Harnessing the richness of the linguistic signal in predicting pragmatic inferences

Sebastian Schuster[*]
Stanford University
sebschu@stanford.edu

Yuxing Chen[*]
Stanford University
yxchen28@stanford.edu

Judith Degen
Stanford University
jdegen@stanford.edu

Abstract

The strength of pragmatic inferences systematically depends on linguistic and contextual cues. For example, the presence of a partitive construction increases the strength of a so-called scalar inference: humans perceive the inference that Chris did not eat all of the cookies to be stronger after hearing *"Chris ate some of the cookies"* than after hearing the same utterance without a partitive, *"Chris ate some cookies"*. In this work, we explore to what extent it is possible to learn associations between linguistic cues and inference strength ratings without direct supervision. We show that an LSTM-based sentence encoder with an attention mechanism trained on a dataset of human inference strength ratings is able to predict ratings with high accuracy ($r = 0.78$). We probe the model's behavior in multiple analyses using corpus data and manually constructed minimal pairs and find that the model learns associations between linguistic cues and scalar inferences, suggesting that these associations are inferable from statistical input.

1 Introduction

An important property of human communication is that listeners can infer information beyond the literal meaning of an utterance. One well-studied type of inference is *scalar inference* (Grice, 1975; Horn, 1984), whereby a listener who hears an utterance with a scalar item like *some* infers the negation of a stronger alternative with *all*:

(1) a. Chris ate some of the cookies.
     b. ⤳ Chris ate some, but not all, of the cookies.

Early accounts of scalar inferences (e.g., Gazdar 1979; Horn 1984; Levinson 2000) considered them to arise by default unless specifically cancelled by the context. However, in a recent corpus study, Degen (2015) showed that there is much

more variability in the strength of scalar inferences from *some* to *not all* than previously assumed. Degen (2015) further showed that this variability is not random and that several lexical, syntactic, and semantic/pragmatic features of context explain much of the variance in inference strength.

Recent Bayesian game-theoretic models of pragmatic reasoning (Goodman and Frank, 2016; Franke and Jäger, 2016), which are capable of integrating multiple linguistic cues with world knowledge, are able to correctly predict listeners' pragmatic inferences in many cases (e.g., Goodman and Stuhlmüller 2013; Degen et al. 2015). These experimental and modeling results suggest that listeners integrate multiple linguistic and contextual cues in utterance interpretation, raising the question how listeners are able to draw these pragmatic inferences so quickly in interaction. This is an especially pressing problem considering that inference in Bayesian models of pragmatics is intractable when scaling up beyond toy domains to make predictions about arbitrary utterances.[1]

One possibility is that language users learn to use shortcuts to the inference (or lack thereof) by learning associations between the speaker's intention and surface-level cues present in the linguistic signal across many instances of encountering a scalar expression like *some*. In this work, we investigate whether it is possible to learn such associations between cues in the linguistic signal and speaker intentions by training neural network models to predict empirically elicited inference strength ratings from the linguistic input. In this

---

[*] Equal contribution.

[1]Recent models of generating pragmatic image descriptions (Andreas and Klein, 2016; Cohn-Gordon et al., 2018) and color descriptions (Monroe et al., 2017) have overcome this issue by approximating the distributions of utterances given a set of potential referents. However, these models require generative models of utterances such as an image captioning model and are therefore limited to scenarios where such a generative model is available.

enterprise we follow the recent successes of neural network models in predicting a range of linguistic phenomena such as long distance syntactic dependencies (e.g., Elman 1990; Linzen et al. 2016; Gulordava et al. 2018; Futrell et al. 2019; Wilcox et al. 2019), semantic entailments (e.g., Bowman et al. 2015; Conneau et al. 2018), acceptability judgements (Warstadt et al., 2018), factuality (Rudinger et al., 2018), and, to some extent, speaker commitment (Jiang and de Marneffe, 2019). In particular, we ask:

1. How well can a neural network sentence encoder learn to predict human inference strength judgments for utterances with *some*?

2. To what extent does such a model capture the qualitative effects of hand-mined contextual features previously identified as influencing inference strength?

To address these questions, we first compare the performance of neural models that differ in the underlying word embedding model (GloVe, ELMo, or BERT) and in the sentence embedding model (LSTM, LSTM+attention). We then probe the best model's behavior through a regression analysis, an analysis of attention weights, and an analysis of predictions on manually constructed minimal sentence pairs.

2 The dataset

We used the annotated dataset reported by Degen (2015), a dataset of the utterances from the Switchboard corpus of telephone dialogues (Godfrey et al., 1992) that contain the word *some*. The dataset consists of 1,362 unique utterances with a noun phrase containing *some* (*some*-NP). For each example with a *some*-NP, Degen (2015) collected inference strength ratings from at least 10 participants recruited on Amazon's Mechanical Turk. Participants saw both the target utterance and ten utterances from the preceding discourse context. They then rated the similarity between the original utterance like (2a) and an utterance in which *some* was replaced with *some, but not all* like (2b), on a 7-point Likert scale with endpoints labeled "very different meaning" (1) and "same meaning" (7). Low similarity ratings thus indicate low inference strength, and high similarity ratings indicate high inference strength.

(2)   a. I like, I like to read *some* of the philosophy stuff.

        b. I like, I like to read *some, but not all,* of the philosophy stuff.

Using this corpus, Degen (2015) found that several linguistic and contextual factors influenced inference strength ratings, including the partitive form *of*, subjecthood, the previous mention of the embedded NP referent, determiner strength, and modification of the head noun.

Partitive: (3a-b) are example utterances from the corpus with and without partitive *some*-NPs, respectively. Values in parentheses indicate the mean inference strength rating for that item. On average, utterances with partitives yielded stronger inference ratings than ones without.

