Interaction with Context During Recurrent Neural Network Sentence Processing

Forrest Davis (fd252@cornell.edu)
Department of Linguistics, Cornell University

Marten van Schijndel (mv443@cornell.edu)
Department of Linguistics, Cornell University

Abstract

Syntactic ambiguities in isolated sentences can lead to increased difficulty in incremental sentence processing, a phenomenon known as a garden-path effect. This difficulty, however, can be alleviated for humans when they are presented with supporting discourse contexts. We tested whether recurrent neural network (RNN) language models (LMs) could learn linguistic representations that are similarly influenced by discourse context. RNN LMs have been claimed to learn a variety of syntactic constructions. However, recent work has suggested that pragmatically conditioned syntactic phenomena are not acquired by RNNs. In comparing model behavior to human behavior, we show that our models can, in fact, learn pragmatic constraints that alleviate garden-path effects given the correct training and testing conditions. This suggests that some aspects of linguistically relevant pragmatic knowledge can be learned from distributional information alone.

Keywords: garden path; neural networks; pragmatics; discourse

Introduction

Without context, syntactic ambiguities can lead to sentence processing difficulties, with garden-path phenomena being one of the most well studied cases. For example:

(1) The horse raced past the barn fell.

In reading (1) in isolation, readers experience confusion at the verb *fell* (known as a garden-path effect), having expected *raced* to be a main verb rather than part of a reduced relative clause (cf. *The horse that was raced past the barn fell*). Embedded in a larger linguistic context, however, this effect can be alleviated (e.g., Trueswell & Tanenhaus, 1991; Spivey-Knowlton, Trueswell, & Tanenhaus, 1993). This alleviation crucially relies on speakers' pragmatic and/or semantic knowledge, leading to questions about how insular syntactic representations are from non-syntactic information. The present study explores if such pragmatic knowledge is acquired, and utilized in a human-like way, by modern recurrent neural network (RNN) language models (LMs), which have been claimed to acquire knowledge of a range of syntactic phenomena.

There have been a number of theoretical accounts attempting to clarify at what level (and at what time in the course of incremental sentence processing) linguistic knowledge beyond syntax is utilized. The garden-path model (e.g., Frazier & Rayner, 1982; Ferreira & Clifton Jr, 1986) posits that syntactic structure is built without consideration of semantic or pragmatic plausibility. Semantics and pragmatics can influence the revision of this structure, but crucially syntax operates first. In contrast, constraint-based approaches (e.g., McClelland, St. John, & Taraban, 1989; Trueswell, Tanenhaus, & Garnsey, 1994) posit that semantic and pragmatic information biases the parser towards certain syntactic structures over others. Similarly, referential theory (e.g., Altmann & Steedman, 1988), posits that the syntactic structure chosen is one where the pragmatic presuppositions are best satisfied. In the absence of context, this amounts to the syntactic alternative that requires the least number of presuppositions in order for it to be interpreted felicitously.

There are theoretical accounts that focus less on details of syntactic structure building. The best known instances of these are information-theoretic surprisal (e.g., Hale, 2001; Levy, 2008) and the "good-enough" theory of parsing (e.g., Ferreira, Bailey, & Ferraro, 2002; Ferreira & Patson, 2007). Theories of surprisal suggest that garden-path effects follow from predictability, where less predictable words are processed more slowly. Thus, parsing *fell* in (1) leads to slowdown in processing because it is not predictable in the local context. The "good-enough" theory proposes that syntactic structures are only generated when necessary, so there is no need for the human parser to maintain competing syntactic structures. Under both of these accounts, pragmatic considerations can influence parser behavior.

