Corpus-Guided Contrast Sets for Morphosyntactic Feature Detection in Low-Resource English Varieties

Tessa Masis *they/them/theirs*

Anissa Neal *she/her/hers*

Lisa Green *she/her/hers*

Brendan O'Connor *he/him/his*

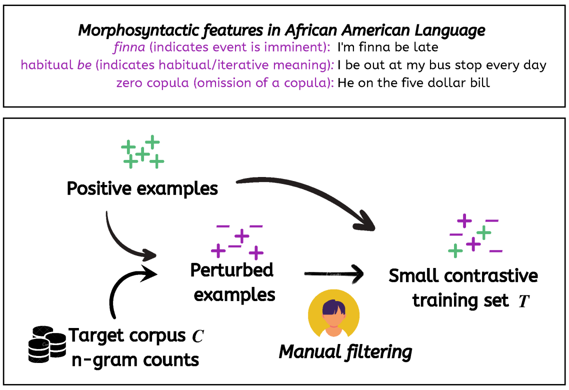
University of Massachusetts Amherst
{tmasis,brenocon}@cs.umass.edu
{anneal,lgreen}@linguist.umass.edu

Abstract

The study of language variation examines how language varies between and within different groups of speakers, shedding light on how we use language to construct identities and how social contexts affect language use. A common method is to identify instances of a certain linguistic feature—say, the zero copula construction—in a corpus, and analyze the feature's distribution across speakers, topics, and other variables, to either gain a qualitative understanding of the feature's function or systematically measure variation. In this paper, we explore the challenging task of automatic morphosyntactic feature detection in low-resource English varieties. We present a human-in-the-loop approach to generate and filter effective contrast sets via corpus-guided edits. We show that our approach improves feature detection for both Indian English and African American English, demonstrate how it can assist linguistic research, and release our fine-tuned models for use by other researchers.

1 Introduction

Linguistic *features*—such as specific phonological, syntactic, or lexical phenomena that may be associated with a language variety—are widely used by sociolinguists to quantify linguistic variation between speakers through feature frequency measurements (Renn and Terry, 2009; Grieser, 2019; Craig and Washington, 2006), even if subject to certain limitations (Green, 2017). Since manual annotation is limited due to the required expert human labor, automatic methods are a valuable alternative (Grieve et al., 2011; Jones, 2015; Eisenstein, 2015; Nguyen et al., 2016). However, accurately detecting morphosyntactic features (e.g. Figure 1) remains an open challenge, especially in informal genres such as transcripts and social media, and in low resource nonstandard languages. We explore fine-tuning pretrained language models (LMs) for utterance-level classification of a feature by training on a *contrast set*—a small collection of positive and negative examples that are highly similar—as recently introduced by Demszky et al. (2021).



Morphosyntactic features in African American Language
finna (indicates event is imminent): I'm finna be late
habitual *be* (indicates habitual/iterative meaning): I be out at my bus stop every day
zero copula (omission of a copula): He on the five dollar bill

Positive examples
Perturbed examples
Target corpus C n-gram counts
Small contrastive training set T
Manual filtering

Figure 1: Top: Example features. Bottom: Our approach to generate contrast sets for feature detection.

Our work makes the following contributions:

- We propose a method for generating morphosyntactically contrastive training data, combining corpus-driven edits and human-in-the-loop filtering (§4).
- We evaluate our method's ability to detect features against new baselines on three datasets, encompassing two Englishes (Indian English (IndE) and African American English (AAE)) and two centuries of speakers, and show that our best method outperforms prior work by up to 16 points in Prec@100 scores (§5).
- For further validation, we confirm and extend the findings of sociolinguistic studies of AAE which use manual feature annotation to examine if feature use aligns with social factors like age and gender (§6).
- Finally, we release training data and models for detecting 10 features in IndE and 17 in AAE.[1]

[1] https://github.com/slanglab/CGEdit

Proceedings of the first workshop on field linguistics, pages 11–25
October 16, 2022.

2 Related Work

Feature detection. Detecting morphosyntactic features in low-resource domains presents significant challenges. Rule-based approaches have used sequences of unigrams and POS tags to identify syntactic features (Blodgett et al., 2016), but many features cannot be defined by sequences and the tags may be unreliable. More recently, machine learning has been used for feature detection by training domain-specific LMs with synthetically augmented data (Santiago et al., 2022), fine-tuning pretrained LMs with contrast sets (Demszky et al., 2021), or manually filtering results from noisy classifiers (Austen, 2017). While prior work has only considered one language variety at a time and primarily evaluated with labeled test sets, we examine performance on multiple language varieties and analyze external sociolinguistic validity.

Contrast set generation. Manual generation of contrast sets has mostly been used for semantic tasks (Staliūnaitė and Bonfil, 2017; Mahler et al., 2017; Gardner et al., 2020), and occasionally for morphosyntactic tasks (Demszky et al., 2021). Unlike these approaches, our proposed method generates a *morphosyntactically diverse* contrast set via a corpus-guided edit system. Data augmentation methods for automatic generation of contrast sets include random edits (Smith and Eisner, 2005; Alleman et al., 2021), which cannot target specific linguistic features, or informed edits (Burlot and Yvon, 2017; Sennrich, 2017; Gulordava et al., 2018; Miao et al., 2020; Ross et al., 2021), which require syntactic or semantic annotations that are not easily available for datasets with nonstandard languages.

