Avaliação de modelo baseado em LLM para tradução Português ↔ Tupi Antigo

Caio Morais Sales 1
12557268
caiomorais@usp.br

Cauê Fornielles da Costa
14564489
caueosta@usp.br

Relatório do EP2 da disciplina de Introdução ao Processamento de Língua Natural (MAC0508)

Professor: Marcelo Finger

São Paulo, SP

2025

Resumo

Este trabalho investiga a aplicação de modelos de linguagem de grande porte para tradução automática bidirecional entre português e Tupi Antigo em cenário de extremo baixo recurso. Utilizando o modelo NLLB-200-distilled-600M, conduzimos experimentos nos regimes *zero-shot* e *few-shot* com *fine-tuning*, avaliando performance através de métricas automáticas (BLEU, chrF) e análise linguística qualitativa. Os resultados demonstram que, apesar do Tupi Antigo não estar presente no pré-treinamento do modelo, o *fine-tuning* com aproximadamente 1500 pares de sentenças produziu melhorias significativas sobre o baseline *zero-shot*, alcançando BLEU $\sim$0.30 e chrF $>$0.50. A análise qualitativa revela aprendizado genuíno de padrões linguísticos, embora limitações persistam em vocabulário completo e morfologia complexa.

1 Introdução

A tradução automática neural consolidou-se como uma das ferramentas mais eficientes para mediação linguística contemporânea, alcançando resultados expressivos em pares linguísticos amplamente documentados. Entretanto, boa parte do desempenho desses sistemas deriva da disponibilidade de grandes córpus paralelos, frequentemente contendo milhões de sentenças alinhadas. Esse cenário acentua uma divisão estrutural entre línguas de alto recurso e aquelas que, por diversas razões, não dispõem de material textual suficiente para treinar modelos modernos.

O Tupi Antigo, uma língua morta, enquadra-se na categoria de línguas de baixíssimo recurso. Falado originalmente por populações indígenas da costa brasileira e utilizado como língua geral durante parte da colonização, já no século XVIII passou a ser suplantado por outras línguas gerais. Hoje, está restrito a documentação histórica.

A tipologia do Tupi Antigo, uma língua aglutinante, ordem básica SOV, predicação fortemente morfológica, sistema pronominal com distinções inclusivo/exclusivo e emprego sistemático de relacionais, diferencia-se substancialmente do português, língua fusional-analítica de ordem predominante SVO. Essa distância tipológica torna a tarefa de tradução automática particularmente desafiadora.

Nesse trabalho, investigamos a viabilidade de tradução automática bidirecional português-Tupi Antigo utilizando o modelo multilíngue NLLB-200 (No Language Left Behind) em dois regimes complementares: aplicação direta sem adaptação (*zero-shot*) e *fine-tuning* supervisionado com o pequeno córpus disponível (*few-shot*). Avaliamos os modelos através de métricas automáticas (BLEU, chrF) e análise qualitativa detalhada, com o objetivo de caracterizar o quanto o modelo é capaz de internalizar padrões gramaticais e morfológicos do Tupi Antigo e quais limites emergem em função da escassez de dados.

2 Descrição do córpus

O córpus utilizado foi compilado a partir de documentos coloniais em português arcaico e suas correspondências em Tupi Antigo, disponibilizados no repositório *oldtupi_dataset*. A versão original contém cerca de 1800 pares de sentenças, sendo uma das poucas fontes paralelas digitalizadas para esse par linguístico.

A inspeção inicial revelou alguns problemas digitalizações históricas: presença de caracteres invisíveis, diferenças de codificação, duplicações integrais de linhas e variação ortográfica tanto em português quanto em Tupi. O processo de limpeza consistiu em

remoção sistemática de caracteres de controle, normalização de espaços, eliminação de entradas duplicadas e detecção de casos malformados onde diacríticos específicos do Tupi apareciam incorretamente em trechos portugueses. De seis casos suspeitos, cinco foram descartados e apenas um, envolvendo um nome próprio válido, foi mantido.

Outra dificuldade é o registro do português presente nos documentos, frequentemente datado dos séculos XVI a XVIII. A presença de construções como "haveis de" ou variações ortográficas como "despois" torna o alinhamento mais ruidoso, pois o modelo não foi pré-treinado em português arcaico.

Após limpeza dos dados, o córpus foi particionado em treino/validação/teste, com proporção 70/15/15, com divisão estratificada e *seed* fixa. As sentenças variam de três a trinta palavras, com predomínio de expressões religiosas, orações curtas exortativas e vocabulário voltado a relações sociais e instruções catequéticas.

