Probing Internal Representations of Multi-Word Verbs in Large Language Models

Hassane Kissane[1,2], Achim Schilling[2,3], Patrick Krauss[2,3]
[1]Chair of English Philology and Linguistics, University Erlangen-Nuremberg, Germany
[2]Cognitive Computational Neuroscience Group, University Erlangen-Nuremberg, Germany
[3]Neuroscience Lab, University Hospital Erlangen, Germany
*

Abstract

This study investigates the internal representations of verb-particle combinations, called multi-word verbs, within transformer-based large language models (LLMs), specifically examining how these models capture lexical and syntactic properties at different neural network layers. Using the BERT architecture, we analyze the representations of its layers for two different verb-particle constructions: phrasal verbs like *give up* and prepositional verbs like *look at*. Our methodology includes training probing classifiers on the model output to classify these categories at both word and sentence levels. The results indicate that the model's middle layers achieve the highest classification accuracies. To further analyze the nature of these distinctions, we conduct a data separability test using the Generalized Discrimination Value (GDV). While GDV results show weak linear separability between the two verb types, probing classifiers still achieve high accuracy, suggesting that representations of these linguistic categories may be *non-linearly separable*. This aligns with previous research indicating that linguistic distinctions in neural networks are not always encoded in a linearly separable manner. These findings computationally support usage-based claims on the representation of verb-particle constructions and highlight the complex interaction between neural network architectures and linguistic structures.

1 Introduction

1.1 The linguistic problem

Multi-word verbs or verb-particle combinations are a linguistic category presented in the English language in which the lexical verb is combined with a particle to form an independent unit. It is called a phrasal verb when the lexical verb is combined with an adverbial particle like *work out*. It is a prepositional verb when the verb is combined with a prepositional particle like *rely on* (Carter and McCarthy, 2006). Usually, the prepositional verbs are followed by a noun phrase. Rather than the nature of the following particle, there are several differences between phrasal verbs and prepositional verbs. One main difference between the two categories is the particle placement in phrasal verbs and the foxed order in prepositional verbs. Where in phrasal verbs, the particle can sometimes be separated from the verb and placed after the object. In contrast, the only grammatical form in prepositional verbs is the V+prepostion+object.

- *Turn off the light. (phrasal)*
- *Turn the light off. (phrasal)*
- *Look at the painting. (prepositional)*
- **Look the painting at.(prepositional)*

Several studies explored the mental storage of these verb-particle constructions, specifically phrasal verbs, to see in which way they are stored and processed in the brain. For instance Cappelle et al. (2010) and further discussed by Pulvermüller et al. (2013) that phrasal verbs are processed as single lexical units, as evidenced by MEG. However, prepositional verbs remain unexplored, which are still treated similarly to phrasal verbs in terms of both the verb and the particle form a single lexical unit called verb, for example the prepositional verbs *look at*, and the phrasal verb *turn off* (Quirk et al., 1985; Carter and McCarthy, 2006). From a constructional point of view, Herbst and Schüller (2008) proposed what is called the valency model for the distinction between phrasal verbs and prepositional verbs, assuming that prepositions function as integral parts of the complement rather than the verb itself in prepositional verbs. This valency-based approach emphasizes the syntactic relationship between the verb and its complements, analyz-

*Correspondence: patrick.krauss@fau.de

Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025), pages 7–13
May 4, 2025 ©2025 Association for Computational Linguistics

ing prepositional verbs like *look at* as the verb *look* and the complement *at*.

1.2 probing-based methods for linguistic tasks

Probing methods analyze the linguistic properties encoded in the representations of the NLP model. Probes are supervised models trained to predict linguistic properties or other categories, such as parts of speech or word meanings, from model representations such as BERT embeddings (Immertreu et al., 2024; Ramezani et al., 2024b). These probes have achieved high accuracy on various linguistic tasks, demonstrating their utility in understanding how models encode features such as syntax and semantics (Conneau et al., 2018). The search classifiers are trained on the activations to identify predefined concepts or linguistic properties, such as syntactic tags or semantic meanings, from the model output embeddings (Hupkes and Zuidema, 2018; Sajjad et al., 2022). Furthermore, layer-wise analysis (Tenney et al., 2019a; Ramezani et al., 2024b; Krauss et al., 2024; Banerjee et al., 2025; Ramezani et al., 2024a) investigates how linguistic knowledge is distributed across the layers of transformer-based models, providing insights into the hierarchical organization of encoded knowledge.

