Automatic Speech Recognition Variance: Consecutive Runs of Low-Resource Languages in Whisper

Laurel Lord* and Mark Newman

Department of Data Sciences, Harrisburg University of Science and Technology, Harrisburg, USA
Email: lalord@my.harrisburgu.edu (L.L.); mnewman@harrisburgu.edu (M.N.)
*Corresponding author

Manuscript received September 15, 2023; revised November 26, 2023; accepted January 19, 2024; published April 26, 2024

***Abstract*—This study employs OpenAI's Whisper to explore the manifestation of variance in an Automatic Speech Recognition (ASR) system. Three trained languages from Whisper's current offerings (English, French, and Haitian Kreyòl) and one untrained (Saint Lucian Kwéyòl) completed thirty consecutive runs each, across five model sizes. Etymologically complex yet orthographically simple, mutually intelligible languages may challenge ASR system capabilities. However, a phonetically similar trained language model generated approximate phonetic transcripts for an untrained one. Despite implicit variance hurdles like non-determinism and data deficiencies, ASR systems may aid in documenting high-orality, low-resource languages.**

***Keywords*—automatic speech recognition, creole, low-resource languages, Whisper**

I. Introduction

OpenAI's mission to ensure "artificial general intelligence benefits all of humanity" [1] is advanced by Whisper, an automatic speech recognition (ASR) system that transcribes audio to text in various file formats [2]. After training on 680,000 hours of web-sourced multilingual and multitask supervised data, Whisper's researchers promoted scaling weakly supervised pre-training in ASR beyond English [2]. Furthermore, they posited that minor increases in Whisper's training data size could enhance ASR performance in low-resource languages [2].

Well-established European natural languages like English (ENG) and French (FRA), present in most ASR systems, benefit from abundant text records from different text domains [3, 4]. Slavery imposed chaos in the Caribbean region [5–15]. The resulting evolution of some creoles may differ from other languages; they can develop informally and achieve local utility, yet lack reinforcement from a formal writing system and widespread use in educational, governmental, and modern commercial contexts [6, 8], [13–15]. Several linguistic institutions monitor creole languages [16–18], yet natural language processing (NLP) research and resources remain scarce. In such cases, a translated Bible text sample is a helpful resource for a low-resource language, as its enduring existence and widespread distribution often foster cross-lingual research in parallel corpora [3].

Haitian Kreyòl (HAT) is an official language of Haiti [15]. It garnered international attention amidst the country's natural disasters and political turmoil [19, 20]. Now it is a default language option on platforms like Whisper [2]. In the country of Saint Lucia, Kwéyòl (ACF) is considered to be a "native" or "national language" but lacks official status [14], [21]. Experts have discussed the mutual intelligibility of these creole languages due to cultural and linguistic similarities, yet ACF remains absent from major language tool platforms [5, 11, 14, 22]. Studies of etymologically complex yet orthographically simple languages with presumed relatedness may challenge current ASR research inadequacies and linguistic disparities. Based on their industry popularity and direct influence on HAT and ACF, assessing an ASR system's variance with a focus on its adaptability to phonetic variations in creole languages can be enhanced by including evaluations of ENG and FRA.

Analysis Questions:

Given that English was a main language used for Whisper's training [2], the following was asked:

1) How does the existing *English* language model variance differ across model sizes and multiple runs of ENG audio?
2) How does the existing non-English (*French*) language model variance differ across model sizes and multiple runs of FRA audio?
3) How does the existing non-English (*Haitian*) creole language model variance differ across model sizes and multiple runs of HAT audio?

How do the non-English (*Haitian*) creole language model variances compare to the outcomes of an untrained mutually intelligible language (ACF), and what discrepancies arise among runs?

II. Literature Review

Transcription Standards and Metrics

The International Phonetic Alphabet (IPA) is a useful standardized cross-lingual tool to represent human speech. Despite its multiple iterations, the IPA offers a comprehensive symbol set that can aid in documenting diacritics and language- or dialect-specific variations [23]. Even languages with phonemic orthography [15] and high orality may gain from IPA use, as it has served to curate orthographic systems that distinguish HAT from FRA [11], [24]. Language learners and researchers benefit from its consistent transcription and pronunciation standards [11, 22, 23]. Yet, IPA-formatted transcriptions may not be ideal final products for some ASR users.

