Evaluating German Transformer Language Models with Syntactic Agreement Tests

Karolina Zaczynska*, Nils Feldhus*, Robert Schwarzenberg, Aleksandra Gabryszak, Sebastian Möller
German Research Center for Artificial Intelligence (DFKI)
{firstname.lastname}@dfki.de

Abstract

Pre-trained transformer language models (TLMs) have recently refashioned natural language processing (NLP): Most state-of-the-art NLP models now operate on top of TLMs to benefit from contextualization and knowledge induction. To explain their success, the scientific community conducted numerous analyses. Besides other methods, syntactic agreement tests were utilized to analyse TLMs. Most of the studies were conducted for the English language, however. In this work, we analyse German TLMs. To this end, we design numerous agreement tasks, some of which consider peculiarities of the German language. Our experimental results show that state-of-the-art German TLMs generally perform well on agreement tasks, but we also identify and discuss syntactic structures that push them to their limits.

1 Introduction

Pre-trained language models, in particular those which are based on the transformer architecture (Vaswani et al., 2017), have immensely improved the performance of various downstream models (see, e.g. Zhang et al. (2020, 2019); Raffel et al. (2019)). To explain their success, numerous introspective experiments have targeted different aspects of TLMs. It was shown, for instance, that they encode syntactic, semantic and world knowledge (Petroni et al., 2019) and present downstream models with a highly contextualized representation of the input tokens (Tenney et al., 2019). For a comprehensive overview of the many studies conducted about arguably the most prominent of language models, BERT (Devlin et al., 2019), we refer the interested reader to the excellent overview paper by Rogers et al. (2020).

With the exception of experiments targeting a multilingual BERT model (Rogers et al., 2020), most of the studies were conducted only for English, however. Other languages are underrepresented. In this work, we narrow the gap for German by analysing the abilities and limits of German TLMs. To the best of our knowledge, we are the first to conduct such an analysis for the German language.

When compared with English, there are considerable syntactic differences in the German language that we consider in this work. For example, the inflection system of the German language is more complex, the range of morpho-syntactic rules needed to form grammatical sentences is larger, and the allowed word order is more diverse. As a consequence, the German language models face specific challenges. The syntactic agreement tests presented in this work include several of them.

Our main contributions are threefold:

1. Utilizing context-free grammars (CFG), we compile a German data set of controlled syntactic correctness tests of various complexities. The motivation and construction of the data set is closely following the one described in Marvin and Linzen (2018), where syntactic tests were conducted for English. In particular, we devise several kinds of subject-verb agreement as well as reflexive anaphora agreement tasks, taking into account peculiarities of the German language. A simple subject-verb agreement task is given in Example 1.1.

* Shared first authorship.

Example 1.1. Decide which of the following sentences is grammatical:

(a) Der Autor *lacht*. (The author laughs.)
(b) * Der Autor *lachen*. (The author laugh.)

2. We use the data set to evaluate two transformer-based language models that were pre-trained on German corpora. During the evaluation, contrary to prior work, we utilize the cross entropy loss to score the syntactic correctness of input sentences. This addresses a problem with the sub-word tokenization of some TLMs that was previously solved by discarding thousands of data points.

3. We conduct a qualitative and quantitative analysis of the experimental results, estimating the abilities and limits of the TLMs tested.

2 Methods

Our work combines and translates the targeted syntactic evaluation of language models by Marvin and Linzen (2018) and the assessment of BERT's syntactic abilities by Goldberg (2019) from English into German. Our methods consist of agreement test generation and model evaluation.

We created the following agreement test, following Marvin and Linzen (2018): Two sentences, a grammatical one and an ungrammatical one, are forwarded through a model. The sentences differ minimally from each other at only one locus of (un)grammaticality, i.e. one word. The model output is monitored and if the output suggests that the model prefers the grammatical one over the ungrammatical one, that instance is counted as a correct classification; otherwise, it is counted as an incorrect classification.

Goldberg (2019) used agreement tests to evaluate BERT models. To account for their bidirectionality, he masked the locus of (un)grammaticality and queried the candidate probabilities for the mask. In Example 1.1, `Der Mann [MASK] .` is forwarded through a BERT model and the candidate probabilities at the position of the mask are determined. If `lacht` receives a higher probability than `lachen`, the task is solved correctly by the language model. The author runs into problems, however, when the candidates are tokenized into multiple sub-word tokens, say `lachen` → [`lach, ##en`]. In this case, the author simply ignores the data point.

