How Good Is the Transcript?

Evaluating ASR & Diarization

David Jurgens and Dallas Card
July 28, 2026

You have your ASR output. Now what?

- Why bother evaluating?
  - Garbage in, garbage out: downstream measures (topic models, sentiment, discourse analysis) inherit transcript errors
  - Errors are rarely random — they correlate with speaker accent, audio quality, topic, and show type
  - Uncritical use of ASR output risks systematically biasing your sample (whose speech gets "heard" accurately)

The Error Chain

- Errors **compound**, and they are **not random with respect to your variables**
- ASR is worse on: accented speech, AAVE, non-English, overlap, poor audio, technical vocabulary, **names**
- Those correlate with exactly the things social scientists study

A 10% WER that falls entirely on one group's speech is **not** a 10% error. It is a **differential measurement error**, and it biases coefficients rather than just widening confidence intervals.

Koenecke, Allison, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. "Racial disparities in automated speech recognition." *PNAS* 117(14), 2020, 7684–7689. DOI 10.1073/pnas.1915768117. https://www.pnas.org/doi/10.1073/pnas.1915768117

Radford, Alec, et al. "Robust Speech Recognition via Large-Scale Weak Supervision." arXiv:2212.04356. https://arxiv.org/abs/2212.04356 — the paper's own per-language WER tables are the source for the non-English part of the list.

Evidence of Disparities

- **Koenecke et al. 2020 (PNAS)** — five commercial ASR systems, matched audio: **WER 0.35 for Black speakers vs. 0.19 for white speakers.** Nearly **double**
- Driven substantially by acoustic modeling of dialect, not by vocabulary
- **Ezema et al. 2025 (CHI)** — the gap **persists in Whisper**, the downstream classifiers reading its output **inherit** it, and fine-tuning on Black speech **reduces but does not eliminate** it
- Implication: any measure derived from transcripts — sentiment, topic, complexity, toxicity — inherits this gap.

Koenecke, Allison, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. "Racial disparities in automated speech recognition." *PNAS* 117(14), 2020, 7684–7689. DOI 10.1073/pnas.1915768117. https://www.pnas.org/doi/10.1073/pnas.1915768117

Design: five commercial systems (Amazon, Apple, Google, IBM, Microsoft) on structured interviews with **73 Black and 42 white speakers** across five US cities, **19.8 hours**, matched on speaker age and gender. The gap held across **every** vendor tested. See `references.md` [S-40].

Ezema, Kelechi, Chelsea Chandler, Rosy Southwell, Niranjan Cholendiran, and Sidney D'Mello. "'It feels like we're not meeting the criteria': Examining and Mitigating the Cascading Effects of Bias in Automatic Speech Recognition in Spoken Language Interfaces." *CHI 2025*. DOI 10.1145/3706598.3714059 — **Whisper** had lower accuracy for Black than white speakers, downstream classifiers inherited the gap, and fine-tuning on Black speech **reduced but did not eliminate** it.

The Standard Metric: Word Error Rate (WER)

- WER = (Substitutions + Deletions + Insertions) / length of reference transcript
- Requires a reference (ground-truth) transcript compared against the hypothesis (ASR output)
- Widely reported, easy to interpret, well understood — the default benchmark metric
- But: computing it isn't just "count the differences" — see next slide

Computing WER: The Alignment Step

- WER requires first **aligning** the two word sequences — finding which hypothesis words correspond to which reference words
- Alignment is done via minimum edit distance (Levenshtein-style dynamic programming)
- Highly dependent on normalization choices (casing, punctuation, number formatting, disfluencies); must be fixed *before* alignment — they change the result substantially
- **Tools:** Several well-tested libraries that handle alignment, normalization, and reporting consistently (e.g., **jiwer** (python package), or NIST's **sclite**)

WER Alignment: A Worked Example

Reference: the quick brown fox **jumps** over **the** lazy dog

ASR output: the quick brown fox **jump** over ~~(deleted)~~ lazy dog **today**

WER calculation:

S = 1, D = 1, I = 1; reference length N = 9

WER = (S + D + I) / N = 3 / 9 ≈ 33%

Note: because insertions aren't capped, WER can exceed 100% when the hypothesis is much longer/noisier than the reference

Reference	ASR output	Result
the	the	correct
quick	quick	correct
brown	brown	correct
fox	fox	correct
jumps	jump	**substitution**
over	over	correct
the	—	**deletion**
lazy	lazy	correct
dog	dog	correct
—	today	**insertion**

Limitations of WER

- Treats all errors as equally bad ("a" → "an" counts the same as dropping a negation like "not")
- Sensitive to normalization choices: punctuation, casing, disfluencies ("um," "uh"), numbers ("$5,000" vs "five thousand dollars"), and more
- Doesn't capture semantic/downstream impact of an error
- Two transcripts with identical WER can differ hugely in usability for research

