The Better Your Syntax, the Better Your Semantics? Probing Pretrained Language Models for the English Comparative Correlative

Leonie Weissweiler[*◇], **Valentin Hofmann**[†*], **Abdullatif Köksal**[*◇], **Hinrich Schütze**[*◇]

[*]Center for Information and Language Processing, LMU Munich
[◇]Munich Center of Machine Learning
[†]Faculty of Linguistics, University of Oxford
{weissweiler,akoksal}@cis.lmu.de
valentin.hofmann@ling-phil.ox.ac.uk

Abstract

Construction Grammar (CxG) is a paradigm from cognitive linguistics emphasising the connection between syntax and semantics. Rather than rules that operate on lexical items, it posits *constructions* as the central building blocks of language, i.e., linguistic units of different granularity that combine syntax and semantics. As a first step towards assessing the compatibility of CxG with the syntactic and semantic knowledge demonstrated by state-of-the-art pretrained language models (PLMs), we present an investigation of their capability to classify and understand one of the most commonly studied constructions, the English comparative correlative (CC). We conduct experiments examining the classification accuracy of a syntactic probe on the one hand and the models' behaviour in a semantic application task on the other, with BERT, RoBERTa, and DeBERTa as the example PLMs. Our results show that all three investigated PLMs are able to recognise the structure of the CC but fail to use its meaning. While human-like performance of PLMs on many NLP tasks has been alleged, this indicates that PLMs still suffer from substantial shortcomings in central domains of linguistic knowledge.

1 Introduction

The sentence "The better your syntax, the better your semantics." contains a construction called the English comparative correlative (CC; Fillmore, 1986). Paraphrased, it could be read as "If your syntax is better, your semantics will also be better." Humans reading this sentence are capable of doing two things: (i) *recognising* that two instances of "the" followed by an adjective/adverb in the comparative as well as a phrase of the given structure (i.e., the syntax of the CC) express a specific meaning (i.e., the semantics of the CC); (ii) *understanding* the semantic meaning conveyed by the CC, i.e., understanding that in a sentence of the given structure, the second half is somehow correlated with the first.

In this paper, we ask the following question: are pretrained language models (PLMs) able to achieve these two steps? This question is important for two reasons. Firstly, we hope that recognising the CC and understanding its meaning is challenging for PLMs, helping to set the research agenda for further improvements. Secondly, the CC is one of the most commonly studied constructions in construction grammar (CxG), a usage-based syntax paradigm from cognitive linguistics, thus providing an interesting alternative to the currently prevailing practice of analysing the syntactic capabilities of PLMs with theories from generative grammar (e.g., Marvin and Linzen, 2018).

We divide our investigation into two parts. In the first part, we examine the CC's syntactic properties and how they are represented by PLMs, with the objective to determine whether PLMs can *recognise* an instance of the CC. More specifically, we construct two syntactic probes with different properties: one is inspired by recent probing methodology (e.g., Belinkov et al., 2017; Conneau et al., 2018) and draws upon minimal pairs to quantify the amount of information contained in each PLM layer; for the other one, we write a context-free grammar (CFG) to construct approximate minimal pairs in which only the word order determines if the sentences are an instance of the CC or not. We find that starting from the third layer, all investigated PLMs are able to distinguish positive from negative instances of the CC. However, this method only covers one specific subtype of comparative sentences. To cover the full diversity of instances, we conduct an additional experiment for which we collect and manually label sentences from C4 (Raffel et al., 2020) that resemble instances of the CC, resulting in a diverse set of sentences that either are instances of the CC or resemble them closely *without* being instances of the CC. Applying the

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10859–10882
December 7-11, 2022

same methodology to this set of sentences, we observe that all examined PLMs are still able to separate the examples very well.

In the second part of the paper, we aim to determine if the PLMs are able to *understand* the meaning of the CC. We generate test scenarios in which a statement containing the CC is given to the PLMs, which they then have to apply in a zero-shot manner. As this way of testing PLMs is prone to a variety of biases, we introduce several mitigating methods in order to determine the full capability of the PLMs. We find that none of the PLMs we investigate perform above chance level, indicating that they are not able to understand and apply the CC in a measurable way in this context.

We make three main **contributions**:

– We present the first comprehensive study examining how well PLMs can recognise and understand a CxG construction, specifically the English comparative correlative.

– We develop a way of testing the PLMs' recognition of the CC that overcomes the challenge of probing for linguistic phenomena not lending themselves to minimal pairs.

– We adapt methods from zero-shot prompting and calibration to develop a way of testing PLMs for their understanding of the CC.[1]

2 Construction Grammar

2.1 Overview

A core assumption of generative grammar (Chomsky, 1988), which can be already found in Bloomfieldian structural linguistics (Bloomfield, 1933), is a strict separation of lexicon and grammar: grammar is conceptualized as a set of compositional and general rules that operate on a list of arbitrary and specific lexical items in generating syntactically well-formed sentences. This dichotomous view was increasingly questioned in the 1980s when several studies drew attention to the fact that linguistic units larger than lexical items (e.g., idioms) can also possess non-compositional meanings (Langacker, 1987; Lakoff, 1987; Fillmore et al., 1988; Fillmore, 1989). For instance, it is not clear how the effect of the words "let alone"(as in "she doesn't eat fish, let alone meat") on both the syntax and the semantics of the rest of the sentence could be inferred from general syntactic rules (Fillmore et al., 1988).. This insight about the ubiquity of stored form-meaning pairings in language is adopted as the central tenet of grammatical theory by Construction Grammar (CxG; see Hoffmann and Trousdale (2013) for a comprehensive overview). Rather than a system divided into non-overlapping syntactic rules and lexical items, CxG views language as a structured system of constructions with varying granularities that encapsulate syntactic and semantic components as single linguistic signs—ranging from individual morphemes up to phrasal elements and fixed expressions (Kay and Fillmore, 1999; Goldberg, 1995). In this framework, syntactic rules can be seen as emergent abstractions over similar stored constructions (Goldberg, 2003, 2006). A different set of stored constructions can result in different abstractions and thus different syntactic rules, which allows CxG to naturally accommodate for the dynamic nature of grammar as evidenced, for instance, by inter-speaker variability and linguistic change (Hilpert, 2006).

2.2 Construction Grammar and NLP

We see three main motivations for the development of a first probing approach for CxG:

– We believe that the active discourse in (cognitive) linguistics about the best description of human language capability can be supported and enriched through a computational exploration of a wide array of phenomena and viewpoints. We think that the probing literature in NLP investigating linguistic phenomena with computational methods should be diversified to include theories and problems from all points on the broad spectrum of linguistic scholarship.

– We hope that the investigation of large PLMs' apparent capabilities to imitate human language and the mechanisms responsible for these capabilities will be enriched by introducing a usage-based approach to grammar. This is especially important as some of the discourse in recent years has focused on the question of whether PLMs are constructing syntactically acceptable sentences for the correct reasons and with the correct underlying representations (e.g. McCoy et al., 2019). We would like to suggest that considering alternative theories of grammar, specifically CxG with

[1] In order to foster research at the intersection of NLP and construction grammar, we will make our data and code available at https://github.com/LeonieWeissweiler/ComparativeCorrelative.

its incorporation of slots in constructions that may be filled by specific word types and its focus on learning without an innate, universal grammar, may be beneficial to understanding the learning process of PLMs as their capabilities advance further.

– Many constructions present an interesting challenge for PLMs. In fact, recent work in challenge datasets (Ribeiro et al., 2020) has already started using what could be considered constructions, in an attempt to identify types of sentences that models struggle with, and to point out a potential direction for improvement. One of the central tenets of CxG is the relation between the form of a construction and its meaning, or to put it in NLP terms, a model must learn to infer parts of the sentence meaning from patterns that are present in it, as opposed to words. We believe this to be an interesting challenge for future PLMs.

2.3 The English Comparative Correlative

The English comparative correlative (CC) is one of the most commonly studied constructions in linguistics, for several reasons. Firstly, it constitutes a clear example of a linguistic phenomenon that is challenging to explain in the framework of generative grammar (Culicover and Jackendoff, 1999; Abeillé and Borsley, 2008), even though there have been approaches following that school of thought (Den Dikken, 2005; Iwasaki and Radford, 2009). Secondly, it exhibits a range of interesting syntactic and semantic features, as detailed below. These reasons, we believe, also make the CC an ideal testbed for a first study attempting to extend the current trend of syntax probing for rules by developing methods for probing according to CxG.

The CC can take many different forms, some of which are exemplified here:

(1) The more, the merrier.

(2) The longer the bake, the browner the colour.

(3) The more she practiced, the better she became.

Semantically, the CC consists of two clauses, where the second clause can be seen as the dependent variable for the independent variable specified in the first one (Goldberg, 2003). It can be seen on the one hand as a statement of a general cause-and-effect relationship, as in a general conditional statement (e.g., (2) could be paraphrased as "If the bake is longer, the colour will be more brown"), and on the other as a temporal development in a comparative sentence (paraphrasing (3) as "She became better over time, and she practiced more over time"). Usage of the CC typically implies both readings at the same time. Syntactically, the CC is characterised in both clauses by an instance of "the" followed by an adverb or an adjective in the comparative, either with "-er" for some adjectives and adverbs, or with "more" for others, or special forms like "better". Special features of the comparative sentences following this are the optional omission of the future "will" and of "be", as in (1). Crucially, "the" in this construction does not function as a determiner of noun phrases (Goldberg, 2003); rather, it has a function specific to the CC and has variously been called a "degree word" (Den Dikken, 2005) or "fixed material" (Hoffmann et al., 2019).

