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

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, Raquel Fernández
Institute for Logic, Language and Computation
University of Amsterdam
{a.j.sinclair|j.w.d.jumelet|zuidema|raquel.fernandez}@uva.nl

Abstract

We investigate the extent to which modern, neural language models are susceptible to syntactic priming, the phenomenon where the syntactic 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 nature of the syntactic knowledge acquired by these models. We introduce a novel metric and release PRIME-LM, a large corpus where we control for various linguistic factors which interact with priming strength. We find that recent large Transformer models indeed show evidence of syntactic priming, but also that the syntactic generalisations learned by these models are to some extent modulated by semantic information. We report surprisingly strong priming effects when priming with multiple sentences, each with different words and meaning but with identical syntactic structure. We conclude that the syntactic priming paradigm is a highly useful, additional tool for gaining insights into the capacities of language models.

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 syntax 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, 2021), where the same questions have been asked about the human mind/brain for centuries.

Building on this tradition, this paper turns to syntactic priming to investigate the degree to which LMs encode structural information in their representations. ‘Syntactic priming’ (first described in Katryn Bock’s Syntactic persistence in language production, 1986) refers to the fact that humans are more likely to produce a sentence of a certain syntactic 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-b), and vice versa be more inclined to produce the double-object target sentence (2-b) after having read (1-b) instead of (1-a):

(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

Evidence for syntactic priming – to the extent that it can be shown to be independent from semantics, lexical overlap and other confounds – is taken as evidence for the cognitive reality of syntactic rules. A large body of work has established the various conditions that underlie priming in human language processing (Bock, 1986; Pickering and Branigan, 1998; Bock and Griffin, 2000; Cleland and Pickering, 2003; Pickering and Ferreira, 2008; Goldwater et al., 2011; Warglien and Gärdenfors, 2015; Pickering et al., 2013; Mahowald et al., 2016). In this paper, we leverage this work to create an elabo-
rate experimental pipeline to assess the priming behaviour of a broad set of language models.

Earlier work on syntactic priming in LMs by Prasad et al. (2019) reformulated priming as a process of model fine-tuning, and thus trained the model weights in between prime and target. In contrast, our approach simply treats priming as prior exposure to linguistic context, compatible with how most modern LMs are trained. This also allows us, in Section 3, to define an efficient novel metric for measuring the effect of priming, and to apply it to a broad spectrum of state-of-art LMs.

For our experiments, we create PRIME-LM, a large-scale corpus for examining priming consisting of ~1.3M prime-target sentence pairs, as we describe in Section 4. In Section 6 and 7 we use this corpus to answer three main research questions: (1) Are modern neural language models susceptible to syntactic priming? (2) Which factors influence the strength of the priming effect? And: (3) what do those factors reveal about the nature of the syntactic generalisations acquired by those models?

2 Background

2.1 Syntactic Priming in Humans

Syntactic 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 tend to reuse aspects of 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. As a given structure is processed, it becomes activated and can therefore persist across utterances.

In human language production, syntactic persistence effects are well attested, both in controlled behavioural experiments and in naturalistic conversational corpora, for a range of different syntactic structures (Bock, 1986; Pickering and Branigan, 1998; Bock and Griffin, 2000; Pickering and Ferreira, 2008; Reitter et al., 2011; Goldwater et al., 2011; Warglien and Gärdenfors, 2015; Pickering et al., 2013; Mahowald et al., 2016, among others). Furthermore, 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), although its presence is by no means restricted to adjacent sentences (Bock and Griffin, 2000; Bock et al., 2007).

Taken together, current psycholinguistic results provide compelling evidence for the autonomy of syntactic information and its susceptibility to priming. This includes experiments showing 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 syntactic persistence effects are robust enough in the absence of semantic and lexical cues. Yet, syntactic priming has been found to be affected by various aspects. For example, some types of lexical repetition between prime and target, such as the head verb for sentences (Pickering and Branigan, 1998) and the head noun for noun phrases (Cleland and Pickering, 2003), have been shown to enhance syntactic priming, suggesting that there is a lexical component involved. In addition, semantic relatedness between prime and target also has a boosting effect on syntactic priming, albeit smaller than the lexical repetition boost (Mahowald et al., 2016).

In the present study, we take inspiration from this tradition to investigate the syntactic priming behaviour of neural language models and thus their ability to encode structural information.