3 Task and Data

3.1 Morphosyntactic feature detection

Given a training set T, target corpus C, and morphosyntactic features F, for each $f \in F$ we model

$$P(f_x = 1 \mid T, x), \tag{1}$$

where $f_x \in \{0, 1\}$ indicates the utterance $x \in C$ contains the feature when $f_x = 1$. An utterance may contain multiple features.

3.2 Language Varieties and Data

We consider two English varieties, IndE and AAE, each with their own target corpora C and feature inventories F; see Appendix A for feature lists.

Indian English. The International Corpus of English (ICE) (Greenbaum and Nelson, 1996) is a collection of national and regional English varieties, and contains IndE material produced after 1989. The ICE-India subcorpus that our study uses is the complete subset of spontaneous spoken dialogues (21,759 utterances). We use manual annotations of 10 syntactic features from Lange (2012).

African American English. We use two unlabeled AAE corpora. The first is the Corpus of Regional African American Language (CORAAL) (Kendall and Farrington, 2021), which contains sociolinguistic interviews with AAE speakers from 1968-2017 from six US sites (152,069 utterances). The second is Born in Slavery: Slave Narratives from the Federal Writers' Project, 1936-38 (FWP) (Library of Congress, 2001), a digital archive containing over 2,300 ex-slave narratives, with speakers from 17 US states (148,018 utterances).[2]

We examine 17 AAE features, sourced from Green (2002) and Koenecke et al. (2020); examples of three features are in Figure 1, and a complete list is in Appendix A. During evaluation, we manually annotated the top 100 utterances per AAE feature, for each corpus, for the Prec@100 scores in §5.

4 CGEdit: Corpus-Guided Edits

4.1 Motivation

Our method starts with a seed set of positive examples illustrating a feature, then uses corpus n-gram statistics to generate proposed negative (and additional positive) examples, which require manual filtering by the user to define the final training set. A major motivation is speed and ease of use—it is easier to filter candidate examples than to manually write all the examples, as in Demszky et al. (2021).

At the same time, we believe negative examples should be intelligently synthesized. A morphosyntactic feature is beholden to its syntactic constraints (i.e. word order, co-occurrence requirements); if a sentence does not follow these constraints then it is not an instance of the feature (Wilson and Mihalicek, 2011, Ch. 5.2). For example, an instance of zero copula must have a noun phrase immediately followed by a predicate and must not have a copula. The positive example in Figure 2 obeys these syntactic constraints while the negatives do

[2] Given authenticity and reliability concerns about FWP (Maynor, 1988; Wolfram, 1990), we primarily use it to evaluate our method, and not to pursue linguistic questions about Early African American English.

not. Unlike previous work which uses constraints to detect or generate positive instances, *we generate negative examples which minimally violate these constraints* to create a contrast set that defines a tight decision boundary. Based on the view that good syntax is largely independent from meaning (Chomsky, 1957), we argue that focusing on syntactic constraint violation is a useful first step. While potentially valuable, semantic-preserving edits are beyond the scope of this work.

4.2 Method

Training data. We briefly describe how the contrast sets are generated (Figure 2; see Appendix B for details). For a single feature, the input is a small set P of 5 positive examples constructed by the authors and an unlabeled target corpus C to compute n-gram statistics. The output is a contrast set T consisting of both P plus semi-synthetic positive and negative examples.

The first step proposes candidate examples by perturbing words in positive examples through corpus-guided local edits. For each overlapping 3-gram t in a positive example p, we perturb it by swapping t for a new 2-, 3-, or 4-gram t' that is both similar to t, and has a high frequency in target corpus C. Similarity is defined as having 0 to 1 subtoken difference between t and t'.[3] This step typically produces 10-50 perturbed examples, which may or may not have the feature. Our corpus-guided edits are effective because they generate plausible sentences with targeted edits, while random edits often propose ungrammatical output.[4]

In the second step, the perturbed examples are manually filtered so that only 2 positive and 3 negative examples are retained for each original p. Both p and the new examples are included in the final training set T. This step takes 30-60 seconds per p, and was performed by the first author.

Models. We fine-tune multiheaded BERT models, where each head is a binary classifier for a single feature (Devlin et al., 2019). We use two sets of models in our experiments, where a set shares a language variety, a feature inventory F, target corpora C (i.e. test set for our results in Table 1), and a BERT variant (*bert-base-uncased* for IndE, *bert-base-cased* for AAE, selected based on preliminary experiments). The only variation between models *within* a set is the approach used to generate the training set T. Models were fine-tuned with cross-entropy loss for 500 epochs using the Adam optimizer, batch size of 64, and learning rate of 10^{-5}, warmed up over the first 150 epochs.[5]

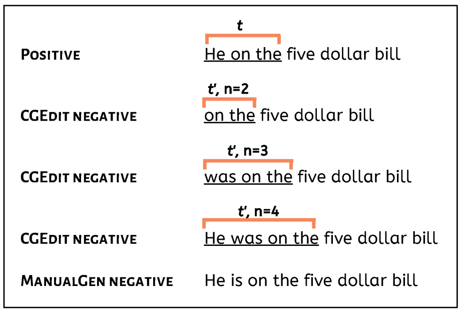


Figure 2: Examples of negative examples generated via our approach, compared to a semantically-matched, manually created example (MANUALGEN).

5 Results and Analysis

Baselines. We compare our approach (which we refer to as CGEDIT) to several baseline methods, all of which take the same seed set of positive examples P then add negative examples to complete the training set. Examples in P were sourced from Demszky et al. (2021) for IndE and crafted by the authors for AAE.