(3)   a. I've seen *some of* them on repeat.   (5.8)

        b. You sound like you have *some* small ones in the background.   (1.5)

Subjecthood: Utterances in which the *some*-NP appears in subject position, as in (4a), yielded stronger inference ratings than utterances in which the *some*-NP appears in a different grammatical position, e.g., as a direct object as in (4b).

(4)   a. *Some* kids are really having it.   (5.9)

        b. That would take *some* planning.   (1.4)

Previous mention: Discourse properties also have an effect on inference strength. A *some*-NP with a previously mentioned embedded NP referent yields stronger inferences than a *some*-NP whose embedded NP referent has not been previously mentioned. For example, (5a) contains a *some*-NP in which *them* refers to previously mentioned *Mission Impossible tape recordings*, whereas *planning* in the *some*-NP in (5b) has not been previously mentioned.

(5)   a. I've seen *some of them* on repeats.   (5.8)

        b. That would take *some planning*.   (1.4)

Modification: Degen (2015) also found a small effect of whether or not the head noun of the *some*-NP was modified such that *some*-NP with unmodified head nouns yielded slightly stronger inferences than those with modified head nouns.

Determiner strength: Finally, the strength of the determiner *some* has traditionally been analyzed as having a weak, indefinite, non-presuppositional reading as well as a strong, quantificational, presuppositional reading (Milsark, 1974; Barwise and Cooper, 1981). While the weak/strong distinction has been notoriously hard to pin down (Horn, 1997), Degen (2015) used strength norms elicited independently for each item, which exploited the presuppositional nature of strong *some*: removing *some (of)* from utterances with weak *some* leads to

higher ratings on a 7-point Likert scale from 'different meaning' to 'same meaning' than removing it from utterances with strong *some*. Items with stronger *some* – e.g., (6a), strength 3.3 – yielded stronger inference ratings than items with weaker *some* – e.g., (6b), strength 6.7.

(6) a. And some people don't vote. (5.2)

  b. Well, we could use some rain up here. (2.1)

The quantitative findings from Degen (2015) are summarized in Figure 3, which shows in blue the regression coefficients for all predictors she considered (see the original paper for more detailed descriptions).

For our experiments, we randomly split the dataset into a 70% training and 30% test set, resulting in 954 training items and 408 test items.

3 Model

The objective of the model is to predict mean inference strength rating i given an utterance (a sequence of words) $U = \{w_1, w_2, ..., w_N\}$. While the original participant ratings were on a Likert scale from 1 to 7, we rescale these values to the interval $[0, 1]$. Figure 1 shows the overall model architecture. The model is a sentence classification model akin to the model proposed by Lin et al. (2017). The model first embeds the utterance tokens using pre-trained embedding models, and then forms a sentence representation by passing the embedded tokens through a 2-layer bidirectional LSTM network (biLSTM) (Hochreiter and Schmidhuber, 1997) with dropout (Srivastava et al., 2014) followed by a self-attention mechanisms that provides a weighted average of the hidden states of the top-most biLSTM hidden states. This sentence representation is then passed through a transformation layer with a sigmoid activation function, which outputs the predicted score in the interval $[0, 1]$. We rescale this predicted value to fall in the original interval $[1, 7]$.

We evaluated three word embedding models: GloVe, a static pre-trained word embedding matrix (Pennington et al., 2014), and pre-trained contextual word embedding models in the form of English ELMo (Peters et al., 2018; Gardner et al., 2018) and English BERT (Devlin et al., 2019; Wolf et al., 2019) models. We used the 100d GloVe embeddings, and we evaluated the 768d uncased BERT-base and 1024d BERT-large models.
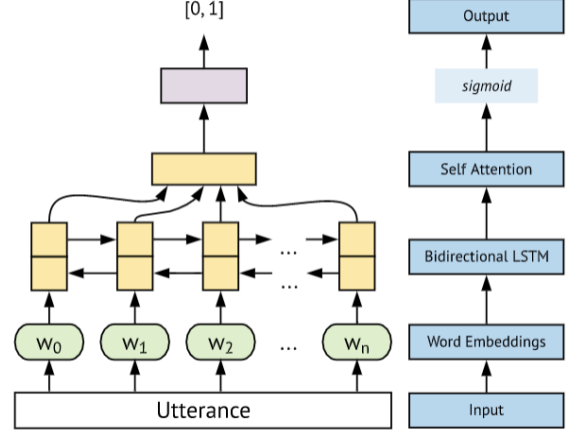

Figure 1: Model architecture.

4 Experiments

4.1 Training

We used 5-fold cross-validation on the training data to optimize the following hyperparameters.
Word embedding model: GloVe, ELMo, BERT-base, BERT-large.
Output layer of word embedding models: $[1, 3]$ for ELMo, $[1, 12]$ for BERT-base, and $[1, 24]$ for BERT-large.
Number of training epochs: $[1, 800]$.
Dimension of LSTM hidden states: $\{100, 200, 400, 800\}$.
Dropout rate in LSTM: $\{0.1, 0.2, 0.3, 0.4\}$.

We first optimized the output layer of the word embedding model for each embedding model while keeping all other parameters fixed. We then optimized the other parameters for each embedding model by computing the average correlation between the model predictions and the human ratings across the five cross-validation folds.