Regardless of its utility, Whisper's developers did not explicitly employ the IPA in their ASR system. Although English language data was prominently featured, the developers adapted a word-deciphering technique for diverse

languages, and leveraged UTF-8's vast string generation capability [2].

Whisper's capabilities have, however, attracted industry interest and scrutiny [25–29]. Acoustic conditions, speaker accents, and environmental variables can impact the variance in an ASR system [2, 28, 29]. Yet, non-determinism can introduce additional variance beyond these base factors. It denotes unpredictable behavior where a system cannot consistently create output for the same inputs under identical conditions [25, 26, 28]. Whisper's manifestation of non-determinism has piqued the interest of some academics [25, 26, 28]. One reviewer cited "high dropouts, repetition, and hallucination" as evidence of non-determinism [26]. Some users stressed that ASR randomness may cause undesirable performance variations, compromising the consistency and quality of generated text in specific applications such as short-form content generation like captioning [25, 28]. However, a few recognized that introducing randomness enhanced general ASR text generation [25–27]. Whisper's developers asserted that ASRs should reliably perform "out of the box" across diverse environments without requiring manual adjustments for specific scenarios [2]. However, its users have acknowledged ameliorative pre-processing [25] and pre-training [27] tasks, as well as fine-tuning [25, 27–29] and post-processing [27, 28] activities.

The Word Error Rate (WER) metric effectively gauges transcription accuracy and variability across various linguistic contexts by quantifying the disparities between an ASR system's output and a reference transcription. ASR developers [2] and critics [25, 27–29] use WER to test the intrinsic ASR systems' goal of low word error rates. WERs offer insight into accuracy and variability in ASR system performance across various model sizes and runs for trained and untrained language models.

III. Materials and Methods

Limited cross-lingual creole resource availability necessitated restricting audio and text data to public files; the Gospel of John, within the King James Version of the Bible, was selected [30–35]. Whisper's models conducted long-form transcription on multilingual audio samples exceeding 30 seconds, with the output serving as a candidate text for comparison with a reference (ground truth) text. To assess the impact of variability on ASR performance, Whisper's WERs for each language and model size were observed. Therefore, in *Analysis 1*, the *English* language model was applied to ENG audio sample files across Whisper's *Tiny*, *Base*, *Small*, *Medium*, and *Large* model sizes. These model sizes align with the options from Whisper's original release [2]. The discrepancy among runs was tested by generating 30 transcriptions for each model size.

In *Analysis 2*, these tasks were replicated using the *French* language model on FRA audio files. The tasks were repeated with Whisper's *Haitian* (HAT) language model, transcribing audio files for two creole languages: HAT (trained) in *Analysis 3* and ACF (untrained) in *Analysis 4*. Fig. 1 shows this process. Uniform data analysis required minor tweaks due to variations in passage introductions across media file types. The data-cleaning process involved lowercasing and removing punctuation yet retaining accents, apostrophes, and hyphens.

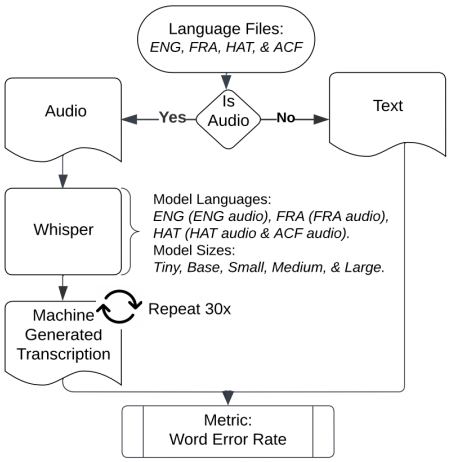

Fig. 1. Diagram illustrating the analysis process.

