

Treinamento e análise de um modelo de tradução automática baseado em *Transformer*

Training and analyzing a Transformer-based machine translation model

Clóvis Henrique Martins Pimentel ^{*1} e Thiago Blanch Pires ^{†1}

¹ Universidade de Brasília, Brasília, DF, Brasil.

Resumo

O presente trabalho possui como objetivo a análise dos modelos de tradução automática baseados em *Transformer*. Em específico, a pesquisa visa o teste da viabilidade do uso de modelos treinados a partir de *corpus* especializado. Para o treinamento do modelo, foi construído um *corpus* paralelo inglês-francês a partir de sete textos da Convenção de 25 de outubro de 1980 sobre os Aspectos Cíveis do Rapto Internacional de Crianças. Os resultados de tradução obtidos pelo modelo treinado foram comparados com aqueles produzidos pelo Google Tradutor. Para a etapa de avaliação foram utilizados os métodos de avaliação automática sacreBLEU e avaliação humana. Os resultados da avaliação automática de frases produzidas pelo modelo treinado foram, em média, mais positivos que aqueles gerados pelo modelo não treinado. A avaliação humana das frases revelou que houve erros de adequação no uso da linguagem específica à matéria da Convenção da Haia de 1980 tanto em frases geradas pelo modelo treinado, quanto em frases geradas pelo modelo do Google Tradutor.

Palavras-chave: Linguística Computacional. Tradutor automático. *Transformer*. *Corpus* paralelo. Avaliação de tradução automática.

Abstract

The objective of this work is to analyze transformer-based machine translation models. It aims to test the feasibility of using trained models based on specialized *corpus*. For the training of such model, a parallel English-French *corpus* was built with seven texts related to the Convention of 25 October 1980 on the Civil Aspects of International Child Abduction. The translation results obtained by the trained model were compared with those produced by Google Translate. For the evaluation stage, sacreBLEU automatic evaluation and human evaluation methods were used. The outcome of the automatic evaluation of sentences produced by the trained model was, on average, higher than those generated by the non-trained model. The human evaluation of the sentences revealed that there were adequacy errors in the use of language specific to the subject matter of the 1980 Hague Convention both in sentences generated by the trained model and in sentences generated by the Google Translate model.

Keywords: Computational Linguistics. Machine translation. Transformer. Parallel *corpus*. Machine translation evaluation.


Linguagem e Tecnologia

DOI: 10.1590/1983-
3652.2024.49118

Seção:
Artigos

Autor Correspondente:
Thiago Blanch Pires

Editor de seção:
Daniervelin Pereira
Editor de layout:
João Mesquita

Recebido em:
8 de dezembro de 2023
Aceito em:
23 de fevereiro de 2024
Publicado em:
8 de abril de 2024

Esta obra tem a licença
"CC BY 4.0".



1 Introdução

Em meados de 2017, com a publicação do artigo *Attention is All You Need*, pela equipe da Google (Vaswani *et al.*, 2017), iniciou-se a mudança de paradigma das técnicas utilizadas no processamento de dados sequenciais, especialmente no campo do Processamento de Linguagem Natural (PLN). Uma nova arquitetura de rede neural denominada *Transformer* possibilitou o aprimoramento significativo da qualidade de traduções automáticas multilíngues, antes amplamente baseada em modelos de processamento de redes neurais recorrentes (RNNs).

Até então, os tradutores automáticos baseados em RNNs eram dominantes, mas ainda apresentavam limitações na captura de dependências de longo prazo, resultando na perda de informações contextuais importantes (Iosifova *et al.*, 2020). Com a chegada do *Transformer*, a tradução automática passou a apresentar resultados significativamente melhores, possibilitando a criação de modelos com maior capacidade de lidar com contextos complexos e produzir traduções mais precisas (*Ibid.*).

*Email: clovismpimentel@gmail.com

†Email: thiagocomaga@gmail.com

Atualmente, é possível encontrar de forma acessível tradutores automáticos que produzam traduções multilíngues de alta qualidade, ainda que com certas restrições em relação a sua confiabilidade (Peña Aguilar, 2023) ou fluência (Banitz, 2020). Por isso, áreas como os estudos de avaliação de tradução automática e da pós-edição de tradução automática são necessários para que se levante os principais problemas e possíveis soluções para o uso desses sistemas.

Embora seja notável a contribuição da pós-edição de tradução automática por meio de trabalhos como o de Koponen (2016) e de O'Hagan (2019), aplicada a áreas diversas como a tradução audiovisual para legendagem e a tradução literária (Costa; Silva, 2020; Koglin *et al.*, 2023), os estudos de avaliação de tradução automática ainda são necessários como etapa de levantamento e análise de dados obtidos automaticamente ou manualmente de resultados de sistema de tradução automática. São necessários, pois, além de revelarem um painel sobre a qualidade dos resultados, subsidiam reflexões críticas e profundas sobre determinados modelos e conjuntos de dados utilizados como treinamento de tais sistemas.

Ainda que os resultados de tradução automática de determinados sistemas, tais como Google Tradutor e DeepL sejam confiáveis para uma ampla gama de gêneros textuais e em vários pares linguísticos, eles podem vir a apresentar resultados menos satisfatórios quando comparados a modelos de tradução automática treinados para um propósito mais específico. O intuito deste trabalho é justamente verificar se há divergência de resultados entre esses dois modelos de tradução automática.

Diante dessa situação, surge a indagação sobre como seria possível desenvolver um tradutor multilíngue, baseado em uma arquitetura *Transformer*, capaz de lidar com o léxico especializado. Assim, o presente trabalho busca preencher a lacuna de desenvolvimento de um modelo de tradução especializado no léxico utilizado em torno da Convenção de 25 de outubro de 1980 sobre os Aspectos Cíveis do Rapto Internacional de Crianças e Adolescentes (HCCH, 1980).

Com efeito, busca-se testar a viabilidade do uso de um modelo de tradução automática baseado em *Transformer*, levando em consideração a sua facilidade de implementação e a qualidade dos resultados gerados. Para a análise da qualidade, compara-se a precisão e adequação lexical dos resultados do modelo treinado em relação a resultados de traduções automáticas feitas por outras plataformas, como o Google Tradutor. Para tanto, a presente investigação compila um *corpus* temático a partir do documento da Convenção de 25 de outubro de 1980 sobre os Aspectos Cíveis do Rapto Internacional de Crianças (HCCH, 1980), que serve de base para o treinamento de um tradutor automático especializado.

Os textos que compõem o *corpus* aqui desenvolvido têm caráter eminentemente jurídico, mais especificamente do Direito Internacional Público (DIP). Gravitam em torno dessa convenção, uma vez que é ela o texto jurídico base para a organização de um sistema internacional de cooperação entre Estados, medidas e ferramentas a serem utilizadas para a correta execução de ações coordenadas internacionalmente.

Em relação à metodologia (detalhada na seção 3 deste artigo), procura-se aplicar o modelo de tradução automática T5¹, introduzido pelo Google Research em 2019 (Raffel *et al.*, 2020) e disponibilizado gratuitamente pela plataforma *HuggingFace*, especializada em processamento de linguagem natural (PLN) e aprendizado de máquina. As versões que compõem o *corpus* constituído pela "Convenção sobre os Aspectos Cíveis do Sequestro Internacional de Crianças", estão em inglês e francês – as duas línguas oficiais da Conferência da Haia de Direito Internacional Privado. Todo o conteúdo do *corpus* foi retirado de textos e traduções oficiais disponibilizados no site dessa Conferência (Pérez-Vera, 1980; HCCH, 1980, 2003a,b, 2005, 2010, 2012).

