Probing for Understanding of English Verb Classes and Alternations in Large Language Models

James V. Bruno, Jiayu Han, David K. Yi, Peter Zukerman *
The University of Washington
{jbruno, jyhan126, davidyi6, pzuk}@uw.edu

Abstract

We investigate the extent to which verb "alternation classes", as described by Levin (1993), are encoded in BERT embeddings using selectively constructed diagnostic classifiers for word and sentence-level prediction tasks. We follow and expand upon the experiments of Kann et al. (2019), which similarly probes whether global embeddings encode frame-selectional properties of verbs. At both the word and sentence level, we find that contextual embeddings produced by BERT not only outperform non-contextual embeddings, but achieve astonishingly high accuracies across various alternation classes. Additionally, we find evidence that different BERT embedding layers achieve varying performance on the task, with the middle and upper layers of BERT performing best overall. Finally, we offer an expanded alternation-class dataset, which augments the original data of Kann et al. 2019 with additional sentence examples.

1 Introduction

We investigate the extent to which verb alternation classes are represented in word and sentence embeddings produced by BERT (Devlin et al., 2018). As first comprehensively cataloged by Levin (1993), verbs pattern together into classes according to the syntactic alternations in which they can and cannot participate. For example, (1) illustrates the *causative-inchoative* alternation. *Break* can be a transitive verb in which the subject of the sentence is the agent and the direct object is the theme, as in example (1a). It can also alternate with the form in (1b), in which the subject of the sentence is the theme and the agent is unexpressed. However, (2) demonstrates that *cut* cannot participate in the same alternation, despite its semantic similarity.

(1) a. Janet broke the cup.
b. The cup broke.

(2) a. Margaret cut the bread.
b. * The bread cut.

(3) demonstrates an alternation of a different class – namely, the *spray-load* class, in which the theme and locative arguments can be syntactically realized as either direct objects or objects of the preposition. *Spray* can participate in the alternation, but as shown in (4), *pour* cannot.

(3) a. Jack sprayed paint on the wall.
b. Jack sprayed the wall with paint.

(4) a. Tamara poured water into the bowl.
b. * Tamara poured the bowl with water.

The alternations in which a verb may participate is taken to be a lexical property of the verb (e.g. Pinker, 1989; Levin, 1993; Levin et al., 1995; Schafer, 2009). Moreover, the alternations should be observable in large corpora of texts, and are therefore available as training data during the Masked-Language-Modeling task used to train neural language models such as BERT. Negative examples such as (2b) and (4b) should be virtually absent from the training data. This leads us to hypothesize that BERT representations may encode whether particular verbs are allowed to participate in syntactic alternations of various classes. Our research questions are as follows:

1. Do BERT's token-level representations encode information about which syntactic frames an individual verb can participate in?

2. At the sentence level, do BERT's layers represent the frame and alternation properties of their main verb?

Through carefully designed experiments, we find that BERT embeddings indeed encode information about verb alternation classes at both the word

*The authors' names appear in alphabetical order.

and sentence level. Furthermore, we find evidence suggesting that different embedding layers encode more information about different verb alternations, with the middle and top layers outperforming the bottom layers on average.

The rest of the paper is organized as follows: after a brief review of related literature in Section 2, we present datasets that are relevant to our experiment in Section 3 including a supplementary dataset that we generated. We then present two experiments to answer our research questions in Sections 4 and 6. Section 5 presents an additional *control task* (Hewitt and Liang, 2019) to test whether our findings are significant. Each of these sections have independent Method and Results sections. Finally, we offer a discussion in Section 7 and overall conclusions in Section 8.

2 Related work

Our work follows Kann et al. (2019), who attempt to predict verb-classes on the basis of GloVe embeddings (Pennington et al., 2014b) and embeddings derived from the 100M-token British National Corpus with the intentionally simple single-directional LSTM architecture from Warstadt et al. 2019. They then use these same embeddings to predict sentence grammaticality through a sentence encoder trained on a "real/fake" sentence classification task. Because their primary research focus has to do with how neural language models can inform learnability (in the sense of human language acquisition), they use language models derived from "an amount of data similar to what humans are exposed to during language acquisition" and intentionally avoid models trained on "several orders of magnitude more data than humans see in a lifetime" (p. 291). They also use a multi-layer perceptron with a hidden layer to predict alternation classes.

