VALUE: Understanding Dialect Disparity in NLU

Caleb Ziems **Jiaao Chen** **Camille Harris**
Jessica Anderson **Diyi Yang**
Georgia Institute of Technology
{cziems, jchen896, charris320}@gatech.edu
{janderson313, dyang888}@gatech.edu

Abstract

English Natural Language Understanding (NLU) systems have achieved great performances and even outperformed humans on benchmarks like GLUE and SuperGLUE. However, these benchmarks contain only textbook Standard American English (SAE). Other dialects have been largely overlooked in the NLP community. This leads to biased and inequitable NLU systems that serve only a sub-population of speakers. To understand disparities in current models and to facilitate more dialect-competent NLU systems, we introduce the VernAcular Language Understanding Evaluation (VALUE) benchmark, a challenging variant of GLUE that we created with a set of lexical and morphosyntactic transformation rules. In this initial release (V.1), we construct rules for 11 features of African American Vernacular English (AAVE), and we recruit fluent AAVE speakers to validate each feature transformation via linguistic acceptability judgments in a participatory design manner. Experiments show that these new dialectal features can lead to a drop in model performance.

1 Introduction

Most of today's research in NLP mainly focuses on 10 to 20 high-resource languages with a special focus on English, though there are thousands of languages and dialects with billions of speakers in the world. NLU systems that are trained on polished or "textbook" Standard American English (SAE) are not as robust to linguistic variation (Belinkov and Bisk, 2018; Ebrahimi et al., 2018). While some recent works have challenged leading systems with adversarial examples like typos (Jones et al., 2020), syntactic rearrangements (Iyyer et al., 2018), and sentence/word substitutions (Alzantot et al., 2018; Jia and Liang, 2017; Ribeiro et al., 2018), fewer have considered the effects of dialectal differences on performance. When language technologies are not built to handle dialectal differences, the benefits of these technologies may not be equitably distributed among different demographic groups (Hovy and Spruit, 2016). Specifically, models tested on African American Vernacular English (AAVE) have been found to struggle with language identification (Jurgens et al., 2017), sentiment analysis (Kiritchenko and Mohammad, 2018), POS tagging (Jørgensen et al., 2016) and dependency parsing (Blodgett et al., 2018), and led to severe racial disparities in the resulting language technologies such as the automated speech recognition used by virtual assistants (Koenecke et al., 2020) the hate speech detection used by online media platforms (Rios, 2020; Halevy et al., 2021).

However, no prior work has systematically investigated these dialect-specific shortcomings across a broad set of NLU tasks, and the effectiveness of low-resource NLP methods for dialectal Natural Language Understanding (NLU) remains largely unexplored. The first barrier to progress is that a standard benchmark for dialectal NLU has not yet been constructed. The second is that no systematic error analyses have yet revealed causal insights about the specific challenges that models face with domain adaptation to different language varieties.

To both understand dialect disparity and facilitate ongoing work on dialect-competent NLU, we introduce a new dialect-specific challenge dataset – the **V**ern**A**cular **L**anguage **U**nderstanding **E**valuation benchmark (**VALUE**). We specifically focus on African American Vernagular English (AAVE), a dialect spoken by nearly 33 million people, and approximately 80% of African Americans in the United States (Lippi-Green, 1997). To facilitate direct comparison with prior work, we build VALUE by directly transforming GLUE (Wang et al., 2019) into synthetic AAVE.

Our AAVE transformation pipeline comes with two key advantages: it is flexible enough to facilitate an interpretable perturbation error analysis, and

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3701 - 3720

the transformation rules are meaning-preserving, which ensures the validity of the transformed NLU tasks. Our pipeline includes a set of linguistically-attested rules for syntax (sentence structure; e.g. negation rules), morphology (word structure; e.g., suffixes), orthography (writing and spelling conventions), and the lexicon (the list of available words and phrases). Because our system is rule-based, we can isolate and systematically test which features most significantly challenge models. While it is also possible to generate pseudo-dialects via end-to-end style transfer (Krishna et al., 2020), these systems often fail to disentangle style from content, and thus also fail to preserve meaning (Lample et al., 2019). We confirm these shortcomings in this work, and affirm the validity of our own meaning-preserving transformation rules via the acceptability judgments of fluent AAVE speakers in a participatory design manner. To sum up, our work contributes the following:

1. **Dialect Transformations:** A set of 11 new linguistic rules for reliably transforming Standard American English (SAE) into African American Vernacular English (AAVE).
2. **VALUE**: An AAVE benchmark dataset with seven NLU tasks.
3. **Synthetic + Gold Standard Data:** Robust validation of synthetic transformations as well as gold standard dialectal data from native AAVE speakers via an iterative participatory design process.
4. **Benchmark Evaluation:** Experiments with RoBERTA baselines plus fine-tuning methods to improve model robustness on dialectal variants.
5. **Dialect-Specific Analysis:** Perturbation analysis that reveals the task-specific challenges of AAVE-specific grammatical features.

2 Related Work

Computational Sociolinguistics of Dialect Prior work on developing NLU models has often used dominant English varieties, Standard American English (SAE), owing to the availability of text datasets for training and testing (Blodgett et al., 2016). Models can marginalize certain groups when trained on datasets that lack linguistic diversity or contain biases against minority language speakers (Blodgett and O'Connor, 2017). Despite these shortcomings, there still has been relatively little attention paid to dialects in the language technologies research communities. Prior studies have mainly focused on distinguishing between English language varieties (Demszky et al., 2021a; Zampieri et al., 2014).

Failure to account for dialects like AAVE can lead to performance degradation of the NLU tools such as Automatic Speech Recognition (ASR) (Dorn, 2019), Language Identification (LID) and dependency parsing tools (Blodgett et al., 2016). Hwang et al. (2020a) also demonstrated the inadequacy of WordNet and ConceptNet in reflecting AAVE and other varieties. Thus there have been several works highlighting the need for AAVE-inclusivity in NLU (Groenwold et al., 2020). Despite its large community of speakers, AAVE is under-represented in current technologies.

Model Robustness and Challenge Datasets Language technologies are not inherently robust to linguistic variation. The performance of neural models is expected to degrade due to sparsity in the presence of non-canonical text (Zalmout et al., 2018; Belinkov and Bisk, 2018; Ebrahimi et al., 2018), as shown empirically for random character, word, and sentence-level permutations (Jones et al., 2020; Alzantot et al., 2018; Jia and Liang, 2017; Ribeiro et al., 2018; Iyyer et al., 2018). This has motivated growing interest in challenging datasets based on adversarial perturbations (Nie et al., 2020; Tan et al., 2020), spurious patterns or correlations (Zhang et al., 2019; McCoy et al., 2019), and counterfactual examples (Gardner et al., 2020; Kaushik et al., 2020). However, the same attention has not been shown to dialects, which vary *systematically* in their syntax, morphology, phonology, orthography, and lexicon (Jurgens et al., 2017). To this end, we introduce the evaluation set by adapting from the in-distribution examples (SAE) to out-of-distribution examples (AAVE) on GLUE benchmarks. Our goal is to develop robust models that have a good performance on test sets in different linguistic variations.

3 Constructing VALUE

We constructed VALUE from the widely-used GLUE benchmark (Wang et al., 2019), which contains NLU tasks such as natural language inference (e.g., MNLI; Bowman et al.), question answering (QNLI; Rajpurkar et al.), and linguistic acceptability (CoLA; Warstadt et al.). For each of the main tasks, we translated the Standard American English (SAE) into a synthetic form of AAVE — a form containing many of AAVE's distinguishing features with extremely high concentration. We implemented these transformations using a set of lexical and morphosyntactic rules derived from a broad survey of the linguistics literature (Collins et al., 2008; Green, 2002; Labov, 1972; Labov et al., 1998; Sidnell, 2002; Stewart, 2014; Thompson, 2016; Wolfram and Schilling, 2015). These features were specifically chosen for their high empirical attestation across regional and generational variants of AAVE.

3.1 Morphosyntactic Translation

This work represents the first attempt to systematically catalogue and operationalize a set of computational rules for inserting AAVE-specific language structures into text. We distill field linguists's observations into procedural code, which operates on specific grammatical conditions from the SAE source. Each grammatical condition is specified by the part of speech tags and syntactic dependency relationships present in the text. Appendix A.1 lists all implementation details for each transformation rule, and we will now enumerate them briefly.

Auxiliaries. AAVE allows copula deletion and other auxiliary dropping (Stewart, 2014; Green, 2002; Labov, 1972; Wolfram and Schilling, 2015). This means the SAE sentence "*We are better than before*" could be rendered in AAVE without the copula as "*We better than before*." We look for the present tense *is* and *are* as well as any tokens with `AUX` part of speech tag to drop (under special conditions listed in more detail in Appendix A.1).

