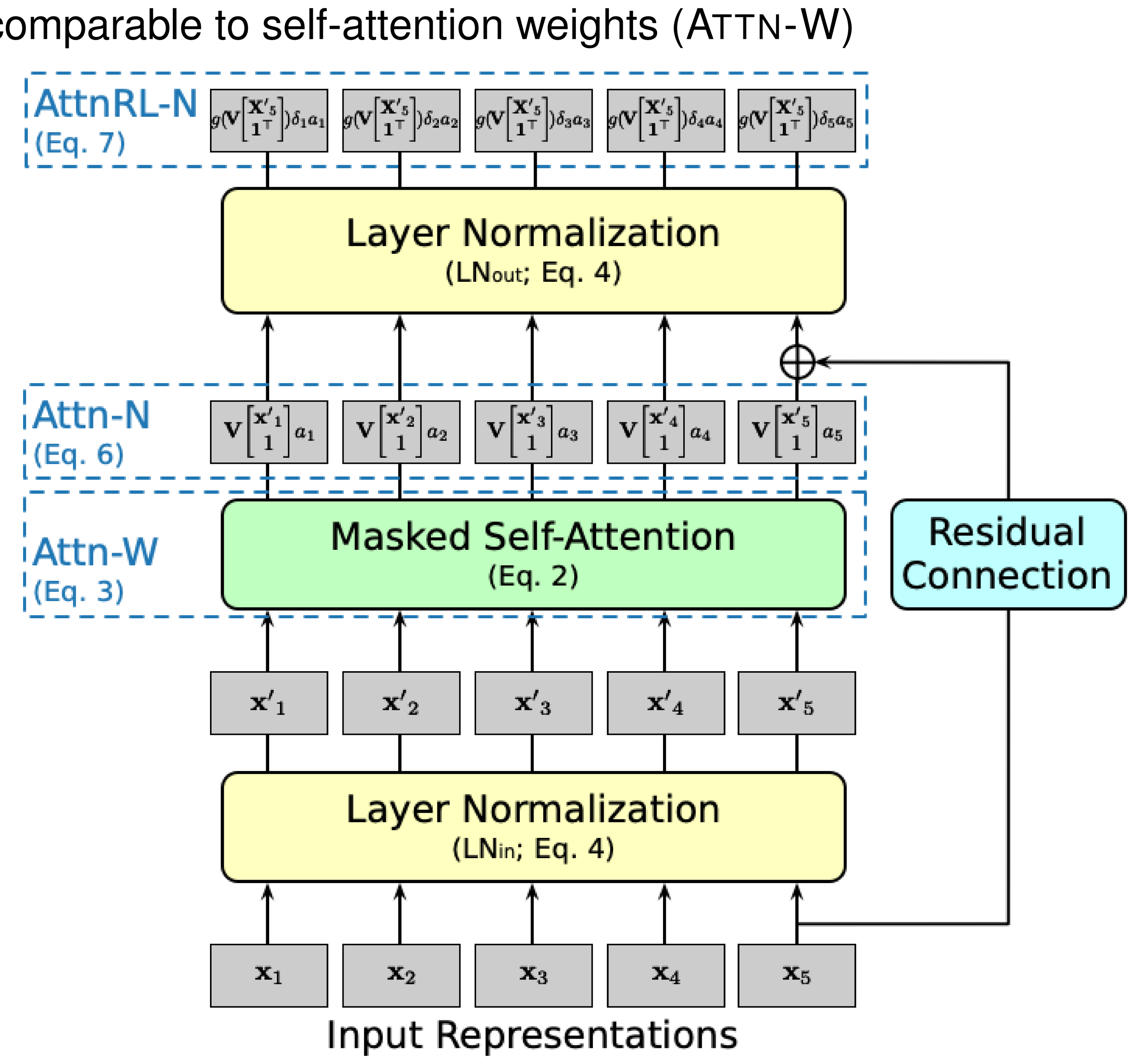
This work defines **entropy- and distance-based predictors** of comprehension difficulty under **different formulations of attention patterns**:



1. Normalized attention entropy (NAE): Entropy of normalized weights over $w_{1..i-1}$ divided by maximum entropy
2. ΔNormalized attention entropy (ΔNAE): Absolute value of change in NAE across consecutive timesteps
3. Manhattan distance (MD): 1-norm of difference in attention weight vectors across consecutive timesteps
4. Earth Mover's Distance (EMD): Minimum amount of 'work' necessary to transform the current attention weight vector to the next

Formulations of GPT-2 [6] Attention Patterns

- Linear nature of the computations in a self-attention block allows the aggregation of representations to be deferred [4, 5]
- Vector norms are normalized to yield weights (ATTN-N, ATTNRL-N) that are comparable to self-attention weights (ATTN-W)



Evaluation on Human Reading Times

- Evaluation on the Natural Stories Corpus [1] and the Dundee Corpus [3]
- Baseline: low-level predictors, unigram surprisal, and GPT-2 surprisal
- Predictors of interest calculated from topmost attention heads of GPT-2

