Representation of Constituents in Neural Language Models: Coordination Phrase as a Case Study

Aixiu An[1] **Peng Qian**[2] **Ethan Wilcox**[3] **Roger Levy**[2]
[1] Université de Paris, LLF, CNRS, aixiu.an@etu.univ-paris-diderot.fr
[2] Department of Brain and Cognitive Sciences, MIT, {pqian, rplevy}@mit.edu
[3] Department of Linguistics, Harvard University, wilcoxeg@g.harvard.edu

Abstract

Neural language models have achieved state-of-the-art performances on many NLP tasks, and recently have been shown to learn a number of hierarchically-sensitive syntactic dependencies between individual words. However, equally important for language processing is the ability to combine words into phrasal constituents, and use constituent-level features to drive downstream expectations. Here we investigate neural models' ability to represent constituent-level features, using coordinated noun phrases as a case study. We assess whether different neural language models trained on English and French represent phrase-level number and gender features, and use those features to drive downstream expectations. Our results suggest that models use a linear combination of NP constituent number to drive CoordNP/verb number agreement. This behavior is highly regular and even sensitive to local syntactic context, however it differs crucially from observed human behavior. Models have less success with gender agreement. Models trained on large corpora perform best, and there is no obvious advantage for models trained using explicit syntactic supervision.

1 Introduction

Humans deploy structure-sensitive expectations to guide processing during natural language comprehension (Levy, 2008). While it has been shown that neural language models show similar structure-sensitivity in their predictions about upcoming material (Linzen et al., 2016; Futrell et al., 2018), previous work has focused on dependencies that are conditioned by features attached to a single word, such as subject number (Gulordava et al., 2018; Marvin and Linzen, 2018) or wh-question words (Wilcox et al., 2018). There has been no systematic investigation into models' ability to compute phrase-level features—features that are attached to a set of words—and whether models can deploy these more abstract properties to drive downstream expectations.

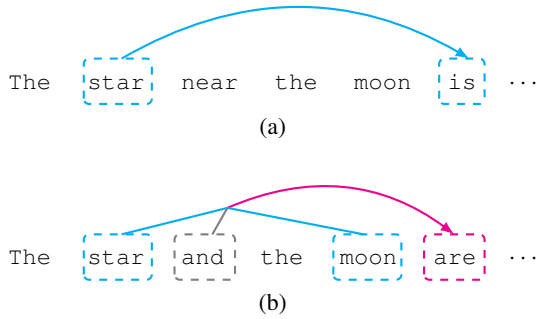


Figure 1: Subject-verb agreement with (a) the head of a noun phrase structure, and (b) the coordination structure.

In this work, we assess whether state-of-the-art neural models can compute and employ phrase-level gender and number features of coordinated subject Noun Phrases (CoordNPs) with two nouns. Typical syntactic phrases are ENDOCENTRIC: they are HEADED by a single child, whose features determine the agreement requirements for the entire phrase. In Figure 1a, for example, the word *star* heads the subject NP *The star*; since *star* is singular, the verb must be singular. CoordNPs lack endocentricity: neither conjunct NP solely determines the features of the NP as a whole. Instead, these feature values are determined by compositional rules sensitive to the features of the conjuncts and the identity of the coordinator. In Figure 1b, because the coordinator is *and*, the subject NP number is plural even though both conjuncts (*the star* and *the moon*) are singular. As this case demonstrates, the agreement behavior for CoordNPs must be driven by more abstract, constituent-level representations, and cannot be reduced to features hosted on a single lexical item.

We use four suites of experiments to assess

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 2888–2899, Hong Kong, China, November 3–7, 2019.

whether neural models are able to build up phrase-level representations of CoordNPs on the fly and deploy them to drive humanlike behavior. First, we present a simple control experiment to show that models can represent number and gender features of non-coordinate NPs **(Non-coordination Agreement)**. Second, we show that models modulate their expectations for downstream verb number based on the CoordNP's coordinating conjunction combined with the features of the coordinated nouns **(Simple Coordination)**. We rule out the possibility that models are using simple heuristics by designing a set of stimuli where a simple heuristic would fail due to structural ambiguity **(Complex Coordination)**. The striking success for all models in this experiment indicates that even neural models with no explicit hierarchical bias, trained on a relatively small amount of text are able to learn fine-grained and robust generalizations about the interaction between CoordNPs and local syntactic context. Finally, we use subject–auxiliary inversion to test whether an upstream lexical item modulates model expectation for the phrasal-level features of a downstream CoordNP **(Inverted Coordination)**. Here, we find that all models are insensitive to the fine-grained features of this particular syntactic context. Overall, our results indicate that neural models can learn fine-grained information about the interaction of Coordinated NPs and local syntactic context, but their behavior remains unhumanlike in many key respects.

2 Methods

2.1 Psycholinguistics Paradigm

To determine whether state-of-the-art neural architectures are capable of learning humanlike CoordNP/verb agreement properties, we adopt the psycholinguistics paradigm for model assessment. In this paradigm the models are tested using handcrafted sentences designed to test underlying network knowledge. The assumption here is that if a model implicitly learns humanlike linguistic knowledge during training, its expectations for upcoming words should qualitatively match human expectations in novel contexts. For example, Linzen et al. (2016) and Kuncoro et al. (2016) assessed how well neural models had learned the subject/verb number agreement by feeding them with the prefix *The keys to the cabinet ....* If the models predicted the grammatical continuation *are* over the ungrammatical continuation *is*, they can be said to have learned the number agreement insofar as the number of the head noun and not the number of the distractor noun, *cabinet*, drives expectations about the number of the matrix verb.

