

Emergent Linguistic Structures in Neural Networks are Fragile

Emanuele La Malfa* Matthew Wicker*[†]

Marta Kwiatkowska*

June 2, 2023

Abstract

Large Language Models (LLMs) have been reported to have strong performance on natural language processing tasks. However, performance metrics such as accuracy do not measure the quality of the model in terms of its ability to robustly represent complex linguistic structures. In this paper, focusing on the ability of language models to represent syntax, we propose a framework to assess the consistency and robustness of linguistic representations. To this end, we introduce measures of robustness of neural network models that leverage recent advances in extracting linguistic constructs from LLMs via probing tasks, i.e., simple tasks used to extract meaningful information about a single facet of a language model, such as syntax reconstruction and root identification. Empirically, we study the performance of four LLMs across six different corpora on the proposed robustness measures by analysing their performance and robustness with respect to syntax-preserving perturbations. We provide evidence that context-free representation (e.g., GloVe) are in some cases competitive with context-dependent representations from modern LLMs (e.g., BERT), yet equally brittle to syntax-preserving perturbations. Our key observation is that emergent syntactic representations in neural networks are brittle. We make the code, trained models and logs available to the community as a contribution to the debate about the capabilities of LLMs.

1 Introduction

Large Language Models (LLMs) exhibit impressive performance in natural language processing (NLP) tasks such as text classification [32], language translation [31], large-scale search engines [62], and even source-code generation for programming languages [60], resulting in a great deal of media attention. However,

*Department of Computer Science, University of Oxford, Oxford, United Kingdom. Corresponding address: emanuele.lamalfa@cs.ox.ac.uk

[†]This work has been done while the author was affiliated with the University of Oxford. Current affiliation: The Alan Turing Institute, London, United Kingdom

current metrics mainly measure the performance of LLMs’ ability to capture statistical patterns of discourse [4, 37], as opposed to their ability to robustly capture and represent complex linguistic patterns of their domains.

Representing the linguistic and grammatical structure underlying the data intuitively plays a cogent role in robust generalization of any linguistic system [9]. Remarkably, LLMs are also arguably capable of accurately representing structures such as syntax trees [38], which has motivated researchers to investigate their linguistic capabilities with ad hoc measures and benchmarks [53, 56]. This indisputable progress is nonetheless counteracted by a series of critiques that show how LLMs are unable to perform basic reasoning [44], have considerable biases [29], are not well aligned to stakeholder values [4], and are brittle in the face of adversarial examples [34]. Such studies make it clear that sustainable, long-term advances in NLP need to be facilitated by appropriate metrics that capture how LLMs represent the complex linguistic patterns underlying their training data [5]. Unfortunately, a naive adaptation of definitions from other domains, e.g., the image domain, is flawed [34].

In deep learning, robustness is often measured in terms of how much a bounded perturbation (e.g., with respect to an ℓ_p -norm) of a test set input affects the output of a network [58]. For NLP, bounded ℓ_p -norm perturbations applied directly to an input do not preserve its semantic meaning or syntactic structure and are therefore linguistically uninteresting. Further, ℓ_p distance measures in the embedding space do not reflect how the input perturbation has affected the representation of key linguistic features, which are extracted by the language model from the input data.

In this paper, we propose a framework to evaluate the syntactic consistency and robustness of linguistic representations that leverages probing tasks [10, 38], namely, neural networks trained directly on the representation embedding to evaluate the representation’s ability to encode a specific linguistic phenomenon, such as the syntax tree of a sentence. To this end, we propose an efficient probing method to perturb the input text so that its syntax (or context) is largely preserved. We validate the perturbations to show that they can serve as an effective proxy of syntax-preserving perturbations. We focus on syntactic robustness, which informs our selection of probing tasks, but note that other tasks can be easily incorporated. To assess robustness, we aim to measure the performance of a language model to probing tasks on the original and perturbed datasets. More specifically, we define a measure of robustness in terms of aggregating (averaging) the worst-case drop of performance of a collection of probing tasks over a given dataset, for a given perturbation budget, which then captures the model’s ability to encode the linguistic phenomena, and is therefore more appropriate for NLP settings.

In principle, our methodology for evaluating robustness of linguistic representations allows us to benchmark LLMs and can be used by others to guide the development of models that optimize for robust syntactic understanding, which we find to be universally lacking. In addition, we demonstrate the ability of the proposed metrics to offer novel insights and perspectives into the workings of LLMs. In particular, we show that, despite conventional wisdom, context-

dependent LLMs (BERT) are just as syntactically brittle as context-free embeddings (Word2Vec), and that deeper latent features provide as much syntactic robustness as shallow features. We also offer a critique of the effect of fine-tuning of such representations. We choose 6 datasets from the English Universal Dependencies [45], which are representative of different linguistic registers, and perform experiments on both standard embeddings (Word2Vec and GloVe) and modern LLMs (BERT and RoBERTa) [11, 18, 32, 40] on 4 representative syntactic probing tasks from [38], i.e., structural probe [21], part-of-speech (POS) tagging, root identification and calculation of the depth of a sentence’s syntax tree. We use two complementary sources of perturbations. The former method is grounded in the utilization of WordNet synonyms [41], which we subsequently augment with constraints designed to uphold the syntactic structure of a sentence while modifying the maximum number of words permissible within a predetermined perturbation threshold. Conversely, the latter approach is centered around word prediction and relies on GPT-2 [50]. WordNet facilitates the selection of perturbations that maintain syntactic integrity, whereas GPT-2, along with other language models, produces substitutions that typically do not retain the original sentence’s syntactic coherence. However, they compensate for this limitation by demonstrating a heightened awareness of contextual factors.

In summary, in this work we make the following contributions:

- Propose measures to evaluate robustness of linguistic representations that leverage probing tasks.
- Develop a methodology for analyzing an LLM’s ability to robustly capture complex syntactical information underlying its training data.
- Demonstrate how our robustness metrics reveal that context-free representations are equally brittle to manipulations as more sophisticated context-dependent representations.
- Provide empirically insightful observations into feature collapse, training duration, and depth of pre-trained LLM heads from the robustness perspective.

In addition to these empirical observations, we draw attention to the brittleness of emergent syntactic representations of language models as a contribution to the debate about the capabilities of LLMs. The code, trained models and logs are made available for reproducibility.¹

2 Related Works

In this section, we first overview the linguistic models we study. Next, we discuss recent methods aimed at extracting the syntactic structure represented by a

¹The code to replicate the experiments of this paper is available at the following repository: <https://github.com/EmanueleLM/emergent-linguistic-structures>.

language model. Finally, we summarise a series of works that revealed weaknesses in language understanding captured by these models.

Linguistic representations Early attempts to represent language were in the form of bag-of-words or binary/one-hot encodings [37]. The success of deep learning led to the increasing reliance on vector representations of language (word embeddings) in NLP tasks [59]. Word embeddings such as Word2Vec[40] and GloVe[46] translate one-hot encoded words and embed them into real-valued vectors such that similar words are mapped to similar vectors. In the past decade, researchers developed linguistic representations whose symbols are independent from each other: we call such representation context-free [37, 40, 46]. Only recently, with the improvement of training procedures and the capability of deep learning models to ‘digest’ massive datasets, representations where each symbol depends on the context in which it appears became possible [47], thus better embodying the distributional hypothesis [54]. We refer to this approach as context-dependent, e.g., Large Language Models (LLM) [11, 47].

Extracting syntactic structures Many works have investigated whether representations embed the structure of a language, with a particular focus on LLMs [17, 23, 38] recently, and context-free representations [30] prior to this. There is an ongoing debate on whether representations can embody complex syntactic structures [20, 56]. Studies include assessing grammaticality of a representation [43] and extracting grammars from representations, with works ranging from linguistics [13, 14] to formal languages [55] and NLP [26, 39].

Alignment in NLP The impressive performance of modern LLMs has led to claims that they have “mastered” language [25]. This claim has been disputed by a series of works seeking to contextualize the results of LLMs, in particular showing their lack of “natural language understanding” [5]. In [6] the authors show that context-free representations can be gender biased. Considerable biases are also found to exist in context-dependent representations in [29]. In addition to bias, [4] highlights the multitude of ways in which LLMs are not well aligned with stakeholder values. In [44], the authors highlight the language models’ failure to perform basic linguistic reasoning tasks, while in [15] their scope limitations.

The works closest to our paper are those that study the robustness of NLP models. While human understanding of linguistic structures is very robust [16], the robustness of NLP models is still far from being achieved [61], as prominent works over-focus on a notion of adversarial robustness [22, 24, 36] that is linguistically flawed [34, 64]. Practically speaking, robustness is measured and guaranteed either in the embedding space, hence w.r.t. bounded ℓ_p -norm changes of a sentence’s embedding representation [22, 35], or through discrete, semantically enhanced replacements [1, 12, 52, 34], which do not capture linguistic structures such as syntax.

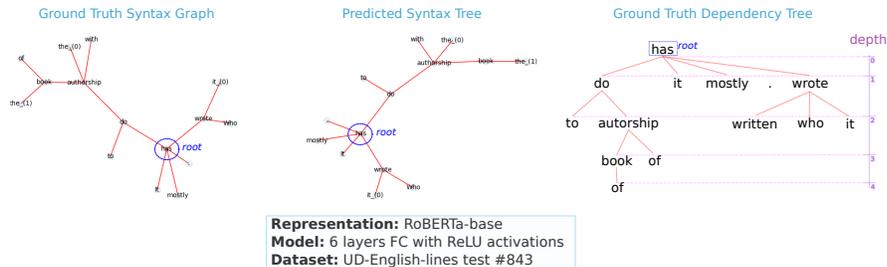


Figure 1: A syntax graph reconstructed via the structural probe task from a RoBERTa representation is shown in the middle; for comparison, the ground truth structure is sketched on the left. On the right, the same structure is displayed as a dependency tree (annotated with additional information so that dependencies and hierarchies between words are made clear) so that other supervised tasks can be instantiated, e.g., identifying the **root**, or computing the **depth** of the tree.

3 Background and Notation

In this section, we present the notation and concepts that we use to frame our methodology. A sentence $s = \{s_1, \dots, s_l\}$ is a finite sequence of $l > 0$ symbols (here words) defined over finite vocabulary Σ . A sense of grammaticality is given by linguistic rules. A linguistic rule assesses the violation of a property by a sentence s and we denote a sentence s satisfying a rule, \mathbf{R} , with $\mathbf{R} \models s$. A language, \mathbf{L} , is defined by an alphabet Σ and a (possibly infinite) set of rules $\mathbf{R} = \{R_1, \dots, R_n\}$.

The application of neural networks to language became possible thanks to a numerical representation of sentences [40]. Given a sentence s comprising $l > 0$ symbols from a language \mathbf{L} , a linguistic representation ψ^θ , where θ are parameters, maps s into a $(l \cdot d)$ -dimensional vector of real numbers, i.e., $\psi^\theta : s \in \mathbf{L} \rightarrow \mathbb{R}^{l \cdot d}$. A linguistic representation ψ^θ is said to be context-free (or independent from the context) when the representation of each word is independent from the other words, i.e., $\psi^\theta(s_i | s \setminus s_i) = \psi^\theta(s_i)$. Otherwise, it is said to be context-dependent (or dependent on the context).

4 Methodology

Given a linguistic embedding ψ^θ and a sentence s , the central question of interest in this work is what information does ψ^θ extract robustly from s ? To answer this question we consider using perturbation-based analysis. Specifically, given another sentence s' that is similar to s , how does $\psi^\theta(s)$ differ from $\psi^\theta(s')$? While such perturbation analysis is reminiscent of adversarial robustness in the image domain [58], we highlight that a naive adaptation to NLP is devoid of the nuance of natural language and inappropriate for this setting [34]. We address this shortcoming with a two-phase framework. Firstly, we seek to gain

insights into the syntactic properties understood by a language model through the use of probing tasks. Secondly, we propose an efficient scheme for computing perturbations that aim to preserve the sentence’s original syntax, and study how such perturbations affect the model insights from the probing task.

4.1 Probing Tasks for Model Introspection

Recently, probing tasks have been introduced as a linguistically relevant measure of a model’s understanding of complex linguistic phenomena, often grouped into surface, syntax, and semantic probing tasks [10]. A probing task is a simple, non-challenging task used to extract linguistically meaningful information about a single facet of a language model, e.g., the subject number task requires us to extract the number of subjects in a sentence from its embedding. Probing tasks are classifiers trained directly on the representation embedding to evaluate the representation’s ability to encode a specific linguistic phenomenon. The key idea here is that the probing task be linguistically specific – testing the representation of a specific phenomenon – and simple enough that strong performance on the task indicates, without bias, that a language model has accurately represented the given linguistic phenomenon. We select four probing tasks, which we design to assess the presence of syntactic structures in linguistic representations, but stress that our framework could be extended to any of the ten probing tasks presented in [10].

