Naturalistic Reading Time Data Support Information Locality

Hailin Hao[1], Himanshu Yadav[2], Elsi Kaiser[1]
[1]University of Southern California
[2]Indian Institute of Technology Kanpur
{hailinha, emkaiser}@usc.edu himanshu@iitk.ac.in

Abstract

Both expectation and locality have been established as key factors in characterizing incremental sentence processing difficulty. Here we investigate the less explored question whether and how expectation and locality interact with each other, using data from naturalistic reading time corpora. We found that the data support the Information Locality hypothesis: Strong expectations can enhance locality effects. We argue that future theory-building in sentence processing should therefore take into consideration both expectation and locality, as well as their potential interaction.

1 Introduction

Characterizing incremental processing difficulty has been a key goal of psycholinguistic research. According to expectation-based theories (e.g., Levy, 2008) the processing difficulty of a word depends on its predictability given the preceding context. Words with higher contextual probability are easier to process. By contrast, locality-based theories (e.g., Gibson, 1998) hold that as the linear distance between two co-dependents increases, the cost of retrieval at the second co-dependent would be higher as the first co-dependent might have undergone decay or interference. Although both theories have received abundant support, it remains an open question how they can be theoretically and empirically reconciled. Research on the interactions of expectations and locality is therefore crucial for building a complete theory of sentence processing, but so far evidence has been limited. To shed light on this issue, we used data from naturalistic reading time (RT) corpora in English to provide broad-coverage evaluations of two hypotheses (Information Locality vs. Prediction Maintenance) that make divergent predictions regarding how expectation and locality interact. The *Information Locality* hypothesis (Futrell, 2019; Futrell, Gibson, and Levy, 2020) states that words that highly predict each other are constrained to be close to each other. Thus, locality should be stronger when expectation is high. In contrast, the *Prediction Maintenance* hypothesis (Husain, Vasishth, & Srinivasan) predicts that strong expectations can cancel locality effects. When two co-dependents highly predict each other, the cost of retrieval at the later co-dependent will be lower, considering that it might already be preactivated.

2 Corpora Evaluations

2.1 Material

We included four datasets in our analysis: Natural Stories Self-Paced Reading (Futrell et al., 2020), Natural Stories A-Maze (Boyce and Levy, 2022), Brown Self-Paced Reading (Smith and Levy, 2013), and Provo Eye-Tracking (Luke and Christianson, 2018).

2.2 Methods

We first parsed the texts from the corpora using the Stanford Neural Dependency Parser (Chen and Manning, 2014), if parses were not provided by the corpora, and extracted all dependencies. Following Futrell (2019), we formalized expectation as Head-Dependent Mutual Information (HDMI; equation in 1), and locality as Dependency Length (DL; number of intervening words). We estimated *p(h,d)* (i.e., frequency that pair occurs together in a dependency) and *p(h)p(d)* (i.e., total frequency of the head and the dependent) for any given pair of word categories using the more fine-grained part-of-speech tags from UD (Nivre et al., 2016).

$$\text{HDMI} = \log \frac{p(h,d)}{p(h)p(d)} \quad (1)$$

We fitted linear mixed effects models on the log transformed RTs of the second co-dependents, with

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 273-274.
Irvine, California, June 27-29, 2024

DL, HDMI, and their interaction, and two word-level factors (word length and frequency) as fixed effects, all scaled. For eye-tracking, first-path duration and total reading times were analyzed. We first analyzed each dataset separately and then ran a meta-analysis on all datasets (for eye-tracking, only total viewing times were included). With the aggregated datasets, we also ran exploratory analyses based on head direction (according to UD standards) and whether the dependency involves only core arguments (i.e., verbs, nouns).

2.3 Results

We found that all datasets show significant locality effects, whereby RTs increase as DL increases, i.e., RTs are slower when the dependency length is longer. Two datasets also show significant expectation effects, whereby RTs decrease as HDMI increases. More importantly, two datasets support show an significant interaction between DL and HDMI, whereby the effects of DL become more positive (i.e., leading to more slowdowns) when HDMI is higher, which supports the *Information Locality* hypothesis. Moreover, no datasets support the *Prediction Maintenance* hypothesis. Our meta-analysis show significant locality and expectation effects, and an interaction that supports the *Information Locality* hypothesis. The exploratory analyses (head direction, core arguments) show locality effects, and the majority show expectation and *Information Locality* effects. Again, no analyses support the *Prediction Maintenance* hypothesis.

3 Conclusion

Using data from naturalistic RT corpora, we provided broad-coverage evidence for the Information Locality hypothesis: Locality effects are enhanced with high expectations. Our results show that naturalistic RT corpora can provide a good source of evidence that corroborates controlled experimentation and can be used to test multiple theoretical predictions against each other. The current study is limited to English, but effects of locality and expectation and their interaction profiles could potentially differ from language to language, due to differences in syntax. Although the availability of cross-linguistic datasets is limited, we emphasize the need for more cross-linguistic work in this area.

References

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

Edward Gibson. 1998. Linguistic complexity: Locality of Syntactic dependencies. *Cognition*, *68*(1):1–76.

Richard Futrell. 2019. Information-theoretic locality properties of natural language, In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 2–15, Paris, France. Association for Computational Linguistics

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP, 2014)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal Dependencies v1: A multilingual treebank collection. 2016. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portoroz, Slovenia. European Language Resources Association.

Samar Husain, Shravan Vasishth, and Narayanan Srinivasan. 2015. Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus. *Journal of Eye Movement Research*, *8*(2):1–12.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko. 2022. The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation 55*: 63–77.

Veronica Boyce and Roger Levy. 2023 A-maze of Natural Stories: Comprehension and surprisal in the Maze task. *Glossa Psycholinguistics 2*:1-34.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition, 128*(2):302–319.

Steven G. Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods 50*:826–833.

Richard Futrell, Edward Gibson, and Roger Levy R. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science, 44*:e12814.