As described in Section 4, we depart from Kann et al. 2019 and build on it by examining the embedding representations of BERT which are derived from a training corpus of 3.3 billion words. We then use an intentionally simple and selective linear diagnostic classifier to probe the representations, as our research questions focuses on the BERT embeddings themselves. We note that Kann et al. (2019) achieved only modest performance in raw prediction accuracy, and only for a limited number of verb classes. While this was a valuable result for their research goals, our hypothesis is that we may achieve higher prediction accuracy due to BERT's more complex architecture and the larger size of its training data.

To our knowledge, attempting to predict verb alternation class membership along the lines of Levin 1993 from BERT representations is novel. However, two very closely related lines of work include the experiments of Warstadt and Bowman (2019), which respectively evaluate the performance of various pretrained language models on the CoLA (Warstadt et al., 2019) and BliMP Warstadt et al. (2020) benchmarks, which include examples from a wide variety of linguistic phenomena (including verb argument structures). We distinguish our experiments from these papers in two major ways. First, we attempt to directly probe the linguistic knowledge of individual BERT embedding layers with a classification probe instead of additionally finetuning BERT to a specific task. Second, we limit our focus to verb alternation classes and present detailed analysis about patterns and trends across different alternations and their corresponding syntactic frames.

3 Data

In our experiments, we use three datasets. First, we leverage the two datasets of Kann et al. (2019). One is the **L**exic**a**l **V**erb-frame **A**lternations dataset (LaVA), which is based on the verbs and alternation classes defined in Levin (1993). It contains a mapping of 516 verbs to 5 alternation classes, which are further subdivided into two syntactic frames for each alternation. The broad categories of the alternation classes are: *Spray-Load*, *Causative-Inchoative*, *Dative*, *There-insertion*, and *Understood-object*. Table 1 [1] provides a statistical description of the syntactic frames and their class distributions. **F**rames and **A**lternations of **V**erbs **A**cceptability (FAVA), the other dataset, is a corpus of $9,413$ semi-automatically generated sentences formed from the verbs in LaVA, along with human grammaticality judgments. The sentences in FAVA are split according to the alternation class of their main verb, and are separated into train ($n = 4838$), development ($n = 968$), and test ($n = 3608$) sets by the authors. Finally, we extend the work of Kann et al. (2019) to generate FAVA-ex, an augmented

[1] A similar table appears in Kann et al. (2019), but we present it again here because of discrepancies that we found in the distribution counts. Most significantly, it appears that the authors flipped the positive and negative counts for the *there-Insertion* and *Understood-Object* alternation classes which carries over to their results.

version of FAVA.

3.1 Data Generation

We provide an extended dataset based on FAVA, FAVA-ex, which widens the scope of verbs and uses 1591 verbs from the 5 broad alternation categories in Levin (1993). The corpus consists of 2,982 semi-automatically generated sentences along with human grammaticality judgments. We annotate the sentences ourselves, and report a Fleiss Kappa for 4-way inter-annotator agreement of 0.507, and a Krippendorff Alpha of 0.508. These metrics suggest that there is moderate agreement between annotators. Each of the verbs generates 2 sentences, one for each version of the alternation, and incorporates variability through a random name and noun schema of over 20 unique names and 50 unique nouns. The sentences are generated using a similar statistical method to Warstadt et al. (2020), maintaining English syntax but losing semantic value. The sentences are then human-curated, ensuring nouns maintain logical semantic agreement. The goal of this dataset is to attempt a wide approach to alternations classes by including a few examples for a multitude of verbs, compared to FAVA's narrow approach. Although FAVA-ex provides less sentences per verb, it covers three times as many verbs found from Levin (1993). The combined dataset of FAVA and FAVA-ex provides both a narrow and broad overview for each alternation class.

4 Experiment 1: Frame Membership from Word Embeddings

4.1 Method

In order to answer the first question, *Do BERT's token-level representations encode information about which syntactic frames an individual verb can participate in?*, we build a diagnostic classifier for each syntactic frame which takes a verb's embedding as input. For example, to probe the *Spray-Load* alternation, we build two binary classifiers: one that predicts whether a verb can participate in the frame exemplified by *spray paint on the wall*, that is, the *locative* frame; and one that predicts whether a verb can participate in the frame exemplified by *spray the wall with paint*, that is, the *preposition* frame.