We first define a generic probing task, which serves as a basis to describe the four syntactic tasks that feature in our robustness framework.

Definition 1 (*Probing Task*) *Given a set $S = \{s^{(1)}, \dots, s^{(n)}\}$ of $n > 0$ sentences from a language \mathbf{L} , each paired with a label $T = \{t^{(1)}, \dots, t^{(n)}\}$, a probing task consists of finding a mapping f from each sentence representation $\psi^\theta(s)$ s.t. $\mathbb{E}_{(s^{(i)}, t^{(i)}) \sim (S, T)} [\mathcal{L}(f(\psi^\theta(s)), t)] > p$, where \mathcal{L} is a measure of performance of such a reconstruction, and p some positive quantity that certifies a given level of performance.*

The first syntactic probing task we propose to study is the *syntax reconstruction* task. An accurate understanding of the information content of a sentence s depends on the reader’s ability to understand the intra-word relationships in s . This is not just true for natural language, but also for programming languages where parse trees are important to understand source code. A syntax tree t is an undirected, acyclic graph $G := (s, A)$, where the words of s are vertices and A is an edge list which contains an edge between two words if they modify each other or are contextually linked, see [37] for more details. There are two standard representations of syntax trees in NLP and linguistics, namely dependency and constituency trees. In the former, each word corresponds to a node and the tree structure reflects the word order, while, in the latter, words themselves are terminal nodes whose order follows the ‘bare phrase structure’ (as per the minimalist program by Noam Chomsky [8]). In this paper, we work with dependency tree representations, but the methodology and the results can be extended to

the constituency representation standard. Formally, the syntax reconstruction probing task is given as:

Definition 2 (*Syntax Reconstruction*) Given a set $S = \{s^{(1)}, \dots, s^{(n)}\}$ of $n > 0$ sentences from a language \mathbf{L} and their syntax-tree representation $T = \{t^{(1)}, \dots, t^{(n)}\}$, syntax reconstruction is a probing task f from S to T that guarantees sufficient performance.

In practice, the syntax probing task consists of extracting, from a sentence representation, the distance between each pair of words, as they are arranged in the dependency parse tree of the sentence itself (see Figure 2): the task is commonly used as a proxy of the capabilities of a representation to recognize the mutual dependency relationships between words in a sentence, represented as a directed graph. Probes are usually linear [38], as one wants to assess how representations encode features that are immediately available to solve the task [44], though there has been recent criticism of the excessive simplicity of linear probes compared to non-linear ones [48, 63].

Using probing tasks to assess the capabilities of a model has become a popular approach with the development of increasingly complex linguistic representations. However, some studies have shown that probes can only reveal the correlation between the traces of a symbolic structure in a representation and its performance on a task [3, 51]. In our work, we use probes to provide evidence of the existence of syntactic structures in linguistic representations, rather than testing their performance on higher-level NLP tasks. We show an example of a dependency syntax tree and its reconstruction in Figure 1.

The second probing task disregards intra-word relationships and focuses on a language model’s ability to identify the part of speech of a given word. Formally, the *part-of-speech (POS) tagging* task is given as:

Definition 3 (*POS-tagging*) Given a set $S = \{s^{(1)}, \dots, s^{(n)}\}$ of $n > 0$ sentences from a language \mathbf{L} and the POS-tags for each sentence, $POS = \{pos^{(1)}, \dots, pos^{(n)}\}$, POS-tag reconstruction is a probing task g from S to POS that guarantees sufficient performance.

This task is commonly used as a proxy of the capabilities of a representation to represent the role of a word in its context: an example is shown in Figure 3. In conjunction, these two tasks allow us to inspect how a language model identifies and semantically links entities in a sentence, thus giving us a comprehensive, linguistically-informed perspective on what is captured by a language model.

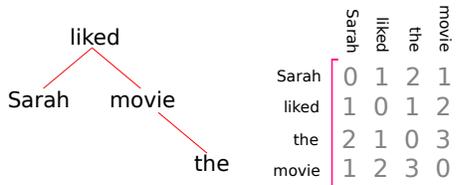


Figure 2: The dependency parse tree of a sentence (left), alongside the matrix of distances between pairs of words in the tree (right).

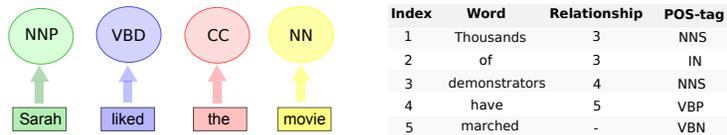


Figure 3: A sentence with its POS tags (left). A sentence in the CONLL format, used to test a model on multiple syntactic tasks (right).

We complete the benchmark with two further syntactic tasks, namely root identification and the tree-depth estimation, which we present below.

Definition 4 (*Root Identification*) Given S and T as in Def. 2, and the root of the tree $R = \{r^{(1)}, \dots, r^{(n)}\}$ where $r^{(i)} \in t^{(i)}$, root identification is a probing task h from S to R that guarantees sufficient performance.

Definition 5 (*Tree-depth Estimation*) Given S and T as in Def. 2, and the depth of the tree $D = \{d^{(1)}, \dots, d^{(n)}\}$ where $d^{(i)} \in \mathbb{N}^+$, tree-depth estimation is a probing task u from S to D that guarantees sufficient performance.

With tasks in Def. 4 and 5, we assess a representation’s capacity to distill single units of information (root and depth), which can be extracted from a tree’s sentence representation. We sketch the two tasks in Figure 1 (right). When compared to the structural probe task, root identification and tree-depth are easier to solve: in fact, they are meant to show to what extent low-order syntax information, as opposed to high-order encoded by structural probe, is encoded in a linguistic representation.

4.2 Syntax-Preserving Perturbation Analysis

The second phase of our methodology involves perturbation-based analysis. It is widely known and confirmed by neuroscience that human language exhibits very robust linguistic representations [7, 16], while NLP models suffer from brittleness against perturbations, which are often easily transferable across models yet difficult to detect [27]. Though many works have shown how brittle NLP models are in the presence of bounded attacks on embedding space [36], such attacks do not necessarily preserve human meaning and are therefore arguably of questionable merit [34]. We define two types of perturbations: the first aims to preserve syntax (referred to as coPOS) and constitutes the backbone of our empirical evaluation; the second exploits context to preserve the semantics (coCO), and is introduced to strengthen our comparison of models’ syntactic robustness. We further add, as baseline, a perturbation method with words randomly sampled from the English vocabulary. We now introduce the coPOS and the coCO perturbation methods, which are illustrated in Figure 4.

Definition 6 (*Consistent POS Substitution*) A consistent POS substitution (coPOS) consists of the replacement of one or more words in a sentence s

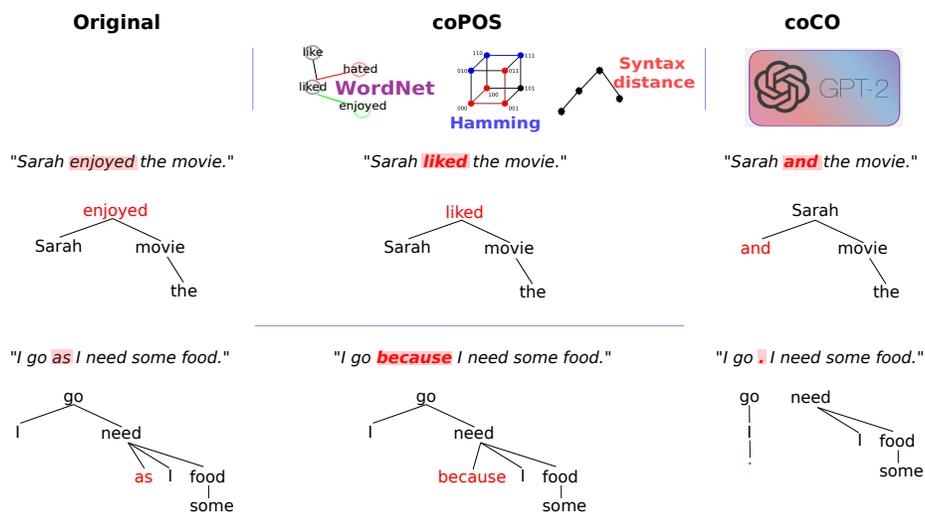


Figure 4: Two examples of coPOS and coCO perturbations applied on clean input texts, and the resulting syntax trees induced by such alterations. Words perturbed are highlighted in red. coPOS perturbations are designed to minimize the probability of disrupting the syntax of a sentence (such as the substitution of ‘as’ with ‘because’), while coCO can possibly disrupt it (e.g., substitution of ‘and’ with a period).

with words that keep unaltered the POS-tag of the perturbed sentence, i.e., if s/s' are the original/perturbed sentence, pos/pos' the ground-truth POS-tag of s/s' , and $s' = sub(s)$, the perturbation procedure, then it holds for coPOS that $sub(s) \implies pos \equiv pos'$.

Ensuring that a perturbation satisfies the coPOS definition enables interpretability of our results. Specifically, a coPOS perturbation is built with the intent to preserve the word’s syntactic role in a sentence, and therefore one can impute any probe misclassifications to a lack of robustness of the linguistic representation. Since guaranteeing that a perturbation always preserves its coPOS tag is challenging due to the intrinsic complexity of natural language, we rely on an efficient algorithmic implementation to generate proxy coPOS perturbations, described in Section 5, which we carefully validate on the datasets used in our experimental evaluation (see Section 5.1).

Definition 7 (*Context Consistent Substitution*) A context consistent substitution (coCO) consists of the replacement of one or more words in a sentence s with a generative model that maintains semantic closeness but does not strictly enforce the substitution to be syntactically coherent. While many alternative methods exist in the literature to generate coCo perturbations, we rely on GPT-2 [50] next word predictions, which serves as a benchmark for syntactically-informed methods such as coPOS. In other words, a substitution w' of a word $w \in s$ is generated by a generative model ϕ conditioned on the context where the word appears, i.e., $w' = \operatorname{argmax}_{w \in V} \phi(s|s \setminus w)$.

Below, we formally define the conditions under which we consider a linguistic model robust: informally, for a linguistic representation to be robust we desire it to accurately solve a family of probing tasks and behave consistently on slight syntax-preserving perturbations of an input text. We assume that coPOS substitutions are used as perturbations, but note that the concept of linguistic robustness can also be instantiated with Def. 7.

First, we introduce the notion of consistency of representations, termed ϵ -robustness.

Definition 8 (*ϵ -robust Representations*) Given a linguistic representation ψ^θ , a set of sentences S , a set of perturbed sentences S' which are coPOS perturbations of S , and a measure of distance $dist : (s, s') \rightarrow \mathbb{R}$ between representations (e.g., ℓ_p -norm, cosine similarity), we say that the representation ψ^θ is ϵ -robust w.r.t. $dist$ if $\forall (s, s') \in (S, S'), \max(dist(\psi^\theta(s), \psi^\theta(s'))) < \epsilon$.

Despite its simplicity, ϵ -robustness is linguistically informed, as all sentences in S' are coPOS to those in S , and thus we can be confident that the perturbations are syntactically consistent for the given probing task. Moreover, this metric can serve as a useful tool for developing robust language models, in the sense of maximizing ϵ while maintaining good performance on the underlying task.

While ϵ -robust representations are desirable, what is more informative is the ability for a representation, ψ^θ , to be robust not just with respect to a distance

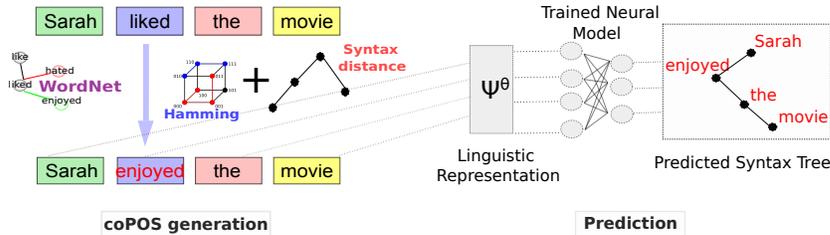


Figure 5: An example of a perturbed sentence s' obtained through a coPOS perturbation. Candidate substitutions are sampled from a pool of WordNet synonyms, from which we select the one that maximizes the Hamming distance and minimizes the syntactic disruption w.r.t. the original input, see Section 5 for details. The perturbation is then fed, through a linguistic representation ψ^θ , to a probe (neural network trained directly on the representation) that in turn predicts its syntax tree.

metric, but with respect to a probing task. Formally, we define a language model ψ^θ to be *syntactically robust* if the performance on multiple proxy tasks is not adversely affected by perturbations that are close in some representation space (e.g., Def. 6).

