# Look at that! BERT can be easily distracted from paying attention to morphosyntax

**Rui P. Chaves**
Department of Linguistics
University at Buffalo, SUNY
rchaves@buffalo.edu

**Stephanie N. Richter**
Department of Linguistics
University at Buffalo, SUNY
snrichte@buffalo.edu

## Abstract

Syntactic knowledge involves not only the ability to combine words and phrases, but also the capacity to relate different and yet truth-preserving structural variations (e.g. passivization, inversion, topicalization, extraposition, clefting, etc.), as well as the ability to infer that these syntactic variations all adhere to common morphosyntactic rules, like subject-verb agreement. Although there is some evidence that BERT has rich syntactic knowledge, our adversarial approach suggests that it is not deployed in a robust and linguistically appropriate way. English BERT can be tricked to miss even quite simple syntactic generalizations, when compared with GPT-2, underscoring the need for stronger priors and for linguistically controlled experiments in evaluation.

## 1 Introduction

BERT (Devlin et al., 2019) has achieved very high-quality results for a wide range of language processing tasks, and there is growing evidence that BERT's internal representations are linguistically rich (Tenney et al., 2019; Lin et al., 2019), and include entire syntax trees implicitly embedded in its deeper levels (Coenen et al., 2019; Clark et al., 2019; Hewitt and Manning, 2019; Jawahar et al., 2019; Manning et al., 2020).[1]

However, syntax is not just about arriving at the correct sentence structure. It is also about the process by which structures are recursively built, and how superficially different constructions are related to each other in deep and complex ways. A model that has acquired rule-like knowledge should be robust and prone to identifying generalizations across related structure types, in contrast to a model that has instead learned superficial heuristics. Indeed, McCoy et al. (2019) shows that BERT can adopt shallow heuristics which end up being valid for frequent patterns, but invalid for less frequent ones. Moreover, Htut et al. (2019) were unable to extract parse trees from BERT's attention heads, even with the gold root annotations, and Yu and Ettinger (2020) found little evidence of composition of phrasal representations.

In this work we focus on whether the large (340 million-parameter) and base (110 million-parameter) English BERT models deploy syntactic constraints in a linguistically appropriate way, using (adversarial) linguistic experiments. Our results suggest that both pre-trained and fine-tuned models fail to robustly deploy basic English morphosyntactic rules about subject-verb agreement and syntactic transformation phenomena despite the rich latent grammatical knowledge that such models supposedly have. In Experiments 1 and 2 we show that BERT is misled by the mere presence of distracting structures that precede (and therefore are completely irrelevant to) the agreeing subject and verb. In Experiment 3 we go further and show that although BERT seems to be able to correctly classify complex syntactic paraphrases, it does so in a completely unnatural way, ignoring the particular words in sentences. We conclude that BERT's linguistic abilities are difficult to deploy in a linguistically motivated way, and are prone to instead be rather brittle and shallow, adding to similar findings by Htut et al. (2019) and Yu and Ettinger (2020).

## 2 Evaluating the robustness of BERT's morphosyntactic knowledge

Goldberg (2019) probed BERT's ability to compute subject-verb agreement dependencies using both naturally-occurring and manually-created stimuli drawn from Linzen et al. (2016), Gulordava et al. (2018), and Marvin and Linzen (2018).

[1]But see Wu et al. (2020) who conclude that what BERT learns is not very similar to linguistic annotated resources.



*Proceedings of the Society for Computation in Linguistics (SCiL) 2021*, pages 28-38.
Held on-line February 14-19, 2021

For example, in relative clauses like (1) Goldberg (2019) extracted the probabilities of the masked token being *is* vs. *are*, and so on.

(1) The game that the guard hates [MASK] bad.

Goldberg (2019) found that (large and base) BERT's ability to accurately predict the correct agreement verb form for the masked slot was very high – higher than that of a state-of-the-art LSTM RNN, and close to human-level performance. Adopting a similar psycholinguistic paradigm, we report experiments designed to test the robustness of BERT's syntactic ability. Throughout, we used the *Transformers* library (Wolf et al., 2020).

### 2.1 Experiment 1: local agreement

Adopting a $2 \times 2 \times 5$ factorial design, we constructed 5040 sentences varying systematically according to: (i) the number inflection of the subject noun (target) adjacent to the respective (masked) verb; (ii) the number inflection of structurally higher nouns (attractors); and (iii) the number of clausal embeddings (E) from 0 to 4. The examples in (2) serve to illustrate the conditions. The attractor nouns are underlined and the target nouns are bold, for ease of exposition. Attractors always had the same number inflection as each other.

