Shaking Syntactic Trees on the Sesame Street: Multilingual Probing with Controllable Perturbations

Ekaterina Taktasheva[*]
HSE University
Moscow, Russia

Vladislav Mikhailov
SberDevices, Sberbank
Moscow, Russia

Ekaterina Artemova
HSE University
Huawei Noah's Ark lab
Moscow, Russia

Abstract

Recent research has adopted a new experimental field centered around the concept of text perturbations which has revealed that shuffled word order has little to no impact on the downstream performance of Transformer-based language models across many NLP tasks. These findings contradict the common understanding of how the models encode hierarchical and structural information and even question if the word order is modeled with position embeddings. To this end, this paper proposes nine probing datasets organized by the type of *controllable* text perturbation for three Indo-European languages with a varying degree of word order flexibility: English, Swedish and Russian. Based on the probing analysis of the M-BERT and M-BART models, we report that the syntactic sensitivity depends on the language and model pre-training objectives. We also find that the sensitivity grows across layers together with the increase of the perturbation granularity. Last but not least, we show that the models barely use the positional information to induce syntactic trees from their intermediate self-attention and contextualized representations.

1 Introduction

An extensive body of works is devoted to analyzing syntactic knowledge of Transformer language models (LMs) (Vaswani et al., 2017; Clark et al., 2019; Goldberg, 2019; Belinkov and Glass, 2019). BERT-based LMs (Devlin et al., 2019) have demonstrated their abilities to encode various linguistic and hierarchical properties (Lin et al., 2019; Jawahar et al., 2019; Jo and Myaeng, 2020) which have a positive effect on the downstream performance (Liu et al., 2019a; Miaschi et al., 2020) and serve as an inspiration for syntax-oriented architecture improvements (Wang et al., 2019; Bai et al., 2021; Ahmad et al., 2021; Sachan et al., 2021). Besides, a variety of pre-training objectives has been introduced (Liu et al., 2020a), with some of them modeling reconstruction of the perturbed word order (Lewis et al., 2020; Tao et al., 2021; Panda et al., 2021).

Recent research has adopted a new experimental direction aimed at exploring the syntactic knowledge of LMs and their sensitivity to word order employing *text perturbations* (Futrell et al., 2018, 2019; Ettinger, 2020). Some studies show that shuffling word order causes significant performance drops on a wide range of QA tasks (Si et al., 2019; Sugawara et al., 2020). However, a number of works demonstrates that such permutation has little to no impact during the pre-training and fine-tuning stages (Pham et al., 2020; Sinha et al., 2020, 2021; O'Connor and Andreas, 2021; Hessel and Schofield, 2021; Gupta et al., 2021). The latter contradict the common understanding on how the hierarchical and structural information is encoded in LMs (Rogers et al., 2020), and even may question if the word order is modeled with the position embeddings (Wang et al., 2020; Dufter et al., 2021).

This has stimulated a targeted probing of the LMs internal representations generated from original texts and their permuted counterparts (Sinha et al., 2021; Hessel and Schofield, 2021). A new type of *controllable* probes has been proposed, designed to test the LMs sensitivity to granular character- and sub-word level manipulations (Clouatre et al., 2021), as well as structured syntactic perturbations (Alleman et al., 2021). Despite the emerging interest in the field, little is investigated for languages other than English, specifically those with flexible word order.

This paper extends the ongoing research on the syntactic sensitivity to three Indo-European languages with a varying degree of word order flexibility: English, Swedish, and Russian. The contributions of this work are summarized as follows. First, we propose nine probing datasets in the languages mentioned above, organized by the type of

[*] etaktasheva@hse.ru

controllable syntactic perturbation: N-gram perturbation (**NgramShift**), shuffling parts of the syntactic clauses (**ClauseShift**) and randomizing word order (**RandomShift**). Despite that randomizing word order has been studied from many perspectives (see Section 2), **NgramShift** differs from similar approaches (Conneau et al., 2018; Ravishankar et al., 2019; Eger et al., 2020; Alleman et al., 2021) in that the N-grams correspond to *only* syntactic phrases (e.g. prepositional or numerical phrases) rather than random word spans. **ClauseShift** is a previously unexplored type of syntactic perturbation adopted from the syntactic tree augmentation method (Şahin and Steedman, 2018). Second, we apply a combination of parameter-free interpretation methods to test the sensitivity of two multilingual Transformer LMs: M-BERT (Devlin et al., 2019), and M-BART (Liu et al., 2020b). We hypothesize that M-BART is more robust to the perturbations as opposed to M-BERT since it is learned to restore the shuffled input during pre-training. We evaluate the discrepancy in the syntactic trees induced by the models from perturbed sentences against the original ones, along with the ability to distinguish between them by judging their linguistic acceptability (Lau et al., 2020). Finally, we analyze the relationship between the models' probe performance and position embeddings (PEs). To the best of our knowledge, it is one of the first attempts to introspect PEs regarding structural probing, particularly in the light of syntactic perturbations. The code and datasets are publicly available[1].

2 Related Work

Syntax Probing Most of the previous studies on the syntactic knowledge of LMs are centered around the concept of probing tasks, where a simple classifier is trained to predict a particular linguistic property based on the model internal representations (Conneau et al., 2018). The scope of the properties ranges from dependency relations (Tenney et al., 2018) to the depth of a syntax tree, and top constituents (Conneau et al., 2018). A variety of probing datasets and benchmarks have been developed. To name a few, Liu et al. (2019a) create a probing suite focused on fine-grained linguistic phenomena, including hierarchical knowledge. SyntaxGym (Gauthier et al., 2020) and LINSPECTOR (Şahin et al., 2020) allow for targeted evaluation of the LMs linguistic knowledge in a standardized and reproducible environment.

These studies have proved that LMs are capable of encoding linguistic and hierarchical information (Belinkov and Glass, 2019; Rogers et al., 2020). However, the probing paradigm has been lately criticized for relying on *supervised* probes, which can learn linguistic properties given the supervision, and make it challenging to interpret the results because of the additional set of parameters (Hewitt and Liang, 2019; Belinkov, 2021). Towards that end, Hewitt and Manning (2019) introduce a *structural* probe to explore a linear transformation of the embedding space, which best approximates the distance between words and depth of the parse tree. The method has proved to infer the hierarchical structure without any linguistic annotation (Kim et al., 2020). Maudslay and Cotterell (2021) propose a *Jabberwocky* probing suite of semantically nonsensical but syntactically well-formed sentences. The results demonstrate that the BERT-based LMs do not isolate semantics from syntax, which motivates further development of the probing field.

Acceptability Judgements Another line of works relies on the concept of acceptability judgments. The CoLA benchmark (Warstadt et al., 2019) and its counterpart for Swedish (Volodina et al., 2021) test LMs ability to identify various linguistic violations. Although Transformer LMs have outperformed the CoLA human solvers on the GLUE leaderboard (Wang et al., 2018), a granular linguistic analysis (Warstadt and Bowman, 2019) shows that the models struggle with long-distance syntactic phenomena as opposed to more local ones. Similar in spirit, BLiMP (Warstadt et al., 2020), and CLiMP (Xiang et al., 2021) allow to evaluate the LMs with respect to the acceptability contrasts, framing the task as ranking sentences in minimal pairs.

Text Perturbations Recent research has adopted a scope of novel approaches to investigating the LMs sensitivity to syntax corruption and input data manipulations. Starting from studies on randomized word order in LSTMs (Hill et al., 2016; Khandelwal et al., 2018; Sankar et al., 2019; Nie et al., 2019), text perturbations have emerged as an audacious experimental direction under the "pre-train & fine-tune" paradigm along with the interpreta-

[1] https://github.com/evtaktasheva/dependency_extraction

tion methods of modern LMs. Si et al. (2019); Sugawara et al. (2020) show that N-gram permutations and shuffled word order in the fine-tuning data cause BERT's performance drops up to 22% on a wide range of QA tasks. In contrast, several works report that models fine-tuned on such perturbed data still produce high confidence predictions and perform close to their counterparts on many tasks, including the GLUE benchmark (Ahmad et al., 2019; Sinha et al., 2020; Liu et al., 2021; Hessel and Schofield, 2021; Gupta et al., 2021). Similar results are demonstrated by the RoBERTa model (Liu et al., 2019b) when the word order perturbations are incorporated into the pre-training objective (Panda et al., 2021) or tested as a part of full pre-training on the perturbed corpora (Sinha et al., 2021). Sinha et al. (2021) find that the randomized RoBERTa models are similar to their naturally pre-trained peer according to parametric probes but perform worse according to the non-parametric ones.

