Effects of Filler-gap Dependencies on Working Memory Requirements for Parsing

William Schuler (schuler@ling.osu.edu)
Department of Linguistics, 1712 Neil Avenue,
Columbus, OH 43210 USA

Abstract

Corpus studies by Schuler, AbdelRahman, Miller, and Schwartz (2010), appear to support a model of comprehension taking place in a general-purpose working memory store, by providing an existence proof that a simple probabilistic sequence model over stores of up to four syntactically-contiguous memory elements has the capacity to reconstruct phrase structure trees for over 99.9% of the sentences in the Penn Treebank Wall Street Journal corpus (Marcus, Santorini, & Marcinkiewicz, 1993), in line with capacity estimates for general-purpose working memory, e.g. by Cowan (2001). But capacity predictions of this simple structure-based model ignore non-structural dependencies, such as long-distance filler-gap dependencies, that may place additional demands on working memory. Distinguishing unattached gap fillers from open attachment sites in syntactically-contiguous memory elements requires this contiguity constraint to be strengthened to a constraint that working memory elements be *semantically* contiguous. This paper presents corpus results showing that this stricter semantic contiguity constraint still predicts working memory requirements in line with capacity estimates such as that of Cowan (2001).

Keywords:

Introduction

It is tempting to think of sentence comprehension as learned manipulations of elements in a general-purpose working memory store. This assumption underlies many established comprehension models (e.g. Johnson-Laird, 1983; Elman, 1991; Gibson, 1991; Just & Carpenter, 1992; Gibson, 1998; Lewis & Vasishth, 2005, making various assumptions about the size and nature of this memory). Corpus studies by Schuler et al. (2010), appear to support this hypothesis by providing an existence proof that a simple probabilistic sequence model over stores of up to four syntactically-contiguous memory elements — simple random variables with discrete domains over pairs of constituent categories at which other structures may attach — has the capacity to reconstruct phrase structure trees for over 99.9% of the sentences in the Penn Treebank Wall Street Journal corpus (Marcus et al., 1993). This is in line with capacity estimates for general-purpose working memory, e.g. by Cowan (2001), but is also compatible with a continuously degrading availability (McElree, 2001), since the model's use of this store degrades very rapidly after one element. To the extent that the Wall Street Journal is comprehensible (and its editors are doing their jobs) this suggests that comprehension can take place in a general-purpose working memory store, using a small set of incomplete constituent states derived from phrase structure.

The structural attachment sites in the memory elements of Schuler et al. (2010) are necessary in order to accurately reconstruct syntactic relations in phrase structure trees. But parsing is only part of comprehension. In order to obtain valid interpretations, a comprehension model must also retain non-structural information like gap fillers in long-distance dependency constructions. Filler-gap dependencies are common (occurring in about 20% of sentences in the Wall Street Journal corpus), and figure prominently in the psycholinguistics literature on memory bounds in parsing (Gibson, 1991; Just & Carpenter, 1992, etc.). But since the introduction of a gap filler adds no unsatisfied structural attachment sites (and thus no need to retain additional memory elements), they are ignored in the capacity predictions of the simple syntax-based model described above. Distinguishing unattached gap fillers from open attachment sites in syntactically-contiguous memory elements requires the syntactic contiguity constraint of Schuler et al. (2010) to be strengthened to a constraint that working memory elements be *semantically* contiguous (that is, linked by roles, in the sense of Gibson, 1991). This would force gap fillers into separate memory elements, requiring additional memory resources to process sentences in which filler-gap dependencies co-occur with structurally-nested constituents.

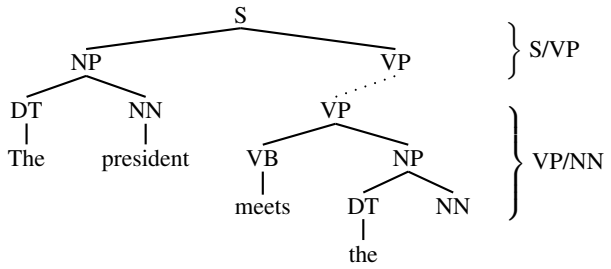

Figure 1: Incomplete constituents in an incremental parse of the sentence 'The president meets the board on Friday,' showing non-immediate dominance between incomplete constituents 'The president ...' and 'meets the ...'.

