Naturalistic Reading Time Data Support Information Locality

Hailin Hao (hailinha@usc.edu)
Department of Linguistics, University of Southern California
Los Angeles, CA 90089-1693 USA

Himanshu Yadav (himanshu@iitk.ac.in)
Department of Cognitive Science
Indian Institute of Technology Kanpur
Kanpur, Uttar Pradesh 208016, India

Elsi Kaiser (emkaiser@usc.edu)
Department of Linguistics, University of Southern California
Los Angeles, CA 90089-1693 USA

Abstract

Both prediction and working memory constraints have been established as key factors in characterizing incremental sentence processing difficulty. Here we investigate the less explored question: Whether and how predictive expectation and working memory interact with each other using data from naturalistic reading time corpora. We provide broad-coverage evaluations of two hypotheses that make divergent predictions regarding the interaction of expectation and memory constraints: the *Information Locality* and *Prediction Maintenance* hypotheses. We first confirmed the predictions of both expectation- and working memory-based theories. Regarding their interactions, we find support the Information Locality hypothesis: Strong mutual predictability can enhance locality effects. We argue that future theory building in sentence processing should therefore take into consideration both prediction and memory constraints, as well as their potential interaction.

Keywords: sentence processing; probabilistic expectations; working memory; information theory; naturalistic methods; corpus analysis

Introduction

The processing behavior exhibited by humans on a sentence comprehension task is assumed to reflect some important cognitive processes operating in real time. A key empirical property of sentence comprehension is that the processing difficulty varies from word-to-word in a sentence. What are the cognitive constraints that determine the processing difficulty of a word given its preceding context? Two broad classes of theoretical proposals exist: the *expectation-based* theories (e.g., Hale, 2001; Jurafsky, 2003; Levy, 2008) and the *working memory-based* theories (e.g., Gibson, 1998; Lewis & Vasishth, 2005). The *expectation-based* theories assume that the processing difficulty at a word reflects the cost of encountering a less-expected structure, the lesser the expectation of the word given the preceding context, the higher is the processing difficulty. A well-established expectation-based theory — the Surprisal theory (Levy 2008) — proposes that the processing difficulty of a word *w* is the negative log of its conditional probability given the preceding context *c*, as formulated in (1). This theory has received considerable support from controlled reading experiments (e.g., Levy & Keller, 2013; Linzen & Jaeger, 2016; Vasishth & Drenhaus, 2011; Wu, Kaiser, & Vasishth, 2018) and also from reading times of arbitrary words in corpora (Shain et al., 2022; Smith & Levy, 2013; Wilcox et al., 2023).

(1) processing difficulty $\propto -\log P(w|c)$

By contrast, *working memory-based* theories generally predict locality effects, whereby increased distance between two co-dependents incurs processing difficulty. For example, in (2), the subject 'doctor' and the verb 'diagnosed' form a subject-verb dependency; while in (2a) they are adjacent, in (2b) they are separated by a relative clause. The working memory-based theories predict that the processing difficulty at the second co-dependent *diagnosed* should be higher in (2b) compared to (2a). This is because to successfully integrate 'diagnosed', one must retrieve its subject 'doctor' from their working memory, and as the linear distance between two co-dependents increases, the memory representation of the early co-dependent becomes weaker due to decay (Gibson, 1998) or interference (Lewis & Vasishth, 2005) from other material in memory, making it more difficult to be retrieved. The theory has received ample empirical support from controlled experimentation and corpora as well (e.g., Bartek et al., 2011; Demberg & Keller, 2008; Grodner and Gibson, 2005; Shain et al., 2016).

(2) a. The **doctor diagnosed** my neighbor.
b. The **doctor** who lived in downtown **diagnosed** my neighbor.

Despite the reasonable empirical coverage of these two classes of theories, neither of these can explain the full range of empirical phenomena observed on sentence comprehension tasks. Expectation and memory-based accounts have been invoked to explain complementary datasets. For example, the locality effect and the similarity-based interference in English are attributed to working memory constraints, but the anti-locality effect in German and Hindi are attributed to predictive expectation. But how do predictive expectation and limited working memory interact? Experimental studies are lacking to explicit address this question. To build a complete, unified theory of sentence processing, it is important to empirically investigate how

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

predictive expectation and working memory constraints interact with each other. The current study uses data from naturalistic RT corpora to test two competing hypotheses about how expectation and memory constraints interact: *Information Locality* hypothesis (Futrell, 2019; Futrell, Levy, Gibson, 2020) and *Prediction Maintenance* hypothesis (Husain, Vasishth, & Srinvasan, 2014). We review them one by one.

