Open Sesame: Getting Inside BERT's Linguistic Knowledge

Yongjie Lin[a,*] and **Yi Chern Tan**[a,*] and **Robert Frank**[b]
[a]Department of Computer Science, Yale University
[b]Department of Linguistics, Yale University
{yongjie.lin, yichern.tan, robert.frank}@yale.edu

Abstract

How and to what extent does BERT encode syntactically-sensitive hierarchical information or positionally-sensitive linear information? Recent work has shown that contextual representations like BERT perform well on tasks that require sensitivity to linguistic structure. We present here two studies which aim to provide a better understanding of the nature of BERT's representations. The first of these focuses on the identification of structurally-defined elements using diagnostic classifiers, while the second explores BERT's representation of subject-verb agreement and anaphor-antecedent dependencies through a quantitative assessment of self-attention vectors. In both cases, we find that BERT encodes positional information about word tokens well on its lower layers, but switches to a hierarchically-oriented encoding on higher layers. We conclude then that BERT's representations do indeed model linguistically relevant aspects of hierarchical structure, though they do not appear to show the sharp sensitivity to hierarchical structure that is found in human processing of reflexive anaphora.[1]

1 Introduction

Word embeddings have become an important cornerstone in any NLP pipeline. Although such embeddings traditionally involve context-free distributed representations of words (Mikolov et al., 2013; Pennington et al., 2014), recent successes with contextualized representations (Howard and Ruder, 2018; Peters et al., 2018; Radford et al., 2019) have led to a paradigm shift. One prominent architecture is BERT (Devlin et al., 2018), a Transformer-based model that learns bidirectional encoder representations for words, on the basis of a masked language model and sentence adjacency training objective. Simply using BERT's representations in place of traditional embeddings has resulted in state-of-the-art performance on a range of downstream tasks including summarization (Liu, 2019), question answering and textual entailment (Devlin et al., 2018). It is still, however, unclear why BERT representations perform well.

A flurry of recent work (Linzen et al., 2016; Gulordava et al., 2018; Marvin and Linzen, 2018; Lakretz et al., 2019) has explored how recurrent neural language models perform in cases that require sensitivity to hierarchical syntactic structure, and study how they do so, particularly in the domain of agreement. In these studies, a pre-trained language model is asked to predict the next word in a sentence (a verb in the target sentence) following a sequence that may include other intervening nouns with different grammatical features (e.g., "the **bear** by the trees *eats*..."). The predicted verb should agree with the subject noun (**bear**) and not the attractors (trees), in spite of the latter's recency. Such analyses have revealed that LSTMs exhibit state tracking and explicit notions of word order for modeling long term dependencies, although this effect is diluted when sequential and structural information in a sentence conflict. Further work by Gulordava et al. (2018) and others (Linzen and Leonard, 2018; Giulianelli et al., 2018) argues that RNNs acquire grammatical competence in agreement that is more abstract than word collocations, although language model performance that requires sensitivity to the phenomena such as reflexive anaphora, non-local agreement and negative polarity remains low (Marvin and Linzen, 2018). Meanwhile, studies evaluating *which* linguistic phenomena are encoded by contextualized representations (Goldberg, 2019; Wolf, 2019; Tenney et al., 2019) successfully demonstrate that purely self-attentive architectures like BERT can

[*] Equal contribution.

[1] The code is available at https://github.com/yongjie-lin/bert-opensesame.

Proceedings of the Second BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 241–253
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

capture hierarchy-sensitive, syntactic dependencies, and even support the extraction of dependency parses (Hewitt and Manning, 2019). However, the way in which BERT does this has been less studied. In this paper, we investigate how and where the representations produced by pre-trained BERT models (Devlin et al., 2018) express the hierarchical organization of a sentence.

We proceed in two ways. The first involves the use of diagnostic classifiers (Hupkes et al., 2018) to probe the presence of hierarchical and linear properties in the representations of words. However, unlike past work, we train these classifiers using a "poverty of the stimulus" paradigm, where the training data admit both linear and hierarchical solutions that can be distinguished by an enriched generalization set. This method allows us to identify what kinds of information are represented most robustly and transparently in the BERT embeddings. We find that as we use embeddings from higher layers, the prevalence of linear/sequential information decreases, while the availability of on hierarchical information increases, suggesting that with each layer, BERT phases out positional information in favor of hierarchical features of increasing complexity.

In the second set of experiments, we explore a novel approach to the study of BERT's self-attention vectors. Past explorations of attention mechanisms, whether in the domain of vision (Olah et al., 2018; Carter et al., 2019) or NLP (Bahdanau et al., 2015; Karpathy et al., 2015; Young et al., 2018; Voita et al., 2018), have largely involved a range of visualization techniques or the study of the general distribution of attention. Our work takes a quantitative approach to the study of attention and its encoding of syntactic dependencies. Specifically, we consider the relationships between verbs and the subjects with which they agree, and reflexive anaphors and their antecedents. Building on past work in psycholinguistics, we consider the influence of distractor noun phrases on the identification of these dependencies. We propose a simple attention-based metric called the *confusion score* that captures BERT's response to syntactic distortions in an input sentence. This score provides a novel *quantitative* method of evaluating BERT's syntactic knowledge as encoded in its attention vectors. We find that BERT does indeed leverage syntactic relationships between words to preferentially attend to the "correct" noun phrase for the purposes of agreement and anaphora, though syntactic structure does not show the strong categorical effects we sometimes find in natural language. This result again points to a representation of syntactically-relevant hierarchical information in BERT, this time through attention weightings.

Our analysis thus provides evidence that BERT's self-attention layers compose increasingly abstract representations of linguistic structure without explicit word order information, and that structural information is expressly favored over linear information. This explains why BERT can perform well on downstream NLP tasks, which typically require complex modeling of structural relationships.