Instead of discarding such sequences, we take inspiration from Marvin and Linzen (2018) and score whole sentences (without masks). However, we still discard cases in which the two candidates have a different amount of sub-words after tokenization, as we see the comparability impaired if the resulting sequences of tokens are of different lengths.

We compute the sentence score with the cross-entropy loss of the forward pass, using the input sequence as the target:

$$\frac{1}{T}\sum_{i=1}^{T}\left(-f(S)_{i,S_i} + \log(\sum_{j=1}^{T}\exp(f(S)_{j,S_j}))\right)$$

where S is a sequence of T positive integer token ids and $f : \mathbb{Z}^N \rightarrow \mathbb{R}^{N\times V}$ a language model mapping N token IDs onto N token probabilities over a vocabulary of size V. We compute Eq. 2 with the grammatical candidate in place and a second time with the ungrammatical candidate in place.

Please note that during the training of a bidirectional language model, the points of interests need to be masked to prevent information leakage (Devlin et al., 2019). In our case, information leakage is not a problem because we compare two whole sequences.

3 Syntactic Agreement Tests

This section describes the syntactic agreement tests we generated to evaluate German TLMs on.

Our tests are inspired by the research of Marvin and Linzen (2018) and Goldberg (2019). In particular, we translate many of their tests on subject-verb agreement (SVA) and reflexive anaphora (RA) agreement from English to German (Section 3.2). In addition, we design tests for syntactic phenomena which are typical of the German language (Section 3.3).

The generated tasks cover a range of difficulties. In German, the subject and the inflected verb agree with regard to person and grammatical number. In the simplest case, the sentences contain only a subject and a verb. In the more challenging cases we added different types of distraction, i.e. either additional non-subjective (pro)nouns as candidates for subjects or other additional lexical material making the sentences more complex.

For our experiments, we consider instances where the grammatical number of non-subjective (pro)nouns matches the one of the subject as well as examples where their grammatical number is different. Furthermore, we distinguish between local and non-local feature agreement, which means, we take into account whether the distractors occur between subject and its corresponding verb or not. The described test scenario allows us to compare the models' performance with regard to the features of the distractor as well as its distance to the relevant verb. Therefore, the designed tests expand the experimental setup of Marvin and Linzen (2018) by going beyond the attractors, i.e. intermissions defined as intervening nouns with the opposite number from the subject (Linzen et al., 2016).

3.1 Dataset

We created a dataset of 12,426 sentences using hand-crafted Context Free Grammars (CFGs) as illustrated in Example 3.1.

Example 3.1. *Context Free Grammar* for creating sentences *S* from a vocabulary *V* to test agreement in a simple sentence:

S −> NP V '.'
NP −> ART N
ART −> 'Die'
N −> 'Autoren' | 'Richterinnen'
V −> 'lachen' | 'reden'

Output: *Die Autoren lachen. / Die Autoren reden. / Die Richterinnen lachen. / Die Richterinnen reden.*

As shown in the example, the CFG creates sentences as output with varying lexical items but with a relatively low variance. However, it allows us to tightly control the generated sentences with respect to the desired tests, in terms of distractor features as well as syntactic structure and correctness of the sentences.

Our data set covers 14 test cases of different challenge levels (Sections 3.2–3.3). The number of sentences ranges from 64 to 2160 with an average of 1035,5 sentences per test case. A sentence is build on average of 6.88 tokens. The vocabulary consists of 88 lexems and 171 word forms. For our corpus, we chose common words to build the sentences, so that the TLM was not confronted with potentially unknown words.

3.2 Established Agreement Tests

In the following, we introduce the agreement tests that we translated from the work of Marvin and Linzen (2018).

We describe three groups of tests ordered by the increasing challenge level: (1) local agreement, no distractors, (2) local agreement, plus distractors, and (3) non-local agreement, plus distractors. Afterwards, we introduce tests designed to target German phenomena specifically.

Local agreement, no distractors We first include cases with local agreement and without a distractor. Sentences consisting of only one subject and verb are what we refer to as *simple sentence* in the following, showcased in Example 3.2.

Example 3.2. *Simple sentence* with only one subject and one verb (the locus of (un)grammaticality is italic, the incorrect variant is preceded by *):
(a) Das Kind *trinkt*.
(b) * Das Kind *trinken*.

