What's in an error?
A detailed look at SRNs processing relative clauses.

Lars Konieczny[1] (lars@cognition.uni-freiburg.de)
Nicolas Ruh[2] (nruh@brookes.ac.uk)
Daniel Müller[1] (daniel@cognition.uni-freiburg.de)

[1]Center for Cognitive Science, Friedrichstr. 50
79098 Freiburg i. Br., Germany

[2]Oxford Brookes University

Abstract

This paper responds to MacDonald and Christiansen's (2002) experience-based account of subject vs. object relative clause processing based on Simple Recurrent Network simulations. They found that object-extracted relative clauses exhibit performance penalties that are absent in subject relative clauses, and more so in less trained networks. Whereas MC argue that their finding reflects a differential amount of word order regularity in subject- vs. object-extractions, a detailed analysis of the word-by-word output-activation pattern suggests that it is caused by the network failing to distinguish verbs from the relative pronoun *that* during early training epochs. This interpretation is supported by other aspects of the activation pattern that indicate incomplete grammar acquisition. Nevertheless, the results point at a viable source of complexity in sentence processing.

Introduction

Relative Clauses and working memory

The contrast between Subject-extracted (2) and Object-extracted relative clauses (1) is the poster child of working-memory oriented psycholinguistics.

(1) The reporter *who the senator attacked* admitted the error (ORC)

(2) The reporter *who attacked the senator* admitted the error (SRC)

Subject-relative clauses (SRCs, 1) are generally easier to process than object-RCs (2), and more notably so for readers with a low reading span (King & Just, 1991). Among the multitude of models, two fundamentally opposing frameworks have been most prominent: retrieval-based working memory models, (eg. Just & Carpenter, 1992, Gibson, 1998, Gordon et al., 2004, Vasishth & Lewis, 2005), and experience-based models, such as probabilistic parsers (Hale 2001, Levy, 2005) and connectionist models, most notably that of MacDonald and Christiansen (2002). Their model is based on Simple Recurrent Networks (Elman, 1990). SRNs acquire implicit grammatical knowledge when they are trained on linguistic corpora. Crucially, they lack a clear distinction between linguistic knowledge, processing, and a knowledge-free notion of a working memory and its capacity. In MCs' SRN-based approach, the complexity difference between Subject and Object-RCs can be attributed to the differential degree of word-order regularity exhibited by SRCs and ORCs. Subject-RCs match the predominant subject-verb-object (SVO) word order of simple main clauses. Object-RCs, on the other hand, show an irregular O-S-V order. Processing SRCs hence benefits from "regular" word order expectations being transferred from main clauses, whereas no such transfer is made for ORCs. Therefore, SRCs are easier to process than ORCs despite the relatively low frequency of relative clauses in general. These predictions were – in principle at least – confirmed by MC's simulations with Simple Recurrent Networks. These networks were trained in three epochs of 10000 random sentences each. Because SRCs were easy even in the earliest training epoch, only ORCs benefited from more training. The resulting grammatical error pattern shows striking resemblance to the reading times of the different span groups of King and Just (1991). MacDonald and Christiansen (2002) hence attribute the differential performance of span-groups to their respective amount of linguistic experience rather than differences in working memory capacity. Basically, they reveal a – this time word-order-based – frequency (amount of training) x regularity (i.e. transfer from predominant order) interaction comparable to what has been demonstrated for other connectionist models in a variety of domains (e.g. Seidenberg & McClelland, 1989).

In this paper, we will show that MC's critical results can be attributed to a fundamental part of speech classification error due to insufficient learning in early epochs. We will argue however that the underlying mechanism of interference by locally coherent predictions might very well be a valid predictor for processing complexity.

SRNs and sentence processing

SRNs have successfully been demonstrated to be capable of implicitly acquiring limited recursive "grammars" (e.g. Elman, 1991; Christiansen & Chater, 1999). They do so by learning to predict the next word when presented with sentences word-by-word at the input. In the SRN architecture, there is a hidden layer that receives combined activation from the input layer and the context layer, which holds the content of the hidden-layer at the previous cycle. Using the standard back-propagation algorithm, the prediction-error, reflecting the deviance of the predicted activation pattern from the actual next word pattern, is used to adjust connection weights throughout the network back to the input layer. Eventually, after thousands of learning cycles, the SRN performs reasonably well even on sentences that it has never seen before. At this point, SRNs can be demonstrated to have classified words into their syntactic categories and possibly even into more fine-grained semantic distinctions (Elman, 1990). SRNs have repeatedly been demonstrated to be able to acquire an implicit recursive grammar (Elman, 1991, Christiansen & Chater, 1999).