Completive *done* and remote time *been*. The phrase "*I had written it.*" can be rendered in AAVE as "*I done wrote it*" using the completive verbal marker *done*. The phrase "*He ate a long time ago*" can be rendered as "*He been ate*" using the remote time *been* (Green, 2002).

Constructions involving the word *ass*. These constructions may be misclassified as obscenity, but they serve a distinct and consistent role in AAVE grammar (Spears et al., 1998). One common form is called the *ass* camouflage construction (Collins et al., 2008), and it can be seen in the phrase "*I divorced his ass*." Here, the word behaves as a metonymic pseudo-pronoun (Spears et al., 1998). Similarly, it can appear reflexively, as in "*Get yo'ass inside*." *Ass* constructions can also serve as discourse-level expressive markers or intensifiers, as in the compound "*We was at some random-ass bar*."

Existential *dey/it*. AAVE speakers can indicate something exists by using what is known as an *it* or *dey* existential construction (Green, 2002). The existential construction in "*It's some milk in the fridge*" is used to indicate "*There is some milk in the fridge*." We identify existential dependencies for this transformation.

Future *gonna* and immediate future *finna*. AAVE speakers can mark future tense with *gon* or *gonna*, as in "*You gon understand*" (Green, 2002; Sidnell, 2002). In the first person, this becomes *I'ma*. In the immediate future, speakers can use *finna* (or variants *fixina, fixna and fitna*), as in "*I'm finna leave*."

Have / got. In the casual speech of AAVE and other dialects, both the modal and the verb form of *have* can be replaced by *got* (Trotta and Blyahher, 2011). *Have to* can become *got to* or *gotta*, and similar for the verb of possession. We simply convert the present-tense *have* and *has* to *got* and ensure that the verb has an object.

Inflection. In AAVE, speakers do not necessarily inflect simple present or past tense verbs differently for number or person (Green, 2002). This means the SAE sentence "*She studies linguistics*" could be rendered in AAVE as "*She* ***study*** *linguistics*." We use the `pyinflect` library to convert all present and simple past verbs into the first person.

Negative concord. This widely-known feature of AAVE (and numerous other dialects) involves two negative morphemes to convey a single negation. (Martin et al., 1998). For example, the SAE sentence "*He doesn't have a camera*" could look more like "*He don't have no camera*" in AAVE. This transformation rule is sensitive to the verb-object dependency structure, and requires that the object is an indefinite noun (Green, 2002).

Negative inversion. This feature is superficially similar to negative concord. Both an auxiliary and an indefinite noun phrase are negated at the beginning of a sentence or clause (Green, 2002; Martin et al., 1998). For example, the SAE assertion that "*no suffering lasts forever*" could be rendered in AAVE as "*don't no suffering last forever*."

Null genitives. AAVE allows a null genitive marking (Stewart, 2014; Wolfram and Schilling, 2015), like the removal of the possessive *'s* in "*Rolanda bed*" (Green, 2002). We simply drop any possessive endings (`POS`) from the text.

Relative clause structures. There is a grammatical option to drop the Wh-pronoun when it is serving as the complementizer to a relative clause, as in "*It's a whole lot of people Ø don't wanna go to hell*" (Green, 2002). In our transformation, we simply drop all lemmas *who* and *that* where the head is a relative clause modifier.

3.2 Lexical and Orthographic Translation

Some of the most recognizable differences between SAE and AAVE are found in the lexicon and orthographic conventions. Because we are not aware of any comprehensive AAVE lexicons, we automatically learn our own SAE to AAVE dictionary from public data, and we will provide this resource in our public repository. This dictionary serves as a mapping between plausible synonyms (e.g., *mash/press*; *homie/friend*; *paper/money*) and orthographic variants (e.g., *da/the*; *wit/with*; *sista/sister*)

In a method inspired by Shoemark et al. (2018), we trained a skip-gram word embedding model[1] (Mikolov et al., 2013) on the public `TwitterAAE` dataset of Blodgett et al. (2016). This dataset contained attested code-switching behavior, which allowed us to extract a *linguistic code* axis c in the embedding space, defined by the average

$$\boldsymbol{c} = \sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i)\in S} \frac{\boldsymbol{x}_i - \boldsymbol{y}_i}{|S|}$$

where S was our seed list of known priors from Shoemark et al. (2018), given in Appendix A.2.

Next, we ranked the candidate word pairs $\boldsymbol{w}_i, \boldsymbol{w}_j$ by $\cos(\boldsymbol{c}, \boldsymbol{w}_i - \boldsymbol{w}_j)$ following Bolukbasi et al. (2016). In this ranking, we consider only the pairs whose cosine similarity passed a threshold δ, where δ was defined by the bottom quartile of the cosine similarities in our seed set S. After automatic filtering, we were left with 2,460 pairs. We hand-filtered this list to remove any semantically dissimilar words, like *fishin/kayakin* or *mom/gramps*. This left us with 1,988 pairs.

Note that these pairs are not one-to-one, but a one-to-many dictionary mapping from SAE to AAVE variants. We provide a sample of this mapping in Table 1. In the final step of the translation, we chose uniformly at random between the AAVE variants to make our substitution. We simply scanned the GLUE dataset and swapped any known tokens from SAE to AAVE.

SAE	AAVE
arguing	*beefing, beefin, arguin*
anymore	*nomore, nomo*
brother	*homeboy*
classy	*fly*
dude	*n*ggah, manee, n*gga*
huge	*bigass*
probably	*prob, prolly, def, probly, deff*
rad	*dope*
remember	*rememba*
screaming	*screamin, yellin, hollering*
sister	*sista, sis*
these	*dese, dem*
with	*wit*

Table 1: A sample of the SAE/AAVE synonym mapping that we learned automatically from corpus data.

3.3 Transformed Datasets

Our transformed tasks are all derived from GLUE. We skip *Diagnostics* because it is not a benchmark, and we do not transform the Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005) because it is proprietary. However, we do transform the remaining seven benchmarks, which include the *single-sentence tasks* (i) Stanford Sentiment Treebank (SST-2) which involves classifying the sentiment of movie reviews as positive or negative, and (ii) Corpus of Linguistic Acceptability (CoLA) which involves deciding whether a sentence is linguistically acceptable or not; the *similarity and paraphrase task* called Semantic Textual Similarity Benchmark (STS-B), which involves predicting the similarity ratings between two sentences; and the *inference tasks* (i) Multi-Genre Natural Language Inference (MNLI) which involves classifying the relationships between two sentences as entailment, contradiction, or neutral, (ii) Question Natural Language Inference (QNLI) which involves predicting whether a given sentence is the correct answer to a given question; and finally (iii) Recognizing Textual Entailment (RTE) which involves predicting an entailment relation between two sentences. Ta-

[1] We used `gensim` word2vec with dimension $d = 200$

Dataset	# data	ass	aux	been	dey/it	got	lexical	neg cncrd	null gen	null relcl	uninflect
CoLA	1,063	9%	15%	6%	2%	2%	51%	4%	3%	3%	17%
MNLI	9,682	30%	20%	9%	4%	5%	69%	4%	11%	10%	23%
QNLI	5,725	16%	42%	2%	1%	3%	50%	1%	10%	4%	17%
QQP	390,690	16%	2%	3%	63%	3%	59%	1%	3%	3%	13%
RTE	3,029	48%	40%	36%	3%	5%	81%	4%	28%	25	40%
SST-2	1,821	31%	25%	5%	3%	4%	64%	4%	14%	15%	39%
STS-B	1,894	1%	∼0	32%	2%	3%	2%	9%	4%	2%	5%
WNLI	146	48%	36%	38%	3%	16%	90%	1%	37%	12%	33%

Table 2: **Dataset statistics** reveal important differences between VALUE datasets, which come in markedly different sizes. The **%** columns reflect the proportion of data points in which the primary sentence or question was modified using the given transformation (e.g. the existential *dey/it*).

ble 2 provides a set of summary statistics for these datasets. It is clear that they come in different sizes, and that the some tasks have been more heavily modified than others. However, most of the sentences in this benchmark have undergone at least one transformation.

4 Speaker Validation and Gold-Standard

4.1 Validating Transformation Rules

Since our morphosyntactic transformations are rule-based rather than data-driven, it is especially important to validate that these rules are aligned with real AAVE speakers' grammaticality judgments.

User-Centered Validation Protocol. We opt for a participatory design process (Schuler and Namioka, 1993) to help ensure that these transformations are usable and meet the language practices of real speakers. We partnered with DataWorks, [2] an initiative started in Georgia Tech's College of Computing that seeks to involve members of underrepresented and economically disadvantaged groups in research and data annotation. All annotators were AAVE speakers and members of the Black community in Atlanta, and they were compensated for their time. Four volunteers from DataWorks partnered in the design of this rule-validation process. Specifically, we co-designed appropriate questions to measure the linguistic and social plausibility of our transformation system.

