Recurrent Neural Network Language Models Always Learn English-Like Relative Clause Attachment

Forrest Davis and **Marten van Schijndel**
Department of Linguistics
Cornell University
{fd252|mv443}@cornell.edu

Abstract

A standard approach to evaluating language models analyzes how models assign probabilities to valid versus invalid syntactic constructions (i.e. is a grammatical sentence more probable than an ungrammatical sentence). Our work uses ambiguous relative clause attachment to extend such evaluations to cases of multiple simultaneous valid interpretations, where stark grammaticality differences are absent. We compare model performance in English and Spanish to show that non-linguistic biases in RNN LMs advantageously overlap with syntactic structure in English but not Spanish. Thus, English models may appear to acquire human-like syntactic preferences, while models trained on Spanish fail to acquire comparable human-like preferences. We conclude by relating these results to broader concerns about the relationship between comprehension (i.e. typical language model use cases) and production (which generates the training data for language models), suggesting that necessary linguistic biases are not present in the training signal at all.

1 Introduction

Language modeling is widely used as pretraining for many tasks involving language processing (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019). Since such pretraining affects so many tasks, effective evaluations to assess model quality are critical. Researchers in the vein of the present study, typically take (pretrained) language models and ask whether those models have learned some linguistic phenomenon (e.g., subject-verb agreement). Often the task is operationalized as: do the models match some human baseline (e.g., acceptability judgments, reading times, comprehension questions) measured as humans experience this linguistic phenomenon (e.g., comparing acceptability ratings of sentences with grammatical/ungrammatical agreement). This approach tacitly assumes that the necessary linguistic biases are in the training signal and then asks whether the models learn the same abstract representations as humans given this signal. The present study casts doubt on the notion that the necessary linguistic biases are present in the training signal at all.

We utilize the, now common, evaluation technique of checking whether a model assigns higher probability to grammatical sentences compared to ungrammatical sentences (Linzen et al., 2016). However, we extend beyond binary grammaticality. Real world applications demand that our models not only know the difference between valid and invalid sentences; they must also be able to correctly prioritize simultaneous valid interpretations (Lau et al., 2017). In this paper, we investigate whether neural networks can in fact prioritize simultaneous interpretations in a human-like way. In particular, we probe the biases of neural networks for ambiguous relative clause (RC) attachments, such as the following:

(1) Andrew had dinner yesterday with the nephew of the teacher *that was divorced*. (from Fernández, 2003)

In (1), there are two nominals (*nephew* and *teacher*) that are available for modification by the RC (*that was divorced*). We refer to attachment of the RC to the syntactically higher nominal (i.e. the nephew is divorced) as HIGH and attachment to the lower nominal (i.e. the teacher is divorced) as LOW.

As both interpretations are equally semantically plausible when no supporting context is given, we might expect that humans choose between HIGH and LOW at chance. However, it has been widely established that English speakers tend to interpret the relative clause as modifying the lower nominal more often than the higher nominal (i.e. they

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1979–1990
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

have a LOW bias;[1] Carreiras and Clifton Jr, 1993; Frazier and Clifton, 1996; Carreiras and Clifton, 1999; Fernández, 2003). LOW bias is actually typologically much rarer than HIGH bias (Brysbaert and Mitchell, 1996). A proto-typical example of a language with HIGH attachment bias is Spanish (see Carreiras and Clifton Jr, 1993; Carreiras and Clifton, 1999; Fernández, 2003).

A growing body of literature has shown that English linguistic structures conveniently overlap with non-linguistic biases in neural language models leading to performance advantages for models of English, without such models being able to learn comparable structures in non-English-like languages (e.g., Dyer et al., 2019). This, coupled with recent work showing that such models have a strong recency bias (Ravfogel et al., 2019), suggests that one of these attachment types (LOW), will be more easily learned. Therefore, the models might appear to perform in a human-like fashion on English, while failing on the cross-linguistically more common attachment preference (HIGH) found in Spanish. The present study investigates these concerns by first establishing, via a synthetic language experiment, that recurrent neural network (RNN) language models (LMs) are capable of learning either type of attachment (Section 4). However, we then demonstrate that these models consistently exhibit a LOW preference when trained on actual corpus data in multiple languages (English and Spanish; Sections 5–7).

In comparing English and Spanish, we show that non-linguistic biases in RNN LMs overlap with interpretation biases in English to appear as though the models have acquired English syntax, while failing to acquire minimally different interpretation biases in Spanish. Concretely, English attachment preferences favor the most recent nominal, which aligns with a general preference in RNN LMs for attaching to the most recent nominal. In Spanish, this general recency preference in the models remains despite a HIGH attachment interpretation bias in humans. These results raise broader questions regarding the relationship between comprehension (i.e. typical language model use cases) and production (which generates the training data for language models) and point to a deeper inability of RNN LMs to learn aspects of linguistic structure from raw text alone.

[1]We use "bias" throughout this paper to refer to "interpretation bias." We will return to the distinction between production bias and interpretation bias in Section 8.

