Subject-verb agreement with Seq2Seq transformers: Bigger is better, but still not best

Michael Wilson and **Zhenghao Zhou** and **Robert Frank**
Yale University
370 Temple Street
New Haven, CT 06511
{michael.a.wilson, herbert.zhou, robert.frank}@yale.edu

Abstract

Past work (Linzen et al., 2016; Goldberg, 2019, a.o.) has used the performance of neural network language models on subject-verb agreement to argue that such models possess structure-sensitive grammatical knowledge. We investigate what properties of the model or of the training regimen are implicated in such success in sequence to sequence transformer models that use the T5 architecture (Raffel et al., 2019; Tay et al., 2021). We find that larger models exhibit improved performance, especially in sentences with singular subjects. We also find that larger pre-training datasets are generally associated with higher performance, though models trained with less complex language (e.g., CHILDES, Simple English Wikipedia) can show more errors when trained with larger datasets. Finally, we show that a model's ability to replicate psycholinguistic results does not correspondingly improve with more parameters or more training data: none of the models we study displays a fully convincing replication of the hierarchically-informed pattern of agreement behavior observed in human experiments.

1 Introduction

In standard English, subjects and present-tense verbs covary in number, called *subject-verb agreement*. Crucially, agreement depends not on linear proximity to the verb, but structural proximity: the head noun of the subject determines correct agreement, not any of its dependents:

(1) a. The label on the bottle is...
b. * The labels on the bottle is...
c. The labels on the bottle are...
d. * The label on the bottles are...

Because of this structure-sensitive property of subject-verb agreement, this phenomenon is a useful grounds for examining the linguistic representations that computational language models learn.

Past work examining the performance of language models on subject-verb agreement has found mixed results. Linzen et al. (2016) and Marvin and Linzen (2018) showed LSTMs do not achieve consistent structure-sensitive generalization on agreement when trained on a language modeling task, though they perform better with explicit supervision related to agreement. Goldberg (2019) examined BERT, an encoder-only transformer model (Devlin et al., 2018), and found much higher subject-verb agreement performance.

These prior studies compared language model probabilities for individual word tokens (e.g., *is* vs. *are*) following a preamble (e.g., *the label on the bottles*) to determine whether singular or plural agreement is more likely. We use a different approach, studying agreement in models trained to map an input (non-agreeing) sequence to an output (agreeing) sequence. This follows a line of work in which grammatical transformation tasks can be used to assess sensitivity to grammatical regularities (McCoy et al., 2020; Mueller et al., 2022; Mulligan et al., 2021). Specifically, we use ablations of the Text-to-Text Transfer Transformer (T5) sequence to sequence (seq2seq) architecture (Raffel et al., 2019; Tay et al., 2021) to examine the effect of model size (number of parameters) and model architecture (where those parameters are located) on agreement behavior. As we shall see, bigger models do better, but some kinds of layers matter more for performance. We also investigate how pre-training data influences model performance, examining T5 models that were pre-trained on different datasets and different amounts of data.

Previous work has demonstrated that pre-training imparts a bias to make use of hierarchical generalizations in at least some seq2seq models on tasks like passivization and question formation in English and German (Mueller et al., 2022). Like these tasks, subject-verb agreement is sensitive to hierarchy and not linear order, as shown in (1).

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 278-288.
Amherst, Massachusetts, June 15-17, 2023

However, unlike passivization and question formation, agreement is not a generalization based on movement.[1] This could potentially impact the models' propensity to form hierarchical generalizations in this domain. Indeed, we find that even though the overall propensity to use grammatical agreement increases with model size, even the largest models we tested showed errors. Moreover, the pattern of these errors does not match patterns of errors found in psycholinguistic studies of agreement errors in humans. People show more sensitivity to structural proximity when making errors, while the models we tested showed more sensitivity to linear proximity. We conclude that the most reliable way to achieve higher performance on agreement in general is with larger models, though even the largest models we tested still do not replicate most human-like patterns of agreement errors, and thus show more evidence of linear rather than hierarchical generalization, at least with regards to agreement behavior.

We note here that we do not have a full explanation of why certain architectural properties and kinds of pre-training data have certain affects on agreement behavior. Rather, our more modest aim is merely to provide a sketch of the empirical landscape in this domain.

2 Methods

2.1 Procedure

Sequence to sequence (seq2seq) language models take a sequence of (tokenized) words as input, and produce a sequence of tokens as output. The model begins generation by producing a beginning of sentence token, and then produces the next most probable token at each generation step given the full input sequence and the previous tokens generated in the output sequence to that point.

To assess agreement behavior in these models, we take advantage of the fact that in English, verbs in the past tense are not marked for number (with the single exception of *was* vs. *were*, which was not included in our test set). Thus, we fine-tune the T5 checkpoints we use on a tense reinflection task (McCoy et al., 2020; Mueller et al., 2022; Petty and Frank, 2021; Mulligan et al., 2021). For example:

Source: "The professor liked the dean. PRES: "

Target: "The professor likes the dean."