	Model	Training data	# tokens
English	LSTM (PTB)	Penn Treebank	$\sim$ 1M
	ActionLSTM (PTB)	Penn Treebank	$\sim$ 1M
	RNNG (PTB)	Penn Treebank	$\sim$ 1M
	LSTM (enWiki)	English Wikipedia	$\sim$ 90M
	LSTM (1B)	1 Billion Word	$\sim$ 800M
French	LSTM (FTB)	French Teebank	$\sim$ 0.5M
	ActionLSTM (FTB)	French Teebank	$\sim$ 0.5M
	RNNG (FTB)	French Teebank	$\sim$ 0.5M
	LSTM (frWaC)	Subset of frWaC	$\sim$ 138M

Table 1: A summary of models tested.

If models are able to robustly modulate their expectations based on the internal components of the CoordNP, this will provide evidence that the networks are building up a context-sensitive phrase-level representation. We quantify model expectations as SURPRISAL VALUES. Surprisal is the negative log-conditional probability $S(x_i) = -\log_2 p(x_i|x_1 \dots x_{i-1})$ of a sentence's i^{th} word x_i given the previous words. Surprisal tells us how strongly x_i is expected in context and is known to correlate with human processing difficulty (Hale, 2001; Levy, 2008; Smith and Levy, 2013). In the CoordNP/Verb agreement studies presented here, cases where the proceeding context sets high expectation for a number-inflected verb form w_i, (e.g. singular 'is') we would expect $S(w_i)$ to be lower than its number-mismatched counterpart (e.g. plural 'are').

2.2 Models Tested

Recurrent Neural Network (RNN) Language Models are trained to output the probability distribution of the upcoming word given a context, without explicitly representing the structure of the context (Sundermeyer et al., 2012; Elman, 1990). We trained two two-layer recurrent neural language models with long short-term memory architecture (Hochreiter and Schmidhuber, 1997) on a relatively small corpus. The first model, referred as '**LSTM (PTB)**' in the following sections, was trained on the sentences from Penn Treebank (Marcus and Marcinkiewicz). The second model, referred as '**LSTM (FTB)**', was trained on the sentences from French Treebank (Abeillé et al., 2003). We set the size of input word embedding and LSTM hidden layer of both models as 256.

We also compare LSTM language models

trained on large corpora. We incorporate two pre-trained English language models: one trained on the Billion Word benchmark (referred as '**LSTM (1B)**') from Jozefowicz et al. (2016), and the other trained on English Wikipedia (referred as '**LSTM (enWiki)**') from Gulordava et al. (2018). For French, we trained a large LSTM language model (referred as '**LSTM (frWaC)**') on a random subset (about 4 million sentences, 138 million word tokens) of the frWaC dataset (Baroni et al., 2009). We set the size of the input embeddings and hidden layers to 400 for the LSTM (frWaC) model since it is trained on a large dataset.

ActionLSTM models the linearized bracketed tree structure of a sentence by learning to predict the next action required to construct a phrase-structure parse (Choe and Charniak, 2016). The action space consists of three possibilities: open a new non-terminal node and opening bracket; generate a terminal node; and close a bracket. To compute surprisal values for a given token, we approximate $P(w_i|w_{1\cdots i-1})$ by marginalizing over the most-likely partial parses found by word-synchronous beam search (Stern et al., 2017).

Generative Recurrent Neural Network Grammars (RNNG) jointly model the word sequence as well as the underlying syntactic structure (Dyer et al., 2016). Following Hale et al. (2018), we estimate surprisal using word-synchronous beam search (Stern et al., 2017). We use the same hyper-parameter settings as Dyer et al. (2016).

The annotation schemes used to train the syntactically-supervised models differ slightly between French and English. In the PTB (English) CoordNPs are flat structures bearing an 'NP' label. In FTB (French), CoordNPs are binary-branching, labeled as NPs, except for the phrasal node dominating the coordinating conjunction, which is labeled 'COORD'. We examine the effects of annotation schemes on model performance in Appendix A. [1]

3 Experiment 1: Non-coordination Agreement

In order to provide a baseline for following experiments, here we assess whether the models tested have learned basic representations of number and gender features for non-coordinated Noun Phrases. We test number agreement in English and French as well as gender agreement in French. Both English and French have two grammatical number feature: SINGULAR (sg) and PLURAL (pl). French has two grammatical gender features: MASCULINE (m) and FEMININE (f).

Condition	Sentence
Npl	The windows is/are
Nsg	The window is/are

Table 2: Conditions of number agreement in Non-coordination Agreement experiment.

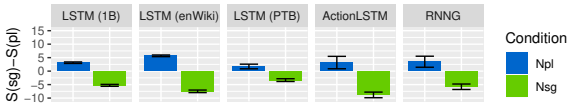
Condition	Sentence
Nm	Les coûts sont importants/importantes the cost.MPL are important.MPL/FPL
Nf	Les dépenses sont importants/importantes the expense.FPL are important.MPL/FPL

Table 3: Conditions of gender agreement in Non-coordination Agreement experiment.

