Memory access during incremental sentence processing causes reading time latency

Cory Shain[1] shain.3@osu.edu
Marten van Schijndel[1] van-schijndel.1@osu.edu
Richard Futrell[2] futrell@mit.edu
Edward Gibson[2] egibson@mit.edu
William Schuler[1] schuler.77@osu.edu

[1]Dept of Linguistics
The Ohio State University

[2]Dept of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Abstract

Studies on the role of memory as a predictor of reading time latencies (1) differ in their predictions about when memory effects should occur in processing and (2) have had mixed results, with strong positive effects emerging from isolated constructed stimuli and weak or even negative effects emerging from naturally-occurring stimuli. Our study addresses these concerns by comparing several implementations of prominent sentence processing theories on an exploratory corpus and evaluating the most successful of these on a confirmatory corpus, using a new self-paced reading corpus of seemingly natural narratives constructed to contain an unusually high proportion of memory-intensive constructions. We show highly significant and complementary broad-coverage latency effects both for predictors based on the Dependency Locality Theory and for predictors based on a left-corner parsing model of sentence processing. Our results indicate that memory access during sentence processing does take time, but suggest that stimuli requiring many memory access events may be necessary in order to observe the effect.

1 Introduction

Any incremental model of sentence processing where an abstract meaning representation is built up word-by-word must involve storage and retrieval of information about previously encountered material from some memory store. The retrieval operations have been hypothesized to be associated with increased processing time (Gibson, 2000; Lewis and Vasishth, 2005; Wu et al., 2010), and this prediction has been borne out in experiments using constructed stimuli (Gibson, 2000; Grodner and Gibson, 2005; Boston et al., 2011; von der Malsburg et al., 2015). However, memory-based latency effects have been null or even negative in broad-coverage reading time experiments using naturally-occurring text data that included baseline controls for n-gram and probabilistic phrase-structure grammar (PCFG) surprisal (Demberg and Keller, 2008; van Schijndel et al., 2013b).

The failure of experimental latency effects to generalize to naturally-occurring data raises doubts about their existence. The effects observed in constructed stimuli could be due to (1) information theoretic phenomena (e.g., surprisal) that such experiments rarely control for, (2) limited syntactic domain (e.g., relative clauses), or (3) 'oddball' effects – i.e. effects related to the semantic strangeness and decontextualized nature of the input, rather than due to difficulty retrieving information from working memory. On the other hand, the lack of positive latency effects in studies using naturally-occurring input could be (1) because of the small number of subjects – ten – in the Dundee corpus (Kennedy et al., 2003) used by e.g. Demberg and Keller (2008) and van Schijndel et al. (2013b) or (2) because naturally-occurring newswire texts might contain too low a proportion of memory-intensive constructions to reveal a generalized memory effect.

In addition to the problem of conflicting results between constructed vs. naturally-occurring stimuli, research on the role of memory in sentence processing must also contend with the open question of where and what kinds of memory effects are predicted during sentence processing. One of the first and most

Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity,
pages 49–58, Osaka, Japan, December 11-17 2016.

well-known memory-based theories of sentence processing is the Dependency Locality Theory (DLT) (Gibson, 2000), which predicts processing difficulty proportional to the number of discourse referents intervening between the current word and any dependencies it shares with words in its preceding context. Lewis and Vasishth (2005), on the other hand, predict difficulty as a function of memory decay during the retrieval operations of an incremental left-corner parser. Note that both of these accounts are locality-based (difficulty is predicted to increase with distance), modeling the notion that decay over time may make it more difficult (and hence time-consuming) to recall items from working memory. However, it is conceivable that processing difficulty may have less to do with locality than simply with whether or not a memory access or recall event has occurred, a hypothesis explored by van Schijndel and Schuler (2013) with mixed results.

The present work seeks to answer these questions by evaluating many plausible implementations of prominent theories of sentence processing as predictors of reading times on the new Natural Stories corpus (Futrell, Gibson, Tily, Vishnevetsky, Piantadosi, & Fedorenko, in prep). The Natural Stories corpus is constructed in order to embed an unusually-high proportion of rare words and memory-intensive constructions in narratives designed to resemble naturally-occurring text. It therefore occupies an intermediary position – which we will refer to as 'constructed-natural' – between isolated constructed stimuli and naturally-occurring text that might help overcome the limitations of both. We evaluate against a strong baseline model that includes controls for *n*-gram and PCFG surprisal. We find clear evidence for (1) a locality effect when constructing dependencies to preceding words and (2) a locality-independent 'reinstatement' effect whenever derivation fragments must be recalled from working memory during the operation of a left-corner parser. We show that both of these effects contribute independently to model fit. Our findings therefore support the existence of memory-based processing difficulty and shed light on the specific role of memory in sentence processing.