Architectural variants. We also evaluated all combinations of two architectural variants: First, we evaluated models in which we included the attention layer (LSTM+ATTENTION) or simply used the final hidden state of the LSTM (LSTM) as a sentence representation. Second, since participants providing inference strength ratings also had access to the preceding conversational context, we also compared models that make predictions based only the target utterance with the *some*-NP and models that make predictions based on target utterances and the preceding conversational context. For the models using GloVe and ELMo, we prepended the conversational context to the target utterance to obtain a joint context and

sentence embedding. For models using BERT, we made use of the fact that BERT had been trained to jointly embed two sentences or documents, and we obtained embeddings for the tokens in the target utterance by feeding the target utterance as the first document and the preceding context as the second document into the BERT encoder. For these models, we discarded the hidden states of the preceding context and only used the output of the BERT encoder for the tokens in the target utterance.

Implementation details. We implemented the model in PyTorch (Paszke et al., 2017). We trained the model using the Adam optimizer (Kingma and Ba, 2015) with default parameters and a learning rate of 0.001, minimizing the mean squared error of the predicted ratings. In the no-context experiments, we truncated target utterances longer than 30 tokens, and in the experiments with context, we truncated the beginning of the preceding context such that the number of tokens did not exceed 150.

Evaluation. We evaluated the model predictions in terms of their correlation r with the human inference strength ratings. For the optimization of hyperparameters and architectural variants, we evaluated the model using 5-fold cross-validation. We then took the best set of parameters and trained a model on all the available training data and evaluated that model on the held-out data.

4.2 Tuning results

We find that the attention layer improves predictions; that contextual word embeddings lead to better results than the static GloVe embeddings; and that including the conversational context does not improve predictions (see Appendix A, for learning curves of all models, and Section 6, for a discussion of the role of conversational context).

Otherwise, the model is quite insensitive to hyperparameter settings: neither the dimension of the hidden LSTM states nor the dropout rate had considerable effects on the prediction accuracy. We do find, however, that there are differences depending on the BERT and ELMo layer that we use as word representations. We find that higher layers work better than lower layers, suggesting that word representations that are influenced by other utterance tokens are helpful for this task.

Based on these optimization runs, we chose the model with attention that uses the BERT-large embeddings but no conversational context for the subsequent experiments and analyses.
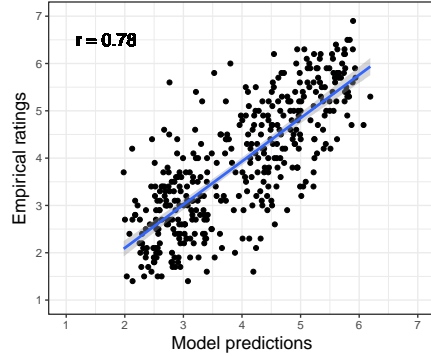

Figure 2: Correlation between empirical ratings and predictions of the BERT-LARGE LSTM+ATTENTION model on held-out test items.

4.3 Test results

Figure 2 shows the correlation between the best model according to the tuning runs (now trained on all training data) and the empirical ratings on the 408 held-out test items. As this plot shows, the model predictions fall within a close range of the empirical ratings for most of the items ($r = 0.78$).[2] Further, similarly as in the empirical data, there seem to be two clusters in the model predictions: one that includes lower ratings and one that includes higher ratings, corresponding to strong and weak scalar inferences, respectively. The only systematic deviation appears to be that the model does not predict any extreme ratings – almost all predictions are greater than 2 or less than 6, whereas the empirical ratings include some cases outside of this range.

Overall, these results suggest that the model can learn to closely predict the strength of scalar inferences. However, this result by itself does not provide evidence that the model learned associations between linguistic cues and inference strength, since it could also be that, given the large number of parameters, the model learned spurious correlations independent of the empirically established cue-strength associations. To rule out the latter explanation, we probed the model's behavior in multiple ways, which we discuss next.

5 Model behavior analyses

Regression analysis. As a first analysis, we investigated whether the neural network model pre-

---

[2]For comparison, we estimated how well the human ratings correlated by re-sampling the human ratings for each item and computing the average correlation coefficient between the original and the re-sampled datasets, which we found to be approximately 0.93.
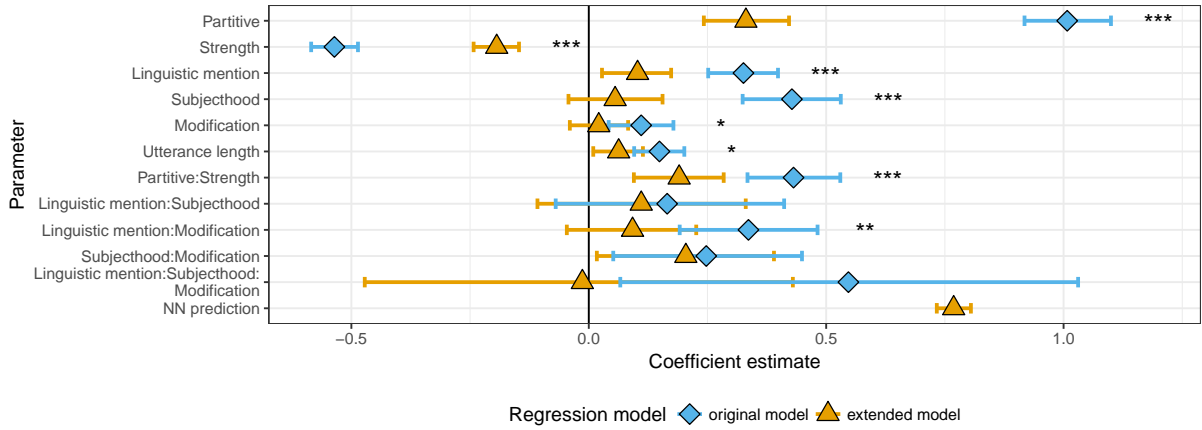
Figure 3: Maximum a posteriori estimates and 95%-credible intervals of coefficients for original and extended Bayesian mixed-effects regression models predicting the inference strength ratings. */**/*** indicate that the probability of the coefficient of the original model having a larger magnitude than the coefficient of the extended model is less than 0.05, 0.01, and 0.001, respectively.