2 Related Work

Much recent work has probed RNN LMs for their ability to represent syntactic phenomena. In particular, subject-verb agreement has been explored extensively (e.g., Linzen et al., 2016; Bernardy and Lappin, 2017; Enguehard et al., 2017) with results at human level performance in some cases (Gulordava et al., 2018). However, additional studies have found that the models are unable to generalize sequential patterns to longer or shorter sequences that share the same abstract constructions (Trask et al., 2018; van Schijndel et al., 2019). This suggests that the learned syntactic representations are very brittle.

Despite this brittleness, RNN LMs have been claimed to exhibit human-like behavior when processing garden path constructions (van Schijndel and Linzen, 2018; Futrell and Levy, 2019; Frank and Hoeks, 2019), reflexive pronouns and negative polarity items (Futrell et al., 2018), and center embedding and syntactic islands (Wilcox et al., 2019, 2018). There are some cases, like coordination islands, where RNN behavior is distinctly non-human (see Wilcox et al., 2018), but in general this literature suggests that RNN LMs encode some type of abstract syntactic representation (e.g., Prasad et al., 2019). Thus far though, the linguistic structures used to probe RNN LMs have often been those with unambiguously ungrammatical counterparts. This extends into the domain of semantics, where downstream evaluation platforms like GLUE and SuperGLUE evaluate LMs for correct vs. incorrect interpretations on tasks targeting language understanding (Wang et al., 2018, 2019).

Some recent work has relaxed this binary distinction of correct vs. incorrect or grammatical vs. ungrammatical. Lau et al. (2017) correlate acceptability scores generated from a LM to average human acceptability ratings, suggesting that human-like gradient syntactic knowledge can be captured by such models. Futrell and Levy (2019) also look at gradient acceptability in both RNN LMs and humans, by focusing on alternations of syntactic constituency order (e.g., heavy NP shift, dative alternation). Their results suggest that RNN LMs acquire soft constraints on word ordering, like humans. However, the alternations in Futrell and Levy, while varying in their degree of acceptability, maintain the same syntactic relations throughout the alternation (e.g., *gave a book to Tom* and *gave Tom a book* both preserve the fact that *Tom* is the

indirect object). Our work expands this line of research by probing how RNN LMs behave when multiple valid interpretations, with crucially different syntactic relations, are available within a single sentence. We find that RNN LMs do not resolve such ambiguity in a human-like way.

There are, of course, a number of other modeling approaches that exist in the current literature; the most notable of these being BERT (Devlin et al., 2019). These transformer models have achieved high performance on a variety of natural language processing tasks, however, there are a number of properties that make them less suitable to this work. One immediate consideration is that of training. We are interested in the behavior of a class of models, so we analyze the behavior of several randomly initialized models. We do not know how representative BERT is of models of its same class, and training more BERT variants is immensely time consuming and environmentally detrimental (Strubell et al., 2019). Additionally, we are interested in probability distributions over individual words given the preceding context, something that is not part of BERT's training as it takes whole sentences as input. Finally, the bidirectional nature of many of these models makes their representations difficult to compare to humans. For these reasons, we restrict our analyses to unidirectional RNN LMs. This necessarily reduces the generalizability of our claims. However, we still believe this work has broader implications for probing what aspects of linguistic representations neural networks can acquire using standard training data.

3 Methods

3.1 Experimental Stimuli

In the present study, we compare the attachment preferences of RNN LMs to those established in Fernández (2003). Fernández demonstrated that humans have consistent RC attachment biases using both self-paced reading and offline comprehension questions. They tested both English and Spanish monolinguals (along with bilinguals) using parallel stimuli across the two languages, which we adopt in the experiments in this paper.[2]

Specifically, Fernández (2003) included 24 items per language, 12 with a singular RC verb (*was*) and 12 with a plural RC verb (*were*). The English and Spanish stimuli are translations of each other, so they stand as minimal pairs for attachment preferences. Example stimuli are given below.

(2) a. Andrew had dinner yesterday with the nephew of the teachers that was divorced.
b. Andrew had dinner yesterday with the nephews of the teacher that was divorced.
c. André cenó ayer con el sobrino de los maestros que estaba divorciado.
d. André cenó ayer con los sobrinos del maestro que estaba divorciado.

The underlined nominal above marks the attachment point of the relative clause (*that was divorced*). (2-a) and (2-c) exhibit HIGH attachment, while (2-b) and (2-d) exhibit LOW attachment. Fernández found that English speakers had a LOW bias, preferring (2-b) over (2-a), while Spanish speakers had a HIGH bias, preferring (2-c) over (2-d).

We ran two experiments per language,[3] one a direct simulation of the experiment from Fernández (2003) and the other an extension (EXTENDED DATA), using a larger set of experimental stimuli. The direct simulation allowed us to compare the attachment preferences for RNN LMs to the experimental results for humans. The extension allowed us to confirm that any attachment preferences we observed were generalizable properties of these models.