The internal representations of LLMs are frequently analyzed using probing approaches. Tenney et al. (2019a) employ probing tasks to investigate the linguistic information that BERT gathers and discover that various layers encode different kinds of linguistic properties. A set of probes is presented by Tenney et al. (2019b) to examine the representations acquired by contextualized word embeddings and to determine the distribution of syntactic and semantic information among layers.

While prior studies (e.g., Cappelle et al. (2010); Pulvermüller et al. (2013)) have suggested that phrasal verbs function as single lexical units, and Herbst and Schüller (2008) proposed a valency-based linguistic distinction for prepositional verbs, it remains unclear whether these distinctions are reflected in the internal representations of neural language models. This study aims to investigate how neural language models encode and differentiate between these two linguistically distinct categories of multi-word verbs. Specifically, we examine whether internal representations capture key syntactic, lexical, and compositional differences. To achieve this, we apply probing classifiers to measure classification accuracy across layers and data separability methods to assess how distinctly these verb categories are organized within the representational space of a neural language model.

2 Methods

2.1 Data

The dataset consists of sentences containing phrasal and prepositional verbs, extracted from the British National Corpus (BNC, 2001). Sequences were selected based on syntactic variability using part-of-speech (PoS) tag patterns:

- *Phrasal verbs were identified using the pattern "V + ADV + Det + N", where the output is like* ***look up the word****.*
- *Prepositional verbs followed the pattern "V + PREP + Det + N", where the the output is like* ***look after the child****.*

The dataset was manually divided to ensure that each verb appearing in the training set does not appear in the test set. This was done to prevent overlap in representation and ensure that the classifiers generalize beyond memorization. The training set includes 1920 examples of phrasal verbs and 2070 of prepositional verbs. The test set contains 522 phrasal verb examples and 623 prepositional verb examples, with a total of 2442 for phrasal verbs and 2693 for prepositional verbs, as shown in Table 1. Since our study focuses on probing analysis rather than optimizing a model, we did not require hyperparameter tuning, which typically demands a development set. Before using the dataset as input for the model, we applied several cleaning steps, which are detailed in Table 2.

2.2 Model (Embedding Extraction)

We use transformer-based model (Vaswani et al., 2017) BERT (Devlin et al., 2019) as the feature extraction model for generating contextual embeddings. Specifically, we use the *bert-base-uncased* version, consisting of 12 layers, each producing 768-dimensional contextual embeddings for input tokens. For each sample, we extract embeddings at two levels:

Token-Level Embeddings For verb-specific analysis, we extract the embedding corresponding to the main token of the verb (e.g., *give* in *give up*). These embeddings focus on the localized representation of the verb within the sentence.

Sentence-Level Embeddings To capture the entire context of the sentence, we compute the average of all token embeddings in the sentence. This

Phrasal	#	Prepositional	#
		Training	
blow_up	52	break_into	147
break_down	134	call_on	138
close_down	54	come_across	168
fill_up	54	do_without	76
find_out	243	get_off	184
finish_off	46	care_for	150
give_away	35	cope_with	150
give_up	239	get_into	150
hand_in	229	get_on	150
hold_up	56	go_into	150
look_up	67	lead_to	148
put_off	57	listen_to	153
shut_down	57	look_at	154
throw_away	58	look_for	152
turn_down	75		
wake_up	31		
take_over	102		
work_out	101		
sort_out	230		
Total	1920	Total	2070
		Test	
take_up	100	depend_on	150
carry_on	184	look_after	154
bring_up	115	deal_with	153
check_out	123	get_over	111
		approve_of	55
Total	522	Total	623
Grand Total	2442	Grand Total	2693

Table 1: Distribution of phrasal and prepositional verbs in training and test sets with their frequencies.

approach aggregates information across all tokens, providing a representation of the sentence without relying solely on the [CLS] token embedding.