dictions explain (some of) the variance explained by the linguistic factors that modulate inference strength. To this end, we used a slightly simplified[3] Bayesian implementation of the linear mixed-effects model by Degen (2015) using the brms (Bürkner, 2017) and STAN (Carpenter et al., 2017) packages and compared this *original* model to an *extended* model that included the output of the above NN model as a predictor. For this comparison, we investigated whether the magnitude of a predictor in the *original* model significantly decreased in the *extended* model that included the NN predictions, based on the reasoning that if the NN predictions already explain the variance previously explained by these manually coded predictors, then the original predictor should explain no or less additional variance. We approximated the probability that the magnitude of the coefficient in the *extended* model including the NN predictor is smaller than the coefficient in the *original* model, $P(|\beta_i^{extended}| < |\beta_i^{original}|)$, by sampling values for each coefficient from the distributions of the *original* and the *extended* models and comparing the magnitude of the sampled coefficients. We repeated this process 1,000,000 times and treated the simulated proportions as approximate probabilities.

An issue with this analysis is that estimating the regression model only on the items in the held-out test set yields very wide credible intervals for

some of the predictors–in particular for some of the interactions–since the model infers these values from very little data. We therefore performed this (and all subsequent) analyses on the entire data, and obtained the NN predictions through 6-fold cross-validation, so that the NN model always made predictions on data that it had not seen during training. This did yield the same qualitative results as the analyses only performed on the held-out test items (see Appendix B) but it also provided us with narrower credible intervals that highlight the differences between the coefficient estimates of the two models.

Figure 3 shows the estimates of the coefficients in the original model and the extended model. We find that the NN predictions explain some or all of the variance originally explained by many of the manually coded linguistic features: the estimated magnitude of the predictors for partitive, determiner strength, linguistic mention, subjecthood, modification, utterance length, and two of the interaction terms decreased in the extended model. These results suggest that the NN model indeed learned associations between linguistic features and inference strength rather than only explaining variance caused by individual items. This is particularly true for the grammatical and lexical features; we find that the NN predictor explains most of the variance originally explained by the partitive, subjecthood, and modification predictors. More surprisingly, the NN predictions also explain a lot of the variance originally explained by the determiner strength predictor, which was empirically determined by probing human inter-

---

[3]We removed by-item random intercepts and by-subject random slopes to facilitate inference. This simplification yielded almost identical estimates as the original model by Degen (2015).
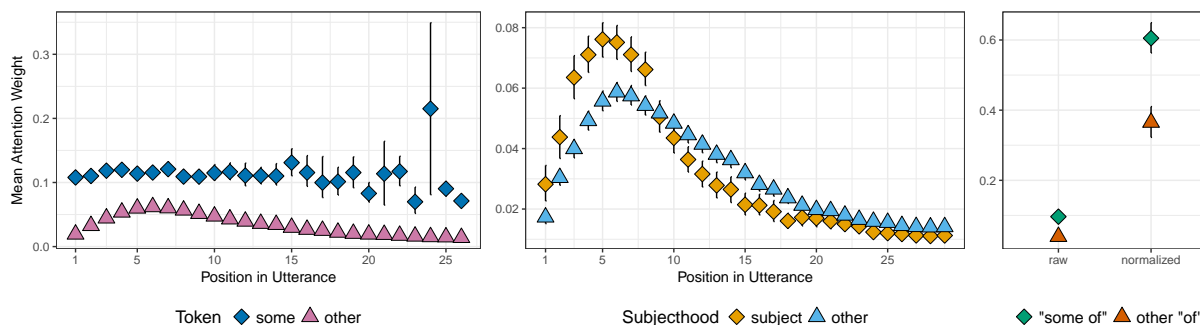
Figure 4: **Left**: Average attention weights at each token position for *some* and other tokens. **Center**: Average attention weights at each token position for utterances with subject and non-subject *some*-NPs. **Right:** Average attention weights of *of*-tokens in partitive *some*-NPs and weights of other *of*-tokens. In the normalized cases, we take only the utterances with multiple *of*-tokens into account and re-normalize the attention weights across all *of*-tokens in one utterance. Error bars indicate 95% bootstrapped confidence intervals.

pretation and is not encoded explicitly in the surface form utterance.[4] One potential explanation for this is that strong and weak *some* have different context distributions. For instance, weak *some* occurs in existential *there* constructions and with individual-level predicates, whereas strong *some* tends not to (Milsark, 1974; McNally and Geenhoven, 1998; Carlson, 1977). Since pre-trained word embedding models capture a lot of distributional information, the NN model is presumably able to learn this association.

Attention weight analysis. As a second type of analysis, we analyzed the attention weights that the model used for combining the token embeddings to a sentence embedding. Attention weight analyses have been successfully used for inspecting and debugging model decisions (e.g., Lee et al., 2017; Ding et al., 2017; Wiegreffe and Pinter, 2019; Vashishth et al., 2019; but see Serrano and Smith, 2019, and Jain and Wallace, 2019, for critical discussions of this approach). Based on these results, we expected the model to attend more to tokens that are relevant for making predictions. Given that many of the hand-mined features that predict inference strength occur within or in the vicinity of the *some*-NP, we should therefore expect the model to attend most to the *some*-NP.