Definition 9 (*Syntactically Robust Representation*) Given an input s , its representation $\psi^\theta(s)$, a set of probing tasks $\{\mathbf{T}_1, \dots, \mathbf{T}_m\}$, a set of mappings $\{f_1(s), \dots, f_m(s)\}$ that take as input the representation $\psi^\theta(s)$ and solve the respective i -th probing task, a set of strictly positive quantities $\{\tau_1, \dots, \tau_m\}$ and a small quantity $\epsilon > 0$, a set of measures of performance on each task $\{\mathcal{L}_1, \dots, \mathcal{L}_m\}$, a consistent perturbation $s' = \text{sub}(s)$, and a measure of distance between representations $\text{dist} : (s, s') \rightarrow \mathbb{R}$, ψ^θ is syntactically robust iff $\forall (\mathbf{T}_i, f_i, \mathcal{L}_i, \tau_i) \in (\mathbf{T}, f, \mathcal{L}, \tau)$, $\text{dist}(\psi^\theta(s), \psi^\theta(s')) < \epsilon \implies \mathcal{L}_i(f_i(\psi^\theta(s)), f_i(\psi^\theta(s'))) < \tau_i$.

5 Algorithm for Evaluating Robustness

In this section, we describe a procedure to assess the robustness of the syntactic structures encoded by a linguistic representation, as formalized in Def. 9. We outline the full algorithm and give step-by-step comments in the Appendix, Section A.

The general framework takes as input a language model ψ^θ , a set of m probing tasks $\{\mathbf{T}_i\}_{i=1}^m$, performance metrics for each task $\{\mathcal{L}_i\}_{i=1}^m$, a perturbation function sub , and two constants τ and k . For each task \mathbf{T} , we sample n sentences and for each sentence s we compute k coPOS (or coCO, alternatively) perturbations s' , where each coPOS perturbation modifies τ -many words. We then use the performance measure corresponding to the task to measure $\mathcal{L}(f(\psi^\theta(s)), f(\psi^\theta(s')))$

and take the perturbed sentence s' that maximizes this quantity to be the approximate (since we sample finitely many replacements only) worst-case perturbation. Then we record the drop in performance that s' causes. Finally, we take the average drop in performance across all n sentences to be an approximate measure of worst-case performance for the language model on the given probing task.

coPOS perturbations Given an $l > 0$ word long sentence s , we formulate a method to obtain a perturbed sentence s' , where $\tau \leq l$ words in s are replaced whilst keeping the syntax of the original input largely preserved.² Our procedure is sketched in Algorithm 1.

We replace each candidate word in s with one drawn from the WordNet synonym graph [41]. We further ensure that a perturbation is, among the input-perturbation pairs generated by a WordNet replacement, the one that minimizes the syntactic distance of the tree representations while maximizing the Hamming distance between the actual sentences, i.e., the number of words that are actually perturbed. The syntactic distance of each pair of inputs and perturbations is computed via the Stanza dependency parser [49], while the Hamming distance between two sentences is the number of word positions in which two words are different.³ In practice, for each input, $b > 0$ sentences are generated by perturbing τ words via WordNet (line 3):

the syntactic distance between the dependency tree of each input/perturbation pair is computed, alongside their Hamming distance (line 6), which could be less than τ if WordNet does not return a viable substitution, and only the sentence that minimizes the syntactic distance while maximizing the Hamming is used to test the representation’s robustness.

While this procedure is designed to preserve syntax between s and s' , the semantics in general is not: though one may want to introduce further constraints on the replacement procedure to ensure the semantics is preserved, our primary intent is to assess robustness against syntax manipulations. We will show in the experiments that, even for these simple proxy syntax-preserving perturbations,

²WordNet synonyms, or any similar technique, are specifically crafted to maintain the syntactic structure of word replacements. However, it is important to acknowledge that no technique can offer an absolute guarantee of preserving syntactic integrity when replacing a word in a sentence with a generic alternative.

³As two sentences in an input-perturbation pair have the same number of words, we do not need to rely on the Levenshtein distance.

Algorithm 1 coPOS perturbations.

Require:

$s, b, \tau, \text{WordNet}(\cdot, \cdot), \text{dist}_{\text{ham}}(\cdot, \cdot),$
 $\text{dist}_{\text{syntax}}(\cdot, \cdot)$

Ensure: A coPOS perturbation.

```

1:  $s^*, d_h^*, d_t^* \leftarrow (s, 0., \text{inf})$ 
2: for  $j \in [1, \dots, b]$  do
3:    $s' \leftarrow \text{WordNet}(s, \tau) \triangleright$  Perturb  $\tau$ 
   random words in  $s$  with synonyms
4:    $d_h \leftarrow \text{dist}_{\text{ham}}(s, s')$ 
5:    $d_t \leftarrow \text{dist}_{\text{syntax}}(s, s')$ 
6:   if  $d_h > d_h^* \wedge d_t < d_t^*$  then
7:      $s^*, d_h^*, d_t^* \leftarrow (s', d_h, d_t)$ 
8:   end if
9: end for
10: return  $s^*$ 

```

a linguistic representation’s performance degrades sensibly and in some cases it is comparable with random guessing, which indicates that this perturbation scheme is powerful enough to benchmark current language models. Further, this method has a clear advantage in terms of simplicity and computational efficiency as multiple word substitutions can be parallelized. We sketch the aforementioned procedure in Figure 5 (left).

coCO perturbations Our results are complemented by experiments with coCo perturbations (Def. 7), which consist of generating τ replacements via a conditioned LLM, as explained in Def. 7. While we employ GPT-2 [50] for generating a replacement, any generative LLM, thus including masked LLMs such as BERT or RoBERTa [11], is suitable for this task. Further implementation details are provided in the Appendix, while the process of coCO perturbation is sketched in Figure 4.

Baseline perturbations Finally, we add a baseline perturbation method that involves substituting $\tau > 0$ words in a sentence with random replacements from the English vocabulary. In this case, the syntactic consistency of a sentence is not guaranteed to be maintained, and thus serves as a base case for our analysis.

5.1 Validating the Perturbation Methods

In this section, we report the results of the validation process of the coPOS and the coCO perturbation methods. For each perturbation method, and for each dataset that we then employ in the experimental evaluation, we calculate the syntactic distance between a sentence syntax tree and a perturbed candidate: the distance between trees is automatically computed as the minimum number of operations of addition and deletion of a node to turn a tree into another, via Stanza [2] dependency parses. While in this work we report the results regarding the distance between dependency trees, as that is the representation provided by the CoNLL format, as well as that employed in [38], our code permits to compute distance between sentences via their constituency representation.

Examples of the perturbed syntax tree of a sentence from the Ud-English-Pud dataset are shown in Figure 6 for each perturbation method. In Figure 7, we report, for each of the 6 dataset used in the experimental evaluation, the syntactic distance between trees, for the coPOS, coCO, and baseline method, with varying perturbation budget τ equal to 1, 2, and 3. We further compute the average distance between pairs of sentences randomly sampled from each dataset, and for which we expect the distance between trees to be higher than for any other method.

As one can notice from Figure 7, the coPOS method induces the least changes in a syntax tree, and it is thus expected to disrupt the performance of the probing tasks only if the representations are inherently brittle. On the other hand, both coCO and baseline are expected to challenge a probe’s capacity to correctly represent a sentence’s syntactic information. We will show with the

proxy coPOS method that probes, and in turn their representations, are very brittle to syntax-preserving manipulations.

Finally, we show examples of perturbations that our methods produce. In Figure 8, we report example sentences that induce, according to Stanza [49], a high degree of disruption in the dependency syntax tree representation, with those produced by the coCO perturbation method the most disrupted for each input/perturbation pair (not counting the baseline, which almost surely disrupts the syntactic tree of a sentence through random replacements). On the other hand, the coPOS perturbation method can be seen to preserve the structure of each sentence. In Figure 9, we present examples of linguistically interesting perturbations, which do not induce the maximum syntactic disruption.

6 Experiments

We implement and empirically validate our framework by demonstrating how it can provide insights into the robustness of language models. We start with details on linguistic representations, datasets, and probing task models. We then discuss context-free and context-dependent linguistic representations, the effect of latent feature depth, and the duration of fine-tuning.⁴ Finally, we summarise the results of applying our framework in these settings.

6.1 Experimental Setting

Datasets and metrics We assess the syntactic robustness of different language models using the probing tasks of *syntax reconstruction*, *POS-tagging*, *root identification* and *tree-depth estimation* on 6 datasets from the Universal Dependencies collection [45]. We chose datasets standardized according to the CONLL format [19], which consists of sentences split into words, with each indexed and annotated with multiple syntactic information such as POS-tags and the relationship with other words/tokens. An example of a sentence in the CONLL format, with relationship tags between words and part-of-speech tags used to build the ground truth for our probing tasks, is given in Figure 3.

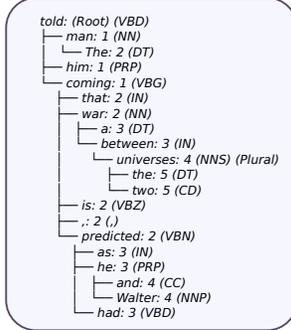
We measure the performance on syntax reconstruction in terms of the ‘undirected unlabeled attachment score’ (UUAS), i.e., the fraction of edges in the ground truth syntax tree that is correctly predicted by a model, the ‘same distance ratio’ (SDR), i.e., the number of times a model correctly guesses the distance between each pair of words in the ground truth syntax tree, and the Spearman correlation (used in [38]), which summarizes the strength of the relationship between the matrix representation of the original vs. reconstructed syntax tree. With regards to POS-tagging, we evaluate a model in terms of the accuracy on estimating the correct POS-tag as shown in Figure 10 (top). On root identification (see Def. 4), we use the accuracy of correct vs. wrong estimates, while

⁴Further details on the implementation, the architectures, the training procedure and the configurations are provided in the Appendix.

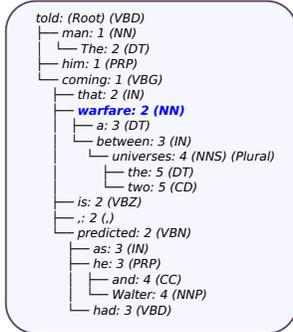
input sentence

'The man told him that a war between the two universes is coming, as he and Walter had predicted;'

original

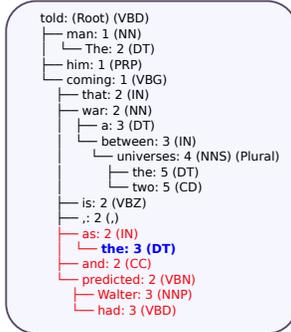


coPOS



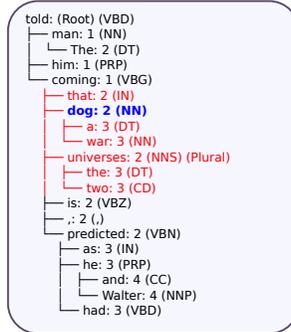
syntactic distance: 0.

coCO



syntactic distance: 0.289

baseline



syntactic distance: 0.263

Figure 6: Comparison of the disruption induced on the dependency syntax tree by different perturbation methods, along with the syntactic distance between trees. The representation of each dependency syntax tree has been compacted to make the effect of the perturbation methods clear, yet it is equivalent to that of Figures 2 and 4. The example sentence belongs to the Ud-English-Pud dataset, and the perturbations are actual perturbations induced by our methods. In blue, the single word that has been perturbed, while in red the perturbation induced by such perturbation on the tree.