Com o *corpus* construído, aplicou-se o modelo de tradução automática preexistente, treinando-o para a obtenção de traduções que utilizassem de forma semanticamente mais precisa expressões com maior recorrência no âmbito do léxico da convenção internacional aqui mencionada (Mais detalhes na seção 3). Ao final, as traduções obtidas foram analisadas e comparadas aos resultados gerados pelo Google Tradutor.

A seguir, a seção 2 discute alguns conceitos técnicos e linguísticos mais relevantes e faz uma

¹ T5 é um modelo codificador-decodificador pré-treinado em uma mistura de tarefas não supervisionadas e supervisionadas, em que cada tarefa é convertida para um formato de *text-to-text*.

revisão de literatura, baseada em estudos que abarcam o tema aqui tratado. A seção 3 descreve com detalhes a metodologia utilizada para a produção dos resultados, que, por sua vez, são analisados na seção 4. Por fim, as considerações finais explicitam a qualidade dos resultados obtidos e ressaltam sua relevância para a área.

2 Revisão de literatura

Baseado no estado-da-arte da tecnologia de tradução automática, aqui ilustrado pelas pesquisas de Lakew, Cettolo e Federico (2018), Banitz (2020), Tian *et al.* (2022), Iosifova *et al.* (2020) e Kimera, Rim e Choi (2022), opta-se pela utilização de um modelo pré-treinado baseado na arquitetura *Transformer* para o processamento de seus dados. A escolha se justifica graças à comprovação da eficácia desse modelo em detrimento do uso de RNNs (Vaswani *et al.*, 2017; Wolf *et al.*, 2020; Raffel *et al.*, 2020). Tal característica é confirmada por Lakew, Cettolo e Federico (2018), que também atesta a superioridade que essa arquitetura possui quando desempenha tarefas de processamento de dados em modelos bilíngues, objeto aqui estudado.

Outro trabalho relevante é o modelo de tradução automática francês-inglês, baseado em uma arquitetura *Transformer*, desenvolvido por Tian *et al.* (2022). O artigo (Tian *et al.*, 2022) trata da superioridade do modelo na performance de tarefas de tradução automática e afirma que modelos baseados em RNNs possuem dois grandes defeitos. O primeiro deles seria a sua limitação no processamento, restritos ao processamento individual de palavras e ocasionando um atraso em relação a sua velocidade de treinamento. O segundo ponto negativo mencionado em relação ao modelo que utiliza RNNs é a incapacidade dessa arquitetura de processar dados de forma precisa quando as frases se tornam demasiadamente longas.

Utilizando então um modelo baseado em *Transformer*, Tian *et al.* (2022) treinaram um modelo de tradução automática francês-inglês. Os resultados do treinamento puderam demonstrar uma acurácia de 80% em suas traduções e informado como maior que aquele produzido por um modelo baseado em RNNs. O trabalho (*Ibid.*) no entanto, se limita a explicitar somente os dados obtidos a partir do processo de treinamento do modelo (perda de validação e de treinamento). O artigo (*Ibid.*) não demonstra os valores de acurácia do modelo RNNs e tampouco procede à avaliação automática ou humana de nenhuma tradução produzida pelo modelo treinado baseado em *Transformer*.

O estudo de Iosifova *et al.* (2020), alinhado a Koponen (2016) e Costa e Silva (2020), parte do pressuposto de que um modelo de tradução automática produz traduções mais adequadas semanticamente quando submetido a um posterior processo de treinamento com dados que possibilitem a melhora do desempenho de uma tarefa específica. Diferentemente daquele (Iosifova *et al.*, 2020), busca-se aqui a utilização do *corpus* paralelo inglês-francês para a melhora na precisão do vocabulário utilizado em traduções sobre o tema da Convenção de Haia de 1980 (HCCH, 1980).

Uma outra pesquisa que se aproxima mais do presente trabalho foi desenvolvida por Kimera, Rim e Choi (2022). Em seu artigo, os autores (*Ibid.*) descrevem a construção de um *corpus* paralelo contendo 41.070 pares de frases em inglês e luganda, posteriormente usado para o treinamento de um modelo baseado em *Transformer*. Após o treinamento de vários modelos testando hiperparâmetros diferentes, em um deles foi possível obter o score BLEU final de 17,47 para traduções de inglês para luganda.

Apesar de ainda contar com um número substancialmente maior de frases alinhadas e capacidade de processamento de dados igualmente superior, tal trabalho (Kimera; Rim; Choi, 2022) lança luz sobre a realidade dos materiais necessários para se produzir um modelo de tradução automática minimamente capaz de melhorar um modelo pré-treinado não especializado. Esse trabalho (*Ibid.*) auxilia na compreensão da relação entre o modo como o modelo é treinado e a aferição dos resultados produzidos pelo modelo. Durante o artigo são demonstrados alguns dados que comprovam a relação direta entre os hiperparâmetros escolhidos pelos autores (como o tamanho do *batch*) e a influência que eles possuem na aferição do score BLEU do modelo. Alterando alguns hiperparâmetros é possível obter scores menores em relação ao score final. O estudo (*Ibid.*) demonstra, então, a definição de parâmetros não ótimos que leva ao score BLEU 13,96.

Semelhante ao que se propõe o presente estudo, o artigo produzido por Kimera, Rim e Choi

(2022), enseja não somente apresentar o *score* do modelo após seu treinamento, mas explora também o resultado de quatro frases traduzidas do inglês para luganda produzidas pelo seu modelo de tradução. Apesar disso, diferentemente do que se propõe neste artigo, o texto de Kimera, Rim e Choi (2022) não apresenta nenhuma avaliação humana consistente de suas traduções e nem pôde comparar, apesar de ter explicitado que esse era um dos objetivos futuros do trabalho, as traduções com resultados gerados pelo Google Tradutor. Tal fato se deve por ainda não existir, até a data da submissão deste artigo, a língua luganda disponível na plataforma *online*.

Cabe, por fim, mencionar o trabalho de Banitz (2020), que realiza um estudo exímio na avaliação de traduções realizadas por dois modelos diferentes, um deles sendo o Google Tradutor e o outro o Systran. Em sua pesquisa, a autora utiliza tanto métodos de avaliação automática (*Translation Error Rate/TER score*)² como métodos de avaliação humana para avaliar as primeiras 24 frases traduzidas a partir de seu *corpus*. Em relação à avaliação automática, a autora⁷esquematisa os *scores* TER obtidos dos resultados de tradução dos dois mecanismos automáticos em uma tabela e os compara explicitando que as traduções produzidas pelo Google Tradutor requerem, ao final, menos pós-edição, uma vez que apresentam uma taxa de erro menor que o outro modelo avaliado. Já em relação à avaliação humana dos resultados, o trabalho de Banitz (2020) utiliza métricas bem delineadas de avaliação de fluência e adequação das frases geradas (assunto explorado na subseção 3.2). Também em uma tabela comparativa dos resultados de tradução gerados pelos dois mecanismos de tradução automática, a autora compara os *scores* atribuídos às frases e aponta os resultados do Google Tradutor como mais satisfatórios. Todavia, a mesma (*Ibid.*) ainda explicita os desafios linguísticos que a tradução automática encontra em relação a erros semânticos, lexicais, sintáticos e morfológicos.

Apesar de não abordar o uso de modelo *Transformer* especificamente treinados para a melhora do resultado de traduções, o trabalho de Banitz (2020) abarca importantes discussões sobre os métodos de avaliação desenvolvidos e utilizados para a avaliação e comparação de traduções automaticamente geradas. É com base nestes aspectos que o trabalho da autora (*Ibid.*) contribui para este artigo, fornecendo grande suporte científico para o método de comparação dos resultados aqui elaborado.