Specifically, the EXTENDED DATA set of stimuli included the English and Spanish stimuli from Carreiras and Clifton Jr (1993) in addition to the stimuli from Fernández (2003), for a total of 40 sentences. Next, we assigned part-of-speech tags to the English and Spanish LM training data using TreeTagger (Schmid, 1999). We filtered the tokens to the top 40 most frequent plural nouns, generating the singular forms from TreeTagger's lemmatization. We then substituted into the test sentences all combinations of distinct nouns excluding reflexives. Then we appended a relative clause with either a singular or plural verb (*was/were* or

[2] All experimental stimuli and models used are available at https://github.com/forrestdavis/AmbiAttach

[3] The vocabulary of the models was constrained to the 50K most frequent words during training. Out-of-vocabulary nominals in the original stimuli were replaced with semantically similar nominals. In English, lid(s) to cover(s) and refill(s) to filler(s). In Spanish, sarcófago(s) to ataúd(es), recambio(s) to sustitución(es), fregadero(s) to lavabo(s), baúl(es) to caja(s), cacerola(s) to platillo(s), and bolígrafo(s) to pluma(s)

estaba/estaban).[4] Finally, each test stimulus in a pair had a LOW and HIGH attachment version for a total of 249600 sentences. An example of four sentences generated for English given the two nouns *building* and *system* is below.

(3) a. Everybody ignored the system of the buildings that was
b. Everybody ignored the systems of the building that was
c. Everybody ignored the system of the buildings that were
d. Everybody ignored the systems of the building that were

Not all combinations are semantically coherent; however, Gulordava et al. suggest that syntactic operations (e.g., subject-verb agreement) are still possible for RNN LMs with "completely meaningless" sentences (Gulordava et al., 2018, p. 2).

3.2 RNN LM Details

We analyzed long short-term memory networks (LSTMs; Hochreiter and Schmidhuber, 1997) throughout the present paper. For English, we used the English Wikipedia training data provided by Gulordava et al. (2018).[5] For Spanish, we constructed a comparable training corpus from Spanish Wikipedia following the process used by Gulordava et al. (2018). A recent dump of Spanish Wikipedia was downloaded, raw text was extracted using WikiExtractor,[6] and tokenization was done using TreeTagger. A 100-million word subset of the data was extracted, shuffled by sentence, and split into training (80%) and validation (10%) sets.[7] For LM training, we included the 50K most frequent words in the vocabulary, replacing the other tokens with '⟨UNK⟩'.

We used the best English model in Gulordava et al. (2018) and trained 4 additional models with the same architecture[8] but different random initializations. There was no established Spanish model architecture, so we took the best Romance model architecture[9] reported in Gulordava et al. (2018) and trained 5 models. All models used in this work were trained for 40 epochs with resultant mean validation perplexities and standard deviations in Table 1.

Language	μ	σ
Synthetic	4.62	0.03
English	51.83	0.96
Spanish	40.80	0.89

Table 1: Mean and standard deviation of LM validation perplexity for the synthetic models used in Section 4, the English models used in Section 5-6, and the Spanish models used in Section 7

3.3 Measures

We evaluated the RNN LMs using information-theoretic surprisal (Shannon, 1948; Hale, 2001; Levy, 2008). Surprisal is defined as the inverse log probability assigned to each word (w_i) in a sentence given the preceding context.

$$\text{surprisal}(w_i) = -\log p(w_i|w_1 ... w_{i-1})$$

The probability is calculated by applying the softmax function to an RNN's output layer. Surprisal has been correlated with human processing difficulty (Smith and Levy, 2013; Frank et al., 2015) allowing us to compare model behavior to human behavior. Each of the experiments done in this work looked at sentences that differed in the grammatical number of the nominals, repeated from Section 3.1 below.

(4) a. Andrew had dinner yesterday with the nephew of the teachers that was divorced.
b. Andrew had dinner yesterday with the nephews of the teacher that was divorced.
(from Fernández, 2003)

In (4-a) the RC verb (*was*) agrees with the HIGH nominal, while in (4-b) it agrees with the LOW nominal. As such, this minimal pair probes the interpretation bias induced by the relativizer (*that*).

We measure the surprisal of the RC verb (*was*) in both sentences of the pair. If the model has a preference for LOW attachment, then we expect that the surprisal will be smaller when the number

[4] Since the unidirectional models are tested at the RC verb, we did not need to generate the rest of the sentence after that verb.

[5] https://github.com/facebookresearch/colorlessgreenRNNs

[6] https://github.com/attardi/wikiextractor

[7] We also created a test partition (10% of our data), which we did not use in this work.

[8] The models had 2 layers, 650 hidden/embedding units, batch size 128, dropout 0.2, and an initial learning rate of 20.

[9] They focused on Italian as a Romance language. The models are the same as English except the batch size is 64.

of the final noun agrees with the number of the RC verb (e.g., surprisal (4-b) < surprisal (4-a)). Concretely, for each such pair we take the difference in surprisal of the RC verb in the case of HIGH attachment (4-a) from the surprisal of the RC verb in the case of LOW attachment (4-b). If this difference (surprisal (4-a) - surprisal (4-b)) is positive, then the LM has a LOW bias, and if the difference is negative, the LM has a HIGH bias.

