Multi-VALUE: A Framework for Cross-Dialectal English NLP

Caleb Ziems 🔥🌲 **William Held** 🔥🐝 **Jingfeng Yang** a
Jwala Dhamala a **Rahul Gupta** a **Diyi Yang**🌲
🌲Stanford University, 🐝Georgia Institute of Technology, a Amazon
{cziems, diyiy}@stanford.edu, {wheld3}@gatech.edu,
{jddhamal, yjfllpyym, gupra}@amazon.com

Abstract

Dialect differences caused by regional, social, and economic factors cause performance discrepancies for many groups of language technology users. Inclusive and equitable language technology must critically be dialect invariant, meaning that performance remains constant over dialectal shifts. Current systems often fall short of this ideal since they are designed and tested on a single dialect: Standard American English (SAE). We introduce a suite of resources for evaluating and achieving English dialect invariance. The resource is called Multi-VALUE, a controllable rule-based translation system spanning 50 English dialects and 189 unique linguistic features. Multi-VALUE maps SAE to synthetic forms of each dialect. First, we use this system to stress tests question answering, machine translation, and semantic parsing. Stress tests reveal significant performance disparities for leading models on non-standard dialects. Second, we use this system as a data augmentation technique to improve the dialect robustness of existing systems. Finally, we partner with native speakers of Chicano and Indian English to release new gold-standard variants of the popular CoQA task. To execute the transformation code, run model checkpoints, and download both synthetic and gold-standard dialectal benchmark datasets, see `http://value-nlp.org/`.

1 Introduction

> "*[Often, speakers] will not be hampered by the lack of language technology in their local language, but by the lack of support for their variety of the contact language.*"
>
> — **Steven Bird** (2022)

Global contact languages like English will continue to have an outsized impact on commerce, economics, wellbeing, and equity worldwide. English, like any other language, is subject to variation across time (Yang, 2000) and between speakers or speaker groups (Eckert, 2017; Holmes and Meyerhoff, 2008). Rather than focusing on social status or political power (Stewart, 1968; Chambers and Trudgill, 1998), linguists define *dialects* as descriptive sets of correlated *features* common across a group of speakers (Nerbonne, 2009). Current pretraining paradigms employ content filters that can exclude text in English dialects other than Standard American and British (Gururangan et al., 2022), which leads to performance gaps for other varieties. These discrepancies in Natural Language Processing (NLP) cause allocational harms for dialectal speakers in downstream applications (Bender et al., 2021), making dialect robustness a critical need for fair and inclusive language technology.

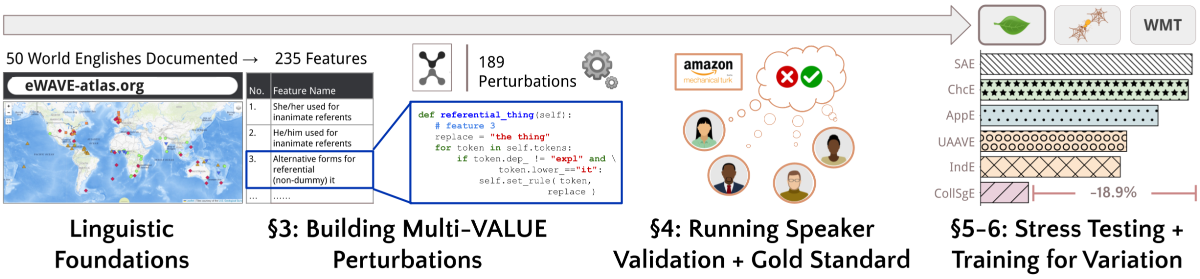
This disparity is clear in a growing body of empirical work on African American English (Ziems et al., 2022; Halevy et al., 2021; Blodgett et al., 2018; Jurgens et al., 2017; Kiritchenko and Mohammad, 2016). However, there does not yet exist a systematic exploration of robustness across multiple Englishes, nor of models' ability to transfer knowledge between varieties with similar features, as in multi-lingual NLP. We need new tools to benchmark and achieve dialect robustness.

We introduce **Multi-VALUE**[1] for English dialect robustness. Our feature-based approach leverages decades of field linguistics research to isolate grammatical constructions (Demszky et al., 2021) that vary in *regional* Englishes (Labov, 1972; Eckert, 1989; Hovy and Yang, 2021). We focus on varieties that (1) are mutually intelligible with Standard American English (SAE); (2) share vocabulary with SAE; and (3) differ from SAE with respect to *morphology* and *syntax*. The third criterion defines the critical axis of variation. The first two criteria ensure that our definition of model robustness aligns with the human ability to understand

🔥Equal contribution.

[1] Multi-VALUE is a **Multi**-dialectal **V**ern**A**cular **L**anguage **U**nderstanding **E**valuation framework (`value-nlp.org`)

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 744–768
July 9-14, 2023 ©2023 Association for Computational Linguistics


WMT
50 World Englishes Documented →
235 Features
eWAVE-atlas.org
189
Perturbations
amazon
mechanical turk
SAE
ChcE
AppE
UAAVE
IndE
CollSgE
-18.9%
Linguistic Foundations
§3: Building Multi-VALUE Perturbations
§4: Running Speaker Validation + Gold Standard
§5-6: Stress Testing + Training for Variation

No.	Feature Name
1.	She/her used for inanimate referents
2.	He/him used for inanimate referents
3.	Alternative forms for referential (non-dummy) it
...	......

```
def referential_thing(self):
    # feature 3
    replace = "the thing"
    for token in self.tokens:
        if token.dep_ != "expl" and \
            token.lower_=="it":
          self.set_rule( token,
                       replace )
```

Figure 1: **The Multi-VALUE pipeline** is grounded in a set of 189 linguistic structures from the dialectology literature (§1). For each structure, we write a perturbation rule to inject it into text (§3). By partnering with native speakers, we validate perturbations and build both synthetic and gold standard benchmarks (§4-5). Finally, we use these resources in (1) stress testing supervised models to reveal dialect disparities and (2) fine-tuning these models on synthetic data to close the performance gap (§6).

other varieties. For example, creoles have their own unique vocabularies and are not easily understood by speakers of other Englishes (Sebba, 1997); they are outside the scope of this study.

First, we provide a controllable **(1) rule-based translation system** for injecting up to 189 features into SAE text. This will allow researchers and practitioners to build *synthetic training data* plus on-demand *dialect stress tests* for nearly any task. We stress test leading models for three challenging tasks and find statistically significant performance gaps. Second, we provide reliable **(2) gold standard benchmarks** for the CoQA task in two widely-spoken varieties: Chicano and Indian English. We find that, by training models on synthetic data, we improve dialectal robustness. Third, we fine-tune and publish **(3) dialect-robust models** on the HuggingFace Hub (Wolf et al., 2020), which can be used directly in downstream applications. Figure 1 demonstrates the full project pipeline.

We recognize five advantages in the Multi-VALUE approach. Our system is

(A) **Interpretable:** supports systematic perturbation analyses

(B) **Flexible:** customized to align with new and evolving dialects by adjusting the *density* of dialectal features, unlike fixed or static datasets.

(C) **Scalable:** allows users to mix and match tasks and dialects at scale without the need for costly human annotation.

(D) **Responsible:** vetted by native speakers to ensure gold standards and synthetic data are dependable for ongoing research.

(E) **Generalizable:** moves the field beyond single-dialect evaluation, which allows researchers to draw more transferrable findings about cross-dialectal NLP performance.

2 Related Work

Dialect Disparity is an issue of equity and fairness (Hovy and Spruit, 2016; Gururangan et al., 2022; Halevy et al., 2021; Blodgett and O'Connor, 2017). There is mounting evidence of dialect disparity in NLP. Hate speech classifiers have known biases against African American English (Davidson et al., 2019; Mozafari et al., 2020; Rios, 2020; Sap et al., 2019; Zhou et al., 2021). Text from regions with a predominantly Black population are more likely to be classified as hate speech (Mozafari et al., 2020; Sap et al., 2019; Davidson et al., 2019). AAVE performance gaps have also been found across a wide range of core NLP tasks like NLI (Ziems et al., 2022), dependency parsing and POS tagging (Blodgett et al., 2018; Jørgensen et al., 2015), plus downstream applications (Lwowski and Rios, 2021). Still, there does not exist a systematic study on cross-dialectal model performance. We aim to fill this gap, expanding the VernAcular Language Understanding Evaluation (VALUE) framework of Ziems et al. (2022). Where VALUE established a uni-dialectal evaluation harness with 11 perturbation rules, Multi-VALUE now supports multi-dialectal evaluation with 189 different perturbations across 50 English dialects. Our empirical study on dialect disparity is also more expansive than prior work as we consider three separate domains: QA, MT, and semantic parsing.

Multilingual NLP studies how to learn common structures that transfer across languages. These strategies may also yield benefits in multi-dialectal settings. Massively multilingual models (Pires et al., 2019; Conneau et al., 2020; Liu et al., 2020;

Xue et al., 2021) exploit the commonalities between many languages at once, rather than merely achieving pairwise transfer (Lin et al., 2019). Additionally, benchmarking across multiple languages can reveal language discrepancies at the modeling level, even without language-specific feature engineering or training data (Bender, 2011; Ravfogel et al., 2018; Ahmad et al., 2019; Tsarfaty et al., 2020). Multi-VALUE aims to bring these advantages to the study of English dialects.

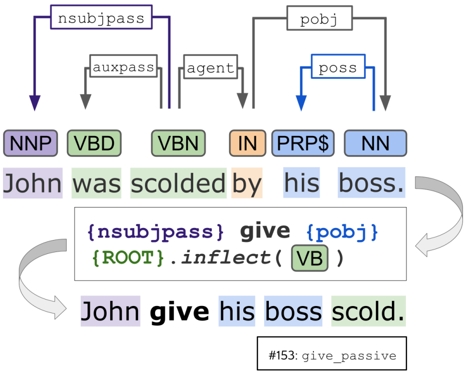


Figure 2: The `give_passive` [#153] perturbation follows this procedure: (1) take the passive subject (`nsubjpass`); (2) insert the verb *give*; (3) insert the object of a preposition that serves as the `agent` of the `ROOT`, (4) insert the `ROOT`, inflected with its base form.

3 Multi-VALUE Perturbations

There is a clear need for dialect robustness (§2). The challenge is that language is subject to *variation* and *change*. This means speakers can contextually modulate the density of features in their grammar, and over time, speakers adopt different features. Shifting language can quickly antiquate training and testing data, and updating such resources can be costly and time-consuming.

In this section, we introduce the first stage of the Multi-VALUE pipeline. We automatically inject structural variation into SAE text using linguistic perturbation rules that alter syntax and morphology but preserve semantics. In this way, perturbations preserve labels. Unlike many black-box translation approaches (Krishna et al., 2020; Sun et al., 2022), label preservation will allow users to convert existing benchmarks directly into dialectal stress tests. Modular, independent perturbation functions give researchers the flexibility to isolate the effects of different features in different combinations.