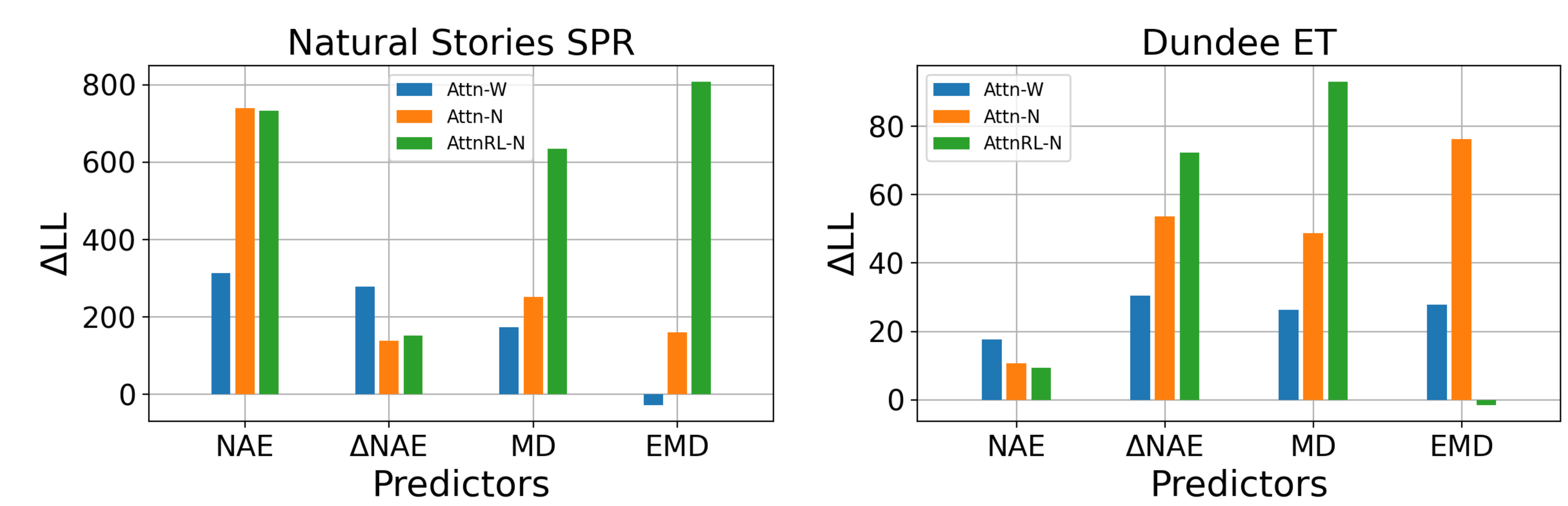


Figure 1: Improvements in regression model log-likelihood from including each predictor on the exploratory (dev) partition.

Corpus	Predictor	Effect Size (p-value)
Natural Stories	ATTN-N+NAE	6.87 ms ($p < 0.001$)
	GPT2SURP	2.56 ms
	ATTNRL-N+MD	6.59 ms ($p < 0.001$)
	GPT2SURP	2.82 ms
Dundee	ATTN-N+NAE	N/A (n.s.)
	GPT2SURP	4.22 ms
	ATTNRL-N+MD	1.05 ms ($p < 0.001$)
	GPT2SURP	3.81 ms

Table 1: Effect sizes per standard deviation on the held-out (test) partition.

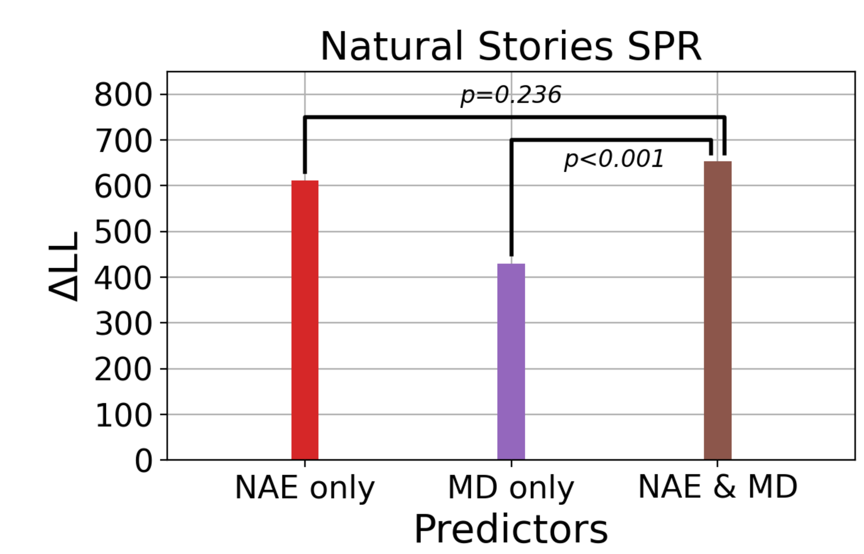


Figure 2: Improvements in log-likelihood on the held-out (test) partition.

