Processing grammatical and ungrammatical center embeddings in English and German: A computational model

Felix Engelmann (felix.engelmann@uni-potsdam.de)
Department of Linguistics, Karl-Liebknecht Str. 24-25,
14476 Potsdam, Germany

Shravan Vasishth (vasishth@uni-potsdam.de)
Department of Linguistics, Karl-Liebknecht Str. 24-25,
14476 Potsdam, Germany

Abstract

Previous work has shown that in English ungrammatical center embeddings are more acceptable and easier to process than their grammatical counterparts (Frazier, 1985; Gibson & Thomas, 1999). A well-known explanation for this preference for ungrammatical structures is based on working-memory overload: the claim is that the prediction for an upcoming verb phrase is forgotten due to memory overload, leading to an illusion of grammaticality (Gibson & Thomas, 1999). However, this memory-overload account cannot explain the recent finding by Vasishth, Suckow, Lewis, and Kern (2008) that in German no illusion of ungrammaticality occurs. We present a simple recurrent network model that can explain both the presence of the grammaticality illusion in English and its absence in German. We argue that the grammaticality illusion emerges as a function of experience with language-specific structures, not working memory constraints as argued for in Gibson and Thomas (1999).

Keywords: sentence comprehension ; center embeddings ; illusion of grammaticality ; working-memory models ; connectionist models

Introduction

Consider the contrast in (1), discussed first by Frazier (1985) (the original observation is attributed by Frazier to Janet Fodor). Although the rules of English grammar allow a sentence like (1a), such a complex structure is perceived by native English speakers to be less acceptable than its ungrammatical counterpart (1b), in which the middle verb phrase, *was cleaning every week*, is missing.

(1) a. The apartment that the maid who the service had sent over was cleaning every week was well decorated.

b. *The apartment that the maid who the service had sent over was well decorated.

The first published study involving this contrast was an offline questionnaire-based experiment by Gibson and Thomas (1999). Their main finding was that ungrammatical sentences such as (1b) were rated no worse than grammatical ones such as (1a). In related work, Christiansen and Macdonald (2009) show that ungrammatical sentences were rated significantly better than the grammatical ones. We will refer to this surprising finding as the *grammaticality illusion*.

At least two competing explanations exist for this illusion. One is due to Gibson and Thomas (1999), who argue that the prediction for the middle verb phrase is forgotten if memory cost exceeds a certain threshold; this explanation relies on the assumption that working memory overload leads to forgetting. The second explanation is due to Christiansen and Chater (1999) and Christiansen and Macdonald (2009), who attribute the illusion to experience (exposure to particular regularities in the syntax of a language) as encoded in a connectionist network. They trained a simple recurrent network (SRN) on right-branching and center-embedding structures and then assessed the output node activations after seeing the ungrammatical sequence NNNVV (i.e., sentences like 1b). The activations showed a clear preference for ungrammatical structures, consistent with empirical data from English speakers.

An important theoretical question is whether these two explanations—the memory-overload account and the experience-based account—can be distinguished. Although the English data is consistent with both explanations, recent work by Vasishth et al. (2008) provides revealing new evidence regarding the grammaticality illusion. Vasishth and colleagues carried out several self-paced reading and eyetracking studies demonstrating that although the English grammaticality illusion can be replicated in online measures like reading time, in German the pattern reverses: readers find the *ungrammatical* sentence (1b) harder to process than its grammatical counterpart (1a). In other words, German readers do not experience the grammaticality illusion.

Specifically, for English Vasishth and colleagues found (across several experiments) longer reading times in the grammatical condition (1a) either at the final verb or the word immediately following it (or in both regions); whereas for German they reported shorter re-reading times in the *grammatical* condition either in the final verb region and/or the region following it.

The absence of the grammaticality illusion in German is interesting because it cannot be explained by the memory-based forgetting account as stated in (Gibson & Thomas, 1999). The explanation due to Christiansen and Chater (1999), however, may be able to explain the German results (in addition to the patterns seen in English): since German relative clauses are always head-final, German readers are exposed to head-final center embeddings much more often than English speakers. This greater exposure to head-final structures could be the reason why German speakers are able to identify the

missing verb but the English speakers are unable to do so.

In this paper, we extend the connectionist model of Christiansen and Chater (1999) to generate predictions for both the English and German structures, and demonstrate that this experience-based account provides a better explanation for the English and German data than an account based on language-independent working-memory constraints.

