Universidade de São Paulo
Instituto de Matemática e Estatística
Curso de Ciências Moleculares

Relatório do Exercício Programa II

Tradução Automática de Baixo Recurso:
Implementação, Treinamento e Avaliação de Modelos Baseados em LLM para Tradução Português ↔ Tupi Antigo

Autores:

Thaís Martins de Sousa[1*] | NUSP 14608786

Gustavo Bernardo Ribeiro[2*] | NUSP 14577174

Disciplina: MAC0508 - Introdução ao Processamento de Língua Natural

Professor: Marcelo Finger

São Paulo

2025

1 - thais_martins@usp.br
2 - gustavo.bernardo@usp.br
* - Igual contribuição dos autores

1. Introdução

A área de *Machine Translation* (MT) teve avanços significativos ao longo da década de 2010 impulsionados tanto pelo desenvolvimento de modelos de redes neurais voltados para tradução (SUTSKEVER et al., 2014; BAHDANAU et al., 2015; GEHRING et al., 2016; VASWANI et al., 2017, apud GUZMÁN et al., 2019) quanto pela disponibilidade crescente de grandes corpora paralelos entre línguas (TIEDEMANN, 2012; SMITH et al., 2013; BOJAR et al., 2017, apud GUZMÁN et al., 2019) , que são conjunto de pares de sentenças, onde cada par contém uma sentença-fonte e sua tradução equivalente. Esses progressos possibilitaram que os sistemas atuais alcançassem desempenho próximo ao humano em línguas que dispõem de grandes quantidades de dados paralelos de treinamento (GUZMÁN et al., 2019). No entanto, o mesmo não acontece para línguas de baixo recurso.

Apesar do termo "low resource language" ainda variar amplamente dentro da comunidade de NLP e não ter uma única definição bem objetiva (NIGATU, 2024),, no contexto de MT considera-se de baixo recurso a língua para a qual não existem dados paralelos suficientes para treinar modelos de grande porte sem *overfitting*. Esse cenário é agravado pela distribuição desigual das tecnologias de NLP: mais de 80% da população mundial não fala inglês, apesar de essa língua concentrar grande parte dos avanços da área, e mesmo as dez línguas mais faladas representam menos de metade da população global (RANZATO, 2021). Como consequência, milhares de línguas permanecem tecnicamente subatendidas, reforçando a necessidade de métodos específicos para tratamento de dados escassos.

O Tupi Antigo se insere nesse contexto como uma língua de baixo recurso, tanto pela ausência de corpora paralelos quanto pelo processo histórico de apagamento linguístico decorrente de políticas coloniais, como o Diretório dos Índios de 1758, do Marquês de Pombal, que proibiu seu uso público e interrompeu sua transmissão intergeracional.

Diante disso, o desenvolvimento de ferramentas computacionais dedicadas torna-se crucial para a preservação e a revitalização do Tupi Antigo. Tais esforços alinham-se às diretrizes da Década Internacional das Línguas Indígenas (2022–2032), estabelecida pela UNESCO, que enfatiza a importância de garantir a sobrevivência e o fortalecimento das línguas originárias no ambiente digital.

2. Descrição do córpus

O corpus utilizado como base para os trabalhos de tradução desse relatório foi obtido no repositório GitHub de Calebe Rezende. Ele foi construído a partir das seguintes obras históricas e de estruturação gramatical do Tupi: *Curso de Tupi Antigo*, de R.A.L. Barbosa; *A Role and Reference Grammar Description of Tupinambá*, de Fabricio Ferraz Gerardi; *Dicionário de Tupi Antigo: a língua indígena clássica do Brasil*, de Eduardo de Almeida Navarro; Poemas: Tupi e Português, de José de Anchieta, com edições bilíngues organizadas por Navarro.

Todos os pares de frases Português Arcaico - Tupi Antigo foram revisados manualmente em trabalho prévio (REZENDE 2025) e a ortografia normalizada seguindo o Dicionário de Tupi Antigo

(NAVARRO 2013). Esse esforço de garantir consistência interna do corpus é essencial para a qualidade do treinamento supervisionado.

Devido a natureza desse corpus, suas construções frasais são em Português Arcaico, do século XVI. Essa característica pode diminuir a acurácia das traduções, dado que todos os modelos a serem usados são majoritariamente pré-treinados com dados da internet, de português contemporâneo. Por isso, nós propomos a modernização da face em português do corpus paralelo, na tentativa de obter melhores resultados de tradução.

