Does BERT *really* agree ? Fine-grained Analysis of Lexical Dependence on a Syntactic Task

Karim Lasri[ə,β] **Alessandro Lenci**[β] **Thierry Poibeau**[ə]
[ə]Lattice (École Normale Supérieure-PSL, CNRS, U. Sorbonne Nouvelle)
[β]University of Pisa
karim.lasri@ens.psl.eu
alessandro.lenci@unipi.it thierry.poibeau@ens.psl.eu

Abstract

Although transformer-based Neural Language Models demonstrate impressive performance on a variety of tasks, their generalization abilities are not well understood. They have been shown to perform strongly on subject-verb number agreement in a wide array of settings, suggesting that they learned to track syntactic dependencies during their training even without explicit supervision. In this paper, we examine the extent to which BERT is able to perform lexically-independent subject-verb number agreement (NA) on targeted syntactic templates. To do so, we disrupt the lexical patterns found in naturally occurring stimuli for each targeted structure in a novel fine-grained analysis of BERT's behavior. Our results on nonce sentences suggest that the model generalizes well for simple templates, but fails to perform lexically-independent syntactic generalization when as little as one attractor is present.

1 Introduction

Every English speaker would judge as grammatical the sentences in (1a)-(1b), but not those in (1c)-(1d), despite that they are all meaningless:

(1) a. Colourless green ideas sleep furiously.
b. Colourless green ideas that cook the door sleep furiously.
c. *Colourless green ideas sleeps furiously.
d. *Colourless green ideas that cook the door sleeps furiously.

At least since Chomsky (1957), data like this has been taken as evidence that natural language grammars contain abstract syntactic rules that (i) are independent of the meaning of lexical items and (ii) obey hierarchical, rather than linear constraints. Number agreement (henceforth NA) between the subject (the **cue**) and the verb (the **target**) of the same clause in English is one of such rules (Corbett, 2003). In fact, (1d) is ungrammatical, even though the closest noun *door* (typically referred to as **attractor**) has the same number as *sleeps*, because the noun belongs to an embedded relative clause. These NA properties have made it one of the preferred test beds to investigate the ability of neural language models (NLMs) to learn abstract, hierarchical syntactic structures (Linzen et al., 2016; Gulordava et al., 2018; Marvin and Linzen, 2018; Goldberg, 2019; Bacon and Regier, 2019; Lakretz et al., 2019). Although recurrent and transformer-based NLMs have been shown to possess syntactic abilities on the task, their nature is not fully understood (Baroni, 2019).

Can NLMs really perform lexically-independent number agreement, regardless of the syntactic structure? To answer this question, we test BERT (Devlin et al., 2019) against the NA task while controlling both the syntactic constructions and the meaningfulness of the stimuli presented to the model.

Our experiments provide two main findings. Contrarily to previous observations that BERT performs fairly well on Gulordava et al.'s (2018) syntactically well-formed but meaningless sentences (Goldberg, 2019), we show that its generalization abilities are not lexically-independent on syntactic constructions where an attractor is present[1]. Though the model has been previously shown to ignore attractors belonging to an embedded clause independent of that containing the target (Goldberg, 2019), we further provide insights on this lexical dependence that reveal the limitations of the model's abilities. Our experiments rather show that the model is actually sensitive to the presence of attractors when semantic and lexical patterns are disrupted in its input sentence.

2 Related work

Linzen et al. (2016) first tested the ability of LSTM language models to solve the NA task, and showed that they capture syntax-sensitive dependencies given targeted supervision. A subsequent study

[1]As in (1b) and (1d) above

Findings of the Association for Computational Linguistics: ACL 2022, pages 2309 - 2315

by Gulordava et al. (2018), showed that LSTMs are able to succeed even on nonce sentences obtained by replacing the lexical content in the used stimuli while keeping the syntactic structure unchanged. This suggested NLMs can acquire grammatical competence that goes beyond meaningful lexical patterns they have seen during training on a language modeling objective. Marvin and Linzen (2018) further tested an LSTM's ability to capture syntactic dependencies on constructed pairs of meaningful manually crafted sentences, so as to test targeted syntactic constructions. Contrarily to previous studies, they showed that there was considerable room for improvement for LSTMs on some challenging syntactic structures.