Para a melhor compressão dos parâmetros apresentados por Banitz (2020), é necessário entender que o erro de tradução, objeto de grande problematização dentro dos estudos de tradução, passa a ser utilizado no contexto da computação, com seu sentido matemático de “cômputo”, dificilmente sendo objeto de questionamentos (Pires, 2017). Assim, os erros de tradução são aqui entendidos como “configurações de incompatibilidades linguísticas (lexical, semântica e sintática) entre o texto de entrada e o texto de saída gerado por uma tradução automática em um dado contexto de produção” (*Ibid.*). White (2003, p. 242), ao abordar o problema da falta de verdade absoluta na tradução, ressalta que o “método de comparação entre o resultado da tradução gerado automaticamente e traduções consideradas ‘corretas’, ainda que humano, é abstrato”.

Após a análise de alguns trabalhos relevantes na área, o presente estudo aborda na próxima seção a metodologia desenvolvida neste artigo para treinar um modelo de tradução automática baseado em *Transformer*, levando em consideração o vocabulário específico da Convenção da Haia de 1980 (HCCH, 1980). A próxima seção delineia técnicas específicas para o alcance de maior precisão, consistência e adaptação do léxico ao campo do conhecimento em questão.

3 Metodologia

Primeiramente, para a confecção do *corpus* especializado no tema da Convenção sobre os Aspectos Civis do Rapto Internacional de Crianças (HCCH, 1980), optou-se pela seleção de textos e traduções, obtidos online de forma gratuita, que fossem reconhecidos como oficiais pela própria Conferência da Haia de Direito Internacional Privado. A instituição é uma organização intergovernamental da área do Direito Internacional Privado (DIP) que administra diversas convenções internacionais, protocolos e instrumentos de *soft law* (regras de valor normativo limitado e que não possuem caráter jurídico obrigatório), com o objetivo de unificar progressivamente as normas dessa área (Rodas; Mônaco,

² Apesar de justificar a escolha desse método de avaliação automática como um método mais intuitivo do “quão bom” é a tradução, a autora explicita as limitações impostas por essa métrica de avaliação como: (i) o método não reflete necessariamente a adequação da tradução gerada e (ii) a métrica depende diretamente da qualidade da tradução-referência, uma vez que qualquer desvio da tradução humana será penalizada (Banitz, 2020).

2007).

Apesar de suas convenções internacionais não possuírem mandatoriamente o valor de lei para os países que assim não as convalidam, as Convenções e instrumentos fornecem clareza e direção em relações transfronteiriças com diversas matérias de DIP, dentre elas o Direito Internacional de Família e Proteção à Criança e Adolescente. É tarefa dessa Conferência uniformizar as normas aplicadas, zelar pelo seu adequado cumprimento e difundi-las entre os países. Com isso, a seleção de textos exclusivamente oficiais, produzidos pela própria Conferência da Haia, atribui ao *corpus* uma maior segurança e uniformidade em seu léxico³.

Em relação à escolha das duas línguas selecionadas para integrar o *corpus* paralelo, a opção pelas línguas inglês e francês se deu pela grande disponibilidade de material produzido nesses idiomas que, até o presente momento, são as duas únicas línguas oficiais dessa organização intergovernamental⁴. Assim, a maioria dos textos possui sua primeira redação em francês, com sua subsequente tradução oficial para o inglês e posterior disponibilização em outros idiomas⁵. Após a seleção dos textos, foram necessárias a limpeza e formatação dos textos e alinhamento semântico dos pares de frases, feitas de forma manual utilizando planilha do Microsoft Excel, conforme Tabela 1.

Tabela 1. Frases em inglês e francês alinhadas de acordo com sua correspondência semântica.

Texto em Inglês	Texto em Francês
Convention on the Civil Aspects of International Child Abduction Concluded 25 October 1980 The States signatory to the present Convention, Firmly convinced that the interests of children are of paramount importance in matters relating to their custody, Desiring to protect children internationally from the harmful effects of their wrongful removal or retention and to establish procedures to ensure their prompt return to the State of their habitual residence, as well as to secure protection for rights of access, Have resolved to conclude a Convention to this effect and have agreed upon the following provisions.	Convention Sur Les Aspects Civils De L'enlèvement International D'enfants Conclue le 25 octobre 1980 Les Etats signataires de la présente Convention, Profondément convaincus que l'intérêt de l'enfant est d'une importance primordiale pour toute question relative à sa garde, Désirant protéger l'enfant, sur le plan international, contre les effets nuisibles d'un déplacement ou d'un non-retour illégitimes et établir des procédures en vue de garantir le retour immédiat de l'enfant dans l'Etat de sa résidence habituelle, ainsi que d'assurer la protection du droit de visite, Ont résolu de conclure une Convention à cet effet, et sont convenus des dispositions suivantes.

Fonte: Elaboração própria.

Foi necessário realizar a limpeza e o alinhamento de quase 5.500 frases pareadas em inglês e francês. Levando em conta todos os textos utilizados, é possível ainda documentar, aproximadamente, 278.000 *tokens* (total de ocorrências) e 15.500 *types* (vocabúlos distintos) presentes entre os textos em inglês e francês.

O próximo passo foi decidir a plataforma que seria utilizada para o processamento desses dados e notou-se que a *HuggingFace* seria a melhor opção para isso. Como justificativa, sublinhamos se tratar de uma ferramenta oferecida de forma gratuita e que disponibiliza uma ampla gama de

³ Cf. <https://www.hcch.net/pt/about>. Acesso em: 17 jul. 2023.

⁴ Cabe destacar que apesar de o site já possuir versões completas em outras línguas como o alemão, português e espanhol, somente essa última será introduzida em 1 de julho de 2024, ao rol de línguas oficiais da Conferência. <https://www.hcch.net/pt/news-archive/details/?varevent=907>. Acesso em: 13 jul. 2023.

⁵ Para a construção do *corpus* foram utilizados, ao todo, 7 pares de textos, cada um com sua versão original em francês e versão em inglês: (i) texto oficial da Convenção sobre os Aspectos Cíveis do Rapto Internacional de Crianças (HCCH, 1980); (ii) Relatório Explicativo de Eliza Pérez-Vera (Pérez-Vera, 1980); e (iii) Guias de Boas Práticas sobre a Convenção da Haia de Subtração Internacional de Crianças e Adolescentes (Partes I a V) (HCCH, 2003a,b, 2005, 2010, 2012). Textos disponíveis em <https://www.hcch.net/pt/publications-and-studies/publications2>. Acesso em: 12 jul. 2023.

modelos pré-treinados e bibliotecas (inclusive a que disponibiliza o modelo *Transformer*) que facilitam o desenvolvimento, treinamento e implantação de modelos de tradução automática. Com todos esses recursos disponíveis, a própria plataforma ensina, por meio de um tutorial utilizando a linguagem de programação *Python*, a implementar a sua biblioteca *Transformer*.

Para que fosse possível utilizar o *corpus* paralelo EN-FR, foi criado um dicionário em arquivo JSON com os pares de frases indexados, conforme exemplo a seguir:

Lista 1. Amostra de dicionário em formato JSON relativo a *corpus* paralelo EN-FR

```
[{"id": "0", "translation": {"en": "Convention on the Civil Aspects of  
International Child Abduction", "fr": "Convention Sur Les Aspects Civils  
De L'enlèvement International D'enfants"}}]
```

Em relação à escolha do modelo de *Transformer* utilizado a partir da biblioteca gratuita disponibilizada pela *HuggingFace*, optou-se por um modelo pré-treinado e que não demandasse uma grande capacidade de processamento computacional para o seu treinamento a partir do *corpus* especializado. A escolha teve como base o modelo “SEBIS/legal_t5_small_multitask_en_fr” de tradução inglês-francês para textos de conteúdo jurídico, disponível gratuitamente na plataforma *HuggingFace*. Assim, foi selecionado o *T5-small*, um modelo de rede neural pré-treinado para tarefas de *text-to-text*, adequado para tarefas de tradução automática aplicadas a máquinas com menor capacidade de processamento (Raffel *et al.*, 2020)⁶.