2.2 Syntactic Sensitivity of Neural LMs

The increasing capacities of neural language models (LMs) 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, i.a.). A common approach to examining LMs is to consider them as ‘psycholinguistic subjects’; by testing hypotheses derived from psycholinguistics we are able to determine to what extent LMs process language similarly to humans (Futrell et al., 2019; Ettinger, 2020; Davis and van Schijndel, 2020).

To assess the linguistic knowledge of LMs, a wide range of tools has 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 or syntactic structure (Tenney et al., 2019b; Hewitt and Manning, 2019; White et al., 2021). Another common approach is the usage of Targeted Syntac-
We capture the effects of priming by measuring the difference in log probabilities 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):**

$$\log P(T_X|P_X) - \log P(T_X|P_Y)$$  \hspace{1cm} [1]

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

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

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

(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:

$$\frac{1}{|P|} \sum_{P_X \in P(T_X)} [\log P(T_X|P_X) - \log P(T_X|P_Y)]$$

where $P(T_X)$ 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 Priming Effect 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) capture the impact of a single prime sentence on two subsequent target sentences of opposing structure, but they note that this metric suffers from confounding factors in the prior target probabilities that need to be regressed out. Priming in their setup is defined as a fine-tuning procedure (van Schijndel and Linzen, 2018), but Kodner and Gupta (2020) demonstrate that this method is prone to various confounding heuristics. Our metric does not suffer from these confounds, as we directly compare the impact of two prime sentences on a single target sentence:

We create a large-scale set of corpora designed to examine the priming behaviour 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 syntactic 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*), while 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*).

**Dative**

$$DO: NP_{agent} \ V \ NP_{recipient} \ NP_{patient}$$

$$PO: NP_{agent} \ V \ NP_{patient} \ F \ NP_{recipient}$$
In 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, 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. This allows us to control for the degree of semantic association between prime and target sentences. To minimise any effects stemming from word frequency factors, we only include USF content words which 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, and hand filter them selecting only those for which the alternation is possible. Given the constraints above, this results in 48 transitive and 22 ditransitive verbs. The ditransitive verbs were manually labelled 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 labelled for what category of nouns they accept, and a set of nouns fulfilling these categories was selected from USF following the same frequency constraints and using the WordNet closure categories of person, social group, social control, institution, physical entity, and object, which we further hand split into the categories non-edible, edible, and drinkable. This results in 119 nouns in total.

From this vocabulary, we are able to generate many realisations of our sentence templates through sampling, respecting the grammatically of the sentences produced. 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. Overall, PRIME-LM contains ~1.3M prime-target pairs.

### 4.3 The Core Corpus

In our 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. 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 since 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.

We create different variants of the core corpus that isolate specific aspects shown to influence syntactic 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 consider a wide range of language models. We confine our experiments to the class of auto-
regressive LMs,\textsuperscript{6} which are trained to predict the next token, in line with human incremental language processing. The main focus of our analysis lies on Transformer models (Vaswani et al., 2017), that constitute the current state of the art in language modelling, and have been shown to produce representations that correlate strongly with human brain signals (Schrimpf et al., 2020). Our approach is, however, not confined to Transformer models, and in our experiments we have also included two LSTM-based models (Hochreiter and Schmidhuber, 1997). 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 two sizes (125M, 1.3B; Black et al., 2021), which is based on GPT3 (Brown et al., 2020);
- **GulordavaLM**, the LSTM LM released by Gulordava et al. (2018);
- **GoogleLM**, a large-scale LSTM LM with CNN character embeddings (Józefowicz et al., 2016).

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), 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 will be publicly released upon publication of the paper.

6 Core Priming Behaviour 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 very strict experimental setup, where we take any level of observed priming effects to stem from abstract structural information.

The results are reported in Figure 1, split by the syntactic structure of the target sentence. First of all, it can be seen that across a large range of models a positive priming effect is present. This in itself is already remarkable: the language modelling objective contains no explicit cues for priming behaviour, and this is hence implicitly being picked up during training. We will now discuss these results in more detail.

**Priming bias** A model that convincingly exhibits susceptibility to priming for a certain alternation (transitive or dative), should obtain a positive Priming Effect on both syntactic alternatives. If a model were to obtain a positive Priming Effect for only one prime structure type, this would indicate that this prime structure is more likely to boost any subsequent target, regardless of its syntactic congruence with the prime. This is the case for DialoGPT-SMALL on transitive, with a negative Priming Effect on active constructions, but a large positive Priming Effect 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. Syntactic priming is hence only reliable if the effect is positive for both syntactic alternatives, as this indicates that the priming behaviour is the result of a higher-level interaction between prime and target.