Local agreement, plus distractors Complex sentences with a local agreement in a *sentential complement* or in an *object relative clause* constitute the next level of difficulty. Those sentences contain two subjects: one in the main clause, and another one in the subordinate clause. In Example 3.3, the latter functions as a sentential complement, in Example 3.4, as an object relative clause. For both types of the subordinate clause, the verb follows the subject directly. The subject of a main clause is the distractor in those cases while the agreement between the subject and the verb of the subordinate clause is our point of interest.

Example 3.3. *SVA in a sentential complement*:
(a) Die Vertreter sagten, dass das Kind *trinkt*.
(b) * Die Vertreter sagten, dass das Kind *trinken*.

Example 3.4. *SVA in an object relative clause*
(a) Der Autor, den die Vertreter *kennen*, lacht.
(b) * Der Autor, den die Vertreter *kennt*, lacht.

Non-local agreement, plus distractors We also tested TLMs on a set of constructions with non-local agreement, induced by potentially distracting words and phrases between the head of the subject and its corresponding verb. With these tasks, we are testing the language model's ability to attend to the subject in sentences across long contexts.

Our first test case is a *SVA across a prepositional phrase* (PP). We created sentences with the

subject modified by a directly following PP, which includes a potentially attracting noun, as in Example 3.5.

Example 3.5. *SVA across a PP*
(a) Der Autor neben den Landstrichen *lacht*.
(b) * Der Autor neben den Landstrichen *lachen*.

Furthermore, we test *SVAs across subject relative clauses* which include one potentially distracting object and verb in between subject and corresponding verb, as in Example 3.6.

Example 3.6. *SVA across a subject relative clause*
(a) Der Autor, der die Architekten liebt, *lacht*.
(b) * Der Autor, der die Architekten liebt, *lachen*.

The same challenge exists for *SVAs across object relative clauses* which also contain potentially distracting chunks and separate the subject and its corresponding verb, as in Example 3.7.

Example 3.7. *SVA across an object relative clause*
(a) Der Autor, den die Vertreter kennen, *lacht*.
(b) * Der Autor, den die Vertreter kennen, *lachen*.

Additionally, we designed various sentences for testing *SVAs across coordinated verbal phrases* (VP), where the subject must agree in person and number with the finite verb included in each VP. In our test, the point of interest is the second verb of the coordination. This kind of structure challenges the model to recognize that the complete subject-verb structure does not end after the first verb, but rather it also includes the second verb. We test the SVA in verbal coordinations of different lengths and various number of distractors.

First, we test the model on sentences consisting of a short and simple VP coordination with no distractors, as illustrated by Example 3.8.

Example 3.8. *SVA in short VP coordinations (i.e. with no distractors)*
(a) Der Autor schwimmt und *lacht*.
(b) * Der Autor schwimmt und *lachen*.

To increase the difficulty level, we inserted noun phrases having a different number as the subject into the coordinated VP. We distinguish between verbal coordinations with a single noun distractor (Example 3.9) and two noun distractors (Example 3.10).

Example 3.9. *SVA in medium VP coordinations (i.e. with a single noun distractor)*
(a) Der Autor redet mit Menschen und *lacht*.
(b) * Der Autor redet mit Menschen und *lachen*.

Example 3.10. *SVA in long VP coordinations (i.e. with two noun distractors)*
(a) Der Autor redet mit Menschen und *verfolgt* die Fernsehprogramme.
(b) Der Autor redet mit Menschen und *verfolgen* die Fernsehprogramme.

3.3 Novel Agreement Tests

In addition to the tests above that we based on previous work, we also designed tasks which target constructs that are more specific to the German language.

First, we test the agreement between verb and its corresponding subject containing an extended modifier, i.e. an adjective modifying a subject and extended by further subordinate nominal or prepositional phrase. The extended modifier is positioned between the determinator and the noun of the subject. In comparison to English, the German language is much more flexible with regard to the number and the types of allowed extensions. To test the impact of nouns used within extended modifiers of a subject we include sentences with simple modifiers as well as with extended modifiers (Example 3.11 and 3.12).

Example 3.11. *SVA with a simple modifier*
(a) Die wartenden Autoren *lachen*.
(b) * Die wartenden Autoren *lacht*.