Information Locality

Information Locality holds that high mutual predictability can enhance locality effects (Futrell, 2019), as in Figure 1. This is a consequence of the Lossy-Context Surprisal theory (Futrell, Gibson, & Levy, 2020; Hahn et al., 2022). Lossy-Context Surprisal is built on Surprisal theory but revised its assumption that context *c* should be perfectly retained and unbounded. Instead, Lossy-Context Surprisal proposed that probabilistic expectations should be constrained by memory limitations such that the context *c* is not perfectly retained. Under Lossy-Context Surprisal, the processing difficulty of a word *w* depends on a *lossy* or *noisy* representation of context *c*', where next word predictability is determined by a posterior $P(w|c')$ over possible non-veridical contexts, as formalized in (3). One way that context can distort to a non-veridical memory representation is via erasure noise, a very common noise model in information theory (Cover & Thomas, 2006), whereby a word in the preceding context is probabilistically deleted. Hahn et al. (2022) implemented a more sophisticated model of erasure noise. The general idea here is that context is not perfectly retained some parts of it may get (partially) deleted. Therefore, if two co-dependents that highly predict each other are separated by extra material in linear order (as in a long-distance dependency), they will not be able to predict each other because by the time the comprehender gets to the second codependent, the first one has been partially forgotten. Similar hypotheses were made by Levy and Keller (2013) and Vasishth and Drenhaus (2011), where they proposed that expectation effects will dominate only when working memory constraints are low.

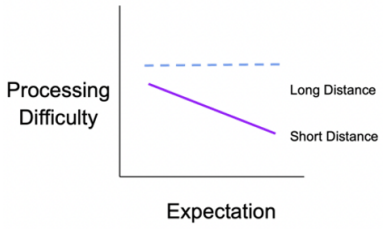

Figure 1: Conceptual predictions of *Information Locality*.

(3) processing difficulty $\propto -\log P(w|c') = -\log \sum_c P(w|c)P(c|c')$

Futrell (2019) showed that word order frequency data in written corpora follow *Information Locality* principles. However, there has not been any human behavioral, real-time processing evidence that supports this prediction.

Prediction Maintenance

Prediction Maintenance predicts that strong predictive expectation can cancel locality effects, as in Figure 2. One of the experiments in Husain, Vasishth, and Srinvasan (2014) manipulated the expectation strength between the two co-dependents in object-verb dependencies in Hindi, and the distance between the two co-dependents. They found strong evidence for a main effect of expectation, as well as in interaction between expectation and distance. While in the low expectation conditions, increased distance caused slowdowns (e.g., locality effects), the opposite pattern was observed in the high expectation conditions. According to the authors' hypothesis, in the high expectation conditions, considering that the second co-dependent is so predictable given the first co-dependent, it could already be pre-activated and integrated when the first co-dependent is processed. This way, there will not be any retrieval cost at second co-dependent. Such retrievals are resource-intensive according to memory-based theories (Gibson, 1998; Lewis & Vasishth, 2005) and responsible for the locality effects in the low expectation conditions.

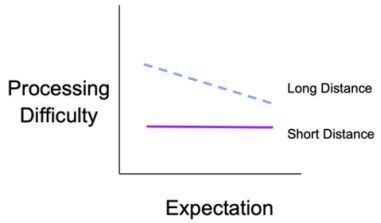

Figure 2: Conceptual predictions of *Prediction Maintenance*.

Empirical evidence for the *Prediction Maintenance* hypothesis, however, is limited. For example, Safavi, Husain, & Vasishth (2016) tested a similar construction in Persian. Despite finding evidence for both expectation and locality effects independently, they found no indication of the two factors interacting with each other. Similarly, testing *wh*-dependencies in Mandarin Chinese, Ming and Wang (in prep) also only observed main effects of expectation and locality but no interaction.

Current Study

As reviewed above, existing empirical evidence regarding how locality and expectation interact is both mixed and limited. The current study therefore aims to provide novel evidence on this issue with data from naturalistic RT corpora.