Recognizing the need to further explore the LMs sensitivity to word order, Clouatre et al. (2021) and Alleman et al. (2021) conduct the interpretation analysis of LMs by means of *controllable* text perturbations. Clouatre et al. (2021) propose two metrics that score local and global structure of sentences perturbed at the granularity of characters and sub-words. The metrics allow identifying that both conventional and Transformer LMs rely on the local order of tokens more than the global one. Alleman et al. (2021) find that BERT builds syntactic complexity towards the output layer and demonstrates a growing sensitivity to the hierarchical phrase structure across layers. In line with these studies, we analyze the syntactic sensitivity of Transformer-based LMs, extending the experimental setup to the multilingual setting.

3 Controllable Perturbations

This work proposes three types of *controllable* syntactic perturbations varying in the extent of sentence corruption. We construct nine probing tasks[2] for three Indo-European languages[3]: English (West Germanic, analytic), Swedish (North Germanic, analytic), and Russian (Balto-Slavic, fusional). Based on the dominant constituent order, all three languages are classified as the SVO (Subject-Verb-Object) languages. Nevertheless, there are some differences between them regarding word order flexibility. Russian is known to exhibit free word order as all of the possible constituent reorderings are acceptable: SOV, OSV, SVO, OVS, VSO, VOS (Bailyn, 2012). English allows for only two of them, namely SVO and OSV (Prince, 1988). Swedish belongs to the verb-second languages, which poses different restrictions on the possible constituent reorderings (Börjars et al., 2003). Each dataset[4] consists of 10k pairs of the corresponding perturbed sentence and its original.

NgramShift tests the LM sensitivity to *local* perturbations taking into account the syntactic structure. We used a set of carefully designed morphosyntactic patterns to perturb N-grams that correspond to *only* syntactic phrases such as numeral phrases, determiner phrases, compound noun phrases, prepositional phrases, etc. Towards this, we applied TF-IDF weighting from scikit-learn library (Pedregosa et al., 2011) to build a ranked N-gram feature matrix from the corpora and further used it for the N-gram inversion. We used the N-gram range $\in [2; 4]$ for each language. Note that the number of words that change their absolute positions is similar for different values of N. Figure 1 illustrates the shift of the head in the prepositional phrase *"to school"* for the sentence *"He did not go to school yesterday"*.

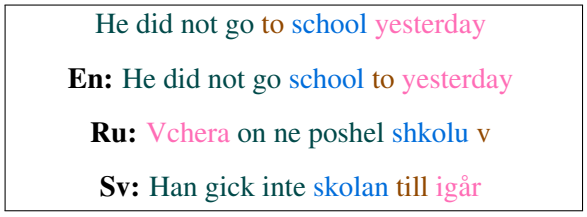


Figure 1: Examples of the N-gram perturbations (**NgramShift**). Languages: **En**=English, **Ru**=Russian, **Sv**=Swedish. The English sentence is translated to the other languages for illustrational purposes.

ClauseShift probes the LM sensitivity to *distant* perturbations at the level of syntactic clauses. We use the syntactic tree augmentation method (Şahin and Steedman, 2019) to rotate sub-trees around the root of the dependency tree of each sentence to form a new synthetic sentence. We then apply a set of manually curated language-specific heuristics to filter out sentences uncorrupted by the rotation procedure. Figure 2 outlines an example of the

[2]We use sentences from the CoNLL 2017 Shared Task on Multilingual Parsing from Raw Texts to Universal Dependencies (Ginter et al., 2017).

[3]https://wals.info

[4]A brief statistics is outlined in Appendix 1.

clause rotation perturbation for the sentence *"He manages to tell her that she has been resurrected"*.

He manages to tell her that she has been resurrected
En: That she has been resurrected he manages to tell her
Sv: Att hon har uppstått han lyckas berätta för henne
Ru: Chto ona byla voskreshena on smog rasskazat' ej

Figure 2: Examples of the clause rotation perturbation (**ClauseShift**). Languages: **En**=English, **Ru**=Russian, **Sv**=Swedish. The English sentence is translated to the other languages for illustrational purposes.

RandomShift tests the LM sensitivity to *global* perturbations obtained by shuffling the word order. This type represents an extreme case of sentence permutation and is useful for comparing the behavior of the models at the scale of the perturbation complexity. An example of the randomized word order perturbation for the sentence *"She wanted to go to London"* is presented in Figure 3.

She wanted to go to London
En: Wanted London go she to to
Sv: Ville London åka hon till att
Ru: Hotela London poehat' ona v

Figure 3: Examples of the word order shuffling (**RandomShift**). Languages: **En**=English, **Ru**=Russian, **Sv**=Swedish. The English sentence is translated to the other languages for illustrational purposes.

4 Experimental Setup

4.1 Models

The experiments are run on two 12-layer multilingual Transformer models released by the HuggingFace library (Wolf et al., 2020):

M-BERT[5] is pre-trained using masked language modeling (MLM) and next sentence prediction objectives, over concatenated monolingual Wikipedia corpora in 104 languages.

M-BART[6] is a sequence-to-sequence model that comprises a BERT encoder and an autoregressive GPT-2 decoder (Radford et al., 2019). The model is pre-trained on the CC25 corpus in 25 languages using text infilling and sentence shuffling objectives, where it learns to predict masked word spans and reconstruct the permuted input. We use only the encoder in our experiments.

[5]Model name: `bert-base-multilingual-cased`.
[6]Model name: `facebook/mbart-large-cc25`.

4.2 Interpretation Methods

Parameter-free Probing We apply two unsupervised probing methods to reconstruct syntactic trees from self-attention (**Self-Attention Probing**) and so-called "impact" (**Token Perturbed Masking**) matrices computed by feeding the MLM models with each sentence s and its perturbed version s'. The trees are induced by Chu-Liu-Edmonds algorithm (Chu, 1965; Edmonds, 1968) used to compute the Maximum Spanning Tree starting from the root of the corresponding gold dependency tree (Raganato and Tiedemann, 2018; Htut et al., 2019; Wu et al., 2020). The probing performance is evaluated by the Undirected Unlabeled Attachment Score (UUAS), which reflects the percentage of words that have been assigned the correct head without taking the direction of relations and dependency labels into account (Klein and Manning, 2004).

Self-Attention Probing (Htut et al., 2019) allows to explore if attention heads encode complete syntactic trees. To this end, each layer-head attention matrix is treated as a weighted directed graph where the vertices represent words in the input sentence and edges are the attention weights. Model-specific special tokens such as `[CLS]`, `[SEP]`, `<s>`, `</s>` are excluded at the pre-processing stage to eliminate their impact on other tokens.

Token Perturbed Masking (Wu et al., 2020) extracts global syntactic information by measuring the impact one word has on the prediction of another in an MLM. The impact matrix is similar to the self-attention matrix as it reflects the inter-word relationships in terms of Euclidean distance, except that it is derived from the outputs of the MLM head. For the sake of space, we refer the reader to Wu et al. (2020) for more details.

Representation Analysis Hessel and Schofield (2021) propose two metrics to compare contextualized representations and self-attention matrices produced by the model for each pair of sentences s and s'. *Token Identifiability (TI)* evaluates the similarity of the LM's contextualized representations of a particular token in s and s'. It is high if the token representations are similar to one another. *Self-Attention Distance (SAD)* measures if

each token in s relates to similar words in s' by computing row-wise Jensen-Shannon Divergence between the two self-attention matrices. It is low if an LM attends to the same words despite the perturbations.

Pseudo-perplexity Pseudo-perplexity (PPPL) is an intrinsic measure that estimates the probability of a sentence with an MLM similar to that of conventional LMs (Salazar et al., 2020). PPPL-based measures have proved to correlate with human ratings (Lau et al., 2017), match or outperform autoregressive LMs (GPT-2) in ranking hypotheses for downstream tasks and the BLiMP benchmark (Salazar et al., 2020), and perform at the human level in acceptability judgments (Lau et al., 2020). We use two PPPL-based measures under implementation[7] by Lau et al. (2020) to infer probabilities of the sentences and their perturbed counterparts. The *MeanLP* and *PenLP* measures are computed as the sum of pseudo-log-likelihood scores for each token in the sentence normalized by the total number of tokens. *PenLP* additionally scales the denominator with the exponent α to penalize the effect of high scores.