(2) a. The **lawyer(s)** [MASK] upset.
[Attractors-pl/sg, E0, Target-sg/pl]

b. The engineer(s) handling the asbestos removal(s) from our building(s) said the **lawyer(s)** [MASK] upset.
[Attractors-pl/sg, E1, Target-sg/pl]

c. The engineer(s) handling the asbestos removal(s) from our building(s) said that I/we thought the **lawyer(s)** [MASK] upset.
[Attractors-pl/sg, E2, Target-sg/pl]

d. The engineer(s) handling the asbestos removal(s) from our building(s) said someone/people claimed that I/we thought the **lawyer(s)** [MASK] upset.
[Attractors-pl/sg, E3, Target-sg/pl]

e. The engineer(s) handling the asbestos removal(s) from our building(s) said that someone/people claimed that I/we thought the audience believed the **lawyer(s)** [MASK] upset.
[Attractors-pl/sg, E4, Target-sg/pl]

| N number | Embedding | S(V-sg) vs. S(V-pl) |
|---|---|---|
| N-sg | 0 | $< 2.2e{-}16$ * |
| N-pl | 0 | $< 2.2e{-}16$ * |
| N-sg | 1 | $= 0.0002$ * |
| N-pl | 1 | $< 6e{-}5$ * |
| N-sg | 2 | $= 0.1$ |
| N-pl | 2 | $< 5e{-}7$ * |
| N-sg | 3 | $= 0.1$ |
| N-pl | 3 | $< 5e{-}8$ * |
| N-sg | 4 | $= 0.5$ |
| N-pl | 4 | $< 1e{-}7$ * |

Table 1: Large uncased BERT verb form surprisal differences ($p$-values), according to target noun number inflection, up to 4 levels of clausal embedding

All items were based on attested sentences from English corpora (COCA, BNC, and Brown), manipulated so that the target subject noun (in bold) is always adjacent to the respective masked verb that must agree with it. Thus, the agreement should be trivial to model. The presence of higher clauses containing other nominals (the attractors, underlined) should have no effect on the inflection of the masked verb since the actual subject is always immediately adjacent to the agreeing masked verb, and attractors are further away.

We fed our items into BERT (base/large uncased) and extracted the softmax activation of 7 singular linking verb forms (i.e., *is*, *was*, *seems*, *gets*, *becomes*, *looks*, and *sounds*) as well as that of the corresponding 7 plural verb forms (i.e., *are*, *were*, *seem*, *get*, *become*, *look*, and *sound*). Next, we transformed the activations to surprisal measures (Hale, 2001; Levy, 2008; Smith and Levy, 2008). Following Wilcox et al. (2018), the surprisal $S(w)$ of a word $w$ was estimated as the log of the inverse probability of $w$ according to the softmax activation $h$ before consuming $w$, given all other words in the sentence:

(3) $S(w) = -log_2(p(w|h))$

A large surprisal value corresponds to low probability; a small surprisal value corresponds to high probability. Although both base and large BERT models obtain perfect accuracy on the baseline items like (2a), the model systematically fails to model subject-verb agreement when the target subject noun is singular, across levels of clausal embedding 2–4 as shown in Table 1. The last column shows the $p$-values of one-tailed $t$-tests pitting the surprisal values of the correct verb forms against those of the incorrect verb forms.

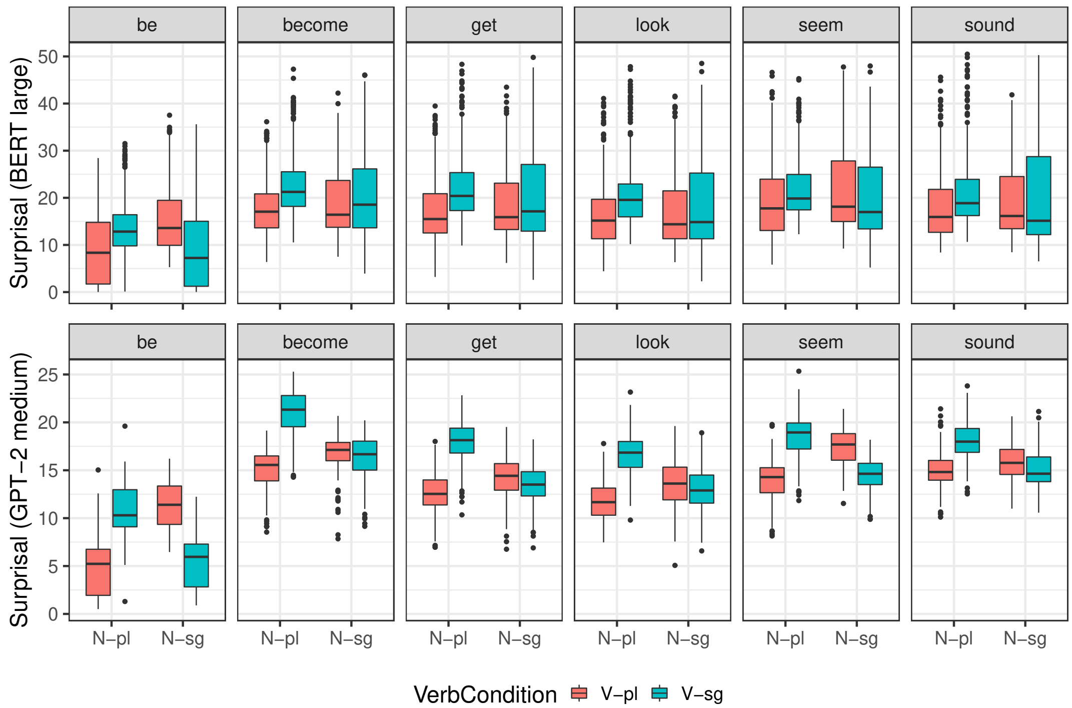


Figure 1: Overall verb form surprisal and target subject noun agreement (top: BERT large; bottom: GPT-2 med.)

