Sometimes We Want Ungrammatical Translations

Prasanna Parthasarathi[1,2], Koustuv Sinha[1,2,3], Joelle Pineau[1,2,3] and Adina Williams[3]
[1] School of Computer Science, McGill University, Canada
[2] Quebec AI Institute (Mila), Canada
[3] Facebook AI Research (FAIR)
{prasanna.parthasarathi, koustuv.sinha, jpineau, adinawilliams}
@{mail.mcgill.ca, mail.mcgill.ca, cs.mcgill.ca, fb.com}

Abstract

Rapid progress in Neural Machine Translation (NMT) systems over the last few years has focused primarily on improving translation quality, and as a secondary focus, improving robustness to perturbations (e.g. spelling). While performance and robustness are important objectives, by over-focusing on these, we risk overlooking other important properties. In this paper, we draw attention to the fact that for some applications, faithfulness to the original (input) text is important to preserve, even if it means introducing unusual language patterns in the (output) translation. We propose a simple, novel way to quantify whether an NMT system exhibits robustness or faithfulness, by focusing on the case of word-order perturbations. We explore a suite of functions to perturb the word order of source sentences without deleting or injecting tokens, and measure their effects on the target side. Across several experimental conditions, we observe a strong tendency towards robustness rather than faithfulness. These results allow us to better understand the trade-off between faithfulness and robustness in NMT, and opens up the possibility of developing systems where users have more autonomy and control in selecting which property is best suited for their use case.

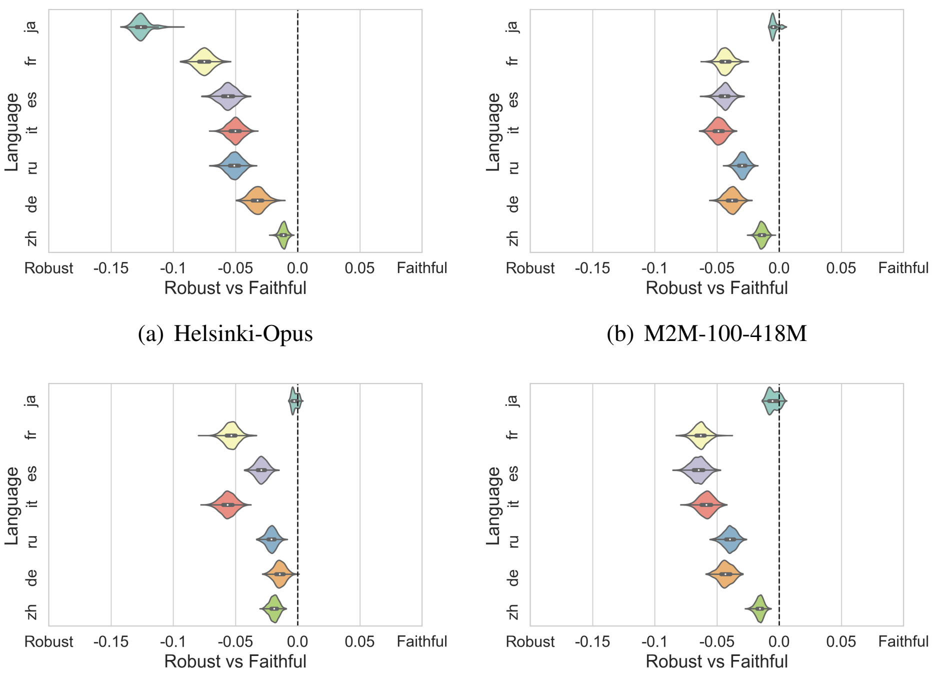
1 Introduction

Recent advances in Neural Machine Translation (NMT) have resulted in systems that are able to effectively translate across many languages (Fan et al., 2020a), and we have already seen many commercial deployments of NMT technology. Yet some studies have also reported that NMT systems can be surprisingly brittle when presented with out-of-domain data (Luong and Manning, 2015), or when trained with noisy input data containing small orthographic (Sakaguchi et al., 2017; Belinkov and Bisk, 2018; Vaibhav et al., 2019; Niu et al., 2020) or lexical perturbations (Cheng et al., 2018). Uncovering these sorts of errors has lead the research community to develop new NMT models that are more *robust* to noisy inputs, using techniques such as targeted data augmentation (Belinkov and Bisk, 2018) and adversarial approaches (Cheng et al., 2020). Unfortunately an approach that (over-)emphasized robustness can lead to "hallucinations"—translating source input to an output that is not *faithful* to the source, and sometimes is even factually incorrect (Vinyals and Le, 2015; Koehn and Knowles, 2017; Wiseman et al., 2017; Nie et al., 2019; Kryscinski et al., 2020; Maynez et al., 2020; Tian et al., 2020; González et al., 2020; Xiao and Wang, 2021). Moreover, such an approach hinges on the key assumption that orthographic, lexical or grammatical variants in the input are *mistakes*, to be *corrected* by the translation system. This ignores the wealth of applications where it may be preferable for a system to offer more *faithfulness* to the original text.

It is worthwhile to consider the diversity of applications where having a faithful translation (opting literal translation over paraphrasing) is desirable. First, consider an automatic language tutoring system: a (human) second-language learner will often produce language that has grammatical mistakes of various types. This learner can be empowered by having a (AI-produced) faithful translation, so that s/he can see what mistakes were made vs. what would be the more common phrasing. Second, recall that many languages, including English, use word order to encode argument structure information (cf. Isabelle et al. 2017): while "the dog bit the man" might be more frequent compared to "the man bit the dog", the latter has a very clear meaning that we may wish to preserve in some (albeit rarer) cases. Third, consider poetry: it is often the case that unusual word order is used to influence rhythm and rhyme. It would be a shame if all our state-of-the-art NMT systems lost such poetic beauty in translation.

In short, by their very design, NMT systems

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3205–3227
November 7–11, 2021. ©2021 Association for Computational Linguistics



(a) Helsinki-Opus

(b) M2M-100-418M

(c) mBART50

(d) M2M-100-1.2B

Figure 1: Machine translation systems (a–d) tend to favor robust or faithful translations as measured by computing the difference between **faithfulness** and **robustness** scores across seven languages (aggregated across all perturbations). Although models with different sizes were analysed, we did not find a strong correlation between the robustness or faithfulness to the model sizes. But, M2M-100-1.2B showed a higher tendency to be robust when compared with M2M-418M or mBART (smaller that M2M-1.2B model).

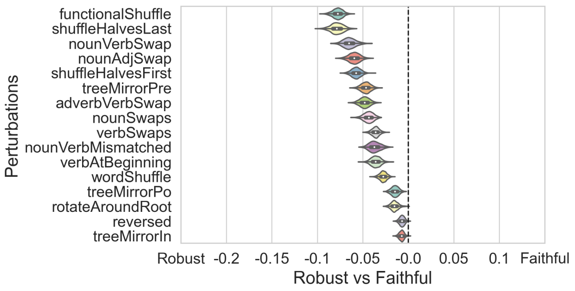


Figure 2: Averaging across languages and NMT systems shows they tend to favor robust translations, although this varies for different perturbations.

preferentially output "normative" language (regardless of whether the nonstandard languages affects spelling, word order, or choice of vocabulary). Isozaki et al. (2010) note that word order is an important problem in distant-language translation. When we increase model robustness (at least with the solutions proposed to date), we generally enforce even stronger tendencies towards the norm, at the expense of diversity of language, of thought, and, perhaps, of our very culture. Although Bisazza et al. (2021)'s observation on word order flexibility only minimally affect the performance of NMT systems is encouraging towards building robust systems, the trade-off on preserving diversity in expression is seldom understood. We believe it will be necessary in future to propose solutions that can explicitly enable a better trade-off between robustness and faithfulness, and can give the user autonomy and control in specifying their preference. It is therefore our goal with this work to draw attention to this important compromise, and to provide tools to detect, quantify, and compare such aspects of NMT systems.

More specifically, this paper is not only the first to deeply analyze the effects of particular perturbations on existing NMT systems, but is the first to investigate their effects in the sphere of generation. We investigate 16 unique perturbations that fall into three categories—*Dependency tree based*, *PoS-tag based* and *Random Shuffles*. We introduce two novel metrics for evaluating machine translation models' preference for robustness or faithfulness. Taking English as the common source, we run a case study with three widely used Transformer-based machine translation models —Helsinki/OPUS machine translation model (Tiedemann and Thottingal, 2020), the multilingual BART model (Liu et al., 2020a), and the Many-to-Many Multilingual translation model (Fan et al., 2020a) (in two sizes)—into 7 target languages from several families (German, French, Spanish, Italian, Russian, Chinese, and Japanese).

Across several experimental conditions, we observe a strong tendency towards robustness rather than faithfulness (Figure 1) that varies somewhat depending on the particular perturbation (Figure 2). More specifically, we observe that (1) state-of-the art NMT systems tend to produce translations that are unaffected by the noisy source (more robust), (2) accuracy (BLEU score) correlates with model robustness, (3) certain perturbations involving part-of-speech-based word reordering tend to further encourage robustness, and (4) results vary by somewhat by target language, with the models producing translations of Japanese that are more faithful than for the other languages (except for Helsinki-OPUS). Overall, our analysis suggests that over-emphasizing accuracy and robustness may limit richer development and broader usefulness of NMT systems.

2 Related Work

The idea to randomly shuffle linguistic elements to evaluate NLP model performance goes back fairly far (Barzilay and Lee, 2004; Barzilay and Lapata, 2008), and has even been used to determine which tasks are "syntax-light" in human sentence processing (Gauthier and Levy, 2019). Recent work on classification tasks, such those on the GLUE benchmark (Wang et al., 2018), has shown that pre-trained Transformer-based models trained with a masked language modeling objective are shockingly insensitive to word order permutations. (Si et al., 2019; Sinha et al., 2020; Pham et al., 2020; Gupta et al., 2021; Sinha et al., 2021). Given these recent findings, we might expect insensitivity to word order permutation in the sphere of generation as well, leading to robust machine translations.

The mismatching of default word orders between target and source has long been an important consideration for multilingual tasks including automatic machine translation. Ahmad et al. (2019) find that word order agnostic models (recurrent neural networks) trained to dependency parse can transfer better than word order sensitive ones (self-attention) to distantly related languages. Also in the context of transfer, Zhao et al. (2020) propose for reference-free MT that the delta between originally ordered and permuted sentences be used as an evaluation technique. Even when considering multilingual sequence labeling tasks in general, Liu et al. (2020c); Kulshreshtha et al. (2020) find that limiting word order information in the multilingual setting can enable models to achieve better zero-shot cross-lingual performance. Taken together, these works also suggest that our models tend to overfit on source word order to the detriment of that of the target, which might lead one to predict that our models will be more robust than they are faithful in our case as well.

