

# Quantification of stylistic differences in human- and ASR-produced transcripts of African American English

Annika Heuser<sup>1</sup>, Tyler Kendall<sup>2</sup>, Miguel Del Rio<sup>3</sup>, Quinn McNamara<sup>3</sup>,  
Nishchal Bhandari<sup>3</sup>, Corey Miller<sup>3</sup>, Migüel Jetté<sup>3</sup>  
<sup>1</sup>University of Pennsylvania, <sup>2</sup>University of Oregon, <sup>3</sup>Rev.com

Full paper  
& code



## Problem

um what're you goin' to do	WER
what are you going to do	3/6
uh what're you gonna do	4/6

- A person could have produced each of the above transcripts for the same audio depending on their transcription style guide, familiarity with the speech variety, etc.  
→ **Inter-transcriber variation** is expected especially on underrepresented varieties of speech like African American English (AAE) [1]
- Status quo: A single human-produced transcript is arbitrarily deemed the gold standard and **stylistic** deviations from it are punished in ASR evaluation

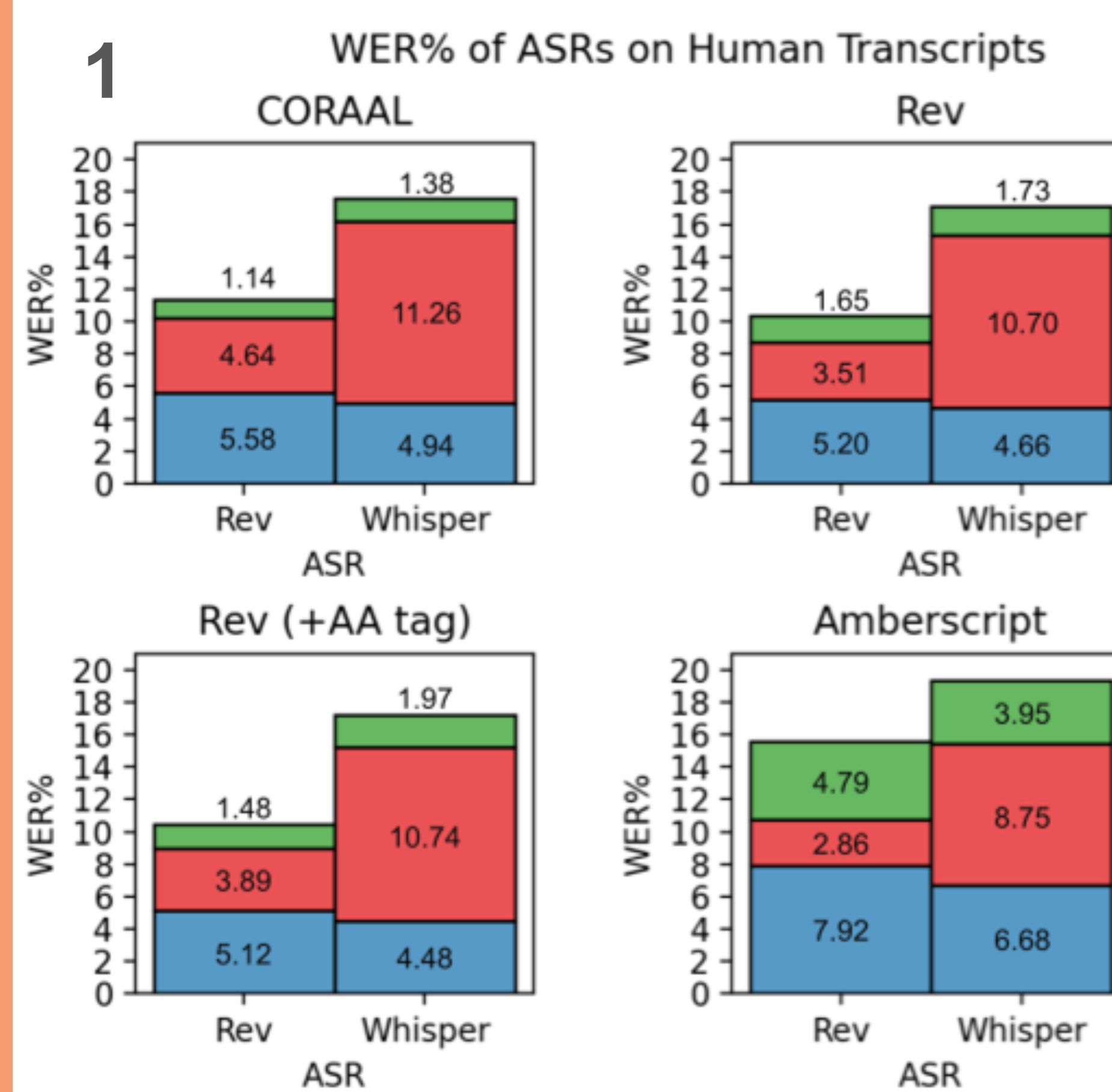
## Approach

- Collected 6 transcript versions of 10 hours of CORAAL [2]: 4 produced by professional human transcribers and 2 by ASR systems
- Operationalized transcription differences as 3 categories that represent hypotheses of potential sources of the differences:
  - verbatim vs non-verbatim (see gold/blue vs pink above)
  - morpho-syntactic features that differentiate AAE from SAE (delineated in [3] and [4])
  - reduction/contraction orthographic representation differences
- Compared WERs across human-produced and ASR-produced transcripts and investigated interactions between WER and the 3 source hypotheses