What distinguishes our work from other syntactic data augmentation methods (Wu et al., 2022) is that our perturbations are grounded in formal language patterns. We operationalize the decades of linguistics research cataloged in the Electronic World Atlas of Varieties of English (eWAVE; Kortmann et al. 2020), a database with 235 features from 75 English varieties, as documented by 87 professional linguists in 175 peer-reviewed publications. eWAVE distinguishes dialects by their unique clusters of linguistic features and the relative *pervasiveness* of each feature.[2] We define a **dialect transformation** as a sequential application of perturbation rules. Decisions to perturb the text follow the eWAVE heuristic probabilities: 100% for obligatory features; 60% for features neither pervasive nor rare; 30% for rare features; 0% for features with no information or an attested absence.

For each rule, we condition the perturbation on morphosyntactic signals from POS tags, noun and verb inflection, and dependency relations using the `spaCy 2.1.0` (Honnibal et al., 2020) and `inflect 5.5.2` libraries. For the *give passive* pertubation above in Figure 2, we search for passive constructions with a past participle `ROOT` (`VBN`), an `nsubjpass` patient, and an agent. We construct the new phrase by inflecting the `ROOT` to its base (`VB`) form and moving it after the entire agentive noun phrase.

Following the eWAVE organizational scheme, we motivate and present our feature perturbations in 12 grammatical categories: (1) Pronouns, (2) Noun Phrases, (3) Tense and Aspect, (4) Mood, (5) Verb Morphology, (6) Negation, (7) Agreement, (8) Relativization, (9) Complementation, (10) Adverbial Subordination, (11) Adverbs and Prepositions, and finally (12) Discourse and Word Order. For a more detailed breakdown, see Appendix A.

Pronouns are critical for tasks like machine translation and summarization, which depend on coreference resolution (Sukthanker et al., 2020). Our pronoun perturbation rules account for linguistic structure and are not merely surface manipulations. For example, we condition on coreference for referential pronouns and on verb frames to identify benefactive datives. In total, we implement 39 of the 47 pronoun features from eWAVE.

Noun Phrases are the focus of fundamental NLP research in semantic role labeling and named entity recognition as well as downstream tasks like

[2] For example, the *give passive* feature #153 is considered pervasive or obligatory in Colloquial Singapore English, while it is rarely observed in Philippine and Tristan da Cunha English, and it is never seen in any other dialect.

sentiment analysis, information extraction, summarization, and question answering (Gildea and Jurafsky, 2000). Multi-VALUE has 31 rules that operate on NP constituents.

Tense and Aspect are two grammatical properties that have to do with time. Together, these categories are known to significantly challenge machine translation (Matusov, 2019; Koehn and Knowles, 2017). With 26 rules, Multi-VALUE introduces different kinds of inflections and auxiliary verbs to indicate when an action, event, or state occurred and how it extends over time.

Mood is important for applications in sentiment analysis and opinion mining, including the detection of biased language (Recasens et al., 2013) and framing strategies in political discourse (King and Morante, 2020; Demszky et al., 2019; Ziems and Yang, 2021). Misunderstandings of modality can also challenge NLU systems on tasks like natural language inference (Gong et al., 2018). There are three modal perturbations in Multi-VALUE.

Verb Morphology is expected to affect model understanding of verb frame semantics (Baker et al., 1998), which could impact performance on semantic role labeling, summarization, and machine translation, among other tasks. We implement 16 related perturbations that change verb suffixes, the forms of verb inflection, and the expression of semantic roles using specialized verbal phrases.

Negation is covered by 16 eWAVE features, 14 of which are implemented in Multi-VALUE. Problems with negation account for many of the failure cases in natural language inference (Hossain et al., 2020) and sentiment analysis (Barnes et al., 2021). Our perturbations introduce negative concord, invariant question tags, and new words for negation.

Agreement is a group of 11 rules which have to do with subject-verb agreement and the omission of copula and auxiliary *be* in different environments. Examples include the invariant present tense in *He speak English* (feature #170), and the existential dummy word in *It's some food in the fridge* (feature #173). Nine of these 11 agreement features are attested in African American English (see Green 2002), which may be linked to the demonstrable performance disparities in AAVE dependency parsing (Blodgett et al., 2018), POS tagging (Jurgens et al., 2017), and NLU tasks (Ziems et al., 2022).

Relativization is a class of perturbations that operates on relativizers, which link relative clauses with their nouns. The purpose of a relative clause is to modify a noun phrase. It's an important construction for NLU because it can contain a presupposition (Joshi and Weischedel, 1977). Our perturbation rules cover all 14 eWAVE features, operating both on individual relativizer words as well as sentence structure to move the relative clause and build correlative constructions, for example.

Complementation is a set of perturbations that turn dependent clauses into the subject or object of the sentence. Like relative clauses, complementation can contain presuppositions and implicatures (Potts, 2002), which are critical for natural language understanding. They can also convey a speaker's degree of certainty (Couso and Naya, 2015), which correlates with biased language and framing strategies. We implement all 11 complementation features that are catalogued in eWAVE.

Adverbial Subordination is a set of perturbations that operate on independent clauses with a "conjunctive adverb." Adverbial conjunctions can express causality (*therefore*), purpose (*so that*), sequence (*then*), contrast (*however*), comparison (*similarly*), and various forms of emphasis (*indeed*). We implement all 5 eWAVE features in this class.

Adverbs and Prepositions are represented by four rules, which can drop prepositions and replace adverbs with their adjectival forms.

Discourse and Word Order has two sides: two discourse features and 9 phrase-based perturbations that move entire constituents in a manner similar to *constituency replacement* (Sutiono and Hahn-Powell, 2022). These rules significantly alter the sentence structure, and in this way radically differ from prior token-level data augmentation techniques like synonym replacement (Wei and Zou, 2019). Phrasal movements include fronting and clefting, subject-auxiliary inversion, and a lack of inversion in questions. We also inject the word *like* to indicate focus or quotation.

4 Scope and Reliability of Multi-VALUE

4.1 Scope

Multi-VALUE's scope is extensive. Out of the 235 features documented in eWAVE, Multi-VALUE covers 189, spanning all 50 recorded English dialects. On average, the feature space for any given

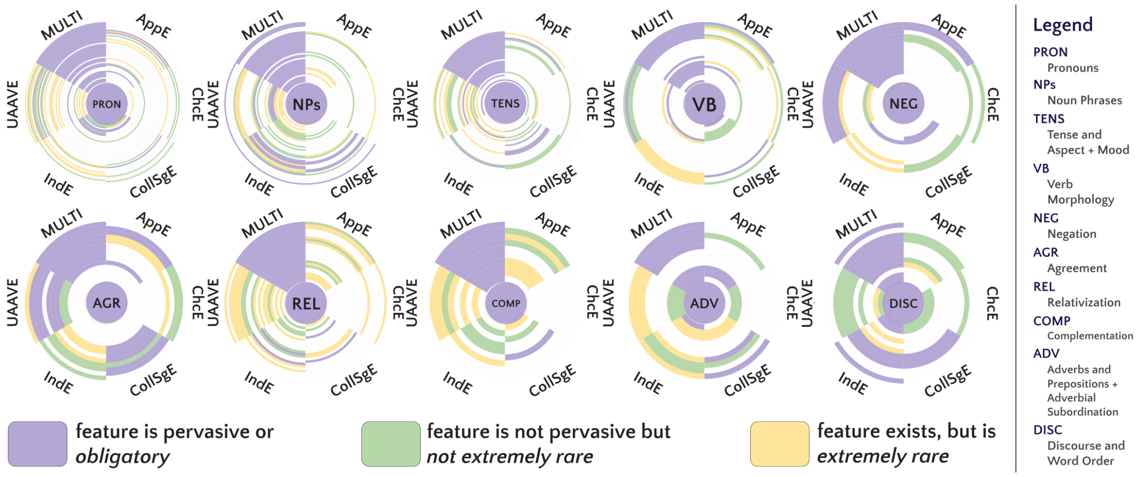


Figure 3: **A comparative distribution of the features in five dialects,** where a dialect is given by a slice of the wheel. Each wheel represents one of the feature groupings in §3 (**Tense and Aspect + Mood** are combined, as are **Adverbial Subordination + Adverbs and Prepositions**). The MULTI slice indicates which of the features are implemented in Multi-VALUE. Rings indicate distinct features and are colored by pervasiveness in each dialect.

dialect is 86.6% implemented, and no dialect is less than 80% implemented (see Appendix A).

4.2 Recruiting Native Speakers for Validation

One key benefit of the Multi-VALUE approach is our ongoing partnership with native speakers to confirm that our theoretically-inspired rules generate plausible and grammatical text. Here, we validate our transformation rules using the linguistic acceptability judgments of native speakers for 10 English dialects.[3] We recruit speakers from Amazon Mechanical Turk and screen them using a Dialect Assessment Survey.[4] This qualification survey ensures that each speaker's empirical language patterns align with the literature on the dialect that they had self-reported. At each turn, the speaker considers a sentence in the target dialect and provides a binary grammaticality judgment about that sentence. Sentences come from published linguistics journals. The survey is efficient[5] as it implements binary search, dynamically selecting the feature that most evenly partitions the space of candidate dialects.

4.3 Validating the Multi-VALUE Pipeline

To validate our perturbation rules, we use the task from Ziems et al. (2022) in which each annotator is shown a pair of sentences: one in SAE, and the other as a dialect transformation: a copy of the first with perturbations corresponding to the target dialect. Annotators see only perturbations corresponding to their native dialect. Annotators mark portions of sentence 1 that were perturbed incorrectly in sentence 2. The interface is shown in in Figure 4 in the Appendix.

FEAT.	ACC.	FEAT.	ACC.	FEAT.	ACC.	FEAT.	ACC.
10	97.4	67	99.1	128	92.7	173	87.5
39	99.7	70	92.9	130	92.9	175	83.3
40	99.8	71	98.8	132	87.7	193	88.7
42	98.1	88	99.4	133	99.5	216	99.7
43	93.2	96	95.5	154	92.9	220	99.4
49	99.6	99	94.7	155	81.8	221	86.7
56	97.3	100	99.9	165	99.1	224	99.5
60	99.6	121	91.8	170	94.9	227	91.2
63	99.0	126	92.3	172	90.0	228	99.8

FEATS.	ACC.
3, 9, 11, 14, 15, 16, 26, 29, 33, 34, 41, 45, 47, 55, 57, 58, 59, 61, 62, 64, 66, 77, 78, 79, 80, 81, 86, 101, 106, 117, 119, 123, 131, 134, 145, 146, 149, 159, 174, 179, 191, 194, 198, 203, 204, 205, 206, 207, 208, 209, 214, 223, 226, 232, 235	100.0

Table 1: **Accuracy of 92 perturbation rules** according to majority vote with at least 5 unique sentence instances. Seventy four rules have >95% accuracy, while sixteen have accuracy in [85,95), and only two are <85% accurate, demonstrating the reliability of our approach.