2 Related Work

Previous studies have explored the influence of memory on reading times using constructed stimuli. For example, Grodner and Gibson (2005) showed memory-related effects in self-paced reading of individual sentences constructed to contain difficult center-embeddings. Other studies have used the (constructed) Potsdam eye-tracking corpus (Kliegl et al., 2004) to investigate the predictivity of ACT-R memory influences on reading times (e.g., Boston et al., 2011; von der Malsburg et al., 2015).[1] As mentioned above, effects found in studies using constructed stimuli presented in isolation might face confounds due to 'oddball' effects or lack of extra-sentential context.

Other work has explored memory and processing using naturally-occurring stimuli (generally, newswire texts). Demberg and Keller (2008) examined the influence of the DLT on reading times using the Dundee eye-tracking corpus (Kennedy et al., 2003). They found some evidence of DLT influences on specific content words, but the effect was weak enough that it was not significant until the analysis was constrained to just nouns and verbs. Even then, the effect is not significant under multiple comparison correction. Other studies have also used the Dundee corpus to test the predicted memory effects of left-corner models of sentence processing (van Schijndel and Schuler, 2013; van Schijndel et al., 2013b), but these studies found a negative correlation between reading times and predicted left-corner memory operations, which is the opposite of what most theories of sentence processing predict. We posit that the weakness of the DLT and the unusual left-corner influence on the Dundee corpus may be caused by the limited number of subjects or by the limited number of complex dependencies in the corpus.

The Natural Stories corpus used in this study is constructed in order to tax working memory resources in the processing of otherwise natural-seeming narratives, so it plausibly mitigates concerns related to oddball effects and lack of context on the one hand and syntactically 'easy' constructions on the other. In this respect, the most similar corpus to ours of which we are aware is Bachrach et al. (2009), which like Natural Stories was constructed to read naturally but included a higher degree of syntactic complexity than is usual in naturally-occurring text. However, compared to Natural Stories, Bachrach et al. (2009)

[1] The present work does not directly test the predictions of ACT-R, but some of the predictors used in this study – especially the distance-weighted left-corner predictors discussed below – make similar predictions to ACT-R.

has substantially fewer subjects (23 vs. 181) and words (3540 vs. 10257). Wu et al. (2010) used the Bachrach et al. (2009) corpus to investigate the correlation between changes in embedding depth and reading times and found a positive effect on latency.[2]

3 Background

This work explores two related models of the relationship between memory and sentence processing: (1) the Dependency Locality Theory, in which memory is predicted to be used to construct syntactic dependencies to words in the preceding context with a cost proportional to the length of the dependency (or dependencies) being constructed, and (2) left-corner theories of sentence processing, such as Lewis and Vasishth (2005) and Schuler et al. (2010), in which certain parser operations require disjoint incomplete signs (referring to discourse referents) to be recalled from working memory. We outline these broader frameworks, along with a number of possible implementations of each, in the remainder of this section.

3.1 Dependency Locality Theory

The Dependency Locality Theory (DLT; Gibson, 2000) predicts a cost for integrating a word into an incomplete parse proportional to the number of discourse referents that intervene in any syntactic dependencies the word shares with words in its preceding context. For simplicity, Gibson (2000) implements this calculation in terms of abstract 'energy units' (EU) and considers all and only nouns (excluding pronouns) and finite verbs to count as discourse referents. Integration cost is the sum of the 'discourse cost' of the word itself (1 for nouns and finite verbs, 0 otherwise) and the distance of any dependencies to preceding words, measured in number of intervening discourse referents. The cost of long-distance dependencies is assessed at the gap site, producing e.g. subject-/object-relative asymmetries (the relative clause verb intervenes in its own dependency for object gaps but does not for subject gaps).