The HIT questions were based on a pair of utterances: (1) the original SAE sentence from the GLUE benchmark, and (2) the transformed AAVE sentence using only the morphosyntactic rules. We highlighted and indexed the portions of utterance (1) that were transformed in utterance (2), and we asked annotators for a binary grammaticality judgment. Separately, we asked for the *social* acceptability using a scale that was co-designed by DataWorkers. Then, for text marked as ungrammatical, annotators provided us with the indices at which transformation errors occurred. The task was hosted on the Amazon Mechanical Turk sandbox platform, but we interfaced with the annotators throughout the entire annotation process to answer any questions.

In early iterations of the task, DataWorkers discussed confusions and disagreements with the authors, and we discovered that the greatest variation in their judgments came not from differences in the speakers' underlying grammars, but rather from their different intuitions about what is *socially* acceptable (alternatively awkward and unnatural) to say in certain social settings. To disentangle these factors, DataWorkers helped us design a 10-point *social acceptability* Likert scale with the following vernacular: *If someone said this in your community, would it be (1) not very cool, (5) a bit sensitive, (7) passing, or (10) cool?*

Separately, we discussed certain orthographic conventions that we had adopted from the linguistics literature. DataWorkers indicated that some of these conventions were disagreeable – especially the spelling for *there are* as *dey* from Green (2002). Some DataWorkers suggested we use the spelling *dey're* instead. Relatedly, the DataWorkers found the *ass constructions* sensitive, given its long history of mischaracterization as an expletive, as well as the broader relationship between such dialect misunderstandings and racial injustice (Rickford and King, 2016; Rickford, 2016). We simply excluded *ass constructions* from the validation. DataWorkers also reported sentences from the original GLUE task were highly offensive (e.g. mentions

[2] https://dataworkforce.gatech.edu/

Transformation	Accuracy (Maj. Vote)	Accuracy (Unanimous)	Size n
Ass constructions	-	-	-
Auxiliaries	96.6	77.4	638
Been / done	95.4	72.7	670
Existential dey/it	91.4	57.9	304
Gonna / finna	95.4	78.7	197
Have / got	96.2	84.8	290
Inflection	97.1	82.3	761
Negative concord	95.9	73.6	584
Negative inversion	95.0	69.3	101
Null genitives	97.9	85.3	573
Relative clause structures	94.1	58.3	489

Table 3: Accuracy of SAE→AAVE transformations and n the number of instances present.

of sexual violence). We used the Perspective API[3] and the offensive language classifier Davidson et al. (2017) to filter out such instances.

Finally, we discussed the visual and interactive elements of the task itself. Workers preferred to see the synthetic AAVE text appear with visual priority above the SAE sentence. We also adjusted the color scheme to maximally distinguish concepts of social and grammatical acceptability. The word *acceptability* itself was triggering for the DataWorkers because it evoked the history of linguistic discrimination against AAVE speakers based on ignorant and prescriptive claims regarding "correct" or "proper" English. For this reason, we modified the prompt to read: *Do the words and the order of the words make sense?* With extensive follow-up meetings, we clarified that *to make sense* means a sentence follows expected and consistent language rules (i.e. a speaker's internal grammar).

Results. In the end, we collected acceptability judgments from three DataWorks workers for each of 2,556 randomly sampled sentence transformation pairs. We observed fair inter-annotator agreement with Krippendorf's α=0.26. Table 3 presents the aggregate judgments for local transformations in each morphological category. Here, we report the transformation accuracy as the proportion of local transformations marked as acceptable by majority vote or unanimous consensus, and we find our transformation rules are strongly validated. Majority vote gives nearly **100% accuracy** for all transformation types. Even under *strict* unanimous consensus, the accuracy exceeds 70% for seven of the 11 transformation types. Overall, this shows the quality of our linguistic transformations.

Error Analysis. Although our transformation rules are generally valid, errors can stem from an overapplication of the rule in restricted contexts. For example, most rules do not apply to idioms or named entities, so if we see a brand name like *Reese's Pieces*, we should not remove the possessive *s*. Other observed challenge cases include the subjunctive mood and subject inversions in questions, the non-standard morphology of certain contractions, as well as co-reference and scoping issues in relative clauses, ellipsis, and long-range dependencies (See Appendix D for more details). These each may introduce their own special cases that could be coded in future iterations. For a more reliable test set, we next construct a gold standard in Section 4.2

4.2 Building a Gold Test Set

Despite the advantages of controllable feature transformations for benchmarking with explainable error analysis, we cannot rely on the synthetic benchmark alone. Synthetic data may not fully capture the social and structural nuances of AAVE, nor speakers' dynamic and contextual use of dialect feature density. This motivates us to build a small test set of Gold Standard AAVE utterances. Here, annotators considered GLUE sentence transformations as before. The DataWorkers could either (1) confirm that synthetic transformation was natural, or alternatively (2) provide us with their own *translation* of the SAE text. Together, datapoints from (1) and (2) construct our Gold Test Set.[4] We provide the distribution of Gold Standard datapoints for each task in Table 4. In future iterations, we will expand the total size of the Gold Test sets for reliable benchmarking.

5 Benchmarking Models on VALUE

In this section, we stress-test current systems on NLU tasks and reveal performance drops on dialect-variants. We investigate the effectiveness of standard training on VALUE and we ablate the dialect test set to understand which dialect features most significantly challenge models.

We have two variants of synthetic AAVE data. In **AAVE (VALUE)**, we apply the full suite of Sec-

[3] https://www.perspectiveapi.com/

[4] **Note:** we did not build a Gold CoLA Test set. The nature of the annotation task would be ambiguous since CoLA itself contains intentionally ungrammatical utterances. It is not clear how annotators should translate ungrammatical SAE into ungrammatical AAVE.

	MNLI	QNLI	QQP	RTE	SST-2	STS-B	WNLI
# Hand-Validated Synthetic Transformations	365	333	355	120	71	160	123
# Natural AAVE Sentence Translations	291	330	314	72	80	104	162
Gold Test Set Total Size	656	663	669	192	151	264	285

Table 4: **Gold Test Set** size for each NLU task.

tion 3 transformations to the standard GLUE tasks. In **AAVE Morph**, we have an ablated variant of VALUE where only the morphosyntactic transformations (Section 3.1) are executed. By testing base SAE models on this data, we can disentangle the challenges associated with vocabulary shift from those associated with structural differences. If the challenges of VALUE were entirely lexical, we would anticipate that any performance disparity could be recovered with domain-specific word embeddings, since prior work has found such embeddings adequately represent the meanings of new words in AAVE corpora (Hwang et al., 2020b).

5.1 Standard Training

The most direct way to prepare models for a particular language variety is to directly train them on a dialect-variant of the task. Using our transformation rules (Section 3), we first augment the GLUE training set with AAVE features and then re-train the models (125M-parameter RoBERTA-base) on the augmented data. Following Liu et al. (2019), the batch size was 16. The maximum learning rate was selected as $5e-4$ and the maximum number of training epochs was set to be either 5 or 10.

5.2 Results

Table 5 compares the performance of RoBERTa models trained and tested on SAE or AAVE-variants of seven natural language understanding tasks in GLUE. Results are given as Matthew's Correlation for CoLA, Pearson-Spearman Correlation for STS-B, and Accuracy for all other tasks, averaged over three random seeds. In most cases, training jointly on GLUE and VALUE (SAE + AAVE) leads to **best performance**. With a single training set, there is an expected pattern: training with the corresponding train set typically leads to best performance on the corresponding test set. With the exception of RTE,[5] base models all suffer a drop in performance when tested on the full AAVE (VALUE) test set compared with the models trained on AAVE or jointly on SAE + AAVE (e.g., a 1.5% drop on SST-2; a 0.9% drop on QNLI compared to SAE + AAVE). Performance gaps of a similar magnitude are observed when we test on the Gold Test set (e.g., a 1.2% drop on SST-2; a 0.8% drop on QNLI). Further effort is needed to make the current NLU models more robust to dialect variations.

We also see that AAVE Morph challenges current models, which suggests that strategies for resolving any performance gap should take dialect morphology and syntax into consideration. Compared to the AAVE column, there is a less severe but still visible drop in AAVE Morph testing: from 94.3 to 93.2 in SST-2, and from 92.6 to 92.0 in QNLI, for instance. Thus we conclude that the challenge with dialects extends beyond a mere difference in the lexicon.