This paper presents coverage results on the same Wall Street Journal corpus showing that, despite their prevalence, long-distance dependencies do not seem to occur in deeply structurally-embedded contexts, and this stricter semantic contiguity constraint still predicts working memory requirements in line with capacity estimates such as that of Cowan (2001). This seems to support the hypothesis that comprehension may take place in a general-purpose working memory store, using a simple notion of semantically-contiguous incomplete constituents as memory elements.

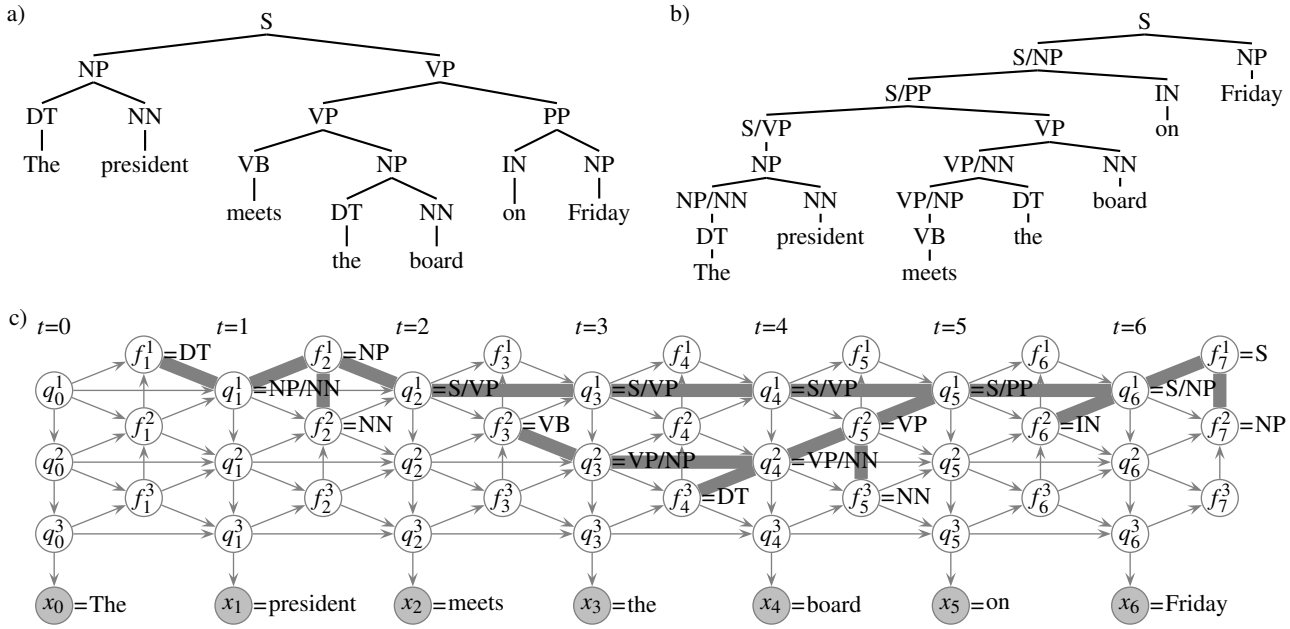

Figure 2: Phrase structure tree for the sentence 'The president meets the board on Friday' (a), transformed into right-corner form (b), then mapped (in dark gray) onto a random variables in a factored sequence model (c) with three hidden levels. Circles denote random variables (over incomplete constituents q_t^d and complete constituents f_t^d at each nesting depth d and time step t), and edges denote conditional dependencies. Shaded circles denote variables with observed values (words in this case).

Background

Schuler et al. (2010) calculate a first approximation of the working memory capacity required to parse the large syntactically-annotated Penn Treebank Wall Street Journal and Switchboard corpora, based on what was intended to be a strict requirement that only completely contiguous syntactic structures could occupy a single working memory element. In particular, each syntactically contiguous chunk is constrained to the form of an *incomplete constituent* state A/B, consisting of a single *active* but unfinished constituent A lacking a single *awaited* constituent B yet to be attached, somewhere in the right progeny of the active constituent. Syntactic relations between these incomplete constituent chunks are underspecified as non-immediate dominance relations between the awaited and active components of successive incomplete constituents (see Figure 1). This can be thought of as a highly-constrained version of the non-immediate dominance relations in Tree Adjoining Grammar (Joshi, 1985) or Description Tree Grammar (Rambow, Weir, & Vijay-Shanker, 1995) in processing models proposed by Stabler (1994) and Mazzei, Lombardo, and Sturt (2007), except that here, all syntactic information other than the categories of active and awaited constituents at the frontier of an incomplete constituent is discarded.

