LREC-COLING 2024

Leveraging syntactic dependencies in disambiguation: the case of African American English

Wilermine Previlon[1], Alice Rozet[1], Jotsna Gowda[1], William Dyer[1], Kevin Tang[1,2], Sarah Moeller[1]

[1] University of Florida, Department of Linguistics, College of Liberal Arts and Sciences

[2] Heinrich Heine University Düsseldorf, Department of English Language and Linguistics, Faculty of Arts and Humanities

African American English

- Spoken primarily by African Americans
- Similarities with Mainstream American English (MAE)
  - MAE: dominant variety of English spoken in the US
- Notable features: the habitual *be*, person/number disagreement, multiple negation, and null copula
- Seen as a lesser or ungrammatical form of English.

African American English and NLP

African American English (AAE) lacks sufficient data for training NLP models

Annotation is time consuming and expensive

Limited language specific NLP tools for the language

How can we develop NLP systems to recognize AAE features and automate the annotation process, reducing reliance on extensive training data?

- Syntactically informed classifier designed for automatic detection and disambiguation of AAE's habitual *be*
- Incorporates dependency parsing to improve POS tagging methods
- Linguistic data comes from two African American oral history corpora
- Success highlights the importance of using linguistic features for low resource languages in NLP

Habitual *Be*

- Indicates a recurring or habitual action (Fasold 1969; Green 2002)
- Typically indicated by “usually” in Mainstream American English
- Understanding AAE mitigates bias in NLP systems (Dacon et al. 2022; Deas et al. 2023; Harris et al. 2022)

I **be** in my office by 7:30

vs

I’m usually in my office by 7:30

Recent Work

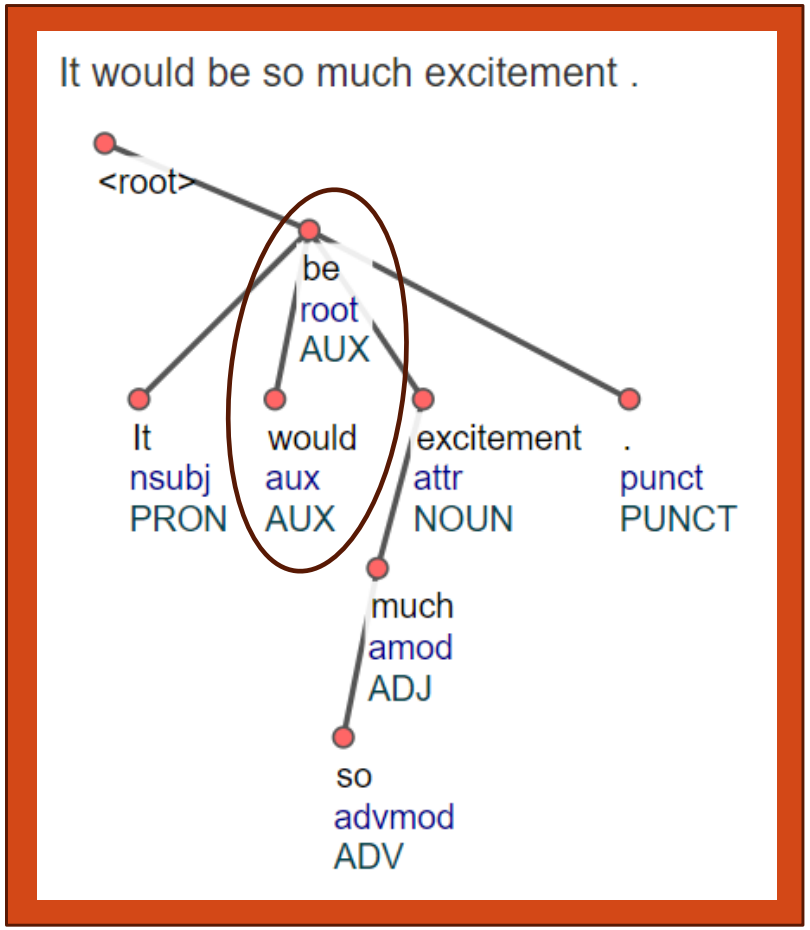
- AAE and NLP
  - Dependency Parsing (Blodgett et al., 2018)
  - POS-tagger (Jørgensen et al., 2016; Dacon, 2022)
  - Hate speech classification (Harris et. al 2022; Sap et al., 2019)
  - Dialectal analysis (Blodgett et al., 2016; Dacon, 2022; Stewart, 2014)
  - Feature detection (Masis et al., 2022; Santiago et al., 2022)
- Roller (2015): Advocates for oral histories as a source of linguistic data
- Santiago et. al (2022): Used POS tags to create machine learning classifier to disambiguate habitual *be*