4.3 Positional Encoding

Various PEs have been proposed to utilize the information about word order in the Transformer-based LMs (Wang et al., 2020; Dufter et al., 2021). Surprisingly, little is known about what PEs capture and how well they learn the meaning of positions. Wang and Chen (2020) among the first present an extensive study on the properties captured by PEs in different pre-trained Transformers and empirically evaluate their impact on the downstream performance for many NLP tasks. In the spirit of this work, we aim at analyzing the impact of the PEs on the syntactic probe performance. Towards this end, we consider the following three configurations of PEs of the M-BERT and M-BART models: (1) **absolute**=frozen PEs; (2) **random**=randomly initialized PEs; and (3) **zero**=zeroed PEs.

5 Results

5.1 Parameter-free Probing

The discrepancy in the syntactic trees induced from the original sentences and their perturbed analogs is measured as the difference between the corresponding UUAS scores (δ UUAS). The lower the δ UUAS, the better is the syntax tree reconstructed from s' with respect to the UUAS score for s.

Self-Attention Probing Figures 4 and 1 in Appendix 2 outline the task-wise heatmaps with the δ UUAS scores achieved by the M-BERT and M-BART models with **absolute** PEs for each layer-head pair, respectively. The models exhibit similar behavior, demonstrating positive correlation between the δ UUAS scores and the granularity of the perturbation. The overall pattern for both models is that they display little to no sensitivity to *local* and *distant* perturbations (**NgramShift**, **ClauseShift**) in contrast to the *global* ones (**RandomShift**). We provide examples of the dependency trees extracted from the self-attention matrices of the M-BERT model for the Swedish **NgramShift** task on Figure 5. The trees from both original (see Figure 5a) and perturbed (see Figure 5b) sentence versions receive the UUAS score of 0.86, demonstrating little changes in the assigned dependency heads under the local perturbation. On the contrary, randomizing word order (**RandomShift**) corrupts the syntactic structure significantly with a δ UUAS score of 0.33 (see Figure 8, Appendix 2).

Token Perturbed Masking The models show similar results to that of in **Self-Attention Probing**, with regards to the perturbation granularity (see Figure 6). In spite of that, the model performance on the **NgramShift** and **ClauseShift** reveal some differences between the encoders. M-BART generally achieves lower and close to zero δ UUAS scores, meaning to better restore the hierarchical information from the perturbed sentences (e.g., **ClauseShift**: [**Sv, Ru**]). We relate this to the fact that M-BART is pre-trained with the sentence shuffling objective.

Language-wise Comparison Another observation is that there are more insensitive attention heads on the Russian tasks, possibly indicating that it is harder to distinguish from the perturbations as opposed to English and Swedish, particularly on the **ClauseShift** task with typically longer and syntactically more complex sentences (see Figures 4, 1, Appendix 2). As for Swedish, which has a similar to English but stricter syntactic structure, M-BART tends to induce correct syntactic trees from the permuted sentences more frequently. This is indicated by negative δ UUAS scores on most tasks.

[7] https://github.com/jhlau/acceptability-prediction-in-context

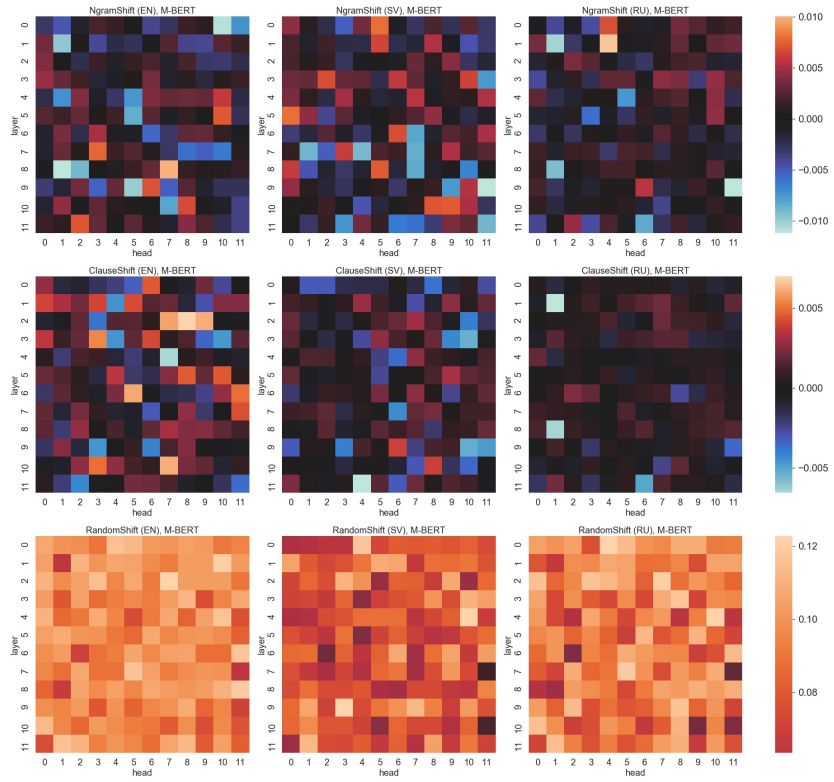


Figure 4: The task-wise heatmaps depicting the δ UUAS scores by M-BERT for each language. Method=**Self-Attention Probing**. PE=**absolute**. X-axis=Attention head index. Y-axis=Layer index. Tasks: **NgramShift** (top); **ClauseShift** (middle); **RandomShift** (bottom). Languages: **En**=English (left); **Sv**=Swedish (middle); **Ru**=Russian (right).

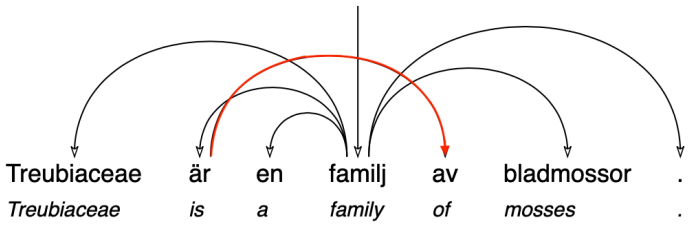
Positional Encoding Analysis of the positional encoding shows that despite the genuine belief that positional information contributes most to syntactic structure encoding, the models do not seem to rely on it as much as might be expected. Figure 2 (see Appendix 2) illustrates the distribution of δ UUAS scores for M-BERT with different PEs on English tasks. The heatmaps show that **zero** and **random** PEs only slightly affects the quality of the probe performance of the self-attention heads.

To analyze the impact of PEs from another perspective, for each pair of (s, s') we compute the Euclidean distance (L2) between the corresponding impact (**Token Perturbed Probing**) and self-attention matrices (**Self-Attention Probing**) described in Section 4.2. The difference in the impact matrices produced by M-BERT model is generally observed only in the setting with **zero** PEs (see Figures 7; Figures 3-4, Appendix 2). In contrast, there is almost no difference between the representations generated by M-BART across all configurations of the PEs (see Figures 5-7, Appendix 2). This behavior is consistent with the head-wise results under **Self-Attention Probing** for all languages.

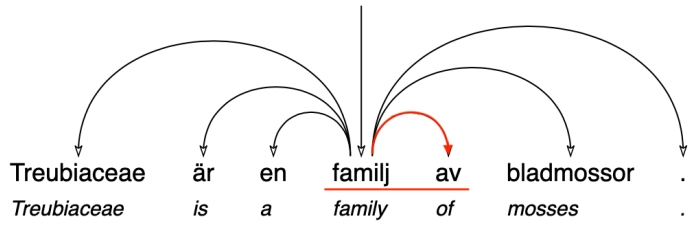
5.2 Representation Analysis

Token Identifiability The overall pattern for both models under the representation analysis is that for *local* and *distant* perturbations TI steadily decreases towards the output layer with rapid increases at layers $[1, 10]$ (see Figure 9, Appendix 3), and high for *global* perturbations (**RandomShift**). TI decreases when the perturbed inputs generate embeddings different from the intact ones. Despite that higher layers in both models are more sensitive, the perturbed representations remain similar to that of the original (Hessel and Schofield, 2021).

