Refining Targeted Syntactic Evaluation of Language Models

Benjamin Newman **Kai-Siang Ang** **Julia Gong** **John Hewitt**
Department of Computer Science
Stanford University
{blnewman,kaiang,jxgong,johnhew}@cs.stanford.edu

Abstract

Targeted syntactic evaluation of subject-verb number agreement in English (TSE) evaluates language models' syntactic knowledge using hand-crafted minimal pairs of sentences that differ only in the main verb's conjugation. The method evaluates whether language models rate each grammatical sentence as more likely than its ungrammatical counterpart. We identify two distinct goals for TSE. First, evaluating the *systematicity* of a language model's syntactic knowledge: given a sentence, can it conjugate arbitrary verbs correctly? Second, evaluating a model's *likely behavior*: given a sentence, does the model concentrate its probability mass on correctly conjugated verbs, even if only on a subset of the possible verbs? We argue that current implementations of TSE do not directly capture either of these goals, and propose new metrics to capture each goal separately. Under our metrics, we find that TSE overestimates systematicity of language models, but that models score up to 40% better on verbs that they predict are likely in context.

1 Introduction

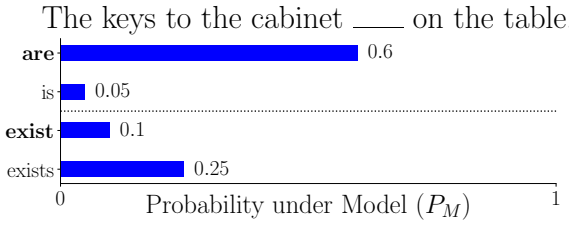
As neural language models have emerged as both broadly useful engineering tools (Devlin et al., 2018; Radford et al., 2019) and potential models of human language processing (Linzen and Leonard, 2018; Ettinger et al., 2018; Futrell et al., 2019), evaluations targeting their syntactic ability have been developed to better understand their capabilities.

One such method for syntactic evaluation tests models' knowledge of English subject-verb (S/V) number agreement (Linzen et al., 2016; Gulordava et al., 2018). These studies consider minimal pairs of sentences, such as *The keys to the cabinet is/are on the table*, that differ only in verb number, and test if models rate grammatical sentences as more probable. The syntactically correct of the two sentences is sampled from natural corpora (Linzen et al., 2016; Kuncoro et al., 2018) or constructed from templates. The use of templates, known as Targeted Syntactic Evaluation (TSE), allows for the fine-grained evaluation of models on specific, often rare, syntactic phenomena (Marvin and Linzen, 2018; Ettinger et al., 2018; Warstadt et al., 2020), but (when evaluating S/V number agreement) relies on researchers hand-specifying a small set of verb lemmas that are substituted into each template.

In this work, we improve the TSE methodology by disentangling its broad objective of evaluating syntactic ability into two distinct goals, and we introduce two variants of TSE to separately capture each goal. These evaluations demonstrate that neural models do not generally consider well-conjugated verbs more likely than their incorrect conjugations, but instead prefer to correctly conjugate verbs they deem likely.

We argue that the objective of evaluating syntactic ability can be decomposed into two goals and that current implementations of TSE do not achieve either of them. The first goal is measuring **systematicity**: for a specific syntactic construction, does the model correctly conjugate arbitrary verbs with the grammatical number of the subject? TSE currently fails to capture this because it evaluates models using only a small set of verbs for each syntactic construction. If models only conjugate these verbs correctly, they receive a high score, even if they conjugate other verbs incorrectly. The second goal is measuring **likely behavior**: when we sample verbs from the model in a specific syntactic construction, will they be properly conjugated? TSE fails to directly capture this because the small set of verbs used in evaluation might differ from the verbs that are likely in context under the model. If models conjugate these hand-specified verbs incorrectly, they receive a low score, even if they correctly conjugate more likely verbs.

To motivate these goals and the misspecification of TSE, consider evaluating a language model on



Metric	Computation	Score
TSE	**are** $>$ is	1.0
EW (systematicity)	**are** $>$ is exists $>$ **exist**	0.5
MW (likely behavior)	$\frac{\textbf{are} + \textbf{exist}}{\textbf{are} + \textbf{exist} + \text{is} + \text{exists}}$	0.7

Table 1: A toy example where a language model puts more probability mass on the correct *are* and the incorrect *exists*, showing how TSE currently may not fully capture a model's syntactic ability. In contrast, we propose EW, which captures this failure of systematicity, and MW, which captures the probability of sampling a correct conjugation. Bolded verbs are correct.

the following two pairs of sentences:

The keys to the cabinet is/are on the table.
The keys to the cabinet exist/exists on the table.

where for simplicity we assert that the only possible verbs are: is/are (`be`) and exists/exist (`exist`). Let the model assign higher probability mass to the correct conjugation for the `be` pair but not for the `exist` pair (Table 1).

First, consider evaluating systematicity. To reflect how TSE chooses a small subset of the possible verbs for evaluation, in this toy example let it choose only `be`. Thus, the model scores 1 out of 1, whereas a test of systematicity should penalize the model for incorrectly conjugating `exist`.

Now, consider evaluating likely behavior. Let this same model generate either of the two correct conjugations (are/exist) with total probability of 0.7 and generate either of the incorrect conjugations with total probability 0.3. Thus, when we sample from the model, it generates a correct conjugation with probability 0.7, but TSE cannot measure this, since it gives a binary score to each verb pair.

The first of our proposed evaluations, **equally-weighted syntactic evaluation** (EW), addresses systematicity. To better approximate a model's ability to conjugate *any* verb, EW expands TSE to consider a much larger set of verbs than given in the templates used by prior work.

The second of our proposed evaluations, **model-weighted syntactic evaluation** (MW), addresses likely behavior. This method computes the probability mass that models put on producing the correct verb conjugation given a particular syntactic context. It rates the syntactic quality of samples—models need not conjugate *all* verbs, but instead be likely to generate *some* well-conjugated verb.

