Structural Persistence in Language Models: Priming as a Window into Abstract Language Representations

Arabella Sinclair[1,2*] **Jaap Jumelet**[2*] **Willem Zuidema**[2] **Raquel Fernández**[2]

[1]School of Natural and Computing Sciences
University of Aberdeen, United Kingdom

[2]Institute for Logic, Language and Computation
University of Amsterdam, The Netherlands

arabella.sinclair@abdn.ac.uk {j.w.d.jumelet|zuidema|raquel.fernandez}@uva.nl

Abstract

We investigate the extent to which modern neural language models are susceptible to structural priming, the phenomenon whereby the structure of a sentence makes the same structure more probable in a follow-up sentence. We explore how priming can be used to study the potential of these models to learn abstract structural information, which is a prerequisite for good performance on tasks that require natural language understanding skills. We introduce a novel metric and release PRIME-LM, a large corpus where we control for various linguistic factors that interact with priming strength. We find that Transformer models indeed show evidence of structural priming, but also that the generalizations they learned are to some extent modulated by semantic information. Our experiments also show that the representations acquired by the models may not only encode abstract sequential structure but involve certain level of hierarchical syntactic information. More generally, our study shows that the priming paradigm is a useful, additional tool for gaining insights into the capacities of language models and opens the door to future priming-based investigations that probe the model's internal states.[1]

1 Introduction

It has become increasingly clear that modern, neural language models (LMs) are capable of representing and learning a broad range of linguistic phenomena (Gulordava et al., 2018; Hewitt and Manning, 2019; Tenney et al., 2019a; Rogers et al., 2020; Warstadt et al., 2020). However, many open questions remain about the extent to which specific LMs have indeed acquired specific linguistic constructions, about whether these models encode an abstract notion of structure in their representations, and about the best ways to even assess the syntactic abilities of these models. A rich literature has emerged in the last few years addressing these questions, often taking inspiration from methodologies developed in theoretical linguistics, psycholinguistics, neurolinguistics, and language acquisition research (Futrell et al., 2019; Ettinger, 2020; Boleda, 2020; Gauthier et al., 2020; Baroni, 2022), where the same questions have been asked about the human mind/brain for centuries.

Building on this tradition, this paper turns to **structural priming** to investigate the degree to which LMs encode abstract structural information independent from the concrete words that make up sentences. This phenomenon refers to the fact that humans are more likely to produce—or to more easily comprehend—a sentence of a certain structure X (the *target*) when they have been exposed before to a sentence of a similar structure X (the *prime*), than if they had been prompted with a sentence of a different structure Y. For example, a native speaker of English will be more inclined to produce the target sentence with a prepositional object in (2-a) after having read sentence (1-a) instead of (1-a), and, vice versa, be more inclined to produce the double-object target sentence (2-b) after having read (1-b) instead of (1-a). Similar effects are also observed in language comprehension.

(1) a. *A teacher cooked a chicken for a worker*

b. *A teacher cooked a worker a chicken*

(2) a. *The guest threw the pot to the lady*

b. *The guest threw the lady the pot*

*Equal contribution.

[1]Our code and data can be found at https://github.com/dmg-illc/prime-lm.

Transactions of the Association for Computational Linguistics, vol. 10, pp. 1031–1050, 2022. https://doi.org/10.1162/tacl_a_00504
Action Editor: Chris Brew. Submission batch: 12/2021; Revision batch: 4/2022; Published 9/2022.

Evidence for structural priming—to the extent that it can be shown to be independent from lexical overlap and other confounds—is taken as evidence for a linguistic structural level of representation that abstracts away from the surface form of sentences. Thus whether or not language models display structural priming can provide insights as to their structural awareness, which is necessary for downstream tasks requiring natural language understanding skills. Previous experiments designed to test structural encoding in LMs are inconclusive. On the one hand, studies on structural probing (Hewitt and Manning, 2019) and on syntactic evaluation tasks (Warstadt et al., 2020) have yielded evidence for its presence. On the other hand, other sets of experiments have indicated that current LMs are surprisingly indifferent to word order (Hessel and Schofield, 2021; Pham et al., 2021; Sinha et al., 2021a) and rely on superficial heuristics when resolving downstream tasks (McCoy et al., 2019; Sinha et al., 2021b). Such unresolved tensions between results—and the active debate about them—highlights the need for developing additional methodologies that isolate structure from the lexico-semantic cues given to the model. In this paper, we leverage findings from structural priming in human language processing to develop a systematic experimental pipeline with the aim of assessing the extent to which pre-trained neural language models learn representations that encode structural information—a prerequisite for their good performance on natural language understanding tasks.

We use the term 'structural priming' (Pickering and Ferreira, 2008) rather than 'syntactic priming' (first described in Katryn Bock's *Syntactic Persistence in Language Production*, 1986) because it comprises priming of abstract structural information that is not restricted to syntactic hierarchical rules, such as the linear positions of semantic roles or the sequential order of parts of speech. In this paper, we focus mostly on the latter and touch upon syntactic rules in Section 7.4.

In Section 3, we define an efficient novel metric for measuring the effect of priming. For our experiments, we create **PRIME-LM**, a large-scale corpus for examining structural priming consisting of ∼1.3M prime-target sentence pairs, as we describe in Section 4. Earlier work on priming in LMs by Prasad et al. (2019) operationalized priming as adaptation or implicit learning and thus fine-tuned the model weights in between prime and target. While our priming effect metric is compatible with priming as adaptation, our experiments in this paper concentrate on priming after recent exposure to linguistic context without updating the model weights. This allows us to assess the structural representational abilities acquired by the models during training and investigate to what extent such structural information remains active at inference time.

In Section 6 and 7 we use our corpus and priming paradigm to answer three main research questions: (1) Are modern neural language models susceptible to structural priming? (2) Which factors influence the strength of the priming effect? (3) What is the nature of the structural representations acquired by those models? Our results show that Transformer language models *do* exhibit structural priming. This finding provides evidence that abstract structural information is encoded by the models to some degree and persists as a model makes predictions about upcoming sentences. The strength of the priming effect is influenced by several factors, including the semantic similarity and the proximity between prime and target, as well as the amount of exposure to a given structure during prompting. Our final experiment moreover reveals that the structural representations encoded by the model may not only be sequential but involve a certain level of hierarchical syntactic structure.

2 Background

2.1 Structural Priming in Humans

Priming is the dominant paradigm in psycholinguistics for investigating the extent to which human language processing involves a level of structural representation independent from other types of linguistic knowledge. The rationale behind this paradigm is that if speakers are sensitive to sentence structure independently from sentence content, then it is reasonable to assume that such structural information is an integral part of the representations built during processing.

In human language processing, structural priming effects are well attested both in comprehension and production (Bock, 1986; Pickering and Branigan, 1998; Bock and Griffin, 2000; Pickering and Ferreira, 2008; Goldwater et al., 2011; Pickering et al., 2013; Reitter and Moore, 2014; Tooley and Bock, 2014, among others). Several studies have shown that the strength of the

priming effect increases after repeated exposure to a given structure (Kaschak et al., 2011; Jaeger and Snider, 2013) and tends to decay if material intervenes between prime and target (Reitter et al., 2011). Other experiments have shown that ungrammatical and semantically incongruent sentences (e.g., *the waitress brunks the book to the monk*) lead to similar priming effects as well-formed sentences (Ivanova et al., 2012, 2017), which suggests that structural persistence effects are robust enough in the absence of semantic and lexical cues.