Example: the Whisper Normalizer

- Normalization before alignment is challenging, and can make a big difference
- Whisper provided two scripts for this:
  - **BasicTextNormalizer** (language-agnostic): lowercases text, strips punctuation, removes diacritics, collapses whitespace
  - **EnglishTextNormalizer** (English-specific, more aggressive): additionally expands contractions ("don't" → "do not"), converts between spelled-out and digit number forms into one standard format, standardizes British/American spelling variants, expands common abbreviations (e.g., "Mr." → "mister"), then strips remaining punctuation
- Caveat: normalization is a design choice that can also mask genuinely wrong content; different studies using different normalizer versions/settings aren't directly comparable

https://github.com/openai/whisper/tree/main/whisper/normalizers

Alternative ASR Metrics: MER, WIL & Semantic Evaluation

- **Match Error Rate (MER)** = (S + D + I) / (S + D + I + C), where C = correct words
- **Word Information Lost (WIL)** = 1 – [C / (C + S + D)] × [C / (C + S + I)]
  - WIL penalizes S > D > I, whereas WER treats all types of errors the same
- **Semantic-similarity-based evaluation** — compare meaning instead of exact strings: embed reference and hypothesis (word/sentence embeddings, e.g., BERTScore-style) and score similarity (cosine distance)
  - Treats meaning-preserving substitutions (e.g., "couch" → "sofa") as low-cost, unlike WER which penalizes them the same as any other error
  - Better aligned with many social science use cases (topic, sentiment, stance) where exact wording matters less than meaning, but does not solve all problems
- Takeaway: WER is most standard; no universally "correct" one

Kim et al.. "Semantic Distance: A New Metric for ASR Performance Analysis Towards Spoken Language Understanding." INTERSPEECH 2021

Diarization Metrics: DER, JER & the Split/Merge Problem

- **Diarization Error Rate (DER)** = (missed speech + false alarm + speaker confusion) / total reference speaker time — time-weighted, so dominant speakers dominate the score
- **Jaccard Error Rate (JER)** — averages error *per speaker* rather than per unit time, giving a rarely-speaking guest equal weight to a dominant host
- **The split/merge problem:** in real multi-speaker conversations, diarization can:
  - **Split** one true speaker into multiple system-detected "speakers" (over-segmentation — e.g., triggered by tone shifts, coughs, or background noise)
  - **Merge** two distinct speakers into a single cluster (under-segmentation — e.g., similar-sounding voices, heavy crosstalk, brief interruptions)
- This failure mode isn't fully captured by a single scalar metric — worth directly inspecting speaker count and cluster purity, not just DER, especially before speaker-level analysis (e.g., "how much did the host vs. guest speak")

Easy vs. Hard Cases

- Easy: single speaker, scripted/read speech, studio-quality recording, standard accent, clean audio
- Hard:
  - Overlapping speech / crosstalk (common in conversational podcasts)
  - Multiple speakers with similar voices
  - Non-standard accents, dialects, code-switching
  - Background music, sound effects, remote/phone-call audio quality
  - Domain-specific jargon, proper nouns, non-English names
- ASR difficulty will vary greatly across podcasts!

Proper Names Present Special Difficulty

- Many names have multiple common spellings, and are therefore inherently ambiguous without context (e.g., Sara vs Sarah)
- If mis-recognized as a name, can be replaced by phonetically similar phrases, e.g., "Helen Forrest" → "how enforced"
- Most difficult will be non-English or unconventionally-spelled names
- Impact on WER will be small (assuming names are not frequently mentioned), but could have a significant impact on analysis
- Practical mitigations: supply known names as a prompt/hotword list when transcribing (Whisper supports an initial prompt), fuzzy-match/entity-link suspicious tokens against a known name list (e.g., episode/guest metadata)

Proper Names Present Special Difficulty

One clip, every failure at once

Ground truth (the show's own title): *Cybersecurity Awareness with Security Professional* **Mack Jackson Jr**

Nineteen configurations, one name, five renderings — one of them right:

Output	Configurations
"Mac Jackson, Jr"	`tiny.en`, `base.en` (×3 precisions), **`large-v3-turbo`**, **`large-v3`** (both)
"Mac Jackson Jr"	`small.en`, `medium.en`, **Parakeet-v3** (both precisions)
"Matt Jackson Jr"	`tiny`, **Parakeet-v2** (both), **Canary-1B** (both)
"Mack Jackson, Jr"	`base`, `distil-large-v3`
✅ "Mack Jackson Jr"	`small` — **1 of 19**

Hallucinations: A Different Kind of Error

- Hallucination = fluent, plausible-sounding output with no basis in the actual audio
- Distinct from ordinary substitution/deletion errors — the model is "making things up" rather than mishearing
- Common patterns: repeated phrases/loops, invented sentences, inserted boilerplate (e.g., "Thanks for watching!")