IV. Results

Fig. 2 and Table 1 exhibit a comprehensive overview of Whisper's model performance, highlighting WER variability across languages and sizes. In Fig. 2, ENG and FRA WERs show little data dispersion. The majority of the ENG and FRA observations cluster around their respective medians, but HAT and ACF WERs display greater skewness.

Table 1. Cross-lingual Whisper WER analysis

Whisper Model Size	Analysis 1: ENG Mean (SD)	Analysis 2: FRA Mean (SD)	Analysis 3: HAT Mean (SD)	Analysis 4: ACF Mean (SD)
Tiny	0.19 (0.02)	0.47 (0.13)	0.98 (0.03)	0.99 (0.09)
Base	0.15 (0.03)	0.39 (0.14)	0.99 (0.13)	1.01 (0.12)
Small	0.11 (0.01)	0.26 (0.05)	0.87 (0.06)	0.99 (0.13)
Medium	0.09 (0.04)	0.22 (0.05)	0.74 (0.04)	0.95 (0.12)
Large	0.08 (0.04)	0.19 (0.03)	0.60 (0.05)	0.86 (0.13)

Within Table 1, the *Large* model displays the lowest WER means, though the standard deviation (SD) results vary. In *Analysis 1*, ENG transcriptions display the lowest WERs at most sizes. Within *Analysis 2*, FRA shows a similar pattern with slightly higher WERs. In Analysis *3*, the *Haitian* model displays significantly higher HAT WERs than ENG and FRA. The *Haitian* model also presents high WERs for ACF in *Analysis 4*.

As model sizes increased, disparities between HAT and ACF became more apparent, despite their initial similarities at smaller sizes. In *Analysis 3,* HAT's *Large* model has a rounded WER mean of 0.60. ACF consistently presented the highest WERs of all languages at all model sizes. Yet, in *Analysis 4*, ACF shows a notable WER decrease between the *Medium* and *Large* model sizes, from 0.95 to 0.86.