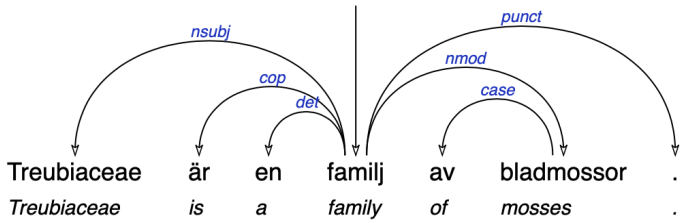
Self-Attention Distance The results by SAD show that both models score significantly lower with **random** and **zero** PEs (see Figure 10, Appendix 3), meaning lower sensitivity to the perturbations supported by the probing results (Section 5.1). This provides evidence that the encoders marginally rely on the positional information to induce the syntactic structure despite the distributions of the self-attention weights for the intact and perturbed sentences may differ according to the Jensen-Shannon divergence.



(a) `original`



(b) `perturbed`



(c) `gold`

Figure 5: Graphical representations of the syntactic trees inferred for the Swedish sentence *Treubiaceae är en familj av bladmossor* 'Treubiaceae is a family of mosses' and its perturbed version. `original`=the original sentence; `perturbed`=the perturbed version; `gold`=gold standard. Task=**NgramShift**. Model=**M-BERT** (Layer: 11; Head: 2). Method=**Self-Attention Probing**. The perturbation is underlined with red, and incorrectly assigned dependency heads are marked with red arrows.

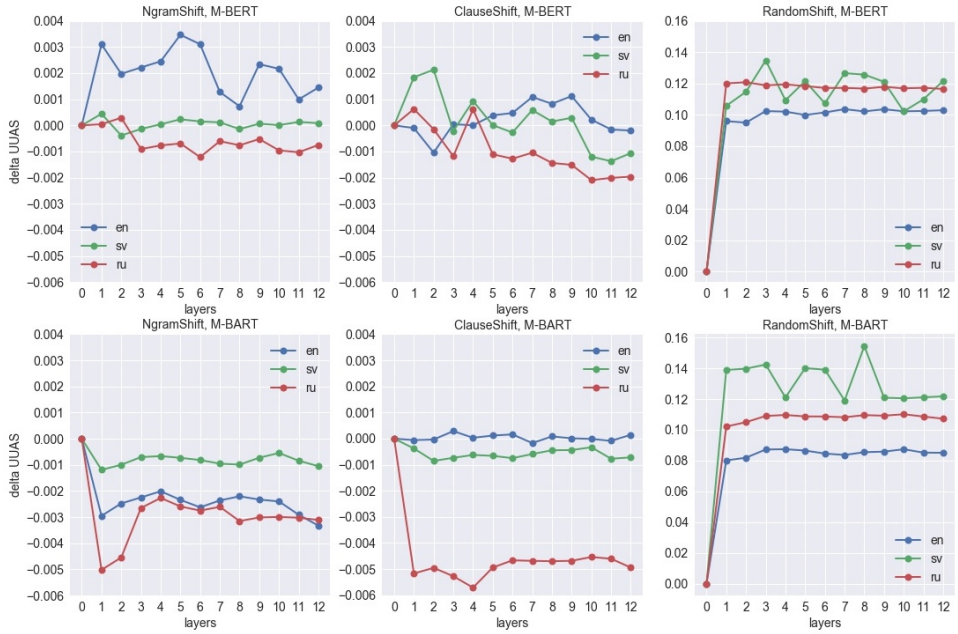


Figure 6: The probing performance in δ UUAS across layers under **Token Perturbed Probing**. PE=**absolute**. The scores are averaged over attention heads at each layer. X-axis=Attention head index. Y-axis=δ UUAS.

5.3 Pseudo-perplexity

Consistent with the results under parameter-free probing (Section 5.1) and representation analysis (Section 5.2), PPPL-based acceptability judgements[8] indicate that the encoders distinguish between the perturbations depending on their granularity. The overall trend is that for all languages the sentence pseudo-log-probability inferred from both LMs decreases with the increase of the perturbation complexity which is demonstrated by higher acceptability scores on **NgramShift**, but significantly lower scores on the **ClauseShift** and **RandomShift** (see Figures 11-12, Appendix 4). The statistical significance of the PPPL distributions is confirmed with Kolmogorov–Smirnov and Wilcoxon signed-rank tests (p-value < 0.01).

6 Discussion

The syntactic sensitivity depends upon language At present, English remains the focal point of prior research in the field of NLP, leaving other languages understudied. Our probing experiments on the less explored languages with different word order flexibility show that M-BERT and M-BART behave slightly differently in Swedish and Russian. While M-BART better restores the corrupted syntactic structure on most of the tasks for Swedish, there are fewer attention heads sensitive to the perturbations in Russian, which is revealed through the examination of head-wise attention patterns of both models. Besides, the encoders receive lower probing performance for Russian that can be contributed to the more complex syntax and flexible word order.

Pre-training objectives can help to improve syntactic robustness Analysis of the M-BERT and M-BART LMs that differ in the pre-training objectives shows that M-BERT achieves higher δ UUAS performance across all languages as opposed to M-BART pre-trained with the sentence shuffling objective. The lower δ UUAS probing performance indicates that M-BART better induces syntactic trees from both perturbed and intact sentences (see Section 5.1). Despite this, the representation and acceptability analysis demonstrate that M-BART is also capable of distinguishing between the perturbations (see Sections 5.2-5.3). A fruitful direction for future work is to analyze more LMs that differ in the architecture design and pre-training objectives.

The LMs are less sensitive to more granular perturbations The results of the parameter-free probing show that M-BERT and M-BART exhibit little to no sensitivity to *local* perturbations within syntactic groups (**NgramShift**) and *distant* perturbations at the level of syntactic clauses (**ClauseShift**). In contrast, the *global* perturbations (**RandomShift**) are best distinguished by the encoders. As the granularity of the syntactic corruption increases, we observe a worse probing performance under all considered interpretation methods. Namely, the results are supported by representation analysis metrics (see Section 5.2) that indicate higher susceptibility to major changes in the sentences structure (**RandomShift, ClauseShift**), and the PPPL-based measures (see Section 5.3) prescribing higher acceptability scores to sentences with more granular perturbations (**NgramShift**). We also find that the sensitivity to the hierarchical corruption grows across layers together with the increase of the perturbation complexity, which is in line with Alleman et al. (2021).

M-BERT and M-BART barely use positional information to induce syntactic trees Previous research has shown that the token embeddings capture enough semantic information to restore the syntactic structure (Vilares et al., 2020; Kim et al., 2020; Rosa and Mareček, 2019). Maudslay and Cotterell (2021) claim that syntactic abilities of BERT-based LMs are overestimated and raise the problem of isolating semantics from syntax. However, more recent studies show that Transformer encoders encode redundant information (Luo et al., 2021), may not sufficiently capture the meaning of positions and be unimportant for downstream tasks (Wang and Chen, 2020), including the setting with perturbed fine-tuning data (Clouatre et al., 2021). In spirit with the latter studies, our results under different PEs configurations reveal that M-BERT and M-BART do not need the precise position information to restore the syntactic tree from their internal representations. The overall behavior is that zeroed (except for M-BERT) or even randomly initialized PEs can result in the probing performance and one with absolute positions. We suppose that despite the absolute positions of words changes during the N-gram permutation and sub-tree rotation procedures, the word order within the clauses remains almost the same as in the intact sentence

[8] We present the results obtained by the *MeanLP* measure which are consistent with those of *PenLP*.

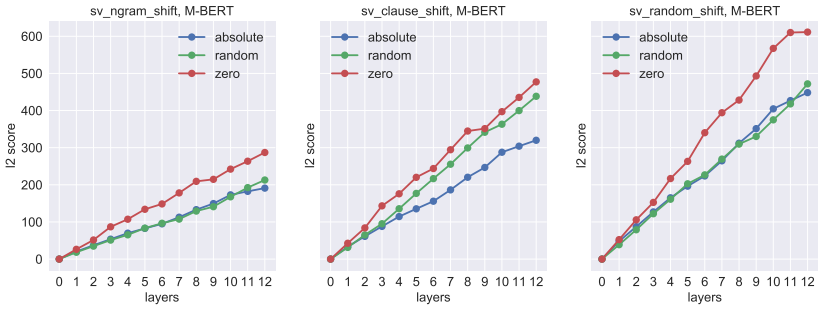


Figure 7: The Euclidean distances between the impact matrices computed by M-BERT with different PEs over each pair of sentences (s, s') for Swedish. The distances are averaged over attention heads at each layer. Method: **Token Perturbed Masking**. Tasks: **NgramShift** (left); **ClauseShift** (middle); **RandomShift** (right)

.

