A Closer Look at the Performance of Neural Language Models on Reflexive Anaphor Licensing

Jennifer Hu
Dept. of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA
jennhu@mit.edu

Sherry Yong Chen
Dept. of Linguistics and Philosophy
Massachusetts Institute of Technology
Cambridge, MA
sychen@mit.edu

Roger Levy
Dept. of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA
rplevy@mit.edu

Abstract

An emerging line of work uses psycholinguistic methods to evaluate the syntactic generalizations acquired by neural language models (NLMs). While this approach has shown NLMs to be capable of learning a wide range of linguistic knowledge, confounds in the design of previous experiments may have obscured the potential of NLMs to learn certain grammatical phenomena. Here we re-evaluate the performance of a range of NLMs on reflexive anaphor licensing. Under our paradigm, the models consistently show stronger evidence of learning than reported in previous work. Our approach demonstrates the value of well-controlled psycholinguistic methods in gaining a fine-grained understanding of NLM learning potential.[1]

1 Introduction

To gain a deeper understanding of the grammatical generalizations acquired by neural language models (NLMs), an emerging line of work seeks to evaluate NLMs as "psycholinguistic subjects" – that is, assessing the extent to which their probability distributions conform to human judgments on linguistic data. This psycholinguistic assessment is typically done by evaluating the model on minimal pairs of sentences, which differ only at a target word or phrase that determines the acceptability of the sentence. If an NLM has learned the linguistic phenomenon in question, then it should assign higher probability to sentences that humans judge to be more acceptable. This approach has shown NLMs to be capable of learning some grammatical phenomena (e.g. subject-verb agreement and filler-gap dependencies) while failing on others (Linzen et al., 2016; Lau et al., 2017; Futrell et al., 2018; Gulordava et al., 2018; Marvin and Linzen, 2018; Tran et al., 2018; Wilcox et al., 2018).

In evaluating these mixed learning outcomes, we raise a broader question that remains largely unaddressed in the field: *What is the standard to which we should be holding artificial language models?* An engineering goal within the machine learning community is to build NLMs that approximate human behavior. In this case, an ideal NLM should achieve high performance even on low-frequency constructions, and the learning signal should be detectable even with coarse experimental paradigms. However, if a scientific goal is to highlight the grammatical phenomena that can be learned from sequential data, then experiments should be designed with the aim to give NLMs a fair shot at displaying successful learning.

We demonstrate the value of robust psycholinguistic methods in serving the latter goal by re-evaluating the performance of neural language models on English reflexive anaphor licensing (RAL). For example, in *John disappointed himself*, the reflexive *himself* can refer to *John*, but in *John knew that Paul disappointed himself*, the reflexive can only refer to *Paul* but not *John*. A priori, we expect RAL to be difficult to learn for sev-

[1] Code and data are available at https://github.com/jennhu/reflexive-anaphor-licensing.

Proceedings of the Society for Computation in Linguistics (SCiL) 2020, pages 382-392.
New Orleans, Louisiana, January 2-5, 2020

eral reasons. From a theoretical perspective, multiple syntactic constraints are simultaneously operative in RAL, which may increase the complexity of the representation that needs to be learned (see Section 2.1). In addition, the appearance of a reflexive is never obligatory based on the preceding context – that is, while a reflexive requires an antecedent NP licensor, an antecedent NP never requires a reflexive downstream (see Section 2.2).

Previous studies have shown NLMs to fail at RAL in various syntactic configurations (Futrell et al., 2018; Marvin and Linzen, 2018). We take a closer look at these previously reported failures, conducting new experiments that control for confounding variables and creating new materials that are compatible with small-vocabulary NLMs. Our experiments detect stronger evidence of learning than reported in previous work, demonstrating the value of robust psycholinguistic methods in studying the potential of NLMs to learn complex syntactic phenomena.

2 Background

2.1 Reflexive anaphor licensing (RAL)

English reflexive anaphors are licensed only when two different structural constraints are both satisfied, which we refer to as LOCALITY and C-COMMAND. These two constraints are independently motivated on theoretical grounds and underlie many syntactic configurations (e.g. Reinhart, 1983; Rizzi, 2013).

LOCALITY stipulates that the matching antecedent must be in the same clause as the reflexive. C-COMMAND requires the matching antecedent to be in a c-commanding relation with the reflexive (Reinhart, 1981; Chomsky, 1993). For present purposes, it is sufficient to define c-command as the following: if a node has any sibling nodes in a syntax tree, then it c-commands its siblings and all of their descendants; if a node has no siblings, then it c-commands everything its parent c-commands.