African American Oral Histories

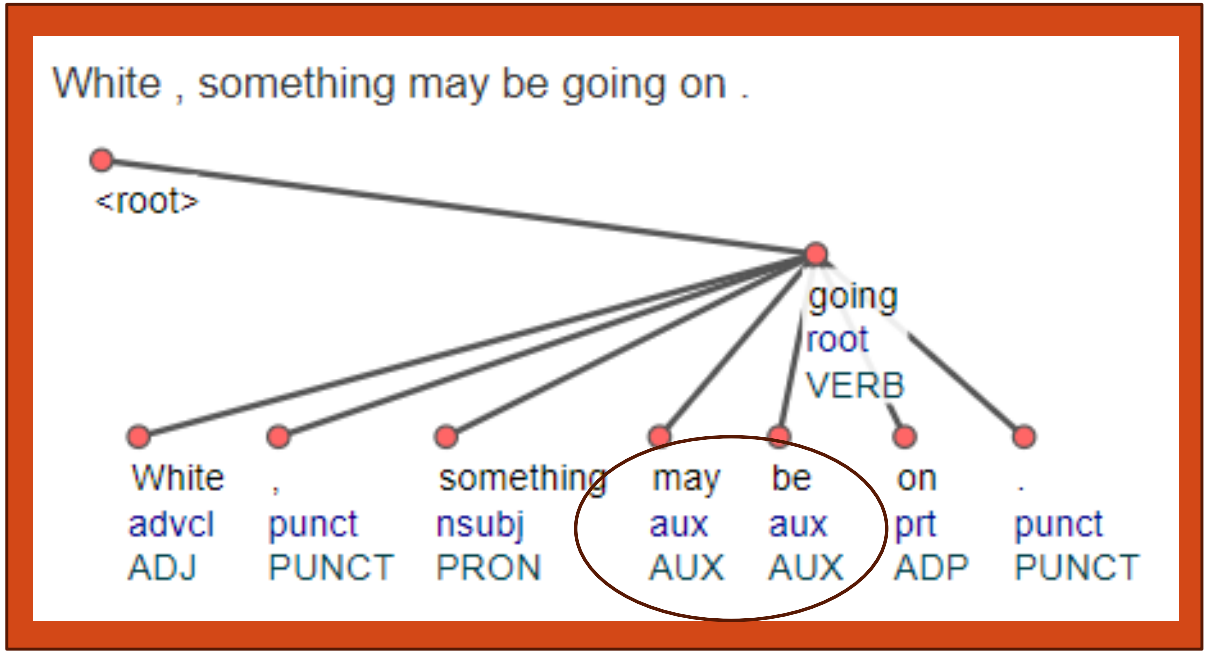
- The Corpus of Regional African American Language (CORAAL)
  - First corpus of African American speech
  - Over 220 sociolinguistic interviews from speakers born between 1888 and 2005
  - (Kendall and Farrington, 2021)

- The Joel Buchanan Archive of African American Oral History
  - Over 700 oral history interviews with African Americans throughout the southern United States
  - Curated by the Samuel Proctor Oral History Program at the University of Florida
  - (University of Florida, 2023)

