Neural Language Models Capture Some, But Not All, Agreement Attraction Effects

Suhas Arehalli
Johns Hopkins University

Tal Linzen
Johns Hopkins University

Abstract

The number of the subject in English must match the number of the corresponding verb (*dog runs* but *dogs run*). Yet in real-time language production and comprehension, speakers often mistakenly compute agreement between the verb and a grammatically irrelevant non-subject noun phrase instead. This phenomenon, referred to as *agreement attraction*, is modulated by a wide range of factors; any complete computational model of grammatical planning and comprehension would be expected to derive this rich empirical picture. Recent developments in Natural Language Processing have shown that neural networks trained only on word-prediction over large corpora are capable of capturing subject-verb agreement dependencies to a significant extent, but with occasional errors. The goal of this paper is to evaluate the potential of such neural word prediction models as a foundation for a cognitive model of real-time grammatical processing. We simulate six experiments taken from the agreement attraction literature with LSTMs, one common type of neural language model. The LSTMs captured the critical human behavior in three of them, indicating that (1) some agreement attraction phenomena can be captured by a generic sequence processing model, but (2) capturing the other phenomena may require models with more language-specific mechanisms.

Keywords: psycholinguistics; computational modeling; agreement attraction; neural language models;

Introduction

In most varieties of English, subjects and their corresponding verbs must share a number feature: *The dog runs*, but *The dogs run*. This constraint, subject-verb agreement, holds regardless of what other noun phrases appear within the sentence. However, real-time human comprehension and production does not always follow this grammatical constraint. Bock and Miller (1991) found that when a noun phrase with a different number feature than the subject (called an **attractor**) appears before the verb position, speakers prompted to produce a verb sometimes produce verbs that agree in number with the attractor rather than the subject itself (*The keys to the cabinet **is*** rather than the grammatically correct *The keys to the cabinet **are***). This phenomenon, called agreement attraction, has been demonstrated to be robust, appearing in both production and comprehension (Bock & Miller, 1991; Pearlmutter et al., 1999; Wagers et al., 2009), and in similar agreement constructions across languages.

The size of the agreement attraction effect has been shown to be sensitive to a variety of syntactic (Bock & Cutting, 1992, *inter alia*) and semantic (Humphreys & Bock, 2005, *inter alia*) factors. A number of theories have been proposed to account for various subsets of results: The Marking & Morphing model (Eberhard, Cutting, & Bock, 2005), feature percolation accounts (Franck, Vigliocco, & Nicol, 2002), and memory retrieval accounts (Wagers et al., 2009, etc). However, none of these theories provide a comprehensive accounts of all of the empirical results.

Recent work in natural language processing has demonstrated that neural language models, particularly LSTM models, were capable of capturing subject-verb agreement (Linzen et al., 2016; Gulordava et al., 2018), and appear to show some human-like attraction errors (Linzen & Leonard, 2018). Unlike prior psycholinguistic models, these language models are broad-coverage models trained over large corpora to predict the following word in a sequence. When combined with a method to link the model's predictions to measures of human behavior, these models can be used to generate predictions for any set of experimental materials.

In this paper, we will use LSTM language models to simulate results from six experiments in the agreement attraction literature: Attractors in prepositional phrases generate a stronger attraction effect than those in relative clauses (Bock & Cutting, 1992); Attractors closer to the verb have a stronger attraction effect, measured in both syntactic (Franck et al., 2002) and linear (Haskell & Macdonald, 2005) distance; Collective subjects with distributive readings having higher rates of plural agreement than those with collective readings (Humphreys & Bock, 2005); Attractors in oblique arguments create a larger attraction effect than those in core arguments (Parker & An, 2018); Attractors outside of the clause where agreement is computed cause attraction effects, and attraction effects make ungrammatical sentences seem grammatical, but do not make grammatical sentences seem ungrammatical (Wagers et al., 2009).