The use of naturalistic RT corpora has several advantages. First, it can provide large-scale broad-coverage evaluations of theories. Controlled experimental studies usually focus on one or two specific constructions and analyze the behavioral or neural reactions associated with one particular region. The use of RT corpora, by contrast, can extend theory testing to

random constructions and random words in the corpora, therefore making the evidence more robust.

Second, the methodology of naturalistic RT corpora is more ecologically valid. Instead of asking participants to read artificially constructed stimuli out of context, in naturalistic RT corpora, participants read naturally occurring texts from stories, newspapers, novels, etc., mimicking everyday language use. Previous work has used such corpora to evaluate the predictions of expectation-based theories (e.g., Demberg & Keller, 2008; Shain et al., 2023) and also memory-based theories (e.g., Demberg & Keller, 2008; Shain et al., 2016; see also Tung & Brennan, 2023, who used naturalistic ERP corpora), but here we test how expectation and memory constraints interact.

Corpus Evaluations

Summary of Datasets

We included four naturalistic datasets in English in our analysis. Below are descriptions of each dataset.

Natural Stories SPR. The Natural Stories SPR dataset (Futrell et al., 2021) contains self-paced reading data from 178 participants, who read 10 naturally occurring narrative or non-fiction pieces. The authors modified the texts so that they include to include low frequency words and syntactic constructions but are still perceived as natural. The dataset contains 485 sentences, 10,256 words, and 1,013,377 responses.

Natural Stories A-Maze. The Natural Stories Maze dataset (Boyce & Levy, 2023) contains Maze task responses from 95 participants using the same materials as in the Natural Stories SPR dataset above. These authors used A-Maze (Boyce, Futrell, & Levy, 2020) to generate high quality forced-choice alternatives for each word in the texts. The dataset contains 485 sentences, 10,256 words, and 97,527 responses.

Brown SPR. The Brown SPR dataset (Smith & Levy, 2013) contains self-paced reading data from 35 participants. The stimuli are short passages from the Brown dataset of American English (Kucera & Francis, 1967). The dataset contains 450 sentences, 7,188 words, and 136,907 responses.

Provo ET. The Provo ET dataset (Luke & Christianson, 2018) contains eye-tracking data from 84 participants who read 55 short passages from various online sources. The dataset contains134 sentences, 2,745 words, and 213,224 distinct fixations to word regions on the screen.

Methods

Dependency Parsing and Extraction. We performed our analysis on extracted dependencies (i.e., head-dependent pairs) from the texts of the datasets. An example UD parse is in Figure 3, which contains a head-final *nsubj* relation (subject-verb dependency), a head-initial *obj* relation (verb-object dependency), and a head-final acl:relcl (RC head-RC verb dependency). Natural Stories SPR provided hand-corrected parses in the Universal Dependency (UD; Nivre et al., 2016) style. Therefore, for Natural Stories SPR and Natural Stories Maze, we directly extracted all dependencies from the provided parses. For Brown SPR and Provo ET, we used the Python implementation of the Stanford neural dependency parser (Qi et al., 2018) to obtain dependency parses of the raw texts and then extracted all dependencies, in the UD style as well.

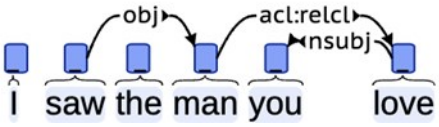

Figure 3. Example of a UD parse (Nivre et al., 2016)

The parser we employed demonstrated strong performance across various standard metrics of automatic parsers. Nevertheless, we did not conduct manual corrections of the parses, which means that the parses might contain errors. However, this potential concern applies to all other findings relying on automatic parsers as well, as noted in previous work (e.g., Boston et al., 2008; Boston et al., 2011 Demberg & Keller, 2008; Levy, 2008; Smith & Vasishth, 2020).