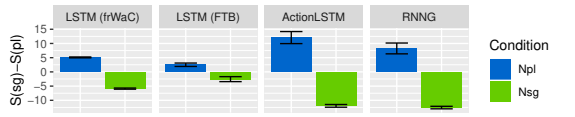
The experimental materials include sentences where the subject NPs contain a single noun which can either match with the matrix verb (in the case of number agreement) or a following predicative adjective (in the case of gender agreement). Conditions are given in Table 2 and Table 3. We measure model behavior by computing the *plural expectation*, or the surprisal of the singular continuation minus the surprisal of the plural continuation for each condition and took the average for each condition. We expect a positive *plural expectation* in the *Npl* conditions and a negative *plural expectation* in the *Nsg* conditions. For gender expectation we compute a *gender expectation*, which is S(feminine continuation) − S(masculine continuation). We measure surprisal at the verbs and predicative adjectives themselves.

The results for this experiment are in Figure 2, with the *plural expectation* and *gender expectation* on the y-axis and conditions on the x-axis. For this and subsequent experiments error bars represent 95% confidence intervals for across-item means. For number agreement, all the models in English and French show positive plural expectation when the head noun is plural and negative plural expectation when it is singular. For gender agreement, however, only the LSTM (frWaC) shows modulation of gender expectation based on the gender of the head noun. This is most likely due to the lower

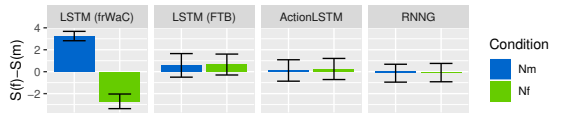
[1] The materials and code for this project can be found in https://github.com/cpllab/rnn_psycholing_coordination.git



(a) English number agreement



(b) French number agreement



(c) French gender agreement

Figure 2: Non-Coordination Agreement experiments for English (number) and French (number and gender).

frequency of predicative adjectives compared to matrix verbs in the corpus.

4 Experiment 2: Simple Coordination

In this section, we test whether neural language models can use grammatical features hosted on multiple components of a coordination phrase—the coordinated nouns as well as the coordinating conjunction—to drive downstream expectations. We test number agreement in both English and French and gender agreement in French.

4.1 Number Agreement

In simple subject/verb number agreement, the number features of the CoordNP are determined by the coordinating conjunction and the number features of the two coordinated NPs. CoordNPs formed by *and* are plural and thus require plural verbs; CoordNPs formed by *or* allow either plural or singular verbs, often with the number features of the noun linearly closest to the verb playing a more important role, although this varies cross-linguistically (Fowler et al., 1992). Forced-choice preference experiments in Keung and Staub (2018) reveal that English native speakers prefer singular agreement when the closest conjunct in an *or*-CoordNP is singular and plural agreement when the closest conjunct is plural. In French, both singular and plural verbs are possible when two singular NPs are joined via disjunction (Goosse and Grevisse, 2016).

In order to assess whether the neural models learn the basic CoordNP licensing for English, we

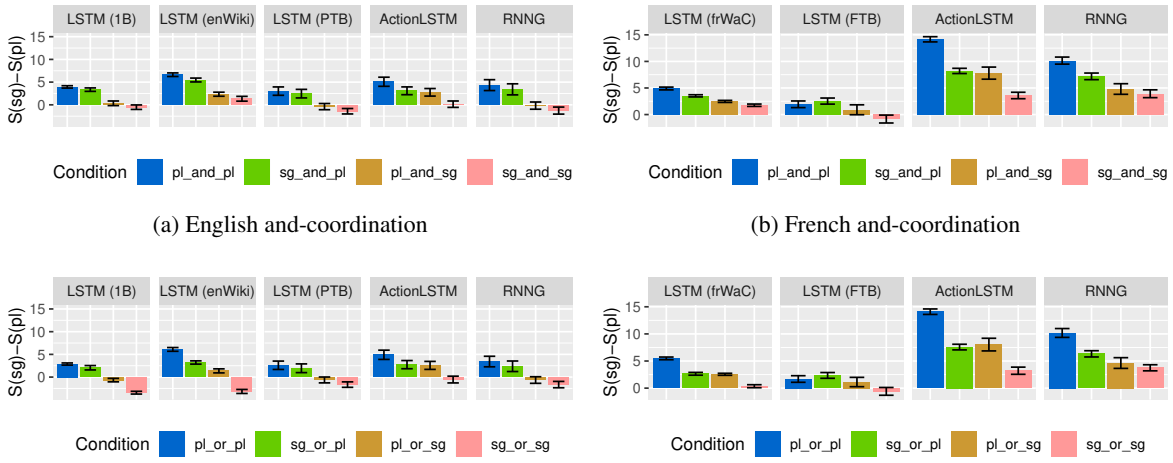
Condition	Sentence
pl_and_pl	The doors and the windows is/are
sg_and_pl	The door and the windows is/are
pl_and_sg	The doors and the window is/are
sg_and_sg	The door and the window is/are
pl_or_pl	The doors or the windows is/are
sg_or_pl	The door or the windows is/are
pl_or_sg	The doors or the window is/are
sg_or_sg	The door or the window is/are

Table 4: Conditions of number agreement in Simple Coordination experiment.

adapted 37 items from Keung and Staub (2018), along the 16 conditions outlined in Table 4. Test items consist of the sentence preamble, followed by either the singular or plural *BE* verb, half the time in present tense (*is/are*) and half the time in past tense *(was/were)*. We measured the plural expectation, following the procedure in Section 3. We created 24 items using the same conditions as the English experiment to test the models trained in French, using the 3^{rd} person singular and plural form of verb *aller*, 'to go' (*va, vont*). Within each item, nouns match in gender; across all conditions half the nouns are masculine, half feminine.