2 Diagnostic Classification

For our first exploration of the kind of linguistic information captured in BERT's embeddings, we apply diagnostic classifiers to 3 tasks: identifying whether a given word is the sentence's **main auxiliary**, the sentence's **subject noun**, and the sentence's n^{th}**-token**. In each task, we assess how well BERT's embeddings encode information about a given linguistic property via the ability of a simple diagnostic classifier to correctly recover the presence of that property from the embeddings of a single word. The three tasks focus on different sorts of information: identifying the main auxiliary and the subject noun requires sensitivity to hierarchical or syntactic information, while the n^{th}-token requires linear information.

For each token in a given sentence, its input representation to BERT is a sum of its token, segment and positional embeddings (Devlin et al., 2018). We refer to these inputs as **pre-embeddings**. Note that by construction, a) the pre-embeddings contain linear but not hierarchical information, and b) BERT cannot generate new linear information that is not already in the input. Thus, any linear information in BERT's embeddings ultimately stems from the pre-embeddings, while any hierarchical information must be constructed by BERT itself.

2.1 Poverty of the stimulus

To classify an embedding as a sentence's main auxiliary or subject noun, the network needs to have represented structural information about a word's role in the sentence. In many cases, such structural information can be approximated lin-

	Main auxiliary task	**Subject noun task**
Training, Development	the cat will sleep the cat will eat the fish that can swim	the bee can sting the bee can sting the boy
Generalization	the cat that *can* meow will sleep the cat that *can* meow will eat the fish that can swim	(compound noun) the *queen* bee can sting (possessive) the *queen*'s bee can sting

Table 1: Representative sentences from the main auxiliary and subject noun tasks. For the latter, the generalization set contains two types of sentences, compound nouns and possessives, which are evaluated on separately. In each example, the correct token is underlined, while the distractor (consistent with the incorrect linear rule) is italicized.

early: the main auxiliary or subject noun could be identified as the first auxiliary or noun in a sentence. Though such a linear generalization may be falsified if given certain complex examples, it will succeed over a large range of simple sentences. Chomsky (1980) argues that the relevant distinguishing examples may be very rare for the case of identifying the main auxiliary (a property that is necessary in order to form questions), and hence this is an instance of the "poverty of the stimulus" that motivates the hypothesis of innate bias toward hierarchical generalizations. However, it seems clear that distinguishing examples are plentiful for the subject noun case. The question we are interested in, then, is whether and how BERT's embeddings, which result from training on a massive dataset, encode hierarchical information.

Pursuing the idea of poverty of the stimulus training (McCoy et al., 2018), we train diagnostic classifiers only on sentences in which the relevant property (main auxiliary or subject noun) is stateable in either hierarchical or sequential terms, i.e., the linearly first auxiliary or noun (cf. Section 2.2). The classifiers are then tested on sentences of greater complexity in which the hierarchical and linear generalizations can be distinguished. Since our classifier is a simple perceptron that can access only one embedding at a time, it cannot compute complex contingencies among the representations of multiple words, and cannot succeed unless such information is already encoded in the individual embeddings. Thus, success on these tasks would indicate that BERT robustly represents the words of a sentence using a feature space where the identification of hierarchical generalizations is easy.

2.2 Dataset

The main auxiliary and subject noun tasks use synthetic datasets generated from context-free grammars (cf. Appendix A.1) that were designed to isolate the relevant syntactic property for a poverty of the stimulus setup. Typical sentences are highlighted in Table 1. In both tasks, the training, development and generalization sets contained 40000, 10000, and 10000 examples respectively.

Main auxiliary In the training and development sets, the main auxiliary (will in Table 1) is always the first auxiliary in the sentence. A classifier that learns the naive linear rule of identifying the first linearly occurring auxiliary instead of the correct hierarchical (syntactic) rule still performs well during training. However, in the generalization set, the subject of each sentence is modified by a relative clause that contains an intervening auxiliary (that *can* meow). Since the main auxiliary is never the first auxiliary in this case, learning the hierarchical rule becomes imperative.

Subject noun In the training and development sets, the subject noun (bee in Table 1) is always the first noun in the sentence. A classifier that learns the linear rule of identifying the first linearly occurring noun does well during training, but only the hierarchical rule gives the right answer at test time. In the generalization set (both compound nouns & possessives cases), the subject noun is the head of the construction (bee) and not the dependent (*queen*). In the possessives case, we note that subword tokenization always produces *'s* as a standalone token, e.g. *queen's* is tokenized into *[queen] ['s]*. Also, we allow sentences to chain an arbitrary number of possessives via nesting.

n^{th}-token For this experiment, we use sentences from the Penn Treebank WSJ corpus. Following the setup of Collins (2002) and filtering for sentences between 10 to 30 tokens BERT tokenization, we obtained training, development and generalization sets of sentences of sizes 21142, 3017 and 2999. We only consider $2 \leq n \leq 9$. In particular, we ignore $n = 1$ since the first token produced by BERT is always trivially [CLS].

2.3 Methods

BERT models In our experiments, we consider two of Google AI's pre-trained BERT models *bert-base-uncased* (**bbu**) and *bert-large-uncased* (**blu**) from a PyTorch implementation.[2] bbu has 12 layers, 12 attention heads and embedding width 768, while blu has 24 layers, 16 attention heads and embedding width 1024.