Measuring Expectation and Locality. To measure expectation, we used Head-Dependent Mutual Information (HDMI; equation in 4), following Futrell (2019). In (4), *p(h,d)* stands for how many times a pairs of words occurs together in a dependency, and *p(h)p(d)* stands for how many times in total the the two words occur in the corpus. HDMI captures the mutual information between two co-dependents and can be used to approximate the degree to which the two co-dependents predict each other. However, as noted in Futrell (2019), obtaining token-specific HDMI requires a giant parsed text, which is feasible. Similar problems of data scarcity were pointed out in Smith and Vasishth (2020). We therefore calculated HDMI for any given pair of word categories, instead of word tokens. For word categories, we use more fine-grained part-of-speech tags from UD. For example, 'doctor' in ex. 2 will be labelled as 'NN' (i.e., all content nouns) and 'diagnosed' as VBD (i.e., past-tense verb).

(4) $HDMI = log \frac{p(h,d)}{p(h)p(d)}$

For locality, we used Dependency Locality (DL), which measures the number of intervening words between two co-dependents (e.g., Niu & Liu, 2022). This measure is more simplistic than the proposed *integration cost* in Dependency Locality Theory (Gibson 1998; 2000), as well as feature similarity measures employed in Cue-based Retrieval (Lewis & Vasishth, 2005; Jäger, Engelmann, & Vasishth, 2017). However, DL is positively correlated with integration cost, or the number of intervening feature-matching words.

Moreover, there are also independent motivations for using DL as a cognitively plausible complexity measure (Futrell, Mahowald, & Gibson, 2015; Liu, 2008; Liu, Xu, & Liang, 2017; Niu & Liu, 2022).

As Futrell (2019) found in text corpora a negative correlation between DL and HDMI (i.e., indicating that word pairs with high HDMI are more likely to be kept together), we checked whether this is true for the texts in our datasets as well. Indeed, across all datasets, there was a negative correlation between HDMI and DL.

Data Analysis

Data Cleaning. We excluded extremely small or large data points in the datasets, with slightly different exclusion measures for different methods. For SPR datasets, we excluded RTs below 50ms or above 3000ms (widely used cutoffs, e.g., Laurinavichyute & von der Malsburg, 2023; Logačev & Vasishth, 2016; Monsalve, Frank, & Vigliocco, 2012;). For the eye-tracking dataset, we excluded data points whose first-pass duration (the duration of the first fixations on the current word, which is usually considered an early measure, e.g., Rayner, 1998) is below 50ms or whose total viewing time (aggregation of the duration of all fixations on the current word) is above 3000ms. For the Maze dataset, we excluded RTs below 100ms or above 5000ms, since RTs in Maze tasks are usually longer (e.g., Boyce, Futrell, & Levy, 2020). We also excluded reading times for words containing punctuation marks (including words at sentence boundaries). For the dependencies, we excluded those labelled in the UD system as *punct* (punctuation), *dep* (unspecified dependency), and *root*, which points to the root of the sentence.

Statistical Modeling. We fitted linear mixed effects models on the log transformed RTs of the second co-dependents, with DL, HDMI, and their interaction, and two word-level factors (word length and frequency) as fixed effects; all predictors were centered. For eye-tracking, first-pass duration and total viewing time were analyzed. We first ran analyses for each dataset separately. For each individual dataset, we also included a random intercept for participants,[1] and random slopes when the model could converge. In addition, we also ran a meta-analysis collapsing three datasets: Natural Stories SPR, Brown SPR, and Provo ET (only the total reading time data). The Natural Stories A-Maze dataset was not included in the meta-analysis because its scale was very different from the remaining of the datasets. For the meta-analysis, we included a random intercept for participants and a random intercept for datasets. With the aggregated datasets, we performed exploratory analyses based on (i) head direction (according to UD standards), since previous work has revealed that head directionality may play an important role in characterizing processing difficulty (e.g., Baumann, 2014; Frazier, 2013; Niu & Liu, 2022; Yamashita, Hirose, & Packard, 2011); and (ii) whether the dependency involves only core arguments (i.e., verbs and nouns), since previous psycholinguistic research has mostly focused on such dependencies (e.g., subject-verb, verb-object dependencies).

Results

We first provide a high-level overview of the results, before detailing the results of each individual analysis. We will only present the effects of DL, HDMI, and their interactions. For the word-level factors, all analyses showed significant effects in the expected direction, whereby RTs decreases as frequency increases, and as word length decreases.

Effects of DL in each individual analysis and the meta-analysis are plotted in Figure 4. The bars represent 95% confidence intervals. As can be seen, all analyses revealed a positive effect of DL on RTs, whereby an increase in DL leads to longer RTs, consistent with the predictions of memory-based theories.