4 Attachment vs. Recency

We begin with a proof of concept. It has been noted that RNN LMs have a strong recency bias (Ravfogel et al., 2019). As such, it could be possible that only one type of attachment, namely LOW attachment, is learnable. To investigate this possibility, we followed the methodology in McCoy et al. (2018) and constructed a synthetic language to control the distribution of RC attachment in two experiments. Our first experiment targeted the question: if all RC attachment is HIGH, how many RCs have to be observed in training in order for a HIGH bias to generalize to unseen data? Our second experiment targeted the question: what proportion of HIGH and LOW attachment is needed in training to learn a bias?

Our synthetic language had RC attachment sentences and filler declarative sentences. The filler sentences follow the phrase structure template given in (5-a), while RC attachment sentences follow the phrase structure template given in (5-b).

(5) a. D N (P D N) (Aux) V (D N) (P D N)
b. D N Aux V D N 'of' D N 'that' 'was/were' V

Material in parentheses was optional and so was not present in all filler stimuli. That is to say, all filler sentences had a subject (abbreviated D N) and a verb (abbreviated V), with the verb being optionally transitive and followed by a direct object (D N). The subject, object, or both could be modified by a prepositional phrase (P D N). The subject and object could be either singular or plural, with the optional auxiliary (Aux) agreeing in number with the subject. There were 30 nouns (N; 60 with plural forms), 2 auxiliaries (Aux; *was/were* and *has/had*), 1 determiner (D; *the*), 14 verbs (V), and 4 prepositions (P). An example filler sentence is given in (6-a), and an example RC sentence is given in (6-b).

(6) a. The nephew near the children was seen by the players next to the lawyer.
b. The gymnast has met the hostage of the women that was eating.

We trained RNN LMs on our synthetic language using the same parameters as the English LMs given in Section 3.2, with 120,000 unique sentences in the training corpus. The resultant RNN LMs were tested on 300 sentences with ambiguous RC attachment, and we measured the surprisal at the RC auxiliary verb (*was/were*), following the methodology given in Section 3.3.

To determine how many HIGH RCs were needed in training to learn a HIGH bias, we first constrained all the RC attachment in the training data to HIGH attachment. Then, we varied the proportion (in increments of 10 RC sentences at a time) of RC sentences to filler sentences during training. We trained 5 RNNs for each training configuration (i.e. each proportion of RCs). This experiment provided a lower bound on the number of HIGH RCs needed in the training data to overcome any RNN recency bias when all RCs exhibited HIGH attachment. When as little as 0.017% (20 sentences) of the data contained RCs with HIGH attachment, the test difference in surprisal between HIGH and LOW attachment significantly differed from zero ($p < 10^{-5}$, BayesFactor (BF) > 100),[10] with a mean difference less than zero ($\mu = -2.24$). These results indicate that the models were able to acquire a HIGH bias with only 20/120000 examples of HIGH RC attachment.

In practice, we would like LMs to learn a preference even when the training data contains a mixture of HIGH and LOW attachment. To determine the proportion of RCs that must be HIGH to learn a HIGH bias, we fixed 10% of the training data as unambiguous RC attachment. Within that 10%, we varied the proportion of HIGH and LOW attachment in 10% increments (i.e. 0% HIGH - 100% LOW, 10% HIGH - 90% LOW, etc). Once again, we trained 5 models on each training configuration and tested those models on 300 test sentences, measuring the surprisal at the RC verb. When

[10] To correct for multiple comparisons, a Bonferroni correction with $m = 6$ was used. Thus, the threshold for statistical significance was $p = 0.0083$. We also computed two-sample Bayes Factors (BF; Rouder et al., 2009) for each statistical analysis using `ttestBF` from the `BayesFactor` R package (Morey and Rouder, 2018). A Bayes Factor greater than 10 is significant evidence for the hypothesis, while one greater than 100 is highly significant.

the training data had 50-100% HIGH attachment, the models preferred HIGH attachment in all the test sentences. Conversely, when the training data had 0-40% HIGH attachment, the models preferred LOW attachment in all test sentences.

Taken together, the results from our synthetic language experiments suggest that HIGH attachment is indeed learnable by RNN LMs. In fact, an equal proportion of HIGH and LOW attachment in the training data is all that is needed for these models to acquire a general preference for HIGH attachment (contra to the recency bias reported in the literature).

5 English Experiments

We turn now to model attachment preferences in English. We trained the models using English Wikipedia. We tested the attachment preferences of the RNN LMs using the original stimuli from Fernández (2003), and using a larger set of stimuli to have a better sense of model behavior on a wider range of stimuli. For space considerations, we only report here results of the EXTENDED DATA (the larger set of stimuli), but similar results hold for the Fernández (2003) stimuli (see Supplemental Materials).

In order to compare the model results with the mean human interpretation results reported by Fernández (2003), we categorically coded the model response to each item for HIGH/LOW attachment preference. If model surprisal for LOW attachment was less than model surprisal for HIGH attachment, the attachment was coded as LOW. See Figure 1 for the comparison between RNNs and humans in English.