Essa modernização foi realizada com auxílio do modelo de linguagem (*gpt-4o-mini*) via API da OpenAI, gerando uma versão contemporânea do lado em português com direcionamento no prompt do sistema para não haver alteração no conteúdo semântico e preservar entidades como nomes próprios, datas e fatos. Alguns exemplos de alteração (ou manutenção) das frases podem ser vistos na **Tabela 1**.

Tabela 1 - *Exemplos de modernização do português: original vs. modernizado*

Frase no Corpus Original (REZENDE 2025)	***Frase modernizada via gpt-4o-mini***
matai-o!	**matem-no!**
se **tu** não **endureceres teu** coração	se **você** não **endurecer seu** coração
tão logo ao ouvir o nome dela, **em outra parte eu me escondo**	**assim que** ouvir o nome dela, **eu me escondo em outra parte.**
por isso mesmo em **tua** grande força **apóio-me**	por isso mesmo em **sua** grande força **me apoio**
embora eles fossem obra **d'ele**, alguns anjos tornaram-se maus	embora eles fossem obra **dele**, alguns anjos tornaram-se maus
eu retomei o bom senso	eu retomei o bom senso

Legenda: *As amostras selecionadas mostram alterações de tratamento do interlocutor, flexão verbal de pessoa, colocação pronominal e grafia. Na última célula há um exemplo de frase que não foi alterada.*
Fonte: *elaboração própria*

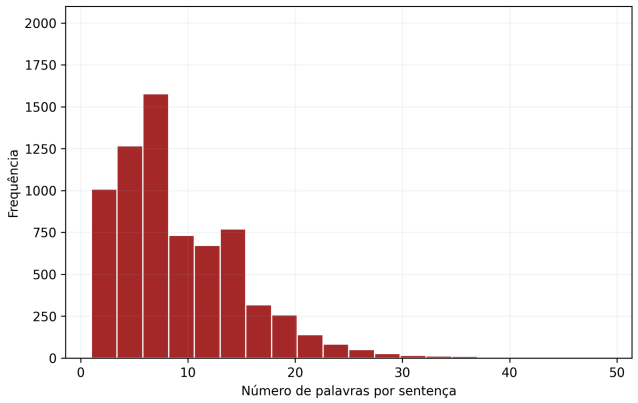
Na seção de Metodologia, descrevemos o procedimento de modernização (instruções de reescrita, parâmetros de geração, etc.) e disponibilizamos o código utilizado no repositório GitHub deste projeto. Também continuamos os experimentos tanto com o Português original quanto com o Português modernizado, permitindo comparação posterior dos resultados e do impacto dessa normalização nos experimentos de tradução.

Tabela 2 - *Medidas estatísticas dos corpora de Português Arcaico e Moderno*

	Mediana	Média	Desvio Padrão
Corpus Português Arcaico (Original)	8.00	9.10	6.01
Corpus Português Moderno	6.00	6.32	6.70

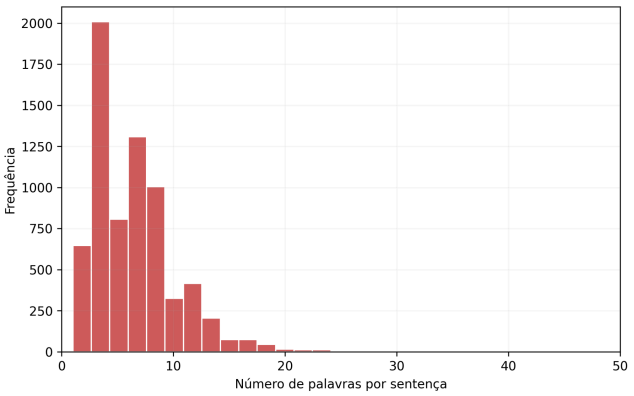
Fonte: *Elaboração própria*

Gráfico 1 - *Palavras por sentença do Português Arcaico*



Fonte: *Elaboração própria*

Gráfico 2 - *Palavras por sentença do Português Moderno*



Fonte: *Elaboração própria*

Analisando ambos os corpora, temos que o corpus de Português Arcaico contém 63270 palavras e distribuição de palavras por sentença de acordo com o **Gráfico 1**. Já o corpus do Português Moderno foi reduzido para 43935 palavras (o que indica compressão das frases) e distribuição de palavras por sentença de acordo com o Gráfico 2. Medidas estatísticas dos corpora podem ser observadas na **Tabela 2.**

3. Metodologia

3.1. Normalização e Modernização do Corpus

Para a normalização do corpus, foi utilizada a biblioteca Pandas para a remoção de pontuação entre colchetes, espaços em branco desnecessários, quebras de linha, numeração de versículos, dos trechos em parênteses do corpus em Português (que não tinham correspondência no lado do Tupi), e conversão das letras maiúsculas para minúsculas.