| Verb | S(V-sg) < S(V-pl) :N-sg | S(V-pl) < S(V-sg) :N-pl |
|---|---|---|
| *be* | $< 2.2e{-}16$* | $< 1.303e{-}15$* |
| *become* | 0.95 | $< 1.697e{-}10$* |
| *seem* | 0.07 | 0.003* |
| *get* | 0.91 | $1.286e{-}8$* |
| *sound* | 0.83 | 0.0001* |
| *look* | 0.93 | $5.605e{-}7$* |

Table 2: BERT large verb surprisal significance

| Verb | S(V-sg) < S(V-pl) :N-sg | S(V-pl) < S(V-sg) :N-pl |
|---|---|---|
| *be* | $< 2.2e{-}16$* | $< 2.2e{-}16$* |
| *become* | 0.04* | $< 2.2e{-}16$* |
| *seem* | $< 2.2e{-}16$* | $< 2.2e{-}16$* |
| *get* | $6.815e{-}5$* | $< 2.2e{-}16$* |
| *sound* | $3.195e{-}5$* | $< 2.2e{-}16$* |
| *look* | 0.02* | $< 2.2e{-}16$* |

Table 3: GPT-2 medium verb surprisal significance

Further analysis indicates that BERT's performance is very sensitive to the verb, as shown in the top panel of Figure 1, and in Table 2. Although BERT's accuracy for *was/were* is statistically significant in the correct direction across all embedding levels and conditions, for all other verbs BERT systematically fails to identify the right verbal agreement form when the subject is N-sg, in all embedding levels. Thus, BERT is good at subject-*be* agreement, but not so good at subject-verb agreement in general, perhaps because past tense indirect speech verbs like *say* are more likely to combine with present tense *be* clauses than with other linking verbs (in both Google N-grams and COCA, the sequence 'X said Y seems' is 2–3 orders of magnitude less frequent than 'X said Y is'). See Marvin and Linzen (2018) for a similar result and explanation for LSTM agreement errors.

For comparison purposes, we ran Experiment 1 on the (small) 124 million and the (medium) 345 million-parameter pre-trained English GPT-2 models (Radford et al., 2019).[2] We computed the surprisal of the critical verbs from their softmax activation after giving the model the preceding words in the sentence, up to the target noun. GPT-2 (small/medium)'s performance was excellent in predicting the correct verb forms. All surprisals were significant in the expected direction, overall and across all embeddings (all $p$'s $< 6.87e{-}7$). See bottom panel of Figure 1 and Table 3. GPT-2 may be more robust in part because it is prone to memorize rare sequences (Carlini et al., 2020).

[2]Although BERT and GPT-2 cannot be directly compared since they were trained with different objectives and on different datasets, the results at the very least suggest that the present subject-agreement task is not fundamentally problematic for other transformer models of like size.

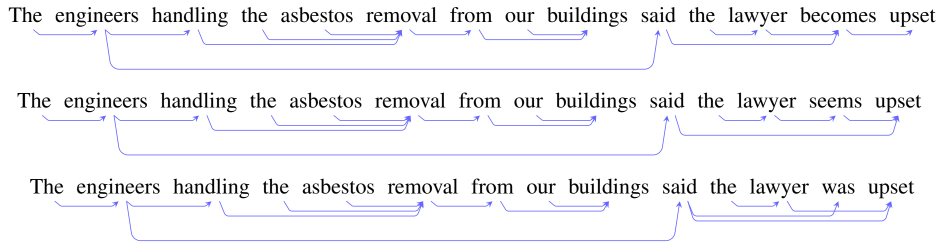


Figure 2: Inconsistent dependency parses obtained via structural probing (BERT large)

BERT's poor performance might have to do with the quality of syntactic representations (Wu et al., 2020). Indeed, using the structural probing of Hewitt and Manning (2019) to examine our items, we found that the subordinate clauses often get different parses depending on the verb, as Figure 2 illustrates for *becomes, seems,* and *was*.

Another reason for BERT's shortcomings is that it may be attending to the wrong parts of the input. Abnar and Zuidema (2020) propose to visualize attention in transformers by viewing the model as an attention graph in which the attention weights are flow capacities. The maximum flow algorithm can then be used to compute the maximum attention flow from any node in any of the layers to any of the input nodes. Thus, the weight of a single path is the minimum value of the weights of the edges in the path. This maximum-flow-value works as an approximation of the attention to input nodes, and can indicate the set of input tokens that are important for the model's final decision. Abnar and Zuidema (2020) show that attention flow is superior to inspecting raw attention weights, especially in deeper layers of the network.[3] We therefore performed an analysis of attention flow for BERT using our items, and the results suggest that the function words in the attractor region have an undue influence on the verb prediction. For example, in Figure 3 we can see that there are at least three words preceding the target subject which contribute a lot of information to the prediction of the masked verb. BERT's hypersensitivity to verb forms and to irrelevant expressions in higher subordinating clauses suggests it has not truly learned subject-verb agreement.