Statistical robustness for our RNN results was determined using the original distribution of surprisal values. Specifically, a two-tailed t-test was conducted to see if the mean difference in surprisal differed from zero (i.e. the model has some attachment bias). This revealed a highly significant ($p < 10^{-5}$, BF > 100) mean difference in surprisal of 0.77. This positive difference indicates that the RNN LMs have a consistent LOW bias, similar to English readers, across models trained with differing random seeds.

There are two possible reasons for this patterning: (1) the models have learned a human-like LOW bias, or (2) the models have a recency bias that favors attachment to the lower nominal. These two hypotheses have overlapping predictions in English. The second hypothesis is perhaps weakened by the results of Section 4, where both attachment types were learnable despite any recency bias. However, we know that other syntactic attachment biases can influence RC attachment in humans (Scheepers, 2003). It could be that other kinds of attachment (such as prepositional phrase attachment) have varying proportions of attachment biases in the training data. Perhaps conflicting attachment biases across multiple constructions force the model to resort to the use of a 'default' recency bias in cases of ambiguity.

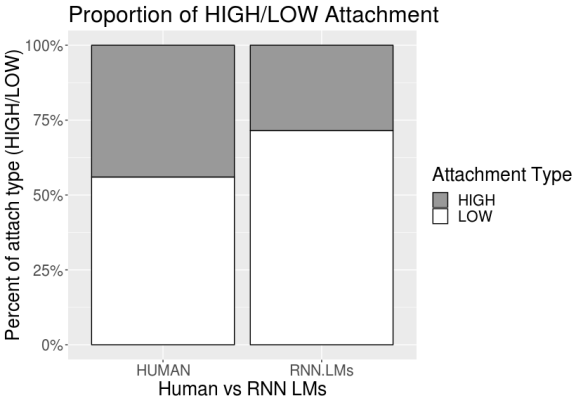

Figure 1: Proportion HIGH vs LOW attachment in English. Human results from the original Fernández (2003) experiment and RNN LM results from EXTENDED DATA (derived from Fernández (2003) and Carreiras and Clifton Jr (1993)).

6 Syntactically blocking low attachment

6.1 Stimuli

To determine whether the behavior of the RNNs is driven by a learned attachment preference or a strong recency bias, we created stimuli[11] using the stimulus template described in Section 3.1 (e.g., (3)). All of these stimuli had only the higher nominal syntactically available for attachment; the lower nominal was blocked by the addition of a relative clause:

(7) a. Everybody ignored the boy that the girls hated that was boring.
b. *Everybody ignored the boys that the girl hated that was boring.

In (7) only (7-a) is grammatical. This follows because *boy(s)* is the only nominal available for mod-

[11] As before, some of these stimuli are infelicitous. We do not concern ourselves with this distinction in the present work, given the results in Gulordava et al. (2018).

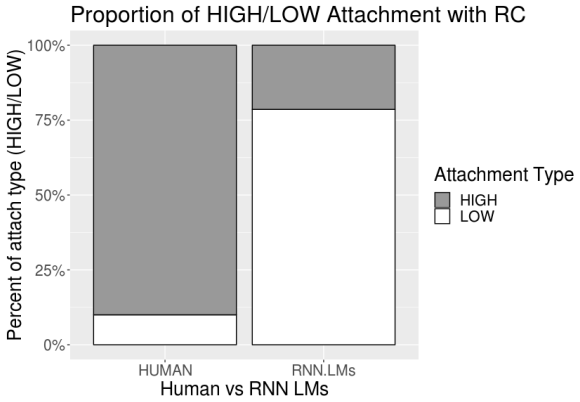

Figure 2: Proportion HIGH vs LOW attachment with syntactically unavailable lower nominal. Human results estimated from Linzen and Leonard (2018) and RNN LM results from the EXTENDED DATA (derived from Fernández (2003) and Carreiras and Clifton Jr (1993)) with the lower nominal blocked.

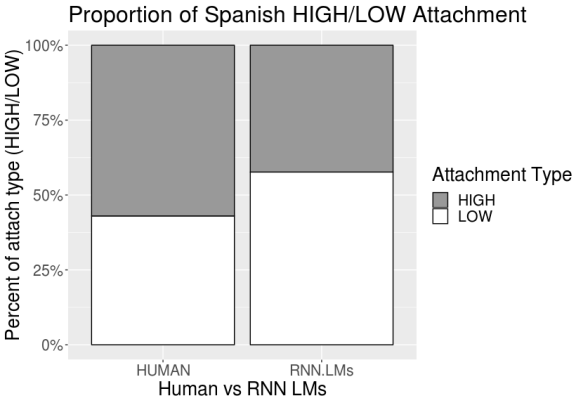

Figure 3: Proportion HIGH vs LOW attachment in Spanish. Human results from the original Fernández (2003) experiment and RNN LM results from the EXTENDED DATA (derived from Fernández (2003) and Carreiras and Clifton Jr (1993)).

ification. In (7-a), the RC verb *was* agrees in number with this nominal, while in (7-b), *was* agrees in number with the now blocked lower nominal *girl* rather than with *boys*. For all such sentence pairs, we calculated the difference in surprisal between (7-a) and (7-b). If their behavior is driven by a legitimate syntactic attachment preference, the models should exhibit an overwhelming HIGH bias (i.e. the mean difference should be less than zero).

