## Syntactic Structure aids Learning of Grammatical Dependencies in Neural Networks

Ethan Wilcox[1], Peng Qian[2], Richard Futrell[3], Miguel Ballesteros[4], Roger Levy[2]
[1]Harvard University, [2]MIT, [3]UC Irvine, [4]IBM
Contact: wilcoxeg@g.harvard.edu

Recurrent Neural Networks (RNNs: Elman, 1990, 1991) have state-of-the-art performance on a range of linguistic tasks (Jozefowicz et al., 2016), but the nature of the representations they learn is poorly understood. A recent line of work assesses the grammatical competence of RNNs by treating them like human subjects in psycholinguistics experiments. In this paradigm, the network is fed hand-crafted sentences designed to draw out behavior that reveals underlying representation. For example, Linzen et al. (2016) and Gulordava et al. (2018) found that RNN language models learn to represent subject/verb number agreement. When they fed the prefix "The keys to the cabinet…" the models robustly preferred "are" over the ungrammatical "is". Here, we investigate whether explicit representations of syntactic structure help such models learn long-distance grammatical dependencies. We comparatively evaluate two different types of RNN language models, one that computes explicit parse trees and one that does not, both trained on the Penn Treebank Corpus. We find that explicit representation of syntax aids the learning of structurally-adjacent dependencies, but that both models have difficulty threading word expectations through embedded clauses.

The two RNN-based models we test are **Long Short Term Memory** networks (LSTMs), sequential models with no obvious hierarchical bias; and **Recurrent Neural Network Grammars** (RNNGs) (Dyer et al., 2016), which are trained on syntactically-annotated data and represent the joint probability of an upcoming word and a syntactic parse. We use the neural network's **surprisal** at a word (-log $p(\mathit{word}|\mathit{context})$) (Hale, 2001; Levy, 2008) in order to investigate the model's expectation for covariance between an upstream licensor and a downstream licensee. A grammatical licensor should set up an expectation for a licensee, reducing the licensee's surprisal compared to minimal pairs with no licensor.

**NPI Licensing.** We find that the LSTM doesn't learn the licensor-NPI relation at all (Fig. 1). However, the RNNG model does: it is sensitive to the polarity of the c-commanding licensor, and although it can be misled by non-c-commanding linearly-proceeding "distractors" (as can humans: Vasishth et al, 2008), the effects from the licensor position are stronger than those from the distractor position.

**Filler-gap dependencies.** We used the technique from Wilcox et al. (2018), who quantify whether large-data LSTM models learn the dependency by calculating the **wh-licensing interaction**, which is the size of the 2x2 interaction between the presence of both a filler and a gap on the total surprisal of a post-gap critical region. RNNGs and LSTMs both learn the flexibility of the constraint (Fig. 2); the RNNG also learns that the dependency is robust to intervening PP and relative clause modification (Fig. 3), and is hierarchically constrained (Fig. 4). However, neither model can thread the dependency through embedded clause modification (Fig. 5). Finally, we investigate whether the models have learned island constraints, looking for a significant reduction in wh-licensing interaction as an indication that the model has learned the island constraint (Fig 6-7). RNNG exhibits more humanlike behavior than the LSTM, but the tests were inconclusive: island-like behavior may merely be sensitivity to general syntactic complexity. Thus, while the syntactic structure in the RNNG aids dependency learning in structurally tree-local contexts, it does not provide enough information for the neural network component to learn fully robust and human-like filler-gap dependencies from 1-million words alone.

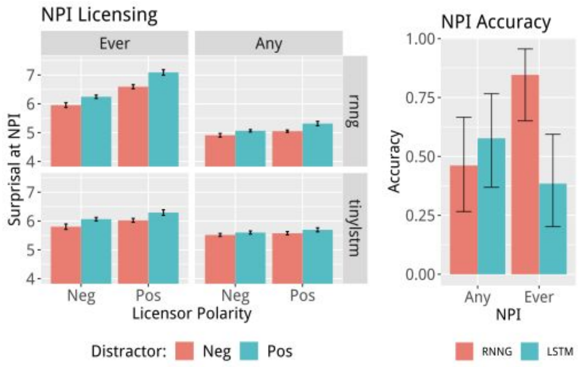

**Fig 1. Negative Polarity Item Licensing** for "any" and "ever." NPI Licensing at left: Y-axis shows surprisal at the NPI, x-axis indicates polarity of the c-commanding licensor, and color indicates distractor polarity. Licensing accuracy at right: Y-axis shows classification accuracy, or % of time NPI surprisal in (b) is lower than in (c) x-axis indicates the NPI tested, and color indicates the model. Error bars represent 95% binomial confidence intervals.