Além da limpeza estruturada do corpus, aplicamos uma etapa de modernização no lado em português, convertendo os segmentos para português brasileiro contemporâneo. Essa etapa foi realizada via API da OpenAI, utilizando o modelo gpt-4o-mini com temperatura 0.0, de modo a garantir comportamento mais determinístico.

A reescrita foi guiada por um prompt restritivo **(Figura 1)**, impondo a preservação do conteúdo semântico e proibindo sumarização, interpretações e alterações em nomes próprios, datas e fatos. Como hiperparâmetros, definimos um máximo de 30 sentenças por batch, para prevenir alucinações, e definimos um limiar de *fallback* de similaridade (`difflib.SequenceMatcher(None, a, b)`) de 0.3, para manter a versão original de frases que o modelo alterasse drasticamente e que, na maior parte dos casos, configuraria um erro. Para rastreabilidade inicial, mantivemos o texto original e o texto modernizado e preservamos o id de cada item. Depois, o arquivo .csv modernizado foi rearranjado como o original e convertido para minúsculas.

Figura 1 - *Prompt com instruções fornecido ao modelo gpt-4o-mini*

```
# ========================
# PROMPTS
# ========================
PROMPT_SISTEMA = """
Você é um assistente especializado em normalizar português antigo para
português brasileiro contemporâneo.

Instruções obrigatórias:
- Reescreva cada texto preservando totalmente o sentido original.
- Não resuma, não interprete, não invente informações novas, nem substitua
trechos por expressões vagas ("de verdade", "tipo assim", etc.).
- Não omita partes do texto: todo conteúdo informativo do original deve
aparecer na versão modernizada.
- Não altere nomes próprios, datas, referências históricas ou fatos.
- Não altere capitalização (maiúsculas/minúsculas) do texto original.
- Não introduza novas frases, exemplos ou explicações.
- Não altere o tempo verbal dos verbos.
- Reescreva com fluência natural em português brasileiro contemporâneo, podendo
ajustar a ordem das palavras e pontuação quando necessário, sem acrescentar
conteúdo.
- Normalize apenas construções arcaicas, ortografia e conectores.
- Padronize tratamento do interlocutor para PT-BR contemporâneo:
  "a ti"→"a você"; formas de "vós" (ex.: "fazei", "dizei") → "vocês" +
  imperativo correspondente ("façam", "digam").
  Em ambiguidade de número, prefira "vocês" (plural).
- Em caso de dúvida sobre o significado de uma expressão arcaica, mantenha a
expressão (corrigindo só ortografia/pontuação), em vez de substituir por
paráfrase livre.
- Não modifique o campo "id" de cada item.
- Retorne apenas um array JSON válido, sem texto adicional, sem comentários e
sem blocos de código.

Formato de saída obrigatório:
[
  {{"id": <mesmo id fornecido>, "modernizado": "<texto modernizado em português
  brasileiro contemporâneo>"}},
  ...
]

Requisitos estritos:
- A saída deve ser SOMENTE JSON.
- Não inclua delimitadores como ```json.
- Não mude a ordem dos itens.
- Garanta que todo texto modernizado esteja corretamente escapado em JSON.
""".strip()
```

Fonte: *Elaboração própria*

Ambas as versões do corpus foram então divididas em subconjuntos de treino, validação e teste, sob a proporção 8-1-1, respectivamente. Na etapa Zero-Shot, apenas os dados de teste foram utilizados.

3.2. Pipeline Zero-Shot

A ideia aqui foi tornar a etapa Zero-Shot como um dos critérios de escolha do modelo a passar para a fase de *fine-tuning*, por isso os experimentos foram realizados tanto com o mBART, quanto com o NLLB e o mT5. Como o tupi antigo não consta no vocabulário oficial de nenhum dos modelos, foi necessária a adoção de estratégias distintas de configuração para cada um:

- **mBART:** Devido à ausência de línguas indígenas sul-americanas neste modelo, utilizamos o código do Português (pt_XX) tanto para a língua de origem quanto para a de destino:

```
df['pred_ida'] = run_mbart(df, 'source_text', "pt_XX", "pt_XX")
df['pred_volta'] = run_mbart(df, 'tupi_clean', "pt_XX", "pt_XX")
```