This task requires the model to convert a sentence where number agreement is absent (i.e., the past tense) to a form where agreement is clearly marked (the present tense), forcing the model to resolve the ambiguity. We measure which form of the present tense verb the model produces.

We fine-tuned all models for $7,812$ weight updates (976.5 epochs) on this tense reinflection task with a learning rate of 5×10^{-5} and a batch size of 128, following Mueller et al. (2022). We saved 15 evenly-spaced checkpoints throughout fine-tuning to use for evaluation.[2]

2.2 Materials

Our fine-tuning dataset consists of $1,098$ examples constructed from sentences randomly drawn from English Wikipedia (`20200501.en`) using Hugging Face's `datasets` library.[3] We parsed the sentences using a transformer-based dependency parser provided by the `spacy` library (`en_core_web_trf`) (Honnibal et al., 2020). These parses allow us to identify the subject of the sentence and the verb, as well as the verb's tense. We created pairs of sentences for fine-tuning as follows: if the verb is in past tense, we treat the sentence as the input, and reinflect the verb into the present tense to produce the desired output; if the verb is in the present tense, we treat it as the desired output, and reinflect it into the past tense to produce the input. For reinflection, we used the `pattern` library (Smedt and Daelemans, 2012), with additional manual corrections. We included only examples that contained no intervening nouns between the main subject and the main verb according to the dependency parses, in order to avoid giving the models evidence during fine-tuning that would disambiguate the correct target of agreement, even inadvertently.[4]

For our test dataset, we created a balanced set of synthetic past-present example pairs using a PCFG. Using synthetic test data allowed us to ensure full

[1] That is, it is a relation that holds between elements in a structure, rather than a relation between structures (as movement is typically defined).

[2] Our code and data are available at: github.com/clay-lab/seq2seq-agreement-attraction-datasets, github.com/clay-lab/seq2seq-agreement-attraction.

[3] We also conducted fine-tuning with larger datasets, up to $10,000$ sentence pairs. Preliminary investigations showed little difference between the results with these larger fine-tuning datasets and the smaller dataset, so we continued to use the smaller dataset.

[4] Preliminary investigations showed that including sentences with interveners where the correct target of agreement was ambiguous in the pre-training data (e.g., *the key to the cabinet is...* is compatible with either a hierarchical or a linear generalization) made little difference to our results.

accuracy of the target forms during testing, since naturally occurring data may contain errors that arise naturally or during parsing. We represent conditions using "S" and "P," with "S" corresponding to a singular noun and "P" corresponding to a plural noun. The linear order of these labels represents their relative linear order in the sentence prior to the verb. For instance, the following is a sentence in the SP condition:

(2) The $\text{students}_{\text{S}}$ near the deans_{P} liked the professor.

Distractor nouns were embedded in either a prepositional phrase (PP) or a subject relative clause (RC), or a combination of two of them, attached to the preceding noun. Thus, there were test sentences for each combination of noun numbers (S, P, SS, SP, PP, PS, SSP, SPS, SPP, PPS, PSP, PSS) and embedding structure (PP, RC (two-noun conditions), PP+PP, PP+RC, RC+RC, RC+PP (three-noun conditions)). The test sentences used 10 nouns in singular and plural forms (*student*, *professor*, *headmaster*, *friend*, *assistant*, *dean*, *advisor*, *colleague*, *president*, *chancellor*), 10 verbs in past and present tense forms (*help*, *visit*, *like*, *bother*, *inspire*, *recruit*, *assist*, *confound*, *accost*, *avoid*), 5 prepositions (*of*, *near*, *by*, *behind*, *with*), the definite article (*the*), and the overt complementizer (*that*). Due to the limited vocabulary and structural simplicity, the S and P conditions each contained only 64 unique sentences each. All other conditions contained 256 unique sentences.

We did not ensure that every sentence had a completely plausible meaning. This is similar to Lasri et al. (2022)'s approach, who examined BERT's performance on subject-verb agreement in sentences without sensible meanings. It is also similar to Newman et al. (2021), who examined how plausibility of a verb in a particular context influenced BERT's ability to predict the syntactically correct form of an agreeing verb. Both studies found that implausible carrier sentences and less plausible verbs in a particular context were associated with a higher rate of errors. While we did not explicitly manipulate plausibility, our results can be similarly interpreted as reflecting models' performance in less than completely natural contexts.

2.3 Evaluation

During preliminary investigations with unconstrained generation of output, we found that the seq2seq models we used often failed to produce output that could be used to determine whether they displayed agreement errors straightforwardly. This was because the models either failed to produce the correct preamble (i.e., the string prior to the main verb); failed to reinflect the verb, leaving it in the past tense; or produced the wrong verb, which made it impossible to parse the output with the CFG used for analysis. For this reason, we used teacher forcing to make the models produce an identical preamble up to the main verb, and then forced them to produce either the singular or plural present tense form of the target verb.[5] This ensures that every output sentence provides information about the model's behavior with regards to agreement, since the output inevtiably reveals whether the model considers the singular or the plural form of the verb more likely given the correct preamble. We ignore the remainder of the output following the main verb for evaluation purposes.

