It's Morphin' Time!
Combating Linguistic Discrimination with Inflectional Perturbations

Samson Tan[§♮], **Shafiq Joty**[‡§], **Min-Yen Kan**[♮], **Richard Socher**[§]
[§]Salesforce Research
[♮]National University of Singapore
[‡]Nanyang Technological University
[§]{samson.tan,sjoty,rsocher}@salesforce.com
[♮]kanmy@comp.nus.edu.sg

Abstract

Training on only perfect Standard English corpora predisposes pre-trained neural networks to discriminate against minorities from non-standard linguistic backgrounds (e.g., African American Vernacular English, Colloquial Singapore English, etc.). We perturb the inflectional morphology of words to craft plausible and semantically similar adversarial examples that expose these biases in popular NLP models, e.g., BERT and Transformer, and show that adversarially fine-tuning them for a single epoch significantly improves robustness without sacrificing performance on clean data.[1]

1 Introduction

In recent years, Natural Language Processing (NLP) systems have gotten increasingly better at learning complex patterns in language by pre-training large language models like BERT, GPT-2, and CTRL (Devlin et al., 2019; Radford et al., 2019; Keskar et al., 2019), and fine-tuning them on task-specific data to achieve state of the art results has become a norm. However, deep learning models are only as good as the data they are trained on.

Existing work on societal bias in NLP primarily focuses on attributes like race and gender (Bolukbasi et al., 2016; May et al., 2019). In contrast, we investigate a uniquely NLP attribute that has been largely ignored: linguistic background.

Current NLP models seem to be trained with the implicit assumption that everyone speaks fluent (often U.S.) Standard English, even though two-thirds (>700 million) of the English speakers in the world speak it as a second language (L2) (Eberhard et al., 2019). Even among native speakers, a significant number speak a dialect like African American Vernacular English (AAVE) rather than Standard English (Crystal, 2003). In addition, these

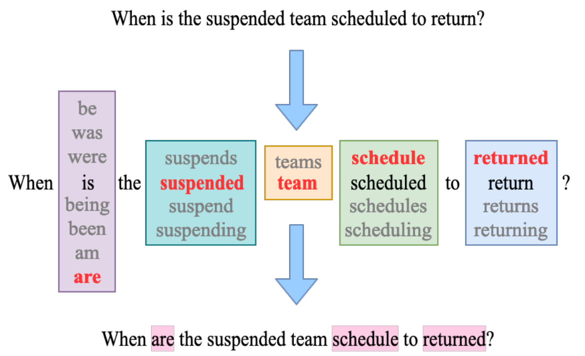


Figure 1: MORPHEUS looks at each noun, verb, or adjective in the sentence and selects the inflected form (marked in **red**) that maximizes the target model's loss. To maximize semantic preservation, MORPHEUS only considers inflections belonging to the same universal part of speech as the original word.

World Englishes exhibit variation at multiple levels of linguistic analysis (Kachru et al., 2009).

Therefore, putting these models directly into production without addressing this inherent bias puts them at risk of committing linguistic discrimination by performing poorly for many speech communities (e.g., AAVE and L2 speakers). This could take the form of either failing to understand these speakers (Rickford and King, 2016; Tatman, 2017), or misinterpreting them. For example, the recent mistranslation of a minority speaker's social media post resulted in his wrongful arrest (Hern, 2017).

Since L2 (and many L1 dialect) speakers often exhibit variability in their production of inflectional morphology[2] (Lardiere, 1998; Prévost and White, 2000; Haznedar, 2002; White, 2003; Seymour, 2004), we argue that NLP models should be robust to inflectional perturbations in order to minimize their chances of propagating linguistic discrimination. Hence, in this paper, we:

[1]Code and adversarially fine-tuned models available at https://github.com/salesforce/morpheus.

[2]Inflections convey tense, quantity, etc. See Appendix A for dialectal examples.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2920–2935
July 5 - 10, 2020.

- Propose MORPHEUS, a method for generating plausible and semantically similar adversaries by perturbing the inflections in the clean examples (Figure 1). In contrast to recent work on adversarial examples in NLP (Belinkov and Bisk, 2018; Ebrahimi et al., 2018; Ribeiro et al., 2018), we exploit morphology to craft our adversaries.
- Demonstrate its effectiveness on multiple machine comprehension and translation models, including BERT and Transformer (Tables 1 & 2).
- Show that adversarially fine-tuning the model on an adversarial training set generated via weighted random sampling is sufficient for it to acquire significant robustness, while preserving performance on clean examples (Table 5).

To the best of our knowledge, we are the first to investigate the robustness of NLP models to inflectional perturbations and its ethical implications.

2 Related Work

Fairness in NLP. It is crucial that NLP systems do not amplify and entrench social biases (Hovy and Spruit, 2016). Recent research on fairness has primarily focused on racial and gender biases within distributed word representations (Bolukbasi et al., 2016), coreference resolution (Rudinger et al., 2018), sentence encoders (May et al., 2019), and language models (Bordia and Bowman, 2019). However, we posit that there exists a significant potential for *linguistic* bias that has yet to be investigated, which is the motivation for our work.

Adversarial attacks in NLP. First discovered in computer vision by Szegedy et al. (2014), adversarial examples are data points crafted with the intent of causing a model to output a wrong prediction. In NLP, this could take place at the character, morphological, lexical, syntactic, or semantic level.

Jia and Liang (2017) showed that question answering models could be misled into choosing a distractor sentence in the passage that was created by replacing key entities in the correct answer sentence. Belinkov and Bisk (2018) followed by demonstrating the brittleness of neural machine translation systems against character-level perturbations like randomly swapping/replacing characters. However, these attacks are not optimized on the target models, unlike Ebrahimi et al. (2018), which makes use of the target model's gradient to find the character change that maximizes the model's error.

Since these attacks tend to disrupt the sentence's semantics, Ribeiro et al. (2018) and Michel et al. (2019) propose searching for adversaries that preserve semantic content. Alzantot et al. (2018) and Jin et al. (2019) explore the use of synonym substitution to create adversarial examples, using word embeddings to find the n nearest words. Eger et al. (2019) take a different approach, arguing that adding visual noise to characters leaves their semantic content undisturbed. Iyyer et al. (2018) propose to create paraphrase adversaries by conditioning their generation on a syntactic template, while Zhang et al. (2019b) swap key entities in the sentences. Zhang et al. (2019a) provide a comprehensive survey of this topic.

Adversarial training. In order to ensure our NLP systems are not left vulnerable to powerful attacks, most existing work make use of adversarial training to improve the model's robustness (Goodfellow et al., 2015). This involves augmenting the training data either by adding the adversaries to or replacing the clean examples in the training set.

Summary. Existing work in fairness mostly focus on tackling bias against protected attributes like race and gender, while those in adversarial NLP primarily investigate character- and word-level perturbations and seek to improve the models' robustness by retraining them from scratch on the adversarial training set. Our work makes use of perturbations in *inflectional morphology* to highlight the linguistic bias present in models such as BERT and Transformer, before showing that simply *fine-tuning* the models for *one* epoch on the adversarial training set is sufficient to achieve significant robustness while maintaining performance on clean data.