Futhermore, we build distinct classifiers for each layer of BERT using the pretrained `bert-base-uncased` model implemented in the `HuggingFace` package (Wolf et al., 2019). For the token-embedding layer (hereafter the *static* embeddings), the verb embedding is formed by averaging the pretrained WordPiece embeddings that correspond to a particular verb. For layers 1–12 (hereafter the *contextual* embeddings), the verb embedding is formed by incorporating contextual information from the sentences in FAVA. Specifically, for each verb, we input the grammatical sentences from FAVA that contain the verb to BERT and average over the WordPiece tokens corresponding to the verb. We then average over the verb representations for all input sentences in each layer to form the "layer-embedding" for the verb.

We choose a Logistic Regression classifier without regularization as our diagnostic probe as implemented in `scikit-learn` (Buitinck et al., 2013) and show that it is sufficiently *selective* in Section 5. Following Kann et al. (2019), we use stratified k-fold cross-validation to split the verbs into 4 equally-sized folds: 3 of which are chosen to be the training set and the remaining fold chosen to be the test set.

Also following Kann et al. (2019), we report Matthews correlation coefficient (MCC) (Matthews, 1975) in addition to accuracy for model evaluation. MCC is better suited to data such as ours, in which there is an extreme majority class bias for all syntactic frames.

4.2 Results

4.2.1 Static Word Embeddings

The results on the experiments with static word embeddings are presented in Table 2, with a comparison to Kann et al. (2019)'s CoLA embeddings[2]. We also provide a comparison to a majority-class baseline. We find that for 3 of the 5 alternation classes, the linear probing classifier trained on BERT embeddings was clearly able to predict the alternation frame better than the MLP classifier trained on CoLA embeddings, particularly in the case of the prepositional variant of dative alternation.

One case in which BERT failed to outperform the original was the *with* variant of the *Spray-Load* alternation, in which performance decreased by $-.01$ in MCC. Another important case is the *Understood-Object* alternation. While the BERT embeddings did outperform the CoLA embeddings, the increases in accuracy over the majority base-

[2]The CoLA and GloVe embeddings results were quite similar. The reader is referred to the original paper for more details.

LEVIN-CLASS	CAUS-INCH		DATIVE		SPRAY-LOAD		there-INSERTION		UNDERSTOOD-OBJECT	
	Inch.	Caus.	Prep.	2-Obj	with	loc.	no-there	there	Refl	No-Refl
Positive	73	124	65	74	101	86	149	50	84	11
Negative	144	0	377	442	242	257	0	192	419	503
Total	217	124	442	516	343	343	149	242	503	514

Table 1: An updated overview of the LaVA dataset based on verb membership class distributions for each syntactic frame. "Postitive" refers to the number of verbs that can participate in the specified syntactic frame, while "Negative" refers to the number of verbs that cannot participate.

	MCC			Accuracy	
	Reference	BERT	Δ	BERT	Baseline
CAUSATIVE-INCHOATIVE					
Inchoative	0.555	0.741	0.186	0.885	0.664
Causative *	0.000	0.000	0.000	1.000	1.000
DATIVE					
Preposition	0.320	0.701	0.381	0.925	0.853
Double-Object	0.482	0.614	0.132	0.913	0.857
SPRAY-LOAD					
With	0.645	0.633	-0.012	0.848	0.706
Locative	0.253	0.525	0.272	0.834	0.749
THERE					
No-There *	0.000	0.000	0.000	1.000	1.000
There	0.459	0.593	0.134	0.876	0.793
UNDERSTOOD OBJECT					
Refl	0.000	0.466	0.466	0.867	0.833
Non-Refl	0.219	0.563	0.344	0.984	0.979

Table 2: Results from Word-Level experiments with static embeddings. Reference MCC is from Kann et al. (2019)'s CoLA experiment. Δ represents the improvement in BERT over the Reference, and *Baseline* refers to a Majority Baseline. * indicates syntactic frames which only have positive examples, which trivially achieve 100% accuracy and 0 MCC for majority class models.

line were quite modest, at 0.034 for the *Reflexive* variant and 0.005 for the *Non-Reflexive* variant.

Moreover, we find significant discrepancies in the class distributions in our table 1 and the original table in Kann et al. (2019) for the *There-insertion* and *Understood-object* alternations[3]. Hence, we note that the metric comparisons for syntactic frames under these alternations are not necessarily meaningful.