(**NgramShift**, **ClauseShift**). That is, the more granular perturbations marginally confuse the LMs when: (i) predicting the masked word under **Token Perturbation Probing** which can be performed using *only* attention (Wang and Chen, 2020), or (ii) judging the acceptability of the sentence where the low token pseudo-log-probability can occur at the juxtaposition of the syntactic groups, and clauses (Alleman et al., 2021). We leave a more detailed exploration of the relationship between PEs and probing analysis for future work.

7 Conclusion

This paper presents an extension of the ongoing research on the controllable text perturbations to the multilingual setting and introspection of positional embeddings in pre-trained LMs. We introduce nine probing datasets for three Indo-European languages varying in their flexibility of the word order: English, Swedish, and Russian. The suite is constructed using language-specific heuristics carefully designed under linguistic expertise and organized by three types of syntactic perturbations: randomization of word order studied by previous research from many perspectives and less explored permutations within syntactic phrases and clauses. The method includes a combination of parameter-free probing methods based on the intermediate self-attention and contextualized representations, novel metrics for representation analysis, and acceptability judgments with pseudo-perplexity. We conduct a line of experiments to probe the syntactic sensitivity of two multilingual Transformers, M-BERT and M-BART, the latter of which is learned to reconstruct the word order during pre-training. The LMs are less sensitive to more granular perturbations and build hierarchical complexity towards the output layer. The analysis of the understudied relationship between the position embeddings and syntactic probe performance reveals that the position information is not necessary for inducing the hierarchical structure, which is a promising direction for a more detailed investigation. The results also show that the syntactic sensitivity may depend on the language and be enhanced by pre-training objectives. We believe there is still room for exploring the sensitivity to word order and syntactic abilities of modern LMs, specifically across a more diverse set of languages and models varying in the architecture design choices.

Acknowledgements

Ekaterina Taktasheva and Ekaterina Artemova are partially supported by the framework of the HSE University Basic Research Program.

References

Wasi Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. 2021. Syntax-augmented multilingual BERT for cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4538–4554, Online. Association for Computational Linguistics.

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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 2440–2452,

Minneapolis, Minnesota. Association for Computational Linguistics.

Matteo Alleman, Jonathan Mamou, Miguel A Del Rio, Hanlin Tang, Yoon Kim, and SueYeon Chung. 2021. Syntactic perturbations reveal representational correlates of hierarchical phrase structure in pretrained language models. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 263–276, Online. Association for Computational Linguistics.

Jiangang Bai, Yujing Wang, Yiren Chen, Yaming Yang, Jing Bai, Jing Yu, and Yunhai Tong. 2021. Syntax-BERT: Improving pre-trained transformers with syntax trees. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3011–3020, Online. Association for Computational Linguistics.

John F Bailyn. 2012. *The Syntax of Russian*. Cambridge University Press.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and alternatives. *arXiv preprint arXiv:2102.12452*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Kersti Börjars, Elisabet Engdahl, Maia Andréasson, Miriam Butt, and Tracy Holloway King. 2003. Subject and Object Positions in Swedish. In *Proceedings of the LFG03 Conference*, pages 43–58. Citeseer.

Yoeng-Jin Chu. 1965. On the Shortest Arborescence of a Directed Graph. *Scientia Sinica*, 14:1396–1400.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Louis Clouatre, Prasanna Parthasarathi, Amal Zouaq, and Sarath Chandar. 2021. Demystifying Neural Language Models' Insensitivity to Word-Order. *arXiv preprint arXiv:2107.13955*.

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

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

Philipp Dufter, Martin Schmitt, and Hinrich Schütze. 2021. Position Information in Transformers: An Overview. *arXiv preprint arXiv:2102.11090*.

Jack Edmonds. 1968. Optimum branchings. *Mathematics and the Decision Sciences, Part*, 1(335-345):25.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2020. How to probe sentence embeddings in low-resource languages: On structural design choices for probing task evaluation. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 108–118, Online. Association for Computational Linguistics.

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

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. RNNs as Psycholinguistic Subjects: Syntactic State and Grammatical Dependency. *arXiv preprint arXiv:1809.01329*.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An Online Platform for Targeted Evaluation of Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76.

Filip Ginter, Jan Hajič, Juhani Luotolahti, Milan Straka, and Daniel Zeman. 2017. CoNLL 2017 shared task - automatically annotated raw texts and word embeddings. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. 2021. BERT & Family Eat Word Salad: Experiments with Text Understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12946–12954.

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

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377, San Diego, California. Association for Computational Linguistics.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. 2019. Do Attention heads in BERT track syntactic dependencies? *arXiv preprint arXiv:1911.12246*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Jae-young Jo and Sung-Hyon Myaeng. 2020. Roles and Utilization of Attention Heads in Transformer-based Neural Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3404–3417.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294, Melbourne, Australia. Association for Computational Linguistics.

Taeuk Kim, Jihun Choi, Daniel Edmiston, and Sanggoo Lee. 2020. Are pre-trained language models aware of phrases? simple but strong baselines for grammar induction. *arXiv preprint arXiv:2002.00737*.

Dan Klein and Christopher Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 478–485, Barcelona, Spain.

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How furiously can colorless green ideas sleep? sentence acceptability in context. *Transactions of the Association for Computational Linguistics*, 8:296–310.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive science*, 41(5):1202–1241.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yongjie Lin, Yi Chern Tan, and RoBERT Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Qi Liu, Matt J Kusner, and Phil Blunsom. 2020a. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: a Robustly Optimized BERT Pre-training Approach. *arXiv preprint arXiv:1907.11692*.

Zihan Liu, Genta I Winata, Samuel Cahyawijaya, Andrea Madotto, Zhaojiang Lin, and Pascale Fung. 2021. On the Importance of Word Order Information in Cross-lingual Sequence Labeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13461–13469.

Ziyang Luo, Artur Kulmizev, and Xiaoxi Mao. 2021. Positional artefacts propagate through masked language model embeddings.

Rowan Hall Maudslay and Ryan Cotterell. 2021. Do Syntactic Probes Probe Syntax? Experiments with Jabberwocky Probing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 124–131.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019. Analyzing Compositionality-Sensitivity of NLI Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6867–6874.

Joe O'Connor and Jacob Andreas. 2021. What Context Features Can Transformer Language Models Use? *arXiv preprint arXiv:2106.08367*.

Subhadarshi Panda, Anjali Agrawal, Jeewon Ha, and Benjamin Bloch. 2021. Shuffled-token detection for refining pre-trained RoBERTa. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 88–93, Online. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, BERTrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Thang M Pham, Trung Bui, Long Mai, and Anh Nguyen. 2020. Out of Order: How Important is the Sequential Order of Words in a Sentence in Natural Language Understanding Tasks? *arXiv preprint arXiv:2012.15180*.

Ellen F Prince. 1988. On Pragmatic Change: the Borrowing of Discourse Functions. *Journal of pragmatics*, 12(5-6):505–518.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

Vinit Ravishankar, Lilja Øvrelid, and Erik Velldal. 2019. Probing multilingual sentence representations with X-probe. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 156–168, Florence, Italy. Association for Computational Linguistics.

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

Rudolf Rosa and David Mareček. 2019. Inducing Syntactic Trees from BERT Representations.

Devendra Sachan, Yuhao Zhang, Peng Qi, and William L. Hamilton. 2021. Do syntax trees help pre-trained transformers extract information? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2647–2661, Online. Association for Computational Linguistics.

Gözde Gül Şahin and Mark Steedman. 2018. Data augmentation via dependency tree morphing for low-resource languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5004–5009, Brussels, Belgium. Association for Computational Linguistics.

Gözde Gül Şahin and Mark Steedman. 2019. Data augmentation via dependency tree morphing for low-resource languages. *arXiv preprint arXiv:1903.09460*.

Gözde Gül Şahin, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych. 2020. LINSPECTOR: Multilingual probing tasks for word representations. *Computational Linguistics*, 46(2):335–385.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Chinnadhurai Sankar, Sandeep Subramanian, Christopher Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37.

Chenglei Si, Shuohang Wang, Min-Yen Kan, and Jing Jiang. 2019. What does bert learn from multiple-choice reading comprehension datasets? *arXiv preprint arXiv:1910.12391*.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *CoRR*, abs/2104.06644.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2020. Unnatural language inference. *arXiv preprint arXiv:2101.00010*.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8918–8927.