RNN LMs have been claimed to acquire syntactic knowledge ranging from subject-verb agreement (e.g., Linzen, Dupoux, & Goldberg, 2016; Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2018), filler-gap dependencies (Wilcox, Levy, Morita, & Futrell, 2018) and center embedding (Wilcox, Levy, & Futrell, 2019). These studies have all tested single sentences focusing on cases of stark grammatical vs. ungrammatical distinctions. This is analogous to the task of single-sentence grammaticality judgments in humans. Pragmatic knowledge, on the other hand, is commonly assumed to rely heavily on reasoning about speaker intent and to utilize extra-linguistic 'world knowledge.' RNN LMs have no such knowledge, having no objective to infer intent for example. As such, they delineate an upper-bound on how far a model can get in acquiring linguistically relevant pragmatic reasoning from only linguistic data.

If nuanced knowledge of both syntax and pragmatics are needed in online human comprehension, we might ask whether these models acquire linguistic representations that are similarly mediated by both syntactic and pragmatic factors. Chaves (2020) has shown that previous studies claiming that RNN LMs acquire knowledge of syntactic islands (e.g., Wilcox et al., 2018) failed to account for experimentally validated fine-grained human grammaticality judgments. By relating model failure to pragmatic rather than syntactic conditions, Chaves claimed that RNN LMs are unable to acquire the full extent of linguistic knowledge necessary for reasoning about incremental sentence processing. The present study examines this claim with respect to garden-path phenomena.

RNN LMs have been shown to exhibit a human-like garden-path effect for isolated sentences (e.g., van Schijndel & Linzen, 2018; Futrell & Levy, 2019; Frank & Hoeks, 2019). In humans, this effect can be alleviated by discourse and pragmatic considerations. This paper addresses whether RNN LMs are truly unable to learn pragmatically conditioned syntactic representations by probing whether they exhibit human-like garden-path alleviation. In broad terms, we study what aspects of discourse structure, or pragmatic context, these models learn, and whether discourse (or pragmatic) representations influence incremental RNN representations. This has implications for our understanding of both how linguistic context influences human sentence processing and of the role that primary linguistic data plays in the acquisition of syntactic and pragmatic knowledge.

Stimuli

We focused on two types of garden-path constructions in this work. The first construction is the main verb vs. reduced-relative (MV/RR) ambiguity exemplified by (1). The second construction (NP/Z), exemplified below, is an ambiguity between a transitive verb *left* with a noun phrase (NP) complement *the party* (as in *the band left the party*) vs. an intransitive verb reading of *left* with zero complement (Z; as in *Even though the band left, the party went on*).

(2) Even though the band left the party went on for at least another two hours.

We manipulated pragmatic knowledge with controlled discourse contexts prior to the presentation of the target sentence. In particular, we embedded MV/RR sentences in two classes of discourse context: referentially supporting contexts and temporally supporting contexts. The stimuli we used (including discourse contexts) were taken from Spivey-Knowlton et al. (1993) for referential contexts and Trueswell and Tanenhaus (1991) for temporal contexts. This allowed us to compare our RNN LM results to those of their human participants. A total of 32 sets of stimuli (16 referential and 16 temporal) were used. All the reduced-relative verbs were ambiguous between both MV/RR readings (i.e. *killed* rather than *slain*). An example of a referential stimulus is given in (3).

(3) a. **Context**
  (i) 1NP - A knight and his squire were attacking a dragon. With its breath of fire, the dragon killed the knight but not the squire.
  (ii) 2NP - Two knights were attacking a dragon. With its breath of fire, the dragon killed one of the knights but not the other.
  b. **Target**
  (i) Reduced - The knight killed by the dragon fell to the ground with a thud.
  (ii) Unreduced - The knight who was killed by the dragon fell to the ground with a thud.

In (3), (3-b-i) is a garden-path sentence, with *killed* being ambiguous between a main verb and a reduced relative interpretation (*killed* in (3-b-ii) is unambiguously embedded in a relative clause). If the context (3-a-ii) is presented to humans before they read (3-b-i), they have a reduced garden-path effect; while the context (3-a-i) followed by (3-b-i) leads to the canonical garden-path effect.