This austere definition still allows the complete specification of phrase structure trees from stores of incomplete constituents arranged in time order (see Figure 2). This correspondence can be defined through a reversible right-corner transform (Schuler et al., 2010), a variant of the left-corner transform of Johnson (1998), associating phrase structure trees (Figure 2a) with memory-minimizing transformed representations (Figure 2b). This is done by associating every top-down sequence of right children between some left child[1] and its rightmost leaf (say, from the root S to the NP 'Friday' in Figure 2a) with a bottom-up sequence of incomplete constituents, each having the original left child as its active component and one of the original right children as its awaited component (producing the sequence S/VP, S/PP, S/NP in Figure 2b). This representation converts right-expanding sequences of complete constituents into left-expanding sequences of incomplete constituents, leaving only center-expanding sequences (alternating expansions of left and right children) to require additional memory resources in a bottom-up time-order traversal.

This memory-minimizing representation can then be mapped to random variables in a sequence model (Figure 2c), with incomplete constituents mapped to store state variables q_t^d and complete constituents mapped to final state variables f_t^d. Connections among these variables define probabilities for partial utterances, in which values are hypothesized for each random variable with probability conditioned on only its adjacent antecedent variables (those connected by outgoing arcs).[2] Each time step in the model (corresponding to columns in Figure 2c) defines a set of incomplete constituents recognised thus far. For example, the store $q_t^{1..D}$ at

[1] For the purpose of this definition, the root of a tree is considered to be a left child (e.g. of a right-branching supra-sentential discourse structure).

[2] The probability of a partial utterance at any time step t, subsuming a store state $q_t^{1..D}$, and the set of observed words $x_{1..t}$ to that time

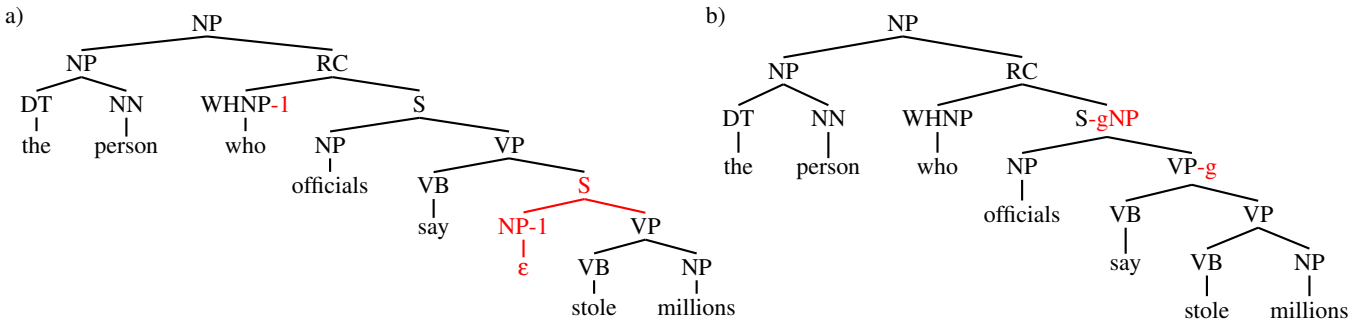

Figure 3: Mapping from the movement-based notation in the Penn Treebank (a) to a purely binary-branching structure (b) on which the right-corner transform is defined.

t=4 corresponds to the set of incomplete constituents in Figure 1 prior to encountering the word 'board'.

Figure 2c shows a particular instantiation of random variables corresponding to one hypothesized analysis of the example sentence, 'The president meets the board on Friday.' In processing, several such analyses are maintained in parallel at each time step, competing with one another probabilistically in subsequent transitions. The parallelism in a probabilistic sequential process such as this one can be maintained using distributed, independent computations of probability mass in a particle filter (Gordon, Salmond, & Smith, 1993), which may resemble distributed processing in human cognition (Levy, 2008b).