To illustrate these two constraints, Figure 1 shows the syntax tree for the sentence *The fathers said the women near the boys saw themselves*. This sentence contains three noun phrases (NPs) that could potentially act as an antecedent for *themselves*, but only one of them satisfies both structural requirements of RAL: (1) the higher subject NP_1 *the fathers* c-commands *themselves* but is not within the local clause, violating LOCALITY; (2) the lower subject NP_2 *the women* c-commands *themselves* locally, licensing the reflexive; (3) the linearly closest NP_3 *the boys* is within the local clause, but violates C-COMMAND since it is inside a prepositional phrase inside NP_2. Thus, NP_2 *the women* is the only possible licensor for the reflexive *themselves*.

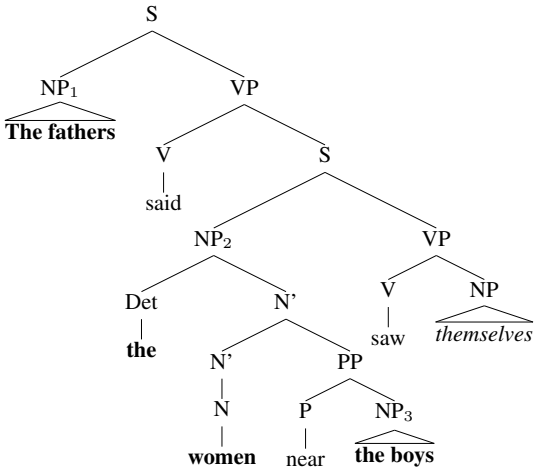

Figure 1: Syntax tree for example sentence. While each NP agrees in number with the reflexive *themselves*, only NP_2 occurs in a position that can license it.

We frame our experiments in terms of the two syntactic constraints involved in RAL, i.e. LOCALITY and C-COMMAND. This is typically done when testing the linguistic knowledge of humans, in order to probe the nature of linguistic generalizations that are being drawn across different types of constructions. In following this convention, we do not intend to claim the NLMs are learning these abstract structural properties per se.

2.2 Distribution of reflexive anaphors

The presence of a reflexive anaphor is never obligatory, in the sense that nothing in the preceding context deterministically predicts an upcoming reflexive. This contrasts with other syntactic dependencies, where the two elements of the dependency mutually require each other. In subject-verb agreement, for example, a subject NP sets the expectation for a downstream verb that agrees in number, and the verb requires a matching subject. This is also the case for less frequent constructions such as filler-gap dependencies, where the appearance of a filler *wh*-word sets the expectation for a gap, and the presence of a gap requires a preceding filler. This property does not hold for reflexive anaphors, as an NP never requires the appear-

ance of a reflexive downstream. Thus, given an upstream reflexive licensor, there is high variance in the downstream contexts.

Furthermore, although we are interested in reflexive anaphors that occur in an argument position, these pronouns can also occur as an intensifier adjoining right next to an NP, as in *The president himself signed my book*. Since the intensifier usage does not obey the same structural constraints, it has a different distribution from the anaphor usage. Both of the factors discussed above pose a challenge for NLMs to learn a robust representation for RAL.

2.3 Paradigms in previous work

Previous work evaluating the ability of neural language models to learn RAL primarily builds upon the paradigms introduced in Marvin and Linzen (2018) and Futrell et al. (2018). Both studies conclude that NLMs fail to learn the appropriate licensing conditions for reflexives.

In particular, Marvin and Linzen (2018) test whether NLMs learn RAL in relative clauses and sentential complements. Consider the following sample items (1) and (2) from their study:

(1) The bankers who the pilot embarrassed hurt *himself / themselves.

(2) The bankers thought the pilot embarrassed himself / *themselves.

In (1), the reflexive *himself* cannot be licensed by *the pilot* because *the pilot* is inside a relative clause, thus violating both LOCALITY and C-COMMAND. In (2), the reflexive *themselves* is embedded in a sentential complement, so the long-distance subject *the bankers* cannot license the reflexive for violating LOCALITY.

As is typical in psycholinguistic evaluation of NLMs, previous RAL studies calculate accuracy as the proportion of trials where the model assigns higher probability to the correct reflexive given the prefix, compared to another reflexive that would make the sentence ungrammatical. Since Marvin and Linzen (2018) and Futrell et al. (2018) test number and gender agreement, respectively, Marvin and Linzen compare the probability of *himself/herself* vs. *themselves*, while Futrell et al. compare the probability of *himself* vs. *herself*.