3 Generating Inflectional Perturbations

Inflectional perturbations inherently preserve the general semantics of a word since the root remains unchanged. In cases where a word's part of speech (POS) is context-dependent (e.g., *duck* as a verb or a noun), restricting perturbations to the original POS further preserves its original meaning.

Additionally, since second language speakers are prone to inflectional errors (Haznedar, 2002; White, 2003), adversarial examples that perturb the inflectional morphology of a sentence should be *less perceivable* to people who interact heavily with non-native speakers or are themselves non-native speakers. Hence, we present MORPHEUS, our proposed method for crafting inflectional adversaries.

	Extractive Question Answering
Original	When is the suspended team scheduled to return?
Adversary	When are the suspended team schedule to returned?
Prediction	**Before:** 2018 **After:** No answer
Original	Who upon arriving gave the original viking settlers a common identity?
Adversary	Who upon arrive give the original viking settler a common identities?
Prediction	**Before:** Rollo **After:** almost no foreign settlers
	Neural Machine Translation
Original	Israeli warplanes struck a target inside the Syrian port city of Latakia Thursday night, a senior administration official confirms to Fox News.
Adversary	Israeli warplanes strikes a target inside the Syrian port city of Latakia Thursday night, a senior administration official confirms to Foxes News.
Prediction	**Before:** Un haut responsable de l'administration confirme à Fox News que des avions de combat israéliens ont frappé une cible à l'intérieur de la ville portuaire syrienne de Lattaquié dans la nuit de jeudi. **After:** Le président de la République, Nicolas Sarkozy, a annoncé jeudi que le président de la République, Nicolas Sarkozy, s'est rendu en République démocratique du Congo.

Table 1: Adversarial examples found for BERT, SpanBERT, and Transformer-big. While not perfectly grammatical, it is plausible for English dialect and second language (L2) speakers to produce such sentences.
(Top) Models trained on SQuAD 2.0 are more fragile than those trained on SQuAD 1.1, and have a bias towards predicting "no answer". Examples are answerable questions and therefore present in both SQuAD 1.1 and 2.0.
(Bottom) Perturbing two inflections caused Transformer-big to output a completely irrelevant sentence. In addition, adversarial examples for ~1.4% of the test set caused the model to output the source (English) sentences.

3.1 MORPHEUS: A Greedy Approach

Problem formulation. Given a target model f and an original input example x for which the ground truth label is y, our goal is to generate the adversarial example x' that maximizes f's loss. Formally, we aim to solve the following problem:

$$x' = \arg\max_{x_c} \mathcal{L}(y, f(x_c)) \qquad (1)$$

where x_c is an adversarial example generated by perturbing x, $f(x)$ is the model's prediction, and $\mathcal{L}(\cdot)$ is the model's loss function. In this setting, f is a neural model for solving a specific NLP task.

Proposed solution. To solve this problem, we propose MORPHEUS (Algorithm 1), an approach that greedily searches for the inflectional form of each *noun, verb,* or *adjective* in x that maximally increases f's loss (Eq. 1). For each token in x, MORPHEUS calls MAXINFLECTED to find the inflected form that caused the greatest increase in f's loss.[3] Table 1 presents some adversarial examples obtained by running MORPHEUS on state-of-the-art machine reading comprehension and translation models: namely, BERT (Devlin et al., 2019), SpanBERT (Joshi et al., 2019), and Transformer-big (Vaswani et al., 2017; Ott et al., 2018).

[3]A task-specific evaluation metric may be used instead of the loss in situations where it is unavailable. However, as we discuss later, the choice of metric is important for optimal performance and should be chosen wisely.

Algorithm 1 MORPHEUS

Require: Original instance x, Label y, Model f
Ensure: Adversarial example x'

```
T ← TOKENIZE(x)
for all i = 1, ..., |T| do
    if POS(T_i) ∈ {NOUN, VERB, ADJ} then
        I ← GETINFLECTIONS(T_i)
        T_i ← MAXINFLECTED(I, T, y, f)
    end if
end for
x' ← DETOKENIZE(T)
return x'
```

There are two possible approaches to implementing MAXINFLECTED: one is to modify each token independently from the others in *parallel*, and the other is to do it *sequentially* such that the increase in loss is accumulated as we iterate over the tokens. A major advantage of the parallel approach is that it is theoretically possible to speed it up by t times, where t is the number of tokens which are nouns, verbs, or adjectives. However, since current state-of-the-art models rely heavily on contextual representations, the sequential approach is likely to be more effective in finding combinations of inflectional perturbations that cause major increases in loss. We found this to be the case in our preliminary experiments (see Table 6 in Appendix D).

Assumptions. MORPHEUS treats the target model as a *black box* and maximally requires only access to the model's logits to compute the loss. As mentioned, task-specific metrics may be used instead of the loss as long as the surface is not overly "flat", like in a step function. Examples of inappropriate metrics are the exact match and F_1 scores for extractive question answering, which tend to be 1 for most candidates but drop drastically for specific ones. This may affect MORPHEUS' ability to find an adversary that induces absolute model failure.

While the black box assumption has the advantage of not requiring access to the target model's gradients and parameters, a limitation is that we need to query the model for each candidate inflection's impact on the loss, as opposed to Ebrahimi et al. (2018)'s approach. However, this is not an issue for inflectional perturbations since each word *usually* has less than 5 possible inflections.

Candidate generation. We make use of `lemminflect`[4] to generate candidate inflectional forms in the GETINFLECTIONS method, a simple process in which the token is first lemmatized before being inflected. In our implementation of GETINFLECTIONS, we also allow the user to specify if the candidates should be constrained to the same universal part of speech.

Semantic preservation. MORPHEUS constrains its search to inflections belonging to the same *universal* part of speech. For example, take the word "duck". Depending on the context, it may either be a verb or a noun. In the context of the sentence "There's a jumping duck", "duck" is a noun andMORPHEUS may only choose alternate inflections associated with nouns.

This has a higher probability of preserving the sentence's semantics compared to most other approaches, like character/word shuffling or synonym swapping, since the root word and its position in the sentence remains unchanged.

Early termination. MORPHEUS selects an inflection if it increases the loss. In order to avoid unnecessary searching, it terminates once it finds an adversarial example that induces model failure. In our case, we define this as a score of 0 on the task's evaluation metric (the higher, the better).

Other implementation details. In order to increase overall inflectional variation in the set of adversarial examples, GETINFLECTIONS shuffles the generated list of inflections before returning it (see Figure 4 in Appendix). Doing this has no effect on MORPHEUS' ability to induce misclassification, but prevents overfitting during adversarial fine-tuning, which we discuss later in Section 6. Additionally, since MORPHEUS greedily perturbs each eligible token in x, it may get stuck in a local maximum for some x values. To mitigate this, we run it again on the reversed version of x if the early termination criterion was not fulfilled during the forward pass.