Goldberg (2019) further tested BERT, a transformer-based model, against stimuli from Linzen et al. (2016), Gulordava et al. (2018) and Marvin and Linzen (2018). He found that BERT substantially outperforms the previously tested LSTM language models.

Newman et al. (2021) have recently tested generalizations beyond Marvin and Linzen's (2018) data by extending the vocabulary at the target verb position. They show that though NLMs' top predictions are generally correct verbforms, the models still struggle on the NA task for infrequent verbs. In addition to testing the effect of meaningfulness by performing replacements at all positions of the sentence similarly to Gulordava et al. (2018), we control for the syntactic constructions from Marvin and Linzen (2018): given a syntactic template, can BERT generalize to *any* syntactically well-formed, but meaningless sentence? If not, when does lexical content matter?

3 General Setup

3.1 The Number Agreement Task

The NA task consists in testing whether a model shows a preference for predictions that do not violate number agreement between a selected verb and its subject. For example, when presenting BERT with sentences (1b) and (1d), we mask the token at the target position, and compare the output probabilities for sleep and sleeps. The model succeeds when it assigns a higher prediction score to the right target form.

3.2 Datasets

We test BERT's ability to solve the NA task using three different, but complementary datasets all consisting of sentences controlled by the syntactic templates described in Table 1:

a) **M&L**. This is the original dataset released by Marvin and Linzen (2018), containing the syntactic constructions we use in this study. We use it to replicate Goldberg's (2019) results as a comparison point. These sentences were designed to respect semantic constraints using a limited, but semantically controlled vocabulary.

b) **WIKI**. For each template in **M&L**, we collected naturally occurring sentences from the Wikidumps used to train BERT, to test whether the model performs better on sequences of words it could have memorized during training. We extracted raw text from the Wikidumps using WikiExtractor[2], and collected sequences of word that corresponded to the sequence of POS tag for each template in **M&L**. The data collection procedure is described in A.1.

c) **NONCE**. For each template in **M&L**, we generated "nonce", meaningless sentences keeping the syntactic structure unaffected[3]. To do so, we replace each word in the sentence with a word of the same lexical category (and same number if applicable) using a large set of words for each POS-tag (see App. A.4), similarly to Gulordava et al.'s (2018) stimuli. When a noun intervenes between the cue and the target (e.g., in condition C from Table 1), it is systematically assigned a different number from the cue, in order to test attraction effects[4]. These nonce sentences are meaningless, therefore they violate selectional restrictions contrarily to **M&L**. They also differ from Gulordava et al.'s (2018) stimuli as we additionally test the effect of the syntactic construction, having separate conditions for each template. This dataset allows us to test the extent to which the model's ability to perform the agreement on nonce sentences is dependent on their syntactic structure. Each set contains 10000 sentences, with balanced proportions of singulars and plurals, making chance level at 50%.

[2] https://github.com/attardi/wikiextractor

[3] We release this data on https://github.com/karimlasri/does-bert-really-agree

[4] That is whether the model succeeds despite the presence of a distractor noun between the cue and target of the agreement.

Struct. ID	Structure description	Example
A	Simple agreement	The **boy** **laughs**/*laugh
B	In a sentential complement	The boy knows the **girls** **play**/*plays
C	Across a prepositional phrase	The **plate** near the glasses **breaks**/*break
D	Across a subject relative clause	The **cat** that chases the mice **runs**/*run
E	In a short verb phrase coordination	The **boy** smiles and **laughs**/*laugh
F	Across an object relative clause	The **mouse** that the cats chase **runs**/*run
G	Within an object relative clause	The mouse that the **cats** **chase**/*chases runs
H	Across an object relative clause (*no that*)	The **mouse** the cats chase **runs**/*run
I	Within an object relative clause (*no that*)	The mouse the **cats** **chase**/*chases runs

Table 1: Agreement structures used in this study. These structures are taken from Marvin and Linzen (2018). The cue is in blue and the target is red. For each target, we display the pair of both the correct and incorrect verb form. In structures C, D, E and H, the attractor is underlined.