Example 3.12. *SVA with an extended modifier*
(a) Die die Pflanze liebenden Autoren *lachen*.
(b) * Die die Pflanze liebenden Autoren *lacht*.

Another agreement test relates to the more diverse word order in German in comparison to English. Example 3.13 illustrates the shift of the direct object *diese Romane* from its standard position in the middle-field (after the finite verb) to the pre-field, and the shift of the subject *der Autor* to the middle-field from its standard position in the pre-field (before the finite verb). This movement would be not possible in English. The German language often allows the shift, since it marks the case of noun phrases by the inflectional suffix of their determiner (e.g. *der Autor* in nominative case vs. *den Autor* in accusative case) and sometimes also by the suffix of the noun itself (e.g. *des Autors* in genitive). That property supports to distinguish subjects (always nominative case) from objects or adjuncts independent from their position in

	distilGBERT	GBERT$_{large}$	# sents
SUBJECT-VERB AGREEMENT			
Simple Sentence	**0.9710**	0.9420	69
In a sentential complement	0.9565	**0.9894**	2160
Short VP coordination	0.7125	**0.7542**	240
Medium VP coordination	0.4813	**0.6188**	480
Long VP coordination	0.5167	**0.5938**	480
Across a PP	0.7968	**0.9005**	2160
Across a subject relative clause	0.6924	**0.9896**	1440
Across an object relative clause	0.7386	**0.9206**	945
In an object relative clause	0.9568	**0.9600**	1575
With a modifier	0.9458	**0.9959**	240
With an extended modifier	0.8917	**0.9583**	480
Pre-field	0.7991	**0.8355**	468
REFLEXIVE ANAPHORA			
Person & number agreement	0.4876	**0.8716**	1737
Case agreement	0.8534	**0.9691**	648

Table 1: Performances (accuracy) of two TLMs on German syntactic agreement tests. Underlined are the five tasks the models performed worst on. Bold-faced are the best scores per task.

a sentence. With this test case, we can evaluate if the model recognizes the subject in sentences correctly, even though the subject-verb-object order is disregarded. We exclude test sentences where the subject and the object have the same inflectional suffixes in nominative and accusative, i.e. an unambiguous distinction between subject and object is not possible solely based on the inflection.

Example 3.13. *Pre-field*
(a) Diese Romane *empfahl* der Autor.
(b) * Diese Romane *empfahlen* der Autor.

Moreover, we created sentences with reflexive verbs, i.e. sentential phrases where the reflexive anaphora (RA) in the accusative case follows the verb and agrees with the subject in the grammatical number and person. The first sentence in Examples 3.14 and 3.15 illustrates the agreement between RA *mich* (accusative case) and the subject *ich* in person (first) and number (singular). We use two different tests: (a) for the recognition of a correct person (Example 3.14), also used by Marvin and Linzen (2018), and (b) for the recognition of a correct case (accusative instead of incorrect dative, Example 3.15). The correct number is always given.

Example 3.14. *Subject RA agreement (person-agreement)*
(a) Ich bedanke *mich*.
(b) * Ich bedanke *sich*.

Example 3.15. *RA in accusative (case-agreement)*
(a) Ich bedanke *mich*.
(b) * Ich bedanke *mir*.

4 Experiments

In this section, we introduce the models we evaluate and in particular highlight their similarities and differences. We probe transformer-based BERT models because they are currently the basis for many state-of-the-art downstream models and very prominent in the community. The model selection was driven and confined by availability. We made use of Wolf et al. (2019)'s transformer package.[1]

The first model which we refer to as GBERT$_{large}$ is a community model provided by the Bavarian State Library.[2] It was trained on multiple German corpora including a recent Wikipedia dump, EU Bookshop corpus, the Open Subtitles corpus, a CommonCrawl corpus, a ParaCrawl corpus and the News Crawl corpus, with 16 GB of training material in total.

The second model which we refer to as distilGBERT was trained on half of the data used

[1] https://github.com/huggingface/transformers (Accessed: 2020-03-05)
[2] https://huggingface.co/dbmdz/bert-base-german-cased (Accessed: 2020-03-05)

to pretrain BERT using distillation with the supervision of $\text{GBERT}_{\text{large}}$ [3].