4.2.2 Contextual Word Embeddings

The results for the contextual word embeddings, that is, the average token-level embeddings extracted from inputting the sentences in the FAVA dataset to BERT, are presented in Table 3. We report a comparison of the best-performing BERT layer with the original performance on the CoLA embeddings (repeated from Table 2). We find that the contextual word embeddings extracted from within the layers of BERT dramatically outperform both the static BERT embeddings and the CoLA embeddings. Furthermore, we note that, with one curious exception, the best performing layers are concentrated around the upper half of the model, from layer 5 to layer 10. The exception is the inchoative frame, whose best-performing layer is layer 1. The full performance across all layers is represented in Figure A5.

[3]We have contacted the original authors, but have yet to hear anything definitive.

5 Control Task

A control task as described by (Hewitt and Liang, 2019) aims to combat the *Probe Confounder Problem*, which highlights the issue of supervised probe classifiers "learning" a linguistic task by combining signals in the data that are irrelevant to the linguistic property of interest. In the context our first experiment, a confounding probe would be problematic since it suggests that good model performance may be attributed to arbitrary signals picked up by the probe, as opposed to the BERT embeddings actually containing linguistic information about the syntactic frames. To mitigate the Probe Confounder Problem, we implement the following control task

	MCC				Accuracy	
	Reference	Layer	Δ	BERT Layer	Layer	Baseline
CAUSATIVE-INCHOATIVE						
Inchoative	0.555	0.959	0.404	1	0.982	0.664
Causative *	0.000	0.000	0.000	-	1.000	1.000
DATIVE						
Preposition	0.320	0.945	0.625	8	0.986	0.853
Double-Object	0.482	0.976	0.494	6	0.994	0.857
SPRAY-LOAD						
With	0.645	0.965	0.320	6	0.985	0.706
Locative	0.253	0.969	0.716	10	0.988	0.749
THERE*						
No-There *	0.000	0.000	0.000	-	1.000	1.000
There	0.459	1.000	0.541	9	1.000	0.793
UNDERSTOOD OBJECT						
Refl	0.000	0.935	0.935	5	0.982	0.833
Non-Refl	0.219	0.903	0.684	7	0.996	0.979

Table 3: Results from Word-Level experiments with contextual embeddings. *As in Table 2, we report what we believe to be the correct labels. * indicates syntactic frames which only have positive examples, which trivially achieve 100% accuracy and 0 MCC.

for the *Spray-Load* "with" syntax frame.

For each verb v_i in lava with a binary label y_i denoting whether v_i can participate in the syntax frame SL-WITH, we independently sample a control behavior $C(v)$ by randomly assigning a binary "membership" value to v_i based on the empirical membership distribution of verbs that participate in the SL-WITH syntax frame. The control task is the function that maps each verb, v_i, to the label specified by the control behavior $C(V_i)$:

$$f_{control}(v_i) = C(v_i)$$

By construction, the control task should only be learnable by supervised probe. Following the experiment design of Hewitt and Liang (2019), we compare the *selectivity* of a linear probe, an Multilayer Perceptron with 1-hidden layer (MLP-1), and an MLP with 2-hidden layers (MLP-2) where the *selectivity* of a model is defined by the difference between its accuracy on the real task (i.e. predicting verb membership for the SL-WITH frame) and the control task. In addition, we explore several "complexity control" methods including limitation of feature dimensionality, reducing the number of training examples, and increasing regularization strength.

5.1 Complexity Hyperparameters

In this section, we describe the complexity control methods in more detail and enumerate the hyperparameters that we tried for each method. The control parameters were chosen based on the three most effective methods from the experiments of Hewitt and Liang (2019). To isolate the effect of each control method, we only change one of the complexity parameters in each experiment.

5.1.1 Limiting Dimensionality

For the Logistic Regression model, we reduce the dimensionality of the feature embeddings by performing a Truncated Singular Value Decomposition and limiting the output matrix to rank k. For the MLP models, we simply limit the size of the hidden layer(s) to k.

Considering the input BERT embeddings which have 768 dimensions, we limit k to the following values: $\{20, 100, 300, 500\}$.

5.1.2 Reducing Proportion of Training Data

Because LaVA is not split into train and test sets, we use 4-fold cross validation as done in Kann et al. (2019) with 3 training folds and one test fold for evaluation on the control task. As an additional constraint, we reduce the number of training samples in each training fold by randomly sampling a proportion p of the samples and discarding the rest.

