Leveraging Syntactic Dependencies in Disambiguation: The Case of African American English

Wilermine Previlon*, **Alice Rozet***, **Jotsna Gowda***, **William Dyer***
Kevin Tang[♣]*, **Sarah Moeller***
*University of Florida
Department of Linguistics, College of Liberal Arts and Sciences
{wprevilon, arozet, jotsna.gowda,wgdyer,smoeller}@ufl.edu
[♣]Heinrich Heine University Düsseldorf
Department of English Language and Linguistics, Faculty of Arts and Humanities
kevin.tang@hhu.de

Abstract

African American English (AAE) has received recent attention in the field of natural language processing (NLP). Efforts to address bias against AAE in NLP systems tend to focus on lexical differences. Whenever the structural uniqueness of AAE is considered, the solution is often to remove or neutralize the differences. This work leverages knowledge about the unique morphosyntactic structures to improve automatic disambiguation of habitual and non-habitual meanings of "be" in naturally produced AAE transcribed speech. Both meanings are employed in AAE but examples of Habitual *be* are rare in the already limited AAE data. Generally, representing contextual syntactic information improves semantic disambiguation of habituality. Using an ensemble of classical machine learning models with a representation of the unique POS and dependency patterns of Habitual *be*, we show that integrating syntactic information improves the identification of habitual uses of "be" by about 65 F_1 points over a simple baseline model of n-grams, and as much as 74 points. The success of this approach demonstrates the potential impact when we embrace, rather than neutralize, the structural uniqueness of African American English.
Keywords: African American English, habitual *be*, syntactic dependencies, semantic disambiguation

1. Introduction

African American English (AAE) is considered a low-resource language, facing the challenge of inadequate annotated data for training natural language processing (NLP) models. While efforts to enhance annotation have shown promise in improving NLP performance on AAE (Dacon, 2022; Masis et al., 2023), the process remains time-consuming, expensive, and often demands expertise that is not always available (Larimore et al., 2021). Furthermore, AAE lacks access to the same host of language-specific NLP tools available to dominant varieties such as Mainstream American English (MAE). Therefore, there is a need to develop NLP systems capable of recognizing AAE features and automating the annotation process, reducing reliance on large amounts of training data (Blodgett et al., 2018; Luca and Streiter, 2003).

This paper describes a syntactically informed classifier designed for the automatic detection and disambiguation of AAE's habitual *be*. Word sense disambiguation usually involves mapping to senses from a lexical resource like the Princeton Wordnet (Princeton University, 2010). Our work achieves the same goal (automatically identifying the correct sense of a word) with one sense of *be*, but the mapping step is missing for the simple reason that WordNet does not include the habitual sense. The use of *be* as a habitual aspect marker in AAE presents challenges for disambiguation due to its infrequency and lexical similarly to more common uses of the word. Building on previous research that explored part-of-speech (POS) tagging as a disambiguation parameter (Santiago et al., 2022), our study delves deeper into the complex syntactic patterns co-occuring with habitual and non-habitual meanings of "be" in African American speech. Using linguistic data from two oral history corpora of AAE speakers, we demonstrate that integrating dependency parsing enhances automatic detection compared to the baseline and previous POS-only approaches. The success of our approach highlights the importance of incorporating unique linguistic features of marginalized language varieties into NLP models.

2. Related Work

Recently, there has been a surge in NLP research for AAE. Studies have explored dependency parsing (Blodgett et al., 2018), POS-tagging (Dacon, 2022; Jørgensen et al., 2016), hate speech classification (Harris et al., 2022; Sap et al., 2019), automatic speech recognition (Koenecke et al., 2020; Martin and Tang, 2020), dialectal analysis (Blodgett et al., 2016; Dacon, 2022; Stewart, 2014) and

Senior and corresponding authors: Kevin Tang and Sarah Moeller

LREC-COLING 2024, pages 10403–10415

feature detection (Masis et al., 2022; Santiago et al., 2022). Projects such as these rely heavily on large amounts of labeled data, however, little research is dedicated to optimizing the disambiguation and annotation process.

When considering methods of mitigating bias in NLP, AAE's unique morphosyntactic structures are often neglected. Semantic context and lexical choice are more commonly accounted for (Barikeri et al., 2021; Cheng et al., 2022; Garimella et al., 2022; Hwang et al., 2020; Kiritchenko and Mohammad, 2018; Maronikolakis et al., 2022; Silva et al., 2021), but when focusing on improving a model's understanding of AAE, research often involves removing its morphological features (Tan et al., 2020) or translating between MAE and AAE (Ziems et al., 2023). In contrast, our work leverages AAE's morphosyntactic differences to improve disambiguation of habitual and non-habitual "be", rather than neutralizing the uniqueness of AAE.

Internet corpora, particularly Twitter data, dominate the data used for NLP research on AAE (Deas et al., 2023; Dacon, 2022; Harris et al., 2022; Blodgett et al., 2016, 2018; Stewart, 2014; Jones, 2015; Jørgensen et al., 2016; Koufakou et al., 2020). AAE is extremely pervasive online, however, it is important to recognize its use in other domains such as everyday speech. To examine organic and authentic utterances, we use a corpus of oral history interviews. These conversations provide extensive linguistic evidence, allowing for reliable statements about language use that are not always possible with other corpora (Fasold, 1969; Roller, 2015). Oral histories also serve as a valuable resource for studying AAE in various contexts, including higher education and healthcare, and for examining variations based on factors like class or gender (Syrquin, 2006; Hudley et al., 2022; Weldon, 2021; Morgan, 1994; Adolphs et al., 2004). Through this, our dataset expands beyond the conventional scope of NLP research on AAE and offers a novel perspective.