Yet, structural priming has been found to be affected by various aspects. For example, priming effects are stronger with lower-frequency than higher-frequency constructions (e.g., Scheepers, 2003; Bernolet and Hartsuiker, 2010; Pickering et al., 2013). Similarly, some types of lexical repetition between prime and target have been shown to enhance structural priming, suggesting that there is a lexical component involved (Pickering and Branigan, 1998; Cleland and Pickering, 2003). Semantic relatedness between prime and target also has a boosting effect, albeit smaller than the lexical repetition boost (Cleland and Pickering, 2003; Mahowald et al., 2016).

In the present study, we take inspiration from this tradition to investigate the priming behaviour of neural language models, which in turn depends on them encoding structural information. Two (not necessarily exclusive) mechanisms have been proposed to account for structural priming in humans: short-term residual activation of structural information across utterances (e.g., Branigan et al., 1999; Wheeldon and Smith, 2003) and long-term adaptation or implicit learning involving changes in the probability of a given structure (Bock et al., 2007; Kaschak et al., 2011; Fine and Jaeger, 2013). Here we focus on the ability of large pre-trained LMs to encode structural information given in the preceding context, similarly to residual activation in humans.

2.2 Structural Sensitivity of Neural LMs

The increasing capacities of neural language models in recent years have led to a surge in research into their representation of language on a fine-grained linguistic level (Alishahi et al., 2019; Tenney et al., 2019a; Rogers et al., 2020, inter alia). A common approach to examining language models is to consider them as '*psycholinguistic subjects*'; by testing hypotheses derived from psycholinguistics we are able to determine to what extent language models process language similarly to humans (Futrell et al., 2019; Ettinger, 2020; Davis and van Schijndel, 2020; Lakretz et al., 2021).

To assess the linguistic knowledge of LMs, a range of tools have been deployed. For instance, by training auxiliary diagnostic classifiers on top of a model's internal states (Hupkes et al., 2018), we can probe whether these states encode certain linguistic properties such as POS tags (Tenney et al., 2019b), syntactic dependencies (Hewitt and Manning, 2019; White et al., 2021), or constructional information (Madabushi et al., 2020; Li et al., 2022). Another common approach is the usage of Targeted Syntactic Evaluations, in which the LM's output probabilities are compared on a minimally different pair of a grammatical and ungrammatical sentence (Linzen et al., 2016; Marvin and Linzen, 2018; Gauthier et al., 2020; Hu et al., 2020). This procedure makes it possible to investigate a model's knowledge of specific linguistic phenomena without probing the model's internal representations, such as negative polarity items (Warstadt et al., 2019; Jumelet et al., 2021), subject-verb agreement (Gulordava et al., 2018; Lakretz et al., 2019), and argument binding (Warstadt et al., 2020).

Taken together, results from probing, Targeted Syntactic Evaluations, and other existing evaluation paradigms can certainly be viewed as providing converging evidence that modern neural LMs learn non-trivial structural, linguistic knowledge, and do not just memorize fragments of texts from the data and simple sequential dependencies. However, although converging, the evidence is not yet conclusive: Each of these evaluation paradigms has also been found to occasionally produce false positives. In probing, for instance, a well-known risk is that probes pick up information represented in the internal states of the language model, but not causally involved in the predictions of the model (Voita and Titov, 2020). In Targeted Syntactic Evaluations, the strength of the evidence depends on the quality of the set of alternative explanations that is considered, which ultimately is a matter of judgements and differs for different linguistic constructions (Vamvas and Sennrich, 2021). Recent studies have provided new challenges, including studies pointing out the indifference of LMs towards word order (Sinha

et al., 2021a, inter alia), their reliance on spurious heuristics (Lovering et al., 2021), and their difficulty in dealing with negation (Ettinger, 2020; Kassner and Schütze, 2020).

Hence, the debate about the abilities of language models to learn structural information in general, as well as their success in learning certain linguistic constructions specifically, is far from over. The research we present in this paper starts from the observation that structural priming may provide a much needed, complementary methodology that, like Targeted Syntactic Evaluations, examines the behavior of a model, but also, like probing, informs us about the nature of the internal states. We will assess a model's representation of a sentence by measuring its consequences in processing the next sentence. Instead of examining how the model deals with specific syntactic properties within a sentence, such as number agreement, we measure its encoding of abstract structure at the overall sentence level and the consequences this has for upcoming sentences. In the next section we explain our approach in detail.

3 Measuring Priming

We capture the effects of priming by measuring the difference in log probability of a target sentence T_{X} given a prime sentence P_{X} of the same syntactic structure X, vs. T_{X} given P_{Y}, a sentence of the exact same semantic and lexical content as P_{X} but differing in syntactic structure Y. We call this metric the ***Priming Effect*** *(PE)*:

$$_{\log}P(T_{\text{X}}|P_{\text{X}}) - _{\log}P(T_{\text{X}}|P_{\text{Y}}) \quad (1)$$

By measuring priming based on a fixed prime-target pair our method is akin to structural priming in comprehension. We condition a target sentence on a prime sentence by concatenating them, separated by a period. The log probability is computed as the sum of token log probabilities of the LM:

$$_{\log}P(T_{\text{X}}|P_{\text{X}}) = \sum_i {}_{\log}P_{\text{LM}}(T_{\text{X}_i}|P_{\text{X}}, T_{\text{X}_{<i}}) \quad (2)$$

For example, the PE of the example in the introduction would be computed as follows:

$$PE_{\text{PO}} = {}_{\log}P(T_{\text{PO}}|P_{\text{PO}}) - {}_{\log}P(T_{\text{PO}}|P_{\text{DO}})$$
$$PE_{\text{DO}} = {}_{\log}P(T_{\text{DO}}|P_{\text{DO}}) - {}_{\log}P(T_{\text{DO}}|P_{\text{PO}})$$

(where P_{PO}, P_{DO}, T_{PO}, T_{DO} denote sentences 1a, 1b, 2a, 2b). To ensure our estimates of the priming effect are robust, we incorporate the procedure of Newman et al. (2021) by pairing each target sentence in a corpus with 10 different prime sentences.

Definition 3.1 (*Priming Effect (PE)*)**.** Measures the effect of priming as the difference in log probabilities:

$$\frac{1}{|\mathcal{P}|} \sum_{P_{\text{X}} \in \mathcal{P}(T_{\text{X}})} [{}_{\log}P(T_{\text{X}}|P_{\text{X}}) - {}_{\log}P(T_{\text{X}}|P_{\text{Y}})]$$

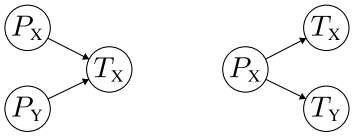
where $\mathcal{P}(T_{\text{X}})$ denotes the set of prime sentences that can be matched with target T_{X}. In our experiments, we report the mean of this metric, taken over large-scale corpora of semantically diverse sentences.

Our PE method is related to various other metrics that are used in the context of priming and statistics in general. When the conditional probabilities are close to 0—as is the case for our corpora with a mean sentence probability around 10^{-18}—this metric approaches the log odds ratio that is used by Mahowald et al. (2016). This allows our scores to be directly comparable to their results on human priming. A more general connection can be made between our metric and Bayes factors (Jeffreys, 1961; Kass and Raftery, 1995), which determine the strength of evidence and are, similar to our metric, also defined as a log probability ratio.