While the failures reported by these studies have been taken as evidence of the limits of NLM learning, they might be attributed to confounding factors in the design of the experiments. As discussed above, previous studies measure accuracy by comparing the probability assigned to different target reflexives given the same context. However, in many standard training corpora, the reflexive pronouns *themselves*, *himself*, and *herself* differ dramatically in frequency, leading to an asymmetry in unigram probabilities (Table 2). This presents a confound, as all models are likely to implicitly factor unigram probabilities when estimating conditional probabilities in context.[2] Thus, even if a model has learned correct generalizations about the relevant features of the context, these generalizations could be obscured by large differences in unigram frequency.

In addition, both Marvin and Linzen (2018) and Futrell et al. (2018) use profession nouns that are almost all stereotypically male (e.g. *banker*, *senator*). However, many of these nouns occur with low frequency in standard training datasets, so existing materials cannot be used to test RAL learning in models with relatively small vocabularies.

To re-evaluate NLM learning potential of RAL, we conduct new experiments that mitigate the issues raised by unigram probability asymmetries and stereotypically gendered nouns. We describe our methods in Section 3.

3 Experimental design

Psycholinguistic evaluation of language models typically measures accuracy as the proportion of trials in which the model correctly assigns higher probability to the grammatical sentence in a minimal pair. This probability differential is affected not only by the expectations set by the context, but also by the unigram probabilities of the target words (in the case of RAL, *themselves*, *himself*, and *herself*). To avoid this issue, we keep the target reflexive fixed and vary the preceding lexical items in each condition.

3.1 Conditions

Each sentence in our test suites has two NPs, a verb, and a target reflexive, as well as material that modulates the syntactic state (e.g. the onset of a relative clause). One NP is in a position that can license a reflexive, and the other NP is not. Our experiments have the following three conditions:

[2] A unigram frequency is one of the easiest things for a neural model to learn, e.g. as the bias term in the output layer.

- **Baseline:** Both NPs match the number feature of the target reflexive. The sentence is grammatical.
- **Distractor:** The NP in the licensing position matches the number of the target, but the other NP mismatches. The sentence is still grammatical, but contains distracting material.
- **Ungrammatical:** The NP in the licensing position mismatches the number of the target. The sentence is ungrammatical.

We choose to test number instead of gender feature agreement (cf. Futrell et al., 2018) because we believe models are more likely to learn a representation of number than gender, as number is more frequently marked than gender in English. There is also evidence of NLMs learning other number-based dependencies such as subject-verb agreement (Linzen et al., 2016).

3.2 Evaluation metric

Our accuracy calculation involves a three-way comparison. For a given item, the model makes a correct prediction if the probability of the target reflexive in the Ungrammatical condition is lower than the probability of the target in *both* the Distractor and Baseline conditions. Accuracy is the proportion of items in the experiment for which the model makes the correct prediction. If the probability of the target is the same across conditions, then the prediction is considered correct with probability ⅓. Under this measure, chance performance is 33.33%, in contrast to the 50% from existing paradigms that compare grammatical vs. ungrammatical constructions.

3.3 Lexical items

Nouns Previous studies on RAL use nouns denoting professions often associated with stereotypical gender, such as *lumberjack* and *hairdresser* (Futrell et al., 2018; Marvin and Linzen, 2018).[3] However, these nouns are not inherently gendered, and manipulating the gender of the reflexive does not change the grammaticality of the sentence. Instead, we use high-frequency nouns with lexicalized gender, such as *man* and *woman*. This allows us to extend our paradigm to models with smaller vocabularies (see Section 4), for which many profession nouns are out-of-vocabulary (e.g. *hairdresser*). This also ensures that our experiments can be replicated with future corpora, as the stereotypical gender of occupations represented in word embeddings can vary across time and cultures (Garg et al., 2018). We selected a total of 10 nouns (5 female and 5 male), with the female and male nouns balanced for frequency of occurrence in the Wikipedia corpus (see Table 2).

Verbs We first manually constructed a list of commonly used reflexive verbs. Using this list, we calculated the relative frequency of their occurrences within a reflexive construction in the Wikipedia corpus, and selected the most frequent ones. We also selected the most frequent verbs by their raw counts in the corpus. A total of 15 verbs were selected using this method.

Counterbalancing To ensure that vocabulary differences in preceding context do not confound the observed effects on the target reflexive, we counterbalance the position of nouns such that each noun occurs in a licensing and a non-licensing position equally often. Consequently, each stimulus item has several variants, where the nouns are equally distributed across positions. Each noun also appears with each of the verbs equally often across items.