3. African American English and Invariant/Habitual *Be*

The habitual *be* is a well-documented linguistic phenomenon in AAE. It is described as an aspectual marker denoting a recurring, or habitual, action (Green, 2002; Fasold, 1969). In contrast to other instances of "be", the habitual *be* is invariant, meaning it never changes to agree with the subject. As shown in the example below, the MAE sentence requires the verb "be" to agree with "I" resulting in "I am" rather than "I be". Additionally, the adverb "usually" is required to indicate the event is recurring. AAE, on the otherhand, does not require "be" to change form nor does it require additional adverbs.

AAE: I be in my office by 7:30.
MAE: 'I am usually in my office by 7:30'

Although relatively infrequent, habitual *be* is employed regularly by AAE speakers (Blodgett et al., 2016) as well as speakers of 90 other Englishes (Kortmann, 2020). No matter its rareness, accurately detecting unique aspects of AAE helps mitigate biases within NLP as incorrect analysis of AAE (Dacon et al., 2022) has led to ungrammatical generations by AI (Deas et al., 2023) and biased toxicity detection (Harris et al., 2022). Thus, further exploration is needed into understand its linguistic patterns from a computational perspective.

Habitual *be* is a term commonly used throughout linguistic literature, however non-habitual instances of an invariant *be* have also been documented. Such instances can indicate future tense ("It be October 15th before we can move in." (Fasold, 1969)), equivalence ("I be the truth" (Alim, 2004)), or emphasis ("New Haven be lit" (Harris, 2019)). Similar sentences appeared in our corpus, containing functional but no structural differences from habitual *be*. As such, these instances cannot be disambiguated using dependency parsing or POS tags. Therefore, we use habitual *be* broadly to include non-habitual instances of the invariant *be* that differ by meaning but not syntactic structure, while non-habitual refers to any non-invariant form of *be*.

4. Data and Annotation

Despite the language's recent popularity, natural spoken data for AAE is limited, resulting in a research gap within AAE-focus corpus linguistic studies and data-dependent NLP (Martin, 2022; Dacon et al., 2022). Only in 2018 was the first corpus of African American speech, the Corpus of Regional African American Language (CORAAL) (Kendall and Farrington, 2021), made available. The Joel Buchanan Archive of African American Oral History (University of Florida, 2023) is another large and growing collection of AAE speech data, containing over 700 oral history interviews with African Americans throughout the southern United States.

Our analysis of the syntactic structures associated with habitual *be* was conducted on a data set of 250 manually annotated instances of "be" in the Joel Buchanan Archive (Moeller et al., to appear). This annotation was performed by a team trained and tested on their abilities to recognize AAE features. Following an established guideline, they annotated oral history transcripts by identifying sentences containing habitual *be* while listening to an

data set	habitual	non-habitual
analysis	132	118
training	465	3,610
test	52	402

Table 1: Unaugmented data. The analysis data was used to discover dependency patterns correlating with habitual *be*. The training/test data are sourced from a different AAE collection than the analysis data.

accompanying audio recording. Our models are trained and tested on a sample of the CORAAL corpus that had previously been manually annotated for habitual *be* (Martin, 2022). The final size of the training and tests sets are shown in Table 1. These splits are identical for all models and experiments.

5. Analysis of Habitual *Be*

Certain morpho-syntactic structures that strongly correlate to the habitual *be* have been described in the literature (Green, 2002; Fasold, 1972). Beyond these established descriptions, our analysis of dependency parses revealed additional syntactic representations that can be harnessed to improve classification. These syntactic structures were programmed as Boolean True/False rules with true indicating whether the dependency structure was found in the sentence. The output of these rules serve as the training input to machine learning classifiers (see Section 6 on Methodology).

5.1. POS Patterns

In their work, Santiago et al. (2022) leveraged linguistic descriptions and POS tagging to develop a rule-based method for identifying many non-habitual instances of "be". This allowed us to filter out these instances, thereby balancing our dataset by increasing the frequency of habitual *be* examples. Implementing this filtering method, we found that Santiago et al. (2022) did not account for patterns of phonetic variation within the transcripts. As a result of this discovery, we modified their POS1 to include the words "wanna", "trynna", "gonna", and "gotta" in addition to the more standardized representations already covered by the filter ("want to", "trying to", "going to", "got to"). We refer the reader to Santiago et al. or Appendix 15.1 for the full list of POS structures used in this method, including our additional previously undocumented "ad-hoc" patterns.

5.2. Dependency Syntax Patterns

Our investigation into NLP representation of the habitual *be* involved a corpus analysis that reveal specific dependency patterns strongly correlated to either the habitual *be* or non-habitual "be". The aforementioned sample of 250 manually annotated sentences was parsed using spaCy (ver: 3.2.2) (Honnibal et al., 2020), with 118 sentences containing a non-habitual "be" and 132 sentences containing a habitual one. From this, two major patterns emerged for each class.