Prasad et al. (2019) model priming as an implicit learning procedure (Chang et al., 2000), instantiated as a fine-tuning-based adaptation process (van Schijndel and Linzen, 2018). The adaptation effect is then obtained by comparing the impact of a single prime structure on two target sentences of opposing structure, comparing their perplexity before and after fine-tuning:

$$PP(T_{\text{X}}) - PP(T_{\text{X}}|P_{\text{X}}) > PP(T_{\text{Y}}) - PP(T_{\text{Y}}|P_{\text{X}})$$

The authors also identify a problem: This metric is proportional to the prior perplexities $PP(T_{\text{X}})$ and $PP(T_{\text{Y}})$. They resolve the issue by regressing out this relationship. This procedure, however, is based on assumptions that do not always hold, namely, that the relationship between the priming metric and the prior perplexities of the two targets is linear and homoscedastic. In our experiments



Priming Effect (Eq. 1) Prasad et al. (2019)

Figure 1: Our Priming Effect metric compares the impact of two prime sentences with different structures on a single target exhibiting one of the structures. Prasad et al. (2019) examine the impact of a single prime structure on two target sentences.

we found neither assumption to hold empirically, and hence we opted to directly compare the impact of two prime sentences on a single target sentence. This way we do not need to regress out confounding effects of prior probabilities, since we are comparing the same quantity (the target sentence) to two primes. The contrast between these metrics is illustrated by the diagrams in Figure 1.

Note that our PE metric could be applied to the priming-as-adaptation paradigm as well, by comparing the target sentence probabilities of two fine-tuned models. In the experiments presented in this paper, we focus on priming as residual activation and thus do not update the model weights, which makes the approach more computationally efficient.

4 The PRIME-LM Corpus

We create a large-scale set of corpora designed to examine the priming behavior of LMs.

4.1 Syntactic Alternations

In the current experiments, we focus on two types of syntactic alternations, *dative* and *transitive*, which allow for the same content to be expressed by two different structures. The dative alternation includes ditransitive verbs whose complements can be expressed by a double-object (DO) structure or a prepositional-object (PO) structure (e.g., *the boss gave the dog a bone* vs. *gave a bone to the dog*). The transitive alternation includes transitive verbs within an active (ACT) or a passive (PASS) structure (e.g., *the actor followed the student* vs. *the student was followed by the actor*).

(3) **Dative**
DO: Dt N_{agent} V Dt $N_{recipient}$ Dt $N_{patient}$
PO: Dt N_{agent} V Dt $N_{patient}$ Pr Dt $N_{recipient}$

(4) **Transitive**
ACT: Dt N_{agent} V Dt $N_{patient}$
PASS: Dt $N_{patient}$ Aux V by Dt N_{agent}

In the transitive case, the active structure is dominant in English (Bock, 1986; Merriam-Webster, 1989). The proportion of use between structures for the dative alternation is less marked, with different studies showing a preference for the direct-object structure (e.g., Bock, 1986; Bresnan et al., 2007).

4.2 Corpus Construction

We construct a set of corpora by filling in the templates in (3) and (4) above. For the content words (nouns and verbs), we exploit the vocabulary present in the University of South Florida (USF) free association norms dataset (Nelson et al., 2004), which contains pairs of cue-target words with their association strength.[1] This allows us to control for the degree of semantic association between prime and target sentences. To minimize any effects stemming from word frequency factors, we only include USF content words that appear in the top 5000 most common words according to the COCA corpus (Davies, 2009).

We identify transitive and ditransitive verbs using vocabulary lists targeted at English language learners,[2] keeping those that are present in USF and meet the frequency constraints (around 80 verbs in total). The ditransitive verbs were manually labeled for the preposition to be used in the PO structure (*to/for*) and the transitive verbs were annotated with their past participle form to be used in the passive construction. In addition, all verbs were manually labeled for some of the noun categories they can take as arguments (e.g., the transitive verb *wash* was annotated as accepting agents of category `person` and patients of category `person` or `object`). Following the same frequency constraints, a set of nouns fulfilling these categories was selected from USF using the WordNet closure categories of *person*, *social_group, social_control, institution*, *physical_entity*, and *object*, which we further hand split into *non-edible*, *edible*, and *drinkable*.[3] This yielded 119 nouns in total.

[1] Corresponding to the percentage of human participants who produced the target word when asked to come up with words related to the cue (`http://w3.usf.edu/FreeAssociation/`).

[2] `http://www.aprendeinglesenleganes.com/resources`, `https://englishpost.org/transitive-verbs-list`, and `https://www.cse.unsw.edu.au/~billw/ditransitive.html`.

[3] To ensure compatibility with the indefinite article *a/an* (see Section 4.3), uncountable nouns were discarded.

From this vocabulary, we are able to generate many realizations of our sentence templates through sampling, respecting the grammaticality of the sentences produced. Three native speakers of English manually examined a subset of sentences for each verb and syntactic alternation to confirm that the sentences produced are well formed. This resulted in the elimination of a few ditransitive verbs for which the DO structure was considered awkward. The final corpus contains 48 transitive and 16 ditransitive verbs.

Using this template-based method, we create a series of corpora that satisfy various semantic and lexical constraints. For each of these corpora we specify a corpus size of 15,000 prime-target pairs per syntactic target structure (DO, PO, ACT, PASS), which are obtained by pairing 1,500 different target sentences with 10 semantically different primes.[4] Overall, PRIME-LM contains ∼1.3M prime-target pairs.

4.3 The *Core* Corpus

PRIME-LM consists of a *core* corpus and a set of variants over this core. In the *core* corpus, we ensure that prime and target sentences (1) include different determiners, either *a/an* or *the*, (2) do not share any nouns nor verbs, and (3) only contain nouns and verbs that are not semantically associated across prime and target according to the USF free association norms dataset.[5] For the PO structure, we additionally make sure that prime and target differ in preposition (*to* vs. *for*), which makes all the prime and target sentences in the dative alternation lexically fully disjoint. For the transitive alternation, this is not possible because the preposition *by* must appear in the PASS structure. Other than that, we completely limit lexical overlap for transitive constructions by using alternate auxiliary verb forms (*is* vs. *was*) for the passive prime and target, and create their active counterparts by using the corresponding tense of the auxiliary to maintain semantic equivalence. All sentences in the dative alternation are in the past simple tense.

As an illustration, below we show two examples from the *core* corpus following the scheme in Figure 1, where P are the prime sentences and T the target:

(5) P_{PO}: *A pilot bought a pie for an attorney*
P_{DO}: *A pilot bought an attorney a pie*
T_{PO}: *The professor sent the tea to the band*

(6) P_{ACT}: *The nurse purchased the beer*
P_{PASS}: *The beer was purchased by the nurse*
T_{PASS}: *An engine is wrapped by a colonel*

We create different variants of the *core* corpus that isolate specific aspects shown to influence structural priming in human sentence processing. They are described in Section 7 together with the corresponding experiments. Example sentences for each of our corpora can be found in Table 1.

5 Language Models

We focus our experiments on the class of *auto-regressive* LMs,[6] which are trained to predict the next token, in line with human incremental language processing. Our methodology can be applied to masked LMs as well; we briefly reflect on this in the discussion (§8). The main focus of our analysis is directed on Transformer models (Vaswani et al., 2017), which constitute the current state of the art in language modeling, and have been shown to produce representations that correlate strongly with human brain signals (Schrimpf et al., 2020).

This is the set of models we consider:

- **GPT2**, in its four sizes (SMALL, MEDIUM, LARGE, XL; Radford et al., 2019), and its *distilled* version (Sanh et al., 2019);
- **DialoGPT**, three GPT2 models of increasing size that have been fine-tuned on dialogue data (Zhang et al., 2020);
- **GPT-Neo** in three sizes (125M, 1.3B, 2.7B; Black et al., 2021), which is based on GPT3 (Brown et al., 2020).

All Transformer LMs are imported with the `transformers` library (Wolf et al., 2020). The extraction of the model probabilities is done using the `diagNNose` library (Jumelet, 2020),

[4] The corpus size of 15,000 was determined based on Cochran's Formula for sample size determination (Cochran, 1977), with a p-value and margin of error of 0.01.

[5] The average cosine similarity across pairs of words in prime and target computed with `word2vec` embeddings by Fares et al. (2017) is 0.2 for both nouns and verbs.