Pre-verb noun(s)	Structures
S	–
P	–
SS, SP; PP, PS	PP; RC
SSS, SSP, SPS, SPP; PPP, PPS, PSP, PSS	PP+PP, PP+RC, RC+PP, RC+RC

Table 1: Summary of test set conditions. The correct target of agreement was always the first noun.

For each example in our test dataset, we record whether the model displayed erroneous agreement, defined as producing the singular form of the verb when the correct target is plural, or vice versa. Our plots show the proportion of errors on the y-axis; thus, higher numbers represent worse performance and lower numbers represent better performance. For each model, we consider results for only the checkpoint that showed the lowest overall proportion of agreement errors.

2.4 Models

We consider several T5 models, drawn from two sources. The first are checkpoints released with Tay et al. (2021), in (3). These models differ in a number of respects with comparison to a "base" model, including the total number of layers (NL), the number of encoder layers (EL), the number of decoder layers (DL), and the number of attention heads (NH).

(3) a. T5 Efficient Tiny, Mini, Small, and Base[6]

[5]This meant that at each generation step, we forced the models to predict only the correct actual token, and used that prediction to feed the next generation step, up to the disambiguating token at the verb.

[6]These models have the following architectures, which vary in several regards relative to T5 Efficient Base. Tiny:

b. Total number of layers (NL): T5 Efficient Base NL02, NL04, NL08, Base (NL12)[7]

c. Number of decoder layers (DL): T5 Efficient Base DL02, DL04, DL06, DL08, Base (DL12)

d. Number of encoder layers (EL): T5 Efficient Base EL02, EL04, EL06, EL08, Base (EL12)

e. Number of attention heads (NH): T5 Efficient Base NH08, Base (NH12), NH16, NH24, NH32

We do not consider other ablations here. This set of models ranges between 16 million parameters on the low end (T5 Efficient Tiny) and 364 million on the high end (T5 Efficient Base NH32). They were all pre-trained on the same dataset drawn from the Colossal Cleaned Common Crawl (C4) corpus, using a span-denoising objective. In total, we considered 19 T5 Efficient models.

To investigate the effects of pre-training data, we used models provided by Aaron Mueller (p.c.). These models each have 63 million parameters, and were pre-trained on a span-denoising objective. Different models were pre-trained on data drawn from different sources, including the CHILDES database (BabyT5), the C4 corpus (C4), Simple English Wikipedia (SimpleWiki), and standard English Wikipedia (WikiT5). The size of the pre-training datasets ranges from 1 million words to 1 billion words, though not every combination of dataset size and source is represented.[8] Altogether, these comprised a separate set of 13 models.

3 Results

3.1 Model size and architecture

First, we consider results for some of the T5 Efficient models (Tay et al., 2021). Figure 1 shows accuracy by condition and number of parameters.

For this and all future statistical results we report, we fit logistic regressions using R's `glm` function (R Core Team, 2022). Throughout the paper, for each family of hypothesis tests, we used the Bonferroni method to correct for multiple comparisons. As shown in (1), performance in most conditions was significantly affected by model size, such that more parameters led to a decreased error rate. The exceptions to this were the single-noun conditions, the PPS PP+PP condition, the PPS RC+PP condition, the PSS PP+PP condition, and the PSS PP+RC condition. In all cases, this appears to be due to the fact that even models with the smallest number of parameters we considered achieved high performance in these conditions, leaving little to no room for further improvement.

We next consider which kinds of parameters have effects. Naturally, increasing the number of layers (for example) increases the number of parameters. But we can also consider whether increasing the number of attention heads without increasing the number of layers is beneficial. Figure 2 shows the overall proportion of errors for the number of encoder layers, decoder layers, total layers, and attention heads per layer.

Both increasing the number of layers, as well as the number of attention heads per layer, significantly improves model performance (NL: $\beta = -0.0780$, $z = -83.9$, $p < 2.2 \times 10^{-16}$; NH: $\beta = -0.0870$, $z = -63.1$, $p < 2.2 \times 10^{-16}$). In addition, increases in the number of encoder layers and in the number of decoder layers both improve performance as well (EL: $\beta = -0.0948$, $z = -64.8$, $p < 2.2 \times 10^{-16}$; DL: $\beta = -0.101$, $z = -68.7$, $p < 2.2 \times 10^{-16}$). We found, however, that increasing the number of encoder layers resulted in a significantly greater increase in performance compared to increasing the number of decoder layers (EL $-$ DL: $\beta = -0.00593$, $z = -2.87$, $p = 0.00415$). The negative slope for the difference indicates that the magnitude of the EL effect is greater than the magnitude of the DL effect. Thus, assigning more parameters to encoding layers when increasing model size appears to carry a greater benefit with regards to overall agreement behavior in our test dataset.