Patterns of Non-Habitual Be. The two primary patterns for the non-habitual "be" are named SynPar1 and SynPar2. Both are centered on the arrangement of auxiliary verbs relative to the "be". SynPar1 occurs when the POS of "be" is tagged as an auxiliary and it has a child with a dependency relation of auxiliary. Synpar2 is similar to SynPar1 but has a *sibling* with a dependency relation of auxiliary. There were 36 instances of SynPar1 and 22 instances of SynPar2. Notably, the auxiliary child or sibling is often a modal, suggesting that modals are unlikely to co-occur with the habitual *be*.

(2) **SynPar1**: It would be so much excitement

(3) **SynPar2**: Those barriers really may not be holding them back.

Patterns of Habitual *Be*. SynPar3 and SynPar4 are the two patterns associated with the habitual *be*. SynPar3 occurs when "be" is labeled as an auxiliary and has a head that is labeled as a verb. This pattern is common among sentences in the form of "<PRONOUN> be <VERB>-ing" as well as cases where "be" precedes a passive verb both of which have been documented in linguistic literature (Fasold, 1969; Green, 2002). SynPar4 occurs when "be" is labeled as a verb, contrasting its more frequent label of auxiliary. This pattern is also consistent with existing literature, which finds habitual *be* often preceding non-verbal predicates.

(4) **SynPar3**: Like a person be lying.

(5) **SynPar4**: They used to say that she always be a spinster

Post-Hoc Syntactic Rules. After creating a preliminary model with the training data, an error analysis revealed five additional instances that might improve the disambiguation of "be". These patterns were not captured by the syntactic patterns described above but appeared frequently in incorrectly labeled sentences. We call them post-hoc due to their development from a bottom-up corpus analysis independent from established linguistic literature or expectations.

(A1) "Don't" preceding "be" indicates habituality except for in imperative sentences.

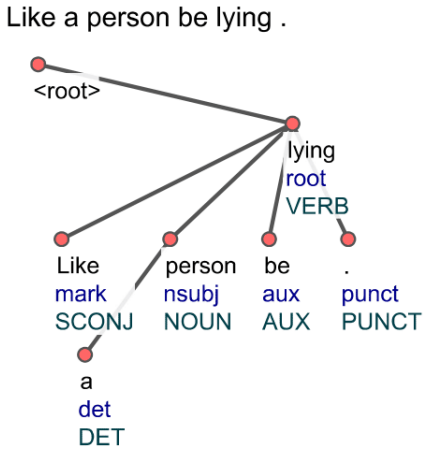

Figure 1: Illustration of SynPar3.

(6) She don't be working out

(A2) When followed by interjections, conjunctions, determinants, proper nouns, and punctuation, "be" tends to be non-habitual.

(7) It wasn't as big as I thought it would be.

(A3) If the "be" is preceded by a pronoun in the sentence, where the words between the pronoun and "be" are not auxiliaries, verbs, or particles, it tends to be habitual.

(8) I just be liking the beat to a hip hop song

(A4) "Be" tends to be habitual when it is followed by a verb ending in -ing and is not preceded by an auxiliary verb, "to", or any of the words following words displaying phonetic variation of the infinitive "to": "gonna", "gotta", "wanna", or "tryna".

(9) Elysa be showing me some work

(A5) When 'be' is preceded by a word ending in -n't that is not "don't", it tends to be non-habitual.

(10) I mean, you can but you wouldn't be too successful with it

6. Methodology

To overcome the issue of limited training data, we approach AAE as a distinct linguistic system with its own set of morphosyntactic rules, exploiting its unique structures to improve semantic disambiguation. Grammatical knowledge about AAE comes from two sources: (1) published linguistic literature about the habitual *be* and (2) linguistic patterns we discovered through our own analysis of AAE speech data. Both sources of information are scripted into True/False statements to identify whether each pattern appears in a sentence containing "be". The output of all the rules is then used as input to a machine learning classifier which labels each instance of "be" as either habitual or non-habitual.

Each sentence containing a "be" is treated as an individual data point. Fewer than 5 sentences contained more than one "be" in the original dataset and so they were omitted. The habitual *be* disambiguation task is treated as a text classification problem where the input is a sequence of True/False values $\vec{x} = (x_1, \ldots, x_n)$ representing properties of each data point. Each data point is classified with one output label ($\hat{y}$) that is either "habitual" or "non-habitual".

Due to limited AAE data, a k-fold approach was adopted, with the reported results representing the average of the 10 folds. The training/test data sets were created with a 90/10 division of the CORAAL data and remained the same for all experiments and models except the Transformer which used a development data taken from the training set that was the same size as the test set. Thus, the Transformer had a 80/10/10 split of training, development and testing.

6.1. Classifiers

In cases of limited training data, deep learning is not an ideal approach (Marcus, 2018). Instead classical machine learning often achieves higher performance (e.g., Jiang et al., 2023). Because of this, we used `scikit-learn` (Pedregosa et al., 2011) to implement an ensemble model that combines logistic regression (LR), multilayer perceptron (MLP), and support vector classification (SVC) modes. This decision was also motivated by Santiago et al. (2022) who demonstrated ensemble models performed better than individual models. The 'balanced' parameter was added to the LR and SVC models because the two classes were severely unbalanced; however, the MLP model has no such parameter. Voting was set to "soft", averaging the models' probability predictions for habituality rather than selecting by frequency of class. N-grams, which served as the input for the baseline model, were generated using `NLTK` (ver: 3.6.5) (Bird et al., 2009). The tools were used consistently in all cases.

