Decoding Parameters as Methodological Choices in ASR-Derived Corpus Construction

Steven Coats
Faculty of Humanities
University of Oulu
Finland
steven.coats@oulu.fi

Abstract

Automatic speech recognition (ASR) is increasingly used for large-scale corpus construction, yet the impact of decoding parameters on resulting datasets remains underexplored. We compare two operational decoding configurations selected by fixed temperature settings (0.0 vs. 0.1) in a Singapore-English-adapted Whisper pipeline. In a paired rerun of 1,322 podcast recordings (774 hours), sampling at temperature 0.1 produced 99,011 more aligned words than beam search at temperature 0.0 (+1.02%; 95% paired-bootstrap interval: +0.89–1.16%). Hand-checked passages show that these differences include both recovered audible material and acoustically unsupported insertions. An additional sampling-only control at temperature 0.01 produced more output than at 0.1, showing that output quantity was not monotonic with temperature. We argue that decoding configuration should be treated as a methodological decision in corpus construction rather than a purely technical setting.

Keywords: ASR, corpus construction, Whisper, decoding configuration, transcription quality

1. Introduction

Recent advances in automatic speech recognition (ASR) have enabled the creation of large-scale spoken corpora from online audio sources with minimal manual intervention. Systems such as wav2vec 2.0 (Baevski et al., 2020), Whisper (Radford et al., 2022), or Qwen3-ASR (Shi et al., 2026) have made it possible to generate time-aligned transcripts at scale, supporting new directions in corpus linguistics and computer-mediated communication (CMC) research. These developments have facilitated the analysis of naturalistic spoken interaction in ways that were previously infeasible (Coto-Solano et al., 2021; Coats, 2025). At the same time, ASR outputs are not neutral representations of speech. Transcription systems may normalize or suppress linguistically relevant features, including discourse particles and non-standard morphosyntax (Lea et al., 2023), leading to a growing recognition that ASR should be treated as a component of corpus design rather than a purely technical preprocessing step.

Building on this perspective, the present paper examines the role of decoding parameters in shaping ASR-derived corpora. We focus on two operational decoding configurations selected by fixed temperature settings (0.0 vs. 0.1), while holding the fine-tuned model and input audio constant. We show that this apparently minor user-facing setting leads to systematic differences in corpus composition and to qualitatively different errors.

We argue that decoding parameters should be treated as methodological choices in corpus construction, with direct implications for downstream linguistic analysis.

2. Decoding Parameters as Corpus Design Choices

Recent work has emphasized that ASR systems do not simply transcribe speech but actively shape the resulting linguistic data. In the context of YCSEP, fine-tuning was shown to substantially alter the recovery of discourse particles and morphosyntactic variation, with consequences for corpus-based analysis (Coats et al., 2025).

This supports a broader view of ASR as a controllable component of corpus construction, where modeling choices directly affect the representation of linguistic phenomena. While prior work has focused primarily on model architecture and training data, less attention has been paid to decoding parameters, which operate at inference time but may have equally important effects.

Decoding temperature controls the degree of stochasticity in token selection. In the faster-whisper implementation used here, zero temperature invokes deterministic beam search, whereas a positive temperature invokes sampling with best-of selection. Temperature therefore selects operational configurations that can produce different transcripts from the same model and audio.

Despite its importance in sequence generation, the role of decoding configuration in corpus construction has not been systematically examined. This paper addresses this gap through a paired comparison of the configurations selected by two

fixed settings.

3. Data and Method

3.1. Decoding settings

Decoding temperature is a standard parameter in neural sequence models that controls the concentration of the probability distribution over possible output tokens. At each decoding step, the model produces a set of logits z_i, which are unnormalized scores for the candidate tokens. These scores are converted into probabilities using a temperature-scaled softmax (e.g. Hinton et al. 2015):

$$p_i(T) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}.$$

Temperature rescales the differences between the logits without changing their ranking. For any two tokens i and k, the ratio of their probabilities is