As a probabilistic sequence model, this model is recurrent (in that it is stationary, using the same model at each time step) and connectionist (in that it is defined entirely in terms of interconnected nodes). But unlike other recurrent connectionist models, which are typically specified at the level of excitatory and inhibitory relations between individual neural elements, probabilistic sequence models may be specified directly over linguistic states (in this case over recognized incomplete constituents at successive time steps). Syntactic probabilities in a sequence model transformed from a probabilistic context free grammar (PCFG) can be defined to generate the same tree and sentence probabilities as the original PCFG (Schuler, 2009). These probabilities are proposed to have a significant role in processing (Jurafsky, 1996; Hale, 2001; Levy, 2008a, among others), and probabilities from partial sequences in this model have been shown to correlate with reading time delays in self-paced reading (Wu, Bachrach, Cardenas, & Schuler, 2010).

step, is:

$$\mathsf{P}(q_t^{1..D} x_{1..t}) \stackrel{\text{def}}{=} \sum_{q_{t-1}^{1..D}} \mathsf{P}(q_{t-1}^{1..D} x_{1..t-1}) \cdot \mathsf{P}(x_t \,|\, q_t^{1..D}) \cdot \left[\sum_{f_t^{1..D}} \prod_{d=1}^{D} \mathsf{P}(f_t^d \,|\, f_t^{d+1} q_{t-1}^d q_{t-1}^{d-1}) \cdot \mathsf{P}(q_t^d \,|\, f_t^{d+1} f_t^d q_{t-1}^d q_t^{d-1}) \right] \quad (1)$$

in which the probability terms within brackets have conditional dependencies defined by the network in Figure 2c, and the probability terms outside the brackets are recursive store state probabilities and evidence probabilities of a Hidden Markov Model (Rabiner, 1990).

Revised Model

The active and awaited components of incomplete constituents retained in this model are the only sites at which subsequently recognised phrase structure will attach, either above or below the incomplete constituent. But in order to extract propositional content from an utterance or read sentence, a comprehension model must also retain semantic referents of unattached gap fillers, which are not necessarily manifested as structural attachment sites. This retention can be accomplished by redefining incomplete constituents to be semantically contiguous as well as syntactically contiguous.

Following Schuler et al. (2010) and Lewis and Vasishth (2005), the model described in this paper will assume a purely binary-branching phrase structure with no empty constituents, in order to simplify the definition of the comprehension process. This can be generated from trace-annotated corpora (see Figure 3a) by eliminating each empty constituent (e.g. NP-1), and the constituent attaching it to the non-empty portion of the tree (the S dominating NP-1), then propagating the trace index up the tree from this gap position to the corresponding filler position (WHNP-1), attaching a '-g' tag to each traversed constituent category indicating that constituent contains a gap but no corresponding filler (see Figure 3b). At the topmost gapped constituent in this traversal, the category of the filler is added to the tag (producing S-gNP), indicating the constituent at which the need for a gap is introduced in a time-order traversal.

In a complete comprehension model, it would be desirable to allow semantic dependencies from referents of fillers to referents of constituents containing gaps to be expressed interactively, so fillers could statistically influence subsequent parsing decisions. This can be done through the conditional dependencies among random variables in the sequence model defined in the previous section. But, since fillers and gaps are not generally adjacent in phrase structure trees, they are not guaranteed to be adjacent in a right-corner transformed tree. For example, the right-corner transform defined in the previous section would transform the simple top-down right-branching sequence of NP, RC, S-gNP, VP-g, etc. in