5.3 Perturbation Analysis

Finally, we run a perturbation analysis (Alvarez-Melis and Jaakkola, 2017) to better understand the impact of each dialectal feature on model performance. For the sake of simplicity, we focus only on MNLI. Specifically, we are interested in cases where the introduction of a particular feature results in a model error. Therefore, we count, for each feature transformation function T, the number of sentence pairs $(\boldsymbol{x}_i^0, \boldsymbol{x}_i^1)$ for which a GLUE-trained RoBERTA model f changes its prediction from a correct inference y_i to an incorrect inference under the transformation. Not all sentence structures allow for new features, so we consider only the subset of pairs for which the transformation is effective in the hypothesis sentence, and where the original GLUE pair had been predicted correctly. Then the ratio r_T is be defined as:

$$r_T = \frac{\left|\{(\boldsymbol{x}_i^0, \boldsymbol{x}_i^1) \in \mathcal{X}_T : f(\boldsymbol{x}_i^0, T(\boldsymbol{x}_i^1)) \neq y_i\right|\}}{|\mathcal{X}_T|}$$

Here $\mathcal{X}_T$ is:

$$\mathcal{X}_T = \{(\boldsymbol{x}_i^0, \boldsymbol{x}_i^1) : T(\boldsymbol{x}_i^1) \neq \boldsymbol{x}_i^1 \wedge f(\boldsymbol{x}_i^0, \boldsymbol{x}_i^1) = y_i\}$$

[5] RTE may be an outlier because of variance due to its small size: only 2.5k data points vs. QNLI with 100k

		Test	Synth. Testing		Gold
	Training	SAE	Morph.	AAVE	Test
CoLA	SAE (GLUE)	56.3	55.7	55.6	-
	AAVE Morph.	56.3	56.0	55.4	-
	AAVE (VALUE)	56.2	55.6	55.8	-
	SAE + AAVE	**57.8**	**56.2**	**56.5**	-
MNLI	SAE (GLUE)	83.6	83.0	82.8	82.1
	AAVE Morph.	82.2	83.3	82.5	82.3
	AAVE (VALUE)	83.1	83.2	83.5	82.9
	SAE + AAVE	**83.8**	**83.6**	**83.6**	**83.3**
QNLI	SAE (GLUE)	**92.8**	92.0	91.4	91.2
	AAVE Morph.	92.5	**92.6**	91.2	91.2
	AAVE (VALUE)	92.5	92.4	91.8	91.8
	SAE + AAVE	**92.8**	92.5	**92.3**	**92.0**
RTE	SAE (GLUE)	66.4	66.4	67.8	67.6
	AAVE Morph.	**68.9**	68.2	**69.7**	**68.8**
	AAVE (VALUE)	67.1	66.1	67.2	67.3
	SAE + AAVE	68.6	**68.9**	69.1	67.6
SST-2	SAE (GLUE)	94.6	93.2	92.4	92.0
	AAVE Morph.	94.0	94.3	92.3	92.1
	AAVE (VALUE)	94.0	93.8	93.0	92.8
	SAE + AAVE	**94.8**	**94.5**	**93.9**	**93.2**
STS-B	SAE (GLUE)	**89.4**	88.3	88.5	88.2
	AAVE Morph.	89.1	88.9	88.0	88.2
	AAVE (VALUE)	88.8	88.7	88.3	88.3
	SAE + AAVE	89.2	**89.0**	**88.9**	**88.5**
QQP	SAE (GLUE)	**90.9**	89.8	89.5	89.2
	AAVE Morph.	90.1	90.2	89.6	89.3
	AAVE (VALUE)	90.3	89.6	89.8	89.6
	SAE + AAVE	90.8	**90.3**	**90.1**	**89.8**

Table 5: **Dialect understanding results** for six tasks (Matthew's Corr. for CoLA; Pearson-Spearman Corr. for STS-B; Accuracy for all others). AAVE Morph is a subset of VALUE in which only the morphosyntactic transformations (Section 3.1) are executed. SAE + AAVE indicates training on the merged GLUE and VALUE train sets. Best performance is given in **bold** and best performance with a single train set is given in gray. The result gaps are significant. With the exception of RTE and STS-B, models trained on SAE (GLUE) suffer a drop in performance when tested on synthetic AAVE (VALUE) or the Gold Test.

and r_T indicates the proportion of inferences that were flipped to an incorrect label in the presence of T. We report this ratio for each feature in Table 6.

The first column in table shows that, when we introduce a *negative inversion* into a Hypothesis sentence for which the GLUE-trained RoBERTa model was originally correct, then in 9.09% of cases, that correct label would be flipped to an incorrect one.[6] The *inflection* rule and *been / done* constructions appear less challenging, but still result in 2.88% and 3.06% of new errors respectively. The remaining table columns indicate the contributions of different model mistakes to the overall r_T ratio. For example, the single error due to negative inversion occurs here when the model mistakes a neutral relationship for entailment (n→e) in the following pair: PREMISE: "*Still, commercial calculation isn't sufficient to explain his stand*" and HYPOTHESIS: "*Won't nothing be enough to explain his strong opinion*". In negative concord environments, we most often see neutral pairs mistakenly labeled as contradictory (n→c), as with the PREMISE: "*Each state is different...*" and HYPOTHESIS: "*You can go from one area of a state to another and not see no resemblance*. For more examples, see Tables 8 and 9 in Appendix C.

6 Why Not Use Style Transfer?

We qualitatively investigated the differences between our rule-based approach and a very well-performing unsupervised dialect style transfer model, STRAP Krishna et al. (2020). To train STRAP, we created a pseudo-parallel corpus using a diverse paraphrase model to paraphrase different styles of text, including SAE and the AAVE text from the `TwitterAAE` corpus Blodgett et al. (2018). Then we fine-tuned a GPT-2 model as the inverse paraphrase function, which learned to reconstruct the various styles. We used the SAE paraphrase model and the AAVE inverse paraphrase model to transfer from SAE to AAVE. In general, we found that STRAP is capable of much greater output diversity. However, in a systematic analysis of dialectal NLU, the first goal is to ensure that the underlying relationships like *entailment* are not distorted. STRAP can distort the meaning of the text with hallucinations and deletion of key details. Our transformation approach preserves the meaning of the text and thus better captures AAVE morphosyntax. See Appendix E for more details.

7 Conclusion

This work introduces the English VernAcular Language Understanding Evaluation (VALUE) benchmark, a challenging variant of GLUE that we created with a set of lexical and morphosyntactic transformation rules. We constructed rules for 11 fea-

[6] This is the highest error rate for any transformation rule. Note that $|\mathcal{X}_T| = 11$ datapoints is a much smaller sample size so the r_T estimate is more variable.

Feature	r_T	c→n	c→e	n→c	n→e	e→c	e→n	$\lvert\mathcal{X}_T\rvert$
Auxiliaries	4.20	0.20	0.07	**1.62**	0.88	0.68	0.74	1,477
Been / done	3.06	0.22	0.00	**1.31**	0.44	0.22	0.88	457
Inflection	2.88	0.33	0.20	0.59	0.46	0.39	**0.92**	1,526
Lexical	5.92	0.67	0.27	1.35	0.57	0.88	**2.18**	4,902
Negative concord	6.88	0.64	0.16	**2.56**	0.16	2.08	1.28	625
Negative inversion	**9.09**	0.00	0.00	0.00	**9.09**	0.00	0.00	11
Relative clause structures	5.86	0.31	0.62	1.23	0.62	0.31	**2.78**	324

Table 6: **Perturbation analysis.** The first column r_T gives the proportion of testing instances where the introduction of a particular dialect feature results in a new model error. This column indicates that *negative inversions* are the most challenging for MNLI. The final column gives the size of the set $\mathcal{X}_T$, which is the denominator in the ratio r_T. The remaining columns indicate the contributions of different error types to the cumulative r_T: the model flips the correct label on the left → into the incorrect label on the right side. **c**: contradiction; **n**: neutral; **e**: entailment.

tures of AAVE, and recruit fluent AAVE speakers to validate each feature transformation via linguistic acceptability judgments in a participatory design manner. Experiments show that the introduction of new dialectal features can lead to a drop in performance. We also test methods for efficiently adapting models to different language varieties, and discuss dialect specific challenges that our current NLP models are struggling with. Our work sheds light on the disparities of language technologies and has key implications for facilitating more dialect-competent NLU systems. Our longer term goals are to expand VALUE to more NLP tasks such as CoQA (Reddy et al., 2019), and to include other dialects such as Indian English (Demszky et al., 2021b; Lange, 2012; Bhatt, 2008) and Singapore English (Wee, 2008).