A group of 72 annotators evaluate a total of 19k sentence pairs, which were drawn from CoQA and other sources. We use CoQA sentences for our Gold Test Sets (§4.4), and for added syntactic diversity, we pull sentences from three `nltk` corpora: Reuters (Rose et al., 2002), Sentiment Analysis (Pang and Lee, 2004) and Movie Reviews (Pang and Lee, 2005). Three annotators evaluate each

[3] Chicano (29 annotators), Colloquial American (13), Indian (11), Appalachian (4), Aboriginal (4), North of England (3), Ozark (3), Southeast American Enclave (3), Urban African American (1), and Black South African English (1).

[4] `https://calebziems.com/resources/value/dialect-quiz.html`

[5] The survey uses an average of 9 questions, but the survey length will depend upon the user's answers.

transformation, marking any pre-highlighted spans where the transformation appeared ungrammatical. This gives us both transformation and perturbation-level evaluations. The majority vote determines the accuracy of the perturbation rule.[6] Perturbation accuracies are given in Table 1. Since there are 55 rules with perfect accuracy, and all perturbation rules achieve above 81%, researchers can feel confident in the linguistic plausibility of the Multi-VALUE transformation pipeline.

4.4 Gold Test Sets

While synthetic Multi-VALUE transformations will be useful for identifying weak points in a model's performance, this does not ensure the model is ready for the real world. We urge practitioners to heavily test user-facing models with numerous in-domain tests. As a first step, we provide reliable gold standard CoQA datasets in Chicano English (ChcE) and Indian English (IndE). Out of 7,983 CoQA questions, our pipeline made changes to 1,726 ChcE questions (21.6%) and 6,825 IndE questions (85.4%). Human annotators considered only transformed questions and provided their own alternative phrasing for transformations they found ungrammatical. Alternatively, they could simply exclude the erroneous perturbations from the question. ChcE had a total transformation accuracy of 82.7% while IndE had 66.1%. The lower IndE accuracy is due to the higher density of features in this dialect. After rephrasing or removing errors, we were left with 1,498 dialect-transformed ChcE questions and 5,289 IndE questions. Together with any unperturbed questions, these gold questions constitute the gold test sets for evaluation in §6.1.

5 Using Multi-VALUE

With our feature rules written (§3) and hand-validated by native speakers (§4), we can use Multi-VALUE to create synthetic data for training dialect-robust models and also for stress testing leading systems on dialect benchmarks. We specifically provide synthetic data for five English dialects: Appalachian (AppE), Chicano English (ChcE), Indian English (IndE), Colloquial Singapore English (CollSgE), and Urban African American English (UAAVE). Three of these dialects are based in the US, where annotators were most abundant for validation, and two are outside the US.

To understand models' ability to transfer knowledge between dialects, we also consider models trained on dialect A and evaluated on dialect B for each dialectal pair (A, B). We can further leverage the strengths of Multi-VALUE as a multi-dialectal augmentation tool by training on a synthetic pseudo-dialect that contains the union of all feature options **(Multi)**. We hypothesize that models trained on multi-(pseudo)-dialectal data will benefit from robustness. While the Multi-VALUE approach could apply over any task with free-form text, we focus on three domains in particular: conversational question answering, semantic parsing, and machine translation. All three are user-facing tasks where language variation may hinder users' access to information, resources, and/or the global economy (Blasi et al., 2022; Faisal et al., 2021).

Conversational Question Answering (CoQA; Reddy et al.2019) is a reading comprehension benchmark with 127k question-answer pairs and 8k passages in seven different genres and domains. We use it because it is a challenging task where dialect-induced errors can compound. The primary challenge is that questions are conversational: they contain coreference and pragmatic relations to prior questions. To transform the publicly available training and development sets, we perturb only questions. This is a natural information-retrieval setting: the user submits queries in a low-resource dialect while the underlying corpus is in SAE.

Semantic Parsing is the task of mapping natural language to formal language. This is a critical skill for dialogue systems, information retrieval, code generation, and other user-facing applications where dialect use is likely. We transform Spider (Yu et al., 2018), a widely-used text-to-SQL benchmark. Again, we transform only the natural language query, leaving both the database tables and the SQL query unchanged to simulate interaction with a dialect user. Unlike the question answering setting where knowledge is encoded in free-text SAE passages, the knowledge and query language in Spider are encoded in formal tables and structured language, both of which are dialect-free. Consequently, any performance discrepancies here will be due to a mismatch between the models' training and testing data rather than a mismatch between the query dialect and that of the knowledge base.

Machine Translation is an interesting test case where challenges can arise from domain mismatch

[6] Accuracy reliably measures strong consensus in the quality of our approach and, unlike kappa scores, it will not suffer from the *prevalence problem* (Eugenio and Glass, 2004).

Model		Test Dialect		
Base	Train Set	SAE	ChcE	IndE
BERT	SAE	77.2	76.7 (-0.5%)	72.3 (-6.7%)$^{-}$
	Multi	76.2 (-1.2%)	76.1 (-1.4%)	75.0 (-2.9%)$^{+-}$
	In-Dialect	77.2	76.5 (-0.9%)	75.1 (-2.7%)$^{+-}$
RoBERTa	SAE	81.8	81.6 (-0.2%)	77.7 (-5.2%)$^{-}$
	Multi	80.6 (-1.5%)$^{-}$	80.5 (-1.6%)$^{-}$	79.7 (-2.7%)$^{+-}$
	In-Dialect	81.8	81.6 (-0.2%)	80.5 (-1.6%)$^{+-}$

Table 2: **Gold QA Evaluation:** F1 Metric on each gold development set of the CoQA benchmark. $^{-}$ and $^{+}$ respectively indicate significantly ($P < 0.05$) worse performance than SAE↦SAE and better performance than SAE↦Dialect by a paired bootstrap test.

(Koehn and Knowles, 2017) due to dialect. We especially anticipate challenges with verb morphology (§3), tense and aspect (§3), and pronouns (§3). We use a standard dataset, WMT19, and evaluate translation from each English Dialect to Chinese, German, Gujurati, and Russian. This simulates a user interacting with translation software using their native dialect.

6 Cross-Dialectal Stress Testing

Here we benchmark current models on dialect variants of the three tasks in §5. For each dataset, we use fixed hyperparameters without early stopping and report all performances on dialect variants of the *evaluation* data, since public test sets are not available for the original datasets. We use the base versions of BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) on dialect variants of the CoQA task, following the Rationale Tagging Multi-Task setup of Ju et al. (2019). For SPIDER, we evaluate BART and T5, since both are near the state of the art in semantic parsing (Xie et al., 2022). For Translation, we evaluate the NLLB Translation Model at two distilled scales: 615M and 1.3B (Costa-jussà et al., 2022). We report hyperparameters and further motivation for model selection in Appendix B.

6.1 Linking Natural and Synthetic Data

While natural data is the gold standard, it is difficult to scale to the number of dialects and tasks we can cover with synthetic data. Thus our broad evaluations are synthetic stress tests. Importantly, we first demonstrate the critical relationship between the gold and synthetic transformations using the gold evaluation sets from §4.4 and the synthetic training data from §5. Table 2 shows the gold standard CoQA results, which should be compared to the synthetic CoQA results in Table 3.

The synthetic stress test results match the gold performance for Chicano English with only small deviations. The Indian English stress tests slightly overestimate the performance drop of an SAE model on Indian English (70.8% synthetic vs. 72.3% natural IndE with BERT; 76.1% vs. 77.7% with RoBERTa). This is expected, as the synthetic feature density may be higher than some annotators naturally use. Synthetic results are a lower bound on performance for a target dialect. For all treatments, the stress tests are directionally correct: treatments that improve performance on the stress test also improve results on the gold data.

Combined with speaker validation of the patterns themselves in §4.3, this shows that Multi-VALUE can be used to reliably measure the effects of modeling choices on dialectal performance.

6.2 Synthetic Stress Tests

We run 3 stress tests to understand worst-case performances on dialect-shifted data across a suite of models and tasks. Evaluation reveals large and statistically significant performance gaps across each task and across all dialects. This highlights, for the first time, the pervasiveness of English dialect disparity beyond any single dialect.

CoQA + Data Augmentation results are shown in Table 3. As predicted in §6.1, Chicano English (ChcE) does not produce a significant drop in performance (-0.7% BERT; -0.3% RoBERTa) since few of its pervasive features are distinct from SAE (the Manhattan distance between feature vectors for ChcE and Colloquial American English is 0.14, or only half the distance as between CollAmE and CollSgE, IndE, and UAAVE.) On the other hand, Singapore English, which is distant from SAE and therefore has many obligatory features, leads to the largest drop (-25.4% BERT; -18.9% RoBERTa). Appalachian, Indian, and Urban African American English each induce significant but smaller RoBERTa performance drops of -3.4%, -7.5%, and -6.7% respectively.

The data augmentation technique described in §5 successfully closes the dialectal performance gap. Across every dialect but Chicano English, we find that we can improve results by training on data that was transformed to the target dialect. Compared to standard RoBERTa, the RoBERTA model trained on **Multi**-dialectal data improves