We conduct these evaluations on four pretrained language models using two template datasets: M&L (Marvin and Linzen, 2018) and BLiMP (Warstadt et al., 2020). Overall, we find that the EW scores are lower than the TSE scores, indicating that the verb choices in these templates overestimate models' systematicity with respect to subject-verb number agreement. This lack of systematicity is particularly apparent when we test verb lemmas that models find unlikely, with scores dropping by up to 40%. In contrast, the MW scores are high, suggesting that language models preferentially conjugate verbs they deem likely. Moreover, this ability improves when the tail of the distribution is truncated, as it is in decoding strategies like nucleus sampling (Holtzman et al., 2020).[1]

2 Methods

To define our metrics, we introduce some notation. TSE has two components: the model M to evaluate, and the set of templates T with interesting syntactic phenomena (e.g., from Marvin and Linzen (2018)). In S/V number agreement, each template contains a context c, including the subject that specifies the correct verb inflection; and the verb lemma ℓ with correct and incorrect inflections in the third person present tense (ℓ_+ and ℓ_-, respectively). M takes c and produces a distribution $P_M(\cdot \mid c)$ over its vocabulary, which we assume includes ℓ_+ and ℓ_-. We then compute a score for each template and average the scores over all templates to get a final score for M. The TSE score for a template can be expressed as:

$$\mathbb{1}\left[P_M(\ell_+ \mid c) > P_M(\ell_- \mid c)\right]. \quad (1)$$

The crux of our proposal is to use a large set of lemmas, $\mathcal{L}$, while drawing contexts c from a predefined set of templates T. We define two evaluation methods using $\mathcal{L}$:

Equally-Weighted (EW) Here we average (1) over all ℓ in $\mathcal{L}$, evaluating systematicity.

[1]Code available at `https://github.com/bnewm0609/refining-tse`

Model-Weighted (MW) Here we compute the total probability of generating a correct inflection conditioned on generating a lemma in $\mathcal{L}$:

$$\frac{\sum_{\ell \in \mathcal{L}} P_M(\ell_+ \mid c)}{\sum_{\ell \in \mathcal{L}} P_M(\ell_+ \mid c) + P_M(\ell_- \mid c)}, \quad (2)$$

evaluating likely behavior. See Table 1 for how these are computed in the toy example.

3 Experiments

Data We use S/V number agreement TSE templates from Marvin and Linzen (2018) and BLiMP (Warstadt et al., 2020) (for BLiMP we use the minimal pairs differing in verb, not subject). For our MW and EW evaluations, we only keep templates with unique contexts (i.e., templates not differing solely in verb lemma). We also ensure that all sentences start with a capital letter (for cased models) and end with a sentence-final period (for bidirectional models).

Our list of English verb lemmas contains 3,562 lemmas and is compiled by combining the top 1,000 most frequent verb lemmas from COCA, extracting all tokens with the `VB` part-of-speech tag in the Penn Treebank (1,951 lemmas), and scraping 3,250 lemmas from the Giant Verb List (Davies, 2008; Marcus et al., 1993; Essay, 2015). [2] Masked LMs may assign a different number of tokens to plural and singular forms of the same lemma, and they may not model joint probabilities over multiple tokens. To enable a fairer comparison between LMs and masked LMs, we only consider lemmas where both inflections are in the wordpiece vocabulary of the models. This choice leaves 980 lemmas for BERT cased, 1159 for BERT uncased, and 1265 for GPT2 and RoBERTa (so results are not comparable between models). This verbal variety situates our work between Gulordava et al. (2018)'s and Marvin and Linzen (2018)'s: our verbs can be infelicitous like the sentences in Gulordava et al. (2018), but our contexts are felicitous. See Section 5 for additional discussion.

Models We evaluate both bidirectional and unidirectional models, including BERT-large-uncased, BERT-large-cased, GPT2-XL, and RoBERTa-large (Devlin et al., 2018; Radford et al., 2019; Liu et al., 2019), all using the Huggingface Tranformers library (Wolf et al., 2020).

[2] The verb lemmas are accessible from the Appendix.

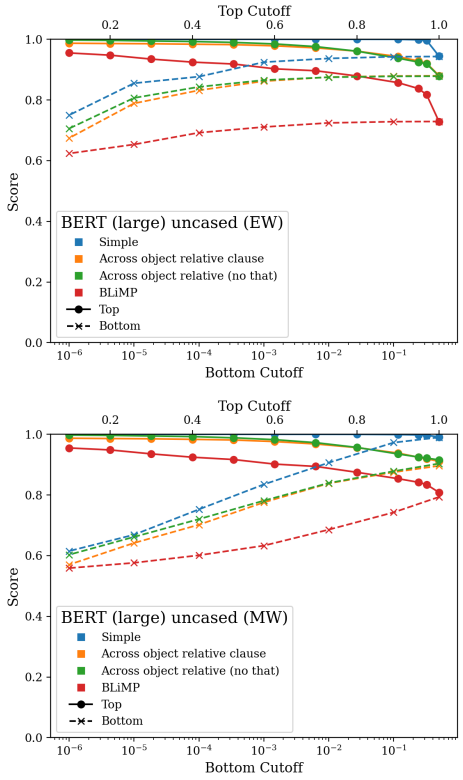


Figure 1: EW and MW scores as a function of Top p and Bottom p cutoffs for a subset of syntactic constructions (colors) using the BERT cased model.

To understand models' performances at the head and tail of their distributions, we additionally restrict $\mathcal{L}$ to the lemmas assigned high and low probabilities.

To consider the high-confidence lemmas, for each template in the dataset, we record the MW and EW scores computed using the inflections that fall into the top p percentile of the model's distribution. We choose $p \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 95, 97, 100\}$, noting that for each p, the distribution we use is the same as the one used by nucleus sampling (with a nucleus of size p).

Analogously, to focus on the low-confidence lemmas, we consider the lemmas where both inflections fall into the bottom p percentile of the model's distribution. Here, we choose $p \in \{50, 10, 1, 0.1, 0.01, 0.001, 0.0001\}$.[3]

4 Results

Our results can be found in Table 2. We find that EW scores are almost always lower than TSE

[3] At times, a cut-off lies within the probability mass on an inflection of interest. In these cases, we linearly interpolate between scores with and without the inflection included.

Templates	BERT cased			BERT uncased			RoBERTa			GPT2		
	MW	EW	TSE	MW	EW	TSE	MW	EW	TSE	MW	EW	TSE
Simple	0.99	0.94	**1.00**	0.98	0.90	**1.00**	0.98	0.93	**1.00**	0.90	0.86	**1.00**
In a sentential complement	0.92	0.67	**0.89**	0.92	0.60	**0.86**	0.92	0.67	**0.88**	0.96	0.65	**0.89**
VP coordination	0.91	0.89	**0.90**	0.93	**0.90**	**0.90**	0.93	0.90	**0.93**	0.89	0.87	**0.97**
Across prepositional phrase	0.91	0.83	**0.93**	0.83	0.75	**0.85**	0.87	0.83	**0.89**	0.84	0.76	**0.96**
Across subject relative clause	0.87	**0.84**	**0.84**	0.88	0.84	**0.85**	0.76	0.72	**0.80**	0.82	0.77	**0.97**
Across object relative clause	0.91	0.88	**0.91**	0.86	0.80	**0.85**	0.88	0.85	**0.91**	0.95	0.89	**0.99**
Across object relative (no that)	0.92	0.88	**0.90**	0.79	0.72	**0.81**	0.86	0.82	**0.89**	0.95	0.89	**0.99**
In object relative clause	0.93	0.95	**0.97**	0.95	0.97	**0.99**	0.89	0.91	**0.97**	0.91	0.88	**0.98**
In object relative (no that)	0.90	0.91	**0.92**	0.81	**0.82**	**0.82**	0.82	0.83	**0.90**	0.91	0.88	**0.97**
BLiMP	0.81	0.73	**0.90**	0.78	0.69	**0.85**	0.70	0.66	**0.78**	0.82	0.75	**0.91**