4 Experiments and Results

4.1 EXP. 1 – Sensitivity to Meaning on a Syntactic Task

In this experiment, we test whether the model's success over the NA task on Marvin and Linzen's (2018) syntactic templates requires satisfying mutual semantic constraints. To do so, we compare the NA task accuracy on **M&L** and **NONCE**. We also use **WIKI** as a comparison point, to observe whether the model succeeds better on sentences it could have memorized during training than on **M&L**'s meaningful but unseen sentences.

The results from Fig. 1 show that even though BERT is quite robust against all templates on stimuli from Marvin and Linzen (2018), it fails on some templates in **NONCE**. Little performance reduction occurs when there is no intervening attractor (A, E, G, I), that is when the cue and target are within the same clause. This shows that the model can solve the NA task in the absence of attractors, even when there is a violation of semantic selectional restrictions. The only exception is when the cue occurs in a sentential complement (B). In the absence of the complementizer *that*, the model might be perturbed by ambiguity, expecting a direct object noun (e.g., *The boy knows the mathematics lessons*). Therefore, we tested two supplementary conditions: one with the overt complementizer (B-2), and another where the verb that introduces the complementizer is constrained to be a stative verb (B-3). The results confirm our hypothesis: BERT carries out the task successfully on **NONCE** when the complementizer makes the sentence syntactically unambiguous, which also suggests that the model relies on heuristics that are partly lexicalized. On the other templates, performance drops close to chance level on **NONCE**. This means that BERT is not able to perform lexically-independent generalizations when the target and the cue are separated by a hierarchically embedded phrase containing an attractor noun. Interestingly, the model often performs better on **WIKI** than on **M&L**, which suggests that memorized lexical patterns can help solve the task in addition to being meaningful.

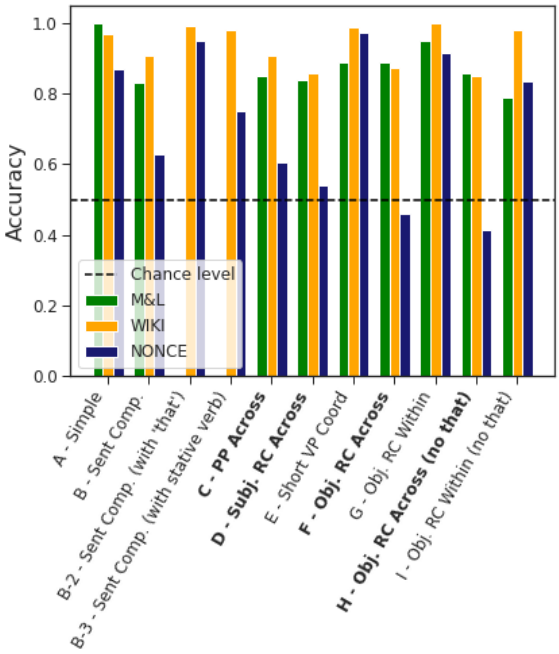

Figure 1: Accuracies on the number agreement task for the retained structures obtained by BERT Base. Templates where an attractor is present are displayed in bold. Note that conditions B-2 and B-3 were not present in the original **M&L** stimuli

4.2 EXP. 2 – Influence of One-Word Replacements

In this experiment, we measure how performance is affected when replacing words at one position at a time in the templates, on **WIKI**. Our goal is to understand whether the performance drop observed