As pointed out by Gibson (2000), this implementation might benefit from modification in light of other cognitive considerations. In this study, we implemented three such modifications related to verb weights (DLT-V), coordination (DLT-C), and preceding modifier dependencies (DLT-M):

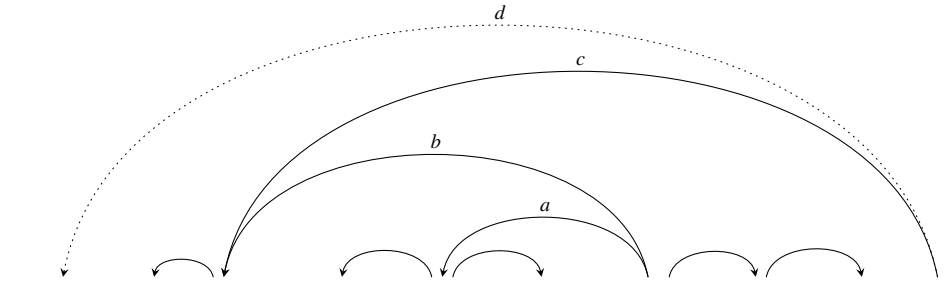
- **DLT-V**: *Verbs are more expensive*. Non-finite verbs receive a cost of 1 (instead of 0) and finite verbs receive a cost of 2 (instead of 1).
- **DLT-C:** *Coordination is less expensive*. Dependencies out of coordinate structures skip preceding conjuncts in the calculation of distance, and dependencies with intervening coordinate structures assign that structure a weight equal to that of its heaviest conjunct.
- **DLT-M:** *Exclude modifier dependencies*. Dependencies to preceding modifiers are ignored.

DLT-V is motivated by the possibility that finite verbs might be more costly to integrate than nouns (since they contain additional information about tense/aspect) and that non-finite verbs might have a non-zero discourse cost. DLT-C is motivated by the fact that coordination can generate very long dependencies that are not particularly difficult to process, suggesting that each sub-referent of a conjunction may be integrated into a conjoined set which is finally integrated at the end of the conjunction. DLT-M is designed to avoid excessive 'double-counting' of material intervening in long modifier dependencies.

These modifications can be applied in any combination, yielding eight distinct implementations of the DLT. Henceforth, we indicate that a modification was applied by suffixing its letter to 'DLT' (e.g., DLT-CM is DLT with the coordination and modifier modifications only). For an illustration of these implementations at work on an example sentence, see Figure 1.

Differences between the four variants illustrated in Figure 1 are especially apparent at the verbs *caught* and *fled* (although note also that *stealing* – a non-finite verb – only has a cost under DLT-V). There are two dependencies between *caught* and preceding words, dependency *a* to the head *and* of its conjoined

[2]While the Wu et al. (2010) embedding depth predictor is derived from automatic parses of their corpus, the present work used hand-corrected syntactic annotations to calculate the left-corner operations required to incrementally construct syntactic structures. Wu et al. (2010) also largely focused on the influence of frequency effects which the present work simply adopts as control predictors.



	Yesterday,	the	person	supervisors	and	co-workers	caught	stealing	millions	fled.
DLT:	0	0	1	1	0	1	5	0	1	10
DLT-V:	0	0	1	1	0	1	7	1	1	13
DLT-C:	0	0	1	1	0	1	4	0	1	8
DLT-M:	0	0	1	1	0	1	5	0	1	5

Figure 1: Integration cost calculations for implementations of DLT on an example sentence. For example, at *caught* there is an object gap and therefore dependencies back to *person* (2 nouns and 1 verb intervene) and *supervisors* (1 noun intervenes) = 4 total cost-accumulating interveners. As a finite verb, *caught* gets a cost of 1+4 = 5. Note that this figure increases if verbs become more expensive (DLT-V) and decreases if coordinates become less expensive (DLT-C). Note also that the cost of *fled* is much lower in DLT-M, since the dotted dependency to *Yesterday* is ignored.

subject, and dependency b to *person*, the modificand of the relative clause. Dependency a spans the intervening word *co-workers*. Dependency b is an object relative dependency. Since the gap site follows *caught*, *caught* is included as an intervener in the dependency, which therefore spans *supervisors and co-workers caught*. The DLT integration cost of *caught* is the sum of the discourse cost of *caught* itself and the costs of dependencies a and b.

These costs vary by implementation. For DLT, *caught* (a finite verb) has a discourse cost of 1, dependency a has a cost of 1 for its single intervening noun *co-workers*, and dependency b has a cost of 3 for *supervisors*, *co-workers*, and *caught*, for a total integration cost of 1 + 1 + 3 = 5. For DLT-V, *caught* is worth 2 EU as a finite verb rather than 1. This increases both its discourse cost and its cost as an intervener in dependency b, thus increasing the integration cost of *caught* by 2 (from 5 to 7). For DLT-C, the cost of the conjoined noun phrase *supervisors and co-workers* is reduced from 2 to 1 (the weight of its heaviest conjunct) as an intervener in dependency b. This reduces the integration cost of *caught* by 1 (from 5 to 4). DLT-M does not affect the cost of *caught*, which has no dependencies to preceding modifiers.