Limitations and Considerations. Researchers and practitioners should keep the following limitations and considerations in mind when using VALUE. Firstly, dialects are not the deterministic speech patterns that our transformation rules might suggest. While speakers of a dialect have *linguistic competence* over systematic and internalized grammar rules, speakers still posses an individual degree of control over which features they will employ (Coupland, 2007). The density of these features can vary, not only along demographic axes of geography, age, and gender (Nguyen et al., 2016), but also with different identity presentations in different social contexts (Bucholtz and Hall, 2005). We use VALUE to stress-test current systems by maximally modifying current resources with feature transformations. The high density of dialectal features may appear exaggerated here. Secondly, linguists have historically studied dialects through oral speech via live interviews (Rickford, 2002). The descriptions of academic references will not always map perfectly to the written domain (see Section 4.1 on the spelling of *dey*). The orthographic conventions of language communities may vary as significantly as do speech patterns. A third and critical concern is the limitation of synthetic data. Synthetic transformations have the advantage of allowing carefully controlled perturbation analysis and scaling up this analysis without the expensive creation of new datasets. However, synthetic data will not fully capture the social and structural nuances of AAVE, nor speakers' dynamic and contextual use of dialect feature density. For this reason, it is important to ultimately test user-facing models on domain-specific and gold-standard dialectal data. We are continuing to expand our gold-standard test set for GLUE tasks. A fourth consideration is the history of linguistic discrimination and the broader relationship between such dialect misunderstandings and racial injustice (Rickford and King, 2016; Rickford, 2016). AAVE has been frequently appropriated and misused by non-Black individuals, especially in online contexts (Reyes, 2005; Ilbury, 2020). To mitigate deployment risks, we ask users to sign a Data Use Agreement (See Ethics Section).

Acknowledgements

The authors would like to thank the DataWorks team, as well as Elizabeth DiSalvo, Rahual Gupta, and Jwala Dhamala for their helpful feedback. CZ is supported by the NSF Graduate Research Fellowship under Grant No. DGE-2039655. DY is supported by the Microsoft Research Faculty Fellowship. This work is funded in part by Amazon Research Award under the Alexa Fairness in AI.

Ethics

Our task comes from the public version of GLUE (Wang et al., 2019). Our annotation efforts revealed non-normative and offensive language in these original datasets, and we caution practitioners to be aware of this. The rules for converting SAE to AAVE are linguistically informed, and are not designed to change the original meaning of the sentence. Due to the participatory design nature of this work, we involved AAVE speakers and volunteers in the task creation and rule validation process. We asked annotators to skip a specific task and take a break if they are overwhelmed with the task. Our annotators were compensated by DataWorks for their time, and volunteered to help build this linguistic resource for their dialects. Note that AAVE is spoken, and our work only involves speakers from Atlanta. We ask that all users sign the following online agreement before using this resource: "*I will not use VALUE for malicious purposes including (but not limited to): deception, impersonation, mockery, discrimination, hate speech, targeted harassment and cultural appropriation. In my use of this resource, I will respect the dignity and privacy of all people.*"

References

David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, Copenhagen, Denmark. 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, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Emily M Bender. 2000. *Syntactic variation and linguistic competence: The case of AAVE copula absence*. stanford university Stanford, California.

Rakesh M Bhatt. 2008. Indian english: Syntax. In *A handbook of varieties of English*, pages 2208–2222. De Gruyter Mouton.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *CoRR*, abs/1707.00061.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter Universal Dependency parsing for African-American and mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Melbourne, Australia. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614.

Chris Collins, Simanique Moody, and Paul M Postal. 2008. An aae camouflage construction. *Language*, pages 29–68.

Nikolas Coupland. 2007. *Style: Language variation and identity*. Cambridge University Press.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *ArXiv preprint*, abs/1703.04009.

Dorottya Demszky, Devyani Sharma, Jonathan Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021a. Learning to recognize dialect features. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2315–2338, Online. Association for Computational Linguistics.

Dorottya Demszky, Devyani Sharma, Jonathan Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021b. Learning to recognize dialect features. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2315–2338, Online. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Rachel Dorn. 2019. Dialect-specific models for automatic speech recognition of African American Vernacular English. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 16–20, Varna, Bulgaria. INCOMA Ltd.

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.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Lisa J Green. 2002. *African American English: a linguistic introduction*. Cambridge University Press.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in transformer-based text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883, Online. Association for Computational Linguistics.

Matan Halevy, Camille Harris, Amy Bruckman, Diyi Yang, and Ayanna Howard. 2021. Mitigating racial biases in toxic language detection with an equity-based ensemble framework. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–11.

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.

Alyssa Hwang, William R. Frey, and Kathleen McKeown. 2020a. Towards augmenting lexical resources for slang and African American English. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 160–172, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Alyssa Hwang, William R. Frey, and Kathleen McKeown. 2020b. Towards augmenting lexical resources for slang and African American English. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 160–172, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Christian Ilbury. 2020. "sassy queens": Stylistic orthographic variation in twitter and the enregisterment of aave. *Journal of Sociolinguistics*, 24(2):245–264.

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.

Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang. 2020. Robust encodings: A framework for combating adversarial typos. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2752–2765, Online. Association for Computational Linguistics.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2016. Learning a POS tagger for AAVE-like language. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1120, San Diego, California. Association for Computational Linguistics.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada. Association for Computational Linguistics.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning the difference that makes A difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Christopher Kennedy. 2007. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and philosophy*, 30(1):1–45.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

William Labov. 1972. *Language in the inner city: Studies in the Black English vernacular*. 3. University of Pennsylvania Press.

William Labov. 1995. The case of the missing copula: The interpretation of zeros in african-american english. *Language*, 1:25–54.

William Labov et al. 1998. Co-existent systems in african-american vernacular english. *African-American English: Structure, History and Use*, pages 110–153.

Guillaume Lample, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Claudia Lange. 2012. *The syntax of spoken Indian English*, volume 45. John Benjamins Publishing.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Rosina Lippi-Green. 1997. What we talk about when we talk about ebonics: Why definitions matter. *The Black Scholar*, 27(2):7–11.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Stefan Martin, Walt Wolfram, et al. 1998. The sentence in african-american vernacular english. *African American English: structure, history, and use*, pages 11–36.

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.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska De Jong. 2016. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. 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.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Angela Reyes. 2005. Appropriation of african american slang by asian american youth 1. *Journal of Sociolinguistics*, 9(4):509–532.

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 R Rickford. 2002. How linguists approach the study of language and dialect. *Ms. January*.

John R Rickford. 2016. *Raciolinguistics: How language shapes our ideas about race*. Oxford University Press.

John R Rickford and Sharese King. 2016. Language and linguistics on trial: Hearing rachel jeantel (and other vernacular speakers) in the courtroom and beyond. *Language*, 92(4):948–988.

Anthony Rios. 2020. Fuzze: Fuzzy fairness evaluation of offensive language classifiers on african-american english. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 881–889.

Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices*. CRC Press.

Philippa Shoemark, James Kirby, and Sharon Goldwater. 2018. Inducing a lexicon of sociolinguistic variables from code-mixed text. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 1–6, Brussels, Belgium. Association for Computational Linguistics.

Jack Sidnell. 2002. African american vernacular english (aave) grammar. 1.7. *Retrieved April*, 19(2009):16.

Arthur K Spears et al. 1998. African-american language use: Ideology and so-called obscenity. *African-American English: Structure, history, and use*, pages 226–250.

Ian Stewart. 2014. Now we stronger than ever: African-American English syntax in Twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 31–37, Gothenburg, Sweden. Association for Computational Linguistics.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

Deanna Thompson. 2016. The morpho-syntax of aspectual stay in aave.

Joe Trotta and Oleg Blyahher. 2011. Game done changed: A look at selected aave features in the tv series the wire. *Moderna språk*, 105(1):15–42.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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

Lionel Wee. 2008. Singapore english: morphology and syntax. In *A handbook of varieties of English*, pages 2250–2264. De Gruyter Mouton.

Walt Wolfram and Natalie Schilling. 2015. *American English: dialects and variation*. John Wiley & Sons.

Nasser Zalmout, Alexander Erdmann, and Nizar Habash. 2018. Noise-robust morphological disambiguation for dialectal Arabic. 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 953–964, New Orleans, Louisiana. Association for Computational Linguistics.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. 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 Details on the Transformation Rules

A.1 Morphosyntactic translation

We conduct morphosyntactic translation as the first step in the pipeline because our methods are based on grammatical rules that are determined by an SAE dependency parse. Here, we provide further details for our methods. We rely on `spaCy` to dependency parse the GLUE text at the sentence level before proceeding.

Inflection. In AAVE, speakers do not inflect simple present or past tense verbs differently for number or person (Green, 2002). This means the SAE sentence "*She studies linguistics*" would be rendered in AAVE as "*She **study** linguistics*." We identify all regular present verbs by their `VBZ` or `VBP` part of speech tag, and regular past verbs by their `VBD` part of speech tag. Then we inflect these verbs to standard first-person `VBP` or `VBD` respectively, using the `pyinflect` library.