However, NMT systems have use cases in diverse applications that require the preservation of word order, local syntax and other linguistic components (Zhang et al., 2020). Translation systems that are contingent on preserving syntax and semantics are used as interpretors to decode the interaction between components of a neural network (Andreas et al., 2017). Further, in practical applications like translating a sentence that is a mixture of two different languages requires the MT systems to strike some balance between preserving L1 syntax and/or word-order and correctly adhering to the grammatical rules of L2 (Renduchintala et al., 2016).

In NLP tasks, where the end-user could be a human, benchmarking the robustness of NLP systems is done by evaluating a model's performance on willfully perturbed examples that could potentially expose fragility of the systems (Goodfellow et al., 2014; Fadaee and Monz, 2020). Towards averting such scenarios, efforts along the lines of building robust models with adversarial training have been a common topic of study in natural language processing (Rajeswar et al., 2017; Wu et al., 2018).

Our word order perturbations also share some points of synergy with work across NLP that aims to devise supplementary heuristics to explicate the inner workings of our machine learning systems. For specific NLP tasks, probe tasks are engineered to measure specific kinds of linguistic knowledge encoded in the systems (Conneau et al., 2018; Sheng et al., 2019; Kim et al., 2019; Jeretic et al., 2020; Parthasarathi et al., 2020; Ribeiro et al., 2020). Swapping the arguments of verbs is a classic way to measure the effects of word order both in humans (Frankland and Greene, 2015; Snell and Grainger, 2017) and in models, largely because changing the order of verbal arguments maintains high word overlap between related examples (Wang et al., 2018; Kann et al., 2019; McCoy et al., 2019); However, although limited word order permutation is applied in this case, it is generally restricted to licit, grammatical sequences of words. When perturbation has been used to evaluate model performance, the utilized perturbation functions have been predominantly fairly simple, including reverse and word shuffle, and usually target only single sentences (Ettinger, 2020; Li et al., 2020; Sinha et al., 2020). For tasks like dialogue prediction that requires multiple input sentences, perturbation functions like reordering the conversation history have been adopted (Sankar et al., 2019). To the best of our knowledge, the set of perturbation functions we propose is the most detailed set explored thus far, perturbing not only tokens, but PoS and dependency structure.

Changing the order of words in the context of NMT also has its roots in classical, syntactically sophisticated models that used parses (of various kinds) to pre-order abstract syntactic representations as an early step in a multi-step translation pipeline from source to target (Collins et al. 2005; Khalilov et al. 2009; Dyer and Resnik 2010; Genzel 2010; Khalilov and Sima'an 2010; Miceli-Barone and Attardi 2013 i.a.). Our approach differs from these approaches in that our main aim is not to incorporate word order changes into the translation pipeline itself, but, instead use them to better understand the behavior of NMT models.

3 Metrics

Let g_x be a sentence where x takes one of two values: e if it is a sentence in the source language (English) or o if it is a gold target sentence. Let $\Phi_{e \to o}$ denote a translation pipeline from the English source (e) to a target language (o) and $t_o \leftarrow \Phi_{e \to o}(g_e)$ for a language $o \in O \;\sim \{$*German* (de), *French* (fr), *Spanish* (es), *Italian* (it), *Russian* (ru), *Japanese* (ja), *Chinese* (zh)$\}$.

Let Ψ denote a perturbation function such that $g_x^- \leftarrow \Psi(g_x)$; then let the translation of perturbed input g_e^- be $t_o^- \leftarrow \Phi_{e \to o}(g_e^-)$.

Let $\kappa(s_i, s_j)$ be a scoring function that rates the similarity between two sentences (s_i and s_j), where $s_i, s_j \in L_x$. The choice of κ can be any of the widely used sentence similarity metrics like BLEU (Papineni et al., 2002a), METEOR (Lavie and Agarwal, 2007), ROUGE (Lin, 2004), or Levenshtein-distance (Levenshtein, 1966). For our purposes, we experiment with BLEU-4, BLEURT (Sellam et al., 2020), BERT-Score (Zhang et al., 2019) and Levenshtein score as choices of κ denoted by a B or L in the superscript respectively (but see §7 for discussion of other κ). The value of κ linearly scales with the similarity between s_i and s_j.

We define three metrics β_1, β_2, and α. β_1 is our measure of **robustness** to perturbation by quantifying the similarity according to κ between the translation of a perturbed sentence in source into the target, and the gold sentence in target: $\beta_1 \leftarrow \frac{1}{N}\sum_{i=1}^{N} \kappa(g_o, t_o^-)_i$, where N denotes the number of samples[1] perturbed by Ψ that we used (see Table 1 in the Appendix for more information on N by perturbation and language).

β_2 is computed as a similarity score between the translation of the perturbed source sentence and applying the same perturbation operation on the target sentence to measure degree of **faithfulness** of translations by machine translation system: $\beta_2 \leftarrow \frac{1}{N}\sum_{i=1}^{N} \kappa(g_o^-, t_o^-)_i$.

The **difficulty** of the perturbation function is measured with α, which scores the similarity between perturbed sentence and the unperturbed sentence in the source language: $\alpha_e \leftarrow \frac{1}{N}\sum_{i=1}^{N} \kappa(g_e, g_e^-)_i$.

β measures the **standard translation performance metric** on any given source-target sentence pair: $\beta \leftarrow \frac{1}{D}\sum_{i=1}^{D} \kappa(g_o, t_o)_i$, where D is the size of the dataset.

4 Perturbations

We propose 16 different functions to perturb the structure of an input sentence. The perturbations can be broadly classified in three categories—*Random Shuffles*, *PoS-tag Based* and *Dependency Tree Based*—comprised of 4, 8, and 4 perturbation functions respectively. The functions vary in complexity and linguistic sophistication so that we can score whether a model translates faithfully or stays robust to the perturbed inputs. We applied all perturbations in seven languages—*de*, *fr*, *ja*, *ru*, *zh*, *it*, and *es*—and describe each perturbation in turn below. See Figure 3 for a selection of examples.

Some of the perturbations we explore are "possible", in the sense that applying them will result (in most cases) in a grammatical sentence

[1] Perturbation functions have certain entry conditions to be applied on a sample. For example, *verbSwaps* mandates that there is at least 2 verbs in the sample. So, in a D size dataset not all samples can be perturbed with all the functions, so we define N independent to D.

TreeMirrorPost: to live a decent place he could n't find Tom said .

TreeMirrorPre: said find place live to a decent he could n't Tom .

TreeMirrorIn: live to place a decent find he could n't said Tom .

RotateAroundRoot: live find said Tom he could n't a decent place to .

WordShuffle: place to could live said decent a Tom n't find he .

Reversed: live to place decent a find n't could he said Tom .

(a)

VerbSwaps: Tom live he find n't said a decent place to could .

NounSwaps: Tom said a decent place could n't find he to live .

NounVerbSwaps: said Tom could he n't a decent place find to live .

NounVerbMismatched: live a decent place find could n't he said to Tom .

ShuffleFirst: he Tom find could said n't a decent place to live .

ShuffleLast: Tom said he could n't find a decent live place to .

(b)

Figure 3: Effect of the different perturbation functions on the sentence — *Tom said he could n't find a decent place to live.* The perturbation functions do not inject new tokens or delete a token to perturb the sentence.

(either in the source language, or in some version of another existing language that is instead supplied with the words of the source). Others are "impossible" (Moro, 2015, 2016). For example, it has been long noticed that human grammar rules operate on hierarchical structure resulting in rules of the form "move the hierarchically closest auxiliary" as opposed to "move the linearly closest auxiliary" when forming questions (Chomsky 1962/2013; Ross 1967; Crain and Nakayama 1987, i.a.). Standard American English exemplifies this: when we form a question from "The man who is tall was happy", we say "Was the man who is tall happy?" not "Is the man who tall was happy?" (McCoy et al. 2020, cf. Chomsky 1957, Ch. 3). To explore more fully the behavior of the NMT models, we include several permutations that neither adhere to the descriptive rules of the source language nor to any grammars across all known human languages (i.e., are "impossible").

4.1 Random Shuffles

The perturbations in the Random bin treat the sentence as though it were a mere sequence of tokens; they reorder the tokens without any reference to their higher order linguistic properties (i.e., PoS or dependency information). Thus, random perturbations can be seen as the most basic type of "impossible" word order perturbation. We use three different random shuffles— *Word-Shuffle*, *Shuffle-First-Half*, *Shuffle-Last-Half* and *Reversed*—none of which result in any recognizable linguistic structure. *Word-Shuffle* shuffles the entire sentence at random (cf. Sinha et al. 2020); for a sentence of length n, there are $(n-1)!$, possible random permutations. *Shuffle-First-* and *Shuffle-Last-Halves* shuffle only the corresponding half of a sentence while keeping the other half unperturbed. *Reversed* reverses the token ordering in a sentence.

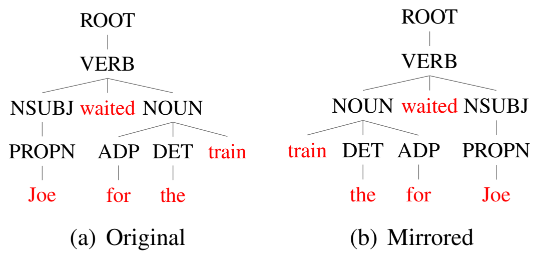
4.2 Part-of-Speech tag Based Perturbations

This set of perturbations uses the PoS tags from a parser to generate perturbations for a sentence, so that we can localize any effects of robustness or faithfulness to particular linguistic categories.

PoS Swaps. When a sentence has more than one token with a particular PoS, the positions of those tokens are exchanged without affecting the rest of the sentence structure.[2] Although the meanings of the sentences are altered, the result generally is grammatical (or near grammatical, see Figure 3(b)), meaning that these swaps are "possible". In this class of permutations, we consider Noun swaps and Verb swaps.

PoS$_X$-PoS$_Y$ Swaps. The position of a token with a particular PoS tag $X \in \{noun, adv\}$ is interchanged with the linearly closest token with PoS tag $Y \in \{verb, adj\}$ leaving the rest of the sentence unperturbed. In this class, we consider Adverb-Verb swaps and Noun-Adjective swaps (which tend to result in grammatical sentences),

[2] Except for cases where person agreement might be affected, for example when verb-swapping "am" for "are" in *I am happy that they are here→ I are happy that they am here.*



(a) Original (b) Mirrored

Figure 4: For tree based perturbations, we mirror the dependency tree and perform InOrder, PreOrder (Root-Left-Right), and PostOrder (Left-Right-Root). (The grammatical relations are excluded for brevity.)

as well as Noun-Verb swaps (which tend to result in ungrammatical sentences).