No entanto, alguns ajustes foram feitos, já que o modelo SEBIS foi treinado a partir de um *corpus* paralelo de 9 milhões de pares de frases, 220 milhões de parâmetros, *batch size* de tamanho 4096, *sequence length* de 512, e *corpus* pré-processado a partir de 88 milhões de frases, com *score* sacreBLEU de 38,063. Os números indicam a necessidade de uma máquina com grande capacidade de processamento, superior ao que se é possível realizar em um computador de uso pessoal. Em decorrência disso, para o presente trabalho, foram redefinidos os valores do *epoch* (número de vezes que todo o conjunto de treinamento é percorrido durante o treinamento) para cinco e dos *batches* (quantidade de blocos de pares de frases analisados pelo modelo a cada iteração) para 16, sendo possível obter, após 3 horas 46 minutos e 36 segundos, um modelo de tradução treinando especializado no vocabulário referente à Convenção sobre os Aspectos Cíveis do Rapto Internacional de Crianças (HCCH, 1980).

Partindo-se então do código-exemplo⁷, que produz seus resultados a partir de uma arquitetura híbrida de RNNs e *Transformer*, o *Google Neural Machine Translation* (GNMT) (Wu *et al.*, 2016) de implementação da biblioteca *Transformer* presente na plataforma *HuggingFace* e utilizando o arquivo JSON previamente mencionado, procedeu-se às adaptações do código para o processamento local do *corpus* paralelo utilizado para o treinamento de um novo modelo especializado de tradução. Aqui, 20% do *corpus* (1.099 frases) foi utilizado para teste e o restante (4.395 frases) para o treinamento do modelo. Após o treinamento do modelo, os resultados obtidos foram comparados com traduções realizadas pelo Google Tradutor⁸.

Buscou-se, assim, avaliar a eficácia do modelo de tradução treinado a partir de um *corpus* especializado em relação aos resultados produzidos pelo tradutor online. Com essa metodologia, foi possível aferir a capacidade do modelo criado a partir de um *corpus* especializado em produzir traduções linguisticamente precisas e contrastar seus resultados com aqueles produzidos pelo GNMT.

3.1 Avaliação Automática

Para a aferição dos resultados optou-se pela utilização de dois métodos: (i) métrica de comparação automática e (ii) avaliação humana, esta última abordada na próxima subseção deste trabalho. Em relação à métrica de comparação automática, utilizou-se a ferramenta de código aberto sacreBleu,

⁶ Esse modelo foi escolhido por sua capacidade de lidar com uma ampla gama de tarefas de PLN, inclusive traduções multilíngues, e, por já se encontrar em uma fase de pré-treinamento, ser necessário somente o treinamento desse modelo em relação ao léxico especializado da Convenção da Haia de 1980 (Raffel *et al.*, 2020).

⁷ Cf. <https://HuggingFace.co/docs/transformers/tasks/translation>. Acesso em: 12 jul. 2023.

⁸ Testes realizados em 29 de junho de 2023, às 17h22min, horário local de Brasília, Brasil.

que desempenha o cálculo do *score* BLEU para avaliar a qualidade das traduções automáticas em comparação com as referências humanas (Post, 2018; Papineni *et al.*, 2002).

A metodologia adotada permite uma avaliação objetiva e sistemática da qualidade das traduções automáticas, comparando-as com referências humanas e levando em consideração a precisão dos *n-gramas* (sequência de *n* itens – por exemplo, palavras – de uma amostra de texto). Contudo, assim como o *score* BLEU, o *sacreBLEU* também possui limitações que devem ser consideradas ao interpretar os resultados aqui obtidos, como a sua falta de compreensão semântica e as limitações impostas pelas referências que ele utiliza para avaliar a tradução⁹.

Em relação à falta de compreensão semântica, o BLEU não avalia a precisão das traduções, não sendo capaz de capturar as nuances semânticas cruciais para traduções que requerem o uso de termos especializados. É possível que uma tradução tenha o mesmo significado que uma referência utilizada pela própria métrica, mas receba um *score* baixo. Da mesma forma, é possível que uma tradução tenha altas pontuações de BLEU, mas transmita um significado diferente ou incorreto em comparação com as referências (caso ilustrado pela Tabela 1, na seção 4).

No caso das limitações impostas pelas referências que a métrica BLEU utiliza para avaliar a tradução, ela depende de frases-referência definidas por humanos para a avaliação. A escolha das referências pode ser subjetiva e não capturar toda a variedade de traduções aceitáveis. Além disso, os resultados gerados nem sempre melhoram com o aumento do número de frases-referência, e estudos recentes comprovam melhor desempenho do *score* quando somente uma frase-referência foi utilizada (Freitag; Grangier; Caswell, 2020).

Ao utilizar a avaliação automática também para avaliar não só o modelo, mas frases produzidas por ele, foram realizadas avaliações de frases individuais produzidas tanto pelo Google Tradutor quanto pelo modelo treinado. Seguindo a técnica utilizada por Freitag, Grangier e Caswell (2020), para cada tradução comparada, selecionou-se a frase correspondente do texto oficial em francês como parâmetro de referência utilizado pelo *sacreBLEU* para a avaliação automática e geração de *scores* individuais. Os resultados foram posteriormente comparados, seguindo modelo de formatação utilizado por Banitz (2020), oportunamente apresentado na seção dedicada à discussão dos resultados.

Por optar, então, pela utilização de uma métrica de avaliação automática de fácil implementação e amplamente difundida, selecionou-se o *sacreBLEU* como parâmetro de avaliação automática. Necessário, entretanto, lembrar que cada métrica de avaliação encontra algum tipo de limitação que deve ser considerada quando utilizada na avaliação automática de modelos de tradução de mesma natureza. Em decorrência disso, o presente trabalho prezou também pelo desenvolvimento da avaliação humana, que se torna imprescindível para a obtenção de uma compreensão mais abrangente da qualidade da tradução gerada pelo modelo treinado, assunto tratado a seguir.

3.2 Avaliação Humana

A base da avaliação humana adotada é constituída especialmente pelos trabalhos de Banitz (2020) e Vilar *et al.* (2006). O primeiro utiliza os parâmetros de fluência e adequabilidade para aferir a qualidade das traduções, enquanto o segundo apresenta a classificação (com respectivas subclasses) de uma série de erros cometidos pela máquina ao realizar uma tradução.

Primeiramente, em relação aos termos utilizados pela autora Banitz (2020), ela explica que a fluência de uma tradução gerada também pode ser entendida como o nível de sua inteligibilidade, compreendendo tanto a correção gramatical quando a escolha de palavras utilizadas na tradução (Douglas *et al.*, 1994; Kalyani *et al.*, 2014). Por sua vez, a adequação pode ser também entendida como a acurácia ou fidelidade da tradução produzida, e se relaciona com o grau no qual a tradução consegue representar o significado original da frase traduzida (Douglas *et al.*, 1994; Kalyani *et al.*, 2014).

Para a metrificação de cada um dos dois parâmetros, Banitz (2020, p. 63) descreve a atribuição de um valor que varia de 1 a 5, cabendo a um humano avaliar a frase traduzida com base nesses

⁹ O *score* também apresenta outras limitações (insensibilidade à ordem das palavras e sintaxe, ênfase excessiva na correspondência do número de *n-gramas*, favorecimento de traduções mais curtas insensibilidade a sinônimos e paráfrases), essas não tão relevantes para a análise dos resultados desse trabalho (Post, 2018).

valores. Para a fluência, a atribuição do valor 1 corresponde a incompreensibilidade da frase e o valor 5 a uma frase perfeitamente inteligível em determinada língua. Em relação à adequação, atribuir o valor 1 a uma frase implica em dizer que o significado expressado nela não se aproxima de forma alguma do significado expressado na frase que se pretendia traduzir. Em contrapartida, o valor 5 expressa que não houve perda de nenhum significado na frase traduzida.