**(a) [Pos Licensor, Pos Distractor ]** *The senator that supported the measure has ever found support..

**(b) [Neg Licensor, Pos Distractor ]** No senator that supported the measure has ever found support..

**(c) [Pos Licensor, Neg Distractor ]*** The senator that supported no measure has ever found support..

**(d) [Neg Licensor, Neg Distractor]** No senator that supported no measure has ever found support from her constituents

**Fig 2 - 7: Filler-Gap Dependencies.** Y-axis is wh-licensing interaction, which measures the strength of the filler-gap dependency in each condition. ✓ indicates high expected wh-licensing interaction, ✘ indicates low expected wh-licensing interaction. Error bars are 95% confidence intervals with within-item means subtracted, as advocated in by Masson and Lotfus (2003).

🟥 RNNG 🟦 LSTM

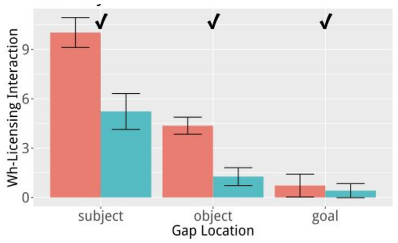

**Fig 2.** Flexibility of the Filler-Gap Dependency.

[Subj] I know who __ gave the gift to Alex yesterday.

[Obj] I know what Mary gave ___ to Alex yesterday.

[Goal] I know who Mary gave the gift to ___ yesterday.

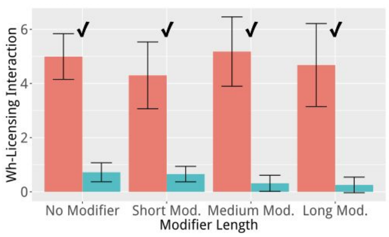

**Fig 3.** Robustness to intervening material

[No Mod] I know who the man insulted __ yesterday.

….

[Long Mod] I know who the man in the straw hat who recently arrived from New York insulted __ yesterday.

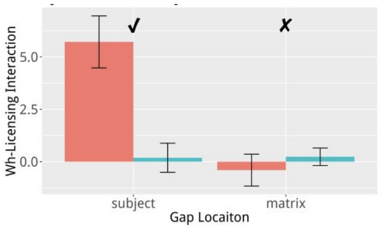

**Fig 4.** Sensitivity to Syntactic Hierarchy

[Subject] The policeman who the criminal shot with his gun __ shocked the jury during the trial.

[Matrix] *The policeman who the criminal shot the politician with his gun shocked __ during the trial.

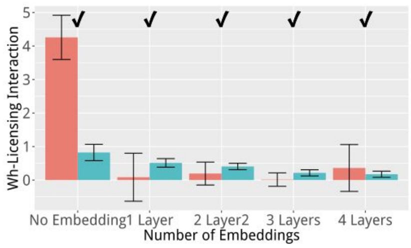

**Fig 5.** Unboundedness of the filler-gap dependency, indicating whether models can thread filler-gap expectation through embedded clauses..

[No Emb] I know who your aunt insulted __ at the party.

...

[4 Layers] I know who the hostess believed the butler reported his friend heard your aunt insulted __ at the party.

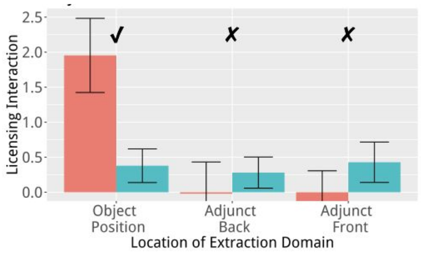

**Fig 6.** Adjunct Islands.

[Object] I know what the librarian placed ___ on the wrong shelf.

[Adjunct Back] I know what the librarian got mad after the patron placed ___ on the wrong shelf.

[Adjunct Front] I know what, after the patron placed __ on the wrong shelf, the librarian got mad.

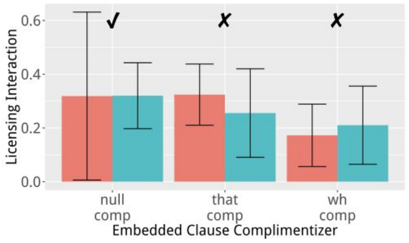

**Fig 7.** Wh Islands.

[Nul-Comp] I know what Alex said Sam bought __ yesterday.

[That-Comp] I know what Alex said that Sam bought __ yesterday.

[Wh-Comp] I know what Alex said whether Sam bought __ yesterday.