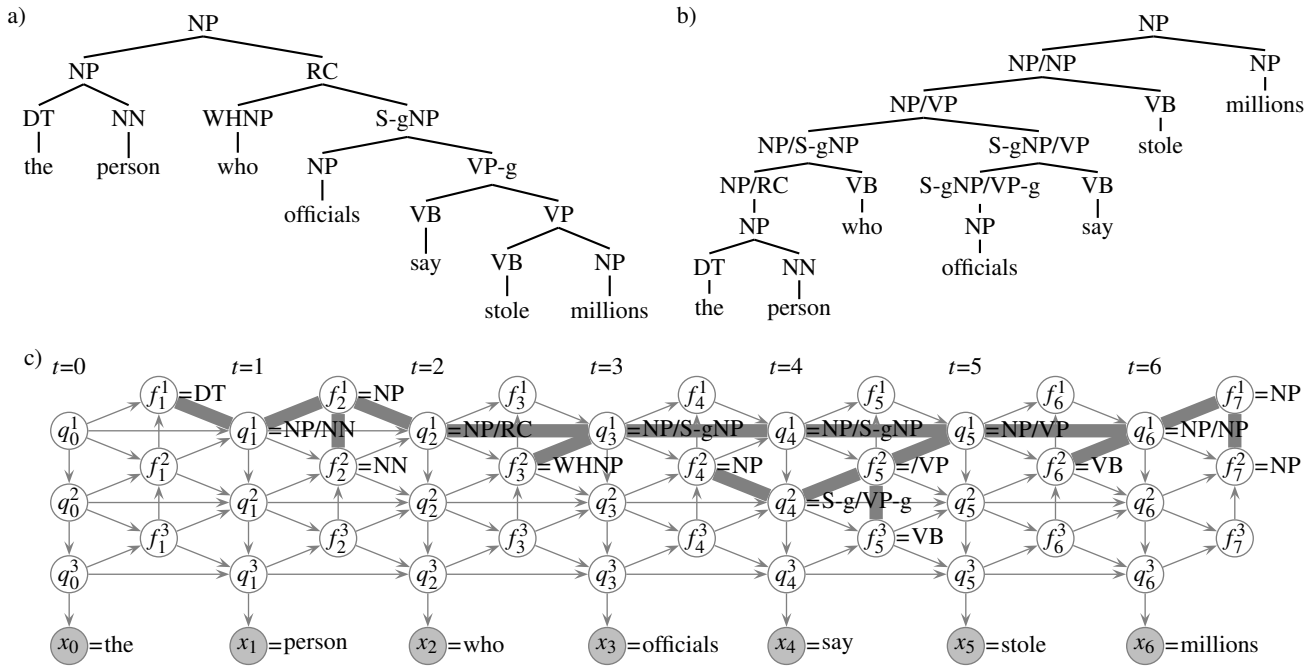

Figure 4: Sample store sequence containing long-distance dependency in a filler-gap construction.

Figure 3b into a simple bottom-up left-branching sequence NP/RC, NP/S-gNP, NP/VP-g, etc. This would leave the random variables associated with the filler and gapped constituent separated by two time steps in the sequence model. Moreover, some of the resulting incomplete constituents (e.g. the NP/VP-g dominating 'the person who officials') would have no fixed semantic dependency between referents of their active and awaited components (between the person and whatever the officials are going to do), violating an assumption that memory elements contain coherent or contiguous chunks.

The right-corner transform is therefore defined to explicitly retain the filler as the awaited component of an incomplete constituent, now both syntactically and semantically contiguous. It does this by halting on this initial gap (see Figure 4a, a copy of Figure 3b), treating the initial gap constituent as a left child, then transforming the sequence of right children below it into a left-expanding sequence of incomplete constituents as a sub-structure (the sub-structure subsuming 'officials say' in Figure 4b). At the last constituent with a gap tag (VP-g in the figure), the transform halts again and terminates this sub-structure with an incomplete constituent (S-gNP/VP in the figure) consisting of the initial gap constituent as the active component and the right child of the current constituent as the awaited component. Finally, this incomplete constituent is connected to the incomplete constituent preceding it (NP/S-gNP subsuming 'the person who') by attaching both as children of another incomplete constituent (NP/VP in the figure), consisting of the active component of this previous incomplete constituent and the awaited component of the incomplete constituent dominating the sub-structure. The transform then resumes constructing incomplete constituents from the bottom up in time order, as described in the previous section. This transformed representation can then be mapped to random variables in a sequence model in Figure 4c as described in the previous section, except that the redundant active component of the final state at the end of the gap sub-structure (the S-gNP at f_5^2) is elided, in order to preserve the model definition and topology.

Implicit in this analysis of fillers is the assumption that each component of an incomplete constituent has a single associated referent. In the pure binary-branching phrase structure tree shown in Figure 4a, the constituent at which the gap is introduced (S-gNP) does not have a single referent. But when this constituent forms the split point between two incomplete constituents (NP/S-gNP and S-gNP/VP in Figure 4b), the awaited component of the upper incomplete constituent may take on the referent of the filler (the referent of the 'NP' in S-gNP, described by 'person' in this example) and the active component of the lower incomplete constituent may take on the referent of the original phrase structure constituent (the referent of the 'S', described by 'say' in this example). This allows the filler to be made available when the gap is encountered, while also potentially allowing subsequent structures to attach above S-gNP/VP to modify the 'say' event.