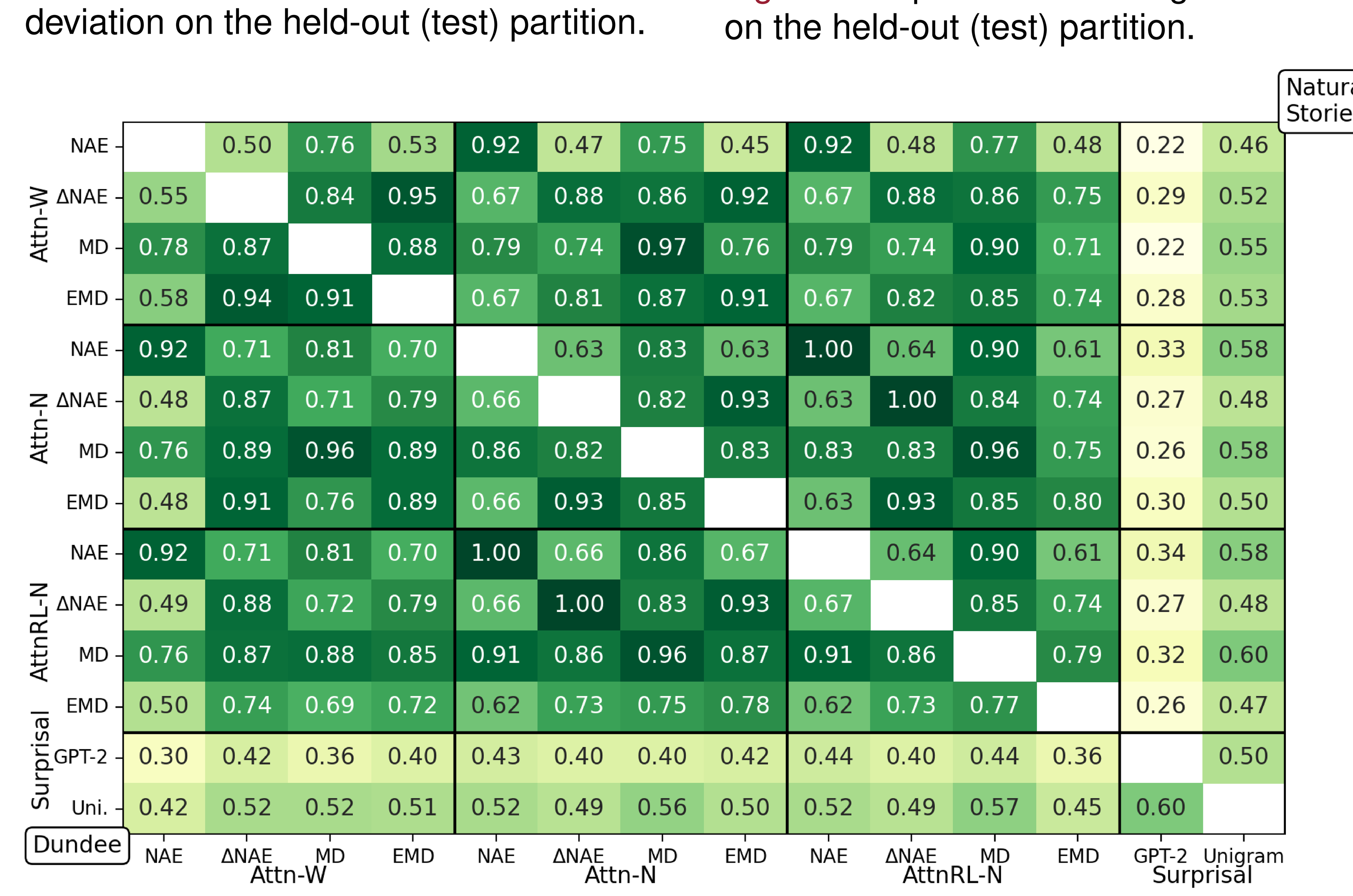


Figure 3: Pearson correlation coefficients between predictors.

Conclusion

Results show robust effects of Transformer attention-based predictors in predicting reading times of broad-coverage naturalistic data

References

[1] R. Futrell, E. Gibson, H. J. Tily, I. Blank, A. Vishnevetsky, S. Piantadosi, and E. Fedorenko. The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 2021.

[2] S. Jain and B. C. Wallace. Attention is not explanation. In *Proc. NAACL*, 2019.

[3] A. Kennedy, R. Hill, and J. Pynte. The Dundee Corpus. In *Proc. ECEM*, 2003.

[4] G. Kobayashi, T. Kuribayashi, S. Yokoi, and K. Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proc. EMNLP*, 2020.

[5] G. Kobayashi, T. Kuribayashi, S. Yokoi, and K. Inui. Incorporating residual and normalization layers into analysis of masked language models. In *Proc. EMNLP*, 2021.

[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

[7] S. H. Ryu and R. L. Lewis. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proc. CMCL*, 2021.

[8] S. H. Ryu and R. L. Lewis. Using Transformer language model to integrate surprisal, entropy, and working memory retrieval accounts of sentence processing. In *35th Annual Conference on Human Sentence Processing*, 2022.

[9] M. van Schijndel and T. Linzen. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 2021.