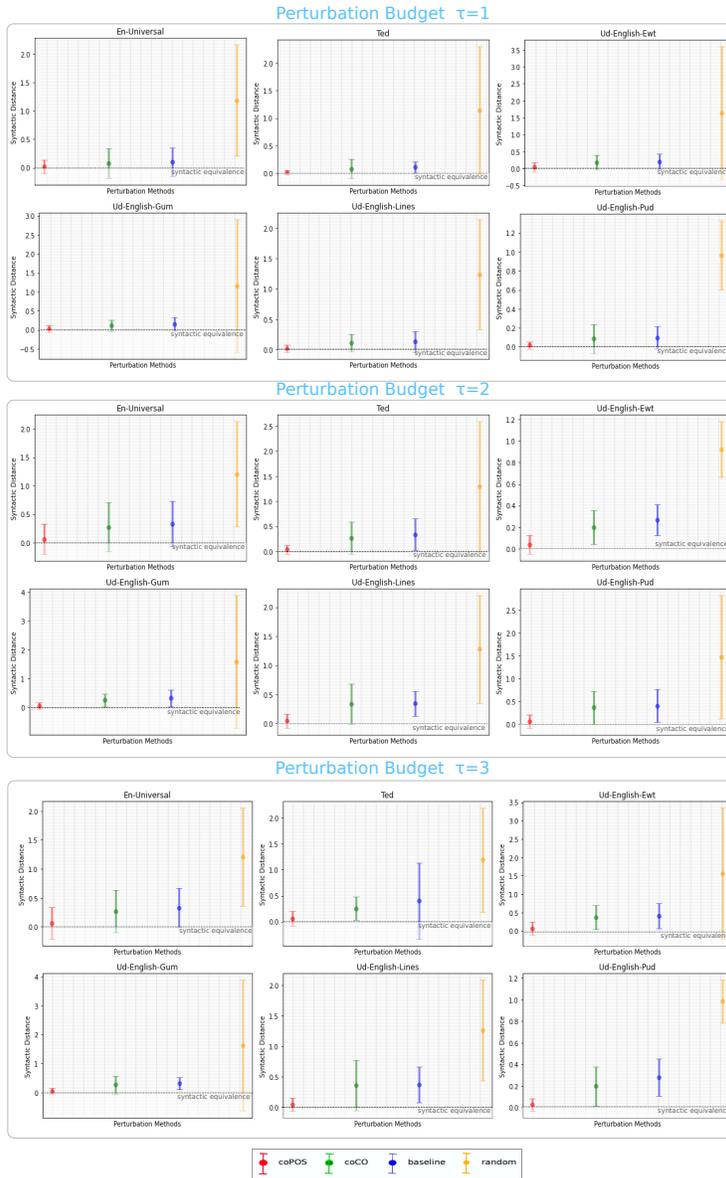


Figure 7: Tree distance, measured with Stanza, between an input and its perturbed version, for different datasets and perturbation budgets: results are averaged over the entire dataset. The coPOS perturbation method (red) induces almost no disruption to a perturbation’s syntax tree, being always close to the level of syntactic equivalence, while injection of random words (blue) and coPOS perturbations (green) both induce some noticeable disruption. The disruption induced by comparing the syntax tree of two randomly picked up sentences that belong to the same dataset is reported for further comparison (orange).

Ted		
[src]	<i>And those are usually the basic scientists , The bottom is usually the surgeons .</i>	
[coCO]	<i>And those are usually the ones that , The bottom line usually the surgeons .</i>	(0.875)
[coPOS]	<i>And those are usually the basic scientist , The bottom is usually the surgeons .</i>	(0.875)
UD-English-Ewt		
[src]	<i>Drs. Ali work wonders .</i>	
[coCO]	<i>Drs. work in the</i>	(2.5)
[coPOS]	<i>Drs. Ali oeuvre wonders .</i>	(2.0)
Ud-English-Gum		
[src]	<i>Environment Canada spokeswoman Sujata Raisinghani told CBC News the department will look into the incident .</i>	
[coCO]	<i>Canada , Sujata Raisinghani told CBC News the agency will be into the incident .</i>	(1.428)
[coPOS]	<i>surround Canada spokeswoman Sujata Raisinghani told CBC News the department will look into the incident .</i>	(0.714)
Un-English-Lines		
[src]	<i>Break with a Banshee by Gilderoy Lockhart</i>	
[coCO]	<i>Break with the best by the Lockhart</i>	(0.428)
[coPOS]	<i>prison-breaking with a banshie by Gilderoy Lockhart</i>	(0.714)
Un-English-Pud		
[src]	<i>However , they were intercepted and had to do battle in Freeman , close to the Hudson River .</i>	
[coCO]	<i>However , they were able and had to do battle with Freeman , close to the Hudson River .</i>	(0.266)
[coPOS]	<i>However , they were intercepted and had to do struggle in Freeman , finis to the Hudson River .</i>	(0.333)
En-Universal		
[src]	<i>Manufacturers Hanover had a loss due to a big reserve addition .</i>	
[coCO]	<i>Manufacturers Hanover , a loss due to a big reserve of .</i>	(11.0)
[coPOS]	<i>producer Hanover had a loss due to a big taciturnity addition .</i>	(10.0)

Figure 8: Examples of sentences and worst-case coCO and coPOS perturbations that are reported in our experiments to highly disrupt the dependency syntax tree according to Stanza [49] (the syntactic distance between the original and perturbed sentence is shown on the right). For each of the 6 CoNLL datasets, we show the original sentence on top. For coCO, perturbed words are highlighted in red and replacements with empty words (which are allowed from the vocabulary) are denoted with a red rectangle ■. For coPOS, perturbed words are highlighted in blue. Results refer to the perturbation regime with $\tau = 3$, i.e., where at most 3 words per-sentence are perturbed.

UD-English-Ewt

[src]	<i>I loved the atmosphere here and the food is good , however the tables are so close together .</i>	
[coCO]	<i>I loved the atmosphere here . the food is good . however the tables are so close together .</i>	(0.684)
[coPOS]	<i>I loved the aura here and the nutrient is good , however the tables are so close together .</i>	(0.0)

Ud-English-Gum

[src]	<i>The purple spheres represent atoms of another element .</i>	
[coCO]	<i>The purple and gold atoms are another element .</i>	(0.666)
[coPOS]	<i>The purple spheres represent atoms of another constituent .</i>	(0.0)

Un-English-Lines

[src]	<i>Can't anyone help you ?</i>	
[coCO]	<i>Can't anyone else you know</i>	(1.0)
[coPOS]	<i>Can't anyone service you ?</i>	(0.0)

Un-English-Pud

[src]	<i>Meanwhile , his place in tribune was occupied by Marco Antonio , who held the position until December .</i>	
[coCO]	<i>Meanwhile , his wife in tribune , occupied by Marco Antonio , who held the position until December .</i>	(0.473)
[coPOS]	<i>Meanwhile , the place in tribune was occupied by the Antonio , who held the position until dec .</i>	(0.157)

En-Universal

[src]	<i>But the report says : The only way sex is sex between uninfected partners .</i>	
[coCO]	<i>But the report says that The only way sex is sex between uninfected partners .</i>	(2.0)
[coPOS]	<i>But the reputation says : The only safe sex is sex between uninfected partners .</i>	(0.0)

Ted

[src]	<i>A windpipe cell already knows it 's a windpipe cell .</i>	
[coCO]	<i>A windpipe is a knows it 's a windpipe cell .</i>	(0.272)
[coPOS]	<i>A trachea cubicle already knows it 's a windpipe cubicle .</i>	(0.0)

Figure 9: Examples of linguistically interesting sentences and perturbations, along with their syntactic distances (right) as calculated with Stanza [49]. For each of the 6 CoNLL datasets, we report the original sentence on top. For coCO, perturbed words are highlighted in red), while for coPOS in blue. Results refer to the perturbation regime with $\tau = 3$, i.e., where at most 3 words per-sentence are perturbed.

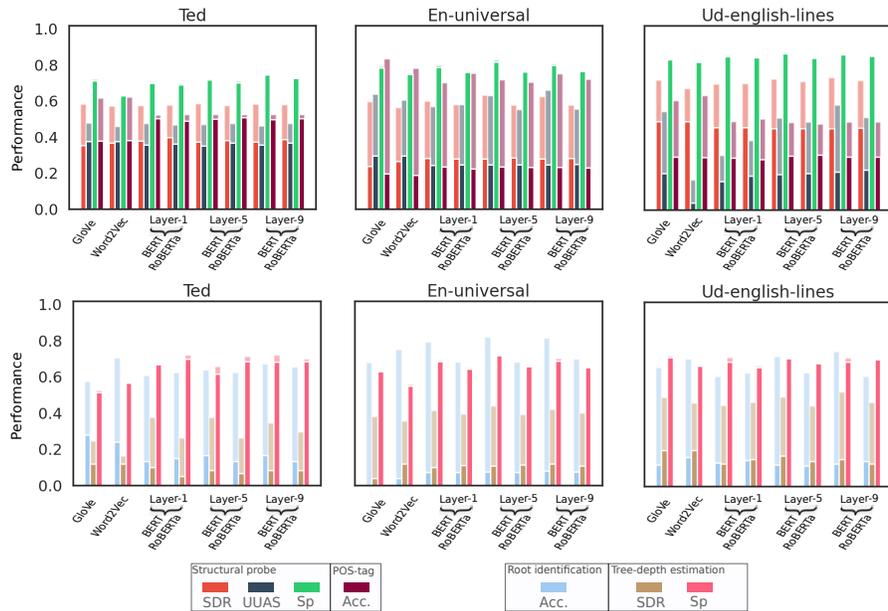


Figure 10: Performance of different linguistic representations on *syntax reconstruction* and *POS-tagging* probing tasks (top) and on *root identification* (accuracy metric) and *tree-depth estimation* (SDR and Spearman metric) probing tasks (bottom). For all plots, the performance of the probing tasks is reported as shaded bars, with the performance for the perturbed representation shown as a solid overlapping bar: the results refer to the case where the coPOS perturbation budget τ is equal to 3 (i.e., at most 3 words per-sentence are perturbed). Regardless of the embedding/representation employed, we observe severe brittleness of the syntactic representations.

on tree-depth estimation (Def. 5) we use the SDR, along with the Spearman correlation, so as not to penalize models that do not infer the label exactly.

Syntactic robustness (Def. 9) is measured in terms of the drop in the performance of a model when targeted with a coPOS or a coCO perturbation, e.g., ΔUUAS represents the drop of the UUAS metric on the syntax robustness task, comparing the performance on the original sentence with its perturbed version. Regarding POS-tags, we convert the drop of accuracy into the more intuitive difference between words correctly guessed on the original sentence compared to its perturbed version, denoted ‘# Words Adv’ in Table 1.

Models and probing tasks We perform our analyses on 4 linguistic representations, of which 2 are context-free, namely GloVe [46] and Word2Vec [40], and 2 context-dependent, namely, BERT [11] and RoBERTa [32]. As LLMs employ deep attention-based architectures, we perform experiments on sentences distilled from the -5^{th} (i.e., the fifth counting from the most external), the -9^{th} and the last (i.e., -1^{th} or the output) hidden layer of a representation. While in [38] it was observed that the most hidden layers are those that perform the best on syntactic tasks, we provide results also for an intermediate and the last hidden layer.

For each probing task (in our setting, syntax reconstruction, POS-tagging, root identification and tree-depth estimation, as per Def. 2, 3, 4 and 5), we stack a deep neural network on top of a linguistic representation ψ^θ , thus obtaining a set of models $\{f_1(s), \dots, f_m(s)\}$: we optimize each model f_i via supervised learning on the i -th task \mathbf{T}_i , leaving the representation’s parameters θ fixed. We note that the measures of performance $\{\mathcal{L}_1, \dots, \mathcal{L}_m\}$ vary from a probing task to another, as we detail in the experimental evaluation.

When training the probing task models, we searched for a common architecture that achieves high performance for each of the four language models. We tested fully-connected, convolutional, and recurrent neural networks, and found that fully-connected (FC) probing models had the best performance across the language models. For each combination of datasets, probing tasks, models, perturbation methods (coPOS, coCO), and for a varying perturbation budget τ , we train a 3-layer deep FC network with a varying number of parameters in the order of 10M. In this sense, both the static and dynamic representations are kept fixed (i.e., their parameters are ‘frozen’ at training time), not to invalidate the scope of the probing tasks and to allow full reproducibility of the results. This experimental evaluation accounts for approximately 900 distinct cases.

6.2 Empirical Evaluation of Syntactic Robustness

We now report the results of our robustness evaluation; in particular, we can quantify the syntactic robustness of the representation ψ^θ according to Def. 9.