Although Zhang and Bowman (2018) recommend training on 1%, 10% and, 100% of the training data, our training data is relatively small and imbalanced (71% of verbs do not participate in the SL-WITH frame). Hence, we experiment with larger values of p: $\{0.1, 0.3, 0.5, 0.7, 0.9\}$

5.1.3 L_2 Regularization

For both the linear and MLP models, we add L_2 regularization with the following strength values:

$\{0.01, 0.1, 0.2, 0.5, 1\}$

5.2 Results

The high-level trends across experiment configurations for model selectivity and accuracy are shown in figure 1 and 2 respectively. More detailed results for the best performing models in each complexity control experiment can be found in Table A1. With the default parameters ($k = 768, p = 1, L_2 = 0$) [4], we find that the linear model outperforms both the MLP-1 and MLP-2 model in selectivity (0.420 v.s. 0.397) with no significant decrease in linguistic task accuracy (0.985 for the linear and MLP-1 models v.s. 0.988 for the MLP-2 model).

Looking at the effect of complexity control methods on model accuracy, we find that limiting dimensionality and L_2 regularization has little impact across all configurations, with the worst model (linear: $k = 20$) achieving an accuracy of 0.983 and the best model (MLP-2: $k = 100$) achieving only a slightly higher accuracy of 0.991. On the other hand, reducing the proportion of data in each training fold appears to have significant impact on model performance. For the linear model, there is a huge discrepancy in accuracy between training on 10% of the data (0.869) and the full training set (0.988). A nearly identical pattern can be observed for both of the MLP models as well.

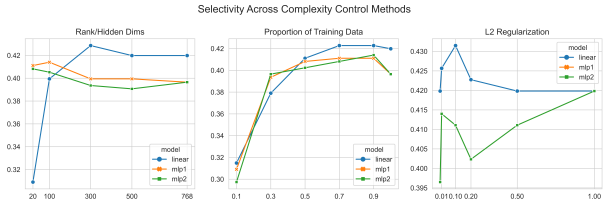

Figure 1: Linguistic Task selectivities for the three complexity control methods.

Comparing selectivity, the linear models outperform both MLPs across all complexity control methods. For dimensionality control, we see a lower selectivity in the linear model for lower values of k ($k = 20, 100$) but the best linear model ($k = 300$) achieved a higher selectivity (0.429) than the best MLP model (0.414). Similarly, the best performing configuration for reduced training samples and L_2 regularization are linear models with $p = 0.9$ (0.423) and $L_2 = 0.1$ (0.431) respectively.

[4] All other parameters are determined by the default values in scikit-learn's Logistic Regression and MLP models (https://scikit-learn.org/stable/supervised_learning.html)

Figure 2: Linguistic Task accuracies for the three complexity control methods.

We arrive at two major conclusions from the control task experiments. The first is that a linear probe is a good choice for our linguistic task since it achieves higher selectivity than the MLP models without substantial loss in model accuracy across a wide range of complexity control methods. The second is that limiting dimensionality, reducing training samples, and L_2 regularization are all effective methods for increasing model selectivity for both the linear and MLP models. However, we conclude that the best configurations were not significantly better (> 0.01 improvement in selectivity) than the default linear model so we did not make any changes to our classification probe. As we only performed these experiments for the SL-WITH frame, a great avenue for future work is to test whether our results extend to the other syntactic frames in the LaVA dataset.

6 Experiment 2: Grammar Judgments from Sentence-embeddings

6.1 Method

In this experiment, we investigate the following hypotheses. First, with regard to this specific linguistic property, is there a difference in different BERT layers? Second, does contextualized sentence embeddings encode more linguistic information than static sentence embeddings?

To test the first hypothesis, for each BERT layer, we build a Logistic Regression classifier with L_1 regularization on the FAVA training set and calculate the MCC score and accuracy of models on the test set. We ignore the held out development set because the probe hyperparameters does not need to be hypertuned. The whole process can be described by the following equation:

$$c_{s_i} = f(\mathbf{W}\mathbf{s}_i + \mathbf{b})$$

where $\mathbf{s}_i$ refers to the embedding of the whole sentence for layer i (by averaging all i layer's hidden states of words in the sentence s), f refers to the logistic regression classifier, $\mathbf{W}$ and $\mathbf{b}$ are the param-

eters of f, and c_{s_i} is a binary value corresponding to whether the sentence is grammatical.