$$\frac{p_i(T)}{p_k(T)} = \exp\left(\frac{z_i - z_k}{T}\right).$$

When $T < 1$, differences between logits are amplified, causing the probability distribution to become more sharply concentrated on the highest-scoring tokens. For a fixed non-uniform set of logits, this generally lowers the entropy of the distribution and reduces the probability that less likely alternatives will be selected. As $T \to 0$, the distribution approaches a point mass on the token with the largest logit, so sampling becomes effectively equivalent to greedy decoding, except in cases involving ties. When $T > 1$, differences between logits are compressed, producing a flatter, higher-entropy distribution. This increases the probability of sampling lower-scoring tokens and therefore introduces greater variation into the decoded output. In the limit $T \to \infty$, the distribution approaches a uniform distribution over the candidate tokens. Temperature therefore controls the stochasticity of sampling-based decoding: lower values make decoding more conservative and repeatable, whereas higher values make it more variable but also increase the risk of selecting implausible tokens. Importantly, temperature affects the output only when tokens are sampled from the distribution. Under strictly greedy decoding, changing T does not change the selected token because dividing all logits by the same positive value leaves their ordering unchanged.

In Whisper, temperature is not typically fixed but used in a fallback schedule: decoding begins at $T = 0.0$, and higher temperatures are applied only if the initial output fails quality checks (e.g. low log-probability or repetitive loops). We instead enforced a fixed scalar temperature in each condition, without fallback. Because faster-whisper changes from beam search at zero to sampling with best-of selection at a positive value, this is a comparison of the two operational configurations selected by the user-facing parameter, not of temperature scaling with the internal search algorithm held constant.

3.2. Test data

The test data comprised 1,322 recordings (774.02 hours) from six Singapore English podcast series in YCSEP. They include conversation, overlapping speech, music, sound effects, editing, and variable recording conditions. Audio was processed with the Singapore-English-adapted Whisper large-v2 model `stcoats/whisper-large-v2-NSC` in a pipeline including transcription, WhisperX alignment, pyannote diarization, and token-level extraction.

Each recording was decoded under two fixed settings ($T = 0.0$ vs. $T = 0.1$):

- Temperature 0.0: deterministic beam search (beam size 5)
- Temperature 0.1: low-temperature sampling (best-of 5)

The waveform, model, language and threshold settings were identical. Each recording was downloaded and diarized once, and the same diarization result was reused; the two transcripts were aligned separately. Previous-text conditioning and VAD filtering were disabled in both conditions.

This setup provides a paired comparison of how the two decoding configurations affect the resulting corpus.

As a sampling-only control, we subsequently decoded the same recordings at $T = 0.01$ with best-of 5, keeping the other transcription settings identical to the $T = 0.1$ condition. This comparison isolates two positive temperatures under the same search procedure.

4. Results

All 1,322 recordings completed under both configurations without processing errors. Table 1 reports the paired corpus totals.

	Aligned words	ASR segments	Derived turns	Turn duration (s)
Temp 0.0	9,666,394	95,789	639,263	2,313,227
Temp 0.1	9,765,405	100,861	634,563	2,301,434

Table 1: Corpus statistics for temperature 0.0 and 0.1.

Temperature 0.1 produced 99,011 more aligned words (+1.02%; 95% paired-bootstrap interval: +0.89–1.16%) and 98,396 more words in the raw ASR segments (+1.01%). These are differences

in output quantity, not by themselves evidence that all additional words represent recovered speech.

The segmentation measures moved in different directions. Temperature 0.1 produced 5,072 more raw ASR segments (+5.29%) but 4,700 fewer derived turns (-0.74%). Derived turns were constructed after alignment, beginning at a speaker-label change or a gap over 0.5 seconds; they are not pyannote diarization turns. Their combined duration also declined by 0.51%. Raw ASR segmentation, aligned duration, and downstream turn construction therefore cannot be treated as interchangeable measures of captured speech.

The aggregate totals do not determine whether particular differences are deletions, insertions, or substitutions. Hand-checking is therefore necessary before interpreting additional output as greater coverage or reduced output as greater precision.

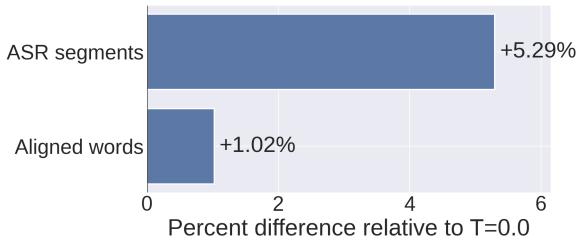

Figure 1: Relative differences in corpus-level output measures for sampling at temperature 0.1 compared with beam search at temperature 0.0.