Methodology: Linguistic Analysis

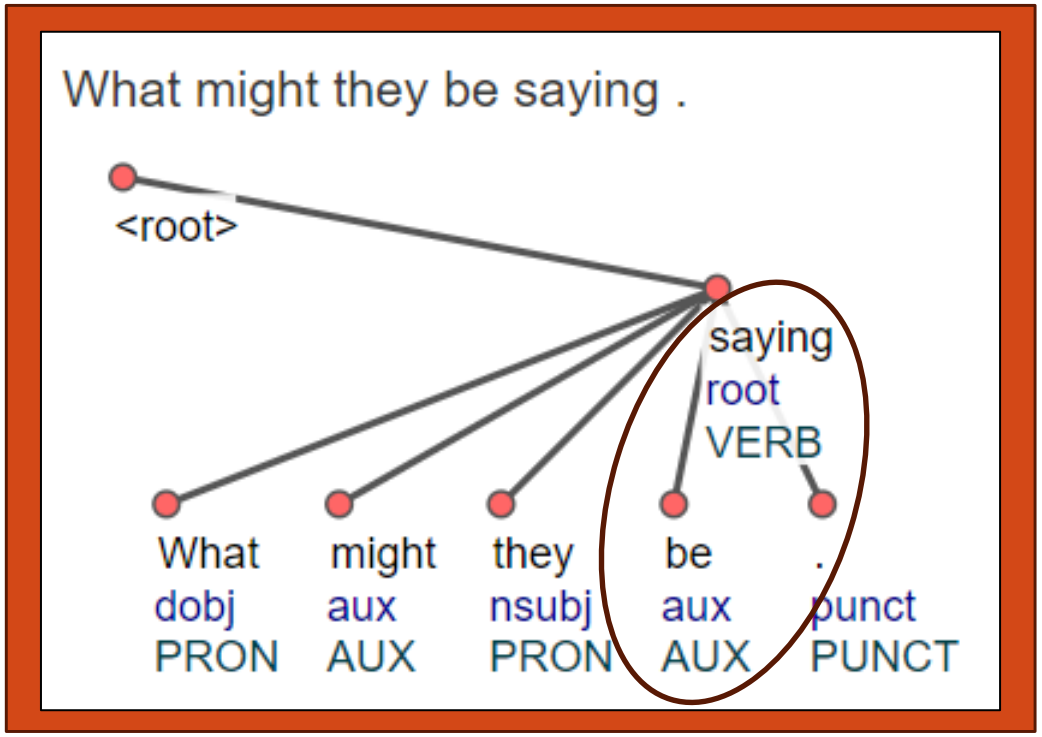


SynPar1: "be" is tagged as an auxiliary and it has a child with a dependency relation of auxiliary