Finally, we use `sacremoses`[5] for tokenization and NLTK (Bird et al., 2009) for POS tagging.

4 Experiments

NLP tasks. To evaluate the effectiveness of MORPHEUS at inducing model failure in NLP models, we test it on two popular NLP tasks: question answering (QA) and machine translation (MT). QA involves language understanding (classification), while MT also involves language generation. Both are widely used by consumers of diverse linguistic backgrounds and hence have a high chance of propagating discrimination.

Baseline. In the below experiments, we include a *random baseline* that randomly inflects each eligible word in each original example.

Measures. In addition to the raw scores, we also report the *relative decrease* for easier comparison across models since they perform differently on the clean dataset. Relative decrease (d_r) is calculated using the following formula:

$$d_r = \frac{\text{score}_{\text{original}} - \text{score}_{\text{adversarial}}}{\text{score}_{\text{original}}} \quad (2)$$

4.1 Extractive Question Answering

Given a question and a passage containing spans corresponding to the correct answer, the model is expected to predict the span corresponding to the answer. Performance for this task is computed using *exact match* or *average* F_1 (Rajpurkar et al., 2016). We evaluate the effectiveness of our attack using average F_1, which is more forgiving (for the target model). From our experiments, the exact match score is usually between 3-9 points lower than the average F_1 score.

SQuAD 1.1 and 2.0. The Stanford Question Answering Dataset (SQuAD) comprises over 100,000 question–answer pairs written by crowdworkers

[4] https://github.com/bjascob/LemmInflect

[5] https://github.com/alvations/sacremoses

Dataset	Model	Clean	Random	MORPHEUS
SQuAD 2.0 Answerable Questions (F_1)	GloVe-BiDAF	78.67	74.00 (−5.93%)	**53.94 (−31.43%)**
	ELMo-BiDAF	80.90	76.81 (−5.05%)	62.17 (−23.15%)
	$\text{BERT}_{\text{SQuAD 1.1}}$	93.14	90.90 (−2.40%)	82.79 (−11.11%)
	$\text{SpanBERT}_{\text{SQuAD 1.1}}$	91.88	91.61 (−0.29%)	82.86 (−9.81%)
	$\text{BERT}_{\text{SQuAD 2}}$	81.19	74.13 (−8.69%)	**57.47 (−29.21%)**
	$\text{SpanBERT}_{\text{SQuAD 2}}$	88.52	84.88 (−4.11%)	69.47 (−21.52%)
SQuAD 2.0 All Questions (F_1)	$\text{BERT}_{\text{SQuAD 2}}$	81.52	78.87 (−3.25%)	**67.24 (−17.51%)**
	$\text{SpanBERT}_{\text{SQuAD 2}}$	87.71	85.46 (−2.56%)	73.26 (−16.47%)
newstest2014 En-Fr (BLEU)	ConvS2S	40.83	27.72 (−32.10%)	**17.31 (−57.60%)**
	Transformer-big	43.16	30.41 (−29.54%)	20.57 (−56.25%)

Table 2: Results for MORPHEUS on QA and NMT models. The subscript in $\text{Model}_{\text{dataset}}$ indicates the dataset used to fine-tune the model. Negated % decrease w.r.t. the scores on clean data are reported in parentheses for easy comparison across models. Bolded values indicate the largest % decrease.

based on Wikipedia articles. SQuAD 1.1 guarantees that the passages contain valid answers to the questions posed (Rajpurkar et al., 2016). SQuAD 2.0 increases the task's difficulty by including another 50,000 unanswerable questions, and models are expected to identify when a passage does not contain an answer for the given question (Rajpurkar et al., 2018). Since the test set is not public, we generate adversarial examples from and evaluate the models on the standard dev set.

In addition, the answerable questions from SQuAD 2.0 are used in place of SQuAD 1.1 to evaluate models trained on SQuAD 1.1. This allows for easy comparison between the performance of the SQuAD 1.1-fine-tuned models and SQuAD 2.0-fine-tuned ones for answerable questions. We found performance on the answerable questions from SQuAD 2.0 to be comparable to SQuAD 1.1.

Models. We evaluate MORPHEUS on Gardner et al. (2018)'s implementation of BiDAF (Seo et al., 2017), a common baseline model for SQuAD 1.1, ELMo-BiDAF (Peters et al., 2018), the `transformers` implementation (Wolf et al., 2019) of BERT, and SpanBERT, a pre-training method focusing on span prediction that outperforms BERT on multiple extractive QA datasets.

4.2 Results and Discussion

From Table 2, we see that models based on contextual embeddings (e.g., ELMo and BERT variants) tend to be more robust than those using fixed word embeddings (GloVe-BiDAF). This difference is likely due to the pre-training process, which gives them greater exposure to a wider variety of contexts in which different inflections occur. Removing the POS constraint further degrades the models' performance by another 10% of the original score, however, this difference is likely due to changes in the semantics and expected output of the examples.

BiDAF vs. BERT. Even after accounting for the performance difference on clean data, the BiDAF variants are significantly less robust to inflectional adversaries compared to the BERT variants. This is likely a result of BERT's greater representational power and masked language modeling pre-training procedure. Randomly masking out words during pre-training could have improved the models' robustness to small, local perturbations (like ours).

BERT vs. SpanBERT. In the context of question answering, SpanBERT appears to be slightly more robust than vanilla BERT when comparing overall performance on the two SQuAD datasets. However, the difference becomes significant if we look only at the SQuAD 2.0-fine-tuned models' performance on answerable questions (7% difference). This indicates that BERT has a stronger bias towards predicting "no answer" when it encounters inflectional perturbations compared to SpanBERT.

SQuAD 1.1 vs. SQuAD 2.0. The ability to "know what you don't know" (Rajpurkar et al., 2018) appears to have been obtained at a great cost. The SQuAD 2.0-fine-tuned models are not only generally less robust to inflectional errors than their SQuAD 1.1 equivalents (6.5% difference), but also significantly less adept at handling answerable questions (12–18% difference). This discrepancy suggests a stronger bias in SQuAD 2.0 models towards predicting "no answer" upon receiving sentences containing inflectional errors (see Table 1).