The Model

Network Architecture, grammar and corpora

We used a simple recurrent network (Elman, 1990) for modeling the effect of experience on forgetting. SRNs have been used previously to model the effect of structural properties in the language on comprehension performance (Christiansen & Chater, 1999; MacDonald & Christiansen, 2002). Since the predictions of an SRN are sensitive to probabilistic constraints in the input structure, they serve well to assess the effect of language-specific properties on learning. Furthermore, the architectural limitations of an SRN and its gradient nature give rise to human-like processing properties that have been explained in terms of working memory capacity limitations and decay in symbolic models. Our claim is that the grammaticality illusion is dependent on experience with word order regularities of the language in question. In order to show this we used a simple artificial language resembling simple sentences and subject- and object-extracted relative clauses. We also held the number of subject- and object-relatives equal in the corpus. In doing so we made sure that the only varying factor between the two training languages was whether its relative clauses are head-final or not.

The Corpora were generated from probabilistic context-free grammars (PCFGs) originally designed by Lars Konieczny (English) and Daniel Müller and Lars Konieczny (German).[1] For generating corpora and likelihood predictions the Simple Language Generator (Rohde, 1999) was used. Every training corpus consisted of 10,000 randomly generated sentences. Test corpora were generated for every condition consisting of 10 test sentences each. The networks described below were built, trained, and tested in the Tlearn simulator (Elman, 1992) on a Windows platform.

Training and Testing Procedure

Prior to training, all networks were initialized with random connection weights in the range of [-0.15, 0.15] and the hidden units received an initial bias activation of 0.5. Each training included 10 individually initialized networks that were trained on 10 different corpora, respectively. The networks were trained for three epochs, where one epoch corresponded to a full run through a corpus.

The SRNs were trained on a word-by-word continuation prediction. Each input word produced an activation distribution over the output nodes which represented lexical entries. In combination with a cross-entropy error calculation (all output activations sum to 1) the activation distribution was comparable to a probability distribution over words.

The SRN's prediction were assessed using grammatical prediction error (Christiansen & Chater, 1999). The GPE algorithm is based on the numerical differences between the PCFG probabilities and the actual output. The GPE value is a difficulty measure for every word in the sentence, which can be used as a reading time predictor (MacDonald & Christiansen, 2002).

Modeling the grammaticality illusion

The SRN trained on English sentences had 31 input and output units and 60 hidden units. Each input and output unit stood for one lexical entry in the lexicon. The lexicon consisted of five nouns, four intransitive and four transitive verbs in singular, plural and past tense forms and one end-of-sentence marker (EOS). At every NP the probability of an RC embedding was 0.1.[2] An RC could be realized as a subject relative (SRC) or an object relative clause (ORC) with equal probability.[3] Probabilities for transitivity and number status were also equal. The longest sentence in the corpus for English had 18 words. The German lexicon contained 21 words, including four verbs and nouns in singular and plural forms, the respective determiners in nominative and accusative, the comma and the EOS marker. In consequence the SRN trained on German had only 21 input and output units. The longest corpus sentence had 41 words, including the obligatory commas in German relative clauses. Both the English and German grammars included a number agreement between subjects and their predicates. In German a number and case agreement between determiner and noun was also included.

Christiansen and Chater (1999) reported node activations for the region after an NNNVV sequence. For better comparison with empirical data we extended their study to obtain GPE values for both conditions on all regions after the missing verb. Consider for example the error values on seeing V1 after the sequence 'N1 N2 N3 V3', which is ungrammatical because V2 is missing. In case the network is not aware of the ungrammaticality, this should be reflected by similar GPE values for both the grammatical and the ungrammatical condition at V1. In order to model that we set the target probability at V1 to the same value as in the grammatical condition. (Meaning the probability distribution is conditioned by the assumption that V2 has actually been seen.) In consequence, an expectation of a V2 at this point would increase the GPE. So, in the ungrammatical condition an SRN with a more accurate grammar representation would produce a higher pre-

[1] Both grammars can be found at http://cognition.iig.uni-freiburg.de/teaching/veranstaltungen/ws03/projekt.htm.

[2] These are the probabilities used by Konieczny in his grammar; MacDonald and Christiansen have used 0.05. The precise number is arbitrary; the essential point is that relative clauses should be less frequent than simple sentences.