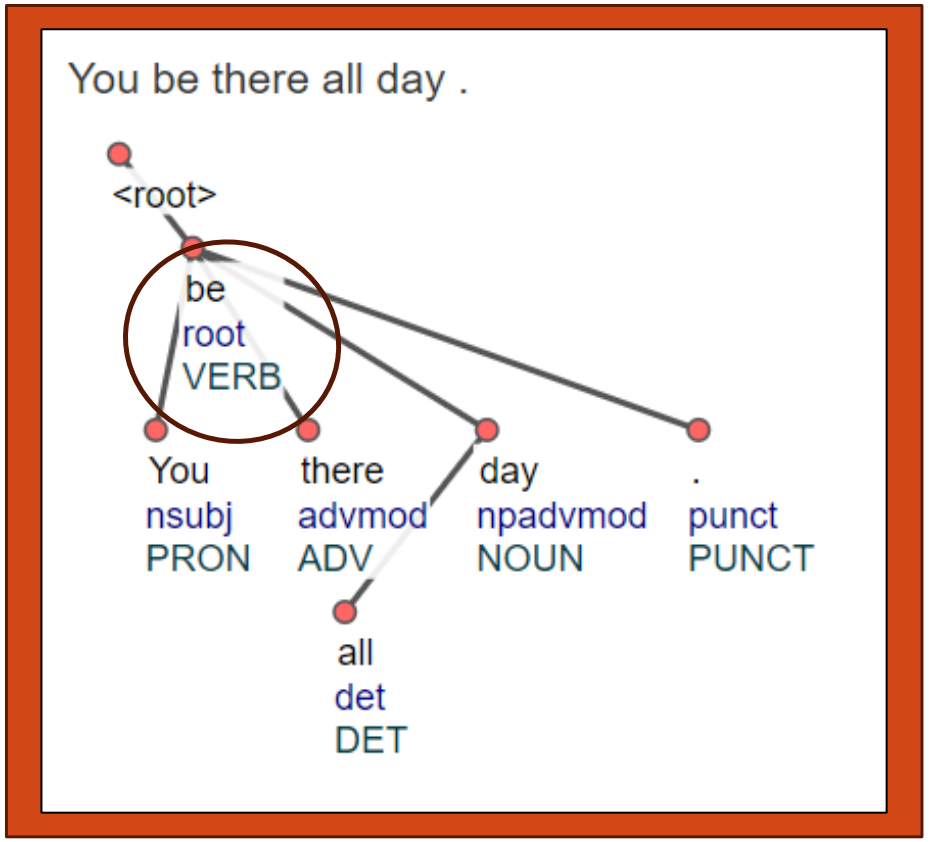


SynPar2: "be" is tagged as an auxiliary and it has a sibling with a dependency relation of auxiliary