The discourse status of the nominal *knight* is the key manipulation. With a main verb reading (leading to a garden-path effect), (3-b-i) presupposes that there exists a unique knight in the preceding context. This is satisfied when the sentence is preceded by (3-a-i). In contexts with only one knight, the relative clause reading is odd because *the knight* is just as informative as *the knight killed by a dragon*, so readers prefer the more concise but equally informative main verb reading (following Grice, 1975). In contexts with more than one knight (3-a-ii), the main verb reading violates the uniqueness presupposition arising from *the*, while the relative clause reading accommodates this presupposition and is informative (it uniquely identifies one of the knights), so there is a greater expectation that the definite nominal will appear modified. This leads to an alleviation of the garden-path effect.

A similar alleviation (though driven by a different discourse requirement) occurs in some temporal contexts as in:

(4) a. **Context**
  (i) Past - Several students were sitting together taking an exam in a large lecture hall earlier today. A proctor noticed one of the students cheating.
  (ii) Future - Several students will be sitting together taking an exam in a large lecture hall later today. A proctor will notice one of the students cheating.
  b. **Target**
  (i) Reduced - The student spotted by the proctor received/will receive a warning.
  (ii) Unreduced - The student who was spotted by the proctor received/will receive a warning.

In (4), we held fixed the number of referents (*several students*). If (4-a-ii) is presented to humans before they read (4-b-i), they do not garden-path; while (4-a-i) followed by (4-b-i) leads to the canonical garden-path effect. As detailed in Trueswell and Tanenhaus (1991), this difference is dependent on the temporal relationship between the discourse context and the target sentence. They hypothesized that the garden-path effect in past contexts is driven by it being less costly to continue the discourse with an additional past event (i.e. a student spotted something in the past) than it is to select the specific discourse referent using a relative clause interpretation (i.e. refer to the student who was spotted by the proctor). With future contexts (4-a-ii), however, the relative clause reading is preferred, because adding a new past event to the discourse with a main verb interpretation requires an additional processing step of adding a new time of reference (i.e. referring to an event before the current discourse).

Finally, we embedded NP/Z sentences in discourse contexts that differed in information status and definiteness. Our design followed Besserman and Kaiser (2016), though their stimuli were not included in their paper. Thus, we took the 20 NP/Z stimuli from Grodner, Gibson, Argaman, and Babyonyshev (2003) and manually created contexts for each.

(5) a. **Indefinite+New**
It was a fun evening. Even though the band left (,) **a party** went on for at least another two hours.

b. **Definite+New**
It was a fun evening. Even though the band left (,) **the party** went on for at least another two hours.

c. **Definite+Old**
A party was organized this evening. Even though the band left (,) **the party** went on for at least another two hours.

As in (3) and (4), each construction had an unambiguous version, in this case disambiguated with a comma inserted after the verb. There are two pragmatic manipulations: whether the potential NP complement is definite (*the party*) or indefinite (*a party*) and whether the potential NP complement is New (as in (5-a) where there is no prior mention of a party) or Old (as in (5-c) where there is a prior mention of party).[1] In brief, Besserman and Kaiser hypothesized that re-analysis from object to subject would be harder (i.e. the garden-path effect would be larger) when the potential NP complements were indefinite and new than for definite and new or definite and old. This follows from findings in corpora that objects tend to be new information and subjects to be given. In other words, (5-a) should be more difficult to process than (5-b), which in turn should be difficult than (5-c).

Modeling Methods

We trained ten RNN LMs with long short-term memory units (LSTMs; Hochreiter & Schmidhuber, 1997)[2] using PyTorch.[3] The models were trained on unannotated text using a language modeling objective of predicting each word given the preceding words (e.g., Elman, 1990). Training was done on an 80 million word subset of the Wikitext-103 corpus (Merity, Xiong, Bradbury, & Socher, 2016). To ensure that our results are robust, each RNN was trained with a different random initialization.