Performance on probing tasks In Figure 10 we observe that, across the structural probe task and the POS-tag, all the models have similar performances, with an average POS-tag accuracy around 0.8 and syntax reconstruction SDR

1.1 Robustness

	Syntax Reconstruction			POS-tagging
	Δ SDR	Δ UUAS	Δ Sp	Δ Acc.
GloVe	0.2066 \pm 0.1243	0.2444 \pm 0.1298	0.0062 \pm 0.0032	6.4897 \pm 3.0856
Word2Vec	0.1553 \pm 0.1161	0.1145 \pm 0.1004	0.0052 \pm 0.0048	6.831 \pm 2.2698
BERT layer -1	0.2011 \pm 0.0784	0.1607 \pm 0.0835	0.0032 \pm 0.0077	3.0803 \pm 2.9857
RoBERTa layer -1	0.193 \pm 0.0808	0.1752 \pm 0.1117	0.0019 \pm 0.0039	4.1293 \pm 3.3021
BERT layer -5	0.2287 \pm 0.0817	0.2451 \pm 0.1212	0.005 \pm 0.003	3.243 \pm 3.0814
RoBERTa layer -5	0.2038 \pm 0.0758	0.2086 \pm 0.1052	0.0024 \pm 0.0379	3.0607 \pm 3.0132
BERT layer -9	0.2307 \pm 0.0838	0.281 \pm 0.1142	0.0035 \pm 0.0037	3.544 \pm 3.2826
RoBERTa layer -9	0.2045 \pm 0.0763	0.2148 \pm 0.1116	0.0017 \pm 0.0023	3.4723 \pm 3.1265

1.2 Robustness

	Root Identification	Tree Depth Estimation	
	Δ Acc.	Δ Acc.	Δ Sp
GloVe	0.4387 \pm 0.1953	0.2663 \pm 0.235	0.0174 \pm 0.0189
Word2Vec	0.5251 \pm 0.1123	0.2582 \pm 0.2712	0.0093 \pm 0.0285
BERT layer -1	0.5015 \pm 0.2135	0.3495 \pm 0.2139	0.0209 \pm 0.0159
RoBERTa layer -1	0.5286 \pm 0.1661	0.3193 \pm 0.2348	0.0261 \pm 0.0126
BERT layer -5	0.6039 \pm 0.1383	0.3314 \pm 0.2401	0.0086 \pm 0.0158
RoBERTa layer -5	0.5211 \pm 0.168	0.2829 \pm 0.2524	-0.0062 \pm 0.0379
BERT layer -9	0.612 \pm 0.1216	0.3256 \pm 0.2423	0.0159 \pm 0.018
RoBERTa layer -9	0.5151 \pm 0.1773	0.3312 \pm 0.2365	-0.0002 \pm 0.0222

1.3 Distance/Similarity Metrics

	ℓ_2 -norm distance	Cosine similarity
GloVe	0.023 \pm 0.0032	0.9352 \pm 0.0132
Word2Vec	0.0038 \pm 0.0005	0.9231 \pm 0.0161
BERT layer -1	0.0299 \pm 0.0036	0.8994 \pm 0.0221
RoBERTa layer -1	0.0103 \pm 0.0013	0.9835 \pm 0.0039
BERT layer -5	0.0388 \pm 0.0045	0.926 \pm 0.0196
RoBERTa layer -5	0.0259 \pm 0.0032	0.9764 \pm 0.0059
BERT layer -9	0.0377 \pm 0.0045	0.9296 \pm 0.0186
RoBERTa layer -9	0.0188 \pm 0.0018	0.9843 \pm 0.0026

Table 1: Relationship between the syntactic robustness metrics for four probing tasks on coPOS perturbations with budget $\tau = 2$ (top and middle row) and the distance between pairs of perturbed and original inputs measured via cosine similarity and ℓ_2 -norm distance (bottom row). The accuracy drop of the POS-tag task is reported as the number of words correctly guessed in both cases. The reported standard deviation is measured by averaging over the 6 training corpora. Whilst the distance (similarity) between inputs and perturbations is low (high), we observe that all embeddings/representations are brittle to syntax-preserving perturbations.

2.1 Robustness

	Syntax Reconstruction			POS-tagging
	Δ SDR	Δ UAS	Δ Sp	Δ Acc.
GloVe	0.2098 ± 0.124	0.2616 ± 0.1411	0.139 ± 0.0106	6.743 ± 3.0337
Word2Vec	0.1655 ± 0.1114	0.1232 ± 0.1014	0.0118 ± 0.014	7.071 ± 2.1687
BERT layer -1	0.208 ± 0.0773	0.1782 ± 0.0862	0.0285 ± 0.0177	3.1592 ± 3.0359
RoBERTa layer -1	0.1989 ± 0.0823	0.1951 ± 0.116	0.02 ± 0.0146	4.2887 ± 3.3774
BERT layer -5	0.235 ± 0.082	0.267 ± 0.1293	0.0261 ± 0.0191	3.286 ± 3.1258
RoBERTa layer -5	0.2093 ± 0.0767	0.2319 ± 0.1117	0.0133 ± 0.0126	4.2887 ± 3.3774
BERT layer -9	0.235 ± 0.0859	0.2988 ± 0.154	0.0162 ± 0.0135	3.613 ± 3.3219
RoBERTa layer -9	0.2109 ± 0.0774	0.2378 ± 0.1227	0.0135 ± 0.012	3.561 ± 3.205

2.2 Robustness

	Root Identification	Tree Depth Estimation	
	Δ Acc.	Δ Acc.	Δ Sp
GloVe	0.4987 ± 0.1827	0.293, 0.2024	0.0417, 0.0134
Word2Vec	0.5785 ± 0.1462	0.2915, 0.2444	0.0174, 0.0184
BERT layer -1	0.5526 ± 0.202	0.3863, 0.2054	0.0838, 0.0494
RoBERTa layer -1	0.5449 ± 0.1804	0.3567, 0.2166	0.0993, 0.0531
BERT layer -5	0.6374 ± 0.1483	0.3937, 0.2247	0.0633, 0.0374
RoBERTa layer -5	0.5448 ± 0.1735	0.3267, 0.2338	0.0672, 0.0215
BERT layer -9	0.6293 ± 0.1408	0.3726, 0.2217	0.054, 0.0512
RoBERTa layer -9	0.549 ± 0.1786	0.3613, 0.2317	0.0471, 0.0326

2.3 Distance/Similarity Metrics

	ℓ_2 -norm distance	Cosine similarity
GloVe	0.0344 ± 0.0018	0.8783 ± 0.0271
Word2Vec	0.0059 ± 0.0002	0.8572 ± 0.0331
BERT layer -1	0.0487 ± 0.0063	0.7951 ± 0.0433
RoBERTa layer -1	0.0195 ± 0.0016	0.9597 ± 0.0064
BERT layer -5	0.0652 ± 0.0058	0.8432 ± 0.0316
RoBERTa layer -5	0.0488 ± 0.0037	0.9416 ± 0.0102
BERT layer -9	0.058 ± 0.004	0.8768 ± 0.0173
RoBERTa layer -9	0.0373 ± 0.0024	0.9557 ± 0.0057

Table 2: Relationship between the syntactic robustness metrics for four probing tasks on coCO perturbations with budget $\tau = 2$ (top and middle row) and the distance between pairs of perturbed and original inputs measured via cosine similarity and ℓ_2 -norm distance (bottom row). The reported standard deviation is measured averaging over the 6 training corpora. The accuracy drop of the POS-tag task is reported as the number of words correctly guessed in both cases. Whilst the distance (similarity) between inputs and perturbations is low (high), we observe that all embeddings/representations are brittle to syntax-preserving perturbations.

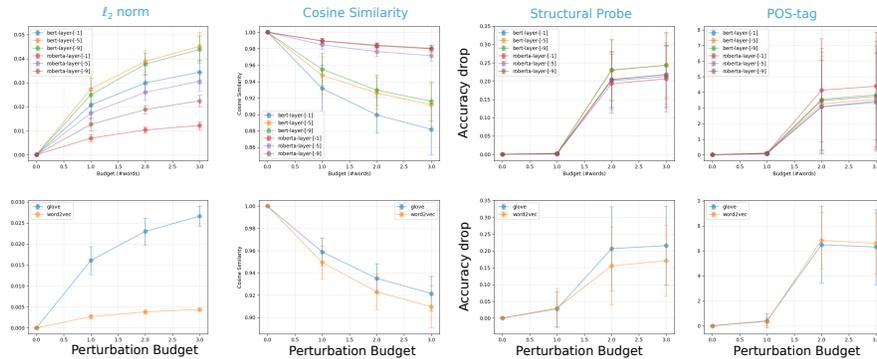


Figure 11: Left: For an increasing perturbation budget τ and the coPOS method, cosine similarity between perturbed and original sentences drops, while the ℓ_2 -norm increases. Right: It is clear that, even with $\tau = 2$ (i.e., at most two words per-sentence are perturbed), the models’ performance already experiences a significant drop (the higher the curve, the worse the model is on a syntactic task). Increasing the perturbation budget only slightly increases a drop of robustness.

around 0.7. Of the context-dependent language models, RoBERTa and BERT are comparable, though we cannot definitively conclude which one is better. Similarly, GloVe slightly outperforms Word2Vec on tree-depth estimation, while Word2Vec is better on the structural probe. Interestingly, we notice that word embeddings are only slightly worse than BERT and RoBERTa. The same trends emerge on root identification and tree-depth estimation, as shown in Figure 10 (bottom): while BERT generally outperforms the competitors and Word2Vec struggles to compete, GloVe is a competitor of both the language models. We conclude with a final remark on the lack of performance gap between GloVe and Word2Vec, as it suggests that pre-training a representation on local and global word co-occurrences [46] does not help syntactic structures to emerge, a controversial yet intriguing discovery.⁵

Robustness on probing tasks In Figure 10, solid bars represent the performance of a trained model subjected to a coPOS perturbation under the perturbation budget of $\tau = 3$: in other words, we generate an approximate worst-case coPOS perturbation given that we can only change at most 3 words in the given sentence. The drop in performance of a model on a task can be inferred via the gap between the solid and shaded bars (the latter corresponding to the unperturbed sentence). We notice that for each metric, excluding the Spearman correlation in tasks 1 and 4, discussed in the Appendix, there is a substantial drop in performance, which suggests that each language model represents a brittle understanding of syntactic concepts. While Spearman is used as a metric in [38],

⁵<http://languagelog.ldc.upenn.edu/myl/PinkerChomskyMIT.html>

in the Appendix, Section D.1, we discuss an interesting finding, validated by substantial empirical evidence, that we believe partially invalidates it as a proper metric to judge the syntactic similarity between trees extracted from linguistic representations. In particular, in the syntax reconstruction task (Figure 10, top) we notice a dramatic decrease in UUAS of more than 50% on any task and any model, while the SDR drop is of around 20 – 30%. Thus, for each language model and dataset, the syntax reconstruction task is now incorrect more often than it is correct. For UUAS, this indicates that our coPOS scheme is able to find syntactically meaningful perturbations which reduce the performance of the model to random guessing. Secondly, we highlight that, for UUAS and SDR, the largest decrease in performance comes for the datasets for which the performance was the highest. This indicates that robust representation of syntax may be at odds with performance. The same considerations are valid for the POS-tag probing task, with the accuracy that drops to a random guess with the perturbation budget τ equal to 1. The coPOS perturbation method degrades the performance on root identification and tree-depth estimation as much as in the previous tasks (also in this case excluding the Spearman correlation, which we remind we discuss in the Appendix). In fact, the performance drops to a random guess on any task and for any representation, which provides evidence that these representations are brittle, and thus not suited to cases of domain shifts.

On the correlation between robustness and sentence similarity In this batch of experiments, we keep track of the farthest distance between an input and its coPOS perturbation (see Def. 8) using the ℓ_2 -norm and the cosine similarity between each pair of input/perturbation. We then measure the performance drop of each model to assess any correlation with the aforementioned measures of distance/similarity. As reported in Table 1, we find that high drops in performance can be caused by perturbations with small ℓ_2 -norm compared to the unperturbed sentence, and conversely high cosine similarity. This confirms that linguistic representations are remarkably brittle even to local perturbations, i.e., those whose representation lies in the proximity of the original input. We replicate the results using the coCO perturbation method (Table 2), which confirm that perturbations extracted via GPT conditioning are farther than the coPOS in the representation space, and equally effective at dismantling a model’s robustness. Similar observations can be made with the baseline perturbation method, as reported in the Appendix, Table 5.