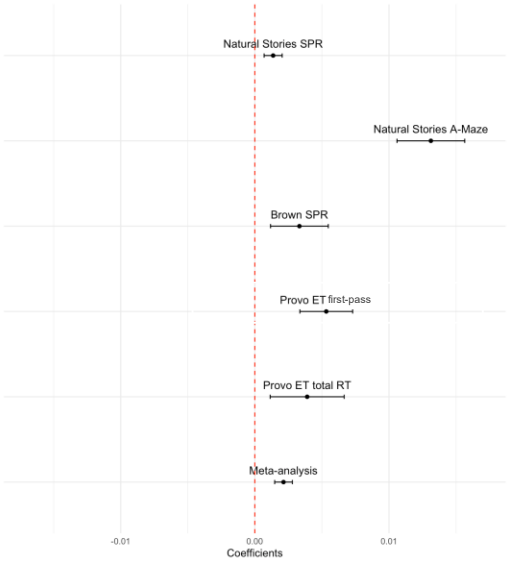

Figure 4. Effects of DL in each analysis (the x-axis represents the model estimates of the co-efficient and the bars represent 95% confidence intervals).

Effects of HDMI in each individual analysis and the meta-analysis are plotted in Figure 5. The error bars represent 95% confidence intervals. As can be seen, the majority of the datasets showed a negative effect of HDMI on RTs, whereby an increase in HDMI leads to shorter RTs, as predicted by expectation-based theories. The meta-analysis is consistent with a negative effect of HDMI.

[1] Per the suggestion of an anonymous reviewer, we also ran analysis with a random intercept for each (lemmatized) word, and the results were qualitatively the same.

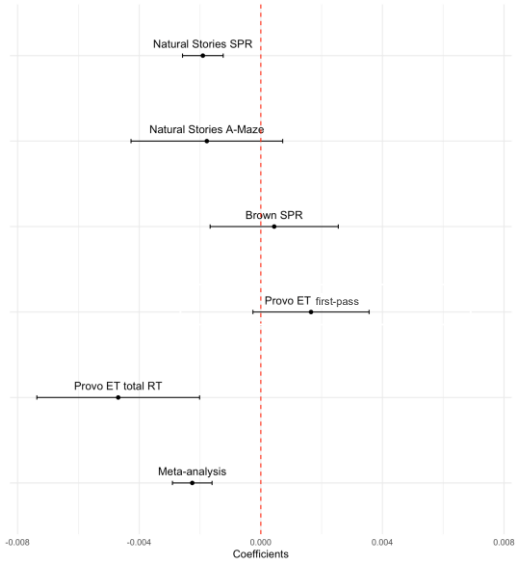

Figure 5. Effects of HDMI in each analysis (the x-axis represents the model estimates of the co-efficient and the bars represent 95% confidence intervals).

Results of the interaction effects between DL and HDMI are plotted in Figure 6. The error bars represent 95% confidence intervals. As can be seen, all datasets showed a positive coefficient for the interaction effects, whereby an increase in DL leads to even more processing slowdowns when HDMI is high, consistent with the *Information Locality* hypothesis. In Natural Stories SPR, Brown SPR, and the meta-analysis, the interaction effects were significant.

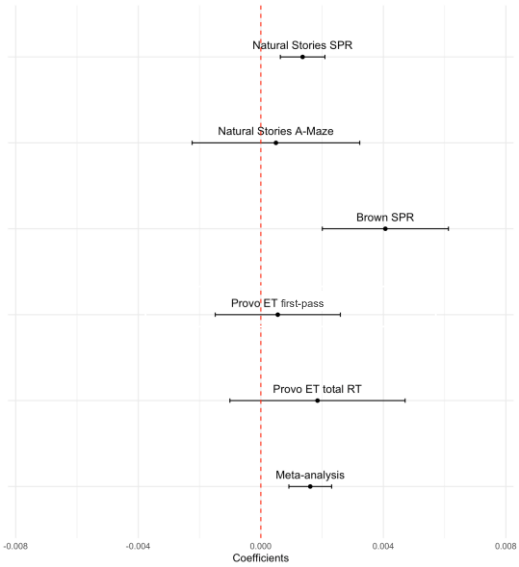

Figure 6. Effects of interaction in each analysis (the x-axis represents the model estimates of the co-efficient and the bars represent 95% confidence intervals)

Below we review the results of each individual analysis and the meta-analysis in detail.