To test the second hypothesis, we generate contextualized sentence embeddings for each classifier using the best layer for each alternation class based on the results of the process described above. Then, we compare the results with the reference model proposed by (Kann et al., 2019), which uses GloVe embeddings, (Pennington et al., 2014a) and the majority baseline model (BASELINE).

We additionally repeat both experiments using the FAVA-ex instead of FAVA to test whether including more training examples helps with model performance.

6.2 Results

With respect to the first hypothesis, a graph of MCC scores appears in Figure (3). (The accuracy scores show the same trend and can be found in Figure A1 of the Appendix.). From the figure, we can see that except for the *Understood-object* category, there are significant distinctions in these different layers. The trend is that lower layers do not encode enough related linguistic knowledge about the whole sentence. Also, four alternation classes achieve the best performance on the 9^{th} or 10^{th} layer, which demonstrates that higher layer will encode more syntactic information that is important for the grammaticality judgement. The outlier is the dative class. The best informative layer for dative is the 6^{th}, suggesting that the dative frames may test for different types of linguistic knowledge than other alternation classes.

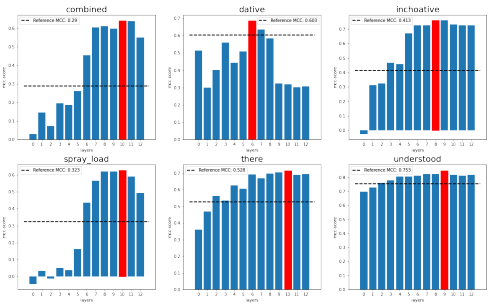

Figure 3: Layer-by-layer MCC score for each alternation class on FAVA

With respect to the second hypothesis, Table 4 shows the MCC scores for the classifiers trained on contextualized embeddings from the best performing layers reported above. As defined by (Kann et al., 2019) an MCC value between 0.5 and 0.7 demonstrates a moderate correlation between predicted and true labels while an MCC greater than 0.7 implies strong correlation. From the table, we see that all MCC values on the BERT embeddings are greater than 0.6, and 6 out of 12 experiments achieve a strong correlation. Also, we can see that MCC values on for BERT are much greater than the those for GloVE. Therefore, we conclude that the contextualized embeddings encode more alternation information than the static embeddings.

Finally, we try to further improve the performance of our classifier probes by augmenting the training data of FAVA with more verbs and sentence examples, as explained in section 3.1, and repeating the probing experiments on FAVA-ex. The results for this experiment, BERT-AUG, also appear in Table 4. Although the additional training data led to modest improvement for some alternation classes, it also led to decreased performance in others. This result is contrary to our expectation, as we assumed including more diverse sentence examples would help the model generalize to unseen sentences.

7 Discussion

For the word level experiments, our linear probe achieved strong correlation (> 0.7 MCC, as defined by Kann et al. (2019)) across all syntactic frames with the strongest performance in the *there* frame (1.00) and the weakest performance in the *Non-Reflexive* frame (0.903). The best embedding layer for each syntactic frame corresponds to a unique layer other than layer 6, which achieved the highest MCC on both the *Double-Object* and *With* frames. When analyzing the average MCC of the contextual layers across all syntactic frames, we find that the middle (5-9, unordered) layers of BERT have the best performance (0.921-0.943), followed closely by the top (10-12) layers (0.913-0.920). On the other hand, the bottom (1-4) layers perform substantially worse (0.829-0.902) than both the top and middle layers. Since understanding of verb alternation structures is a task that primarily requires syntactic knowledge, this roughly aligns with the hypothesis of Jawahar et al. (2019) that syntactic features are encoded in the middle layers of BERT while semantic features are encoded in the higher layers.

For the sentence-level experiments, we see a similar outcome wherein the upper-middle layers of BERT achieved the best performance on average. Also, we find similar patterns as Kann et al. (2019) for the difficulty of predicting sentences for each

	COMB.	CAUSATIVE-INCHOATIVE	DATIVE	SPARY-LOAD	THERE	UNDERSTOOD
MCC						
BERT	0.643(10)*	0.760(8)	**0.686**(6)	0.628(10)	**0.716**(10)	0.848(9)
BERT-AUG	**0.649**(10)	**0.772**(6)	0.672(6)	**0.630**(8)	0.710(8)	**0.849**(9)
GLOVE	0.290	0.603	0.413	0.323	0.528	0.753
ACCURACY						
BERT	84(10)	92(8)	**88.5** (6)	**82.1**(10)	**89** (10)	**92.5**(9)
BERT-AUG	**84.1**(10)	**92.4**(6)	87.2(6)	81.9(8)	88.1(8)	**92.5**(9)
GLOVE	64.6	85.4	76	66.2	72.9	87.4
BASELINE	66.6	77.6	82.1	60.3	77.5	53.7