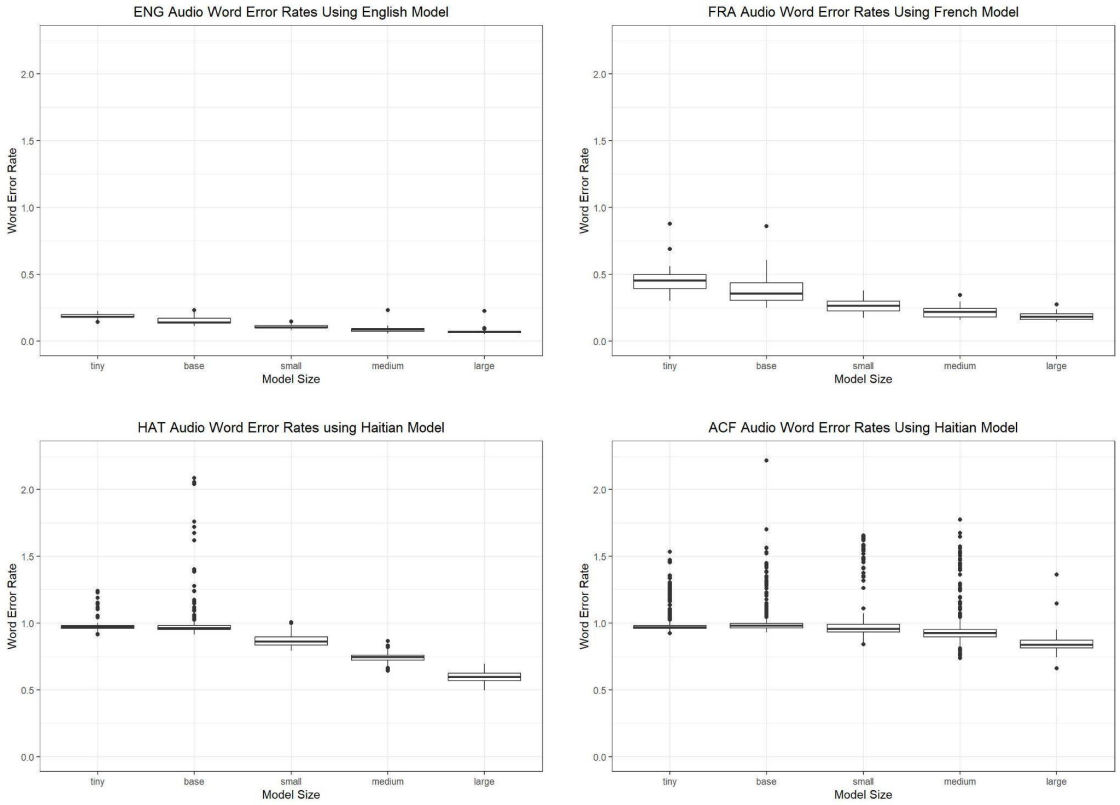

Fig. 2. Box plots displaying Whisper's variability across consecutive cross-lingual runs via aggregate WERs.

V. Discussion

Overall, analyzing consecutive runs across languages and model sizes provides insights into Whisper's ASR system variability, albeit not utilizing IPA. Whisper's developers acknowledged that they had less than 1000 hours of training data on most non-English languages [2]; approximately 74 hours of HAT audio for "Translation" and 1 hour for "Multilingual Speech Recognition" were observed in their training dataset statistics [2]. Despite using the *Large Haitian* model, on average, 60% of the transcribed HAT words differed from the ground truth. Unlike HAT and ACF in Fig. 2, ENG and FRA exhibited lower mean WERs across sizes, with observations clustering around their medians. Their consistent output suggests superior precision and stability in *English* and *French* language model performances. ACF, bearing ENG and FRA etymological components and HAT similarities, still observed a WER of 86% from the *Large Haitian* model. Such high creole WERs expose Whisper's data deficiencies that increase variability.

Fig. 2's ACF WER distribution showcases extended tails, which indicate a wide range of potential outlier values beyond the central tendency and signify variability and unpredictability in ASR performance. Whisper's developers promoted its capacity to generate any UTF-8 string instead of a limited set of graphemes, but this required intricate text standardization rules; thus, their proposed solution bears imperfections that can at times lead to unpredictable outputs. The ACF WER distribution illustrates the underlying ASR non-determinism.

High ACF WERs may stem from limitations in Whisper's *Haitian* model, prompting doubts about assumed ACF and HAT mutual intelligibility. The limited effectiveness of cross-lingual transfer learning could be attributed to poor diversity in the ASR model's training data for covering creole speech variations, leading to performance variability. However, these languages, with shared phonetic patterns from a common historical origin or linguistic borrowing, may present words that sound alike but have diverged in meaning. An ASR system that transfers knowledge without addressing false friends, especially in languages with significant homonymy [10, 13] and polysemy may encounter transcription errors. Models may not discern similar-sounding words with different meanings, affecting their generalization [27]. High WERs from so little input can also spotlight their underlying orthographic differences, as HAT's current writing style [24] may use fewer accented characters than ACF [10, 11].

Although using the same work, cross-lingual ASR transcription variance may also emerge from distinct translation style choices. A document detailing the ACF's reference text translation process may imply potential variability in cross-lingual ASR output [13]. The authors utilized a "meaning-based" or "dynamic equivalent" translation style, employing vocabulary choices for naming people and places that balanced conveying source material meanings with the novelty of ACF pronunciations [13]. Less-than-pristine audio can further hinder accurate transcriptions of uncommon vocabulary or speech patterns, possibly outlining disparities in audio quantity and quality. Therefore, this cross-lingual study's use of small, religion-focused datasets may impact ASR variance.

The temperature parameter in ASR systems' settings introduces randomness during generation by influencing token probability distribution [2, 26]. As Whisper's ASR creativity would not be the goal, tuning would primarily serve to control the decoding process, alleviating challenges in long-form transcription. Instead of employing Whisper's default settings, strategic ASR system adjustments should

enhance determinism and reduce output variability.