Similar considerations govern the variation in integration cost of *fled*, which has dependencies c (to *person*) and d (to *Yesterday*). For DLT, the integration cost of *fled* is 4 (for dependency c) + 5 (for dependency d) + 1 (discourse cost of *fled*) = 10. This increases to 13 for DLT-V because the finite verb *caught*, which intervenes in both dependencies c and d, is upweighted from 1 to 2, along with *fled* itself. Because the cost of *supervisors and co-workers*, which also intervenes in both c and d, decreases from 2 to 1 for DLT-C, the DLT-C integration cost of *fled* is reduced by 2 (from 10 to 8). DLT-M ignores the preceding modifier dependency d, resulting in an integration cost of 4 (dependency a) + 1 (discourse cost of *fled*) = 5.

3.2 Left-corner parsing

Many sentence processing models (Johnson-Laird, 1983; Abney and Johnson, 1991; Gibson, 1991; Resnik, 1992; Stabler, 1994; Lewis and Vasishth, 2005) are defined in terms of left-corner parsing operations (Aho and Ullman, 1972; van Schijndel et al., 2013a), which assemble local dependencies between signs using a minimal store of incomplete derivation fragments. Left-corner parsers account for sequences of words $x_1 \ldots x_T$ as stacked-up derivation fragments a/b, each consisting of a top sign a

lacking a bottom sign b yet to come. When a left-corner parser consumes a word, it makes decisions to fork off and/or join up these derivation fragments. When the current word x_t satisfies the missing bottom sign b of a derivation fragment a/b, the parser replaces a/b in the memory store with a, indicating a completed prediction. Because this does not increase the number of derivation fragments in the memory store, we call this a no-fork (–F) operation:

$$\frac{a/b \quad x_t}{a}\, b \rightarrow x_t. \tag{–F}$$

When the current word x_t does not satisfy the missing sign of a derivation fragment, then x_t is added to the memory store as part of a new derivation fragment c. As this increases the number of stored derivation fragments, we call this a yes-fork operation (+F):[3]

$$\frac{a/b \quad x_t}{a/b \quad c}\, b \xrightarrow{+} c \ldots ; \;\; c \rightarrow x_t. \tag{+F}$$

The other class of parser operations is join operations. In these operations, the parser decides whether to connect two previously disjoint derivation fragments. When the sign c satisfies the missing sign of the fragment a/b while predicting b', we rewrite the memory store with a single fragment a/b'. This reduces the number of derivation fragments in the memory store, so we call it a yes-join (+J) operation:

$$\frac{a/b \quad c}{a/b'}\, b \rightarrow c\, b'. \tag{+J}$$

Conversely, when memory contains a fragment a/b and a sign c, but c does not satisfy a/b, we make the appropriate left-corner predictions from c while keeping it as a separate memory item (no-join, –J):

$$\frac{a/b \quad c}{a/b \quad a'/b'}\, b \xrightarrow{+} a' \ldots ; \;\; a' \rightarrow c\, b'. \tag{–J}$$

These two binary decisions have four possible outcomes in total: the parser can fork only (which increases the number of derivation fragments by one), join only (which decreases the number of derivation fragments by one), both fork and join (which keeps the number of derivation fragments the same), or neither fork nor join (which also preserves the number of derivation fragments). The experiments described in this paper also use a variant of a left-corner parser (van Schijndel et al., 2013b) which introduces additional derivation fragments to carry referents involved in non-local dependencies such as filler-gap constructions (see Figure 2).

As in the case of the DLT, there are a number of ways in which the memory predictions of this left-corner parsing model could be implemented. In this study, we consider three families of predictors:

- **EMBD:** *End of embedded region.* Flag integration operations where disjoint derivation fragments are merged in working memory. EMBD includes –F+J operations that reduce the stack as well as the closure of long-distance dependency carriers for gapping and heavy-shift.

- **NoF:** *'No fork' (–F) operation.* Flag parser operations that recall and transition the top sign of a derivation fragment once the bottom sign has been completed, including –F+J operations (integrations) as well as –F–J operations, in which the current derivation fragment is given a new top sign but is not integrated with another fragment. NoF models the notion that memory is required to access and update the top sign of the attentionally-focused fragment, and is not sensitive to carrier fragments.

- **REINST:** *Reinstatement operation.* Flag if either a long-distance dependency has terminated or if a –F operation has taken place (i.e. the union of EMBD and REINST flags).