For the Transformer, we implemented the *fairseq* (Ott et al., 2019) version with parameter modifications that have been shown to be successful in low-resource NLP tasks (Wu et al., 2021)[1].

[1] 4 encoder-decoder layers, 4 self-attention heads, 256 embedding size, 1,024 hidden size of feed-forward layer, layer normalization before self-attention, decoding left-to-right in a greedy fashion.

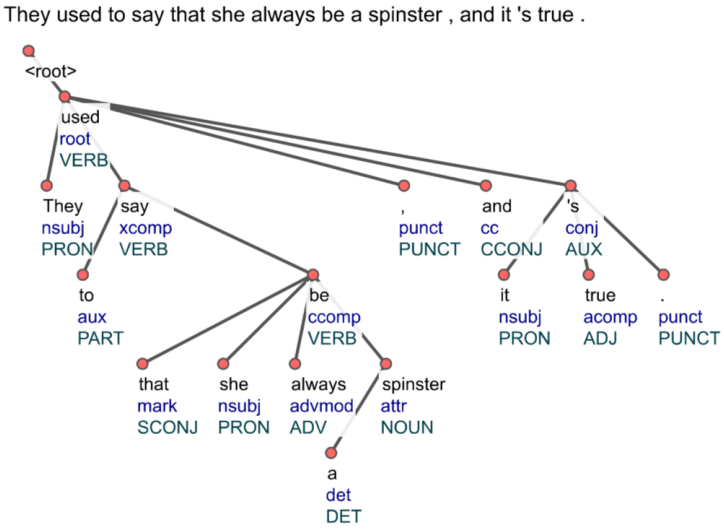

Figure 2: Illustration of SynPar4.

6.2. Experiments

We conducted multiple experiments to investigate how varying amounts of syntactic information affect the system's ability to recognize the habitual *be*. The experiments derive from the hypothesis that incorporating, rather than neutralizing, AAE linguistic structures into NLP systems will improve classification rates, thereby overcoming limitations of scarce and unbalanced AAE data.

Part Of Speech (POS). The POS experiment leverages previous work by Santiago et al. (2022) (see Section 5.1) and incorporates POS patterns described in published literature for both habitual and non-habitual instances of "be". For example, given that many instances of habitual *be* are preceded by a pronoun and followed by verbs ending in "-ing", an absence of these patterns suggest a non-habitual use of "be". These patterns were translated into Boolean True/False rules, e.g. "the word preceding 'be' is NOT a pronoun and the word following 'be' is NOT verb ending in -ing" and the output of the rules serve as the input to the models. See Appendix 15.1 for these patterns.

Dependency Patterns (DEP). This approach is similar to the POS experiments but tests whether AAE-informed syntactic patterns increase classification rates. As described in Section 5.2, the presence or absence in each sentence of the four syntactic environments, two each correlating to the habitual and the non-habitual *be*, serve as input to the classifier model .

Post-hoc Rules (PH). In Section 5.2, we mention post-hoc rules that were uncovered during an error analysis of the combined POS+DEP models. In this fifth experiment, we also translate these rules to values that indicate the presence or absence of these structures in each sentence. The results are used to evaluate the impact that our error analysis insights have on classification.

Interaction of Patterns (INT). After conducting error analysis of the previous experiments using augmented data (see below), we discovered that some patterns interacted with each other in a way that indicated habituality beyond their individual scopes. Semantic disambiguation requires a modeling of how the presence or absence of the different structures in one sentence impact disambiguation of habitual *be*.

The first interaction we discovered is that sentences that do not contain any of the POS rules described in Section 5.1 are much more likely to be habitual. Thus, an additional rule (R1) was created to flag the lack of any POS pattern.

Secondly, multiple syntactic patterns may appear in one sentence, which can lead to conflicting likelihoods of habituality that the individual patterns indicate. In these cases, we found that the presence of one pattern often dominates, meaning that when one pattern is present in the sentence, its indication of habituality or non-habituatliy overpowers the conflicting indications of other structures that are present in the sentence. For example, our DEP rule called SynPar3 (see Section 5.2) indicates habituality; however, if a sentence has been tagged by *any* POS rule (aka, R1 == 0), the "be" in the sentence is more likely to be non-habitual. From this, we decided to trigger a change of the Boolean value of the less dominant pattern to make it appear as though the structure

flagged by that rule does not occur in the sentence. A side effect is that the classifier sees the less dominant rules as occurring less frequently than they actually do in the data; however we hypothesized this would make the dominating rule a stronger predictor of habituality, thus reducing the conflict between the two patterns in one sentence that indicate opposed values of habituality.

Baseline (+ngrms) A baseline model trained on simple textual data was used to compare against linguistically informed models. The input for this baseline model are bigram and unigram representations of the window of eight words surrounding each "be" (four before and four after). This method generated a total of 453 ngrams for each sentence. Any data outside this window was ignored, as it is unlikely to have a correlation with habituality.

We attempted to augment subsequent experiments with this baseline, however we found that adding all 453 ngrams reduced performance. It is likely that additional information overwhelmed the model while adding little useful information. To avoid this confusing weight towards ngrams, we instead added only the predicted habituality values from the baseline model to other experiments.