The results for this experiment can be seen in Figure 3, with the results for English on the left and French on the right. The results for *and* are on the top row, *or* on the bottom row. For all figures the y-axis shows the plural expectation, or the difference in surprisal between the *singular* condition and the *plural* condition. Turning first to **English-*and*** (Figure 3a), all models show plural expectation (the bars are significantly greater than zero) in the *pl_and_pl* and *sg_and_pl* conditions, as expected. For the *pl_and_sg* condition, only the LSTM (enWiki) and ActionLSTM are greater than zero, indicating humanlike behavior. For the *sg_and_sg* condition, only the LSTM (enWiki) model shows the correct plural expectation. For the **French-*and*** (Figure 3b), all models show positive plural expectation in all conditions, as expected, except for the LSTM (FTB) in the *sg_and_sg* condition.

Examining the results for **English-*or***, we find that all models demonstrate humanlike expectation in the *pl_or_pl* and *sg_or_pl* conditions. The LSTM (1B), LSTM (PTB), and RNNG models show zero or negative singular expectation for the *pl_or_sg* conditions, as expected. However the LSTM (enWiki) and ActionLSTM models show positive plural expectation in this condition, indi-



(a) English and-coordination

(b) French and-coordination

(c) English or-coordination

(d) French or-coordination

Figure 3: Comparison of models' expectation preferences for singular vs. plural predicate in English and French Simple Coordination experiments.

cating that they have not learned the humanlike generalizations. All models show significantly negative plural expectation in the *sg_or_sg* condition, as expected. In the **French-*or*** cases, models show almost identical behavior to the *and* conditions, except the LSTM (frWaC) shows smaller plural expectation when singular nouns are linearly proximal to the verb.

These results indicate moderate success at learning coordinate NP agreement, however this success may be the result of an overly simple heuristic. It appears that expectation for both plural and masculine continuations are driven by a linear combination of the two nominal number/gender features transferred into log-probability space, with the earlier noun mattering less than the later noun. A model that optimally captures human grammatical preferences should show no or only slight difference across conditions in the surprisal differential for the *and* conditions, and be greater than zero in all cases. Yet, all the models tested show gradient performance based on the number of plural conjuncts.

4.2 Gender Agreement

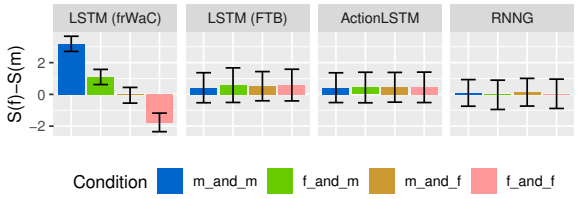
In French, if two nouns are coordinated with *et* (*and*-coordination), agreement must be masculine if there is one masculine element in the coordinate structure. If the nouns are coordinated with *ou* (*or*-coordination), both masculine and feminine agreement is acceptable (Corbett, 1991; Wechsler and Zlatić, 2003). Although linear proximity effects have been tested for a number of languages that employ grammatical gender, as in e.g. Slavic languages (Willer et al., 2018), there is no systematic study for French.

Condition	Sentence
m_and_m	Les prix et les coûts sont importants/importantes the price.MPL and the cost.MPL are important.MPL/FPL
f_and_m	Les recettes et les coûts sont importants/importantes the revenues.FPL and the cost.MPL are important.MPL/FPL
m_and_f	Les prix et les dépenses sont importants/importantes the price.MPL and the expense.FPL are important.MPL/FPL
f_and_f	Les recettes et les dépenses sont importants/importantes the revenues.FPL and the expense.FPL are important.MPL/FPL

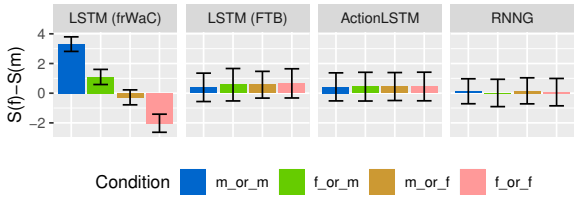
Table 5: Conditions for the and-coordination experiment. (Items for or-coordination are the same except that we change the coordinator to *ou*.)

To assess whether the French neural models learned humanlike gender agreement, we created 24 test items, following the examples in Table 5, and measured the masculine expectation. In our test items, the coordinated subject NP is followed by a predicative adjective, which either takes on masculine or feminine gender morphology.

Results from the experiment can be seen in Figure 4. No models shows qualitative difference based on the coordinator, and only the LSTM (frWaC) shows significant behavior difference between conditions. Here, we find positive masculine expectation in the *m_and_m* and *f_and_m* conditions, and negative masculine expectation in the *f_and_f* condition, as expected. However, in the *m_and_f* condition, the masculine expectation is not significantly different from zero, where we would expect it to be positive. In the *or-*



(a) French and-coordination



(b) French or-coordination

Figure 4: Comparison of models' expectation preferences for Feminine v.s. Masculine predicative adjectives in French.

coordination conditions, following our expectation, masculine expectation is positive when both conjuncts are masculine and negative when both are feminine. For the LSTM (FTB) and ActionLSTM models, the masculine expectation is positive (although not significantly so) in all conditions, consistent with results in Section 3.