3.4 Logic of experiments

In Experiment 1, we first perform a loose replication of Marvin and Linzen (2018) by adapting their materials into our experimental paradigm. The experiment includes relative clause and sentential complement constructions, which we test in Experiments 1a and 1b, respectively. To construct the materials, we crossed 10 nouns with 7 matrix verbs from the original Marvin and Linzen study, resulting in a total of 70 items per pronoun.

As discussed in Section 2.3, one issue with previous studies is the choice to use lexical items with stereotypical gender. In subsequent experiments, we create new test suites with materials using lexicalized gender. In Experiments 2a and 2b, we use our new materials to test relative clause and sentential complement constructions, respectively, for comparison with Experiments 1a and 1b.

Since the relative clause construction tests both LOCALITY and C-COMMAND and the sentential complement construction only tests LOCALITY, we test prepositional phrases in Experiment 3 to isolate the effect of C-COMMAND. We cross 4

[3] RNNs have been shown to learn NP stereotypical gender (Rudinger et al., 2018).

	Condition	Example sentence
LOCALITY & C-COMMAND		
Relative clause (M&L)	Grammatical	The bankers who the pilot embarrassed hurt themselves
	Ungrammatical	*The bankers who the pilot embarrassed hurt herself
Relative clause (Exp. 1a)	Baseline	The {banker, pilot} that the {pilot, banker} embarrassed hurt herself
	Distractor	The {banker, pilot} that the {pilots, bankers} embarrassed hurt herself
	Ungrammatical	*The {bankers, pilots} that the {pilot, banker} embarrassed hurt herself
Relative clause (Exp. 2a)	Baseline	The {mother, girl} that the {girl, mother} liked saw herself
	Distractor	The {mother, girl} that the {girls, mothers} liked saw herself
	Ungrammatical	*The {mothers, girls} that the {girl, mother} liked saw herself
LOCALITY ONLY		
Sentential complement (M&L)	Grammatical	The bankers thought the pilot hurt herself
	Ungrammatical	*The bankers thought the pilot hurt themselves
Sentential complement (Exp. 1b)	Baseline	The {banker, pilot} said that the {pilot, banker} hurt herself
	Distractor	The {bankers, pilots} said that the {pilot, banker} hurt herself
	Ungrammatical	*The {banker, pilot} said that the {pilots, bankers} hurt herself
Sentential complement (Exp. 2b)	Baseline	The {mother, girl} said that the {girl, mother} saw herself
	Distractor	The {mothers, girls} said that the {girl, mother} saw herself
	Ungrammatical	*The {mother, girl} said that the {girls, mothers} saw herself
C-COMMAND ONLY		
Prepositional phrase (Exp. 3)	Baseline	The {mother, girl} near the {girl, mother} saw herself
	Distractor	The {mother, girl} near the {girls, mothers} saw herself
	Ungrammatical	*The {mothers, girls} near the {girl, mother} saw herself

Table 1: Sample stimuli for *herself* in our experiments and the original Marvin and Linzen ("M&L") study.

nouns with 15 verbs, resulting in 60 items for each pronoun in each of Experiments 2 and 3.[4] Table 1 shows sample items for Experiments 1-3 along with corresponding items from the original Marvin and Linzen (2018) study.

4 Language models

We evaluate RAL in six neural language models, as well as a baseline n-gram model. Together, the models cover a range of vocabulary sizes, architectures, and inductive biases (Table 2). Our goal here is not to draw general conclusions about certain architectures or training regimes, but to present results across a diverse set of models, including those that were previously untestable due to experimental design.

GRNN and JRNN Recurrent neural networks (RNNs; Elman, 1990; Mikolov et al., 2010) perform well in language modeling, with long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997; Sundermeyer et al., 2012) being the most popular variant. We test two LSTMs that differ significantly in vocabulary size and have been shown to learn syntactic dependencies to varying degrees of success. The Gulordava et al. (2018) LSTM ("GRNN") was trained on a subset of English Wikipedia with 90M training tokens. The Jozefowicz et al. (2016) LSTM ("JRNN") was trained on the One Billion Word Benchmark (Chelba et al., 2013). JRNN additionally has convolutional neural network character input embeddings.