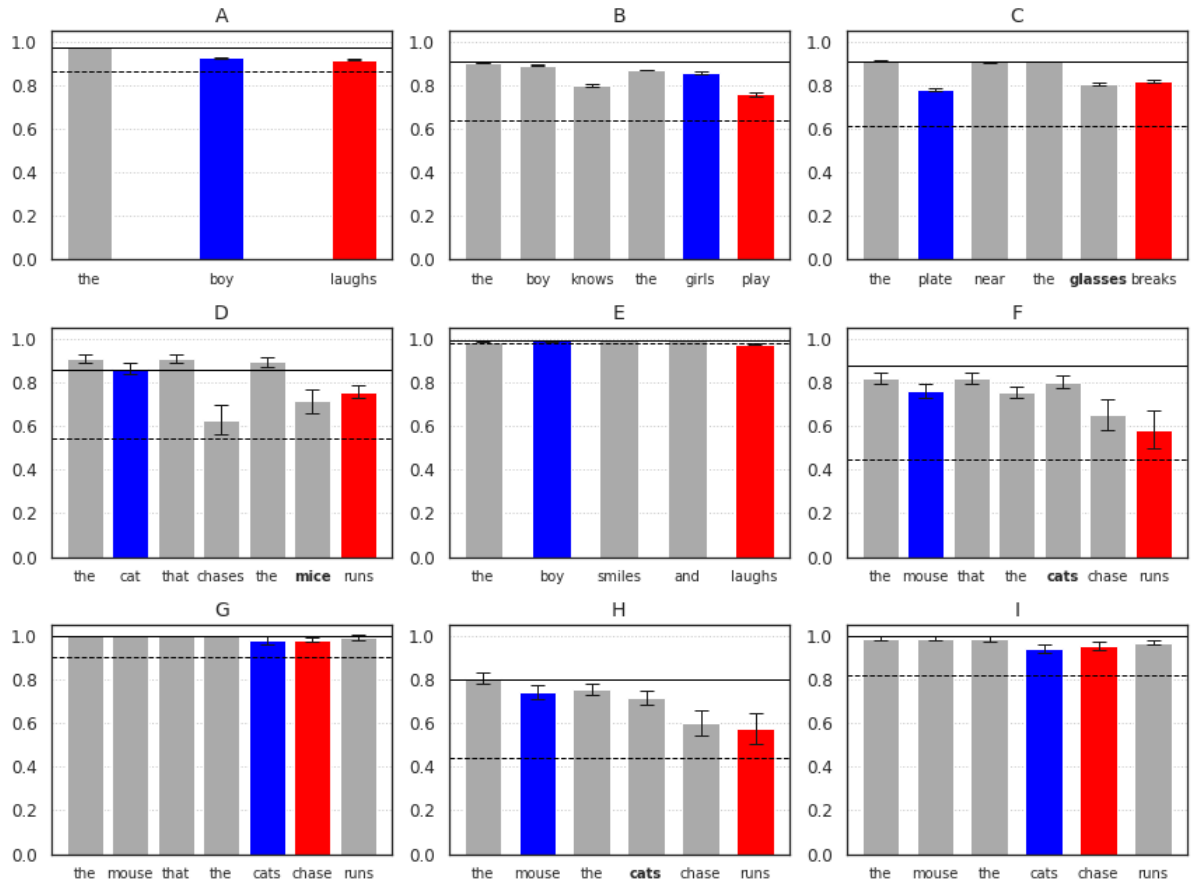

Figure 2: Accuracies on the NA task after one-word replacement. Each column represents the model's performance after intervening at the position exemplified by the word displayed in the x-axis. Attractors are represented in bold. Replacements are performed over sentences from **WIKI**. For each syntactic template, the performance on **WIKI** (continuous line) and **NONCE** (dashed line) is represented as a comparison point. The cue's replacement is represented in blue and the target's in red.

in EXP. 1 is due to the lexical content filling specific syntactic positions in our templates. In particular, we wish to understand whether most of the effect is due to replacing the cue, the target, the attractor (if present) or words in none of those three categories.

The results in Fig. 2 show that in sentences with no attractor (A, E, G, I), one-word replacement results in low performance drops, consistently with observations from EXP. 1. When the stimuli contain an embedded phrase containing an attractor, replacing the target itself, but also words close to the target verb (in D, F and H) can significantly harm performance. The cue is linearly distant from the target in sentences with attractors, and its replacement has little impact on performance. We observe that replacing the attractor replacement also has a limited impact on the task, as templates D and H show. We note a general tendency that replacing closest words results in higher performance drop than replacing farther ones, including verbs in embedded clauses. This suggests that the model's ability to deal with attractors is not due solely to hierarchical, lexically independent generalizations acquired during training. Instead, our observations show that the model is also sensitive to the content of syntactically-independent intervening material linearly close to the target verb.