Hallucinations: A Different Kind of Error

- Whisper generates; when the audio has no speech, it can still emit fluent text
- **Koenecke et al. 2024, "Careless Whisper" (ACM FAccT)** — hallucinations in ~**1%** of transcribed segments, disproportionately for speakers with **aphasia**; ~38% carried explicit harms (invented violence, fabricated authority, false attribution)
- **Barański et al. 2025 — non-speech audio reliably induces hallucination**: music, applause, silence, room tone
- Podcasts are full of exactly that: intro music, ad breaks, stings, silences

Koenecke, Allison, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. "Careless Whisper: Speech-to-Text Hallucination Harms." *ACM FAccT 2024*. DOI 10.1145/3630106.3658996 · arXiv:2402.08021. https://arxiv.org/abs/2402.08021

Barański, Mateusz, Jan Jasiński, Julitta Bartolewska, Stanisław Kacprzak, Marcin Witkowski, and Konrad Kowalczyk. "Investigation of Whisper ASR Hallucinations Induced by Non-Speech Audio." *ICASSP 2025*, Hyderabad. arXiv:2501.11378. https://arxiv.org/abs/2501.11378

Detecting & Mitigating Hallucinations

- Voice activity detection (VAD) pre-filtering to skip non-speech segments before transcription
- Repetition/n-gram loop detection as a red flag
- Confidence/logprob thresholds where available
- Spot-checking a random sample against audio, especially at segment boundaries and silence
- **Open problem**: No fully automated solution exists here

Using Show-Provided Transcripts as a Reference

- Tempting: many shows publish their own transcripts — free, scalable, no manual annotation needed
- Potential upside: could serve as a large-scale reference for evaluating ASR without costly human transcription
- In addition, show transcripts often include **speaker names** (host/guest identified by name), which diarization alone cannot give you

Transcript Example

00:08 **Our sponsor: Qlik**

MS: Data Stories brought to you by Qlik, who allows you to explore the hidden relationships within your data that leads to meaningful insights. Let your instincts lead the way to create personalized visualizations and dynamic dashboards with QlikSense, which you can download for free at qwik.e/datastories. That's qlik.e/datastories.

00:50 **Welcome from Moritz**

MS: Hey everyone, it's a new Data Stories. So this time, it's only me, and you know, because Enrico's on vacations, but I have a good set of guests here. Two very special guests and what we want to discuss today is a topic that's um, to me is a very close to my heart because it affects me everyday and it's about the actual practice and the business of doing data visualizations for other people. LIke how do you actually manage that process, how do you find new projects, how do you um, go about figuring out what a client needs and yeah how do you come up with good end results, like all these practical considerations that go into that.

01:34 **Introducing Jan Willem Tulp and Mahir Yavuz**

MS: And yeah, for that topic, I have two really special guests who are experts in the area. One is Jan WIllem Tulp, he's a freelance information designer from the Netherlands.

JT: Hi, hi Moritz.

MS: And Mahir Yavuz who we had on um, a few dozens of episodes before? So it's been a few years and he's now, can you say what your job is?

MY: Creative director of data science and visualization at RGA in New York City.