In general this priming bias is not as severe, and

![Figure 1: 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.](image-url)
at most slightly skewed towards a particular construction. The double-object constructions score higher on average than their prepositional counterparts, but several models (e.g., GPT2-LARGE and DialoGPT-SMALL) net positive Priming Effects on both constructions. For the transitive alternation, the passive structures score higher on average than the active ones, but again various models obtain positive Priming Effects on both alternatives (e.g., all GPT2 models and DialoGPT-LARGE).

Model size We can investigate the impact of model size by comparing the results of the three model types that have been released in multiple sizes (GPT2, GPT-Neo, DialoGPT). Larger models may have more potential for encoding finer-grained structural information. If model size were to have a positive effect on syntactic priming this might manifest itself in two ways: either (1) the Priming Effect increases for both syntactic 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 both GPT2 and GPT-Neo larger model sizes slightly increase the bias towards passive in the transitive alternation, and decrease the bias towards double object in the dative case. The DialoGPT results are less consistent across model sizes: here we can see the priming biases flip completely as the size increases. From this we conclude that sensitivity to syntactic priming is not strictly driven by model size, and is likely to depend on a combination of factors related to model architecture and training data.

LSTMs The LSTM results demonstrate that these two models do not exhibit any clear priming behaviour, with the Priming Effect being close to 0. For future work it will be an exciting direction to compare LSTM and Transformer models that have been trained on the same amount of data, in order to investigate whether the recurrent nature of LSTMs makes them more susceptible to priming.

Best model The models that exhibit more susceptibility to syntactic priming across all four structure 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 specialised experimental conditions within the syntactic 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 robust to priming or to influence it in human language processing. For each experimental condition, we present a specialised corpus followed by an analysis of the priming effects exhibited by GPT2-large on this data, comparing them to the model’s behaviour on the core corpus. Examples from the core and specialised conditions can be found in Table 1.

7.1 Structural Autonomy

We begin by testing whether the effects found on the core corpus are robust to manipulations concerned with the syntactic complexity and the semantic plausibility of the sentences used as stimuli. These two aspects help to diagnose to what extent any structural information encoded by the model is autonomous from semantics.

7.1.1 Complex NPs

Our core corpus only includes bare noun phrases, consisting of a determiner and a noun (e.g., the tea). 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 NPs are somewhat more complex (e.g., the awful tea). The rationale behind this manipulation is the following: if the structure of a sentence is
Table 1: Example sentences for the core corpus and each condition described in §7.1 and 7.2. The same manipulations illustrated here for the Active construction hold for the other three syntactic constructions.

| Corpus     | Condition   | Prime (ACT)                     | Target (ACT)                      |
|------------|-------------|---------------------------------|-----------------------------------|
| Core       | —           | A worker broke a plate.         | The prince kissed the lady.       |
| Complex NP | Prime Complex | A sexy physician chased a minister. | The administration forgot the attorney. |
|            | Target Complex | The lady judged the leader.     | A terrible school embraced a business. |
|            | Both Complex | An awful author protected a son. | The attractive adult raised the log. |
| Implausible | —           | A meal drank a telephone.       | The hero followed the physician.  |
| Semantic Similarity | Verb Only | The chief struck the mayor.     | A bishop beat a kid.              |
|            | All Nouns   | An actor broke a glass.         | The actress wrapped the bottle.   |
|            | All Words   | The student drank the wine.     | A professor prepared a beer.      |
| Lexical Overlap | Random Noun | The pilot smelled the chicken.   | A girl prepared a chicken.        |
|            | Main Verb   | A father used a computer.       | The woman used the iron.          |
|            | Function Words | The soldier wanted the tea.    | The manager carried the book.     |
|            | All Nouns   | The king smelled the wine.      | A king drank a wine.              |

Results: The results are shown in Figure 2A. The first thing to note is that the presence of complex phrases does not cancel out the Priming Effect: in all cases, the effect remains positive. There is however a slight decrease for several conditions. This suggests that some of the priming effects observed on the core corpus may be driven by the consistently simple structures present in that data. Yet, the fact that the priming effect remains positive (even when the complexity of the prime and targets do not match) suggests that there is some degree of abstract structural information commonly encoded for both bare and complex NPs, which is carried over to influence the prediction of the target.²

7.1.2 Semantic Implausibility

In order to test to what extent reliance on semantic information is driving the priming effects observed in the core corpus, we construct a version of the corpus where the prime sentences are specifically designed to be semantically implausible. Gulordava et al. (2018) employed 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 hierarchical syntactic structure, in line with Chomsky’s (1957) proposal that structural grammaticality is distinct from meaningfulness in the human language faculty. Here we further test this hypothesis by analysing whether the LM is susceptible to syntactic priming effects when the prime sentence is nonsensical. As mentioned in §2.1, humans do exhibit syntactic 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’.