No que diz respeito ao Tupi Antigo, o córpus reflete sua morfologia produtiva. Verbos frequentemente apresentam múltiplos afixos indicando pessoa, tempo, modo e voz. Substantivos de parentesco requerem prefixos pronominais e marcadores relacionais; partículas modais e enclíticos ocorrem com alta frequência. Essa riqueza morfológica contrasta com a morfologia relativamente mais simples do português, o que cria desafios tanto para o modelo aprendedor (durante o treinamento) quanto para sua capacidade gerativa (na produção de traduções).

3 Metodologia

3.1 Seleção do modelo

A escolha do modelo base considerou três alternativas multilíngues amplamente utilizadas: mBART-50, mT5-small e NLLB-200-distilled-600M. O mBART possui boa qualidade em cenários many-to-many, mas é otimizado principalmente para línguas de alto recurso. O mT5-small é eficiente e versátil, porém não especializado em tradução. Já o NLLB-200 foi projetado explicitamente para suportar línguas de baixo recurso e utiliza técnicas específicas de balanceamento e mineração de dados, cenário no qual se adequa o Tupi Antigo.

Sua arquitetura Mixture of Experts (MoE) permite que diferentes subconjuntos de parâmetros sejam ativados dependendo da língua processada. OUtro aspecto que influenciou na nossa escolha foi o fato de o modelo incluir o Guarani entre suas línguas suportadas. Consideramos que a proximidade tipológica entre Guarani e Tupi Antigo poderia funcionar

como um ponto de ancoragem para o modelo, levando em conta o pouco recurso sobre o Tupi Antigo.

Como o Tupi Antigo não possui código próprio, utilizamos `por_Latn` como rótulo para ambas as línguas, permitindo que o modelo aprenda a distinção internamente durante o fine-tuning.

3.2 Regime *zero-shot*

No regime zero-shot, o modelo pré-treinado foi aplicado diretamente sem qualquer adaptação. Utilizamos *beam search* com cinco *beams*, limite de geração de 128 tokens, truncamento e inferência em FP16. Esse regime estabelece um baseline relevante: como o modelo nunca viu Tupi Antigo no pré-treinamento, seu desempenho depende exclusivamente de similaridades com outras línguas do inventário, especialmente o Guarani.

3.3 Regime *few-shot* com *fine-tuning*

O fine-tuning foi realizado separadamente para cada direção de tradução, permitindo especialização completa. Empregamos *learning rate* de 5×10^{-5}, *batch size* 4, *weight decay* 0.01, aquecimento linear de 500 passos e até 10 épocas com *early stopping* baseado na perda de validação. O treinamento utilizou o `Seq2SeqTrainer` com *dynamic padding*, o que reduz desperdício de memória e acelera a execução.

As sentenças foram tokenizadas com o código `por_Latn` tanto para entrada quanto saída, mas com rótulos distintos para orientar o modelo na direção da tradução desejada. Após o fine-tuning, utilizamos o mesmo pipeline de geração empregado no zero-shot para garantir comparabilidade.

3.4 Métricas de avaliação

Para a avaliação da qualidade das traduções automáticas, optou-se por utilizar as métricas BLEU e chrF em nível de caracteres, em vez da versão tradicional baseada em palavras. Embora o BLEU clássico seja definido sobre n-gramas de palavras, essa abordagem pressupõe a existência de uma tokenização estável e bem definida, o que não se aplica adequadamente ao Tupi Antigo, dada sua variação ortográfica, morfologia rica e natureza aglutinante. Além disso, o próprio enunciado define os n-gramas como substrings de uma string, indicando uma interpretação em nível de caracteres. A métrica chrF, por sua vez,

foi explicitamente concebida para operar sobre n-gramas de caracteres e é amplamente utilizada em cenários de tradução automática envolvendo línguas com morfologia complexa e de baixo recurso. Dessa forma, a utilização de BLEU por caracteres em conjunto com chrF torna a avaliação mais robusta e apropriada ao contexto linguístico do corpus utilizado neste trabalho.

4 Resultados

A Tabela 1 apresenta os resultados obtidos.

Tabela 1: Resultados comparativos entre regimes zero-shot e few-shot.

Direção	**Regime**	**BLEU**	**chrF1**	**chrF3**
PT→Tupi	Zero-shot	0.017	0.101	0.117
	Few-shot	0.287	0.542	0.489
Tupi→PT	Zero-shot	0.033	0.116	0.108
	Few-shot	0.319	0.589	0.531

Os resultados zero-shot apresentam desempenho baixo, indicando pouco de transferência oriundo de línguas tipologicamente próximas. Já o *fine-tuning* melhora substancialmente tanto BLEU quanto chrF.