We find that LSTMs are able to capture the critical human behavior in three of them, indicating that (1) some agreement attraction phenomena can be captured by a simple sequence model without any built-in language-specific mechanisms, but (2) capturing the other phenomena may require models with such mechanisms.

Methods

Models

Language models take as input a sequence of words, and probabilistically predict the next word. The neural language models we use in this paper operate by generating a vector representation of the input sequence using an recurrent neural network architecture, and predicting the next word using a softmax classifier over the model's vocabulary. In our simulations, we will use a language model based on the Long Short-Term Memory (LSTM, Hochreiter & Schmidhuber, 1997) architecture, a standard architecture in language technologies. We trained five such models with different random initializations and orderings of training examples. Following Gulordava et al. (2018), each model was a 2-layer LSTM with 650 hidden units in each layer, trained for 12 epochs over 80 million words extracted from English Wikipedia. The models all achieved perplexities between 55.16 and 55.31 over the development set (an additional 10 million words from English Wikipedia). The models were trained with the code publicly provided by the authors.[1]

Linking Language Models to Human Behavior

Given the first $k-1$ words of a sentence, each language model outputs a probability distribution over the k-th word in the sentence. For the six experiments we wish to model, we must select a linking function to transform this probability distribution into some measure comparable to the human data reported in the original study. The behavioral data elicited in the experiments is of two types: reading times from a self-paced reading task and the proportion of ungrammatical verbs produced from a sentence completion task. Prior work has demonstrated that surprisal is a effective linking hypothesis between probabilistic models of language and reading times (Hale, 2001, *inter alia*). Here, we will report surprisals averaged over all words in the sentence.

In prior work, Linzen et al. (2016) evaluated language models on experimental data generated from natural text, and compared the probability of the singular and plural form of the verb that appeared in the original text. In this paradigm, the model is evaluated as though it had produced the verb form with the higher probability.

For our simulations, we will use a slightly modified version of the paradigm of Linzen et al. (2016). For each input, we we will compare a singular and plural form of the verb *be*, and have the model probabilistically select the singular and plural form in proportion to the probabilities the it assigns to those forms. The simulations will be run over a relatively small number of experimental items for each experiment, and some of the effects we seek to simulate are subtle, so this modification allows for the simulations to be more sensitive to variation in the probability distribution provided by the model.

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

Statistical Analysis

For each statistical analysis presented below, we first constructed a mixed-effects model with the relevant fixed effects and a maximal random effects structure. If the model did not converge, we incrementally pruned the random effects until convergence was reached. For all mixed models discussed below, this process resulted in random intercepts for each item and each of the 5 LSTM models.

For studies where the response variable is surprisal, we used linear mixed-effects regressions as our analysis technique. For studies where the response variable is the probability, we used beta regression, which assumes the response variable lies between 0 and 1. To test the significance of each relevant fixed effect, we report the results of the corresponding Wald test on the resulting model.

Simulation Results

Syntactic Effects

Attractors in PPs vs. RCs Bock and Cutting (1992) asked whether the syntactic environment containing the attractor affected the strength of the attraction effect, particularly whether an attractor within a subordinate clause could attract as much as one in the matrix clause. To do this, they conducted a sentence completion task where participants were provided preambles containing a subject with a post-nominal modifier and were asked to repeat the preamble and finish the sentence. Each sentence was varied along three dimensions: whether the subject was singular or plural, whether the number of the attractor matched that of the subject, and whether the post-nominal modifier was a prepositional phrase (PP) or relative clause (RC) (see Ex. 1 and 2).

(1) The demo tape(s) from the popular rock singer(s)...

(2) The demo tape(s) that promoted the rock singer(s)...

The two critical findings we wish to simulate are (i) a number asymmetry—plural attractors had a much stronger attraction effect than singular attractors—and (ii) a stronger attraction effect from attractors in PPs than those in RCs. The human results taken from Bock and Cutting (1992) can be seen in Fig. 1a.