Varying the perturbation budget While we have already shown that a small perturbation budget τ exposes a representation to effective performance-degrading attacks, we now investigate the relationship of the distance/similarity metrics and robustness with respect to a varying number of coPOS perturbations. In Figure 11, both the cosine similarity and the ℓ_2 -norm behave monotonically as τ increases. Deeper LLM’s layers are less affected by an increased perturbation budget, with BERT less prone than RoBERTa to maintaining the internal consistency of its representation of the original and perturbed sentences

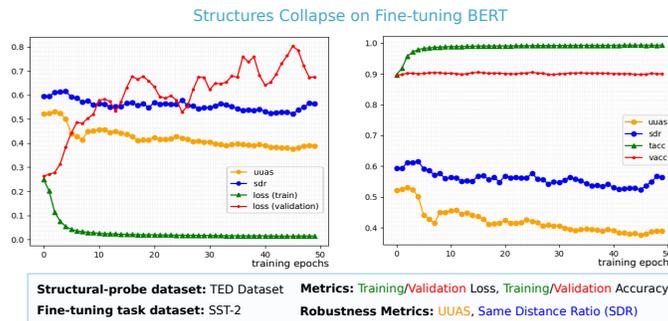


Figure 12: BERT model fine-tuned (and finally, overfitted) on the SST dataset, while its representations are used to train a model to solve the structural probe task. Train and validation losses (left) and accuracies (right) pertain to the fine-tuning task (SST), while SDR and UUAS show the performance on the structural probe. The syntactic metrics degrade as the fine-tuning process proceeds, yet severe over-fitting does not harm syntactic structures.

(Figure 11, top-left). In word embeddings (Figure 11, bottom-left), while the trends of cosine similarity between GloVe and Word2Vec are similar, the l_2 -norm of Word2Vec does not change as much as for GloVe, a sign that in this representation words lie very close to each other w.r.t. the Euclidean distance. When it comes to performance drop (Figure 11, right), static and dynamic representations are comparable as they are both not worse than those induced by a single coPOS perturbation, i.e., with $\tau = 1$. The performance drop on the coCO and the baseline method are reported in the Appendix, Figures 14 and 15.

On linear vs. non-linear probes In our experiments, we use ReLU-activated probing tasks, which constitute a stronger model than classical linear probes [10]. There have been arguments recently in favour of non-linear probes, motivated by the fact that linguistic structures are not necessarily encoded linearly by linguistic representations [48, 63]. While we choose to report the results that concern non-linear probes, we conducted the same experiments with linear probes, for which results are released in the code.

Experiments with linear probes are meant to (i) provide a robustness comparison to the ReLU setting; and (ii) contribute to the debate on the effectiveness of linear probes as tools to investigate the internals of LLMs, since non-linear probes attached to linguistic representations may be on one hand more accurate, yet can incur in overfitting and potentially hinder the expressiveness capabilities of a linguistic representation [44].

The effect of fine-tuning and overfitting on syntactic structures We finally conduct an analysis whose primary intent is to understand the effect of counter-fitting [42] and fine-tuning on syntactic structures, respectively for

context-free and context-dependent representations. We train a counter-fitted version of GloVe and we fine-tune, and finally overfit, a BERT-based representation on the SST-2 dataset [57]: differently from previous experiments, we update the weights of the language model (i.e., we do not keep them frozen) to investigate the existence of some form of catastrophic forgetting of the syntactic structures encoded in the model. By performing the robustness analysis introduced in this paper, we observe that fine-tuning negatively affects the performance of both shallow and deep context-dependent representations; despite this, excessive fine-tuning is not significantly harmful as the performance does not collapse even after many epochs the model has overfitted on the training set. The task’s validation loss is informative to prevent overfitting on the structural probe task, while accuracy on the fine-tuning task can be misleading and hide syntax collapse. We sketch the training dynamics, along with the accuracy of the model on the classification task and the structural probe metrics, in Figure 12. We also observe that any metric of a counter-fitted model has inferior performance on any task and any dataset while being equally brittle to coPOS perturbations, thus limiting the utility of counter-fitting on models aimed at capturing different aspects of human linguistics: performance and robustness are in line with those of standard GloVe embedding (i.e., Table 1 and Figure 10); we report an extended evaluation of this phenomenon in the Appendix.

Justification for the linguistic structures collapse In light of the empirical evidence provided in this paper, we conclude with some hypotheses on why good performance of linguistic representations, whether LLMs or standard word embeddings, comes at the cost of brittleness on high-order syntactic tasks. Certainly, the robustness-performance trade-off accounts for the frailty of overfitted probes [33]. On the other hand, the absence *stricto sensu* of adversarial attacks, replaced by coPOS and coCO perturbations, forces us to second-guess the existence of such structures. In high-dimensional spaces, vectors (i.e., words) become progressively harder to distinguish, while at the same time the high-dimensionality allows one to optimize a decision boundary that is overfitted on the training set, but fails poorly on slight variations of an input. The hypothesis that we put forward in this paper, with the aim to stimulate discussion among NLP researches as well as linguists, is that linguistic structures emerge as a process of fitting between static sentences and their syntax trees, granted by rich linguistic representations which nonetheless collapse as soon as the input distribution allows for word substitution, a shift against which human linguistic structures are indeed extremely robust.

7 Conclusion and Future Works

In this work we studied a notion of syntactic robustness for linguistic representations. Robustness is a desirable property of such models, yet we have exhibited the risk of taking their inherent robustness for granted. We gave empirical evidence of severe brittleness of both LLMs and word embeddings when perturbed

via the coPOS and coCO method with restrained perturbation budget. We complemented this by an analysis of the dynamics of robustness w.r.t. overfitting and counter-fitting. In future, we aim to extend this empirical framework to tasks where the syntactic information is not explicitly provided, and further exploring formally such ‘syntactic subspaces’ where common linguistic features lie: the aim is the foundation of a ‘linguistics for Large Language Models’.

8 Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (FUN2MODEL, grant agreement No. 834115).

References

- [1] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*, 2018.
- [2] John Bauer, Chloé Kiddon, Eric Yeh, Alex Shan, and Christopher D. Manning. Semgrex and ssurgeon, searching and manipulating dependency graphs. In *Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023)*, pages 67–73, Washington, D.C., March 2023. Association for Computational Linguistics.
- [3] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.
- [4] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.
- [5] Emily M Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, 2020.
- [6] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357, 2016.
- [7] Noam Chomsky. *Syntactic structures*. De Gruyter Mouton, 2009.
- [8] Noam Chomsky. 153A Minimalist Program for Linguistic Theory. In *The Minimalist Program*. The MIT Press, 12 2014.

- [9] Noam Chomsky et al. *Language and mind*. Cambridge University Press, 2006.
- [10] Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2126–2136. Association for Computational Linguistics, 2018.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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 4171–4186. Association for Computational Linguistics, 2019.
- [12] Xinshuai Dong, Anh Tuan Luu, Rongrong Ji, and Hong Liu. Towards robustness against natural language word substitutions. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.
- [13] Jonathan Dunn. Computational learning of construction grammars. *Language and cognition*, 9(2):254–292, 2017.
- [14] Jonathan Dunn and Harish Tayyar Madabushi. Learned construction grammars converge across registers given increased exposure. In Arianna Bisazza and Omri Abend, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning, CoNLL 2021, Online, November 10-11, 2021*, pages 268–278. Association for Computational Linguistics, 2021.
- [15] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, 2020.
- [16] Edward Gibson, Richard Futrell, Steven P Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. How efficiency shapes human language. *Trends in cognitive sciences*, 23(5):389–407, 2019.
- [17] Yoav Goldberg. Assessing bert’s syntactic abilities. *arXiv preprint arXiv:1901.05287*, 2019.
- [18] Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. Exposing the implicit energy networks behind masked language models via metropolis-hastings. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

- [19] Jan Hajic, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, M Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18, 2009.
- [20] Jack Hessel and Alexandra Schofield. 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, 2021.
- [21] John Hewitt and Christopher D Manning. 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, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.
- [22] Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. Achieving verified robustness to symbol substitutions via interval bound propagation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4081–4091. Association for Computational Linguistics, 2019.
- [23] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.
- [24] Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. Certified robustness to adversarial word substitutions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4127–4140. Association for Computational Linguistics, 2019.
- [25] Steven Johnson and Nikita Izhev. A.i. is mastering language. should we trust what it says?, Apr 2022.
- [26] Taeuk Kim, Jihun Choi, Daniel Edmiston, and Sang-goo Lee. Are pre-trained language models aware of phrases? simple but strong baselines for grammar induction. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

- [27] Volodymyr Kuleshov, Shantanu Thakoor, Tingfung Lau, and Stefano Ermon. Adversarial examples for natural language classification problems. *arxiv pre-print*, 2018.
- [28] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.
- [29] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *International Conference on Machine Learning*, pages 1078–1088. PMLR, 2020.
- [30] Tomasz Limisiewicz and David Marecek. Syntax representation in word embeddings and neural networks - A survey. In Martin Holena, Tomáš Horváth, Alica Kelemenová, Frantisek Mráz, Dana Pardubská, Martin Plátek, and Petr Sosík, editors, *Proceedings of the 20th Conference Information Technologies - Applications and Theory (ITAT 2020), Hotel Tyrapol, Oravská Lesná, Slovakia, September 18-22, 2020*, volume 2718 of *CEUR Workshop Proceedings*, pages 40–50. CEUR-WS.org, 2020.
- [31] Xiaodong Liu, Kevin Duh, Liyuan Liu, and Jianfeng Gao. Very deep transformers for neural machine translation. *arXiv:2008.07772*, 2020.
- [32] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*, 2019.
- [33] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.
- [34] Emanuele La Malfa and Marta Kwiatkowska. The king is naked: On the notion of robustness for natural language processing. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11047–11057. AAAI Press, 2022.
- [35] Emanuele La Malfa, Rhiannon Michelmore, Agnieszka M. Zbrzezny, Nicola Paoletti, and Marta Kwiatkowska. On guaranteed optimal robust explanations for NLP models. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 2658–2665. ijcai.org, 2021.

- [36] Emanuele La Malfa, Min Wu, Luca Laurenti, Benjie Wang, Anthony Hartshorn, and Marta Kwiatkowska. Assessing robustness of text classification through maximal safe radius computation. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2949–2968. Association for Computational Linguistics, 2020.
- [37] Christopher Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.
- [38] Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, 2020.
- [39] David Marecek and Rudolf Rosa. From balustrades to pierre vinken: Looking for syntax in transformer self-attentions. In Tal Linzen, Grzegorz Chrupala, Yonatan Belinkov, and Dieuwke Hupkes, editors, *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 263–275. Association for Computational Linguistics, 2019.
- [40] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.
- [41] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.
- [42] Nikola Mrksic, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasic, Lina Maria Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve J. Young. Counter-fitting word vectors to linguistic constraints. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 142–148. The Association for Computational Linguistics, 2016.
- [43] Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. Refining targeted syntactic evaluation of language models. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3710–3723. Association for Computational Linguistics, 2021.
- [44] Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In Anna Korhonen, David R. Traum,

- and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4658–4664. Association for Computational Linguistics, 2019.
- [45] Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1659–1666, 2016.
- [46] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- [47] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- [48] Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. *arXiv preprint arXiv:2004.03061*, 2020.
- [49] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020.
- [50] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [51] Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? *arXiv preprint arXiv:2005.00719*, 2020.
- [52] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*, 2020.
- [53] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.
- [54] Herbert Rubenstein and John B Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

- [55] Yikang Shen, Zhouhan Lin, Athul Paul Jacob, Alessandro Sordani, Aaron C. Courville, and Yoshua Bengio. Straight to the tree: Constituency parsing with neural syntactic distance. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1171–1180. Association for Computational Linguistics, 2018.
- [56] Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2888–2913. Association for Computational Linguistics, 2021.
- [57] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.
- [58] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.
- [59] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394, 2010.
- [60] Priyan Vaithilingam, Tianyi Zhang, and Elena L. Glassman. Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In Simone D. J. Barbosa, Cliff Lampe, Caroline Appert, and David A. Shamma, editors, *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022, Extended Abstracts*, pages 332:1–332:7. ACM, 2022.
- [61] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2153–2162. Association for Computational Linguistics, 2019.

- [62] Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the Web Conference 2021*, pages 1785–1797, 2021.
- [63] Jennifer C White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. A non-linear structural probe. *arXiv preprint arXiv:2105.10185*, 2021.
- [64] Ying Xu, Xu Zhong, Antonio Jose Jimeno Yepes, and Jey Han Lau. Elephant in the room: An evaluation framework for assessing adversarial examples in nlp. *arXiv preprint arXiv:2001.07820*, 2020.

A Algorithm for Evaluating Robustness

Average distance of the farthest representation As a measure of distance between sentences, we rely on the ℓ_2 -norm and the cosine similarity, whose usage is widespread in NLP robustness [22, 24], noting that other measure are also possible [12, 28, 35]. We provide a sketch of the procedure in Alg. 2. It estimates the average distance of the farthest perturbation in the representation space of ψ^θ , w.r.t. a measure of distance like an ℓ -p norm. The procedure easily accounts for measures of similarity, like the cosine similarity, by changing *max* with *min* at line 9.