To test this, we first explored whether the model attended on average more to *some* than to other tokens in the same position. Further, we exploited the fact that subjects generally occur at the beginning of a sentence. If the model attends to the vicinity of the *some*-NP, the average attention

weights should be higher on early positions for utterances with a subject *some*-NP compared to utterances with a non-subject *some*-NP, and conversely for late utterance positions. We thus compared the average attention weights for each position in the utterance across utterances with subject versus non-subject *some*-NPs. To make sure that any effects were not only driven by the attention weight of the *some*-tokens, we set the attention weights of the token corresponding to *some* to 0 and re-normalized the attention weights for this analysis. Further, since the attention weights are dependent on the number of tokens in the utterance, it is crucial that the average utterance length across the two compared groups be matched. We addressed this by removing outliers and limiting our analysis to utterances up to length 30 (1,028 utterances), which incidentally equalized the number of tokens across the two groups. (While these exclusions resulted in tiny quantitative difference in the average attention weights, the qualitative patterns are not affected.)

The left panel of Figure 4 shows the average attention weight by position for *some* versus other tokens. The model assigns much higher weight to *some*. The center panel of Figure 4 shows the average attention weight by position for subject vs. non-subject *some*-NP utterances. The attention weights are generally higher for tokens early in the utterance,[5] but the attention weights of utterances with a subject *some*-NP are on average higher for tokens early in the utterance compared to utterances with the *some*-NP in non-subject positions. Both of these findings provide evidence

---

[4]As explained above, Degen (2015) obtained strength ratings by asking participants to rate the similarity of the original utterance and an utterance without the determiner *some (of)*.

[5]This is in part an artifact of shorter utterances which distribute the attention weights among fewer tokens.

that the model assigns high weight to the tokens within and surrounding the *some*-NP.[6]

In a more targeted analysis to assess whether the model learned to use the partitive cue, we examined whether the model assigned higher attention to the preposition *of* in partitive *some*-NPs compared to when *of* occurred elsewhere. As utterance length was again a potential confound, we conducted the analysis separately on the full set of utterances with raw attention weights and on a subset that included only utterances with at least two instances of *of* (128 utterances), in which we renormalized the weights of *of*-tokens to sum to 1.

Results are shown in the right panel of Figure 4. The attention weights were higher for *of* tokens in partitive *some*-NPs, suggesting that the model learned an association between partitive *of* in *some*-NPs and inference strength.

Minimal pair analysis. As a final analysis, we constructed artificial minimal pairs that differed along several factors of interest and compared the model predictions. Such methods have been recently used to probe what kind of syntactic dependencies different types of recurrent neural network language models are capable of encoding (e.g., Linzen et al. 2016; Gulordava et al. 2018; Chowdhury and Zamparelli 2018; Marvin and Linzen 2018; Futrell et al. 2019; Wilcox et al. 2019), and also allow us to probe whether the model is sensitive to controlled changes in the input.

We constructed a set of 25 initial sentences with *some*-NPs. For each sentence, we created 32 variants that differed in the following four properties of the *some*-NP: subjecthood, partitive, prenominal modification, and post-nominal modification. For the latter three features, we either included or excluded *of the* or the modifier, respectively. To manipulate subjecthood of the *some*-NP, we created variants in which *some* was either the determiner in the subject NP as in (7a) or in the object-NP as in (7b). We also created passive versions of each of these variants (7c-d). Each set of sentences included a unique main verb, a unique
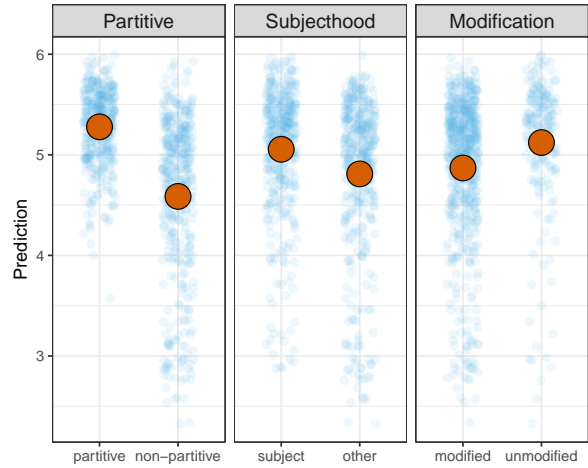

Figure 5: Average model predictions on manually constructed sentences, grouped by presence of partitives, by grammatical function of the *some*-NP, and by presence of nominal modifiers. Semi-transparent dots show predictions on individual sentences.

pair of NPs, and unique modifiers. The full list of sentences can be found in Appendix C.

(7)   a. Some (of the) (organic) farmers (in the mountains) milked the brown goats who graze on the meadows.

b. The organic farmers in the mountains milked some (of the) (brown) goats (who graze on the meadows).

c. The brown goats who graze on the meadows were milked by some (of the) (organic) farmers (in the mountains).

d. Some (of the) (brown) goats (who graze on the meadows) were milked by the organic farmers in the mountains.

Figure 5 shows the model predictions for the manually constructed sentences grouped by the presence of a partitive construction, the grammatical function of the *some*-NP, and the presence of a modifier. As in the natural dataset from Degen (2015), sentences with a partitive received higher predicted ratings than sentences without a partitive; sentences with subject *some*-NPs received higher predicted ratings than sentences with non-subject *some*-NPs; and sentences with a modified head noun in the *some*-NP received lower predictions than sentences with an unmodified *some*-NP. All these results provide additional evidence that the model learned the correct associations. This is particularly remarkable considering the train-test mismatch: the model was trained on noisy transcripts of spoken language that contained many disfluencies and repairs, and was subsequently tested on clean written sentences.

---

[6]The regression analysis suggests that the model learned to make use of the subjecthood cue and previous work on probing behavior of contextual word representations has found that such models are capable of predicting dependency labels, including subjects (e.g., Liu et al., 2019). We therefore also hypothesize that the representations of tokens that are part of a subject *some*-NP contain information about the subjecthood status. This in return could be an important feature for the output layer of the model and therefore be providing additional signal for the model to attend to these tokens.