MANUALGEN: The approach used in Demszky et al. (2021). This method involves manually generating negatives by modifying positive examples so they are (1) semantically-similar Mainstream American English versions, and (2) do not have the feature (see Figure 2); see discussion in §4.1. Next, we also test two methods to completely automatically generate negative examples:

AUTOGEN: This approach automatically generates negative examples by dividing a positive example p into n-grams and shuffling the n-grams.

AUTOID: Automatic identification randomly chooses unlabeled examples from target corpus C as the negatives. The assumption that unlabeled examples are negatives with class label noise underpins contrastive learning (Chen et al., 2020) and PU learning methods (Bekker and Davis, 2020).

Overall results. Table 1 presents performance

[3] Specifically, the set difference between subtoken sets $set(t)$ and $set(t')$ must have cardinality 0 or 1; thus a 2-gram t' represents a (sub)token deletion, a 4-gram an insertion, and perturbations may change order as well. Since only a single 3-gram is changed, the resulting perturbed utterance has a low edit distance to the original.

[4] While our n-gram swapping heuristic is straightforward, generating from a C-specific language model could be an interesting alternative in future work.

[5] Early experiments indicated that class-balanced loss did not improve scores.

	ICE-India			CORAAL	FWP
Approach	ROC-AUC	AP	Prec@100	Prec@100	Prec@100
AutoGen	68.94	12.63	16.93	-	-
AutoID	74.90	15.24	17.87	-	-
ManualGen	86.83	25.77	31.63	57.88	58.71
AutoID + ManualGen	76.34	19.95	24.30	-	-
CGEdit	84.92	27.48	32.50	**67.41**	68.00
ManualGen + CGEdit	**88.76**	**29.32**	**35.67**	64.94	**74.35**

Table 1: Area under precision-recall curve (ROC-AUC), average precision (AP), and precision@100 in percentages for feature detection on all three corpora. Results are averages over all features (10 in ICE-India, 17 in CORAAL and FWP). Reported scores for ICE-India are averaged from three runs with different random seeds. Best scores are bolded.

of the proposed approach against baselines and prior work. AutoGen and AutoID perform the worst across metrics. CGEdit outperforms ManualGen, the best prior work on this task, by up to 10 points in Prec@100 scores for both AAE datasets, CORAAL and FWP. Combining the training sets of ManualGen and CGEdit yielded the best performance, consistently outperforming ManualGen by about 4 points across metrics in ICE-India and by about 10-15 points in Prec@100 scores for both CORAAL and FWP. These gains can't simply be attributed to more training data, as combining AutoID and ManualGen training sets did not improve performance.

Better performance on AAE corpora may be due to a few variables: a higher number of AAE features means a larger total training set; larger AAE corpora mean more target corpus n-grams; the selected AAE features may be easier to distinguish or more prevalent than the IndE ones. Discrepancies between CORAAL and FWP are likely due to different feature prevalences.

Results by feature. Feature difficulty is similar across approaches; invariant features are easier to detect (i.e. focus *itself* in IndE; *finna* in AAE), while features with long-distance dependencies are more difficult (i.e. double object construction in AAE). See Appendix C for complete results.

6 Replicating Prior Sociolinguistic Work

We recreate three recent studies of CORAAL where original authors manually annotated AAE morphosyntactic features and analyzed correlations between feature frequency and speaker metadata (i.e. gender, region, socioeconomic status). We used the combined ManualGen + CGEdit model and Classify & Count (CC, summing hard classifications (Bella et al., 2010)) to calculate per-speaker feature frequency.[6] The same subsets of features and CORAAL were used as in previous work when possible; detailed results are in Appendix C.

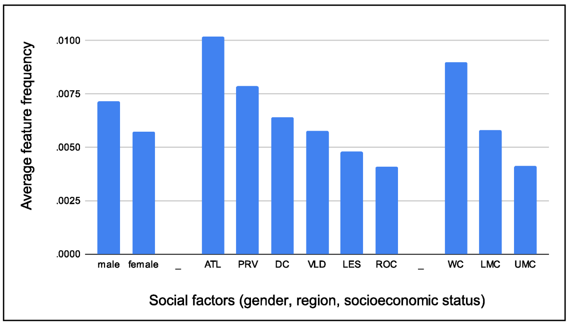


Figure 3: African American English feature variation by speaker's social factor, across all of CORAAL. Regions are Atlanta, GA; Princeville, NC; Washington, DC; Valdosta, GA; Lower East Side, NY; Rochester, NY; socioeconomic classes are Working Class, Lower Middle Class, Upper Middle Class.

Koenecke et al. (2020) annotated 35 morphosyntactic features in 150 utterances. We confirm their conclusions that average feature frequency was lowest in Rochester, followed by DC, then Princeville; and lower among male speakers than female.

Cukor-Avila and Balcazar (2019) looked at 3 features over 14,506 utterances. They qualitatively found considerable variation in feature use between speakers, even when within the same age group. We confirm this quantitatively: standard deviation between speakers within an age group is larger than standard deviation between age group means.

Grieser (2019) examined 14 features over 18,553 utterances. We confirm findings that age and so-

[6] In early experiments we tested the Saerens et al. (2002) EM algorithm and PCC (Bella et al., 2010) to improve frequency estimation, but found few improvements.

cioeconomic status are negatively correlated with feature use. Grieser also found that being male was weakly correlated with feature use; interestingly, our results agree when we look at all 17 features or all of CORAAL, but not when we look at the same feature and data subsets as Grieser. This may indicate how small sample size (in terms of both features and datasets) can skew results.

See Figure 3 for average frequencies of our 17 features in all 152,069 utterances of CORAAL, broken down by several social factors of the speaker. Feature detection at this scale is only possible with automatic methods, and allows researchers to draw more reliable conclusions about language use.