As we alluded to earlier, this is particularly troubling: since SQuAD 2.0 presents a more realistic

SQuAD 2.0 Answerable Questions (F_1)			
Original	Transfer	Clean	MORPHEUS
GloVe-BiDAF	$\text{BERT}_{\text{SQuAD 1.1}}$	93.14	89.67
	$\text{SpanBERT}_{\text{SQuAD 1.1}}$	91.88	90.75
	$\text{BERT}_{\text{SQuAD 2}}$	81.19	72.21
	$\text{SpanBERT}_{\text{SQuAD 2}}$	88.52	81.95
$\text{BERT}_{\text{SQuAD 1.1}}$	GloVe-BiDAF	78.67	71.33
	$\text{SpanBERT}_{\text{SQuAD 1.1}}$	91.88	88.68
	$\text{BERT}_{\text{SQuAD 2}}$	81.19	69.68
	$\text{SpanBERT}_{\text{SQuAD 2}}$	88.52	80.11
$\text{SpanBERT}_{\text{SQuAD 1.1}}$	GloVe-BiDAF	78.67	71.41
	$\text{BERT}_{\text{SQuAD 1.1}}$	93.14	87.48
	$\text{BERT}_{\text{SQuAD 2}}$	81.19	70.05
	$\text{SpanBERT}_{\text{SQuAD 2}}$	88.52	77.89
SQuAD 2.0 All Questions (F_1)			
Original	Transfer	Clean	MORPHEUS
$\text{BERT}_{\text{SQuAD 2}}$	$\text{SpanBERT}_{\text{SQuAD 2}}$	87.71	82.49
$\text{SpanBERT}_{\text{SQuAD 2}}$	$\text{BERT}_{\text{SQuAD 2}}$	81.52	75.54

Table 3: Transferability of our adversarial examples.

scenario than SQuAD 1.1, it is fair to conclude that such models will inadvertently discriminate against L2 speakers if put into production as is.

Transferability. Next, we investigate the transferability of adversarial examples found by MORPHEUS across different QA models and present some notable results in Table 3. The adversarial examples found for GloVe-BiDAF transfer to a limited extent to other models trained on SQuAD 1.1, however, they have a much greater impact on $\text{BERT}_{\text{SQuAD 2}}$ and $\text{SpanBERT}_{\text{SQuAD 2}}$ (3–4x more).

We observe a similar pattern for adversarial examples found for $\text{SpanBERT}_{\text{SQuAD 1.1}}$. Of the two, BERT is more brittle in general: the $\text{SpanBERT}_{\text{SQuAD 1.1}}$ adversaries have a greater effect on $\text{BERT}_{\text{SQuAD 2}}$'s performance on answerable questions than on $\text{SpanBERT}_{\text{SQuAD 2}}$'s.

Discussion. One possible explanation for the SQuAD 2.0 models' increased fragility is the difference in the tasks they were trained for: SQuAD 1.1 models expect all questions to be answerable and only need to contend with finding the right span, while SQuAD 2.0 models have the added burden of predicting whether a question is answerable.

Therefore, in SQuAD 1.1 models, the feature space corresponding to a possible answer ends where the space corresponding to another possible answer begins, and there is room to accommodate slight variations in the input (i.e., larger individual spaces). We believe that in SQuAD 2.0 models, the need to accommodate the `unanswerable` prediction forces the spaces corresponding to the possible answers to shrink, with `unanswerable` spaces potentially filling the gaps between them. For SQuAD 2.0 models, this increases the probability of an adversarial example "landing" in the space corresponding to the `unanswerable` prediction. This would explain the effectiveness of adversarial fine-tuning in Section 6, which intuitively creates a "buffer" zone and expands the decision boundaries around each clean example.

The diminished effectiveness of the transferred adversaries at inducing model failure is likely due to each model learning slightly different segmentations of the answer space. As a result, different small, local perturbations have different effects on each model. We leave the in-depth investigation of the above phenomena to future work.

4.3 Machine Translation

We now demonstrate MORPHEUS' ability to craft adversaries for NMT models as well, this time *without* access to the models' logits. The WMT'14 English-French test set (newstest2014), containing 3,003 sentence pairs, is used for both evaluation and generating adversarial examples. We evaluate our attack on the `fairseq` implementation of both the Convolutional Seq2Seq (Gehring et al., 2017) and Transformer-big models, and report the BLEU score (Papineni et al., 2002) using `fairseq`'s implementation (Ott et al., 2019).

From our experiments (Table 2), ConvS2S and Transformer-big appear to be extremely brittle even to inflectional perturbations constrained to the same part of speech (56–57% decrease). In addition, some adversarial examples caused the models to regenerate the input verbatim instead of a translation: 1.4% of the test set for Transformer-big, 3% for ConvS2S (see Table 9 in the Appendix for some examples). This is likely due to the joint source/target byte–pair encoding (Sennrich et al., 2016) used by both NMT systems to tackle rare word translation.

We experimented with both BLEU and chrF (Popović, 2015) as our optimizing criterion[6] and achieved comparable results for both, however, MORPHEUS found more adversarial examples that caused the model to output random sentences about Nicolas Sarkozy when optimizing for chrF.

5 Human Evaluation

To test our hypothesis that inflectional perturbations are likely to be relatively natural and semantics preserving, we randomly sample 130 adversar-

[6]We use the `sacrebleu` implementation (Post, 2018).

	Native U.S. English Speakers		Unrestricted	
Plausibility				
	SQuAD 2.0	newstest2014	SQuAD 2.0	newstest2014
Native	11.58%	25.64%	22.82%	32.56%
L2 Speaker	**42.82%**	**42.30%**	**53.58%**	**52.82%**
Beginner	31.79%	23.33%	17.17%	10.25%
Non-human	13.84%	8.71%	6.41%	4.35%
Semantic Equivalence				
	SQuAD 2.0	newstest2014	SQuAD 2.0	newstest2014
Highly Likely	**52.82%**	**62.30%**	33.84%	**40.76%**
Likely	20.51%	18.71%	**36.15%**	33.84%
Somewhat Likely	11.02%	7.94%	22.82%	19.48%
Somewhat Unlikely	6.92%	6.15%	5.38%	4.35%
Unlikely	3.58%	3.07%	1.53%	1.28%
Highly Unlikely	5.12%	1.79%	0.25%	0.25%

Table 4: Human judgements for adversarial examples that caused a significant degradation in performance.

ial examples[7] from each dataset and ask 3 Amazon Mechanical Turk workers to indicate (1) whether the sentences could have been written by a native speaker, L2 speaker, beginner learner[8], or no human; and (2) the likelihood of the original and adversarial examples sharing the same meaning. To ensure the quality of our results, only Turkers who completed >10,000 HITs with a ≥99% acceptance rate could access our task. For comparison, we also report ratings by native U.S. English speakers, who were selected via a demographic survey and fluency test adapted from Hartshorne et al. (2018). Workers were paid a rate of at least \$12/hr.[9]

Table 4 shows that Turkers from our unrestricted sample judged ~95% of our adversaries to be plausibly written by a human and 92% generally likely to be semantically equivalent to the original examples 92% of the time, hence validating our hypothesis. Qualitative analysis revealed that "is/are"→"am/been" changes accounted for 48% of the implausible adversaries.

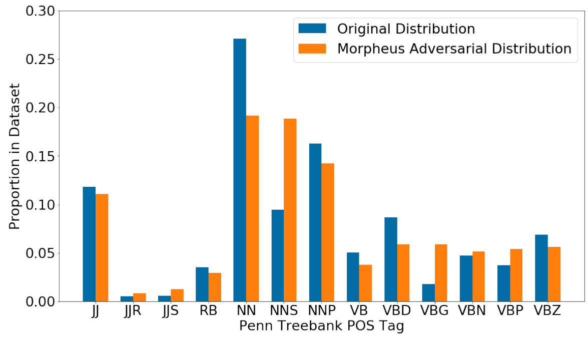
Discussion. We believe that non-native speakers may tend to rate sentences as more human-like for the following reasons:

- Their exposure to another language as a native speaker leads them to accept sentences that mimic errors made by L2 English speakers who share their first language.
- Their exposure to the existence of these above-mentioned errors may lead them to be more forgiving of other inflectional errors that are uncommon to them; they may deem these errors as plausibly made by an L2 speaker who speaks a different first language from them.
- They do not presume mastery of English, and hence may choose to give the higher score when deciding between 2 choices.