Results from the LSTM simulation can be seen in Fig. 1b. A beta mixed-effects regression found that the attraction effect was significantly stronger in PPs than in RCs in the LSTMs ($|z| = 2.86$, $p < 0.005$), as well as stronger when the attractors are plural rather than singular ($|z| = 5.81$, $p < 0.001$). Both results are consistent with the human results. Note that our simulation results differ from those of Linzen and Leonard (2018), who found the opposite error pattern for the PP/RC comparison. This may be due to a difference in task: Linzen and Leonard (2018) trained their models to directly predict number features, while the models evaluated here are trained on language modeling. Language modeling demands that successful models represent a more varied set of syntactic features (predicting a verb both inside and outside an RC, for example), as well as provides more

training examples per sentence (one example per word rather than one per sentence), both of which may contribute to the differences between results.

Syntactic vs. Linear Distance Franck et al. (2002) tested whether the probability of an agreement attraction error is more strongly affected by syntactic or linear distance between the attractor and the verb. Using the same sentence completion task as Bock and Cutting (1992), they provided participants with sentences with two potential attractors in prepositional phrases within the subject. Crucially, the second prepositional phrase modified the first attractor, causing it to simultaneously be syntactically further from the verb position (as it is more deeply embedded) but linearly closer in the surface form (see Ex. 3). They constructed eight versions of each item, with all possible combinations of the number of the subject and the two attractors. They found that the syntactically closer attractor had a stronger attraction effect than the linearly closer one, indicating that the mechanism underlying attraction in humans is sensitive to the hierarchical structure of the sentence. Their results can be seen in Fig. 2a.

(3) The threat(s) [$_{PP}$ to the president(s) [$_{PP}$ of the company(s)]]...

Results from the LSTM simulation can be seen in Fig. 2b. A beta mixed effects regression found there were significant attraction effects for the intermediate (LSTM: $\beta = 0.34$, $|z| = 7.48$, $p < 0.001$) and local (LSTM: $\beta = 1.14$, $|z| = 21.71$, $p < 0.001$) attractors. Here, the model predictions run counter to the human results: The effect of linear distance is stronger than that of hierarchical distance.

Linear Distance in Coordination Haskell and Macdonald (2005) attempted to measure the effect of linear distance on agreement attraction, controlling for effects of syntactic distance. Again using a sentence completion paradigm, they achieve this by coordinating two disjuncts that differ in number:

(4) Can you ask Brenda if the boy or the girls...

(5) Can you ask Brenda if the boys or the girl...

They then measured the rate of plural agreement with the subject between the two disjunct orderings: Under the assumption that the two disjuncts are the same syntactic distance away from the verb, any difference in plural agreement rates must be due to the linear ordering of the disjuncts. They found that participants did produce plurals at a higher rate when the plural disjunct was closer to the position at which the verb is to be produced than when it was further, indicating that linear distance does have an effect independent of syntactic distance. The human results can be seen in Fig. 3a.

Results from the LSTM simulation can be seen in Fig. 3b. A significant effect of order was found in the beta mixed effects regression ($|z| = 4.10$, $p < 0.001$), consistent with the linear order effect observed in our simulation of Franck et al. (2002).

Semantic Effects

Notional Number and Distributivity While the previous experiments we simulated focused on manipulating the grammatical number of attractors, Humphreys and Bock (2005) tests the effect of **notional number** on agreement, particularly whether attaining a distributive or collective reading of noun phrase influences agreement. A noun phrase's notional number reflects how the phrase's referent is conceptualized: For instance, in Ex. 6, the presence of the preposition *on* are expected to bias participants toward a distributive reading (each gang member on their own motorcycle), and thus create a notionally plural NP, whereas in Ex. 7, participants should be biased toward a collective reading (a group of gang members next to group of motorcycles) and singular notional number:

(6) The gang on the motorcycles...

(7) The gang near the motorcycles...