Transformer-XL and BERT Next, we test two models based on the Transformer architecture (Vaswani et al., 2017). Transformer-XL ("TransXL"; Dai et al., 2019) reuses the hidden states obtained in previous segments, which facilitates modeling of long-term dependencies. BERT (Devlin et al., 2018) is bi-directional, in that it is trained to predict the identity of masked words based on the preceding and following context.[5] Both models were trained on document-level corpora instead of shuffled sentences: WikiText-103

[4] To counterbalance the position of the nouns, there are 6 variants of each item (2 per condition) for *himself* and *herself*, and 12 variants of each item (4 per condition) for *themselves*.

[5] We use the small, uncased version of BERT ($\text{BERT}_{\text{BASE}}$) with no fine-tuning after the initial pre-training tasks.

Model	Architecture	Training data	Training tokens	Vocab size	*themselves*	*himself*	*herself*
BERT	Transformer	BooksCorpus, Wikipedia	3.3B	30K	-	-	-
TransXL	Transformer	WikiText-103	103M	267K	9K	20K	5K
JRNN	LSTM	1B Word Benchmark	1B	800K	103K	124K	34K
GRNN	LSTM	Wikipedia	90M	50K	10K	17K	4K
TinyLSTM	LSTM	PTB §2-21 (terminals)	950K	23K	114	95	12
RNNG	RNNG	PTB §2-21 (trees)	950K	23K	114	95	12
5-gram	n-gram	Wikipedia	90M	50K	10K	17K	4K

Table 2: Language models evaluated in our experiments, along with raw frequency counts of reflexives in the training data. Pre-training data was not publicly released for BERT.

(Merity et al., 2017) for TransXL, and a combination of BooksCorpus (Zhu et al., 2015) and Wikipedia for BERT. Recent work has shown BERT to perform well on reflexive constructions (Goldberg, 2019).

RNNG and TinyLSTM The last two neural models in our test suite have identical vocabularies but differing inductive biases: a recurrent neural network grammar ("RNNG"; Dyer et al., 2016) and a vanilla LSTM ("TinyLSTM"). Both models were trained on the 1-million-word English Penn Treebank §2-21 (Marcus et al., 1993), but TinyLSTM is only trained on the terminal word sequences, while RNNG is trained on the full annotations, which contain complete constituency parses. This minimal difference allows us to observe the effect of structural supervision, which has been shown to be beneficial in acquiring certain grammatical dependencies (Kuncoro et al., 2017; Wilcox et al., 2019). Crucially, the vocabulary of these models is too small to acommodate the lexical items used in previous RAL studies.

***n*-gram** As a baseline, we test a 5-gram model trained on the same Wikipedia data as GRNN. We use Kneser-Ney smoothing to perform backoff.

4.1 Computing word probabilities

In practice, we calculate accuracy (see Section 3.2) by comparing differentials in log probability space at the target pronoun. To obtain the log probability of word w_i assigned by the LSTMs and Transformer models, we compute

$$\log_2 p(w_i|h_{i-1}), \quad (1)$$

where h_{i-1} is the model's hidden state before observing w_i. This probability is calculated from the model's softmax activation.

To obtain the log probability of w_i in the RNNG, we follow the method used in Hale et al. (2018). We use word-synchronous beam search (Stern et al., 2017) to find the most likely incremental parses, and sum their forward probabilities to approximate $P(w_1, \ldots, w_{i_{+1}})$ and $P(w_1, \ldots, w_{i-1})$. We use 100 for the action beam size and 10 for the word beam size.

In contrast to the other models in our test suite, BERT is bi-directional. To obtain the log probability of w_i, we first feed BERT a sentence with w_i masked out and obtain the word predictions for the masked position. This gives us a probability distribution over words. In practice, since the target reflexive in our items always occurs directly before the final token '.', we do not expect the right context to modulate predictions about the target differently across conditions.

5 Results

5.1 Experiment 1: Marvin and Linzen (2018)

The original materials of Marvin and Linzen (2018) use profession nouns that are stereotypically male. Since these nouns are out-of-vocabulary for RNNG and TinyLSTM, we run this experiment only on the large-vocabulary models (BERT, TransXL, JRNN, GRNN, 5-gram).

Exp. 1a: M&L relative clause We first investigate RAL learning in the relative clause construction (see Table 1). Here, the NP inside the relative clause cannot license the reflexive, as such a relationship would violate both LOCALITY and C-COMMAND. Our design differs from Marvin and Linzen (2018) in that we hold the reflexive anaphor constant while varying the context, with the position of the nouns counterbalanced.