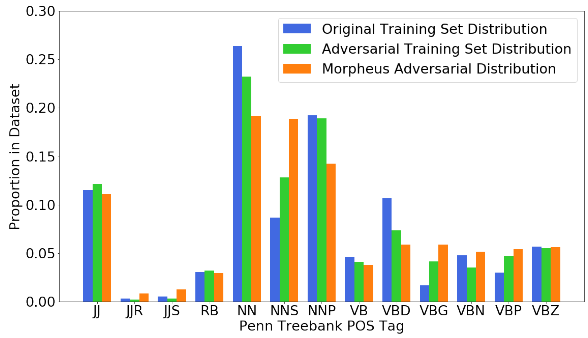
[7]Only adversarial examples that degraded the F_1 score by >50 and the BLEU score by >15 were considered.

[8]We define a beginner as one who has just started learning the language, and an L2 speaker to be an experienced speaker.

[9]Each task was estimated to take 20-25s to be comfortably completed, but they were routinely completed in under 20s.



(a) SQuAD 2.0 dev set



(b) SQuAD 2.0 training set

Figure 2: Comparison of inflectional distributions for $\text{SpanBERT}_{\text{SQuAD 2}}$. The adversarial distributions include only examples that degrade model performance. To make the best use of limited space, we omit the `RBR`, `RBS`, and `NNPS` tags since they do not vary much across distributions. Full figures in Appendix D.

6 Adversarial Fine-tuning

In this section, we extend the standard adversarial training paradigm (Goodfellow et al., 2015) to make the models robust to inflectional perturbations. Since directly running MORPHEUS on the entire training dataset to generate adversaries would be far too time-consuming, we use the findings from our experiments on the respective dev/test sets (Section 4) to create representative samples of good adversaries. This significantly improves robustness to inflectional perturbations while maintaining similar performance on the clean data.

We first present an analysis of the inflectional distributions before elaborating on our method for generating the adversarial training set.

SpanBERT$_{\text{SQuAD 2}}$ (F_1)						
	Original			Adversarially Fine-tuned		
Dataset	Clean	MORPHEUS	Epoch	Clean	MORPHEUS$_{\text{orig}}$	MORPHEUS$_{\text{adv}}$
SQuAD 2.0 Ans	88.52	69.47 (−21.52%)	1	**86.80**	85.17 (−1.87%)	82.76 (−4.65%)
			4	86.15	**84.93 (−1.41%)**	**82.92 (−3.74%)**
SQuAD 2.0 All	87.71	73.26 (−16.47%)	1	86.00	84.72 (−1.48%)	82.41 (−4.17%)
			4	**87.08**	**85.93 (−1.32%)**	**84.71 (−2.72%)**
Transformer-big (BLEU)						
	Original			Adversarially Fine-tuned		
Dataset	Clean	MORPHEUS	Epoch	Clean	MORPHEUS$_{\text{orig}}$	MORPHEUS$_{\text{adv}}$
newstest2014	43.16	20.57 (−56.25%)	1	39.84	**31.79 (−20.20%)**	**31.43 (−21.10%)**
			4	**40.60**	31.99 (−21.20%)	30.82 (−24.08%)

Table 5: Results from adversarially fine-tuning SpanBERT$_{\text{SQuAD 2}}$ and Transformer-big. MORPHEUS$_{\text{orig}}$ refers to the initial adversarial examples, while MORPHEUS$_{\text{adv}}$ refers to the new adversarial examples obtained by running MORPHEUS on the robust model. Relevant results from Table 2 reproduced here for ease of comparison.

6.1 Distributional Analysis

Figure 2a illustrates the overall distributional differences in inflection occurrence between the original and adversarial examples found by MORPHEUS for SQuAD 2.0. Note that these distributions are computed based on the Penn Treebank (PTB) POS tags, which are finer-grained than the universal POS (UPOS) tags used to constrain MORPHEUS' search (Section 4). For example, a UPOS `VERB` may be actually be a PTB `VBD`, `VBZ`, `VBG`, etc.

We can see obvious differences between the global inflectional distributions of the original datasets and the adversaries found by MORPHEUS. The differences are particularly significant for the `NN`, `NNS`, and `VBG` categories. `NNS` and `VBG` also happen to be uncommon in the original distribution. Therefore, we conjecture that the models failed (Section 4) because MORPHEUS is able to find the contexts in the training data where these inflections are uncommon.

6.2 Adversarial Training Set Generation

Since there is an obvious distributional difference between the original and adversarial examples, we hypothesize that bringing the training set's inflectional distribution closer to that of the adversarial examples will improve the models' robustness.

To create the adversarial training set, we first isolate all the adversarial examples (from the dev/test set) that caused any decrease in F_1/BLEU score and count the number of times each inflection is used in this adversarial dataset, giving us the inflectional distribution in Figure 2a.

Next, we randomly select an inflection for each eligible token in each *training example*, weighting the selection with this inflectional distribution instead of a uniform one. To avoid introducing unnecessary noise into our training data, only inflections from the same UPOS as the original word are chosen. We do this 4 times per training example, resulting in an adversarial training set with a clean–adversarial ratio of $1:4$. This can be done in linear time and is *highly scalable*. Algorithm 2 in Appendix C details our approach and Figure 2b depicts the training set's inflectional distribution before and after this procedure.

Fine-tuning vs. retraining. Existing adversarial training approaches have shown that retraining the model on the augmented training set improves robustness (Belinkov and Bisk, 2018; Eger et al., 2019; Jin et al., 2019). However, this requires substantial compute resources. We show that fine-tuning the pre-trained model for just a *single* epoch is sufficient to achieve significant robustness to inflectional perturbations yet still maintain good performance on the clean evaluation set (Table 5).

6.3 Experiments

SpanBERT. Following Joshi et al. (2019), we fine-tune SpanBERT$_{\text{SQuAD 2}}$ for another 4 epochs on our adversarial training set. Table 5 shows the effectiveness of our approach for SpanBERT$_{\text{SQuAD 2}}$.

After just a single epoch of fine-tuning, SpanBERT$_{\text{SQuAD 2}}$ becomes robust to most of the initial adversarial examples with a $<$ 2-point drop in performance on the clean dev set. More importantly, running MORPHEUS on the robust model fails to significantly degrade its performance.