Chongyang Tao, Shen Gao, Juntao Li, Yansong Feng, Dongyan Zhao, and Rui Yan. 2021. Learning to organize a bag of words into sentences with neural networks: An empirical study. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1682–1691, Online. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2018. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

David Vilares, Michalina Strzyz, Anders Søgaard, and Carlos Gómez-Rodríguez. 2020. Parsing as Pretraining. 34:9114–9121.

Elena Volodina, Yousuf Ali Mohammed, and Julia Klezl. 2021. DaLAJ-a Dataset for Linguistic Acceptability Judgments for Swedish: Format, Baseline, Sharing. *arXiv preprint arXiv:2105.06681*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Benyou Wang, Lifeng Shang, Christina Lioma, Xin Jiang, Hao Yang, Qun Liu, and Jakob Grue Simonsen. 2020. On Position Embeddings in BERT. In *International Conference on Learning Representations*.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. 2019. Structbert: incorporating language structures into pre-training for deep language understanding. *arXiv preprint arXiv:1908.04577*.

Yu-An Wang and Yun-Nung Chen. 2020. What do position embeddings learn? an empirical study of pre-trained language model positional encoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6840–6849, Online. Association for Computational Linguistics.

Alex Warstadt and Samuel R. Bowman. 2019. Grammatical analysis of pretrained sentence encoders with acceptability judgments. *CoRR*, abs/1901.03438.

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

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

Appendix

1 Dataset Statistics

	Language	NgramShift	ClauseShift	RandomShift
num. tokens	**Ru**	105.8k	199.7k	95.6k
	En	128.5k	198.6k	111.1k
	Sv	134.1k	192.9k	100.7k
unique tokens	**Ru**	25.2k	46.1k	27.8k
	En	19.2k	25.1k	22.8k
	Sv	23.2k	25.7k	17.8k
tokens / sentence	**Ru**	10.9	19.9	10.5
	En	12.9	19.9	11.1
	Sv	13.4	19.3	10.1

Table 1: A brief statistics of the controlled perturbation datasets. Languages: **Ru**=Russian, **En**=English, **Sv**=Swedish.

2 Parameter-free Probing

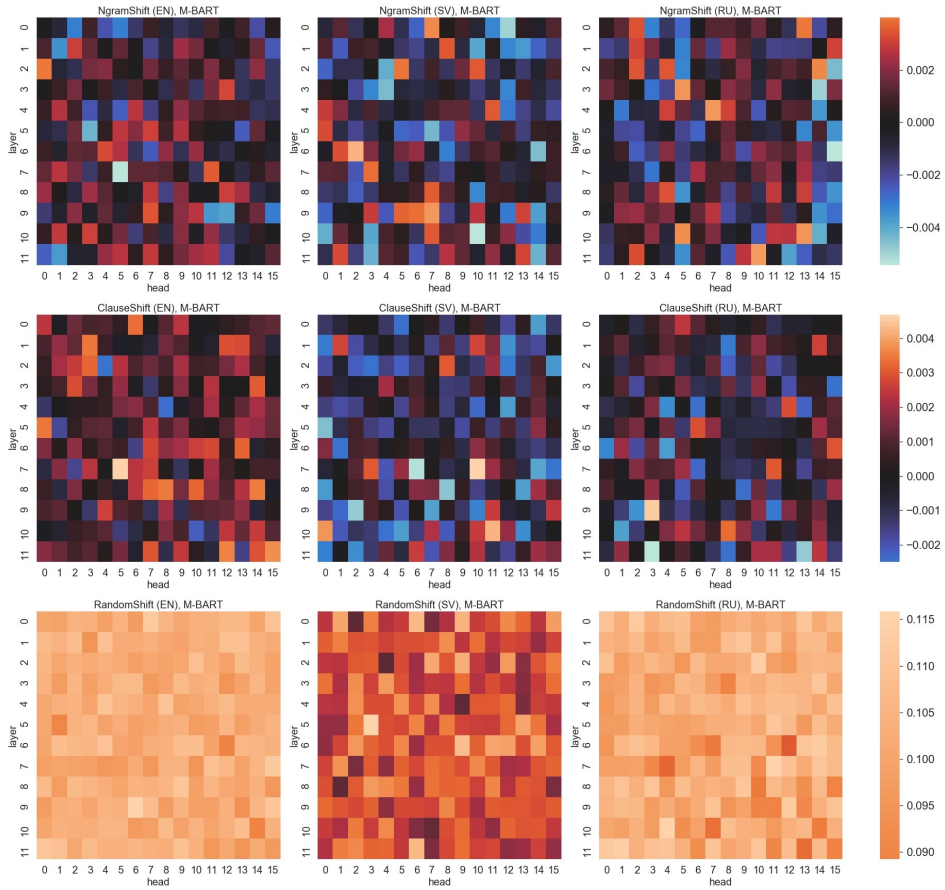


Figure 1: The task-wise heatmaps depicting the δ UUAS scores by M-BART for each language. Method=**Self-Attention Probing**. PE=**absolute**. X-axis=Attention head index. Y-axis=Layer index. Tasks: **NgramShift** (top); **ClauseShift** (middle); **RandomShift** (bottom). Languages: **En**=English (left); **Sv**=Swedish (middle); **Ru**=Russian (right).

	Language	M-BERT	M-BART
NgramShift	**En**	0.32; 0.33	0.31; 0.32
	Sv	0.30; 0.31	0.30; 0.31
	Ru	0.36; 0.38	0.37; 0.38
ClauseShift	**En**	0.20; 0.21	0.20; 0.21
	Sv	0.20; 0.21	0.20; 0.21
	Ru	0.20; 0.21	0.20; 0.21
RandomShift	**En**	0.37; 0.38	0.37; 0.37
	Sv	0.38; 0.41	0.37; 0.39
	Ru	0.39; 0.42	0.40; 0.41

Table 2: The UUAS scores by **Self-Attention Probing** method. The minimum and maximum values are given (min; max). Languages: **Ru**=Russian, **En**=English, **Sv**=Swedish.

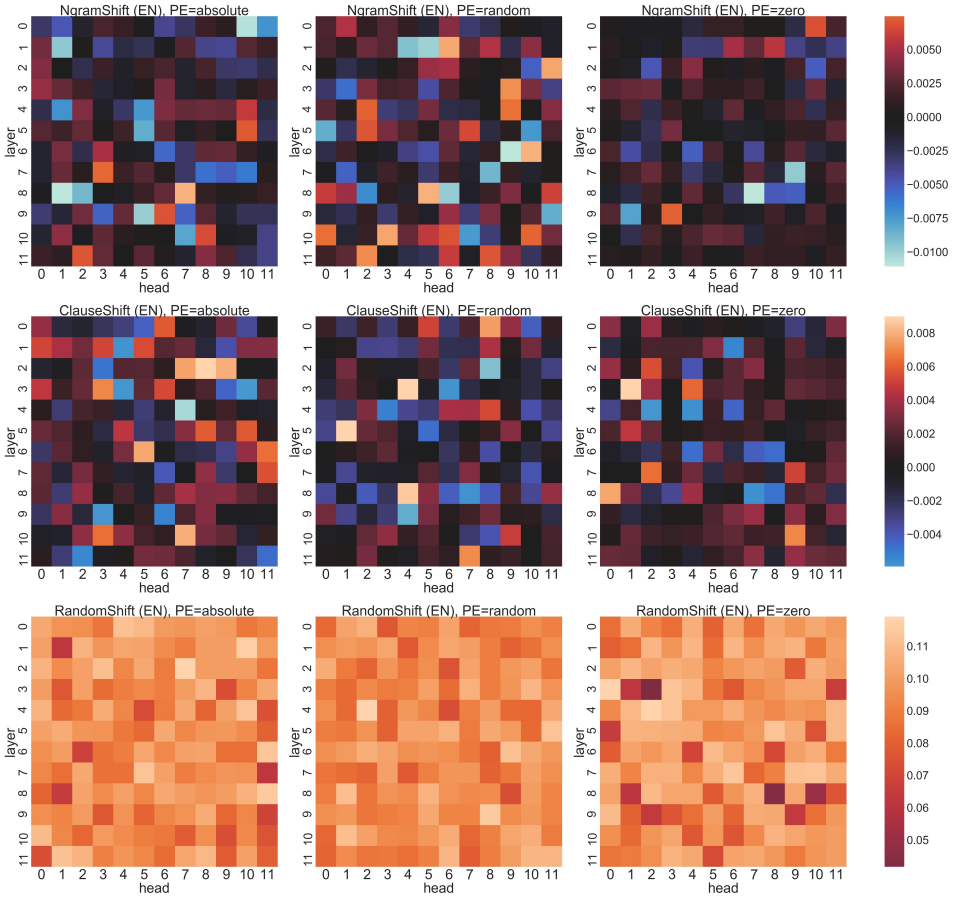


Figure 2: The task-wise heatmaps depicting the δ UUAS scores by M-BERT for each language. Method=**Self-Attention Probing**. PE: **absolute** (left); **random** (middle); **zero** (right). X-axis=Attention head index. Y-axis=Layer index. Tasks: **NgramShift** (top); **ClauseShift** (middle); **RandomShift** (bottom).