Table 2: MW, EW, and TSE evaluations on various models and syntactic constructions (See Warstadt et al. (2020); Marvin and Linzen (2018) for descriptions). MW is colored differently because its score is based directly on the model's probability mass, while EW and TSE are based on 0/1 judgements, so they are not directly comparable.

scores, indicating that TSE overestimates systematicity. On the other hand, higher MW scores reveal that sampling from the models is likely to result in correct conjugations.

A potential confounder for unidirectional LMs (GPT2) is that they only receive the left context and subject verb pairs sometimes look like noun phrases. For example, a sentence starting with *The officer* can be continued by *experiences joy* or by *experience is overwhelming*. This is not an issue when there are phrases or clauses between the subject and verb, and it may not occur for other English syntactic phenomena or in other languages.

To investigate the extent to which models perform well on likely lemmas and poorly on unlikely lemmas, we plot these scores for the top and bottom p percentiles in Figure 1. In general, the models perform better on lemmas that they assign high probability to in both evaluations.

For example, consider the BERT cased model assessed on object relative clause constructions. The MW plot illustrates that sampling from the top 60% of the distribution will produce a grammatical output with 97% probability, while sampling from the entire distribution only does so with 91% probability. The EW plot shows that the model attains a score under 80% when assessed on verbs in the bottom 0.001% of the model's probability mass, even though considering verbs in the top 90% of the model's probability mass would yield a score over 94%. These observations extend previous work on nucleus sampling, showing that cutting off the tails of the distribution generates more syntactically correct outputs (Holtzman et al., 2020).

There are two additional factors to keep in mind for these plots. First, the heads and tails of the distributions often contain very few lemmas eligible for use in score calculation. Second, models often assign probability mass to other lemma inflections (e.g. the past tense) that do not allow us to assess models' S/V number agreement ability. See the Appendix for related plots.

4.1 Qualitative Results

Earlier, we motivated MW with the consideration that the lemmas TSE uses might be unlikely, and therefore give an unrealistic depiction of models' likely syntactic behavior. Two examples where this happens and leads to a deceptively low score on a template for a model (here BERT-large-cased) are in Table 3.

In the first column, the lemma set used by TSE contains `like`, `hate`, and `love`, and the model puts more probability on *like* than *likes*, leading to a TSE score of 0.67. However, the most probable lemmas are `meet`, `encounter`, `see`, and `face`, all of which the model conjugates correctly.

In the second column, there is another example where the MW score rewards models for correct conjugations while TSE does not. Like the last example, the lemma set used by TSE contains `like`, `hate`, and `love`, and `like` is conjugated incorrectly. However, the more probable lemmas `pilot`, `control`, `employ`, `train`, `use`, `include`, `have`, `order`, `command`, and `feature` are all conjugated correctly.

5 Related Work

Evaluating Models Some previous work has focused on using minimal pairs to evaluate syntactic representations of models. Goldberg (2019) and Wolf (2019) assess the syntactic abilities of large transformers like BERT and GPT2, while Kuncoro et al. (2018), Tran et al. (2018) and Kim et al.

The senators that the skater [mask] are young.		The pilots that the executive [mask] are tall.	
meets	0.20	pilots	0.088
encounters	0.057	controls	0.059
sees	0.057	employs	0.025
meet	0.048	trains	0.023
encounter	0.023	uses	0.022
see	0.019	includes	0.019
##s	0.018	has	0.017
faces	0.013	orders	0.015
saw	0.012	commands	0.014
met	0.010	features	0.013

Table 3: Example sentences and the top 10 most probable subwords.

(2019) evaluate architectures designed to capture syntax (e.g., Ordered Neurons (Shen et al., 2019) and Recurrent Neural Network Grammars (Dyer et al., 2016)). In these cases, minimal pair evaluations should align with models' performance as language models, which is measured by our MW score.

Psycholinguistics Recent work has also applied experimental procedures from psycholinguistics to compare human and neural model language processing (Futrell et al., 2019). Experiments investigating garden path sentences' surprisal, S/V number agreement, and other specific syntactic phenomena reveal that models and humans have different patterns of errors and processing (Linzen and Leonard, 2018; Ettinger et al., 2018; Wilcox et al., 2020; van Schijndel and Linzen, 2020). Many of these phenomena are rare, so evaluations with templated minimal pairs complement perplexity as a metric for evaluating models' syntactic generalization (Hu et al., 2020). When measuring syntactic ability on arbitrary lemmas, our EW metric would be preferred.

Lexical Choice in Syntactic Evaluation Prior work also considered how the lexical items in minimal pairs affect the syntactic evaluation of models. Marvin and Linzen (2018) note that certain verbs are preferentially conjugated correctly (they observe RNNs conjugate `be` correctly more often than `swim`) and claim that this is due to unigram frequency of the verbs. Similarly, we observe that models succeed on our MW metric indicating that they correctly inflect verbs with high in-context probability under the model.

Relatedly, Yu et al. (2020) investigate the nouns used in TSE minimal pairs and find that language model performance at subject-verb number agreement is uncorrelated with unigram probability of the noun. We instead focus on model-estimated in-context probability of the verb in minimal pairs, finding that model performance increases with the model probability.

Finally, Gulordava et al. (2018) argue that the results of syntactic evaluations are influenced by semantic associations between tokens, so they remove these associations by substituting each token with a different randomly selected token with the same syntactic role. Although the resulting minimal pairs are infelicitous, models are still able to predict the correct inflection with above-chance accuracy. Our methods are similar in that some of the verbs in our evaluation set are infelicitous, however the contexts we use are semantically coherent. Rather than avoiding semantic effects by creating infelicitous contexts, we marginalize them out by using a large set of verb lemmas. This makes our metrics less stringent than those of Gulordava et al. (2018), but captures a potentially more realistic setting where we expect our models to perform systematically.

