Documento de trabajo · Revisión interna | v1.0 | 2026-11-20

# La elección de los textos de prueba cambia la calidad medida de la traducción del inglés al criollo haitiano

Comparación de puntuaciones publicadas para frases generales, textos religiosos y relatos educativos

Adrian Dunkley | info@caribbeanaiassociation.com

## Resumen

Un sistema de traducción puede obtener puntuaciones distintas según los textos que se usen para evaluarlo. Comparamos los resultados publicados de OPUS del inglés al criollo haitiano en Tatoeba, JW300 y MIT-Haiti. Después comparamos dos modelos solo en MIT-Haiti. Al sustituir Tatoeba por relatos educativos, OPUS perdió 30,5 puntos BLEU y 23,4 puntos chrF. Cambiar de modelo en el corpus educativo modificó las puntuaciones en 7,3 y 8,1 puntos respectivamente. Estas cifras identifican un riesgo en el diseño de pruebas para proyectos escolares de traducción. No indican qué traducciones comprenden los estudiantes. [1, 2]

## Pregunta de investigación

¿Qué magnitud tiene el cambio de puntuación al sustituir el corpus de prueba, comparado con el cambio entre dos modelos que traducen el mismo corpus educativo del inglés al criollo haitiano?

## Datos y métodos

Mantuvimos una sola dirección de traducción. La ficha de OPUS da puntuaciones para Tatoeba y JW300. Lent y sus colaboradores evaluaron OPUS y CreoleM2M con las mismas 1.559 líneas del MIT-Haiti del inglés al criollo haitiano. Reunieron el material educativo multilingüe a partir de 60 relatos de la plataforma MIT-Ayiti. Utilizamos sus resultados agregados. Las frases y traducciones siguen con sus proveedores de origen. [1, 2, 3]

BLEU mide la coincidencia de secuencias de palabras y penaliza las traducciones demasiado cortas. chrF compara secuencias de caracteres mediante precisión y exhaustividad. La ficha de OPUS expresa chrF como fracción. Multiplicamos esas cifras por 100 para usar la escala de CreoleVal. Ninguna de las métricas da el porcentaje de frases traducidas correctamente. [1, 2, 4, 5]

Calculamos tres diferencias por métrica. Restamos a la puntuación de OPUS en Tatoeba su puntuación educativa y repetimos el cálculo con JW300. Luego restamos la puntuación educativa de OPUS a la de CreoleM2M en esos mismos textos. Comparamos cada métrica por separado en su propia escala.

## Resultados

OPUS obtuvo 45,2 puntos BLEU en Tatoeba, 38,3 en JW300 y 14,7 en MIT-Haiti. En chrF obtuvo 59,2, 54,5 y 35,8. Ambos indicadores bajaron al pasar de cualquiera de los dos corpus de comparación al educativo. Entre Tatoeba y el corpus educativo, las diferencias fueron 30,5 puntos BLEU y 23,4 chrF. Con JW300 fueron menores: 23,6 y 18,7 puntos. [1, 2]

En MIT-Haiti, CreoleM2M alcanzó 22,0 puntos BLEU y 43,9 chrF. En esas mismas líneas superó a OPUS por 7,3 puntos BLEU y 8,1 chrF. Para cada métrica, las dos diferencias entre corpus superaron la diferencia entre modelos. Esta relación se mantuvo al cambiar Tatoeba por JW300 como referencia. [1, 2]

## Puntuaciones BLEU del inglés al criollo haitiano según corpus y modelo

puntos BLEU




Una puntuación mayor indica más coincidencias con la referencia. Las dos últimas barras usan el mismo material educativo. La tabla incluye también chrF, que compara caracteres. Fuentes: CreoleVal, tabla 5 [1], y ficha de OPUS [2].

## Discusión

Una escuela que evalúe programas de traducción necesita pruebas sobre los materiales que utilizará. Una puntuación en frases generales puede dar una impresión equivocada de lo que hará el mismo programa con relatos educativos. Incluso al usar JW300 como referencia, cambiar de corpus tuvo mayor efecto en las puntuaciones que cambiar de sistema en los relatos.

Una evaluación posterior podría entregar pasajes completos a docentes haitianos y plantear dos preguntas. ¿Conservó la traducción el concepto? ¿Podría un estudiante seguir la instrucción? Leer el pasaje completo permitiría detectar errores entre frases que una evaluación aislada podría pasar por alto.

## Limitaciones

Los corpus difieren en el tema, la selección de frases y la redacción de las referencias. No se puede atribuir al contenido educativo por sí solo la diferencia observada. Las fuentes no establecen que el programa y los ajustes de puntuación fueran idénticos en las tres pruebas.

Las fuentes publican métricas agregadas. Sin traducciones emparejadas, identificadores de relatos y referencias alineadas en este análisis, no calculamos intervalos de incertidumbre ni contamos errores perjudiciales. Estos sistemas históricos y relatos educativos no establecen la calidad de las herramientas actuales, los libros de texto ni su uso en clase.

## Reproducir el análisis

Copia las filas Tatoeba y JW300 de la ficha de OPUS del inglés al criollo haitiano [2]. Multiplica chrF por 100.

En la tabla 5 de CreoleVal [1], extrae las dos filas con inglés como origen y criollo haitiano como destino. Conserva las dos puntuaciones.

En cada métrica, resta la puntuación de OPUS en MIT-Haiti a cada una de las otras tres filas. Da la diferencia con un decimal e indica la métrica.

## Tabla de datos

| Corpus de prueba | Modelo | Puntos BLEU | Puntos chrF (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 |

## Referencias

[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.

## Créditos de los datos

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