Auxiliaries. In AAVE, auxiliaries with negated heads can be replaced by *ain't* (Green, 2002), and we make this conversion first. Copula deletion and optional auxiliary dropping are also grammatical in AAVE (Stewart, 2014; Green, 2002; Labov, 1972; Wolfram and Schilling, 2015). This means the SAE sentence "*We are better than before*" would be rendered in AAVE without the copula as "*We better than before*." The question "*Did you see that?*" could be rendered without the auxiliary verb as "*You see that?*" Similarly, the phrase "*I have seen him*" could go without the auxiliary, as in "*I seen him*."

We treat the dropped copula as a separate case. Since it only applies to the present tense *is* and *are*, we search for these tokens and check that the environment is one where *contraction* would be allowed in SAE. We ensure the copula is not negated and that it has an object dependant; that is neither a clausal complement nor the head of a clausal complement. We have confirmed that these decisions all account for the fact that copula deletion is disallowed in non-finite contexts, imperatives, ellipsis, inversion environments, or complement and subject extraction environments (Bender, 2000; Labov, 1995).

To account for other auxiliary dropping, we drop tokens with the `AUX` part of speech tag. We do not drop modals (tag `MD`), the future tense marker *will*, or any token whose head is a copula or an open clausal complement (`xcomp`).

Existential *dey/it*. AAVE speakers can indicate something exists by using what is known as an *it* or *dey* existential construction (Green, 2002). The existential construction in "*It's some milk in the fridge*" is used to mean "*There is some milk in the fridge*." We make this transformation by searching the text for expletive or pleonastic nominals (`expl` dependencies) and substituting these tokens with either *it* or *dey* with equal probability.

Negative concord. This phenomenon, also called *multiple negation* or *pleonastic negation*, is "the use of two negative morphemes to communicate a single negation," a widely-known feature of AAVE (Martin et al., 1998). For example, the SAE sentence "*He doesn't have a camera*" would become "*He don't have no camera*." To capture this transformation, we search the text for `neg` dependents of verbal heads. Then we negate the object dependents[7] of that verbal head. In these constructions, the negation can only be marked on auxiliaries and indefinite nouns (Green, 2002), but not definite nouns. We check for indefiniteness by ensuring that the object itself has only indefinite determiner children (a/an), and that the object is not a proper noun (`NNP`), nor a personal pronoun (`PRP`), nor is it an adjective modifier (`amod`). We also ensure that the object is not already a Negative Polarity Item (e.g. *nobody*, *nothing*).

Negative inversion. This AAVE feature is superficially similar to negative concord. Both an auxiliary and an indefinite noun phrase are negated at the beginning of a sentence or clause (Green, 2002; Martin et al., 1998). For example, the SAE assertion that "*no suffering lasts forever*" would be rendered in AAVE as "*don't no suffering last forever*." Since there was no auxiliary already present to front and negate, the syntax required obligatory *do* support. When the statement contains an auxiliary, the auxiliary verb will be fronted and negated instead, as in the transformation from "*Nobody can hear you*" to "*Can't nobody hear you.*" We operationalize these rules using the dependency parse. Specifically, we identify the span of the given clause by traveling up the dependency tree until we hit a ROOT, conjunction (`conj`), or complement dependency; then we use tree traversal from that origin to find the smallest index in the clause. In this way, we confirm that the negation is clause-initial.

[7] 'dobj', 'iobj', 'obj', 'pobj', 'obl', 'attr'

Relative clause structures. AAVE speakers most frequently use the complementizer *that* to introduce relative clauses, rather than using Wh-pronouns like *who, where, when* (Martin et al., 1998). There is also a grammatical option to drop the complementizer altogether. For example, "*There are a whole lot of people who don't want to go to hell*" could become in AAVE, "*It's a whole lot of people don' wanna go to hell*" (Green, 2002). In our transformation, we simply drop all lemmas *who* and *that* where the head is a relative clause modifier (`relcl`).

Null genitives. AAVE allows a null genitive marking (Stewart, 2014; Wolfram and Schilling, 2015). For example, "*Rolanda's bed isn't made*" can be rendered "*Rolanda bed don't be made up*" (Green, 2002). To capture this pattern, we simply drop any possessive endings (`POS`) from the text.

Completive *done* and remote time *been*. The phrase "*I had written it.*" can be rendered in AAVE as "*I done wrote it*" using the completive verbal marker dən. The phrase "*He ate a long time ago*" can be rendered as "*He been ate*" using the remote time BIN (Green, 2002). To operationalize this construction, we search for simple past verbs (`VBD`) with temporal noun phrase adverbial modifier children (`npadvmod`), like the *yesterday* in *I ate yesterday.* Then appended either *done* or *been* preverbally, each with equal probability. We also consider past participle verbs (`VBN`), and we replace the *have* auxiliaries with *done*/*been*.

***Ass* constructions.** These constructions may be mis-classified as obscenity, but they serve a distinct and consistent role in AAVE grammar (Spears et al., 1998). One common form called the *ass camouflage construction* (Collins et al., 2008) can be seen in the phrase "*I divorced his ass.*" Here it behaves as a metonymic pseudo-pronoun (Spears et al., 1998). Similarly, the form can appear reflexively, as in "*Get yo'ass inside.*" *Ass* constructions can also serve as discourse-level expressive markers or intensifiers, as in the compound "*We was at some random-ass bar.*" To operationalize the former, we substitute the appropriate *ass* construction for any personal pronoun (`PRP`) that was the object of a verb. To operationalize the latter, we transform adjective modifiers (`amod`). Not all adjectives can participate in this construction, however. That is why we consider only *gradable* adjectives (Kennedy, 2007), or adjectives that accept comparative and superlative modifiers and morphology. For example, *cold* can become *colder, very cold, coldest*, so a *cold-ass day* is an acceptable phrase in AAVE. Non-gradable or absolute adjectives like *finished* and *American* cannot participate; it is not acceptable to say *this finished-ass project* or *that American-ass woman* in AAVE.

Future *gonna* and immediate future *finna*. In AAVE, the future tense is marked by *gon* or *gonna* instead of *will*, as in "*You gon understand*" (Green, 2002; Sidnell, 2002). In the first person, this becomes *I'ma*. In the immediate future, speakers can use *finna* (or variants *fixina, fixna and fitna*), as in "*I'm finna leave.*" Although they are morphosyntactic, we treat these cases with simple lexical substitution.

Have / got. In the casual speech of AAVE and other dialects, both the modal and the verb form of *have* can be replaced by *got* (Trotta and Blyahher, 2011). *Have to* can become *got to* or *gotta*, and similar for the verb of possession. We simply convert the present-tense *have* and *has* to *got* and ensure that the verb has an object.

A.2 Lexical and orthographic translation

The seed list from Shoemark et al. (2018) contained the (1) *the/tha*, (2) *with/wit*, (3) *getting/gettin*, (4) *just/jus*, (5) *and/nd*, (6) *making/makin*, (7) *when/wen*, (8) *looking/lookin*, (9) *something/somethin*, (10) *going/goin*.

B Lightweight Training

Directly training new models for every language variety is expensive in both compute time and storage space. This motivates a lightweight fine-tuning strategy inspired by the state-of-the-art prefix-tuning method (Li and Liang, 2021). Specifically, we freeze the models trained on SAE. Then, for each dialect d, we fine-tune a transformation matrix M_d. When training the dialect-specific model, we append M_d to the embeddings e_i^d of each input sequence x_i^d. The matrix M_d is the only parameter that needs to be trained and stored besides the base model. During inference on dialect d, we can directly fetch base SAE model and the corresponding transformation matrix M_d to form the dialect-specific model and make predictions. Besides efficient domain adaptation, one additional advantage may be improved out-of-domain generalization (Li and Liang, 2021). Following Li and

Liang (2021), we used a batch size of 16. The prefix length was set to 50; the maximum learning rate was $5e-4$; the maximum number of training epochs was 5.

Results are given in Table 7. The second row for each task is labeled Prefix Tuning, and it gives the results of our lightweight fine-tuning approach. Prefix tuning demonstrates reasonable performance, but, with the exception of SST-2 sentiment analysis, Prefix Tuning fails to match the performance of full AAVE (VALUE) training. Thus there is still a need for more effective and efficient domain adaption methods for dialects like AAVE.