[3] We did not encode the well-known difference in probability of occurrence between SRCs and ORCs because we were not modeling this difference; this assumption does not affect the results presented here.

diction error than an SRN wrongly predicting V1 instead of V2.

For the English case, the GPE values would be lower in the ungrammatical condition. This effectively means that the SRN is unable to make correct predictions based on long-distance dependencies, but bases its predictions on rather locally consistent sequences. For example after seeing V3 the network only predicts one more verb because the observation of N1 is too weakly encoded in the hidden representations to influence the predictions. Consequently, on V1 the error for the ungrammatical condition should be lower because in the grammatical condition V1 is the third verb which is inconsistent with the SRN's predictions. The preference for the ungrammatical structure should continue on the post-V1 regions because a locally coherent context with two verbs is easier to handle than a context of three verbs.

We first tested whether the SRN makes the same predictions as previous work on the English grammatical and ungrammatical structures (Christiansen & Macdonald, 2009).

Simulation 1: English

The SRN, which was trained on the English corpus, was tested on the grammatical and the ungrammatical condition after one, two, and three epochs.

The grammar we used was more complex than Christiansen and Chater's, but structurally compatible. Therefore we expected that we would replicate their findings for English. In particular, the GPE values for the V1 and post-V1 regions should receive lower values in the ungrammatical condition (see corpus example 2b).

(2) a. The judge that the reporters that the senators understand praise attacked the lawyers .

b. *The judge that the reporters that the senators understand attacked the lawyers .

Results for simulation 1 In order to compare the results for the English self-paced reading and eyetracking experiments in Vasishth et al. (2008) the assessed regions in the simulation were the three verbs V3, V2, V1 and the post-V1 region. The V2 region contains no datapoint in the ungrammatical condition because the verb is dropped in the testing stimuli.

Figure 1 shows GPE values for the SRNs trained and tested on the English grammar after one, two and three epochs of training. The pattern corresponded to the empirical results; the SRNs predicted an advantage for ungrammatical structures at V1 and post-V1. No effect was predicted on V3 because no difference in stimuli and probability between the conditions is present at this point.

Simulation 2: German

We turn next to the simulations for German center embeddings. German relative clauses differ from English in at least two respects (a third difference is the morphology of the relative pronoun; but we do not discuss this difference here due to space constraints). First, German relative clauses are obligatorily head final; second, commas are obligatory in German

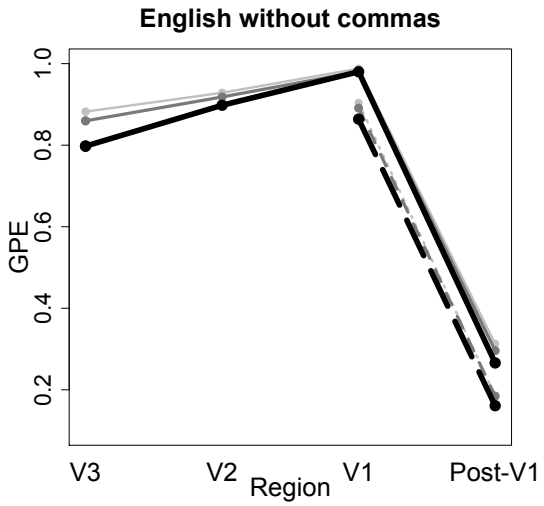

Figure 1: Simulation 1. English double-embedded object relative clauses. The figure shows the GPE values (for three epochs) for the three verbs and the subsequent region of the grammatical and ungrammatical conditions. The dotted line shows the ungrammatical condition. Epochs 3, 2, and 1 are colored black, dark grey and light grey, respectively.

relative clauses (see 3 for an example). We return to the role of commas later in the paper.

(3) a. Der Polizist , den der Mensch , den der Passant verspottet , ruft , trifft den Jungen .

b. *Der Polizist , den der Mensch , den der Passant verspottet , trifft den Jungen .

Results of simulation 2 Figure 2 summarizes the findings. First, in the regions V2 and V1, the GPEs were lower compared to the English sentences. Second, in contrast to the English case, the comparison by conditions did not reveal any difference on the main verb (V1). Finally, a small but significant preference for the grammatical structure was found on the post-V1 region ($p < 0.001$).

Discussion

The English and German center-embedding simulations suggest that experience with head-final structures may furnish a better explanation for the grammaticality illusion in English (and its absence in German) than working-memory based accounts such as Gibson and Thomas'. Both the English and German reading patterns found in the literature can be modeled by the SRN, whereas the working-memory based explanation can only explain the English results.