The data set, the CFGs with the list of lexical items and the code for the experiments are publicly available. [4]

5 Results & Discussion

The coarse-grained results of our experiments are listed in Table 1. We note that both models perform well across the majority of tasks. This is in line with previous work that demonstrated that BERT models are capable of solving syntactic agreement tasks. As shown by Goldberg (2019) for English, for instance, our most successful German BERT model, $\text{GBERT}_{\text{large}}$, also scores above 80% or 90% in most of the tasks, whereas the LSTM-LMs probed by Marvin and Linzen (2018) achieve scores not above 74%.

We observe that $\text{GBERT}_{\text{large}}$ outperforms $_{\text{distil}}$GBERT in thirteen out of fourteen tasks. For example, in the case of *SVA across an object relative clause*, $\text{GBERT}_{\text{large}}$ achieved a score of 92.06%, whereas $_{\text{distil}}$GBERT's score is lower by around 18 percentage points. Based on these observations, we assume the higher amount of German training data, that $\text{GBERT}_{\text{large}}$ was trained on, is the distinguishing factor.

There is a big overlap between the most challenging stress tests. Four out of five tests align when sorted in ascending order (worst performance first, underscored in Table 1). To analyse the stress tests further, in Table 2, we subdivide cases between singular and plural subjects and distractors.

We expected high accuracies for the cases with local agreement. Our results show that all those cases, which are *Simple Sentence*, *SVA in a sentential complement*, *SVA in an object relative clause* and *SVA with a simple modifier*, have a score above 94 percent for both models.

Regarding the German-specific syntactic constructs, we observe that both models perform well. The movement of the subject from pre-field to middle-field does not seem to cause any bigger problems, both $_{\text{distil}}$GBERT and $\text{GBERT}_{\text{large}}$ have an accuracy around 80%.

As can be seen in Tables 1 and 2, VP coordination probing cases were a big challenge for both models. For example, $_{\text{distil}}$GBERT only achieves an overall accuracy of 0.4813 on *SVA in a medium VP coordination* and 0.5167 on *SVA in a long VP coordination*, while $\text{GBERT}_{\text{large}}$ achieves 0.6188 and 0.5938, respectively. In these aspects, our results deviate considerably from the findings of Goldberg (2019) who reported that the English BERT models performed well on long VP tasks, too. The respective syntactic constructs may thus be particularly challenging for the BERT models in the German language. Interestingly, according to Table 2, $\text{GBERT}_{\text{large}}$ performs with an accuracy of 1.0 for long VPs with a singular subject. We note that the most challenging sentences for both models in all of the VP coordination cases were the ones with a plural subject.

In contrast to the aforementioned VP coordinations, *SVA across an object relative clause* for both models and *SVA across a subject relative clause* for $_{\text{distil}}$GBERT show a better accuracy for sentences when the subject is plural. We assume that for some cases the grammatical number of the subject is a more influential aspect for the result than the number of the distractor. We didn't expect this given that we used the same lexemes within one case to ensure comparability between the results.

We expected that sentences in which the grammatical number of the distractor deviates from the number of the relevant verb (singular-plural and plural-singular) have a lower accuracy. This, however, applies only to a few cases, like *SVA across an object relative clause* and *Pre-field*. Thus, the TLMs appear to be mostly robust against those distractors.

Inferring sound causes for why some syntactic constructs push the models to their limit would require a thorough statistical analysis of the data and probably even an introspective analysis of the model. We leave it to future work to conduct such an analysis.

6 Related Work

There is a huge body of related literature on the syntactic evaluation of language models. For more background, we refer the interested reader to the works cited in the influential contribution by Marvin and Linzen (2018) and Goldberg (2019).

Gulordava et al. (2018) assessed subject-verb agreement with an emphasis on syntactic over

[3] https://github.com/huggingface/transformers/blob/master/examples/distillation/README.md (Accessed: 2020-05-21)