7 Discussion and Future Work

We propose a corpus-driven and manually-filtered approach to generate contrast sets for morphosyntactic feature detection in low-resource language varieties, which may be useful for novel sociolinguistic analysis in future work. This approach may be extendable to datasets with other nonstandard language varieties (e.g. ICE with 14 English varieties (Greenbaum and Nelson, 1996), QADI with 18 Arabic varieties (Abdelali et al., 2021), Corpus del Español with 21 Spanish varieties (Davies, 2016), or Masakhane's African language collection, currently under development (∀ et al., 2020)), in addition to social media corpora, which are largely unlabeled and could benefit from automatic methods.

Additionally, while we only examined automatic identification of noisy negatives, future work might explore automatic identification of reliable negatives by using an apt word representation and distance function to obtain unlabeled examples which are least similar to the positives (Bekker and Davis, 2020). Other extensions might consider adding manual filtering to an automatic identification approach, such as filtering through and identifying the nearest unlabeled examples that are true negatives, instead of identifying reliable (e.g. distant) negatives.

8 Ethical Considerations and Broader Impact

Our objective is to expand the linguistic coverage of NLP tools to include marginalized language varieties, so that they may also benefit from the linguistic analysis made possible by methodological innovation. We hope to aid both sociolinguistic and corpus linguistic researchers studying nonstandard language use.

Since language varieties, including the ones examined in this study, may correlate with the national origin or ethnicity of the speaker and linguistic feature frequency may correlate with social factors, such as gender or socioeconomic status, there is a risk of automatic feature detection being used to infer personal information about a speaker (Kröger et al., 2022; Chancellor et al., 2019; Veronese et al., 2019). Our study has sought to show that there is a correlation between language use and social factors, but does not support any claims about the accuracy or ethics of using linguistic feature frequency to predict a given social factor.

There is not a one-to-one mapping of feature frequency to ethnicity, socioeconomic status, or any other social factor. Two speakers with the same set of social factors may exhibit different feature frequencies; life circumstances do not deterministically produce linguistic competence. In addition, linguistic competence does not deterministically produce feature frequency. Every speaker has the ability to style-shift and thus use linguistic features to varying degrees for a given context, exhibiting a range of feature frequencies throughout their spoken interactions (Sharma, 2017, 2018). There are many factors that may influence observed feature frequency, including pragmatic context, register, topic, relationship between the speakers, relationship to one's own identity, and so on. This complex relationship between language production and external factors should be considered when using this technology.

Acknowledgements

We would like to thank Devyani Sharma for help with accessing the ICE-India corpus and Claudia Lange's annotations, with permission. We would also like to thank the UMass AAE group as well as anonymous reviewers from the First Field Matters Workshop for their helpful comments and feedback. This work was supported by National Science Foundation grants 1845576 and 2042939; any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. QADI: Arabic Dialect Identification in the Wild. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Matteo Alleman, Jonathan Mamou, Miguel A Del Rio, Hanlin Tang, Yoon Kim, and SueYeon Chung. 2021. Syntactic Perturbations Reveal Representational Correlates of Hierarchical Phrase Structure in Pretrained Language Models. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 263–276, Online. Association for Computational Linguistics.

Martha Austen. 2017. "Put the Groceries Up": Comparing Black and White Regional Variation. *American Speech*, 92(3):298–320.

Jessa Bekker and Jesse Davis. 2020. Learning from Positive and Unlabeled Data: A Survey. *Mach. Learn.*, 109(4):719–760.

Antonio Bella, Cesar Ferri, José Hernández-Orallo, and María José Ramírez-Quintana. 2010. Quantification via Probability Estimators. In *2010 IEEE International Conference on Data Mining*, pages 737–742.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of Machine Translation Systems. In *Proceedings of the Second Conference on Machine Translation*, pages 43–55, Copenhagen, Denmark. Association for Computational Linguistics.

Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio, and Munmun De Choudhury. 2019. A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 79–88, New York, NY, USA. Association for Computing Machinery.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

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

Holly K Craig and Julie A Washington. 2006. *Malik goes to school: Examining the language skills of African American students from preschool-5th grade*. Psychology Press.

Patricia Cukor-Avila and Ashley Balcazar. 2019. Exploring Grammatical Variation in the Corpus of Regional African American Language. *American Speech*, 94(1):36–53.

Mark Davies. 2016. Corpus del Español: Web/Dialects.

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 *NAACL-HLT (1)*, pages 4171–4186.

Jacob Eisenstein. 2015. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19(2):161–188.

∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

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

Lisa Green. 2017. Beyond Lists of Differences to Accurate Descriptions. In *Data Collection in Sociolinguistics*, pages 281–284. Routledge.

Lisa J Green. 2002. *African American English: A Linguistic Introduction*. Cambridge University Press.

Sidney Greenbaum and Gerald Nelson. 1996. The International Corpus of English (ICE) Project. *World Englishes*, 15(1):3–15.

Jessica A. Grieser. 2019. Investigating Topic-Based Style Shifting in the Classic Sociolinguistic Interview. *American Speech*, 94(1):54–71.

Jack Grieve, Douglas Biber, Eric Friginal, and Tatiana Nekrasova. 2011. *Variation Among Blogs: A Multi-dimensional Analysis*, pages 303–322. Springer Netherlands, Dordrecht.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless Green Recurrent Networks Dream Hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Taylor Jones. 2015. Toward a Description of African American Vernacular English Dialect Regions Using "Black Twitter". *American Speech*, 90(4):403–440.