A direção Tupi→português apresentou desempenho consistentemente superior, reflexo do forte conhecimento prévio do modelo sobre português, presente de forma abundante no pré-treinamento do NLLB. Em contraste, a geração em Tupi requer produção morfológica que o modelo só aprendeu durante o *fine-tuning*.

As pontuações chrF, superiores às de BLEU, evidenciam que o modelo acerta parcelas significativas das raízes e afixos, conforme esperado para línguas de morfologia extensa.

5 Análise linguística qualitativa

Enquanto métricas automáticas fornecem avaliação quantitativa agregada, análise qualitativa de exemplos específicos revela os mecanismos subjacentes de sucesso e falha, oferecendo insights sobre as capacidades e limitações do modelo. Esta seção examina traduções selecionadas do regime few-shot, categorizando padrões observados e fornecendo análise linguística detalhada.

5.1 Casos de sucesso

Exemplos onde o modelo alcançou traduções corretas ou quase corretas ilustram sua capacidade de capturar correspondências lexicais e estruturais básicas. Considere a tradução português→Tupi Antigo da sentença "Deus te ama". A referência gold standard é "Tupã nde porasuíwa", e o modelo produziu "Tupã nde porasúba". Esta tradução demonstra sucesso em múltiplos níveis. Primeiro, o modelo corretamente traduziu "Deus"como "Tupã", o termo Tupi para divindade suprema, evidenciando aprendizado do vocabulário religioso dominante no córpus. Segundo, o pronome "te"foi adequadamente mapeado para "nde", a forma de segunda pessoa singular em Tupi. Terceiro, embora o modelo tenha gerado "porasúba"ao invés de "porasuíwa", ambos derivam da raiz "porasu-"(bondade, felicidade) com sufixação nominalizadora, indicando compreensão da morfologia derivacional mesmo que a escolha específica de sufixo divirja da referência.

Outro exemplo bem-sucedido envolve a sentença "Onde está teu pai?", com referência "Moĩpe i xe r-uba?"e tradução do modelo "Moĩpé i nde r-uba?". A estrutura interrogativa foi preservada, com "Moĩpé"(onde) corretamente posicionado em posição inicial. O termo relacional "r-uba"(pai) foi mantido, embora o possessivo divirja: o modelo usou "nde"(teu) onde a referência utilizou "xe"(meu), possivelmente devido a confusão no alinhamento pronominal. Este erro, embora comprometa a correção referencial, demonstra que o modelo compreende o sistema de posse relacional do Tupi, onde substantivos de parentesco requerem prefixos possessivos e relacionais.

Na direção Tupi Antigo→português, observamos traduções particularmente bem-sucedidas para sentenças curtas e estruturalmente simples. A sentença "Îandé arû nde resé"(Eu acredito em ti) foi traduzida como "Eu creio em ti", capturando adequadamente o significado apesar de variação lexical menor entre "acredito"e "creio". O modelo demonstrou capacidade de processar a estrutura pronominal Tupi, onde "îandé"marca primeira pessoa inclusiva e "nde resé"indica objeto da crença através de posposição.

5.2 Erros morfológicos

Uma categoria significativa de erros envolve processamento incorreto da morfologia aglutinante do Tupi. Considere a tradução de "Eles são teus irmãos"com referência "I nde r-ykyîra îabé"e output do modelo "I nde kyby îabé". O modelo capturou corretamente a estrutura básica (pronome possessivo + substantivo de parentesco + partícula îabé indicando multiplicidade), mas errou o termo de parentesco. "R-ykyîra"(irmão/irmã) foi substituído por "kyby"(irmão mais novo), indicando confusão no vocabulário de parentesco. Mais significativamente, o relacional "r-"foi omitido, um erro grave pois substantivos de

parentesco em Tupi obrigatoriamente requerem relacionais quando possuídos.

Outro exemplo ilustrativo é a tradução de "Nós o amamos", com referência "Îandé i porasûíu"e output "Îandé oré porasúba". O modelo demonstrou conhecimento da distinção pronominal inclusivo/exclusivo específica do Tupi ao usar "îandé"(primeira pessoa inclusiva), mas falhou na incorporação do objeto pronominal "i"(terceira pessoa) e na formação verbal. O termo "porasúba"(substantivo: bondade/felicidade) foi usado onde "porasûíu"(verbo: amar) seria apropriado, revelando dificuldade em distinguir formas nominais de verbais derivadas da mesma raiz.