Natural Stories SPR. This dataset showed a significant positive effect of DL has on RTs as well (p<0.001), whereby RTs increase as DL increases, consistent with the predictions of memory-based theories. In addition, there was significant effect of HDMI on RTs (p<0.001), whereby RTs decrease as HDMI increase, as predicted by memory-based theories. Moreover, we found a significant positive interaction between DL and HDMI (p<0.001), which confirmed the predictions of the *Information Locality* hypothesis.

Natural Stories A-Maze. In this dataset, we found a significant positive effect of DL on RTs (p<0.001), again confirming locality-based theories. However, we did not find significant effects of HDMI (p=0.180) or interaction (p=0.757), therefore supporting neither *Information Locality* nor *Prediction Maintenance*.

Brown SPR. In this dataset, we found that DL has a significant positive effect on RTs (p<0.05), which confirmed the predictions of locality-based theories, but no significant effect of HDMI on RTs (p=0.69). However, there was a positive significant interaction between DL and HDMI (p<0.001), supporting the *Information Locality* hypothesis.

Provo ET. In this dataset, first-pass durations showed significant positive effects of DL on RTs (p<0.001), in line with locality-based theories. However, no significant effects of HDMI (p=0.095) or interaction (p=0.601) were found. For total reading times, we found significant positive effects of DL (p<0.01) and also significant negative effects of HDMI (p<0.001) on RTs, confirming the predictions of locality-based theories and expectation-based theories, respectively. However, there is no significant interaction between HDMI and DL (p=0.193).

Meta-analysis & Exploratory Analysis. The analysis based on the aggregated datasets showed a significant positive effect of DL on RTs (p<0.001), a significant negative effect of HDMI on RTs (p<0.001), as well as a significant interaction between the two factors (p<0.001), whereby effects of DL are stronger as HDMI gets higher. Therefore, the meta-analysis confirmed the predictions of locality-based theories, expectation-based theories, as well as *Information Locality*. In our exploratory analyses (i.e., analysis of head-initial dependencies, analysis of head-final dependencies, analysis of dependencies involving only core arguments, and analysis of dependencies involving non-core arguments), all of them showed a significant positive effect of DL on RTs and the majority showed a significant negative effect of HDMI on RTs. Moreover, we found significant interactions between the DL and HDMI in all exploratory analyses (p<0.01), such that effects of DL are stronger as HDMI gets higher.

Discussion

Sentence comprehension in humans is shown to involve two key processes, working memory constraints and prediction. How do these two processes interact and together determine the sentence processing behavior? Using data from four large-scale naturalistic reading time corpora, we tested two competing hypotheses: (i) the *Information Locality* hypothesis, which maintains that words with high mutual predictability are constrained to be linearly close to each other, and (ii) the *Prediction Maintenance* hypothesis, which assumes that words with high mutual predictability can overcome certain working memory demands allowing them to be relatively far apart in the sentence. These two hypotheses make divergent predictions about the interaction of working memory constraints and predictability. The information locality predicts that locality effects are enhanced when mutual predictability is high, while prediction maintenance predicts that locality effects are weakened when a word is highly predictable.

Two out of the four tested datasets showed evidence for *Information Locality*, while none of them supported *Prediction Maintenance*. A meta-analysis of ensembled data from three datasets provided conclusive evidence for the *Information Locality* hypothesis. The results support an information-theoretic account of sentence processing, the lossy-context surprisal theory, which assumes that sentence comprehension is driven by moment-by-moment predictions of the upcoming sentence material, based on a non-veridical, distorted memory representation of the actual linguistic input. The distortions from the original input to a potentially non-veridical memory representation occur due to working memory limitations and are governed by information theoretic principles.

This is the first large scale targeted evaluation of hypotheses about how working memory and prediction interact, a bottleneck problem in developing a complete theory of sentence processing, and we conclusively show that strong predictability between words does not necessarily attenuate working memory demands rather it constrains them to optimize on fewer working memory resources. This is because high predictability between words is useful only when words are readily accessible in memory during comprehension. However, we do acknowledge that Demberg and Keller (2008) analyzed such interactions using the Dundee Eye-Tracking Corpus (Kennedy & Pynte, 2005), they focused on the effects of DL and surprisal, and did not test any hypotheses regarding the interaction effects.