3 Syntax

Our investigation of PLMs' knowledge of the CC is split into two parts. First, we probe for the PLMs' knowledge of the syntactic aspects of the CC, to determine if they recognise its structure. Then we devise a test of their understanding of its semantic aspects by investigating their ability to apply, in a given context, information conveyed by a CC.

3.1 Probing Methods

As the first half of our analysis of PLMs' knowledge of the CC, we investigate its syntactic aspects. Translated into probing questions, this means that we ask: can a PLM recognise an instance of the CC? Can it distinguish instances of the CC from similar-looking non-instances? Is it able to go beyond the simple recognition of its fixed parts ("The `COMP-ADJ/ADV`, the ...") and group all ways of completing the sentences that are instances of the CC separately from all those that are not? And to frame all of these questions in a syntactic probing framework: will we be able to recover, using a logistic regression as the probe, this distinguishing information from a PLM's embeddings?

The established way of testing a PLM for its syntactic knowledge has in recent years become minimal pairs (e.g., Warstadt et al., 2020, Demszky et al., 2021). This would mean pairs of sentences which are indistinguishable except for the fact that one of them is an instance of the CC and the other is not, allowing us to perfectly separate a model's knowledge of the CC from other confounding factors. While this is indeed possible for simpler syntactic phenomena such as verb-noun number agreement, there is no obvious way to construct minimal pairs for the CC. We therefore construct minimal pairs in two ways: one with artificial data based on a context-free grammar (CFG), and one with sentences extracted from C4.

3.1.1 Synthetic Data

In order to find a pair of sentences that is as close as possible to a minimal pair, we devise a way to modify the words following "The X-er" such that the sentence is no longer an instance of the construction. The pattern for a positive instance is "The ADV-er the NUM NOUN VERB", e.g., "The harder the two cats fight". To create a negative instance, we reorder the pattern to "The ADJ-er NUM VERB the NOUN", e.g., "The harder two fight the cats". The change in role of the numeral from the dependent of a head to a head itself, made possible by choosing a verb that can be either transitive or intransitive, as well as the change from an adverb to an adjective, allows us to construct a negative instance that uses the same words as the positive one, but in a different order.[2] In order to generate a large number of instances, we collect two sets each of adverbs, numerals, nouns and verbs that are mutually exclusive between training and test sets. To investigate if the model is confused by additional content in the sentences, we write an CFG to insert phrases before the start of the first half, in between the two halves, and after the second half of the CC (see Appendix, Algorithms 1 and 2 for the complete CFG).

While this setup is rigourous in the sense that positive and negative sentences are exactly matched, it comes with the drawback of only considering one type of CC. To be able to conduct a more comprehensive investigation, we adopt a complementary approach and turn to pairs extracted from C4 (see Appendix, Tables 6 and 7, for examples of training and test data). These cover a broad range of CC patterns, albeit without meeting the criterion that positive and negative samples are exactly matched.

3.1.2 Corpus-based Minimal Pairs

While accepting that positive and negative instances extracted from a corpus will automatically not be minimal and therefore contain some lexical overlap and context cues, we attempt to regularise our retrieved instances as far as possible. To form a first candidate set, we POS tag C4 using spaCy (Honnibal and Montani, 2018) and extract all sentences that follow the pattern "The" (DET) followed by either "more" and an adjective or adverb, or an adjective or adverb ending in "-er", and at any point later in the sentence again the same pattern. We discard examples with adverbs or adjectives that were falsely labelled as comparative, such as "other". We then group these sentences by their sequence of POS tags, and manually classify the sequences as either positive or negative instances. We observe that sentences sharing a POS tag pattern tend to be either all negative or all positive instances, allowing us to save annotation time by working at the POS tag pattern level instead of the sentence level. To make the final set as diverse as possible, we sort the patterns randomly and label as many as possible. In order to further reduce interfering factors in our probe, we separate the POS tag patterns between training and test sets (see Appendix, Table 8, for examples).

3.1.3 The Probe

For both datasets, we investigate the overall accuracy of our probe as well as the impact of several factors. The probe consists of training a simple logistic regression model on top of the mean-pooled sentence embeddings (Vulić et al., 2020). To quantify the impact of the length of the sentence, the start position of the construction, the position of its second half, and the distance between them, we construct four different subsets D_f^{train} and D_f^{test} from both the artificially constructed and the corpus-based dataset. For each subset, we sample sentences such that both the positive and the negative class is balanced across every value of the feature within a certain range of values. This ensures that the probes are unable to exploit correlations between a class and any of the above features. We create the dataset as follows

$$D_f = \bigcup_{v \in f_v} \bigcup_{l^* \in L} S(D, v, l^*, n^*),$$

where f is the feature, f_v is the set of values for f, $L = \{positive, negative\}$ are the labels, and S is a function that returns n^* elements from D that have value v and label l^*.

To make this task more cognitively realistic, we aim to test if a model is able to generalise from shorter sentences, which contain relat-

[2] Note that an alternative reading of this sentence exists: the numeral "two" forms the noun phrase by itself and "The harder" is still interpreted as part of the CC. The sentence is actually a positive instance on this interpretation. We regard this reading as very improbable.

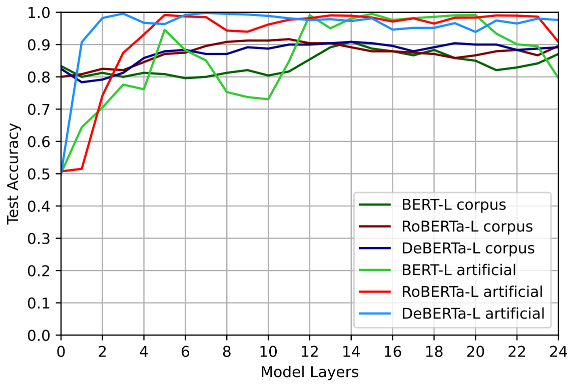

Figure 1: Overall accuracy per layer for D_{length}. All shown models are the large model variants. The models can easily distinguish between positive and negative examples in at least some of their layers.

ively little additional information besides the parts relevant to the classification task, to those with greater potential interference due to more additional content that is not useful for classification. Thus, we restrict the training set to samples from the lowest quartile of each feature so that f_v becomes $[v_f^{\min}, v_f^{\min}+\frac{1}{4}(v_f^{\max}-v_f^{\min})]$ for D_f^{train} and $[v_f^{\min}, v_f^{\max}]$ for D_f^{test}. We report the test performance for every value of a given feature separately to recognise patterns. For the artificial syntax probing, we generate 1000 data points for each value of each feature for each training and test for each subset associated with a feature. For the corpus syntax probing, we collect 9710 positive and 533 negative sentences in total, from which we choose 10 training and 5 test sentences for each value of each feature in a similar manner. To improve comparability and make the experiment computationally feasible, we test the "large" size of each of our three models, using the Huggingface Transformers library (Wolf et al., 2019). Our logistic regression probes are implemented using Scikitlearn (Pedregosa et al., 2011).

3.2 Probing Results

3.2.1 Artificial Data

As shown in Figure 1, the results of our syntactic probe indicate that all models can easily distinguish between positive and negative examples in at least some of their layers, independently of any of the sentence properties that we have investigated. We report full results in the Appendix in Figures 2, 3, and 4. We find a clear trend that DeBERTa performs better than RoBERTa, which in turn performs better than BERT across the board.

As DeBERTa's performance in all layers is nearly perfect, we are unable to observe patterns related to the length of the sentence, the start position of the CC, the start position of the second half of the CC, and the distance between them. By contrast, we observe interesting patterns for BERT and RoBERTa. For D_{length}, and to a lesser degree D_{distance} (which correlates with it), we observe that at first, performance goes down with increased length as we would expect—the model struggles to generalise to longer sentences with more interference since it was only trained on short ones. However, this trend is reversed in the last few layers. We hypothesize this may be due to an increased focus on semantics in the last layers (Peters et al., 2018; Tenney et al., 2019), which could lead to interfering features particularly in shorter sentences.

3.2.2 Corpus Data

In contrast, the results of our probe on more natural data from C4 indicate two different trends: first, as the positive and negative instances are not identical on a bag-of-word level, performance is not uniformly at 50% (i.e., chance) level in the first layers, indicating that the model can exploit lexical cues to some degree. We observe a similar trend as with the artificial experiment, which showed that DeBERTa performs best and BERT worst. The corresponding graphs can be found in the Appendix in Figures 5, 6, and 7.

Generally, this additional corpus-based experiment validates our findings from the experiment with artificially generated data, as all models perform at 80% or better from the middle layers on, indicating that the models are able to classify instances of the construction even when they are very diverse and use unseen POS tag patterns.

Comparing the average accuracies on D_{length} for both data sources in Figure 1, we observe that all models perform better on artificial than on corpus data from the fifth layer on, with the notable exception of a dip in performance for BERT large around layer 10.