After 4 epochs, the performance on the clean SQuAD 2.0 dev set is almost equivalent to the original SpanBERT$_{\text{SQuAD 2}}$'s, however this comes at a slight cost: the performance on the answerable questions is slightly lower than before. In fact, if performance on answerable questions is paramount, our results show that fine-tuning on the adversarial training set for 1 epoch would be a better (and more cost effective) decision. Retraining SpanBERT adversarially did not result in better performance.

We also found that weighting the random sampling with the adversarial distribution helped to improve the robust model's performance on the answerable questions (refer to Table 7 in Appendix).

Transformer-big. Similarly, model robustness improves dramatically (56.25% to 20.20% decrease) after fine-tuning for 1 epoch on the adversarial training set with a ~3 BLEU point drop in clean data performance (Table 5). Fine-tuning for a further 3 epochs reduced the difference but made the model less robust to new adversarial examples.

We also experimented with using randomly sampled subsets but found that utilizing the entire original training set was necessary for preserving performance on the clean data (see Table 8 in Appendix).

6.4 Discussion

Our anonymous reviewers brought up the possibility of using grammatical error correction (GEC) systems as a defense against inflectional adversaries. Although we agree that adding a GEC model before the actual NLU/translation model would likely help, this would not only require an extra model—often another Transformer (Bryant et al., 2019)—and its training data to be maintained, but would also double the resource usage of the combined system at inference time.

Consequently, institutions with limited resources may choose to sacrifice the experience of minority users rather than incur the extra maintenance costs. Adversarial fine-tuning only requires the NLU/translation model to be fine-tuned once and consumes no extra resources at inference time.

7 Limitations and Future Work

Although we have established our methods' effectiveness at both inducing model failure and robustifying said models, we believe they could be further improved by addressing the following limitations:

1. MORPHEUS finds the distribution of examples that are adversarial for the target model, rather than that of *real* L2 speaker errors, which produced some unrealistic adversarial examples.
2. Our method of adversarial fine-tuning is analogous to curing the symptom rather than addressing the root cause since it would have to be performed for each domain-specific dataset the model is trained on.

In future work, we intend to address these limitations by directly modeling the L2 and dialectal distributions and investigating the possibility of robustifying these models further upstream.

8 Conclusion

Ensuring that NLP technologies are inclusive, in the sense of working for users with diverse linguistic backgrounds (e.g., speakers of World Englishes such as AAVE, as well as L2 speakers), is especially important since natural language user interfaces are becoming increasingly ubiquitous.

We take a step in this direction by revealing the existence of linguistic bias in current English NLP models—e.g., BERT and Transformer—through the use of inflectional adversaries, before using adversarial fine-tuning to significantly reduce it. To find these adversarial examples, we propose MORPHEUS, which crafts plausible and semantically similar adversaries by perturbing an example's inflectional morphology in a constrained fashion, without needing access to the model's gradients. Next, we demonstrate the adversaries' effectiveness using QA and MT, two tasks with direct and wide-ranging applications, before validating their plausibility and semantic content with real humans.

Finally, we show that, instead of retraining the model, fine-tuning it on a representative adversarial training set for a single epoch is sufficient to achieve significant robustness to inflectional adversaries while preserving performance on the clean dataset. We also present a method of generating this adversarial training set in linear time by making use of the adversarial examples' inflectional distribution to perform weighted random sampling.

Acknowledgments

We would like to express our gratitude to Lav Varshney, Jason Wu, Akhilesh Gotmare, and our anonymous reviewers for their insightful feedback on our paper, and friends who participated in our pilot studies. Samson is supported by Salesforce and the Singapore Economic Development Board under its Industrial Postgraduate Programme.

References

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. 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*, Vancouver, BC, Canada.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

David Crystal. 2003. *English as a Global Language*. Cambridge University Press.

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.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2019. *Ethnologue: Languages of the World*, 22 edition. SIL International.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Steffen Eger, Gözde Gül Sahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding nlp systems. 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 1634–1647, Minneapolis, Minnesota. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia. PMLR.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations*, San Diego, California.

J. Hartshorne, B. Tenenbaum, J., and S. Pinker. 2018. A critical period for second language acquisition: Evidence from 2/3 million english speakers. *Cognition*, 177:263–277.

Belma Haznedar. 2002. Missing surface inflection in adult and child l2 acquisition. In *Proceedings of the 6th Generative Approaches to Second Language Acquisition Conference*, pages 140–149, Somerville, Massachusetts. Cascadilla Proceedings Project.

Alex Hern. 2017. Facebook translates 'good morning' into 'attack them', leading to arrest. *The Guardian*.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is bert really robust? natural language attack on text classification and entailment. *arXiv e-prints*, arXiv:1907.11932.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. SpanBERT: Improving pre-training by representing and predicting spans. *arXiv e-prints*, arXiv:1907.10529.

Braj B. Kachru, Yamuna Kachru, and Cecil Nelson, editors. 2009. *The Handbook of World Englishes*. Wiley-Blackwell.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv e-prints*, arXiv:1909.05858.

Donna Lardiere. 1998. Case and tense in the 'fossilized' steady state. *Second Language Research*, 14(1):1–26.

Jacob RE Leimgruber. 2009. *Modelling variation in Singapore English*. Ph.D. thesis, Oxford University.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Michel, Xian Li, Graham Neubig, and Juan Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3103–3114, Minneapolis, Minnesota. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Philippe Prévost and Lydia White. 2000. Missing surface inflection or impairment in second language acquisition? evidence from tense and agreement. *Second Language Research*, 16:103–133.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

John Rickford and Sharese King. 2016. Language and linguistics on trial: Hearing rachel jeantel (and other vernacular speakers) in the courtroom and beyond. *Language*, 92:948–988.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the Annual Meeting of the North American Association of Computational Linguistics (NAACL)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations*, Toulon, France.

Harry Seymour. 2004. The challenge of language assessment for african american english-speaking children: A historical perspective. *Seminars in Speech and Language*, 25:3–12.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations*, Banff, AB, Canada.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Lydia White. 2003. Fossilization in steady state l2 grammars: Persistent problems with inflectional morphology. *Bilingualism: Language and Cognition*, 6:129 – 141.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv e-prints*, arXiv:1910.03771.

Walt Wolfram. 2004. The grammar of urban African American Vernacular English. *Handbook of varieties of English*, 2:111–32.

Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. 2019a. Adversarial attacks on deep learning models in natural language processing: A survey. *arXiv e-prints*, arXiv:1901.06796.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019b. PAWS: paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

A Examples of Inflectional Variation in English Dialects

African American Vernacular English (Wolfram, 2004)

- They seen it.
- They run there yesterday.
- The folks was there.

Colloquial Singapore English (Singlish) (Leimgruber, 2009)

- He want to see how we talk.
- It cover up everything in the floss. It's not nice. It look very cheap.
- I want to shopping only.