[3] Here $b \xrightarrow{+} c \ldots$ indicates one or more grammar rule applications yielding a category c followed by zero or more other categories.

a)

$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$	$t = 8$	$t = 9$
S/S	S/S	S/S	S/S	S/S	S/S	S/S	S/S	S/VP
	NP/N	NP/RC	NP/RC	NP/RC	NP/VPgap	NP/Sgap	NP/NP	
			NP/NPconj	NP/NP				
Yesterday,	the	person	supervisors	and	coworkers	caught	stealing	millions	fled.

b)

$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$	$t = 8$	$t = 9$
S/S	S/S	S/S	S/S	S/S	S/S	S/S	S/S	S/VP
	NP/N	NP/RC$_p$	NP/RC$_p$	NP/RC$_p$	NP/RC$_p$	NP/RC$_p$	NP/NP	
		RC/RC	RC/RC	RC/RC	RC/VPgap	RC/Sgap		
			NP/NPconj	NP/NP				
Yesterday,	the	person	supervisors	and	coworkers	caught	stealing	millions	fled.

c)

	Yesterday,	the	person	supervisors	and	coworkers	caught	stealing	millions	fled.
Depth:	0	1	2	3	4	4	3	3	2	1
EMBD:	0	0	0	0	0	1	0	1	1	1
NoF:	0	0	1	0	0	1	0	0	1	1
REINST:	0	0	1	0	0	1	0	1	1	1

Figure 2: (a) Partial analyses of the sentence *Yesterday, the person supervisors and coworkers caught stealing millions fled,* in a left-corner parser, showing stacked-up derivation fragments (vertical axis) over time (horizontal axis). (b) Partial analyses of the same sentence using additional derivation fragments to carry referents involved in non-local dependencies (in this case, for a referent p of *the person*). (c) Calculation of predictors EMBD ('end of embedding', sensitive to carrier fragments), NoF ('no fork', ignores carrier fragments) and REINST ('reinstatement', union of EMBD and NoF) over the example sentence. Tokens are flagged when stack depth reduces (e.g., *millions*). NoF and REINST (but not EMBD) also flag when the awaited sign is encountered without joining, as is the case for *person* just before forking off the relative clause. Note that NoF does not flag *stealing*, the end of a carrier fragment, while EMBD and REINST do.

For each of these three families, we consider both boolean and distance-weighted variants. In the case of EMBD, the distance-weighted predictor EMBD-LEN is the length of the embedded region being integrated. In the case of NoF, the distance-weighted predictor NoF-LEN is the distance since the last no-fork operation at that depth level, modeling decay since the last time that the top sign of the attentionally-focused derivation fragment was recalled from working memory. The distance-weighted version of REINST (REINST-LEN) is the max of these two measures. REINST-LEN is essentially an implementation of the ACT-R retrieval cost of Lewis and Vasishth (2005), Boston et al. (2011), and von der Malsburg et al. (2015). To maintain comparability with the DLT, we implement three types of distance in each family: number of words, number of DLT discourse referents, and number of DLT-V (verb-reweighted) discourse referents. We therefore consider twelve distinct implementations of left-corner memory cost.

4 Experimental setup

4.1 The Natural Stories corpus

The Natural Stories corpus (Futrell, Gibson, Tily, Vishnevetsky, Piantadosi, & Fedorenko, in prep) is a set of 10 texts written to sound fluent while containing many low-frequency and marked syntactic constructions, especially subject- and object-extracted relative clauses, clefts, topicalized structures, extraposed relative clauses, sentential subjects, sentential complements, local structural ambiguity, and idioms. Self-paced reading time data was collected over these texts from 181 native English speakers.

One reason that previous corpus studies might have failed to find locality and integration effects is that the texts might not have included the low-frequency constructions where such effects emerge. Naturalistic texts, such as the newspaper columns forming the Dundee corpus, are produced and edited to be understood, so they will not frequently contain the kinds of low-frequency, hard-to-process events that bring out differences between processing models. The Natural Stories corpus is designed to exercise such models using constructions which are known to be difficult, providing an opportunity for memory effects to emerge where they have been obscured otherwise. The Natural Stories corpus contains 848,207 reading events. To control for edge effects, we filtered out all tokens occurring at sentence start and end, leaving 768,023 events. These were then divided into an exploratory corpus of 255,554 events and a confirmatory corpus of 512,469 events.[4]

4.2 Memory predictor implementations

The 8 DLT predictors and 12 left-corner predictors discussed in §3 were implemented over gold-standard trees in the Generalized Categorial Grammar (GCG) framework of Nguyen et al. (2012). Source trees for the entire corpus were hand-corrected by a single expert annotator from an automatic reannotation from gold-standard Penn Treebank style representations, which are distributed with the Natural Stories corpus. The GCG framework was chosen because it contains an implicit representation of syntactic dependencies and because it can be used to calculate incremental representations of the memory store of a left-corner parser. This allowed us to compute all predictors under consideration from source trees.