4 Semantics

4.1 Probing Methods

4.1.1 Usage-based Testing

For the second half of our investigation, we turn to semantics. In order to determine if a model has understood the meaning of the CC, i.e., if it has understood that in any sentence, "the `COMP` .... the

No.	Purpose	Approach	Sentence Schema
S1	Base		The ADJ1-er you are, the ADJ2-er you are. The ANT1-er you are, the ANT2-er you are. NAME1 is ADJ1-er than NAME2. Therefore, NAME1 is [MASK] than NAME2.
S2	Bias Test	Recency	The ANT1-er you are, the ANT2-er you are. The ADJ1-er you are, the ADJ2-er you are. NAME1 is ADJ1-er than NAME2. Therefore, NAME1 is [MASK] than NAME2.
S3		Vocabulary	The ADJ1-er you are, the ANT2-er you are. The ANT1-er you are, the ADJ2-er you are. NAME2 is ADJ1-er than NAME2. Therefore, NAME1 is [MASK] than NAME2.
S4		Name	The ADJ1-er you are, the ADJ2-er you are. The ANT1-er you are, the ANT2-er you are. NAME2 is ADJ1-er than NAME1. Therefore, NAME2 is [MASK] than NAME1.
S5	Calibration	Short	NAME1 is ADJ1-er than NAME2. Therefore, NAME1 is [MASK] than NAME2.
S6		Name	The ADJ1-er you are, the ADJ2-er you are. The ANT1-er you are, the ANT2-er you are. NAME1 is ADJ1-er than NAME2. Therefore, NAME3 is [MASK] than NAME4.
S7		Adjective	The ADJ1-er you are, the ADJ2-er you are. The ANT1-er you are, the ANT2-er you are. NAME1 is ADJ3-er than NAME2. Therefore, NAME1 is [MASK] than NAME2.

Table 1: Overview of the schemata of all test scenarios used for semantic probing

COMP" implies a correlation between the two halves, we adopt a usage-based approach and ask: can the model, based on the meaning conveyed by the CC, draw a correct inference in a specific scenario? For this, we construct general test instances of the CC that consist of a desired update of the belief state of the model about the world, which we then expect it to be able to apply. More concretely, we generate sentences of the form "The ADJ1-er you are, the ADJ2-er you are.", while picking adjectives at random. To this general statement, we then add a specific scenario with two random names: "NAME1 is ADJ1-er than NAME2." and ask the model to draw an inference from it by predicting a token at the masked position in the following sentence: "Therefore, NAME1 is [MASK] than NAME2." If the model has understood the meaning conveyed by the CC and is able to use it in predicting the mask, we expect the probability of ADJ2 to be high. To provide the model with an alternative, we add a second sentence, another instance of the CC, using the antonyms of the two adjectives. This sentence is carefully chosen to have no impact on the best filler for [MASK], but also for other reasons explained in Section 4.1.2. The full test context is shown in Table 1, S1. This enables us to compare the probability of ADJ2 for the mask token directly with a plausible alternative, ANT2. One of our test sentences might be "The stronger you are, the faster you are. The weaker you are, the slower you are. Terry is stronger than John. Therefore, Terry will be [MASK] than John", where we compare the probabilities of "faster" and "slower".

Note that success in our experiment does not necessarily indicate that the model has fully understood the meaning of the CC. The experiment can only provide a lower bound for the underlying understanding of any model. However, we believe that our task is not unreasonable for a masked language model in a zero-shot setting. It is comparable in difficulty and non-reliance on world knowledge to the NLU tasks presented in LAMBADA (Paperno et al., 2016), on which GPT-2 (117M to 1.5B parameters) has achieved high zero-shot accuracy (Radford et al., Table 3). While we investigate masked language models and not GPT-2, our largest models are comparable in size to the sizes of GPT-2 that were used (340M for BERT_L, 355M for RoBERTa_L, and 1.5B parameters for DeBERTa-XXL_L), and we believe that this part of our task is achievable to some degree.

4.1.2 Biases

In this setup, we hypothesise several biases that models could exhibit and might cloud our assessment of its understanding of the CC, and devise a way to test their impact.

Firstly, we expect that models might prefer to repeat the adjective that is closest to the mask token. This has recently been documented for prompt-based experiments (Zhao et al., 2021). Here, this adjective is ANT2, the wrong answer. To test the influence this has on the prediction probabilities, we construct an alternative version of our test context in which we flip the first two sentences so that the correct answer is now more recent. The result can be found in Table 1, S2.

Secondly, we expect that models might assign higher probabilities to some adjectives, purely based on their frequency in the pretraining corpus, as for example observed by Holtzman et al. (2021). To test this, we construct a version of the test context in which `ADJ2`/`ANT2` are swapped, which means that we can keep both the overall words the same as well as the position of the correct answer, while changing which adjective it is. The sentence is now S3 in Table 1. If there is a large difference between the prediction probabilities for the two different versions, that this means that a model's prediction is influenced by the lexical identity of the adjective in question.

Lastly, a model might have learned to associate adjectives with names in pretraining, so we construct a third version, in which we swap the names. This is S4 in Table 1. If any prior association between names and adjectives influences the prediction, we expect the scores between S4 and S1 to differ.

4.1.3 Calibration

After quantifying the biases that may prevent us from seeing a model's true capability in understanding the CC, we aim to develop methods to mitigate it. We turn to calibration, which has recently been used in probing with few-shot examples by Zhao et al. (2021). The aim of calibration is to improve the performance of a model on a classification task, by first assessing the prior probability of a label (i.e., its probability if no context is given), and then dividing the probability predicted in the task context by this prior; this gives us the conditional probability of a label given the context, representing the true knowledge of the model about this task. In adapting calibration, we want to give a model every possible opportunity to do well so that we do not underestimate its underlying comprehension.

We therefore develop three different methods of removing the important information from the context in such a way that we can use the prediction probabilities of the two adjectives in these contexts for calibration. The simplest way of doing this is to remove both instances of the CC, resulting in S5 in Table 1. If we want to keep the CC in the context, the two options to remove any information are to replace either the names or the adjectives with new names/adjectives. We therefore construct two more instances for calibration: S6 and S7 in Table 1.

For each calibration method, we collect five examples with different adjectives or names. For a given base sample S_b, we calculate P_c, the calibrated predictions, as follows:

$$P_c(a|S_b) = P(a|S_b)/[\sum_{i=1}^{i=5}(P(a|C_i)/5)]$$

where C_i is the i-th example of a given calibration technique, a is the list of adjectives tested for the masked position, and the division is applied elementwise. We collect a list of 20 adjectives and their antonyms manually from the vocabulary of the RoBERTa tokenizer and 33 common names and generate 144,800 sentences from them. We test BERT (Devlin et al., 2019) in the sizes base and large, RoBERTa (Liu et al., 2019) in the sizes base and large, and DeBERTa (He et al., 2020) in the sizes base, large, xlarge and xxlarge.

	Accuracy		Decision Flip		
	S1	S2	S2	S3	S4
BERT_B	37.65	64.64	26.98	75.69	02.70
BERT_L	36.85	67.21	30.44	73.31	02.32
RoBERTa_B	61.60	52.84	09.91	76.18	02.76
RoBERTa_L	55.71	68.00	14.33	79.47	04.33
DeBERTa_B	49.72	49.80	00.91	99.66	01.07
DeBERTa_L	50.88	51.40	07.04	94.83	02.23
DeBERTa_XL	47.73	49.33	05.46	89.28	02.51
$\text{DeBERTa}_\text{XXL}$	47.34	48.72	03.59	82.09	01.13

Table 2: Selected accuracies and results for the semantic probe. We report the average accuracy on the more difficult sentences in terms of recency bias (S1) and the easier ones (S2), as well as the percentage of decisions flipped by changing from the base S1 to the sentences testing for recency bias (S2), vocabulary bias (S3), and name bias (S4). RoBERTa and DeBERTa perform close to chance on S1 and S2 accuracy, indicating that they do not understand the meaning of CC. BERT's performance is strongly influenced by biases (recency, lexical identity), also indicating that it has very limited if any understanding of CC.

4.2 Results

In Table 2, we report the accuracy for all examined models. Out of the three variations to test biases, we report accuracy only for the sentence testing the recency bias as we expect this bias to occur systematically across all sentences: if it is a large effect, it will always lead to the sentence where the correct answer is the more recent one being favoured. To assess the influence of each bias beyond accuracy, we report as decision flip the percentage of sentences for which the decision (i.e., if the correct adjective had a higher probability than the incorrect one) was changed when considering the alternative sentence that was constructed to test for bias. We report full results in Appendix, Table 4.

Looking at the accuracies, we see that RoBERTa's and DeBERTa's scores are close to 50% (i.e., chance) accuracy for both S1 and S2. BERT models differ considerably as they seem to suffer from bias related to the order of the two CCs, but we can see that the average between them is also very close to chance. When we further look at the decision flips for each of the biases, we find that there is next to no bias related to the choice of names (S4). However, we can see a large bias related to both the recency of the correct answer (S2) and the choice of adjectives (S3). The recency bias is strongest in the BERT models, which also accounts for the difference in accuracies. For RoBERTa and DeBERTa models, the recency bias is small, but clearly present. In contrast, they exhibit far greater bias towards the choice of adjective, even going as far as 99.66% of decisions flipped by changing the adjective for DeBERTa base. This suggests that these models' decisions about which adjective to assign a higher probability is almost completely influenced by the choice of adjective, not the presence of the CC. Overall, we conclude that without calibration, all models seem to be highly susceptible to different combinations of bias, which completely obfuscate any underlying knowledge of the CC, leading to an accuracy at chance level across the board.

We therefore turn to our calibration methods, evaluating them first on their influence on the decision flip scores, which directly show if we were able to reduce the impact of the different types of bias. We report these only for order and vocabulary bias as we found name bias to be inconsequential. We report the complete results in Appendix, Tables 4 and 5. We see that across all models, while all three calibration methods work to reduce some bias, none does so consistently across all models or types of bias. We report the impact of all calibration methods on the final accuracies of the three largest models in Table 3. Even in cases where calibration has clearly reduced the decision flip score, we find that the final calibrated accuracy is still close to 50%. This indicates that despite the effort to retrieve any knowledge that the models have about the CC, they are unable to perform clearly above chance, and we have therefore found no evidence that the investigated models understand and can use the semantics of the CC.