Humphreys and Bock (2005) tested the effect of this notional number manipulation using a sentence completion paradigm. They found that participants were more likely to produce plural agreement in trials with the distributive-biasing preposition than in those with collective-biasing prepositions (see Fig. 4a).

Results from the LSTM simulation can be found in Fig. 4b. A beta mixed-effects regression found no significant effect of preposition choice. This can indicate one of two things: A failure to understand the lexical semantics of *on* and *near* (mapping the words to the collective and distributive readings), or a failure in understanding the influence of notional number on agreement (mapping collective and distributive readings to differing rates of plural agreement). An evaluation scheme targeted directly at the model's semantics could help distinguish between these hypotheses in future work.

Agreement Attraction in Comprehension

Argument Status Attractors can appear in both core arguments, which are required for the interpretation of the verb (*...sat **the girls***), and oblique arguments, which are not (*...sat near **the girls***). Core arguments, due to their importance to interpretation, have been argued to be encoded more carefully than their less-critical oblique counterparts (Van Dyke & McElree, 2011). Parker and An (2018) predict that this would lead to stronger attraction effects from attractors in oblique arguments, as the poor encoding may cause the attractor to be more easily confused with the subject. Parker and An (2018) tested this hypothesis in a self-paced reading study using materials such as the following:

(8) The waitress who sat the girl(s) unsurprisingly was/were...

(9) The waitress who sat near the girl(s) unsurprisingly was/were...

Participants read one of eight versions of each item, four grammatical and four ungrammatical. The attraction effect

can thus be realized in two ways: as a facilitatory effect in ungrammatical sentences (i.e., a speed-up in the ungrammatical sentences with an attractor) or as an inhibitory effect in grammatical sentences (i.e., a slowdown in grammatical sentences with an attractor). In both cases, the difference between the grammatical and ungrammatical reading times—the slowdown associated with ungrammaticality—should be smaller when the sentence contains an attractor whose number does not match the subject. This is the measure of the attraction effect we aim to simulate. Parker and An (2018) found that this effect was smaller when the attractor appeared in a core argument like in Ex. 8 than when it appeared in an oblique argument as in Ex. 9. Human results are shown in Fig. 5a.

Simulation results from the LSTMs can be found in Fig. 5b. A linear mixed-effects regression did not find a significant interaction between attractor number and the argument status of the phrase containing the attractor NP. The failure to simulate this effect again leads us to two possible explanations: Either the model fails to encode argument status, or it fails to use that encoding when computing agreement. Just as with Humphreys and Bock (2005), future work could help distinguish between these two possibilities. In particular, one could see if a linear classifier could distinguish between core and oblique arguments based on the representations the models construct. If so, we can conclude that the models encode argument status.

Clause-External Attractors and the Grammaticality Asymmetry

Bock and Cutting (1992) found that attractors that appeared in a subordinate clause (*...The key [that was in **the cabinet(s)**]...*) caused weaker attraction effects than those that appeared in the same clause as the verb. One might predict that the same effect might happen in reverse: An attractor in the matrix clause would have little effect on agreement occurring within a subordinate clause (*...**The key(s)** [the cabinet holds]...*). Wagers et al. (2009) found a clause-external attraction effect in humans, and argued that the fact that the attractor is not part of of the relevant subject NP in these kinds of sentences causes significant trouble for some accounts of attraction in humans. They used materials such as 10, where the head of the matrix clause subject (*musician(s)*) acts as the attractor. Each item was varied in the number of the RC subject, the attractor, and the verb.

(10) The musician(s) [who the reviewer(s) praise(s) so highly] will probably win a Grammy.

Wagers et al. (2009) additionally found that while there were significant facilitatory effects of attraction in ungrammatical sentences (ungrammatical sentences with attractors that didn't match the number of the subject were read faster than those with attractors that did match the number of the subject), there were no inhibitory effects on grammatical sentences (grammatical sentences were read at the same speed regardless of the number of the attractor; see Fig. 6a). They refer to this effect as the **grammaticality asymmetry**. We will discuss the implications of this asymmetry in the discussion.