Accuracy scores from the original study and

	BERT	TransXL	JRNN	GRNN	TinyLSTM	RNNG	5-gram
LOCALITY & C-COMMAND							
Relative clause (M&L)	0.80†	–	–	0.55*	–	–	0.50*
Relative clause (Exp. 1a)	0.76 ± 0.057	0.74 ± 0.059	0.41 ± 0.067	0.70 ± 0.062	–	–	0.33
Relative clause (Exp. 2a)	0.70 ± 0.067	0.70 ± 0.067	0.68 ± 0.068	0.45 ± 0.073	0.16 ± 0.053	0.24 ± 0.062	0.33
LOCALITY ONLY							
Sentential complement (M&L)	0.98†	–	–	0.86*	–	–	0.50*
Sentential complement (Exp. 1b)	0.95 ± 0.029	0.91 ± 0.038	0.96 ± 0.026	1.00 ± 0	–	–	0.33
Sentential complement (Exp. 2b)	0.98 ± 0.022	0.92 ± 0.039	0.97 ± 0.026	0.99 ± 0.013	0.82 ± 0.057	0.88 ± 0.047	0.33
C-COMMAND ONLY							
Prepositional phrase (Exp. 3)	0.99 ± 0.008	0.71 ± 0.067	0.69 ± 0.063	0.75 ± 0.063	0.43 ± 0.072	0.62 ± 0.071	0.33

Table 3: Accuracy scores for each experiment, with 95% confidence intervals shown below where applicable. Accuracy is computed at the item-level for each pronoun, then averaged across all pronouns. Chance accuracy is 33.33%, except for entries marked with † or *, where chance is 50%. The BERT results marked with † come from Goldberg (2019), while the GRNN and 5-gram results marked with * come directly from Marvin and Linzen (2018). These results are also not directly comparable to each other due to the bi-directionality of BERT; see Goldberg (2019) and Wolf (2019) for details.

our Experiment 1 are reported in Table 3 (top two rows). Accuracy is computed at the item-level for each pronoun, then averaged across all pronouns. Under our evaluation method, GRNN shows considerable improvement over what was reported in Marvin and Linzen (2018), while the 5-gram model remains at chance. While our metrics are not strictly comparable, the original study reports near-chance accuracy ($55\% \sim 50\%$), while we report accuracy well above chance ($70\% \gg 33.33\%$). BERT achieves slightly lower accuracy under our paradigm than was reported in Goldberg (2019) (76% vs. 80%); note, however, that our chance baseline is lower.

Exp. 1b: M&L sentential complement Next, we investigate RAL learning in the sentential complement construction. Here, the long-distance subject cannot license the reflexive embedded in a sentential complement, because such a relationship would violate LOCALITY (while satisfying C-COMMAND). As in Exp. 1a, our approach differs from Marvin and Linzen (2018) in that we hold the reflexive anaphor constant while varying the context, with the position of the nouns counterbalanced.

All large-vocabulary neural models perform near ceiling in our paradigm, despite our metric having a lower baseline. GRNN achieves 100% accuracy, showing a marked improvement over previously reported results (Table 3). Overall, the models exhibit the correct trend for the sentential complement construction (Exp. 1b), but the pattern is less clear for the relative clause construction (Exp. 1a). One possible explanation is that in a relative clause, the licensing NP is linearly farther away from the reflexive than the distracting NP; a global preference for linear proximity may have obscured learning of structural adjacency.

5.2 Experiment 2

The materials used in Marvin and Linzen (2018) (and our Experiment 1) involve items with stereotypically gendered nouns. This raises two potential issues: (1) gender biases may overshadow number mismatch effects, and (2) the materials can only be used to evaluate models with reasonably large vocabularies. As in Experiment 1, the design of Experiment 2 differs from Marvin and Linzen (2018) in that we hold the reflexive anaphor constant while varying the context. In addition, we create new materials using nouns with lexicalized gender rather than stereotypical gender. This allows us to evaluate all seven models in our test suite.