Results

This structural analysis of fillers as awaited components of new incomplete constituents allows them to be explicitly retained and associated with gap constituents in comprehension, despite the fact that they are not semantically related

a)

memory load	words	coverage
0 store elements	39,882	4.76%
1 store element	465,977	60.25%
2 store elements	290,550	94.85%
3 store elements	41,587	99.79%
4 store elements	1,745	99.99%
5 store elements	24	100.00%
TOTAL	839,765	100.00%

b)

memory load	words	coverage
0 store elements	39,882	4.76%
1 store element	464,044	60.01%
2 store elements	291,426	94.71%
3 store elements	42,454	99.77%
4 store elements	1,934	99.99%
5 store elements	25	100.00%
TOTAL	839,765	100.00%

c)

memory load	words	coverage
0 store elements	39,882	4.76%
1 store element	360,806	47.82%
2 store elements	264,265	79.36%
3 store elements	112,936	92.84%
4 store elements	40,090	97.62%
5 store elements	13,455	99.23%
6 store elements	4,440	99.76%
7 store elements	1,400	99.92%
8 store elements	499	99.98%
9 store elements	109	99.99%
TOTAL	837,882	100.00%

Table 1: Percent coverage of right-corner transformed Wall Street Journal Treebank sections 2–21, using original transform (a), and revised transform (b), and corpus randomly sampled from PCFG, using revised transform (c).

to the intervening words. But the addition of new incomplete constituents to hold fillers places additional demands on working memory. If filler-gap dependencies regularly co-occur with ordinary structural nesting, the capacity requirements of this model may become incompatible with independent estimates of general-purpose working memory capacity. In order to determine whether this model is still plausible, its capacity requirements were evaluated on the Wall Street Journal corpus (Marcus et al., 1993).

First, the Schuler et al. (2010) study was replicated as a baseline. Sections 2–21 of the Penn Treebank Wall Street Journal corpus were binarized, right-corner transformed, and mapped to elements in a bounded memory store as described in Schuler et al. (2010). Coverage of this corpus, in words that can be processed by a recognizer using one to five memory elements, is shown in Table 1a. These results show that a simple syntax-based chunking into incomplete constituents, using the right-corner transform defined in Schuler et al. (2010), allows a vast majority of words in the Wall Street Journal corpus (over 99.7%) to be recognized using three or fewer elements of memory, with no sentences requiring more than five elements, as predicted by a general-purpose working memory model.[3]

[3]This measure is more fine-grained than Schuler et al. (2010), which counted only the maximum depth at each sentence.

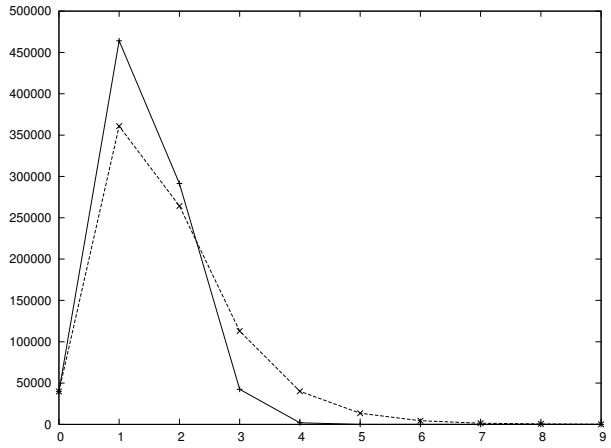

Figure 5: Number of words at each nesting depth for attested corpus (solid line) and randomly sampled corpus (dotted line).

The study was then repeated using the modified right-corner transform described in the previous section, to allow the model to explicitly retain gap fillers as awaited components of incomplete constituents. Results are shown in Table 1b. Although results show increased requirements at capacity 3 and 4, there is no substantial increase at capacity 5 or beyond, suggesting that long-distance filler-gap dependencies do not occur in memory-taxing portions of syntactic structure.