Nevertheless, Whisper's *Haitian* model captured the essence of some ACF sounds, resulting in serviceable phonetic transcriptions. Other creole languages with greater mutual intelligibility with HAT might fare better with Whisper's current offerings. The slight improvement between *Medium* and *Large* model sizes suggests that for untrained low-resource audio, selecting a phonetically similar Whisper language option in the largest available model size is the most practical current choice.

The experiment offers valuable cross-lingual creole insight, underscoring the value of religious texts as a basis for comparing languages with limited audio and text data. Yet, a significant obstacle arose from the dichotomy between the abundance of resources in high-resource languages and the challenge of aligning quality resources for the desired cross-lingual analysis. This well-known text amassed considerable content options for widely spoken languages, but procuring proper pairs proved problematic.

VI. Conclusion

This study examined ASR variance via consecutive runs, gauging Whisper's adaptability to phonetic variations in trained and untrained languages. Default settings revealed Whisper's *Large* model as having the reliably lowest variability. Yet, notable variability arose in reportedly mutually intelligible languages; attempts at creole transfer learning revealed phonetic similarities but stark orthographic differences. Whisper's approximate transcriptions nonetheless portend advances in transcribing untrained, low-resource languages. Future studies on Whisper may continue exploring ASR pipeline tasks and potential variations like integrating data from diverse genres, languages, or model sizes. However, it is crucial to acknowledge that Whisper's lack of a user-friendly IPA transcription option may impede further utility and ASR system comparisons. Ultimately, future studies on phonetically similar languages may aid in mitigating ASR variability.

Acknowledgments

The authors acknowledge the SIL International team, mother-tongue translators, and other contributors for their meticulous translation and free publication of this study's Kwéyòl reference text. They also commend composers of public multilingual audio-visual resources and hosting platforms such as ScriptureEarth.org and eBible.org.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

L.L. proposed and conducted the initial research and collected and adjusted the data. M.N. processed the data. L.L. and M.N. collaborated on the analysis and approved the final version.

References

[1] OpenAI. (2018). *OpenAI Charter.* [Online]. Available: https://openai.com/charter

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. International Conference on Machine Learning*, PMLR, 2023, pp. 28492-28518.

[3] C. Christodouloupoulos and M. Steedman, "A massively parallel corpus: the Bible in 100 languages," *Language Resources and Evaluation*, vol. 49, pp. 375-395, 2015.

[4] J. Nivre, "Towards a universal grammar for natural language processing," in *Proc. International Conference on Intelligent Text Processing and Computational Linguistics*, Springer International Publishing, 2015, pp. 3-16.

[5] L. D. Carrington. (1988). Creole discourse and social development," International Development Research Centre (IDRC). [Online]. Available: http://hdl.handle.net/10625/35009

[6] P. Baker, "Creativity in creole genesis," *Creolization and Language Change*, Max Niemeyer Verlag, pp. 65-84, 1994.

[7] H. Simmons-McDonald, "Cultural preservation and language reclamation: The St. Lucian paradox," *Caribbean Quarterly*, vol. 52, no. 4, pp. 57-73, 2006.

[8] L. D. Carrington, "The status of creole in the caribbean," *Caribbean* Quarterly, vol. 45, no. 2-3, pp. 41-51, 1999.

[9] D. B. Frank, "We don't speak a real language: Creoles as misunderstood and endangered languages," in *Symposium on Endangered Languages*, National Museum of Language, College Park, Maryland, 2007, vol. 25.

[10] D. B. Frank, Ed., *Kwéyòl Dictionary*. Ministry of Education, Govt. of Saint Lucia, 2001.

[11] J. E. Mondesir. (1992). *Dictionary of St. Lucian Creole: Part 1: Kwéyòl - English, Part 2: English – Kwéyòl*. [Online]. Available: https://doi.org/10.1515/9783110877267

[12] A. S. Hilaire, "Postcolonialism, identity, and the French language in St. Lucia," *New West Indian Guide/Nieuwe West-Indische Gids*, vol. 81, no. 1-2, pp. 55-77, 2008.

[13] D. B. Frank, "Lexical challenges in the St. Lucian Creole Bible translation project," in *Proc. Twelfth Biennial Conference of the Society for Caribbean Linguistics*, Castries, St. Lucia, 1998, pp. 1-16.