[6] Also known as *causal* or *left-to-right* language models, predicting the probability of the next token solely on prior context.

which provides support for efficient activation extraction. Our implementation allows our priming procedure to be efficiently tested on any kind of language model and to be easily reproducible. All our code and corpora are available at `https://github.com/dmg-illc/prime-lm`.

Why Should LMs Exhibit Structural Priming? Since structural repetition is present in human language use and common in corpora (Dubey et al., 2008), LMs have, in theory, the potential to learn such structural dependencies during training. It is not reasonable to expect that models which have been trained on shuffled sentences will exhibit priming, however, because such models will not be able to adequately carry over a linguistic signal (structural or otherwise) from the prime sentence to the target.[7] As mentioned in the Introduction and in Section 2.2, several studies have suggested that structural information is being encoded by large language models; yet, other studies showing that LMs are often insensitive to permutations in word order (e.g., Kodner and Gupta, 2020; Sinha et al., 2021b) cast doubt on these results. Thus, while there is potential for LMs pre-trained on unshuffled data to encode structural dependencies that are detectable with our priming paradigm, whether they will in fact do so remains an open question, since the language modeling objective (next word prediction) contains no explicit cues for structural information. This is precisely the question we address in this work.

Priming Behavior To interpret our results we distinguish between three types of behaviour: (i) *symmetrical priming* occurs when a model obtains positive PEs for both constructions within an alternation: The model has fully picked up on the structural congruence between prime and target; (ii) *asymmetrical priming* occurs when a model obtains a positive PE for one construction, and a PE close to zero for its counterpart;[8] and (iii) *biased priming* occurs when a model obtains a positive PE for one construction, but a negative PE for its counterpart. A priming bias indicates that a prime of the preferred structure is more likely

[7]In our experiments we had initially incorporated two LSTM LMs (Józefowicz et al., 2016; Gulordava et al., 2018), and indeed due to their shuffled training corpus we did not observe any notable PE. We are not aware of any available LSTM LM trained on unshuffled data.

[8]Such asymmetries are common in humans (Bock, 1986; Gries, 2005; Segaert et al., 2016).

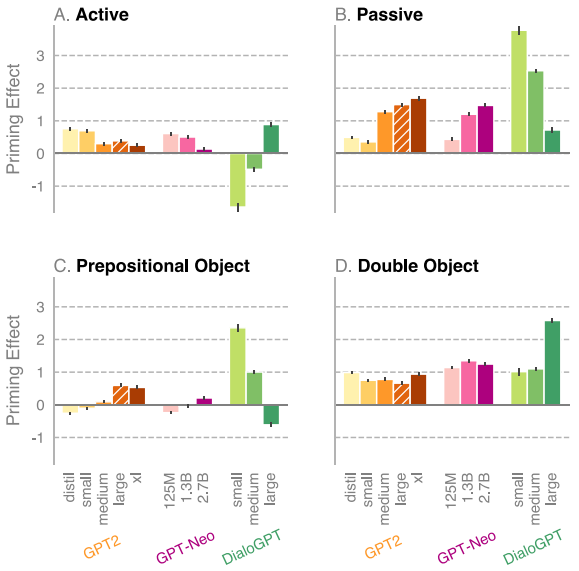


Figure 2: *Priming Effect* results of all models on the *core* corpus, across the four syntactic structures. Error bars denote 99% confidence intervals of the mean. The GPT2-LARGE model that will be explored in more detail in §7 has been highlighted.

to boost any subsequent target that we consider, regardless of its structural congruence with the prime. Hence, we take symmetrical and, to some extent, asymmetrical priming behavior to represent evidence for the structural priming effect we are interested in.[9]

6 Core Priming Results across LMs

We initially test all LMs described in the previous section on our *core* corpus, designed to control for lexical overlap and semantic similarity. This provides a clean experimental setup, where the only element shared between prime and target is the abstract sequential structure. The results are reported in Figure 2, split by the structure type of the target sentence. It can be seen that across many models a positive PE is present. We will now discuss these results in more detail.

There are two models that exhibit symmetrical priming for both transitive and dative alternations: GPT2-LARGE and GPT2-XL. The other GPT2 models exhibit symmetrical priming for transitive as well, but exhibit moderate asymmetrical priming

[9]This is analogical to, for example, subject-verb agreement: A model that always prefers a plural verb, regardless of the subject number, can't be said to understand the task. A model that scores 100% on plural verb prediction, but randomly for singular verbs, has an *asymmetric* understanding of the task.

Corpus	Condition	Prime (ACT)	Target (PASS)
Core	—	*The boy judged the adult.*	*A cousin is forgotten by a secretary.*
Semantic Similarity	Verb Only	*The chief **struck** the mayor.*	*A bishop was **beaten** by a hero.*
	All Nouns	*An **actor** broke a **glass**.*	*The **bottle** was wrapped by the **actress**.*
	All Words	*The **student drank** the **wine**.*	*A **beer** was **prepared** by a **professor**.*
Lexical Overlap	Random Noun	*The girl smelled the **chicken**.*	*A **chicken** was prepared by a pilot.*
	Main Verb	*A woman **used** a computer.*	*The iron was **used** by the father.*
	Function Words	***The** soldier wanted **the** pie.*	***The** book was carried by **the** manager.*
	All Nouns	*The **king** smelled the **wine**.*	*A **wine** was drunk by a **king**.*
Implausible Prime	—	*The **newspaper grabbed** the **pot**.*	*A key is removed by an attorney.*
Structural Complexity	Prime Complex	*A lady **with a red bag** chased a minister.*	*The juice was purchased by the child.*
	Target Complex	*The physician judged the leader.*	*A **rich** school was embraced by a business.*
	Both Complex	*The **bad** adult **with the hat** raised the knife.*	*A son was helped by an author **from Cuba**.*

Table 1: Example sentences for the core corpus and each condition described in §7.1, §7.2, and §7.4. The same manipulations illustrated here for the ACT and PASS also hold for the dative alternation.

behavior for dative, with priming occurring only for double-object structure. DialoGPT-SMALL exhibits biased priming for transitive constructions: a negative PE on active constructions, but a large positive PE for passive constructions. This shows that for this particular model a passive prime boosts the probability of an active target more than an active prime does, resulting in a negative effect.

Model Size We can investigate the impact of model size by comparing the results of the different sizes of the models we consider.[10] Larger models may have more potential for encoding finer-grained structural information (see, e.g., Hu et al., 2020). If model size were to have a positive effect on structural priming this might manifest itself in two ways: either (1) the PE increases for both structural alternatives, or (2) the priming bias towards one structure decreases. We do not see evidence of (1). As for (2) regarding bias, results differ between transitive and dative. For the GPT2 models the asymmetrical priming towards double objects is decreased, resulting in symmetrical priming for both GPT2-LARGE and GPT2-XL. For the DialoGPT results on transitive we can see that the severe bias towards passive decreases as model size is increased, resulting in symmetrical priming behaviour for DialoGPT-LARGE. For dative constructions, however, the larger model size gives rise to a priming bias towards double objects: in this case increasing model size actually has a detrimental effect on the model's priming behaviour. From this we conclude that sensitivity to structural priming is partly driven by model size, but is likely to depend on a more intricate combination of factors related to model architecture and training data, which needs to be investigated further in future work.

Best Model The models that exhibit more susceptibility to structural priming across all four construction types are GPT2-LARGE and GPT-2-XL. For GPT2-LARGE the congruent conditional probability $P(T_X|P_X)$ was larger than the incongruent one $P(T_X|P_Y)$ 60.5% of the time for active, 81.0% for passive, 65.4% for prepositional object, and 72.1% for double object. In the subsequent experiments we will focus our analysis on GPT2-LARGE and use more specialized experimental conditions within the priming paradigm to dig deeper into the potential of the model for encoding structural information.