[4] https://github.com/DFKI-NLP/gevalm/

		$_{\text{distil}}$GBERT	GBERT$_{\text{large}}$	# sents
SUBJECT-VERB AGREEMENT				
Simple sentence	-sg	0.9744	0.8974	39
	-pl	0.9667	1.0	30
In a sentential complement	-sgsg	1.0	0.9593	540
	-plpl	0.8926	1.0	270
	-sgpl	0.9407	1.0	1080
	-plsg	0.9963	0.9963	270
Short VP coordination	-sg	0.8917	0.9667	120
	-pl	0.5333	0.5417	120
Medium VP coordination	-sgsg	0.7667	0.95	120
	-plpl	0.2333	0.3167	120
	-sgpl	0.775	0.9667	120
	-plsg	0.15	0.2417	120
Long VP coordination	-sgsg	0.5917	1.0	120
	-plpl	0.2	0.2167	120
	-sgpl	0.4917	1.0	120
	-plsg	0.7833	0.1583	120
Across a prepositional phrase	-sgsg	0.7593	0.8667	540
	-plpl	0.7759	0.9426	540
	-sgpl	0.7907	0.8333	540
	-plsg	0.8611	0.9593	540
Across a subject relative clause	-sgsg	0.4222	0.9944	360
	-plpl	1.0	0.975	360
	-sgpl	0.3638	0.9889	360
	-plsg	0.9833	1.0	360
Across an object relative clause	-sgsg	0.4148	0.963	270
	-plpl	0.9667	0.9481	270
	-sgpl	0.4889	0.7481	135
	-plsg	0.9593	0.937	270
In an object relative clause	-sgsg	0.9911	1.0	450
	-plpl	0.9511	0.9422	450
	-sgpl	0.88	0.9822	225
	-plsg	0.9667	0.9267	450
With a simple modifier	-sg	0.975	1.0	120
	-pl	0.9167	0.9917	120
With an extended modifier	-sgsg	0.9417	0.9667	120
	-plpl	0.8	0.9583	120
	-sgpl	0.9083	0.9667	120
	-plsg	0.9167	0.9417	120
Pre-field	-sgsg	0.7167	0.975	120
	-plpl	1.0	0.9417	120
	-sgpl	0.8833	0.7333	120
	-plsg	0.5741	0.6759	108
REFLEXIVE ANAPHORA				
Person & number agreement	-simple	0.3611	0.6389	72
	-longer	0.3492	0.7841	315
	-SentCompl	0.5267	0.9045	1350
Case agreement	-simple	0.9444	1.0	18
	-longer	0.7222	0.7889	90
	-SentCompl	0.8722	0.9981	540

Table 2: Fine-grained results of our experiments. Double-case specifications, e.g. "-plsg", are to be read as the tested agreement being in plural form, while the distractor is in singular form.

semantic preference. McCoy et al. (2019) created a data set with entailment tests. Bacon and Regier (2019) extended Goldberg (2019) to 26 languages, excluding German, and found out that with a higher number of distractors and long-range dependencies, BERT achieves lower accuracies for the syntactic agreement tests.

As mentioned above, we also recommend the overview paper by Rogers et al. (2020) on studies of BERT models specifically. Apart from the experiments cited in this work that evaluate multilingual models, such as MBERT, we are not aware of any study dedicated to the agreement analysis of German BERT models.

Rönnqvist et al. (2019), nevertheless, tested multilingual BERT models on their ability of hierarchical understanding of German sentences and with a cloze test for which an arbitrary (grammatically correct) word was masked and needed to be filled in again.

7 Conclusion

We conducted a broad analysis of German BERT models, targeting their syntactic abilities. We translated agreement tests from English to German and also designed tasks that reflect syntactic phenomena that are typical for the German language. The data set we generated and the accompanying grammars are publicly available.

Furthermore, we utilized the cross-entropy loss to score whole natural sentences and this way mitigated a problem with sub-word tokenization. Our source code is open source, too.

Our experimental results show that the German models perform very well on most of the agreement tasks. However, we also identified syntactic stress tests that models in other languages appear to solve much better. We plan to replace the synthetic sentences with real language samples in the future, to better reflect the diversity of the German language in our experiments.

Acknowledgements

We would like to thank Leonhard Hennig for his valuable feedback. This work has been supported by the German Federal Ministry of Education and Research as part of the project XAINES.

References

Geoff Bacon and Terry Regier. 2019. Does BERT agree? Evaluating knowledge of structure dependence through agreement relations. *arXiv:1908.09892 [cs]*. ArXiv: 1908.09892.

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.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv: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.

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.

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.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1910.10683*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *arXiv preprint arXiv:2002.12327*.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36, Turku, Finland. Linköping University Electronic Press.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. *arXiv:1905.06316 [cs]*. ArXiv: 1905.06316.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2019. Semantics-aware bert for language understanding. *arXiv preprint arXiv:1909.02209*.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2020. Retrospective reader for machine reading comprehension. *arXiv preprint arXiv:2001.09694*.