- **NLLB:** Neste modelo, foi utilizado o código do guarani (grn_Latn) para o tupi antigo. Devido à proximidade filogenética da família Tupi-Guarani esperava-se que isso contribuísse para que o modelo gerasse construções morfológicas válidas, mesmo sem treinamento específico.

```
df['pred_ida'] = run_nllb(df, 'source_text', "por_Latn", "grn_Latn")
df['pred_volta'] = run_nllb(df, 'tupi_clean', "grn_Latn", "por_Latn")
```

- **mT5:** Neste modelo, foi dado o prompt em linguagem natural "translate Portuguese to Tupi:" em ambos os casos, delegando ao modelo a tarefa de inferir o objetivo semântico diretamente a partir do contexto textual.

```
df['pred_ida'] = run_mt5(df,'source_text',"translate Portuguese to Tupi:")
df['pred_volta']= run_mt5(df,'tupi_clean',"translate Tupi to Portuguese:")
```

3.3. Pipeline One-Shot (Fine-Tuning)

Nesta etapa, realizamos o treinamento supervisionado (fine-tuning) do modelo *No Language Left Behind* (NLLB) [6], da Meta AI, com o objetivo de aprimorar sua capacidade de traduzir entre tupi antigo e português brasileiro. O processo descrito a seguir baseia-se na implementação presente no script de treinamento (`treino.py`).

Após a preparação do corpus, o modelo NLLB-200 (via *HuggingFace*) e seu tokenizer foram carregados com as *language tags* necessárias. Foram utilizadas as tags linguísticas padrão do NLLB, sendo `por_Latn` para o português e `grn_Latn` como identificador substituto para o tupi antigo, uma vez que essa língua não possui código oficial no modelo. Durante o treinamento, todo o conteúdo em tupi foi associado à tag `grn_Latn`, permitindo que o modelo aprendesse a tratá-la como uma língua distinta dentro do vocabulário do NLLB.

Todas as sentenças foram tokenizadas com limite máximo de comprimento e *padding* dinâmico, preservando a correspondência entre textos de entrada e saída. O treinamento foi conduzido com o método *Seq2SeqTrainer*, configurado com os hiperparâmetros abaixo:

- número de épocas: 10
- taxa de aprendizado: 5e-5
- tamanho do batch: 4
- frequência de avaliação: cada época

Durante cada época, o modelo gerava traduções para as frases de validação, e a função de perda guiava a atualização dos pesos. Após a conclusão do *fine-tuning*, a avaliação quantitativa foi realizada em uma etapa separada (`gerar_predicoes.py`). O modelo treinado foi utilizado para gerar traduções inéditas para dois conjuntos de teste distintos:

1. **Teste Original:** Textos em Português Arcaico *(traducoes_para_avaliacao_antigo.csv)*
2. **Teste Moderno:** Textos modernizados via GPT-4o-mini *(traducoes_para_avaliacao_moderno.csv)*

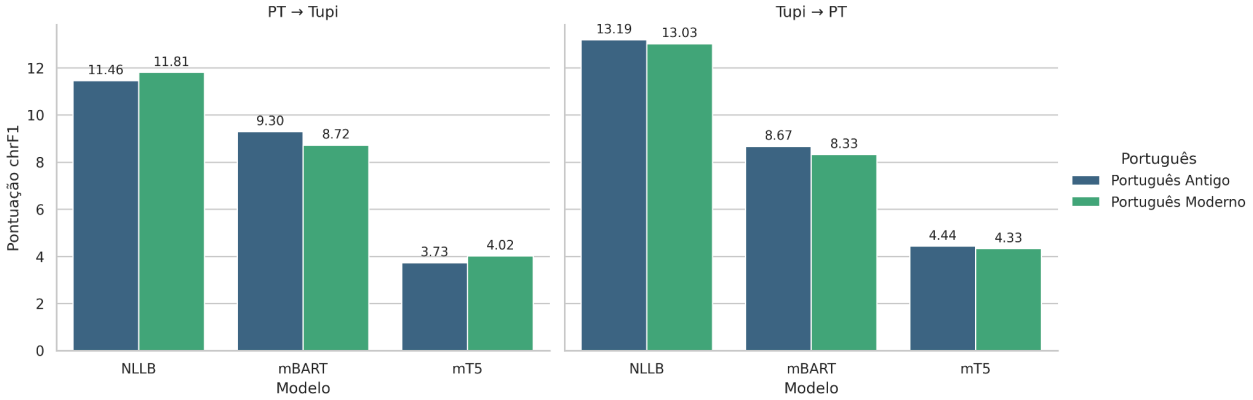
As métricas de desempenho (BLEU, chrF1 e chrF3) foram calculadas comparando essas predições com os gabaritos de referência, conforme consolidado no arquivo *RELATORIO_FINAL_NLLB_FEWSHOT.csv.*