6 Conclusion

As neural models have proven successful at NLP tasks and as potential psycholinguistic models, we continue to refine our understanding of how and whether they capture human-like language faculties. TSE provides a useful framework to address this question, but its current implementation focuses on a limited group of hand-chosen verbs, so it inaccurately reflects models' syntactic generalization abilities. In response, we propose two minimal pair evaluations: equally-weighted and model-weighted syntactic evaluation. The first focuses on *systematicity* by expanding the set of verbs TSE considers, and illustrates that language models still struggle with S/V agreement for unlikely verbs. The second focuses on *likely behavior* by computing the probability of producing a correctly conjugated verb, and illustrates that despite systematic shortcomings, language models generate syntactically valid utterances with high probability. By introducing these metrics, we hope to arrive at a clearer picture of the syntactic abilities of language models.

7 Ethical Considerations

The metrics we propose have been developed specifically with corpora using Standard American English in order to evaluate models' abilities to understand Standard American English syntax. This focus means that models performing well under these evaluations may perform poorly in other English dialects, and they may not understand all syntactic systems, for example in other languages. Finally, our MW metric concerns itself with how models are likely to preform during generative processes (such as beam search or sampling). Performing well on this metric means models will be able to generate more human-like text which has potential downstream harms such as misinformation generation or other inauthentic behavior in situations where written language is the medium used for communication.

Acknowledgments

The authors would like to thank the reviewers for their helpful feedback, along with Tal Linzen, Chris Manning, Rishi Bommasani, Kawin Ethayarajh, Lisa Li, Nelson Liu, Yasuhide Miura, Aaron Mueller, and Tianyi Zhang for their invaluable comments and discussions. JH was supported by an NSF Graduate Research Fellowship under grant number DGE- 1656518, and by Two Sigma under their 2020 PhD Fellowship Program.

References

Mark Davies. 2008. *The corpus of contemporary American English*. BYE, Brigham Young University.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California. Association for Computational Linguistics.

Pattern Based Writing: Quick & Easy Essay. 2015. Giant verb list: 3,250 verbs plus spelling rules and irregular verbs marked.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Yoon Kim, Alexander Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, and Gábor Melis. 2019. Unsupervised recurrent neural network grammars. 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 1105–1117, Minneapolis, Minnesota. Association for Computational Linguistics.

Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. 2018. LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Melbourne, Australia. Association for Computational Linguistics.

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

Tal Linzen and Brian Leonard. 2018. Distinct patterns of syntactic agreement errors in recurrent networks and humans. In *CogSci 2018*, pages 690–695.

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

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

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.

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

Marten van Schijndel and Tal Linzen. 2020. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *PsyArXiv*.

Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. 2019. Ordered neurons: Integrating tree structures into recurrent neural networks. In *International Conference on Learning Representations*.

Ke Tran, Arianna Bisazza, and Christof Monz. 2018. The importance of being recurrent for modeling hierarchical structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4731–4736, Brussels, Belgium. Association for Computational Linguistics.

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

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, page 1707–1713.

Thomas Wolf. 2019. Some additional experiments extending the tech report "assessing BERTs syntactic abilities" by Yoav Goldberg. Technical report, Technical report.

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. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Charles Yu, Ryan Sie, Nicolas Tedeschi, and Leon Bergen. 2020. Word frequency does not predict grammatical knowledge in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4040–4054, Online. Association for Computational Linguistics.

A Additional Plots

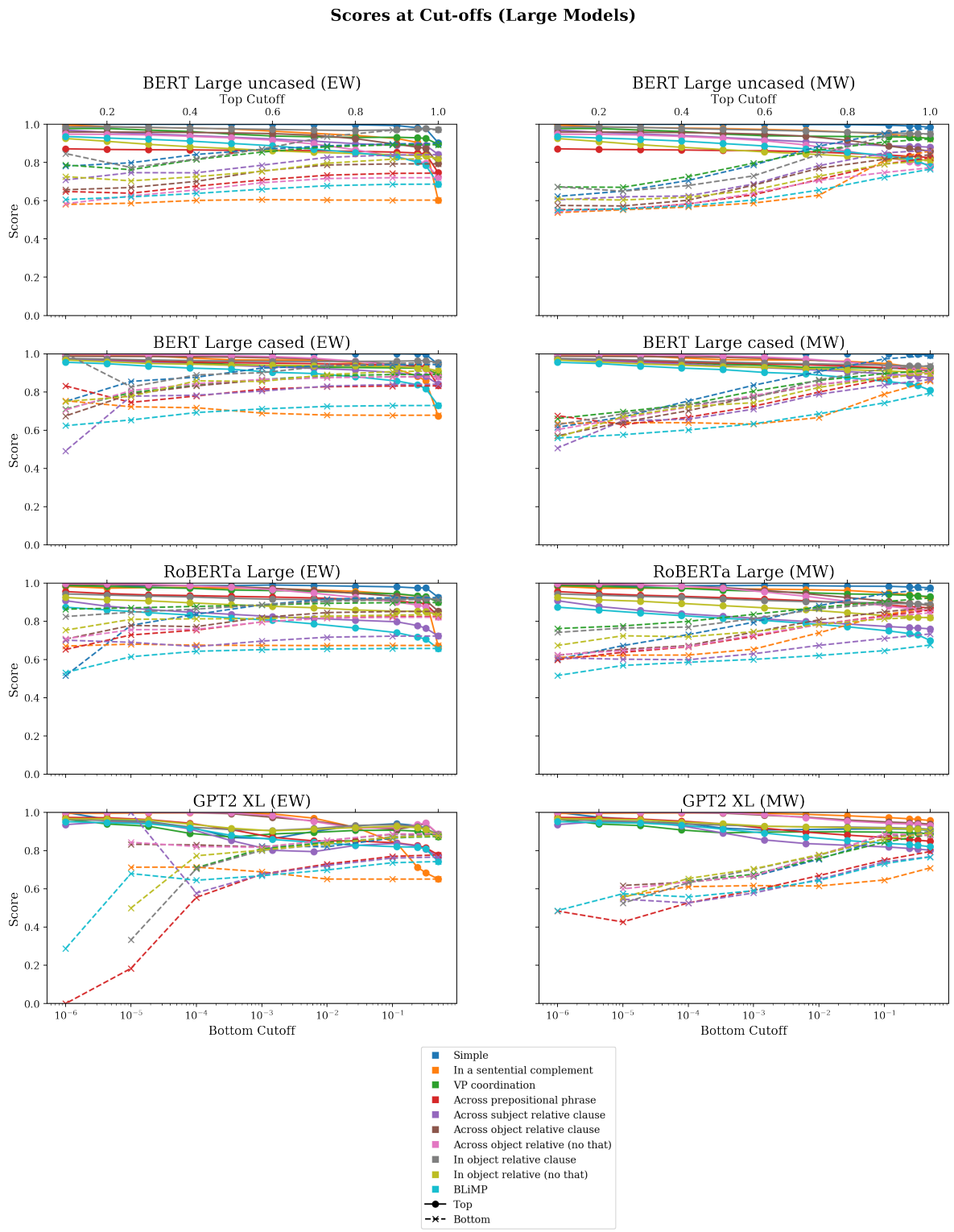


Figure 2: The plots above show the EW and MW scores as a function of Top-p and Bottom-p percentile cutoffs for BERT-base-uncased, BERT-large-cased, RoBERTa-large and GPT2-XL. In general, as the percentile increases, so does the score, though RoBERTa and BERT's EW scores are quite stable.