5 Discussion

Previous NA studies have led Baroni (2019) to claim that "the linguistic proficiency of neural networks extends beyond shallow pattern recognition". Though it is undeniable that BERT does generalize beyond its input and is able to carry out the NA task on the simplest templates, our experiments also suggest that these generalizations can be lexically dependent. When naturally occurring lexical patterns are replaced with syntactically well-formed,

but meaningless combinations, the model's syntactic ability seems to be heavily compromised, contrary to Goldberg (2019)'s reported results on the Gulordava et al. (2018) stimuli.

Moreover, most disruption is caused by replacing the words closest to the target within the embedded phrase, that in principle should not affect the agreement relation. These two facts together indicate that some of BERT's syntactic abilities are limited to specific word sequences that the model could have memorized during training, including words that are linearly close but belong to a different embedded phrase or clause. Furthermore, the fact that the model improves its performance on data it has been trained on (i.e., the **WIKI** dataset) over other meaningful, unseen sentences (i.e., the **M&L** dataset) is further evidence that at least part of its alleged generalization abilities might be just the effect of memorization.

We can surmise that the model relies on a variety of heuristics acquired during training to approximate syntactic generalizations, in line with Finlayson et al. (2021), who found two distinct mechanisms to accomplish agreement in Transformer-based architectures. We find that those heuristics can therefore tend to be highly lexicalized, similarly to Newman et al. (2021) who showed that generalization is not systematic by testing a wide range of verbs. This is confirmed by BERT's sensitivity to the main verb when there is no overt complementizer[5], which prevents it from solving the NA task. This suggests that the model has acquired semi-lexicalized syntactic information about verb subcategorization preferences.

Although BERT's ability to approximate syntactic rules is probably more brittle than previously argued, this should not lead to rejecting its ability to learn natural language grammar. For instance, constructionist approaches (Hoffman and Trousdale, 2013) have argued since long against a purely abstract grammar detached from lexical meaning, despite what the data in (1) have often been claimed to prove. The alternative view is a grammar consisting of constructions that differ for their level of abstractness and lexicalization. BERT's lexically-driven behavior could therefore be consistent with this less abstract conceptions of syntax. Finally, given previous experiments (Laurinavichyute and von der Malsburg, 2022), we can speculate that humans could also similarly manifest patterns of errors driven by semantic, or lexical interferences from words linearly close to the target. Though such patterns seem to differ between language models and humans (Linzen and Leonard, 2018), this in turn leads us to questioning our expectations regarding the syntactic abilities of neural language models.

[5]cf. sentence type B *no that*

6 Conclusion

In this paper, we have shown that BERT's ability to solve the NA task on meaningless sentences strongly depends on the stimuli's syntactic template. While the model is able to perform lexically-independent generalization in simple settings, it fails when the agreement relation crosses an embedded phrase containing an attractor. We further provide insights on this lexical dependence, showing that the model relies mostly on the lexical content at the closest positions to the target of the agreement, though they belong to an independent embedded phrase.

In the future, we want to get a better understanding of the mechanisms underlying the observed syntactic abilities of Transformers, and in particular what makes some heuristics involved to solve a syntactic task lexically dependent. A more detailed analysis of the influence played by lexical combinations will help us understand the nature of the heuristics the model uses to solve complex NA cases involving one or more attractors. Moreover, we wish to compare BERT's predictions with human judgments on our meaningless sentences.

7 Acknowledgements

This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the"Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

References

Geoff Bacon and Terry Regier. 2019. Does BERT agree? evaluating knowledge of structure dependence through agreement relations. *ArXiv*, abs/1908.09892.

Marco Baroni. 2019. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791).