Figure 1 summarizes these differences. The largest change is the 5.29% increase in raw ASR segments; aligned words increase by 1.02%, whereas derived turns and their combined duration decline. The configuration change thus affects both textual output and its downstream segmentation, but not as a uniform increase across measures.

These differences can affect the linguistic structure of the corpus. Token counts and segmentation differ systematically between versions, while features already vulnerable to ASR normalization—such as discourse particles—may be affected by either omitted or unsupported output.

Tables 2 and 3 present two hand-checked excerpts. Neither transcript is treated as a complete gold reference; the examples isolate one audible omission and one unsupported insertion.

The temperature 0.01 output also contained the audible exchange in Table 2, but did not contain the unsupported continuation in Table 3. The first example therefore primarily distinguishes sampling from beam search, whereas the second records a divergence between the two sampling temperatures. These examples do not establish corpus-wide deletion or insertion rates.

The sampling-only control further qualified the primary comparison. Temperature 0.01 produced 10,051,644 aligned words, compared with 9,765,405 at temperature 0.1; output at 0.1 was 2.85% lower (95% paired-bootstrap interval: 2.68–3.01% lower). Transcript quantity therefore did not increase monotonically with temperature.

Setting	Transcript excerpt
Temp 0.0	that was our plug lah …how do you like being back in the studio like back in the office and all …ya ya ya maybe it's harder for us to have very heated arguments …
Temp 0.1	that was our plug lah …that was a plug woah we totally skip the small talk ah I actually like it man it's a slightly different dynamic to the podcast …but I think it's refreshing lah …ya ya ya maybe it's harder for us to have very heated arguments …

Table 2: Hand-checked example of an audible passage present only in the temperature 0.1 output (blue). Both outputs contain other transcription errors.

Setting	Transcript excerpt
Temp 0.0	his most famous film is Ex-Machina lah …he wrote *The Beach* …
Temp 0.1	his most famous film is Ex-Machina lah which is a movie that is set in the Middle East and it's about a guy named Alex-Garland who is a terrorist who is trying to escape from the u u u u u u u u u u u u u u …he wrote *The Beach* …

Table 3: Hand-checked example of an acoustically unsupported temperature 0.1 insertion (red), followed by a repetition loop.

In sum, the results show that the decoding configurations produce measurably different textual and structural representations of the same audio.

5. Discussion

The results reinforce the view that ASR should be understood as a component of corpus design rather than as a neutral preprocessing tool. Previous work has shown that fine-tuning can alter the representation of linguistically meaningful features in ASR-derived corpora (Coats et al., In review). The present study extends this argument by showing that operational decoding configurations, with both the model and input audio held constant, can produce systematic differences in the resulting corpus.

The ideal outcome for corpus construction would be a transcript containing all and only the material supported by the speech signal: no omissions, hallucinations, or substitutions. None of the configurations necessarily achieves this ideal. Sampling at temperature 0.1 produces about one percent more text than beam search, but output quantity

alone cannot distinguish recovered speech from insertions. The hand-checked examples demonstrate that both occur in particular passages. The temperature 0.01 control further shows that the relationship between temperature and output quantity is not monotonic. Without a corpus-wide gold transcript, the present study cannot estimate general deletion and insertion rates or declare any configuration more accurate overall.

One possible explanation for the greater output at temperature 0.01 than at 0.1 is that, at ambiguous boundaries, the broader distribution at 0.1 increases the probability of sampling lower-ranked end or timestamp tokens, or other tokens that subsequently lead to earlier termination. The reverse would apply where an end token already had the highest logit. Because the experiment did not retain complete token distributions or discarded candidate paths, the precise mechanism cannot be determined from the resulting transcripts alone.

The consequences of omissions and insertions depend on the intended use of the corpus. Omissions reduce the amount of observable linguistic material and may disproportionately remove quiet, reduced, overlapping, disfluent, or otherwise acoustically difficult speech. This is particularly problematic for analyses concerned with interactional phenomena, discourse markers, hesitation, repetition, turn-taking, or non-canonical forms, since precisely these features may be especially vulnerable to deletion. Hallucinations create a different problem: they introduce tokens and constructions for which there is no corresponding evidence in the audio. Such material may inflate word frequencies, create spurious lexical or grammatical patterns, distort measures of repetition or fluency, and generate apparently meaningful examples that were never produced by a speaker.