Tyler Kendall and Charlie Farrington. 2021. The Corpus of Regional African American Language. Eugene, OR: The Online Resources for African American Language Project.

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

Jacob Leon Kröger, Leon Gellrich, Sebastian Pape, Saba Rebecca Brause, and Stefan Ullrich. 2022. Personal information inference from voice recordings: User awareness and privacy concerns. *Proceedings on Privacy Enhancing Technologies*, 2022(1):6–27.

Claudia Lange. 2012. *The Syntax of Spoken Indian English*, volume 45. John Benjamins Publishing.

Manuscript Division Library of Congress. 2001. Born in Slavery: Slave Narratives from the Federal Writers' Project, 1936 to 1938.

Taylor Mahler, Willy Cheung, Micha Elsner, David King, Marie-Catherine de Marneffe, Cory Shain, Symon Stevens-Guille, and Michael White. 2017. Breaking NLP: Using Morphosyntax, Semantics, Pragmatics and World Knowledge to Fool Sentiment Analysis Systems. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 33–39, Copenhagen, Denmark. Association for Computational Linguistics.

Natalie Maynor. 1988. Written Records of Spoken Language: How Reliable Are They. *Methods in Dialectology*, pages 109–20.

Zhengjie Miao, Yuliang Li, Xiaolan Wang, and Wang-Chiew Tan. 2020. *Snippext: Semi-Supervised Opinion Mining with Augmented Data*, page 617–628. Association for Computing Machinery, New York, NY, USA.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational Sociolinguistics: A Survey. *Comput. Linguist.*, 42(3):537–593.

Jennifer Renn and J. Michael Terry. 2009. Operationalizing Style: Quantifying the Use of Style Shift in the Speech of African American Adolescents. *American Speech*, 84(4):367–390.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E. Peters, and Matt Gardner. 2021. Tailor: Generating and Perturbing Text with Semantic Controls. *CoRR*, abs/2107.07150.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. 2002. Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. *Neural Computation*, 14(1):21–41.

Harrison Santiago, Joshua Martin, Sarah Moeller, and Kevin Tang. 2022. Disambiguation of morpho-syntactic features of African American English – the case of habitual be.

Rico Sennrich. 2017. How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

Devyani Sharma. 2017. Scalar effects of social networks on language variation. *Language Variation and Change*, 29(3):393–418.

Devyani Sharma. 2018. Style dominance: Attention, audience, and the 'real me'. *Language in Society*, 47(1):1–31.

Noah A. Smith and Jason Eisner. 2005. Contrastive Estimation: Training Log-Linear Models on Unlabeled Data. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, page 354–362, USA. Association for Computational Linguistics.

Ieva Staliūnaitė and Ben Bonfil. 2017. Breaking Sentiment Analysis of Movie Reviews. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 61–64, Copenhagen, Denmark. Association for Computational Linguistics.

Alexandre Veronese, Alessandra Silveira, and Amanda Nunes Lopes Espiñeira Lemos. 2019. Artificial intelligence, Digital Single Market and the proposal of a right to fair and reasonable inferences: a legal issue between ethics and techniques. *UNIO – EU Law Journal*, 5(2):75–91.

C Wilson and V Mihalicek. 2011. *Language Files: Materials for an Introduction to Language and Linguistics*. Columbus: Ohio State University Press.

Walt Wolfram. 1990. Re-Examining Vernacular Black English. *Language*, 66(1):121–133.

A Feature inventories

Level	IndE Feature	Example utterance
Noun phrase	Non-initial existential *there*	library facility was not there
	Focus *itself*	We are feeling tired now itself
	Focus *only*	I like dressing up I told you at the beginning only
Verb phrase	Zero copula	Everybody (is) so worried about the exams
Sentence level	Left dislocation	we elders, we don't have much time to converse
	Resumptive subject pronoun	the father, sometimes he is unemployed
	Resumptive object pronoun	also pickles, we eat it with this jaggery and lot of butter
	Topicalized object (argument)	brothers and sisters you have
	Topicalized non-argument constituent	with your child you have come
	Invariant tag *no/na/isn't it*	both works same hours, isn't it?

Table 2: Features of Indian English used in our study.

Level	Grammatical domain	AAE Feature	Example utterance
Noun phrase	Pronominal case	Zero possessive -*'s*	go over my grandmama('s) house
Verb phrase	Copula deletion	Zero copula	she (is) the folk around here
	Tense marking	Double marked/overregularized	she likeded me the best
	Aspect marking	Habitual *be*	I just be liking the beat
		Resultant *done*	you done lost your mind
	Other verbal markers	*finna*	she's finna have a baby
		come	she come grabbing me on my shirt
		Double modal	he might could really get our minds
	Negation	Negative concord	I ain't doing nothing wrong
		Negative auxiliary inversion	don't nobody know what I had
		Non-inverted negative concord	nobody don't say nothing
		Preverbal negator *ain't*	I ain't doing nothing wrong
Sentence level	Subject-verb agreement	Zero 3rd p sg present tense -*s*	I don't know if it count(s)
		is/was-generalization	they is die hard Laker fans
	Number marking	Zero plural -*s*	about four or five month(s)
	Ditransitive constructions	Double-object construction	I got me my own car
	Interrogative constructions	*Wh*-question	what they was doing?

Table 3: Features of African American English used in our study.