7 Impact of Specific Factors

The next battery of experiments isolates various factors that have been shown to be of influence to priming in human language processing. For each experimental condition, we present a specialized corpus followed by an analysis of the priming effects exhibited by GPT2-LARGE on this data, comparing them to the model's behavior on the core corpus. Examples from the *core* and specialized conditions can be found in Table 1.

[10] Note that the different sizes of a model are trained on the same amount of data; only the number of parameters is affected.

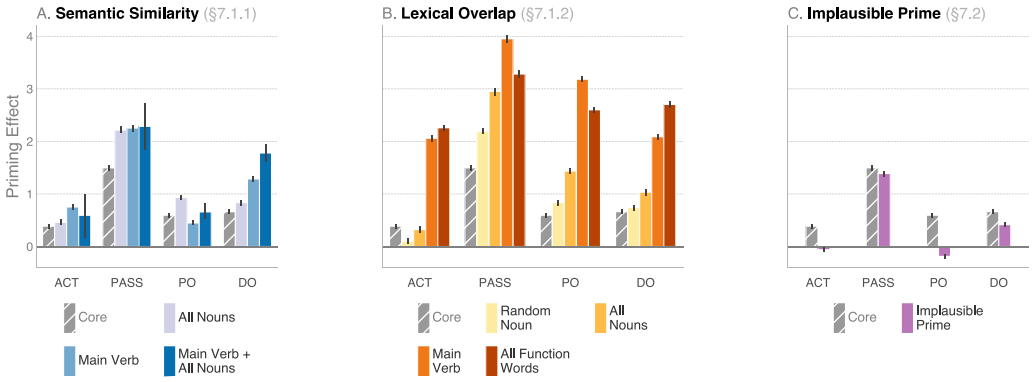


Figure 3: Results for GPT2-LARGE on the experiments described in and §7.1 and §7.2: **A.** measures the impact of semantic similarity between prime and target, **B.** the impact of lexical overlap between prime and target, and **C.** whether priming is affected by the semantic implausibility of the prime.

7.1 Lexical Dependence

In the *core* corpus, prime and target sentences are semantically unrelated, which ensures that priming effects cannot stem from the model assigning higher probabilities to words that are similar or identical to those present in the prime sentence. In the following two experiments we relax this constraint to investigate the extent to which lexical semantic similarity and lexical repetition across prime and target have an impact on structural priming effects.

7.1.1 Semantic Similarity

We create versions of the *core* corpus where prime and target sentences have different degrees of lexical semantic similarity. Concretely, a pair of words sharing the same semantic role in the prime and target is considered semantically similar if they (a) are associated according to the USF norms, and (b) have a cosine similarity (computed with embeddings from Fares et al., 2017) equal or higher than the 90%-percentile of the distribution of similarities in the *core* corpus.[11]

In human experiments, semantic similarity has been found to boost priming (Goldwater et al., 2011), both in nouns (Cleland and Pickering, 2003) and in verbs (Pickering and Branigan, 1998). We isolate the effects of verb and noun similarity by creating conditions where (1) only the verb, (2) all nouns, or (3) all content words are semantically similar across prime and target sentences. These additional constraints result in a more limited set of possible sentence pairs for condition (3), and thus in a reduced corpus of 228 (transitive) and 1648 (dative) prime-target pairs rather then 15,000.[12]

Results We find greater PE across constructions in this setup compared to the core corpus, although this is less pronounced for the PO structure. As can be seen in Figure 3A, a semantically similar verb in prime and target leads to an increase of the PE, comparable to the condition where all nouns are similar. With the exception of DO, we do not observe an additive effect: When all content words are similar, the PE is not substantially higher than when only the verb is similar.

7.1.2 Lexical Overlap

Lexical overlap between prime and target in the core corpus was avoided in both content and function words. Here we systematically introduce lexical repetition across prime and target sentences. We create versions of the core corpus where lexical overlap takes place with respect to only (1) one of the nouns at random but with the same semantic role across prime and target (*agent, patient, recipient*, see §4.1), (2) all nouns, (3) the verb, and (4) all function words (i.e., any determiners, prepositions, and auxiliary verbs are shared across prime and target, without content words being shared).

Results As can be seen in Figure 3B, overall the presence of lexical overlap greatly boosts structural priming effects. For all constructions, verb overlap leads to higher priming effects than repeating one noun or even all nouns. Surprisingly,

[11] This results in a cosine similarity threshold of ∼0.4.

[12] In this case, to maximize the number of unique pairs, we allow a varying number of primes to target, rather than observing the 10-to-1 prime-target setup of the other corpora.

overlap of function words has the highest boosting effect for ACT and DO.[13] To place these results into context, we calculate the PE when prime and target are identical sentences. Language models are known to fall prone to repeatedly generating the same sentence (Foster and White, 2007; Fu et al., 2021); hence this value can be considered a ceiling. We obtain a PE of 2.5 for ACT, 7.2 for PASS, 9.2 for PO, and 10.1 for DO constructions. None of the lexical overlap conditions we consider reaches the corresponding ceiling.

7.2 Semantic Implausibility

In this experiment, we test whether the effects found in the *core* corpus are robust to manipulations concerned with the semantic plausibility of the sentences used as stimuli. This helps to diagnose to what extent any structural information encoded by the model is autonomous from semantics. To this end, we construct a version of the corpus where the prime sentences are specifically designed to be semantically implausible. Gulordava et al. (2018) used a similar method in their study of long-distance agreement dependencies, finding that RNN's ability to predict number agreement was robust to nonsensical sentences. The authors interpret this result as evidence that the networks track abstract structure, in line with Chomsky's (1957) proposal that grammaticality is distinct from meaningfulness in the human language faculty. Here we further test this hypothesis by analyzing whether the LM is susceptible to structural priming effects when the prime sentence is nonsensical. As mentioned in §2.1, humans do exhibit structural priming effects when prompted with incongruent sentences (Ivanova et al., 2012, 2017). We construct semantically implausible primes via sampling nouns at random among noun categories that do not respect the verb selectional restrictions. This results in grammatically correct, yet nonsensical sentences such as *'the iron threw the hero to the chocolate'*. The same constraints regarding absence of semantic similarity and lexical overlap between prime and target present in the core corpus apply here as well.

Results The results of this experiment are shown in Figure 3C. We find here that the PE exhibits asymmetrical priming behavior, indicating that the prime structure itself is more likely to boost any subsequent target regardless of shared structural properties. The PE disappears and becomes negative for the ACT and PO constructions, while for PASS and DO it decreases when compared to the results on the core corpus, but remains positive. While some degree of abstract structural information present in the nonsensical sentences may be exploited to predict the target construction, the asymmetrical behaviour suggests that structural encoding is not fully independent from semantic plausibility.

7.3 Activation Strength

In the following two experiments, we test whether structural priming effects are affected by the proximity of prime to target and by increased exposure to the priming structure. We maintain the strict setting of our core corpus, where prime and target are semantically and lexically unrelated, thus testing to what extent the activation of abstract structural information across sentences is affected by recency and cumulativity factors.

7.3.1 Recency

To vary the proximity of prime to target, we create a set of *padding* sentences, using intransitive verbs, personal pronouns, and different auxiliary verbs to those used in our core corpus, including modal auxiliary verbs (e.g., *you might come, he did remain, they should appear*). These sentences were designed to contain frequent vocabulary with no lexical overlap nor semantic similarity to the prime and target sentences in the core corpus. A context in this setting consists of a sequence of 4 sentences, within which the priming sentence will vary in position relative to the target. This setup ensures that any priming observed is not influenced by the total length of the context, but solely by the position of the prime. In this condition, the PE is computed as follows:

$$_{\log}P(T_{\text{X}}|P_{\text{Z}}^{*}\ P_{\text{X}}\ P_{\text{Z}}^{*}) - {}_{\log}P(T_{\text{X}}|P_{\text{Y}}) \qquad (3)$$

where P_{Z} denotes the sequence of intransitive padding sentences.