5 Experiment 3: Complex Coordination

One possible explanation for the results presented in the previous section is that the models are using a 'bag of features' approach to plural and masculine licensing that is opaque to syntactic context: Following a coordinating conjunction surrounded by nouns, models simply expect the following verb to be plural, proportionally to the number of plural nouns.

In this section, we control for this potential confound by conducting two experiments: In the *Complex Coordination Control* experiments we assess models' ability to extend basic CoordNP licensing into sententially-embedded environments, where the CoordNP can serve as an embedded subject. In the *Complex Coordination Critical* experiments, we leverage the sentential embedding environment to demonstrate that when the CoordNPs cannot plausibly serve as the subject of the embedded phrase, models are able to suppress the previously-demonstrated expectations set up by these phrases. These results demonstrate that models are not following a simple strategy for predicting downstream number and gender features, but are building up CoordNP representations on the fly, conditioned on the local syntactic context.

5.1 Complex Coordination Control

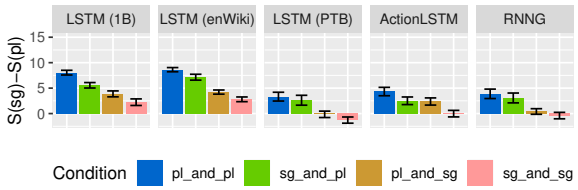
Following certain sentential-embedding verbs, CoordNPs serve unambiguously as the subject of the verb's sentence complement and should trigger number agreement behavior in the main verb of the embedded clause, similar to the behavior presented in 4.1. To assess this, we use the 37 test items in English and 24 items in French in section 4.1, following the conditions in Table 6 (for number agreement), testing only *and* coordination. For gender agreement, we use the same test items and conditions for *and* coordination in Section 4.2, but with the Coordinated NPs embedded in a context similar to (1). As before, we derived the plural expectation by measuring the difference in surprisal between the singular and plural continuations and the gender expectation by computing the difference in surprisal between the masculine and feminine predicates.

(1) Je croyais que les prix et les
I thought that the.PL price.MPL and the.PL
dépenses étaient importants/importantes.
expense.FPL were important.MPL/FPL
I thought that the prices and the expenses were important.

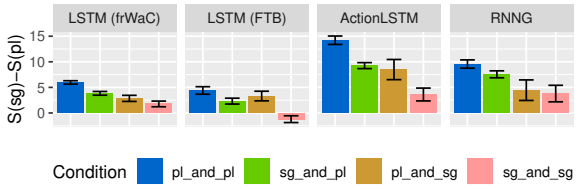
Condition	Sentence
pl_and_pl	I think that the doors and the windows is/are
sg_and_pl	I think that the door and the windows is/are
pl_and_sg	I think that the doors and the window is/are
sg_and_sg	I think that the door and the window is/are

Table 6: Conditions of number agreement in Complex Coordination Control experiment.

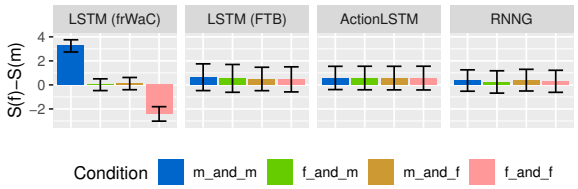
The results for the control experiments can be seen in Figure 5, with English number agreement on the top row, French number agreement in the middle row and French gender agreement on the bottom. The y-axis shows either plural or masculine expectation, with the various conditions along the x-axis. For English number agreement, we find that the models behave similarly as they do for simple coordination contexts. All models show significant plural expectation when the closest noun is plural, with only two models demonstrating plural expectation in the *sg_and_sg* case. The French number agreement tests show simi-



(a) English number agreement

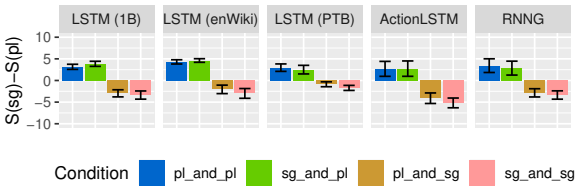


(b) French number agreement

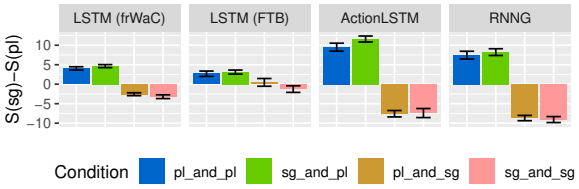


(c) French gender agreement

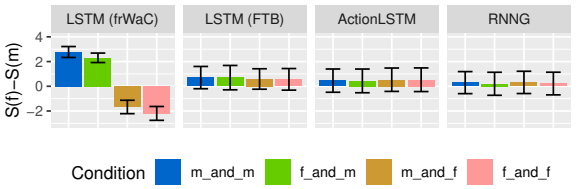
Figure 5: Comparison of model's expectation preferences in the Complex Coordination Control experiments.