B More Details on Human Evaluation

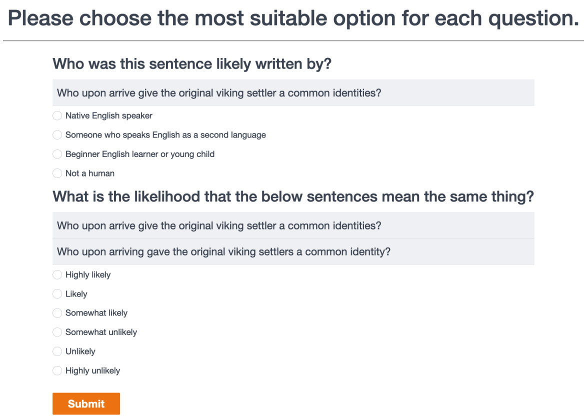


Figure 3: Amazon Mechanical Turk UI.

Figure 3 contains a screenshot of the UI we present to crowd workers. We intentionally prime Turkers by asking if the sentence could be written by an L2 speaker instead of directly asking for acceptability/naturalness ratings in order to ensure that they consider these possibilities.

We also do not use the Semantic Textual Similarity evaluation scheme (Agirre et al., 2013); during preliminary pilot studies, we discovered that annotators interpreted certain words in the scheme (e.g., "information", "details", and "topics") considerably differently, introducing substantial noise into an already subjective judgement task.

Possible limitations. It is possible that seeing the original sentence could affect the worker's judgment of the perturbed sentence's plausibility. However, we argue that this is not necessarily negative since seeing the original sentence would make it easier to spot perturbations that are just outright wrong (i.e., a human will not make that error regardless of their level of fluency).

C Adversarial Training Set Generation

Algorithm 2 RandomInflect

Require: Original instance x, hyperparameter k
  Adversarial distribution $\mathcal{D}_{adv}$
Ensure: Adversarial training dataset X'_x for x
  $X'_x \leftarrow \{x\}$
  for $i = 1$ to k **do**
    $T \leftarrow \text{TOKENIZE}(x)$
    for all $i = 1, \ldots, |T|$ **do**
      if $\text{POS}(T_i) \in \{\text{NOUN, VERB, ADJ}\}$ **then**
        $I \leftarrow \text{GETINFLECTIONS}(T_i)$
        $T_i \leftarrow \text{RANDOMWEIGHTED}(I, D_{adv})$
      end if
    end for
    $x' \leftarrow \text{DETOKENIZE}(T)$
    $X'_x \leftarrow X'_x \cup \{x'\}$
  end for
  return X'_x

D Tables and Figures

SpanBERT$_{\text{SQuAD 2}}$ (F$_1$)			
Dataset	Clean	Morpheus$_{\text{seq}}$	Morpheus$_{\text{parallel}}$
SQuAD 2.0 Ans	88.52	**69.47 (-21.52%)**	74.38 (-15.97%)
SQuAD 2.0 All	87.71	**73.26 (-16.47%)**	76.64 (-12.62%)
Transformer-big (BLEU)			
Dataset	Clean	Morpheus$_{\text{seq}}$	Morpheus$_{\text{parallel}}$
newstest2014	43.16	**20.57 (-56.25%)**	20.85 (-51.69%)

Table 6: Results of the parallel and sequential approaches to implementing MORPHEUS on SpanBERT$_{\text{SQuAD 2}}$ and Transformer-big.

SpanBERT$_{\text{SQuAD 2}}$ (F$_1$)			
Weighted	Dataset	Clean	Morpheus$_{\text{orig}}$
Yes	SQuAD 2.0 Ans	86.80	85.17 (-1.87%)
	SQuAD 2.0 All	86.00	84.72 (-1.48%)
No	SQuAD 2.0 Ans	84.52	83.15 (-1.62%)
	SQuAD 2.0 All	87.12	86.03 (-1.25%)

Table 7: Comparison of results from using weighted vs. uniform random sampling to the create adversarial training set for fine-tuning SpanBERT$_{\text{SQuAD 2}}$

Transformer-big (BLEU)			
Subset	Original	Clean	Morpheus$_{\text{orig}}$
$\frac{1}{20}$	43.16	30.90	24.95
$\frac{1}{4}$	43.16	36.59	29.46
Full	43.16	40.60	31.99

Table 8: Results from adversarially fine-tuning Tranformer-big on different subsets of the original training set.

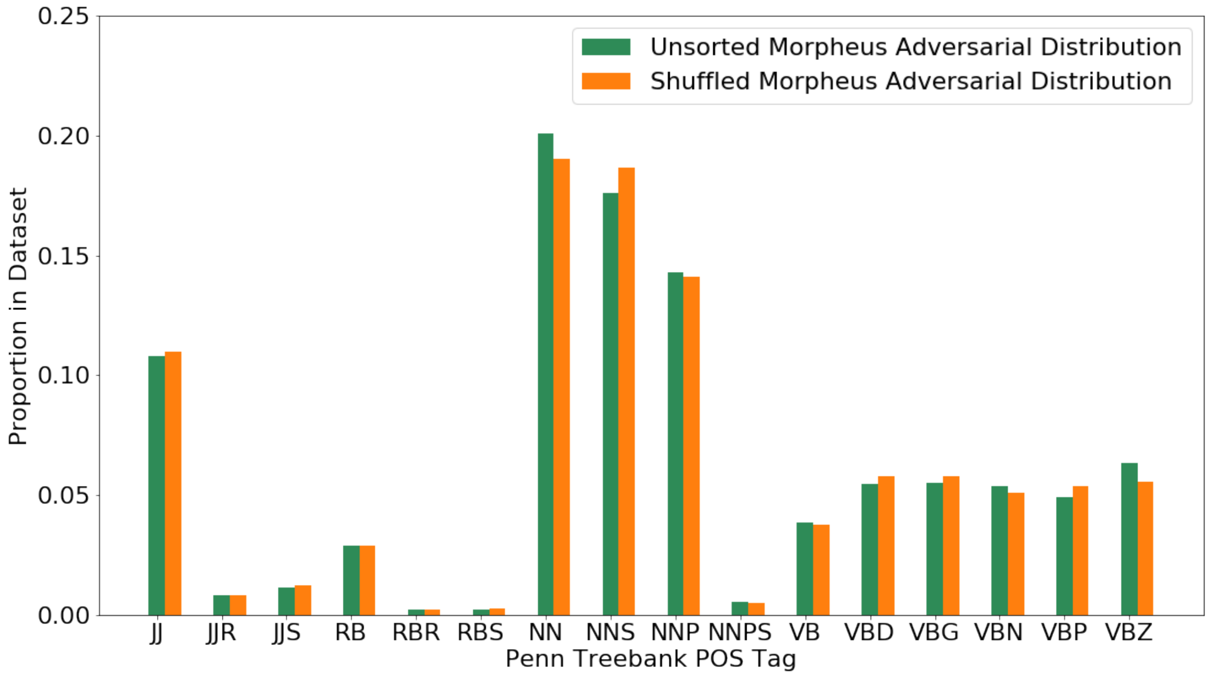
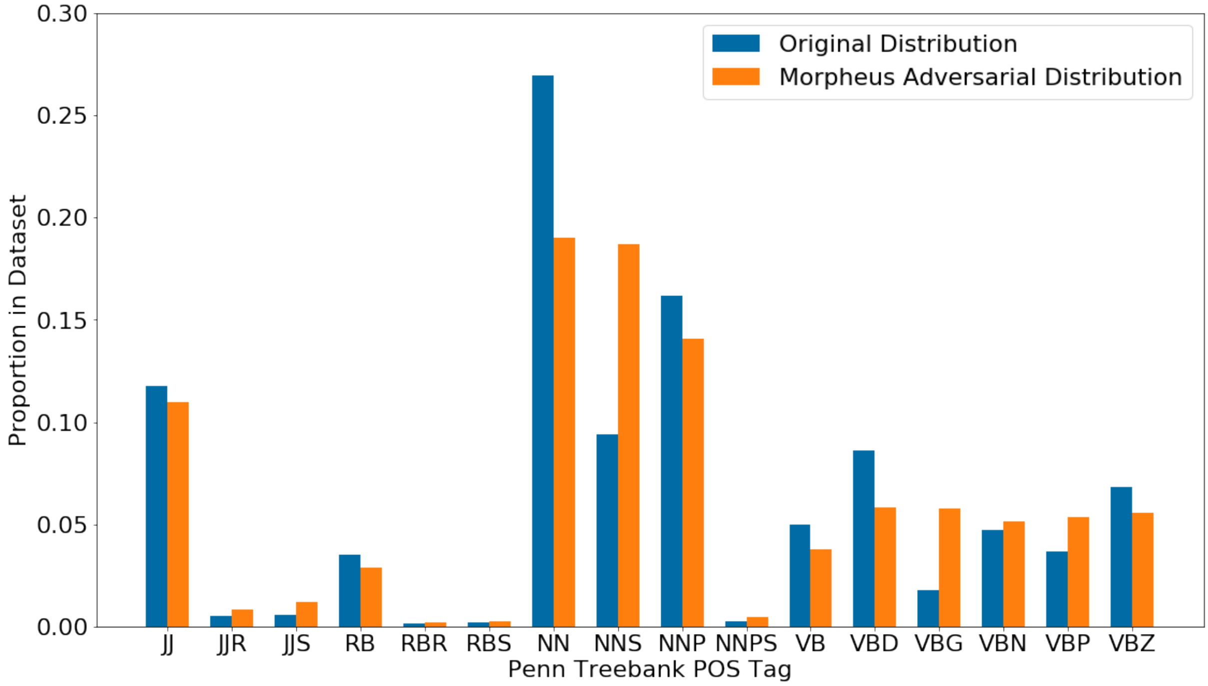