Table 1: Mean and standard deviation of LM validation perplexity for the models trained on ordered text and models trained on text shuffled by sentence.

Model Type	μ	σ
Ordered	27.82	0.12
Shuffled	32.56	0.13

To manipulate the pragmatic knowledge acquired by our RNN LMs, we trained five of the models on the training data after shuffling the data by sentence, which removed the discourse context while leaving syntactic structures intact, and which is actually a common approach in the computational literature (e.g., Gulordava et al., 2018; Jozefowicz, Vinyals, Schuster, Shazeer, & Wu, 2016). The five remaining models were trained on the original, unshuffled data (i.e. discourse contexts were preserved).

We evaluated the quality of the LMs used in this work by calculating model perplexity on the validation data given in the Wikitext-103 corpus. Due to memory constraints, we divided the validation data into individual Wikipedia articles, for a total of 60 articles.[4] Each article was passed to each model (shuffled and ordered) as a continuous chunk, so any cross-sentence dependencies within each article were available for the model to use. We report the standard by-word perplexity in Table 1. Training on ordered data lead to an average decrease in perplexity of 4.74 (i.e. models trained on ordered data performed better).

Measures

We used information-theoretic surprisal as our dependent measure (Hale, 2001; Levy, 2008).

$$S(w_i) = -\log_2 \mathrm{p}(w_i | w_1 ... w_{i-1}) \quad (1)$$

This measure tells us how probable a word is according to each model given the preceding context. Surprisal is used throughout the computational linguistics literature on language modeling and within the recent literature on modeling garden-path effects using RNN LMs. A larger surprisal value is correlated with greater reading times, and thus more surprisal indicates a larger garden-path effect.

For the MV/RR stimuli, we calculated the surprisal at their respective disambiguating regions. Given the human results, we expected that surprisal would be affected by preceding discourse context. To quantify the contextual effects, we took the difference in surprisal at the disambiguating region in the reduced target sentence (e.g., (3-b-i)) when preceded by the garden-path supporting context (e.g., one referent context (3-a-i)) and the same region in the reduced target sentence

[1] Indefinite+Old is regarded as infelicitous. It was not tested in Besserman and Kaiser (2016) or in the present study.

[2] The models had two LSTM layers with 400 hidden units each, 400-dimensional word embeddings, a dropout rate of 0.2 and batch size 128, and was trained for 40 epochs (with early stopping).

[3] The models and code for this paper can be found at `https://github.com/forrestdavis/GardenPath`

[4] Due to size of 3 of the 60 articles, we had to further split those for a total of 65 chunks.

when preceded by the garden-path alleviating context (e.g., two referents context (3-a-ii)). In other words, S(region|1NP) - S(region|2NP) and S(region|Past) - S(region|Future). This was done by stimulus, to control for stimulus-specific differences (e.g., lexical semantics).

Similarly, for the NP/Z stimuli, given the human results, we expected that definiteness and information status would have a significant effect on surprisal values at the disambiguating region. No significant interaction effect was reported in Besserman and Kaiser (2016), so we looked at the two main effects individually. For the effect of definiteness, we held fixed information status and took the difference in surprisal at the disambiguating region with an indefinite potential NP complement (as in (5-a)) from the same region with a definite potential NP complement (as in (5-b)). For the effect of information status, we held fixed definiteness, taking the difference in surprisal at the disambiguating region when the potential NP complement is discourse new (as in (5-b)) from when it is discourse old (as in (5-c)). In other words, we derived two measures: S(region|Indefinite+New) - S(region|Definite+New) and S(region|Definite+New) - S(region|Definite+Old).