Model	Test	-	S5	S6	S7
BERT_{L}	S1	36.85	31.91	47.21	44.03
	S2	67.13	73.48	54.39	64.45
	S3	36.46	43.43	47.79	44.36
$\text{RoBERTa}_{\text{L}}$	S1	55.72	58.37	65.08	69.53
	S2	68.01	74.53	62.73	77.76
	S3	55.36	52.02	65.28	69.23
$\text{DeBERTa}_{\text{XXL}}$	S1	47.35	53.56	54.92	54.12
	S2	48.73	52.85	54.03	53.81
	S3	47.57	49.36	55.25	53.59

Table 3: Effect of our three calibration methods compared to no calibration, for the three largest models. We report the accuracy scores for the base sentence (S1), recency bias (S2), and vocabulary bias (S3). The results indicate that, even if we try to address bias through calibration, the models are unable to perform clearly above chance. We have therefore found no evidence that the models understand the semantics of the CC.

4.2.1 Problem Analysis

Different conclusions might be drawn as to why none of these models have learned the semantics of the CC. They might not have seen enough examples of it in their training corpus to have formed a general understanding. Given the many examples that we were able to find in C4, and the overall positive results from the syntax section, we find this to be unlikely. Alternatively, it could be argued that models have never had a chance to learn what the CC means because they have never seen it applied, and do not have the same opportunities as humans to either interact with the speaker to clarify the meaning or to make deductions using observations in the real world. This is in line with other considerations about large PLMs acquiring advanced semantics, even though it has for many phenomena been shown that pretraining is sufficient (Radford et al., 2019). Lastly, it might be possible that the type of meaning representation required to solve this task is beyond the current transformer-style architectures. Overall, our finding that PLMs do not learn the semantics of the CC adds to the growing body of evidence that complex semantics like negation (Kassner and Schütze, 2020) is still beyond state-of-the-art PLMs.

5 Related Work

5.1 Construction Grammar in NLP

CxG has only recently and very sparsely been investigated in neural network-based NLP. Tayyar Madabushi et al. (2020) use a probe to show that while a probe on top of BERT contextual embeddings is able to mostly correctly classify if two sentences contain instances of the same construction, injecting this knowledge into the model by adding it to pretraining does not improve its performance. Our work differs from this study in that we delve deeper into what it means to understand a construction on a semantic level, and take careful precautions to isolate the recognition of the construction at the syntax level from confounding factors. Li et al. (2022) recreate the experiments of Bencini and Goldberg (2000) and Johnson and Goldberg (2013) on argument structure constructions, by creating artificial sentences with four major argument structure types and a random combination of verbs, to investigate whether PLMs prefer sorting by construction or by main verb. Tseng et al. (2022) choose items from a Chinese construction list and investigate PLM's predictions when masking the open slots, the closed slots, or the entire construction. They find that models find closed slots easier to predict than open ones. Other computational studies about CxG have either focused on automatically annotating constructions (Dunietz et al., 2017) or on the creation and evaluation of automatically built lists of constructions (Marques and Beuls, 2016; Dunn, 2019).

5.2 Probing

Our work also bears some similarity to recent work in generative grammar-based syntax probing of large PLMs in that we approximate the minimal pairs-based probing framework similar to Wei et al. (2021), Marvin and Linzen (2018) or Goldberg (2019). However, as we are concerned with different phenomena and investigating them from a different theoretical standpoint, the syntactic half of our work clearly differs.

The semantic half of our study is closest to recent work on designing challenging test cases for models such as Ribeiro et al. (2020), who design some edge cases for which most PLMs fail. Despite the different motivation, the outcome is very similar to a list of some particularly challenging constructions.

6 Conclusion

We have made a first step towards a thorough investigation of the compatibility of the paradigm of CxG and the syntactic and semantic capabilities exhibited by state-of-the-art large PLMs. For this, we chose the English comparative correlative, one of the most well-studied constructions, and investigated if large PLMs have learned it, both syntactically and semantically. We found that even though they are able to classify sentences as instances of the construction even in difficult circumstances, they do not seem to be able to extract the meaning it conveys and use it in context, indicating that while the syntactic aspect of the CC is captured in pretraining, the semantic aspect is not. We see this an indication that major future work will be needed to enable neural models to fully understand language to the same degree as humans.

Limitations

As our experimental setup requires significant customisation with regards to the properties of the specific construction we investigate, we are unable to consider other constructions or other languages in this work. We hope to be able to extend our experiments in this direction in the future. Our analysis is also limited—as all probing papers are—by the necessary indirectness of the probing tasks: we cannot directly assess the model's internal representation of the CC, but only construct tasks that might show it but are imperfect and potentially affected by external factors.

Acknowledgements

This work was funded by the European Research Council (#740516). The second author was also supported by the German Academic Scholarship Foundation. The third author was also supported by the German Federal Ministry of Education and Research (BMBF, Grant No. 01IS18036A). We thank the reviewers for their extremely helpful comments. We are also very grateful to David Mortensen and Lori Levin for helpful discussions and comments.

References

Anne Abeillé and Robert D Borsley. 2008. Comparative correlatives and parameters. *Lingua*, 118(8):1139–1157.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Giulia ML Bencini and Adele E Goldberg. 2000. The contribution of argument structure constructions to sentence meaning. *Journal of Memory and Language*, 43(4):640–651.

Leonard Bloomfield. 1933. *Language*. Holt, Rinehart & Winston, New York, NY.

Noam Chomsky. 1988. Generative grammar. *Studies in English linguistics and literature*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Peter W Culicover and Ray Jackendoff. 1999. The view from the periphery: The english comparative correlative. *Linguistic inquiry*, 30(4):543–571.

Dorottya Demszky, Devyani Sharma, Jonathan H. Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. Learning to recognize dialect features. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2021*.

Marcel Den Dikken. 2005. Comparative correlatives comparatively. *Linguistic Inquiry*, 36(4):497–532.

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.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. Automatically tagging constructions of causation and their slot-fillers. *Transactions of the Association for Computational Linguistics*, 5:117–133.

Jonathan Dunn. 2019. Frequency vs. association for constraint selection in usage-based construction grammar. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 117–128, Minneapolis, Minnesota. Association for Computational Linguistics.

Charles J Fillmore. 1986. Varieties of conditional sentences. In *Eastern States Conference on Linguistics*, volume 3, pages 163–182.

Charles J. Fillmore. 1989. Grammatical construction: Theory and the familiar dichotomies. In Rainer Dietrich and Carl F. Graumann, editors, *Language processing in social context*, pages 17–38. North-Holland, Amsterdam.

Charles J. Fillmore, Paul Kay, and Mary C. O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64(3):501–538.

Adele Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, Chicago, IL.

Adele Goldberg. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press, Oxford, UK.

Adele E Goldberg. 2003. Constructions: A new theoretical approach to language. *Trends in cognitive sciences*, 7(5):219–224.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Martin Hilpert. 2006. A synchronic perspective on the grammaticalization of Swedish future constructions. *Nordic Journal of Linguistics*, 29(2):151–173.

Thomas Hoffmann, Jakob Horsch, and Thomas Brunner. 2019. The more data, the better: A usage-based account of the english comparative correlative construction. *Cognitive Linguistics*, 30(1):1–36.

Thomas Hoffmann and Graeme Trousdale. 2013. *The Oxford handbook of construction grammar*. Oxford University Press.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2018. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Eiichi Iwasaki and Andrew Radford. 2009. Comparative correlatives in english: A minimalist-cartographic analysis.

Matt A Johnson and Adele E Goldberg. 2013. Evidence for automatic accessing of constructional meaning: Jabberwocky sentences prime associated verbs. *Language and Cognitive Processes*, 28(10):1439–1452.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Paul Kay and Charles J. Fillmore. 1999. Grammatical constructions and linguistic generalizations: The *What's X doing Y?* construction. *Language*, 75(1):1–33.

George Lakoff. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago Press, Chicago, IL.

Ronald W. Langacker. 1987. *Foundations of cognitive grammar: Theoretical prerequisites*. Stanford University Press, Stanford, CA.

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7410–7423, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tânia Marques and Katrien Beuls. 2016. Evaluation strategies for computational construction grammars. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1137–1146, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT) 2018*.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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. In *International Conference on Learning Representations (ICLR) 7*.

Yu-Hsiang Tseng, Cing-Fang Shih, Pin-Er Chen, Hsin-Yu Chou, Mao-Chang Ku, and Shu-Kai Hsieh. 2022. CxLM: A construction and context-aware language model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6361–6369, Marseille, France. European Language Resources Association.

Ivan Vulić, Edoardo M. Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of* *the Association for Computational Linguistics*, 8:377–392.