		Synthetic Testing	
	Training	SAE	AAVE
CoLA	AAVE (VALUE)	**56.2**	**55.8**
	Prefix Tuning	17.0	17.1
MNLI	AAVE (VALUE)	**83.1**	**83.5**
	Prefix Tuning	82.1	81.5
QNLI	AAVE (VALUE)	**92.5**	**91.8**
	Prefix Tuning	86.7	86.0
RTE	AAVE (VALUE)	**67.1**	**67.2**
	Prefix Tuning	54.5	54.1
SST-2	AAVE (VALUE)	94.0	93.0
	Prefix Tuning	**94.6**	**93.1**
STS-B	AAVE (VALUE)	**88.8**	**88.3**
	Prefix Tuning	27.8	25.1
QQP	AAVE (VALUE)	**90.3**	**89.8**
	Prefix Tuning	88.7	88.0

Table 7: **Lightweight tuning results** for six tasks (Matthew's Corr. for CoLA; Pearson-Spearman Corr. for STS-B; Accuracy for all others). Prefix Tuning fails to match the performance of full AAVE (VALUE) training

C Detailed Examples from the Perturbation Analysis

For each transformation type, we provide an example of each error category in Tables 8 and 9 when applicable. Here, we will briefly discuss our observations. For *aux-dropping*, the most common error is to confuse neutral relationships for contradictions (**n**→**c**). The model may fail to link the subject with the predicate of the HYPOTHESIS without the overt copula. We also notice an entailment relation mistaken for a contradiction when the loss of the auxiliary verb renders *mine mutts* syntactically ambiguous. Its position in the sentence suggests a Noun Phrase where the possessive pronoun *mine* is used in place of the possessive adjective *my*.[8] For *completive done* and *remote time been*, **n**→**c** is the most common error due again, possibly due to a failure to link subject and predicate. However, the converse error **c**→**n** may be triggered for similar reasons, as in *The woman done never spoke before.* For both the *inflection* rules and the *lexical* changes, the most common error is to mistake an entailment relationship for a neutral one. This may be due to the fragmenting of common subsequences and an overall reduced lexical similarity between the PREMISE and the HYPOTHESIS. Both lexical overlap and subsequence matching are well-known heuristics for NLI (McCoy et al., 2019). Finally, we recognize that some errors may arise from semantically ambiguous transformations. For example, in the **c**→**n** Lexical error, the word *right* was swapped for the alternative spelling *rite*, which is misleading in the context of church, since *rite* typically refers to a religious or ceremonial act. The transformation is not technically erroneous, but the setting renders it unfairly ambiguous.

D Detailed Error Analysis

Here, we provide a more detailed error analysis for our transforamtion system, organize by transformation rules.

- *Broader issues.* GLUE contains examples from journalism and news, which tends to use a more academic or formal register. Some of the annotators were not accustomed to seeing language variation in the body copy of a newspaper and so they identified stylistic errors that may have been grammatically well-formed. On the other hand, GLUE also contains purposeful disfluencies, which harm the performance of the syntactic parsers in our pipeline.

- *Existential dey/it.* Some annotators held that this feature should not be present in questions, so the example *is there a place?* should not be swapped with *is it a place?*

[8] This usage can be seen in sources such as the KJV Bible and in older hymns like the 1862 *Mine Eyes Have Seen the Glory*.

- *Auxiliaries.* Copula dropping is not allowed in questions or in cases of ellipsis like "*I like Bill's new wine, but Max's old ∅ even better.*" Similarly, with long-range dependencies in appositive phrases, we should not drop that be-verb, as done here: "*Hyperthymesia, or hyperthymesic syndrome, ∅ a disorder.*"

- *Been / done.* Errors appeared in this rule's handling of negation, like in "*This was a persistent problem which has not been solved*" becoming "*This was a persistent problem which done not been solved.*"

- *Have / got.* Some annotators found this transformation unacceptable in the subjunctive mood. For example, "*If the United States has a female president, ...* became "*If the United States got a female president.*"

- *Inflection.* Some errors with inflection occur when the POS tagger erroneously marks a noun as a root verb or similar. Also, inflection rules should not apply to idioms like "*just goes to show...*"

- *Negative concord.* Again, this rule should not apply to idiomatic or phrasal constituents. Negative concord also does not apply to finite nouns, including demonstratives, like in this error, swapping "*Couldn't nothing in Siegel's work explain this perception*" for "*Couldn't nothing in Siegel's work explain **no** perception.*"

- *Null genitive.* The possessive *'s* should not be dropped when doing so would lead to syntactic ambiguity. For example, "*How I see someone's deleted Instagram account?*" would become ambiguous in "*How I see someone deleted Instagram account?*" A similar situation arises with the complex object NP in "*to grab the old lady at the end of my aisle**'s** walker.*"

Through iterative discussion with the DataWorkers throughout the participatory design process, we also gleaned the following insights. We recognize that it is possible for a linguistic transformation to alter the connotative or social meaning of a text without altering the denotative semantic meaning. In the following example, the phrase "done cut her off" is linguistically acceptable. Furthermore, the truth conditional meanings of (1) and (2) are equivalent. However, the social connotations differ.

> (1) **GLUE**: Beth didn't get angry with Sally, who had cut her off, because she stopped and counted to ten.
>
> (2) **VALUE**: Beth ain't get angry with Sally, done cut her off, because she stopped and counted to ten.

One undergraduate annotator explained that *"Done cut her off"* would be used if, for instance, the speaker was getting mad while explaining the details, and just threw that small piece of information in their speech.

Annotators also identified examples in which certain features could apply, but the rules had not yet implemented in our system. For example, the habitual *be* (Stewart, 2014; Green, 2002; Wolfram and Schilling, 2015; Labov et al., 1998) would apply in the sentence "*Guatemala **be** accepting the Pet passport as proof of vaccination.*" In order to capture these missing features and widen the diversity of our test sets, we build the Gold Standard Test sets in Section 4.2.

E Qualitative Comparison Between VALUE and Style Transfer

Examples 1-5 show STRAP's creative phrase-level transformations, like converting "absurdly suspicious" to the phrase "weird as hell." However, STRAP can distort the text by hallucinations like in Examples 6-7. In (8), STRAP removes the name of the subject *Cochran*, a valuable detail. The neural style-transfer can also produce erratic behaviors as we see in 9 and 10. VALUE on the other hand appears to better capture AAVE morphosyntax. We see future *gon* in 11, negative concord in 12, copula dropping in 13, uninflection in 14, and an ass intensifier in 15, none of which are represented in the STRAP output. These are the primary advantages of our approach: (1) integrity of underlying constructs, and (2) linguistically attested features that can be systematically analyzed.

Transf.	**Error**	PREMISE	HYPOTHESIS
Auxiliaries	**c→n**	Energy-related activities are the primary source of U.S. man-made greenhouse gas emissions.	*Producing cars is the main source of US greenhouse gas emissions.* ↪ Producing cars the main source of US greenhouse gas emissions.
	c→e	Search out the House of Dionysos and the House of the Trident with their simple floor patterns, and the House of Dolphins and the House of Masks for more elaborate examples, including Dionysos riding a panther, on the floor of the House of Masks.	*The floor patterns of the House of the Trident are very intricate.* ↪ The floor patterns of the House of the Trident very intricate.
	n→c	To the west of the city at Hillend is Midlothian Ski Centre, the longest artificial ski slope in Europe.	*The Midlothian Ski Centre is the only artificial ski slope in Scotland.* ↪ The Midlothian Ski Centre the only artificial ski slope in Scotland.
	n→e	Mykonos has had a head start as far as diving is concerned because it was never banned here (after all, there are no ancient sites to protect).	*Protection of ancient sites is the reason for diving bans in other places.* ↪ Protection of ancient sites the reason for diving bans in other places.
	e→c	oh yeah all all mine are uh purebreds so i keep them in	*none of mine are mutts* ↪ none of mine mutts
	e→n	This particular instance of it stinks.	*It is a terrible situation.* ↪ It a terrible situation.
Been / done	**c→n**	She had spoken with no trace of foreign accent.	*The woman had never spoken before.* ↪ The woman done never spoken before.
	n→c	For more than 26 centuries it has witnessed countless declines, falls, and rebirths, and today continues to resist the assaults of brutal modernity in its time-locked, color-rich historical center.	*Modernity has made no progress in the historical center.* ↪ Modernity done made no progress in the historical center.
	n→e	(And yes, he has said a few things that can, with some effort, be construed as support for supply-side economics.)	*He has begun working on construing the things as support for supply-side economics.* ↪ He done begun working on construing the things as support for supply-side economics.
	e→c	Detroit Pistons they're not as good as they were last year	*Detroit Pistons played better last year* ↪ Detroit Pistons done played better last year
	e→n	I don't know what I would have done without Legal Services, said James.	*James said Legal Services helped him a lot.* ↪ James said Legal Services been helped him a lot.
Inflection	**c→n**	Once or twice, but they seem more show than battle, said Adrin.	*Adrin said they were amazing warriors.* ↪ Adrin said they was amazing warriors.
	c→e	The story of the technology business gets spiced up because the reality is so bland.	*Reality is so bland that the garbage business gets spiced up.* ↪ Reality is so bland that the garbage business get spiced up.
	n→c	The air is warm.	*The arid air permeates the surrounding land.* ↪ The arid air permeate the surrounding land.
	n→e	Long ago–or away, or whatever–there was a world called Thar?? and another called Erath.	*Thar and Erath were not the only worlds in existence then.* ↪ Thar and Erath was not the only worlds in existence then.
	e→c	The disputes among nobles were not the first concern of ordinary French citizens.	*Ordinary French citizens were not concerned with the disputes among nobles.* ↪ Ordinary French citizens was not concerned with the disputes among nobles.
	e→n	Perched on a steep slope, high in the Galilean hills, Safed (known also as Tzfat, Tsfat, Sefat, and Zefat) is a delightful village-town of some 22,000 people.	*Safed is a village that goes by numerous other names.* ↪ Safed is a village that go by numerous other names.