B Approach descriptions

B.1 Proposed approach

A positive example p is defined as $(x_1, x_2, ..., x_n)$ where x_i is a subtoken. For each positive example p:

1. A 3-gram instance t in p is defined as (x_i, x_{i+1}, x_{i+2}). For each 3-gram instance t in p:
   (a) For each $n \in \{2, 3, 4\}$, find the 3 most frequent n-grams from the corpus where, for each n-gram t', the set difference between $set(t)$ and $set(t')$ is at most one subtoken.
   (b) Create perturbed examples by swapping t for t'. These perturbed examples may or may not have the feature.
2. Randomly order the perturbed examples.
3. Manually filter and label the perturbed examples; examples that pass the filter should not have invalid subtoken combinations, positive examples should unambiguously have the feature, and negative examples should unambiguously not have the feature. Examples that pass the filter (positive or negative) may be ungrammatical. Stop after 2 positives and 3 negatives have passed the filter. Including the original positive example p, you should have 3 positives and 3 negatives.

We provide here an example of our approach. For the feature zero copula, we are given $p =$ He on the five dollar. We generate:

Perturbed example
He on the last five
He on the five
on the other five dollar
He on the five hundred dollar
He was on the dollar
on the five dollar
the on five dollar
He and five on the dollar
He was on the five dollar
He on the five dollar bill
He beating on the five dollar
He on the dollar
He on the other dollar
He on five dollar
He the five dollar
He on five dollar bill
was on the five dollar

The manually filtered contrast set looks like:

Example	Label
He on the five dollar	1
He on the last five	1
He on the five	1
on the other five dollar	0
He was on the dollar	0
on the five dollar	0

B.2 Manual generation

Given a positive example p, manually construct a negative example by modifying p so they are (1) semantically-similar MAE versions, and (2) do not have the feature.

B.3 Automatic generation

For each positive example p:

1. Randomly choose n-gram order, where n is some value 0 < n < length(p) - 1.
2. Split positive example into sequential non-overlapping n-grams from left to right. If length of sentence isn't a multiple of n, then the remaining words form an additional m-gram (m < n).
3. Randomly shuffle the list of n-grams.
4. Repeat steps 1-3 until you have three distinct shuffled negative examples per positive example.[7]

B.4 Automatic identification

Randomly choose unlabeled examples from target corpus and label them as the negative examples. Five negatives are chosen per positive example.[8]

C Extended results and figures

Tables 4, 5, and 6 are per-feature results for Indian English features in ICE-India. Tables 7 and 8 are per-feature results for African American English features in CORAAL and FWP. Tables 9, 10, and 11 are standard deviation scores for Indian English features in ICE-India. Figures 4, 5, and 6 are detailed results from replicating prior sociolinguistic work.

[7] Number of negatives per positive was a tuned hyperparameter.

[8] Number of negatives per positive was a tuned hyperparameter.

	ROC-AUC					
Feature	AutoG.	AutoID	MnlG.	AutoID +MnlG.	CGEdit	MnlG. +CGEdit
Non-init. exist. *there*	91.14	90.47	89.88	89.74	95.46	89.03
Focus *itself*	94.08	98.02	98.70	97.58	99.49	99.89
Focus *only*	85.38	97.00	98.94	95.40	96.72	99.02
Zero copula	53.28	61.82	73.75	67.77	73.79	75.61
Left dislocation	64.17	70.18	93.13	69.32	89.92	93.14
Res. subject pronoun	72.81	70.03	93.60	67.92	88.32	89.94
Res. object pronoun	67.49	70.46	86.87	78.24	86.44	88.93
Topic. object (arg.)	63.20	59.17	76.72	54.28	72.08	81.30
Topic. non-arg. const.	44.90	55.48	69.24	55.55	59.99	79.54
Invar. tag *no/na/isn't it*	52.96	76.37	87.46	87.55	86.95	91.24
Macro average	68.94	74.90	86.83	76.34	84.92	88.76

Table 4: ROC-AUC results on ICE-India, averaged over 3 runs.

	AP					
Feature	AutoG.	AutoID	MnlG.	AutoID +MnlG.	CGEdit	MnlG. +CGEdit
Non-init. exist. *there*	46.56	41.32	53.16	51.84	61.11	59.56
Focus *itself*	39.99	40.16	74.76	72.76	78.12	75.14
Focus *only*	24.23	32.74	40.04	28.12	41.10	44.31
Zero copula	01.78	04.96	02.05	04.19	03.88	02.95
Left dislocation	02.78	05.70	25.78	09.47	23.07	26.63
Res. subject pronoun	03.68	03.57	21.72	07.55	20.64	20.50
Res. object pronoun	00.24	01.58	02.47	00.93	02.96	05.66
Topic. object (arg.)	02.04	15.95	06.99	02.13	06.00	10.16
Topic. non-arg. const.	01.11	02.53	03.78	02.26	02.65	06.10
Invar. tag *no/na/isn't it*	03.89	04.96	26.95	20.26	37.26	42.18
Macro average	12.63	15.24	25.77	19.95	27.48	29.32

Table 5: AP results on ICE-India, averaged over 3 runs.