In order to determine whether these nesting limits arise from purely syntactic statistical tendencies (e.g. toward right-side branching in English) or from bounded-memory effects, the above results were compared to memory load results for a corpus of phrase structure trees that were randomly generated from the PCFG estimated from the relative frequency of each branch in the binarized Treebank corpus, with no sensitivity to nesting depth (shown in Table 1c and Figure 5). The capacity requirements of the randomly generated corpus form a relatively mild peak at d=1 and exhibit a gradual exponential decay at each higher capacity requirement (the dashed line in Figure 5). This is to be expected form a pure PCFG model, since additional nestings at each level are generated with the same probability. In contrast, the capacity requirements for the actual sentences in the Wall Street Journal corpus using the revised model of filler-gap dependencies described in this paper (solid lines) peak much more sharply at d=1, and then fall off much more rapidly. The difference between these distributions is statistically significant over sentences ($p < .01$) using a two-tailed t-test at all depths below d=5, for which non-zero counts exist in both corpora, except for the crossover point at d=2 which is not significant. This suggests that bounded-memory effects play a significant role in syntactic structure beyond what can be explained by syntactic preferences in a PCFG, even when accounting for memory

required to connect filler-gap dependencies. This may argue in favor of the use of factored sequence models in place of pure PCFG models as a source of incremental probability estimates in modeling comprehension.

Discussion

This paper has presented predicted capacity requirements for incremental parsing with a stricter condition for semantically contiguous memory elements showing that the required working memory capacity does not significantly increase when a more sensitive semantic contiguity constraint is introduced. The model is similar to Lewis and Vasishth (2005), except that the focus is on the estimation model rather than the time course.

This result covers filler-gap dependencies, but may still ignore other types of semantic dependencies that cause discontiguities. One source of such discontiguities may be quantifier scope raising. The definition of chunks in the right-corner sequence model described here can be readily extended to incremental calculation of quantifier scope by allowing semantic structures for quantifiers and restrictors (the semantics of noun phrases) to attach to exposed active constituents less deeply nested in an analysis, while the syntax transitions the syntactic state of the lowest incomplete constituent. Like filler-gap dependencies, this might predict additional required memory capacity to maintain semantically contiguous, monotonically growing chunks. But the effect on predicted capacity requirements must await corpora annotated with quantifier scope.

References

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195–225.

Gibson, E. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown*. Unpublished doctoral dissertation, Carnegie Mellon.

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

Gordon, N. J., Salmond, D. J., & Smith, A. F. M. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, *140*(2), 107–113.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the north american chapter of the association for computational linguistics* (pp. 159–166). Pittsburgh, PA.

Johnson, M. (1998). Finite state approximation of constraint-based grammars using left-corner grammar transforms. In *Proceedings of COLING/ACL* (pp. 619–623). Montreal, Canada.

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

Joshi, A. K. (1985). How much context sensitivity is necessary for characterizing structural descriptions: Tree adjoining grammars. In L. K. D. Dowty & A. Zwicky (Eds.), *Natural language parsing: Psychological, computational and theoretical perspectives* (pp. 206–250). Cambridge, U.K.: Cambridge University Press.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science: A Multidisciplinary Journal*, *20*(2), 137–194.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122-149.

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

Levy, R. (2008b). Modeling the effects of memory on human online sentence processing with particle filters. In *Proceedings of NIPS* (pp. 937–944). Vancouver, BC, Canada.

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

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Mazzei, A., Lombardo, V., & Sturt, P. (2007). Dynamic tag and lexical dependencies. *Research on Language and Computation*, *5*, 309–332.

McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology, Leanrning Memory and Cognition*, *27*(3), 817–835.

Rabiner, L. R. (1990). A tutorial on hidden Markov models and selected applications in speech recognition. *Readings in speech recognition*, *53*(3), 267–296.

Rambow, O., Weir, D., & Vijay-Shanker, K. (1995). D-tree grammars. In *Proceedings of the 33rd annual meeting of the association for computational linguistics (ACL'95)* (pp. 151–158).

Schuler, W. (2009). Parsing with a bounded stack using a model-based right-corner transform. In *Proceedings of naacl* (pp. 344–352). Boulder, Colorado.

Schuler, W., AbdelRahman, S., Miller, T., & Schwartz, L. (2010). Broad-coverage incremental parsing using human-like memory constraints. *Computational Linguistics*, *36*(1), 1–30.

Stabler, E. (1994). The finite connectivity of linguistic structure. In *Perspectives on sentence processing* (pp. 303–336). Lawrence Erlbaum.

Wu, S., Bachrach, A., Cardenas, C., & Schuler, W. (2010). Complexity metrics in an incremental right-corner parser. In *Proceedings of the 48th annual meeting of the association for computational linguistics (ACL'10)* (pp. 1189–1198).