6 Context, revisited

In the tuning experiments above, we found that including the preceding conversational context in the input to the model did not improve or lowered prediction accuracy. At the same time, we found that the model is capable of making accurate predictions in most cases without taking the preceding context into account. Taken together, these results suggest either that the conversational context is not necessary and one can draw inferences from the target utterance alone, or that the model does not make adequate use of the preceding context.

Degen (2015) did not systematically investigate whether the preceding conversational context was used by participants judging inference strength. To assess the extent to which the preceding context in this dataset affects inference strength, we re-ran her experiment, but without presenting participants with the preceding conversational context.[7] If the context is irrelevant for drawing inferences, then mean inference strength ratings should be very similar across the two experiments, suggesting the model may have rightly learned to not utilize the context. If the presence of context affects inference strength, ratings should differ across experiments, suggesting that the model's method of integrating context is ill-suited to the task.

The new, no-context ratings correlated with the original ratings ($r = 0.68$, see Appendix D) but were overall more concentrated towards the center of the scale, suggesting that in many cases, participants who lacked information about the conversational context were unsure about the strength of the scalar inference. Since the original dataset exhibited more of a bi-modal distribution with fewer ratings at the center of the scale, this suggests that the broader conversational context contains important cues to scalar inferences.

For our model, these results suggest that the representation of the conversational context is inadequate, which highlights the need for more sophisticated representations of linguistic contexts beyond the target utterance.[8] We further find that the model trained on the original dataset is worse at predicting the no-context ratings ($r = 0.66$) than

the original ratings ($r = 0.78$), which is not surprising considering the imperfect correlation between ratings across experiments, but also provides additional evidence that participants indeed behaved differently in the two experiments.

7 Conclusion and future work

We showed that neural network-based sentence encoders are capable of harnessing the linguistic signal to learn to predict human inference strength ratings from *some* to *not all* with high accuracy. Further, several model behavior analyses provided consistent evidence that the model learned associations between previously established linguistic features and the strength of scalar inferences. Taken together, these results suggest that it is possible to learn associations between linguistic features and scalar inferences from statistical input consisting of a relatively small set of utterances.

In an analysis of the contribution of the conversational context, we found that humans make use of the preceding context whereas the models we considered failed to do so adequately. Considering the importance of context in drawing both scalar and other inferences in communication (Grice, 1975; Clark, 1992; Bonnefon et al., 2009; Zondervan, 2010; Bergen and Grodner, 2012; Goodman and Stuhlmüller, 2013; Degen et al., 2015), the development of appropriate representations of larger context is an exciting avenue for future research.

One further interesting line of research would be to extend this work to other pragmatic inferences. Recent experimental work has shown that inference strength is variable across scale and inference type (Doran et al., 2012; Van Tiel et al., 2016). We treated *some* as a case study in this work, but none of our modeling decisions are specific to *some*. It would be straightforward to train similar models for other types of inferences.

Lastly, the fact that the attention weights provided insights into the model's decisions suggests possibilities for using neural network models for developing more precise theories of pragmatic language use. Our goal here was to investigate whether neural networks can learn associations for already established linguistic cues but it would be equally interesting to investigate whether such models could be used to discover new cues, which could then be verified in experimental and corpus work, potentially providing a novel model-driven approach to experimental and formal pragmatics.

---

[7]We recruited 680 participants on Mechanical Turk who each judged 20 or 22 items, yielding 10 judgments per item.

[8]The representation of larger linguistic context is also important for span-based question-answer (QA) systems (e.g., Hermann et al., 2015; Chen, 2018; Devlin et al., 2019) and adapting methods from QA to predicting scalar inferences would be a promising extension of the current model.

References

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*.

Jon Barwise and Robin Cooper. 1981. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4(2):159–219.

Leon Bergen and Daniel J. Grodner. 2012. Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(5):1450–60.

Jean-François Bonnefon, Aidan Feeney, and Gaëlle Villejoubert. 2009. When some is actually all: Scalar inferences in face-threatening contexts. *Cognition*, 112(2):249–258.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*.

Paul-Christian Bürkner. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.

Greg N. Carlson. 1977. A unified analysis of the english bare plural. *Linguistics and Philosophy*, 1(3):413–456.

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Danqi Chen. 2018. *Neural Reading Comprehension and Beyond*. Ph.D. thesis, Stanford University.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*.

Herbert H Clark. 1992. *Arenas of language use*. University of Chicago Press.

Reuben Cohn-Gordon, Noah D. Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2018)*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Holger Schwenk, Ves Stoyanov, Adina Williams, and Samuel R. Bowman. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*.

Judith Degen. 2015. Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, 8:11–1.

Judith Degen, Michael Henry Tessler, and Noah D. Goodman. 2015. Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society (CogSci 2015)*, pages 548–553.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*.

Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*.

Ryan Doran, Gregory Ward, Meredith Larson, Yaron McNabb, and Rachel E. Baker. 2012. A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language*, 88(1):124–154.

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

Michael Franke and Gerhard Jäger. 2016. Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, 35(1):3–44.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*.

G. Gazdar. 1979. *Pragmatics: Implicature, Presupposition and Logical Form*. Academic Press.

John J. Godfrey, Edward C. Holliman, and Jane Mc-Daniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 92)*.

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Noah D. Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5(1):173–184.

Herbert P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Vol. 3, Speech Acts*, pages 41–58. Academic Press, New York.

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

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28 (NeurIPS 2015)*.

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

Laurence Horn. 1997. All John's children are as bald as the king of France: Existential import and the geometry of opposition. In *CLS 33*, pages 155–179.