Em relação ao trabalho produzido por Vilar *et al.* (2006), esse apresenta uma série de categorias e subcategorias de erros passíveis de identificação quando da análise de uma tradução gerada automaticamente. Este trabalho, no entanto, opta por focar em uma subcategoria específica apresentada pelos autores: o erro aferido a partir da utilização errônea de palavras na tradução, especialmente quando levado em consideração o sentido da palavra presente no texto gerado. Isso pode ocorrer, de acordo com os autores, por uma escolha lexical errada ou por uma desambiguação incorreta. O presente estudo se atém ao fator relacionado a uma escolha lexical errada para comparar as traduções geradas pelo modelo treinado e aquelas geradas pelo Google Tradutor.

Para a execução da investigação proposta, o primeiro autor deste trabalho, bacharel em Direito e em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação (LEA-MSI), com experiência profissional em DIP e atuação perante a Autoridade Central Administrativa Federal (ACAF)¹⁰, utilizou os parâmetros anteriormente mencionados para proceder à avaliação humana da matéria aqui explorada. Partindo dessa experiência, é possível constatar que os fenômenos linguísticos gerados por sistemas de tradução automática têm relação direta na atuação da interface entre bacharel em LEA-MSI, já que se informa de fundamentos linguísticos, relações multilíngues e computacionais, desenvolvendo sua formação em tarefa especializada e combinada com o léxico especializado em Direito. Assim, conforme Pires (2020, p. 73), essa formação acadêmica pode proporcionar que tais profissionais atuem “acadêmica e profissionalmente de forma a olhar e desempenhar, com viés crítico e prático, o fenômeno linguístico que se coloca na relação complexa entre o humano e a máquina”.

Desta feita, utilizando os métodos de avaliação humana aqui expostos e as métricas de avaliação automática anteriormente discutidos, passa-se a apresentar e discutir, na próxima seção, os resultados obtidos por meio deste estudo.

4 Resultados

Para o treinamento do modelo de tradução automática, o *corpus* paralelo construído foi composto por quatorze textos, sendo sete pares de textos em inglês-francês. Os textos selecionados foram todos retirados do site oficial da Conferência da Haia sobre Direito Internacional Privado. Um desses pares de textos representa o texto da própria Convenção de 25 de outubro de 1980 sobre os Aspectos Cíveis do Rapto Internacional de Crianças e Adolescentes (HCCH, 1980). Os outros pares de textos selecionados possuem relação direta com essa Convenção, sendo um relatório explicativo sobre a própria convenção e outros cinco guias de boas práticas também relacionados à implementação dessa convenção internacional (Pérez-Vera, 1980; HCCH, 1980, 2003a,b, 2005, 2010, 2012).

Após a coleta dos textos, foram computados 134.943 *tokens* e 6.453 *types* presentes nos textos em inglês e 143.285 *tokens* e 8.932 *types* em francês. Ao final foram alinhadas 5.494 frases em inglês-francês. A partir disso, o modelo treinado recebeu o *score* sacreBLEU de 7,6467.

Levando em consideração o tamanho do *corpus* usado para treino neste modelo (5.494 pares de frases), e os hiperparâmetros ajustados de forma a ser possível o processamento em uma máquina com configurações feitas primariamente para o uso doméstico, vemos que o valor do resultado da avaliação automática encontra-se alinhado com os estudos previamente mencionados aqui (Lakew; Cettolo; Federico, 2018; Banitz, 2020; Tian *et al.*, 2022; Kimera; Rim; Choi, 2022).

Apesar de apresentar um *score* de aproximadamente 7,65, o modelo tende a render melhor desempenho em sua tarefa de acordo com aquilo para o qual foi treinado. Assim, deve-se também avaliar individualmente a pontuação de alguns de seus resultados de frases traduzidas, para que seja possível compreender o que esse valor atribuído ao modelo, apesar de baixo, representa quando da realização

¹⁰ Órgão, no Brasil, incumbido da adoção de providências para o adequado cumprimento das obrigações impostas pela Convenção de Haia de 1980 sobre os Aspectos Cíveis da Subtração Internacional de Crianças. (cf. Decreto nº 11.348, de 1º de janeiro de 2023; Decreto nº 3.413, de 14 de abril de 2000.)

de tarefas de tradução automática sobre o tema aqui exposto.

Para a análise dos resultados da avaliação automática, foram identificadas 20 frases em inglês (retiradas do par inglês-francês presente no *corpus*) sustentadas pelos dois modelos de tradução por meio do número de sua chave correspondente no *corpus*. Os *scores* sacreBLEU gerados para cada tradução em francês proveniente de cada modelo seguem discriminados em colunas paralelas às chaves de identificação das frases. Para a obtenção dos *scores*, cada tradução gerada, tanto pelo modelo treinado quanto pelo Google Tradutor, foi comparada com a frase-referência em francês presente no *corpus*. A investigação, então, passou a analisar os *scores* individuais dessas frases traduzidas a partir do *corpus* construído, conforme Tabela 2.

Tabela 2. Comparação dos *scores* sacreBLEU atribuídos às frases geradas pelo modelo treinado e pelo Google Tradutor.

Chave	Score Modelo Treinado	Score sacreBLEU Google Tradutor
316	32,5	32,6
384	29,5	58,1
635	71,9	50,7
796	22,8	23,2
852	8,3	6,3
950	20,6	12,6
965	21	63,9
1013	82,4	91,2
1166	7,3	7,1
1377	22	26,3
1390	33	7,8
1399	49,2	11,4
1411	49,8	36,1
1418	23,4	10
1437	5,7	14,1
1451	14,4	9,4
1455	29	24,6
1471	37,5	33,5
1486	41,4	34,8
1520	31,9	27,2
Média	31,7	29,0

Fonte: Elaboração própria.

A partir dos resultados obtidos pela avaliação automática, pode-se perceber que o modelo treinado ainda é capaz de obter *scores* maiores em relação ao modelo de tradução disponibilizado pela Google. O resultado da média dos *scores* obtidos dentre as traduções geradas pelo modelo treinado foi de aproximadamente 31,66. Em contrapartida, o resultado da média dos *scores* obtidos dentre as traduções geradas pelo Google Tradutor foi de 29,045. A diferença não é grande, mas dificilmente o seria, caso levássemos em consideração o tamanho do *corpus* paralelo utilizado para o treinamento do modelo de tradução automática.

Apesar de o foco desta pesquisa ser o treinamento de um modelo de tradutor automático baseado em *Transformer*, uma breve análise das traduções geradas demonstra-se relevante para a compreensão geral do *score* sacreBLEU do modelo treinado. Para a avaliação humana, foi imperativo que essa tenha sido realizada por um profissional que atua na área de tradução e com experiência em matéria de DIP, sendo capaz de transitar entre os dois campos do conhecimento, aplicando teoria e prática tradutória ao léxico especializado que requer a matéria de subtração internacional de crianças e adolescentes dentro do DIP.

Após a análise humana das traduções geradas baseada nos parâmetros já descritos na subseção 3.2, pode-se afirmar que, apenas duas traduções geradas apresentaram erros, sendo uma delas gerada pelo modelo treinado e a outra pelo Google Tradutor. Todas as outras não apresentam erros relevantes que influenciasssem na adequação e fluência da frase, podendo ser consideradas boas traduções. Em outras palavras, textos que aplicam de forma acertada termos jurídicos e jargões próprios da área relativa à matéria de subtração internacional de crianças e adolescentes.