Results: The results of this experiment are shown in Figure 2B. The Priming Effect disappears for the

²Future work could investigate whether these results hold for more complex phrases, such as NPs with relative clauses.
ACT and PO constructions, while for PASS and DO it decreases when compared to the results on the core corpus, but remains positive. This suggests that, in some cases, some degree of abstract structural information present in the nonsensical sentences is exploited to predict the target construction, but that the strength of this syntactic signal partly depends on semantic clues.

7.2 Lexical Dependence

In all previous experimental conditions, prime and target sentences were semantically unrelated, which ensured that priming effects could not stem from the model assigning higher probabilities to words that were 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 priming effects.

7.2.1 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 are 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.\footnote{This results in a cosine similarity threshold of \(\sim 0.4\).}

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 than 15,000.

**Results** We find greater Priming Effect across constructions in this setup compared to the core corpus, although this is less pronounced for the PO structure. 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). As can be seen in Figure 3A, a semantically similar verb in prime and target leads to an increase of the Priming Effect, 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 Priming Effect is not substantially higher than when only the verb is similar. This underlines the importance of the verb as the syntactic head of the sentence.

7.2.2 Lexical Overlap

Lexical overlap between prime and target in the core corpus was avoided in both content and function word. 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 syntactic priming effects. This is in line with evidence found for humans (Mahowald et al., 2016). For all structures, verb overlap leads to higher priming effects than repeating one noun or even all nouns. Surprisingly, overlap of function words has the highest boosting effect for ACT and DO. This contrasts with psycholinguistic evidence suggesting that structural priming is not led by function-word overlap.

\[\text{In this case, to maximise the number of unique pairs, we allow a varying number of primes to target, rather than observing the 10-to-1 prime-target setup of the other corpora.}\]
Figure 4: Results for GPT2-large on the experiments described in §7.3: A. cumulative effects on priming, by increasing the number of prime sentences before a target; B. 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 ($P_Z$).

7.3 Activation Strength

Finally, in the following two experiments, we test whether syntactic priming effects are affected by increased exposure to the priming structure and by the proximity of prime to target. For these experiments, we return to 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 syntactic information is affected by cumulativity and recency factors. We hypothesise that with cumulative exposure to the prime structure, priming effects will be boosted. Conversely, increasing the distance between prime and target by introducing linguistic material with unrelated syntactic structures should lead to a reduction of the priming effects.

7.3.1 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 which vary between 1 and 5 sentences in length. In this case, the Priming Effect is measured as follows:

$$\log P(T_X | P_X^+) - \log P(T_X | P_Y)$$ [2]

in other words, the Priming Effect of a sequence of congruent primes $P_X^+$ is expressed with relation to the log probability of a single incongruent prime sentence $P_Y$; if we were to change the number of incongruent primes as well, it would be less straightforward to compare the priming effects in between the different number of priming sentences. This formulation isolates the impact that the number of congruent primes has on a subsequent prediction.

**Results** As shown in Figure 4A, for all constructions the Priming Effect increases monotonically as the number of congruent prime sentences increases, in line with cumulative short-term syntactic priming effects observed in human language production (Kaschak et al., 2011, 2014). A positive effect of cumulativity was also observed by Prasad et al. (2019) with their model-fine-tuning approach. This result is a particularly strong indication of syntactic information being encoded by the model, as the abstract syntactic structure is the only element shared across the multiple priming sentences.

7.3.2 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 Priming Effect is computed as follows:

$$\log P(T_X | P_X^+ P_X) - \log P(T_X | P_Y P_Y)$$ [3]

where $P_Z$ denotes the sequence of intransitive padding sentences. The Priming Effect can hence be interpreted as the difference in probability when the congruent prime $P_X$ has been swapped with its incongruent counterpart $P_Y$, in the context of $P_Z$.