To control for memory-independent information theoretic effects, for each word in the corpus we also computed 5-gram forward probabilities from the Gigaword 4.0 corpus (Graff and Cieri, 2003) using the KenLM toolkit (Heafield et al., 2013) and PCFG surprisal using the van Schijndel et al. (2013a) parser.

It is an open question as to when during processing the effects in question will occur. For example, while readers may slow down when they encounter the final word of a center-embedding region, it is also possible that they would not slow down until the following word, when the need for integration is confirmed. In addition, self-paced reading (SPR) data are known to sometimes produce later effects (Kaiser, 2014; Jegerski, 2014). We therefore calculate variants of each of these 20 predictors in four spillover positions, yielding 80 possible main effects.

4.3 Statistical evaluation

Each of our 80 predictors was evaluated via likelihood ratio test of two linear mixed-effects (LME) models fitted to the exploratory dataset: a baseline model with the main fixed effect omitted, and a test model with the main fixed effect included. All models included sentence position, word length, 5-gram forward surprisal, and total PCFG surprisal as fixed effects, along with by-subject random slopes for each of these, a by-subject random slope for the main effect, and random intercepts for each subject and word. To control for sentence-level confounds, we additionally included a by-subject random slope and random intercept for sentence ID. To facilitate convergence and maintain comparability between predictors, all predictors were centered and z-transformed prior to fitting.

The likelihood ratio test assumes normally-distributed data, so we used the Box and Cox (1964) transform ($\lambda \approx -0.63$) to assure that the data match these assumptions as closely as possible.[5] Significant improvement to model fit for a given main effect indicates that it predicts reading times independently of all controls. The most significantly predictive effects on the exploratory corpus were selected for evaluation on the confirmatory corpus (see § 4.1 for discussion of the exploratory/confirmatory partition).

5 Results

Exploratory results revealed highly significant effects for a number of predictors. The 13 most significant of these were on the word following the target (spillover-1 (S1) position). This might suggest that listeners wait for confirmation of their syntactic analysis before attempting to retrieve items from working

[4]Code to reproduce this experiment is distributed through the ModelBlocks and NaturalStories repositories on Github.com.

[5]The Box and Cox (1964) transformation is $y' = \frac{y^\lambda - 1}{\lambda}$. We selected $\lambda \approx -0.63$ via likelihood maximization.

		Exploratory corpus				Confirmatory corpus			
		β	β-ms	t-value	p-value	β	β-ms	t-value	p-value
Best	**NoF-S1**	1.23e-4	1.29	6.66	1.45e-10	1.46e-4	1.54	8.15	2.33e-14
	DLT-CM-S1	1.11e-4	1.16	5.85	1.42e-8	9.63e-5	1.10	6.48	4.87e-10
Canon	**REINST-S1**	1.17e-4	1.23	6.33	1.60e-9	1.35e-4	1.43	8.01	5.77e-14
	DLT-S1	8.04e-5	0.846	4.51	1.03e-05	6.04e-05	0.634	4.50	1.11e-05

Table 1: Evaluation results. **Upper:** Best left-corner (NoF-S1) and DLT (DLT-CM) predictors. **Lower:** Canonical DLT and left-corner (REINST) predictors. **Left:** Results on exploratory corpus. **Right:** results on confirmatory corpus. Column β contains the LME effect estimate per SD of the independent variable, which is valid over Box and Cox (1964)-transformed reading times. Column β-ms is a back-transformation of β into milliseconds using the equation β-ms = $(\lambda\bar{y}' + \lambda\beta + 1)^{1/\lambda} - (\lambda\bar{y}' + 1)^{1/\lambda}$, where $\bar{y}'$ is the mean of the transformed reading times (1.55 in our data). Because Box and Cox (1964) introduces non-linearity, β-ms is only valid at the back-transformed mean, holding all other effects at their means.