Simple Part-Of-Speech Window (+win). Next, we test whether syntactic information that has not been informed by AAE distinctive structures might improve disambiguation. We extract the POS tags of the same eight-word window surrounding "be" and added them to the input. For example, the POS window for the AAE sentence in example (1) in Section 2 would be `pronoun preposition pronoun noun`. The tags are considered "simple" in that the model is given the POS of each word without being told which patterns correlate to a habitual and non-habitual "be".

Data Augmentation (+aug) To reduce the bias of the highly imbalanced training data as seen in Table 1, we replicate Santiago et al. (2022)'s data augmentation methodology which had a positive impact on disambiguation of habitual *be*. This method increases the quantity of training data available for our model. The augmented sentences were subjected to the same filtering mechanism used by Santiago et al. The filtering mechanism eliminate sentences that do not align with habitual patterns. This filtering process ensures that the training data is not only increased overall, but that the augmented data also increases the size of the habitual class, futher improving the balance of training data.

7. Results

Each experiment was replicated on a Transformer (Vaswani et al., 2017) and the results of the two models were compared. After generating results, we examined the mean and range of all the F_1-scores for habitual and non-habitual instances, as well as their weighted average. The results of the experiments are displayed in Tables 2 and 3. Row 1 shows the simple baseline which provides a comparison to training a model with only textual n-grams. The baseline+win is the output of the baseline model with the addition of a sequential 8-word window of the POS tags that occur around the "be". Rows 2-6 compare models trained on various amounts of syntactically-informed representations of AAE sentences containing "be". These rows display the combinations of the part-of-speech structures correlating to habituality (POS), the dependency structures correlating to habituality (DEP), the post-hoc rules that further refine the capture of dependency patterns (PH), and the simplication of co-occuring patterns that conflict in habituality (INT). The columns indicate the addition of information that is not syntactically informed. These include the output of the simple baseline n-grams models (+ngrms), an 8-word window of POS tags surrounding the "be" (+win), and the results of synthetic data augmentation with limited filtering of the augmented data (+aug).

The simple baseline model feeds in only the most basic information, unigrams and bigrams, to the model. This baseline is insufficient with the limited available natural data. The baseline Ensemble did much better than the Transfomer, reaching a weighted average of .86 F_1-score, yet only .29 F_1-score for habitual *be*. However, feeding in a simpler representation of the data brings a large jump in performance. We created the simple representation by generating the parts of speech tags for the same window of eight words surrounding each "be" that the ngrams represented. This unsophisticated syntactic representation (+win) improved the overall F_1-score to .92 with the Ensemble and .94 with the Transformer. Most notably, the habitual *be* F_1-score increased to .66 with the Ensemble and made a nearly seventy-five point jump with the Transformer.

The other models tested the impact of representing the sentences by their syntactic structures. For comparison, we also tested these models with the addition of the baseline's output and the window of POS tags that we used in the baseline model. In each test, we added more syntactic information. Model 2 (POS) had solely the POS rules described in literature and in Santiago et al. (2022). Model 3 (DEP) only represented the dependency patterns. Model 4 (POS+DEP) combined the POS

and the dependency patterns. Model 5 had the post-hoc rules A1-A5 as well, and Model 6 was the final model where the conflict of co-occuring patterns was resolved. Though the overall F_1-score increases slowly with each subsequent model, the biggest changes are visible when observing the habitual F_1-score in Table 3. The model had so few habitual instances to train on relative to the non-habitual instances, and therefore had a harder time classifying them. Ultimately, the best results without data augmentation come from Model 5 with the window of POS tags, which displays a 0.96 weighted average F_1-score and a 0.83 habitual F_1-score with the Ensemble model. This is a surprising result; we expected that Model 6, with the added rule interactions, would perform the best.

Data augmentation increased the number of habitual sentences relative to non-habitual. With these additional sentences, both models had more data to train on and became a better predictor of habituality, increasing the maximum habitual F_1-score from 0.83 to 0.95 with the Transformer. With the Ensemble, this method did lead to a decrease in the weighted average F_1-score overall.

It is clear that leveraging the syntactic information that distinguishes the two meanings of "be" is a successful approach when training data for AAE is limited. The best models are proficient in classifying sentences as habitual or non-habitual, as evidenced by the highest F_1-score of 0.95 for habitual and the weighted average F_1-score, considering both habitual and non-habitual classes, which also reached 0.96 by the Ensemble model. The syntactically informed models outperformed the simple baseline and the baselines with added types of information (+win and +aug) that are not informed by the unique syntax-semantic interface of AAE. Table 3 shows a general trend that as more structured linguistic information is added, the disambiguation of habitual *be* improves. The Transformer seems to converge with additional data and minimal syntactic information while the Ensembles shows a more gradual improvement. Interestingly, the POS patterns tend improve the Transformer's ability to recognize habitual *be* more than the dependency patterns while with the Ensemble the dependency patterns tend to help.

The augmented data which balanced the training data by increasing the proportion of habitual sentences shows the positive impact of additional training data. The additional training data made no significant difference on the overall performance shown in Table 2, but it yields noticeable improvement on the habitual class, as seen in Table 3. This holds for both the Ensemble of classical machine learning models and the Transformer, but the effect is slightly stronger for the latter. The augmented baseline Transformer performs nearly as well as the best model, suggesting that deep learning has a higher dependence on data abundance. It's worth noting that while data augmentation bolstered our results, the utilization of the filtering mechanism in only the augmented data might have introduced some bias into the results.