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Algorithm 1 Context-Free Grammar for Artificial Data Creation Training Set

S → SPOS | SNEG
SPOS → POS1 PUNCT POS2 '.' | POS1 INSERT PUNCT POS2 '.'
SNEG → NEG1 PUNCT NEG2 '.' | NEG1 INSERT PUNCT NEG2 '.'
PUNCT → ',' | ';' | ''
CORE_POS → ADV_I 'the' NUM NOUN VERB
CORE_NEG → ADV_I NUM VERB 'the' NOUN
POS_UPPER → '0 The' CORE_POS
POS_LOWER → '0 the' CORE_POS
NEG_UPPER → '0 The' CORE_NEG
NEG_LOWER → '0 the' CORE_NEG
POS1 → POS_UPPER | POS_UPPER ADD | START POS_LOWER | START POS_LOWER ADD
POS2 → POS_LOWER | POS_LOWER ADD
NEG1 → NEG_UPPER | NEG_UPPER ADD | START NEG_LOWER | START NEG_LOWER ADD
NEG2 → NEG_LOWER | NEG_LOWER ADD
INSERT → INSERT1 | INSERT2
INSERT2 → ADDITION BETWEEN_ADD_AND_SENT SENT
PRON → 'we' | 'they'
ADDITION → ', and by the way ,' | ', and I want to add that' | ', and' PRON 'just want to say that' | ', and then' PRON 'said that' | ', and then' PRON 'said that'
SAY → 'say' | 'think' | 'mean' | 'believe'
BETWEEN_ADD_AND_SENT → PRON SAY 'that' | PRON SAY 'that' | PRON SAY 'that' | PRON SAY 'that'
LOC_SENT → PRON 'said this in' LOC 'too'
LOC → CITY 'and' LOC | CITY
CITY → 'Munich' | 'Washington' | 'Cologne' | 'Prague' | 'Istanbul'
SENT → 'this also holds in other cases' | 'this is not always true' | 'this is always true' | 'this has only recently been the case' | 'this has not always been the case' | 'this has always been the case'
INSERT1 → 'without stopping' | 'without a break' | 'without a pause' | 'uninterrupted' |
START → 'Nowadays ,' | 'Nowadays' | 'Therefore ,' | 'Therefore' | 'We can' CANWORD 'that' | 'It is' KNOWNWORD 'that' | 'It follows that' | 'Sometimes' | 'Sometimes ,' | 'It was recently announced that' | 'People have told me that' | 'I recently read in a really interesting book that' | 'I have recently read in an established , well-known newspaper that' | 'It was reported in a special segment on TV today that'
CANWORD → 'say' | 'surmise' | 'accept' | 'state'
KNOWNWORD → 'clear' | 'known' | 'accepted' | 'obvious'
ADD → TEMP | UNDER1 | TEMP UNDER1 | UNDER1 TEMP
ADV_I → ADV | ADV 'and' ADV
TEMP → TEMP1 TEMP2
TEMP1 → 'before' | 'after' | 'during'
TEMP2 → 'the morning' | 'the afternoon' | 'the night'
UNDER1 → 'under the' UNDER2
UNDER2 → 'bed' | 'roof' | 'sun'
VERB → 'push' | 'attack' | 'chase' | 'beat' | 'believe' | 'boil' | 'box' | 'burn' | 'call' | 'date'
NOUN → 'lions' | 'pandas' | 'camels' | 'pigs' | 'horses' | 'sheep' | 'chickens' | 'foxes' | 'cows' | 'deer'
ADV → 'worse' | 'earlier' | 'slower' | 'deeper' | 'bigger' | 'smaller' | 'flatter' | 'weaker' | 'stronger' | 'louder'
NUM → 'twelve' | 'thirteen' | 'fourteen' | 'fifteen' | 'sixteen' | 'seventeen' | 'eighteen' | 'nineteen' | 'twenty' | 'twenty-one'

Algorithm 2 Context-Free Grammar for Artificial Data Creation Test Set

S → SPOS | SNEG
SPOS → POS1 PUNCT POS2 '.' | POS1 INSERT PUNCT POS2 '.'
SNEG → NEG1 PUNCT NEG2 '.' | NEG1 INSERT PUNCT NEG2 '.'
PUNCT → ',' | ';' | ''
CORE_POS → ADV_I 'the' NUM NOUN VERB
CORE_NEG → ADV_I NUM VERB 'the' NOUN
POS_UPPER → '0 The' CORE_POS
POS_LOWER → '0 the' CORE_POS
NEG_UPPER → '0 The' CORE_NEG
NEG_LOWER → '0 the' CORE_NEG
POS1 → POS_UPPER | POS_UPPER ADD | START POS_LOWER | START POS_LOWER ADD
POS2 → POS_LOWER | POS_LOWER ADD
NEG1 → NEG_UPPER | NEG_UPPER ADD | START NEG_LOWER | START NEG_LOWER ADD
NEG2 → NEG_LOWER | NEG_LOWER ADD
INSERT → INSERT1 | INSERT2
INSERT2 → ADDITION BETWEEN_ADD_AND_SENT SENT
PRON → 'I' | 'you'
ADDITION → ', and by the way ,' | ', and I want to add that' | ', and' PRON 'just want to say that' | ', and then' PRON 'said that' | ', and then' PRON 'said that'
SAY → 'say' | 'think' | 'mean' | 'believe'
BETWEEN_ADD_AND_SENT → PRON SAY 'that' | PRON SAY 'that' | PRON SAY 'that' | PRON SAY 'that'
LOC_SENT → PRON 'said this in' LOC 'too'
LOC → CITY 'and' LOC | CITY
CITY → 'London' | 'New York' | 'Berlin' | 'Madrid' | 'Paris'
SENT → 'this also holds in other cases' | 'this is not always true' | 'this is always true' | 'this has only recently been the case' | 'this has not always been the case' | 'this has always been the case'
INSERT1 → 'without stopping' | 'without a break' | 'without a pause' | 'uninterrupted' |
START → 'Nowadays ,' | 'Nowadays' | 'Therefore ,' | 'Therefore' | 'We can' CANWORD 'that' | 'It is' KNOWNWORD 'that' | 'It follows that' | 'Sometimes' | 'Sometimes ,' | 'It was recently announced that' | 'People have told me that' | 'I recently read in a really interesting book that' | 'I have recently read in an established , well-known newspaper that' | 'It was reported in a special segment on TV today that'
CANWORD → 'say' | 'surmise'
KNOWNWORD → 'clear' | 'known'
ADD → TEMP | UNDER1 | TEMP UNDER1 | UNDER1 TEMP
ADV_I → ADV | ADV 'and' ADV
TEMP → TEMP1 TEMP2
TEMP1 → 'before' | 'after' | 'during'
TEMP2 → 'the day' | 'the night' | 'the evening'
UNDER1 → 'under the' UNDER2
UNDER2 → 'bridge' | 'stairs' | 'tree'
VERB → 'slam' | 'break' | 'bleed' | 'shake' | 'smash' | 'throw' | 'strike' | 'shoot' | 'swallow' | 'choke'
NOUN → 'cats' | 'dogs' | 'girls' | 'boys' | 'men' | 'women' | 'people' | 'humans' | 'mice' | 'alligators'
ADV → 'faster' | 'quicker' | 'harder' | 'higher' | 'later' | 'longer' | 'shorter' | 'lower' | 'wider' | 'better'
NUM → 'two' | 'three' | 'four' | 'five' | 'six' | 'seven' | 'eight' | 'nine' | 'ten' | 'eleven'