[14] A. Syea, *French Creoles: A Comprehensive and Comparative Grammar*, Taylor & Francis, 2017.

[15] B. Migge, I. Léglise, and A. Bartens, Eds., "Creoles in education: A discussion of pertinent issues," *Creoles in Education: An Appraisal of Current Programs and Projects*, John Benjamins Publishing Company, 2010, pp. 1–30.

[16] S. M. Michaelis, P. Maurer, M. Haspelmath, and M. Huber, *Atlas of Pidgin* and *Creole Language Structures Online*, Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013.

[17] G. Simon. (2023). Welcome to the 26th edition. *Ethnologue*. [Online]. Available: https://www.ethnologue.com/ethnoblog/welcome-26th-edition/

[18] M. S. Dryer and M. Haspelmath, *WALS Online*, Max Planck Institute for Evolutionary Anthropology, 2013.

[19] C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan, "Findings of the 2011 workshop on statistical machine translation," in *Proc. the Sixth Workshop on Statistical Machine Translation*, 2011, pp. 22-64.

[20] W. Lewis, "Haitian Creole: How to build and ship an mt engine from scratch in 4 days, 17 hours, & 30 minutes," in *Proc. the 14th Annual Conference of the European Association for Machine Translation*, 2010.

[21] D. B. Frank. (1993). Political, religious, and economic factors affecting language choice in St. Lucia. *102*, pp. 39-56, Walter de Gruyter. [Online]. Available: https://doi.org/10.1515/ijsl.1993.102.39

[22] J. A. Holm, *Pidgins and Creoles: Volume 2, Reference Survey*, Cambridge, United Kingdom: Cambridge University Press, 1988.

[23] C. Anderson, T. Tresoldi, T. Chacon, A. M. Fehn, M. Walworth, R. Forkel, and J. M. List, "A cross-linguistic database of phonetic transcription systems," *Yearbook of the Poznan Linguistic Meeting*, vol. 4, no. 1, pp. 21-53, 2018.

[24] A. Valdman. (1988). *Ann pale kreyol: An Introductory Course in Haitian Creole.* [Online]. Available: https://eric.ed.gov/?id=ED356617

[25] M. Kadlčík, A. Hájek, J. Kieslich, and R. Winiecki, "A Whisper transformer for audio captioning trained with synthetic captions and transfer learning," 2023, *arXiv:2305.09690*

[26] GDELT. (2022). Experiments With Whisper ASR: Model Parameters & Non-Determinism: Temperature_increment_on _fallback. *The GDELT Project*. [Online]. Available: https://blog.gdeltproject.org

[27] R. Ma, M. Qian, M. J. F. Gales, and K. M. Knill, "Adapting an Unadaptable ASR System," in *Proc. Interspeech,* 2023.

[28] A. K. Rai, S. D. Jaiswal, and A. Mukherjee, "A deep dive into the disparity of word error rates across thousands of NPTEL MOOC videos," 2023, *arXiv:2307.10587*

[29] H. Saadany, C. Orasan, and C. Breslin, "Better transcription of UK Supreme Court Hearings," 2022, *arXiv:2211.17094*
[30] University of Michigan. (2023). Bible: King James Version. [Online]. Available: https://quod.lib.umich.edu/k/kjv/browse. html
[31] B. Bob. (2016). *The Gospel of John KJV Audio Bible with Text.* [Online]. Available: https://www.youtube.com/@bobb.2882
[32] eBible.org. (2023). Bible List. [Online]. Available: https://ebible.org/find/
[33] JesusSecondComing. (2023). The Gospel of John in French. [Online]. Available: https://www.youtube.com/@JesusSecondComing
[34] C. Marc [charlesmarc2012]. (2012). No. 1 Bib La Nan Kreol Ayisien / Haitian Audio Bible," [Video]. [Online]. Available: https://www.youtube.com/@charlesmarc2012
[35] ScriptureEarth. (2023). In language page of scripture earth - the bible in Saint Lucian Creole French. [Online]. Available: https://www.scriptureearth.org/