Model		Test Dialect						
Base	Train Set	SAE	AppE	ChcE	CollSgE	IndE	UAAVE	Average
BERT Base	SAE	77.2	74.4 (-3.8%)$^{-}$	76.6 (-0.7%)	61.5 (-25.4%)$^{-}$	70.8 (-9%)$^{-}$	71.2 (-8.4%)$^{-}$	71.9 (-7.3%)
	AppE	76.3 (-1.1%)	76.4 (-1%)$^{+}$	76.1 (-1.4%)	64.7 (-19.3%)$^{-+}$	72.8 (-6%)$^{-+}$	73.2 (-5.4%)$^{-+}$	73.3 (-5.3%)
	ChcE	76.8 (-0.5%)	74.7 (-3.3%)$^{-}$	76.5 (-0.8%)	63.6 (-21.3%)$^{-+}$	71.6 (-7.8%)$^{-}$	71.4 (-8.1%)$^{-}$	72.4 (-6.5%)
	CollSgE	75.7 (-1.9%)$^{-}$	74.1 (-4.2%)$^{-}$	75.5 (-2.2%)$^{-}$	74.7 (-3.3%)$^{-+}$	73.6 (-4.8%)$^{-+}$	73.4 (-5.1%)	74.5 (-3.6%)
	IndE	76.0 (-1.5%)	75.4 (-2.4%)$^{-}$	75.7 (-2%)$^{-}$	63.2 (-22%)$^{-+}$	75.1 (-2.7%)$^{-+}$	74.1 (-4.1%)$^{-+}$	73.3 (-5.3%)
	UAAVE	76.1 (-1.4%)	75.6 (-2%)$^{-+}$	76.0 (-1.5%)$^{-}$	64.6 (-19.5%)$^{-+}$	74.5 (-3.6%)$^{-+}$	75.3 (-2.5%)$^{-+}$	73.7 (-4.7%)
	Multi	76.2 (-1.2%)	75.6 (-2%)$^{-+}$	76.1 (-1.3%)	73.7 (-4.7%)$^{-+}$	74.9 (-3.1%)$^{-+}$	75.1 (-2.7%)$^{-+}$	75.3 (-2.5%)
	In-Dialect	77.2	76.4 (-1%)$^{+}$	76.5 (-0.8%)	74.7 (-3.3%)$^{-+}$	75.1 (-2.7%)$^{-+}$	75.3 (-2.5%)$^{-+}$	75.9 (-1.7%)
RoBERTa Base	SAE	81.8	79.1 (-3.4%)$^{-}$	81.5 (-0.3%)	68.8 (-18.9%)$^{-}$	76.1 (-7.5%)$^{-}$	76.6 (-6.7%)$^{-}$	77.3 (-5.8%)
	AppE	82.0 (0.3%)	81.8^{+}	81.8	71.2 (-14.9%)$^{-+}$	79.0 (-3.5%)$^{-+}$	79.6 (-2.8%)$^{-+}$	79.2 (-3.2%)
	ChcE	81.7 (-0.1%)	79.3 (-3.1%)$^{-}$	81.5 (-0.4%)	68.8 (-18.9%)$^{-}$	76.5 (-7%)$^{-}$	77.3 (-5.9%)$^{-}$	77.5 (-5.5%)
	CollSgE	81.5 (-0.4%)	80.1 (-2.2%)$^{-}$	81.2 (-0.7%)	80.2 (-2%)$^{-+}$	79.4 (-3%)$^{-+}$	78.7 (-3.9%)$^{-+}$	80.2 (-2%)
	IndE	81.1 (-0.8%)	80.5 (-1.5%)$^{-+}$	80.9 (-1.1%)	67.2 (-21.7%)$^{-}$	80.3 (-1.9%)$^{-+}$	79.2 (-3.3%)$^{-+}$	78.2 (-4.6%)
	UAAVE	81.6 (-0.2%)	81.1 (-0.9%)$^{+}$	81.5 (-0.3%)	69.2 (-18.2%)$^{-}$	79.6 (-2.7%)$^{-+}$	81.1 (-0.9%)$^{+}$	79.0 (-3.5%)
	Multi	80.6 (-1.5%)$^{-}$	80.4 (-1.7%)$^{-+}$	80.5 (-1.6%)$^{-}$	78.5 (-4.2%)$^{-+}$	79.7 (-2.7%)$^{-+}$	80.0 (-2.2%)$^{-+}$	80.0 (-2.3%)
	In-Dialect	81.8	81.8^{+}	81.5 (-0.4%)	80.2 (-2%)$^{-+}$	80.3 (-1.9%)$^{-+}$	81.1 (-0.9%)$^{+}$	81.1 (-0.9%)

Table 3: **Dialect QA Stress Test:** F1 Metric on each VALUE-transformed development set of the CoQA benchmark. $^{-}$ and $^{+}$ indicate significantly ($P < 0.05$) worse performance than SAE$\mapsto$SAE and better performance than SAE$\mapsto$Dialect by a paired bootstrap test.

Evaluation		Input Dialect						
Model	Metric	SAE	AppE	ChcE	CollSgE	IndE	UAAVE	Avg.
BART-base	Exact Match ACC	49.3	45.2 (-8.3%)$^{-}$	48.5 (-1.6%)$^{-}$	41.9 (-15.0%)$^{-}$	40.5 (-17.8%)$^{-}$	45.0 (-8.7%)$^{-}$	45.1 (-8.5%)
	Execution ACC	51.0	47.3 (-7.3%)$^{-}$	50.3 (-1.4%)	44.1 (-13.5%)$^{-}$	42.3 (-17.1%)$^{-}$	46.1 (-9.6%)$^{-}$	46.9 (-8.0%)
BART-large	Exact Match ACC	67.9	63.6 (-6.3%)$^{-}$	65.5 (-3.5%)$^{-}$	60.3 (-11.2%)$^{-}$	61.2 (-9.9%)$^{-}$	62.3 (-8.2%)$^{-}$	63.5 (-6.5%)
	Execution ACC	70.5	65.2 (-7.5%)$^{-}$	68.2 (-3.3%)$^{-}$	63.0 (-10.6%)$^{-}$	62.8 (-10.9%)$^{-}$	64.5 (-8.5%)$^{-}$	65.4 (-7.2%)
T5-base	Exact Match ACC	58.7	54.3 (-7.5%)$^{-}$	57.4 (-2.2%)$^{-}$	50.0 (-14.8%)$^{-}$	49.1 (-16.4%)$^{-}$	53.1 (-9.5%)$^{-}$	53.8 (-8.3%)
	Execution ACC	59.8	56.0 (-6.4%)$^{-}$	58.5 (-2.2%)$^{-}$	51.6 (-13.7%)$^{-}$	51.3 (-14.2%)$^{-}$	54.6 (-8.7%)$^{-}$	55.3 (-7.5%)
T5-3b	Exact Match ACC	71.7	65.3 (-8.9%)$^{-}$	69.7 (-2.8%)$^{-}$	60.7 (-15.3%)$^{-}$	62.9 (-12.3%)$^{-}$	68.5 (-4.5%)$^{-}$	66.5 (-7.3%)
	Execution ACC	75.6	69.3 (-8.3%)$^{-}$	73.4 (-2.9%)$^{-}$	64.9 (-14.2%)$^{-}$	66.5 (-12.0%)$^{-}$	66.9 (-11.5%)$^{-}$	69.4 (-8.2%)

Table 4: **Dialect SPIDER Stress Test:** Evaluation on each VALUE-transformed evaluation set of the SPIDER benchmark. We finetune BART and T5 on SPIDER and evaluate for both Exact Match and Execution accuracy. $^{-}$ indicates a significant performance drop ($P < 0.05$) compared to SAE performance by a bootstrap test.

average cross-dialectal performance by 2.7 points. However, multi-dialectal training causes a drop of 1.2 points on SAE, reminiscent of interference in multilingual models (Wang et al., 2019, 2020).

We performed a **Qualitative Error Analysis** on 30 errors for each transformed dialect. In each error, models trained on SAE flipped from a correct answer in SAE to an incorrect answer in one of the dialect-transformed COQA sets. Fully validated perturbations in tense, inflection, plural marking, phrasal order, and the deletion of pragmatically-recoverable pronouns, prepositions, and auxiliaries all lead to significant errors. As expected, these errors can cascade down the conversation, leading to model failure on later *unperturbed* questions as well. In some cases, erroneous answers still belong to the correct class, like flipping from *yes* to *no* in the presence of *negative concord*. Suprisingly, transformations also frequently cause the model to respond with an erroneous *class*, like giving a noun phrase or prepositional phrase to a yes/no question under perturbations like *clefting* and the omission of auxiliary *did*, *is*, and *wh*-words.

Our analysis also suggests that the noticeably larger drop in performance on Singapore English might be largely due to the higher density of two perturbation types: preposition omissions (feature #198), and the *one relativizer* (feature #216). Future work can use perturbation analyses (Ziems et al., 2022) to quantitatively measure these sources of error.

Semantic Parsing Table 4 shows that SAE models significantly underperform on all dialectal stress tests, both in terms of Exact Match Accuracy and Execution Accuracy. For both BART and T5, the largest performance gaps appear when we test on the two non-American dialects, CollSgE and IndE (-15.3% and -12.3% exact match accuracy for T5-3b). The semantic parsing performance gaps here are as large as those in conversational question answering. This supports our claim that the discrepancies are caused by model mismatch, rather

Evaluation		Source Dialect						
# Param.	Target	SAE	AppE	ChcE	CollSgE	IndE	UAAVE	Avg.
615M	Chinese	22.5	21.2 (-6.1%)$^{-}$	21.7 (-3.6%)$^{-}$	17.0 (-24.5%)$^{-}$	18.7 (-16.8%)$^{-}$	19.8 (-12.3%)$^{-}$	20.1 (-10.6%)
	German	39.6	34.3 (-13.41%)$^{-}$	37.8 (-4.65%)$^{-}$	22.3 (-43.60%)$^{-}$	26.8 (-32.32%)$^{-}$	30.5 (-23.1%)$^{-}$	31.9 (-19.5%)
	Gujurati	21.7	18.6 (-14.5%)$^{-}$	20.4 (-6.2%)$^{-}$	13.4 (-38.4%)$^{-}$	16.6 (-23.4%)$^{-}$	17.2 (-20.7%)$^{-}$	18.0 (-17.2%)
	Russian	27.8	24.6 (-11.4%)$^{-}$	26.7 (-4.0%)$^{-}$	17.2 (-38.1%)$^{-}$	20.8 (-25.4%)$^{-}$	21.7 (-22.1%)$^{-}$	23.1 (-16.8%)
1.3B	Chinese	23.2	21.5 (-7.4%)$^{-}$	22.5 (-3.3%)	17.8 (-23.5%)$^{-}$	19.4 (-16.6%)$^{-}$	19.8 (-15.0%)$^{-}$	20.7 (-11.0%)
	German	42.6	37.5 (-11.9%)$^{-}$	40.6 (-4.6%)$^{-}$	25.3 (-40.6%)$^{-}$	29.4 (-31.0%)$^{-}$	34.2 (-19.7%)$^{-}$	34.9 (-18.0%)
	Gujurati	24.0	20.7 (-13.8%)$^{-}$	22.9 (-4.5%)$^{-}$	15.5 (-35.4%)$^{-}$	18.5 (-22.8%)$^{-}$	19.7 (-17.8%)$^{-}$	20.2 (-15.7%)
	Russian	31.7	28.5 (-10.1%)$^{-}$	30.3 (-4.4%)	20.3 (-36.0%)$^{-}$	24.5 (-22.6%)$^{-}$	25.3 (-20.2%)$^{-}$	26.7 (-15.5%)

Table 5: **Dialect Translation Stress Test:** SacreBLEU Score (Post, 2018) on each VALUE-transformed validation set of the WMT19 benchmark at 2 distilled scales of the NLLB Translation model (Costa-jussà et al., 2022). $^{-}$ indicates a significant performance drop ($P < 0.05$) compared to SAE performance by a bootstrap test.

than solely a mismatch between the dialect of the question and that of the knowledge base.