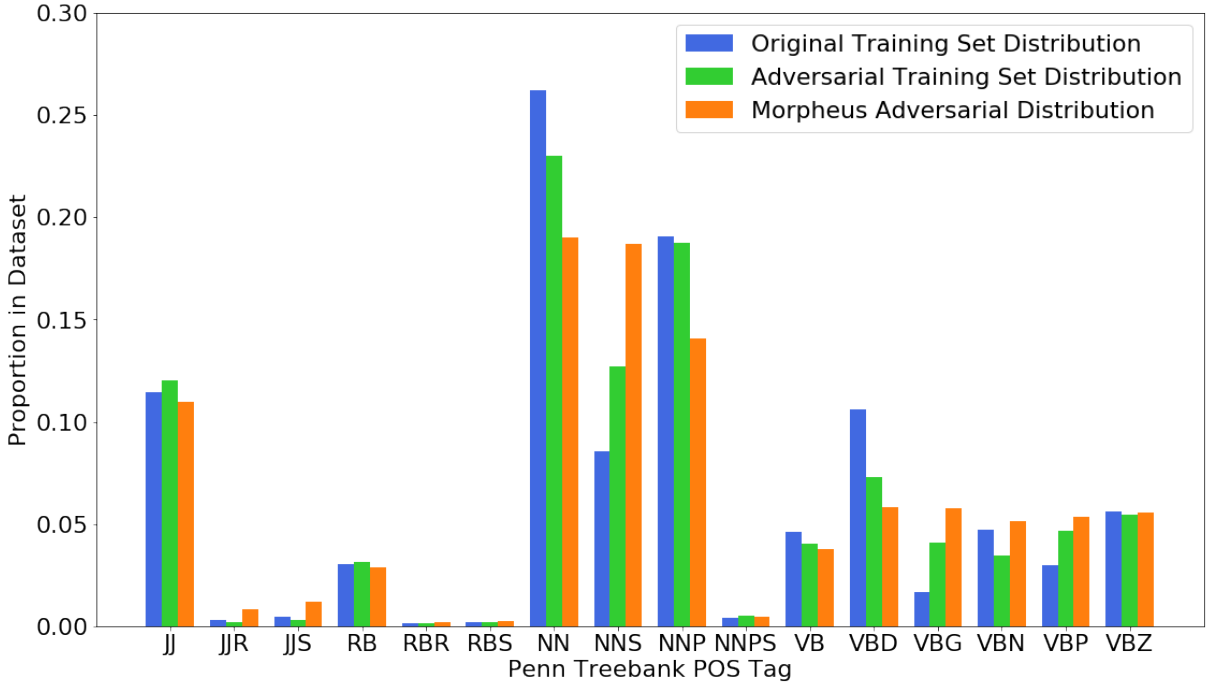
Figure 4: Effect of shuffling the inflection list on the adversarial distribution. We observe that shuffling the inflection list induces a more uniform inflectional distribution by reducing the higher frequency inflections and boosting the lower frequency ones.

Original Source	According to Detroit News, the queen of Soul will be performing at the Sound Board hall of MotorCity Casino Hotel on 21 December.
Adversarial Source	Accorded to Detroit News, the queen of Soul will be performing at the Sound Board hall of MotorCity Casino Hotel on 21 December.
Original Translation	Selon Detroit News, la reine de Soul se produira au Sound Board Hall de l'hôtel MotorCity Casino le 21 décembre.
Original Source	Intersex children pose ethical dilemma.
Adversarial Source	Intersex child posing ethical dilemma.
Original Translation	Les enfants intersexuels posent un dilemme éthique.
Original Source	The Guangzhou-based New Express made a rare public plea for the release of journalist Chen Yongzhou.
Adversarial Source	The Guangzhou-based New Expresses making a rare public plea for the release of journalist Chen Yongzhou.
Original Translation	Le New Express, basé à Guangzhou, a lancé un rare appel public en faveur de la libération du journaliste Chen Yongzhou.
Original Source	Cue stories about passport controls at Berwick and a barbed wire border along Hadrian's Wall.
Adversarial Source	Cue story about passport controls at Berwick and a barbed wires borders along Hadrian's Walls.
Original Translation	Cue histoires sur le contrôle des passeports à Berwick et une frontière de barbelés le long du mur d'Hadrien.

Table 9: Some of the adversaries that caused Transformer-big to output the source sentence instead of a translation.



(a) SQuAD 2.0 dev set



(b) SQuAD 2.0 training set

Figure 5: Full versions of Figure 2