PoS$_{\text{Noun}}$-PoS$_{\text{Verb}}$ Mismatched Swaps. While Noun-Verb Swap replaces each noun with the verb closest to it, the mismatched swap exchanges the position of a noun with the verb *farthest* from it, which results in displacing all verbs and nouns from their original positions.

Functional Shuffle. Functional tokens (i.e., conjunctions, prepositions and determiners) are re-ordered so that they occupy the original position of another functional token in the perturbed sentence.

Verb-At-Beginning. This perturbation moves a verb to the beginning of the sentence as a prefix without disturbing the remaining relative positions within the text. If the sentence has multiple verbs, the first verb found when parsing the sentence will be moved to the beginning.

4.3 Dependency Tree Based

The dependency tree structure of a sentence conveys its grammatical structure. Perturbing the dependency tree in a language like English—which expresses verb-argument relationships largely via word order— could have several effects: the semantics of the sentence will be changed, and the base word order might now be indicative of a different family of languages. Therefore, we investigate dependency tree perturbations with an eye towards determining whether perturbations that result in sentence structures from another family (e.g., Japanese) will be more faithfully translated.

Tree Mirror (Pre/Post/In). While an In-Order traversal of a sentence's dependency tree (Figure 4) provides the right parse of the sentence, we perform Pre-Order, Post-Order and In-Order traversals on the mirrored dependency tree. Although the perturbed sentences largely preserve each word's position with respect to its local neighbors, since they are ungrammatical, their meanings (if there are any) are much harder to understand.

Rotate Around Root. The sentence is perturbed by rotating the tree around its root and then subsequently performing an In-Order traversal.

4.4 Distribution

We observe in Figure 5 that the dependency tree-based perturbation functions have less overlap with the PoS tag-based perturbations across languages, but higher intra-category similarity scores.

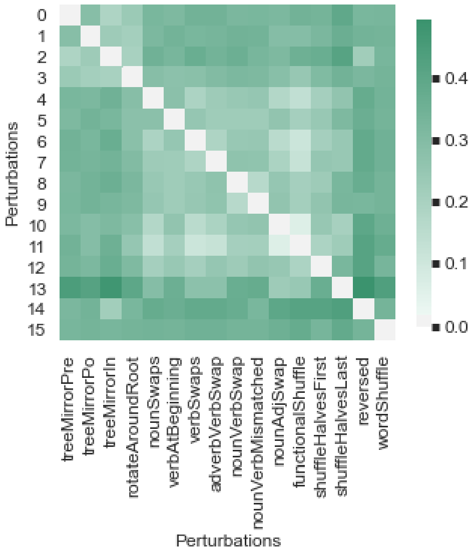


Figure 5: $\kappa\left(\Psi_i(s), \Psi_j(s)\right)$ highlight the differences between the three categories of perturbations in English. The trend is similar across the other languages (Figure 14 (Appendix)).

Similarly the PoS tag-based functions have understandably higher similarity with other PoS tag-based functions than with Shuffle or Dependency tree perturbation functions.

5 Experiments

We experiment with some of the state of the art translation models — OPUS translation models (Tiedemann and Thottingal, 2020), MBART (Liu et al., 2020b), Facebook's M2M (Fan et al., 2020b) (both 418M and 1.2B models). We construct the perturbed dataset using the eval set of OPUS corpus (Tiedemann and Thottingal, 2020) in 7 different languages paired with English as source —French (fr), German (de), Russian (ru), Japanese (ja), Chinese (zh), Spanish (es), and Italian (it). Our experiments[3] have a twofold objective: (1) compute the robustness (β_1) and faithfulness (β_2) of the translations in different languages when the input is perturbed, and (2) analyse the β_1 and β_2 scores with different levels of perturbations.

6 Results

6.1 Faithfulness vs. Robustness

For each language paired with English, we perturb the source English and the gold target language with the perturbation functions proposed in §4. We measure β_1 and β_2 with BLEU-4 (Papineni et al., 2002b), BLEURT (Sellam et al., 2020) and BERT-Score (Zhang et al., 2019) as the choice for κ. As BERT-Score and BLEURT were forgiving to the flaws[4] in predictions towards being robust, we base our analysis with BLEU-4 as the choice of κ.

We observe that β_1 scores are generally higher than β_2 scores across the perturbation functions and across all the languages, indicating that the translation system is largely unfazed when presented with *unnatural*, *ungrammatical* input (see Figure 1)[5]. Given these results, the model acts as though it makes an intermediate "hallucination" that somehow either recreates the unperturbed input before translating it, or "hallucinates" an unperturbed target without much reference to the perturbed source.

6.2 Patterns in β_1 and β_2, and Length

Given our results, we would like to know whether there are any particular properties of particular examples or of permutations which lead models to be more or less robust. Towards that end, we observe the correlations between (a) β vs β_1/β_2 (b) β_1 vs β_2, and (c) β_1/β_2 vs $Length$ of source sentence.

β vs β_1/β_2. We find that our β_1 does correlate with BLEU-4 on the translation of the original, unperturbed gold English sentence and gold target language. We show correlations of β_1 and β_2 with β in Figure 7. The Spearman's rank correlation between β_1 and β is larger than between β_2 and β; in the former we observe a medium strength effect and in the latter a small effect, although language does play a role (e.g., Chinese has the largest β_1 correlation with BLEU, but among the smallest β_2 correlation with BLEU).

β_1 vs β_2. Figure 6(a) shows that the correlation between robustness and faithfulness to be present, but weak. By definition, the model can either be faithful or robust and when it is both, then that suggests only a higher α_e or a lower perturbation difficulty. Usually this occurs when sentences are very short—for short sentences, fewer permutations are possible, and different permutation functions are more likely to collapse onto the same word orders.

β_1/β_2 vs *Length*. The length of the source sentence has different effects on the scores depending largely on language. But, it is intuitive to understand that the model is better able to fix a word order perturbation when the sentences are short, resulting in higher β_1 score for shorter sentences. The opposite is true for β_2 where longer sentences generally have higher β_2 score.

There is some relationship between which permutation function generated a permuted example and its α_E score (Figure 12). The top 5 permutation functions with high α_E scores—{*shuffleHalvesLast*, *shuffleHalvesFirst*, *verbAtBeginning*, *nounVerbSwap*, *nounVerbMismatched*}—and with low α_E scores—{*treeMirrorPost*, *wordShuffle*, *reversed*, *treeMirrorIn*, *treeMirrorPre*}. The mix of examples from different perturbation categories at different levels of α_E score, as well as the fact that β_1 scores are higher than β_2, suggests that models' attempting to correct the perturbed input may not be because they understand language, but instead it might be due to correlations between certain n-grams in the sentence. We also observe that β_1 decreases with increasing α_E^L, which also supports this argument.

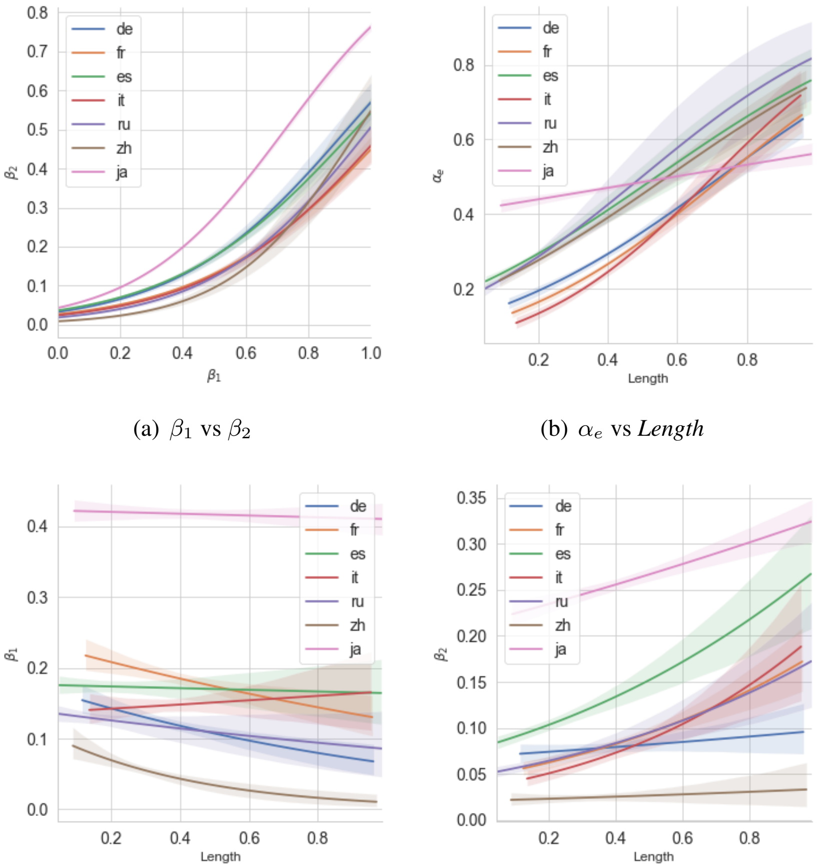
7 Discussion

Languages Vary. One way to think about the models' tendency towards behaving robustly is to take them to be hallucinating an unperturbed response even when the word order of the original is perturbed. The difference between β_1 and β_2 (Figure 1) shows a ranking across languages, and with perturbation functions. Among the languages analysed, Japanese in Helsinki is generally more robust than the other languages. However, we note that our findings could also be attributed to the strength of the translation system—Japanese in Helsinki has the highest performance (Table 2) and the strong

[3] The code for reproducing the metrics and perturbation functions can be found in the code repository here.

[4] Figure 8, 9, and 10 in Appendix C.

[5] More discussion can be found in Figure 13 in Appendix D.



(a) β_1 vs β_2

(b) α_e vs *Length*

(c) β_1 vs *Length*

(d) β_2 vs *Length*

Figure 6: We observe the length of the source sentences to differently correlate with the two scores. The robustness score, β_1, is higher for shorter source sentences, while the opposite is true for β_2 suggesting that the model's ability to see through the syntactic errors has a limitation on the length. Also, the model being able to stay faithful in longer sentences can be explained with higher α_e hinting at their lower difficulty.