Training For each task, we train a simple perceptron with a sigmoid output to perform binary classification on individual token embeddings of a sentence, based on whether the underlying token possesses the property relevant to the task. This is similar to the concept of diagnostic classification by Hupkes et al. (2018); Giulianelli et al. (2018).

$$\begin{array}{ccccc} & \text{the} & \text{cat} & \text{will} & \text{sleep} \\ \text{(BERT)} & \downarrow & \downarrow & \downarrow & \downarrow \\ & e_1 & e_2 & e_3 & e_4 \\ \text{(Classifier)} & \downarrow & \downarrow & \downarrow & \downarrow \\ & \widehat{y}_1 & \widehat{y}_2 & \widehat{y}_3 & \widehat{y}_4 \end{array}$$

Each input sentence is tokenized and processed by BERT, and the resulting embeddings $\{e_i\}$ are individually passed to the classifier f_θ to produce a sequence of logits $\{\widehat{y}_i\}$. Supervision is provided via a one-hot vector of indicators $\{y_i\}$ for the specified property. For example, in the main auxiliary task, the above example would have $y_1 = y_2 = y_4 = 0$ and $y_3 = 1$, since the third word is the main auxiliary. The contribution of each example to the total cross-entropy loss is:

$$\mathcal{L}_\theta = -\sum_i (y_i \log \widehat{y}_i + (1 - y_i) \log(1 - \widehat{y}_i)) \quad (1)$$

Each classifier is trained for a single epoch using the Adam optimizer (Kingma and Ba, 2014) with hyperparameters $lr = 0.001, \beta_1 = 0.9, \beta_2 = 0.999$. We freeze BERT's weights throughout training, which allows us to take good classification performance as evidence that the information relevant to the task is being encoded in BERT's embeddings in a salient, easily retrievable manner.

Evaluation For each example at test time, after computing the logits we obtain the index of the classifier's most confident guess within the sentence:

$$i^* = \arg\max_i \widehat{y}_i \quad (2)$$

The average y_{i^*} across the test set is reported as the classification accuracy.

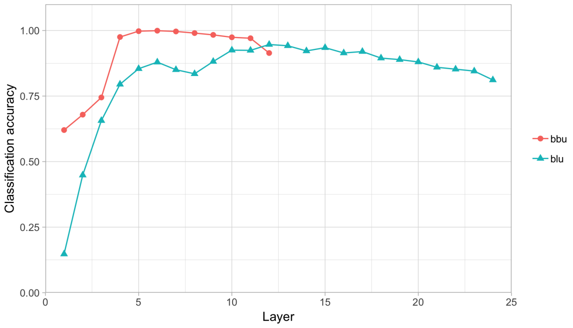

Figure 1: Layerwise accuracy of diagnostic classifiers on the generalization set of the main auxiliary task.

Layerwise diagnosis One key aspect of our experiments is the training of layer-specific classifiers for *all* layers. This yields a *layerwise* diagnosis of the information content in BERT's embeddings, providing a glimpse into how BERT internally manipulates and composes linguistic information. We also train classifiers on the pre-embeddings, which can be considered as the "zero-th" layer of BERT and hence act as useful baselines for content present in the input.

2.4 Results

Main auxiliary Classifiers for both models achieved near-perfect accuracy across all layers on the development set. In Figure 1, we observe that on the generalization set, the classifiers for both models can identify the main auxiliary with over 85% accuracy past layer 5, and bbu in particular obtains near-perfect accuracy from layers 4 to 11.

As discussed in Section 2.2, the classifiers were only given training examples where the main auxiliary was also the first auxiliary in the sentence. Although the linear rule "pick the first auxiliary" is compatible with the training data, the classifier nonetheless learns the more complex but correct hierarchical rule "pick the auxiliary of the main clause". By our argument from Section 2.1, this suggests that BERT embeddings encode syntactic information relevant to whether a token is the main auxiliary, as a feature salient enough to be recoverable by our simple diagnostic classifier.

We found that almost all instances of classification errors involved the misidentification of the linearly first auxiliary (within the relative clause) as the main auxiliary, e.g. *can* instead of will in Table 1. We believe that this stems from the significance of part-of-speech information for language modeling. As a result, any word of a different POS will not be chosen by the classifier.

[2] https://github.com/huggingface/pytorch-pretrained-BERT

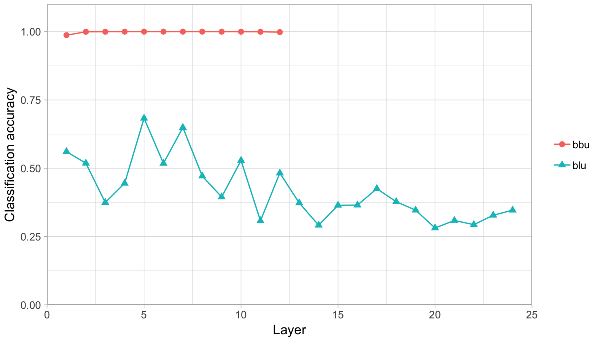

Figure 2: Layerwise accuracy of diagnostic classifiers on the compound noun generalization set of the subject noun task.

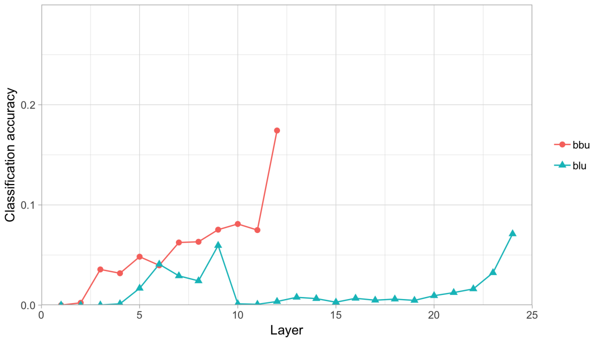

Figure 3: Layerwise accuracy of diagnostic classifiers on the possessive generalization set of the subject noun task.