Average worst-case syntactic robustness Complementary to Algorithm 2, Algorithm 3 permits to evaluate syntax robustness of a linguistic representation ψ^θ . For each pair of a model and a probing task (f_i, \mathbf{T}_i) , we draw a pool of sentences S_i from a CONLL corpus of choice (lines 2, 4); then, for each sentence $s \in S_i$, we compute a set of coPOS perturbations (lines 7, 8). The ratio between the number of sentences in S_i and the perturbations depends on the budget parameter k , as well as the number of words per sentence τ that are perturbed; e.g., with $\tau = k = 1$, each sentence in S_i is perturbed once via a single-word substitution. We rely on WordNet [41] and its graph of synonyms to draw, for each sentence s , a substitute s' that is syntax-preserving. We exemplify this process in Figure 5. We then quantify the drop of performances of f_i on the original vs. perturbed input representations via the performance measure \mathcal{L}_i (line 10). As we aim for a measure of robustness against perturbations, we return for each sentence $s \in S_i$ the worst-case drop induced by any of the s' generated previously, averaged over the number of test cases (lines 11, 13 and 15). We can now pair the measure of robustness with the ϵ -distance between S and the set of worst-case perturbations S' (Def. 8) as the largest deviation of a pair of input/perturbation w.r.t. the representation ψ^θ .

B Experiments

As described in the experimental section, we perform our analyses on 4 linguistic representations, namely GloVe [46], Word2Vec [40], BERT [11] and RoBERTa [32].

We searched for the best architecture in terms of performance on the four tasks on the 6 corpora of interest: we found that the convolutional architecture performs poorly on the tasks, while recurrent architectures are competitive with fully connected, but introduce additional computation overhead and the intrinsic inductive bias of the recurrent gates.

We applied the least amount of pre-processing to the input texts, i.e., we split sentences into words – both for the context-free and the context-dependent representations. As the size of the input matrix for the *syntax reconstruction* task grows quadratically with the input length, we cut (pad) the sentences longer (shorter) than 20 words.

Algorithm 2 Estimate the average distance of the farthest perturbations w.r.t. a representation ψ^θ .

Require: $\psi^\theta(\cdot), S, sub(\cdot), k, dist(\cdot, \cdot)$

Ensure: Average distance of the farthest representation of $\psi_{s \sim S}^\theta(s)$ against

```

     $sub_{s' \sim S'}(s')$ 
1:  $rob = 0$ .
2: for  $s \in S$  do
3:    $x \leftarrow \psi^\theta(s)$ 
4:    $worst = 0$ 
5:   for  $j$  in  $[1, \dots, k]$  do
6:      $s' \leftarrow sub(s)$ 
7:      $x' \leftarrow \psi^\theta(s')$   $\triangleright$  Obtain the representation of a perturbed input
8:      $d = dist(x, x')$   $\triangleright$  Calculate the distance between input and
        perturbation
9:      $worst = max(worst, d)$   $\triangleright$  Worst-case as farthest perturbation
10:  end for
11:   $rob += worst$ 
12: end for
13: return  $rob/|S|$   $\triangleright$  Average over each worst-case.

```

Algorithm 3 Estimate the average worst-drop of robustness of ψ^θ on probing tasks \mathbf{T} .

Require: $\psi^\theta(\cdot), \{\mathbf{T}_1, \dots, \mathbf{T}_m\}, \{f_1(s), \dots, f_m(s)\}, \{\mathcal{L}_1, \dots, \mathcal{L}_m\}, sub(\cdot), \tau, k$

Ensure: Average worst-drop of robustness of $\psi_{s \sim S}^\theta(s)$ against $sub_{s' \sim S'}(s')$ on each task $\{\mathbf{T}_1, \dots, \mathbf{T}_m\}$

```

1:  $Drop = \{\}$   $\triangleright$  Will contain the average worst-case drop per task  $\mathbf{T}_i$ 
2: for  $i \in [1, \dots, m]$  do
3:    $drop = 0$   $\triangleright$  Average worst-case drop of robustness
4:    $S_i \leftarrow data(\mathbf{T}_i)$   $\triangleright$  Get data from each task
5:   for  $s \in S_i$  do
6:      $d = 0$ .
7:     for  $j$  in  $[1, \dots, k]$  do
8:        $s' \leftarrow sub(s, \tau)$   $\triangleright \tau$  words are perturbed to obtain  $s'$  from  $s$ 
9:        $x, x' \leftarrow \psi^\theta(s), \psi^\theta(s')$   $\triangleright$  Input/perturbation pairs
10:       $\Delta d = \mathcal{L}_i(f_i(x), f_i(x'))$   $\triangleright$  Drop of robustness between input and
        perturbation
11:       $d = max(d, \Delta d)$   $\triangleright$  Get the case that minimizes syntax robustness
12:    end for
13:     $drop += d$ 
14:  end for
15:   $Drop \leftarrow^{\pm} drop/|S_i|$ 
16: end for
17: return  $Drop$ 

```

As regards the training, models have been trained until they start overfitting the data, i.e., we end the procedure via early-stop.

For further details on the architectures, e.g., the number of parameters, the initialization, etc., we provide the logs of each experiment as part of the code to reproduce the experiments.

C Logs for all the experiments of the paper (and more)

We make the code available, which alongside the instructions to run it can be found at <https://github.com/EmanueleLM/emergent-linguistic-structures>. We also provide all the logs to reconstruct the results we present in this paper. Files are stored as plain text in the `.zip` archive, under the `robust-linguistic-structures\verify\MLM_internals\syntax-integrity\results` folder.

D Experiments: Additional Results

D.1 On Spearman correlation’s drop as a measure of robustness

While the Spearman correlation metric is used in [38] to measure the capacity of a model to represent a sentence’s syntax tree, we found two reasons why that this metric is not indicative of the robustness of a model. To illustrate the first reason, we present an example: consider the syntax tree presented in Figure 13 (left), which encodes the distances between the sentence $[0, 1, 2, 3]$. Its distance metric between nodes is also reported in Figure 13 (right). Now, suppose that a perturbation makes the tree change from that on the left to that on the right, with the relative distance matrix reported on the right. The Spearman coefficient between the two trees is, according to a standard implementation⁶, 0.65; however, we could not consider a model, which outputs the second tree as a result of a sentence manipulation, robust as the difference between the syntax structures is not reflected in the value of the coefficients.

Furthermore, we discovered that for all the experiments, both the coPOS and coCo perturbation methods induce a shift of each word-to-word distance towards negative values. One might argue that the shift can preserve the information of the original tree, e.g., any word-to-word distance is for example shifted by a negative constant c : in that case, one could revert the predicted tree with a simple mathematical operation. Unfortunately, this is not possible, as the shift varies sensibly from word to word while maintaining the correlation between the original and the predicted trees high.

⁶<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html>

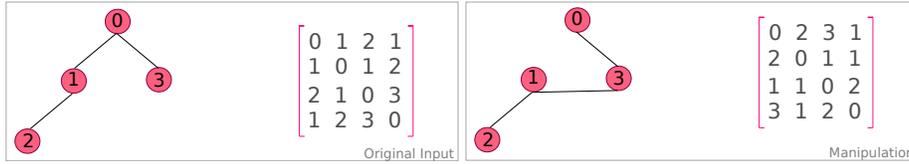


Figure 13: Example of Spearman correlation and its lack of consistency when judging syntax manipulations.

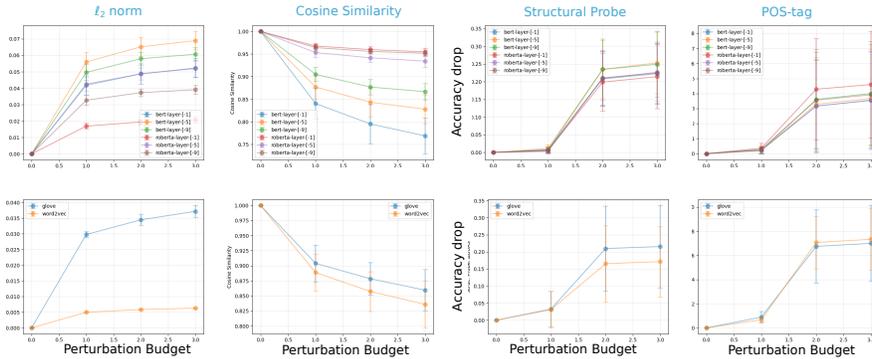


Figure 14: Left: for an increasing perturbation budget τ and the coco method, cosine similarity between perturbed and original sentences drops, while the ℓ_2 -norm increases. Right: It is clear that, even with $\tau = 2$ (i.e., at most two words per-sentence are perturbed), the models’ performance already experiences a significant drop (the higher the curve, the worse the model is on a syntactic task). Increasing the perturbation budget does not lead to a significant drop of robustness.

D.2 coCO and Baseline Perturbations

In Figure 14, we report the cosine similarity and the ℓ_2 drop between pairs of inputs and perturbations (left), and the relative drop of performances on the probing tasks (right), when inputs are targeted by coCO perturbations. The same results, but relative to the baseline perturbation method, are reported in Figure 15. Results are comparable to those with coPOS perturbations, thus strengthening the hypothesis that linguistics structures, if present, are brittle to slight, syntax-preserving perturbations.

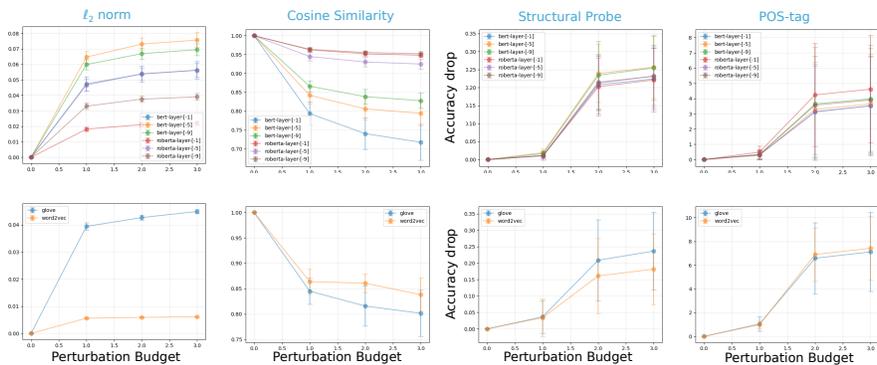


Figure 15: Left: For an increasing perturbation budget τ and the baseline method, cosine similarity between perturbed and original sentences drops, while the ℓ_2 -norm increases. Right: It is clear that, even with $\tau = 2$ (i.e., at most two words per-sentence are perturbed), the models’ performance already experiences a significant drop (the higher the curve, the worse the model is on a syntactic task). Increasing the perturbation budget does not lead to a significant drop of robustness.

D.3 The Effect of Fine-tuning and Over-fitting on Syntactic Structures

We report the performances of GloVe counter-fitted models on the four probing tasks in Tables 4 and 3. As it can be observed, the performance on each task is inferior to standard GloVe embedding, suggesting that counter-fitting is harmful to the syntactic structures encoded in the representation. When targeted with perturbations, the syntactic structures of GloVe counter-fitted models collapse, thus, we can conclude that counter-fitting does not improve the syntactic capabilities of a model, yet it is equivalently as brittle as GloVe.

3.1 Robustness

	Syntax Reconstruction			POS-tagging
	Δ SDR	Δ UAS	Δ Sp	Δ Acc.
TED	0.167	0.0872	0.0011	4.470
En-Universal	0.2360	0.29931	0.0056	9.456
Ud-English-ewt	0.1463	0.3124	0.0106	6.951
Ud-English-gum	0.1678	0.3252	0.0025	3.253
Ud-English-lines	0.2599	0.2981	0.0022	7.402
Ud-English-pud	0.0574	0.00612	0.0010	5.746

3.2 Robustness

	Root Identification	Tree Depth Estimation	
	Δ Acc.	Δ Acc.	Δ Sp
TED	0.3753	0.0727	0.0165
En-Universal	0.5288	0.2215	-0.0011
Ud-English-ewt	0.816	0.7751	0.0102
Ud-English-gum	0.389	0.3065	0.0028
Ud-English-lines	0.4493	0.314	0.0037
Ud-English-pud	0.3894	0.2583	-0.0189

3.3 Distance/Similarity Metrics

	Cosine similarity	ℓ_2 -norm distance
TED	0.881	0.006
En-Universal	$\approx 1.$	0.0033
Ud-English-ewt	0.9295	0.005
Ud-English-gum	$\approx 1.$	0.005
Ud-English-lines	$\approx 1.$	0.0058
Ud-English-pud	0.881	0.0058

Table 3: Robustness of GloVe counter-fitted models, with an analysis, w.r.t. each dataset (one per row), of the relationship between the syntactic robustness metrics for coCO perturbations with budget $\tau = 2$ (top and middle) and the distance between pairs of perturbations and original inputs (bottom). The accuracy drop of the POS-tag task is reported in number of words correctly guessed. Results confirm that counter-fitting does not improve robustness at any level (in this case, syntactic robustness).