Exp. 2a: Relative clause As in Exp. 1a, we first test RAL learning in the relative clause construc-

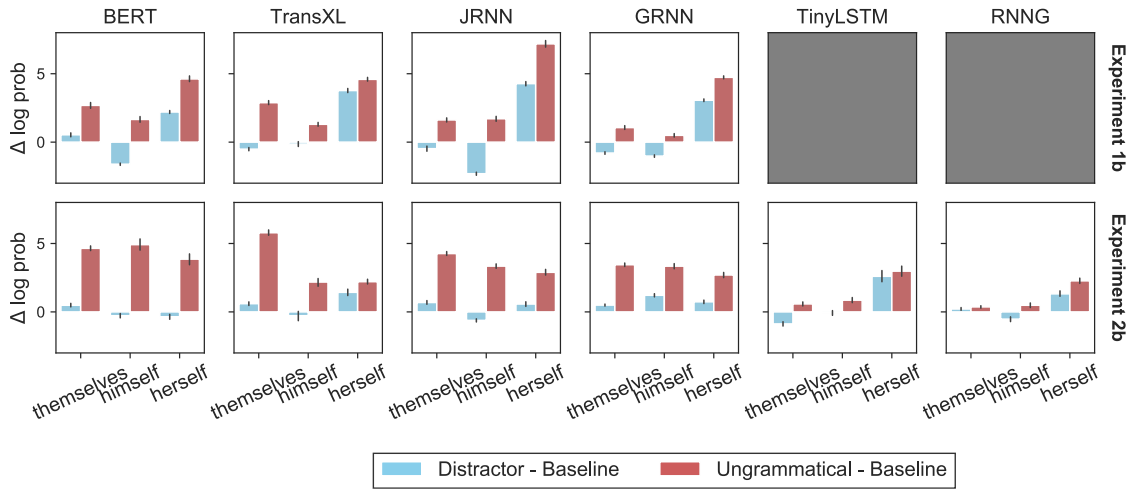

Figure 2: Negative log probability differential at target reflexive in sentential complement construction. Error bars are bootstrapped 95% confidence intervals. **Blue bars:** Distractor-Baseline differential at target reflexive. **Red bars:** Ungrammatical-Baseline differential at target reflexive. If the models learn the correct generalization for RAL, then the red bars should be both positive and higher than the blue bars. **Top (Exp. 1b:)** Distractor-Baseline differential is significantly higher at *herself* than *himself* or *themselves*. The stimuli contain materials that are out-of-vocabulary for TinyLSTM and RNNG. **Bottom (Exp. 2b):** For the large-vocabulary models, the Distractor-Baseline differential is comparable across pronouns. For the small-vocabulary models, the differential is significantly higher at *herself*.

tion using our new set of materials. Accuracy scores are high for most of the large-vocabulary neural models (BERT, TransXL, JRNN) and above chance for GRNN, but at or below chance for the other models (Table 3).

Exp. 2b: Sentential complement In Experiment 3, we test the sentential complement construction using our materials. As shown in Table 1, we place the reflexive inside a complement clause, such that either both c-commanding NPs match the number feature of the reflexive (Baseline), or there is one mismatching NP either in the non-local subject position (Distractor) or the local subject position (Ungrammatical).

All large-vocabulary neural models perform near ceiling (Table 3). The small-vocabulary models RNNG and TinyLSTM achieve lower accuracy, but RNNG outperforms TinyLSTM.

5.3 Experiment 3

Since previous studies have focused on the relative clause and sentential complement constructions, C-COMMAND has not been tested separately from LOCALITY. In Experiment 3, we hold LOCALITY constant while manipulating C-COMMAND by placing a distractor NP inside a non-c-commanding PP modifier in the local subject NP. No clausal boundary is introduced. As in Experiment 2, our approach differs from Marvin and Linzen (2018) in that we hold the reflexive anaphor constant while varying the context, and we use nouns with lexicalized gender.

Accuracy scores are reported in the bottom section of Table 3. Performance is well above chance for all neural models except TinyLSTM. RNNG shows a clear advantage over TinyLSTM (62% vs. 43%).

5.4 Asymmetry between *himself* & *herself*

Thus far, we have reported accuracy scores averaged across the three reflexive pronouns (Table 3). The three pronouns are weighted equally in the reported numbers, as accuracy is computed at the level of each item.

Next, we investigate differences in performance across reflexive anaphors. Figure 2 shows the results of this cross-pronoun comparison for Experiments 1b and 2b, which both use the sentential complement construction (LOCALITY only). Blue bars show the Distractor-Baseline log probability differential at the target reflexive. Red bars show the Ungrammatical-Baseline log probability differential at the target reflexive. If the models learn the correct generalization for RAL, then the red

bars should be both positive (i.e. above baseline) and higher than the blue bars.

In Experiment 1b, which uses profession nouns that are primarily associated with men,[6] the Distractor-Baseline differential (blue bars) is significantly higher at *herself* than at *himself* or *themselves*. In contrast, in Experiment 2b, which uses nouns with lexicalized gender, there is only a significant difference between the Distractor-Baseline differentials at *himself* and *herself* for the small-vocabulary models TinyLSTM and RNNG.