Laurence R. Horn. 1984. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In Deborah Schiffrin, editor, *Meaning, Form, and Use in Context: Linguistic Applications*, pages 11–4. Georgetown University Press, Washington, D.C.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Do you know that florence is packed with visitors? evaluating state-of-the-art models of speaker commitment. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL 2019)*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2015)*.

Jaesong Lee, Joong-Hwi Shin, and Jun-Seok Kim. 2017. Interactive visualization and manipulation of attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Stephen C Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*.

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.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*.

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 (EMNLP 2018)*.

Louise McNally and Veerle Van Geenhoven. 1998. Redefining the weak/strong distinction.

Gary Milsark. 1974. *Existential sentences in English*. Ph.D. thesis, MIT.

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NeurIPS Autodiff Workshop 2017*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNL 2014)*.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2018)*.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2018)*, pages 731–744.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Bob Van Tiel, Emiel Van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar diversity. *Journal of Semantics*, 33(1):137–175.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across NLP tasks. *arXiv preprint arXiv:1909.11218.*

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471.*

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2019. What syntactic structures block dependencies in RNN language models? In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci 2019)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rmi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771.*

Arjen Zondervan. 2010. *Scalar Implicatures or Focus: An Experimental Approach*. LOT Publications, Utrecht.

A Hyperparameter tuning

Figure 6 shows the learning curves averaged over the 5 cross-validation tuning runs for models using different word embeddings. As these plots show, the attention layer improves predictions; contextual word embeddings lead to better results than the static GloVe embeddings; and including the conversational context does not improve predictions and in some cases even lowers prediction accuracy.
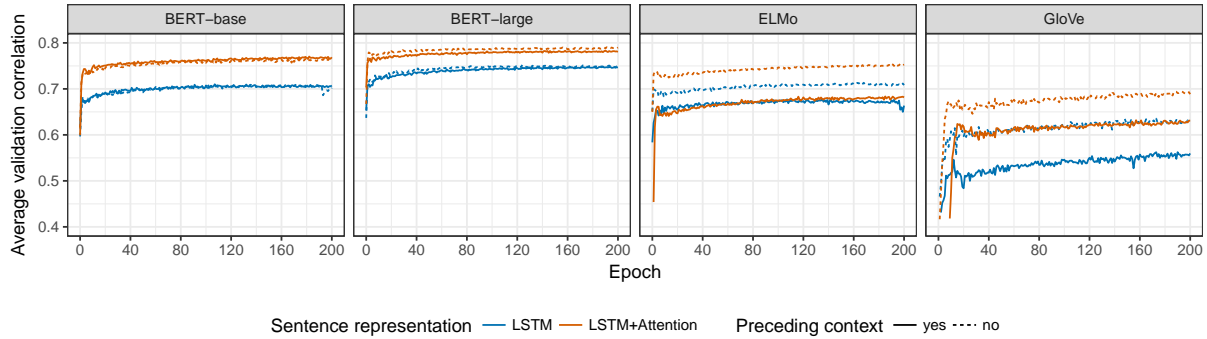

Figure 6: Correlation between each model's predictions on valuation set and empirical means, by training epoch.

B Regression analysis on held-out test data

Figure 7 shows the estimates of the predictors in the original and extended Bayesian mixed-effects models estimated only on the held-out test data. We find the same qualitative effects as in Figure 3, but since these models were estimated on much less data (only 408 items), there is a lot of uncertainty in the estimates and therefore quantitative comparisons between the coefficients of the different models are less informative.
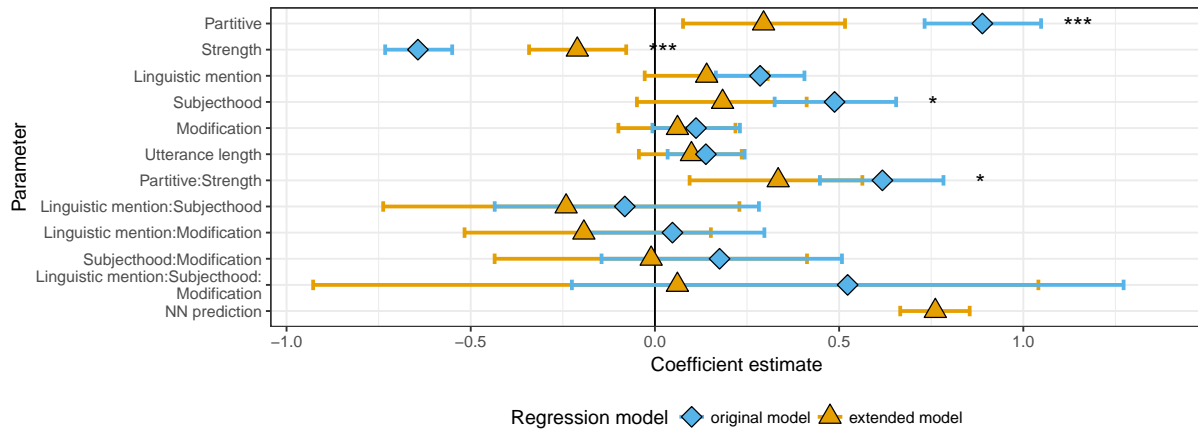

Figure 7: Maximum a posteriori estimates and 95%-credible intervals of coefficients for original and extended Bayesian mixed-effects regression models predicting the inference strength ratings on the held-out test set. */**/*** indicate that the probability of the coefficient of the original model having a larger magnitude than the coefficient of the extended model is less than 0.05, 0.01, and 0.001, respectively.

C   List of manually constructed sentences

Tables 1 and 2 show the 25 manually created sentences for the analyses described in the minimal pairs analysis in Section 5. As described in the main text, we created 16 variants of the sentence with the *some*-NP in subject position (sentences in the left column), and 16 variants of the sentence with the *some*-NP in object position (sentences in the right column), yielding in total 800 examples.