### 2.2 Experiment 2: long-distance agreement

In this second experiment, we go one step further and probe the ability of BERT and GPT-2 to

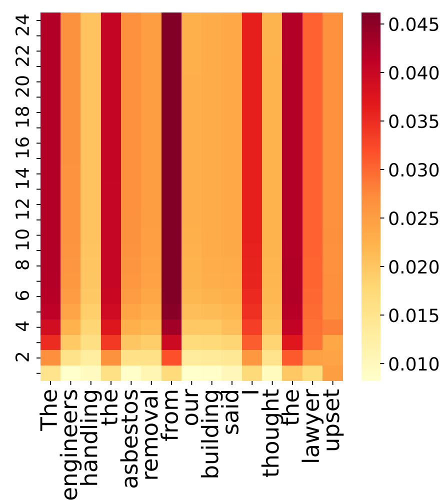


Figure 3: In large uncased BERT, attention flow to the masked verb comes from words in the higher clauses

compute subject-agreement when the target noun and the verb are separated by a long-distance dependency. For example, in (4) the singular noun *lawyer* is the agent of the verb in bold font. Therefore, the two must agree in person and number, as if the noun had been realized at the '_' gap site instead, three clauses deep into the sentence.

(4) a. It was the **lawyer** who [the witness stated [that Alex said [ _ **was**/***were** upset]]].

b. It was the **lawyers** who [the witness stated [that Alex said [ _ **were**/***was** upset]]].

According to Jawahar et al. (2019) and Da Costa and Chaves (2020), BERT can handle subject-verb agreement, even in long-distance dependencies like (4). To test the robustness of BERT's ability to compute agreement in such constructions in a more adversarial way, we adopted a similar design to Experiment 1, but with a few differences. The attractor was the matrix subject, the complexity of which had three conditions: 'none', 'simple', and 'complex'. In (5) we illustrate items

[3] Although some research suggests attention does not explain BERT's predictions; see Rogers et al. (2020).

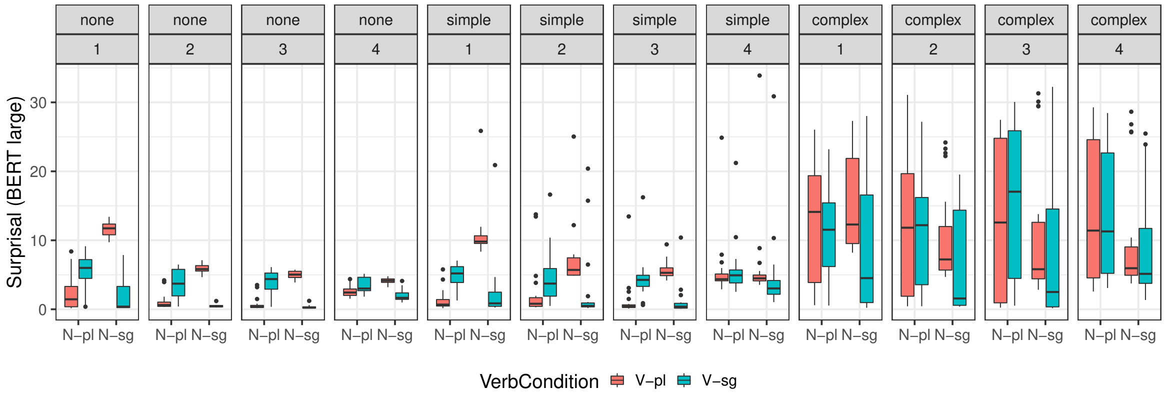


Figure 4: Bert large (uncased) verb form surprisal given a singular/plural extracted subject filler phrase, separated by embedded clauses (1 to 4), and with matrix subjects of varying complexity (none, simple, complex)

in the 'simple' condition. Finally, the inflection in the matrix attractor subject nouns (underlined) was always mismatched with that of the target subject nouns (in bold).

(5) a. Some employer(s) decided that it was the **cousin(s)** who I think [MASK] upset.
[simple, D-pl/sg, E1, Target-pl/sg]

b. Some employer(s) decided that it was the **cousin(s)** who I think you said [MASK] upset.
[simple, D-pl/sg, E2, Target-pl/sg]

c. Some employer(s) decided that it was the **cousin(s)** who I think you said you thought [MASK] upset.
[simple, D-pl/sg, E3, Target-pl/sg]

d. Some employer(s) decided that it was the **cousin(s)** who people believe I think you said you thought [MASK] upset.
[simple, D-pl/sg, E4, Target-pl/sg]

The 'none' condition items simply lacked the matrix subordinator clause (i.e. the 'none' counterparts of (5) lacked *some employer(s) decided that*). For all items in the experiment, the material embedding clauses between the (target) extracted subject and the masked verb were very simple, with pronominal or bare nominal subjects (e.g., ... *who people believe I think you said ...*) and never exceeded 8 words. In (6) are items in the 'complex' condition. The masked verbs were *was* and *were*, for a total of 960 stimuli (20 per condition).