Our results do not imply that working memory plays no role in these constructions; rather, our claim is that experience plays a dominant role. A plausible way to reconcile the two accounts into one composite theory would have experience modulating working-memory overload. These details are or-

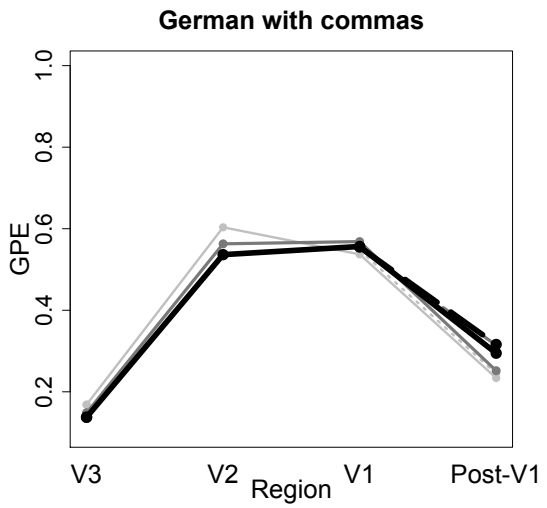

Figure 2: Simulation 2. German double-embedded object relative clauses. The figure shows the GPE values (for three epochs) for the three verbs and the subsequent region of the grammatical and ungrammatical conditions.

thogonal to our main finding, which is that experience determines whether English and German readers can correctly maintain predictions for upcoming verbs.

The role of commas in processing English center embeddings

One objection to this experience-based explanation for the grammaticality illusion (and its absence) is that the difference between English and German center embeddings could be related to the obligatory presence of commas in German. The commas in German relative clauses could lead to a strategy that is not available in the English structures previously studied. For example, readers could simply be counting the number of commas in German, and this could make it easier for them to detect ungrammaticality.

If commas alone (and not the head-final nature of relative clauses) are responsible for the patterns observed in German, then two straightforward predictions are that: (a) adding commas to English relative clauses should result in a German-like pattern for English sentences; and (b) removing commas from German relative clauses should result in an English-like pattern for German sentences.

Prediction (a) can be evaluated empirically but prediction (b) cannot because, as mentioned earlier, commas are obligatory in German relative clauses. As it turns out, Vasishth et al. (2008) tested the prediction for English and found that the presence of commas in English does not change the pattern; the grammaticality illusion persists.

The question we address next is: What does the SRN model predict for English RCs when commas are present?

Simulation 3: English with commas

For the simulation we extended the English grammar with appropriate comma insertions and trained the SRNs on the resulting corpora. In English non-restrictive object relative clauses (ORCs), commas would appear after nouns in the beginning of the sentence and after the verbs in the end. In a double-embedded ORC there would be a comma after V3 and V2. Thus, the grammatical/ungrammatical sequence pair is N,N,NV,V,V vs. N,N,NV,V. See (4) for examples.

For the SRN the comma effectively appears as a word category with only one token which attaches to nouns or verbs and is not involved in long-distance dependencies. Hence, the activation pattern representing it should not be too complex. In fact the learning of comma usage in ORCs can be reduced to a counting recursion problem of the pattern *aabb* instead of *abba*. As discussed in (Christiansen & Chater, 1999), counting recursion is the easiest of the three recursion types for both humans and connectionist networks. Thus, it is very likely that the inclusion of commas facilitates processing in the grammatical condition, lowering the respective GPE values.

(4) a. The lawyer , who the senator , who the judges attack , understands , praises the reporters .

b. *The lawyer , who the senator , who the judges attack , praises the reporters .

Results for simulation 3 See Figure 3 for the results after one, two and three epochs. Compared to simulation 1, there was a global improvement for both conditions, i.e., the GPEs were lower in each region. On V1 training had more effect in the ungrammatical than in the grammatical condition, resulting in a preference for the ungrammatical structure on V1 (as in simulation 1). On post-V1 training affected the grammatical condition more, however, not resulting in a grammaticality preference.

In summary, the SRN model suggests that although the insertion of commas in English helps to make better predictions overall, training effects seem to be driven by rather local consistency (Tabor, Galantucci, & Richardson, 2004), (Konieczny & Mueller, 2007), affecting the ungrammatical condition more than the grammatical one.