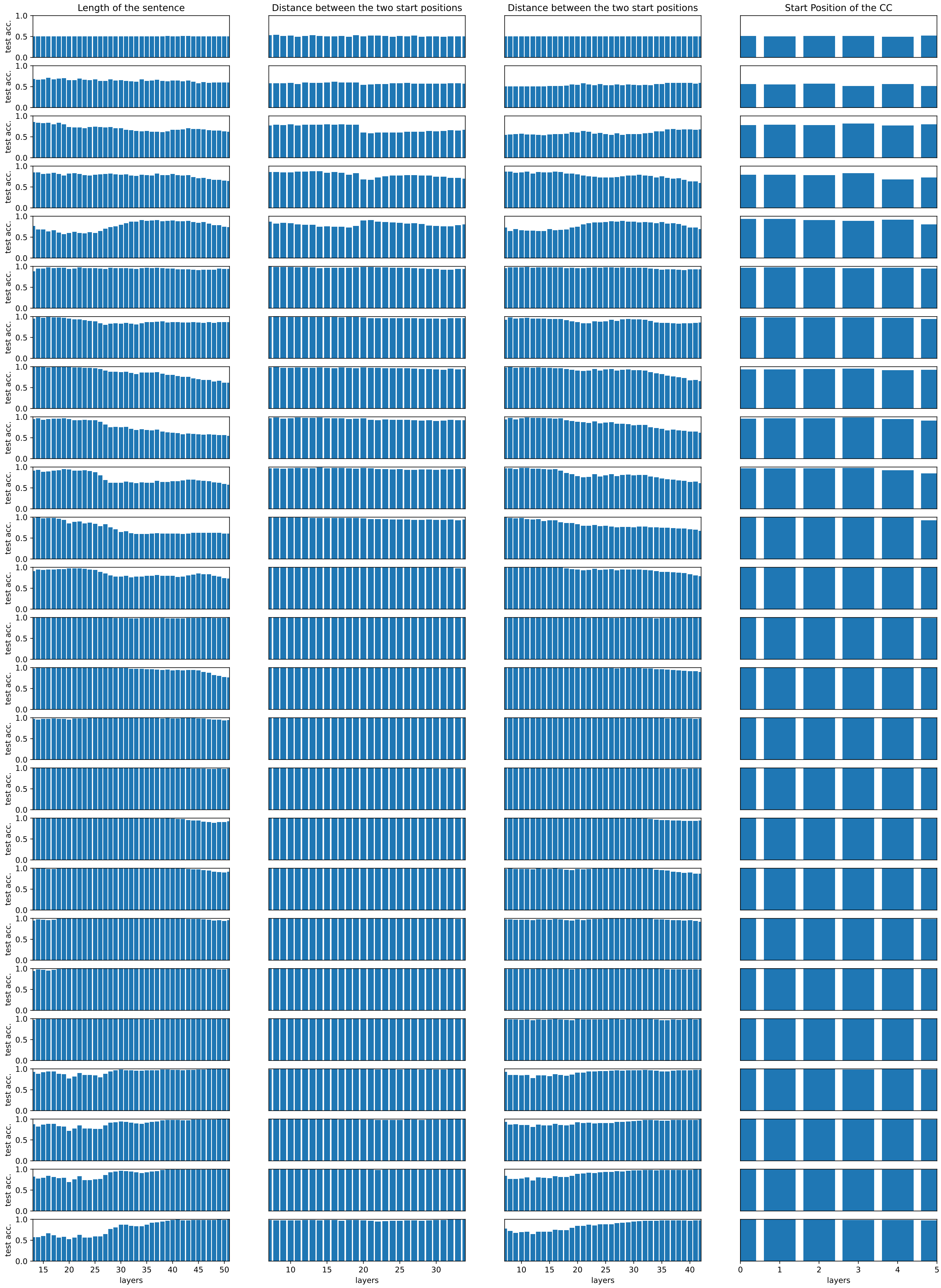

Figure 2: Full results for $\text{BERT}_{\text{LARGE}}$ on artificial data. Columns indicate the variable that the training and test set controls for.

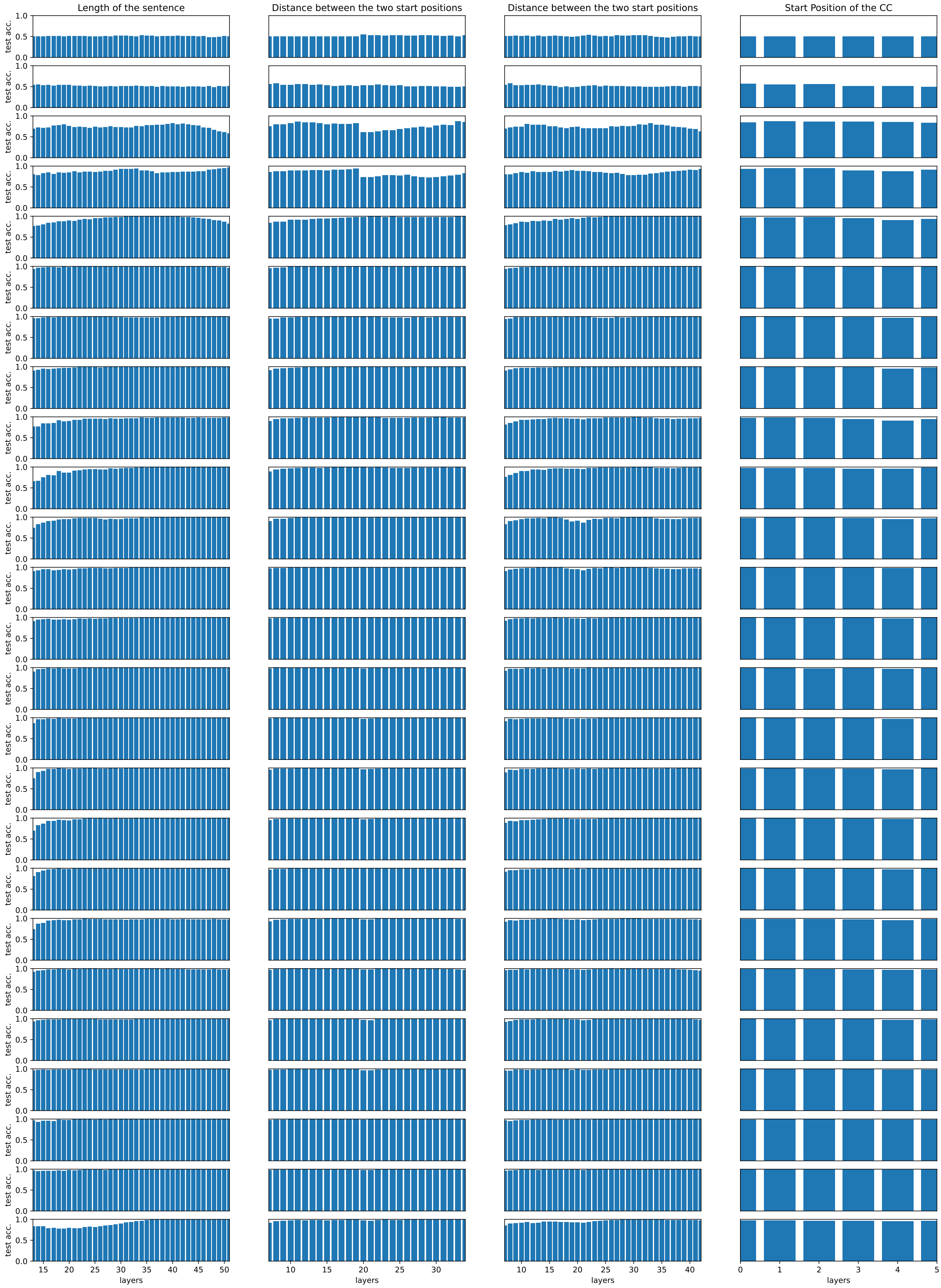

Figure 3: Full results for RoBERTa$_{\text{LARGE}}$ on artificial data. Columns indicate the variable that the training and test set controls for.

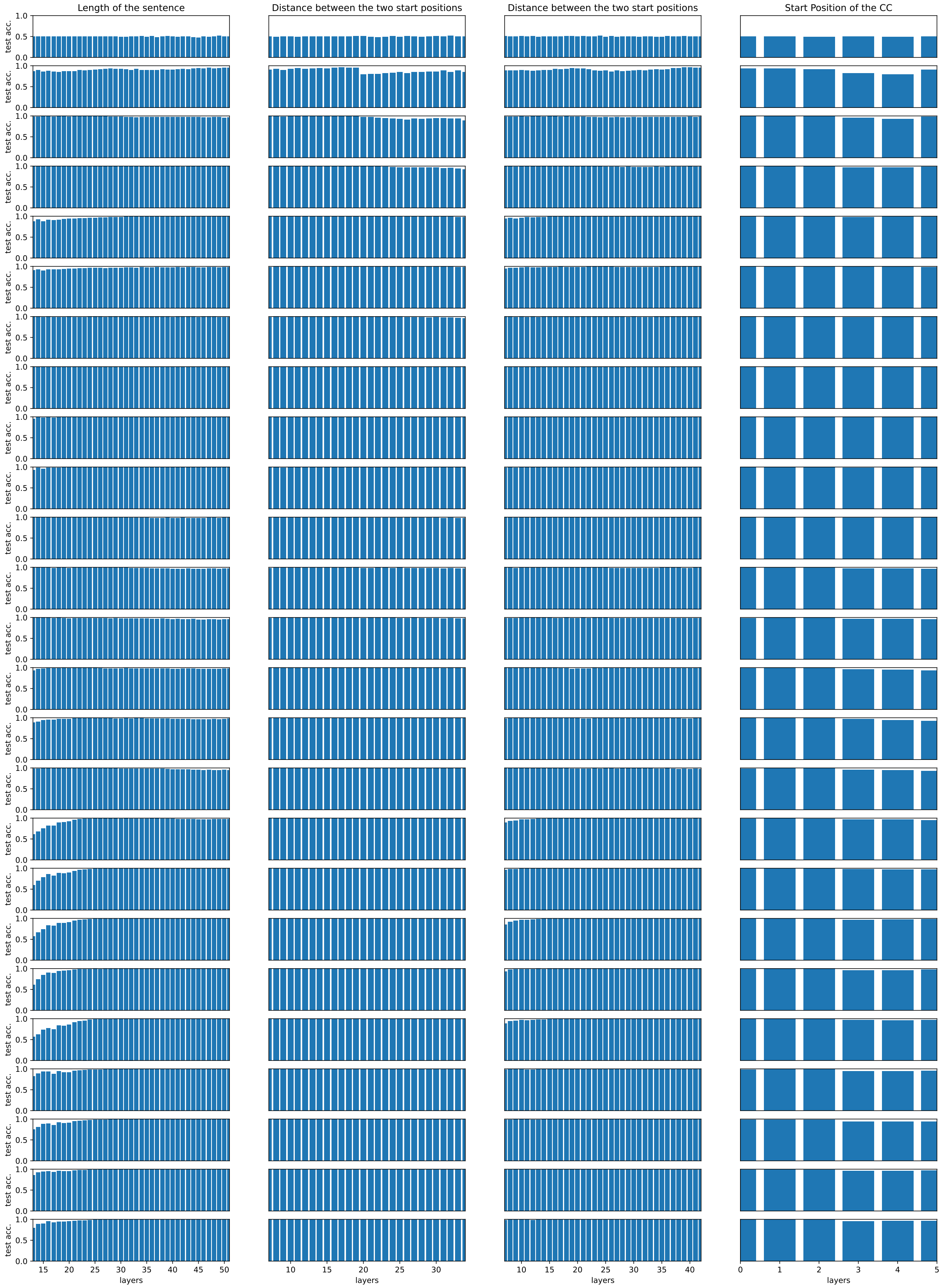

Figure 4: Full results for DeBERTa$_{\text{LARGE}}$ on artificial data. Columns indicate the variable that the training and test set controls for.