(6) a. The guy(s) who did not even warm up with the team prior to the game said it was the **cousin(s)** who I think [MASK] upset.
[complex, D-pl/sg, E1, Subj-pl/sg]

b. The guy(s) who did not even warm up with the team prior to the game said it was the **cousin(s)** who I think you said [MASK] upset.
[complex, D-pl/sg, E2, Subj-pl/sg]

c. The guy(s) who did not even warm up with the team prior to the game said it was the **cousin(s)** who I think you said you thought [MASK] upset.
[complex, D-pl/sg, E3, Subj-pl/sg]

d. The guy(s) who did not even warm up with the team prior to the game said it was the **cousin(s)** who people believe I think you said you thought [MASK] upset.
[complex, D-pl/sg, E4, Subj-pl/sg]

As in Experiment 1, the presence of a subordinating clause and the complexity of its subject phrase should be irrelevant for the subject-verb agreement dependency between the target noun and the masked verb, located in the subordinate clause. If BERT can model the subject-agreement dependencies in question, then the surprisal of a singular verb like *was* in the presence of a plural subject filler phrase should be higher than that of *was* in the presence of a singular subject filler phrase. Conversely, the surprisal of a plural verb like *were* in the presence of a singular subject filler phrase should be higher than that of *were* in the presence of a plural subject filler phrase. The results are in Figure 4. BERT models subject-agreement correctly for the baseline condition (E1 through E4) and for the simple condition up to E3. This is consistent with prior studies (Jawahar et al., 2019; Da Costa and Chaves, 2020). However, BERT systematically fails to identify the cor-

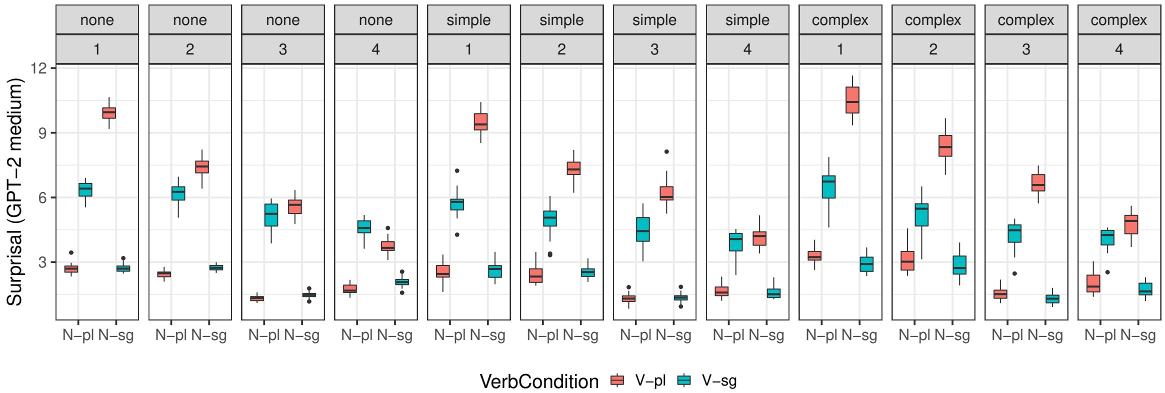


Figure 5: GPT-2 medium verb form surprisal given a singular/plural extracted subject filler phrase, separated by embedded clauses (1 to 4), and with matrix subjects of varying complexity (none, simple, complex)

| **Verb** | **BERT (large)** | **GPT-2 (medium)** |
|---|---|---|
| *be* | 80% | 100% |
| *become* | 50% | 72% |
| *get* | 53% | 92% |
| *look* | 56% | 82% |
| *seem* | 68% | 100% |
| *sound* | 65% | 87% |

Table 4: Verb form prediction accuracy (E4, complex)

rect verb inflection form in the complex condition. In other words, agreement patterns like (5d) and (6) are not correctly captured.

Analogously to Experiment 1, we ran the same stimuli through GPT-2 (small and medium). We computed the surprisal of *was* and *were* after feeding the model the preceding words in the sentence. GPT-2's performance was excellent, with all conditions statistically significant in the expected direction (all $p$'s $< 2.57e{-}15$). GPT-2 was again immune to the presence of attractors, as Figure 5 illustrates. As in Experiment 1, BERT is far more sensitive to the verb, as can be seen by the accuracy results for the language models shown in Table 4. These results are consistent with other research suggesting that GPT-2 is linguistically more robust than BERT, such as Warstadt et al. (2020) and Da Costa and Chaves (2020).