Simulation results can be found in Fig. 6b. Significant attraction effects were found in a linear mixed effects analysis of simulation results ($|t| = 9.88,\ p < 0.001$), indicating that a clause-external noun phrase can act as an attractor. In addition, a significant interaction was found between grammaticality and subject-attractor match conditions (such that the attraction effect was significantly larger for ungrammatical sentences than grammatical; $|t| = 4.14,\ p < 0.001$), patterning like the grammaticality asymmetry. However, despite a difference in effect size, significant attraction effects were found in in both ungrammatical and grammatical sentences (Ungrammatical: $|t| = 9.88$, $p < 0.001$; Grammatical: $|t| = 4.07$, $p < 0.001$).

Discussion

We have simulated six experiments from the agreement attraction literature using LSTM language models. Our simulations replicated a number of the findings from the human experiments, but failed to replicate others. Space constraints preclude detailed discussion of the relationship of each of these results to theoretical debates; instead, we will focus on two particularly salient aspects of our findings.

Syntactic Position of the Attractor The LSTMs showed greater attraction effects when the attractor was in a prepositional phrase (PP) than in a relative clause (RC), matching the human pattern (Bock & Cutting, 1992). At the same time, whatever syntactic distance effects they exhibited were weaker than the effects of linear distance, as evidenced by (1) the failure to simulate the human pattern in which syntactically close, but linearly distant attractors resulted in greater attraction effects than the reverse (Franck et al., 2002), and (2) the stronger attraction effect of linearly closer attractors that are matched for syntactic distance (Haskell & Macdonald, 2005). This pattern of results is in line with the Clause Packaging Hypothesis (Bock & Cutting, 1992), under which attraction effects are weakened when the attractor is separated from the verb by a clausal boundary. While this factor crucially differentiates the PPs and RCs in Bock and Cutting (1992), it does not distinguish attractors in the doubly-embedded PP structures of Franck et al. (2002). This more coarse-grained syntactic factor, the number of clausal boundaries crossed, combined with linear distance effects, explain the results of the simulation. To derive the human pattern of results, language models with greater sensitivity to hierarchical structure may be necessary: the processing of the LSTM allows the model to encode some syntactic information (clausal boundaries), but not representations sufficient to match human processing behaviors (i.e., the effect of syntactic distance observed by Franck et al., 2002).

Grammaticality asymmetry As in human studies, our models showed a significant interaction between the grammaticality of a sentence (whether the subject and its verb match in number) and the attraction effect (whether the attractor matches or does not match the subject number). Specifically, the facilitatory attraction effect (the mismatched attractor causing ungrammatical sentences to less surprising) was stronger than the inhibitory effect (the mismatched attractor causing grammatical sentences to be more surprising). According to Wagers et al. (2009), this asymmetry demonstrates that attraction effects in humans must be explained through retrieval mechanisms, where attraction emerges from a failure to retrieve the correct number feature from memory, rather than through encoding mechanisms, where attraction emerges from an incorrect encoding of the subject's number feature in memory. This is because an encoding failure can only result in a symmetric attraction effect: For example, such an account must make the same predictions for the number feature of the subject *The key to the cabinets* regardless of whether it is followed by *is* or *are*. Thus, if a fallible encoding procedure caused this subject to be marked as plural some percentage of the time, this incorrect marking would lead readers to misjudge *The key to the cabinets is...* as ungrammatical just as often as they would misjudge *The key to the cabinets are...* as grammatical. Prior work has shown that LSTMs seem to use an maintain an encoding of the subject's number to process agreement (Lakretz et al., 2019); the fact that this asymmetry emerged from what appears to be an encoding model of agreement is therefore particularly intriguing and should motivate future work.