Table 4: Results from Sentence-Level experiments. "BERT" are models trained on the original FAVA datasets; "BERT-AUG" are models trained on the datasets augmented with FAVA-ex; "GloVe" denotes the reference probing model in (Kann et al., 2019); Bolded values show the best result for each alternation class. *As in Table 4, '()' indicates which BERT layer achieved the best performance for that alternation.

alternation class, achieving strong correlation for *Understood-object* and *There-insertion* and moderate correlations for *Causative-Inchoative*, *Dative*, and *Spray-Load*. There are a few other interesting phenomena we noticed. First, sentences that are wrongly predicted for individual class models are also predicted incorrectly for the combined model. Second, we observe that there are at least four sentences for each verb in FAVA, and if the model fails to predict one sentence correctly, it tends to predict them all incorrectly. In this paper, we introduce FAVA-ex to attempt to introduce a wider range of verbs with fewer sentences to address this issue, but the additional training data only yielded marginal improvements for the *Combined*, *Causative-Inchoative*, and *Understood-object* classes. This indicates that there is little to no linguistic knowledge gleaned from the additional sentences in FAVA-ex.

While we are optimistic about our results, there are several limitations to our experiments. First, we only analyze five different alternation classes which is a small subset of the 83 classes presented in Levin (1993). In addition – although our control task ensures that our classifier probe is relatively *selective* for the first experiment, it may not necessarily generalize well to the second experiment or even other syntactic frames. In the future, we hope to expand our experiments to more alternation classes as well as the scope of the control task.

8 Conclusion and Future Work

Overall, our results support the hypothesis that BERT embeddings encode linguistic information about verb alternation classes at both the word and sentence levels. Furthermore, when considering contextual embeddings instead of just BERT's static WordPiece embeddings, BERT outperform the static embeddings used by Kann et al. (2019) for both experiments across all alternation classes. Additionally, we find that the upper-middle layers achieve the best performance for most of the experiments. While there are numerous factors that may be responsible for BERT's performance, we hypothesize that the improvement can largely be attributed to the attention and positional encoding mechanisms of BERT embeddings since we only saw modest improvements in MCC when using the WordPiece embeddings but significant improvements for the contextual embeddings.

From the data perspective, supporting more alternations aside from the 5 main classes, as well as sub-alternations within classes, remains to be explored. Polishing the FAVA-ex dataset to extend to all verbs in Levin (1993), as well as more than 2 examples per verb, might lead to significant improvements in model accuracy.

Beyond the expansion of verb alternation classes, there are several interesting adaptations that can be made in our experiment methodology. For example, instead of limiting our analysis to the BERT base model, it may be interesting to compare our results with other encoder models to determine how varying architectures, training approaches, and modeling objectives may affect performance on the linguistic tasks. Moreover, while we attempt to control the *Probe Confounder Problem* by building a selective probe, there is no guarantee the classifier probes do not pick up on arbitrary signals in the training data that improve performance. A promising solution may be to remove the supervised probes altogether and redesign the experiments as unsupervised tasks as done in Warstadt et al. (2020).

References

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

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

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743.

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

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 287–297.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Beth Levin, Malka Rappaport Hovav, and Samuel Jay Keyser. 1995. *Unaccusativity: At the syntax-lexical semantics interface*, volume 26. MIT press.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014a. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014b. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Steven Pinker. 1989. *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Florian Schafer. 2009. The causative alternation. *Language and linguistics compass*, 3(2):641–681.

Alex Warstadt and Samuel R. Bowman. 2019. Linguistic analysis of pretrained sentence encoders with acceptability judgments.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: A benchmark of linguistic minimal pairs for English. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 409–410, New York, New York. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.

A Appendix

Figure (A1) - (A4) show the MCC and accuracy values on different layers in BERT on the FAVA datasets and the augmented FAVA datasets.