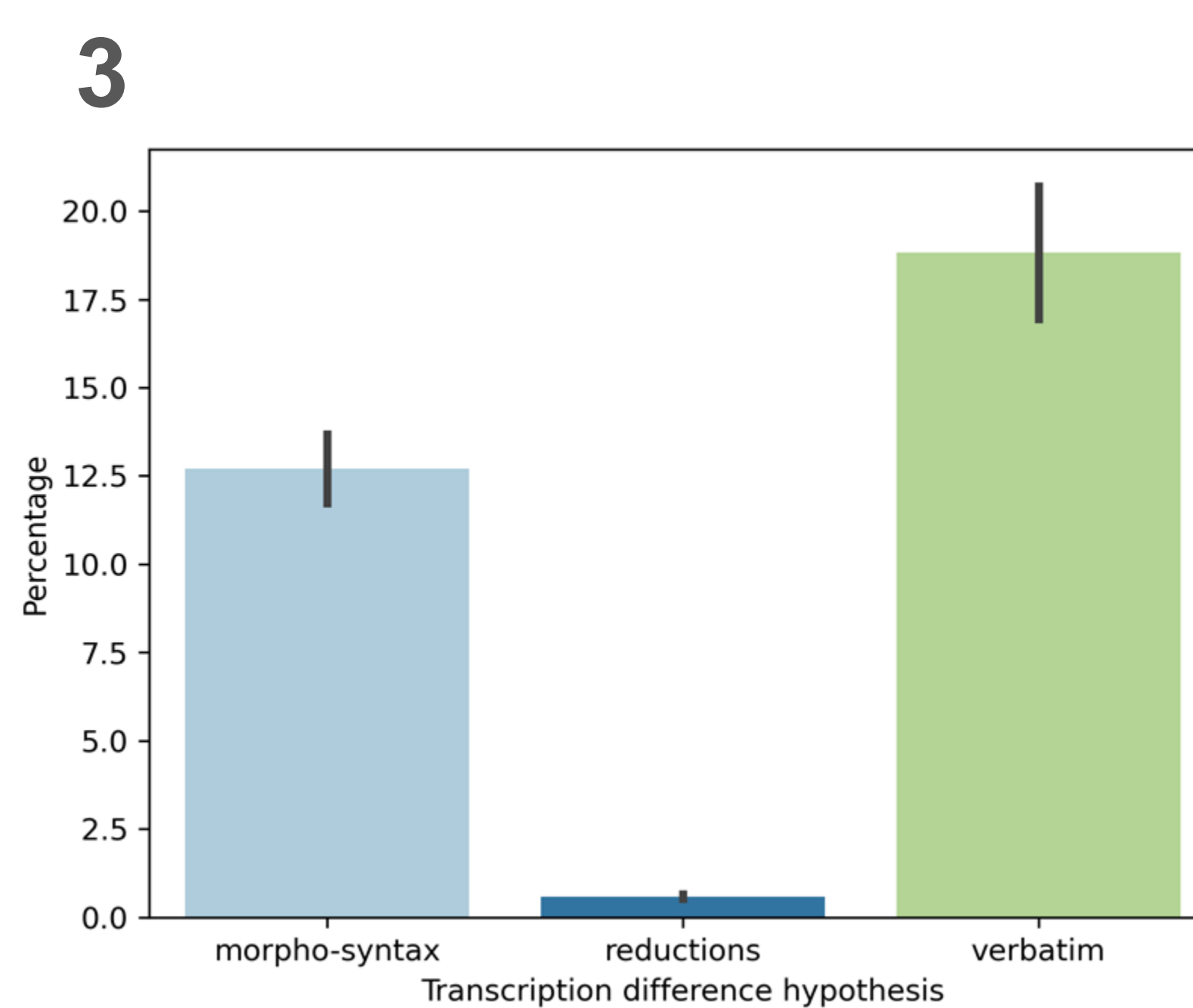
## Transcript versions

- Human-produced:**
  - CORAAL** - created by linguists, released with the audio
  - Rev** - requested human transcription for the CORAAL audio from Rev.com
  - Rev (+AA tag)** - same as Rev but added the accent tag “Other: African American” to potentially increase transcriber familiarity with AAE
  - Amberscript** - another transcription company, using a different style guide
- ASR systems:** **Rev** and **Whisper** [5]

## Results



• **Figure 1 (left):** Rev ASR performance no longer consistent when evaluated against Amberscript transcript, resulting in the WER jumping up 5% and decreasing the gap to Whisper



• **Figure 2 (right):** Human-produced transcripts vary by WERs between 10 and 20%

• **Figure 3 (above):** Verbatim and morpho-syntactic categories accounted for ~30% of the total differences

## Conclusions and Contributions

- WER between human transcripts comparable to ASR WER → single transcript WER is not sufficient to characterize ASR performance
- We found that transcription difference hypotheses provide useful supplementary metrics
- We suggest that multiple reference transcripts may be necessary for more accurate evaluation

## References

- [1] M. Bucholtz, “The politics of transcription,” *Journal of Pragmatics*, vol. 32, no. 10, pp. 1439–1465, 2000. [2] T. Kendall and C. Farrington, “The corpus of regional African American Language,” Version 2023.06, 2023. [3] J. R. Rickford, *African American Vernacular English*. Blackwell Publishers, 1999, ch. Phonological and Grammatical Features of African American Vernacular English (AAVE). [4] A. K. Spears, “Rickford’s list of African American English grammatical features: an update,” in *The Routledge companion to the work of John R. Rickford*. Routledge, 2019, pp. 79–89. [5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International Conference on Machine Learning*. PMLR, 2023.