Results

Context-Free Garden Path Effects

Spivey-Knowlton et al. (1993), Trueswell and Tanenhaus (1991), and Besserman and Kaiser (2016) reported greater reading times with reduced sentences (e.g., (3-b-i)) than with unreduced sentences (e.g., (3-b-ii)). For the MV/RR stimuli, we summed the surprisals over the entire relative clause (e.g., *killed by the dragon*, *spotted by the proctor*) which was read significantly slower in the human experiments as a reduced relative (i.e. when not preceded by *who was*) than an unreduced relative clause. For NP/Z, we expected the models to have greater surprisal at the disambiguating region (e.g., *went on* in (5)) in the reduced case compared to the unreduced case (as in van Schijndel & Linzen, 2018). This served as a sanity check and confirmation of the previous RNN LM literature on garden-path sentences.

As expected, we observed a decontextualized garden-path effect for both the MV/RR data (Figure 1a) and NP/Z data (Figure 1b). We conducted a two way ANOVA test in R, with model type (ordered vs. shuffled training data) and sentence type (reduced vs. unreduced) as main effects. The results of this confirmed the observed pattern, with sentence type highly significant and model type also significant.[5] The interaction, however, was not significant ($p = 0.28$). The mean difference in surprisal was greater for the models trained on shuffled data than for the models trained on ordered data, in line with the lower overall perplexity for models trained on ordered data. The lack of interaction suggests that this performance advantage did not affect the garden-path effect as a whole.

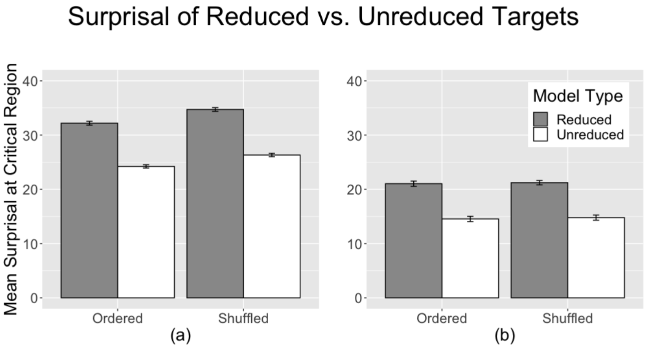

Figure 1: Mean RNN LM surprisals for (a) reduced versus unreduced MV/RR target stimuli with surprisal summed over the relative clause region, and (b) NP/Z target stimuli with surprisal summed over the disambiguating region. Greater difference between reduced and unreduced correspond to a greater garden-path effect. Error bars are 95% confidence intervals. Stimuli are from Spivey-Knowlton et al. (1993), Trueswell and Tanenhaus (1991), and Grodner et al. (2003).

Referential Contexts

We turn now to context mediated effects, beginning with referentially supporting contexts (exemplified with (3)). We predicted, based on the human results from Spivey-Knowlton et al. (1993), that preceding contexts with two referents of the same type as the subject of the target sentence (e.g., context with two knights followed by the target *The knight killed by the dragon ...*) would have less of a garden-path effect than those with one referent ((3-a-ii) followed by (3-b-i) vs. (3-a-i) followed by (3-b-i)).

Specifically, we calculated the surprisal at *by* which partially disambiguates the reduced-relative reading. The fully disambiguating main verb (e.g., *fell*) did not exhibit an effect of discourse context in the human experiments, so we did not look at that region in the present study. As detailed above, we measured S(by|1NP) - S(by|2NP). The distributions, broken into model type as before, are given in Figure 2a.

Contrary to our prediction, contexts with two referents did not significantly reduce the surprisal at *by* when compared to contexts with only one referent regardless of training condition (i.e. ordered vs. shuffled). However, if we calculated surprisal over the verb+*by* region, which showed a significant context effect in Spivey-Knowlton et al. (1993), we did see an effect of context, with two referent contexts decreasing the surprisal over this region. This effect was only marginally significant ($p = 0.004$) after Bonferroni correction for models trained on ordered data, and there was no significant effect ($p = 0.51$) for models trained on shuffled data. Additionally, there was a marginally significant ($p = 0.004$) difference between model types. These results suggest that whatever referenetial alleviation might have been learned from text was only learned by models trained on ordered data. The presence of

[5] We corrected for multiple comparisons using family-wise Bonferroni correction. All effects we report as significant had $p < 0.00001$.