As a measure of the grammatical viability of the network's predictions, output vectors are compared to grammaticality vectors calculated from the underlying context free grammar used to generate the training set. Each unit corresponds to a lexicon entry (word) and carries its grammatical probability in the context of the previous words in the sentence. For instance, if there are two grammatical continuations, both equally likely, the corresponding units both have a probability of 0.5 and should hence receive 50% of the output activation each.

Deviation from this activity pattern increases the *grammatical prediction error* (GPE). The GPE is a global error measure (i.e. the specific errors on each output unit are collapsed into a single value) ranging from zero to one, with zero meaning a perfect prediction of all grammatical continuations, and one meaning that all activation is on ungrammatical units. To achieve this, the GPE is computed from *hits* (summed activation on correctly predicted, grammatical nodes), *false alarms* (summed activation on incorrectly predicted, ungrammatical nodes) plus *misses* (sum of differences of desired and actual activity on grammatical nodes, if positive, weighted by the amount of total output activation), as specified in (3).

$$(3) \quad GPE = 1 - \frac{hits}{hits + falsealarms + misses}$$

A GPE decreasing over several training epochs reflects the network's ongoing acquisition of implicit grammatical knowledge.

MC used the GPE to predict on-line processing load, with GPEs being directly proportional to reading times. Unfortunately, they restricted their analyses to global error (GPE) patterns. However, the GPE as a global measure can reflect two independent properties of the networks: *i.* how well the networks have learnt the grammar underlying the training corpora, and *ii.* on-line processing difficulty. MC clearly focused on the second aspect, implicitly presuming that grammar acquisition even after the earliest training epoch has reached a mature enough stage to be compared to adult participants in reading studies. However, until more fine-grained analyses have been carried out, the source of the errors remains obscure.

What's in an error?

False alarm activation can indicate a. the lack of adequate knowledge about word categories and the constructions they can appear in, or b. the interference induced by locally coherent continuations, ignoring the global context they appear in. We will show that strong but globally inconsistent local dependencies can distract from globally grammatical predictions, even in networks that have sufficiently learnt to classify words along syntactic categories.

We present detailed analyses of a. the output activation patterns in our replication of MacDonald and Christiansen's SRNs, and b. multi-dimensional scaling results of average hidden layer activations[1].

SRN simulation

The SRNs were built from thirty-one units each in the input and the output layer, and sixty units each in the hidden and the context layer. Like MC (2002), we trained ten SRNs with ten different corpora. The corpora were generated from a 30 word vocabulary plus the *end of sentence* marker (EOS) fed into a probabilistic context free grammar. Ten percent of the NPs were modified by relative clauses[2], regardless of their position in the sentence. Half of the RCs were SRCs (25% transitive and 25% intransitive) and the other half ORCs (transitive only). RCs were both center-embedded or right branching. One half of the verbs were in the present tense, the other half in the past tense. The present tensed verbs agreed in number (singular or plural) with their clausal subject, past tensed verbs fit with both singular and plural subjects.

[1] We did not have access to MCs networks and data except for the summarized output activities. We therefore had to replicate their results before we could start analyzing hidden layer activities.

[2] The probabilities differ slightly from those published in the article, because we rather used the numbers of the actual original grammar generator that M. Christiansen has provided to us. Our test revealed the same basic activation patterns with either set of values.

Each training corpus contained 10,000 sentences, resulting in an epoch of about 55,000 sweeps (words) on average. The learning rate was set to .1, and there was no momentum. Cross-entropy was used to calculate the error used by the backpropagation learning algorithm. The test sentences were not included in the training corpus.

Results

There are two positions of interest with high GPEs: the embedded verb in ORCs and the matrix verb in both ORCs and SRCs. The most interesting spot in ORCs is the embedded verb, where the largest portion of experience-based variance was obtained in MC's networks, motivating the *frequency x regularity* interpretation.

Embedded verb in ORCs

In ORCs, the embedded verb follows a "NP–that–NP" sequence. After the first training epoch, the element most active here, quite surprisingly, is the *end of sentence* (EOS, see figure 1). This prediction is clearly ungrammatical, because neither the matrix clause nor the RC received a verb yet. In the second epoch, the prediction of an EOS has been strongly reduced, while the correct predictions of verbs with the right number marking were increased. This trend continues until the third epoch, where there is virtually no activity left for EOS. As for the verbs, it should be easy to establish the agreement between the NP and the verb, since both are adjacent in ORCs, as they are in main clauses. Surprisingly, it takes three epochs to learn this dependency to an adequate extent.