6.2 Results

As before, the differences in surprisal were calculated for each pair of experimental items. If the difference was greater than zero, the attachment was coded as LOW. The results categorically coded for HIGH/LOW attachment are given in Figure 2, including the results expected for humans given the pattern in Linzen and Leonard (2018).[12] A two-tailed t-test was conducted to see if the mean difference in surprisal differed from zero. The results were statistically significant ($p < 10^{-5}$, BF > 100). The mean difference in surprisal was 1.15, however, suggesting that the models still had a LOW bias when the lower nominal was syntactically unavailable for attachment. This is in stark contrast to what one would expect if these models had learned the relationship between syntactic constituents and relative clause attachment. A possible alternative to the recency bias explanation is that RNN LMs might learn that there is a general LOW attachment bias in English and overgeneralize this pattern even in cases where one of the nominals is syntactically unavailable.

7 The case of default HIGH bias: Spanish

Our English analyses suggest that RNN LMs either learn a general English LOW attachment preference that they apply in all contexts, or that they have a 'default' recency bias that prevents them from learning HIGH attachment preferences with more complex, naturalistic training data. In the case of the former, we would expect that models trained on a language whose speakers generally prefer HIGH attachment should be able to learn HIGH attachment. Spanish has a well-attested HIGH bias in humans (Carreiras and Clifton Jr, 1993; Carreiras and Clifton, 1999; Fernández, 2003) offering a way to distinguish between competing recency bias and over-generalization accounts. That is, if the models can learn a HIGH bias when trained on Spanish data, we should be able to conclude that the general LOW bias in English is being overgeneralized by the RNNs to corner cases where HIGH bias should be preferred.

7.1 Results

As before, the differences in surprisal were calculated for each pair of experimental items. If the difference was greater than zero, the attachment was coded as LOW. Two sample t-tests were conducted to see if the mean difference in surprisal differed

[12] Linzen and Leonard (2018) conducted experiments probing the agreement errors for subject-verb agreement with intervening RCs (and prepositional phrases). Our work is concerned with agreement between an object and its modifying RC. As such, their task serves as an approximate estimate of the errors we would expect for humans.

significantly from zero for both the direct simulation of Fernández (2003) and the EXTENDED DATA that included the stimuli derived from Carreiras and Clifton Jr (1993). The results categorically coded for HIGH/LOW attachment for the extended stimulus set are given in Figure 3, alongside the human results reported in Fernández (2003).

For the direct simulation, the mean did not differ significantly from 0 (BF $< 1/3$). This suggests that there is no attachment bias for the Spanish models for the stimuli from Fernández (2003), contrary to the human results. For the extended set of stimuli, the results were significant ($p < 10^{-5}$, BF > 100) with a mean difference greater than zero ($\mu = 0.211$). Thus, rather than a HIGH bias, as we would expect, the RNN LMs once again had a LOW bias.

8 Discussion

In this work, we explored the ability of RNN LMs to prioritize multiple simultaneous valid interpretations in a human-like way (as in *John met the student of the teacher that was happy*). While both LOW attachment (i.e. *the teacher was happy*) and HIGH attachment (i.e. *the student was happy*) are equally semantically plausible without a disambiguating context, humans have interpretation preferences for one attachment over the other (e.g., English speakers prefer LOW attachment and Spanish speakers prefer HIGH attachment). Given the recent body of literature suggesting that RNN LMs have learned abstract syntactic representations, we tested the hypothesis that these models acquire human-like attachment preferences. We found that they do not.

We first used a synthetic language experiment to demonstrate that RNN LMs are capable of learning a HIGH bias when HIGH attachment is at least as frequent as LOW attachment in the training data. These results suggest that any recency bias in RNN LMs is weak enough to be easily overcome by sufficient evidence of HIGH attachment. In English, the RNNs exhibited a human-like LOW bias, but this preference persisted even in cases where LOW attachment was ungrammatical. To test whether the RNNs were over-learning a general LOW bias of English, we tested whether Spanish RNNs learned the general HIGH bias in that language. Once again, RNN LMs favored LOW attachment over HIGH attachment. The inability of RNN LMs to learn the Spanish HIGH attachment preference suggests that the Spanish data may not contain enough HIGH examples to learn human-like attachment preferences.