Conclusion

The simulations we have presented serve two purposes. First, they allow us to benchmark an off-the-shelf statistical sequence learner on its ability to capture agreement attraction. Phenomena captured by the LSTMs can be characterized as emerging from domain-general sequence processing mechanisms, whereas those that do not may demand additional mechanisms to explain them. This knowledge can help to guide psycholinguistic theory building with respect to agreement: understanding which effects must be explained with language-specific mechanisms and which can be reduced to general purpose sequence processing can help understand which results theories should be held accountable for.

Of course, demonstrating that an effect can emerge from a trained neural network model does not provide an explanation of the phenomenon. It does, however, allow us to characterize what kinds of inductive biases are needed for a learner to learn to exhibit that effect. Our simulations used LSTM language models as a simple, relatively unsophisticated learner, but by comparing profiles of effects simulated across multiple model architectures, one can identify what kinds of biases might lead to human-like behaviors. These results thus also serve as a starting point for developing more sophisticated neural network models of agreement, identifying in which areas the simple LSTM model used here needs more sophisticated machinery, whether it be additional mechanisms or processing constraints like noise or memory limitations. The results presented here represent a promising sign for future work, demonstrating that relatively simple models can learn to exhibit a number of signatures of human agreement processing from large corpora, but also that much work needs to be done in developing more human-like neural models of processing.

References

Bock, K., & Cutting, J. C. (1992). Regulating mental energy: Performance units in language production. *JML*, *31*(1).

Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, *23*(1).

Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making Syntax of Sense: Number Agreement in Sentence Production. *Psychological Review*, *113*(3).

Franck, J., Vigliocco, G., & Nicol, J. (2002). Subject-verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes*, *17*(4).

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *NAACL 2018*. New Orleans, Louisiana: ACL.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *NAACL 2001*.

Haskell, T. R., & Macdonald, M. C. (2005). Constituent Structure and Linear Order in Language Production: Evidence From Subject-Verb Agreement. *JEP: LMC*, *31*(5).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8).

Humphreys, K. R., & Bock, K. (2005). Notional Number Agreement in English. *Psychonomic Buttetin & Review*, *12*(4).

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., & Baroni, M. (2019). The emergence of number and syntax units in LSTM language models. In *NAACL 2019*. Minneapolis, Minnesota: ACL.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *TACL*, *4*.

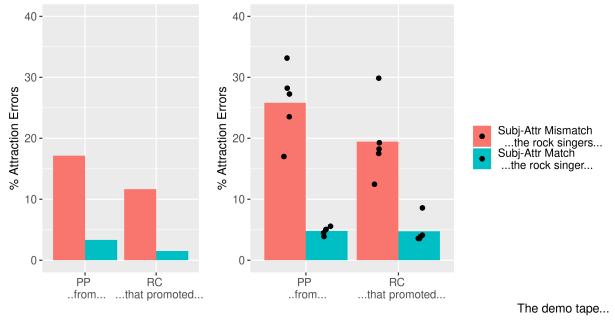
Linzen, T., & Leonard, B. (2018). Distinct patterns of syntactic agreement errors in recurrent networks and humans. In *40th Annual Conference of the Cognitive Science Society.*

Parker, D., & An, A. (2018). Not all phrases are equally attractive: Experimental evidence for selective agreement attraction effects. *Frontiers in Psychology*, *9*.

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement Processes in Sentence Comprehension. *JML*, *41*(3).

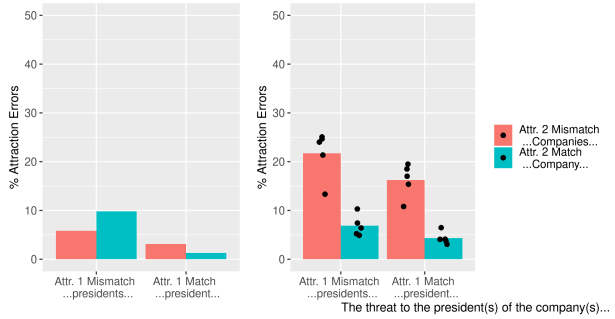
Van Dyke, J., & McElree, B. (2011). Cue-dependent interference in comprehension. *JML*, *65*(3).