memory. It could also be an artifact of the aforementioned tendency for effects to be delayed in self-paced reading (SPR) experiments. The most significant DLT predictor was DLT-CM-S1 (the DLT with coordination and modifier modifications in S1 position), and the most significant left-corner predictor was NoF-S1 (the no-fork boolean predictor in S1 position). These predictors were therefore selected for confirmatory evaluation, along with S1 'canonical' predictors for each family (unmodified DLT and boolean REINST). These four predictors (and no others) were then evaluated on the confirmatory corpus, with results given in Table 1. The confirmatory results indicate that all four effects generalize robustly to new data, with all achieving p-values well below the Bonferroni-corrected significance threshold of $p = 0.0125$ for four comparisons. The left-corner predictors have a higher order of significance than the DLT predictors and larger effect estimates. Because the main effects are z-transformed, β values are per standard deviation. Over our entire data set, noF = 1 is 2.51 SD greater than noF = 0,[6] so a recall event is predicted to produce a delay approximately 2.5 times larger than β (3.88ms in β-ms). DLT-CM ranges between 0 and 13.19 SD (DLT-CM = 12) in our data, with 92% of events $\leq$ 1.1 SD (DLT-CM = 1) and 99% $\leq$ 4.4 SD (DLT-CM = 4). The effective predictions for noF-S1 are therefore larger than those for most instances of DLT-CM-S1 $>$ 0, but at extreme values DLT-CM predicts larger effects.

6 Discussion

The principal contribution of this work is to give the first strong evidence of memory effects in broad-coverage sentence processing. The constructed-natural Natural Stories corpus used here reduces the likelihood of confounds due to lack of context or oddball sentences to which studies using constructed stimuli are vulnerable, as well as the likelihood of confounds due to lack of memory-intensive syntax or small numbers of subjects to which studies using naturalistic stimuli are vulnerable. Our rigorous baseline model, which includes controls for n-gram and PCFG surprisal, helps ensure that the observed effects are not due to other plausible sources of processing difficulty. Despite these controls, our evaluation results are highly significant.

In order to evaluate whether our DLT and boolean left-corner predictors could both be driven by a single effect, we ran a four-way LME comparison of models on the exploratory corpus (1) with both DLT-CM-S1 and NoF-S1 ablated, (2) with one or the other of the effects ablated, and (3) with neither ablated. Both effects significantly improved over the baseline on their own, and the joint model significantly improved over both effects individually, indicating that neither effect is reducible to the other.[7]

Our distance-weighted left-corner predictors (especially REINST-LEN) are very similar to the ACT-R retrieval predictors of e.g. Lewis and Vasishth (2005), Boston et al. (2011), and von der Malsburg et al. (2015). Many of our distance-weighted left-corner predictors showed positive effects on the exploratory

[6]This is because memory recalls are predicted for a small proportion ($\approx$20%) of all events.

[7]DLT-CM-S1 over baseline: p = 7.24e-12; NoF-S1 over baseline: p = 7.93e-10; both over NoF-S1: p = 6.33e-13; both over DLT-CM-S1: p = 6.87e-11. Note that in order to achieve convergence in all models, we removed controls for sentence ID from the model specification.

corpus in spillover-1 position (e.g., NoF-LEN-S1, β = 5.98e-05, p = 9.43e-4). However, these effects were substantially weaker than those of the effects selected for confirmatory evaluation. The strength of the DLT predictors in comparison to the left-corner predictors on our exploratory data suggests that our DLT effect is not simply capturing the effects of decay in left-corner parsing, further supporting the independence of the DLT effect from effects related to left-corner parsing.

Given this, we now consider some important differences in the predictions of both frameworks. First, the left-corner effects predict processing difficulty exclusively on the basis of syntactic tree configurations, while the DLT effects predict difficulty on the basis of a combination of syntax and semantics, namely (1) asymmetries in referential status between nouns/verbs and other words, and (2) word-to-word dependencies. Also, the DLT computes memory costs as soon as both words in the dependency arc are encountered, while the distance-weighted left-corner predictors compute memory costs at the right edges of subtrees in the incremental parse. They therefore not only compute cost differently but predict the cost to be incurred at different words.

The DLT also differs in important ways from our distance-weighted left-corner predictors (like REINST-LEN). The distance-weighted left-corner predictors track recency of activation of derivation fragments by measuring the distance to the most recent word in a fragment. This distance is a function of the size of the attentionally-focused derivation fragment rather than that of the fragment being recalled from memory. By computing total dependency length, the DLT can also index the complexity (in nouns and verbs) of the stored derivation fragment, depending on the location of its head. More complex derivations might be more difficult to retrieve, a possibility which none of the left-corner predictors are designed to account for.