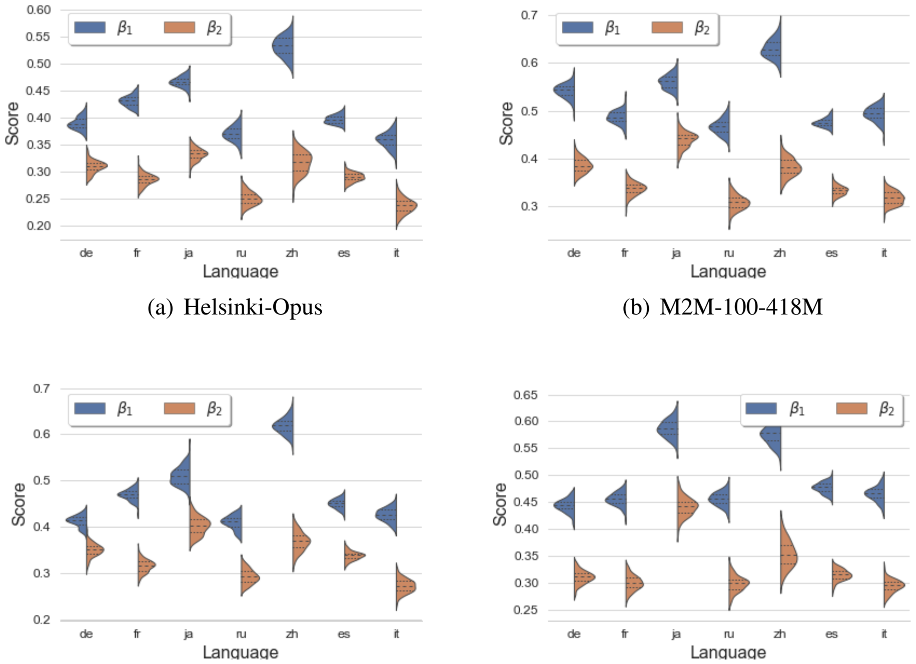
correlation between β and β_1 support the argument. Also, the weak β_1 and β_2 scores of Chinese translation model could also be attributed to the general poor performance of the translation systems for the language (Table 2 shows that the β scores of the Chinese model are too low).

Perturbation Functions. Among the perturbation functions, *FunctionalShuffle* evoked the most robust generation across all languages while models were most faithful on *TreeMirrorIn* and *Reversed*. Recall, however, the fact that all languages fall to the left of 0 in Figure 1 and 2 means that all models are reasonably robust. More work is needed to suggest clear ways of training a model to control its faithfulness or robustness. We believe our perturbation methods can be used to guide model selection by helping to determine just how faithful or robust a model should be based on specific downstream requirements.

Across Models. Although models have different numbers of parameter, we observe in Figure 1 that the models are in general more robust than faithful. The performance of the non-Helsinki models suggests slightly higher NMT performance could be attributed to the greater representational capacity of the model. In Figure 1 we observed the robustness to correlate largely with the NMT performance (β).

Alternate choices for κ. To further understand the role of metric on our results, we explored a few other translation metrics, including BERT-Score and BLEURT. But, we found that these metrics [6] overlook minor errors towards being robust to perturbed sentences. It makes it unclear whether that is the model's tendency or the metric that is improving the robustness. Hence, we found BLEU to be a more stable metric for the study.

[6] Figure 8, 9, 10 and 11 show a comparison between β_1 and β_2 computed with the different metrics.



(a) Helsinki-Opus

(b) M2M-100-418M

(c) mBART50

(d) M2M-100-1.2B

Figure 7: The correlation of $\rho(\beta_1^B, \beta^B)$ as β_1^B and $\rho(\beta_2^B, \beta^B)$ as β_2^B shows that the robustness of the translation system has a strong correlation with the performance of the machine translation system. The faithful translations have a weak correlation, indicating that the easier to translate examples are difficult for the model to do word-to-word translations on.

Unnatural translations. Although rare, examples for which reordering the source results in a better target translation do exist. Similarly to the prediction flips observed by Sinha et al., a fraction of the translations have β_1 scores greater than β[7]. This suggests that the model might require the source sentences to be in a particular order to attain the expected translation. Our work opens up potential avenues for probing datasets for flips as a way to measure "unnaturalness" of models' translation algorithms.

Conclusion. Overall, it is important to understand how NMT systems behave on such malformed input—should a model be robust and risk "hallucinating" an input, or should it be faithful, taking the input at face-value, and provide word-by-word translations. Particular examples might differ in whether a robust or a strongly faithful approach is warranted; for example, we wouldn't want to badly translate poetry that was using nonstandard word order for creative effect. Our novel metrics and perturbation functions allow one to quantify how systems strike a balance between robustness and faithfulness in NMT, both on the corpus level and at the level of particular examples.

Acknowledgements

We thank Dieuwke Hupkes, Douwe Kiela, Robin Jia, Molly FitzMorris, Adi Renduchintala, Mona Diab, Hagen Blix, Saujas Vaduguru and Radhika Govindarajan for the suggestions and discussions that helped shape this work. Further, we thank McGill University and ComputeCanada for the computing resources.

References

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.

Jacob Andreas, Anca Dragan, and Dan Klein. 2017. Translating neuralese. In *Proceedings of the 55th*

[7] Table 3, 4, 5, and 6 in Appendix explain this in detail.

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 232–242, Vancouver, Canada. Association for Computational Linguistics.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 113–120, Boston, Massachusetts, USA. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Arianna Bisazza, Ahmet Üstün, and Stephan Sportel. 2021. On the difficulty of translating free-order case-marking languages. *arXiv preprint arXiv:2107.06055*.

Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020. Advaug: Robust adversarial augmentation for neural machine translation. *arXiv*.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1766, Melbourne, Australia. Association for Computational Linguistics.

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

Noam Chomsky. 1962/2013. 7. the logical basis of linguistic theory. In *Eight decades of general linguistics*, pages 123–236. Brill.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, Michigan. Association for Computational Linguistics.

Alexis Conneau, Germán 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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2126–2136. Association for Computational Linguistics.

Stephen Crain and Mineharu Nakayama. 1987. Structure dependence in grammar formation. *Language*, pages 522–543.

Chris Dyer and Philip Resnik. 2010. Context-free reordering, finite-state translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 858–866, Los Angeles, California. 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.

Marzieh Fadaee and Christof Monz. 2020. The unreasonable volatility of neural machine translation models. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 88–96, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020a. Beyond English-Centric Multilingual Machine Translation. *arxiv*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2020b. Beyond english-centric multilingual machine translation. *arXiv*.

Steven M Frankland and Joshua D Greene. 2015. An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proceedings of the National Academy of Sciences*, 112(37):11732–11737.

Jon Gauthier and Roger Levy. 2019. Linking artificial and human neural representations of language. 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 529–539, Hong Kong, China. Association for Computational Linguistics.

Dmitriy Genzel. 2010. Automatically learning source-side reordering rules for large scale machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 376–384, Beijing, China. Coling 2010 Organizing Committee.

Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. 2020. Type B reflexivization as an unambiguous testbed for multilingual multi-task gender bias. In *Proceedings of the 2020 Conference on Empirical Methods* *in Natural Language Processing (EMNLP)*, pages 2637–2648, Online. Association for Computational Linguistics.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv*.

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. 2021. BERT & Family Eat Word Salad: Experiments with Text Understanding. *arXiv*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 287–297.

Maxim Khalilov, José A. R. Fonollosa, and Mark Dras. 2009. Coupling hierarchical word reordering and decoding in phrase-based statistical machine translation. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation (SSST-3) at NAACL HLT 2009*, pages 78–86, Boulder, Colorado. Association for Computational Linguistics.

Maxim Khalilov and Khalil Sima'an. 2010. Source reordering using MaxEnt classifiers and supertags. In *Proceedings of the 14th Annual conference of the European Association for Machine Translation*, Saint Raphaël, France. European Association for Machine Translation.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Saurabh Kulshreshtha, Jose Luis Redondo Garcia, and Ching-Yun Chang. 2020. Cross-lingual alignment methods for multilingual BERT: A comparative study. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 933–942, Online. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2020. Contextualized perturbation for textual adversarial attack. *arXiv*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.

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

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*.

Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Zhaojiang Lin, and Pascale Fung. 2020c. On the importance of word order information in cross-lingual sequence labeling. *arXiv preprint arXiv:2001.11164*.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domain. In *International Workshop on Spoken Language Translation*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140.

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

Antonio Valerio Miceli-Barone and Giuseppe Attardi. 2013. Pre-reordering for machine translation using transition-based walks on dependency parse trees. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 164–169, Sofia, Bulgaria. Association for Computational Linguistics.

Andrea Moro. 2015. *The boundaries of Babel: The brain and the enigma of impossible languages*. MIT press.

Andrea Moro. 2016. *Impossible languages*. MIT Press.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan. 2020. Evaluating robustness to input perturbations for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8538–8544, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics.

Prasanna Parthasarathi, Joelle Pineau, and Sarath Chandar. 2020. How to evaluate your dialogue system: Probe tasks as an alternative for token-level evaluation metrics. *arXiv*.

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*.

Sai Rajeswar, Sandeep Subramanian, Francis Dutil, Christopher Pal, and Aaron Courville. 2017. Adversarial generation of natural language. In *arXiv*.

Adithya Renduchintala, Rebecca Knowles, Philipp Koehn, and Jason Eisner. 2016. Creating interactive macaronic interfaces for language learning. In *Proceedings of ACL-2016 System Demonstrations*, pages 133–138, Berlin, Germany. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

John Robert Ross. 1967. *Constraints on variables in syntax*. Ph.D. thesis, Massachusetts Institute of Technology.

Keisuke Sakaguchi, Kevin Duh, Matt Post, and Benjamin Van Durme. 2017. Robsut wrod reocginiton via semi-character recurrent neural network. In *AAAI*.

Chinnadhurai Sankar, Sandeep Subramanian, Christopher Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. *arXiv*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *EMNLP-IJCNLP*.

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. *arXiv preprint arXiv:2104.06644*.

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

Joshua Snell and Jonathan Grainger. 2017. The sentence superiority effect revisited. *Cognition*.

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P Parikh. 2020. Sticking to the facts: Confident decoding for faithful data-to-text generation. *arXiv preprint arXiv:1910.08684*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Vaibhav Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. 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 1916–1920, Minneapolis, Minnesota. Association for Computational Linguistics.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv*.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Lijun Wu, Yingce Xia, Fei Tian, Li Zhao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. Adversarial neural machine translation. In *Asian Conference on Machine Learning*. PMLR.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. *arXiv*.

Tianfu Zhang, Heyan Huang, Chong Feng, and Xiaochi Wei. 2020. Similarity-aware neural machine translation: reducing human translator efforts by leveraging high-potential sentences with translation memory. *Neural Computing and Applications*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.

A Packages and Tools

We use Python 3.7, pytorch 1.7.1, transformers 4.2.2 for the experiments. For tokenization and parsing, we use Spacy 3.0.0 (Honnibal et al., 2020)[8] for all the languages.