Machine Translation stress test results are shown in Table 5. Except for ChcE, performance drops significantly across all dialects for each language. Interestingly, the size of the average dialectal performance gap is higher when the target language is structurally *more similar* to English: the largest average drop is from English↦German (-19.5% on 615M; -18.0% on 1.3B) and the smallest average drop is from English↦Chinese (-10.6% on 615M; -11.0% on 1.3B). This result cannot be explained simply as a reflection of the model's SAE translation performance. If it were, we might expect a smaller performance gap for Gujurati, a low-resource Indo-European language, since it has low SAE translation performance (21.7 SacreBLEU on 615M), but in fact, English↦Gujurati has the second *largest* dialectal translation performance gap (-17.2% on 615M; -15.7% on 1.3B). Our explanation is that Gujurati has syntax that is more similar to English.

Despite both the 1.3B and 615M NLLB models being distilled from the same larger model, we see that the dialectal gap is smaller for German, Gujurati, and Russian. This suggests that model compression may affect low-resource dialects more heavily than SAE, similar to multi-lingual findings for low-resource languages (Ahia et al., 2021).

7 Conclusion

In this work, we introduced Multi-VALUE – a dialect robustness evaluation framework that is interpretable, flexible, scalable, responsible, and generalizable. The rule-based methods form a transparent syntactic translation system that can flexibly adjust to the shifting feature space of living dialects. Additionally, the transformation rules are reliably sourced from over a decade of linguistics literature and vetted by native speakers. After showing that these transformations predict human-translated dialect benchmark performance, we used them to build dialect benchmarks and training data at scale, without the need for additional annotation efforts. By training and evaluating in a cross-dialectal manner, we demonstrated how Multi-VALUE can be used for more generalizable findings about model performance and dialect transferability.

Multi-VALUE can facilitate a wide range of NLP tasks and applications, such as measuring the relationships between dialect similarity and generalization performance, the scaling laws of dialect disparity, as well as inspiring algorithms on better dialect transfer. Overall, we anticipate that Multi-VALUE will continue to support the development of more fair and equitable language technologies.

8 Limitations

Lexical variation is not our focus because it is not well-described by systematic, scalable, and generalizable rules. One can derive lexical distributions from data, but many low-resource dialects lack corpora on which to base these insights. This is an important problem for future research.

Multi-VALUE's strength is its extensive coverage of English morphosyntactic patterns that have been documented in eWAVE by over 80 linguists. Such comprehensive resources are not available for other languages, but we encourage continued collaborations between computer scientists and linguists to build these resources for dialect-robust NLP systems across languages. As it stands, the current iteration of Multi-VALUE provides global value by serving a global contact language, English, and its 50 most documented varieties.

Despite the scope and precision of eWAVE for

English, its catalog ultimately derives from linguists' oral interviews with native speakers, and here we can identify some additional limitations. First, the orthographic conventions that linguists use to encode spoken dialect may not always align with the speakers' own writing conventions and usage. Second, our approach can only cover the variation that linguists observe frequently enough to document, and in canonical forms in which they are documented. This means we may not fully capture variation within each feature.

Finally, dialects should not be treated like deterministic speech patterns, but rather like a range of grammatical options or switches that may be turned on and off and adjusted for frequency in various social and personal contexts. Dialects do not always fit into nicely prescribed categories.

9 Ethical Considerations

This work makes use of human subjects for annotation. All procedures were subject to ethical review and were approved by the authors' institution. Consent was gathered in accordance with the authors' institution guidelines and annotators had access to a data use statement when giving consent.

The purpose of Multi-VALUE is to provide tools which enable researchers and practitioners to understand and mitigate dialectal bias in their models. We will release these tools responsibly, ensuring that users sign a Data Use Agreement that forbids the use of Multi-VALUE for deception, impersonation, mockery, discrimination, hate speech, targeted harassment and cultural appropriation.

In the agreement, researchers and practitioners will also acknowledge the Limitations of this work (§8), that Multi-VALUE may not fully or accurately represent the natural usage patterns of all sub-communities of speakers. Multi-VALUE is designed to be easily updatable and configurable such that it can be extended by and for specific sub-communities and updated as dialects evolve over time.

Acknowledgements

We are thankful to the members of SALT Lab for their helpful feedback on the draft. Caleb Ziems is supported by the NSF Graduate Research Fellowship under Grant No. DGE-2039655. Part of this work was funded by an Amazon Faculty Research Award on Alexa Fairness in AI to DY.

References

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 2021. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3316–3333, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Jeremy Barnes, Erik Velldal, and Lilja Øvrelid. 2021. Improving sentiment analysis with multi-task learning of negation. *Natural Language Engineering*, 27(2):249–269.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829.

Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505.

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

Jack K Chambers and Peter Trudgill. 1998. *Dialectology*. Cambridge University Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv preprint*, abs/2207.04672.

María José López Couso and Belén Méndez Naya. 2015. Epistemic/evidential markers of the type verb+ complementizer: Some parallels from english and romance. In *New directions in grammaticalization research*, pages 93–120. John Benjamins.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. 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 2970–3005, Minneapolis, Minnesota. Association for Computational Linguistics.

Dorottya Demszky, Devyani Sharma, Jonathan Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. 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.

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.

Penelope Eckert. 1989. *Jocks and burnouts: Social categories and identity in the high school*. Teachers college press.

Penelope Eckert. 2017. Age as a sociolinguistic variable. *The handbook of sociolinguistics*, pages 151–167.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: A second look. *Computational linguistics*, 30(1):95–101.

Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2021. SD-QA: Spoken dialectal question answering for the real world. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3296–3315, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Gildea and Daniel Jurafsky. 2000. Automatic labeling of semantic roles. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 512–520, Hong Kong. Association for Computational Linguistics.

Yichen Gong, Heng Luo, and Jian Zhang. 2018. Natural language inference over interaction space. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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

Suchin Gururangan, Dallas Card, Sarah K Drier, Emily K Gade, Leroy Z Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A Smith. 2022. Whose language counts as high quality? measuring language ideologies in text data selection. *ArXiv preprint*, abs/2201.10474.

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.

Janet Holmes and Miriam Meyerhoff. 2008. *The handbook of language and gender*. John Wiley & Sons.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

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.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18, Beijing, China. Association for Computational Linguistics.

Aravind K. Joshi and Ralph Weischedel. 1977. Computation of a subclass of inferences: Presupposition and entailment. *American Journal of Computational Linguistics*, pages 1–54. Microfiche 63.

Ying Ju, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang, and Yunfeng Liu. 2019. Technical report on conversational question answering. *ArXiv preprint*, abs/1909.10772.

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.

Liza King and Roser Morante. 2020. Must children be vaccinated or not? annotating modal verbs in the vaccination debate. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5730–5738, Marseille, France. European Language Resources Association.

Svetlana Kiritchenko and Saif Mohammad. 2016. The effect of negators, modals, and degree adverbs on sentiment composition. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 43–52, San Diego, California. Association for Computational Linguistics.

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

Bernd Kortmann, Kerstin Lunkenheimer, and Katharina Ehret, editors. 2020. *eWAVE*.

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.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Brandon Lwowski and Anthony Rios. 2021. The risk of racial bias while tracking influenza-related content on social media using machine learning. *Journal of the American Medical Informatics Association*, 28(4):839–849.

Evgeny Matusov. 2019. The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland. European Association for Machine Translation.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.

John Nerbonne. 2009. Data-driven dialectology. *Language and Linguistics Compass*, 3(1):175–198.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. 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.

Christopher Potts. 2002. The lexical semantics of parenthical-as and appositive-which. *Syntax*, 5(1):55–88.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer nlp. *ArXiv preprint*, abs/2205.12586.

Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. 2018. Can LSTM learn to capture agreement? the case of Basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107, Brussels, Belgium. Association for Computational Linguistics.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria. 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.

Anthony Rios. 2020. Fuzze: Fuzzy fairness evaluation of offensive language classifiers on african-american english. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 881–889. AAAI Press.

Tony Rose, Mark Stevenson, and Miles Whitehead. 2002. The Reuters corpus volume 1 -from yesterday's news to tomorrow's language resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Mark Sebba. 1997. *Contact languages: Pidgins and creoles*. Bloomsbury Publishing.

William Stewart. 1968. A sociolinguistic typology for describing national multilingualism. *Readings in the Sociology of Language*, 3:531–545.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.

Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2022. Dialect-robust evaluation of generated text.

Arie Sutiono and Gus Hahn-Powell. 2022. Syntax-driven data augmentation for named entity recognition. In *Proceedings of the First Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning*, pages 56–60, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Maggie Tallerman. 2019. *Understanding syntax*. Routledge.

Reut Tsarfaty, Dan Bareket, Stav Klein, and Amit Seker. 2020. From SPMRL to NMRL: What did we learn (and unlearn) in a decade of parsing morphologically-rich languages (MRLs)? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7396–7408, Online. Association for Computational Linguistics.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. 2019. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11293–11302.

Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhengxuan Wu, Isabel Papadimitriou, and Alex Tamkin. 2022. Oolong: Investigating what makes crosslingual transfer hard with controlled studies. *ArXiv preprint*, abs/2202.12312.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *ArXiv preprint*, abs/2201.05966.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Charles D Yang. 2000. Internal and external forces in language change. *Language variation and change*, 12(3):231–250.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding dialect disparity in NLU. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland. Association for Computational Linguistics.

Caleb Ziems and Diyi Yang. 2021. To protect and to serve? analyzing entity-centric framing of police violence. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 957–976, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Implementation Details

In Table 6, we give summary statistics for the number of features implemented for each of the 50 focus dialects, and the number of such features which were validated by native speakers. On average, the feature space for any given dialect is 86.6% implemented, and no dialect is less than 80% implemented. The reason we did not cover 100% of the eWAVE catalogue is that some features operate with information unavailable to us. For example, in SAE, aspect and mood may not be marked morphosyntactically; these features are outside the scope of current methods. Similarly, we are unable to inject distinct pronouns for groups of 2, 3, and 4+ people [#37], as group size information may not be contained in the focus utterance.

In Tables 7-18, we detail our Multi-VALUE implementations with an enumeration of our implemented dialects and features and examples of each. In the VAL ACC. column we give the validation accuracy (§4.3) as well as tags **ChcE** or **IndE** to indicate if the feature appears in the gold Chicano or Indian English CoQA dataset respectively.

A.1 Pronouns

There are 47 pronoun features in eWAVE, and we cover 39 of them (83%). While simple regular expressions can cover some pronoun mappings, this is not always possible since English maps the same surface forms to different grammatical roles.[7] We overcome this problem by conditioning rules on pronouns' syntactic roles. We also condition on coreference for referential pronouns [29], and on verb frames to identify benefactive datives [9]. Furthermore, we swap the morphology of possession [20], change reflexive marking [11-16], swap animate pronouns for inanimate objects [1-2], and include additional elements like reduplication [40]. In summary, our pronoun perturbation rules account for linguistic structure and are not merely surface manipulations.

A.2 Noun Phrases

Among our 31 noun phrase perturbations, we regularize or modify plural morphology [49] and comparison strategies [80], to drop or modify articles [60], construct phrases for possession [75], and adjust the tree adjoining order to create adjective postfixes [87].