Our study focuses on *bert-base-uncased* as a widely used transformer model, but we acknowledge that different LLM architectures may encode linguistic categories differently. Future research could extend this analysis to other models, such as *roberta-base* or *bert-large*, to assess whether the observed patterns generalize across architectures.

2.3 Classification Models

Logistic Regression (LR) is a linear model used in modeling the probabilities of possible outcomes given an input variable.

Support Vector Machines (SVM) perform well on smaller datasets by optimizing data transformations based on predefined classes. They are based on the principle of Structural Risk Minimization from Statistical Learning Theory (Boser et al., 1992). In their fundamental form, SVMs learn linear discrimination that separates positive examples from negative ones with a maximum margin. This margin, defined by the distance of the hyperplane to the nearest positive and negative examples, has proven to have good properties in terms of generalization bounds for the induced classifiers.

2.4 Generalized Discrimination Value (GDV)

We used the GDV to calculate cluster separability as published and explained in detail in Schilling et al. (2021). Briefly, we consider N points $\mathbf{x_{n=1..N}} = (x_{n,1}, \cdots, x_{n,D})$, distributed within D-dimensional space. A label l_n assigns each point to one of L distinct classes $C_{l=1..L}$. In order to become invariant against scaling and translation, each dimension is separately z-scored and, for later convenience, multiplied with $\frac{1}{2}$:

$$s_{n,d} = \frac{1}{2} \cdot \frac{x_{n,d} - \mu_d}{\sigma_d}. \quad (1)$$

Here, $\mu_d = \frac{1}{N}\sum_{n=1}^{N} x_{n,d}$ denotes the mean, and $\sigma_d = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_{n,d} - \mu_d)^2}$ the standard deviation of dimension d. Based on the re-scaled data points $\mathbf{s_n} = (s_{n,1}, \cdots, s_{n,D})$, we calculate the *mean intra-class distances* for each class C_l

$$\bar{d}(C_l) = \frac{2}{N_l(N_l-1)} \sum_{i=1}^{N_l-1} \sum_{j=i+1}^{N_l} d(\mathbf{s}_i^{(l)}, \mathbf{s}_j^{(l)}), \quad (2)$$

and the *mean inter-class distances* for each pair of classes C_l and C_m

$$\bar{d}(C_l, C_m) = \frac{1}{N_l N_m} \sum_{i=1}^{N_l} \sum_{j=1}^{N_m} d(\mathbf{s}_i^{(l)}, \mathbf{s}_j^{(m)}). \quad (3)$$

Here, N_k is the number of points in class k, and $\mathbf{s}_i^{(k)}$ is the i^{th} point of class k. The quantity $d(\mathbf{a}, \mathbf{b})$ is the euclidean distance between $\mathbf{a}$ and $\mathbf{b}$. Finally, the Generalized Discrimination Value (GDV) is calculated from the mean intra-class distances

$$\langle \bar{d}(C_l) \rangle = \frac{1}{L} \sum_{l=1}^{L} \bar{d}(C_l) \quad (4)$$

and the mean inter-class distances

$$\langle \bar{d}(C_l, C_m) \rangle = \frac{2}{L(L-1)} \sum_{l=1}^{L-1} \sum_{m=l+1}^{L} \bar{d}(C_l, C_m) \quad (5)$$

as follows:

$$\text{GDV} = \frac{1}{\sqrt{D}} \left[\langle \bar{d}(C_l) \rangle - \langle \bar{d}(C_l, C_m) \rangle \right] \quad (6)$$

<table>
<tr><th>Character</th><th>Pre-processing step</th></tr>
<tr><td>Punctuations
(!"#$%&'()*+,-./:;<=>?@[\]\^_
\`{}|\~)</td><td>Removed</td></tr>
<tr><td>Leading and trailing whitespaces</td><td>Removed</td></tr>
<tr><td>Extra whitespaces</td><td>Replaced with a single space</td></tr>
<tr><td>Uppercase characters</td><td>Converted to lowercase</td></tr>
</table>

Table 2: Text pre-processing steps applied to the dataset.

whereas the factor $\frac{1}{\sqrt{D}}$ is introduced for dimensionality invariance of the GDV with D as the number of dimensions.