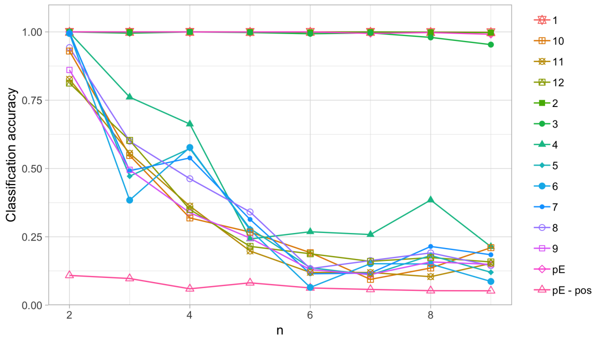

Figure 4: Layerwise accuracy of diagnostic classifiers on the generalization set of the n^{th} token task, for the bbu model only. Note that each line corresponds to a particular layer's embeddings as we vary $2 \leq n \leq 9$. pE denotes pre-embeddings and pE – pos denotes pre-embeddings without the positional component.

Subject noun As with the previous task, classifiers for both models achieved near-perfect accuracy across all layers on the development set. On the compound noun generalization set (Figure 2), while bbu achieved near-perfect accuracy in later layers, blu consistently performed poorly. bbu's performance suggests that in the classifier successfully learns a generalization that excludes the first noun of a compound, as opposed to the naive linear rule "pick the first noun". As before, this suggests that BERT encodes syntactic information in its embeddings. However, blu's performance is unexpected: it consistently predicts the object noun when it makes errors. In contrast, on the possessive generalization set (Figure 3), both models perform poorly. We offer an explanation for this distinctive performance in Section 2.5.

n^{th} token Since this property is entirely determined by the linear position of a word in a sentence, it directly measures the amount of positional information encoded in the embeddings. Here we have two baselines characterizing both extremes: the normal pre-embeddings (denoted pE) and a variant (pE – pos) where we exclude the positional component from its construction. Since BERT cannot introduce any new positional information, we expect these two to represent upper and lower bounds on the amount of positional information present in BERT's embeddings.

In Figure 4, we see a dramatic difference in performance on pE (one of the topmost lines) compared to pE – pos (bottommost line). We note that performances across all 12 layers fall between these two extremes, confirming our intuitions from earlier. Specifically, the classifiers for layers 1 – 3 have near-perfect accuracy on identifying an arbitrary n^{th} token ($2 \leq n \leq 9$). However, from layer 4 onwards, the accuracy drops sharply as n increases. This suggests that the positional component of pre-embeddings is the primary source of positional information in BERT, and BERT (bbu) discards a significant amount of positional information between layers 3 and 4, possibly in favor of hierarchical information.

2.5 Further Analysis

Main auxiliary In Figure 1, we observe that classification accuracy increases sharply in the first 4 layers, then plateaus before slowly decreasing. This mirrors a similar layerwise trend observed by Hewitt and Manning (2019). We postulate that the embeddings reach their "optimal level of abstraction" with respect to their ability to predict the main auxiliary halfway through BERT (about layer 6 for bbu, 12 for blu). At layer 1, the embedding for each token is a highly localized representation that contains insufficient sentential context to determine whether it is the main auxiliary of the sentence. As layer depth increases, BERT composes increasingly abstract representations via self-attention, which allows it to extract information from other tokens in the sentence. At some point, the representation becomes abstract enough to represent the hierarchical concept of a "main auxiliary", causing an early increase in classification accuracy. However, as depth increases further, the representations become so abstract that finer linguistic features are increasingly difficult to recover, e.g., a token embedding at the sentence-vector level of abstraction may longer be capable of identifying itself as the main auxiliary, accounting for the slowly deteriorating performance towards later layers.

Subject noun Given the similarity of the main auxiliary and the subject noun classification tasks, we might expect them to exhibit similar trends in performance. In Figure 2, we observe a similar early increase in diagnostic classification accuracy for the bbu embeddings. The lack of significant performance decay on higher layers possibly reflects the salience of the subject noun feature even at the sentence-vector level of abstraction. Strangely, blu performed poorly, even worse than chance (50%). We are unable to explain why this happens and leave this for future research.

On the possessive generalization set, the poor performance of both models seems to contradict the hypothesis that BERT has learned an abstract hierarchical generalization to classify subject nouns. We conjecture that BERT's issues in the possessive case stem from the ambiguity of the *'s* token, which can function either as a possessive marker or as a contracted auxiliary verb (e.g."She's sleeping"). If BERT takes a possessive occurrence of *'s* as the auxiliary verb, the immediately preceding noun can be (incorrectly) analyzed as the subject. If so, this would suggest that BERT does not represent the syntactic structure of the entire sentence in a unified fashion, but instead uses local cues to constituency. In Figure 3, the gradually increasing but still poor performance towards later layers in both models suggests that the embeddings might be trending toward a more abstract representation, but do not ultimately achieve it.

n^{th} **token** For each layer $k \geq 3$, Figure 4 shows an asymmetry where the classifier for layer k performs worse at identifying the n^{th} token as n increases. We believe that this may be an artifact of the distributional properties of natural language: the distribution of words that occur at the start of a sentence tends to be concentrated on a small class of parts of speech that can occur near the beginning of constituents that can begin a sentence. As n increases, the class of possible parts is no longer a function of the beginning of the sentence, and as a result becomes more uniform. As a result, it is easier for a classifier to predict whether a given word is the n^{th} token when n is small, since it can make use of easily accessible part-of-speech information in the embeddings to limit its options to only the tokens likely to occur in a given position.

3 Diagnostic Attention

Our second exploration of BERT's syntactic knowledge focuses on the encoding of grammatical relationships instead of the identification of elements with specific structural properties. We consider two phenomena: **reflexive anaphora** and **subject-verb agreement**. For each, we determine the extent to which BERT attends to linguistically relevant elements via the self-attention mechanism. This gives us further information about *how* hierarchy-sensitive syntactic information is encoded.