A.3 Tense and Aspect

Tense and aspect perturbations include alternative inflections and auxiliaries to mark tense [117], including immediate vs. distant future [119], as well as perfect aspect [99].

A.4 Mood

Multi-VALUE includes perturbations that inject double modals [121] and quasi-modals [126], change verb inflections under modal scope [123], and introduce auxiliaries to mark the sequential or irrealis mood [106].

A.5 Verb Morphology

Verb morphology features include levelling certain finite and non-finite verb forms [130] adding suffixes for transitive verbs [143], and building *serial verb phrases* (Tallerman, 2019) to mark passive constructions [153], indirect objects [148], or the movement of direct objects [150].

A.6 Negation

Multi-VALUE includes rules for building phrases with negative concord [154], and forms of negation with the negation words *never*, *no*, *not*, *no more* or *ain't*, as well as special invariant tags for questions [166].

A.7 Agreement

We implement the invariant present tense [170], as well as the existential dummy *it* [173].

A.8 Relativization

These perturbations modify the form of the relativizer [186-190], as well as drop [193] or introduce new shadow pronouns [194], such as double relativizers [191] and phrasal forms [192]. Our perturbations also operate on the sentence structure by forming correlative constructions [196], deleting stranded prepositions [198], and moving the relative clause before the head noun [199].

A.9 Complementation

These perturbations can change the form of the complementizer [200, 201], delete [208, 209] or introduce additional complementizer words [203, 204], build existential constructions from complementizer phrases [205, 206], and modify the verb in the non-finite clause complement [210].

[7] For example, *her* is both the accusative in "give it to her" and the noun modifier in "her cart," while the masculine pronouns in "give it to him" and "his cart" differ. This problem was observed but not solved in the rule-based perturbation augmentation of Qian et al. (2022).

A.10 Adverbial Subordination

Our perturbation rules introduce clause-final conjunctions [211, 212] and double conjuctions [214, 215], and remove the adverb in verb-chaining constructions [213], which together represent the five adverbial subordination features in eWAVE.

A.11 Adverbial Prepositions

In this section, we drop prepositions [216] and replace adverbs with their adjectival forms [220, 221]. We also include the word *too* as a qualifier [222].

A.12 Discourse and Word Order

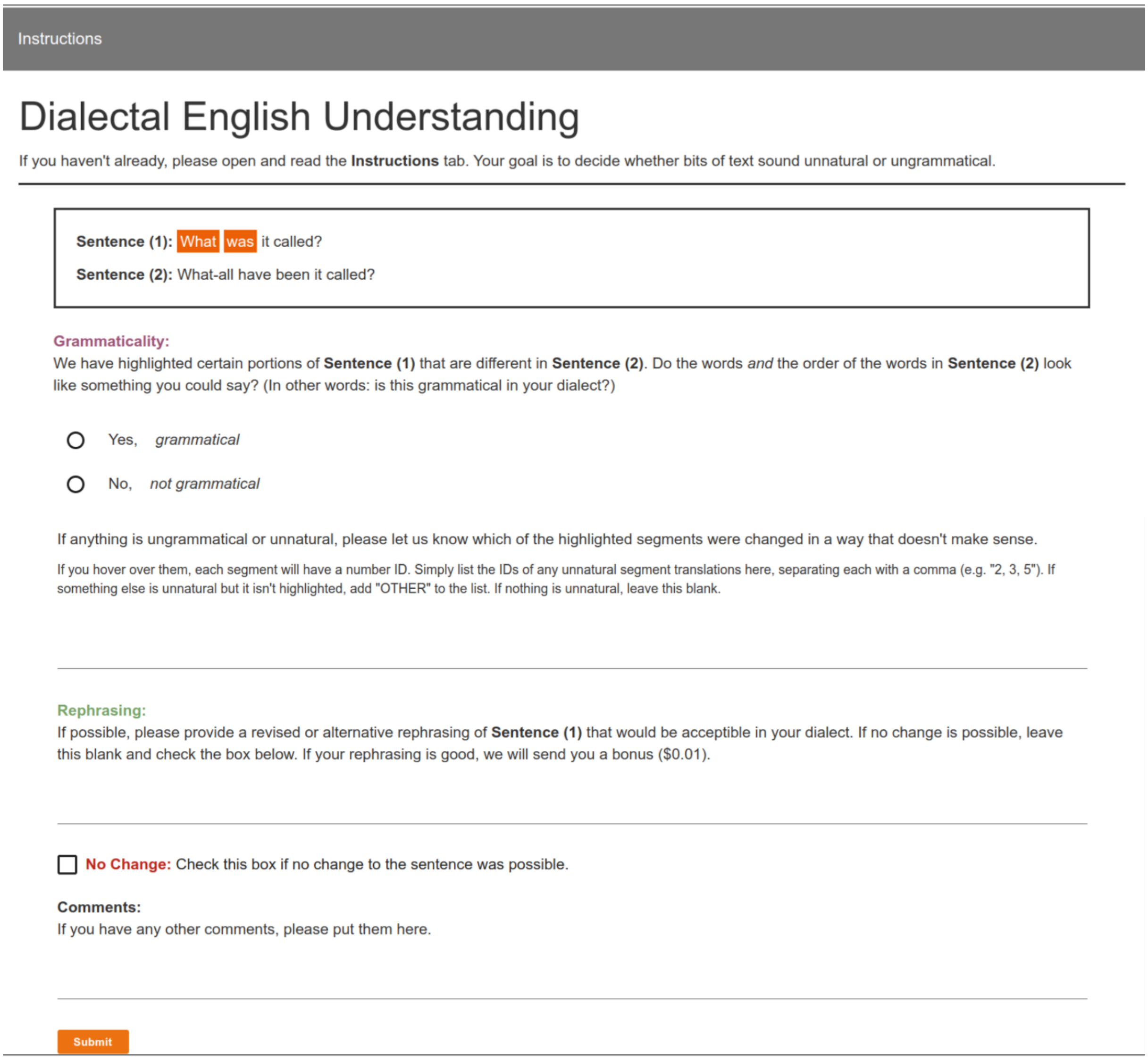
In discourse, we insert the word *like* as a focus [234] or quotation marker [235]. Our phrase-based perturbations include fronting and clefting [223, 224], subject–auxiliary inversion in both negation phrases [226] and indirect questions [227], and a lack of inversion in certain questions [228, 229]. These rules significantly alter the sentence structure, and in this way radically differ from prior token-level data augmentation techniques like synonym replacement (Wei and Zou, 2019). Our approach here is most similar to *constituency replacement* (Sutiono and Hahn-Powell, 2022).

B Models & Hyperparameters

CoQA We use the base versions of BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) on dialect variants of the CoQA task, following the Rationale Tagging Multi-Task setup of Ju et al. (2019) to adapt these models to the CoQA setup which includes *Yes, No,* and *Unknown* responses in addition to extractive answers. Each model was trained on an Nvidia GeForce RTX 2080 Ti for approximately 6 hours. For each model and dialect, we fine-tune using AdamW (Loshchilov and Hutter, 2019) for 2 epochs with a batch size of 16 and a learning rate $3e - 5$.

Semantic Parsing. Following Xie et al. (2022), for T5-base we adopted the AdamW optimizer, while Adafactor was used for T5-3B and the two BART models. We used NVIDIA A100 to train these models with T5-3b, BART-large, T5-base, and BART-base using 8 GPUs for 52 hours, 4 GPUs for 32 hours, 4 GPUs for 4 hours, 4 GPU for 13 hours respectively. We set the learning rate at 5e-5 for T5 models and 1e-5 for BARTs. We fixed the batch size at 32 when fine-tuning T5-BASE and BARTs. As for the extremely large T5-3B, we configured a batch size of 64 to speed up convergence and utilised DeepSpeed to save memory. Linear learning rate decay was used for all models.

Machine Translation. We evaluate the NLLB Translation Model at two distilled scales: 615M and 1.3B (Costa-jussà et al., 2022). Evaluation was done on an Nvidia GeForce RTX 2080 Ti and takes less than 10 minutes. The NLLB model is designed for many-to-many translation with low-resource language communities and is trained on a large corpus mined from the internet, rather than exclusively human aligned translations. We choose this model to give us an estimate of the performance of large scale translation products available to users.



Instructions

Dialectal English Understanding

If you haven't already, please open and read the **Instructions** tab. Your goal is to decide whether bits of text sound unnatural or ungrammatical.

Sentence (1): What was it called?

Sentence (2): What-all have been it called?

Grammaticality:
We have highlighted certain portions of **Sentence (1)** that are different in **Sentence (2)**. Do the words *and* the order of the words in **Sentence (2)** look like something you could say? (In other words: is this grammatical in your dialect?)

○ Yes, *grammatical*

○ No, *not grammatical*

If anything is ungrammatical or unnatural, please let us know which of the highlighted segments were changed in a way that doesn't make sense.

If you hover over them, each segment will have a number ID. Simply list the IDs of any unnatural segment translations here, separating each with a comma (e.g. "2, 3, 5"). If something else is unnatural but it isn't highlighted, add "OTHER" to the list. If nothing is unnatural, leave this blank.

Rephrasing:
If possible, please provide a revised or alternative rephrasing of **Sentence (1)** that would be acceptible in your dialect. If no change is possible, leave this blank and check the box below. If your rephrasing is good, we will send you a bonus ($0.01).

□ **No Change:** Check this box if no change to the sentence was possible.

Comments:
If you have any other comments, please put them here.

Figure 4: **MTurk Validation Task Interface.** Workers consider sentence pairs and evaluate whether the synthetic sentence is an acceptable dialectal form of the gloss given by the natural SAE sentence.