Table 8: **Example perturbation errors** for *aux-dropping, been/done,* and *inflection* transformations. The HYPOTHESIS was transformed from the original *SAE* ↪ Synthetic AAVE.

Transf.	Error	PREMISE	HYPOTHESIS
Lexical	**c→n**	Oh, sorry, wrong church.	*It was the right church.* ↪ It was the rite church
	c→e	The story of the technology business gets spiced up because the reality is so bland.	*Reality is so bland that the garbage business gets spiced up.* ↪ Reality is so bland that the garbage bizness gets spiced up.
	n→c	And who should decide?	*No one is willing to make the decision.* ↪ No one is willin to make the decision.
	n→e	At the eastern end of Back Lane and turning right, Nicholas Street becomes Patrick Street, and in St. Patrick's Close is St. Patrick's Cathedral.	*Back Lane and Nicholas Street are longer than Patrick Street.* ↪ Bacc Lane and Nicholas Street r longer than Patrick Street.
	e→c	Even analysts who had argued for loosening the old standards, by which the market was clearly overvalued, now think it has maxed out for a while.	*Some analysts wanted to make the old standards easier.* ↪ Sum analysts wanted to make the old standards easier.
	e→n	8 million in relief in the form of emergency housing.	*Emergency housing relief totaled 8 million dollars.* ↪ Emergency housing relief totaled 8 million dollas.
Negative Concord	**c→n**	Agencies may perform the analyses required by sections 603 and 604 in conjunction with or as part of any other agenda or analysis required by other law if such other analysis satisfies the provisions of these sections.	*The agency is free to decide not to perform the analyses covered in section 603.* ↪ The agency is free to decide not to perform no analyses covered in section 603.
	c→e	and clean up is is uh is a joy uh a little soap and water and air dry them and you don't have to worry about that	*You don't need any soap for the clean up.* ↪ You don't need no soap for the clean up.
	n→c	Each state is different, and in some states, intra-state regions differ significantly as well.	*You can go from one area of a state to another and not see a resemblance.* ↪ You can go from one area of a state to another and not see no resemblance
	n→e	Bars with views and live music include Sky Lounge in the Sheraton Hotel and Towers, Tsim Sha Tsui; and Cyrano in the Island Shangri-La in Pacific Place.	*It is not far to go to find good drink, gorgeous views, and live music.* ↪ It ain't far to go to find good drink, gorgeous views, and live music.
	e→c	Regulators may not be totally supportive of a more comprehensive business model because they are concerned that the information would be based on a lot of judgment and, therefore, lack of precision, which could make enforcement of reporting standards difficult.	*Being totally supportive of a more comprehensive business model is not something regulators may do.* ↪ Being totally supportive of a more comprehensive business model ain't something regulators may do.
	e→n	uh well no i just know i know several single mothers who absolutely can't afford it they have to go with the a single uh what i mean a babysitter more more or less	*They simply don't have the money to put into that sort of thing.* ↪ They simply don't have no money to put into that sort of thing.
Relative Clause Structures	**c→n**	He unleashed a 16-day reign of terror that left 300 Madeirans dead, stocks of sugar destroyed, and the island plundered.	*He unleashed a large debate over the 16-day reign that ended in a peaceful protest.* ↪ He unleashed a large debate over the 16-day reign ended in a peaceful protest.
	c→e	Shoot only the ones that face us, Jon had told Adrin.	*Shoot the ones that face us, Adrin told Jon* ↪ Shoot the ones face us, Adrin told Jon
	n→c	The conspiracy-minded allege that the chains also leverage their influence to persuade the big publishers to produce more blockbusters at the expense of moderate-selling books.	*Most people who read a book, tend to watch a film adaptation of it.* ↪ Most people read a book, tend to watch a film adaptation of it.
	n→e	Wear a nicely ventilated hat and keep to the shade in the street.	*The street has plenty of shade for those who want it.* ↪ The street has plenty of shade for those want it.
	e→c	External Validity The extent to which a finding applies (or can be generalized) to persons, objects, settings, or times other than those that were the subject of study.	*External Validity gets its name from the fact that what's being studied are people, things, and individuals who are outside of the study.* ↪ External Validity gets its name from the fact that what's being studied are people, things, and individuals are outside of the study.
	e→n	Perched on a steep slope, high in the Galilean hills, Safed (known also as Tzfat, Tsfat, Sefat, and Zefat) is a delightful village-town of some 22,000 people.	*Safed is a village that goes by numerous other names.* ↪ Safed is a village goes by numerous other names.

Table 9: **Example perturbation errors** for *lexical* transformations as well as *negative concord* and *wh-dropping* (*relative clause structures*). The HYPOTHESIS was transformed from the original *SAE* ↪ Synthetic AAVE.

	GLUE	VALUE	STRAP
1.	Really, Boris, you are absurdly suspicious.	Really, Boris, you absurdly suspicious.	you look **weird as hell** Boris
2.	In alliance with the Jesuits, he weeded out corrupt clerics and what he considered too soft Umiliati order of Catholic laymen.	In alliance with da Jesuits, he weeded out corrupt clerics and wht he considered too soft-ass Umiliati order of Catholic laymen.	Solo son and a corrupt priest quickly rid his self of corrupt members of the Umiliati community, which he considered **too lenient**.
3.	Jon did not expect better reactions from most of the town folk.	Jon did not expect betta reactions from most of the town folk.	**Most ppl** in the city didnt show any better reaction
4.	No, she replied with perfect truth.	No, she replied wid perfect truth.	**she was very accurate with her answer**
5.	I was bewildered, but I knew that it was no good asking him to explain.	I was bewildered, but I knew dat it was no gud asking his ass to explain.	I’m confused but I know **this can’t help but make me more frustrated**
6.	my wife used to plant a few snow peas i don’t really care for snow peas well she just plants a few for herself	my wife used to plant a few snow peas i don’t really care fa snow no peas well she jus plant a few for herself	She use to put some peas **in the house** that **she don’t really like peas**
7.	The real market for filters isn’t Mom and Dad–it’s Dilbert’s boss.	The real market fa filters ain’t moma and Dad–it’s Dilbert boss.	**Dilbert not a momma or a dad** lmao
8.	Cochran pontificated windily for the camera.	Cochran pontificated windily for da camera.	**Pontificated** loud af in the camera
9.	yeah but then sometimes you know yeah um it just went up the first of this month i’m paying uh seventy nine dollars a week for now	yea but then sometimes u know yeah um it just been went up the first of this month i’m payin uh seventy nine dollars a week for now	**yea, yea, yea, yea, yea, yea, yea, yea, yea, yea**. You know it’s like that sometimes.
10.	Less classic and expensive, more funky and youthful attire is offered at virtually every doorstep along MelroseAvenue between La Brea and Fairfax.	Less classic and expensive, more funky and youthful attire is offered at virtually every doorstep along MelroseAvenue between La Brea and Fairfax.	On **tt-lanta-west** and mailwood in **La Brea-fen,La Brea-fen,La Brea-fen**,the few doors leading to Fairfax
11.	They will lose over $250,000.	They **gon lose** over $250,000.	They will lose over $250,000.
12.	I don’t want them to follow in my footsteps, as I didn’t go to college, but I want them to go.	I don’t want them to follow in my footsteps, as **I ain’t go to no college**, but I want them to go.	I don’t want them to follow in my footsteps, as I didn’t go to college, but I want them to go.
13.	The size is universal	Da **size universal**	The size is universal
14.	It coarsens and inflames dialogue in a way that tends to prevent exposure of which ideas are right and which are wrong.	It **coarsen** and **inflame** dialogue in **a way tend** to prevent exposure of which ideas right and which wrong.	Inadequate discussion of which ideas are correct and which are wrong is likely to prevent exposure to which ideas are correct
15.	Whatever they had done to patch him up hadn’t been very successful, apparently.	Whatever they done to patch **his ass** up **ain’t been** very successful, apparently.	not in a good way they seem to have it down his way

Table 10: A comparison between sampled sentences from the original GLUE MNLI, and VALUE and STRAP transformed AAVE text. We see STRAP’s creative phrase-level transformations (Examples 1-5), but also how STRAP can hallucinate and alter the meaning of the sentence (6-7), remove valuable details (8) and produce erratic behaviors (9-10). VALUE on the other hand appears to better capture AAVE morphosyntax (11-15).