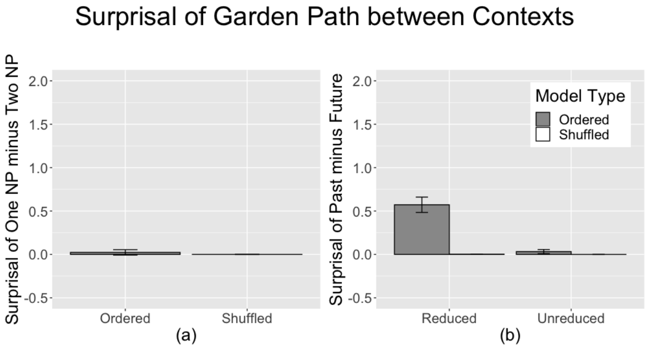

Figure 2: Differences between RNN LM surprisals at the critical region (*by*) when preceded by a garden-path supporting context ((a) contexts with a single referent and (b) past contexts) and when preceded by a garden-path alleviating context ((a) contexts with two referents and (b) future contexts). For (b), differences between critical region in reduced relative clause and unreduced relative clause are given. Error bars are 95% confidence intervals. Positive values correspond to a garden-path alleviation effect. Stimuli are from Spivey-Knowlton et al. (1993) and Trueswell and Tanenhaus (1991).

an RNN effect for the entire verb+*by* region and not for just *by* suggests that the alleviation is concentrated on the reduced-relative verb rather than distributed over the verb+*by* region analyzed in previous human experiments.[6]

Temporal Contexts

We turn now to temporally supporting contexts (exemplified with (4)), and their relationship to garden-path alleviation in RNN LMs. We predicted, based on the human results from Trueswell and Tanenhaus (1991), that preceding past contexts would result in less of a garden-path effect than future contexts ((4-a-ii) followed by (4-b-i) vs. (4-a-i) followed by (4-b-i)). As with the referential contexts, we calculated the surprisal values at *by*.[7] As detailed above we conditioned this on context, so we measured S(by|Past) - S(by|Future)).

Trueswell and Tanenhaus (1991) reported that a difference in reading times in the future context is observed for the reduced relative clauses and not the unreduced relative clauses. We made the prediction that only the reduced target sentences would exhibit a context effect, with future contexts reducing the surprisal. The surprisal values for *by* should be similar if the target sentence has an unreduced relative clause. The surprisal distributions, broken into model type as before, are given in Figure 2b.

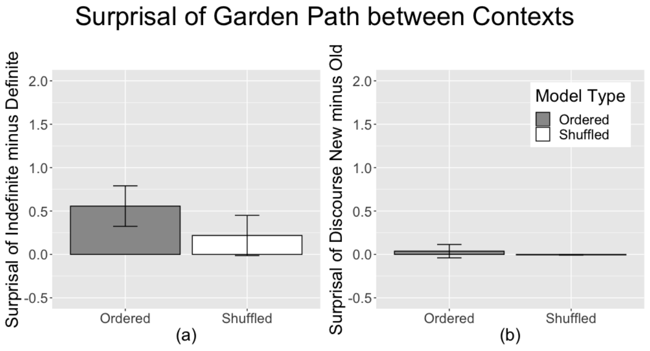

Figure 3: Differences between RNN LM surprisals at the disambiguating region when potential NP complement in NP/Z garden-path stimuli is indefinte and definite in (a) or discourse new and old in (b). Error bars are 95% confidence intervals. Positive values correspond to a garden-path alleviation effect. Stimuli are from Grodner et al. (2003).

As predicted, future contexts significantly reduced the surprisal at *by*, when compared to past contexts. The surprisal values differed significantly between model types. We saw an effect only for the models trained on ordered data, while those trained on shuffled data did not significantly differ from zero. Moreover, as predicted, only the reduced target sentences exhibited the context based alleviation. In unreduced relative clauses, the difference in surprisal between past contexts and future contexts did not differ significantly from zero given Bonferroni correction ($p = 0.03$).