	Prec@100					
Feature	AutoG.	AutoID	MnlG.	AutoID +MnlG.	CGEdit	MnlG. +CGEdit
Non-init. exist. *there*	78.33	74.00	86.00	82.00	84.33	87.00
Focus *itself*	15.67	18.67	28.00	25.00	28.00	28.00
Focus *only*	34.33	41.33	48.33	39.67	45.00	48.33
Zero copula	03.33	01.67	03.33	05.00	03.00	05.33
Left dislocation	08.33	18.33	46.33	27.00	42.67	42.00
Res. subject pronoun	09.67	13.67	39.00	24.67	36.00	31.67
Res. object pronoun	00.00	01.00	03.67	01.67	04.67	08.33
Topic. object (arg.)	05.67	03.00	15.00	06.67	12.33	19.33
Topic. non-arg. const.	01.33	01.00	07.33	06.33	07.00	13.67
Invar. tag *no/na/isn't it*	12.67	06.00	39.33	25.00	62.00	73.00
Macro average	16.93	17.87	31.63	24.30	32.50	35.67

Table 6: Prec@100 results on ICE-India, averaged over 3 runs.Prec@100 results on CORAAL. Note that if there are less than 100 instances of a certain feature (e.g. *finna* occurs only 35 times in this dataset, confirmed via keyword search), then its Prec@100 score will have an upper bound of less than 1.

	Prec@100		
Feature	Mnlg.	CGEdit	Mnlg. +CGEdit
Zero possessive -*'s*	030.0	071.0	088.0
Zero copula	089.0	100. 0	100.0
Double marked	024.0	031.0	045.0
Habitual *be*	100.0	100.0	100.0
Resultant *done*	089.0	097.0	097.0
finna	035.0	035.0	035.0
come	011.0	016.0	015.0
Double modal	014.0	014.0	013.0
Negative concord	100.0	096.0	077.0
Neg. auxiliary inversion	078.0	096.0	089.0
Non-inverted neg. concord	009.0	010.0	012.0
Preverbal negator *ain't*	100.0	100.0	100.0
Zero 3rd p sg pres. tense -*s*	096.0	100.0	098.0
is/was-generalization	063.0	100.0	100.0
Zero plural -*s*	017.0	062.0	059.0
Double-object construction	050.0	030.0	018.0
Wh-question	079.0	088.0	058.0
Macro average	057.9	067.4	064.9

Table 7: Prec@100 results on CORAAL. Note that if there are less than 100 instances of a certain feature (e.g. *finna* occurs only 35 times in this dataset, confirmed via keyword search), then its Prec@100 score will have an upper bound of less than 1.

	Prec@100		
Feature	Mnlg.	CGEdit	Mnlg. +CGEdit
Zero possessive -*'s*	011.0	042.0	026.0
Zero copula	097.0	099.0	100.0
Double marked	053.0	049.0	095.0
Habitual *be*	078.0	099.0	097.0
Resultant *done*	093.0	100.0	100.0
finna	000.0	000.0	000.0
come	001.0	050.0	082.0
Double modal	004.0	005.0	004.0
Negative concord	100.0	100.0	100.0
Neg. auxiliary inversion	093.0	100.0	100.0
Non-inverted neg. concord	015.0	024.0	056.0
Preverbal negator *ain't*	100.0	100.0	100.0
Zero 3rd p sg pres. tense -*s*	100.0	100.0	100.0
is/was-generalization	100.0	100.0	100.0
Zero plural -*s*	024.0	070.0	096.0
Double-object construction	036.0	028.0	020.0
Wh-question	093.0	090.0	088.0
Macro average	058.7	068.0	074.4

Table 8: Prec@100 results on FWP. Note that if there are less than 100 instances of a certain feature (e.g. *finna* occurs 0 times in this dataset, confirmed via keyword search), then its Prec@100 score will have an upper bound of less than 1.

	ROC-AUC Standard Deviation					
Feature	AutoG.	AutoID	MnlG.	AutoID +MnlG.	CGEdit	MnlG. +CGEdit
Non-init. exist. *there*	03.29	00.69	00.65	07.39	01.89	08.42
Focus *itself*	03.38	00.54	00.42	00.45	00.47	00.03
Focus *only*	06.40	01.59	00.66	01.25	02.74	00.48
Zero copula	04.63	03.80	07.95	04.71	06.87	01.04
Left dislocation	07.90	01.83	01.24	16.00	01.62	00.78
Res. subject pronoun	04.62	07.10	00.39	17.13	04.77	05.24
Res. object pronoun	04.73	06.15	05.66	07.79	01.77	00.70
Topic. object (arg.)	06.20	02.88	10.93	06.49	04.89	05.39
Topic. non-arg. const.	03.25	05.52	03.87	01.79	05.57	03.31
Invar. tag *no/na/isn't it*	07.64	04.35	03.04	01.59	10.77	04.97
Macro average	05.20	03.45	03.48	06.46	04.14	03.04

Table 9: Standard deviation of ROC-AUC results on ICE-India over 3 runs.

	AP Standard Deviation					
Feature	AutoG.	AutoID	MnlG.	AutoID +MnlG.	CGEdit	MnlG. +CGEdit
Non-init. exist. *there*	09.52	03.07	04.32	15.13	09.13	08.09
Focus *itself*	09.87	11.30	03.44	08.26	04.30	08.19
Focus *only*	08.36	02.62	05.68	08.01	04.74	00.43
Zero copula	01.79	05.45	01.22	02.07	01.50	01.36
Left dislocation	00.80	01.31	04.90	05.84	01.36	00.78
Res. subject pronoun	00.70	03.12	07.30	05.54	08.82	04.91
Res. object pronoun	00.07	01.77	00.72	00.89	00.65	01.83
Topic. object (arg.)	01.29	25.05	02.46	00.57	01.31	01.18
Topic. non-arg. const.	00.13	01.93	00.99	00.96	00.93	00.39
Invar. tag *no/na/isn't it*	00.73	03.02	13.96	07.02	25.90	16.98
Macro average	03.33	05.86	04.50	05.43	05.86	04.41

Table 10: Standard deviation of AP results on ICE-India over 3 runs.