4. Resultados

4.1 Zero-Shot

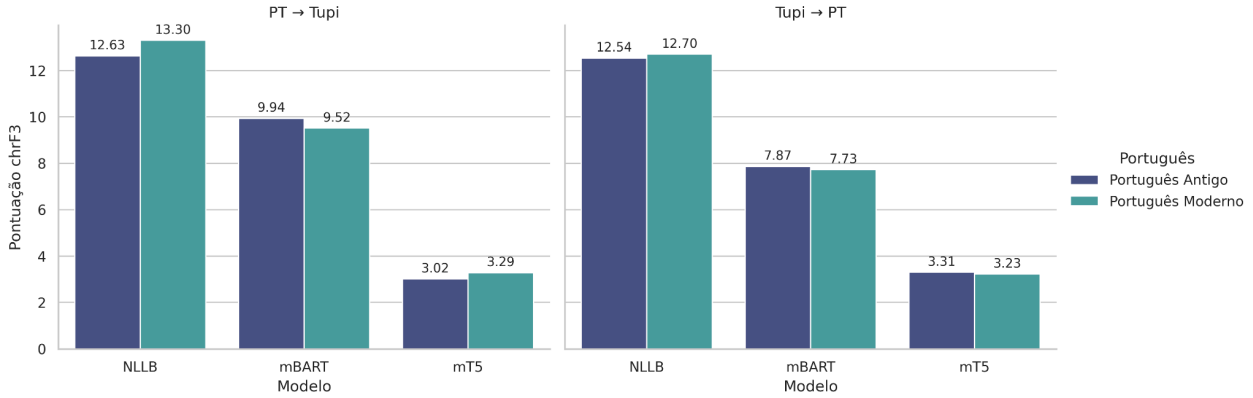
Podemos ver nos **Gráficos 3, 4 e 5,** as métricas chrF1, BLEU e chrF3 obtidas para cada combinação de modelo (NLLB, mBART e mT5), corpus (Português Arcaico e Tupi Antigo ou Português Moderno e Tupi Antigo) e direção de tradução (Tupi → Português ou Português → Tupi).