By contrast, the boolean left-corner predictors like NoF-S1 capture phrase-structural information that is absent from the DLT, flagging moments in processing at which derivation fragments stored in memory are predicted to be accessed (which do not necessarily correspond to endpoints of word-to-word dependencies). The fact that NoF-S1 is much more successful on our data than its distance-weighted counterparts suggests that memory-related processing difficulty may be more a function of whether a memory access event has occurred than of recency of activation. Finally, the fact that NoF-S1 is slightly more successful than REINST-S1 – which is identical to NoF-S1 except that it also flags the ends of long-distance dependency carriers – shows that sensitivity to long-distance dependencies does not improve our left-corner predictor. It is therefore possible that the storage and retrieval of incomplete derivation fragments differ mechanistically from the resolution of long distance dependencies.

Acknowledgements

This work was supported by grants from the National Science Foundation: Graduate Research Fellowship Program Award DGE-1343012 to MvS; Doctoral Dissertation Research Improvement Award 1551543 to RF; Linguistics Program Award 1534318 to EG.

References

Steven P. Abney and Mark Johnson. 1991. Memory requirements and local ambiguities of parsing strategies. *J. Psycholinguistic Research*, 20(3):233–250.

Alfred V. Aho and Jeffery D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling, Vol. 1: Parsing*. Prentice-Hall, Englewood Cliffs, New Jersey.

Asaf Bachrach, Brian Roark, Alex Marantz, Susan Whitfield-Gabrieli, Carlos Cardenas, and John D.E. Gabrieli. 2009. Incremental prediction in naturalistic language processing: An fMRI study.

Marisa Ferrara Boston, John T. Hale, Shravan Vasishth, and Reinhold Kliegl. 2011. Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, 26(3):301–349.

G. E. P. Box and D. R. Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society, B*, 26:211–234.

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

Edward Gibson. 1991. *A computational theory of human linguistic processing: Memory limitations and processing breakdown*. Ph.D. thesis, Carnegie Mellon.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first mind articulation project symposium*, pages 95–126, Cambridge, MA. MIT Press.

David Graff and Christopher Cieri, 2003. *English Gigaword LDC2003T05*.

Daniel J. Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input. *Cognitive Science*, 29:261–291.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria, August.

Jill Jegerski. 2014. Self-paced reading. In Jill Jegerski and Bill VanPatten, editors, *Research methods in second language psycholinguistics*, pages 20–49. Routledge, New York.

Philip N. Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press, Cambridge, MA, USA.

Elsi Kaiser. 2014. Experimental paradigms in psycholinguistics. In Robert J. Podesva and Devyani Sharma, editors, *Research methods in linguistics*, pages 135–168. Cambridge University Press, Cambridge.

Alan Kennedy, James Pynte, and Robin Hill. 2003. The Dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

R. Kliegl, E. Grabner, M. Rolfs, and R. Engbert. 2004. Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16(1):262–284.

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

Luan Nguyen, Marten van Schijndel, and William Schuler. 2012. Accurate unbounded dependency recovery using generalized categorial grammars. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING '12)*, pages 2125–2140, Mumbai, India.

Philip Resnik. 1992. Left-corner parsing and psychological plausibility. In *Proceedings of COLING*, pages 191–197, Nantes, France.

William Schuler, Samir AbdelRahman, Tim Miller, and Lane Schwartz. 2010. Broad-coverage incremental parsing using human-like memory constraints. *Computational Linguistics*, 36(1):1–30.

Edward Stabler. 1994. The finite connectivity of linguistic structure. In *Perspectives on Sentence Processing*, pages 303–336. Lawrence Erlbaum.

Marten van Schijndel and William Schuler. 2013. An analysis of frequency- and memory-based processing costs. In *Proceedings of NAACL-HLT 2013*. Association for Computational Linguistics.

Marten van Schijndel, Andy Exley, and William Schuler. 2013a. A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, 5(3):522–540.

Marten van Schijndel, Luan Nguyen, and William Schuler. 2013b. An analysis of memory-based processing costs using incremental deep syntactic dependency parsing. In *Proc. of CMCL 2013*. Association for Computational Linguistics.

Titus von der Malsburg, Reinhold Kliegl, and Shravan Vasishth. 2015. Determinants of scanpath regularity in reading. *Cognitive Science*.

Stephen Wu, Asaf Bachrach, Carlos Cardenas, and William Schuler. 2010. Complexity metrics in an incremental right-corner parser. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, pages 1189–1198.