In post-hoc analyses of the Spanish Wikipedia training corpus and the AnCora Spanish newswire corpus (Taulé et al., 2008), we find a consistent production bias towards LOW attachment among the RCs with unambiguous attachment. In Spanish Wikipedia, LOW attachment is 69% more frequent than HIGH attachment, and in Spanish newswire data, LOW attachment is 21% more frequent than HIGH attachment.[13] This distributional bias in favor of LOW attachment does not rule out a subsequent HIGH RC bias in the models. It has been established in the psycholinguistic literature that attachment is learned by humans as a general abstract feature of language (see Scheepers, 2003). In other words, human syntactic representations of attachment overlap, with prepositional attachment influencing relative clause attachment, etc. These relationships could coalesce during training and result in an attachment preference that differs from any one structure individually. However, it is clear that whatever attachment biases exist in the data are insufficient for RNNs to learn a human-like attachment preference in Spanish. This provides compelling evidence that standard training data itself may systematically lack aspects of syntax relevant to performing linguistic comprehension tasks.

We suspect that there are deep systematic issues leading to this mismatch between the expected distribution of human attachment preferences and the actual distribution of attachment in the Spanish training corpus. Experimental findings from psycholinguistics suggest that this issue could follow from a more general mismatch between language production and language comprehension. In particular, Kehler and Rohde (2015, 2018) have provided empirical evidence that the production and comprehension of these structures are guided by different biases in humans. Production is guided by syntactic and information structural considerations (e.g., topic), while comprehension is influenced by those considerations plus pragmatic and discourse factors (e.g., coherence relations). As such, the biases in language production are a proper subset of those of language comprehension. As it stands now, RNN LMs are typically trained on production data

[13] https://github.com/UniversalDependencies/UD_Spanish-AnCora

(that is, the produced text in Wikipedia).[14] Thus, they will have access to only a subset of the biases needed to learn human-like attachment preferences. In its strongest form, this hypothesis suggests that no amount of production data (i.e. raw text) will ever be sufficient for these models to generalizably pattern like humans during comprehension tasks.

The mismatch between human interpretation biases and production biases suggested by this work invalidates the tacit assumption in much of the natural language processing literature that standard, production-based training data (e.g., web text) are representative of the linguistic biases needed for natural language understanding and generation. There are phenomena, like agreement, that seem to have robust manifestations in a production signal, but the present work demonstrates that there are others, like attachment preferences, that do not. We speculate that the difference may lie in the inherent ambiguity in attachment, while agreement explicitly disambiguates a relation between two syntactic units. This discrepancy is likely the reason that simply adding more data doesn't improve model quality (e.g., van Schijndel et al., 2019; Bisk et al., 2020). Future work needs to be done to understand more fully what biases are present in the data and learned by language models.

Although our work raises questions about mismatches between human syntactic knowledge and the linguistic representations acquired by neural language models, it also shows that researchers can fruitfully use sentences with multiple interpretations to probe the linguistic representations acquired by those models. Before now, evaluations have focused on cases of unambiguous grammaticality (i.e. ungrammatical vs. grammatical). By using stimuli with multiple simultaneous valid interpretations, we found that evaluating models on single-interpretation sentences overestimates their ability to comprehend abstract syntax.

Acknowledgments

We would like to thank members of the NLP group and the C.Psyd lab at Cornell University, and the Altmann and Yee labs at University of Connecticut, who gave feedback on an earlier form of this work. We would also like to thank the three anonymous reviewers and Yonatan Belinkov. Special thanks go to Dorit Abusch and John Whitman for invaluable suggestions and feedback, and Laure Thompson for comments on an earlier draft.

[14] Some limited work has explored training models with human comprehension data with positive results (Klerke et al., 2016; Barrett et al., 2018).

References

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 302–312, Brussels, Belgium. Association for Computational Linguistics.

Jean-Philippe Bernardy and Shalom Lappin. 2017. Using deep neural networks to learn syntactic agreement. *Linguistic Issues in Language Technology (LiLT)*, 15.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. *arXiv preprint arXiv:2004.10151*.

Marc Brysbaert and Don C Mitchell. 1996. Modifier attachment in sentence parsing: Evidence from dutch. *The Quarterly Journal of Experimental Psychology Section A*, 49(3):664–695.

Manuel Carreiras and Charles Clifton. 1999. Another word on parsing relative clauses: Eyetracking evidence from Spanish and English. *Memory & Cognition*, 27(5):826–833.

Manuel Carreiras and Charles Clifton Jr. 1993. Relative clause interpretation preferences in Spanish and English. *Language and Speech*, 36(4):353–372.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Chris Dyer, Gábor Melis, and Phil Blunsom. 2019. A critical analysis of biased parsers in unsupervised parsing. *arXiv preprint arXiv:1909.09428*.

Émile Enguehard, Yoav Goldberg, and Tal Linzen. 2017. Exploring the syntactic abilities of RNNs with multi-task learning. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 3–14. Association for Computational Linguistics.

Eva M. Fernández. 2003. *Bilingual sentence processing: Relative clause attachment in English and Spanish*. John Benjamins Publishing, Amsteradam.

Stefan L Frank and John Hoeks. 2019. The interaction between structure and meaning in sentence comprehension: Recurrent neural networks and reading times. *PsyArXiv preprint:10.31234*.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain & Language*, 140:1–11.