These results again suggest that training on ordered data is crucial for exhibiting temporal alleviation of the garden-path effect. The lack of a context effect for unreduced relative clauses suggests that the observed behavior is driven by alleviating the ambiguity of the relative clause, rather than just an increased likelihood for the relative clause given a Future context.

Information Status and Definiteness Effects

Finally we turn to effects of information status (given vs. new) and definiteness on NP/Z garden-paths (exemplified with (5)). We predicted, based on the human results from Besserman and Kaiser (2016), that discourse old potential NP complements would lead to reduced garden-path effects (as in (5-c)) and definite potential NP complements would lead to reduced garden-path effects (as in (5-b) and (5-c)). As detailed above, we took two measurements: S(region|Indefinite+New) - S(region|Definite+New) and S(region|Definite+New) - S(region|Definite+Old). Surprisal distributions for each condition are given in Figures 3a and 3b respectively.

For the models trained on ordered data, definite NPs did significantly reduce surprisal, while discourse old NPs did not

[6] In Spivey-Knowlton et al. (1993), they ran an additional experiment with single word presentations and found an effect of context on the reduced-relative verb as well. They, however, included another manipulation, whether the reduced-relative verb was possible as a main-verb (e.g., *killed*) or not (e.g., *slain*). We did not include this manipulation in the present study, so cannot directly compare our results to the human ones.

[7] Trueswell and Tanenhaus (1991) reported only a by-subject effect and not a by-item effect for context type for the fully disambiguating main verb (e.g., received) so we did not look at this region.

reach significance in reduction of surprisal. Models trained on shuffled data showed no alleviation effects. This again suggests that models trained on ordered data were able to exhibit human-like garden-path alleviation effects, at least along the dimension of definiteness.

Discussion

Recent work has suggested that RNN LMs are unable to acquire fine-grained pragmatic representations (Chaves, 2020). The present study points to two crucial components missing from these previous experiments: explicit discourse context prior to the target sentence and models trained on ordered data. Models trained on both ordered and shuffled data in this study exhibited the canonical garden-path effect, but only ordered data led to the acquisition of pragmatically conditioned representations. These findings highlight the significance of different training conditions in comparing model performance to human experimental findings. Moreover, language modeling, like human experimentation, that focuses only on isolated sentences may miss factors, like presupposition failure, that better account for processing mechanisms. In attempting to compare model behavior and human behavior, crucial explanatory factors are compounded by the fact that humans bring a wealth of language experience and expectations to an experiment that are difficult to quantify in comparison to a model's initial state. Providing explicit discourse contexts in the experimental manipulation for both models and humans is a key component in evaluating and differentiating model and human linguistic representations.

Turning to the specific results in this study, we saw that temporal contexts had the largest effects on garden-path alleviation, definiteness had a lesser effect, referential contexts were mixed in their effect, and information status had none. We might, then, extract the generalization that tense is more robustly learned (in the sense that it can influence model representations) than grammatical features constrained to nominals (as in uniqueness, definiteness, and relative clause modification). Additionally, we failed to replicate the information status results in Besserman and Kaiser (2016). Perhaps the effect is weaker than their work suggests, or perhaps the pragmatic knowledge needed to exhibit this effect is not learnable from linguistic data alone. It is worth noting that they themselves report only a numerical effect for the hierarchy of Definite+Old over Indefinite+New and Definite+New. Finally, in the case of referential alleviation, we replicated the context alleviation effect in the verb+*by* region from the human experiments, but it seems that the effect in RNNs was driven by the reduced-relative verb alone. These results raise the possibility that the human findings in the verb+*by* region may also be driven by priming of the semantic content of the reduced-relative verb rather than solely by the previously accepted explanation of increased expectation of subject modification.