3.1 Quantifying intrusion effects via attention

Subject-verb and antecedent-anaphor dependencies both involve a dependent element, which we call the *target* (the verb or the anaphor) and the element on which it depends, which we call the *trigger* (the subject or the antecedent that provides the interpretation). A considerable body of work in psycholinguistics has explored how humans process such dependencies in the presence of elements that are not in relevant structural positions but which linearly intervene between the trigger and target. Dillon et al. (2013) aim to quantify this intrusion effect in human reading for the two dependencies we explore here. Under the assumption that higher reading time and eye movement regressions indicate an intrusion effect, they conclude that intruding noun-phrases have a substantial effect on the processing of subject-verb agreement, but not antecedent-anaphor relations.

We adapt this idea in measuring intrusion effects in BERT. We propose a simple and novel metric we term the "confusion score" for quantifying intrusion effects using attention. This quantita-

Subject-Verb Agreement

Condition	Relative Clause	DN Number Match	Example Sentence	Mean Confusion Score
A1	✗	✓	the cat near *the dog* does sleep	0.97
A2	✗	✗	the cat near *the dogs* does sleep	0.93
A3	✓	✓	the cat that can comfort *the dog* does sleep	0.85
A4	✓	✗	the cat that can comfort *the dogs* does sleep	0.81

Reflexive Anaphora

Condition	Relative Clause	DN_o Gender Match	DN_r Gender Match	Example Sentence	Mean Confusion Score
R1	✗	✓	NA	the lord could comfort *the wizard* by himself	1.01
R2	✗	✗	NA	the lord could comfort *the witch* by himself	0.92
R3	✓	NA	✓	the lord that can hurt *the prince* could comfort himself	0.99
R4	✓	NA	✗	the lord that can hurt *the princess* could comfort himself	0.89
R5	✓	✓	✓	the lord that can hurt *the prince* could comfort *the wizard* by himself	1.57
R6	✓	✓	✗	the lord that can hurt *the princess* could comfort *the wizard* by himself	1.52
R7	✓	✗	✓	the lord that can hurt *the prince* could comfort *the witch* by himself	1.49
R8	✓	✗	✗	the lord that can hurt *the princess* could comfort *the witch* by himself	1.39

Table 2: Representative sentences from the subject-verb agreement and reflexive anaphora datasets for each condition, and corresponding mean confusion scores. DN_o: distractor noun as object. DN_r: distractor noun in relative clause.

tive metric allows us to measure the preferable attention of transformer-based self-attention on one entity as opposed to another. Formally, suppose $X = \{x_i\}_{i=1}^n$ are linguistic units of interest, i.e. candidate triggers for the dependency, and Y is the dependency target. For each layer l and attention head a, we sum the self-attention weights from the indices of x_i (since each x_i may consist of multiple words) on attention head a of layer $l-1$ to Y on layer l, and take the mean over A attention heads:

$$\text{attn}_l(x_i, Y) = \frac{1}{A}\sum_{a=1}^{A}\sum_{x_{ij}\in x_i}\text{attn}_{la}(x_{ij}, Y) \quad (3)$$

We finally define the **confusion score** on layer l as the binary cross entropy of the normalized attention distribution between $\{x_i\}$ given Y as follows:

$$\text{conf}_l(X, Y) = -\log\frac{\text{attn}_l(x_1, Y)}{\sum_{i=1}^{n}\text{attn}_l(x_i, Y)} \quad (4)$$

Note that this equation assumes that each dependency has a unique trigger x_1: verbs agree with a single subject, and anaphors take a single noun phrase as their antecedent.

Our study focuses on the examples of the forms shown in Table 2. For subject-verb agreement, there are two types of examples: with the distractor within a PP (A1 and A2) and with the distractor within a RC (A3 and A4). Past psycholinguistic work has shown that distractor noun phrases within PPs give rise to greater processing difficulty than distractors within RCs (Bock and Cutting, 1992). For each type, we compare confusion in the case of distractors that share features with the subject, the true trigger of agreement, (A1 and A3) with those that do not (A2 and A4). Our expectation is that distractors that do not share features with the target of agreement will yield less confusion.

For reflexive anaphora, because of the possibility of ambiguity, we also consider sentences that include a noun phrase that is a structurally possible antecedent. For example, condition R1 has the subject *the lord* as its antecedent, but the object noun phrase *the wizard* is also grammatically possible. In contrast, for R2, the mismatch in gender features prevents the object from serving as an antecedent, which should lead to lower confusion. Sentences R3 and R4 include a distractor noun phrase within a RC. Since this noun phrase does not c-command the anaphor, it is grammatically inaccessible and should therefore contribute less, if at all, to confusion. Sentence types R5 through R8 include both the relative modifier and the object noun phrase, and systematically vary the agreement properties of the two distractors.

We hypothesize that attention weights on each linguistic unit indicate the relative importance of that entity as a trigger of a linguistic dependency. As a result, the ideal attention distribution should put all of the probability mass on the antecedent noun phrase for reflexive anaphora or on the subject noun phrase for agreement, and zero on the distractor noun phrases. As a baseline, a uniform distribution over two noun phrases, one the actual target and the other a distractor, would lead to a confusion score of $-\log \frac{1}{2} = 1$; with two distractors, the uniform probability baseline would be $-\log \frac{1}{3} = 1.6$.

3.2 Dataset

We construct synthetic datasets using context-free grammars (shown in Appendix A.1) for both subject-verb agreement and reflexive anaphora and compute mean confusion scores across multiple sentences. This allows us to control for semantic effects on the confusion score. All datasets for each condition contain 10000 examples.

In the subject-verb agreement datasets, we vary 1) the type of subordinate clause (prepositional phrase, PP; or relative clause, RC), and 2) the number on the distractor noun phrase. All conditions should be unambiguous, since only the head noun phrase can agree with the auxiliary.