Abbr	# Feat.	% Feat.	# Val.	% Val.	Dialect
AborE	89	83.2%	57	53.3%	Aboriginal English
AppE	65	85.5%	51	67.1%	Appalachian English
AusE	54	90.0%	40	66.7%	Australian English
AusVE	47	83.9%	34	60.7%	Australian Vernacular English
BahE	107	83.6%	70	54.7%	Bahamian English
BlSAfE	95	88.0%	71	65.7%	Black South African English
CamE	76	87.4%	62	71.3%	Cameroon English
CFE	49	90.7%	39	72.2%	Cape Flats English
ChIsE	47	94.0%	33	66.0%	Channel Islands English
ChcE	30	93.8%	28	87.5%	Chicano English
CollAmE	57	83.8%	44	64.7%	Colloquial American English
CollSgE	67	89.3%	52	69.3%	Colloquial Singapore English (Singlish)
EAAVE	96	89.7%	61	57.0%	Earlier African American Vernacular English
EA	46	85.2%	32	59.3%	East Anglian English
FlkE	44	89.8%	30	61.2%	Falkland Islands English
FijiE	39	88.6%	36	81.8%	Acrolectal Fiji English
CollFijiE	95	85.6%	68	61.3%	Pure Fiji English (basilectal FijiE)
GhE	58	92.1%	49	77.8%	Ghanaian English
HKE	74	91.4%	61	75.3%	Hong Kong English
IndE	90	90.0%	82	82.0%	Indian English
InSAfE	75	83.3%	58	64.4%	Indian South African English
IrE	75	81.5%	54	58.7%	Irish English
JamE	69	88.5%	47	60.3%	Jamaican English
KenE	50	90.9%	45	81.8%	Kenyan English
LibSE	86	84.3%	58	56.9%	Liberian Settler English
MalE	68	89.5%	57	75.0%	Malaysian English
MaltE	72	86.7%	59	71.1%	Maltese English
ManxE	55	83.3%	40	60.6%	Manx English
NZE	44	88.0%	37	74.0%	New Zealand English
NfldE	84	85.7%	53	54.1%	Newfoundland English
NigE	45	88.2%	37	72.5%	Nigerian English
North	77	85.6%	47	52.2%	English dialects in the North of England
O&SE	30	81.1%	19	51.4%	Orkney and Shetland English
OzE	56	86.2%	43	66.2%	Ozark English
PakE	48	87.3%	42	76.4%	Pakistani English
PhilE	92	85.2%	71	65.7%	Philippine English
RAAVE	136	82.9%	88	53.7%	Rural African American Vernacular English
ScE	44	80.0%	30	54.5%	Scottish English
SEAmE	108	80.6%	75	56.0%	Southeast American enclave dialects
SLkE	29	82.9%	23	65.7%	Sri Lankan English
StHE	113	85.0%	78	58.6%	St. Helena English
SE	46	93.9%	33	67.3%	English dialects in the Southeast of England
SW	73	89.0%	46	56.1%	English dialects in the Southwest of England
TznE	41	93.2%	35	79.5%	Tanzanian English
TdCE	92	82.9%	64	57.7%	Tristan da Cunha English
UAAVE	118	83.7%	79	56.0%	Urban African American Vernacular English
UgE	65	86.7%	52	69.3%	Ugandan English
WelE	76	80.9%	53	56.4%	Welsh English
WhSAfE	41	83.7%	35	71.4%	White South African English
WhZimE	61	88.4%	46	66.7%	White Zimbabwean English

Table 6: **Multi-VALUE Implemented Dialects.** We've implemented 50 English dialects as shown in this table. We list the number of implemented features (# Feat), the proportion of that dialect's catalogued eWAVE features implemented (% Feat), the number of validated features (# Val), and the proportion of that dialect's catalogued eWAVE features validated (% Val). All dialects are at or above 80% implemented and above 51.4% validated. Gold **ChcE** and **IndE** indicate that we also release a Gold CoQA dev set in Chicano and Indian English.

	Function	SAE	Transform	Val Acc.
1	she_inanimate_objects	It's a good bike	She's a good bike	
2	he_inanimate_objects	The driver's license? She wasn't allowed to renew it right?	The driver's license? She wasn't allowed to renew 'im right?	
3	referential_thing	Christmas dinner? I think it's better to wait until after she's had it.	Christmas dinner? I think it's better to wait until after she's had the thing.	100.0
4	pleonastic_that	It's raining.	Thass raining.	
5	em_subj_pronoun	This old woman, she started packing up.	This old woman, 'em started packing up.	
6	em_obj_pronoun	We just turned it around.	We just turned 'im around.	
7	me_coordinate_subjects	Michelle and I will come too.	Me and Michelle will come too.	
8	myself_coordinate_subjects	My husband and I were late.	My husband and myself were late.	
9	benefactive_dative	I have to get one of those!	I have to get me one of those!	ChcE 100.0
10	no_gender_distinction	Susan is a nurse but she does not like to put drips on patients.	Susan is a nurse but he does not like to put drips on patients.	IndE 97.4
11	regularized_reflexives	He hurt himself.	He hurt hisself.	100.0
12	regularized_reflexives_object_pronouns	I'll do it myself.	I'll do it meself.	
13	regularized_reflexives_aave	They look after themselves.	They look after theyselves.	
14	reflex_number	We cannot change ourselves.	We cannot change ourself.	IndE 100.0
15	absolute_reflex	and he and the bull were tuggin' and wrestlin'	and himself and the bull were tuggin' and wrestlin'	IndE 100.0
16	emphatic_reflex	They brought it by themselves.	They brought it by their own self.	ChcE 100.0
18	my_i	my book	I book	
19	our_we	our farm	we farm	
20	his_he	his book	he book	
21	their_they	their book	they book	
22	your_you	your book	you book	
23	your_yalls	Where are your books?	Where are y'all's books?	
24	his_him	his book	him book	
25	their_them	their book	them book	
26	my_me	my book	me book	100.0
27	our_us	our book	us book	
29	me_us	Show me the town!	Show us the town!	100.0
30	non_coordinated_subj_obj	Do you want to come with us?	Do you want to come with we?	
31	non_coordinated_obj_subj	They can ride all day.	Them can ride all day.	
33	nasal_possessive_pron	her, his, our; hers, ours, ours	hern, hisn, ourn; hersn, oursn, ourns	100.0
34	yall	you	y'all	ChcE IndE 100.0
35	you_ye	Sure it's no good to you in England.	Sure it's no good to ye in England.	
39	plural_interrogative	Who came?	Who-all came?	99.7
40	reduplicate_interrogative	Who's coming today?	Who-who's coming today?	IndE 99.8
41	anaphoric_it	Things have become more expensive than they used to be.	Things have become more expensive than it used to be.	IndE 100.0
42	object_pronoun_drop	I got it from the store.	I got from the store.	IndE 98.1
43	null_referential_pronouns	When I come back from my work I just travel back to my home.	When I come back from my work just travel back to my home.	ChcE IndE 93.2
45	it_dobj	As I explained to her, this is not the right way.	As I explained it to her, this is not the right way.	IndE 100.0
46	it_is_referential	It is very nice food.	Is very nice food.	
47	it_is_non_referential	Okay, it's time for lunch.	Okay, is time for lunch.	IndE 100.0

Table 7: Pronouns (Section 3)

	FUNCTION	SAE	TRANSFORM	VAL ACC.
49	regularized_plurals	wives, knives, lives, leaves	wifes, knifes, lifes, leafs	IndE 99.6
50	plural_preposed	shooting birds	shooting alla bird	
51	plural_postposed	The boys	Da boy dem	
55	mass_noun_plurals	furniture, machinery, equipment, evidence, luggage, advice, mail, staff	furnitures, machineries, equipments, evidences, luggages, advices, mails, staffs	IndE 100.0
56	zero_plural_after_quantifier	It's only five miles away.	It's only five mile away.	ChcE IndE 97.3
57	plural_to_singular_human	The three girls there don't want to talk to us.	The three girl there don't want to talk to us.	IndE 100.0
58	zero_plural	Some apartments are bigger.	Some apartment are bigger.	IndE 100.0
59	double_determiners	This common problem of ours is very serious.	This our common problem is very serious.	IndE 100.0
60	definite_for_indefinite_articles	She's got a toothache.	She's got the toothache	IndE 99.6
61	indefinite_for_definite_articles	The moon was very bright last night.	A moon was very bright last night.	IndE 100.0
62	remove_det_definite	He's in the office.	He's in office.	IndE 100.0
63	remove_det_indefinite	Can I get a better grade?	Can I get better grade?	IndE 99.0
64	definite_abstract	I stayed on until Christmas.	I stayed on until the Christmas.	IndE 100.0
65	indefinite_for_zero	We received good news at last.	We received a good news at last.	
66	indef_one	What happened? Oh, a dog bit me.	What happened? Oh, one dog bit me.	IndE 100.0
67	demonstrative_for_definite_articles	They have two children. The elder girl is 19 years old.	They have two children. That elder girl is 19 years old.	IndE 99.1
68	those_them	I don't have any of those qualifications.	I don't have any of them qualifications.	
70	proximal_distal_demonstratives	this book that is right here vs. those books that are over there	this here book vs. them there books	ChcE 92.9
71	demonstrative_no_number	These books are useful for my study.	This books are useful for my study.	IndE 98.8
73	existential_possessives	I have a son.	Son is there.	
74	possessives_for_post	This is my mother's house.	This is the house for my mother.	
75	possessives_for_pre	Long time ago he was my sister's husband.	Long time he was for my sister husband.	
76	possessives_belong	the woman's friend	woman belong friend	
77	null_genitive	my cousin's bike	my cousin bike	IndE 100.0
78	double_comparative, double_superlative	That is so much easier to follow.	That is so much more easier to follow.	IndE 100.0
79	synthetic_superlative	He is the most regular guy I know.	He is the regularest guy I know.	IndE 100.0
80	analytic_superlative	one of the prettiest sunsets	one of the most pretty sunsets	IndE 100.0
81	more_much	The situation is more serious than I thought.	The situation is much serious than I thought.	IndE 100.0
82	comparative_as_to	She is bigger than her sister.	She is bigger as her sister.	
84	comparative_than	They like football more than basketball.	They like football than basketball.	
85	comparative_more_and	He has more clothes than all of us.	He has more clothes and all of us.	
86	zero_degree	He is one of the most radical students that you can ever find.	He is one of the radical students that you can ever find.	IndE 100.0
87	adj_postfix	A big and fresh fish is my favorite.	A fish big and fresh is my favorite.	

Table 8: Noun Phrases (Section 3)

	FUNCTION	SAE	TRANSFORM	VAL ACC.
88	progressives	I like her hair style right now.	I am liking her hair style.	IndE 99.4
95	standing_stood	He was standing on the corner.	He was stood on the corner.	
96	that_resultative_past_participle	There is a car that broke down on the road.	There is a car broken down on the road.	95.5
97	medial_object_perfect	He has written a letter.	He has a letter written.	
98	after_perfect	She has just sold the boat.	She's after selling the boat.	
99	simple_past_for_present_perfect	I've eaten the food. So can I go now?	I ate the food. So can I go now?	ChcE 94.7
100	present_perfect_for_past	We were there last year.	We've been there last year.	IndE 99.9
101	present_for_exp_perfect	I've known her since she was a child.	I know her since she was a child.	IndE 100.0
102	be_perfect	They haven't left school yet.	They're not left school yet.	
103	do_tense_marker	I knew some things weren't right.	I did know some things weren't right.	
104	completive_done	Sharon has read the whole book.	Sharon done read the whole book.	
105	completive_have_done	He has talked about me.	He has done talked about me.	
106	irrealis_be_done	If you love your enemies, they will eat you alive in this society.	If you love your enemies, they be done eat you alive in this society.	100.0
107	perfect_slam	I have already told you not to mess up	I slam told you not to mess up.	
108	present_perfect_ever	I have seen the movie.	I ever see the movie.	
109	perfect_already	Have you eaten lunch?	Did you eat already?	
110	completive_finish	I have eaten.	I finish eat.	
111	past_been	I told you.	I been told you.	
112	bare_perfect	We had caught the fish when the big wave hit.	We had catch the fish when the big wave hit.	
114	future_sub_gon	He will come with us.	He gon' come with us.	
115	volition_changes	You want to go.	You waan go.	
116	come_future	I am about to cook your meal.	I am coming to cook your meal.	
117	present_for_neutral_future	Next week, I will be leaving the States and going to Liberia.	Next week, I leaving the States, I going to Liberia.	IndE 100.0
118	is_am_1s	I am going to town.	I's going to town.	
119	will_would	I will meet him tomorrow.	I would meet him tomorrow.	IndE 100.0
120	if_would	If I were you I would go home now.	If I would be you I would go home now.	