Consequently, no configuration should be preferred solely because it produces more or less text. For information retrieval, additional output may expose relevant material, but unsupported additions also create false hits. For lexical, grammatical, interactional, or quantitative corpus analysis, insertions may be especially damaging because they can be mistaken for genuine linguistic evidence; omissions can be equally consequential when the analysis depends on low-salience or interactionally important material. The appropriate decoding configuration must therefore be evaluated against manually checked target-domain speech and in relation to the intended downstream analysis.

These findings also have practical implications for corpus-building workflows. Decoding configuration should not be treated as a minor implementation detail or left at an undocumented default. Corpus creators should report the model, decoding strategy, temperature, fallback behavior, segmentation procedure, and any post-processing applied. The present study identifies paired corpus-level differences and illustrates two hand-checked passages, rather than providing a complete gold-standard error analysis. The adapted model was trained on comparatively clean two-person NSC conversations, whereas YCSEP contains overlap, production effects, music, and noise; the observed differences may therefore interact with acoustic and domain mismatch. Future work should compare standard Whisper and the adapted model under their normal fallback schedules and evaluate error types on a manually checked sample. Where feasible, original audio should remain linked to the transcript, and ASR-derived corpora should be versioned when decoding configuration changes.

The differences are not negligible variation around an otherwise identical transcript: they amount to approximately 99,000 aligned words and thousands of segment-boundary differences. A corpus generated with one decoding configuration may therefore support different lexical searches, discourse samples, and interactional measurements from a corpus generated from the same audio with another configuration. Determining which differences constitute structured linguistic bias requires a larger gold-standard sample, but the possibility is sufficient to make decoding configuration methodologically relevant.

We consequently argue that decoding parameters should be explicitly reported, justified, and, where possible, empirically evaluated in corpus-based research using ASR. Rather than relying uncritically on default settings, researchers should treat decoding configuration as part of the design space of the corpus itself. The aim is not to accept a permanent choice between omissions and hallucinations, but to identify, document, and mitigate the error profile introduced by the selected workflow.

6. Conclusion

This paper has shown that operational decoding configurations can materially affect the structure and content of an ASR-derived corpus. Sampling at temperature 0.1 produced 1.02% more aligned words than beam search at temperature 0.0, while hand-checked examples revealed both recovered audible material and unsupported insertion. A sampling-only control showed that this difference was not a monotonic effect of temperature.

We propose that decoding parameters should be treated as first-order methodological decisions in corpus construction. Future work should explore additional parameters and develop standardized practices for ASR-based corpus creation.

As ASR becomes increasingly central to corpus linguistics, greater attention to these issues will be

essential for ensuring the validity and reproducibility of research findings.

7. Bibliographical References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. arXiv preprint arXiv:2006.11477 [cs.CL].

Steven Coats. 2025. An automatic pipeline for processing streamed content: New horizons for corpus linguistics and phonetics. In Louis Cotgrove, Laura Herzberg, and Harald Lüngen, editors, *Exploring digitally-mediated communication with corpora: Methods, analyses, and corpus construction*, pages 257–274. De Gruyter.

Steven Coats, Carmelo Alessandro Basile, Cameron Morin, and Robert Fuchs. 2025. The YouTube Corpus of Singapore English Podcasts. *English World-Wide: A Journal of Varieties of English*, 46(3):274–298.

Steven Coats, Carmelo Alessandro Basile, Cameron Morin, and Robert Fuchs. In review. Fine-tuning ASR for corpus linguistics: Singapore English.

Rolando Coto-Solano, James N. Stanford, and Sravana K. Reddy. 2021. Advances in completely automated vowel analysis for sociophonetics: Using end-to-end speech recognition systems with DARLA. *Frontiers in Artificial Intelligence*, 4.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 [stat.ML].

Colin Lea, Zifang Huang, Jaya Narain, Lauren Tooley, Dianna Yee, Dung Tien Tran, Panayiotis Georgiou, Jeffrey P Bigham, and Leah Findlater. 2023. From User Perceptions to Technical Improvement: Enabling People Who Stutter to Better Use Speech Recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, Hamburg Germany. ACM.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356 [eess.AS].

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. 2026. Qwen3-ASR technical report. arXiv preprint arXiv:2601.21337 [cs.CL].