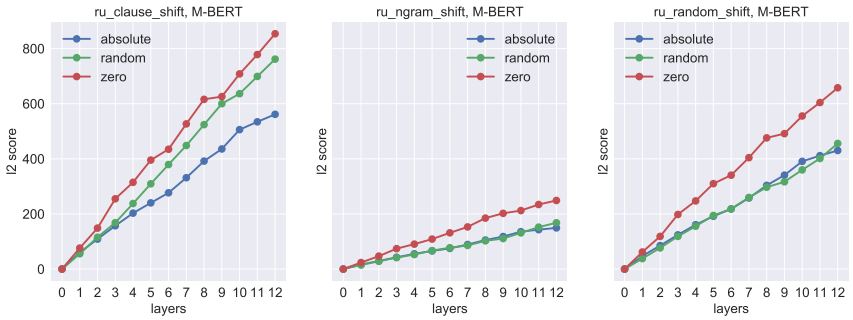


Figure 3: The Euclidean distance between the impact matrices computed by M-BERT with different PEs over each pair of sentences (s, s') for Russian. The distances are averaged over attention heads at each layer. Method: **Token Perturbed Masking**. Tasks: **NgramShift** (left); **ClauseShift** (middle); **RandomShift** (right)

.

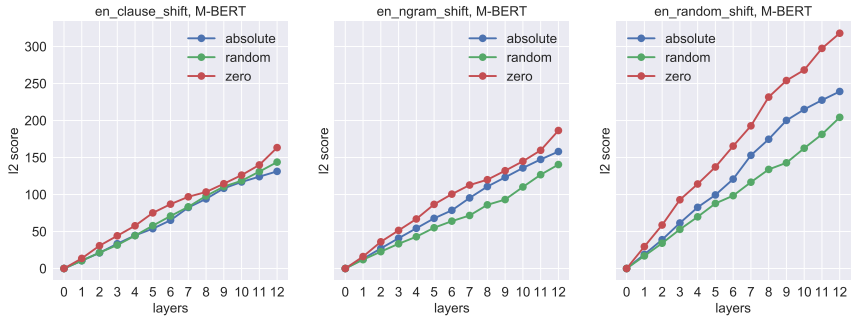


Figure 4: The Euclidean distance between the impact matrices computed by M-BERT with different PEs over each pair of sentences (s, s') for English. The distances are averaged over attention heads at each layer. Method: **Token Perturbed Masking**. Tasks: **NgramShift** (left); **ClauseShift** (middle); **RandomShift** (right)

.

Figure 5: The Euclidean distance between the impact matrices computed by M-BART with different PEs over each pair of sentences (s, s') for Swedish. The distances are averaged over attention heads at each layer. Method: **Token Perturbed Masking**. Tasks: **NgramShift** (left); **ClauseShift** (middle); **RandomShift** (right)

.

Figure 6: The Euclidean distance between the impact matrices computed by M-BART with different PEs over each pair of sentences (s, s') for English. The distances are averaged over attention heads at each layer. Method: **Token Perturbed Masking**. Tasks: **NgramShift** (left); **ClauseShift** (middle); **RandomShift** (right)

.

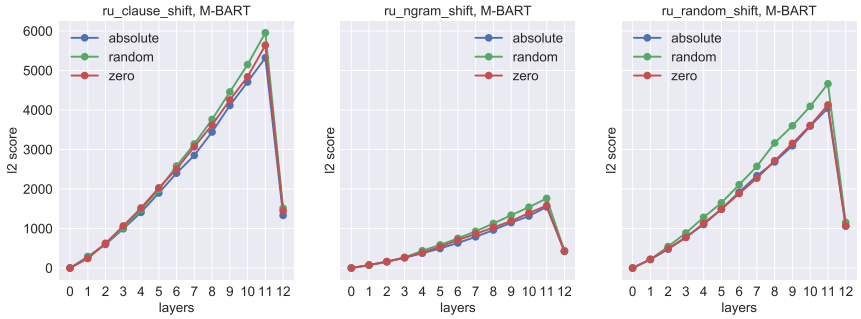
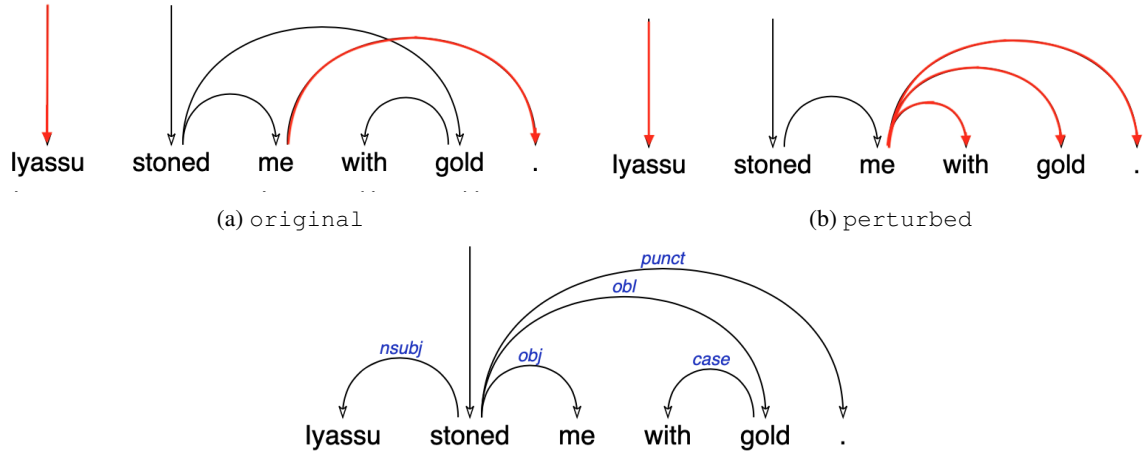


Figure 7: The Euclidean distance between the impact matrices computed by M-BART with different PEs over each pair of sentences (s, s') for Russian. The distances are averaged over attention heads at each layer. Method: **Token Perturbed Masking**. Tasks: **NgramShift** (left); **ClauseShift** (middle); **RandomShift** (right)

.

	Language	M-BERT	M-BART
NgramShift	**En**	0.36; 0.44	0.31; 0.44
	Sv	0.37; 0.46	0.43; 0.53
	Ru	0.4244; 0.52	0.46; 0.54
ClauseShift	**En**	0.23; 0.29	0.23; 0.29
	Sv	0.25; 0.33	0.24; 0.33
	Ru	0.23; 0.28	0.22; 0.28
RandomShift	**En**	0.39; 0.47	0.39; 0.47
	Sv	0.46; 0.53	0.43; 0.53
	Ru	0.46; 0.54	0.43; 0.54

Table 3: The UUAS scores by **Token Perturbed Masking** probe. The minimum and maximum values are given (min; max). Languages: **Ru**=Russian, **En**=English, **Sv**=Swedish.