(a) English number agreement



(b) French number agreement



(c) French gender agreement

Figure 6: Comparison of model's expectation preferences in the Complex Coordination Critical experiments.

lar results, with all models except LSTM (FTB) demonstrating significant plural prediction in all cases. Turning to French gender agreement, only the LSTM (frWaC) shows sensitivity to the various conditions, with positive masculine expectation in the *m_and_m* condition and negative expectation in the *f_and_f* condition, as expected. These results indicate that the behavior shown in Section 4.1 extends to more complex syntactic environments—in this case to sentential embeddings. Interestingly, for some models, such as the LSTM (1B), behavior is *more humanlike* when the CoordNP serves as the subject of an embedded sentence. This may be because the model, which has a large number of hidden states and may be extra sensitive to fine-grained syntactic information carried on lexical items (Futrell et al., 2018), is using the complementizer, *that*, to drive more robust expectations.

5.2 Complex Coordination Critical

In order to assess whether the models' strategy for CoordNP/verb number agreement is sensitive to syntactic context, we contrast the results presented above to those from a second, critical experiment. Here, two coordinated nouns follow a verb that cannot take a sentential complement, as in the examples given in Table 7. Of the two possible continuations—*are* or *is*—the plural is only grammatically licensed when the second of the two conjuncts is plural. In these cases, the plural continuation may lead to a final sentence where the first noun serves as the verb's object and the second introduces a second main clause coordinated with the first, as in *I fixed the doors and the windows are still broken.* For the same reason, the singular-verb continuation is only licensed when the noun immediately following *and* is singular.

We created 37 test items in both English and French, and calculated the plural expectation. If the models were following a simple strategy to drive CoordNP/verb number agreement, then we should see either no difference in plural expectation across the four conditions or behavior no different from the control experiment. If, however, the models are sensitive to the licensing context, we should see a contrast based solely on the number features of the second conjunct, where plural expectation is positive when the second conjunct is plural, and negative otherwise.

Condition	Sentence
pl_and_pl	I fixed the doors and the windows is/are
sg_and_pl	I fixed the door and the windows is/are
pl_and_sg	I fixed the doors and the window is/are
sg_and_sg	I fixed the door and the window is/are

Table 7: Conditions of number agreement in Complex Coordination Critical experiment.

Condition	Sentence preamble
Vpl_Npl	What are the doors and
Vpl_Nsg	What are the door and
Vsg_Nsg	What is the door and

Table 8: Conditions in Inverted Coordination experiment.

Experimental items for a critical gender test were created similarly, as in Example (2). As with plural agreement, gender expectation should be driven solely by the second conjunct: For the *f_and_m* and *m_and_m* conditions, the only grammatical continuation is one where the adjectival predicate bears masculine gender morphology. Conversely, for the *m_and_f* or *f_and_f* conditions, the only grammatical continuation is one where the adjectival predicate bears feminine morphology. As in 4.1, we created 24 test items and measured the gender expectation by calculating the difference in surprisal between the masculine and feminine continuations.

(2) Nous avons accepté les prix et les
we have accepted the.PL price.MPL and the
dépenses étaient importants/importantes.
expense.FPL were important.MPL/FPL
We have accepted the prices and the expenses were important.

The results from the critical experiments are in Figure 6, with the English number agreement on the top row, French number agreement in the middle and gender expectation on the bottom row. Here the y-axis shows either plural expectation or masculine expectation, with the various conditions are on the x-axis. The results here are strikingly different from those in the control experiments. For number agreement, all models in both languages show strong plural expectation in conditions where the second noun is plural (blue and green bars), as they do in the control experiments. Crucially, when the second noun is singular, the plural expectation is significantly negative for all models (save for the French LSTM (FTB) *pl_and_sg* condition). Turning to gender agreement, only the LSTM (frWaC) model shows differentiation between the four conditions tested. However, whereas the *f_and_m* and *m_and_f* gender expectations are not significantly different from zero in the control condition, in the critical condition they pattern with the purely masculine and purely feminine conditions, indicating that, in this syntactic context, the model has successfully learned to base gender expectation solely off of the second noun.

These results are inconsistent with a simple 'bag of features' strategy that is insensitive to local syntactic context. They indicate that both models can interpret the same string as either a coordinated noun phrase, or as an NP object and the start of a coordinated VP with the second NP as its subject.

6 Experiment 4: Inverted Coordination

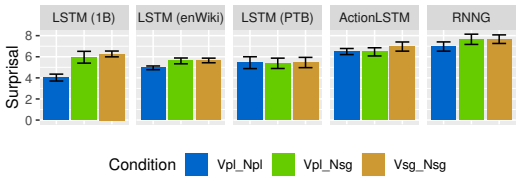
In addition to using phrase-level features to drive expectation about downstream lexical items, human processors can do the inverse—use lexical features to drive expectations about upcoming syntactic chunks. In this experiment, we assess whether neural models use number features hosted on a verb to modulate their expectations for upcoming CoordNPs.

To assess whether neural language models learn inverted coordination rules, we adapted items from Section 4.1 in both English (37 items) and French (24 items), following the paradigm in Table 8. The first part of the phrase contains either a plural or singular verb and a plural or singular noun. In this case, we sample the surprisal for the continuations *and* (*or* is grammatical in all conditions, so it is omitted from this study). Our expectation is that 'and' is less surprising in the *Vpl_Nsg* condition than in the *Vsg_Nsg* condition, where a CoordNP is not licensed by the grammar in either French or English (as in **What is the pig and the cat eating?*). We also expect lower surprisal for *and* in the *Vpl_Nsg* condition, where it is obligatory for a grammatical continuation, than in the *Vpl_Npl* condition, where it is optional.