Note that the GDV is invariant with respect to a global scaling or shifting of the data (due to the z-scoring), and also invariant concerning a permutation of the components in the N-dimensional data vectors (because the euclidean distance measure has this symmetry). The GDV is zero for completely overlapping, non-separated clusters, and it becomes more negative as the separation increases. A GDV of -1 signifies already a very strong separation.

3 Results

Token-based classification

The results of the lexical verb token classification task using logistic regression and linear SVM have shown distinct trends across the 12 layers of the BERT model Figure 1. For the Logistic Regression classifier, accuracy starts at 0.87 in the input layer, remains stable around 0.84 - 0.80 through layers 2 to 4, and then increases significantly, reaching 0.99 in layer 6 before slightly decreasing in the late layers of the model. Similarly, the linear SVM classifier achieves an accuracy of 0.63 in the input layer. Then 0.84, through layers 2 to 4. To start improving from layer 5 onward, reaching its highest accuracy of 0.99 at layer 8. Both classifiers show the best accuracy in the middle layers (layers 6–9, suggesting that these layers encode the most significant linguistic features to distinguish between phrasal verbs and prepositional verbs. Therefore, the accuracies decrease slightly in the late layers, which indicates a shift towards task-specific representations less suited for this classification task.

Sentence-Based Classification

For the sentence-based classification task, both Logistic Regression and Linear SVM show distinct trends across the 12 BERT layers Figure 1. However, the accuracies in the token-based classification were higher than those based on sentence embeddings. The Logistic Regression classifier begins with an accuracy of 0.80 in the input layer and improves to 0.85 in layers 6 and 7. Then, accuracy decreases in the late layers, dropping to 0.69 in layers 11 and 12. Similarly, the linear SVM classifier starts with an accuracy of 0.77 and 0.76 in the input and first layer respectively, to 0.84 in layer 6, then decreases to the lowest accuracy of 0.66 in the final layer of the model. With these results, it is suggested that the middle layers (layers 5–7) of the model are the best to capture linguistic information at sentence-level representations to distinguish phrasal verbs and prepositional verbs, while the higher layers, likely focused on task-specific semantics, encode features less suited to these properties predictions.

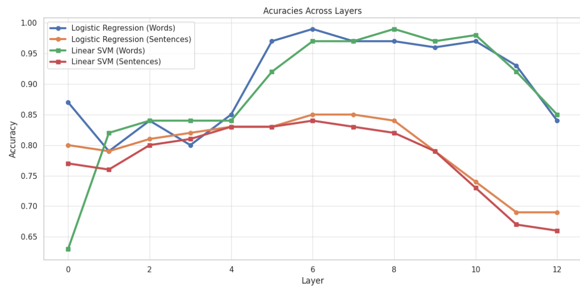

Figure 1: Classification accuracies

GDV Values for Data Separability

The GDV calculations for both token-based and sentence-based embeddings has shown non-strong separation between between the phrasal and prepositional verbs across BERT layers Figure 2. For the token-based embeddings, GDV values start at equivalent of 0.00 in the input layer which is responsible for converting tokens into dense vector representations before they are processed by the transformer layers. Then, the GDV has shown an improvement (less negative) across the layers, reaching their strongest separability at layers 3 and 4 with a value of -0.049 and -0.048 respectively.

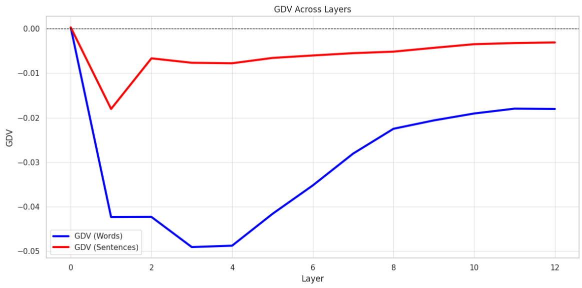

Figure 2: GDV values for data separability between the two multi-word verbs classes across BERT layers

This improvement indicates that BERT's middle layers may encode more discriminative features for distinguishing between the two verb types in word embeddings.