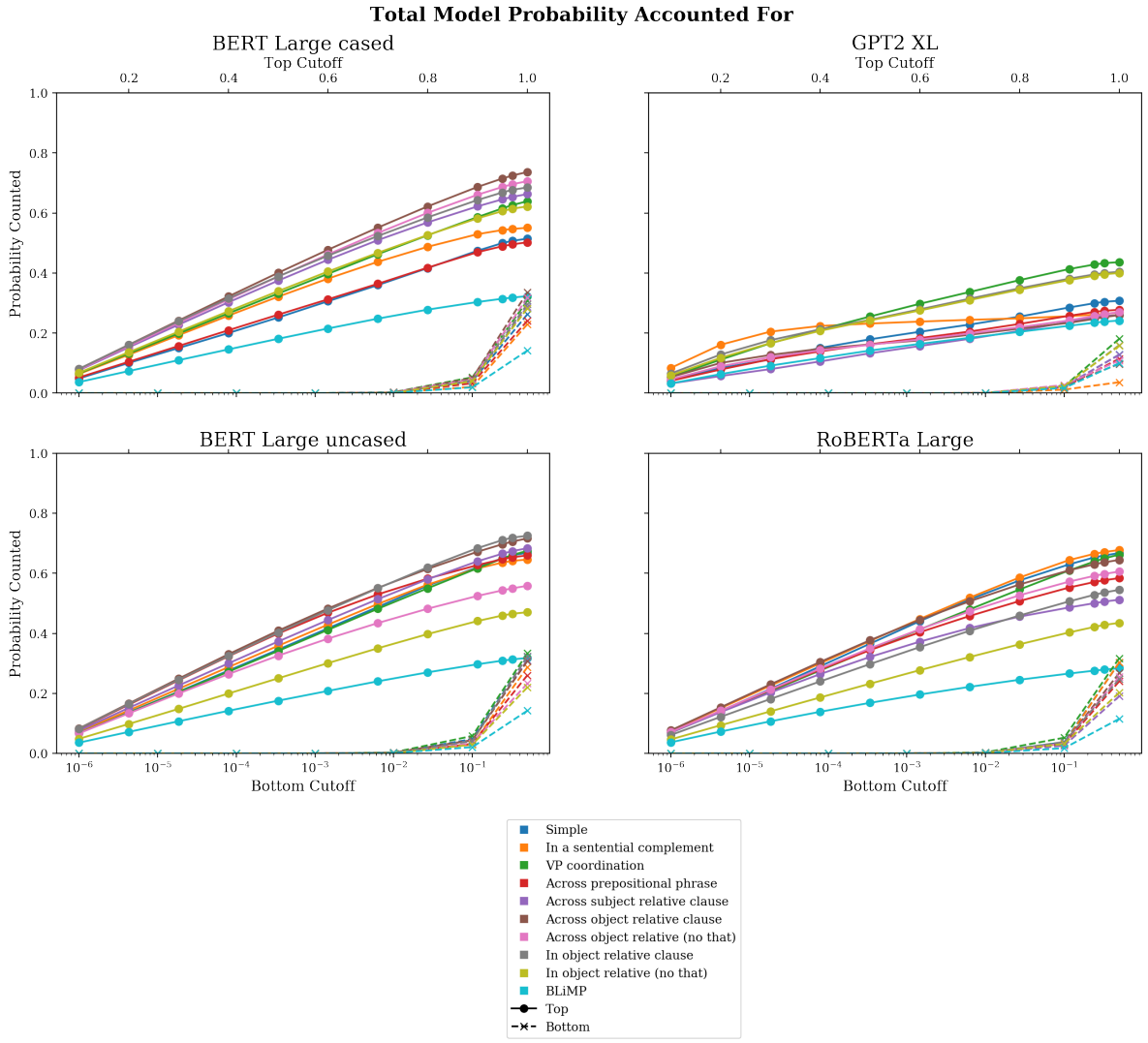


Figure 3: Above are plots of the proportion of the models' probability mass take up by the inflections of the verb lemmas we used. The y-axis is the total probability mass the lemmas account for and the x-axis is the percentile cutoff. Note that even when considering all of the lemmas, (at $p = 100\%$) there is probability mass not covered by our inflections. This probability mass is often put on other inflections of verbs (e.g. past-tense verbs) or other vocabulary items.