Results The results of this experiment are shown in Figure 4A, which shows that increasing the proximity between prime and target has a highly positive impact on the strength of priming. Interestingly, the PE for the transitive cases is still

[13] This contrasts with psycholinguistic evidence suggesting that structural priming is not led by function-word priming in humans (Bock, 1989; Tree and Meijer, 1999).

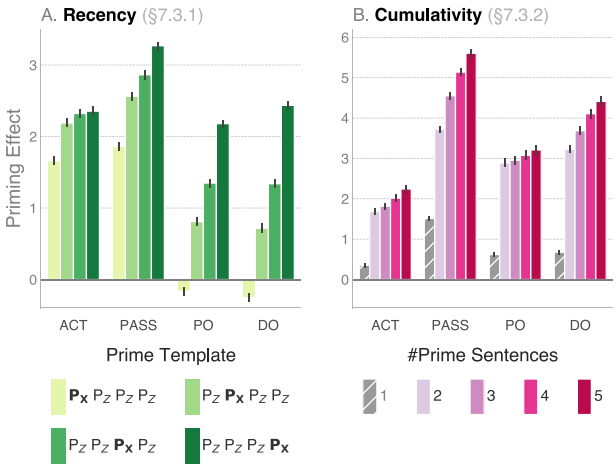


Figure 4: Results for GPT2-LARGE on the experiments described in §7.3: **A.** recency effect on priming, by increasing the distance between prime and target with additional intransitive sentences: each bar denotes a different position of the prime (P_X), surrounded by intervening sentences (P_Z); **B.** cumulative effects on priming, by increasing the number of prime sentences before a target.

relatively high even when the distance between prime and target is at its largest, whereas for the dative cases the PE has dropped drastically. This may indicate that the syntactic configuration of a transitive sentence is not corrupted as much by the intermediate intransitive sentence as the configuration of a dative sentence.

7.3.2 Cumulativity

To investigate the effect of cumulativity, we create a version of the *core* corpus where for each target sentence we sample multiple primes and concatenate them, resulting in priming contexts that vary between 1 and 5 sentences in length. All prime sentences in the prompt satisfy the semantic constraints with respect to the target that were outlined in §4. In this case, the PE is measured as follows:

$$\log P(T_X|P_X^+) - \log P(T_X|P_Y) \quad (4)$$

in other words, the PE of a sequence of congruent primes P_X^+ is expressed with relation to the log probability of a *single* incongruent prime sentence P_Y.

Results As shown in Figure 4B, for all constructions the PE increases monotonically as the number of congruent prime sentences increases. This resonates with the potential of large LMs for few-shot learning: The multiple priming sentences appear to act as ‘‘demonstrations’’ (in the sense of Brown et al., 2020) of a given structure, which presumably increases the activation of that type of structural information. This result is a yet another indication of structural information being encoded by the model and remaining active across sentences, as the main feature that is repeated across the multiple primes is the shared abstract structure.

7.4 Structural Complexity

Finally, we test whether the priming effects present in the *core* corpus are robust to different degrees of structural complexity between prime and target. In our core corpus, congruent prime and target sentences are constructed from the same sequence of parts of speech (see §4.1). Results by Reitter and Keller (2007) suggest that, for humans, short-term priming via residual activation is better explained by assuming hierarchical representations. In this experiment, we test whether the structural information encoded by the model is limited to sequential abstract structure or rather involves hierarchical syntactic representations.

To gain more insight on the nature of the structural information represented by the model, we construct a version of the corpus where some of the noun phrases are more complex than simply ‘‘`Dt N`’’ (e.g., *the awful tea from Spain*). The rationale behind this manipulation is the following: If the structure of a sentence is represented in terms of something akin to a hierarchical phrase-structure rule such as *VP* $\rightarrow$ *NP NP* or *VP* $\rightarrow$ *NP PP* rather than as a sequence of part-of-speech categories, then it should not matter whether prime and target differ with respect to the internal structure of the sub-constituents—we should observe a similar degree of priming whether the noun phrases are complex or not. Evidence suggests that this is indeed the case for humans (Tree and Meijer, 1999; Pickering and Branigan, 1998; Branigan et al., 2006).

We create a version of the core corpus where the noun phrases may contain a prenominal adjective, a prepositional phrase, neither or both in order to introduce varying degrees of complexity. We use a total of 164 adjectives manually labeled for compatibility with the different noun categories. The prepositional phrases are constructed with either *with* or *from*. For the *with* case, we select a set of

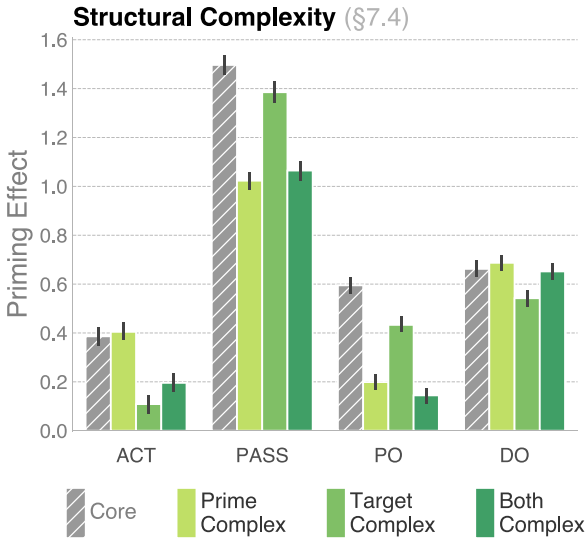


Figure 5: Results for GPT2-LARGE on the experiment described in §7.4, measuring the impact of increasing the complexity of one noun phrase per sentence in prime and target.

27 suitable nouns within the WordNet categories of *clothing, device*, or *container*. This results in noun phrases such as ''`Dt (A) N` *with* `Dt (A) N`''. For the *from* case, we use 23 country names, resulting in noun phrases such ''`Dt (A) N` *from* `N`''. All the additional vocabulary adheres to the same selection procedure as in §4, with prime and target being semantically unrelated. We test three conditions: (1) only the prime sentence has a complex NP, (2) only the target sentence does, (3) both prime and target have a complex NP—ensuring different NP structures across prime and target. In all three settings, any semantic role (*agent, patient*, or *recipient*) can be modified to become complex and there is at most one complex NP per sentence.

Results The results are shown in Figure 5. The first thing to note is that the presence of noun phrases of varying complexity across prime and target does not cancel out the PE: In all cases, the effect remains positive, although there is a decrease for several conditions. We also observe *asymmetrical* priming effects, for example, for transitive with complex prime (e.g., active is unaffected, whereas the PE for passive is clearly reduced). This suggests that some of the effects observed on the core corpus may be driven by the consistently simple sequential structures present in that data. Yet, the fact that the priming effect remains positive suggests that there is some degree of hierarchical structural information commonly encoded for both simple and complex NPs, which is carried over to influence the prediction of the target.

8 Discussion and Conclusions

In this paper, we investigated three main questions: (1) Are modern neural LMs susceptible to structural priming? (2) Which factors influence the strength of the priming effect? (3) What is the nature of the structural representations acquired by those models? To answer these questions, we designed a series of carefully curated large-scale corpora, proposed a metric to measure the degree to which a model is susceptible to priming, and ran a series of experiments on several Transformer LMs. This methodology constitutes a new way of assessing the representational abilities of LMs via examining their behavior in controlled setups, which complements tools like Targeted Syntactic Evaluations and the adaptation-based priming measure by Prasad et al. (2019).