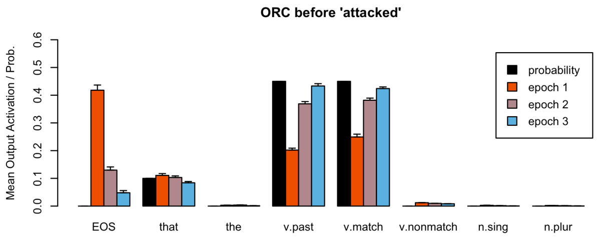

Figure 1: Mean output activations and grammatical probabilities at the embedded verb in ORCs, for three training epochs. Whiskers indicate standard errors.

Matrix verb

The second position at which a *sentence-type* x *experience* interaction was established in MCs simulations is the matrix verb. Moreover, GPEs on the matrix verb were high for both SRCs and ORCs. The results seem to fit King and Just's (1991) reading data in as much as reading times were also highest at this point in both SRCs and ORCs, with a slight advantage for SRCs. Nevertheless, while reading times at the matrix verb after ORCs showed the highest variability for readers of different span groups, the GPEs for ORCs in the network simulations varied not nearly as much at the matrix verb as on the embedded verb[3].

We examined the activation patterns at the matrix verb after both SRCs and ORCs, since both exhibit extremely high GPEs (between about .55 and .88).

SRCs. The detailed output vector analysis revealed that the GPE is based on one major false alarm component. In SRCs (figure 2), after a *verb-NP* sequence, the high GPE was based on false activation of the EOS, which did not change substantially over epochs.

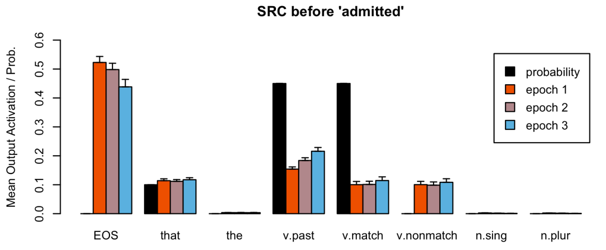

Figure 2: Mean output activations and grammatical probabilities at the matrix verb after SRCs, for three training epochs.

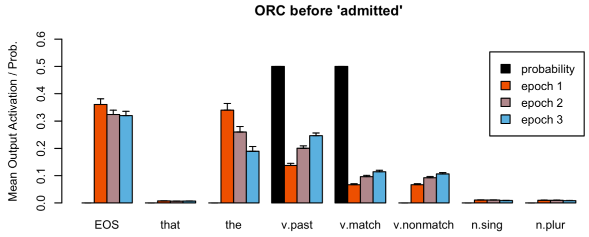

Figure 3: Mean output activations and grammatical probabilities at the matrix verb after ORCs, for three training epochs.

ORCs. After ORCs, following a *NP-verb* sequence, the only grammatical continuation is the matrix verb. Activation on all other words is a false alarm. Note that in the first epoch, the sum of false alarms is about 80%. The activation pattern reveals that the high GPE was due to one of the following two major false alarm components:

1. The false prediction of a determiner, indicating the prediction of another NP following the verb. This error dramatically decreased over the three epochs, but was still present even in the third epoch.
2. The false activation of EOS, which even grew slightly in the third epochs.

[3] However, the reading data on the matrix verb can be explained by a spill-over from the embedded verb, something that can quite regularly be observed in reading data. This dissimilarity between reading and simulation data should therefore not be taken too seriously.

Discussion

Embedded verbs

The activation patterns reveal that the high GPE at the embedded verb in ORCs during the first and the second epochs is mainly due to an ungrammatical prediction of an EOS. The remaining activation of the verbs shows that the networks have, at the same time, learned intra-RC number agreement, if not perfectly. How can this pattern of results be explained?

The EOS prediction is also high after SRCs following the sequence *NP-that-verb-NP*. Note that about half of the sentences end after the RC, namely when the RC modifies the Object-NP in transitive main clauses. The high false EOS prediction might thus be due to locally predicting the sentence ending, despite the context of a Subject-NP modifying center-embedded sentence.