B Sample statistics

C β_1 vs β_2

Figure 8, Figure 9, and Figure 10 show the comparison of the β_1 and β_2 scores across the different perturbations on the different translation tasks.

D α_e

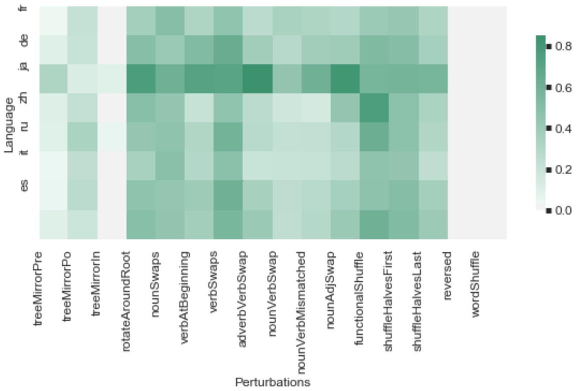


Figure 12: α_e^L

E Measured MT Performances

F $\beta_1 > \beta$

In some corner cases, we observed the β_1 to be greater than β. This suggests that the model, at least in those cases, opts an unnatural understanding of the syntax for the translation.

[8] https://spacy.io/

Perturbations	**de**	**fr**	**ru**	**ja**	**es**	**it**	**zh**
TreeMirrorPre	3869	3732	3201	1580	7004	3009	155
TreeMirrorPost	3862	3726	3199	1525	7001	3009	147
TreeMirrorIn	3862	3726	3199	1525	7001	3009	147
VerbAdvSwaps	944	831	747	1297	1287	615	1649
VerbSwaps	2019	2084	1496	4376	3714	1582	3703
NounAdjSwaps	508	967	631	985	1863	600	469
FuncShuffle	1197	1274	383	7004	2666	666	229
NounVerbSwaps	3777	3624	2821	4798	6664	2687	6746
NounVerbMis	3005	2989	2623	4102	5448	2189	5932
ShuffleLastHalf	3905	4002	3213	4997	7083	3030	7084
VerbAtBeginning	3584	3410	3939	1817	7135	3729	7084
RotateAroundRt	3904	4002	3212	4997	7082	3030	7074
WordShuffle	3905	4002	3213	4997	7083	3030	7084
ShuffleFirstHalf	3905	4002	3213	4997	7083	3030	7084
NounSwaps	3747	2242	2954	4912	5936	1934	6545
Reversed	3904	4002	3212	4997	7082	3030	7074
Total	$5k$	$5k$	$5k$	$5k$	$10k$	$5k$	$10k$

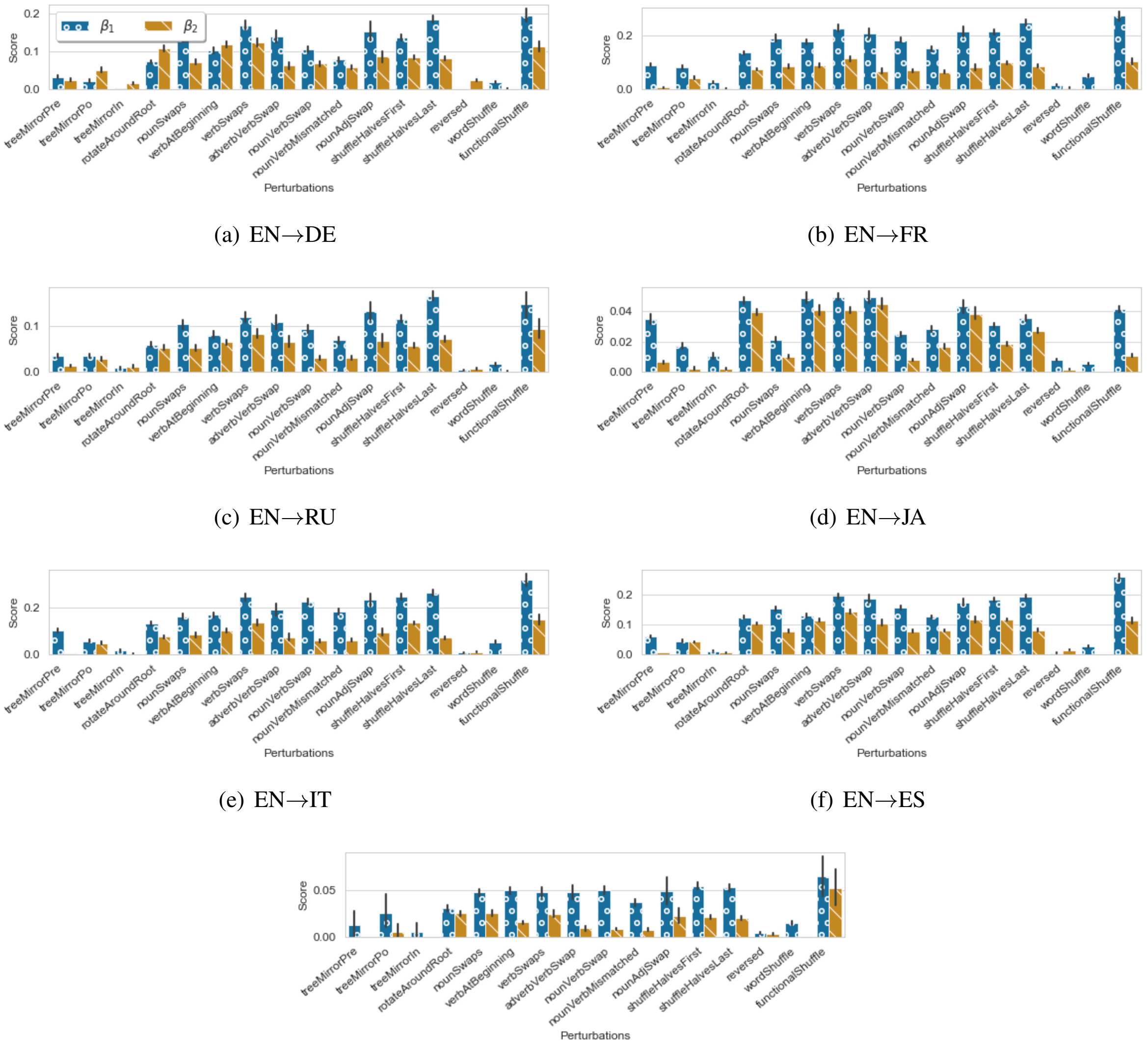
Table 1: The distribution of samples under different perturbation functions across the different languages. The trend shows that there might be some parts-of-speech that are minority – Adjective, Adverb – across the languages. This does not affect the analysis in the paper.

Language	**Helsinki-OPUS**	**mBART**	**M2M_100_418M**	**M2M_100_1.2B**
German	$0.40 \pm 7.77 \times 10^{-6}$	$0.30 \pm 7.10 \times 10^{-6}$	$0.25 \pm 7.96 \times 10^{-6}$	$\mathbf{0.34} \pm 8.80 \times 10^{-6}$
Russian	$0.39 \pm 9.51 \times 10^{-6}$	$0.24 \pm 8.00 \times 10^{-6}$	$0.23 \pm 8.36 \times 10^{-6}$	$\mathbf{0.28} \pm 8.53 \times 10^{-6}$
French	$0.45 \pm 7.66 \times 10^{-6}$	$0.35 \pm 7.15 \times 10^{-6}$	$0.30 \pm 6.89 \times 10^{-6}$	$\mathbf{0.37} \pm 8.33 \times 10^{-6}$
Japanese	$0.69 \pm 4.01 \times 10^{-6}$	$0.07 \pm 1.64 \times 10^{-6}$	$0.07 \pm 1.77 \times 10^{-6}$	$\mathbf{0.10} \pm 2.72 \times 10^{-6}$
Italian	$0.39 \pm 9.74 \times 10^{-6}$	$\mathbf{0.37} \pm 9.67 \times 10^{-6}$	$0.30 \pm 9.93 \times 10^{-6}$	$0.35 \pm 9.52 \times 10^{-6}$
Spanish	$0.47 \pm 8.34 \times 10^{-6}$	$0.30 \pm 7.47 \times 10^{-6}$	$0.34 \pm 7.75 \times 10^{-6}$	$\mathbf{0.39} \pm 9.96 \times 10^{-6}$
Chinese	$0.08 \pm 2.95 \times 10^{-6}$	$0.09 \pm 3.25 \times 10^{-6}$	$0.07 \pm 2.96 \times 10^{-6}$	$\mathbf{0.10} \pm 5.07 \times 10^{-6}$

Table 2: Performances in BLEU-4 (β) of our NMT models. We can see that the models have a poor performance on Japanese and Chinese datasets with an only exception of Helsinki-OPUS model having 0.69 BLEU on Japanese. This could be attributed to the fact that the validation data are from OPUS and the distributions between the train and validation set on Japanese language are too close and unique. This explains the poor performance on Japanese by the other models. Also, we observed the size of the model to affect linearly the performance of the model (comparing models mBART, 418M and 1.2B).

Language	**Helsinki-OPUS**	**mBART**	**M2M_100_418M**	**M2M_100_1.2B**
German	**514**	334	373	399
Russian	**643**	382	388	512
French	**693**	601	516	592
Japanese	**608**	0	5	16
Italian	**914**	644	408	509
Spanish	**575**	558	410	527
Chinese	501	**560**	322	230

Table 3: Number of flips by language and model. We found no relation between the number of flips a model might exhibit when presented with perturbed data to its size or performance in NMT task (β). At this point we think this is just a noise and might have more to do with the dataset than the models themselves.