Our results in Section 6 showed that on our *core* corpus, where we control for lexical overlap and semantic similarity between prime and target, *most* of the language models we test exhibit *some* degree of priming for *most* of the constructions we study. This is important, as it opens up the possibility of using priming to investigate what influences the learned representations of these language models.

In Section 7, we focused on GPT2-LARGE to conduct a series of subsequent experiments to dig deeper into the impact of different factors on the model's susceptibility to priming. In line with psycholinguistic accounts of residual activation, we found that the effects of priming decrease with the distance between prime and target and increase with the amount of exposure to a given structure. Our results indicate that the structural information being encoded is not fully autonomous from semantics: The Priming Effect is highly boosted by semantic similarity and lexical overlap between the words used in prime and target. Such boosting effects are well known to be present in human language processing as well. Furthermore, the Priming Effect partly disappears with semantically implausible prime sentences, suggesting that semantic plausibility is an important cue for the encoding of structure, arguably more so than in human language processing. Finally, we showed

that priming effects remain positive in the presence of phrases with differing degrees of complexity across prime and target. This offers some insight into the nature of the representations learned by the model: It suggests that, in addition to abstract sequential structure, some degree of hierarchical syntactic information is being represented.

The current work does not reveal, for the various conditions tested, what the mechanics of the boosting or suppressing effects are. For example, we do not know whether the boosts from lexical overlap or semantic similarity are the result of an improved match with the same structural representations, or of independent factors that influence priming behaviour. Similarly, the precise interplay between semantic plausibility and structural encoding remains unclear. Overall, the pattern of results calls for further investigation using interpretability methods, such as probing and feature attributions, which we plan to pursue in future work.

An additional aspect that requires further study is the role of the training data and its statistics, for example, regarding the frequency of the different constructions under investigation and the impact this may have on priming asymmetries within an alternation, and on priming behaviour more generally. An important future step to disentangle the factors that may give rise to priming behavior would involve training a range of different model types on the same data. This way it becomes possible to interpret the role that model architecture, model size, training objective, and corpus statistics play in shaping the behavior of the model. An important class of models to include in such studies are Masked Language Models. We conducted a preliminary experiment on three such models, which resulted in biased priming behavior for all (see Figure 6). We postulate that these models may rely less on the structure of a prime because their bi-directional nature allows them to take the entire target sentence into account. However, in order to adequately determine that this is entirely due to their training objective, and not due to external factors stemming from corpus statistics, future work could control for this with newly trained models.

Our study reveals novel details about the potential of LMs to represent structural information and the persistence of this information when making predictions about upcoming sentences. But more generally, we believe our findings also

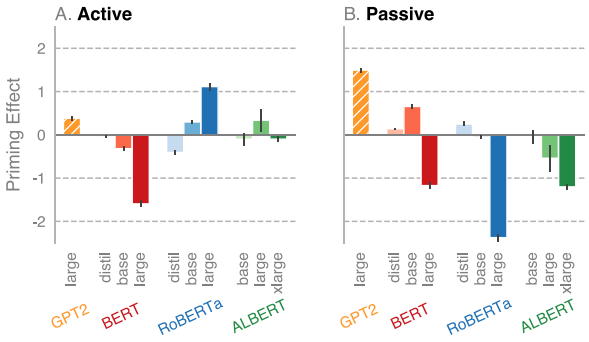


Figure 6: Priming Effects for three masked language models on the *core* corpus: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020). To compute sentence probabilities we utilize the pseudo-log-likelihood of Salazar et al. (2020), masking out one token at a time. Results for dative yield a similar pattern.

demonstrate the usefulness of the priming paradigm for investigating such questions. Even more generally, they illustrate the benefits of repurposing experimental paradigms from psycholinguistics to investigate the knowledge acquired by large neural language models. In that sense, the current paper complements exciting recent work that borrows other paradigms from linguistics and psycholinguistics, including grammaticality judgments, few shot learning, and cloze tests (Gauthier et al., 2020; Brown et al., 2020; Baroni, 2022; Lovering et al., 2021). That is, while syntactic priming offers one window into abstract language representations in neural language models, linguistics offers a whole row of windows that are starting to reveal an exciting vista.

Acknowledgments

We would like to thank the anonymous reviewers for their extensive and thoughtful feedback and suggestions, which greatly improved our work, as the action editor for his helpful guidance. We would also like to thank members of the ILLC past and present for their useful comments and feedback, specifically, Dieuwke Hupkes, Mario Giulianelli, Sandro Pezzelle, and Ece Takmaz. Arabella Sinclair worked on this project while affiliated with the University of Amsterdam. The project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455).

References

Afra Alishahi, Grzegorz Chrupala, and Tal Linzen. 2019. Analyzing and interpreting neural networks for NLP: A report on the first BlackboxNLP Workshop. *Natural Language Engineering*, 25(4):543–557. https://doi.org/10.1017/S135132491900024X

Marco Baroni. 2022. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. In Shalom Lappin, editor, *Algebraic Systems and the Representation of Linguistic Knowledge*, To appear. Abingdon-on-Thames: Taylor and Francis.

Sarah Bernolet and Robert J. Hartsuiker. 2010. Does verb bias modulate syntactic priming? *Cognition*, 114(3):455–461. https://doi.org/10.1016/j.cognition.2009.11.005

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow. GitHub repository. https://doi.org/10.18653/v1/2022.bigscience-1.9

Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355–387. https://doi.org/10.1016/0010-0285(86)90004-6

Kathryn Bock. 1989. Closed-class immanence in sentence production. *Cognition*, 31(2):163–186. https://doi.org/10.1016/0010-0277(89)90022-X

Kathryn Bock, Gary S. Dell, Franklin Chang, and Kristine H. Onishi. 2007. Persistent structural priming from language comprehension to language production. *Cognition*, 104(3):437–458. https://doi.org/10.1016/j.cognition.2006.07.003

Kathryn Bock and Zenzi M. Griffin. 2000. The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General*, 129(2):177. https://doi.org/10.1037/0096-3445.129.2.177

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234. https://doi.org/10.1146/annurev-linguistics-011619-030303

Holly P. Branigan, Martin J. Pickering, and Alexandra A. Cleland. 1999. Syntactic priming in written production: Evidence for rapid decay. *Psychonomic Bulletin & Review*, 6(4):635–640. https://doi.org/10.3758/BF03212972

Holly P. Branigan, Martin J. Pickering, Janet F. McLean, and Andrew J. Stewart. 2006. The role of local and global syntactic structure in language production: Evidence from syntactic priming. *Language and Cognitive Processes*, 21(7–8):974–1010. https://doi.org/10.1080/016909600824609

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. In *Cognitive Foundations of Interpretation*, pages 69–94. KNAW.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutsk, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Franklin Chang, Gary S. Dell, Kathryn Bock, and Zenzi M. Griffin. 2000. Structural priming as implicit learning: A comparison of models of sentence production. *Journal of Psycholinguistic Research*, 29(2):217–230. https://doi.org/10.1023/A:1005101313330

Noam Chomsky. 1957. *Syntactic Structures*. De Gruyter Mouton. https://doi.org/10.1515/9783112316009

Alexandra A. Cleland and Martin J. Pickering. 2003. The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language*, 49(2):214–230. https://doi.org/10.1016/S0749-596X(03)00060-3

William G. Cochran. 1977. *Sampling Techniques*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York; London.

Mark Davies. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190. https://doi.org/10.1075/ijcl.14.2.02dav

Forrest Davis and Marten van Schijndel. 2020. Discourse structure interacts with reference but not syntax in neural language models. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 396–407, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.conll-1.32

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.