Observando a Tabela 3, pode-se perceber que, a partir da frase em inglês, a tradução gerada pelo modelo Transformer treinado foi adequada e fluente. Houve, de fato, um erro (Vilar *et al.*, 2006) correspondente à ausência do artigo indefinido *une*, porém que não afeta a fluência e adequação do restante da frase (Banitz, 2020). Essa foi a única diferença que impediu que a frase traduzida ficasse idêntica à frase francesa oficial usada como referência (presente no *corpus* paralelo). A frase recebeu a pontuação de 32,5.

Tabela 3. Resultado das traduções do modelo treinado e do Google Tradutor.

Frase EN para tradução	FR Referência	Resultado Modelo	Tradução	Resultado Tradução Google
Convention of cooperation among authorities [Convenção de cooperação entre autoridades]	Une convention de coopération entre autorités [Uma convenção de cooperação entre autoridades]	Convention de coopération entre les autorités [Uma convenção de cooperação entre as autoridades]	Convention de coopération entre collectivités [Uma convenção de cooperação entre as autoridades territoriais]	

Fonte: Elaboração própria.

Por outro lado, a tradução do Google Tradutor gerou um resultado considerado fluente, porém inadequado para o contexto, uma vez que utiliza o termo “collectivités” para se referir aos órgãos federais de cooperação internacional relativa à Convenção Internacional de Haia de 1980 (HCCH, 1980) – as Autoridades Centrais. No entanto, o termo “collectivité” faz menção a departamentos administrativos reconhecidos na França, consideradas entidades territoriais coletivas, com poder de governo estabelecido. A terminologia engloba qualquer área que possua uma forma de governo local eleito e autoridade reguladora local e não faz referência à Autoridade Central, órgão incumbido da adoção de medidas para o cumprimento das obrigações impostas pela Convenção de Haia de 1980 (HCCH, 1980).

Ao utilizar a classificação de Vilar *et al.* (2006), este erro cometido pelo modelo de tradução do Google é claramente uma escolha lexical errada, afetando toda a adequação da tradução (Banitz, 2020). Ainda assim, essa tradução gerada pelo Google Tradutor recebeu uma pontuação de 32,6, infimamente maior que a tradução mais correta gerada pelo modelo treinado.

Em relação à segunda frase que apresentou um erro, ela foi gerada pelo modelo treinado com base em *corpus* especializado. Como demonstrado na Tabela 4, o modelo traduziu erroneamente a palavra “Requesting” para “requisites”, trocando totalmente o sentido da frase em inglês¹¹. Assim, apesar de mantida a fluência da frase, não há adequação do termo utilizado (Banitz, 2020) e a escolha por esse léxico distorce o sentido da frase no idioma de entrada, podendo ser caracterizado como um resultado proveniente de uma escolha equivocada de léxico (Vilar *et al.*, 2006). O *score* sacreBLEU obtido para essa tradução foi de 5,7.

Na análise da Tabela 4, em relação a tradução do Google dessa mesma frase, constata-se que ela foi igualmente fluente. Contudo, a sentença pode ser considerada mais adequada semanticamente ao contexto por ter utilizado corretamente o termo “requérantes” para se referir às Autoridades Centrais Requerentes (*Requesting*); termo também utilizado na frase-referência oficial em francês. O *score* acompanhou essa lógica correta e atribuiu, à tradução automática do Google, 14,1 pontos.

A partir do exame de todos esses resultados gerados, é possível perceber que um modelo treinado

¹¹ Importante aqui ressaltar que a Autoridade Central pode figurar tanto como Requerente (*Requérant/Requesting*) como Requerida (*Requise/Requested*). Esse título é importante para identificar onde o processo se inicia e qual o papel dessas Autoridades para a recuperação da criança ou adolescente.

Tabela 4. Resultado das traduções do modelo treinado e do Google Tradutor.

Frase EN para tradução	Frase FR Referência	Resultado Modelo	Tradução	Resultado Tradução Google
Requesting Central Authorities are often under pressure from applicants (usually left-behind parents) to provide daily reports of progress [Autoridades Centrais requerentes geralmente estão sob pressão dos demandantes (geralmente os genitores que foram deixados para trás) para fornecer relatórios diários de progresso]	Les demandeurs (généralement les parents privés de leur enfant) exercent souvent des pressions sur les Autorités centrales pour qu'elles leur fournissent des rapports de suivi journaliers [Os demandantes (geralmente os pais privados de seus filhos) geralmente fazem pressão nas Autoridades Centrais para que elas lhes forneça relatórios diários de monitoramento]	Les Autorités centrales requises sont souvent soumises à des pressions de la part des demandeurs (habituellement des parents laissés derrière eux) pour fournir des rapports quotidiens sur les progrès accomplis [As autoridades centrais requeridas estão frequentemente sob pressão dos demandantes (geralmente os pais deixados para trás) para fornecer relatórios diários de progresso]	Les Autorités centrales requérantes subissent souvent des pressions de la part des demandeurs (généralement des parents délaissés) pour qu'elles fournissent des rapports quotidiens sur les progrès [As Autoridades Centrais Requerentes são frequentemente pressionadas pelos demandantes (geralmente pais deixados para trás) a fornecer relatórios diários de progresso]	

Fonte: Elaboração própria.

com apenas 5.494 pares de frases consegue desempenhar tarefas de tradução automática baseada em vocábulo especializado com uma qualidade um ligeiramente melhor que os outros modelos.

No que tange às avaliações automática e manual, é possível depreender dos resultados que a disparidade entre os resultados obtidos pelo modelo treinado e pela ferramenta de tradução do Google não diferem muito entre si. A avaliação automática ressalta a proximidade de termos e estruturas utilizadas por ambas as traduções, ao passo que a avaliação manual ratifica a qualidade das mesmas ressaltando apenas um equívoco no uso de léxico especializado gerado por cada ferramenta de tradução automática.

5 Considerações finais

Este artigo teve como objetivo geral o treinamento e análise de modelos de tradução automática baseados em *Transformer*, cujos resultados demonstram que a análise foi realizada a contento. Em um de seus objetivos específicos, o trabalho buscou o teste da viabilidade do uso de modelos de tradução automática baseados em *Transformer*. A partir disso, foi possível perceber que a implementação do modelo foi facilitada pelo uso de código aberto e de bibliotecas disponibilizadas de forma gratuita pela plataforma *HuggingFace*, fatores que contribuíram significativamente para a etapa inicial de escrita do código utilizado.

Em etapas mais avançadas da implementação do modelo, encontrou-se dificuldades de treinamento do modelo devido aos limites computacionais de capacidade de processamento de dados pela máquina utilizada. Como ilustrado com o modelo SEBIS, modelos de tradução automática baseados em *Transformer* que contam com bons scores, apesar de melhor aproveitarem a capacidade de processamento da máquina, ainda necessitam de uma quantidade significativa de dados para que possam ser treinados e, consequentemente, de máquinas que possam lidar de forma eficiente com esses dados. Com menor capacidade de processamento, o modelo pré-treinado T5-small, os hiperparâmetros definidos e o tempo necessário para o treinamento se tornaram grandes limitadores para o desenvolvimento de um modelo de tradução automática que utiliza um léxico mais preciso e adequado em relação ao termos utilizados por essa área do DIP.

Em relação aos erros apontados na fase de avaliação de modelo, em específico na avaliação humana, constatou-se que ambos os modelos de tradução, seja ele treinado ou não, estão suscetíveis a falhas. Dessa forma, cabe melhorar o modelo *Transformer* treinado para que seus resultados sejam ainda melhores. A revisão do corpus, o aumento no número de pares de frases alinhadas, o investimento

em máquinas e técnicas de processamento de dados mais avançadas, a dedicação de maior tempo para o treinamento do modelo e o teste de mais alguns hiperparâmetros ajustáveis no momento de treinar o modelo são só alguns dos fatores concretos apontados por esse trabalho que garantem uma melhora no modelo e, conseqüentemente, nas traduções geradas.