To obtain a rough estimate for a human baseline accuracy, we recruited 52 participants via the Amazon Mechanical Turk marketplace, and asked them to select the correct singular/plural *be* verb form for the 40 masked items in the complex E4 condition (interspersed with 40 filler sentences). The human accuracy was 77% overall (74% for the plural verb condition and 80% for the singular verb condition). So, although BERT does very well for *be*, its accuracy plunges unreasonably for rarer verbs, which argues against a rule-like generalization for the computation of agreement.

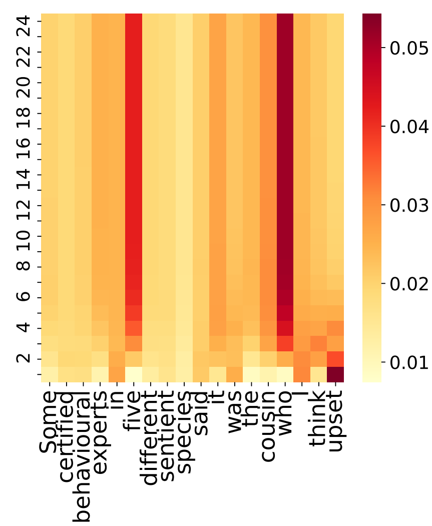


Figure 6: Attention flow targeting masked verb

Post hoc analysis of BERT's attention flow suggests that the matrix subject plays an undue role in the prediction of the masked verb in the embedded clause, as illustrated in Figure 6. The only linguistically relevant noun for the verb agreement should be *cousin*, but for BERT, parts of the complex higher subject (e.g. *five*) play a bigger role.

### 2.3 Experiment 3: syntactic paraphrase task

To further probe BERT's morphosyntactic abilities, we fine-tuned it with two datasets, separately. The first dataset consisted of 30K syntactic paraphrase sentence pairs from PAWS (Zhang et al., 2019), where half were true paraphrases. Of the 30K pairs, 15% were randomly extracted as a test set. The second dataset was our own, and consisted of syntactic paraphrases exhibiting a much wider range of complex syntactic transformations, including passivization, object/subject clefts, inversion, and multiple combinations thereof. One remarkable property of filler-gap dependencies is that they can compound with a wide variety of other syntactic transformations, while preserving truth-conditional equivalence. Thus, the sentences in (7) are paraphrases of each other: in any given context, they are either all true or all false.

(7) a. I think Sam saw Alex.
(active)

b. Who I think Sam saw was Alex.
(active + object *wh*-cleft)

c. I think Alex was who Sam saw.
(active + (reversed) object *wh*-cleft)

d. I think Alex was seen by Sam.
(passive)

e. Who I think was seen by Sam was Alex.
(passive + subject *wh*-cleft)

f. Alex was who I think was seen by Sam.
(passive + (reversed) subject *wh*-cleft)

By reversing the nominal arguments, we can create sentences that are not paraphrases of (7) but which have very high lexical overlap, e.g., *I think Alex saw Sam* or *Who I think was seen by Alex was Sam*. To our knowledge, the compounding of syntactic transformations has not been systematically included in any paraphrase training dataset, including datasets created with the specific goal of generating text satisfying certain structural requirements (Iyyer et al., 2018; Wiseman et al., 2018; Colin and Gardent, 2018; Zhang et al., 2019). In all datasets we know of, any syntactic transformations are clause-bounded, i.e., locally restricted to word reordering within the same clause. This is illustrated in the sentence pairs in (8), from the PAWS (Zhang et al., 2019) dataset.

(8) a. When comparable rates of flow can be maintained, the results are high. The results are high when comparable flow rates can be maintained.

b. Lebilaango is a city in the central Somalia region of Hiran. Lebilaango is a town in the central Somalia region of Hiran.

### 2.3.1 Syntactic paraphrase generation

We generated our paraphrase dataset as follows. First, 11761 animate NPs (with 1 to 38 words comprising each NP, averaging 4.6 words) were extracted from various parsed corpora (COCA, BNC, etc.), along with 1033 two-argument verbs. 84 adjunct PPs, 47 parenthetical insertions, and 87 sentence embedding strings were also variably included in any combination (including none).

Next, we randomly generated over 6K sentence seeds (roughly) of the form $NP_1$ VB $NP_2$, where $NP_{1,2}$ are unique nominal phrases. For each of these 6K sentence seeds, we generate 28 total sentences (active/passive vs. (reversed/nonreversed) it-/*wh*-cleft, subj/obj extraction, embedded/nonembedded) using a controlled generator that ensures the correct structure and inflection of the phrasal components. These sentences are divided into two groups of 14 sentences; each group is identical, except the ordering of the NPs is swapped relative to the other group. Thus, all sentences within one group are paraphrases, yet are reciprocal non-paraphrases across groups. Examples of these paraphrases and non-paraphrases are in (9a) and (9b) respectively.