(a) EN→DE

(b) EN→FR

(c) EN→RU

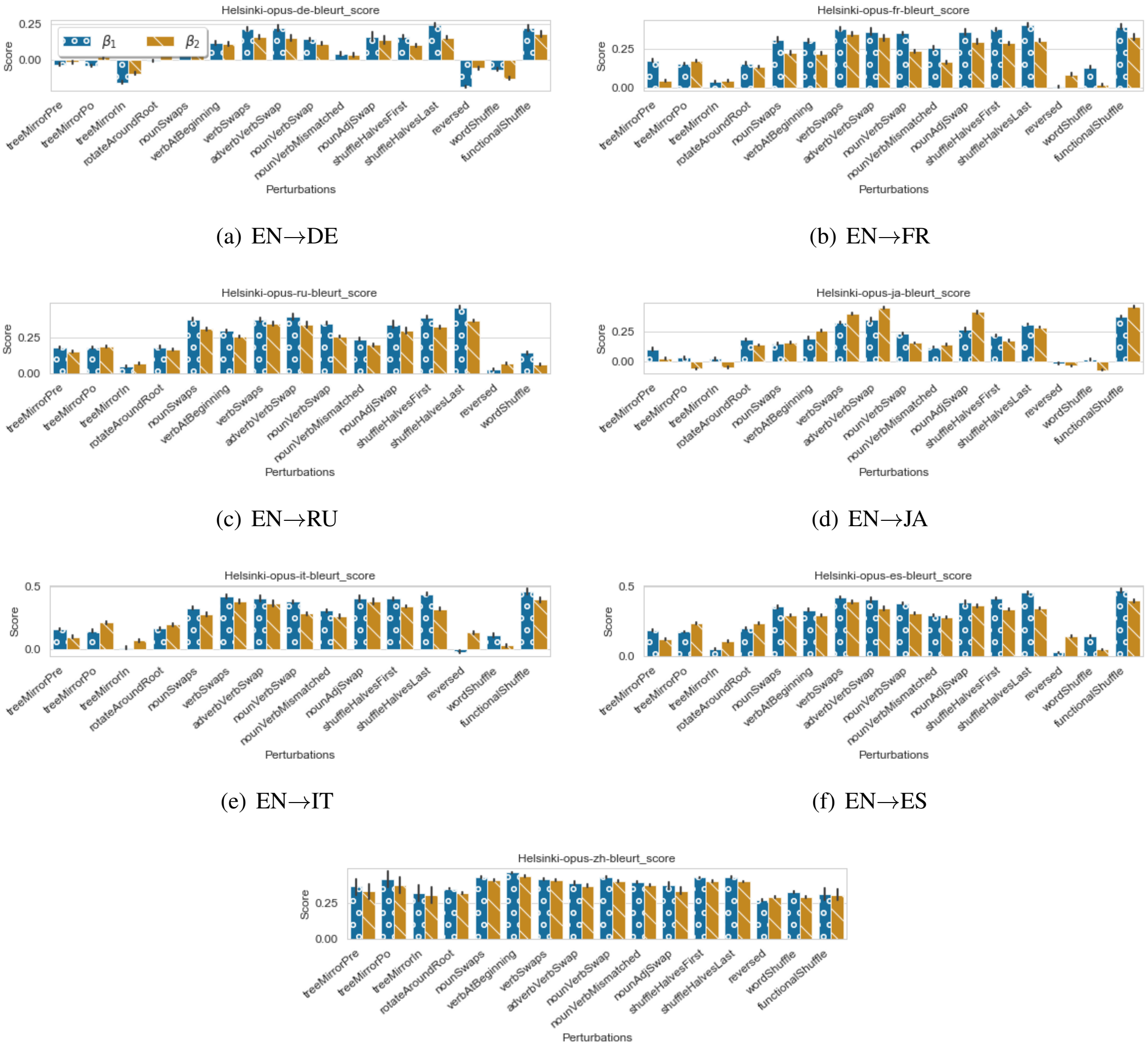
(d) EN→JA

(e) EN→IT

(f) EN→ES

(g) EN→ZH

Figure 8: We present the results only averaged from reactions to perturbations across the 4 models to showcase the trend of β_1 scores being generally higher than β_2 scores across the different perturbations in different languages. The scores computed using BLEU-4 records the differences by better showcasing that harder perturbations having lower β_1 and β_2 scores, while on the other perturbations the models being robust is highlighted well.



(a) EN→DE

(b) EN→FR

(c) EN→RU

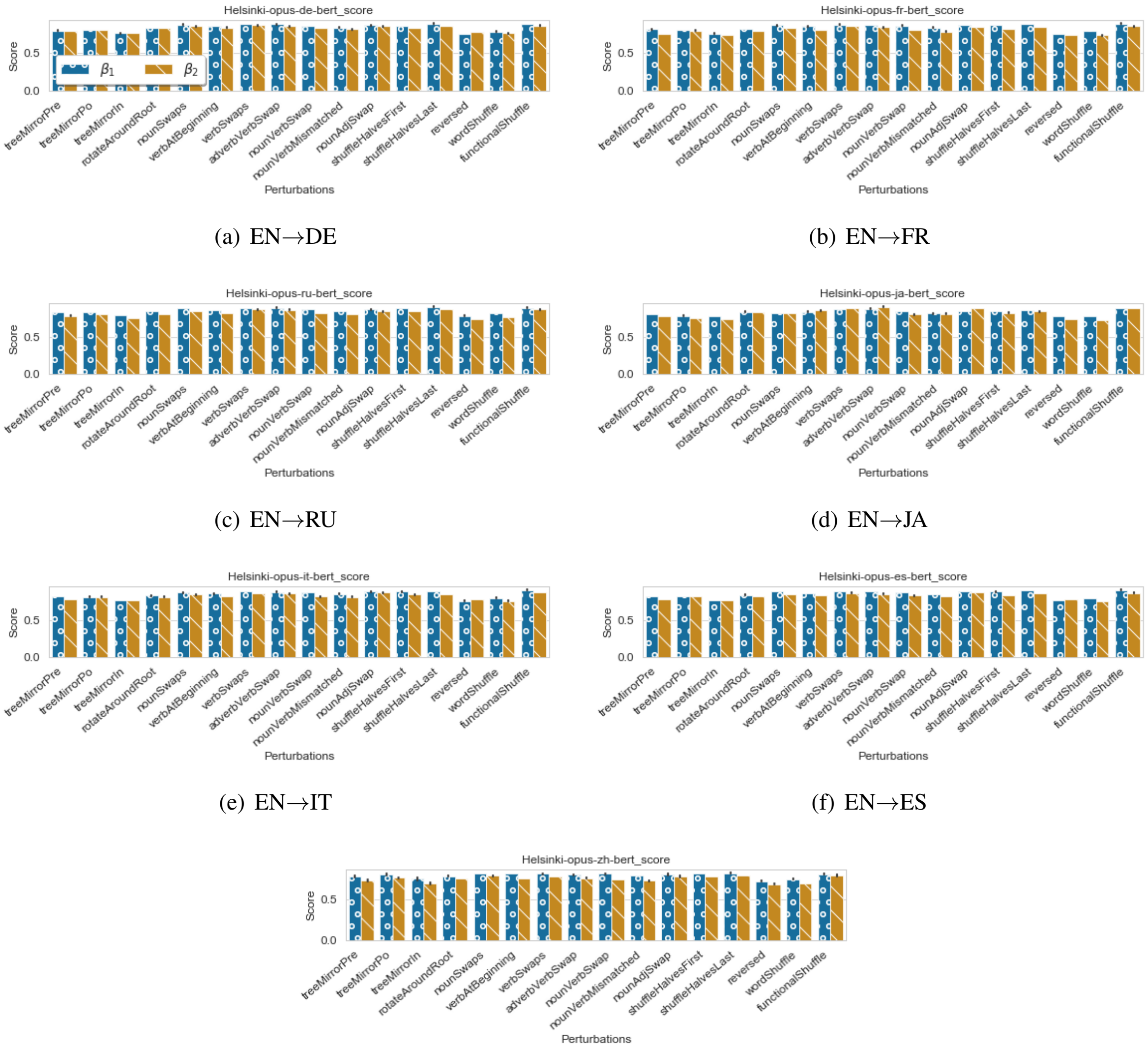
(d) EN→JA

(e) EN→IT

(f) EN→ES

(g) EN→ZH

Figure 9: The BLEURT scores as the choice of κ were mildly forgiving of the perturbations; indicating an intrinsic robustness. Although this did not affect the general trend in most cases as compared to BLEU-4, this was not a suitable metric for measuring faithfulness and robustness of the models.



(a) EN→DE

(b) EN→FR

(c) EN→RU

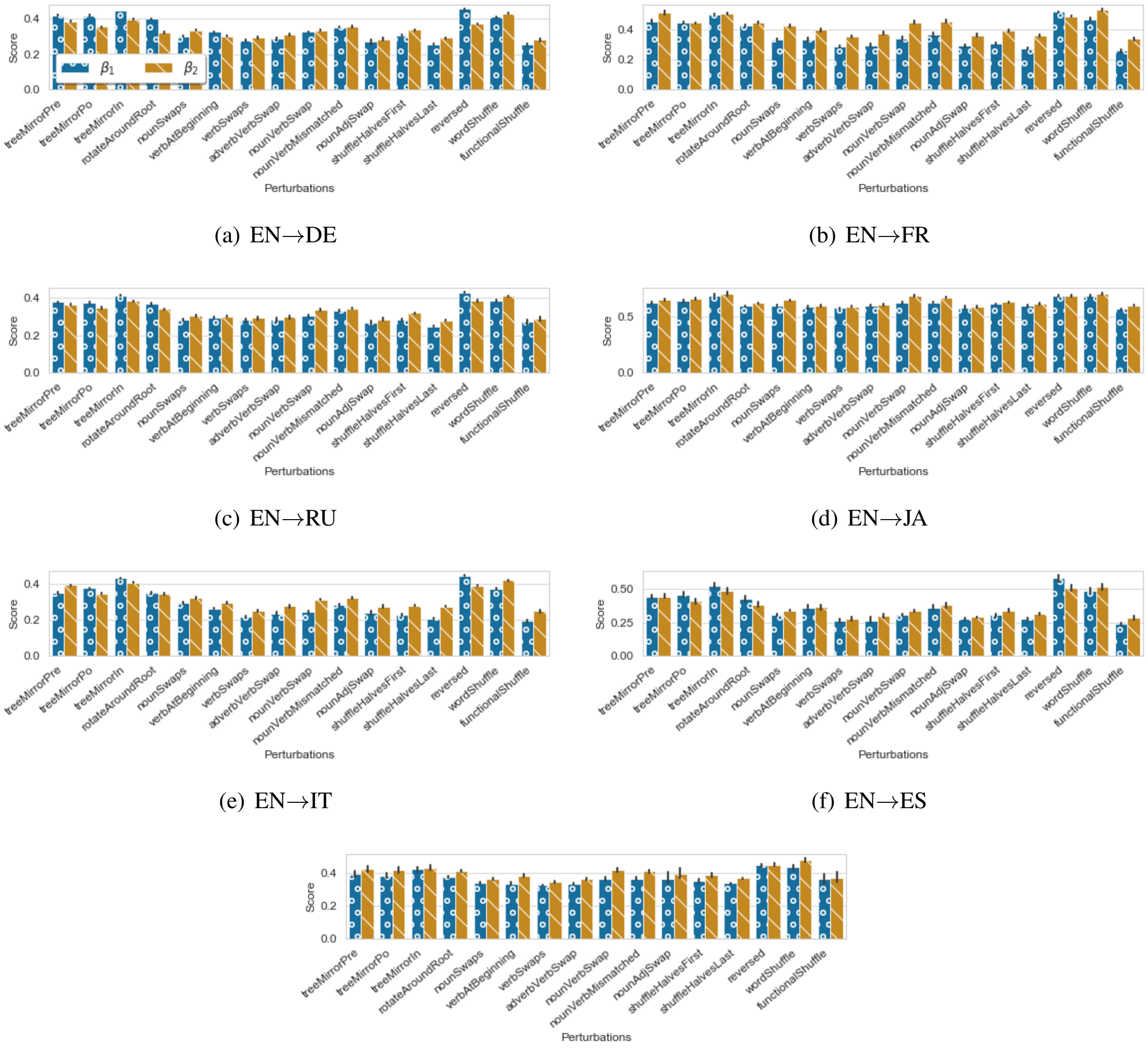
(d) EN→JA

(e) EN→IT

(f) EN→ES

(g) EN→ZH

Figure 10: The BERT-score can be observed to be too forgiving of the perturbations in the text thereby not having any difference to the scores across languages. The sheer lack of discriminating perturbed vs unperturbed makes BERT-score a less suitable candidate for the task.