This effect could in principle have two sources. One possibility is obvious: increasing the number of encoder layers provides greater benefits with regards to our tense-reinflection task and/or subject-verb agreement. But another possibility is that models with fewer decoder layers show less reduction in performance compared to models with more decoder layers, leaving less room for improvement as the number of decoder layers is increased. To investigate this, we can compare the intercepts of the regressions. We found that the intercept for the encoder-layer model was -0.661, while the intercept for the decoder-layer model was -0.618. This

NL04 (EL04, DL04), NH04; Mini: NL04 (EL04, DL04), NH08; Small: NL06 (EL06, DL06), NH08; Base: NL12 (EL12, DL12), NH12.

[7] Using the convention from Tay et al. (2021), the number by "NL" signifies half the total number of layers; e.g., NL02 means there are 2 encoder layers and 2 decoder layers (4 total).

[8] BabyT5: 1M, 5M; C4: 1M, 10M, 100M, 1B; SimpleWiki: 1M, 10M, 100M; WikiT5: 1M, 10M, 100M, 1B.

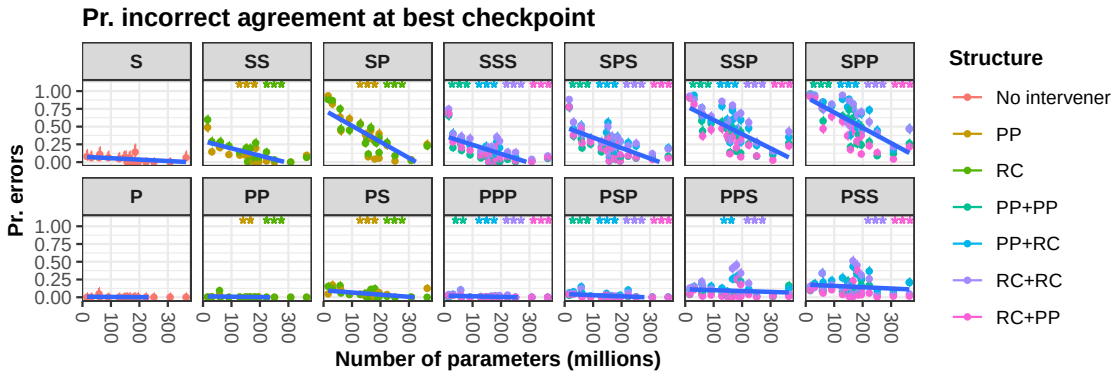


Figure 1: Accuracy by number of parameters and condition. Bars represent 95% CIs on the beta distribution. Colored stars indicate significance of the corresponding condition with Bonferroni-corrected α.

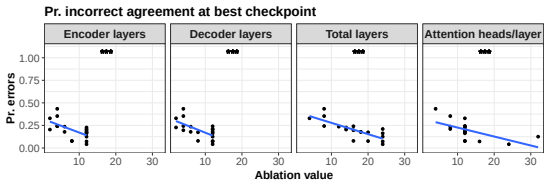


Figure 2: Accuracy by ablation type. Stars indicate significance with Bonferroni-corrected α.

indicates that the models with fewer encoder layers are *less* likely to make errors than the models with fewer decoder layers, and this difference is significant (EL − DL intercept: $\beta = 0.0435$, $z = 2.53$, $p = 0.01$). Thus, we find evidence that the difference reflects a genuine advantage for increased number of encoder layers on our task. We have no ready explanation for why this should be (in principle, agreement could be determined in either the encoder or the decoder, or equally in both). Nevertheless, we find this result interesting given the current focus of the field on decoder-only models like LLaMA (Touvron et al., 2023) and GPT (Brown et al., 2020; OpenAI, 2023). Our results suggest that for some tasks, it may be possible to more efficiently achieve higher performance with a model that incorporates an encoder.

When considering effects of this sort by condition (which we do not plot), we again found that in most conditions, increases in the relevant number of layers/heads led to improved performance. However, there were exceptions, summarized in table 2. In all other conditions, there were improvements in performance associated with increasing the parameters of each type. The single clear pattern is that performance in the plural subject conditions is less often improved by increasing model size, again likely due to the low error rate in these conditions to begin with. It is unclear to us why this should be the case; we recorded the number of singular and

Ablation(s)	Noun(s)	Structure(s)
DL, TL	S	–
EL, DL, TL, NH	P	–
NH	PP	PP, RC
NH	PS	PP
EL, DL	PPP	PP+PP
NH	PPP	PP+RC
EL, DL, TL, NH	PPS	PP+PP
DL, NH	PPS	PP+RC
EL, DL, TL	PPS	RC+PP
TL	PPS	RC+RC
EL, DL, TL, NH	PSS	PP+PP
DL, TL, NH	PSS	PP+RC
DL	PSS	RC+PP, RC+RC

Table 2: Summary of conditions where no improvement associated with various ablations was found.

plural subjects in our fine-tuning data and found that 89% of subjects were singular, while 11% were plural, which if anything should be expected to produce higher accuracy in the singular subject conditions. For instance, if the model simply assigns higher probability to the more frequent form, this should be correct most of the time in the singular-subject condition. One possibility (suggested by a reviewer) is that when there are conflicting signals about agreement, the models default to the morphologically unmarked plural form.