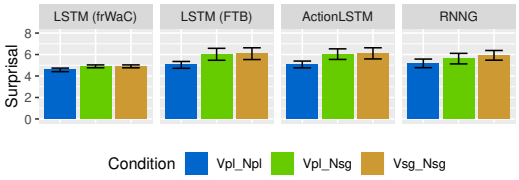
For French experimental items, the question is embedded into a sentential-complement taking verb, following Example (3), due to the fact that unembedded subject-verb inverted questions sound very formal and might be relatively rare in the training data.

(3) Je me demande où vont le
I myself ask where go.3PL the.MSG
maire et
mayor.MSG and

The results for both languages are shown in Figure 7, with the surprisal at the coordinator on the y-axis and the various conditions on the x-axis. No model in either language shows a signficant difference in surprisal between the *Vpl_Nsg* and *Vpl_Npl* conditions or between the *Vpl_Nsg* and *Vsg_Nsg* conditions. The LSTM (1B) shows significant difference between the *Vpl_Nsg* and *Vpl_Npl* conditions, but in the opposite direction than expected, with the coordinator less surprising in the latter condition. These results indicate that the models are unable to use the fine-grained context sensitivity to drive expectations for CoordNPs, at least in the inversion setting.



(a) English



(b) French

Figure 7: Comparison of models' surprisals of *and*-coordination in Inverted Coordination experiment.

7 Discussion

The experiments presented here extend and refine a line of research investigating what linguistic knowledge is acquired by neural language models. Previous studies have demonstrated that sequential models trained on a simple regime of optimizing the next word can learn long-distance syntactic dependencies in impressive detail. Our results provide complimentary insights, demonstrating that a range of model architectures trained on a variety of datasets can learn fine-grained information about the interaction of CoordNPs and local syntactic context, but their behavior remains unhumanlike in many key ways. Furthermore, to our best knowledge, this work presents the first psycholinguistic analysis of neural language models trained on French, a high-resource language that has so far been under-investigated in this line of research.

In the **simple coordination** experiment, we demonstrated that models were able to capture some of the agreement behaviors of humans, although their performance deviated in crucial aspects. Whereas human behavior is best modeled as a 'percolation' process, the neural models appear to be using a linear combination of NP constituent number to drive CoordNP/verb number agreement, with the second noun weighted more heavily than the first. In these experiments, supervision afforded by the RNNG and ActionLSTM models did not translate into more robust or humanlike learning outcomes. The **complex coordination** experiments provided evidence that the neural models tested were not using a simple 'bag of features' strategy, but were sensitive to syntactic context. All models tested were able to interpret material that had similar surface form in ways that corresponded to two different tree-structural descriptions, based on local context. The **inverted coordination** experiment provided a contrasting example, in which models were unable to modulate expectations based on subtleties in the syntactic environment.

Across all our experiments, the French models performed consistently better on subject/verb number agreement than on subject/predicate gender agreement. Although there are likely more examples of subject/verb number agreement in the French training data, gender agreement is syntactically mandated and widespread in French. It remains an open question why all but one of the models tested were unable to leverage the numerous examples of gender agreement seen in various contexts during training to drive correct subject/predicate expectations.

Acknowledgments

This project is supported by a grant of Labex EFL ANR-10-LABX-0083 (and Idex ANR-18-IDEX-0001) for AA and MIT–IBM AI Laboratory and the MIT–SenseTimeAlliance on Artificial Intelligence for RPL. We would like to thank the anonymous reviewers for their comments and Anne Abeillé for her advice and feedback.

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for french. In *Treebanks*, pages 165–187. Springer.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Do Kook Choe and Eugene Charniak. 2016. Parsing as language modeling. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2331–2336, Austin, Texas. Association for Computational Linguistics.

Corbett. 1991. *Gender*. Cambridge University Press.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A Smith. 2016. Recurrent neural network grammars. In *Proceedings of NAACL-HLT*, pages 199–209.

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

Jessica Ficler and Yoav Goldberg. 2016. Coordination annotation extension in the penn tree bank. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 834–842.

Henry Ramsey Fowler, Jane E Aaron, and Murray McArthur. 1992. *The little, brown handbook*. HarperCollins.

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

André Goosse and Maurice Grevisse. 2016. *Le bon usage*. De Boeck Superieur.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

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.

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. 2018. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2727–2736.

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*.

Lap-Ching Keung and Adrian Staub. 2018. Variable agreement with coordinate subjects is not a form of agreement attraction. *Journal of Memory and Language*, 103:1–18.

Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, Graham Neubig, and Noah A Smith. 2016. What do recurrent neural network grammars learn about syntax? *arXiv preprint arXiv:1611.05774*.

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.

Mitchell P Marcus and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2).

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.

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

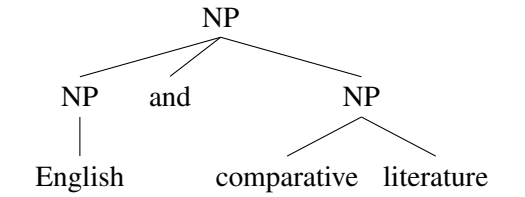
Mitchell Stern, Daniel Fried, and Dan Klein. 2017. Effective inference for generative neural parsing. *arXiv preprint arXiv:1707.08976*.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

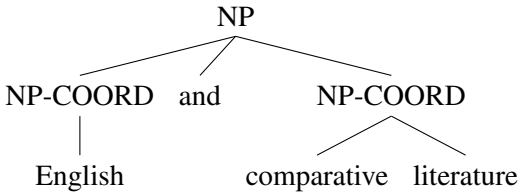
Stephen Wechsler and Larisa Zlatić. 2003. *The many faces of agreement*. Stanford: CSLI Publications.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221.