The results are important for theories of sentence processing because they provide a strong empirical basis for how working memory constraints and prediction should be incorporated in a model of sentence processing. For example, our results strongly suggest that the predictions of the upcoming sentence material are constrained by working memory limitation. Predictions, despite being ubiquitous, are not made based on a perfectly retained representation of context (see also Apurva & Husain, 2021). Rather, the quality of context is oftentimes degraded, and expectations are comprised, especially in cases where two words that highly predict each other are separated in linear time. Our corpus evaluations, therefore, offered important insights for future theory building in sentence processing. Theoretical accounts should move beyond explaining only expectation-based or only memory-based effects. Our results will inspire more complete models of sentence processing that can explain multiple empirical phenomena in sentence comprehension.

Although it is not the focus of the paper, we also provided broad-coverage evidence for the cognitive plausibility of dependency length (DL) and head-dependent mutual information (HDMI) measures. Across the datasets, we found robust effects of DL on RTs. This is the first large-scale evidence that a simple measure like DL, can reasonably capture linguistic complexity of an utterance (see Niu & Liu for an evaluation based on only the Natural Stories SPR). In future work, we plan to explore more complicated measures of locality, such as intervening number of references (e.g., Gibson, 1998), and intervening number of heads (e.g., Yadav et al., 2020 and Yadav, Mittal, & Husain 2022). Regarding HDMI, although its effect is less robust, there is evidence that it can be used to approximate prediction-based effects. As mentioned before, we did not calculate token-level HDMI, which means that its effects may not capture very fine-grained expectation effects. Future work should explore this.

A major limitation of our study is that it used only English data: the effects of working memory constraints and prediction have been shown to differ across languages (Stone et al., 2020, Husain et al., 2014, Safavi et al., 2016, Mertzen et al. 2020). For example, there is no evidence for prediction-locality interaction from a controlled experiment on German (Stone et al., 2020) and data from a Hindi experiment support prediction maintenance (Husain, Vasishth, & Srinvasan 2014). It is critical to evaluate the *Information Locality* and *Prediction Maintenance* hypotheses on data from typologically different languages. It could be the case that the underlying mechanism of interaction between working memory and prediction is invariant across languages, but languages differ in the extent of predictability, e.g., prediction strategy might be more robust and reliable in verb-final languages (Konieczny 2000; Yamashita, 2000).

Conclusion

The current work reveals new insights about the cognitive processes that underlie real time sentence comprehension: The dependency between words with high mutual predictability becomes costlier to resolve as the working memory demand incurred by the dependency increases. Sentence comprehension capitalizes on mutually predictable words kept under low working-memory load. To our knowledge, this is the first large-scale empirical investigation that focusses on how the key processes involved in sentence comprehension interact in real time. Our work contributes to developing a unified theory of sentence processing that can explain memory-based as well as expectation-based effects observed across typologically different languages.

References

Apurva, & Husain, S. (2021). Revisiting anti-locality effects: Evidence against prediction-based accounts. *Journal of Memory and Language*, *121*, 104280.

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.

Baumann, P. (2014). Dependencies and hierarchical structure in sentence processing. In: Bello, P., Guarini, M., McShane, M., Scassellati, B. (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society. Cognitive Science Society*, Austin, pp. 152–157

Boston, M. F., Hale, J. T., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam sentence corpus. *Journal of Eye Movement Research*, **2**(1), 1–12.

Boston, M. F., Hale, J. T., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, 26(3), 301–349.

Boyce, V., Futrell, R., & Levy, R. P. (2020). Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language, 111*, 1-13.

Boyce, V., & Levy, R. (2023). A-maze of natural stories: Comprehension and Surprisal in the maze task. *Glossa Psycholinguistics, 2*(1).

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. Hoboken, NJ: John Wiley & Sons.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition, 109*(2), 193–210.

Frazier, L. (2013). Syntax in sentence processing. In R. P. G. van Gompel (ed.), *Sentence Processing* (vol. Current issues in the psychology of language), pp. 21–50. New York NY: Psychology Press: Taylor & Francis Group.