After we ran a normality test on the classification accuracies and GDV scores across BERT layers, we found that the data was not normally distributed, which justified the use of Spearman's rank correlation as a non-parametric test Figure 3. The analysis showed no statistically significant correlation between the variables. For words, the correlation between Logistic Regression and GDV was r_s = 0.32, p = 0.285, and between Linear SVM and GDV, r_s = 0.26, p = 0.383. For sentences, the correlation between Logistic Regression and GDV was r_s = -0.44, p = 0.128, while Linear SVM and GDV showed a negative correlation of r_s = -0.52, p = 0.069, approaching significance. Overall, no strong or significant associations were observed.

4 Discussion

Our findings indicate that while probing classifiers and GDV provide some investigations into how BERT encodes differences between linguistic categories, they may not fully capture the complexities of linguistic representations. Particularly, when distinguishing between phrasal and prepositional verbs based on token-level embeddings. As proposed by Goldberg (1995), lexico-semantic elements convey a portion of linguistic information, but they do not embody all structural and functional aspects present in a text. This limitation is particularly relevant in the case of the investigated cases in the study, where distinctions often emerge from interactions between lexical, syntactic, and semantic factors rather than being determined by individual token representations.

Our study focuses on bert-base-uncased as a widely used transformer model, but we acknowledge that different LLM architectures may encode linguistic categories differently. Future research could extend this analysis to other models, such as roberta-base or bert-large, to assess whether the observed patterns generalize across architectures.

This perspective aligns with the constructionist approach to language processing (Madabushi et al., 2020), which challenges the traditional separation of lexical and grammatical elements. Instead, it proposes a continuum of constructions—where linguistic representations arise from learned pairings of form and meaning rather than being strictly lexical or grammatical only. From this point, phrasal and prepositional verbs might be better understood as integrated constructions, rather than purely lexical or syntactic units. Consequently, probing classifiers, which primarily capture lexical or semantic properties in token-based classification tasks, may fail to fully account for the grammatical and contextual information that shapes the representation of these constructions. This is evident in the mismatch between classification accuracies and GDV values, suggesting that different methods may capture other dimensions of representation.

Several studies have discussed the limitations of probing classifiers (Belinkov, 2022; Sajjad et al., 2022). One major limitation is the disconnect between probing accuracy and the original model's internal processing. While probing classifiers can detect correlations between model embeddings and linguistic features, they do not necessarily indicate whether the model actively uses these properties for linguistic processing. This limitation is apparent in our findings: while probing classifiers achieved high accuracies, GDV analysis showed weak linear separability between phrasal and prepositional verbs, as indicated by the lack of significant correlation between classification accuracies and GDV values.

This observed Disagreement between classifier accuracies and GDV values aligns with previous research suggesting that internal representations in neural networks and large language models are not necessarily linearly separable (Hewitt and Liang, 2019; Kissane et al., 2024; Zhang and Bowman, 2018; Banerjee et al., 2025; Hildebrandt et al., 2025; Krauss et al., 2024; Ramezani et al., 2024b). Since GDV measures linear separability, it does not capture non-linearly structured representations. In contrast, probing classifiers can still detect non-linearly separable distinctions, allowing them to identify linguistic categories that may be encoded

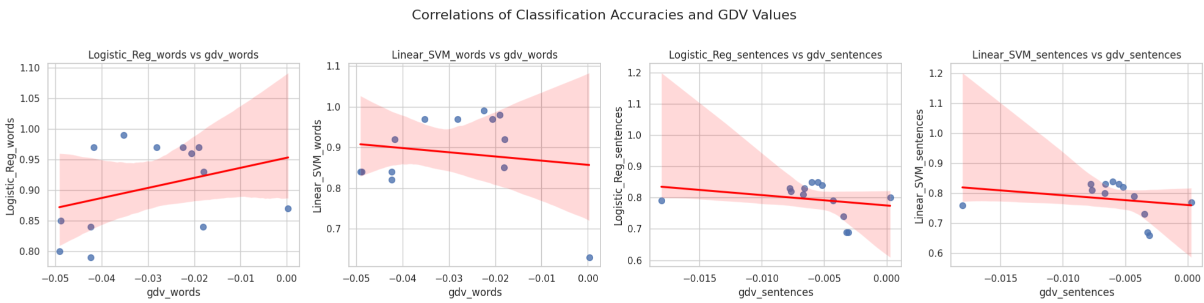

Figure 3: The correlation between probing classifier accuracy and Generalized Discrimination Value (GDV) scores across BERT layers. While probing classifiers achieve high accuracy in distinguishing phrasal and prepositional verbs, GDV values remain close to zero, indicating that these categories are not linearly separable in BERT's representation space.