Não obstante, este trabalho iniciou o desenvolvimento de um *corpus* paralelo inglês-francês específico relativo à Convenção da Haia de 1980 (HCCH, 1980), sendo desenvolvido um conjunto de dados eminentemente promissor que serve de referência para pesquisas futuras e aplicações em trabalhos que busquem desenvolver técnicas semelhantes. O trabalho de organização, limpeza e anotação dos textos selecionados aqui realizado e documentado demonstram o rigor metodológico empreendido que, por sua vez, fornece os alicerces necessários para a revisão e expansão do *corpus* relativo à convenção (HCCH, 1980).

O trabalho ainda serviu como forma de avaliação da medida na qual a construção de um *corpus* específico e a sua implementação em um modelo de tradução se torna mais satisfatória do que uma tradução feita em um modelo com léxico não especializado. É importante ressaltar o fator das dimensões envolvidas no trabalho, uma vez que quanto maior a quantidade de texto disponível para alimentar o treinamento modelo, maior a qualidade dos resultados esperada. Dessa forma, apesar de ter seu treinamento baseado em um *corpus* que conta com somente 5.494 pares de palavras, o modelo de tradução desenvolvido se demonstra extremamente eficiente e desempenha, na maioria das vezes, melhor do que um modelo de tradução não treinado.

Em diversas etapas, que partem desde a concepção e elaboração de um *corpus* paralelo em meio digital, e na anotação e programação de dados não-estruturados até a análise humana da tradução automática realiza, foi fundamental a formação prévia do primeiro autor em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação, uma área do saber que tornou possível o desenvolvimento do trabalho e da análise dos resultados obtidos (Pires, 2020).

Para que a pesquisa aqui desenvolvida continue sendo relevante e sirva de base para acadêmicos não só ligados ao bacharelado em LEA-MSI, como também profissionais e pesquisadores ligados a outras áreas, como linguística, estudos da tradução, e ciência da computação, por exemplo, propõe-se a ampliação deste estudo tanto em tarefas de pós-edição, quanto em avaliações conjuntas (*shared tasks*) (Freitas, 2022). Por meio de pesquisas encabeçadas por especialistas em linguística e estudos da tradução, temos como objetivo a constante melhora do modelo aqui desenvolvido (Freitas, 2022).

Por fim, considera-se que os resultados aqui relatados estejam alinhados com aqueles produzidos por pesquisas semelhantes, com real possibilidade de melhora a partir do desdobramento, ampliação e aprofundamento de investigação de fatores especificamente apontados ao longo do trabalho.

Referências

- BANITZ, Brita. Machine translation: a critical look at the performance of rule-based and statistical machine translation. *Cadernos de Tradução*, v. 40, n. 1, p. 54–71, jan. 2020. ISSN 2175-7968. DOI: 10.5007/2175-7968.2020v40n1p54. Disponível em: <https://periodicos.ufsc.br/index.php/traducao/article/view/2175-7968.2020v40n1p54>. Acesso em: 22 mar. 2024.
- COSTA, Cynthia Beatrice; SILVA, Igor A. Lourenço Da. On the Translation of Literature as a Human Activity par Excellence: Ethical Implications for Literary Machine Translation. *Aletria: Revista de Estudos de Literatura*, v. 30, n. 4, p. 225–248, dez. 2020. ISSN 2317-2096, 1679-3749. DOI: 10.35699/2317-2096.2020.22047. Disponível em: <https://periodicos.ufmg.br/index.php/aletria/article/view/22047>. Acesso em: 22 mar. 2024.
- DOUGLAS, Arnold; BALKAN, L.; LEE HUMPHREYS, R.; MEIJER, S.; SADLER, L. *Machine Translation: An Introductory Guide*. [S. l.]: NCC Blackwell, 1994. ISBN 9781855542464.
- FREITAG, Markus; GRANGIER, David; CASWELL, Isaac. BLEU might be Guilty but References are not Innocent. In: WEBBER, Bonnie; COHN, Trevor; HE, Yulan; LIU, Yang (ed.). *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, nov. 2020. p. 61–71. DOI: 10.18653/v1/2020.emnlp-main.5. Disponível em: <https://aclanthology.org/2020.emnlp-main.5>. Acesso em: 22 mar. 2024.

FREITAS, Cláudia. *Linguística Computacional*. São Paulo: Parábola Editorial, 2022.

HCCH. *Convention of 25 October 1980 on the Civil Aspects of International Child Abduction*. [S. l.: s. n.], 1980. Disponível em: <https://assets.hcch.net/docs/e86d9f72-dc8d-46f3-b3bf-e102911c8532.pdf>. Acesso em: 22 mar. 2024.

HCCH. *Guide to Good Practice under the Hague Convention of 25 October 1980 on the Civil Aspects of International Child Abduction: Part I – Central Authority Practice*. [S. l.: s. n.], 2003. Disponível em: <https://www.hcch.net/en/instruments/conventions/publications1/?dtid=3&cid=24>. Acesso em: 22 mar. 2024.

HCCH. *Guide to Good Practice under the Hague Convention of 25 October 1980 on the Civil Aspects of International Child Abduction: Part II – Implementing Measures*. [S. l.: s. n.], 2003. Disponível em: <https://www.hcch.net/en/instruments/conventions/publications1/?dtid=3&cid=24>. Acesso em: 12 jul. 2023.

HCCH. *Guide to Good Practice under the Hague Convention of 25 October 1980 on the Civil Aspects of International Child Abduction: Part III – Preventive Measures*. [S. l.: s. n.], 2005. Disponível em: <https://www.hcch.net/en/instruments/conventions/publications1/?dtid=3&cid=24>. Acesso em: 12 jul. 2023.

HCCH. *Guide to Good Practice under the Hague Convention of 25 October 1980 on the Civil Aspects of International Child Abduction: Part IV – Enforcement*. [S. l.: s. n.], 2010. Disponível em: <https://www.hcch.net/en/instruments/conventions/publications1/?dtid=3&cid=24>. Acesso em: 12 jul. 2023.

HCCH. *Guide to Good Practice under the Hague Convention of 25 October 1980 on the Civil Aspects of International Child Abduction: Part V – Mediation*. [S. l.: s. n.], 2012. Disponível em: <https://www.hcch.net/en/instruments/conventions/publications1/?dtid=3&cid=24>. Acesso em: 12 jul. 2023.

IOSIFOVA, Olena; IOSIFOV, Ievgen; ROLIK, Oleksandr; SOKOLOV, Volodymyr. Techniques Comparison for Natural Language Processing. *MoMLeT&DS*, v. 2631, n. 1, p. 57–67, 2020. Disponível em: <https://core.ac.uk/reader/328802590>. Acesso em: 22 mar. 2024.

KALYANI, Aditi; KUMUD, Hemant; SINGH, Shashi Pal; KUMAR, Ajai; DARBARI, Hemant. *Evaluation and Ranking of Machine Translated Output in Hindi Language using Precision and Recall Oriented Metrics*. [S. l.]: arXiv, abr. 2014. DOI: 10.48550/arXiv.1404.1847. Disponível em: <http://arxiv.org/abs/1404.1847>. Acesso em: 22 mar. 2024.

KIMERA, Richard; RIM, Daniela N.; CHOI, Heeyoul. Building a Parallel Corpus and Training Translation Models Between Luganda and English. *Journal of KIISE*, v. 49, n. 11, p. 1009–1016, nov. 2022. ISSN 2383-630X, 2383-6296. DOI: 10.5626/JOK.2022.49.11.1009. Disponível em: <http://arxiv.org/abs/2301.02773>. Acesso em: 22 mar. 2024.