J Gold Willer, Boban Arsenijević, Mia Batinić, Michael Becker, Nermina Čordalija, Marijana Kresić, Nedžad Leko, Franc Lanko Marušič, Tanja Milićev, Nataša Milićević, et al. 2018. When linearity prevails over hierarchy in syntax. *Proceedings of the National Academy of Sciences of the United States of America*, 115(3):495–500.



(a) PTB-control



(b) PTB-coord

Figure 8: Comparison of annotation schemes of coordination structure.

A The Effect of Annotation Schemes

This section further investigates the effects of CoordNP annotation schemes on the behaviors of structurally-supervised models. We test whether an explicit COORD phrasal tag improves model performance. We trained two additional RNNG models on 38,546 sentences from the Penn Treebank annotated with two different schemes: The first, **RNNG (PTB-control)** was trained with the original Penn Treebank annotation. The second, **RNNG (PTB-coord)**, was trained on the same sentences, but with an extended coordination annotation scheme, meant to employ the scheme employed in the FTB, adapted from Ficler and Goldberg (2016). We stripped empty categories from their scheme and only kept the NP-COORD label for constituents inside a coordination structure. Figure 8 illustrates the detailed annotation differences between two datasets. We tested both models on all the experiments presented in Sections 3-6 above.

Turning to the results of these six experiments: We see little difference between the two models in the *Non-coordination agreement* experiment. For the *Complex coordination control* and *Complex coordination critical* experiments, both models are largely the same as well. However, in the *Simple and-coordination* and *Simple or-coordination* experiments the values for all conditions are shifted upwards for the RNNG PTB-coord model, indicating higher over-all preference for the plural continuation. Furthermore, the range of values is reduced in the RNNG PTB-coord model, compared to the RNNG PTB-control model. These results indicate that adding an explicit COORD phrasal label does not drastically change model performance: Both models still appear to be using a linear combination of number features to drive plural vs. singular expectation. However, the explicit representation has made the interior of the coordination phrase more opaque to the model (each feature matters less) and has slightly shifted model preference towards plural continuations. In this sense, the PTB-coord model may have learned a generalization about CoordNPs, but this generalization remains unlike the ones learned by humans.

	PTB			FTB		
Condition	sg	pl	total	sg	pl	total
pl_and_pl	0	67	67	1	116	116
sg_and_pl	0	72	72	0	50	50
pl_and_sg	0	11	11	0	30	30
sg_and_sg	7	213	220	5	288	293
pl_or_pl	0	2	2	0	14	14
sg_or_pl	0	0	0	0	0	0
pl_or_sg	0	1	1	0	1	1
sg_or_sg	5	1	6	5	8	13

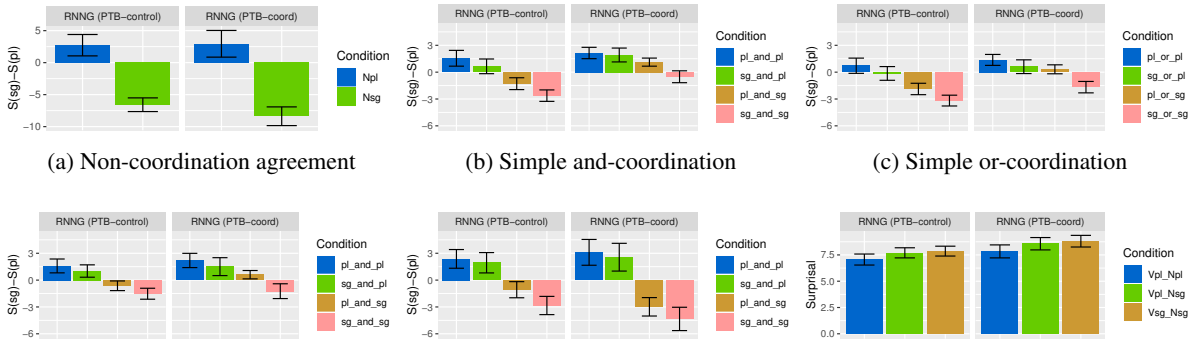
Table 9: Frequency of number agreement patterns in PTB and FTB.

Condition	m	f	total
m_and_m	38	0	38
m_and_f	10	1	11
f_and_m	9	0	9
f_and_f	0	16	16
m_or_m	1	0	1
m_or_f	0	0	0
f_or_m	1	0	1
f_or_f	0	1	1

Table 10: Frequency of gender agreement patterns in FTB.

B PTB/FTB Agreement Patterns

We present statistics of subject/predicate agreement patterns in the Penn Treebank (PTB) and French Treebank (FTB) in Table 9 and 10.



(a) Non-coordination agreement (b) Simple and-coordination (c) Simple or-coordination

(d) Complex coordination control (e) Complex coordination critical (f) Inverted coordination

Figure 9: Comparison between RNNGs trained on PTB data with original annotation vs. fine-grained annotation of coordination structure.