Gráfico 3 - *Performance dos modelos em Zero-Shot (Métrica ChrF1)*



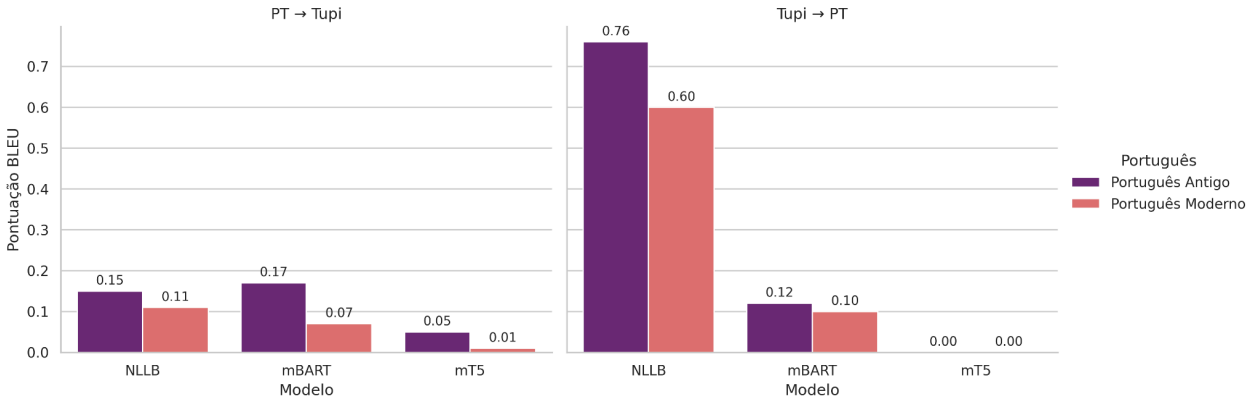
Fonte: *Elaboração própria*

Gráfico 4 - *Performance dos modelos em Zero-Shot (Métrica ChrF3)*



Fonte: *Elaboração própria*

Gráfico 5 - *Performance dos modelos em Zero-Shot (Métrica BLEU)*



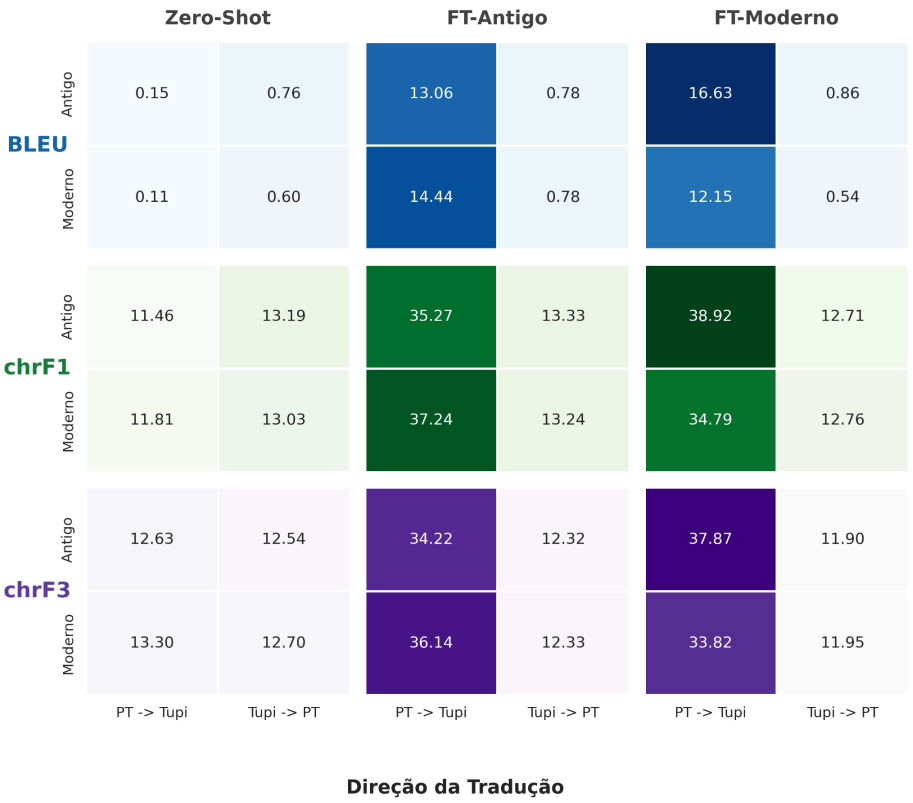
Fonte: *Elaboração própria*

Pode-se notar que a diferença de performance entre os modelos é significativa e consistente, sendo que o modelo NLLB demonstra desempenho superior em quase todas as métricas, independente da escolha de corpus. Esse resultado é um dos critérios que levou o modelo a ser selecionado para a etapa do few-shot e será melhor discutido na seção de discussões.

4.2 Few-Shot

Como esperado, os resultados do modelo melhoraram após processo de fine-tuning, tanto com o corpus de português antigo (o original), quanto com a versão modernizada, como pode ser visto no **Gráfico 6:**

Gráfico 6 - *Comparação da versão Zero-Shot e das versões Few-Shot do modelo NLLB*



Legenda: *Os gráficos apresentam os valores das métricas BLEU (azul), chrF1 (verde) e chrF3 (roxo) para as duas direções de tradução (PT→Tupi e Tupi→PT) e para as variantes de Tupi Antigo e Tupi Moderno. Em todas as métricas, cores mais escuras indicam melhor desempenho, enquanto cores mais claras indicam resultados mais baixos, permitindo visualizar os ganhos obtidos em cada configuração de treinamento.*

Fonte: *elaboração própria*

Para além das métricas, também achamos interessante trazer alguns exemplos de sentenças traduzidas pelo modelo para uma análise não apenas quantitativa mas também qualitativa dos modelos. Os exemplos de tradução na direção Tupi → PT podem ser vistos na **Tabela 4,** enquanto os exemplos PT → Tupi podem ser vistos na **Tabela 5.**

Tabela 3 - Exemplos de tradução do modelo FT-Antigo na direção Tupi → PT

Fenômeno	Entrada (Tupi)	Referência (PT Antigo)	NLLB (Treino Antigo)	NLLB (Treino Moderno)
Empréstimo / Inversão	*ladainhas ra'angaba*	*o proferir das ladainhas*	*ra'angaba ladainhas*	*ra'angaba ladainhas*
Viés de Gênero	*tupã raîyra*	*filha de deus*	*filho de deus*	*filho de deus*
Falha em Sentença Longa (Loop)	*oîaby bépe abá aîpó tupã nhe'enga...*	*o homem também transgride aquela palavra de deus...*	*será que será que será que será que...*	*será que será que será que será que...*
Interrogativa	*ma'epe ereîpotar?*	*o que você quer?*	*o que queres?*	*o que você quer?*