8. Discussion & Error Analysis

The best models achieve a mean F_1-score of 0.95 for the habitual class, marking a significant improvement over both the system reported in Santiago et al. (2022) and the baseline. Notably, models informed by AAE syntax consistently outperform baseline models lacking such information. This trend is evident in the gradual addition of POS tags, dependency patterns, and post-hoc syntactic rules, demonstrating that the inclusion of syntactic information enhances the model's performance in disabmiguating AAE's habitual *be*.

Moreoever, there is potential for a comprehensive data-driven approach that leverages relevant AAE syntactic structures. For instance, one could exhaustively generate syntactic rules that, once computed for their feature importance, can then be used as features in a classifier. Given our limited resources, this was not feasible and we instead opted to perform a small corpus analysis focusing on relevant dependency patterns.

Our analysis confirms the prevalence of patterns described by linguistic scholars of AAE while also revealing previously unreported patterns. This underscores the importance of leveraging existing linguistic literature and expert knowledge in the development of NLP systems for low-resource languages.

Our findings corroborate previous research, such as Santiago et al. (2022), demonstrating the effectiveness of data augmentation in enhancing NLP performance for low-resource tasks. Specifically, augmentation had a significant impact on the habitual class, ranging from 0.11 to 0.19 (cf. +aug and +win in Table 3). However, for the non-habitual class, augmentation lead to a small decrease in F_1-scores (not reported in the tables). The POS+DEP+PH+INT models had a mean F_1-score of 0.97 without augmentation that dropped to 0.95 after its implementation. A similar decrease can be seen with the POS+DEP+PH models, dropping from an F_1-score of 0.98 to 0.94. While this may be seen as a disadvantage of data augmentation, it is a reasonable performance trade-off for higher classification rates within the minority class.

9. Conclusion

This paper outlined a precise disambiguation model for the habitual *be* in AAE, enabling NLP

		Ensemble				Transformer			
			+ngrms	+win	+aug		+ngrms	+win	+aug
1	baseline	0.86	n/a	0.92	0.80	0.83	n/a	0.94	0.93
2	POS	0.92	0.92	0.94	0.93	0.93	0.92	0.94	0.95
3	DEP	0.94	0.94	0.94	0.88	0.94	0.94	0.95	0.90
4	POS+DEP	0.93	0.94	0.95	0.92	0.95	0.95	0.95	0.95
5	POS+DEP+PH	0.95	0.95	**0.96**	0.94	0.95	0.95	0.95	0.95
6	POS+DEP+PH+INT	0.95	0.95	0.95	0.94	0.94	0.95	0.95	0.95

Table 2: The models' performance on the habitual/non-habitual disambiguation of "be" using the weighted average F_1-scores by the number of true instances for each class. The F_1-scores averaged over 10 folds. Going from top to bottom, the second column indicates an increasing amount of syntactic information added to the models (from only baseline to adding INT). Going across the columns of Ensembles/Transformer from left to right, the second row indicates information added to the models that is not unique to AAE (from none to adding +aug). See Section 6.2 for the meaning of the abbreviations.

		Ensemble				Transformer			
			+ngrms	+win	+aug		+ngrms	+win	+aug
1	baseline	0.29	n/a	0.66	0.78	0.01	n/a	0.74	0.92
2	POS	0.72	0.72	0.75	0.93	0.72	0.72	0.77	**0.95**
3	DEP	0.74	0.73	0.74	0.87	0.72	0.74	0.79	0.89
4	POS+DEP	0.76	0.76	0.79	0.92	0.78	0.78	0.77	**0.95**
5	POS+DEP+PH	0.82	0.81	0.83	0.94	0.79	0.78	0.79	**0.95**
6	POS+DEP+PH+INT	0.81	0.80	0.81	0.94	0.75	0.78	0.79	0.94

Table 3: Classification of the habitual class. F_1-scores averaged over 10 folds. See the caption of Table 2 for explanations of the table.

systems to accurately identify instances of AAE within extensive corpora. This capability not only facilitates advancements in AAE-centered NLP systems, but also holds potential benefits for the African American community across various fields, including health (Lee et al., 2022; Yoon et al., 2023; Davis et al., 2024), education (Wolfe, 2019; Samuel Proctor Oral History Program) and psychology (Berger and Packard, 2022). Although the habitual *be* is a relatively infrequent phenomenon, its detection is crucial for enhancing NLP tools. Oftenly overlooked in lingustic research and tool development, infrequent phenomena play a significant role in mitigating biases toward minority languages. Additionally, the habitual *be* is found in 90 different English varieties (Kortmann et al., 2020) at varying frequencies, suggesting that our work can benefit and serve other minority English varieties as well.

10. Limitations & Future Directions

Better baseline The baseline n-gram model was produced using approximately 400 unigrams and bigrams surrounding the "be". In the later models, we instead used the habituality predictions of the n-gram model as an input, since the number of n-gram features overshadowed the influence of the rule-based features. Future work could instead use the estimated probability of the non-habitual class generated by the baseline n-gram model. Additionally, an alternative baseline model can involve pretrained English embeddings.

Better data To reduce the bias of the highly imbalanced training data, data augmentation was shown to have a positive effect. Further work can further reduce the bias by applying Santiago et al. (2022)'s filtering mechanisms directly to the original (not augmented) data, as well as the augmented data. This filtering method uses the POS patterns that are based on the published descriptive literature (see Section 5.1) but not any that were derived from a bottom-up corpus analysis or a error analysis of a simpler habitual *be* disambiguation model. The advantage of basing the filter only on these patterns is that they are completely precise. These POS patterns are only found with non-habitual sentences.