In the reflexive anaphora datasets, we vary 1) the presence of a RC, 2) the gender match between the RC's noun phrase and the reflexive 3) the presence of an object noun phrase, and 4) the gender match between the object noun and the reflexive. All nouns are singular. Conditions R1, R5, R6 are ambiguous conditions, as they include an object noun phrase that matches the reflexive in gender. In other conditions, only the head noun phrase is the possible antecedent: the object mismatches in features and the noun phrase within the RC is grammatically inaccessible.

3.3 Methods

We use Equation 4 to compute the confusion score on each layer for the target in each sentence in our dataset. As in Section 2.3, this yields a *layerwise* diagnosis of confusion in BERT's self-attention mechanism. We also compute the mean confusion score across all layers.[3] In our experiments, we compute confusion scores using bbu only.

Note that in conditions R1, R5 and R6, there are two possible antecedents of the reflexive. We nonetheless use Equation 4 to calculate confusion scores relative to a single antecedent (the subject).

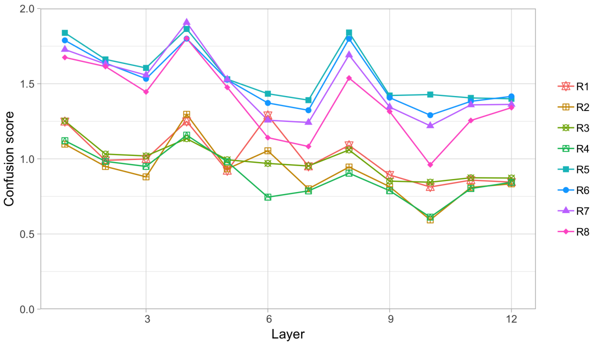

Figure 5: Layerwise confusion scores for each reflexive anaphora condition listed in Table 2. Conditions R1 to R4 have one distractor noun phrase, but conditions R5 to R8 have two distractor noun phrases.

Coefficient	Estimate	p-value
Subject-Verb Agreement		
Intercept	$1.33 \pm 1.32\mathrm{e}{-3}$	$< 2\mathrm{e}{-16}$
Relative Clause	$-0.12 \pm 1.03\mathrm{e}{-3}$	$< 2\mathrm{e}{-16}$
$\mathrm{DN_r}$ Number Match	$0.03 \pm 1.03\mathrm{e}{-3}$	$< 2\mathrm{e}{-16}$
Layer	$-0.06 \pm 1.50\mathrm{e}{-4}$	$< 2\mathrm{e}{-16}$
Reflexive Anaphora		
Intercept	$0.63 \pm 1.24\mathrm{e}{-3}$	$< 2\mathrm{e}{-16}$
$\mathrm{DN_o}$ Gender Match	$0.60 \pm 9.09\mathrm{e}{-4}$	$< 2\mathrm{e}{-16}$
$\mathrm{DN_o}$ Gender Mismatch	$0.50 \pm 9.09\mathrm{e}{-4}$	$< 2\mathrm{e}{-16}$
$\mathrm{DN_r}$ Gender Match	$0.57 \pm 9.09\mathrm{e}{-4}$	$< 2\mathrm{e}{-16}$
$\mathrm{DN_r}$ Gender Mismatch	$0.49 \pm 9.09\mathrm{e}{-4}$	$< 2\mathrm{e}{-16}$
Layer	$-0.03 \pm 9.72\mathrm{e}{-5}$	$< 2\mathrm{e}{-16}$

Table 3: Regression estimates and p-values for the coefficient effects under reflexive anaphora and subject-verb agreement. All effects are statistically significant.

To compute the significance of the presence of different types of distractors and of feature mismatch of the distractors, we run a linear regression to predict confusion score. For subject-verb agreement, the baseline value is the confusion at layer 1 of a sentence with a PP and a mismatch in number on the distractor noun (condition A2 in Table 2). For reflexive anaphora, the baseline is the confusion at layer 1 of a sentence with no RC and no object noun (e.g. "the lord comforts himself").

3.4 Results

Subject-verb agreement Since sentence types A1 to A4 are all unambiguous, ideal confusion scores should be zero. However, Table 2 indicates that the mean confusion scores are instead closer to the uniform probability baseline confusion score of 1, suggesting that BERT's self-attention mechanism is far from able to perfectly model syntactically-sensitive hierarchical information. Nonetheless, from Table 3, we see that BERT's attention mechanism is in fact sensitive

[3] We built on Vig (2019)'s BERT attention visualization library `https://github.com/jessevig/bertviz` to implement the attention-based confusion score.

to subtleties of linguistic structure: a distractor within a PP causes more confusion than one within a relative clause (i.e., the presence of the relative has a negative coefficient in the linear model), in agreement with past psycholinguistic work (Bock and Cutting, 1992). Moreover, the presence of matching distractors has a significant positive effect on confusion scores. These findings therefore suggest that BERT representations are sensitive to different types of syntactic embedding as well as the values of number features in computing subject-verb agreement dependencies.

Reflexive anaphora From Table 2, we see the major effect of the number of distractor noun phrases: mean confusion scores for conditions with one distractor (R1-R4) are lower than those with two distractors (R5-R8). If BERT were perfectly exploiting grammatical structure, we should expect the presence of a grammatically inaccessible distractor noun within a relative clause not to add to confusion. Thus, we might expect R5 and R6 to have mean confusion scores comparable to R1, as both include single grammatically viable distractor. However, they both have higher mean confusion scores than R1 (the same is true for R7/R8 vs. R2). Moreover, conditions R2 to R4 and R7 to R8 should have confusion scores of zero, since the head noun phrase is the only grammatically possible antecedent. This, however, is not so. Taken together, we might conclude that BERT attends unnecessarily to grammatically inaccessible or grammatically mismatched distractor noun phrases, suggesting that it does not accurately model reflexive dependencies.