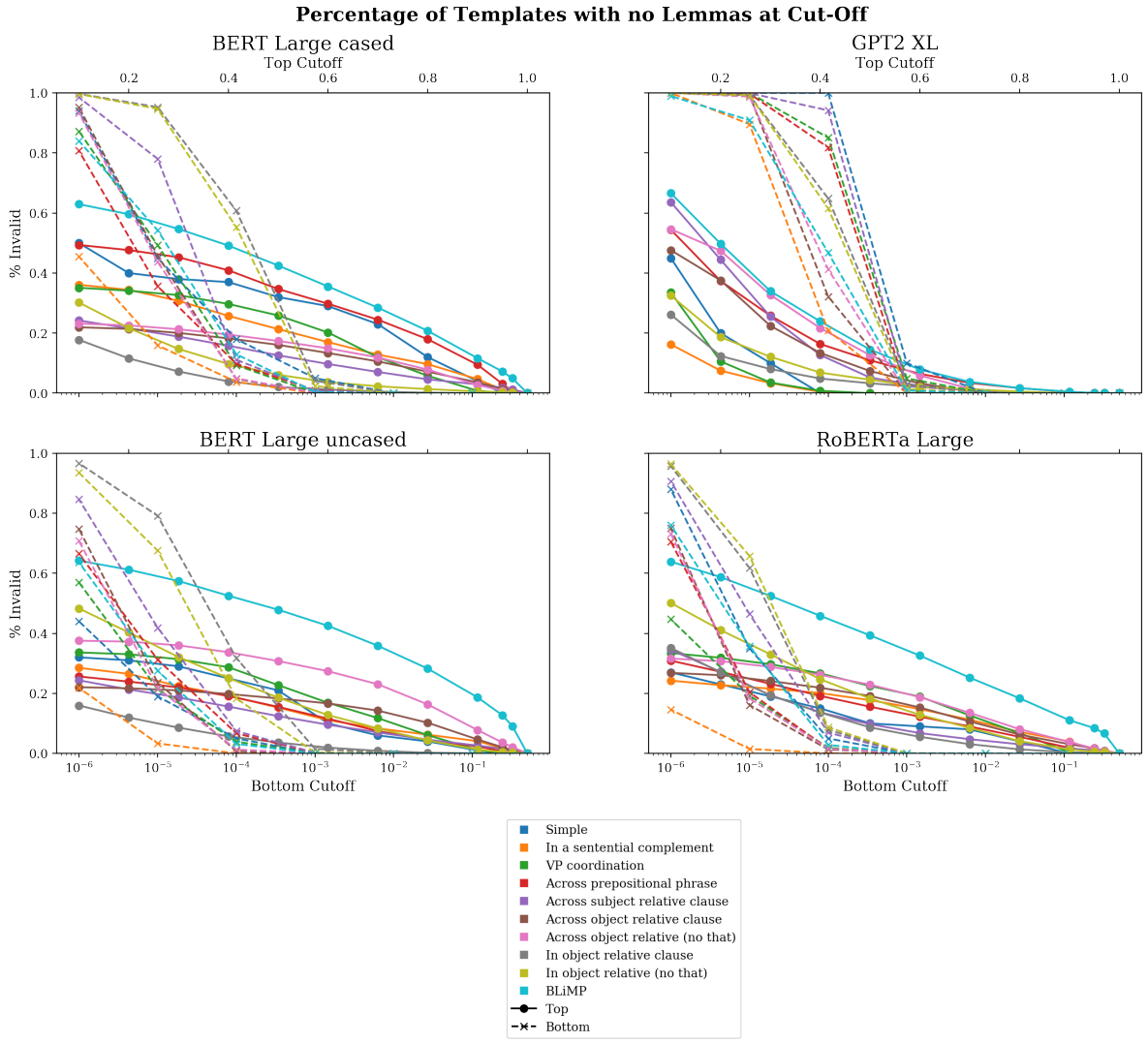


Figure 4: Above is the proportion of the templates in the datasets where models assign no probability mass to lemmas in the top or bottom $p\%$ of the their distributions. The y-axis is the proportion of lemmas that are rejected (i.e. values closer to one mean that the scores are calculated based on fewer templates). The x-axis is again the percentile cutoff. Note that the bottom-most cutoffs often has a large proportion of invalid lemmas, so these scores are based on fewer lemmas.

B A Subset of Verb Lemmas

The 1971 lemmas from COCA and The Penn Treebank we used are available below (Davies, 2008; Marcus et al., 1993):

abandon, abate, abide, abolish, absorb, abuse, accede, accelerate, accept, access, acclaim, accommodate, accomodate, accompany, accomplish, accord, account, accrue, accuse, achieve, acknowledge, acquire, acquit, act, adapt, add, address, adhere, adjust, administer, admit, adopt, adorn, advance, advantage, advise, affect, afford, aggravate, agonize, agree, aid, aim, air, alert, alienate, allay, allege, alleviate, allocate, allow, allowed, ally, alter, amalgamate, amass, amaze, amble, amend, amortize, amount, amplify, analyze, anchor, announce, answer, antagonize, anticipate, apologize, appeal, appear, appease, append, applaud, apply, appoint, appraise, appreciate, approach, approve, are, argue, arise, arm, arouse, arrange, arrest, arrive, articulate, ask, assassinate, assemble, assert, assess, assign, assimilate, assist, associate, assuage, assume, assure, atone, attach, attack, attain, attempt, attend, attention, attest, attract, attribute, auction, audit, audition, augment, authorize, autograph, average, avert, avoid, await, back, backfire, bail, balance, balk, balloon, ban, band, banish, bank, bankroll, bankrupt, bar, bargain, base, bash, batter, battle, bear, beat, become, bedevil, beef, begin, behave, belie, believe, bellow, belong, bend, benefit, bet, bid, bill, bite, blackmail, blame, blast, bleed, blend, bless, blink, block, blow, blunder, blunt, blur, board, boast, bode, bog, bolster, bomb, boom, boost, born, borrow, bother, bottle, bounce, bow, brace, brag, branch, brave, brazen, breach, break, breathe, breed, brew, bribe, bridge, brief, bring, broadcast, broaden, browbeat, brush, buck, buckle, budge, buffer, buffet, build, bumble, bump, buoy, burn, bury, butt, buttress, buy, buy-back, buzz, bypass, calculate, call, calm, cancel, cap, capitalize, capture, care, careen, caricature, carry, cash, cast, catapult, catch, cater, cause, caution, cease, cede, celebrate, cement, centralize, certify, challenge, change, channel, characterize, charge, chart, chase, chat, chauffeur, cheat, check, cheer, chew, chill, chisel, choke, choose, chop, churn, cinch, circulate, circumvent, cite, claim, clamp, clarify, clash, classify, clean, cleanse, clear, click, climb, cling, clip, clobber, close, cloud, clutter, coach, coax, code, co-exist, cohere, co-host, coincide, collaborate, collapse, collect, combat, combine, come, command, commemorate, commend, comment, commercialize, commit, communicate, compare, compel, compensate, compete, compile, complain, complement, complete, complicate, comply, compound, comprise, compromise, compute, computerize, con, conceal, concede, conceive, concentrate, concern, conclude, condemn, condone, conduct, confer, confirm, confiscate, conform, confront, confuse, congratulate, connect, connote, conquer, consent, conserve, consider, consist, console, consolidate, constitute, constrain, construct, construe, consult, consume, contact, contain, contemplate, contemporize, contend, content, continue, contract, contrast, contribute, control, convene, convert, convey, convict, convince, cook, cool, cooperate, coordinate, cope, co-produce, copy, corner, corral, correct, correspond, co-sponsor, cost, cough, count, countenance, counter, counteract, counterprogram, court, cover, crack, craft, crank, crash, crawl, creak, create, credit, crest, criminalize, crimp, criticize, crop, cross, crumble, crunch, crush, cry, cuff, curb, cure, curl, curry, curtail, cushion, cut, dabble, damage, damp, dampen, dance, dare, dash, date, deal, debate, debunk, debut, deceive, decide, declare, decline, decrease, deduct, default, defeat, defect, defend, defer, define, deflect, defraud, defuse, degenerate, delay, delegate, deliberate, delight, deliver, demand, demilitarize, demobilize, democratize, demolish, demonstrate, denude, deny, depart, depend, depict, deposit, depress, deprive, derail, deregulate, describe, desert, deserve, design, designate, desist, destabilize, destroy, detect, deter, deteriorate, determine, detract, develop, devise, devote, dial, dictate, die, differ, differentiate, dig, digest, dignify, dilute, diminish, dip, direct, disagree, disappear, disappoint, disarm, disassemble, disassociate, disband, discard, discern, disclose, discomfit, disconnect, discontinue, discount, discourage, discover, discredit, discuss, disdain, disengage, disguise, dish, dislike, dismantle, dismember, dismiss, disobey, disparage, dispatch, dispel, dispense, displace, display, dispose, disprove, dispute, disqualify, disregard, disrupt, dissipate, dissociate, dissolve, dissuade, distance, distinguish, distort, distract, distribute, disturb, diverge, diversify, divert, divest, divide, do, doctor, document, dole, dominate, don, donate, double, doubt, downsize, draft, drag, drain, drape, draw, dream, dress, drift, drill, drink, drive, drop, drown, drum, dry, dump, duplicate, dwarf, earmark, earn, ease, eat, eavesdrop, ebb, echo, eclipse, economize, edge, edit, educate, effect, elaborate, elect, eliminate, elon-