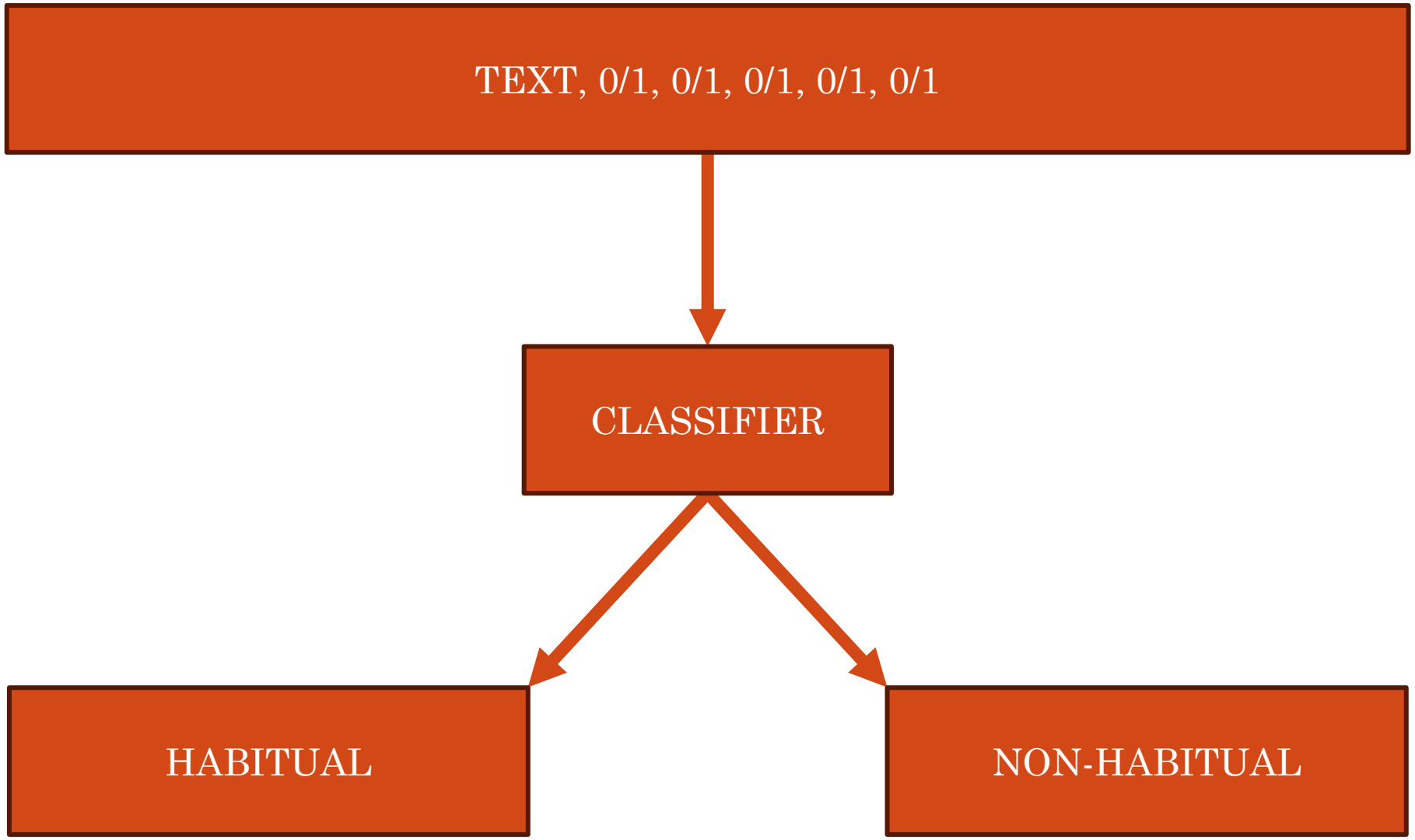
Methodology: Linguistic Analysis



SynPar3: "be" is tagged as an auxiliary and it has a head that is tagged as a "VERB"



SynPar4: "be" is tagged as a verb


TEXT, 0/1, 0/1, 0/1, 0/1, 0/1
CLASSIFIER
HABITUAL
NON-HABITUAL

Experiments

- Baseline: bigrams and unigrams in 8 word window around ‘be”
1. Part-of-Speech (POS)
2. Dependency Patterns (DEP)
3. Post-Hoc Rules (PH)
4. Interactions of Patterns (INT)
5. Simple Part-Of-Speech Window (+win)
6. Data Augmentation (+aug)

Results

AAE-informed Data

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

Classification of the habitual class

General Linguistic Data

Conclusion

- Leveraged the unique features of AAE's habitual *be* to develop a classifier based on linguistics rather than extensive amounts of data.
- 0.83 F1-score without data augmentation improves Santiago et al. (2022) 0.652 F1-score with data augmentation.
- Models informed by AAE syntax out-perform those that don't
- It is important to leverage existing linguistic literature and expert knowledge in the development of NLP systems for low resource languages.

Thank You

References

- Blodgett, S. L., Green, L., & O'Connor, B. (2016). Demographic Dialectal Variation in Social Media: A Case Study of African-American English. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1119–1130. https://doi.org/10.18653/v1/D16-1120
- Dacon, J. (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*, 55–63. https://doi.org/10.18653/v1/2022.hcinlp-1.8
- Deas, N., Grieser, J., Kleiner, S., Patton, D., Turcan, E., & McKeown, K. (2023). *Evaluation of African American Language Bias in Natural Language Generation* (arXiv:2305.14291). arXiv. https://doi.org/10.48550/arXiv.2305.14291
- Fasold, R. W. (1969). Tense and the Form Be in Black English. *Language*, *45*(4), 763–776. https://doi.org/10.2307/412334
- Green, L. J. (2002). *African American English: A Linguistic Introduction*. Cambridge University Press.
- Harris, C., Halevy, M., Howard, A., Bruckman, A., & Yang, D. (2022). Exploring the Role of Grammar and Word Choice in Bias Toward African American English (AAE) in Hate Speech Classification. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 789–798. https://doi.org/10.1145/3531146.3533144
- Jørgensen, A., Hovy, D., & Søgaard, A. (2016). Learning a POS tagger for AAVE-like language. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1115–1120. https://doi.org/10.18653/v1/N16-1130