Fonte: Elaboração própria

Tabela 4 - Exemplos de tradução do modelo FT-Antigo na direção Tupi → PT

Fenômeno	Entrada (PT Antigo)	Referência (Tupi)	NLLB (Treino Antigo)	NLLB (Treino Moderno)
Genitivo/Posse	*quinhão de Deus dízimo*	*tupã potaba*	*tupã-apyra-pûera*	*tupã-apyra-tyba*
Empréstimo Lexical	*o proferir das ladainhas*	*ladainhas ra'angaba*	*ladainhas nhe'enga*	*ladainhas nhe'enga*
Morfologia Verbal	*amarram-lhe as mãos fortemente, fazendo-lhe mal*	*îapopûaratã, i moangaîpapa*	*oîoapyk-atã sesé i moxy-moxy*	*i pó-pyrangeté, i moxymo*
Construção Complexa	*e por outra parte, semelhantemente, o próprio mar será mais terrível do que é seu costume, fazendo as ondas grande estrondo*	*kokoty paranã aé rame'ĩ o abaetéramo erimba'e gûekoagûera sosé yapenunga ryapugûasuramo*	*kokoty paranã anhẽ bé i poxy-eté o abaeté, yby-pyrungusu*	*kokoty paranã bé i abaeté sekóû, ybaté-katu-rondyka*

5. Discussão

5.1. Dos Modelos

Os resultados dos experimentos demonstraram que o modelo NLLB superou consistentemente o mBART e o mT5 no cenário Zero-Shot. Essa superioridade do modelo pode ter se dado por diversos motivos, como o fato do NLLB ser o mais novo dos 3 modelos testados (tendo sido lançado em 2022, enquanto as versões multilíngues do BART e do T5 são de 2020) e ter sido construído justamente com o objetivo de dar assistência a línguas de baixíssimo recurso (NLLB TEAM, 2022). Isso também pode decorrer do viés dos dados de pré-treinamento de línguas de baixo recurso, frequentemente compostos por textos religiosos (como a Bíblia e tradução de missionários), que mantêm um registro formal e arcaico. Assim, a estrutura sintática do corpus original de Anchieta pode ter ativado representações latentes mais próximas do "Guarani bíblico" aprendido pelo modelo do que a linguagem coloquial moderna.

Sua arquitetura conta com um transformer do tipo MoE (Mixture of Experts) estabilizado, o que permite que ele se especialize nas línguas raras sem aumento expressivo de custo computacional (SANSEVIERO et al., 2023). Além disso, o NLLB é um modelo treinado desde o início como Encoder-Decoder many-to-many, ou seja, que atua em todas as direções de tradução (NLLB TEAM, 2022) [6]. Por fim, apesar de nenhum dos modelos ter pré-treino em Tupi, o NLLB tem para o Guarani, língua próxima ao Tupi e que, portanto, mantém algumas similaridades morfológicas com ela.

Baseado nesse desempenho e nessas características estruturais, o NLLB foi o modelo escolhido para ir para a etapa de *fine-tuning*, na qual, pudemos observarna qual pudemos observar um salto expressivo nas métricas de desempenho. Conforme ilustrado no Gráfico 6, o fine-tuning elevou o BLEU e o chrF3 em ambas as direções, demonstrando que o modelo foi capaz de adaptar seus pesos para a morfologia específica do Tupi, superando as limitações do conhecimento zero-shot baseado apenas no Guarani.

5.2. Dos Corpora

A etapa de modernização do *corpus* original, realizada via modelo GPT-4o-mini, resultou em uma diminuição perceptível na contagem total de palavras do conjunto de dados. A hipótese é que essa redução decorre da adaptação sintática de estruturas redundantes ou enfáticas típicas do português quinhentista para formas contemporâneas mais concisas. Por exemplo, construções analíticas complexas ou com mesóclises e pronomes redundantes (como *"a ti é que matam?"*) tendem a ser simplificadas na modernização (para *"é você que matam?"* ou *"matam você?"*), reduzindo a extensão da sentença sem perda de conteúdo semântico.

Ao comparar os resultados das métricas (BLEU e chrF) no regime *Zero-Shot*, observou-se que a variação de desempenho entre o uso do *corpus* Original (Arcaico) e o Modernizado foi, em geral, inferior a 1%. Esta diferença marginal sugere que se trata apenas de uma flutuação estatística, e não de um ganho real de desempenho. Isso indica que, na ausência de treinamento específico (*fine-tuning*), o conhecimento

prévio dos modelos (mBART/NLLB) não é robusto o suficiente para capitalizar sobre a fluidez do português moderno. O "ruído" causado pelo desconhecimento total da língua Tupi domina a taxa de erro, tornando irrelevante, nesta etapa, a escolha do estilo do *corpus* de entrada.