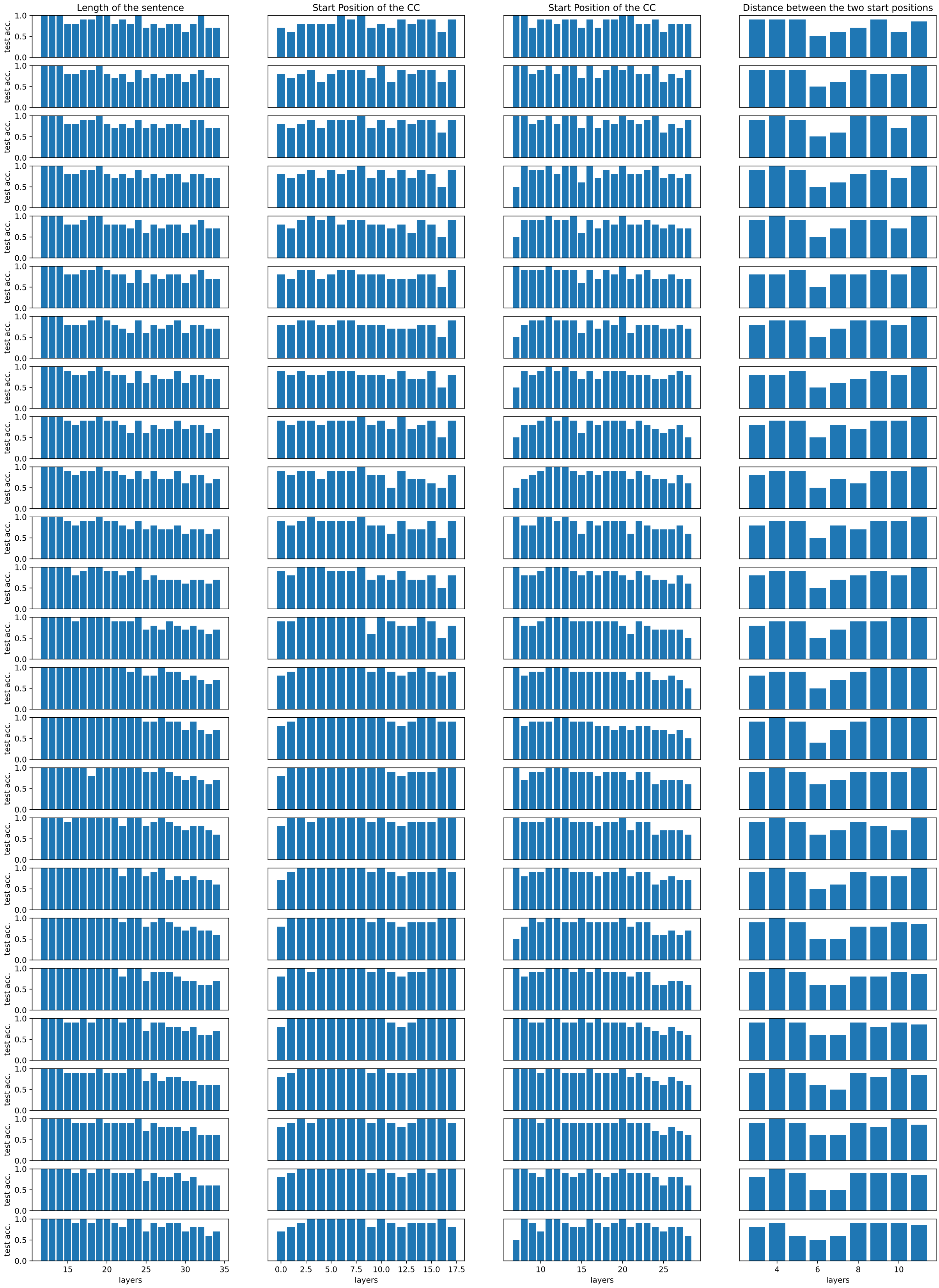

Figure 5: Full results for $\text{BERT}_{\text{LARGE}}$ on corpus data. Columns indicate the variable that the training and test set controls for.

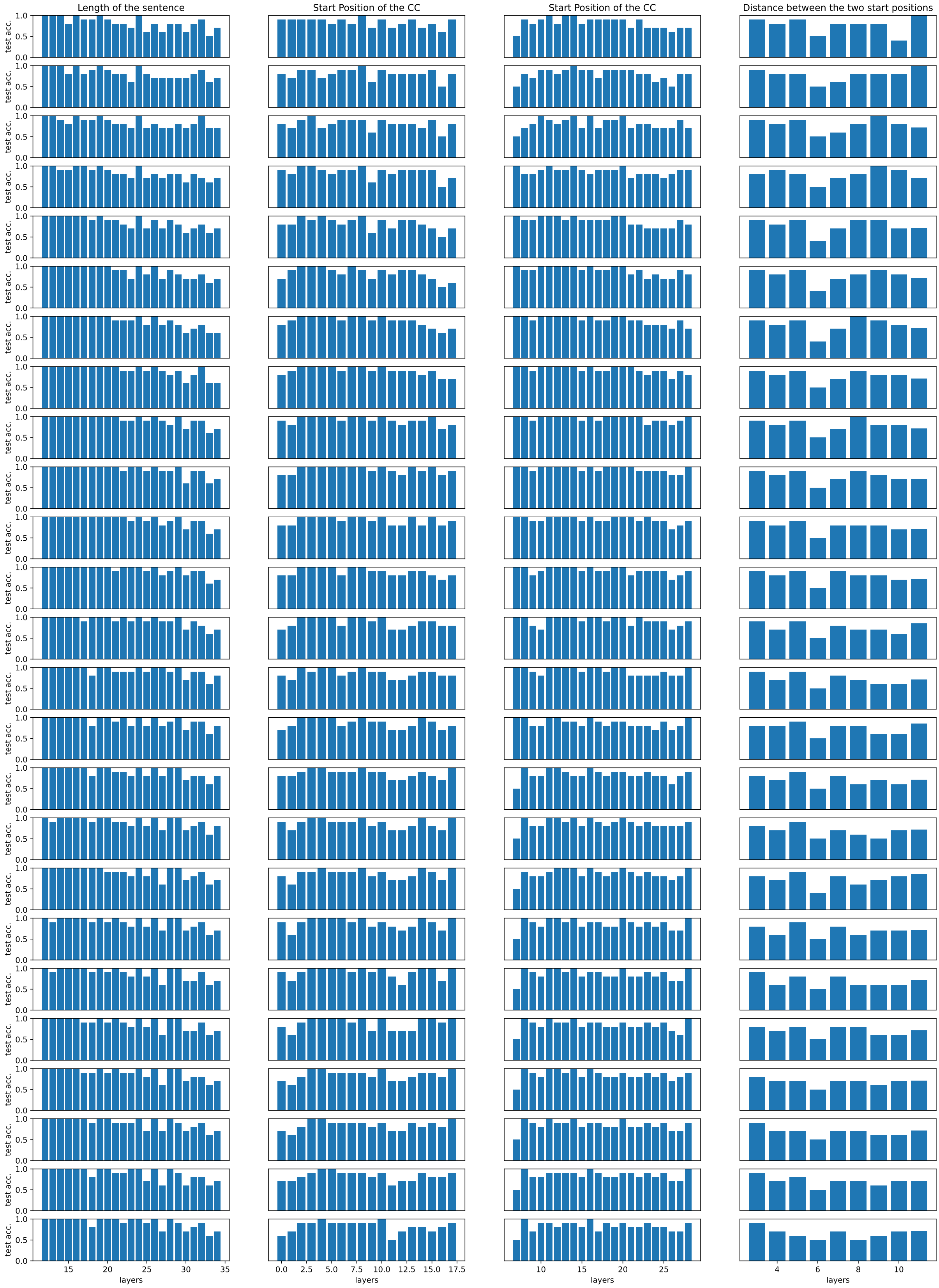

Figure 6: Full results for $\text{RoBERTa}_{\text{LARGE}}$ on corpus data. Columns indicate the variable that the training and test set controls for.