Table 9: Tense and Aspect (Section 3)

	FUNCTION	SAE	TRANSFORM	VAL ACC.
121	`double_modals`	We could do that.	We might could do that.	ChcE 91.8
123	`present_modals`	I wish I could get the job.	I wish I can get the job.	IndE 100.0
126	`finna_future, fixin_future`	They're about to leave.	They're fixin to leave town.	ChcE 92.3

Table 10: Mood (Section 3)

	FUNCTION	SAE	TRANSFORM	VAL ACC.
128	`regularized_past_tense`	He caught the ball.	He catched the ball.	ChcE IndE 92.7
129	`bare_past_tense`	They came and joined us.	They come and joined us.	
130	`past_for_past_participle`	He had gone.	He had went.	ChcE 92.9
131	`participle_past_tense`	I saw it.	I seen it.	IndE 100.0
132	`bare_past_tense`	Here are things you ordered yesterday.	Here are things you order yesterday.	ChcE IndE 87.7
133	`double_past`	They didn't make it this time.	They didn't made it this time.	IndE 99.5
134	`a_ing`	Where are you going?	Where are you a-goin?	100.0
135	`a_participle`	You've killed your mother.	You've a-killed your mother.	
143	`transitive_suffix`	You can see the fish.	You can see 'im fish.	
145	`got_gotten`	I hope you've got your topic already.	I hope you've gotten your topic already.	100.0
146	`verbal_ing_suffix`	I can drive now.	I can driving now.	IndE 100.0
147	`conditional_were_was`	If I were you	If I was you	
148	`serial_verb_give`	I bought rice for you.	I buy rice give you.	
149	`serial_verb_go`	Grandfather sends us to school.	Grandfather send us go school.	100.0
150	`here_come`	Bring the book here.	Take the book bring come.	
153	`give_passive`	John was scolded by his boss	John give his boss scold.	

Table 11: Verb Morphology (Section 3)

	FUNCTION	SAE	TRANSFORM	VAL ACC.
154	`negative_concord`	I don't want any help.	I don't want no help.	ChcE IndE 92.9
155	`aint_be`	That isn't fair.	That ain't fair.	ChcE 81.8
156	`aint_have`	I hadn't seen them yet.	I ain't seen them yet.	
157	`aint_before_main`	something I didn't know about	something I ain't know about	
158	`dont`	He doesn't always tell the truth.	He don't always tell the truth.	
159	`never_negator`	He didn't come.	He never came.	100.0
160	`no_preverbal_negator`	I don't want any job or anything.	I no want any job or anything.	
161	`not_preverbal_negator`	The baby didn't eat food and cried a lot.	The baby not ate food and cried a lot.	
162	`nomo_existential`	There is not any food in the refrigerator.	No more food in the refrigerator.	
163	`wasnt_werent`	John was there, but Mike wasn't	John was there, but Mike weren't	
164	`invariant_tag_amnt`	I believe I am older than you. Is that correct?	I am older than you, amn't I?	
165	`invariant_tag_non_concord`	I believe you are ill. Is that correct?	You are ill, isn't it?	IndE 99.1
166	`invariant_tag_can_or_not`	Can I go home?	I want to go home, can or not?	
167	`invariant_tag_fronted_isnt`	I can go there now can't I?	Isn't, I can go there now?	

Table 12: Negation (Section 3)

	FUNCTION	SAE	TRANSFORM	VAL ACC.
170	`uninflect`	He speaks English.	He speak English.	ChcE IndE 94.9
171	`generalized_third_person_s`	Every Sunday we go to church.	Every Sunday we goes to church.	
172	`existential_there`	There are two men waiting in the hall.	There's two men waiting in the hall.	ChcE IndE 90.0
173	`existential_it`	There's some milk in the fridge.	It's some milk in the fridge.	ChcE 87.5
174	`drop_aux_be_progressive`	You are always thinking about it.	You always thinking about it.	IndE 100.0
175	`drop_aux_be_gonna`	He is gonna go home and watch TV.	He gonna go home and watch TV.	ChcE IndE 83.3
176	`drop_copula_be_NP`	He is a good teacher.	He a good teacher.	
177	`drop_copula_be_AP`	She is smart.	She smart.	
178	`drop_copula_be_locative`	She is at home.	She at home.	
179	`drop_aux_have`	I have seen it before.	I seen it before.	IndE 100.0
180	`were_was`	You were hungry but he was thirsty.	You was hungry but he was thirsty. OR: You were hungry but he were thirsty.	

Table 13: Agreement (Section 3)

	Function	SAE	Transform	Val Acc.
186	who_which	He's the man who looks after the cows.	He's the man which looks after the cows.	
187	who_as	The man who was just here.	The man as was just here.	
188	who_at	This is the man who painted my house.	This is the man at painted my house.	
189	relativizer_where	My father was one of the founders of the Underground Railroad, which helped the slaves to run away to the North	My father was one o de founders o' de Underground Railroad where help de slaves to run way to de North.	
190	who_what	This is the man who painted my house.	This is the man what painted my house.	
191	relativizer_doubling	But these, these little fellahs who had stayed before	But these, these little fellahs that which had stayed befo'	IndE 100.0
192	analytic_whose_relativizer	This is the man whose wife has died.	This is the man that his wife has died. OR: This is the man what his wife has died.	
193	null_relcl	The man who lives there is friendly.	The man lives there is friendly.	ChcE IndE 88.7
194	shadow_pronouns	This is the house which I painted yesterday.	This is the house which I painted it yesterday.	IndE 100.0
195	one_relativizer	The cake that John buys is always very nice to eat.	The cake John buy one always very nice to eat.	
196	correlative_constructions	The ones I made are the good ones.	The one I made, that one is good.	
197	linking_relcl	Unless you are going to get 88, but some universities are not going to give those marks	Unless you are going to get 88 which some universities are not going to give those marks	
198	preposition_chopping	You remember the swing that we all used to sit together on?	You remember the swing that we all used to sit together?	IndE 100.0
199	reduced_relative	There is nothing like food cooked by Amma!	There is nothing like Amma cooked food!	

Table 14: Relativization (Section 3)

	Function	SAE	Transform	Val Acc.
200	say_complementizer	We hear that you were gone to the city.	We hear say you gone to the city.	
201	for_complementizer	You mean your mother allows you to bring over boyfriends?	You mean your mother allows you for bring over boyfriends?	
202	for_to_pupose	We always had gutters in the winter time to drain the water away.	We always had gutters in the winter time for to drain the water away.	
203	for_to	He had the privilege to turn on the lights.	He had the privilege for to turn on the lights. OR: He had the privilege for turn on the lights.	100.0
204	what_comparative	I'm taller than he is.	I'm taller than what he is.	IndE 100.0
205	existential_got	There's no water in the toilet.	Got no water in the toilet.	100.0
206	existential_you_have	There are some people who don't give a damn about animals.	You have some people they don't give a damn about animals.	IndE 100.0
207	that_infinitival_subclause	He wanted me to go with him.	He wanted that I should go with him.	IndE 100.0
208	drop_inf_to	They were allowed to call her.	They were allowed call her.	100.0
209	to_infinitive	He made me do it.	He made me to do it.	IndE 100.0
210	bare_ccomp	When mistress started whooping her, she sat her down.	When mistress started whoop her, she sat her down.	

Table 15: Complementation (Section 3)

	Function	SAE	Transform	Val Acc.
211	clause_final_though_but	There's nothing wrong with this box though.	There's nothing wrong with this box, but.	
212	clause_final_really_but	I don't know what else she can do, really.	I don't know what else she can do, but.	
213	chaining_main_verbs	If you stay longer, they have to charge more.	Stay longer, they have to over-charge.	
214	corr_conjunction_doubling	Despite being instructed on what to do, he still made some misakes.	Despite being instructed on what to do still yet he made some misakes.	IndE 100.0
215	subord_conjunction_doubling	Although you are smart, you are not appreciated	Although you are smart, but you are not appreciated	

Table 16: Adverbial Subordination (Section 3)

	Function	SAE	Transform	Val Acc.
216	null_prepositions	I'm going to town.	I'm going town.	IndE 99.7
220	degree_adj_for_adv	That's really nice and cold	That's real nice and cold	ChcE IndE 99.4
221	flat_adj_for_adv	She speaks so softly.	She speaks so soft.	ChcE IndE 86.7
222	too_sub	They are very nice. We had a good time there.	They are too nice. We had a good time there.	

Table 17: Adverbs and Prepositions (Section 3)

	FUNCTION	SAE	TRANSFORM	VAL ACC.
223	`clefting`	A lot of them are looking for more land.	It's looking for more land a lot of them are.	100.0
224	`fronting_pobj`	I drive to town every Saturday.	To town every Saturday I drive.	**IndE** 99.5
226	`negative_inversion`	Nobody showed up.	Didn't nobody show up.	100.0
227	`inverted_indirect_question`	I'm wondering what you are going to do.	I'm wondering what are you going to do.	**ChcE** **IndE** 91.2
228	`drop_aux_wh`	When is she coming?	When she coming?	**IndE** 99.8
229	`drop_aux_yn`	Do you get the point?	You get the point?	**IndE** 99.9
230	`doubly_filled_comp`	Who ate what?	What who has eaten?	
231	`superlative_before_matrix_head`	The thing I like most is apples.	The most thing I like is apples.	
232	`double_obj_order`	She would teach it to us.	She'd teach us it.	**IndE** 100.0
234	`acomp_focusing_like`	It was really cheap.	It was like really cheap.	**ChcE** **IndE** 91.2
235	`quotative_like`	And my friend said "No way!"	And my friend was like "No way!"	

Table 18: Discourse and Word Order (Section 3)

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 8

☑ A2. Did you discuss any potential risks of your work?
Section 9

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3

☑ B1. Did you cite the creators of artifacts you used?
Section 5

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 9

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 9

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The released datasets are derivatives of CoQA. Our Morphosyntactic patterns could not add additional information about individuals. The annotators were anonymized in accordance with the ethics review body of the authors' institution.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 5

C ☑ Did you run computational experiments?

Section 6.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix C

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 6

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We used Bootstrap tests for significance for each run. We state that this is the bootstrap of a single run in the caption of each table.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Figures 4 and 5

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 4

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 9

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section 9

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 4