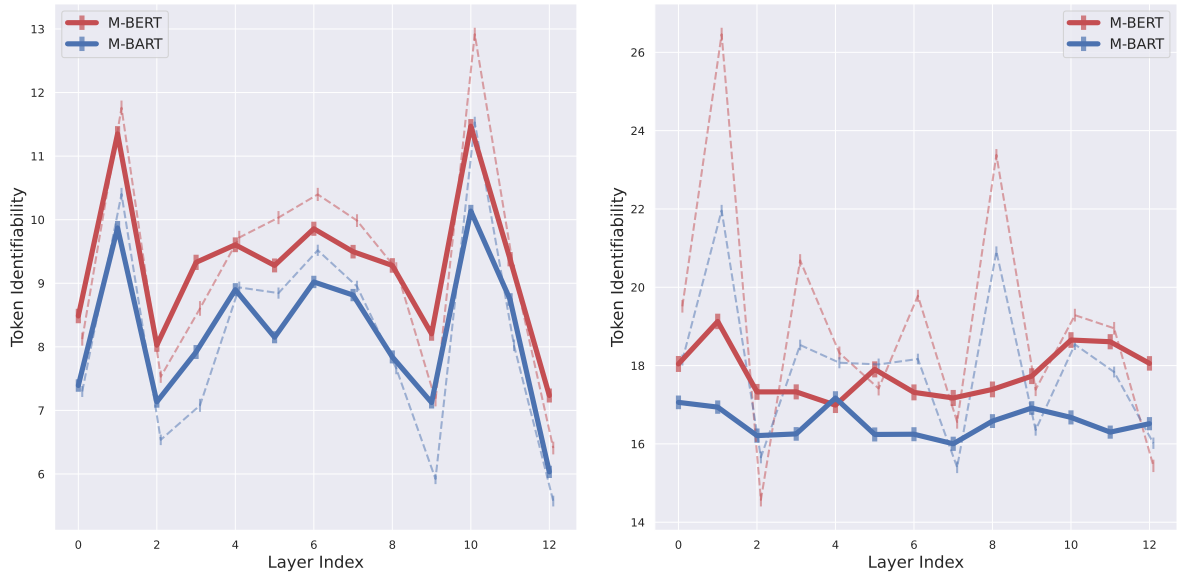


(a) `original` (b) `perturbed`

(c) `gold`

Figure 8: Graphical representations of the syntactic trees inferred for the English sentence *Iyassu stoned me with gold* and its perturbed version. `original`=the original sentence; `perturbed`=the perturbed version; `gold`=gold standard. Task=**RandomShift**. Model=**M-BERT** (Layer: 11; Head: 2). Method=**Self-Attention Probing**. The perturbation is underlined with red, and incorrectly assigned dependency heads are marked with red arrows.

3 Representation Analysis



(a) **NgramShift** (b) **ClauseShift**

Figure 9: Token identifiability (TI) by layer for M-BERT and M-BART on the **NgramShift** (left) and **ClauseShift** (right) tasks for Russian. Dashed lines represent the scores computed over the intact sentences. X-axis=Layer index. Y-axis=TI.

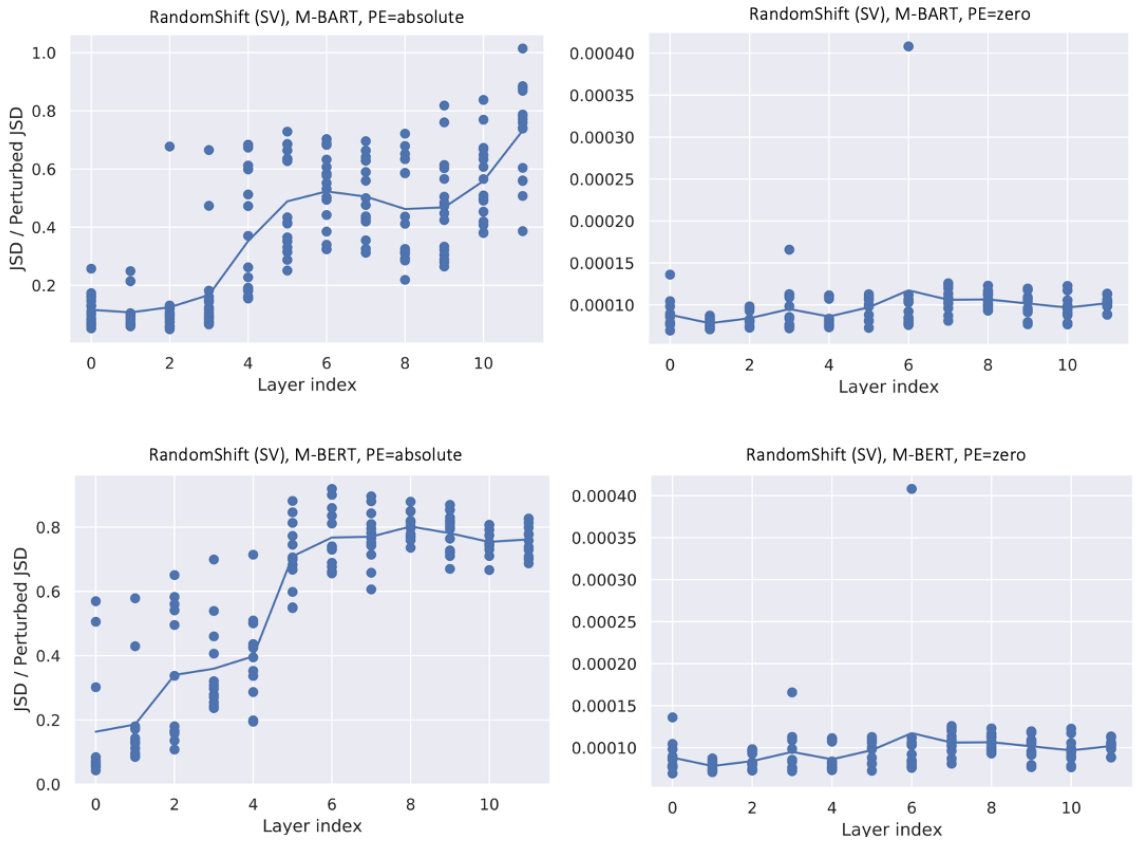


Figure 10: Self-Attention Distance (SAD) by layer for M-BART and M-BERT with absolute (left) and zeroed (right) positional embeddings on the **RandomShift** task for Swedish. X-axis=Layer index. Y-axis=SAD.

4 Acceptability Judgements

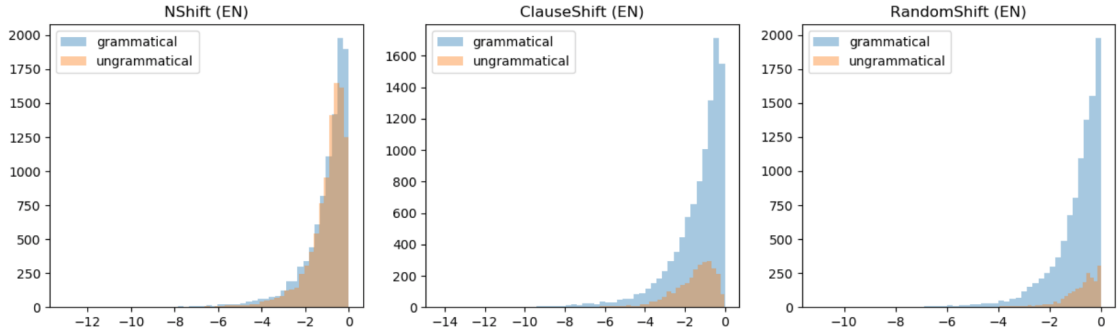


Figure 11: The *MeanLP* distributions for the perturbed (ungrammatical) and intact (grammatical) sentences by M-BART. Tasks: **NgramShift** (left); **ClauseShift** (middle); **RandomShift** (right).

.

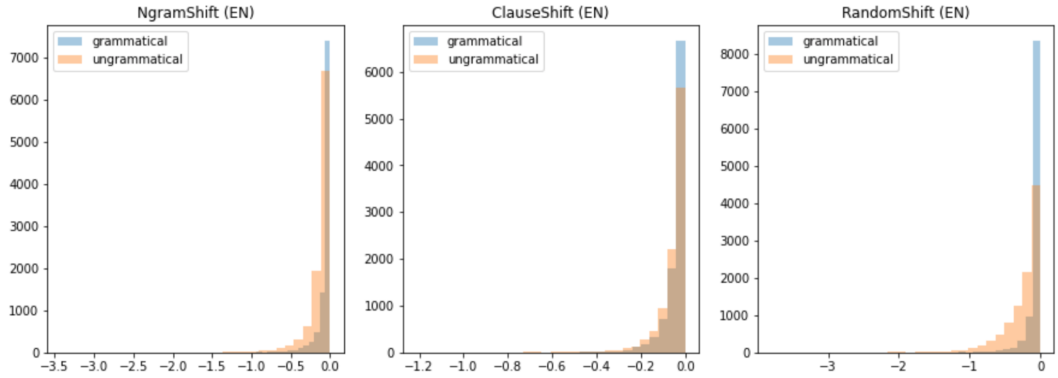


Figure 12: The *MeanLP* distributions for the perturbed (ungrammatical) and intact (grammatical) sentences by M-BERT. Tasks: **NgramShift** (left); **ClauseShift** (middle); **RandomShift** (right).

.