Amit Dubey, Frank Keller, and Patrick Sturt. 2008. A probabilistic corpus-based model of syntactic parallelism. *Cognition*, 109(3):326–344. https://doi.org/10.1016/j.cognition.2008.09.006

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48. https://doi.org/10.1162/tacl_a_00298

Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 271–276, Gothenburg, Sweden. Association for Computational Linguistics.

Alex B. Fine and T. Florian Jaeger. 2013. Evidence for implicit learning in syntactic comprehension. *Cognitive Science*, 37(3):578–591. https://doi.org/10.1111/cogs.12022

Mary Ellen Foster and Michael White. 2007. Avoiding repetition in generated text. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 33–40, Saarbrücken, Germany. DFKI GmbH. https://doi.org/10.3115/1610163.1610170

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2021. A theoretical analysis of the repetition problem in text generation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI* 2021, *Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12848–12856. AAAI Press.

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. https://doi.org/10.18653/v1/N19-1004

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. Syntaxgym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76. https://doi.org/10.18653/v1/2020.acl-demos.10

Micah B. Goldwater, Marc T. Tomlinson, Catharine H. Echols, and Bradley C. Love. 2011. Structural priming as structure-mapping: Children use analogies from previous utterances to guide sentence production. *Cognitive Science*, 35(1):156–170. https://doi.org/10.1111/j.1551-6709.2010.01150.x

Stefan Th. Gries. 2005. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research*, 34(4):365–399. https://doi.org/10.1007/s10936-005-6139-3

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*

(NAACL-HLT), pages 1195–1205. Association for Computational Linguistics. https://doi.org/10.18653/v1/n18-1108

Jack Hessel and Alexandra Schofield. 2021. How effective is BERT without word ordering? Implications for language understanding and data privacy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 204–211, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-short.27

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4129–4138. Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1419

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.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926. https://doi.org/10.1613/jair.1.11196

Iva Ivanova, Holly P. Branigan, Janet F. McLean, Albert Costa, and Martin J. Pickering. 2017. Do you what I say? People reconstruct the syntax of anomalous utterances. *Language, Cognition and Neuroscience*, 32(2):175–189. https://doi.org/10.1080/23273798.2016.1236976

Iva Ivanova, Martin J. Pickering, Holly P. Branigan, Janet F. McLean, and Albert Costa. 2012. The comprehension of anomalous sentences: Evidence from structural priming. *Cognition*, 122(2):193–209. https://doi.org/10.1016/j.cognition.2011.10.013

T. Florian Jaeger and Neal E. Snider. 2013. Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, 127(1):57–83. https://doi.org/10.1016/j.cognition.2012.10.013

Harold Jeffreys. 1961. *Theory of Probability*, 3rd edition. International series of monographs on physics. Clarendon Press.

Rafal Józefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *CoRR*, abs/1602.02410.

Jaap Jumelet. 2020. diagNNose: A library for neural activation analysis. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 342–350, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.blackboxnlp-1.32

Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess NPI licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4958–4969, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.439

Michael P. Kaschak, Timothy J. Kutta, and John L. Jones. 2011. Structural priming as implicit learning: Cumulative priming effects and individual differences. *Psychonomic Bulletin & Review*, 18(6):1133–1139. https://doi.org/10.3758/s13423-011-0157-y

Robert E. Kass and Adrian E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795. https://doi.org/10.1080/01621459.1995.10476572

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.698

Jordan Kodner and Nitish Gupta. 2020. Overestimation of syntactic representation in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1757–1762, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.160

Yair Lakretz, Dieuwke Hupkes, Alessandra Vergallito, Marco Marelli, Marco Baroni, and Stanislas Dehaene. 2021. Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, 213:104699. https://doi.org/10.1016/j.cognition.2021.104699

Yair Lakretz, Germán Kruszewski, Théo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. 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 11–20. https://doi.org/10.18653/v1/N19-1002

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7410–7423, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.512

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. https://doi.org/10.1162/tacl_a_00115

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. *CoRR*, abs/1907.11692.

Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. 2021. Predicting inductive biases of pre-trained models. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 4020–4032. International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.355

Kyle Mahowald, Ariel James, Richard Futrell, and Edward Gibson. 2016. A meta-analysis of syntactic priming in language production. *Journal of Memory and Language*, 91:5–27. https://doi.org/10.1016/j.jml.2016.03.009

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. https://doi.org/10.1016/j.jml.2016.03.009

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1334

Inc. Merriam-Webster. 1989. *Webster's Dictionary of English Usage*. Springfield, MA: Merriam-Webster.

Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407. https://doi.org/10.3758/BF03195588

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. 2021. Refining targeted

syntactic evaluation of language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3710–3723, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.290

Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. 2021. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1145–1160, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.98

Martin J. Pickering and Holly P. Branigan. 1998. The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, 39(4):633–651. https://doi.org/10.1006/jmla.1998.2592

Martin J. Pickering and Victor S. Ferreira. 2008. Structural priming: A critical review. *Psychological Bulletin*, 134(3):427. https://doi.org/10.1037/0033-2909.134.3.427

Martin J. Pickering, Janet F. McLean, and Holly P. Branigan. 2013. Persistent structural priming and frequency effects during comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(3):890. https://doi.org/10.1037/a0029181

Grusha Prasad, Marten van Schijndel, and Tal Linzen. 2019. Using priming to uncover the organization of syntactic representations in neural language models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/K19-1007

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

David Reitter and Frank Keller. 2007. Against sequence priming: Evidence from constituents and distituents in corpus data. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*.

David Reitter, Frank Keller, and Johanna D. Moore. 2011. A computational cognitive model of syntactic priming. *Cognitive Science*, 35(4):587–637. https://doi.org/10.1111/j.1551-6709.2010.01165.x

David Reitter and Johanna D. Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76:29–46. https://doi.org/10.1016/j.jml.2014.05.008

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866. https://doi.org/10.1162/tacl_a_00349

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.240

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (NeurIPS)*.

Christoph Scheepers. 2003. Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. *Cognition*, 89(3):179–205. https://doi.org/10.1016/S0010-0277(03)00119-7

Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. 2020. Artificial neural networks accurately predict language processing in the brain. *bioRxiv*.

Katrien Segaert, Linda Wheeldon, and Peter Hagoort. 2016. Unifying structural priming effects on syntactic choices and timing of sentence generation. *Journal of Memory and Language*, 91:59–80. https://doi.org/10.1016/j.jml.2016.03.011

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021a. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.230

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021b. UnNatural Language Inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7329–7346. Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.569

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. https://doi.org/10.18653/v1/P19-1452

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 *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

Kristen M. Tooley and Kathryn Bock. 2014. On the parity of structural persistence in language production and comprehension. *Cognition*, 132(2):101–136. https://doi.org/10.1016/j.cognition.2014.04.002

Jean E. Fox Tree and Paul J. A. Meijer. 1999. Building syntactic structure in speaking. *Journal of Psycholinguistic Research*, 28(1):71–90. https://doi.org/10.1023/A:1023239604158

Jannis Vamvas and Rico Sennrich. 2021. On the limits of minimal pairs in contrastive evaluation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 58–68, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.blackboxnlp-1.5

Marten van Schijndel and Tal Linzen. 2018. A neural model of adaptation in reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4704–4710, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1499

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pages 5998–6008.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.14

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, et al. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. 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 2870–2880. https://doi.org/10.18653/v1/D19-1286

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel Bowman. 2020. BLiMP: A benchmark of linguistic minimal pairs for English. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 351–352. https://doi.org/10.1162/tacl_a_00321

Linda Wheeldon and Mark Smith. 2003. Phrase structure priming: A short-lived effect. *Language and Cognitive Processes*, 18(4):431–442. https://doi.org/10.1080/01690960244000063

Jennifer C. White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. 2021. A non-linear

structural probe. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 132–138, Online.Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.30