gate, emasculate, embark, embarrass, embellish, embrace, emerge, emote, empathize, emphasize, emphaticize, employ, empty, emulate, enable, enact, encapsulate, encompass, encounter, encourage, end, endanger, endeavor, endorse, endure, enforce, engage, engineer, enhance, enjoin, enjoy, enlarge, enlist, enroll, ensue, ensure, entail, enter, entice, entitle, entrench, entrust, envision, equal, equate, equip, eradicate, erase, erect, erode, erupt, escalate, escape, establish, estimate, evade, evaluate, evaporate, even, evolve, exacerbate, exaggerate, examine, exceed, except, exchange, exclude, excorciate, excuse, execute, exempt, exercise, exhaust, exhibit, exist, exit, exorcise, expand, expect, expedite, expel, experience, expire, explain, explode, exploit, explore, export, expose, express, expunge, extend, extinguish, extort, extract, extricate, fabricate, face, facilitate, factor, fade, fail, fall, falsify, familiarize, fantasize, fare, farm, fashion, favor, fear, feature, feed, feel, fend, ferret, ferry, fester, fetch, field, fight, figure, file, fill, finance, find, fine, fine-tune, finger, finish, fire, firm, fit, fix, flash, flatten, flaunt, flay, flee, flinch, flip, float, flood, flounder, flourish, flow, fluctuate, flush, fly, focus, fog, foil, fold, follow, fool, foot, force, forecast, foresee, forfeit, forge, forget, forgive, forgo, form, formulate, foster, frame, franchise, free, freeze, freight, fret, frighten, frolic, frustrate, fuck, fudge, fuel, fulfill, function, fund, funnel, furnish, further, gain, galvanize, gamble, garden, garner, gasp, gather, gauge, gender, generalize, generate, get, give, glamorize, glaze, glean, glide, gloat, gloss, glut, go, gon, gore, govern, grab, grace, grant, grapple, grasp, grimace, ground, group, grow, growth, guarantee, guard, guess, guide, gut, guzzle, hack, halt, halve, hammer, hamper, hand, handicap, handle, hang, happen, harass, harm, hasten, hate, haul, haunt, have, head, heal, hear, heat, hedge, heed, heighten, help, herald, hesitate, hide, highlight, hinder, hinge, hint, hire, hit, hoe, hold, holler, homer, hone, honor, hope, host, house, hum, hurry, hurt, identify, idle, ignite, ignore, illuminate, illustrate, imagine, impact, impair, impede, implement, implicate, imply, import, impose, impound, impress, imprison, improve, impugn, incarcerate, inch, include, incorporate, increase, increased, incur, indemnify, indicate, indict, induce, indulge, industrialize, infiltrate, inflame, inflate, inflict, influence, inform, infringe, infuriate, infuse, ingest, ingratiate, inhibit, initiate, inject, injure, innovate, insist, inspect, inspire, install, instill, institute, insulate, insure, integrate, intend, intensify, interconnect, interest, interfere, interpret, intervene, interview, intimate, introduce, invade, invent, invest, investigate, invite, invoke, involve, irk, iron, isolate, issue, itemize, jack, jeopardize, join, joke, jolt, judge, juggle, jump, junk, justify, keen, keep, key, kick, kidnap, kill, kiss, knock, know, kowtow, label, lack, lag, land, languish, lash, last, laugh, launch, launder, lay, lead, lean, leap, leapfrog, learn, lease, leave, lecture, legislate, legitimize, lend, lengthen, lessen, let, level, levy, liberalize, license, lie, lift, light, lighten, like, limit, line, linger, link, liquefy, liquidate, list, listen, live, load, loan, lobby, locate, lock, lodge, log, look, loom, loose, loosen, loot, lose, love, lower, lunch, lure, mail, maintain, make, man, manage, mandate maneuver, manipulate, manufacture, march, mark, market, marry, marvel, massage, master, match, materialize, matter, mature, maul, maximize, mean, measure, mediate, meet, meld, melt, mention, merge, merit, mesh, migrate, militate, milk, mimic, mince, mind, minimize, mirror, misinterpret, misrepresent, miss, mistreat, mitigate, mix, moan, mobilize, model, moderate, modernize, modify, mollify, monitor, monopolize, mop, mortgage, motivate, mount, move, mow, mull, multiply, muse, muster, mute, nail, name, narrow, navigate, naysay, need, negotiate, net, network, nod, nominate, normalize, notch, note, notice, notify, nudge, nullify, nurture, obfuscate, object, obscure, observe, obtain, obviate, occasion, occupy, occur, offend, offer, offset, omit, ooze, open, operate, oppose, opt, order, organize, originate, oust, outfit, outflank, outfly, outlast, outline, outpace, outperform, outsell, outshine, out-smart, outstrip, outtrade, outweigh, overbid, overcome, overemphasize, overhaul, overlook, overpower, overpurchase, overreact, override, overrule, oversee, overstate, overthrow, overturn, overwhelm, owe, own, pace, pack, package, paint, pair, pale, palm, pan, panic, parachute, parallel, parcel, park, parry, part, partake, participate, pass, patronize, pave, pay, peak, pedal, peddle, peer, penalize, penetrate, perform, permit, perpetuate, persist, persuade, peruse, petition, phase, pick, piece, pile, pin, pinch, ping, pinpoint, pit, pitch, placate, place, plague, plan, plant, play, plea, plead, please, plot, plow, pluck, plug, plummet, plunge, plur, pocket, point, police, polish, poll, pollinate, pollute, ponder, pop, popularize, populate, portend, portray, pose, position, possess, post, postpone, pot, pounce, pound, pour, practice, pray, precede, preclude, predict, predispose,