(a) EN→DE

(b) EN→FR

(c) EN→RU

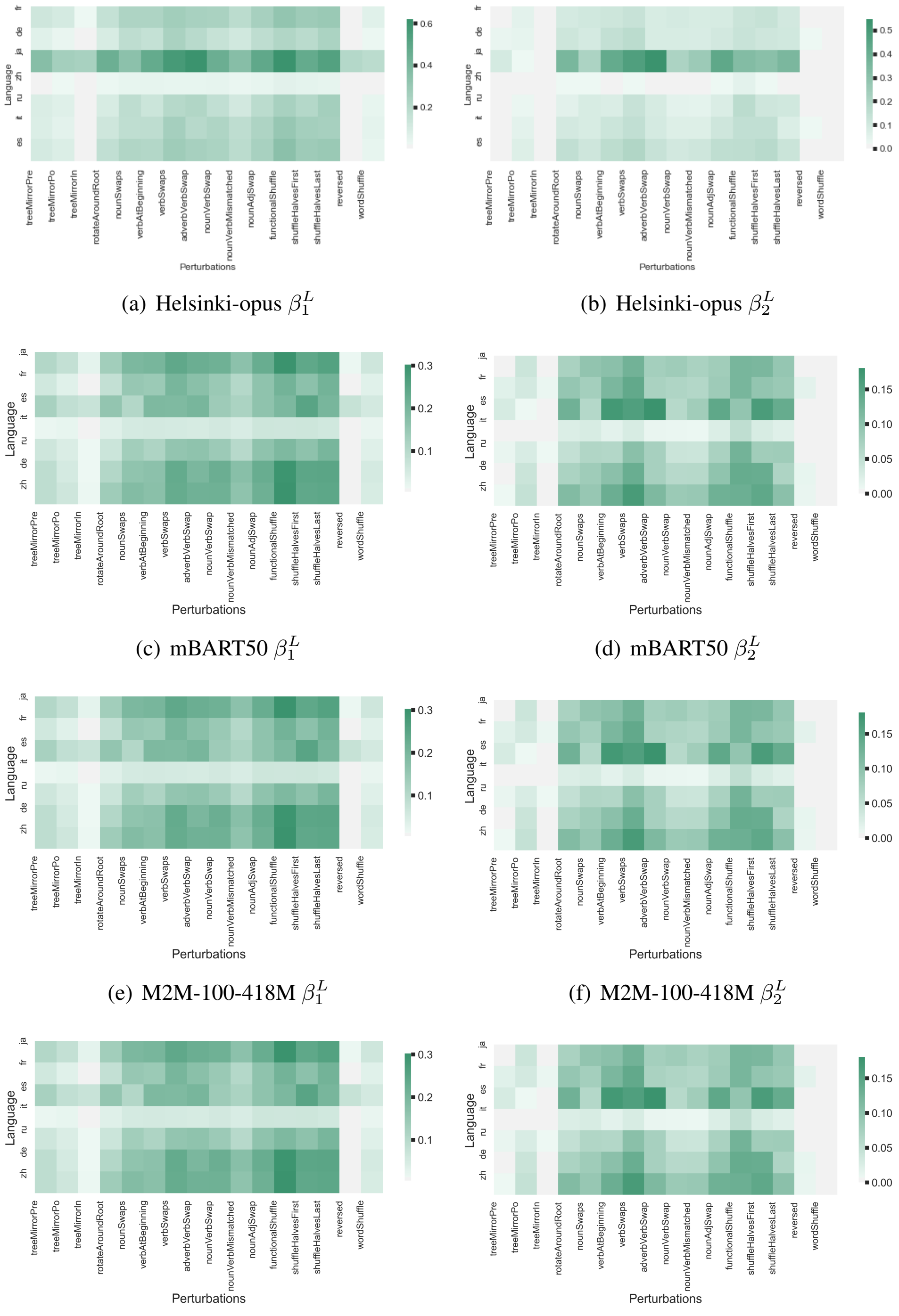
(d) EN→JA

(e) EN→IT

(f) EN→ES

(g) EN→ZH

Figure 11: Levenshtein scores did not provide the sufficient discrimination between β_1 and β_2, making it less suitable for the task.



(a) Helsinki-opus β_1^L

(b) Helsinki-opus β_2^L

(c) mBART50 β_1^L

(d) mBART50 β_2^L

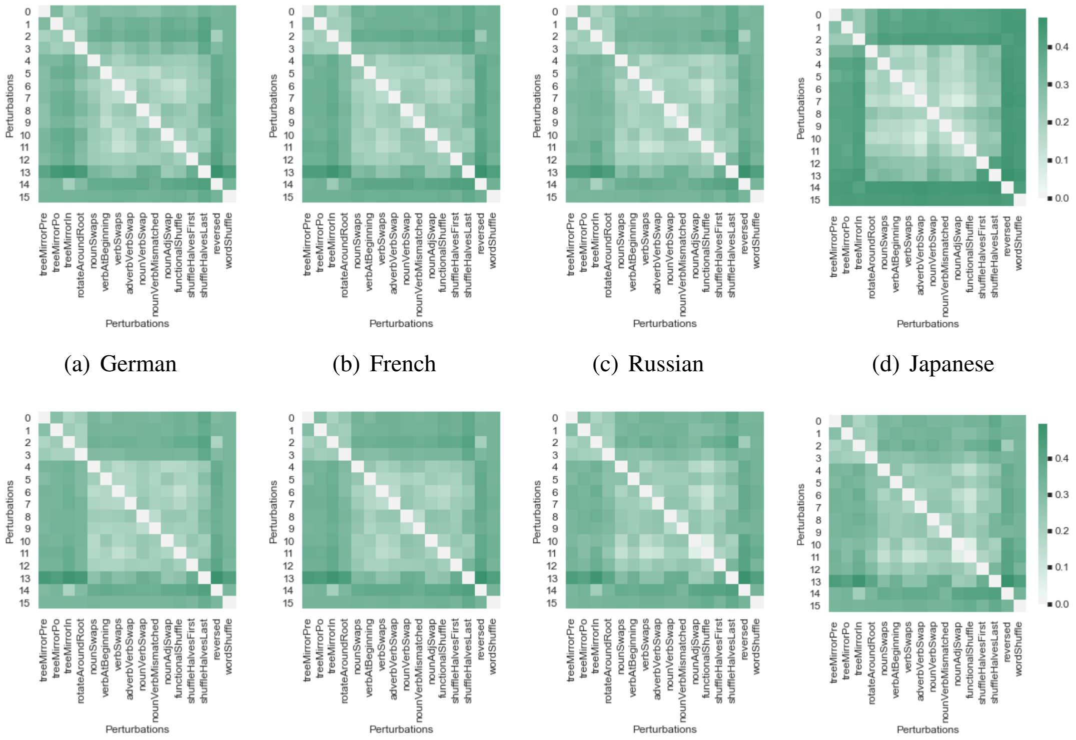
(e) M2M-100-418M β_1^L

(f) M2M-100-418M β_2^L

(g) M2M-100-1.2B β_1^L

(h) M2M-100-1.2B β_2^L

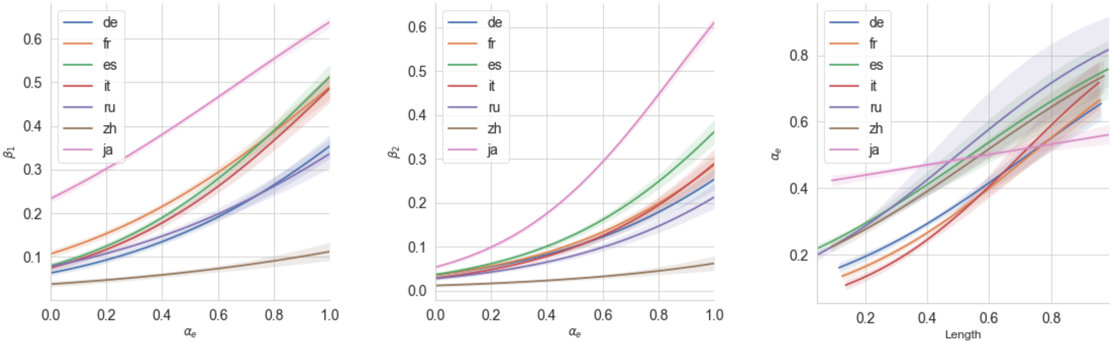
Figure 13: Models ignore precise word order they are presented with: Compare the heat maps showing higher β_1 than β_2 values on average across languages. Models tend to recover more when faced with *PoS tag-based* perturbations: Figure 12 generally shows darker shades for PoS tag-based perturbations than for the others. This means that models find it harder to ignore word order for sentences perturbed with *Dependency tree-based* and *Random* perturbations than with *PoS tag-based* ones.



(a) German (b) French (c) Russian (d) Japanese

(e) Italian (f) Spanish (g) Chinese (h) English

Figure 14: The heatmap illustrates average of Levenshtein distances between different perturbations. The map shows interesting patterns that naturally differentiate the dependency tree based, PoS-based, and random perturbation categories. It is interesting to observe the pattern being consistent across the different languages.