Another possibility is that this behavior is due to an artifact of how the models tokenize certain verbs we used in our test set. In some cases, the models tokenize a singular verb as two tokens (e.g., *like* and *s* for *likes*). Due to how we used teacher-forcing, this meant that the models were forced to predict identical tokens up until the disambiguating token, which for a word like *like(s)* would be the token following *like*. After this, the models were forced to predict either the singular continuation, *s*, or a token that was the beginning of a word (indicated in the sentence piece tokenizer as tokens that begin with a special unicode character). This

regimen may have masked cases where the models predictions were poor before the verb, leading the model to enter a state where it was being forced to choose the best continuation for a sequence that it considers low probability to begin with. In this case, the following token may have been chosen erroneously, but in the plural conditions, this would still look like the model had correctly predicted the plural verb. Distinguishing between the possibilities will require further investigation.

3.2 Amount and kind of pre-training data

We next consider the effects of pre-training data on agreement behavior while holding model size constant. We consider T5 models with 63M parameters, pre-trained on CHILDES (MacWhinney, 2000), Simple English Wikipedia (simple.wikipedia.org), English Wikipedia (en.wikipedia.org), and C4 (Raffel et al., 2019). Figure 3 shows the proportion of errors by dataset type and size for each condition.

Due to the limited number of models we had available for each source of pre-training data (2 for CHILDES, 3 for Simple English Wikipedia, and 4 each for C4 and English Wikipedia), we classified models as having been pre-trained on either simple English (CHILDES, Simple English Wikipedia) or standard English (C4, English Wikipedia). We fit logistic regressions using the `glmer` function from R's `lme4` library (Bates et al., 2015) with random intercepts and slopes for each individual source of data, with p-values obtained using the `lmerTest` library (Kuznetsova et al., 2017). To address statistical concerns, we used the $\log_{10}$ of the dataset size in words as a predictor.

When predicting errors across all conditions, we found a significant main effect of dataset size ($\beta = -0.18633$, $z = -6.979$, $p < 0.001$), indicating improved performance as the size of the pre-training dataset increases. However, there was no effect of language complexity (i.e., simple vs. standard English) ($\beta = -0.06758$, $z = -0.375$, $p = 0.707$), nor any interaction between complexity and size ($\beta = 0.01620$, $z = 0.465$, $p = 0.642$).

As before, the effect of dataset size was significant in most conditions for both types of models. However, as (3) shows, for models pre-trained on simple English, more data led to a higher error rate in the SP, SSP, and SPP conditions. In contrast, for models pre-trained on standard English, all effects found went in the expected direction. We would urge caution in over-interpreting these results, since even the largest of the datasets we consider here, at 1 billion words, is much smaller than the C4 dataset used to pre-train the T5 Efficient models we consider earlier, which consists of approximately 156 billion tokens (Dodge et al., 2021). While the unit of measurement used to report the size of these datasets differs, it seems clear that the full C4 corpus is roughly 100 times larger than the largest dataset used to pre-train these models. A fuller study of properties of the different corpora used may shed light on this behavior, though this is beyond the scope of this paper.

Nevertheless, we find it interesting that in some cases larger datasets led to increased errors, which may be due to a kind of overfitting to the simpler data that made the models less robust to longer sentences with multiple nouns prior to the main verb. However, notably, these conditions all have singular subjects and plural interveners, which is known to lead to increased agreement errors in people. This leads us to a consideration of whether the kinds of agreement errors the models make are in general like those people make.

3.3 Agreement attraction

Psycholinguistic studies have found some linguistic contexts lead to more agreement errors than others. A common feature of contexts that lead to more of these errors is the presence of a noun that linearly intervenes between the head noun of the subject (the correct target) and the verb that has a different number feature from the correct target. This is a feature in most of our conditions. For example, more agreement errors are produced after preambles like (4b) than after preambles like (4a) (Bock and Cutting, 1992).

(4) a. The key to the cabinet...
b. The key to the cabinets...

Intuitively, the reason (4b) prompts more errors than (4a) is due to the plural noun, *cabinets*. The noun interferes with the correct target of agreement, *key*, leading to increased production of an incorrect plural verb. This kind of error is referred to as *agreement attraction*.

Recent work has examined to what extent language models replicate patterns of human language use (e.g., Arehalli and Linzen, 2020; Brennan et al., 2020; Hao et al., 2020; Wilcox et al., 2021). It is possible the errors of the models we investigate reflect a human-like understanding of agreement. This could be true if errors are disproportionately