Lyn Frazier and Charles Clifton. 1996. *Construal*. MIT Press, Cambridge, Mass.

Richard Futrell and Roger Levy. 2019. Do RNNs learn human-like abstract word order preferences? In *Proceedings of the Society for Computation in Linguistics*, volume 2, pages 50–59.

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

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

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

Andrew Kehler and Hannah Rohde. 2015. Pronominal reference and pragmatic enrichment: A bayesian account. In *CogSci*.

Andrew Kehler and Hannah Rohde. 2018. Prominence and coherence in a bayesian theory of pronoun interpretation. *Journal of Pragmatics*.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533, San Diego, California. 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*, 41:1202–1241.

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

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

Tal Linzen and Brian Leonard. 2018. Distinct patterns of syntactic agreement errors in recurrent networks and humans. In *Proceedings of the 2018 Annual Meeting of the Cognitive Science Society*, pages 690–695. Cognitive Science Society.

R Thomas McCoy, Robert Frank, and Tal Linzen. 2018. Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Richard D. Morey and Jeffrey N. Rouder. 2018. *BayesFactor: Computation of Bayes Factors for Common Designs*. R package version 0.9.12-4.2.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. 2019. Using priming to uncover the organization of syntactic representations in neural language models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. 2019. Studying the inductive biases of RNNs with synthetic variations of natural languages. In *Proceedings of NAACL-HLT*.

Jeffrey N. Rouder, Paul L. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson. 2009. Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2):225–237.

Christoph Scheepers. 2003. Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. *Cognition*, 89(3):179–205.

Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to German. In *Natural language processing using very large corpora*, pages 13–25. Springer.

Claude Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for catalan and spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.

Andrew Trask, Felix Hill, Scott E Reed, Jack Rae, Chris Dyer, and Phil Blunsom. 2018. Neural arithmetic logic units. In *Advances in Neural Information Processing Systems*, pages 8035–8044.

Marten van Schijndel and Tal Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2018. What Syntactic Structures block Dependencies in RNN Language Models? In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*.

Ethan Wilcox, Roger Levy, and Richard Futrell. 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*.

A Fernández (2003) Replications

A.1 English

We compute RNN surprisal for each experimental item from Fernández (2003) as detailed in Section 3.3 in the paper. The results coded for HIGH/LOW attachment are given in Figure 4, including the results for humans reported by Fernández (2003). While these categorical results enable easier comparison to the human results reported in the literature, statistical robustness was determined using the original distribution of surprisal values. Specifically, a two-tailed t-test was conducted to see if the mean difference in surprisal differed from zero (i.e. the model has some attachment bias). The result is highly significant ($p < 10^{-5}$, Bayes Factor (BF) > 100) with a mean surprisal difference of $\mu = 0.66$. This positive difference suggests that the RNN LMs have a LOW bias, similar to English readers.

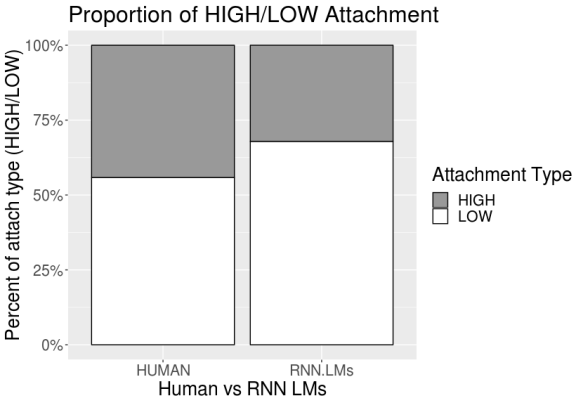

Figure 4: Proportion HIGH vs LOW attachment in English. Human results from the original Fernández (2003) experiment and RNN LM results from the stimuli from Fernández (2003).

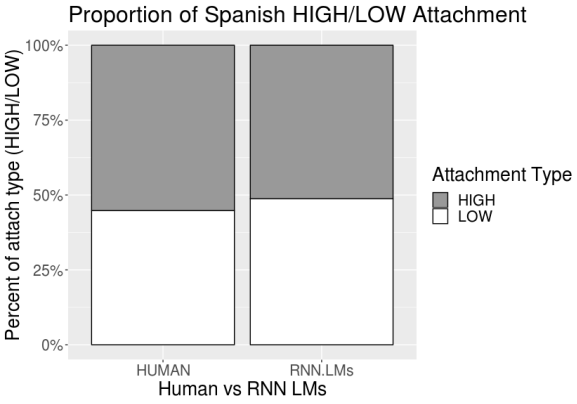

Figure 5: Proportion HIGH vs LOW attachment in Spanish. Human results from the original Fernández (2003) experiment and RNN LM results from the stimuli from Fernández (2003).

A.2 Spanish

The results coded for HIGH/LOW attachment for the Spanish replication are given in Figure 5, including the human results reported by Fernández (2003). The mean did not differ significantly from 0 (BF $< 1/3$). This suggests that there is no attachment bias for the Spanish models for the stimuli from Fernández (2003), contrary to the human results.