Feeding back into descriptive linguistics Our features were built upon descriptive linguistic analyses. While we reported the overall performance of the models, a detailed analysis of the importance of each feature was not performed. The newly discovered syntactic patterns and their feature importance can serve as valuable tools to inform the grammatical description of AAE and linguistic theories in general.

Generalisation Our models were trained only on sentences containing one "be", however this project can be extended to analyze sentences containing multiple. Room is also left for a comparative analysis of the performance of a model trained on sentences with multiple "be's" versus sentences with singular.

Inter-annotator agreement Due to the time consuming process of manual annotation, each document was annotated only once. Therefore, we were not able to evaluate the quality of our annotations using inter-annotator agreement.

11. Acknowledgements

This research is part of the project "Reanimating African American Histories of the Gulf South" which was supported by the National Endowment for the Humanities (20200715-PW). We thank the anonymous reviewers for their valuable feedback.

12. Data and code availability

Data and code are available at https://github.com/wilermine/CoLing-LREC-HabitualBe.

13. Contribution statement

SM and KT are the senior and corresponding authors. We follow the CRediT taxonomy[2]. Conceptualization: SM, KT; Data curation: WP, JG, AR; Formal Analysis: WP, JG, AR, SM; Funding acquisition: SM, KT; Investigation: WP, JG, AR, SM, KT; Methodology: SM, KT; Project administration: SM; Resources: SM; Software: WP, JG, AR, SM, KT; Supervision: SM, KT; Validation: JG, AR, SM; Visualization: X; and Writing – original draft: WP, JG, AR, SM, KT and Writing – review & editing: WP, AR, SM, KT.

14. Bibliographical References

Svenja Adolphs, Brian Brown, Ronald Carter, Paul Crawford, Opinder Sahota, Mariya Limerick, Alison Pilnick, Linda Gibson, and Stacy Johnson. 2004. Applying corpus linguistics in a health care context. *Journal of Applied Linguistics*, 1.

H. Samy Alim. 2004. *You know my steez: An ethnographic and sociolinguistic study of styleshifting in a Black American Speech Community*. Duke University Press for the American Dialect Society.

Soumya Barikeri, Anne Lauscher, Ivan Vulic, and Goran Glavas. 2021. Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1941–1955. Association for Computational Linguistics.

Jonah Berger and Grant Packard. 2022. Using natural language processing to understand people and culture. *American Psychologist*, 77(4):525–537.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, Beijing.

Su Lin Blodgett, Lisa Green, and Brendan T. 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, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1119–1130. The Association for Computational Linguistics.

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.

Peter Broadwell, Nicole Davis, and Sunmoo Yoon. 2022. Using Artificial Intelligence to Develop a Lexicon-Based African American Tweet Detection Algorithm to Inform Culturally Sensitive Twitter-Based Social Support Interventions for African American Dementia Caregivers. In *Studies in Health Technology and Informatics*. IOS Press.

Lu Cheng, Nayoung Kim, and Huan Liu. 2022. Debiasing word embeddings with nonlinear geometry. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 1286–1298. International Committee on Computational Linguistics.

[2] https://credit.niso.org/

Jamell Dacon. 2022. Towards a deep multi-layered dialectal language analysis: A case study of African-American English. *Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing*.

Jamell Dacon, Haochen Liu, and Jiliang Tang. 2022. Evaluating and Mitigating Inherent Linguistic Bias of African American English through Inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1442–1454, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Alexis Davis, Joshua L. Martin, Eric Cooks, Melissa J. Vilaro, Danyell Wilson-Howard, Kevin Tang, and Janice Krieger. 2024. From English to "Englishes": A process perspective on enhancing the linguistic responsiveness of culturally tailored cancer prevention interventions. *JMIR Preprints*.

Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American language bias in natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824, Singapore. Association for Computational Linguistics.

Ralph W. Fasold. 1969. Tense and the Form Be in Black English. *Language*, 45(4):763–776. Publisher: Linguistic Society of America.

Ralph W. Fasold. 1972. *Tense marking in Black English: A linguistic and social analysis*. Center for Applied Linguistics.

Aparna Garimella, Rada Mihalcea, and Akhash Amarnath. 2022. Demographic-aware language model fine-tuning as a bias mitigation technique. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Volume 2: Short Papers, Online only, November 20-23, 2022*, pages 311–319. Association for Computational Linguistics.

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

Alysia Nicole Harris. 2019. *The Non-Aspectual Meaning of African American English 'Aspectual' Markers*. Ph.D. thesis, Yale University.

Camille Harris, Matan Halevy, Ayanna Howard, Amy Bruckman, and Diyi Yang. 2022. Exploring the role of grammar and word choice in bias toward African American English (AAE) in hate speech classification. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 789–798, New York, NY, USA. Association for Computing Machinery.

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

Anne H Charity Hudley, Christine Mallinson, and Mary Bucholtz. 2022. *Talking college: Making space for Black language practices in Higher Education*. Teachers College Press.

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

Zhiying Jiang, Matthew Yang, Mikhail Tsirlin, Raphael Tang, Yiqin Dai, and Jimmy Lin. 2023. "low-resource" text classification: A parameter-free classification method with compressors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6810–6828, Toronto, Canada. 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.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2016. Learning a POS tagger for aave-like language. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1115–1120. The Association for Computational Linguistics.