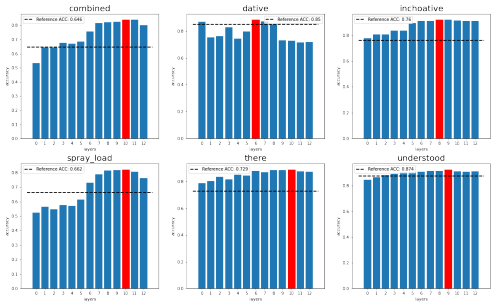

Figure A1: Layer-by-layer accuracy for each alternation class on FAVA

	Dimensions (k)	Training Prop. (p)	L_2 Reg.	Accuracy	Selectivity
DEFAULT PARAMS					
Linear	768	1.0	0.0	0.985	**0.420**
MLP-1	768	1.0	0.0	0.985	0.397
MLP-2	768	1.0	0.0	0.988	0.397
LIMITING DIMENSIONS					
Linear	300	1.0	0.0	0.985	**0.429**
MLP-1	100	1.0	0.0	0.985	0.414
MLP-2	20	1.0	0.0	0.983	0.408
REDUCING TRAINING SAMPLES					
Linear	768	0.9	0.0	0.988	**0.423**
MLP-1	768	0.9	0.0	0.983	0.411
MLP-2	768	0.9	0.0	0.985	0.414
L_2 REGULARIZATION					
Linear	768	1.0	0.1	0.985	**0.431**
MLP-1	768	1.0	1.0	0.988	0.420
MLP-2	768	1.0	1.0	0.988	0.420

Table A1: Results from the Complexity Control Experiments. For each experiment, only the best performing configuration for each model is reported.

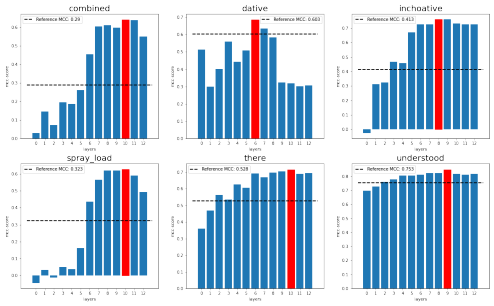

Figure A2: Layer-by-layer MCC for each alternation class on FAVA

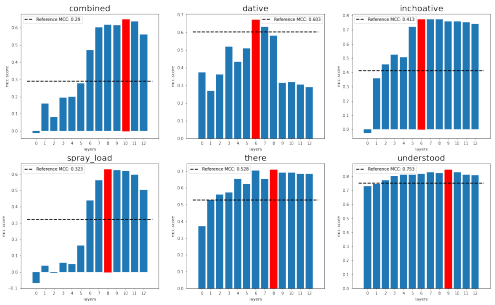

Figure A4: Layer-by-layer MCC for each alternation class on augmented FAVA

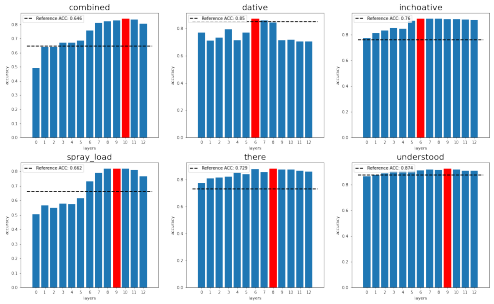

Figure A3: Layer-by-layer accuracy for each alternation class on augmented FAVA

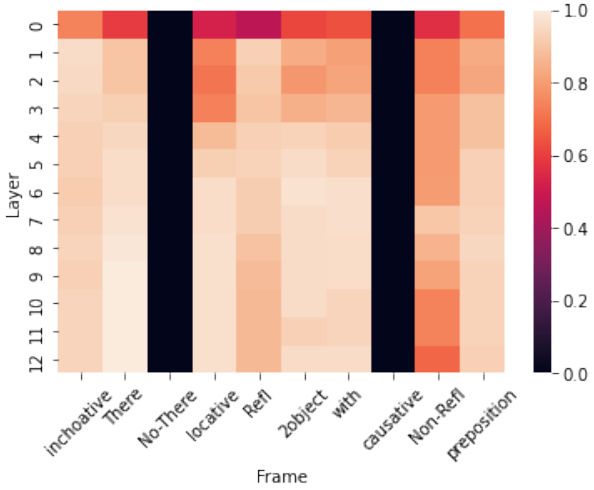

Figure A5: Heatmap of performance on experiment 1 for each BERT layer. Layer 0 corresponds to the static WordPiece layer.