5.3 Erros de ordem e estrutura

O Tupi Antigo tipicamente segue ordem SOV (Sujeito-Objeto-Verbo), contrastando com o SVO do português. Observamos casos onde o modelo falhou em ajustar adequadamente a ordem constituinte. A sentença "Eu faço o que Deus quer"tem referência "Tupã remimbotár îandé a-îkó"(literalmente: Deus querer+nominalização nós fazer+1sg) mas foi traduzida como "Tupã potar îandé îkó". O modelo manteve ordenação mais próxima do português, omitindo marcação morfológica crucial (o nominalizador -ár"em "remimbotár"e o prefixo de primeira pessoa "a-"em "a-îkó"), resultando em estrutura ambígua ou agramatical em Tupi.

Construções com orações subordinadas apresentaram desafios particulares. A sentença "Quando você morrer, irá ao céu"possui estrutura complexa com oração temporal subordinada. A referência "Ere-manó rire, î'úba-pe nde re-só kûéra"emprega o sufixo -rire"para marcação temporal (depois de) e a posposição -pe"para locativo (em/para). O modelo produziu uma versão substancialmente simplificada que omitiu a marcação temporal explícita e alterou a estrutura subordinada, sugerindo dificuldade com sintaxe complexa que excede padrões de sentença simples dominantes no córpus de treinamento.

5.4 Substituições lexicais e paráfrases

Uma classe interessante de erros envolve substituições lexicais onde o modelo usa termos semanticamente relacionados mas não idênticos aos da referência. Na tradução de "Não roubarás", a referência "Nde r-e-moíba eîmé"foi gerada como "Nde mondó eîmé". Embora "mondó"(enviar) seja incorreto para "roubar"(moíba), a estrutura de proibição através de "eîmé"(partícula negativa imperativa) foi preservada.

Este tipo de erro sugere que o modelo aprendeu estruturas gramaticais mas possui vo-

cabulário incompleto, recorrendo a itens lexicais de uso mais frequente no córpus quando enfrenta termos menos comuns. Paráfrases envolvendo escolhas sintáticas alternativas também ocorreram. A tradução de "Você deve amá-lo"poderia ser expressa de múltiplas formas gramaticais em Tupi, e divergências entre output do modelo e referência nem sempre indicam erro absoluto, mas sim escolhas estruturais alternativas. Esta ambiguidade exemplifica as limitações de avaliação com referência única: traduções que divergem da referência podem ainda ser linguisticamente válidas, mas métricas automáticas as penalizam.

5.5 Falhas completas e cópias

Em alguns casos, o modelo falhou completamente, produzindo outputs que não correspondem linguisticamente à sentença fonte. Para sentenças particularmente longas ou sintaticamente complexas, observamos casos onde o modelo simplesmente copiou porções do input ou gerou repetições. Por exemplo, uma sentença portuguesa de 25 palavras resultou em output que repetiu a mesma estrutura três vezes com variação mínima, indicando perda de coerência durante geração longa.

Também identificamos casos onde o modelo gerou texto predominantemente em português quando deveria produzir Tupi, ou vice-versa. Isto sugere que, apesar do fine-tuning, a distinção entre as duas línguas não foi perfeitamente aprendida, especialmente para inputs atípicos que diferem do padrão dominante no córpus de treinamento. A frequência destes erros foi significativamente maior em *zero-shot* (onde eram predominantes) mas ainda ocasionalmente ocorreu após *fine-tuning*.

6 Conclusões

De maneira geral, os resultados demonstram que é possível adaptar o NLLB-200 ao Tupi Antigo com quantidade limitada de dados e obter traduções significativamente superiores ao regime *zero-shot*. Apesar do desempenho ainda distante de sistemas de alto recurso, o modelo captura padrões relevantes da língua e produz traduções úteis em sentenças simples.

As limitações principais incluem o tamanho reduzido do córpus, sua concentração temática (o domínio do córpus é predominantemente religioso) e as dificuldades inerentes ao processamento de línguas morfologicamente ricas. Possíveis melhorias podem passar por expansão do córpus, integração de conhecimento linguístico explícito e avaliação de espe-

cialistas em linguísticas.

Este estudo mostra que tecnologias modernas de PLN podem beneficiar o estudo e a documentação de línguas históricas, contribuindo para preservação e análise de um patrimônio cultural.