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

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans,*

Louisiana, USA, June 5-6, 2018, pages 43–53. Association for Computational Linguistics.

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

Bernd Kortmann. 2020. *Syntactic Variation in English*, chapter 16. John Wiley Sons, Ltd.

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

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, SocialNLP@NAACL 2021, Online, June 10, 2021*, pages 81–90. Association for Computational Linguistics.

Donghee N. Lee, Myiah J. Hutchens, Thomas J. George, Danyell Wilson-Howard, Eric J. Cooks, and Janice L. Krieger. 2022. Do they speak like me? exploring how perceptions of linguistic difference may influence patient perceptions of healthcare providers. *Medical Education Online*, 27(1):2107470. PMID: 35912473.

Ernesto William De Luca and Oliver Streiter. 2003. *Example-based NLP for Minority Languages: Tasks, Resources and Tools*.

Gary Marcus. 2018. Deep learning: A critical appraisal. *CoRR*, abs/1801.00631.

Antonis Maronikolakis, Philip Baader, and Hinrich Schütze. 2022. Analyzing hate speech data along racial, gender and intersectional axes. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 1–7, Seattle, Washington. Association for Computational Linguistics.

Joshua L. Martin. 2022. *Automatic Speech Recognition Systems, Spoken Corpora, and African American Language: An Examination of Linguistic Bias and Morphosyntactic Features*. Ph.D. thesis, University of Florida.

Joshua L. Martin and Kevin Tang. 2020. Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be". In *Proc. Interspeech 2020*, pages 626–630.

Tessa Masis, Chloe Eggleston, Lisa J Green, Taylor Jones, Meghan Armstrong, and Brendan O'Connor. 2023. Investigating morphosyntactic variation in African American English on Twitter. *Proceedings of the Society for Computation in Linguistics*, 6(1):392–393.

Tessa Masis, Anissa Neal, Lisa Green, and Brendan O'Connor. 2022. Corpus-guided contrast sets for morphosyntactic feature detection in low-resource English varieties. In *Proceedings of the first workshop on NLP applications to field linguistics*, pages 11–25, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Sarah Moeller, Alexis Davis, Wilermine Previlon, Michael Bottini, and Kevin Tang. to appear. Compiling spoken and transcribed corpus of African American language from oral histories.

Marcyliena Morgan. 1994. Theories and politics in African American English. *Annual Review of Anthropology*, 23:325–345.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Princeton University. 2010. "About WordNet.", WordNet. `https://wordnet.princeton.edu/`. Accessed: 2024-03-25.

Katja Roller. 2015. Towards the 'oral' in oral history: using historical narratives in linguistics. *Oral History*, 43(1):73–84.

Samuel Proctor Oral History Program. AALGS Curriculum. Accessed: 2023-10-20.

Harrison Santiago, Joshua Martin, Sarah Moeller, and Kevin Tang. 2022. Disambiguation of

morpho-syntactic features of African American English – the case of habitual be. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 70–75, Dublin, Ireland. Association for Computational Linguistics.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1668–1678. Association for Computational Linguistics.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, Online. Association for Computational Linguistics.

Ian Stewart. 2014. Now we stronger than ever: African-American English syntax in Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 31–37. The Association for Computer Linguistics.

M. Stone. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133.

Anna F. Syrquin. 2006. Registers in the academic writing of African American college students. *Written Communication*, 23(1):63–90.

Samson Tan, Shafiq Joty, Lav Varshney, and Min-Yen Kan. 2020. Mind your inflections! Improving NLP for non-standard Englishes with Base-Inflection Encoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5647–5663, Online. Association for Computational Linguistics.

Textual Optics Lab and History of Black Writing. History of Black Writing Novel Corpus. Accessed: 2023-10-20.

University of Florida. 2023. Joel Buchanan archive.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Long Beach, California, USA. Curran Associates Inc.

Tracey L. Weldon. 2021. *Middle-Class African American English*. Cambridge University Press.

Erin Wolfe. 2019. Natural language processing in the humanities: A case study in automated metadata enhancement. *Code4Lib Journal*, (46).

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1901–1907. Association for Computational Linguistics.

Sunmoo Yoon, Peter Broadwell, Frederick F. Sun, Maria De Planell-Saguer, and Nicole Davis. 2023. Application of Topic Modeling on Artificial Intelligence Studies as a Foundation to Develop Ethical Guidelines in African American Dementia Caregiving. In *Studies in Health Technology and Informatics*. IOS Press.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

15. Appendices

15.1. Rules to filter non-habitual "be"

- POS1: If the word immediately preceding "be" is a modal, adjective, or "to".
- POS2: If the word immediately following "be" is an adjective, while the word immediately preceding "be" is not a personal pronoun nor a noun.
- POS3: If the word immediately following "be" is a preposition or subordinating conjunction, while the word immediately preceding "be" is a singular present verb.
- POS4: If the word immediately preceding "be" is a noun, and the word immediately preceding that noun is an adjective

- POS5: If the word immediately preceding “be” is an adverb, and the word immediately following “be” is either a personal pronoun or determiner.
- POS6: If the word immediately preceding “be” is an adverb, and either the word immediately preceding the adverb is a verb, or modal
- If the word immediately following “be” is a verbal noun, while the word immediately preceding is not a personal pronoun nor a noun.