(9) a. [CLS] Tim thought that a career expert was going to tell the nearest stranger who he was, what he had done, and what he proposed to do here. [SEP] A career expert was who Tim thought was going to tell the nearest stranger who he was, what he had done, and what he proposed to do here.
(paraphrases)

b. [CLS] James claimed each registered nurse, midwife or health visitor surprised some workers at the newspaper. [SEP] Who James claimed was surprised by some workers at the newspaper was each registered nurse, midwife or health visitor.
(non-paraphrases)

Training and test sets were created by randomly sampling two sentences from each block of 28 sentences, with replacement, for all items. Sentences coming from the same group were labeled '1' (therefore, paraphrases); those coming from opposite groups were labeled '0' (therefore, non-paraphrases). In total, 5292 sentence pairs for training were randomly generated at a time, as well as 588 sentence pairs for testing.

Throughout, no NPs or (non-sentence embedding) verbs are repeated, although there are NPs with similar words. This was intended to force BERT to deploy abstract syntactic knowledge, rather than rely on token distribution. The generated sets consist of 5880 total sentence pairs, 50% of which were paraphrases like (9a), labeled as ‘1’, and 50% of which that were not, like (9b), labeled as ‘0’. Sentence pairs were separated with [SEP] tokens, and [CLS] tokens were inserted at the beginning of every input sentence pair.

We fine-tuned base uncased BERT for binary classification multiple times, and on multiple re-generated training/test datasets, for 4–6 epochs, on a Tesla T4 GPU, using Adam (Kingma and Ba, 2015) ($\eta$ =5$e$−5, $\varepsilon$ =1$e$−8), and batch size of 32.

#### 2.3.2 Fine-tuning results

For our syntactic paraphrase dataset, BERT tended to converge by epoch 5, and for the 25K PAWS training data, it tended to converge by epoch 7. In both cases, the fine-tuned models respectively achieved over 98% and 92% accuracy on their test sets. The high accuracy on our paraphrase dataset is particularly striking given the (apparent) syntactic complexity of the task, and given the minimal lexical overlap between training and test stimuli.

We next fed the models fine-tuned with our complex paraphrases with sentence pairs that were more complex than those which the model was exposed to during training. Recall that in the original training and test sets, there was at most one clausal embedding between the filler phrase and the gap site, as in (9). We thus hand-constructed 365 sentence pairs (50% true paraphrases) with the same multifactorial design as the original dataset, except that all items either had 1, 2, or 3 clausal embeddings (E) between the *wh*-phrase and the gap site. An E3 paraphrase pair is seen in (10).

(10) [CLS] Who we hoped the reporters wrote that Mary thought the girl in red slapped yesterday was the cashier from the corner store. [SEP] We hoped the reporters wrote that Mary thought the girl in red slapped the cashier from the corner store yesterday.

Surprisingly, BERT performs very well, obtaining a mean accuracy of 91%, across all three levels of clausal embedding, even though the stimuli contained longer dependencies than those with which the model was originally fine-tuned with. Using Bertviz (Vig, 2019), we extracted attention values from the fine-tuned models for various sentences and found that deeper layers usually showed evidence that the fronted *wh*-phrase specifically attends to the verb that it is an argument of. Figure 7 illustrates this, where *who* attends to the verb *slapped*. Such a set of results is impressive, but a closer look reveals a more complex picture.

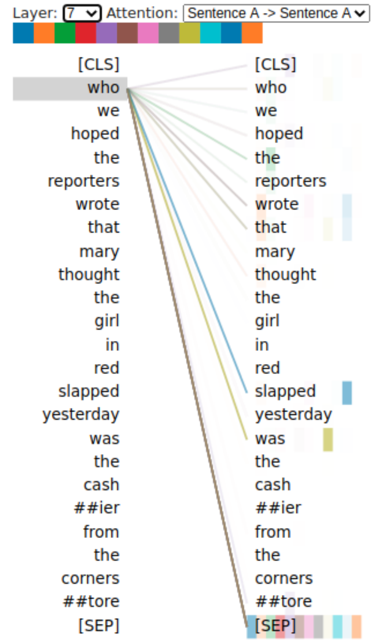


Figure 7: Fine-tuned base uncased BERT’s attention

#### 2.3.3 Adversarial testing results

Adversarial examples are inputs that are designed to cause poor performance in machine learning models, and have been used in various kinds of natural language processing tasks including text classification, machine translation, and question answering. Using a word-level adversarial approach, we found that replacing one word for another with an incompatible meaning had little effect on the models’ behaviour. For example, in the word-flip in (11), the second occurrence of the noun *nun* was replaced with *cop*. Yet, the fine-tuned BERT models still deemed the sentence pair a paraphrase. Replacing the verb *insulted* with *met* similarly yielded no change in classification.

(11) [CLS] It was a nun who a student insulted. [SEP] A student insulted a cop.

We took the original test sets that were used to evaluate the fine-tuning stage and changed at random one noun, adjective, or verb, in only one of the sentences in each pair, creating new test sets of the same size as the original, but now entirely consisting of non-paraphrases, as in (12).