Futrell, R. (2019). Information-theoretic locality properties of natural language. In X. Chen & R. Ferrer-i-Cancho (Eds.), *Proceedings of the first workshop on quantitative syntax (Quasy, SyntaxFest 2019)* (pp. 2–15). Paris, France: Association for Computational Linguistics.

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science, 44*(3).

Futrell, R., Gibson, E., Tily, H., Blank, I., Vishnevetsky, A., Piantadosi, S., & Fedorenko, E. (2020). The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 1–15.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341.

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

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, Bbain: Papers from the first mind articulation project symposium* (pp. 95–126). Cambridge, MA: MIT Press.

Grodner, D., & Gibson, E. (2005). Some consequences of the serial nature of linguistic input. *Cognitive Science, 29*(2), 261–290.

Husain, S., Vasishth, S., & Srinivasan, N. (2014). Strong expectations cancel locality effects: Evidence from Hindi. *PLoS ONE, 9*(7).

Hahn, M., Futrell, R., Levy, R., & Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, *119*(43).

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies 2001 - NAACL '01*.

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research, 45*, 153–168.

Konieczny, L. (2000). Locality and Parsing Complexity. *Journal of Psycholinguistic Research, 29*(6), 628-645.

Kučera, H. & Francis, W. N. (1967). *Computational analysis of present-day American English* Providence, R.I.: Brown University Press .

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

Levy, R. P., & Keller, F. (2013). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language, 68*(2), 199–222.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29*(3), 375–419.

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and bayesian meta-analysis. *Journal of Memory and Language, 94*, 316–339.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 39– 95). Cambridge, MA: MIT Press.

Linzen, T., & Jaeger, T. F. (2015). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science, 40*(6), 1382–1411.

Haitao L. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), 159–191.

Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, *21*, 171–193.

Niu, R., & Liu, H. (2022). Effects of syntactic distance and word order on Language Processing: An investigation

based on a psycholinguistic treebank of English. *Journal of Psycholinguistic Research*, *51*(5), 1043–1062.

Mertzen, D., Laurinavichyute, A., Dillon, B., Engbert, R., & Vasishth, S. (2020). Crosslinguistic evidence against interference from extra-sentential distractors. *PsyArXiv*.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1659–1666). Portoroz, Slovenia: European Language Resources Association.

Qi, P., Dozat, T., Zhang, Y., & Manning, C. D. (2018). Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 160–170).

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Safavi, M. S., Husain, S., & Vasishth, S. (2016). Dependency resolution difficulty increases with distance in Persian separable complex predicates: Evidence for expectation and memory-based accounts. *Frontiers in Psychology*, 7.

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

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2022). Large-scale evidence for logarithmic effects of word predictability on reading time. *PsyArXiv*.

Shain, C., van Schijndel, M., Futrell, R., Gibson, E., & Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*.

Stone, K., von der Malsburg, T., & Vasishth, S. (2020). The effect of decay and lexical uncertainty on processing long-distance dependencies in reading. *PeerJ, 8*, e10438.

Tung, T.-Y., & Brennan, J. R. (2023). Modeling retrieval interference during naturalistic comprehension. Poster at *the 36th Annual Conference on Human Sentence Processing*.

Vasishth, S., & Drenhaus, H. (2011). Locality in German. *Dialogue & Discourse, 2*(1), 59–82.

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. (2023). Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*.

Wu, F., Kaiser, E., & Vasishth, S. (2017). Effects of early cues on the processing of Chinese relative clauses: Evidence for experience-based theories. *Cognitive Science, 42*(S4), 1101–1133.

Xiang, M., & Wang, S. (Under Revision). Locality and expectation in Chinese wh-in-situ dependencies.

Yadav, H., Mittal, S., & Husain, S. (2022). A reappraisal of dependency length minimization as a linguistic universal. *Open Mind*, *6*, 147–168.

Yadav, H., Vaidya, A., Shukla, V., & Husain, S. (2020). Word order typology interacts with linguistic complexity: A cross-linguistic corpus study. *Cognitive Science, 44*(4), Article e12822.

Yamashita, H. (2000). Structural computation and the role of morphological markings in the processing of Japanese. *Language and Speech*, *43*(4), 429–455.

Yamashita, H., Hirose, Y., & Packard, J. L. (2011). *Processing and producing head-final structures*. Springer.