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *JML*, *61*(2).



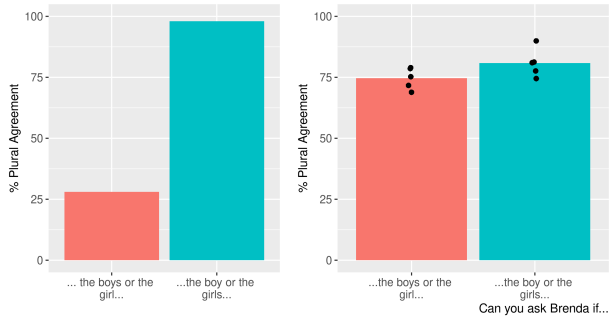
(a) Human Results (b) Simulation

Figure 1: Human and simulation PP/RC attraction results for Bock and Cutting (1992). Dots represent individual model means; bars represent averages over 5 models.



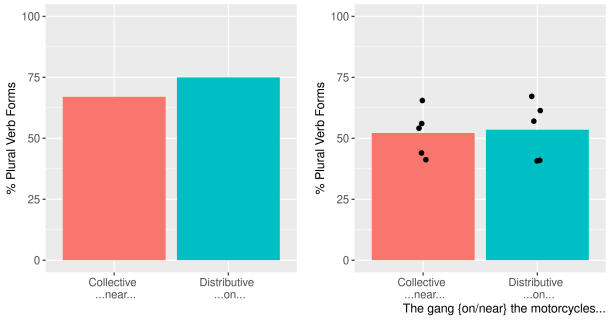
(a) Human Results (b) Simulation

Figure 2: Human and simulation results for Franck et al. (2002). Dots represent individual model means; bars represent averages over 5 models.



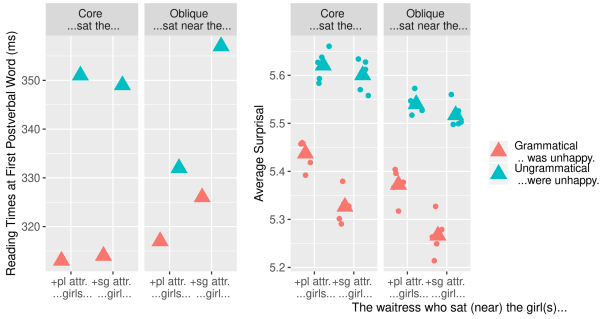
(a) Human Results (b) Simulation

Figure 3: Human and simulation results for Haskell and Macdonald (2005). Dots represent individual model means; bars represent averages over 5 models.



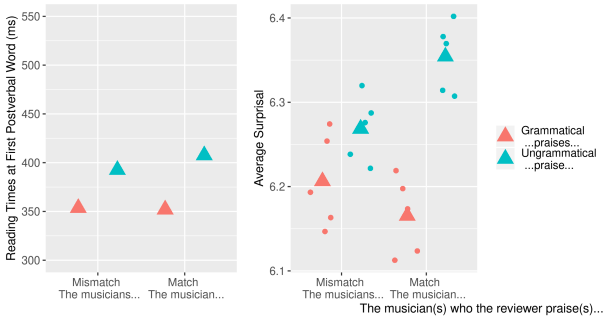
(a) Human Results (b) Simulation

Figure 4: Human and simulation results for Humphreys and Bock (2005). Dots represent individual model means; bars represent averages over 5 models.



(a) Human Results (b) Simulation

Figure 5: Human and simulation results for Parker and An (2018). Dots represent individual model means; triangles represent averages over 5 models.



(a) Human Results (b) Simulation

Figure 6: Human and simulation results for Wagers et al. (2009). Dots represent individual model means; triangles represent averages over 5 models.