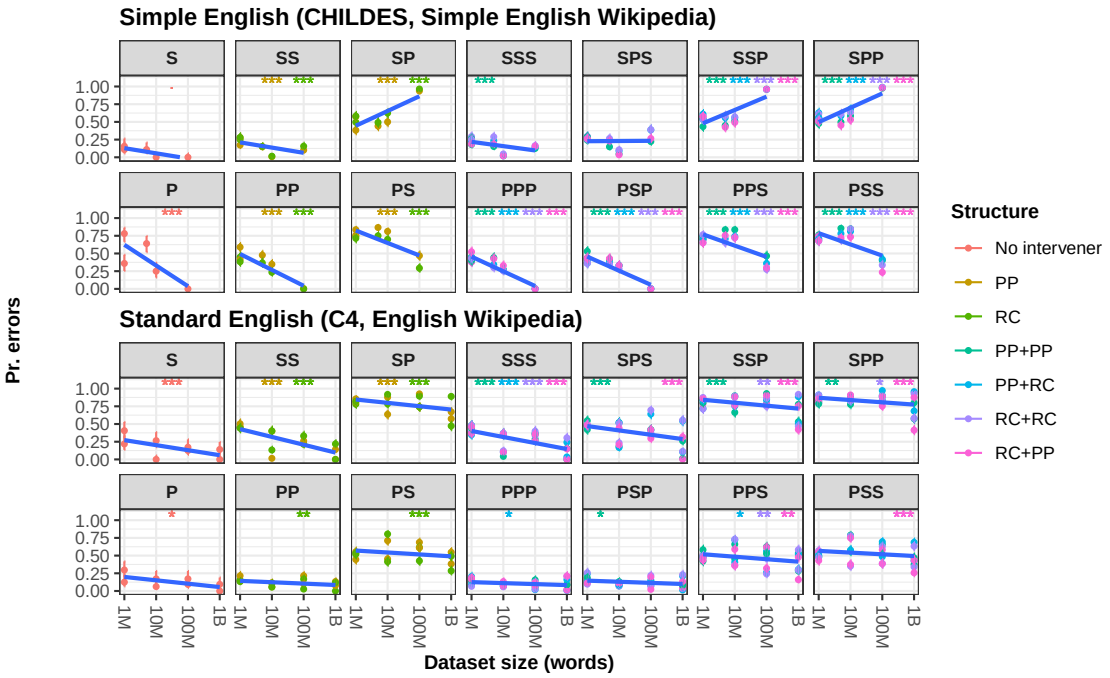


Figure 3: Accuracy by dataset size, type, and condition at each model's best overall checkpoint. Colored stars indicate significance of the corresponding condition with Bonferroni-corrected α.

concentrated in contexts where people make relatively more agreement errors. Arehalli and Linzen (2020) investigated this question with LSTMs pretrained on English Wikipedia. They used preambles taken from psycholinguistic studies of agreement attraction, and measured the models' predictions for *is* or *are* as the following token. Their models replicated some but not all agreement attraction effects. Like people, their LSTMs showed more attraction for distractors in PPs than distractors in RCs, effects of adjacency in coordinate structures, and sensitivity to clause-external distractors. However, unlike people, they were more influenced by linear adjacency than structural proximity, and showed no effect of notional number nor of argument vs. adjunct status of the distractor. We examine the singular-plural asymmetry, structure (PP vs. RC) and linear adjacency (e.g., SPS vs. SSP) to determine how similarly the T5 models we tested behave compared to people.

3.3.1 Singular-plural asymmetry

Bock and Cutting (1992) found that people produce more agreement errors after (4b) than after (5).

(5) The keys to the cabinet...

In other words, more errors arise with singular subjects and plural interveners (SP) than with plural subjects and singular interveners (PS).

Figure 4 shows the difference in the proportion of agreement errors for the SP and PS conditions by model. A positive value indicates more errors in SP than in the PS conditions, and thus a singular-plural asymmetry that goes in the same direction as observed in psycholinguistic experiments.

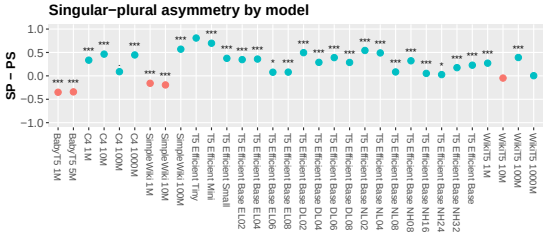


Figure 4: Singular-plural asymmetry by model for the two-noun conditions. Stars indicate significance obtained from χ^2 tests comparing accuracy across the two conditions with Bonferroni-corrected α.

Most models show the same asymmetry as people; the exceptions are BabyT5, SimpleWiki 1M and 10M, and WikiT5 10M (with only the latter difference not statistically significant). The overall pattern is not so surprising given fig. (1), but this shows the differences by model.

3.3.2 Structural context of distractor

In addition to the morphologically-based singular-plural asymmery, Bock and Cutting (1992) also

showed that people were more likely to make errors when the intervener was embedded in a PP (6a) compared to when it was embedded in an RC (6b), a structural asymmetry.

(6) a. The student in the classes...

b. The student who failed the classes...

Figure 5 shows the difference in the proportion of agreement errors for the PP and RC two-noun conditions. A positive value indicates more errors in the PP conditions than in the RC conditions, and thus a PP-RC asymmetry that matches the results of Bock and Cutting (1992).