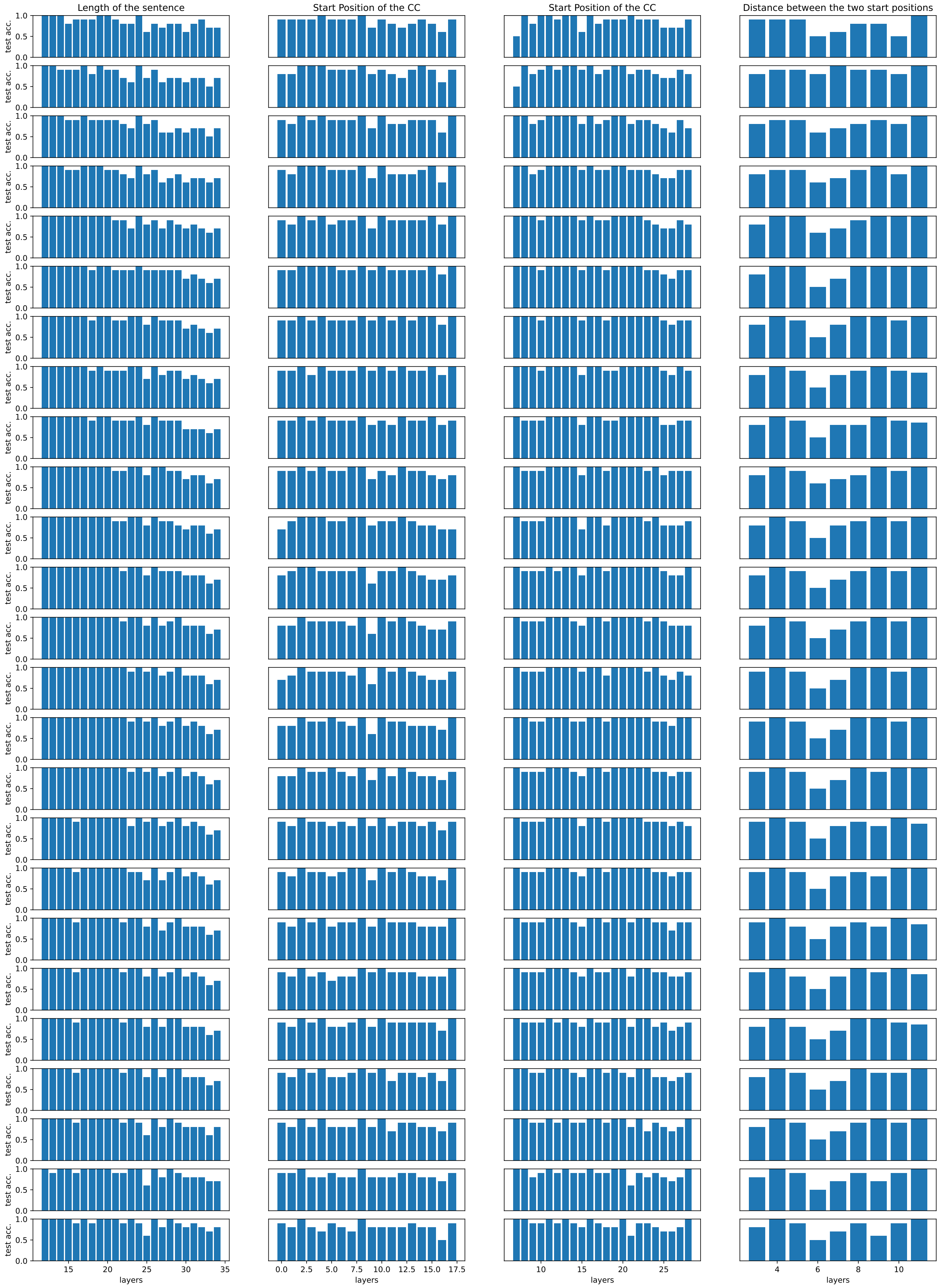

Figure 7: Full results for DeBERTa$_{\text{LARGE}}$ on corpus data. Columns indicate the variable that the training and test set controls for.

Model	Test Scenario	-	S5	S6	S7
BERT_{B}	S1	37.65%	37.62%	44.39%	47.9%
	S2	64.64%	62.79%	56.66%	55.41%
	S3	38.04%	44.78%	44.09%	48.29%
BERT_{L}	S1	36.85%	31.91%	47.21%	44.03%
	S2	67.13%	73.48%	54.39%	64.45%
	S3	36.46%	43.43%	47.79%	44.36%
$\text{RoBERTa}_{\text{B}}$	S1	61.6%	58.76%	42.13%	62.32%
	S2	52.85%	51.35%	71.33%	60.25%
	S3	62.21%	55.17%	43.04%	62.76%
$\text{RoBERTa}_{\text{L}}$	S1	55.72%	58.37%	65.08%	69.53%
	S2	68.01%	74.53%	62.73%	77.76%
	S3	55.36%	52.02%	65.28%	69.23%
$\text{DeBERTa}_{\text{B}}$	S1	49.72%	49.72%	49.86%	49.2%
	S2	49.81%	48.67%	49.7%	49.06%
	S3	50.28%	50.19%	49.97%	50.0%
$\text{DeBERTa}_{\text{L}}$	S1	50.88%	49.86%	50.03%	49.39%
	S2	51.41%	48.09%	47.21%	48.04%
	S3	50.58%	49.94%	50.41%	49.42%
$\text{DeBERTa}_{\text{XL}}$	S1	47.73%	45.08%	43.31%	43.67%
	S2	49.34%	46.27%	45.58%	41.74%
	S3	47.9%	49.14%	42.68%	45.58%
$\text{DeBERTa}_{\text{XXL}}$	S1	47.35%	53.56%	54.92%	54.12%
	S2	48.73%	52.85%	54.03%	53.81%
	S3	47.57%	49.36%	55.25%	53.59%

Table 4: Accuracies for the semantic probe with our three calibration methods compared to no calibration. We report the average accuracy on the more difficult sentences in terms of recency bias (S1),the easier ones (S2), and vocabulary bias (S3). Our calibration tecniques are short (S5), name (S6), and adjective (S7).

Model	Test Scenario	-	S5	S6	S7
$BERT_B$	S2	26.99%	25.22%	14.75%	10.77%
	S3	75.69%	23.51%	86.33%	91.05%
	S4	2.71%	-	-	-
$BERT_L$	S2	30.44%	41.8%	13.37%	22.24%
	S3	73.31%	25.94%	88.65%	85.97%
	S4	2.32%	-	-	-
$RoBERTa_B$	S2	9.92%	8.67%	31.13%	10.86%
	S3	76.19%	22.04%	79.03%	74.75%
	S4	2.76%	-	-	-
$RoBERTa_L$	S2	14.34%	17.82%	15.94%	15.86%
	S3	79.48%	43.54%	64.78%	57.27%
	S4	4.34%	-	-	-
$DeBERTa_B$	S2	0.91%	11.77%	7.13%	10.8%
	S3	99.67%	56.44%	96.52%	94.94%
	S4	1.08%	-	-	-
$DeBERTa_L$	S2	7.04%	7.85%	14.31%	14.28%
	S3	94.83%	43.18%	85.75%	79.86%
	S4	2.24%	-	-	-
$DeBERTa_{XL}$	S2	5.47%	7.87%	13.48%	18.78%
	S3	89.28%	45.44%	68.48%	65.94%
	S4	2.51%	-	-	-
$DeBERTa_{XXL}$	S2	3.59%	3.09%	17.02%	17.21%
	S3	82.1%	79.06%	63.43%	59.81%
	S4	1.13%	-	-	-

Table 5: Decision flip scores for the semantic probe with our three calibration methods compared to no calibration. We report the percentage of decisions flipped by changing from the base S1 to the sentences testing for recency bias (S2), vocabulary bias (S3), and name bias (S4). Our calibration tecniques are short (S5), name (S6), and adjective (S7).

Sentence	Label
Nowadays , the bigger the eighteen sheep date , the louder and bigger the twelve horses beat under the sun .	Positive
The flatter the fourteen lions push , the deeper and smaller the sixteen deer burn under the roof .	Positive
The deeper the sixteen cows beat ; the flatter and earlier the twenty cows attack .	Positive
Therefore , the worse the sixteen sheep believe after the morning without a pause , the smaller the thirteen cows box after the morning under the sun .	Positive
The flatter the fourteen lions push , the deeper and smaller the sixteen deer burn under the roof .	Positive
Sometimes , the worse and earlier seventeen believe the deer , and we just want to say that they mean that this has always been the case , the flatter twenty-one attack the foxes before the afternoon under the roof .	Negative
Nowadays , the smaller sixteen box the camels , and by the way , they mean that this is always true ; the weaker thirteen date the cows .	Negative
Therefore the earlier and weaker fourteen chase the deer , the stronger and earlier thirteen boil the chickens during the night .	Negative
The weaker and worse fifteen box the lions during the morning under the sun , the worse twenty push the cows .	Negative
It follows that the worse twelve date the pigs without a break the flatter and louder nineteen call the pigs under the sun .	Negative

Table 6: Examples of artificial training data

Sentence	Label
The harder and longer the three cats throw , the harder and shorter the ten dogs shake .	Positive
I have recently read in an established , well-known newspaper that the later the ten mice strike ; the later and better the seven men smash under the tree during the night .	Positive
The shorter the ten girls break without a pause ; the later the ten boys bleed under the tree .	Positive
It was recently announced that the better and later the five women break ; the quicker the six mice smash under the tree during the evening .	Positive
The faster the seven humans choke under the stairs after the evening , and I just want to say that I think that this is not always true , the lower and higher the two boys swallow .	Positive
The higher nine strike the women without a pause the shorter ten choke the girls .	Negative
We can say that the longer and faster four strike the men under the stairs before the evening , the harder four throw the dogs after the day under the bridge .	Negative
The quicker and higher eight bleed the people , and then I said that you believe that this also holds in other cases ; the longer seven break the girls after the night .	Negative
The shorter four smash the people before the night , and by the way , you think that this is always true ; the harder three bleed the people .	Negative
The longer seven shoot the women without stopping , the faster ten strike the mice after the night under the bridge .	Negative

Table 7: Examples of artificial test data

Sentence	**Label**
" The higher up the nicer ! "	Positive
She thinks the more water she drinks the better her skin looks .	Positive
It becomes an obsession lightly because the more fish you catch the higher your adrenaline flows .	Positive
It is worth noting , however , that the more specific you are the better .	Positive
In other words , the more videos you make the greater your audience reach .	Positive
Subtract the smaller from the larger . "	Negative
The way the older guys help out the younger guys is fantastic .	Negative
In this procedure the lower lip is pulled ventrally to expose the lower incisors .	Negative
The 5th bedroom is on the lower floor with easy access to the lower bath .	Negative
Note the distinctive bend of the larger vein adjacent to the smaller vein at the top .	Negative

Table 8: Examples of corpus data