Noam Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.

G. Corbett. 2003. Agreement: Terms and boundaries. In *The Role of Agreement in Natural Language: TLS 5 Proceedings*, pages 109–122.

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.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. 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 1828–1843, Online. Association for Computational Linguistics.

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

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, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Hoffman and Graeme Trousdale, editors. 2013. *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. 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 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.

Anna Laurinavichyute and Titus von der Malsburg. 2022. Semantic attraction in sentence comprehension. *Cognitive Science*, 46(2):e13086.

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.

Tal Linzen and Brian Leonard. 2018. Distinct patterns of syntactic agreement errors in recurrent networks and humans. *CoRR*, abs/1807.06882.

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.

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.

Struct. ID	Structure description	Example
A	Simple agreement	The **window** **fails**/*fail
B	In a sentential complement	The prisons insist the **surprise** **happens**/*happen
C	Across a prepositional phrase	The **gift** in the origins **reflects**/*reflect
D	Across a subject relative clause	The **passion** that identifies the sellers **binds**/*bind
E	In a short verb phrase coordination	The **pepper** falls and **pulls**/*pull
F	Across an object relative clause	The **bombings** that the tune picks **flows**/*flow
G	Within an object relative clause	The rhyme that the **elders** **need**/*needs happens
H	Across an object relative clause (*no that*)	The **decrees** the cage examine **happen**/*happens
I	Within an object relative clause (*no that*)	The lyric the **beetles** **quote**/*quotes scores

Table 2: Randomly picked examples of generated sentences for each tested structure.

A Appendix - Data collection

A.1 Wikipedia data collection

For each of the structures described in 1, we represent the construction by its sequence of lexical categories. We then extract sequences of words from Wikipedia for each of the constructions that match the pattern. To do so, we read Wikipedia linearly and store naturally occurring token sequences that match our constructions, based on the same vocabulary that we use to generate our **NONCE** sentences, described in A.4.

A.2 Data Generation procedure

Generated sentences are built from the sequence of POS-tags describing each construction. We randomly pick one word from our dictionaries at each position of the sequence, as in (Gulordava et al., 2018). When a noun intervenes between the cue and the target (e.g., in condition C from Table 1), it is systematically assigned a different number from the cue, in order to test attraction effects[6]. We chose to only use neutral determiners along with possessives to avoid clashes between a noun's and its determiner's numbers. Datasets contain 10000 samples, and for Exp. 2, we reproduced the experiments 10 times for each replacement to produce error bars. Our data is balanced, which means each dataset contains 5000 singulars and 5000 plurals. Randomly picked examples are displayed in Table 2.

A.3 Data Generation Vocabulary Collection and Preprocessing

Nouns and verbs were collected from Linzen et al. (2016)'s dataset. As the NA task setting requires looking at predicted scores for the masked target forms, we only keep verbs for which both forms are present in BERT's vocabulary as an unsplit token. Similarily to Goldberg (Goldberg, 2019), we filter out sentences where the target is a present form of the verb 'be' as this verb is too frequent in corpora and is treated differently from other verbs. Our data generation procedure and vocabulary are publicly available at `https://github.com/karimlasri/does-bert-really-agree`.

A.4 Used Vocabulary

Determiners and possessives. 'my', 'your', 'his', 'her', 'its', 'our', 'their', 'the'

Relativizer/complementizer. 'that'

Nouns. We use 2636 noun pairs for which both the singular and plural forms are part of BERT's vocabulary.

Verbs. We use 444 verb pairs, for which both singular and plural forms are present in BERT's vocabulary.

Stative verbs in Condition B-3. We use the following stative verbs for the (B-3) condition: ('believes', 'believe'), ('considers', 'consider'), ('doubt', 'doubt'), ('hears', 'hear'), ('knows', 'know'), ('realises', 'realise'), ('says', 'say'), ('supposes', 'suppose'), ('thinks', 'think'), ('understands', 'understand'), ('wishes', 'wish')

[6]That is whether the model succeeds despite the presence of a distractor noun between the cue and target of the agreement.