Back to the embedded verb in ORCs. Here, *NP-that-NP*, and *...that-NP* is certainly not a good sentence ending. Two simple hypotheses can be ruled out fairly quickly. First, since about half of the sentences end with an NP, it might be just the NP that makes a good EOS in early training. Secondly, the prediction of the EOS might just reflect that with each additional word, the likelihood of an EOS increases.

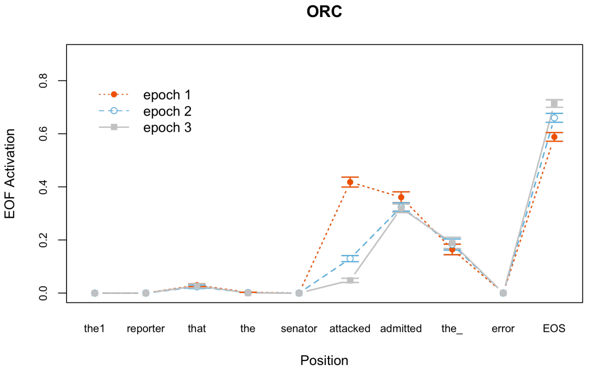

Figure 4: Output activaton of EOS at each position in sentences with ORCs, for three training epochs.

Figure 4 shows the activation of an EOS throughout the entire sentence. There is clearly little activation after the first NP, ruling out the first hypothesis. Moreover, there is a clear peak at the embedded verb, the matrix verb and the following determiner, whereas the subsequent noun shows almost zero EOS activation. An implicit counting mechanism that predicts increasing EOS activity with each step further downstream can hence be ruled out.

We want to pursue a third hypothesis: The network has not yet classified the relPro *that* correctly after the first epoch and confuses it with verbs. Note that the sequence *NP-that-NP* shares some distributional properties with regular transitive main clauses. The training corpora contained both simple main clauses and sentences with one or more RCs, most of which were center-embedded, i.e. modifying the first NP. All sentences started with an NP. The next word could either be a verb, or the relative pronoun *that*. Both were often followed by another NP, as *i.* transitive verbs in main clauses are followed by the direct object, and *ii.* the relative pronoun is followed by the subject-NP in ORCs. Due to this distributional resemblance, it seems reasonable that in early epochs, the networks are bad in distinguishing *NP-verb-NP* sequences from *NP-that-NP* sequences, or, to put it more simply, they confuse the relative pronoun with transitive verbs, at least in the local context of one NP to the left and one NP to the right. Hence, at the acquired level of grammatical knowledge, the EOS appears to be a feasible continuation for *NP-that-NP*, since it appears to mark the end of a simple transitive SVO main clause[4].

With more training, the networks slowly adapt to the fact that the relPro and verbs are not distributionally equivalent when a wider context is taken into account.

To substantiate this claim, we analyzed the hidden layer activities for all words in the corpus. There have been several proposals for analyzing distributed representations in neural networks, such as cluster analysis (Hinton, 1988), principle component & phrase state analysis (Elman, 1989), skeleton analysis (Mozer & Smolensky, 1989), contribution analysis (Sanger, 1989), which make the networks' representations and behavior more transparent. Since we are interested in how the SRNs have classified words, we analyzed hidden layer activities for each word averaged over test runs of one thousand random sentences. We present *multi-dimensional scaling* (MDS) data illustrating the internal grouping of words and indicating scaled euclidean distances between individual words (word groups). All stress values were below 0.1.

If the confusion of relPros and verbs is responsible for false EOS prediction, the hidden layer activations of relPros and verbs should be more similar in the first epoch than in later epochs.

Results

As figures 5 and 6 illustrate, euclidean distances between the relPro *that* and transitive verbs change considerably between epochs. The relPro is thus much more similar to verbs, especially transitive verbs, after the first training epoch than it is after the third, where

[4] There is even more distributional overlap between relpros and verbs: In SRCs, the relative pronoun *that* is immediately followed by a verb. However, even this local sequence is locally consistent with the verb classification of *that*, since in sentences with ORCs, the matrix verb immediately follows the embedded verb (it even follows a NP-verb sequence!). NP-that-verb-NP sequences are hence locally consistent with both the verb reading of *that*, since there is a NP-verb-verb-NP sequence contained in sentences with ORCs, and with the correct relative pronoun reading of *that*.

that builds an outlier categorie of its own. The hidden layer activities support the confusion hypothesis: After the first epoch, average activities of relPros resemble those of verbs much more than after the second and the third epoch.