KOGLIN, Arlene; MOURA, Willian Henrique Cândido; MATOS, Morgana Aparecida de; SILVEIRA, João Gabriel Pereira da. Quality assessment of machine-translated post-edited subtitles: an analysis of Brazilian translators' perceptions. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, v. 22, dez. 2023. ISSN 2295-5739. DOI: 10.52034/lans-tts.v22i.765. Disponível em: <https://lans-tts.uantwerpen.be/index.php/LANS-TTS/article/view/765>. Acesso em: 22 mar. 2024.

KOPONEN, Maarit. Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *The Journal of Specialised Translation*, p. 131–148, 2016. Disponível em: <https://www.phoenix3.ir/Article?id=34>. Acesso em: 22 mar. 2024.

LAKEW, Surafel Melaku; CETTOLO, Mauro; FEDERICO, Marcello. A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation. In: BENDER, Emily M.; DERCZYNSKI, Leon; ISABELLE, Pierre (ed.). *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, ago. 2018. p. 641–652. Disponível em: <https://aclanthology.org/C18-1054>. Acesso em: 22 mar. 2024.

O'HAGAN, Minako (ed.). *The Routledge Handbook of Translation and Technology*. 1. ed. Abingdon, Oxon ; New York, NY : Routledge, 2020. |: Routledge, ago. 2019. ISBN 9781315311258. DOI: 10.4324/9781315311258. Disponível em: <https://www.taylorfrancis.com/books/9781315311241>. Acesso em: 22 mar. 2024.

PAPINENI, Kishore; ROUKOS, Salim; WARD, Todd; ZHU, Wei-Jing. Bleu: a Method for Automatic Evaluation of Machine Translation. In: ISABELLE, Pierre; CHARNIAK, Eugene; LIN, Dekang (ed.). *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, jul. 2002. p. 311–318. DOI: 10.3115/1073083.1073135. Disponível em: <https://aclanthology.org/P02-1040>. Acesso em: 27 mar. 2024.

PEÑA AGUILAR, Argelia. Challenging machine translation engines: Some Spanish-English linguistic problems put to the test. *Cadernos de Tradução*, v. 43, n. 1, p. 1–26, dez. 2023. ISSN 2175-7968. DOI: 10.5007/2175-7968.2023.e85397. Disponível em: <https://periodicos.ufsc.br/index.php/traducao/article/view/85397>. Acesso em: 22 mar. 2024.

PÉREZ-VERA, Elisa. HCCH | Explanatory Report on the 1980 HCCH Child Abduction Convention. In: ACTS and Documents of the Fourteenth Session (1980), tome III, Child abduction. [S. l.: s. n.], 1980. Disponível em: <https://www.hcch.net/en/publications-and-studies/details4/?pid=2779>. Acesso em: 22 mar. 2024.

PIRES, Thiago Blanch. *Ampliando olhares sobre a tradução automática online : um estudo exploratório de categorias de erros de máquina de tradução gerados em documentos multimodais*. Jun. 2017. Tese de doutorado – Universidade de Brasília, Brasília. Disponível em: <http://repositorio2.unb.br/jspui/handle/10482/23727>. Acesso em: 22 mar. 2024.

PIRES, Thiago Blanch. A avaliação de tradução automática na atuação do bacharel em LEA-MSI. In: PEREIRA, Fernana Alencar (ed.). *Línguas Estrangeiras Aplicadas: trajetórias e possibilidades*. Campinas: Pontes Editores, 2020. p. 61–75.

POST, Matt. *A Call for Clarity in Reporting BLEU Scores*. [S. l.]: arXiv, set. 2018. DOI: 10.48550/arXiv.1804.08771. Disponível em: <http://arxiv.org/abs/1804.08771>. Acesso em: 22 mar. 2024.

RAFFEL, Colin; SHAZEER, Noam; ROBERTS, Adam; LEE, Katherine; NARANG, Sharan; MATENA, Michael; ZHOU, Yanqi; LI, Wei; LIU, Peter J. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. [S. l.]: arXiv, set. 2020. DOI: 10.48550/arXiv.1910.10683. Disponível em: <http://arxiv.org/abs/1910.10683>. Acesso em: 22 mar. 2024.

RODAS, João Grandino; MÔNACO, Gustavo Ferraz de Campos. *A Conferência da Haia de direito internacional privado: a participação do Brasil*. Brasília: Fundação Alexandre de Gusmão, 2007.

TIAN, Taoling; SONG, Chai; TING, Jin; HUANG, Hongyang. A French-to-English Machine Translation Model Using Transformer Network. *Procedia Computer Science*, v. 199, p. 1438–1443, jan. 2022. ISSN 1877-0509. DOI: 10.1016/j.procs.2022.01.182. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050922001831>. Acesso em: 22 mar. 2024.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N.; KAISER, Lukasz; POLOSUKHIN, Illia. *Attention Is All You Need*. [S. l.]: arXiv, ago. 2017. DOI: 10.48550/arXiv.1706.03762. Disponível em: <http://arxiv.org/abs/1706.03762>. Acesso em: 22 mar. 2024.

VILAR, David; XU, Jia; D'HARO, Luis Fernando; NEY, Hermann. Error Analysis of Statistical Machine Translation Output. In: CALZOLARI, Nicoletta; CHOUKRI, Khalid; GANGEMI, Aldo; MAEGAARD, Bente; MARIANI, Joseph; ODIJK, Jan; TAPIAS, Daniel (ed.). *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA), maio 2006. Disponível em: http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf. Acesso em: 22 mar. 2024.

WHITE, J. S. How to evaluate machine translation. In: SOMERS, Harold (ed.). *Computers and Translation: A translator's guide*. Amsterdam/Philadelphia: John Benjamins Publishing, 2003. v. 35. p. 211–244.

WOLF, Thomas; DEBUT, Lysandre; SANH, Victor; CHAUMOND, Julien; DELANGUE, Clement; MOI, Anthony; CISTAC, Pierrick; RAULT, Tim; LOUF, Rémi; FUNTOWICZ, Morgan; DAVISON, Joe; SHLEIFER, Sam; PLATEN, Patrick von; MA, Clara; JERNITE, Yacine; PLU, Julien; XU, Canwen; SCAO, Teven Le; GUGGER, Sylvain; DRAME, Mariama; LHOEST, Quentin; RUSH, Alexander M. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. [S. l.]: arXiv, jul. 2020. DOI: 10.48550/arXiv.1910.03771. Disponível em: <http://arxiv.org/abs/1910.03771>. Acesso em: 22 mar. 2024.

WU, Yonghui; SCHUSTER, Mike; CHEN, Zhifeng; LE, Quoc V.; NOROUZI, Mohammad; MACHEREY, Wolfgang; KRIKUN, Maxim; CAO, Yuan; GAO, Qin; MACHEREY, Klaus; KLINGNER, Jeff; SHAH, Apurva; JOHNSON, Melvin; LIU, Xiaobing; KAISER, Łukasz; GOUWS, Stephan; KATO, Yoshikiyo; KUDO, Taku; KAZAWA, Hideto; STEVENS, Keith; KURIAN, George; PATIL, Nishant; WANG, Wei; YOUNG, Cliff; SMITH, Jason; RIESA, Jason; RUDNICK, Alex; VINYALS, Oriol; CORRADO, Greg; HUGHES, Macduff; DEAN, Jeffrey. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. [S. l.]: arXiv, out. 2016. DOI: 10.48550/arXiv.1609.08144. Disponível em: <http://arxiv.org/abs/1609.08144>. Acesso em: 22 mar. 2024.

Contribuições dos autores

Clóvis Henrique Martins Pimentel: Conceituação, Curadoria de dados, Análise formal, Investigação, Metodologia, Administração de projetos, Recursos, Programas, Validação, Visualização, Escrita – rascunho original, Escrita – revisão e edição; **Thiago Blanch Pires**: Conceituação, Administração de projetos, Recursos, Visualização, Escrita – rascunho original, Escrita – revisão e edição.