pre-empt, prefer, premiere, prepare, prepay, pre-register, presage, prescribe, present, preserve, press, pressure, pretend, pre-try, prevail, prevent, price, privatize, probe, proceed, process, proclaim, prod, produce, profile, profit, program, progress, prohibit, project, prolong, promise, promote, prompt, prop, propel, propose, prosper, protect, protest, prove, provide, provoke, prune, publicize, publish, pull, pummel, pump, punch, punish, purchase, pursue, push, put, puzzle, qualify, quantify, quarrel, quell, question, quiet, quit, quiz, quote, rage, raid, rain, raise, rally, ramp, range, rank, rat, ratify, rationalize, rattle, rave, reach, react, read, readmit, reaffirm, realign, realize, reap, reappraise, rearm, rearrange, reason, reassert, reassess, reassume, reassure, reauthorize, rebound, rebuild, rebut, recall, recapture, receive, reckon, reclaim, recognize, recommend, reconcile, reconnect, reconsider, reconstruct, record, recoup, recover, recraft, recreate, recycle, redden, redeem, redefine, redesign, redevelop, redial, rediscover, redo, redound, redraw, reduce, re-emerge, re-enter, reestablish, re-establish, re-evaluate, re-examine, refer, refinance, refine, reflect, refocus, refocuses, reform, refrain, refund, refurbish, refuse, refute, regain, regard, regenerate, register, regret, regroup, regulate, rehabilitate, reignite, reimburse, reimpose, rein, reinforce, reinstate, reinvent, reinvest, reinvigorate, reject, rejoin, rejuvenate, rekindle, relate, relaunch, relax, release, relieve, relinquish, relish, relocate, rely, remain, remake, remark, remedy, remember, remind, remove, renege, renegotiate, renew, renounce, renovate, rent, reopen, reorganize, repair, repatriate, repay, repeal, repeat, repel, replace, replaster, replenish, replicate, reply, repond, report, reposition, repossess, represent, reproduce, repurchase, request, require, reschedule, rescind, rescue, research, resell, resemble, resent, reserve, reset, reshape, reshuffle, reside, resign, resist, resolve, resort, respect, respond, rest, restart, restate, restore, restrain, restrict, restructure, result, resume, resurrect, retail, retain, rethink, retire, retreat, retrieve, retrofit, retry, return, reunite, revamp, reveal, reverberate, reverse, review, revise, revisit, revive, revoke, revolutionize, revolve, reward, rewrite, rid, ride, ring, rise, risk, rival, rock, roil, roll, roost, root, rotate, round, row, rub, rubber-stamp, ruin, rule, run, rush, sabotage, sacrifice, safeguard, safety, salvage, sap, satisfy, saturate, save, savor, say, scale, scan, scape, scare, schedule, school, scold, score, scorn, scour, scout, scrap, scream, screen, scrutinize, scurry, scuttle, seal, search, secure, seduce, see, seek, seem, segregate, seize, select, self-reinsure, sell, send, sense, sensitize, separate, sequester, serve, service, set, settle, sever, shadow, shake, shall, shape, share, sharpen, shave, shed, shell, shield, shift, shine, ship, shirk, shock, shoe-horn, shoot, shop, shore, shorn, short, shorten, shoulder, shout, shove, show, shower, shrink, shun, shut, shy, side, sidetrack, sift, sign, signal, signify, simplify, sing, sink, sit, ski, skid, skim, skip, skipper, slack, slam-dunk, slash, sleep, slide, slip, slog, slow, slump, smash, smell, smile, smoke, smooth, smother, smuggle, snatch, sniff, soak, soar, sob, socialize, sock, soften, solicit, solidify, solve, soothe, sort, sound, sour, sow, spare, spark, spawn, speak, specialize, specify, speculate, speed, spell, spend, spill, spin, split, sponsor, spot, spotlight, spread, spring, sprout, spruce, spur, spurn, spy, square, squeeze, stabilize, stack, staff, stage, stain, stall, stamp, stampede, stanch, stand, standardize, star, stare, start, starve, stash, state, staunch, stave, stay, steal, steer, stem, step, sterilize, stick, stifle, stimulate, stipulate, stir, stock, stockpile, stomach, stop, store, strafe, straighten, strain, stray, streamline, streetspeak, strengthen, stress, stretch, strike, strip, stroll, structure, struggle, study, stumble, subcontract, subject, sublet, submit, subordinate, subpoena, subscribe, subsidize, substantiate, substitute, subtract, subvert, succeed, suck, sue, suffer, suffice, suggest, suit, sum, summarize, summon, supersede, supervise, supplement, supply, support, suppose, suppress, surface, surge, surpass, surprise, surrender, surround, survey, survive, suspect, suspend, sustain, swallow, swamp, swap, sway, swear, sweat, sweeten, swell, swing, switch, synthesize, tackle, tag, take, talk, tamper, tandy, tap, target, tarnish, taste, tax, teach, team, tear, tell, tend, tender, term, terminate, test, test-drive, testify, thank, the, think, thrash, threaten, thrive, throw, thwart, tick, tie, tighten, tilt, time, tiptoe, tolerate, top, topple, torment, torpedo, torture, toss, total, totter, touch, tough, toughen, tour, tout, tower, trace, track, trade, traduce, trail, train, transact, transfer, transform, translate, transmit, transplant, transport, trash, travel, tread, treat, trend, trick, trickle, trigger, trim, triple, trivialize, trust, try, tumble, turn, twitch, unblock, uncover, undercut, undergo, underlie, underline, undermine, underperform, underpin, underscore, understand, undertake, underwrite, undo, undulate, unfold, unite, unleash, unload, unmask, unplug, unravel, unveil, unwind, update, upgrade, uphold, upset, urge, use,

usurp, utilize, vacate, vacillate, vacuum, value, vanish, vary, vault, veer, vent, venture, verify, veto, view, violate, visit, visualize, vitiate, voice, void, volunteer, vote, wad, wade, wage, wail, wait, waive, wake, walk, wall, wan, wander, wane, want, ward, warm, warn, warrant, wash, waste, watch, water, weaken, wear, weather, wedge, weigh, weight, welcome, were, whack, whip, widen, will, wimp, win, wind, wipe, wire, wish, withdraw, withhold, withstand, witness, wonder, woo, work, worry, worsen, wound, wrap, wreak, wreck, wrest, wrestle, wring, write, yank, yield, zero, zip, zoom

The rest of the lemmas from Essay (2015) are available here: https://patternbasedwriting.com/1/Giant-Verb-List-3250-Verbs.pdf.