in high-dimensional space. Therefore, the low GDV scores do not suggest that BERT fails to encode multi-word verb distinctions, but rather that these representations may require non-linear transformations to be fully distinguished. This Point up the need for comprehensive analytical methods when investigating how LLMs structure linguistic knowledge and suggests that linear separability should not be the only one criterion for assessing learned representations.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): KR 5148/3-1 (project number 510395418), KR 5148/5-1 (project number 542747151), and GRK 2839 (project number 468527017) to PK, and grant SCHI 1482/3-1 (project number 451810794) to AS.

References

Awritrojit Banerjee, Achim Schilling, and Patrick Krauss. 2025. Exploring narrative clustering in large language models: A layerwise analysis of bert. *arXiv preprint arXiv:2501.08053*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

BNC. 2001. The British National Corpus, version 2 (BNC World). Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Title*, COLT '92, page 144–152, New York, NY, USA. Association for Computing Machinery.

Bert Cappelle, Yury Shtyrov, and Friedemann Pulvermüller. 2010. Heating up or cooling up the brain: MEG evidence that phrasal verbs are lexical units. *Brain and Language*, 115(3):189–201.

R. Carter and M. McCarthy. 2006. *Cambridge Grammar of English: A Comprehensive Guide*. Cambridge Grammar of English Series. Cambridge University Press.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Thomas Herbst and Susen Schüller. 2008. *Introduction to Syntactic Analysis: A Valency Approach*. Narr Francke Attempto Verlag.

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, Hong Kong, China. Association for Computational Linguistics.

Fabian Hildebrandt, Andreas Maier, Patrick Krauss, and Achim Schilling. 2025. Refusal behavior in large language models: A nonlinear perspective. *arXiv preprint arXiv:2501.08145*.

Dieuwke Hupkes and Willem H. Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure (extended abstract). In *International Joint Conference on Artificial Intelligence*.

Mathis Immertreu, Achim Schilling, Andreas Maier, and Patrick Krauss. 2024. Probing for consciousness in machines. *arXiv preprint arXiv:2411.16262*.

Hassane Kissane, Achim Schilling, and Patrick Krauss. 2024. Analysis and visualization of linguistic structures in large language models: Neural representations of verb-particle constructions in bert. *arXiv preprint arXiv:2412.14670*.

Patrick Krauss, Jannik Hösch, Claus Metzner, Andreas Maier, Peter Uhrig, and Achim Schilling. 2024. Analyzing narrative processing in large language models (llms): Using gpt4 to test bert. *arXiv preprint arXiv:2405.02024*.

Harish Tayyar Madabushi, Laurence Romain, Petar Milin, and Dagmar Divjak. 2020. Cxgbert: Bert meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032. International Committee on Computational Linguistics.

Friedemann Pulvermüller, Bert Cappelle, and Yury Shtyrov. 2013. 397 brain basis of meaning, words, constructions, and grammar. In *The Oxford Handbook of Construction Grammar*. Oxford University Press.

Randolph Quirk, Sidney Greenbaum, and Geoffrey Leech. 1985. *A Comprehensive Grammar of the English Language*. Longman.

Pegah Ramezani, Achim Schilling, and Patrick Krauss. 2024a. Analysis of argument structure constructions in a deep recurrent language model. *arXiv preprint arXiv:2408.03062*.

Pegah Ramezani, Achim Schilling, and Patrick Krauss. 2024b. Analysis of argument structure constructions in the large language model bert. *arXiv preprint arXiv:2408.04270*.

Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. 2022. Neuron-level interpretation of deep NLP models: A survey. *Transactions of the Association for Computational Linguistics*, 10:1285–1303.

Achim Schilling, Andreas Maier, Richard Gerum, Claus Metzner, and Patrick Krauss. 2021. Quantifying the separability of data classes in neural networks. *Neural Networks*, 139:278–293.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

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

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

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.