Note: The diagrams in the figure illustrate the results for GPT2-large for the experiments described in §7.3: A. cumulative effects on priming, by increasing the number of prime sentences before a target; B. 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 ($P_Z$).
Results  The results of this experiment are shown in Figure 4B. Similarly to what has been observed in human language production experiments (Chang et al., 2000), increasing the distance between prime and target has a highly negative impact on the strength of priming. For the ACT case the effect is less clear than for the other three constructions: here the Priming Effect does not increase monotonically as the prime moves closer to the target.

8 Discussion and Conclusions

In this paper, we investigated three main questions: (1) Are modern neural LMs susceptible to syntactic priming? (2) Which factors influence the strength of the priming effect? And: (3) what do those factors reveal about the nature of the syntactic generalisations 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 computational experiments on a range of different neural LMs. This methodology constitutes a new way for investigating the nature of the internal representations learned by the multi-billion parameter space of this type of language model.

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 syntactic constructions we study. This is important, as it opens up the possibility of using priming to investigate 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. We showed that the priming effect may remain positive in the presence of phrases with different degrees of complexity and semantically implausible prime sentences. This provides further evidence that some degree of abstract syntactic structure is being represented by the model. Our results also indicate that the structural information being encoded is not fully autonomous from semantics: the priming effect partly disappears with semantically implausible sentences, and it is highly boosted by semantic similarity and lexical overlap between the words used in prime and target. The latter boosting effects are well known to be present in human language processing as well.

The current work does not reveal 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 syntactic representations, or the result of an independent path through which information flows. The pattern of results does call for further investigation using interpretability methods, which we will plan to pursue in future work.

Our study thus reveals novel details about the nature of the representations learned by specific language models about specific syntactic constructions. But more generally, we believe our findings also demonstrate the usefulness of the priming paradigm for investigating such questions. Even more generally, they illustrate the benefits of re-purposing 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, 2021; 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.
References

Afra Alishahi, Grzegorz Chrupala, and Tal Linzen. 2019. Analyzing and interpreting neural networks for nlp: A report on the first black-boxnlp workshop. Natural Language Engineering, 25(4):543–557.

Marco Baroni. 2021. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. arXiv preprint arXiv:2106.08694.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow.

J Kathryn Bock. 1986. Syntactic persistence in language production. Cognitive psychology, 18(3):355–387.

Kathryn Bock. 1989. Closed-class immanence in sentence production. Cognition, 31(2):163–186.

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.

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.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. Annual Review of Linguistics, 6(1):213–234.

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.

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, BenjaminChess, JackClark, Christopher Berner, SamMcCandlish, AlecRadford, IlyaSutskever, and DarioAmodei. 2020. Language models are few-shot learners. CoRR, abs/2005.14165.

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.

Noam Chomsky. 1957. Syntactic structures. De Gruyter Mouton.

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.

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.

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.

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.

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.
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.

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.

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.

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 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pages 1195–1205. Association for Computational Linguistics.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4129–4138. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural Computation, 9(8):1735–1780.

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.

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.

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.

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.

H. 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.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? on the ability of lstms to understand negative polarity items. In Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018, pages 222–231. Association for Computational Linguistics.
Michael P Kaschak, Timothy J Kutta, and Jacqueline M Coyle. 2014. Long and short term cumulative structural priming effects. *Language, cognition and neuroscience*, 29(6):728–743.

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.

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

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.

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.

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.

Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. 2021. *Predicting inductive biases of pre-trained models*. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

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.

Inc. Merriam-Webster. 1989. *Webster’s Dictionary of English Usage*. Springfield, Mass.: 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.

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.

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.

Martin J Pickering and Victor S Ferreira. 2008. Structural priming: a critical review. *Psychological bulletin*, 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.

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.

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

David Reitter, Frank Keller, and Johanna D Moore. 2011. A computational cognitive model of syntactic priming. *Cognitive science*, 35(4):587–637.
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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR, abs/1910.01108.

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.

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.

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.

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 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.

Jean E Fox Tree and Paul JA Meijer. 1999. Building syntactic structure in speaking. Journal of Psycholinguistic Research, 28(1):71–90.

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 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008.

Massimo Warglien and Peter Gärdenfors. 2015. Meaning negotiation. In Applications of conceptual spaces, pages 79–94. Springer.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Bli, 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.

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.

Jennifer C. White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. 2021. A Non-Linear Structural Probe. arXiv:2105.10185 [cs]. ArXiv: 2105.10185.

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.

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.