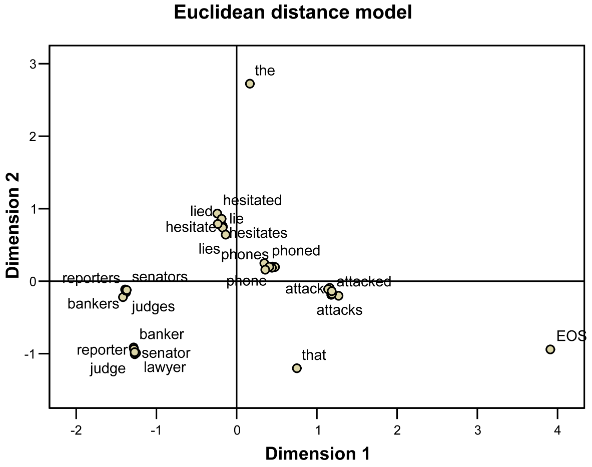

Figure 5: MDS plot of average hidden layer activations after epoch 1

In fact, relPros resemble transitive verbs more than intransitive verbs. These data clearly suggest that the biggest part of what is gained from training is the substantially better classification of the relPro. On the other hand it is also clear that relPros are not generally classified as verbs even in the first epoch. However, the hidden layer analyses reflect averaged hidden layer activities at the moment when the word is at the input, not after the entire NP-that-NP sequence.

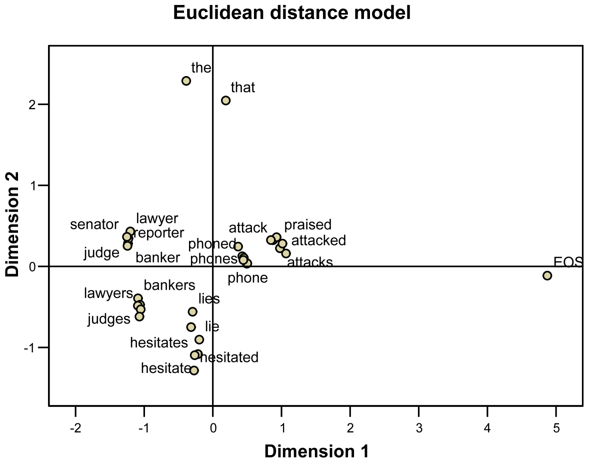

Figure 6. MDS plot of average hidden layer activations after epoch 3

Discussion of matrix verb results

The results on the matrix verbs strongly suggest that the GPE is mainly based on one or two false alarm components for SRCs and ORCs, respectively. In both sentences, the EOS is a major false alarm component.

In SRCs, the EOS-prediction follows a …-verb$_{\textit{transitive}}$-NP sequence. Expecting an EOS here is locally legitimized by the word order in transitive main clauses, which end here in the majority of the cases. The false EOS prediction appears to be stable, and would probably survive even more training epochs, even though the activation of correct verbs is continuously growing throughout the epochs. These data suggest that the main reason for long reading times on matrix verbs in center embedded sentences is that readers, even the most experienced ones, expect the sentence to end here about as much as they expect a correct matrix verb. In the absence of further empirical data, we resort to questioning this empirical prediction on the grounds of plausibility. We are convinced that adult readers, even less experienced ones, would be quite surprised if the sentence ended after a simple center-embedded RC.

In ORCs, both false alarm predictions of the determiner and the EOS prediction follow a …-NP-verb$_{\textit{transitive}}$ sequence. In this local context, the prediction of the determiner is legitimized by the word order in simple transitive main clauses, where verbs are followed by an NP. As in SRCs, this prediction indicates that, to a substantial degree, the networks ignore the fact that the RC is sub-ordinate. Contrary to the stable EOS prediction in SRCs however, the determiner prediction shrinks over time, indicating that the networks learned to widen their contextual window. The decreasing amount of false alarm activation is responsible for the global GPE reduction at the matrix verb. Although it appears odd that adult readers would run into this local trap, this result is modestly consistent with MCs *frequency* x *regularity* interpretation.

The false prediction of an EOS at this position seems a bit puzzling at first glance. The embedded verbs used here are transitive, as they have to combine with an object-NP in the test sentences. Even if the networks pursue a main clause analysis, they should predict a NP, but rule out an EOS. However, half of the transitive verbs used (*phones, phone, phoned, understands, understand, understood*) were also used as intransitives. It seems likely that the averaged GPEs are based on false predictions due to these verbs. A more detailed analysis, distinguishing strictly transitive and optional transitive verbs could clarify this issue. Also note that the false prediction of an EOS increases with experience. So the most experienced networks, and hence high span readers, are predicted to not really be surprised if the sentence ends after a center-embedded ORC. Once again, we are skeptical about this hypothesis.