(a) β_1 vs α_e (b) β_2 vs α_e (c) $Length$ vs α_e

Figure 15: Models tend to be more robust and more faithful for easier perturbations (α_e is higher). The longer sentences having higher α_e has more to do with most of our perturbation functions targeting specific sentence constituents, leaving majority of the sentence unperturbed. [$Length$ is normalized with the length of the longest sentence in every language $+1$ to compute a value between $[0, 1)$.]

mBART	g_e	g_e^-	β	β_1	Ψ
de	The problem was too much for me.	was The problem too much for me .	0.00	0.61	nounVerbSwp
	They don't even know why.	do They n't even know why .	0.43	0.56	nounVerbSwp
	Tom took part in the race.	Tom took part race in the.	0.00	1.00	nounVerbMis
fr	That's what makes me nervous.	That makes what 's me nervous .	0.37	0.54	verbSwaps
	If you cannot come, I'll eat alone.	you If not can come , I 'll eat alone .	0.27	0.61	shuffleHFirst
	It's been raining since last night.	It 's been raining since night last .	0.36	0.64	nounAdjSwp
es	Is there a shorter road to get there?	a shorter road there Is to get there ?	0.00	0.53	nounVerbSwap
	Just ignore what Tom said.	Just ignore what said Tom.	0.36	0.59	shuffleHLast
	Do you want to play football with us?	play you Do to want football with us ?	0.50	1.00	verbSwaps
it	What do you think about her?	her about What do you think ?	0.00	0.61	treeMirrorPo
	I've read every page except the last one.	I 've read every page except the one last.	0.36	0.59	nounAdjSwap
	He threw a stone at the dog.	threw He a stone at the dog .	0.42	1.00	wordShuffle
ru	I'll tell him this afternoon.	I 'll tell him this afternoon .	0.00	0.54	nounSwaps
	Have you ever seen a car accident?	a car accident seen Have you ever ?	0.50	1.00	rotateArouRt
	I'm sure that you'll succeed.	succeed 'm I sure that you 'll .	0.43	0.64	verbAtBegin
zh	How heavy is your suitcase?	your suitcase How heavy is ?	0.00	0.76	treeMirrorPo
	That dog runs very fast.	fast very runs dog That .	0.00	0.61	reversed
	Tom is hiding a terrible secret.	hiding is Tom a terrible secret .	0.41	0.54	nounVerbMis

Opus	g_e	g_e^-	β	β_1	Ψ
de	Did you bring a hair dryer?	a hair dryer Did you bring ?	0.00	0.54	treeMirrorPo
	It's a river that has never been explored.	It 's a river that has explored been never.	0.42	0.59	nounVerbSwap
	I may go to Boston next month.	go may I to Boston next month .	0.37	0.52	nounVerbMis
fr	Yes, my name is Karen Smith.	Karen Smith Yes , my name is .	0.00	0.61	treeMirrorPo
	Why didn't you call me last night?	did you n't Why call me last night ?	0.50	1.00	shuffleHFirst
	Our fridge doesn't work anymore.	does Our fridge n't work anymore .	0.00	0.54	nounVerbSwap
es	Have you ever been on TV?	been Have you ever on TV ?	0.34	0.62	verbAtBegin
	I'm looking forward to your coming to Japan.	I coming looking forward to your 'm to Japan .	0.45	0.51	verbSwaps
	We left him some cake.	We some left cake him .	0.0	0.54	wordShuffle
it	Have you tried online dating?	you Have tried online dating ?	0.45	0.76	nounVerbSwap
	What did you do this morning?	What this do you morning did ?	0.00	1.00	wordShuffle
	She was able to read the book.	read She was able to the book .	0.35	0.65	verbAtBegin
ru	Tom knew that I was lonely.	Tom knew that lonely was I .	0.43	0.64	nounAdjSwap
	He said he would come tomorrow.	come he said would He tomorrow .	0.47	1.00	nounVerbMis
	You can stay if I want to.	You can stay if to want I.	0.45	0.54	shuffleHLast
ja	Joseph said to them, "It is like I told you, saying, 'You are spies!'	Joseph saying to them , " It are like I said you , is , ' You told spies ! '	0.48	1.00	verbSwaps
	Don't be overcome by evil, but overcome evil with good.	Do n't overcome overcome by evil , but be evil with good .	0.42	1.00	verbSwaps
	How amiable are thy tabernacles, O LORD of hosts!	hosts of LORD O , tabernacles thy are amiable How !	0.42	0.65	reversed
zh	He has completely lost all sense of duty.	He has lost completely all sense of duty.	0.45	0.54	verbAdvSwap
	We have a white cat.	We have a cat white .	0.35	0.84	nounAdjSwap
	The main question is how does Tom feel.	The main question does how is Tom feel.	0.47	0.61	verbSwaps

Table 4: Samples from across different languages and perturbations where the models translated better when the source sentence was perturbed (a lá Sinha et al. 2020). Although such flips made only a small fraction, we observed the unnaturalness understanding of the syntactic structure in translation task.

M2M-418	g_e	g_e^-	β	β_1	Ψ
de	Do you know who they are?	you Do who know are they ?	0.38	0.54	nounVerbSwp
	You remind me of Tom.	remind me of Tom You .	0.00	0.51	treeMirrorPre
	That architect builds very modern houses.	That architect builds very houses modern .	0.00	0.47	nounVerbMis
fr	Can I get you a cup of tea?	you a cup of tea get Can I ?	0.46	0.53	rotateArndRt
	I use the Internet as a resource for my research.	use I the Internet as a resource for my research .	0.30	0.67	verbAtBegin
	There's a serious problem.	serious There 's a problem .	0.49	0.51	wordShuffle
es	You are not a dog. Are you a cat?	You are not a dog . Are a cat you ?	0.34	0.59	nounSwaps
	The cat jumps on top of the table.	of cat jumps on top The the table .	0.00	0.61	funcShuffle
	Does Tom enjoy watching horror movies?	horror movies Does Tom enjoy watching ?	0.29	0.56	wordShuffle
it	I was very tired last night.	very tired last night I was .	0.00	0.61	treeMirrorPo
	You shouldn't be alone.	You be n't should alone .	0.00	1.00	verbShuffles
	She published two collections of short stories.	She published two collections stories short of.	0.37	0.68	shuffleHLast
ru	They're still not safe.	still not safe 're They .	0.38	0.54	treeMirrorIn
	Let me talk with Tom.	talk Let me with Tom .	0.00	0.76	verbAtBegin
	Go away! I hate you!	away Go ! you I hate !	0.00	0.64	treeMirrorPo
zh	I love music.	love I music .	0.00	1.00	verbAdvSwap
	He paid double fare.	paid He double fare .	0.00	1.00	verbAtBegin
	I doubt that I'm a good writer.	I doubt that a good writer 'm I.	0.43	0.60	nounSwaps

M2M-1.2	g_e	g_e^-	β	β_1	Ψ
de	Tom is not happy to be here.	Tom is not happy here be to.	0.34	1.00	ShuffleHLast
	You should give up smoking.	You give should up smoking .	0.00	1.00	verbSwaps
	I know who you are.	I know you who are .	0.43	1.00	nounSwaps
fr	Tom drowned in the ocean.	drowned in ocean the Tom .	0.00	0.84	treeMirrorPre
	She saw it, too.	saw She it , too .	0.00	0.54	verbAtBegin
	Of course you can stay.	Of course stay can you .	0.43	0.64	nounVerbMis
es	Being able to use a computer is advantageous.	Being able to a computer use is advantageous .	0.38	0.64	nounVerbMis
	He never forgets to pay a bill.	He never bill forgets to pay a .	0.00	0.54	wordShuffle
	Never betray the trust of your friends.	betray trust of friends your the Never .	0.0	0.54	treeMirrorPre
it	Tom isn't a member of our club.	n't a member of our club is Tom .	0.35	0.56	rotateArndRt
	I think she's 40 years old.	think I 's she 40 years old .	0.37	0.68	nounVerbSwp
	There's enough food for all of you.	There 's enough food for of all you .	0.37	0.59	funcShuffle
ru	I saw Tom this morning.	Tom I saw this morning .	0.00	1.00	shuffleHFirst
	He said he would come tomorrow.	said come tomorrow he would He .	0.47	1.00	treeMirrorPre
	I will be busy next week.	week next busy be will I .	0.00	0.64	reversed
zh	You remind me of Tom.	me remind Tom of You .	0.00	0.51	nounSwaps
	That dog runs very fast.	That dog runs fast very.	0.38	0.81	shuffleHLast
	This photo was taken in Nara.	taken was This photo in Nara .	0.47	0.61	nounVerbMis

Table 5: Samples from across different languages and perturbations where the models translated better when the source sentence was perturbed. Although such flips made only a small fraction, we observed the unnaturalness in the understanding of the syntactic structure in translation task. This is similar to the observations made by Sinha et al. (2020).

Perturbations	**de**	**fr**	**ru**	**ja**	**es**	**it**	**zh**
treeMirrorPre	82	127	**138**	142	9	123	3
treeMirrorPo	43	76	**107**	93	4	84	4
treeMirrorIn	15	9	28	25	3	**34**	1
rotateAroundRoot	92	172	**186**	150	76	121	162
nounSwaps	101	113	98	106	40	109	**142**
verbAtBeginning	241	180	277	**351**	17	274	234
verbSwaps	109	87	**140**	122	83	75	76
adverbVerbSwap	62	34	72	**76**	25	58	52
nounVerbSwap	163	273	232	**321**	58	220	198
nounVerbMismatched	95	152	179	**200**	28	105	145
nounAdjSwap	24	74	74	**76**	14	39	14
functionalShuffle	51	**90**	75	51	58	26	6
shuffleHalvesFirst	186	316	303	**348**	74	255	210
shuffleHalvesLast	286	279	**354**	301	71	300	219
reversed	4	10	16	9	12	27	**38**
wordShuffle	63	73	**114**	101	22	71	106

Perturbations	**Helsinki-Opus**	**mBART50**	**M2M-100-418M**	**M2M-100-1.2B**
treeMirrorPre	**196**	121	144	163
treeMirrorPo	**158**	92	72	89
treeMirrorIn	**34**	26	27	28
rotateAroundRoot	**356**	245	167	191
nounSwaps	**268**	198	119	124
verbAtBeginning	**522**	396	326	330
verbSwaps	**282**	174	111	125
adverbVerbSwap	**143**	94	61	81
nounVerbSwap	**529**	364	293	279
nounVerbMismatched	**304**	263	152	185
nounAdjSwap	**102**	77	61	75
functionalShuffle	**159**	73	32	93
shuffleHalvesFirst	**615**	423	329	325
shuffleHalvesLast	**525**	393	372	520
reversed	**32**	30	25	29
wordShuffle	**180**	102	126	142

Table 6: The distribution count of flips by every perturbation functions across the languages and models show that Helsinki-Opus recording the highest flips. While the trend is similar across the models the maximum flips in Opus models could be attributed to the experiments being done on the validation set of the datasets the Opus models were trained on. Although it is not clear whether the specific overlap between the train and dev sets cause the flips and we leave that to the future work. Among the languages,*ru* and *ja* accounted for majority of the flips and we hypothesize that it could be some artifact of the dataset that causes it more than the model itself.