Importantly, the grammaticality illusion persists for English even when commas are present. This is consistent with the empirical findings for non-restrictive English relative clauses: Vasishth et al. (2008) also found in a self-paced reading study that the comma cue did not affect the grammaticality illusion in English.

The above findings raise an interesting question for German: is the reversal of the grammaticality illusion in German due only to the head-final nature of relative clauses, or do commas also play a role in determining the outcome? The only way to empirically disentangle the effect of head-finality and commas in German would be to examine a language such as Hindi, which also has head-final relative clauses but does not require commas.

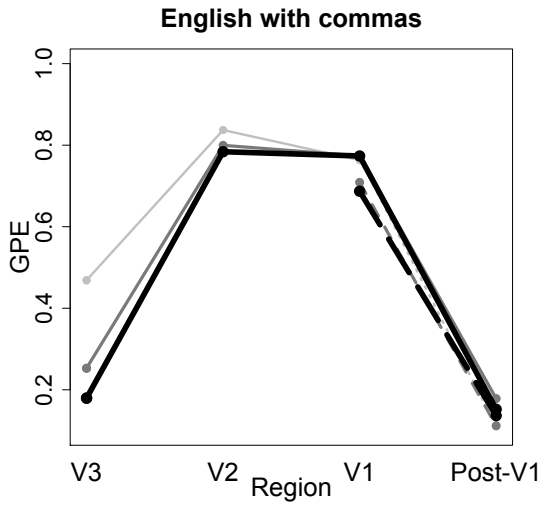

Figure 3: Simulation 3. The figure shows the GPEs (for the three epochs) of English center embeddings with commas.

Until such empirical evidence becomes available we cannot definitively answer the question about the role of commas, head-finality their interaction with experience. The SRN model can however generate predictions regarding the role of commas versus head-finality in German. We simulated the acquisition of experience with German head-final relative clauses which do not have any commas at all; in effect, we can simulate the learning of Hindi-type relative clauses in German. If commas are (partly) responsible for the reversal of the grammaticality illusion in German, then we should see an English-like pattern; if head-finality alone is the critical factor, then we should see a preference for grammatical structures even when commas are absent. This simulation is presented next.

Simulation 4: German without commas

In German, the presence of commas could have a facilitating effect because the counting-recursion pattern *aabb* is not only applicable in the ORC as in English but also in the SRC (both are head-final structures in German, unlike English). Consequently, the SRN trained on the German corpus should be very skilled on center-embedding recursion and comma counting-recursion and hence will have much lower error rates for the grammatical condition.

Thus, in German the removal of commas should make the SRN's predictions more error-prone. The verb-finality regularity in German, however, could still result in better predictions for the grammatical condition in German than in English. In order to test these predictions, simulation 4 tested SRNs trained on a comma-free German grammar.

Results of Simulation 4 The GPE values of the simulation involving German without commas (Figure 4) show a similar pattern as in English without commas. In the first epoch, an ungrammaticality preference was found in a small effect on V1 and a very pronounced effect on the region following it. After completion of training, V1 and post-V1 show a similar sized preference for the ungrammatical structure. Surprisingly, the regularity of verb-final structures does not seem to support correct predictions in German any more than in English. Rather, the more regular application of commas in German has a very facilitating effect on both conditions, slightly more on the grammatical.

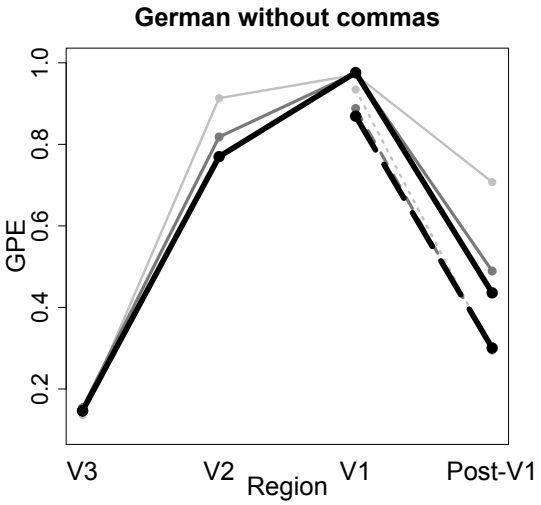

Figure 4: Simulation 4. The GPEs for German center embeddings without commas.