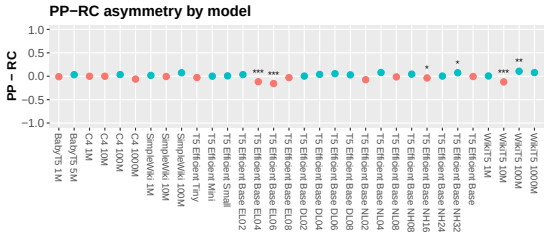


Figure 5: PP-RC asymmetry by model for the two-noun conditions. Stars indicate significance obtained from χ^2 tests with Bonferroni-corrected α.

In this case, 12 of the 32 models showed a numerical asymmetry in the opposite direction compared to people.Of these, only the differences for T5 Efficient Base EL04, EL06, and NH24; and WikiT5 10M are statistically significant. Even for those models with the expected asymmetry, it is less pronounced than the singular-plural asymmetry is in most models, with only two models showing a significant difference in the expected direction (T5 Efficient Base NH32 and WikiT5 100M).

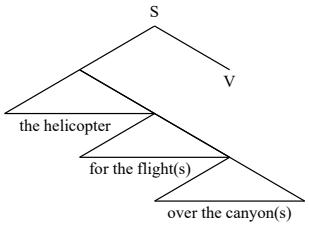
3.3.3 Linear vs. structural proximity

People are more likely to produce agreement attraction errors for distractors that are structurally closer to the verb compared to distractors that are linearly closer but structurally more distant. Franck et al. (2002) found that preambles like (7a) led to more errors than preambles like (7b).

(7) a. The helicopter for the flights over the canyon...

b. The helicopter for the flight over the canyons...

c.



As shown in (7c), the noun that mismatches the subject in number is structurally closer to the verb in (7a) than in (7b). Figure 6 shows three asymmetries that are relevant to this question.

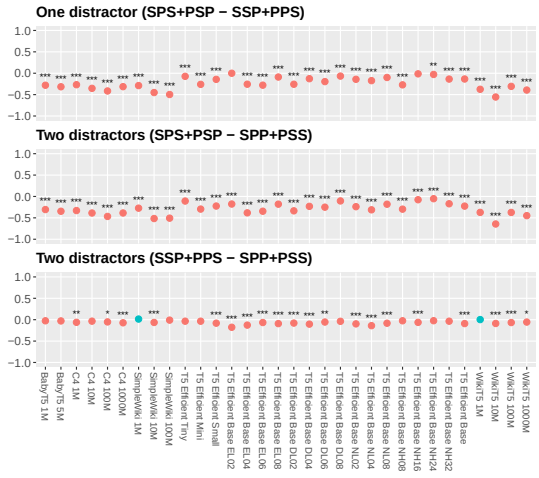


Figure 6: Comparison of multiple-distractor conditions. Stars indicate significance obtained from χ^2 tests comparing accuracy across the two conditions with Bonferroni-corrected α.

The top plot shows the accuracy difference between structural vs. linear closeness for the single-distractor conditions, e.g., SPS (structurally close) and SSP (linearly close). All differences are < 0, indicating that the models' performance is worse when distractors are linearly closer to the verb, with differences for all but two models (T5 Efficient Base EL02 and NH16) being statistically significant. The middle plot shows the difference between the conditions with a single structurally close distractor (e.g., SPS) and conditions with structurally and linearly close distractors (e.g., SPP). Though the SPP and PSS conditions contain structurally and linearly close distractors, Franck et al. (2002) found attraction errors were highest in the single, structurally close distractor conditions, such that, e.g., SPS led to more errors than SPP. The models fail to replicate this pattern, showing worse performance in the multiple distractor conditions than in the single distractor conditions, since all differences are < 0. All of these differences are statistically significant. Finally, the lowest row shows the difference between the single, linearly close distractor conditions and the multiple distractor conditions. A negative value means that the model shows more attraction with two distractors compared to one, unlike Franck et al. (2002)'s results. Nearly all of the models behave this way; the sole exceptions are SimpleWiki 1M and WikiT5 1M. However, the

negative differences for BabyT5 1M and 5M; C4 10M; SimpleWiki 100M; T5 Efficient Tiny, Mini, Base DL08, Base NH08, Base NH24, and Base NH32 are not statistically significant; neither of the positive differences are statistically significant.

In general, unlike what Franck et al. (2002) found, the models are more likely to make attraction errors when distractors are linearly adjacent to the verb compared to when they are structurally adjacent, and they are more likely to make errors when there are multiple distractors that intervene between the subject and the main verb.