Some of the attentive waiters at the gallery opening poured the white wine that my friend really likes.	The attentive waiters at the gallery opening poured some of the white wine that my friend really likes.
Some of the experienced lawyers in the firm negotiated the important terms of the acquisition.	The experienced lawyers in the firm negotiated some of the important terms of the acquisition.
Some of the award-winning chefs at the sushi restaurant cut the red salmon from Alaska.	The award-winning chefs at the sushi restaurant cut some of the red salmon from Alaska.
Some of the brave soldiers who were conducting the midnight raid warned the decorated generals who had served in a previous battle.	The brave soldiers who were conducting the midnight raid warned some of the decorated generals who had served in a previous battle.
Some of the eccentric scholars from the local college returned the old books written by Camus.	The eccentric scholars from the local college returned some of the old books written by Camus.
Some of the entertaining magicians with top hats shuffled the black cards with dots.	The entertaining magicians with top hats shuffled some of the black cards with dots.
Some of the convicted doctors from New York called the former patients with epilepsy.	The convicted doctors from New York called some of the former patients with epilepsy.
Some of the popular artists with multiple albums performed the fast songs from their first album.	The popular artists with multiple albums performed some of the fast songs from their first album.
Some of the angry senators from red states impeached the corrupt presidents from the Republican party.	The angry senators from red states impeached some of the corrupt presidents from the Republican party.
Some of the underfunded researchers without permanent employment transcribed the recorded conversations that they collected while doing fieldwork.	The underfunded researchers without permanent employment transcribed some of the recorded conversations that they collected while doing fieldwork.
Some of the sharp psychoanalysts in training hypnotized the young clients with depression.	The sharp psychoanalysts in training hypnotized some of the young clients with depression.
Some of the harsh critics from the Washington Post read the early chapters of the novel.	The harsh critics from the Washington Post read some of the early chapters of the novel.
Some of the organic farmers in the mountains milked the brown goats who graze on the meadows.	The organic farmers in the mountains milked some of the brown goats who graze on the meadows.
Some of the artisanal bakers who completed an apprenticeship in France kneaded the gluten-free dough made out of spelt.	The artisanal bakers who completed an apprenticeship in France kneaded some of the gluten-free dough made out of spelt.
Some of the violent inmates in the high-security prison reported the sleazy guards with a history of rule violations.	The violent inmates in the high-security prison reported some of the sleazy guards with a history of rule violations.

Table 1: Manually constructed sentences used in the minimal pair analyses. Sentences in the left column have a *some*-NP in subject position; sentences on the right have a *some*-NP object position.

Some of the eager managers in the company instructed the hard-working sales representatives in the steel division about the new project management tool.	The eager managers in the company instructed some of the hard-working sales representatives in the steel division about the new project management tool.
Some of the brilliant chemists in the lab oxidized the shiny metals extracted from ores.	The brilliant chemists in the lab oxidized some of the shiny metals extracted from ores.
Some of the adventurous pirates on the boat found the valuable treasure that had been buried in the sand.	The adventurous pirates on the boat found some of the valuable treasure that had been buried in the sand.
Some of the mischievous con artists at the casino tricked the elderly residents of the retirement home.	The mischievous con artists at the casino tricked some of the elderly residents of the retirement home.
Some of the persistent recruiters at the conference hired the smart graduate students who just started a PhD as interns.	The persistent recruiters at the conference hired some of the smart graduate students who just started a PhD as interns.
Some of the established professors in the department supported the controversial petitions that were drafted by the student union.	The established professors in the department supported some of the controversial petitions that were drafted by the student union.
Some of the muscular movers that were hired by the startup loaded the adjustable standing desks made out of oak onto the truck.	The muscular movers that were hired by the startup loaded some of the adjustable standing desks made out of oak onto the truck.
Some of the careful secretaries at the headquarter mailed the confidential envelopes with the bank statements.	The careful secretaries at the headquarter mailed some of the confidential envelopes with the bank statements.
Some of the international stations in South America televised the early games of the soccer cup.	The international stations in South America televised some of the early games of the soccer cup.
Some of the wealthy investors of the fund excessively remunerated the successful brokers working at the large bank.	The wealthy investors of the fund excessively remunerated some of the successful brokers working at the large bank.

Table 2: Manually constructed sentences used in the minimal pair analyses (continued).

D Results from no-context experiment

Figure 8 shows the correlation between the mean inference strength ratings for each item in the experiment from Degen (2015) and the mean strength ratings from the new no-context experiment, discussed in Section 6.
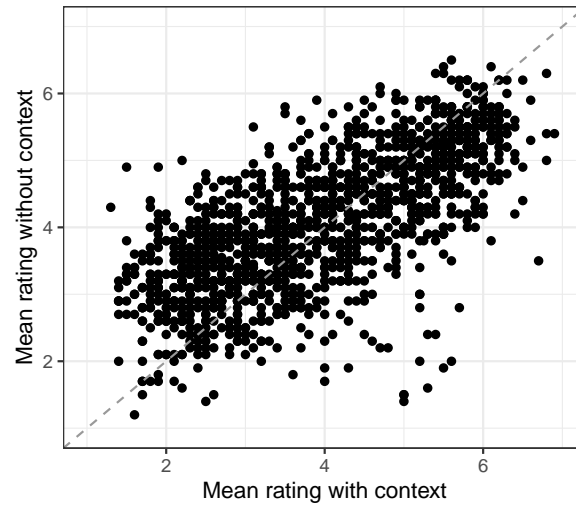

Figure 8: Mean inference strength ratings for items without context (new) against items with context (original), $r = .68$.