General Discussion

The results of simulation 1 (English without commas) and 2 (German with commas) were consistent with existing empirical data from both offline studies and online (self-paced reading and eyetracking) studies (Gibson & Thomas, 1999; Christiansen & MacDonald, 1999; Vasishth et al., 2008; Christiansen & Macdonald, 2009): the grammaticality illusion occurs in English but not in German.

These simulations demonstrate that the inherent architectural constraints of SRNs correctly predict both the grammaticality illusion in English double-embedded ORCs, as well as the absence of the illusion in German. In addition, the SRN model also makes the correct predictions regarding the effect of commas in English relative clauses: although commas reduce the GPEs, the grammaticality illusion persists in English. This is consistent with the evidence presented by Vasishth et al. (2008). Finally, we showed that in German head-finality alone does not explain the absence of the grammaticality illusion; commas appears to be crucial for the patterns observed.

Conclusion

This paper investigated the explanatory power of a particular implementation of the experience-based account for the

grammaticality illusion. The well-known SRN modeling approach of MacDonald and Christiansen (2002), Christiansen and Macdonald (2009) was adopted to test its predictions on the forgetting effect in complex center-embedding.

The grammaticality illusion was predicted for English but not for German, consistent with human data. However, further simulations revealed the comma insertion as an important factor for the German pattern.

A caveat is necessary here. An SRN trained on a simple grammar obviously does not learn exactly the same constraints as humans do. These simulations are rather approximations that are suggestive of the role that experience plays in modulating memory processes. An important issue with the SRNs' predictions is their dependency on local coherence. Interestingly, however, there is evidence that even human readers rely on local coherence in certain structures (Tabor et al., 2004). Another finding is that the simulations reported by Christiansen and Chater (1999), and also the comma issue in simulations presented here, showed that the SRN handles counting-recursion better than other types. That may be the reason for the strong facilitating effect of comma insertion compared to head-finality.

More broadly, this work argues in favor of a uniform account of language-specific differences that are grounded in experience and that emerge as a consequence of architectural constraints. This account is broadly consistent with a range of recent work that characterizes processing modulated by experience (Hale, 2001). At the same time, it is clear that working-memory centered accounts capture a great deal of the empirical base that purely experience-based accounts cannot explain. Some examples are: the presence of both similarity-based interference and similarity-based facilitation effects (Logačev & Vasishth, 2009), the interaction of interference with locality (Van Dyke & Lewis, 2003) and with antilocality (Vasishth & Lewis, 2006). Thus, it appears that a principled composition experience as well as working-memory constraints is necessary to explain the range of empirical phenomena in sentence processing.

Acknowledgements

We are very grateful to Lars Konieczny for permission to use the grammar developed in his lab.

References

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*(2), 157–205.

Christiansen, M. H., & Macdonald, M. (2009). *A usage-based approach to recursion in sentence processing.* (Submitted)

Christiansen, M. H., & MacDonald, M. C. (1999). *Processing of recursive sentence structure: Testing predictions from a connectionist model.* (Manuscript in preparation)

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Elman, J. L. (1992). *Tlearn simulator.* Software available at: http://crl.ucsd.edu/innate/tlearn.html.

Frazier, L. (1985). Syntactic complexity. In D. R. Dowty, L. Kartunnen, & A. M. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 129–189). Cambridge University Press.

Gibson, E., & Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, *14*(3), 225–248.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *North American Chapter Of The Association For Computational Linguistics*, 1–8.

Konieczny, L., & Mueller, D. (2007). Local coherence interpretation in written and spoken language. In *Proceedings of the 20th annual CUNY conference on human sentence processing.* La Jolla, CA.

Logačev, P., & Vasishth, S. (2009). Morphological ambiguity and working memory. In P. de Swart & M. Lamers (Eds.), *Case, word order, and prominence: Psycholinguistic and theoretical approaches to argument structure.* Springer.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, *109*(1), 35–54.

Rohde, D. L. T. (1999). The simple language generator: Encoding complex languages with simple grammars (Tech). *Mellon University, Department of Computer Science*, 99–123.

Tabor, W., Galantucci, B., & Richardson, D. (2004, May). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*(4), 355–370.

Van Dyke, J. A., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*(3), 285–316.

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, *82*(4), 767-794.

Vasishth, S., Suckow, K., Lewis, R., & Kern, S. (2008). *Short-term forgetting in sentence comprehension: Crosslinguistic evidence from head-final structures.* (Submitted to Language and Cognitive Processes)