Nonetheless, if we look more closely at the effects of the different factors through the linear model reported in Table 3, we once again find evidence for a sensitivity to both syntactic structure and grammatical features: the presence of grammatically accessible distractors has a (slightly) larger effect on confusion than grammatically inaccessible distractors (i.e., $\mathrm{DN_o}$ vs. $\mathrm{DN_r}$), particularly when the distractor matches in features with the actual antecedent.

3.5 Further Analysis

Layerwise diagnosis Figure 5 and Table 3 show that confusion is negatively correlated with layer depth for reflexive anaphora. Confusion scores for subject-verb agreement exhibit a similar trend. This provides additional evidence for our conjecture that BERT composes increasingly abstract representations containing hierarchical information, with an optimal level of abstraction. Notably, the observed sensitivity of BERT's self-attention values to grammatical distortions suggests that BERT's syntactic knowledge is in fact encoded in its attention matrices. Finally, it is worth noting that confusion for both reflexives and subject-verb agreement showed an increase at layer 4. Strikingly, this was the level at which linear information was found, through diagnostic classifiers, to be degraded. We leave for the future an understanding of the connection between these.

4 Conclusion

In this paper, we investigated how and to what extent BERT representations encode syntactically-sensitive hierarchical information, as opposed to linear information. Through diagnostic classification, we find that positional information is encoded in BERT from the pre-embedding level up through lower layers of the model. At higher layers, information becomes less positional and more hierarchical, and BERT encodes increasingly complex representations of sentence units.

We propose a simple and novel method of observing, for a given syntactic phenomenon, the intrusion effects of distractors on BERT's self-attention mechanism. Through such diagnostic attention, we find that BERT does encode aspects of syntactic structure that are relevant for subject-verb agreement and reflexive dependencies through attention weights, and that this information is represented more accurately on higher layers. We also find evidence that BERT is responsive to matching of grammatical features such as gender and number. However, BERT's attention is only incompletely modulated by structural and featural properties, and attention is sometimes spread across grammatically irrelevant elements.

We conclude that BERT composes increasingly abstract hierarchical representations of linguistic structure using its self-attention mechanism. To further understand BERT's syntactic knowledge, further work can be done to (1) investigate or visualize layer-on-layer changes in BERT's structural and positional information, particularly between layers 3 and 4 when positional information is largely phased out, and (2) retrieve the increasingly hierarchical representations of BERT across layers via the self-attention mechanism.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference for Learning Representations*.

Kathryn Bock and J. Cooper Cutting. 1992. Regulating mental energy: Performance units in language production. *Journal of Memory and Language*, 31(1):99–127.

Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. 2019. Activation atlas. `https://distill.pub/2019/activation-atlas` [Accessed: 19 Apr 2019].

Noam Chomsky. 1980. Rules and representations. *Behavioral and Brain Sciences*, 3(1):1–15.

Michael Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 1–8. Association for Computational Linguistics.

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

Brian Dillon, Alan Mishler, Shayne Sloggett, and Colin Phillips. 2013. Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69(2):85–103.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *Computing Research Repository*, abs/1901.05287.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

John Hewitt and Christopher Manning. 2019. A structural probe for finding syntax in word representations. In *International Conference for Learning Representations*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and diagnostic classifiers reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907926.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *Computing Research Repository*, abs/1506.02078.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *Computing Research Repository*, abs/1412.6980.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. *Computing Research Repository*, abs/1903.07435.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Tal Linzen and Brian Leonard. 2018. Distinct patterns of syntactic agreement errors in recurrent networks and humans. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 690–695.

Yang Liu. 2019. Fine-tune BERT for extractive summarization. In *Computing Research Repository*, volume abs/1903.10318.

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.

R Thomas McCoy, Robert Frank, and Tal Linzen. 2018. Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 2093–2098.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The building blocks of interpretability. https://distill.pub/2018/building-blocks [Accessed: 19 Apr 2019].

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *Computing Research Repository*, abs/1802.05365.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf [Accessed: 19 Apr 2019].

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Jesse Vig. 2019. Visualizing attention in transformer-based language models. *Computing Research Repository*, abs/1904.02679.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Thomas Wolf. 2019. Some additional experiments extending the tech report “Assessing BERT’s Syntactic Abilities” by Yoav Goldberg. https://huggingface.co/bert-syntax/extending-bert-syntax.pdf [Accessed: 19 Apr 2019].

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75.

A Appendix

A.1 Context-free grammars for dataset generation

S	→	NP$_{\text{M}}$ VP$_{\text{M}}$
NP$_{\text{M}}$	→	Det N | Det N Prep Det Nom | Det N RC
NP$_{\text{O}}$	→	Det Nom | Det Nom Prep Det Nom | Det Nom RC
VP$_{\text{M}}$	→	Aux VI | Aux VT NP$_{\text{O}}$
RC	→	Rel Aux VI | Rel Det Nom Aux VT | Rel Aux VT Det Nom
Nom	→	N | JJ Nom
Det	→	the | some | my | your | our | her
N	→	bird | bee | ant | duck | lion | dog | tiger | worm | horse | cat | fish | bear | wolf | birds | bees | ants | ducks | lions | dogs | tigers | worms | horses | cats | fish | bears | wolves
VI	→	cry | smile | sleep | swim | wait | move | change | read | eat
VT	→	dress | kick | hit | hurt | clean | love | accept | remember | comfort
Aux	→	can | will | would | could
Prep	→	around | near | with | upon | by | behind | above | below
Rel	→	who | that
JJ	→	small | little | big | hot | cold | good | bad | new | old | young

Figure 6: Context-free grammar for the main auxiliary dataset.