4.1 Robustness

	Syntax Reconstruction			POS-tagging
	Δ SDR	Δ UAS	Δ Sp	Δ Acc.
TED	0.167	0.0872	0.0011	4.470
En-Universal	0.2360	0.29931	0.0056	9.456
Ud-English-ewt	0.1463	0.3124	0.0106	6.951
Ud-English-gum	0.1678	0.3252	0.0025	3.253
Ud-English-lines	0.2599	0.2981	0.0022	7.402
Ud-English-pud	0.0574	0.00612	0.0010	5.746

4.2 Robustness

	Root Identification	Tree Depth Estimation	
	Δ Acc.	Δ Acc.	Δ Sp
TED	0.3753	0.0727	0.0165
En-Universal	0.5288	0.2215	-0.0011
Ud-English-ewt	0.816	0.7751	0.0102
Ud-English-gum	0.389	0.3065	0.0028
Ud-English-lines	0.4493	0.314	0.0037
Ud-English-pud	0.3894	0.2583	-0.0189

4.3 Distance/Similarity Metrics

	Cosine similarity	ℓ_2 -norm distance
TED	0.925	0.0037
En-Universal	0.919	0.0033
Ud-English-ewt	0.9295	0.0035
Ud-English-gum	$\approx 1.$	0.0036
Ud-English-lines	$\approx 1.$	0.0037
Ud-English-pud	0.9518	0.0038

Table 4: Robustness of GloVe counter-fitted models, with an analysis, w.r.t. each dataset (one per row), of the relationship between the syntactic robustness metrics for coPOS perturbations with budget $\tau = 2$ (top and middle) and the distance between pairs of perturbations and original inputs (bottom). The accuracy drop of the POS-tag task is reported in number of words correctly guessed. Results confirm that counter-fitting does not improve robustness at any level (in this case, syntactic robustness).

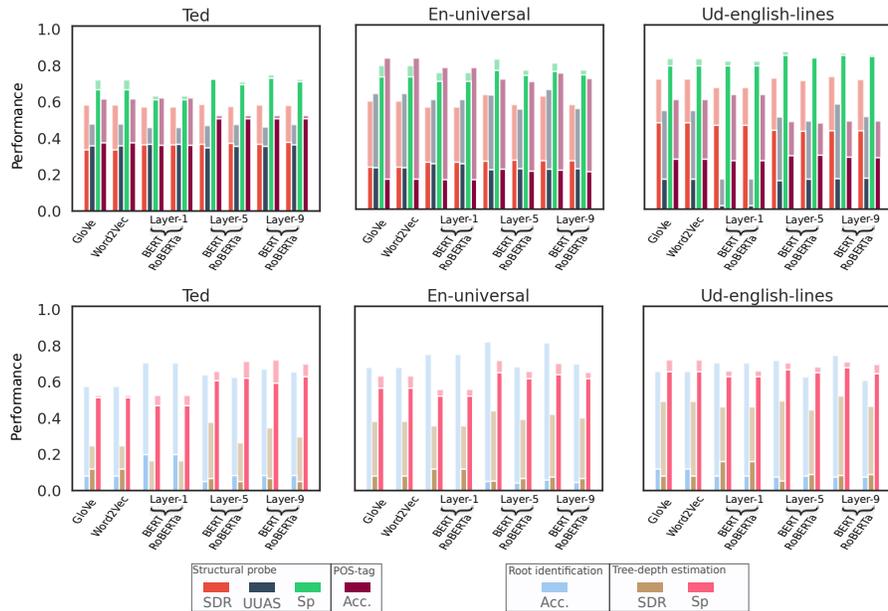


Figure 16: Performance, against baseline perturbations, of different linguistic representations on *syntax reconstruction* and *POS-tagging* probing tasks (top) and on *root identification* (accuracy metric) and *tree-depth estimation* (SDR and Spearman metric) probing tasks (bottom). For all plots, the performance of the probing tasks is reported as shaded bars, with the performance for the perturbed representation shown as a solid overlapping bar: the results refer to the case where the baseline perturbation budget τ is equal to 3 (i.e., at most 3 words per-sentence are perturbed). Regardless of the embedding/representation employed, we observe severe brittleness of the syntactic representations.

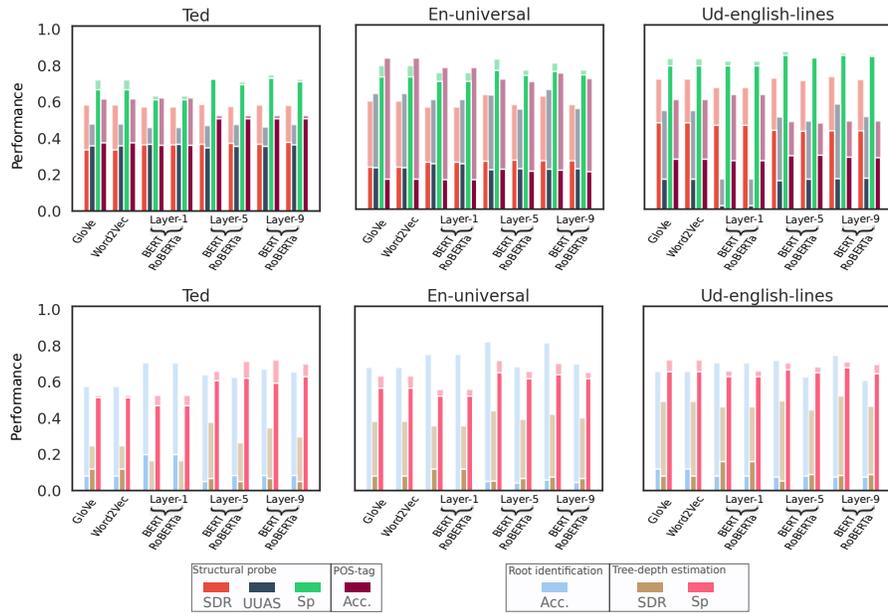


Figure 17: Performance, against coCO perturbations, of different linguistic representations on *syntax reconstruction* and *POS-tagging* probing tasks (top) and on *root identification* (accuracy metric) and *tree-depth estimation* (SDR and Spearman metric) probing tasks (bottom). For all plots, the performance of the probing tasks is reported as shaded bars, with the performance for the perturbed representation shown as a solid overlapping bar: the results refer to the case where the coCO perturbation budget τ is equal to 3 (i.e., at most 3 words per-sentence are perturbed). Regardless of the embedding/representation employed, we observe severe brittleness of the syntactic representations.

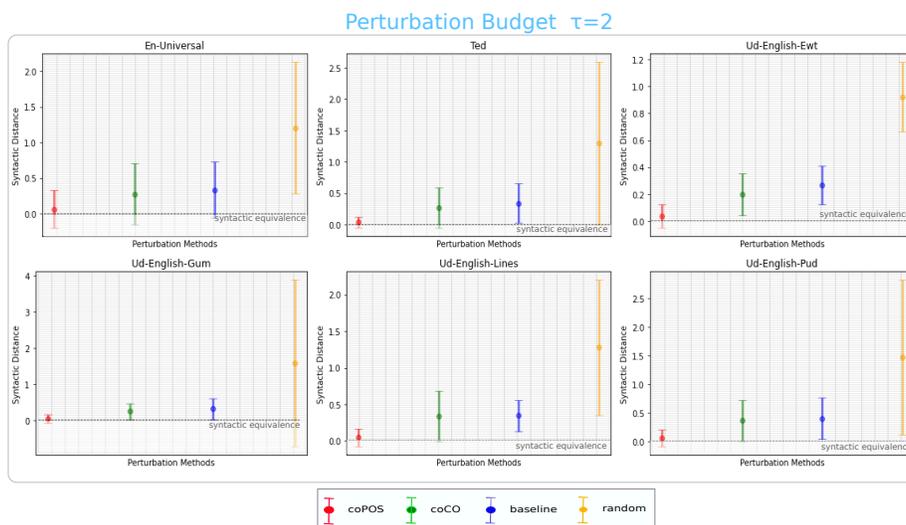


Figure 18: Tree-distance, measured with Stanza, between an input and its perturbed version, for different datasets and perturbation budget $\tau = 2$. The coPOS perturbation method (red) induces almost no disruption to a perturbation’s syntax tree, while injection of random words (blue) and coPOS perturbations (green) both induce some noticeable disruption. The disruption induces by comparing the syntax tree of two random sentences is reported for comparison (orange).

5.1 Robustness

	Syntax Reconstruction			POS-tagging
	Δ SDR	Δ UAS	Δ Sp	Δ Acc.
GloVe	0.2087 ± 0.123	0.2822 ± 0.1498	0.0545 ± 0.0097	6.583 ± 2.9971
Word2Vec	0.161 ± 0.1137	0.139 ± 0.0996	0.035 ± 0.0079	6.8977 ± 2.2552
BERT layer -1	0.208 ± 0.0803	0.1787 ± 0.0827	0.0289 ± 0.0156	3.1208 ± 3.0138
RoBERTa layer -1	0.2026 ± 0.0805	0.1954 ± 0.1129	0.013 ± 0.0154	4.2408 ± 3.3794
BERT layer -5	0.2395 ± 0.0804	0.2599 ± 0.1287	0.023 ± 0.0196	3.289 ± 3.1118
RoBERTa layer -5	0.2123 ± 0.0764	0.2303 ± 0.1094	0.0137 ± 0.0098	3.1448 ± 3.1161
BERT layer -9	0.2345 ± 0.0938	0.2828 ± 0.162	0.0204 ± 0.0126	3.6504 ± 3.7005
RoBERTa layer -9	0.2151 ± 0.0773	0.2372 ± 0.1189	0.009 ± 0.0106	3.568 ± 3.2097

5.2 Robustness

	Root Identification	Tree Depth Estimation	
	Δ Acc.	Δ Acc.	Δ Sp
GloVe	0.4853 ± 0.1776	0.2796 ± 0.2265	-0.0108 ± 0.1222
Word2Vec	0.6118 ± 0.1342	0.3115 ± 0.2495	0.0407 ± 0.0156
BERT layer -1	0.5466 ± 0.2041	0.3883 ± 0.2186	0.103 ± 0.0784
RoBERTa layer -1	0.5557 ± 0.1783	0.3586 ± 0.2312	0.0818 ± 0.0306
BERT layer -5	0.6466 ± 0.1421	0.3896 ± 0.2246	0.0317 ± 0.0575
RoBERTa layer -5	0.5464 ± 0.1778	0.3252 ± 0.2351	0.0225 ± 0.0704
BERT layer -9	0.6462 ± 0.1392	0.3927 ± 0.2461	0.0314 ± 0.0464
RoBERTa layer -9	0.5563 ± 0.1834	0.3485 ± 0.2493	0.0344 ± 0.046

5.3 Distance/Similarity Metrics

	ℓ_2 -norm distance	Cosine similarity
GloVe	0.0427 ± 0.001	0.8156 ± 0.0391
Word2Vec	0.0059 ± 0.0002	0.8606 ± 0.0184
BERT layer -1	0.0538 ± 0.0052	0.7401 ± 0.0418
RoBERTa layer -1	0.0211 ± 0.0014	0.9519 ± 0.006
BERT layer -5	0.0731 ± 0.0041	0.8059 ± 0.0299
RoBERTa layer -5	0.0539 ± 0.004	0.9302 ± 0.0122
BERT layer -9	0.0669 ± 0.0035	0.8382 ± 0.0199
RoBERTa layer -9	0.0375 ± 0.0021	0.9553 ± 0.0051

Table 5: Relationship between the syntactic robustness metrics for four linear probing tasks on baseline perturbations with budget $\tau = 2$ (top and middle row) and the distance between pairs of perturbed and original inputs measured via cosine similarity and ℓ_2 -norm distance (bottom row). The accuracy drop of the POS-tag task is reported as the number of words correctly guessed in both cases. The reported standard deviation is measured by averaging over the 6 training corpora. Whilst the distance (similarity) between inputs and perturbations is low (high), we observe that all embeddings/representations are brittle to syntax-preserving perturbations.