Working paper · Internal review | v1.0 | 2026-11-20

Test-set choice changes the measured quality of English-to-Kreyòl translation

A comparison of published scores for general sentences, religious text and educational stories

Adrian Dunkley | info@caribbeanaiassociation.com

Abstract

A translation system can earn different scores when evaluators change the reading material. We compared published English-to-Haitian Creole results for the OPUS model across Tatoeba, JW300 and MIT-Haiti. We then compared two models on MIT-Haiti alone. Replacing Tatoeba with educational stories reduced the OPUS score by 30.5 BLEU points and 23.4 chrF points. Switching models on the educational corpus changed the respective scores by 7.3 and 8.1 points. These summary statistics identify a test-design risk for school translation projects. They cannot establish which translations students understand. [1, 2]

Research question

How large is the observed change in translation scores when the English-to-Kreyòl test corpus changes, compared with the change between two models on the same educational corpus?

Data and methods

We selected one language direction throughout. The OPUS model card reports English-to-Haitian Creole scores on Tatoeba and JW300. Lent and colleagues report results for OPUS and CreoleM2M on the same 1,559 English-to-Kreyòl lines from MIT-Haiti. They assembled the multilingual educational material from 60 stories on the MIT-Ayiti platform. We used their aggregate scores and left the sentences and translations with the original providers. [1, 2, 3]

BLEU measures overlap in sequences of words, with a penalty for translations that are too short. chrF measures overlap in sequences of characters using precision and recall. The OPUS card prints chrF as a fraction. We multiplied those fractions by 100 to match the scale used in CreoleVal. Neither metric gives the percentage of sentences translated correctly. [1, 2, 4, 5]

For each metric we calculated three differences. We subtracted the OPUS educational score from its Tatoeba score, repeated that calculation with JW300, and subtracted the OPUS educational score from the CreoleM2M educational score. We kept each metric on its original scale and compared both metrics separately.

Results

For OPUS, the BLEU scores were 45.2 on Tatoeba, 38.3 on JW300 and 14.7 on MIT-Haiti. Its chrF scores were 59.2, 54.5 and 35.8. Moving from either comparison corpus to the educational corpus reduced both scores. The Tatoeba-to-education differences were 30.5 BLEU points and 23.4 chrF points. Substituting JW300 yielded smaller differences of 23.6 and 18.7 points. [1, 2]

On MIT-Haiti, CreoleM2M scored 22.0 BLEU and 43.9 chrF. Its lead over OPUS on the same lines was 7.3 BLEU points and 8.1 chrF points. Both between-corpus differences exceeded the between-model difference for each metric. This ordering held when we changed the comparison corpus from Tatoeba to JW300. [1, 2]

English-to-Kreyòl BLEU scores by test set and model



Higher scores indicate greater reference overlap. The final two bars use the same educational test material. The data table also reports chrF, a character-based measure. Sources: CreoleVal, Table 5 [1], and the OPUS model card [2].

Discussion

A school choosing software needs evidence for the material it will translate. The score from a collection of general sentences could give a misleading picture of the same model’s performance on educational stories. Even an alternative comparison drawn from JW300 left a larger score difference across test sets than the difference between the two systems on the stories.

A follow-up assessment could give Haitian educators full passages and ask two separate questions. Did the translation preserve the concept? Could a student follow the instruction? Keeping the passage as a unit would let reviewers detect errors in references between sentences that a sentence-level comparison may miss.

Limitations

Each corpus differs in more than its subject, including its selection of sentences and reference wording. The differences therefore do not isolate a causal effect of educational content. Original reporting does not establish identical scoring software and settings across all three test sets.

The source reports aggregate metrics. Without paired translations, story identifiers and aligned references in this analysis, we did not calculate uncertainty intervals or count harmful mistranslations. These historical systems and educational stories do not establish the quality of current tools, textbooks or classroom use.

Reproduce the analysis

Take the Tatoeba and JW300 rows from the English-to-Haitian OPUS model card [2]. Multiply chrF by 100.

From Table 5 of CreoleVal [1], extract the two rows with English as source and Haitian Creole as target. Retain both scores.

For each metric, subtract the OPUS MIT-Haiti score from each of the other three rows. Report differences to one decimal place and retain the metric name.

Data table

Test corpus	Model	BLEU points	chrF points (0–100)
Tatoeba	OPUS	45.2	59.2
JW300	OPUS	38.3	54.5
MIT-Haiti	OPUS	14.7	35.8
MIT-Haiti	CreoleM2M	22.0	43.9

References

[1] Lent, H. et al. (2024). CreoleVal: Multilingual Multitask Benchmarks for Creoles. Transactions of the Association for Computational Linguistics, Section 4.2 and Table 5.

[2] Helsinki-NLP. OPUS-MT English-to-Haitian model card, benchmark scores for Tatoeba and JW300.

[3] Platfòm MIT-Ayiti. Educational resources in Haitian Creole.

[4] Papineni, K. et al. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation.

[5] Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation.

Data credits

Lent et al. (2024), Table 5. CAIA comparisons. CC BY 4.0