S	→	NP$_{\text{M}}$ VP
NP$_{\text{M}}$	→	Det MNom | Det MNom Prep Det Nom | Det MNom RC
NP$_{\text{O}}$	→	Det Nom | Det Nom Prep Det Nom | Det Nom RC
VP	→	Aux VI | Aux VT NP$_{\text{O}}$
RC	→	Rel Aux VI | Rel Det Nom Aux VT | Rel Aux VT Det Nom
Nom	→	N | JJ Nom
MNom	→	MNom1 | MNom2
MNom1	→	N | JJ MNom1
MNom2	→	N | JJ MNom2 | NS Poss MNom2 | Nadj+MN
Det	→	the | some | my | your | our | her
Poss	→	’s
NS	→	bird | bee | ant | duck | lion | dog | tiger | worm | horse | cat | fish | bear | wolf
N	→	bird | bee | ant | duck | lion | dog | tiger | worm | horse | cat | fish | bear | wolf | birds | bees | ants | ducks | lions | dogs | tigers | worms | horses | cats | fish | bears | wolves
Nadj+MN	→	worker bee | worker ant | race horse | queen bee | german dog | house cat
VI	→	cry | smile | sleep | swim | wait | move | change | read | eat
VT	→	dress | kick | hit | hurt | clean | love | accept | remember | comfort
Aux	→	can | will | would | could
Prep	→	around | near | with | upon | by | behind | above | below
Rel	→	who | that
JJ	→	small | little | big | hot | cold | good | bad | new | old | young

Figure 7: Context-free grammar for the subject noun dataset.

S	$\rightarrow$	NP$_{sg\_Agr}$ Aux$_{sg}$ VI | NP$_{pl\_Agr}$ Aux$_{pl}$ VI
NP$_{sg\_Agr}$	$\rightarrow$	Det N$_{sg}$ | Det N$_{sg}$ Prep Det N | Det N$_{sg}$ Prep RC$_{sg}$
NP$_{pl\_Agr}$	$\rightarrow$	Det N$_{pl}$ | Det N$_{pl}$ Prep Det N | Det N$_{pl}$ Prep RC$_{pl}$
RC$_{sg}$	$\rightarrow$	Rel Aux$_{sg}$ VI | Rel Aux$_{sg}$ VT Det N | Rel Det N$_{sg}$ Aux$_{sg}$ VT | Rel Det N$_{pl}$ Aux$_{pl}$ VT
N	$\rightarrow$	N$_{sg}$ | N$_{pl}$
RC$_{pl}$	$\rightarrow$	Rel Aux$_{pl}$ VI | Rel Aux$_{pl}$ VT Det N | Rel Det N$_{sg}$ Aux$_{sg}$ VT | Rel Det N$_{pl}$ Aux$_{pl}$ VT
Aux$_{sg}$	$\rightarrow$	does | Modal
Aux$_{pl}$	$\rightarrow$	do | Modal
Det	$\rightarrow$	the | some | my | your | our | her
N$_{sg}$	$\rightarrow$	bird | bee | ant | duck | lion | dog | tiger | worm | horse | cat | fish | bear | wolf
N$_{pl}$	$\rightarrow$	birds | bees | ants | ducks | lions | dogs | tigers | worms | horses | cats | fish | bears | wolves
VI	$\rightarrow$	cry | smile | sleep | swim | wait | move | change | read | eat
VT	$\rightarrow$	dress | kick | hit | hurt | clean | love | accept | remember | comfort
VS	$\rightarrow$	think | say | hope | know
VD	$\rightarrow$	tell | convince | persuade | inform
Modal	$\rightarrow$	can | will | would | could
Prep	$\rightarrow$	around | near | with | upon | by | behind | above | below
Rel	$\rightarrow$	who | that

Figure 8: Context-free grammar for the subject-verb agreement dataset.

S	$\rightarrow$	NP$_{M\_Ant}$ Aux VT Refl$_M$ | NP$_{F\_Ant}$ Aux VT Refl$_F$ | NP$_{M\_Ant}$ Aux VT Det N$_F$ by Refl$_M$ | NP$_{F\_Ant}$ Aux VT Det N$_M$ by Refl$_F$ | NP$_{M\_Ant}$ Aux VT Det N$_M$ by Refl$_M$ | NP$_{F\_Ant}$ Aux VT Det N$_F$ by Refl$_F$
NP$_{M\_Ant}$	$\rightarrow$	Det N$_M$ | Det N$_M$ RC
NP$_{F\_Ant}$	$\rightarrow$	Det N$_F$ | Det N$_F$ RC
N	$\rightarrow$	N$_M$ | N$_F$
RC	$\rightarrow$	Rel Aux VI | Rel Det N Aux VT | Rel Aux VT Det N
Refl$_M$	$\rightarrow$	himself
Refl$_F$	$\rightarrow$	herself
Det	$\rightarrow$	the | some | my | your | our | her
N$_F$	$\rightarrow$	girl | woman | queen | actress | sister | wife | mother | princess | aunt | lady | witch | niece | nun
N$_M$	$\rightarrow$	boy | man | king | actor | brother | husband | father | prince | uncle | lord | wizard | nephew | monk
VI	$\rightarrow$	cry | smile | sleep | swim | wait | move | change | read | eat
VT	$\rightarrow$	dress | kick | hit | hurt | clean | love | accept | remember | comfort
VS	$\rightarrow$	think | say | hope | know
VD	$\rightarrow$	tell | convince | persuade | inform
Aux	$\rightarrow$	can | will | would | could
Prep	$\rightarrow$	around | near | with | upon | by | behind | above | below
Rel	$\rightarrow$	who | that

Figure 9: Context-free grammar for the reflexive anaphora dataset.