Uma hipótese crítica a ser considerada sobre a metodologia de modernização é o risco de propagação de erro em cascata. A adição de uma etapa extra de processamento (Tradução Arcaico → GPT-4o → Moderno → Modelo de Tradução) introduz uma camada adicional onde desvios semânticos podem ocorrer. Embora a modernização vise reduzir a perplexidade do modelo, ela pode inadvertidamente alterar nuances de significado do texto jesuítico original. O fato de o *Zero-Shot* ter apresentado resultados idênticos (< 1% de variação) pode indicar um ponto de equilíbrio: o ganho de fluidez trazido pela modernização foi anulado pelo ruído ou imprecisões introduzidos por essa etapa extra de tradução.

6. Conclusão

Através do experimento realizado, foi possível perceber a maior eficiência do modelo NLLB-200 no regime Zero-Shot, possivelmente pela presença do Guarani, língua proximal do Tupi, no seu pré-treinamento. e também por ser um modelo arquitetado para tradução de baixo recurso. Também foi possível perceber a melhora deste modelo após a realização do fine tuning, utilizando ambos os datasets contendo o português arcaico quanto o moderno.

Uma limitação metodológica deste trabalho foi o uso de *proxies* linguísticos (grn_Latn e pt_XX) devido à ausência do Tupi Antigo no vocabulário dos modelos. Um passo futuro crucial seria a expansão léxica do tokenizador, redimensionando a matriz de embeddings para incluir um novo token de língua dedicado (ex: tpi_Latn). Isso permitiria realizar um fine-tuning focado em alinhar representações vetoriais exclusivas para o Tupi, criando efetivamente um "canal" neural exclusivo para a língua indígena.

Uma outra limitação diz respeito às métricas utilizadas (BLEU e chrF), que baseiam-se na sobreposição superficial de caracteres e n-gramas. Para línguas de baixo recurso com alta variabilidade morfológica, essas métricas podem penalizar traduções que são semanticamente corretas, mas lexicalmente distintas da referência. Trabalhos futuros devem incorporar métricas baseadas em embeddings contextuais, avaliando a similaridade no espaço vetorial latente, permitindo mensurar se o modelo capturou o significado da sentença em Tupi, mesmo que a estrutura frasal difira do gabarito.

7. Referências

1. **REZENDE, Calebe Macena.** *Tupi Antigo: desenvolvimento de ferramentas computacionais para tradução e preservação cultural: tradução direta e via pivotação tradução por língua proximal.* 2025. Dissertação (Mestrado em Ciências da Computação) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2025.
2. **NAVARRO, Eduardo de Almeida.** *Dicionário de tupi antigo: a língua indígena clássica do Brasil.* São Paulo: Global, 2013.

3. **GUZMÁN, Francisco et al.** *The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English.* 2019. arXiv:1902.01382. Disponível em: https://arxiv.org/abs/1902.01382. Acesso em: 1 dez. 2025.
4. **RANZATO, Marc'Aurelio.** *Stanford CS224N: NLP with Deep Learning | Winter 2020 | Low Resource Machine Translation*. 2021. Vídeo (1h15). Canal: Stanford Online. Disponível em: https://www.youtube.com/watch?v=mp95Z5yM92c. Acesso em: 1 dez. 2025.
5. **NIGATU, Hellina Hailu; TONJA, Atemafu Lambebo; ROSMAN, Benjamin; SOLORIO, Thamar; CHOUDHURY, Monojit.** *The Zeno's paradox of 'low-resource' languages.* In: **CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING**, 2024, Miami, Florida. Proceedings… Miami: Association for Computational Linguistics, 2024. p. 17753–17774. Disponível em: https://aclanthology.org/2024.emnlp-main.983/. Acesso em: 5 dez. 2025.
6. **NLLB TEAM.** *No Language Left Behind: scaling human-centered machine translation.* 2022. arXiv:2207.04672. Disponível em: https://arxiv.org/abs/2207.04672. Acesso em: 1 dez. 2025.
7. **SANSEVIERO, Omar; TUNSTALL, Lewis; SCHMID, Philipp; MANGRULKAR, Sourab; YOUNES B.; CUENCA, Pedro**. *Mixture of Experts explained.* 11 dez. 2023. Disponível em: https://huggingface.co/blog/moe. Acesso em: 1 dez. 2025.