	Prec@100 Standard Deviation					
Feature	AutoG.	AutoID	MnlG.	AutoID +MnlG.	CGEdit	MnlG. +CGEdit
Non-init. exist. *there*	08.02	07.00	04.00	12.90	04.16	03.61
Focus *itself*	03.51	04.04	00.00	31.19	00.00	00.00
Focus *only*	06.03	04.16	06.43	07.13	05.57	05.51
Zero copula	01.15	01.53	02.08	01.30	03.00	01.53
Left dislocation	04.04	06.66	05.20	34.27	05.51	02.65
Res. subject pronoun	04.51	03.21	14.73	21.81	17.69	07.09
Res. object pronoun	00.00	00.00	01.15	02.89	00.58	02.52
Topic. object (arg.)	03.79	02.65	05.20	06.48	02.31	03.51
Topic. non-arg. const.	00.58	00.00	03.21	07.57	03.00	03.79
Invar. tag *no/na/isn't it*	04.16	04.36	16.20	38.91	25.51	17.09
Macro average	03.58	03.36	06.15	16.45	06.73	04.73

Table 11: Standard deviation of Prec@100 results on ICE-India over 3 runs.

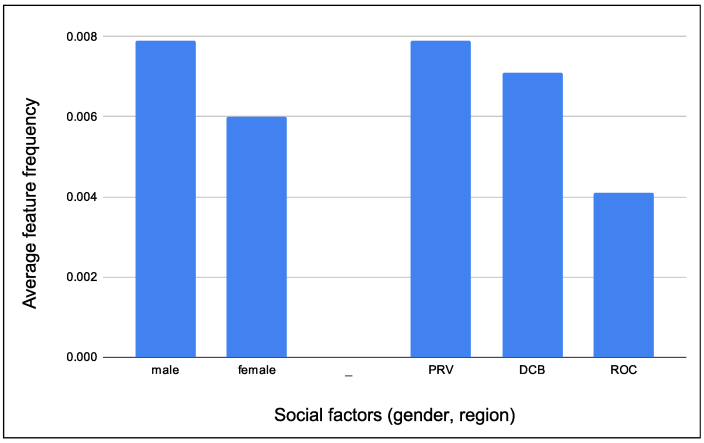


Figure 4: Confirming results from Koenecke et al. (2020). Examined 17 features over entire DCB, PRV, and ROC subcorpora. We find higher feature frequencies among male speakers than female speakers; and highest feature frequency in Princeville, followed by DC, and then Rochester.

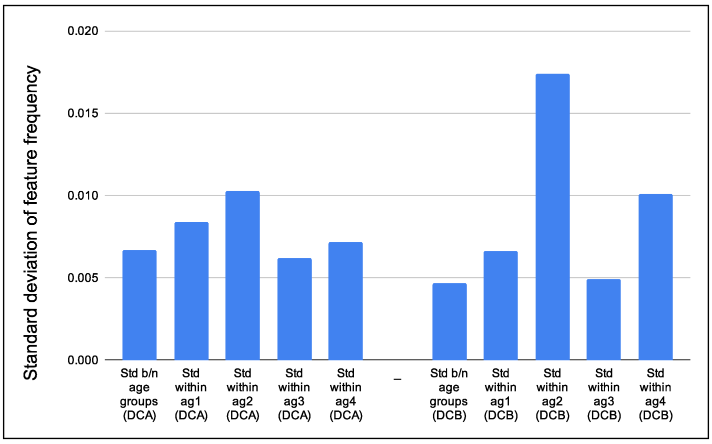


Figure 5: Confirming results from Cukor-Avila and Balcazar (2019). Examined 3 features over files specified in their study from DCA and DCB subcorpora. Ag1 corresponds to ages less than 20, ag2 corresponds to ages 20-29, ag3 corresponds to 30-50, and ag4 corresponds to 50+. We find that standard deviation between speakers in an age group is equal to or larger than standard deviation between age groups.

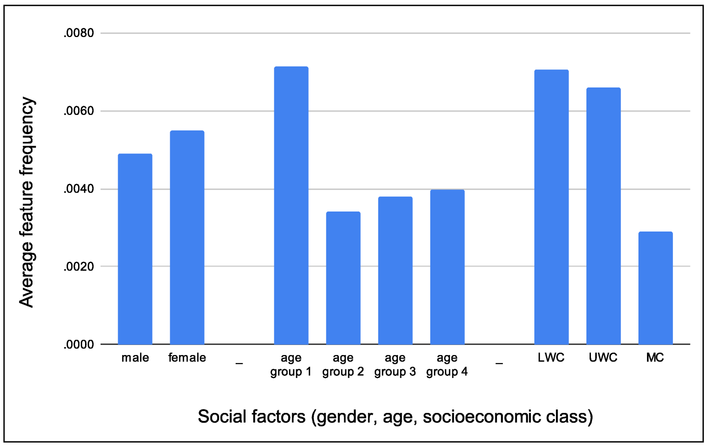


Figure 6: Confirming results from Grieser (2019). Examined 14 features over files specified in their study from DCA subcorpus. Age group 1 corresponds to ages less than 20, age group 2 corresponds to ages 20-29, age group 3 corresponds to 30-50, and age group 4 corresponds to 50+; the socioeconomic classes, from left to right, are Lower Working Class, Upper Working Class, and Middle Class. We find that age and socioeconomic status are negatively correlated with feature use. We find that men have a slightly lower average feature frequency; however, when looking at all of CORAAL for all of our features, we confirm that men have a higher average feature frequency. This is perhaps an example of how small sample size can skew results.