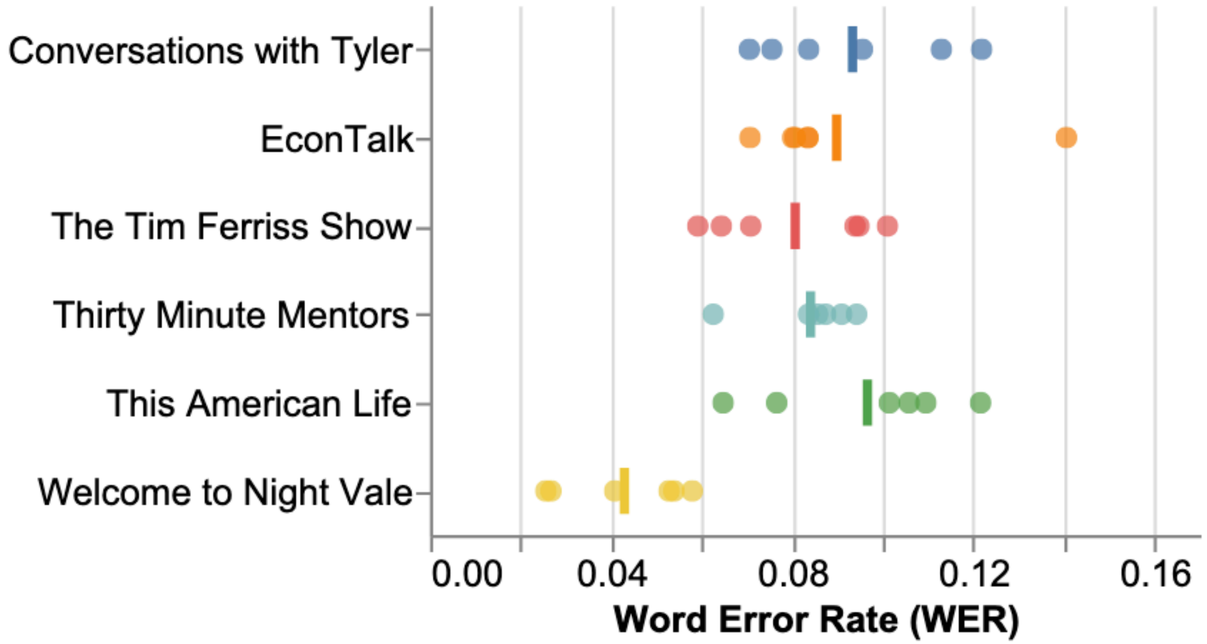
https://datastori.es/transcripts/

Problems With Show Transcripts as Ground Truth

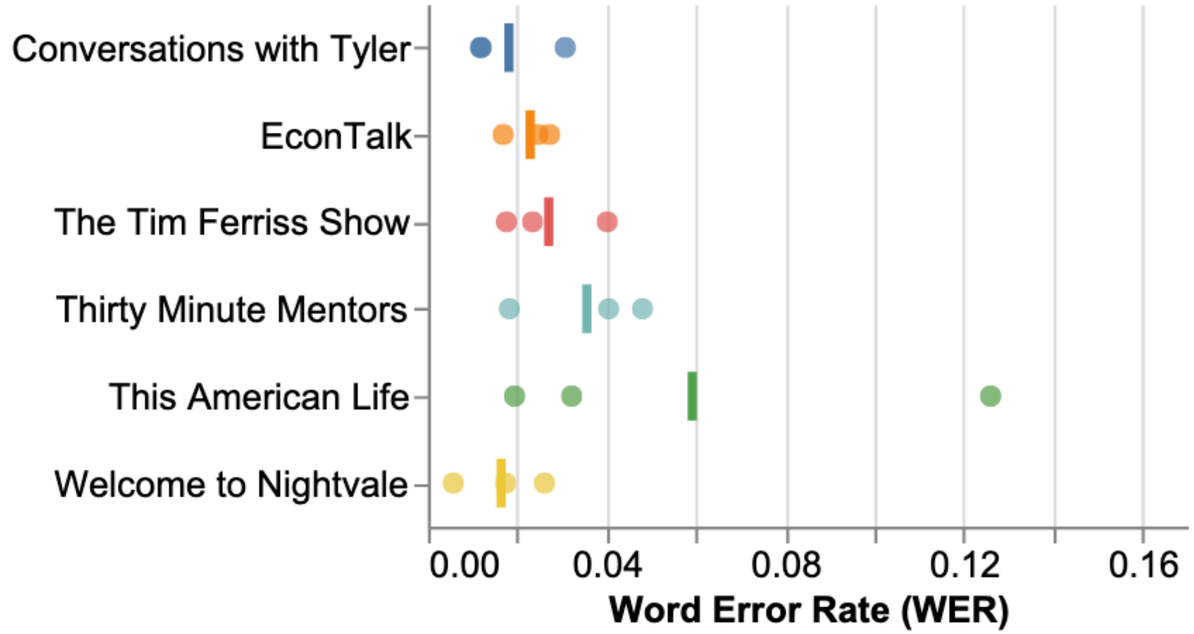
- **Provenance is typically unknown:** you usually can't tell how a show's transcript was actually produced — professional human transcription, in-house ASR, an outsourced vendor
- **Often edited for readability**, not verbatim (removed disfluencies, cleaned-up grammar, summarized crosstalk)
- **Frequently missing timestamps** needed for alignment/diarization evaluation
- **Inconsistent coverage**: ads often removed, intros/outros paraphrased, segments skipped

Problems With Show Transcripts as Ground Truth

WER against official transcript



Manually verified WER



Litterer, Jurgens, and Card. "Mapping the Podcast Ecosystem with the Structured Podcastx Research Corpus". ACL 2025

Recap & Bridge

- WER is the standard metric, but computing it correctly requires proper alignment and normalization — use a library, don't hand-roll it
- MER, WIL, and semantic-similarity metrics each address a different blind spot of WER; diarization needs its own metrics (DER, JER) and a watchful eye for split/merged speakers
- Podcast audio sits at the hard end of the ASR difficulty spectrum — and proper names are a persistent weak point regardless of overall audio quality
- Hallucinations are a distinct, dangerous failure mode — especially with Whisper on noisy/non-speech audio
- Show transcripts are a convenient but provenance-unknown reference — useful especially for speaker names, but verify before trusting for error rates
- Next: Annotating podcast transcripts