(12) a. [CLS] A senior water scientist was who the Ecuadorian government educated. [SEP] Who a senior water scientist was energized by was the Ecuadorian government.

b. [CLS] It was a viable minority candidate who a retired wind was quizzed by. [SEP] who a viable minority candidate quizzed was a retired player.

c. [CLS] A hard worker with a quick mind was who was equipped by silly authorities. [SEP] It was a hard worker with a quick mind who French authorities equipped.

The word replacement candidates were found by using spaCy's `en_core_web_lg` word similarly model, with the requirement that the word replacement and the original work have a similarity score somewhere between .15 and .70, and have the same part-of-speech. We found that 100% of sentence pairs that were originally paraphrases and suffered a word flip were systematically still classified as paraphrases, with an overall accuracy drop to at-chance performance. In sum, the fine-tuned BERT model was completely insensitive to word substitutions. We subsequently discovered that switching the order of subject+verb sequences was enough to lead our models to systematically miss-classify the input. Thus, BERT deems pairs like (13) as paraphrases, despite the ordering mismatch between the underlined strings.

(13) a. [CLS] It was those three players who Mary thought we denied that the kids replaced yesterday. [SEP] We denied Mary thought that the kids replaced those three players.

b. [CLS] Someone claimed Mia thought that Sam wrote that a lady in red mentioned these scientists yesterday. [SEP] It was a lady in red who someone claimed Sam wrote that Mary thought that mentioned these scientists.

The BERT model fine-tuned on the 25K PAWS dataset was just as susceptible to word-level adversarial attacks on its test set, with a drop to 55% accuracy. Even though the PAWS dataset included word reversals and synonym replacements, as in (8b), it is possible that the high amount of lexical overlap between paraphrase sentence pairs easily misled BERT into adopting fallible syntactic heuristics, like those explicitly identified by McCoy et al. (2019) on a similar BERT fine-tuning task, whereby the model is distracted by parts of the input, and deems others of little importance.

## 3 Conclusions

Although there is evidence that BERT's internal representations are syntactically rich (Tenney et al., 2019; Lin et al., 2019; Coenen et al., 2019; Clark et al., 2019; Hewitt and Manning, 2019; Jawahar et al., 2019), there is also a growing body of research suggesting that BERT does not deploy that knowledge effectively (Htut et al., 2019; Yu and Ettinger, 2020). In particular, our results indicate that BERT deploys shallow (though usually effective) processing heuristics when processing English morphosyntactic dependencies. We found that BERT fails to correctly model surprisingly trivial subject-verb agreement patterns, such as those in Experiment 1, in contrast to GPT-2. This should not be taken to mean that GPT-2 is linguistically richer than BERT, since its superior performance may simply be the result of the model's tendency to memorize rare sequences (Carlini et al., 2020). Further research is needed to probe GPT-2's linguistic representations.

Our results suggest BERT misses significant abstract syntactic generalizations, and likely deploys excessively shallow heuristics like those discussed in McCoy et al. (2019). This outcome underscores the need for multi-pronged linguistically controlled experiments in the evaluation of language models, as advocated by Warstadt et al. (2019), among others.

All data and code are at https://osf.io/ad62v/.

## 4 Authorship contribution statement

R.C. designed, ran, and analyzed all experiments, and created the plots. S.R. implemented the syntactic paraphrase generator (Experiment 3).

## 5 Acknowledgements

We thank Erika Bellingham, Liz Soper, and the SCiL anonymous reviewers for their comments, suggestions, and criticism. All remaining errors and shortcomings remain with the authors.

## References


Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? An analysis of BERT's attention. In *Black-BoxNLP*, page pp.11. Association for Computational Linguistics.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. In *Visualization or Exposition Techniques for Deep Networks*, page pp.8.

Emilie Colin and Claire Gardent. 2018. Generating syntactic paraphrases. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 937–943, Brussels, Belgium. Association for Computational Linguistics.

Jillian K. Da Costa and Rui P. Chaves. 2020. Assessing the ability of transformer-based neural models to represent structurally unbounded dependencies. In *Society for Computation in Linguistics*, volume 3, page 189–198.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. Unpublished ms. `https://arxiv.org/pdf/1901.05287.pdf`.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL*, pages 1195–1205.

John T. Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL-2001, Pittsburg, PA*, pages 159–166. ACL.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. Association for Computational Linguistics.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R. Bowman. 2019. Do attention heads in bert track syntactic dependencies? Unpublished manuscript.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of NAACL-HLT 2018*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. 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.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Ms.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. In *Transactions of the Association for Computational Linguistics*.

Nathaniel J. Smith and Roger Levy. 2008. Optimal processing times in reading: a formal model and empirical investigation. In *Proceedings of the thirtieth annual conference of the Cognitive Science Society*, pages 595–600.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. https://arxiv.org/abs/1905.05950.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: A benchmark of linguistic minimal pairs for English. In *Proceedings of the Society for Computation in Linguistics*, volume 3.

Ethan Wilcox, Roger P. Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the Workshop on Analyzing and Interpreting Neural Networks for NLP*.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2018. Learning neural templates for text generation. *CoRR*, abs/1808.10122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176. Association for Computational Linguistics.

Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.