In all cases, locally coherent continuations have distracted the network from the global necessity of a matrix verb at this position. More generally speaking, locally consistent false alarms were identified as the main source of processing difficulty.

Conclusion

We have argued that when predictions derived from connectionist models are presented, global error measures must be accompanied with detailed analyses of the output activation vectors to understand the source of the errors in the networks. A detailed analysis of false alarm components can hint at substantial acquisition deficits at the current stage of learning and at simulation artifacts caused by the choice of the grammar that the training corpora are generated from. In the present case, MC's networks were shown to make unrealistic continuation predictions based on classification errors (the relative pronoun *that* is considered a verb). However, identifying a flaw in a particular simulation hardly renders a general hypothesis invalid. Experience is a likely source of both construction specific complexity and inter-individual variation, and empirical support is beginning to materialize. For instance, Wells, Christiansen, Race, Acheson, and MacDonald (2009) showed that processing of relative clauses, and especially of ORCs, can be improved by training with RCs.

The activation analyses also revealed that the main source of complexity is the distraction induced by *locally coherent* continuations. Are adult language processors distracted by such false alarm predictions? Again, empirical support is beginning to surface. Tabor, Galantucci and Richardson (2004) provided data indicating that locally coherent but globally incoherent fragments can distract attention from the globally valid analysis in ambiguities. Konieczny (2005) revealed that syntactic errors produced by adding locally coherent words to a sentence were harder to detect than errors induced by locally incoherent words. Konieczny, Müller, Hachmann, Schwarzkopf and Wolfer (2009) showed in visual-world eyetracking experiments that local coherences are being interpreted during speech processing. Despite their misleading results, MC's approach helped identifying a fundamental processing phenomenon: interference by local coherences. Empirical data showing local coherence effects in real language processers provides support for the connectionist framework as a whole.

References

Christiansen, M. H. & Chater N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science, 23,* 157-205.

Daneman, M. & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Behavior*, *19*, 450-466.

Elman, J.L. (1989). *Representation and structure in connectionist models.* Technical Report CRL-8903. Center for Research in Language, University of California, San Diego.

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, *14*, 179-211.

Elman, J. L. (1991).Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*, 195-224.

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

Gordon, P. C.; Hendrick, R.; Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *27(6)*, 1411-1423.

Hale, J. (2001) *A Probabilistic Earley Parser as a Psycholinguistic Model.* Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics. 159-166.

Hinton, G.E. (1988). *Representing part-whole hierarchies in connectionist networks.* Technical Report CRG-TR-88-2, Connectionist Research Group, University of Toronto.

Just, M. A. & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122-149.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory & Language, 30,* 580-602.

Levy, R (2005). *Probabilistic models of word order and syntactic discontinuity.* Stanford University.

Konieczny, L. (2005). *The psychological reality of local coherences in sentence processing.* Proceedings of the 27th Annual Conference of the Cognitive Science Society. Stresa, Italy, August 2005.

Konieczny, L., Müller, D., Hachmann, W., Schwarzkopf, S., & Wolfer, S. (2009). *Local syntactic coherence interpretation. Evidence from a visual world study.* Paper presented at the 31st Annual Conference of the Cognitive Science Society.

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

MacDonald, M. C. & Christiansen, M. H. (2002). Reassessing working memory: A comment on Just & Carpenter (1992) and Waters & Caplan (1996). *Psychological Review*, *109*, 35-54.

Mozer, M. C. and Smolensky, P. (1989). Skeletonization: A technique for trimming the fat from a network via relevance assessment. In: Touretzky, D. S. (ed.), *Advances in Neural Information Processing Systems 1*, 107-115. San Mateo, CA.

Sanger, D. (1989*). Contribution analysis: a technique for assigning responsibilities to hidden units in connectionist networks.* Boulder, Colorado.

Seidenberg, M. & McClelland, C. (1989). A distributed model of word recognition and naming. *Psychological Review, 96,* 523-568.

Tabor, W., Galantucci, B., Richardson, D. (2004). Effects of Merely Local Syntactic Coherence on Sentence Processing. *Journal of Memory and Language*, *50(4)*, 355-370.

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology 58*, 250-271