We hypothesize that this can be attributed to the choice of vocabulary items. In the Distractor condition of Experiment 1, the distracting noun is plural and has stereotypically male gender (e.g. *senators*). The features of this noun partially match with *himself* (in stereotypical gender but not number), but match in neither feature with *herself*, leading to a higher Distractor-Baseline differential for *herself*. This is not an issue in Experiments 2 and 3, where all nouns match in gender feature with the target reflexive across conditions. However, training data with a low number of occurrences of *herself* can still lead to a high Distractor-Baseline differential, as is the case in Experiment 3 for TinyLSTM and RNNG.

This pattern may also result from a more general asymmetry between gender stereotypes: encountering *herself* after a stereotypically male noun is more surprising than encountering *himself* after a stereotypically female noun. Interestingly, asymmetry also manifests in human production biases, where gendered pronoun production and interpretation are not mutually calibrated (Boyce et al., 2019).

6 Discussion

In this paper, we used new experiments to re-evaluate the performance of neural language models on reflexive anaphor licensing. Our methods address issues in previous studies, such as unigram probability asymmetries between target pronouns and the choice to use nouns with stereotypical gender, which may have led to an underestimation of learning signal. The results suggest that NLMs are learning more about RAL than they have previously been given credit for, and demonstrates the value of robust psycholinguistic methods in highlighting the potential of NLMs to learn complex syntactic phenomena.

The value of our approach extends beyond RAL. If we seek to understand the linguistic generalizations that NLMs can *potentially* acquire, then we must design our experiments to give NLMs a fair shot at displaying successful learning, regardless of the phenomenon under study.

Of course, the generalizations acquired by NLMs may not be well characterized in linguistic terms such as LOCALITY and C-COMMAND, but rather properties of the data that are irrelevant to structural considerations. Further experiments will be required to deepen our understanding of the generalizations underlying the successes and failures of these models on this and other evaluation tasks. More generally, future work in this domain should carefully address hypotheses about language learning, keeping in mind complementary questions that arise from engineering and scientific agendas.

[6] 11 out of these 12 nouns are stereotypically male according to United States Census data (Bureau of Labor Statistics, 2017).

Acknowledgments

We would like to thank Tal Linzen, Peng Qian, and the anonymous reviewers for their insightful comments. J.H. is supported by an NSF Graduate Research Fellowship.

References

Veronica Boyce, Titus von der Malsburg, Till Poppels, and Roger Levy. 2019. Remember 'him', forget 'her': Gender bias in the comprehension of pronomial referents. In *32nd Annual CUNY Conference on Human Sentence Processing*.

Bureau of Labor Statistics. 2017. Labor force statistics from the current population survey.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. Technical report, Google.

Noam Chomsky. 1993. *Lectures on government and binding: The Pisa lectures*. 9. Walter de Gruyter.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. *CoRR*, abs/1602.07776.

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

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *CoRR*, abs/1809.01329.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *CoRR*, abs/1901.05287.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan R. Brennan. 2018. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 2727–2736, Melbourne, Australia. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, Graham Neubig, and Noah A. Smith. 2017. What do recurrent neural network grammars learn about syntax? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1249–1258, Valencia, Spain. Association for Computational Linguistics.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 5:1202–1247.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 521–535.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *Proceedings of ICLR*.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, pages 1045–1048, Makuhari, Chiba, Japan.

Tanya Reinhart. 1981. Definite NP anaphora and c-command domains. *Linguistic Inquiry*, 12(4):605–635.

Tanya Reinhart. 1983. *Anaphora and semantic interpretation*. Routledge.

Luigi Rizzi. 2013. Locality. *Lingua*, 130:169–186.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Mitchell Stern, Daniel Fried, and Dan Klein. 2017. Effective inference for generative neural parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1695–1700. Association for Computational Linguistics.

Martin Sundermeyer, Ralf Schluter, and Hermann Ney. 2012. LSTM neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Ke Tran, Arianna Bisazza, and Christof Monz. 2018. The importance of being recurrent for modeling hierarchical structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4731–4736, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about fillergap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballestros, and Roger Levy. 2019. Structural supervision improves learning of non-local grammatical dependencies. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3302–3312, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf. 2019. Some additional experiments extending the tech report "Assessing BERT's syntactic abilities" by Yoav Goldberg. Technical report, HuggingFace, Inc.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 19–27, Washington, DC. IEEE Computer Society.