Models trained on both ordered and shuffled data exhibited the single-sentence garden-path effect, but only models trained on ordered data exhibited additional pragmatic constraints. The difference in model behavior based on training condition suggests that syntactic aspects of the garden path effect are distinct from the acquisition of the pragmatic knowledge used to alleviate them.

In the introduction we asked which aspects of discourse structure these models learn and whether these representations influence on-line syntactic representations. Our results suggest that components of tense, uniqueness, and definiteness can be learned from linguistic data alone, without any structures pre-defined as such. This strengthens work that has shown that distributional data carry rich semantic knowledge (e.g., Lupyan & Lewis, 2019; Lewis, Zettersten, & Lupyan, 2019), and suggests that more pragmatic knowledge is contained in language statistics than is commonly assumed. Further work on what aspects of pragmatics RNN LMs can and cannot learn will provide a bound on what linguistic phenomena is possible to learn without extra-linguistic knowledge.

Acknowledgements

We would like to thank members of the C.Psyd lab at Cornell University, the Cornell NLP group, and the Altmann and Yee labs at University of Connecticut, who gave feedback on earlier forms of this work. We would also like to thank the anonymous reviewers for their helpful suggestions and comments.

References

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*(3), 191–238.

Besserman, A., & Kaiser, E. (2016). The effects of discourse cues on garden path processing. In *Proceedings of CogSci.*

Chaves, R. P. (2020). What Don't RNN Langauge Models Learn About Filler-Gap Dependencies? In *Proceedings of SCiL* (Vol. 3, pp. 20–30).

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

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current directions in psychological science*, *11*(1), 11–15.

Ferreira, F., & Clifton Jr, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, *25*(3), 348–368.

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1-2), 71–83.

Frank, S. L., & Hoeks, J. (2019). The interaction between structure and meaning in sentence comprehension: Recurrent neural networks and reading times. *PsyArXiv preprint:10.31234*.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology*, *14*(2), 178–210.

Futrell, R., & Levy, R. (2019). Do RNNs learn human-like abstract word order preferences? In *Proceedings of SCiL* (Vol. 2, pp. 50–59).

Grice, H. P. (1975). Logic and Conversation. *Syntax and Semantics, Speech Acts*, *3*, 41–58.

Grodner, D., Gibson, E., Argaman, V., & Babyonyshev, M. (2003). Against repair-based reanalysis in sentence comprehension. *Journal of Psycholinguistic Research*, *32*(2), 141–166.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL.*

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of NAACL* (pp. 1–8).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410.*

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

Lewis, M., Zettersten, M., & Lupyan, G. (2019). Distributional semantics as a source of visual knowledge. *Proceedings of the National Academy of Sciences*, *116*(39), 19237–19238.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

Lupyan, G., & Lewis, M. (2019). From words-as-mappings to words-as-cues: the role of language in semantic knowledge. *Language, Cognition and Neuroscience*, *34*(10), 1319-1337.

McClelland, J. L., St. John, M., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and cognitive processes*, *4*(3-4), SI287–SI335.

Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). *Wikitext-103* (Tech. Rep.). Salesforce.

Spivey-Knowlton, M. J., Trueswell, J. C., & Tanenhaus, M. K. (1993). Context effects in syntactic ambiguity resolution: Discourse and semantic influences in parsing reduced relative clauses. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *47*(2), 276.

Trueswell, J. C., & Tanenhaus, M. K. (1991). Tense, temporal context and syntactic ambiguity resolution. *Language and Cognitive Processes*, *6*(4), 303–338.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, *33*(3), 285–318.

van Schijndel, M., & Linzen, T. (2018). Modeling garden path effects without explicit hierarchical syntax. In *Proceedings of CogSci.*

Wilcox, E., Levy, R., & Futrell, R. (2019). Hierarchical representation in neural language models: Suppression and recovery of expectations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP.*

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP* (pp. 211–221).