A potential confound is that the locus of attachment may be ambiguous in our synthetic data. While Franck et al. (2002) controlled for this by word choice (as shown in (7c), where the alternative "high-attachment" parse of the final modifier would be semantically anomalous), our synthetic test dataset did not. As such, the "correct" parse of the three-noun conditions is potentially ambiguous. Nevertheless, due to how our PCFG was defined, high-and low-attachment parses of the final modifier should be equally plausible. Despite this, we still found significant differences for most models when the distractor was linearly adjacent to the the verb, and when there were multiple distractors. This suggests to us that the models' performance is typically significantly influenced by linear adjacency, since we might have otherwise expected at worst chance performance. Furthermore, Franck et al. (2002) found that for people, there was little difference between the single, structurally close distractor conditions (e.g., SPS and PSP) and the multiple distractor conditions (e.g., SPP and PSS), while the models show significantly higher error rates with multiple distractors. Thus, despite the potential ambiguity, most models behave consistently differently from people in this regard.[9]

4 Conclusion

We examined pre-trained T5 models to determine how model size, architecture, dataset size, and dataset type affected subject-verb agreement on a tense reinflection task. We found that bigger models performed better, especially in singular-subject conditions. In contrast, model performance was already high even for small models in the plural-subject conditions. Increasing the number of layers as well as the number of attention heads per layer result in improvements, though adding encoder layers was associated with greater improvement than adding decoder layers.

When considering the type and amount of pre-training data, we found increasing the amount of pre-training data improved agreement accuracy overall. However, for the models trained on simple English text (CHILDES, Simple English Wikipedia), bigger training datasets led to **worse** performance in singular-subject conditions with linearly-adjacent distractors (e.g., SP, SSP, SPP), despite leading to better performance in plural subject conditions. In contrast, for models trained on standard English (C4, English Wikipedia), more pre-training data uniformly led to increased performance (when performance with small datasets was not already high).

The models did not consistently display patterns reminiscent of agreement attraction. While most models showed a number asymmetry matching what has been found in psycholinguistic work, other asymmetries found in agreement attraction errors were not present. Unlike the LSTMs examined in Arehalli and Linzen (2020) and unlike the results of Bock and Cutting (1992), only some of the transformer models we considered produced more errors in PP than in RC conditions. However, similarly to Arehalli and Linzen (2020)'s LSTMs, the transformer models still showed more attraction for linearly adjacent distractors compared to structurally closer distractors, in addition to showing worse performance with multiple distractors.

Our results show both the advantages and limitations of increasing the size of models and datasets. While increases in both of these independently lead to better performance on subject-verb agreement, an indirect indicator of hierarchical knowledge of language, not even the largest models we considered, nor those pre-trained on the largest amounts of data, display fully human-like behavior. Instead, they were still susceptible to linear interference to a much greater degree than people are (cf. Petty and Frank, 2021). It appears these perennial issues of hierarchical vs. linear generalization with regards to language modeling remain a concern for transformers even now.

[9] We have also conducted preliminary investigations on the models' performance using a span-denoising task on the actual stimuli used in Franck et al. (2002), and found that even on those stimuli, the models display essentially the same sensitivity to linear over structural proximity, though we have not yet conducted statistical tests.

Acknowledgments

We would like to thank Aaron Mueller for sharing his T5 models with us. We also thank the members of the Computational Linguistics at Yale lab and three anonymous reviewers for suggestions and feedback. This work was made possible by support from the National Science Foundation grant BCS-1919321.

References

Suhas Arehalli and Tal Linzen. 2020. Neural language models capture some, but not all, agreement attraction effects. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 370–376.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Kathryn Bock and J. Cooper Cutting. 1992. Regulating mental energy: Performance units in language production. *Journal of Memory and Language*, 31:99–127.

Jonathan R. Brennan, Chris Dyer, Adhiguna Kuncoro, and John T. Hale. 2020. Localizing syntactic predictions using recurrent neural network grammars. *Neuropsychologia*, 146.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Julie Franck, Gabriella Vigliocco, and Janet Nicol. 2002. Subject-verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes*, 17(4):371–404.

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

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *Proceedings of the Cognitive Modeling and Computational Linguistics (CMCL) Workshop (EMNLP)*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26.

Karim Lasri, Alessandro Lenci, and Thierry Poibeau. 2022. Does BERT really agree ? fine-grained analysis of lexical dependence on a syntactic task. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2309–2315, Dublin, Ireland. Association for Computational Linguistics.

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.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, third edition, volume II: The Database. Lawrence Erlbaum Associates, Mahwah, NJ.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *CoRR*, abs/1808.09031.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140.

Aaron Mueller, Robert Frank, Tal Linzen, Luheng Wang, and Sebastian Schuster. 2022. Coloring the blank slate: Pre-training imparts a hierarchical inductive bias to sequence-to-sequence models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1352–1368, Dublin, Ireland. Association for Computational Linguistics.

Karl Mulligan, Robert Frank, and Tal Linzen. 2021. Structure here, bias there: Hierarchical generalization by jointly learning syntactic transformations. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 125–135, Online. Association for Computational Linguistics.

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. 2021. Refining targeted syntactic evaluation of language models. In *Proceedings of the 2021 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, pages 3710–3723, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Jackson Petty and Robert Frank. 2021. Transformers generalize linearly. *CoRR*, abs/2109.12036.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *Journal of Machine Learning Research*, 13(66):2063–2067.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. 2021. Scale efficiently: Insights from pre-training and fine-tuning transformers. *CoRR*, abs/2109.10686.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Ethan Wilcox, Pranali Vani, and Roger Levy. 2021. A targeted assessment of incremental processing in neural language models and humans. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 939–952, Online. Association for Computational Linguistics.